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 kì 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 cám 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

List of Figures

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

Related Work

CheckList: [X] Sơ lược ý chính [X] Điều chỉnh [.] Translate [ ] Complete

Nowadays, research works related to the problem of converting sign language into text have been proposed by many researchers from all over the world, from many different approaches and perspectives. In which, two main approaches can be mentioned as follows: - Glove based approaches: With this approach, it requires deaf and mute people to wearing a sensor glove. When user has any different action or gesture, these sensor will be recorded. After that, data from sensor will analyze by analyzer component and return the output for user. - Vision based approaches: With this approach, image processing algorithms will be applied to be able to determine hand position, gestures and movements of the hand. The user will not have to wear necessary equipment like glove based approaches, which is convenient for user. However, with using library or algorithms of image processing, we need to deal with worst quality output, which is greatly affected by this algorithms. With both approaches above, there is has some problems, that is, they can only recognize a very small number of words. These words are mostly words with different hand shapes that can be classified like that. However, in sign language, there will be many words that use the same hand shape but will differ in many characteristics, such as position and orientation. To our knowledge, there is currently no model that can handle the conversion of sign language flexibly and conveniently for the deaf-mute, helping them to communicate effectively. natural to the common man. Therefore, by applying appropriate technologies, the authors carry out this graduation thesis with the goal of breaking down the barriers between deaf-mute people and normal people, helping them to become self-sufficient. more confident in daily communication.

Ngày nay, các công trình nghiên cứu liên quan đến vấn đề chuyển đổi ngôn ngữ ký hiệu thành văn bản đã được nhiều nhà nghiên cứu từ khắp nơi trên thế giới đề xuất , theo nhiều hướng tiếp cận và góc nhìn khác nhau. Trong đó có thể kể đến 2 hướng tiếp cận chính như sau: - Hướng tiếp cận sử dụng găng tay cảm biến: Đây là hướng tiếp cận mà người sử dụng sẽ đeo 1 chiếc găng tay được trang bị các cảm biến chuyển động chuyên dùng. Khi người sử dụng có các hành động hay cử chỉ khác nhau sẽ được các cảm biến này ghi nhận, sau đó qua một bộ phân tích và sẽ trả về kết quả cho người dùng -Hướng tiếp cận sử dụng xử lý hình ảnh: Trong hướng tiếp cận này, các thuật toán về xử lý hình ảnh sẽ được áp dụng để có thể xác định được vị trí bàn tay, các cử chỉ, chuyển động của bản tay như thế nào . Người sử dụng sẽ không phải mang các trang bị cần thiết như hướng tiếp cận sử dụng găng tay, thuận tiện cho người sử dụng. Tuy nhiên, độ hiệu quả của các thuật toán xử lý ảnh hướng rất nhiều đến chất lượng đầu ra.

Với cả hai cách tiếp cận trên đều có một đặc điểm chung, đó là đều chỉ có thể nhận diện được một số lượng rất ít từ vựng. Các từ vựng này hầu hết là các từ có sự khác nhau về hình dạng bàn tay thì mới có thể phân loại được như thế. Tuy nhiên, trong ngôn ngữ ký hiệu , sẽ có rất nhiều

từ sử dụng chung một hình dạng bàn tay nhưng sẽ khác nhau về nhiều đặc điểm , ví dụ như vị trí và hướng. Theo hiểu biết của chúng em thì hiện nay vẫn chưa có một mô hình nào có thể xử lý được việc chuyển đổi ngôn ngữ ký hiệu một cách linh hoạt và thuận tiện cho người câm-điếc, giúp họ có thể giao tiếp được một cách tự nhiên với người bình thường. Chính vì thế, bằng cách vận dụng những công nghệ phù hợp, nhóm tác giả tiến hành thực hiện đề tài luận văn tốt nghiệp này hướng đến mục tiêu phá bỏ các rào cản giữa người câm-điếc và người bình thường, giúp họ tự tin hơn trong việc giao tiếp hằng ngày.

In this category requires signers to wear a sensor glove or a colored glove. The task will be simplified during segmentation process by wearing glove. The drawback of this approach is that the signer has to wear the sensor hardware along with the glove during the operation of the system.

+ Hướng tiếp cận sử dụng xử lý hình ảnh

Chapter 3

Theoretical Background

Lorem ipsum

CheckList: [...] Mô hình Convolution Neural Network - CNN [] Media Pipe -> Lấy từ eureka [...] Distance Matrix [...] BeamSearch with CTC decode

3.1 Convolution Neural Network - CNN

Convolution Neural Networks are a special class of Neural Networks. They are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. CNN mainly consist of Convolution Layers, Pooling Layers, Activation Layers and Fully Connected Layers. ConvNet architectures make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture. These then make the forward function more efficient to implement and vastly reduce the amount of parameters in the network. Some of the main uses of CNN can be mentioned as: image classification, object detection, semantic segmentation, face recognition, etc.

Insert picture about CNN https://scholarworks.iupui.edu/bitstream/handle/1805/24768/FINAL

The figure above shows an example of convolution neural network, which is taking an image as input and then extracting features from it through various layers and then finally predicting the class of the object in the given image

3.1.1 Architecture

Convolutional Neural Networks have a different architecture than regular Neural Networks. Regular Neural Networks transform an input by putting it through a series of hidden layers. Every layer is made up of a set of neurons, where each layer is fully connected to all neurons in the layer before. Finally, there is a last fully-connected layer — the output layer — that represent the predictions. Convolutional Neural Networks are a bit different. First of all, the layers are organised in 3 dimensions: width, height and depth. Further, the neurons in one layer do not connect to all the neurons in the next layer but only to a small region of it. Lastly, the final output will be reduced to a single vector of probability scores, organized along the depth dimension. Insert image of diff architecture : https://www.freecodecamp.org/news/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050/ CNN can be divided into two parts: + The hidden layers/ Feature extraction part In this part, the network will perform a series of convolutions and pooling operations during which the features are detected. If you had a picture of a zebra, this is the part where the network would recognise its stripes, two ears, and four legs.

+ The Classification part Here, the fully connected layers will serve as a classifier on top of these extracted features. They will assign a probability for the object on the image being what the algorithm predicts it is. Insert image of CNN arc: https://www.freecodecamp.org/news/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050/

3.1.2 Feature extraction part

Convolutional Layer

Convolution Layer is the core building block of a Convolutional Network that does most the computational heavy lifting. A convolution is executed by sliding the filter over the input. At every location, a matrix multiplication is performed and sums the result onto the feature map. This process of extracting features from image happens throughout the CNN's convolutional layers. This process is illustrated in Fig...

https://scholarworks.iupui.edu/bitstream/handle/1805/24768/FINAL

When the feature map is made, we can pass each value in the feature map through a non-linearity function, such as ReLU, sigmoid, etc. Before it becomes the input of the next convolution layer.

Because the size of the feature map is always smaller than the input, we have to do something to prevent our feature map from shrinking. This is where we use padding. A layer of zero-value pixels is added to surround the input with zeros, so that our feature map will not shrink. In addition to keeping the spatial size constant after performing convolution, padding also improves performance and makes sure the kernel and stride size will fit in the input. => Need picture

Pooling Layers

After a convolution layer, it is common to add a pooling layer in between CNN layers. The function of pooling is to continuously reduce the dimensionality to reduce the number of parameters and computation in the network. This shortens the training time and controls overfitting.

There are mainly two types of Pooling Layers in a CNN: Max Pooling and Average Pooling. The functionality of these two types of layers are demonstrated in Figure ... .Max Pooling restores the maximum value from the segment of the picture covered by the Kernel. Whereas, Average Pooling restores the average of the multitude of values from the bit of the picture covered by the Kernel. => Insert Picture

Activation Layers

Neural networks in general and CNNs in particular rely on a non-linear "trigger" function to signal distinct identification of likely features on each hidden layer. CNNs may use a variety of specific functions, such as rectified linear units (ReLUs) and continuous trigger (non-linear) functions—to efficiently implement this non-linear triggering. Insert picture of some function like ReLU, tanh ....

3.1.3 Classification part

Fully connected layers

The last layers of a CNN are fully connected layers. Neurons in a fully connected layer have full connections to all the activations in the previous layer. This part is in principle the same as

a regular Neural Network.

Figure ... illustrate the way of input value stream into the fully connected layer. Because these fully connected layer can only accept 1 Dimensional data. So, we need convert our 3D data to 1D data. After pass through some FCL, we will get the result is the data classification.

Insert picture ...

3.2 Media Pipe

Lấy từ eureka bỏ vào

3.3 Distance Matrix

A distance matrix is a table that shows the distance between pairs of objects. For example, in the figure ..., we can see the distance of A and B is 16, B and C is 37 and so on. In the diagonal of table is the distance of object from itself, so the value as we can see is 0. Distance matrices are sometimes called dissimilarity matrices.

Insert picture of distance matrix

3.3.1 Create Distance Matrix

A distance matrix is computed from a raw data table (Figure ...).

In the example below, we can use high school math (Pythagoras) to work out that distance between A and B. Chèn công thức vào đây

We can use same formula with more than two variables, and this is known as the Euclidean distance.

In result, we have the distance matrix represented like Figure ... Chèn bảng kết quả vào

3.4 Beam search with CTC decoder

CheckList: [] BeamSearch [] CTC recap [] Combination [] Pseudo code

3.4.1 Beam Search

In computer science, beam search is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set. Beam search is an optimization of best-first search the reduces its memory requirements. Best-first search is a graph search which orders all partial solutions (states) according to some heuristic. But in beam search, only a predetermined number of best partial solutions are kept as candidates.

Để có thể dễ hình dung, chúng tôi sẽ trình bày mã giả của giải thuật beam-search như hình dưới đây. Ta có thể thấy ....

Insert beam search picture pseudocode

3.4.2 CTC decoder

Copy CTC decode recap

3.4.3 Beam Search with CTC decoder

....

Chapter 4

Design and Solution

4.1 System Structure

Lorem ipsum

4.2 Detail Implementation

Lorem ipsum

Chapter 5

Summary

5.1 Thesis Status

Lorem ipsum

5.2 Future Development

Lorem ipsum