

VIETNAM NATIONAL UNIVERSITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING



GRADUATION THESIS PROPOSAL

**USING MACHINE LEARNING METHODS
IN TRANSLATING SIGN LANGUAGE
INTO VIETNAMESE**

Council: Software Engineering

Instructor: Assoc. Prof. Quan Thanh Tho

—o0o—

Student: Võ Tuấn Khanh (1810220)

Nguyễn Trí Nhân (1810390)

Ho Chi Minh City, December 2021

Declaration Of Authenticity

TODO: Viết sơ sơ về việc nội dung báo cáo không phải là false, ăn cắp này kia nọ ví dụ:

Nhận diện hướng nhìn trong ảnh (Nhận diện vật thể trong ảnh) không phải là một đề tài mới nhưng vẫn là một thách thức bởi: trong các ứng dụng: việc nhận diện hướng nhìn của con người qua hình ảnh đòi hỏi kết quả chính xác cao, ở Việt Nam, hiện tại không thực sự có nhiều nghiên cứu chuyên sâu về đề tài. Trong quá trình nghiên cứu đề tài có rất nhiều kiến thức không nằm trong chương trình giảng dạy ở bậc Đại học tuy vậy chúng tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của tiến sĩ Nguyễn Đức Dũng. Nội dung nghiên cứu và các kết quả đều là trung thực và chưa từng được công bố trước đây. Các số liệu được sử dụng cho quá trình phân tích, nhận xét được chính tôi thu thập từ nhiều nguồn khác nhau và sẽ được ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, tôi cũng có sử dụng một số nhận xét, đánh giá và số liệu của các tác giả khác, cơ quan tổ chức khác. Tất cả đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào, tôi xin hoàn toàn chịu trách nhiệm về nội dung luận văn của mình. Trường đại học Bách Khoa thành phố Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện.

Acknowledgment

TODO: Viết sau cùng -> về việc cảm ơn này kia

ví dụ:

Để hoàn thành kì đề cương luận văn này, tôi tỏ lòng biết ơn sâu sắc đến tiến sĩ Nguyễn Đức Dũng đã hướng dẫn tận tình trong suốt quá trình nghiên cứu.

Chúng tôi chân thành cảm ơn quý thầy, cô trong khoa Khoa Học Và Kỹ Thuật Máy Tính, trường đại học Bách Khoa thành phố Hồ Chí Minh đã tận tình truyền đạt kiến thức trong những năm chúng tôi học tập ở trường. Với vốn kiến thức tích lũy được trong suốt quá trình học tập không chỉ là nền tảng cho quá trình nghiên cứu mà còn là hành trang để bước vào đời một cách tự tin.

Cuối cùng, tôi xin chúc quý thầy, cô dồi dào sức khỏe và thành công trong sự nghiệp cao quý.

Abstract

TODO: Viết sau cùng

ví dụ:

Nội dung chính của luận văn nhằm tìm hiểu, nghiên cứu xây dựng hệ thống nhận diện hướng nhìn thông qua ảnh chụp dựa trên những công trình, công nghệ mới được nghiên cứu và phát triển trong những năm gần đây của lĩnh vực Deep Learning. Trong quá trình nghiên cứu, tôi đã tiến hành tổng hợp, đánh giá ưu và nhược điểm của cách phương pháp, công nghệ đã và đang được nghiên cứu, sử dụng. Tiếp cận vấn đề theo nhiều hướng khác nhau, tôi thực hiện một số phương pháp sử dụng học sâu (CNN) để phát hiện hướng nhìn của con người qua hình ảnh. Bên cạnh việc hoàn thành nội dung của đề tài, nhóm chúng tôi đã nghiên cứu thêm một số phần để từ đó đặt nền móng cho các nghiên cứu sau này. Phần còn lại của luận văn tập trung vào việc đánh giá mô hình, kết quả đạt được, đồng thời phân tích ưu nhược điểm của mô hình thực hiện và thảo luận những vấn đề mà mô hình còn gặp phải. Cuối cùng, nhóm chúng tôi đề xuất hướng phát triển tiếp theo của đề tài trong tương lai.

Contents

| | | |
|----------|------------------------------------|----------|
| 1 | Introduction | 1 |
| 1.1 | Problem statement | 1 |
| 1.2 | Goals | 2 |
| 1.3 | Scopes | 2 |
| 1.4 | Thesis structure | 2 |
| 2 | Theoretical Background | 3 |
| 3 | Design and Solution | 4 |
| 3.1 | System Structure | 4 |
| 3.2 | Detail Implementation | 5 |
| 3.2.1 | Hand pattern recognition | 5 |
| 3.2.2 | Direction determination | 5 |
| 3.2.3 | Location detection | 6 |
| 3.2.4 | Word decoder | 7 |
| 3.2.5 | Text to speech | 7 |
| 4 | Summary | 8 |
| 4.1 | Thesis Status | 8 |
| 4.2 | Future Development | 8 |

List of Figures

| | | |
|---|---|---|
| 1 | Overview of the system modules | 4 |
| 2 | Structure of convolutional neural network | 5 |
| 3 | Word “You” (bạn) in sign language | 5 |
| 4 | Word “I” (tôi) in sign language | 6 |
| 5 | View from the camera module | 7 |

List of Tables

Chapter 1

Introduction

1.1 Problem statement

TODO: Recheck problem statement

TLDR: It is hard for the deaf and mute to communicate with normal people. And there is not many ways for them to express their thought.

“Each deaf person is a separate world, and they feel more self-deprecating and alone when they do not interact and share with others. They still have the desire to contribute to society”, said Mr. Do Hoang Thai Anh, Vice Chairman of the Hanoi Deaf Association.

Language is a universal key that not only connects people but also builds up our society. Any disability that affects the ability to communicate is a significant disadvantage, especially for people with disabilities. They cannot integrate, have fun, learn, and communicate like ordinary people because they cannot express their thoughts, ideas, and desires to develop society as we do. That burden usually makes them fall into poverty, live a dependent life, and be exploited, apart from society. Hence, it is challenging for them to have beautiful lives.

In 2020, Vietnam had more than 2.5 million people who are deaf and mute, yet, only a tiny portion of them took part in education, had the chance to be understood, and integrated with society.

According to UNICEF, “Households with members with disabilities are often poorer, children with disabilities are at risk of having less education than their peers, and employment opportunities for people with disabilities are also lower than those without disabilities. Even though people with disabilities are beneficiaries of the policy, and poverty is not a burden to accessing health facilities, very few people with disabilities (2.3%) have access to functional rehabilitation services when being sick or injured. Besides, there still exist inequalities in living standards and social participation for people with disabilities [6]. Many organizations are founded to support, help, and create better living conditions for people with disabilities to develop. However, this work still has many difficulties and inadequacies as there is no formal school or class. Moreover, there is no specific profession for this group of people, and the number of translators who know sign language is insufficient, while they take an essential role in helping the people with disabilities connect with society.

A quote from Cavett Robert, “Life is a grindstone, and whether it grinds you down or polishes you up is for you and you alone to decide.” However, it is challenging for these people to go to school and have an excellent education. They have their desires and dreams, but our resources and efforts are not enough to make them a polished grindstone. Furthermore, sign language shares the same property as any other spoken language; each different region and territory has a different way of expressing sign language. These unseen differences make com-

munication, self-expression, and information exchange even more complex and challenging for humanity.

In short, we must admit that understanding and breaking the language barrier is extremely necessary and urgent because the deaf and mute, like many other ordinary people, deserve to be assisted, understood, and acknowledged. Furthermore, we believe our system is the resolve to problems of the deaf and hard of hearing.

1.2 Goals

TODO: Write Goals

TLDR: It is crucial to find out a way that help we connect more easily, the deaf and mute can convey their thoughts much comfortably.

Mục tiêu của đề tài là nghiên cứu, hiểu và hiện thực một số phương pháp học sâu để phát hiện hướng nhìn của con người qua hình ảnh.

Một số vấn đề đặt ra:

- Làm thế nào để giải quyết bài toán trên?
- Cách tiếp cận như thế nào?
- Những công nghệ nào đã và hiện đang được sử dụng?
- Hướng cải tiến?...

Như vậy để thực hiện theo đúng mục tiêu của đề tài cần xác định một số công việc phải giải quyết như sau:

- Tìm kiếm và thu thập dữ liệu phù hợp với nội dung đề tài.
- Tìm hiểu các phương pháp tiếp cận đã được hiện thực
- Lựa chọn mô hình phù hợp
- Lên kế hoạch hiện thực, phát triển hệ thống nhận diện huấn luyện và kiểm thử.

1.3 Scopes

TODO: Write Scopes

TLDR: In this case study, we will build a system including an app and camera module to translate at least 100 words from sign language into Vietnamese.

1.4 Thesis structure

This proposal includes four sections and each will convey the related works and output when doing this thesis.

| Chapter | Content |
|---------|--|
| 1 | A brief introduction about plan and objectives of thesis |
| 2 | Introduction of theoretical background as foundation knowledge that are applied in the project |
| 3 | Solution and design approach for problem statement of project |
| 4 | Summary of the thesis status and future plan |

Chapter 2

Theoretical Background

Lorem ipsum

Chapter 3

Design and Solution

3.1 System Structure

Overall, the whole system includes three parts of hardware modules, naming camera module, user's smartphone, and the server.

Our sign language translating AI system includes six main modules: hand pattern recognition, direction determination, location detection, action detection, word decoder, and text to speech (Figure 4). Firstly, the system continuously captures the hand's motion, processes it with the hand landmark model, and then puts it into those modules. Each of them has a unique role, and after combining the first four modules' results (hand pattern, direction, location, and action detection), the word decoder module will take the output data and bring out the corresponding result. Then, the result will show up on the main screen (Figure 16); meanwhile, the phone will speak out that word. In the below sections, we will discuss each module's role and how it works.

TODO: Replace with new structure

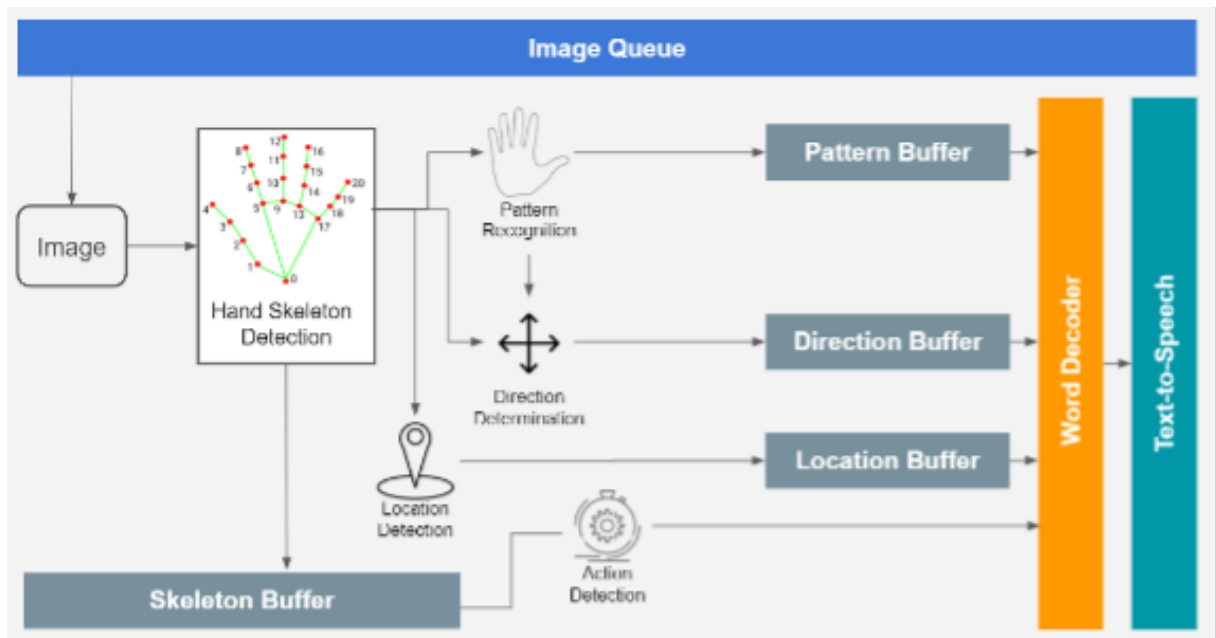


Figure 1: Overview of the system modules

3.2 Detail Implementation

3.2.1 Hand pattern recognition

Hand pattern recognition is the first and basic module of this system. While a person with disabilities does signs of sign language, his hands perform a series of different movements, where their hand may be spread out, clenched, or his fingers pointing out at something. Therefore, the role of this module is to recognize the pattern of the hands. Then combining the outcome with other modules, the system can give out the final result.

This module uses the output of the hand landmark model, which is a matrix size of 21. After calculating all the values in that matrix, we get a new matrix representing the distance between those 21 coordinates. Using the distance matrix as the input of CNN [2] with the designed structure (see Figure 2), as seen in Figure 1, will tell us the pattern of the hand at the moment it is captured.



Figure 2: Structure of convolutional neural network

3.2.2 Direction determination

The directions of the hand include four directions, i.e., right, left, up, down, front, and back. Each hand's pattern combined with different directions leads to a different meaning. For example, the pattern that points at someone means the word “you”; on the other hand, when we point it to ourselves, it means the word I (see Figure 3 and Figure 4).



Figure 3: Word “You” (bạn) in sign language



Figure 4: Word “I” (tôi) in sign language

To determine the hand’s direction, we use the hand landmark model provided in MediaPipe (see section 2 TK). The inception here is that we calculate the distance between the tip of the index finger and the wrist (called TK), then project it to the axis Ox , Oy , Oz , respectively. After that, we take each of those coordinates and compare them with the others. Finally, the one with the immense value will tell which axis the hand is on; besides, with the direction from the wrist to the tip of the index finger projected on that corresponding axis, we will know which direction the hand is.

For instance, a hand is known to be pointing toward the left direction. The value of the distance, when projected on the axis Ox , will be the biggest one among the three projected values. Then, calculate the vector drawn from the wrist to the tip of the index finger; we will know the direction of the hand itself.

3.2.3 Location detection

Locations of hand vary, is the hand put at forehead, mouth or the chest level, and so on. Every hand pattern that goes with every location will result in different words. Nevertheless, it is hard for the AI to know the hand’s location with only one camera, and its view is from above (see Figure 5). However, we came up with some solutions to this issue.

Firstly, we will take pictures of the hand and calculate the size of the hand in every frame in order to know whether that hand is getting bigger or smaller. Hence, if that hand is smaller than before, it means the hand is getting far away from the camera, and its location is somewhere at the chest level or the stomach level.



Figure 5: View from the camera module

Nonetheless, the above solution still has an issue: every man's hand has a different size, and the system does not know the correct position of the hand. Therefore, another solution is to use a wide-angle camera and set it away from the forehead. With this solution, the camera can have a much broader view. However, since we only have a normal-angle camera, we could not try out this solution and confirm its suitability.

3.2.4 Word decoder

TODO: Finish Word decoder

3.2.5 Text to speech

Besides displaying the translated sign language in text form, we included a text-to-speech module to know the result without looking into the screen. This module makes use of a free API provided by Google, named Text-to-Speech [4]. It converts arbitrary strings, words, and sentences into the sound of a person speaking the same things.

Chapter 4

Summary

4.1 Thesis Status

Lorem ipsum

4.2 Future Development

Lorem ipsum