HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Trần Quang Hưng

# NGHIÊN CỨU, THIẾT KẾ VÀ THỬ NGHIỆM HỆ THỐNG CHUYỂN VĂN BẢN TIẾNG VIỆT THÀNH NGÔN NGỮ KÝ HIỆU SỬ DỤNG MÔ HÌNH 3D HỖ TRỢ NGƯỜI KHIẾM THÍNH

## ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

**(Theo định hướng ứng dụng)**

HÀ NỘI – NĂM 2025

Escaneado con CamScanner

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

---

**Trần Quang Hưng**

# NGHIÊN CỨU, THIẾT KẾ VÀ THỬ NGHIỆM HỆ THỐNG CHUYỂN VĂN BẢN TIẾNG VIỆT THÀNH NGÔN NGỮ KÝ HIỆU SỬ DỤNG MÔ HÌNH 3D HỖ TRỢ NGƯỜI KHIẾM THÍNH

CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN
MÃ SỐ: 8.48.01.04 (Hệ thống thông tin)

**ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT**

**(Theo định hướng ứng dụng)**

**NGƯỜI HƯỚNG DẪN KHOA HỌC**

**TS. VŨ HOÀI NAM**

HÀ NỘI – NĂM 2025

Escaneado con CamScanner

# LỜI CAM ĐOAN

Tôi xin cam đoan rằng toàn bộ nội dung trình bày trong đề án này là kết quả của quá trình nghiên cứu, tổng hợp và biên soạn từ các nguồn tài liệu có kiểm chứng, dưới sự hướng dẫn của giảng viên chuyên môn.

Tôi hoàn toàn chịu trách nhiệm về tính trung thực và chính xác của các thông tin được trình bày trong tài liệu này.

**Tác giả đề án**

Trần Quang Hưng

Escaneado con CamScanner

# LỜI CẢM ƠN

Trước hết, em xin gửi lời cảm ơn chân thành và sâu sắc đến thầy Vũ Hoài Nam đã tận tình hướng dẫn, góp ý và đồng hành trong suốt quá trình thực hiện đề án này. Những chỉ dẫn quý báu từ thầy/cô là nền tảng quan trọng giúp em hoàn thiện nội dung nghiên cứu một cách nghiêm túc và có hệ thống.

Em cũng xin trân trọng cảm ơn các thầy cô giảng viên khoa Công nghệ thông tin 1 và Khoa Đào tạo Sau Đại học đã hỗ trợ chuyên môn, cung cấp tài liệu, cũng như tạo điều kiện thuận lợi để em triển khai và hoàn thành đề án đúng tiến độ.

Cuối cùng, em xin cảm ơn tất cả những cá nhân và tập thể đã góp phần hỗ trợ về tinh thần cũng như chuyên môn trong suốt quá trình thực hiện.

Escaneado con CamScanner

# MỤC LỤC

LỜI CAM ĐOAN ..... i

LỜI CẢM ƠN ..... ii

DANH MỤC CÁC KÍ HIỆU VÀ CHỮ VIẾT TẮT ..... vi

DANH MỤC HÌNH ẢNH ..... vii

DANH MỤC BẢNG ..... viii

MỞ ĐẦU ..... 1

CHƯƠNG 1: TỔNG QUAN VỀ HỆ THỐNG CHUYỂN VĂN BẢN THÀNH NGÔN NGỮ KÝ HIỆU ..... 5

1.1 Khái quát về ngôn ngữ ký hiệu và nhu cầu giao tiếp của người khiếm thính ..... 5

1.1.1 Định nghĩa và bản chất của ngôn ngữ ký hiệu ..... 5

1.1.2 Các đặc trưng ngôn ngữ học quan trọng ..... 6

1.1.3 Đa dạng và biến thể ngôn ngữ ký hiệu ..... 6

1.1.4 Đặc điểm giao tiếp của người khiếm thính và nhu cầu thiết yếu ..... 7

1.1.5 Vấn đề thực tế trong giao tiếp và tiếp cận thông tin ..... 8

1.1.6 Hệ quả cho thiết kế giải pháp công nghệ ..... 9

1.1.7 Tác động giáo dục và xã hội ..... 10

1.2 Các tiếp cận công nghệ trong chuyển đổi văn bản sang ngôn ngữ ký hiệu ..... 11

1.2.1 Tiền xử lý ngôn ngữ và ánh xạ ngôn ngữ ..... 12

1.2.2 Phương pháp sinh chuyển động ký hiệu ..... 13

1.2.3 Thách thức và giới hạn kỹ thuật ..... 14

1.2.4 Dữ liệu, công cụ và nguồn lực ..... 16

1.2.5 Xu hướng phát triển công nghệ ..... 18

Escaneado con CamScanner

1.3 Ứng dụng mô hình 3D và hệ thống trực quan hóa trong truyền tải ký hiệu ..... 19
1.3.1 Ứng dụng thực tiễn và bối cảnh triển khai ..... 20
1.3.2 Lợi ích của mô hình 3D và ý nghĩa trực quan ..... 22
1.3.3 Thành phần kỹ thuật và chức năng hệ thống trực quan hóa ..... 23
1.3.4 Hạn chế và thách thức khi áp dụng 3D cho ký hiệu ..... 24
CHƯƠNG 2: XÂY DỰNG HỆ THỐNG CHUYỂN VĂN BẢN SANG MÔ HÌNH 3D ..... 26
2.1. Xác định yêu cầu xử lý và hướng tiếp cận ..... 26
2.1.1 Yêu cầu chức năng chính ..... 26
2.1.2 Yêu cầu dữ liệu và annotation ..... 27
2.1.3 Yêu cầu xử lý và luồng pipeline hệ thống ..... 28
2.1.4 Yêu cầu phi chức năng ..... 29
2.1.5 Tổng quát, tính khả thi về hướng tiếp cận đề xuất ..... 30
2.1.6 Các rủi ro chính và biện pháp giảm thiểu ..... 31
2.2. Quy trình xử lý: từ văn bản đầu vào đến mô hình 3D ký hiệu ..... 32
2.2.1 Tiếp nhận và phân tích văn bản đầu vào ..... 32
2.2.2 Ánh xạ từ văn bản sang ký hiệu ..... 33
2.2.3 Chuyển ký hiệu sang mô hình khung xương ..... 34
2.2.4 Xây dựng mô hình 3D ký hiệu ..... 36
2.2.5 Kết xuất và lưu trữ video ký hiệu 3D ..... 38
2.2.6 Tích hợp vào giao diện hiển thị ..... 40
2.3. Thiết kế kiến trúc hệ thống và các thành phần chức năng chính ..... 40
2.3.1. Kiến trúc tổng thể ..... 40

Escaneado con CamScanner

2.3.2. Các thành phần chức năng chính ..... 42
2.3.3. Luồng xử lý trong kiến trúc ..... 44
2.4. Thiết kế giao diện người dùng và kết nối với hệ thống mô phỏng ..... 45
2.4.1. Nguyên tắc thiết kế ..... 45
2.4.2. Các thành phần chính của giao diện ..... 46
2.4.3. Kết nối với hệ thống mô phỏng ..... 47
2.4.4 Lợi ích và ý nghĩa ..... 48
CHƯƠNG 3: THỬ NGHIỆM MÔ HÌNH CHUYỂN ĐỔI VĂN BẢN SANG KÝ HIỆU 3D ..... 50
3.1. Quy trình xử lý đầu vào và xây dựng mô phỏng chuyển động 3D ..... 50
3.1.1. Các bước xử lý chính ..... 50
3.1.2. Ý nghĩa của quy trình ..... 51
3.2. Thử nghiệm chuyển đổi văn bản sang trình tự ký hiệu 3D ..... 52
3.2.1 Chuẩn bị thử nghiệm ..... 53
3.2.2 Quy trình thực hiện ..... 53
3.2.3 Kết quả thử nghiệm và quan sát thực tế ..... 54
3.3. Đánh giá kết quả chuyển đổi và khả năng biểu đạt của mô hình ..... 59
3.3.1 Kết quả quan sát ..... 60
3.3.2 Các hạn chế còn tồn tại ..... 61
3.3.3 Đánh giá và định hướng cải tiến ..... 63
KẾT LUẬN ..... 65
DANH MỤC CÁC TÀI LIỆU THAM KHẢO ..... 66

Escaneado con CamScanner

# DANH MỤC CÁC KÍ HIỆU VÀ CHỮ VIẾT TẮT

| Từ viết tắt | Tiếng Anh | Tiếng Việt |
|---|---|---|
| NLP | Natural Language Processing | Xử lý ngôn ngữ tự nhiên |
| VSL | Vietnamese Sign Language | Ngôn ngữ ký hiệu tiếng Việt |
| ASL | American Sign Language | Ngôn ngữ ký hiệu người Mỹ |
| BSL | British Sign Language | Ngôn ngữ ký hiệu người Anh |
| POS tagging | Parts of Speech (PoS) tagging | Gắn thẻ các loại lời nói |
| RNN | Recurrent Neural Networks | Mạng nơ-ron hồi quy |
| LSTM | Long short-term memory | Bộ nhớ dài hạn ngắn hạn |
| GAN | Generative Adversarial Networks | Mạng đối nghịch sinh sản |
| VAE | Variational Autoencoders | Bộ mã hóa tự động biến thiên |
| RMSE | Root Mean Squared Error | Sai số trung bình bình phương |
| DTW | Dynamic Time Warping | Sự cong vênh thời gian động |
| SUS | System Usability Scale | Thang đo khả năng sử dụng hệ thống |
| FID | Fréchet Inception Distance | Khoảng cách khởi đầu của Fréchet |
| UI/UX | User Interface/User Experience | Giao diện người dùng/Trải nghiệm người dùng |

Escaneado con CamScanner

# DANH MỤC HÌNH ẢNH

Hình 1.1 Ảnh minh họa động tác ký hiệu từ "mỳ tôm" ..... 5
Hình 1.2 Bản đồ thế giới với các loại ngôn ngữ ký hiệu ..... 7
Hình 1.3 Lớp học dạy ngôn ngữ ký hiệu ..... 10
Hình 1.4 Hình ảnh minh họa tokenization ..... 12
Hình 1.5 Công cụ trích xuất khung cử chỉ, động tác OpenPose ..... 16
Hình 1.6 Mô hình SMPLifyX sẽ xây dựng mô hình 3D ..... 17
Hình 2.1 Hình ảnh dataset từ nguồn QIPEDC ..... 27
Hình 2.2 Pipeline tóm tắt các bước quy trình xử lý ..... 32
Hình 2.3 Mô hình quy trình tiếp nhận và tiền xử lý văn bản ..... 32
Hình 2.4 Biểu đồ ánh xạ từ với dữ liệu từ ký hiệu ..... 33
Hình 2.5 Minh họa lấy mô hình khung xương từ dữ liệu ..... 34
Hình 2.6 Mô hình hoạt động của OpenPose ..... 35
Hình 2.7 Minh họa kết quả chuyển đổi video 2D sang mô hình 3D ..... 37
Hình 2.8 Chuỗi chuyển dạng từ dữ liệu xương 2D sang mô hình 3D của SMPLify-X 37
Hình 2.9 Mô hình trích xuất mô hình 3D & lưu trữ video ký hiệu 3D ..... 39
Hình 2.10 Mô hình kiến trúc tổng thể của hệ thống ..... 41
Hình 2.11 Sơ đồ các thành phần chức năng chính ..... 43
Hình 2.12 Kiến trúc luồng xử lý của hệ thống ..... 44
Hình 2.13 Thiết kế giao diện chính của hệ thống ..... 46
Hình 3.1 Kết quả render mô hình 3D ..... 55
Hình 3.2 Kết quả chế độ xem video mô hình 3D ..... 56
Hình 3.3 Một số ví dụ về hạn chế kết quả huấn luyện ..... 61

Escaneado con CamScanner

# DANH MỤC BẢNG

Bảng 3.1 Số liệu từ quá trình huấn luyện dữ liệu dataset ..... 57
Bảng 3.2 Thống kê nguyên nhân lỗi hỏng dữ liệu huấn luyện ..... 57
Bảng 3.3 So sánh các nguyên nhân lỗi trong SMPLify-X ..... 58
Bảng 3.4 Kết quả thống kê so sánh độ rung với huấn luyện FPS khác nhau ..... 59

Escaneado con CamScanner

# MỞ ĐẦU

**Lý do chọn đề tài:**

Trong xã hội hiện đại, nhu cầu tiếp cận thông tin một cách công bằng và toàn diện cho mọi đối tượng, đặc biệt là người khuyết tật, ngày càng được chú trọng. Trong đó, cộng đồng người khiếm thính – những người gặp khó khăn trong việc tiếp nhận thông tin bằng ngôn ngữ nói – luôn phải đối mặt với nhiều rào cản trong học tập, làm việc và sinh hoạt hàng ngày. Ngôn ngữ ký hiệu là phương tiện giao tiếp chính của họ, tuy nhiên, phần lớn các thông tin trên internet và trong hệ thống giáo dục, hành chính hiện nay vẫn chủ yếu ở dạng văn bản hoặc âm thanh, gây ra sự thiếu đồng đều trong tiếp cận tri thức.

Trong bối cảnh công nghệ thông tin và trí tuệ nhân tạo phát triển mạnh mẽ, việc xây dựng một hệ thống có khả năng chuyển văn bản tiếng Việt thành ngôn ngữ ký hiệu là một hướng tiếp cận hiệu quả và thiết thực nhằm hỗ trợ người khiếm thính hòa nhập sâu rộng hơn vào đời sống xã hội. Đặc biệt, việc kết hợp công nghệ mô hình 3D trong việc biểu đạt cử chỉ ngôn ngữ ký hiệu không chỉ giúp truyền đạt thông tin rõ ràng hơn mà còn tạo ra một trải nghiệm sinh động, trực quan, dễ tiếp nhận cho người sử dụng.

Tuy nhiên, hiện nay tại Việt Nam, các hệ thống chuyển ngữ ngôn ngữ ký hiệu vẫn còn rất hạn chế về số lượng, độ chính xác cũng như mức độ tương tác [1]. Việc nghiên cứu, thiết kế và thử nghiệm một hệ thống chuyển đổi văn bản tiếng Việt thành ngôn ngữ ký hiệu sử dụng mô hình 3D không chỉ mang ý nghĩa thực tiễn trong việc hỗ trợ người khiếm thính, mà còn có giá trị khoa học trong lĩnh vực xử lý ngôn ngữ tự nhiên, mô hình hóa chuyển động và ứng dụng trí tuệ nhân tạo. Đây chính là lý do mang tính cấp thiết và nhân văn khiến đề tài được lựa chọn để nghiên cứu, phát triển và ứng dụng

Escaneado con CamScanner

**Tổng quan về vấn đề nghiên cứu:**

Việc nghiên cứu chuyển đổi văn bản tiếng Việt thành ngôn ngữ ký hiệu có ý nghĩa quan trọng trong việc hỗ trợ người khiếm thính tiếp cận thông tin và giao tiếp xã hội thuận tiện hơn. Ngôn ngữ ký hiệu là một hệ thống giao tiếp độc lập dựa trên cử chỉ, biểu cảm và chuyển động cơ thể, đòi hỏi sự kết hợp giữa ngôn ngữ học, xử lý ngôn ngữ tự nhiên (NLP), thị giác máy tính và mô hình hóa 3D. Trên thế giới đã có nhiều hệ thống tiên tiến như SignAll hay Virtual Sign ứng dụng AI và công nghệ 3D cho việc chuyển đổi văn bản sang ngôn ngữ ký hiệu, tuy nhiên ở Việt Nam, các nghiên cứu trong lĩnh vực này còn hạn chế, chủ yếu dừng lại ở việc xây dựng cơ sở dữ liệu hoặc giảng dạy ký hiệu cơ bản. Những thách thức lớn hiện nay bao gồm: thiếu dữ liệu ký hiệu tiếng Việt chuẩn hóa, thiếu mô hình ngôn ngữ phù hợp với ngữ pháp ký hiệu, và sự hạn chế trong việc thể hiện biểu cảm, cử động chân thực trong mô hình 3D. Đề tài này nhằm xây dựng hệ thống chuyển đổi văn bản tiếng Việt thành chuỗi ký hiệu có nghĩa, biểu diễn bằng mô hình 3D sinh động, tự nhiên, có khả năng tích hợp vào các nền tảng hỗ trợ học tập và giao tiếp, từ đó góp phần giải quyết những hạn chế còn tồn tại trong nghiên cứu hiện nay

**Mục đích nghiên cứu:**

Mục đích chính của đề tài là nghiên cứu, thiết kế và thử nghiệm hệ thống chuyển đổi văn bản tiếng Việt thành ngôn ngữ ký hiệu sử dụng mô hình 3D, nhằm hỗ trợ người khiếm thính tiếp cận thông tin một cách trực quan, hiệu quả. Hệ thống cho phép tự động chuyển văn bản thành chuỗi ký hiệu, biểu diễn bằng mô hình hoạt họa 3D sinh động, dễ hiểu và phù hợp ngữ cảnh.

Escaneado con CamScanner

Về lý luận, đề tài góp phần xây dựng cơ sở khoa học cho việc ánh xạ giữa tiếng Việt và ngôn ngữ ký hiệu, bao gồm phân tích ngữ nghĩa, cú pháp và ngữ cảnh. Đồng thời, đề tài đóng góp cho lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt, đặc biệt trong chuyển đổi ngôn ngữ sang hình ảnh động.

Về thực tiễn, đề tài hướng đến phát triển một hệ thống phần mềm có thể tích hợp vào các nền tảng học tập và giao tiếp cho người khiếm thính. Hệ thống chạy trên nền web hoặc máy tính, cho phép nhập văn bản và hiển thị ký hiệu bằng mô hình 3D. Kết quả mong muốn là thu hẹp khoảng cách tiếp cận thông tin và mở rộng ứng dụng trong giáo dục, truyền thông và dịch vụ công.

**Đối tượng và phạm vi nghiên cứu:**

Đối tượng nghiên cứu của đề tài là hệ thống chuyển đổi văn bản tiếng Việt thành ngôn ngữ ký hiệu, tập trung vào ba thành phần chính: văn bản đầu vào, cơ chế chuyển đổi sang chuỗi ký hiệu, và mô hình hóa hoạt ảnh 3D thể hiện các cử chỉ ký hiệu. Người khiếm thính – nhóm hưởng lợi chính – cũng là yếu tố quan trọng trong việc đánh giá tính hiệu quả và ứng dụng của hệ thống.

Phạm vi nghiên cứu về không gian tập trung tại Việt Nam, sử dụng tiếng Việt và ngôn ngữ ký hiệu Việt Nam (VSL), hướng đến ứng dụng trong trường học, trung tâm hỗ trợ người khuyết tật và các nền tảng học trực tuyến trong nước.

Về thời gian, nghiên cứu được thực hiện trong khuôn khổ một đề tài luận văn, chủ yếu phát triển nguyên mẫu thử nghiệm với tập dữ liệu nhỏ, số lượng văn bản và ký hiệu giới hạn. Phạm vi sẽ được mở rộng trong các nghiên cứu tiếp theo

**Phương pháp nghiên cứu:**

Escaneado con CamScanner

Để đạt được mục tiêu, đề tài tập trung vào việc nghiên cứu và xây dựng quy trình chuyển đổi văn bản tiếng Việt thành mô hình 3D ký hiệu, phục vụ các nhu cầu và ngữ cảnh như truyền tải thông tin cho người khiếm thính và hỗ trợ học tập, tra cứu ngôn ngữ ký hiệu. Phương pháp phân tích – tổng hợp tài liệu được sử dụng để thu thập cơ sở lý thuyết về ngôn ngữ ký hiệu, xử lý ngôn ngữ tự nhiên tiếng Việt, và các kỹ thuật biểu diễn chuyển động 3D, từ đó xác định hướng triển khai phù hợp.

Quy trình thực hiện đề tài được mô hình hóa theo các bước: tiếp nhận và phân tích văn bản đầu vào; tách và ánh xạ từ, cụm từ sang ký hiệu tương ứng; xây dựng trình tự ký hiệu dựa trên dữ liệu ký hiệu 3D có sẵn; xuất mô phỏng 3D để người dùng quan sát dưới dạng hoạt ảnh.

Hệ thống được thử nghiệm với tập văn bản tiếng Việt phổ biến nhằm kiểm tra khả năng chuyển đổi và mức độ rõ ràng của ký hiệu. Việc đánh giá kết quả được thực hiện thông qua phân tích so sánh giữa nội dung văn bản và trình tự ký hiệu sinh ra, kết hợp với quan sát tính tự nhiên, dễ hiểu của hoạt ảnh 3D trong các tình huống giả lập.

Escaneado con CamScanner

# CHƯƠNG 1: TỔNG QUAN VỀ HỆ THỐNG CHUYỂN VĂN BẢN THÀNH NGÔN NGỮ KÝ HIỆU

## 1.1 Khái quát về ngôn ngữ ký hiệu và nhu cầu giao tiếp của người khiếm thính

### *1.1.1 Định nghĩa và bản chất của ngôn ngữ ký hiệu*



**Hình 1.1 Ảnh minh họa động tác ký hiệu từ "mỳ tôm"**

Ngôn ngữ ký hiệu là một hệ thống giao tiếp hoàn chỉnh, hình thành và phát triển như một dạng ngôn ngữ tự nhiên sử dụng hình ảnh cơ thể thay cho âm thanh. Cách truyền đạt chủ yếu dựa vào chuyển động tay, vị trí của tay trong không gian, hướng bàn tay, đường di chuyển và nhịp độ động tác, kết hợp với các tín hiệu bổ trợ như nét mặt, ánh mắt hay tư thế phần thân trên. Khi các yếu tố này được phối hợp theo quy tắc nhất định, chúng tạo thành những đơn vị mang nghĩa tương đương với từ vựng hoặc cụm ý trong ngôn ngữ âm thanh. Điều cần nhấn mạnh là ngôn ngữ ký hiệu không phải bản sao hay biến thể của ngôn ngữ nói, mà tồn tại như một hệ thống độc lập với cơ chế cấu trúc riêng.

Ở góc độ nghiên cứu ngôn ngữ, dạng ngôn ngữ này thể hiện những mô hình tổ chức rất đặc thù, đặc biệt ở cách vận dụng không gian ba chiều để biểu đạt ý nghĩa. Nhờ khả năng khai thác không gian, người sử dụng có thể biểu thị mối liên hệ giữa các khái niệm, thể hiện nhiều lớp nghĩa, hoặc nhấn mạnh thông tin bằng các dấu hiệu phi thủ. Ngôn ngữ ký hiệu cũng sử dụng các khuôn mẫu câu đặc trưng như cấu trúc chủ đề – bình luận, hệ thống classifier cùng những biểu đạt mang tính hình tượng. Nhờ vậy,

Escaneado con CamScanner

người dùng có thể diễn đạt các nội dung trừu tượng, cảm xúc và suy nghĩ theo cách sinh động và linh hoạt.

### *1.1.2 Các đặc trưng ngôn ngữ học quan trọng*

Ngữ pháp không gian là đặc điểm nổi bật của ngôn ngữ ký hiệu, sử dụng không gian ba chiều để thể hiện quan hệ ngữ nghĩa và cú pháp. Người ký hiệu có thể chỉ vào một vị trí để biểu thị đối tượng và tiếp tục tham chiếu lại bằng cùng vị trí đó. Cách này giúp tăng tính trực quan và diễn đạt các quan hệ phức tạp mà ngôn ngữ nói khó thể hiện.

Một đặc trưng khác là cấu trúc câu theo mô hình chủ đề – bình luận. Thay vì trật tự SVO như ngôn ngữ nói, ngôn ngữ ký hiệu thường đưa chủ đề lên trước rồi mới đến phần bình luận. Ví dụ, "con chó" được nêu trước, sau đó mới mô tả hành động "chạy", giúp tổ chức thông tin rõ ràng và linh hoạt.

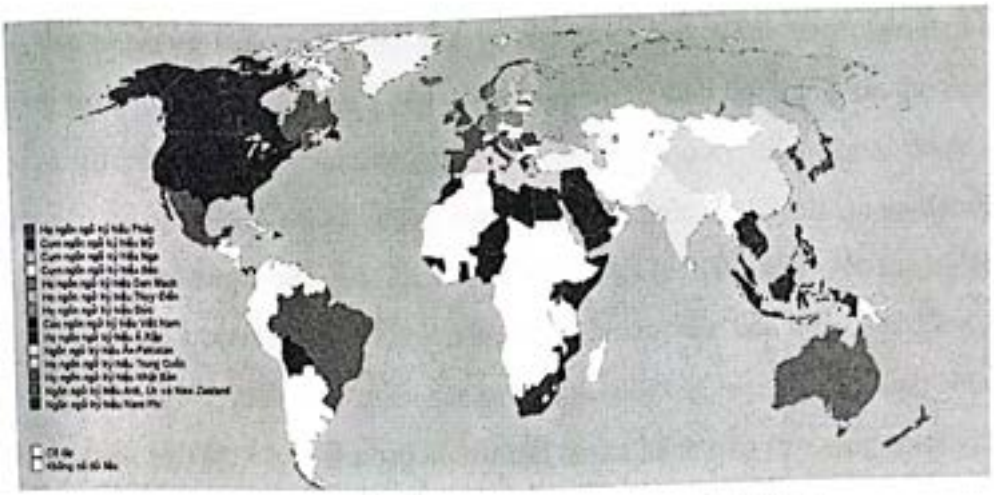
Classifier (phân loại) là các động tác hoặc hình dạng tay mô tả đặc tính, hình dạng, chuyển động của đối tượng. Chúng giúp tạo ra mô tả trực quan, sinh động và chi tiết, hỗ trợ người tiếp nhận dễ hình dung nội dung.

Tín hiệu phi-thủ như biểu cảm khuôn mặt, ánh mắt, cử chỉ đầu không chỉ thể hiện cảm xúc mà còn mang chức năng ngữ pháp, giúp phân biệt câu hỏi, mệnh lệnh hay khẳng định. Chúng làm phong phú ngôn ngữ ký hiệu và tăng khả năng biểu đạt.

Ngôn ngữ ký hiệu còn mang tính biểu tượng và ẩn dụ. Nhiều ký hiệu mang tính iconic, mô phỏng trực tiếp đối tượng hoặc hành động. Theo thời gian, chúng trở nên quy ước hóa, trừu tượng và đa nghĩa, cho thấy sự phát triển linh hoạt như các ngôn ngữ nói.

### *1.1.3 Đa dạng và biến thể ngôn ngữ ký hiệu*

Escaneado con CamScanner



**Hình 1.2 Bản đồ thế giới với các loại ngôn ngữ ký hiệu**

Ngôn ngữ ký hiệu không tồn tại dưới dạng duy nhất, mà được hình thành thành nhiều hệ thống khác nhau tùy theo quốc gia hoặc khu vực, ví dụ như ASL, BSL hay VSL. Ngay trong cùng một quốc gia cũng có thể thấy nhiều biến thể mang tính địa phương, gắn với cộng đồng người dùng tại từng vùng. Ở Việt Nam, VSL cũng được chia thành nhiều cách sử dụng khác nhau, thể hiện qua khác biệt về cấu trúc động tác, hình dạng tay hoặc phong cách thể hiện. Vì vậy, khi thiết kế công cụ chuyển đổi hay mô phỏng ký hiệu, việc tôn trọng đặc thù của từng biến thể là yêu cầu thiết yếu.

Đối với cộng đồng người điếc, hoạt động giao tiếp và tương tác đời sống diễn ra theo những phương thức riêng. Điều này dẫn đến nhu cầu tiếp nhận và xử lý thông tin không giống với người nghe. Dù các công nghệ hỗ trợ trực quan và những nền tảng ký hiệu đã cải thiện đáng kể khả năng tiếp cận thông tin, quá trình hòa nhập xã hội vẫn đối mặt với nhiều rào cản. Khó khăn có thể đến từ sự khác biệt trong hệ thống ký hiệu mà mỗi cá nhân được tiếp xúc trong gia đình hoặc cộng đồng, hoặc từ sự thay đổi của môi trường giao tiếp. Nắm được các yếu tố này giúp làm rõ mức độ phức tạp khi phát triển giải pháp hỗ trợ, đồng thời nhấn mạnh tầm quan trọng của việc thiết kế một hệ thống thân thiện và công bằng hơn cho người điếc.

### *1.1.4 Đặc điểm giao tiếp của người khiếm thính và nhu cầu thiết yếu*

Escaneado con CamScanner

Với nhiều cá nhân mất thính lực, hệ thống ký hiệu được tiếp nhận từ nhỏ và đóng vai trò như phương tiện giao tiếp đầu tiên, giúp họ lĩnh hội thông tin theo cách tự nhiên và trực quan. Ngược lại, những người bắt đầu học ký hiệu ở giai đoạn muộn thường gặp trở ngại do phải làm quen với một phương thức biểu đạt hoàn toàn mới.

Một bộ phận người điếc sử dụng linh hoạt cả ký hiệu lẫn chữ viết, tuy nhiên khả năng đọc viết giữa các cá nhân không đồng nhất, khiến việc tiếp nhận nội dung chính xác trong môi trường giáo dục, chăm sóc sức khỏe hoặc các loại giấy tờ hành chính trở nên khó khăn. Vì vậy, các giải pháp hỗ trợ cần tạo ra kênh tiếp cận song song giữa ký hiệu và phụ đề để bảo đảm tính minh bạch và dễ hiểu.

Nhu cầu giao tiếp của cộng đồng này trải rộng trong nhiều lĩnh vực: học tập, khám chữa bệnh, dịch vụ công và sinh hoạt thường ngày. Ở trường học, họ cần bài giảng đi kèm người hỗ trợ hoặc hệ thống phụ đề. Tại cơ sở y tế, cần có sự tham gia của phiên dịch hoặc công nghệ hỗ trợ ngôn ngữ. Trong các dịch vụ công, thông tin nên được cung cấp thông qua hình thức dễ tiếp nhận đối với người điếc. Mỗi tình huống đều yêu cầu một cách tiếp cận ngôn ngữ và kỹ thuật phù hợp để đảm bảo hiệu quả truyền đạt.

Bên cạnh yếu tố ngôn ngữ, khía cạnh xã hội và việc hòa nhập cộng đồng cũng giữ vai trò quan trọng. Nếu thiếu các phương tiện hỗ trợ, người điếc dễ gặp hạn chế trong tiếp cận thông tin, học vấn và dịch vụ thiết yếu, từ đó ảnh hưởng đến cơ hội việc làm và sự tham gia xã hội. Những hệ thống chuyển đổi nội dung chữ viết sang hình thức ký hiệu có thể góp phần xoá bớt rào cản này, tạo điều kiện để họ tiếp cận thông tin tốt hơn, nâng cao khả năng học tập và mở rộng cơ hội nghề nghiệp.

### *1.1.5 Vấn đề thực tế trong giao tiếp và tiếp cận thông tin*

Trong đời sống hằng ngày, người điếc thường đối mặt với nhiều trở ngại khi trao đổi và tiếp nhận nội dung, chủ yếu vì thiếu nguồn tài liệu bằng ký hiệu và vì các

Escaneado con CamScanner

dịch vụ hỗ trợ hiện có chưa phù hợp với nhu cầu sử dụng. Những hạn chế này tác động trực tiếp đến việc học, cơ hội nghề nghiệp và mức độ tham gia vào cộng đồng.

Một trong những vấn đề nổi bật là sự khan hiếm tài liệu trình bày bằng ký hiệu. Phần lớn nội dung trên internet vẫn được thể hiện dưới dạng chữ hoặc âm thanh, khiến việc tiếp cận trở nên khó khăn. Dịch vụ phiên dịch qua video do con người thực hiện cũng có giới hạn, vừa tốn thời gian vừa khó mở rộng, nên không thể phục vụ kịp cho các bối cảnh như lớp học, bệnh viện hoặc thủ tục hành chính.

Chất lượng phiên dịch cũng là điểm dễ phát sinh sai lệch. Việc dựa hoàn toàn vào người phiên dịch có thể dẫn đến khác biệt về độ chính xác tùy vào kinh nghiệm của từng cá nhân. Các hệ thống tự động nếu chưa đạt mức hoàn thiện cao thường gặp khó trong việc tái hiện các yếu tố ngoài tay như biểu cảm hoặc cử chỉ bổ trợ, khiến thông điệp bị thiếu sắc thái và giảm độ tự nhiên.

Bên cạnh đó, sự phong phú về cách dùng và tình huống giao tiếp trong ký hiệu làm cho quá trình chuyển đổi từ câu chữ sang động tác trở nên phức tạp. Cùng một câu nhưng cách biểu đạt có thể thay đổi tùy mục tiêu truyền tải, ví dụ dùng cho học tập, trò chuyện thân mật hoặc thông báo công khai. Vì vậy, công nghệ xử lý cần khả năng phân tích linh hoạt theo bối cảnh và ý nghĩa để thông điệp được truyền đi một cách rõ ràng và chính xác.

### *1.1.6 Hệ quả cho thiết kế giải pháp công nghệ*

Khi thiết kế công nghệ cho người khiếm thính, đặc biệt trong chuyển đổi và giao tiếp qua ngôn ngữ ký hiệu, cần xem xét yếu tố ngôn ngữ học, biểu đạt, dữ liệu và tính linh hoạt. Những yếu tố này quyết định hiệu quả, sự phù hợp và khả năng sử dụng trong đời sống hàng ngày.

Về ngôn ngữ học, hệ thống không chỉ dịch từng từ mà phải xử lý đúng ngữ nghĩa, cú pháp và prosody (nhịp điệu, sắc thái) của ngôn ngữ ký hiệu. Điều này giúp duy trì sự tự nhiên và chính xác khi truyền đạt thông tin.

Escaneado con CamScanner

Về tính biểu đạt, ngôn ngữ ký hiệu phụ thuộc cả vào tay và tín hiệu phi-thủ như biểu cảm, nhún vai, tư thế. Hệ thống cần mô phỏng đầy đủ để đảm bảo thông điệp không bị thiếu ý nghĩa hay cảm xúc.

Về dữ liệu, cần cơ sở dữ liệu đa dạng gồm video ký hiệu, annotation, gloss và mô tả chuyển động. Đây là nền tảng huấn luyện mô hình, giúp hệ thống hiểu rõ ngữ pháp và ngữ nghĩa để dịch chính xác.

Cuối cùng, hệ thống phải có tính tùy chỉnh theo người dùng, như điều chỉnh tốc độ, phân đoạn nội dung, thêm phụ đề, chọn phương ngữ hoặc giọng điệu. Sự linh hoạt này giúp người dùng tiếp cận thông tin dễ dàng hơn và nâng cao hiệu quả giao tiếp.

### *1.1.7 Tác động giáo dục và xã hội*



**Hình 1.3 Lớp học dạy ngôn ngữ ký hiệu**

Việc sử dụng ký hiệu trong môi trường học tập, trong các hoạt động hỗ trợ cộng đồng và trong lĩnh vực truyền thông góp phần tạo nên một không gian thân thiện hơn cho người khiếm thính, đồng thời thúc đẩy tính hòa nhập trong xã hội. Nhờ hình thái biểu đạt rõ ràng và dễ nhận biết, người điếc có thể giao tiếp tự tin hơn và tiếp nhận

Escaneado con CamScanner

thông tin trong nhiều tình huống khác nhau. Điều này giúp họ tiếp cận kiến thức và các dịch vụ quan trọng tương tự những người khác trong cộng đồng.

Ở mảng giáo dục, ký hiệu giữ vai trò như một phương tiện truyền đạt kiến thức hiệu quả. Khi bài giảng được số hóa và bổ sung phần mô phỏng động tác, khả năng tiếp thu của học sinh tăng cao, đồng thời tạo điều kiện cho việc tự học ngoài giờ học chính. Các tài liệu trực quan, chẳng hạn video hoặc mô hình 3D tái hiện động tác, cũng cho phép học sinh xem lại và tra cứu dễ dàng, hỗ trợ củng cố kiến thức một cách tự nhiên hơn.

Trong khu vực dịch vụ công và chăm sóc sức khỏe, việc áp dụng ký hiệu giúp người điếc tiếp nhận hướng dẫn, cảnh báo hoặc thông báo quan trọng một cách chính xác. Điều này đặc biệt hữu ích trong các tình huống cần phản ứng nhanh như thảm họa, bệnh dịch hoặc các sự cố bất ngờ. Ngoài ra, nhiều thủ tục hành chính hoặc quy trình phục vụ cộng đồng còn có thể được minh họa bằng động tác, giúp giảm bớt khó khăn phát sinh từ rào cản giao tiếp.

Trong lĩnh vực truyền thông và phát sóng, việc hiển thị ký hiệu ngay trên bản tin, chương trình trực tiếp hoặc các cuộc họp trực tuyến làm tăng mức độ tiếp cận của người khiếm thính đối với thông tin thời sự. Điều này không chỉ giúp họ theo dõi nội dung một cách đầy đủ hơn mà còn góp phần xây dựng môi trường truyền thông công bằng, hạn chế sự chênh lệch về khả năng tiếp cận giữa các nhóm dân cư.

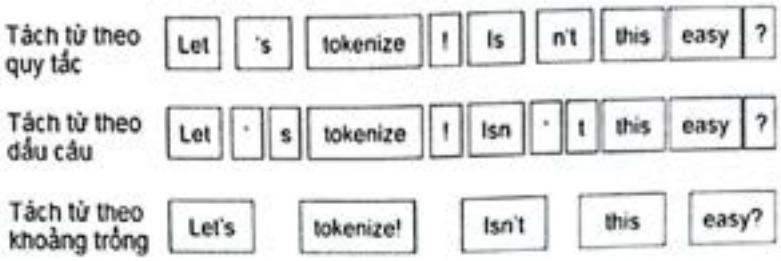
## 1.2 Các tiếp cận công nghệ trong chuyển đổi văn bản sang ngôn ngữ ký hiệu

Chuyển đổi văn bản thành ngôn ngữ ký hiệu là một bài toán đa ngành, kết hợp ngôn ngữ học, xử lý ngôn ngữ tự nhiên, thị giác máy tính và đồ họa/hoạt họa 3D. Về phương pháp luận, các tiếp cận chính có thể được chia thành ba nhóm lớn: (1) phương pháp dựa trên quy tắc/to từ điển, (2) phương pháp dựa trên dữ liệu (corpus-driven / học máy), và (3) các hệ thống lai (hybrid) kết hợp hai hướng trên. Mỗi hướng có lợi thế,

Escaneado con CamScanner

hạn chế và các kỹ thuật phụ trợ riêng, phần dưới đây phân tích chi tiết từng thành phần của pipeline và các lựa chọn phương pháp kỹ thuật.

### *1.2.1 Tiền xử lý ngôn ngữ và ánh xạ ngôn ngữ*




**Let's tokenize! Isn't this easy?**

**Hình 1.4 Hình ảnh minh họa tokenization**

Trong hệ thống dịch ngôn ngữ ký hiệu từ văn bản, tiền xử lý và ánh xạ ngôn ngữ là bước nền tảng. Quy trình này giúp chuyển đổi văn bản sang ký hiệu chính xác, giữ được ngữ nghĩa và ngữ dụng, đồng thời tạo cơ sở cho các mô hình dịch tự động xử lý tình huống phức tạp.

Tiền xử lý bắt đầu bằng chuẩn hóa văn bản: loại bỏ ký tự đặc biệt, sửa dấu câu, chuẩn hóa chữ hoa/thường. Sau đó hệ thống tách câu, tách từ và gán nhãn từ loại để nhận diện cú pháp, hỗ trợ hiểu cấu trúc trước khi dịch sang ký hiệu. Một thách thức quan trọng là xử lý tên riêng. Chúng thường không có ký hiệu trực tiếp nên cần chuyển thành fingerspelling, tức đánh vần từng chữ cái bằng động tác tay. Đây là yếu tố quan trọng để bản dịch tự nhiên và chính xác.

Sau tiền xử lý, văn bản được chuyển thành gloss – chuỗi ký hiệu phản ánh ý nghĩa gốc, kèm chỉ dẫn về nhịp điệu, biểu cảm và vị trí không gian. Gloss giúp tách ngữ nghĩa khỏi động tác, tạo điều kiện cho nhiều phương pháp sinh chuyển động khác nhau. Ánh xạ văn bản sang gloss có thể dùng từ điển ký hiệu cho trường hợp đơn giản, hoặc mô hình cú pháp và học máy như seq2seq để xử lý ngữ nghĩa phức tạp, đảm bảo dịch chính xác.

Escaneado con CamScanner

Theo các nghiên cứu gần đây, mô hình PhoBERT đạt độ chính xác 93% trong tác vụ phân loại văn bản tiếng Việt [2], trong khi các benchmark năm 2025 cho thấy một số mô hình embedding tiếng Việt mới có thể đạt F1-score trên 90% cho Named Entity Recognition (NER). Những con số này chứng minh rằng các kỹ thuật NLP hiện đại có thể hỗ trợ mạnh mẽ cho việc ánh xạ gloss, giúp hệ thống dịch ký hiệu đạt độ tin cậy cao hơn.

Cuối cùng, xử lý ngữ cảnh và ngữ dụng là cần thiết để thông điệp phù hợp mục tiêu giao tiếp. Các mô-đun ngữ cảnh giúp chọn chuỗi gloss đúng với tình huống, nâng cao tính chính xác và tự nhiên của bản dịch.

### *1.2.2 Phương pháp sinh chuyển động ký hiệu*

Trong xây dựng hệ thống dịch ngôn ngữ ký hiệu, việc lựa chọn ký hiệu và sinh chuyển động là yếu tố then chốt. Quá trình này không chỉ ánh xạ chính xác từ văn bản sang ký hiệu mà còn xử lý các yếu tố như động tác tay, vị trí, biểu cảm khuôn mặt và sự nối tiếp mượt mà giữa các ký hiệu. Nền tảng của quy trình là xây dựng từ điển ký hiệu kèm metadata chi tiết, bao gồm tay sử dụng, hình dáng, hướng, vị trí, chuyển động, dấu hiệu phi-thủ và thời lượng. Metadata này đảm bảo hệ thống chọn đúng ký hiệu và phản ánh đầy đủ ngữ nghĩa.

Từ điển ký hiệu cần có nhãn rõ ràng để phân loại các nhóm như lexical signs, classifier signs và fingerspelling. Lexical signs là ký hiệu chuẩn cho từ vựng; classifier signs mô tả đặc điểm hoặc hành động, đòi hỏi thông tin hình học chi tiết; fingerspelling dùng để đánh vần tên riêng hoặc từ mượn. Việc ghi chú chính xác giúp bản dịch tự nhiên và chính xác hơn.

Để tạo ra động tác tay và biểu cảm tự nhiên, có nhiều phương pháp khác nhau, từ sử dụng video ghi sẵn đến hoạt họa và học máy. Mỗi phương pháp có ưu, nhược điểm riêng, tùy thuộc vào yêu cầu chất lượng, tính linh hoạt và khả năng mở rộng:

Escaneado con CamScanner

Truy hồi video: dùng kho clip ký hiệu đã ghi sẵn để ghép thành câu hoàn chỉnh [3]. Ưu điểm là hình ảnh và biểu cảm rất thực tế, nhưng hạn chế ở chỗ chuyển tiếp không mượt, khó đồng bộ và kho dữ liệu hữu hạn. Một số kỹ thuật như time-warping hay blending có thể cải thiện nhưng chưa khắc phục triệt để.

Hoạt họa khung xương: biểu diễn ký hiệu bằng keyframe hoặc quỹ đạo khớp, sau đó nội suy để tạo chuyển động mượt. IK và FK được dùng để điều chỉnh khớp, phù hợp cho avatar 3D vì dễ điều khiển và tiết kiệm bộ nhớ. Thực tế triển khai thường dùng khoảng 20–30 keyframes cho mỗi ký hiệu để cân bằng giữa độ mượt và chi phí tính toán. Tuy nhiên, biểu cảm khuôn mặt và chi tiết phức tạp vẫn khó tái hiện [3].

Học máy: mô hình sequence-to-sequence chuyển gloss hoặc embedding văn bản thành chuỗi vị trí khớp [4]. Các kiến trúc như RNN, LSTM, Transformer, GAN, VAE hay diffusion models có thể sinh chuyển động đa dạng và mượt. Ưu điểm là khả năng mở rộng, nhưng cần dữ liệu lớn và chi tiết, nếu không sẽ thiếu các yếu tố tinh tế như biểu cảm hay hình dáng tay.

Mô hình lai: kết hợp nhiều phương pháp để tận dụng ưu điểm, ví dụ dùng video cho ký hiệu có sẵn, học máy hoặc hoạt họa cho phần thiếu, hoặc kết hợp học máy với quy tắc cho fingerspelling.

Ngoài ra, khi chuyển đổi nhiều ký hiệu trong câu, hệ thống cần xử lý co-articulation để các động tác nối tiếp mượt mà. Nếu chỉ ghép clip thô sẽ gây giật, mất tự nhiên. Do đó, cần có mô-đun chuyển tiếp hoặc mô hình hóa tích hợp để tạo chuỗi động tác liền mạch, đảm bảo tính chính xác và dễ hiểu trong giao tiếp.

### *1.2.3 Thách thức và giới hạn kỹ thuật*

Quá trình chuyển từ pose 2D sang mô hình 3D của avatar trong ngôn ngữ ký hiệu là một nhiệm vụ phức tạp, kết hợp nhiều kỹ thuật xử lý hình ảnh và học máy. Mục tiêu không chỉ tái tạo chính xác chuyển động cơ thể mà còn thể hiện đúng cử chỉ tay, biểu cảm khuôn mặt và các chi tiết tinh tế.

Escaneado con CamScanner

Bước đầu là thu nhận pose 2D, thường từ video mocap (marker-based/markerless) hoặc ước tính keypoints bằng công cụ như OpenPose, vốn có khả năng phát hiện tới 135 landmarks trên cơ thể, bàn tay, mặt và bàn chân [5]. Tương tự, MediaPipe Pose cung cấp 33 landmarks 3D cho toàn bộ cơ thể. Các tập dữ liệu video ký hiệu thường được ghi ở 25–30 FPS, mỗi clip dài trung bình 3–5 giây với hàng nghìn frame, đảm bảo độ chi tiết cho quá trình huấn luyện.

Tiếp theo là lifting sang pose 3D. Các mô hình như SMPL và SMPL-X được sử dụng rộng rãi; riêng SMPL-X có tới 10.475 vertices và 54 joints, bao gồm cả khớp cho ngón tay và khuôn mặt [6]. Kỹ thuật SMPLifyX được dùng để "fit" keypoints 2D thành mesh 3D hoàn chỉnh, đảm bảo độ chính xác cao và khả năng tái hiện chi tiết.

Trong giai đoạn retargeting, chuyển động từ skeleton gốc được điều chỉnh cho avatar đích. Thách thức là xử lý xung đột khớp, tránh xuyên mô hình và giữ nguyên handshape. Rigging chi tiết cho tay và ngón tay (với hàng chục DOF – Degrees of Freedom) cùng blendshapes cho khuôn mặt giúp avatar tự nhiên và sống động. Cuối cùng là rendering và visualization. Render mesh 3D cho độ chân thực cao nhưng tốn tài nguyên; avatar stylized hoặc low-poly giúp giảm chi phí. Đồng bộ thời gian giữa pose và gloss/timing rất quan trọng để ký hiệu chính xác.

Trong ngôn ngữ ký hiệu, các tín hiệu phi-thủ như biểu cảm khuôn mặt, ánh mắt và cử động cơ thể giữ vai trò quan trọng trong việc truyền đạt ý nghĩa và cấu trúc cú pháp. Một khó khăn chính là cần chú thích chi tiết cho từng yếu tố như biểu cảm, ánh mắt hay động tác cơ thể. Các tín hiệu này khó nhận diện và mô phỏng, do đó cần các mô-đun riêng để xử lý và sinh chuyển động cho mặt và thân, nhằm đảm bảo cử chỉ vừa chính xác vừa sống động. Ngoài ra, việc đồng bộ giữa tay và tín hiệu phi-thủ quyết định prosody, tức tốc độ, nhịp điệu và ngắt quãng trong câu. Nếu bỏ qua yếu tố phi-thủ, hệ thống dễ mất tự nhiên và gây khó khăn cho người học.

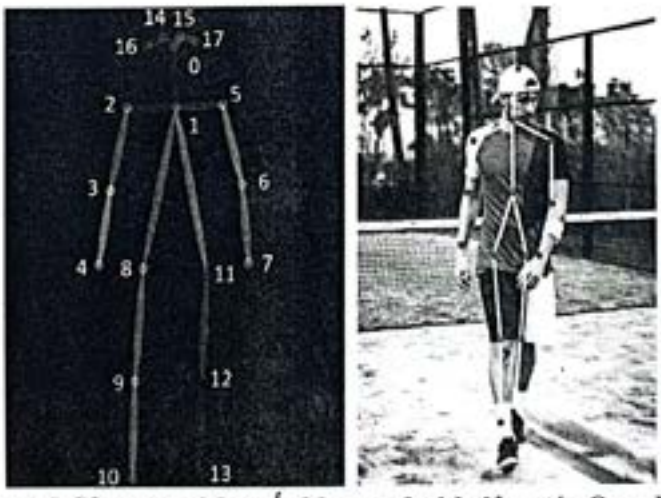
Các tính năng đặc thù như fingerspelling, classifier signs và spatial referencing cũng đặt ra nhiều thách thức. Fingerspelling được dùng để diễn đạt tên riêng hoặc từ không có ký hiệu, đòi hỏi mô-đun chuyên biệt để tái hiện chính xác. Classifier signs

Escaneado con CamScanner

mô tả hình dạng, kích thước, vị trí và chuyển động của đối tượng, cần dữ liệu chi tiết về quỹ đạo và động tác, thường dựa vào mocap hoặc mô hình sinh chuyển động. Spatial referencing cho phép chỉ định và tham chiếu lại vị trí đối tượng trong không gian ba chiều, hệ thống cần quản lý tốt để duy trì mối quan hệ không gian và ngữ cảnh.

Để hệ thống ngôn ngữ ký hiệu tự động hoạt động hiệu quả, cần đánh giá chất lượng và hiệu suất. Các phương pháp đánh giá thường gồm ngôn ngữ học, chất lượng chuyển động và các chỉ số tự động. Một hướng phổ biến là dùng chỉ số tự động để đo ngữ nghĩa và cú pháp, đánh giá mức độ khớp giữa văn bản gốc và chuỗi ký hiệu. Song song đó, đánh giá chất lượng động tác dựa trên mức độ tự nhiên, sự trôi chảy và đồng bộ. Các kỹ thuật phân tích khung xương hoặc vector chuyển động giúp kiểm tra khả năng tái tạo chuyển động mượt và sát thực tế. Ngoài ra, có thể dùng chỉ số định lượng như RMSE cho vị trí khớp, DTW cho chuỗi thời gian, hoặc FID để đánh giá phân phối chuyển động. Cuối cùng, đánh giá end-to-end kiểm tra toàn bộ chuỗi xử lý, đối chiếu đầu ra với mục tiêu ban đầu để đảm bảo ngữ nghĩa được duy trì và thông tin truyền đạt đầy đủ, chính xác.

### *1.2.4 Dữ liệu, công cụ và nguồn lực*

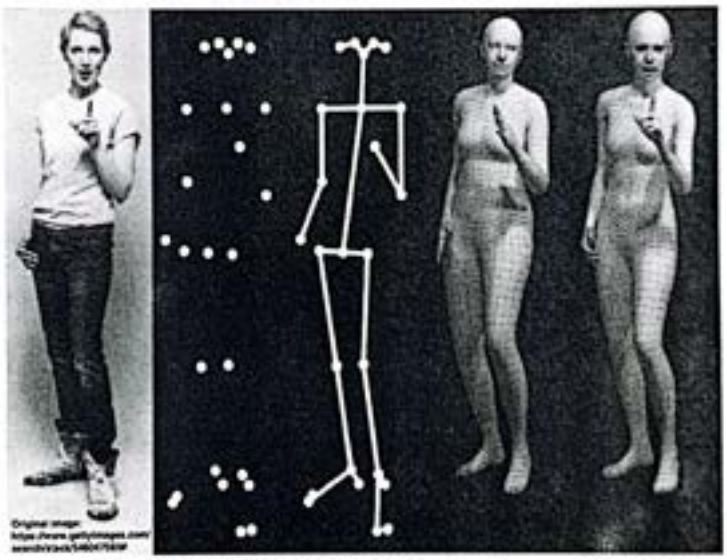



**Hình 1.5 Công cụ trích xuất khung cử chỉ, động tác OpenPose**

Escaneado con CamScanner

Để phát triển hệ thống ngôn ngữ ký hiệu tự động, dữ liệu chất lượng là yếu tố cốt lõi. Thách thức lớn hiện nay là sự thiếu hụt tài nguyên cho tiếng Việt, đặc biệt các bộ dữ liệu ký hiệu có gán nhãn và mô phỏng chuyển động. Video ký hiệu gán nhãn (gloss) là nền tảng huấn luyện, trong khi dữ liệu mocap và 2D pose giúp tối ưu mô hình sinh chuyển động, tạo động tác tự nhiên và mượt mà.

Với tiếng Việt, nguồn dữ liệu còn hạn chế, gây khó khăn cho phát triển công cụ tự động. Một ví dụ là QIPEDC dataset, hiện có khoảng 4.000 video, được thực hiện bởi người điếc từ 3 vùng miền Bắc Trung Nam, mỗi video dài trung bình 5–10 giây, tương đương hơn 5 giờ dữ liệu ký hiệu [7]. Đây là nguồn dữ liệu quan trọng nhưng vẫn còn nhỏ so với các bộ dữ liệu quốc tế như WLASL (hơn 21.000 video, độ dài từ 0.3-8 giây, trung bình 2 giây) [8]. Có thể tận dụng dữ liệu từ các ngôn ngữ ký hiệu khác như ASL hoặc BSL, nhưng cần thận trọng vì khác biệt ngữ nghĩa và cấu trúc có thể ảnh hưởng đến độ chính xác khi áp dụng cho tiếng Việt.



**Hình 1.6 Mô hình SMPLifyX sẽ xây dựng mô hình 3D**

Ngoài dữ liệu, công cụ và thư viện cũng rất quan trọng. OpenPose, MediaPipe hỗ trợ nhận diện keypoints từ video 2D; SMPLifyX giúp fitting mô hình 3D cơ thể;

Escaneado con CamScanner

Unity và Blender hỗ trợ render và animation chất lượng cao. Các pipeline render bằng Python cũng hữu ích cho việc tùy chỉnh và mở rộng mô phỏng.

### *1.2.5 Xu hướng phát triển công nghệ*

Việc xây dựng hệ thống ký hiệu tự động gặp phải nhiều khó khăn, từ hạn chế về dữ liệu chất lượng đến độ phức tạp trong việc mô phỏng chuyển động. Một trở ngại lớn là thiếu các video ký hiệu được gắn nhãn đầy đủ (gloss, tag phi-thủ, thời gian). Giải pháp khả thi là thu thập và gán nhãn dữ liệu một cách tỉ mỉ, kết hợp chuyên môn của các nhà nghiên cứu ngôn ngữ và kỹ thuật công nghệ. Hiện nay, các bộ dữ liệu quốc tế như WLASL (~21.083 video) [8] và SignAvatars (70.000 chuỗi chuyển động) [9] cho thấy quy mô cần thiết để huấn luyện mô hình ổn định.

Một thách thức khác là đảm bảo các chuyển tiếp giữa ký hiệu liền mạch. Hiện tượng co-articulation dễ tạo cảm giác cứng nhắc, vì vậy có thể áp dụng mô hình học sâu theo thời gian hoặc các module blending nhằm tạo ra chuyển động mượt mà hơn. Các nghiên cứu về Diffusion-based Continuous Sign Language Generation (CVPR 2025) cho thấy mô hình diffusion có thể cải thiện độ mượt và tính cá nhân hóa, với FID giảm hơn 15% so với baseline Transformer [10].

Trong thời gian tới, một hướng phát triển quan trọng của công nghệ chuyển đổi văn bản sang ký hiệu là xây dựng các mô hình xử lý liền mạch từ đầu vào đến đầu ra. Các hệ thống này tạo chuyển động ký hiệu trực tiếp từ câu chữ mà không cần nhiều tầng xử lý trung gian. Việc ứng dụng các kiến trúc học sâu thế hệ mới như Transformer hoặc mô hình diffusion được kỳ vọng sẽ rút ngắn chuỗi xử lý, đồng thời cải thiện mức ổn định của kết quả. Ví dụ, mô hình DiffSLT (2024) đã chứng minh khả năng dịch gloss-free với BLEU-4 đạt 23.5, cao hơn các baseline cũ [11].

Một hướng quan trọng khác là nâng cao độ chi tiết khi tái hiện các yếu tố không dùng tay, gồm biểu cảm khuôn mặt, hướng nhìn hoặc tư thế thân trên. Các mô hình chuyên biệt dành cho những thành phần này sẽ thể hiện tốt hơn sắc thái và các lớp

Escaneado con CamScanner

nghĩa phụ trong từng ký hiệu. Để làm được điều đó, việc xây dựng bộ dữ liệu có nhãn đầy đủ về cảm xúc và chuyển động vi mô là yêu cầu then chốt, giúp mô phỏng trở nên sinh động và gần với thực tế hơn. Ngoài ra, các hệ thống trong tương lai có thể kết hợp hai phương pháp: truy xuất các đoạn ký hiệu có sẵn với chất lượng cao và tạo mới chuyển động bằng mô hình sinh. Cách làm này giúp tận dụng sự chân thực từ video gốc, đồng thời giữ được sự linh hoạt khi cần tạo ra các ký hiệu hiếm gặp, qua đó giảm bớt gánh nặng thu thập dữ liệu quy mô lớn về lâu dài.

Công nghệ mô phỏng ký hiệu cũng hứa hẹn được cải thiện khi liên kết với nhiều loại đầu vào khác nhau, chẳng hạn như hệ thống chuyển âm thanh thành văn bản hoặc tín hiệu hình ảnh hỗ trợ phân tích bối cảnh. Khi tích hợp đa nguồn, hệ thống có thể phản ứng nhạy hơn trước tình huống thực tế và cung cấp khả năng hiểu thông tin thông qua nhiều kênh cùng lúc. Một xu thế khác được quan tâm là điều chỉnh hệ thống theo sở thích của từng cá nhân. Các nền tảng có thể tự động ghi nhận tốc độ trình chiếu ưa dùng, kiểu nhân vật hiển thị, hoặc biến thể ký hiệu của từng vùng mà người dùng thường tiếp xúc. Điều này cho phép trải nghiệm phù hợp hơn với mục đích học tập, tra cứu hoặc giao tiếp của mỗi người.

Cuối cùng, việc xây dựng và chuẩn hoá tập dữ liệu dành riêng cho ký hiệu tiếng Việt là nhiệm vụ thiết yếu. Bộ dữ liệu này nên bao gồm gloss, thông tin ngữ dụng, đặc tả không gian và các tín hiệu phi thủ, phục vụ cho quá trình huấn luyện mô hình đạt độ chính xác cao, dễ mở rộng và thích hợp cho nhiều lĩnh vực như giáo dục, truyền thông và ứng dụng đời sống.

## 1.3 Ứng dụng mô hình 3D và hệ thống trực quan hóa trong truyền tải ký hiệu

Mô hình 3D và hệ thống trực quan hóa đóng vai trò then chốt trong việc đưa ngôn ngữ ký hiệu từ dạng lý thuyết/ghi hình tĩnh tới trải nghiệm tương tác, trực quan

Escaneado con CamScanner

và có thể mở rộng. Phần này đi sâu vào các khía cạnh ứng dụng, cấu trúc chức năng, lợi ích, giới hạn và những cân nhắc thiết kế khi sử dụng 3D để truyền tải ký hiệu.

### *1.3.1 Ứng dụng thực tiễn và bối cảnh triển khai*

Ngôn ngữ ký hiệu, với khả năng truyền đạt thông tin thông qua hình ảnh và cử chỉ trực quan, ngày càng được ứng dụng rộng rãi trong các lĩnh vực như giáo dục, truyền thông, dịch vụ công và nhiều hoạt động xã hội khác. Các công nghệ hiện đại, ví dụ như hệ thống chuyển đổi văn bản thành ký hiệu qua video hoặc mô hình hoạt hình, đang giúp người khiếm thính tiếp cận thông tin nhanh chóng và trực quan hơn.

Trong giáo dục, việc ứng dụng ký hiệu hỗ trợ học sinh khiếm thính tiếp thu kiến thức dễ dàng hơn. Theo thống kê, hiện có khoảng 2–3 triệu người Mỹ sử dụng ngôn ngữ ký hiệu, trong đó 500.000 người dùng ASL như ngôn ngữ chính. Các bài giảng số kết hợp mô phỏng 3D minh họa cử chỉ giúp người học nắm vững từ vựng, ngữ pháp và hội thoại, đồng thời các chức năng như tua lại, điều chỉnh tốc độ hay theo dõi chi tiết từng động tác giúp việc tự học trở nên thuận tiện. Ở lĩnh vực thông tin công cộng, hệ thống ký hiệu hỗ trợ truyền tải tin tức quan trọng, hướng dẫn khẩn cấp hoặc thủ tục hành chính qua video trên website, màn hình công cộng hoặc ứng dụng, đảm bảo người khiếm thính nhận được thông tin kịp thời. Hiện nay, WHO ước tính có khoảng 430 triệu người trên thế giới mắc khiếm thính mức độ nặng cần phục hồi chức năng. Trong các sự kiện trực tuyến hoặc truyền hình trực tiếp, công nghệ ký hiệu có thể được tích hợp để cung cấp bản chuyển ngữ theo thời gian thực, có thể kết hợp với phiên dịch viên, giúp người khiếm thính tham gia dễ dàng hơn. Trong truyền thông và giải trí, việc lồng ghép ký hiệu vào bản tin, chương trình hay nội dung số giúp tạo ra môi trường tiếp nhận thông tin bình đẳng, hạn chế rào cản giao tiếp. Trong lĩnh vực y tế và dịch vụ xã hội, việc chuyển đổi thông tin về thuốc men, hướng dẫn điều trị hoặc chỉ dẫn khẩn cấp sang ký hiệu giúp người khiếm thính tiếp cận dịch vụ an toàn, đồng thời nắm bắt thông tin sức khỏe một cách chính xác. Ngoài ra, công nghệ ký hiệu còn đóng vai trò

Escaneado con CamScanner

quan trọng trong nghiên cứu và lưu trữ các biến thể ngôn ngữ ký hiệu, góp phần bảo tồn giá trị văn hóa và thúc đẩy phát triển nghiên cứu ngôn ngữ ký hiệu.

Ngoài ra, thị trường công nghệ hỗ trợ người khiếm thính cũng đang phát triển mạnh. Theo báo cáo năm 2025, thị trường công nghệ trợ giúp toàn cầu (assistive technology) đạt giá trị hơn 30 tỷ USD, với tốc độ tăng trưởng trung bình hàng năm khoảng 7–10% trong giai đoạn 2025–2032. Trong đó, các thiết bị hỗ trợ giao tiếp và phần mềm cho người khiếm thính chiếm tỷ trọng đáng kể, phản ánh nhu cầu ngày càng cao về giải pháp công nghệ giúp giảm rào cản thông tin và tăng khả năng hòa nhập xã hội.

Bên cạnh ứng dụng, việc triển khai công nghệ ký hiệu cũng đòi hỏi chiến lược phù hợp. Có ba cách tiếp cận chính để render chuyển động và biểu hiện trong ngôn ngữ ký hiệu: pre-rendered clips, real-time rendering, và hybrid approach. Pre-rendered clips dùng video đã ghi sẵn cho từng ký hiệu, sau đó ghép thành câu. Ưu điểm là hiệu suất cao, hình ảnh ổn định, phù hợp cho bài học cố định. Nhược điểm là thiếu linh hoạt khi cần thay đổi tốc độ, ngữ điệu, khó chuyển tiếp mượt và tốn dung lượng lưu trữ. Real-time rendering tạo chuyển động trực tiếp từ skeleton hoặc 3D mesh, cho phép linh hoạt và cá nhân hóa, thích hợp cho ứng dụng tương tác như hội thảo trực tuyến hay trợ giúp thời gian thực. Tuy nhiên, nó đòi hỏi tài nguyên tính toán lớn và có thể gây độ trễ. Hybrid approach kết hợp cả hai: dùng clip sẵn cho ký hiệu phổ biến, còn ký hiệu ít gặp hoặc chuyển tiếp sẽ render thời gian thực. Cách này cân bằng giữa chất lượng, hiệu suất và tính linh hoạt, phù hợp cho hệ thống cần vừa ổn định vừa mở rộng.

Để xác định mức độ hoạt động của một hệ thống chuyển đổi sang ký hiệu dưới dạng số, cần tiến hành các bước kiểm tra và phân tích nhằm xác nhận chất lượng đầu ra. Những thước đo này cho biết mức chính xác, độ dễ tiếp nhận và hiệu suất truyền tải thông tin. Quá trình đánh giá thường tập trung vào ba nhóm: khả năng nắm bắt nội dung, độ mượt của chuỗi động tác và mức độ thân thiện khi sử dụng. Việc đánh giá khả năng tiếp nhận nội dung nhằm xác định người dùng, bao gồm người dùng quen với ký hiệu và cá nhân mất thính lực, có diễn giải đúng ý từ mô hình 3D hay không. Người

Escaneado con CamScanner

tham gia được xem đoạn mô phỏng rồi mô tả lại thông tin họ hiểu. Một số chỉ số được sử dụng gồm tỉ lệ diễn giải chính xác, thời gian xử lý và mức sai lệch so với ý nghĩa ban đầu.

Kiểm tra chất lượng động tác tập trung vào sự tự nhiên của các cử chỉ. Phương pháp định tính thường khảo sát cảm nhận về độ mượt, sự linh hoạt và mức dễ hiểu. Ở hướng định lượng, các kỹ thuật đo lường như RMSE hoặc DTW được dùng để đối chiếu vị trí các khớp với dữ liệu chuẩn nhằm đánh giá độ khớp của chuyển động. Khía cạnh liên quan đến trải nghiệm thao tác xem xét mức tiện dụng của giao diện và cảm nhận tổng thể của người dùng. Các bộ câu hỏi như SUS, cùng trao đổi trực tiếp hoặc phân tích hành vi sử dụng như số lần tương tác, tua lại hoặc thay đổi tốc độ, giúp nhận diện mức độ hài lòng và đề xuất điều chỉnh để nâng cao chất lượng hệ thống.

### *1.3.2 Lợi ích của mô hình 3D và ý nghĩa trực quan*

So với video 2D, mô hình 3D mang lại nhiều lợi ích vượt trội trong phát triển hệ thống ngôn ngữ ký hiệu. Nó cho phép tái sử dụng, mở rộng, phân tích chi tiết và cá nhân hóa, mang lại trải nghiệm trực quan và linh hoạt hơn.

Ưu điểm lớn nhất là khả năng tái sử dụng và mở rộng. Một clip 3D có thể điều chỉnh góc nhìn, kích thước avatar, dễ chỉnh sửa và nâng cấp, giảm phụ thuộc vào video cố định. Ngoài ra, mô hình 3D cũng hỗ trợ phân tích chi tiết, cho phép phóng to các phần như bàn tay để quan sát handshape hay quỹ đạo chuyển động. Điều này giúp người học nắm rõ các động tác khó thấy trong video 2D. Một lợi ích khác là xử lý tốt co-articulation, tạo chuyển tiếp mượt mà giữa các ký hiệu nhờ khung xương và kỹ thuật nội suy, giúp động tác tự nhiên và dễ hiểu hơn. Một minh chứng thực nghiệm: hệ thống avatar-3D dựa trên văn bản/giọng nói sang ký hiệu cho Indian Sign Language (ISL) đã đạt tỷ lệ đúng ký hiệu khoảng 89.5% (Sign Error Rate = 10.50%) khi chuyển từ câu tiếng Anh sang chuỗi ký hiệu + avatar 3D [12].

Escaneado con CamScanner

Bên cạnh đó, mô hình 3D còn cho phép cá nhân hóa, như điều chỉnh tốc độ, kích thước, phong cách avatar (thực tế hoặc stylized), phù hợp với nhu cầu và ngữ cảnh sử dụng. Cuối cùng, mô hình 3D hỗ trợ tích hợp annotation vào timeline, đồng bộ phụ đề, highlight từ và metadata. Điều này giúp người học theo dõi ngữ nghĩa, ghi chú và tra cứu hiệu quả hơn. Một đánh giá từ khảo sát của người học sử dụng avatar/VR để học ký hiệu cho biết ~72% người tham gia cho rằng "hand gestures và handshapes rõ hơn trong môi trường 3D so với video", và ~78% nói rằng họ dễ theo dõi chuyển động trong 3D vì có thể quan sát từ nhiều góc nhìn khác nhau.

Thiết kế trải nghiệm học tập và giao tiếp cho người khiếm thính cũng là một lợi ích quan trọng của mô hình 3D. Các yếu tố chính gồm mức độ hiển thị chi tiết, đồng bộ văn bản, tùy chọn phương ngữ, kiểm soát tốc độ và trợ năng bổ sung. Hệ thống nên cho phép phóng to vùng tay, hiển thị skeleton hoặc vector chuyển động để người học dễ theo dõi, đặc biệt với ký hiệu phức tạp. Đồng bộ văn bản và ký hiệu giúp người dùng hiểu rõ mối liên hệ giữa từ và ký hiệu. Highlight từ khi ký hiệu diễn ra và cung cấp phụ đề song song sẽ hỗ trợ người học, nhất là người mới.

Tùy chọn phương ngữ cần thiết vì ngôn ngữ ký hiệu khác nhau theo vùng. Cho phép chọn biến thể phù hợp giúp người học làm quen và giao tiếp tốt hơn trong thực tế. Kiểm soát tốc độ giúp người học điều chỉnh theo khả năng tiếp thu, đặc biệt hữu ích cho người học chậm hoặc muốn xem lại nhiều lần. Cuối cùng, hệ thống cần có trợ năng bổ sung như giao diện high-contrast cho người khó phân biệt màu, phím tắt và điều khiển bàn phím để đảm bảo mọi người đều sử dụng thuận tiện.

### *1.3.3 Thành phần kỹ thuật và chức năng hệ thống trực quan hóa*

Khi thiết kế hệ thống trực quan hóa ngôn ngữ ký hiệu, một số thành phần kỹ thuật then chốt cần đảm bảo khả năng truyền tải chính xác và hiệu quả. Chúng bao gồm biểu diễn chuyển động, tái tạo chi tiết tay và mặt, các mô-đun chuyển tiếp, cùng tối ưu hóa hiệu suất và giao diện người dùng.

Escaneado con CamScanner

Biểu diễn chuyển động dựa trên skeleton/keypoints để mô phỏng động tác tay, cánh tay và cơ thể. Chuỗi tư thế (pose sequence) giúp theo dõi chuyển động liên tục, trong khi các mô hình tham số hoặc mesh 3D nâng cao chất lượng hiển thị và độ mượt.

Tái tạo chi tiết tay và mặt là thách thức lớn vì handshape có nhiều DOF. Hệ thống cần dữ liệu chi tiết để mô phỏng chính xác ngón tay, đồng thời dùng blendshapes và animation cho biểu cảm khuôn mặt, ánh mắt và thân trên, giúp tăng sắc thái và cảm xúc.

Module chuyển tiếp (co-articulation/blending) xử lý nối giữa các ký hiệu, nội suy vị trí để tạo chuyển động mượt và đồng bộ nhịp điệu, khắc phục vấn đề time-warping.

Rendering và xuất bản có thể dùng clip render sẵn cho tốc độ, hoặc on-the-fly rendering cho tương tác thời gian thực. Các kỹ thuật như caching, LOD và streaming giúp giảm độ trễ và tối ưu tài nguyên.

Cuối cùng, giao diện người dùng UI/UX cần trực quan, hỗ trợ điều khiển cơ bản (play/pause, tốc độ, phân đoạn) và nhiều góc nhìn (close-up, side view, full-body). Các chế độ hỗ trợ như slow-motion, step-by-step hay comparison mode giúp người học dễ theo dõi và so sánh ký hiệu

### *1.3.4 Hạn chế và thách thức khi áp dụng 3D cho ký hiệu*

Việc đưa mô phỏng ba chiều vào trình bày động tác ký hiệu đặt ra nhiều vấn đề cần giải quyết. Các khó khăn này không chỉ nằm ở mặt công nghệ, mà còn liên quan đến việc chuẩn bị tư liệu, tối ưu hiệu suất và khả năng tái hiện chuyển động sao cho tự nhiên. Dù mô hình 3D đem lại ưu điểm trực quan, việc triển khai trong thực tế vẫn còn gặp nhiều giới hạn.

Một trong những trở ngại quan trọng nhất nằm ở khâu xây dựng dữ liệu. Để tạo được một hệ thống hoạt động ổn định và có độ chính xác chấp nhận được, người thực hiện cần thu thập lượng lớn tư liệu được mô tả chi tiết. Các đoạn video phải được chú

Escaneado con CamScanner

giải rõ ràng về nhịp động tác, vị trí bàn tay, cấu hình ngón tay, nét mặt, hướng cơ thể và những tín hiệu bổ trợ khác. Đây là việc tiêu tốn thời gian, công sức và đòi hỏi nhiều bước rà soát, vì chỉ cần một sơ suất nhỏ cũng đủ khiến mô phỏng trở nên sai lệch. Các bộ dữ liệu motion capture hiện có vẫn còn hạn chế. Ví dụ, MC-TRISLAN corpus chỉ mới cung cấp khoảng 18 giờ dữ liệu ký hiệu tiếng Czech với chú giải chi tiết [13], trong khi bộ SignAvatars mới đạt quy mô 70.000 chuỗi chuyển động, 8,34 triệu frames 3D SMPL-X [9]. Điều này cho thấy nguồn dữ liệu cho các ngôn ngữ ký hiệu khác ngoài ASL/BSL vẫn còn rất nhỏ.

Một thách thức khác xuất phát từ hiện tượng khiến người xem khó chịu khi mô hình 3D trông gần giống thật nhưng vẫn còn "lệch". Nếu avatar hướng đến phong cách giống người thật nhưng thiếu độ mượt ở chuyển động hoặc biểu cảm, người sử dụng dễ cảm thấy không thoải mái. Ngược lại, tạo hình quá đơn giản lại làm giảm khả năng truyền tải sắc thái trong ký hiệu. Xác định mức độ "giống thật" phù hợp để vừa dễ quan sát vừa đảm bảo chính xác là một bài toán khó.

Ngoài ra, sự khác biệt lớn giữa cơ chế diễn đạt của ký hiệu và ngôn ngữ nói cũng tác động trực tiếp đến quá trình xây dựng hệ thống. Ký hiệu không tuân theo kiểu chuyển đổi từng từ, mà dùng những quy tắc sắp xếp và rút gọn riêng. Vì vậy, bộ xử lý ngôn ngữ phải điều chỉnh lại thứ tự các động tác, lược bỏ phần thừa và giữ nguyên ý tổng thể. Nếu không có bước này, chuỗi ký hiệu tạo ra sẽ thiếu tự nhiên hoặc truyền đạt sai.

Một yếu tố khác không thể bỏ qua là yêu cầu về năng lực xử lý đồ họa. Việc dựng, tính toán và hiển thị chuyển động ba chiều theo thời gian thực có thể tiêu tốn khá nhiều tài nguyên, đặc biệt trên các thiết bị không có GPU mạnh. Điều này khiến việc triển khai trên máy cấu hình thấp trở nên khó khăn. Do đó, mô phỏng cần được tối ưu để giảm mức tiêu thụ tài nguyên, hạn chế độ trễ và bảo đảm chuyển động trơn tru trong các tình huống phát trực tiếp hoặc tương tác liên tục.

Escaneado con CamScanner

# CHƯƠNG 2: XÂY DỰNG HỆ THỐNG CHUYỂN VĂN BẢN SANG MÔ HÌNH 3D

## 2.1. Xác định yêu cầu xử lý và hướng tiếp cận

Mục tiêu của phần này là làm rõ những yêu cầu kỹ thuật và chức năng cần có để triển khai một hệ thống chuyển văn bản tiếng Việt thành mô phỏng ngôn ngữ ký hiệu dạng video 3D, đồng thời chỉ ra hướng tiếp cận tổng quát (pipeline) tức các bước "đi từ gì sang gì" phù hợp với nguồn lực và phạm vi đề tài. Phần này bao gồm yêu cầu chức năng, yêu cầu dữ liệu, yêu cầu xử lý (pipeline), yêu cầu phi chức năng (hiệu năng, lưu trữ, khả năng mở rộng), đề xuất chiến lược tiếp cận và rủi ro cần quản lý.

### *2.1.1 Yêu cầu chức năng chính*

Hệ thống mô phỏng ký hiệu 3D cần thực hiện những chức năng cốt lõi nhằm đảm bảo tương tác mượt mà và trải nghiệm người dùng thuận tiện. Các khả năng chính bao gồm xử lý dữ liệu đầu vào, chuyển đổi từ ngữ sang video hoặc mô hình mesh 3D, điều khiển hiển thị và quản lý kho ký hiệu.

Văn bản đầu vào có thể là câu, cụm từ hoặc đoạn, được nhập trực tiếp hoặc lấy từ phụ đề đã chuẩn hóa. Hệ thống sẽ phân tích và ánh xạ từng đơn vị từ sang video hoặc mesh 3D tương ứng, dựa trên dữ liệu huấn luyện từ 2D.

Kết quả đầu ra là video ký hiệu 3D hoặc mô hình mesh hiển thị trên trình xem. Video có thể ghép thành câu hoặc đoạn, kèm thông tin metadata để đồng bộ với văn bản.

Quá trình ánh xạ cần đảm bảo chính xác từng từ hoặc cụm từ với video/mesh trong kho dữ liệu đã được đánh nhãn.

Các công cụ điều khiển hiển thị gồm phát/tạm dừng, tua lại, chia đoạn, điều chỉnh tốc độ và chọn góc nhìn. Trình xem video cũng cho phép chỉnh độ sáng, tương

Escaneado con CamScanner

phản, bão hòa, màu sắc/tint và cường độ tint. Người dùng có thể bật hoặc tắt phụ đề đồng thời, hoặc nghe văn bản bằng tính năng Text-to-Speech.

Việc đồng bộ hóa hiển thị bao gồm highlight từ hoặc đoạn văn bản đang ký hiệu, cùng với timecode để dễ theo dõi và định vị nội dung trong video.

Kho dữ liệu hỗ trợ các thao tác thêm, sửa, xóa và lưu trữ thông tin metadata (từ/cụm từ, ID,...). Hệ thống cũng cần cơ chế ghép nối mượt giữa các video/mesh để chuyển động được liền lạc và tự nhiên.

### *2.1.2 Yêu cầu dữ liệu và annotation*




**Hình 2.1 Hình ảnh dataset từ nguồn QIPEDC**

Để tạo ra một nền tảng mô phỏng ký hiệu 3D vận hành ổn định, bộ dữ liệu sử dụng cho mô hình cần được xây dựng một cách cẩn trọng. Toàn bộ quy trình thu thập, tinh chỉnh và sắp xếp dữ liệu giữ vai trò quyết định đối với độ chính xác của hệ thống. Nguồn tư liệu không chỉ phải rõ ràng, chính xác mà còn cần đủ đa dạng, có độ phân giải tốt và dễ mở rộng khi bổ sung thêm ký hiệu mới.

Tập tài liệu thô ban đầu bao gồm các đoạn video 2D, mỗi đoạn thể hiện một đơn vị nghĩa (gloss/từ). Từ các video này, mô hình trích xuất các điểm đặc trưng (keypoints) bằng OpenPose, sau đó tái dựng lại hình dạng ba chiều thông qua SMPLify-X. Tùy thuộc vào từng ký hiệu, có thể tồn tại nhiều phiên bản thể hiện khác

Escaneado con CamScanner

nhau ví dụ khác biệt vùng miền, mức độ nhanh chậm hay biến đổi ở nét mặt nhằm phản ánh đúng sự đa dạng trong sử dụng ký hiệu ngoài đời thực.

Trong nghiên cứu này, phần lớn nguồn tư liệu được lấy từ kho QIPEDC, một bộ dữ liệu đã được chuẩn hoá và phù hợp cho mục đích học tập cũng như nghiên cứu liên quan đến cộng đồng khiếm thính tại Việt Nam.

Về mặt gán nhãn, mỗi video thường chỉ đi kèm tập keypoints 2D để dùng làm dữ liệu đầu vào cho quá trình tái dựng mesh 3D bằng SMPLify-X. Sau khi dựng xong, mô hình được gắn với ký hiệu tương ứng dựa trên tên định danh, ví dụ đặt tên tệp theo gloss (chào.mp4, chào_mesh). Ngoài ra còn có tập tư thế trung gian gồm chuỗi keypoint 2D/3D kèm mốc thời gian, hỗ trợ hệ thống trong các bước ghép chuyển động (blending) và chuyển đổi tư thế giữa các ký hiệu (retargeting) khi hiển thị trên viewer 2D hoặc môi trường xem mô hình 3D.

Metadata của hệ thống được thiết kế đơn giản nhất có thể, chủ yếu dựa vào định danh từng ký hiệu. Cách bố trí này giúp truy xuất file nhanh chóng mà không phải xử lý thêm các lớp nhãn phức tạp.

Bên cạnh yêu cầu về dung lượng, bộ dữ liệu cũng cần đủ rộng để bao phủ đầy đủ vốn từ và các biến thể liên quan. Quá trình đánh giá chất lượng được thực hiện với sự hỗ trợ của người thành thạo ký hiệu, nhằm đảm bảo chuyển động được tái dựng đúng ngữ nghĩa và giữ sự tự nhiên khi mô phỏng. Nội dung được kiểm tra qua hai kiểu hiển thị: video (có thể điều chỉnh ánh sáng, màu, phụ đề hoặc TTS) và mô hình 3D (tùy chọn xoay, phóng to, thay đổi góc quan sát), giúp việc thẩm định và tương tác trở nên trực quan hơn.

### *2.1.3 Yêu cầu xử lý và luồng pipeline hệ thống*

Hệ thống mô phỏng ngôn ngữ ký hiệu 3D cần một pipeline chặt chẽ để chuyển đổi văn bản thành video hoặc mô hình ký hiệu. Các bước phải đảm bảo tính chính xác,

Escaneado con CamScanner

hiệu quả và khả năng mở rộng, từ xử lý dữ liệu đầu vào đến sinh chuyển động và hiển thị kết quả.

Tiền xử lý văn bản chủ yếu tập trung vào chuẩn hóa chữ viết, xử lý dấu câu và tách từ. Đầu ra là danh sách từ/ngữ được chuẩn hóa để ánh xạ trực tiếp sang tập video hoặc mesh đã được huấn luyện.

Trích xuất pose 2D được thực hiện bằng OpenPose, ước lượng keypoints từ video huấn luyện. Các keypoints này bao gồm cơ thể, tay và khuôn mặt, đóng vai trò trung gian cho bước tái tạo 3D.

Chuyển đổi sang pose 3D sử dụng SMPLifyX để fitting keypoints 2D thành mesh 3D hoàn chỉnh. Kết quả là mô hình 3D chi tiết có thể giữ lại dưới dạng mesh hoặc tiếp tục xử lý để sinh video.

Sinh video/mesh: từ mesh 3D, hệ thống có thể render thành video hoặc lưu giữ mesh để hiển thị trong viewer 3D. Điều này cho phép linh hoạt trong việc lựa chọn định dạng đầu ra.

Ánh xạ và hiển thị: các video hoặc mesh được gắn nhãn và ánh xạ trực tiếp với từ/ngữ trong văn bản. Khi hiển thị, hệ thống có thể phát video hoặc cho phép người dùng quan sát mô hình 3D trong viewer.

Quản lý dữ liệu: toàn bộ video và mesh được lưu trữ kèm metadata để đồng bộ giữa văn bản và ký hiệu. Cơ chế caching được áp dụng nhằm đảm bảo truy xuất nhanh chóng trong giao diện người dùng.

### *2.1.4 Yêu cầu phi chức năng*

Bên cạnh các chức năng chính, hệ thống mô phỏng ký hiệu 3D cần đáp ứng những tiêu chí phi chức năng nhằm đảm bảo vận hành ổn định và lâu dài. Các yếu tố này liên quan đến hiệu suất, khả năng mở rộng, bảo trì, lưu trữ và hỗ trợ người dùng, giúp trải nghiệm mượt mà và dễ tiếp cận.

Escaneado con CamScanner

Hiệu năng và độ trễ là điểm then chốt. Tùy thuộc vào nhu cầu sử dụng, hệ thống có thể vận hành theo thời gian thực hoặc xử lý theo lô. Một chiến lược hợp lý là pre-render các ký hiệu phổ biến, đồng thời xử lý động cho các trường hợp hiếm, vừa tối ưu tốc độ vừa giảm hiện tượng tắc nghẽn.

Về lưu trữ, kho video pre-rendered có dung lượng lớn, do đó cần áp dụng cơ chế caching, nén dữ liệu và định dạng hiệu quả để tiết kiệm không gian mà vẫn giữ chất lượng hình ảnh và chuyển động.

Khả năng mở rộng phải cho phép thêm mới từ vựng, biến thể ngữ cảnh hoặc clip một cách dễ dàng. Cấu trúc dữ liệu và API cần được thiết kế linh hoạt để không ảnh hưởng đến hiệu suất khi hệ thống được phát triển thêm.

Bảo trì đòi hỏi các công cụ quản lý video, chú giải và metadata đơn giản, trực quan, giúp việc cập nhật và điều chỉnh thuận tiện, đồng thời duy trì tính nhất quán và ổn định khi mở rộng hệ thống.

Cuối cùng, trợ năng là yếu tố quan trọng. Giao diện cần hỗ trợ phím tắt, chế độ high-contrast, phụ đề đầy đủ và tối ưu cho người khiếm thính, nhằm mang lại trải nghiệm thân thiện và dễ thao tác.

### *2.1.5 Tổng quát, tính khả thi về hướng tiếp cận đề xuất*

Để xây dựng hệ thống mô phỏng ngôn ngữ ký hiệu 3D hiệu quả và dễ duy trì, cần một chiến lược tổng quát, kết hợp giữa pre-render và sinh chuyển động động nhằm tối ưu hiệu suất và đảm bảo linh hoạt.

Chiến lược chủ đạo của hệ thống là hybrid: sử dụng video pre-rendered cho từ vựng phổ biến và mesh 3D cho các từ ít gặp hoặc biến thể ngữ cảnh. Khi không có video phù hợp, hệ thống sẽ sử dụng mô hình 3D của fingerspelling.

Hệ thống được thiết kế với pipeline tự động, bao gồm quá trình tiền xử lý video 2D, ánh xạ keypoints 2D sang mô hình 3D qua SMPLify-X, và kết quả là video 3D hoặc mô hình mesh. Quá trình này tối ưu hóa quy trình và dễ dàng mở rộng khi cần.

Escaneado con CamScanner

Metadata điều khiển trong hệ thống đơn giản và dựa trên việc ánh xạ trực tiếp giữa các video/mesh và từ vựng, với mỗi tệp video hoặc thư mục mesh tương ứng với một từ ký hiệu. Hệ thống sử dụng thông tin này để dễ dàng tìm kiếm và đồng bộ hóa video/mesh với các từ vựng trong giao diện người dùng.

Giao diện người dùng (UI) sẽ đơn giản, gồm textbox nhập văn bản, nút "generate", timeline hiển thị clip, và các điều khiển tốc độ. Hệ thống sẽ có tùy chọn 3D viewer để người dùng có thể trực tiếp quan sát và tương tác với mô hình 3D (xoay, phóng to/thu nhỏ).

Phía backend cung cấp API để ánh xạ văn bản–gloss và truy xuất video/mesh, giúp người dùng dễ dàng tạo và xem nội dung ký hiệu.

### *2.1.6 Các rủi ro chính và biện pháp giảm thiểu*

Quá trình phát triển hệ thống chuyển đổi văn bản sang ký hiệu 3D tiềm ẩn một số rủi ro có thể ảnh hưởng đến chất lượng sản phẩm và trải nghiệm của người dùng, tuy nhiên, các biện pháp phòng ngừa có thể được áp dụng để duy trì độ ổn định và hiệu quả.

Thiếu dữ liệu là một rủi ro quan trọng, đặc biệt khi số lượng video hoặc mô hình 3D được gán nhãn đầy đủ còn hạn chế. Giải pháp là xây dựng tập từ vựng nền tảng kèm theo các clip và mô hình 3D đã được xử lý trước, đảm bảo sẵn sàng sử dụng trong hệ thống.

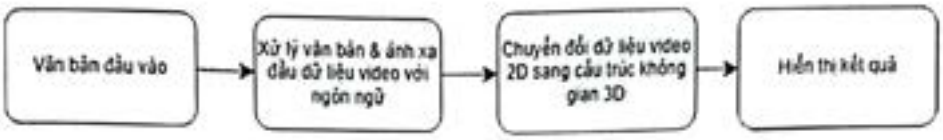
Hiện tượng co-articulation, tức là chuyển động bị gượng khi ghép nối các đoạn video hoặc mô hình 3D, có thể xảy ra. Để khắc phục, hệ thống cần đảm bảo việc nối các phân đoạn này liền mạch, hạn chế giật hoặc bất thường trong chuyển động.

Một thách thức khác là sự thiếu hụt tín hiệu phi-thủ như biểu cảm gương mặt, ánh mắt, vốn rất quan trọng trong giao tiếp ký hiệu. Trong khi hệ thống hiện tại chưa xử lý đầy đủ các yếu tố này, có thể bổ sung phụ đề hoặc sử dụng âm thanh tổng hợp từ văn bản (text-to-speech) để hỗ trợ truyền đạt nghĩa và duy trì khả năng hiểu thông điệp.

Escaneado con CamScanner

Cuối cùng, hiệu suất và dung lượng lưu trữ cũng là vấn đề. Kho dữ liệu lớn có thể dẫn đến chậm trễ và tốn không gian lưu trữ. Giải pháp bao gồm nén video và tổ chức lưu trữ hợp lý, giúp giảm tải cho hệ thống và duy trì khả năng mở rộng trong tương lai.

## 2.2. Quy trình xử lý: từ văn bản đầu vào đến mô hình 3D ký hiệu




**Hình 2.2 Pipeline tóm tắt các bước quy trình xử lý**

Quy trình xử lý trong hệ thống chuyển đổi văn bản sang mô hình 3D ngôn ngữ ký hiệu được thiết kế theo chuỗi các bước liên hoàn, đảm bảo đi từ dữ liệu ngôn ngữ đầu vào (text) đến đầu ra trực quan hóa bằng video 3D ký hiệu. Các bước chính bao gồm:

### *2.2.1 Tiếp nhận và phân tích văn bản đầu vào*




**Hình 2.3 Mô hình quy trình tiếp nhận và tiền xử lý văn bản**

Giai đoạn đầu tiên của hệ thống là thu thập phần nội dung chữ mà người dùng gửi lên qua giao diện hay thông qua API. Khi dữ liệu được tiếp nhận, mô đun tiền xử lý sẽ được kích hoạt nhằm chuẩn bị văn bản trước khi chuyển sang các bước xử lý chuyên sâu hơn.

Trước tiên, hệ thống thực hiện quá trình làm sạch: chuẩn hoá định dạng, loại bỏ ký tự dư thừa, điều chỉnh dấu câu và đưa toàn bộ chuỗi chữ về trạng thái thống nhất, giúp việc phân tích diễn ra trôi chảy. Tiếp theo, quy trình tách từ được áp dụng để phân

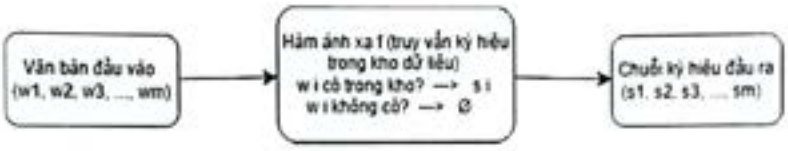
Escaneado con CamScanner

rã câu thành các đơn vị ngôn ngữ riêng rẽ. Đây là thao tác quan trọng đối với tiếng Việt, nơi nhiều cụm từ có thể đi liền nhau mà không có dấu tách rõ ràng. Song song với đó, những thành phần không mang giá trị ngữ nghĩa hoặc gây nhiễu cũng được loại bỏ để giảm bớt khối lượng xử lý cho các mô đun phía sau.

Mục tiêu của bước này là khoanh vùng tập từ vựng cốt lõi sẽ được chuyển sang ký hiệu, từ đó hỗ trợ quá trình ánh xạ diễn ra mạch lạc và ít sai lệch hơn. Khi khâu tiền xử lý được thực hiện tốt, chất lượng dữ liệu đầu vào được cải thiện đáng kể, góp phần tăng độ tin cậy cho toàn bộ quy trình chuyển đổi tiếp theo.

### *2.2.2 Ánh xạ từ văn bản sang ký hiệu*

Giai đoạn tiếp theo trong quy trình xử lý là chuyển nội dung chữ viết thành các đơn vị biểu đạt bằng ký hiệu. Việc này được thực hiện dựa trên kho tư liệu ký hiệu thu thập từ bộ video QIPEDC.




**Hình 2.4 Biểu đồ ánh xạ từ với dữ liệu từ ký hiệu**

Tại bước này, từng từ hoặc cụm từ trong câu sẽ được dò tìm trong kho dữ liệu để xác định ký hiệu tương ứng. Nếu áp dụng ánh xạ một–một, mỗi từ $w_i$ trong câu đầu vào $Sentence = (w_1, w_2, \dots, w_m)$ sẽ trực tiếp ánh xạ thành ký hiệu $s_i$ thông qua hàm ánh xạ f:

$$s_i = f(w_i),\ i = 1 \dots m \tag{2.1}$$

Trong trường hợp từ $w_i$ không có ký hiệu tương ứng trong kho dữ liệu, ta đánh dấu bằng Ø:

$$f(w_i) = \emptyset \tag{2.2}$$

Ở mức tổng thể, câu văn bản gốc sẽ được chuyển thành chuỗi ký hiệu:

Escaneado con CamScanner

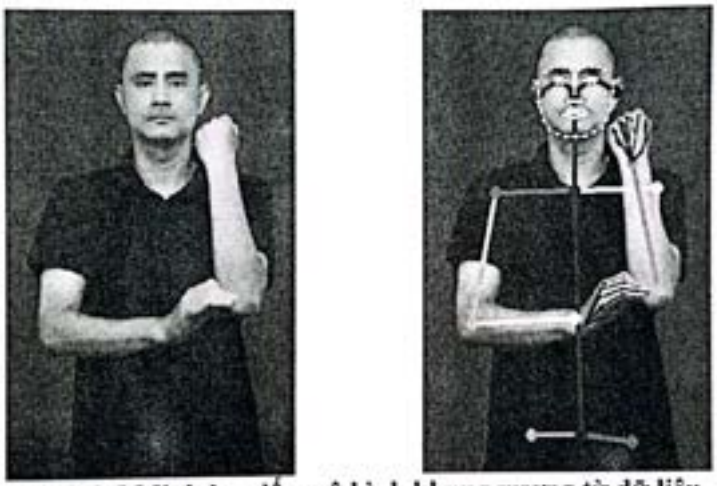
$$SignSequence = (s_1, s_2, \dots, s_m) \quad (2.3)$$

Trong đó $W = \{w_1, w_2, \dots, w_m\}$ là tập từ trong câu đầu vào, còn $S = \{s_1, s_2, \dots, s_n\}$ là tập các ký hiệu trong kho dữ liệu.

Khi nội dung không thể diễn đạt chính xác theo từng từ riêng lẻ, toàn bộ cụm từ hoặc câu được xem như một khối và ánh xạ thành chuỗi ký hiệu phù hợp với cách diễn đạt đặc trưng của ngôn ngữ ký hiệu. Phương pháp này giúp kết quả hiển thị tự nhiên, tránh sai lệch ý nghĩa và đảm bảo mạch lạc khi trình bày.

Sau bước xử lý, hệ thống tạo ra một dãy ký hiệu được sắp xếp theo trình tự thể hiện. Chuỗi này là cơ sở để mô phỏng lại nội dung bằng ký hiệu theo đúng ý nghĩa và đảm bảo sự mạch lạc khi hiển thị.

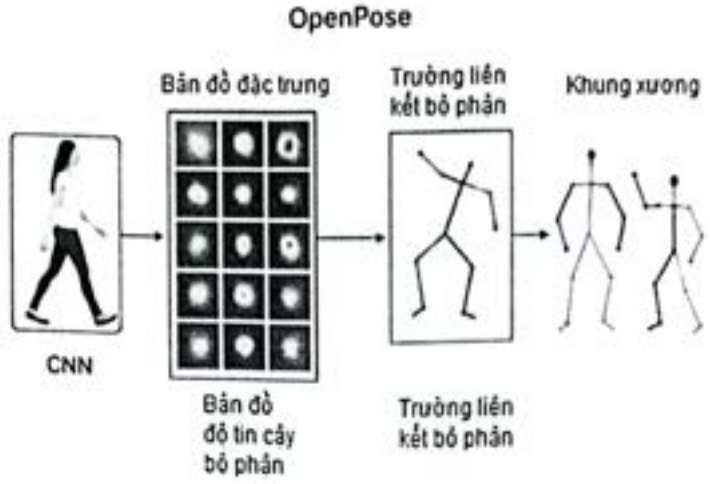
### *2.2.3 Chuyển ký hiệu sang mô hình khung xương*



**Hình 2.5 Minh họa lấy mô hình khung xương từ dữ liệu**

Bước tiếp theo trong quy trình là biến các ký hiệu thành mô hình khung xương (skeleton/pose). Mỗi ký hiệu được biểu diễn thông qua dữ liệu chuyển động, được trích xuất từ các video ký hiệu đã có trong tập dữ liệu chuẩn.

Escaneado con CamScanner




**Hình 2.6 Mô hình hoạt động của OpenPose**

OpenPose là hệ thống ước lượng tư thế người theo thời gian thực, có thể phát hiện nhiều người trong ảnh hoặc video và trích xuất các khớp (keypoints) 2D của toàn thân, tay, mặt, chân. Mỗi người được biểu diễn bằng một skeleton gồm các tọa độ khớp cùng độ tin cậy (confidence score), dữ liệu này có thể dùng cho rigging hoặc mô phỏng 3D.

Pipeline hoạt động như sau: ảnh hoặc frame video được đưa vào CNN để trích xuất feature maps, sau đó chia thành hai nhánh song song: tạo Confidence Maps (heatmaps xác suất khớp) và Part Affinity Fields (vector chỉ hướng kết nối giữa các khớp). Các output này được refine qua nhiều stage để tăng độ chính xác, sau đó áp dụng thuật toán ghép khớp (bipartite matching) để nối các khớp thành skeleton cho từng người. Kết quả cuối cùng là danh sách khớp với tọa độ 2D và skeleton hoàn chỉnh, sẵn sàng dùng cho animation hoặc dựng mô hình 3D.

Về mặt kỹ thuật, bước này sử dụng các công cụ nhận dạng hình thể, chẳng hạn OpenPose, để xác định tọa độ các khớp xương từ từng khung hình của video. Với mỗi khớp (j) tại khung hình (f), OpenPose sinh ra một cặp tọa độ 2D:

$$s_{j(f)} = \left(x_j^f, y_j^f\right), j = 0, \ldots, N-1 \qquad (2.4)$$

Escaneado con CamScanner

Trong đó (N) là số lượng khớp của mô hình (ví dụ BODY_25 có (N = 25)). Các tọa độ này thường được chuẩn hoá để loại bỏ ảnh hưởng của kích thước khung hình hoặc vị trí người ký hiệu trong video:

$$x'_{j(f)} = \frac{\left(x_j^f - \mu_x\right)}{\sigma_x},\ y'_{j(f)} = \frac{y_j^f - \mu_y}{\sigma_y} \tag{2.5}$$

Với $\mu_x, \mu_y, \sigma_x, \sigma_y$ lần lượt là giá trị trung bình và độ lệch chuẩn của toàn bộ toạ độ khớp trong video.

Đối với các hệ thống tái dựng 3D (SMPLify-X hoặc các mô hình tương tự), quá trình fitting 3D có thể mô tả bằng hàm tối ưu hóa để tìm bộ tham số pose/shape $\theta$ sao cho khi chiếu khớp 3D xuống 2D, chúng khớp với điểm 2D từ OpenPose:

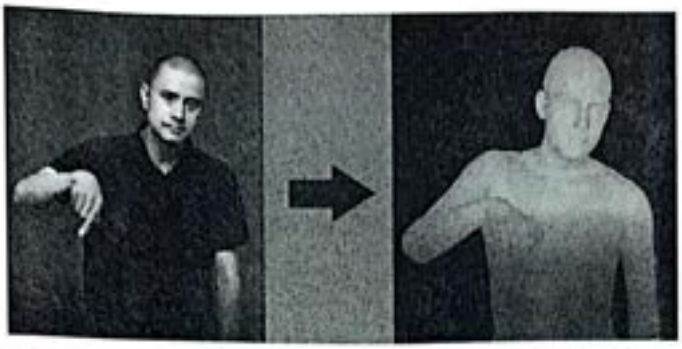
$$\min_\theta L(\theta) = \sum_{j=0}^{N-1} \left\| Proj\left(J_j(\theta)\right) - s_j(f)^2 \right\| + \lambda\, R(\theta) \tag{2.6}$$

Trong đó $J_j(\theta)$ là tọa độ khớp 3D dự đoán, $Proj$ là phép chiếu phối cảnh, còn $R(\theta)$ là regularization giúp giữ cho chuyển động tự nhiên.

Chuỗi tọa độ này mô tả đầy đủ cử chỉ và động tác của người ký hiệu, cho phép tạo ra một mô hình động học trung thực cho từng ký hiệu. Kết quả là tập dữ liệu skeleton 2D hoặc 3D, tùy theo mức độ chi tiết cần thiết của hệ thống, tương ứng với từng ký hiệu đã được xác định ở bước ánh xạ trước đó.

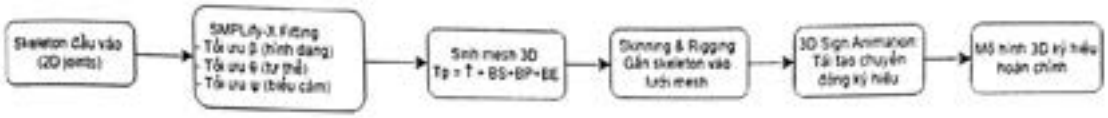
### *2.2.4 Xây dựng mô hình 3D ký hiệu*

Escaneado con CamScanner



**Hình 2.7 Minh họa kết quả chuyển đổi video 2D sang mô hình 3D**

Đến bước thứ tư trong quy trình, ta xây dựng mô hình 3D ký hiệu thông qua quá trình mesh animation. Sau khi thu được chuỗi skeleton từ bước trước, bước này tập trung vào việc chuyển đổi dữ liệu khớp xương thô thành một mô hình 3D sống động.




**Hình 2.8 Chuỗi chuyển dạng từ dữ liệu xương 2D sang mô hình 3D của SMPLify-X**

Tính đến kỹ thuật, quá trình này áp dụng SMPLify-X, một mô hình parametric thống kê, để chuyển từ skeleton 2D/3D sang mesh 3D. Mô hình được định nghĩa bởi hàm:

$$M(\beta,\theta,\psi): \mathbb{R}^{|\beta|} \times \mathbb{R}^{|\theta|} \times \mathbb{R}^{|\psi|} \rightarrow \mathbb{R}^{3N} \tag{2.7}$$

Trong đó:

- $\beta$ là tham số hình dạng vector
- $\theta$ là tham số tư thế vector
- $\psi$ là biểu thức vecto/tham số khuôn mặt
- $N$ là số đỉnh của lưới

Mesh được biến đổi theo công thức:

$$T_{p(\beta,\theta,\psi)} = \bar{T} + BS(\beta) + BP(\theta) + BE(\psi) \tag{2.8}$$

Escaneado con CamScanner

Với là template mesh trung bình, BS(β) là shape blend-shape, BP(θ) là pose blend-shape, và BE(ψ) là expression blend-shape. Mesh cuối cùng được trích xuất theo:

$$M(\beta,\theta,\psi) = W\left(T_{p(\beta,\theta,\psi)}, J(\beta), \theta, W_{weights}\right) \quad (2.9)$$

Để khớp mesh với dữ liệu skeleton đầu vào, SMPLify-X tối ưu bộ tham số (β,θ,ψ) bằng cách minim hóa hàm mất mát tổng quát:

$$E(\beta,\theta,\psi) = E_J + \lambda_{\{\theta_b\}E_{\{\theta_b\}}} + \lambda_{\{pose\}E_{\{pose_{prior}\}}} + \lambda_{\{shape\}E_{\{shape_{prior}\}}} + \cdots$$

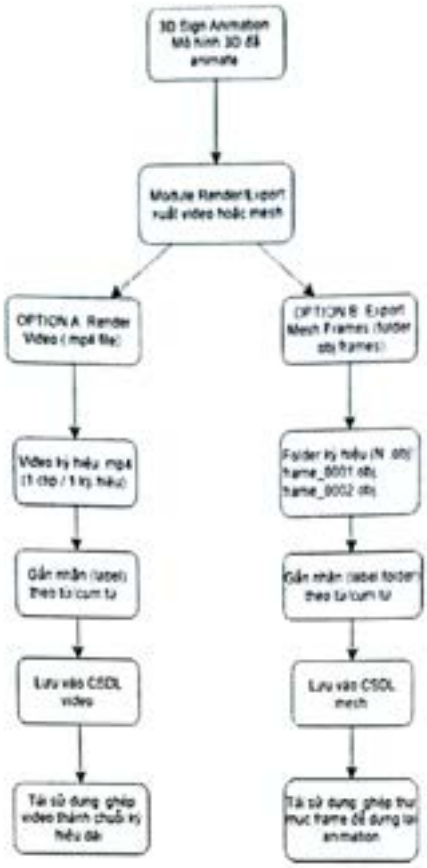
(2.10)

Trong đó $E_J$ là reprojection loss giữa khớp 3D chiếu xuống 2D và keypoints thực từ skeleton, các thành phần còn lại là priors để đảm bảo mesh và pose tự nhiên.

Sau khi tham số được tối ưu, quá trình rigging gắn skeleton vào mesh 3D được thực hiện, tiếp theo là animation để tái tạo chuyển động ký hiệu mượt mà. Mục tiêu cuối cùng là biến dữ liệu chuyển động thô thành hình ảnh trực quan, giúp người xem dễ dàng hiểu và nhận biết các ký hiệu trong ngôn ngữ ký hiệu. Đây là bước quan trọng, tạo nền tảng cho việc mô phỏng ký hiệu trong môi trường 3D.

### *2.2.5 Kết xuất và lưu trữ video ký hiệu 3D*

Escaneado con CamScanner



**Hình 2.9 Mô hình trích xuất mô hình 3D & lưu trữ video ký hiệu 3D**

Ở bước thứ năm, ta kết xuất và lưu trữ video ký hiệu 3D. Sau khi các ký hiệu đã được hoạt họa trên mô hình 3D, chúng sẽ được kết xuất (render) thành các video ngắn, mỗi video đại diện cho một ký hiệu cụ thể. Quá trình kết xuất đảm bảo rằng các chuyển động của ký hiệu được hiển thị mượt mà và rõ ràng, giúp người xem dễ dàng nhận diện.

Các video này sau đó sẽ được lưu trữ trong cơ sở dữ liệu, với mỗi video được gắn nhãn cụ thể theo từ hoặc cụm từ tương ứng trong văn bản. Việc gắn nhãn này giúp cho hệ thống dễ dàng tra cứu và truy xuất các video khi cần thiết.

Điều quan trọng ở bước này là đảm bảo khả năng tái sử dụng các video đã kết xuất. Các video nhỏ có thể được ghép nối lại để tạo thành chuỗi ký hiệu dài hơn, phục vụ cho việc mô phỏng các câu văn bản phức tạp. Việc lưu trữ và tổ chức hợp lý các

Escaneado con CamScanner

video này là yếu tố quan trọng trong việc tối ưu hóa hiệu suất và khả năng mở rộng của hệ thống.

### *2.2.6 Tích hợp vào giao diện hiển thị*

Bước thứ sáu của quy trình tập trung vào việc kết nối hệ thống với giao diện trình chiếu. Khi người dùng nhập văn bản, hệ thống đầu tiên phân tích và chia nội dung thành các từ hoặc cụm từ. Việc này giúp nhận diện rõ ràng cấu trúc câu và chuẩn bị cho bước ánh xạ ký hiệu chính xác.

Tiếp theo, hệ thống tìm kiếm các video 3D tương ứng với từng từ hoặc cụm từ trong kho dữ liệu. Những video này đã được chuẩn bị trước đó và sẽ phục vụ để tái hiện chuyển động ký hiệu phù hợp với từng phần nội dung.

Cuối cùng, các video 3D được ghép nối theo đúng trình tự để tạo thành một chuỗi động tác liền mạch. Quá trình này đảm bảo các ký hiệu được thể hiện mượt mà, chính xác và phản ánh đầy đủ nội dung của câu gốc. Kết quả là một video 3D trực quan, giúp người xem, đặc biệt là người khiếm thính hoặc người học ký hiệu, dễ dàng theo dõi và nắm bắt thông tin.

## 2.3. Thiết kế kiến trúc hệ thống và các thành phần chức năng chính

Để triển khai một hệ thống chuyển văn bản thành mô hình 3D ngôn ngữ ký hiệu, cần thiết kế một kiến trúc tổng thể bao gồm nhiều thành phần chức năng liên kết với nhau. Kiến trúc này vừa đảm bảo tính mô-đun hóa (modularity), giúp dễ mở rộng và bảo trì, vừa đảm bảo luồng xử lý được diễn ra tuần tự từ đầu vào văn bản đến đầu ra video 3D ký hiệu.

### *2.3.1. Kiến trúc tổng thể*

Escaneado con CamScanner




**Hình 2.10 Mô hình kiến trúc tổng thể của hệ thống**

Toàn bộ hệ thống được tổ chức thành bốn lớp chức năng chính, mỗi lớp đảm nhiệm một phần việc trong chuỗi xử lý chuyển nội dung tiếng Việt sang chuyển động ký hiệu dạng 3D. Các khối này hoạt động liên tục, bổ trợ cho nhau nhằm tạo ra chuỗi ký hiệu có độ chính xác cao, đáp ứng nhu cầu sử dụng của người khiếm thính cũng như những người đang học ngôn ngữ ký hiệu.

Lớp đầu tiên là bộ phận tiếp nhận và xử lý dữ liệu đầu vào (Input Processing Layer). Tại đây, văn bản người dùng cung cấp được đưa vào quy trình tiền xử lý ngôn ngữ: làm sạch dữ liệu, chuẩn hoá cách viết, tách từ và loại bỏ những thành phần gây nhiễu. Sau giai đoạn này, hệ thống thu được danh sách từ/cụm từ đã qua tinh chỉnh, phù hợp cho bước chuyển đổi sang ký hiệu.

Escaneado con CamScanner

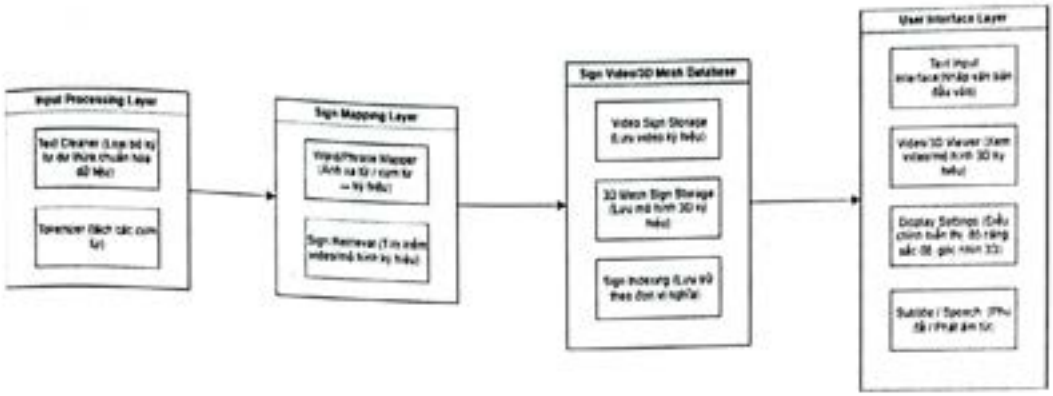
Kế tiếp là lớp chuyển đổi từ vựng sang ký hiệu (Sign Mapping Layer). Giai đoạn này đảm nhiệm việc đối chiếu từng đơn vị ngôn ngữ với mục tương ứng trong kho dữ liệu ký hiệu. Kho dữ liệu có thể bao gồm video, mô hình 3D hoặc các dạng biểu diễn động tác khác. Mỗi mục trong kho đóng vai trò như một "đơn vị ký hiệu" mà hệ thống lựa chọn để ghép lại thành chuỗi đầy đủ. Độ chính xác của bước này đặc biệt quan trọng, nhất là với những từ đa nghĩa hoặc các câu có cấu trúc dài và phức tạp.

Sau khi hoàn thành việc ánh xạ, quá trình được chuyển sang lớp mô phỏng (3D Simulation Layer). Nếu hệ thống sử dụng dạng video, nhiệm vụ chính là phát lại nội dung đã dựng sẵn. Trong trường hợp dùng mô hình 3D, nhân vật ảo được tải vào môi trường trình diễn và thực hiện chuỗi động tác tương ứng. Do dữ liệu đã được chuẩn hóa trước đó, giai đoạn này chủ yếu tập trung vào việc tái hiện ký hiệu một cách mượt mà.

Lớp cuối cùng là phần giao diện và hiển thị (Visualization & UI Layer) – nơi người dùng tương tác trực tiếp. Tại đây, người dùng có thể nhập nội dung, quan sát chuỗi ký hiệu được tạo ra, đồng thời điều chỉnh cách thức trình diễn. Với dạng video, hệ thống cung cấp các công cụ điều khiển như phát, tạm dừng hoặc thay đổi thông số hiển thị (độ sáng, độ tương phản hay sắc độ). Nếu sử dụng mô hình 3D, người dùng có thể xoay nhân vật, thay đổi khoảng cách nhìn hoặc điều chỉnh góc quan sát để xem rõ từng động tác. Hệ thống cũng có thể bổ sung phụ đề hoặc đọc lại văn bản nhằm hỗ trợ việc tiếp nhận thông tin.

### *2.3.2. Các thành phần chức năng chính*

Escaneado con CamScanner



**Hình 2.11 Sơ đồ các thành phần chức năng chính**

Hệ thống chuyển đổi nội dung văn bản sang biểu đạt ngôn ngữ ký hiệu dưới dạng 3D được xây dựng từ nhiều module, mỗi module đảm nhiệm một vai trò riêng để xử lý dữ liệu và hiển thị mô phỏng phù hợp với đầu vào của người dùng.

Khối xử lý ban đầu (Text Preprocessor) đảm nhiệm việc tinh chỉnh văn bản đầu vào. Giai đoạn này bao gồm làm sạch ký tự dư thừa, chuẩn hoá cú pháp và tách văn bản thành các đơn vị nhỏ hơn. Khi bước này hoàn tất, dữ liệu sẽ được chuyển sang quy trình ánh xạ.

Tiếp theo, module chuyển đổi từ sang ký hiệu (Sign Mapper) đối chiếu từng đơn vị từ hoặc cụm từ với mô phỏng tương ứng trong hệ thống. Việc đối chiếu này dựa trên kho dữ liệu đã xây dựng trước, bao gồm các đoạn video ký hiệu hoặc các mô hình 3D được tạo sẵn cho từng từ vựng.

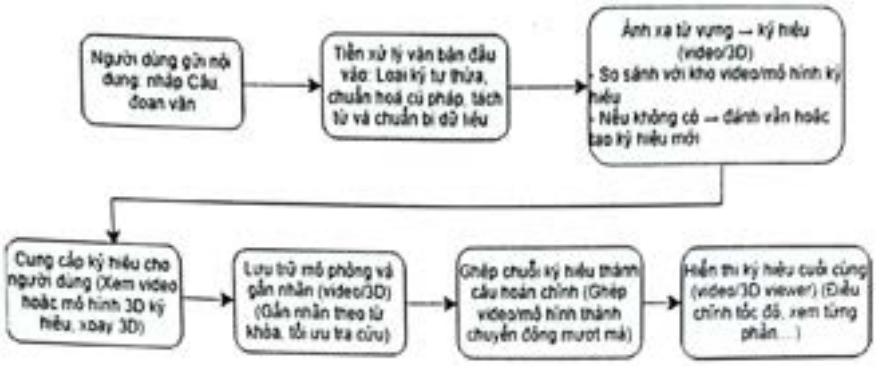
Kho lưu trữ mô phỏng (Sign Video/3D Mesh Database) là nơi hệ thống tập hợp các video ký hiệu và mô hình 3D được dựng từ dữ liệu 2D thông qua SMPLify-X. Tất cả được sắp xếp theo từng đơn vị nghĩa để bộ ánh xạ có thể truy xuất nhanh chóng khi cần.

Module giao tiếp với người dùng (User Interface) là khu vực mà người dùng thao tác trực tiếp. Tại đây, người dùng nhập văn bản, theo dõi kết quả mô phỏng ký hiệu 3D và tuỳ chỉnh nhiều thông số hiển thị như độ sáng, độ tương phản hoặc sắc độ

Escaneado con CamScanner

của video. Ngoài ra còn hỗ trợ phụ đề, phát âm từ văn bản, và trình xem 3D giúp người dùng xoay, phóng to – thu nhỏ hoặc quan sát chi tiết mô hình một cách linh hoạt.

### *2.3.3. Luồng xử lý trong kiến trúc*




**Hình 2.12 Kiến trúc luồng xử lý của hệ thống**

Quá trình vận hành của hệ thống chuyển chữ viết thành ký hiệu 3D diễn ra theo một chuỗi bước liên tục, mỗi bước đảm nhiệm một chức năng riêng, góp phần tạo ra kết quả cuối cùng có độ tin cậy cao và hoạt động ổn định.

Ban đầu, người dùng gửi nội dung cần chuyển đổi thông qua giao diện hoặc API. Dữ liệu đầu vào có thể là câu đơn, đoạn văn hay tập hợp thuật ngữ tùy trường hợp sử dụng.

Khi văn bản được tiếp nhận, mô đun tiền xử lý được kích hoạt, gồm các tác vụ như loại bỏ ký tự dư thừa, chỉnh quy tắc dấu câu và thực hiện tách từ. Sau giai đoạn này, hệ thống thu được tập từ đã được làm sạch và chuẩn hóa để phục vụ quá trình ánh xạ.

Ở bước kế tiếp, từng đơn vị từ vựng hoặc cụm từ được so sánh với kho video ký hiệu và mô hình 3D có sẵn. Nếu không tìm thấy bản tương ứng, hệ thống có thể chuyển sang phương án đánh vần bằng bảng chữ cái hoặc tạo ký hiệu thay thế.

Escaneado con CamScanner

Khi việc đối chiếu hoàn tất, hệ thống cung cấp video hoặc mô hình 3D tương ứng cho người sử dụng. Tại đây, người xem có thể xoay mô hình, thu phóng hoặc thay đổi góc nhìn để quan sát chi tiết hơn.

Các mô phỏng này được lưu trữ lại và gắn nhãn theo từ khóa tương ứng, giúp việc tra cứu và tái sử dụng trong các lần xử lý sau trở nên nhanh và nhẹ hơn.

Ở giai đoạn cuối, những đoạn video hoặc chuỗi chuyển động rời rạc được ghép thành câu ký hiệu hoàn chỉnh với chuyển động mượt mà. Người dùng có thể xem toàn bộ phần ký hiệu dưới dạng video hoặc trực tiếp trong trình xem 3D, kèm theo các tuỳ chọn như điều chỉnh tốc độ phát hoặc xem từng phần nhỏ để phục vụ mục đích nghiên cứu chi tiết.

## 2.4. Thiết kế giao diện người dùng và kết nối với hệ thống mô phỏng

Trong một hệ thống chuyển đổi văn bản thành ngôn ngữ ký hiệu 3D, giao diện người dùng (UI) đóng vai trò cầu nối giữa người dùng cuối (người khiếm thính, giáo viên, hoặc người học ngôn ngữ ký hiệu) và các thành phần xử lý ngôn ngữ, mô phỏng 3D. Một thiết kế giao diện hợp lý không chỉ giúp việc nhập liệu và quan sát kết quả trở nên trực quan, mà còn nâng cao khả năng tiếp cận và ứng dụng thực tiễn của hệ thống.

### *2.4.1. Nguyên tắc thiết kế*

Khi phát triển giao diện cho hệ thống chuyển văn bản sang hoạt cảnh ký hiệu 3D, việc tuân theo các tiêu chí thiết kế phù hợp là rất cần thiết để người dùng thao tác thuận tiện và có trải nghiệm trơn tru. Một điểm then chốt là bố cục phải rõ ràng, dễ định hướng. Các khu vực chức năng nên được tổ chức gọn gàng, ưu tiên các nhiệm vụ chính như nhập nội dung và xem kết quả mô phỏng, giúp người dùng nhanh chóng nắm được cách sử dụng mà không bị quá tải thông tin.

Ngoài sự tinh giản, giao diện cũng cần tính đến nhu cầu của nhiều nhóm người khác nhau. Thiết kế phải thân thiện với người khiếm thính, người giảm thính lực cũng

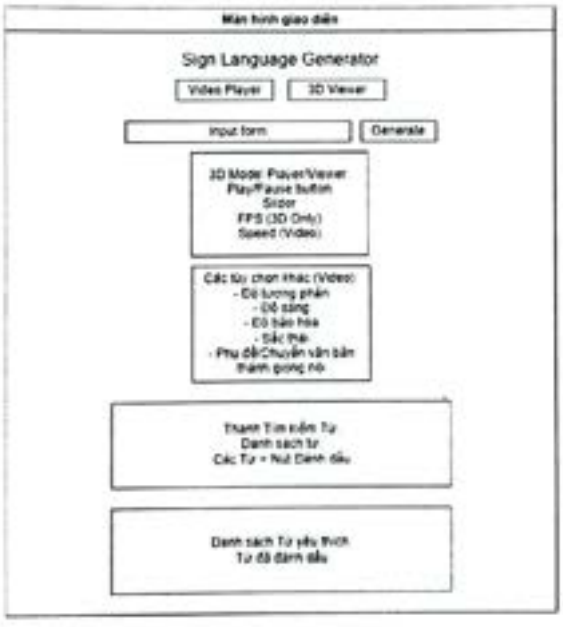
Escaneado con CamScanner

như những người đang học ngôn ngữ ký hiệu. Điều này đòi hỏi lựa chọn bảng màu dễ nhìn, phông chữ dễ theo dõi và cách sắp xếp các thành phần sao cho người dùng có thể quan sát và tương tác thoải mái. Việc tối ưu các yếu tố này giúp mở rộng khả năng tiếp cận và cải thiện trải nghiệm cho đa dạng đối tượng.

Một nguyên tắc quan trọng nữa là tốc độ xử lý. Khi người dùng nhập vào một đoạn văn, hệ thống cần tạo ra mô phỏng 3D trong thời gian ngắn để đảm bảo thao tác không bị ngắt quãng. Nếu phản hồi chậm, người dùng dễ cảm thấy gián đoạn hoặc khó chịu.

Bên cạnh đó, giao diện phải hoạt động ổn định trên nhiều nền tảng khác nhau. Việc triển khai dưới dạng trang web hoặc ứng dụng độc lập giúp người dùng có thể sử dụng trên máy tính, máy tính bảng hoặc điện thoại. Khả năng thích ứng này nâng cao tính linh hoạt và đặc biệt hữu ích cho cộng đồng khiếm thính cũng như người học ký hiệu trong nhiều bối cảnh sử dụng.

### *2.4.2. Các thành phần chính của giao diện*




**Hình 2.13 Thiết kế giao diện chính của hệ thống**

Escaneado con CamScanner

Giao diện hệ thống chuyển đổi văn bản sang ngôn ngữ ký hiệu 3D gồm nhiều thành phần quan trọng để tạo trải nghiệm mượt mà và dễ sử dụng. Khung nhập văn bản là nơi người dùng nhập hoặc dán nội dung cần chuyển đổi, đóng vai trò khởi đầu cho toàn bộ quá trình.

Vùng hiển thị mô phỏng 3D trình bày các ký hiệu được sinh ra từ văn bản. Người dùng có thể tương tác qua các điều khiển cơ bản như phát/tạm dừng, tua lại, hoặc xem từng ký hiệu riêng lẻ. Ngoài ra, hệ thống sẽ tích hợp tùy chọn 3D viewer cho phép trực tiếp quan sát và xoay, phóng to/thu nhỏ mesh 3D được render, mang lại trải nghiệm trực quan và sinh động hơn so với chỉ xem video.

Thanh điều khiển và tùy chọn cho phép điều chỉnh tốc độ phát (chậm, bình thường, nhanh), thay đổi góc nhìn (nếu hỗ trợ) hoặc bật/tắt phụ đề song song. Các chức năng này giúp người học theo dõi và hiểu ký hiệu dễ dàng hơn.

Khu vực hỗ trợ hiển thị danh sách từ/cụm từ tiếng Việt cùng ký hiệu tương ứng, giúp người học đối chiếu và tra cứu nhanh, tăng hiệu quả trong việc tiếp cận ngôn ngữ ký hiệu.

### *2.4.3. Kết nối với hệ thống mô phỏng*

Để phần giao diện có thể trình bày kết quả nhanh và chính xác, việc thiết lập một cơ chế trao đổi dữ liệu ổn định giữa UI và bộ xử lý phía sau là yếu tố then chốt. Cách kết nối này giúp toàn bộ quy trình vận hành trơn tru, đồng thời bảo đảm người sử dụng nhận được phản hồi gần như ngay lập tức sau khi cung cấp nội dung văn bản.

Khi người dùng nhập dữ liệu vào màn hình giao tiếp, thông tin được chuyển tới bộ phận xử lý ban đầu. Tại bước này, văn bản được làm sạch, chuẩn hoá cấu trúc, tách thành các đơn vị từ và phân tích ngữ cảnh nhằm tạo ra chuỗi ký hiệu phù hợp. Đầu ra của giai đoạn này là tập ký hiệu đã được mã hoá để phục vụ những bước chuyển đổi tiếp theo.

Escaneado con CamScanner

Ngay sau đó, giao diện tiến hành gửi yêu cầu đến hệ thống lưu trữ để lấy về các video hoặc mô hình 3D tương ứng với chuỗi ký hiệu vừa được tạo thành. Kho dữ liệu này chứa những đoạn mô phỏng ngắn được tạo sẵn cho từng ký hiệu riêng biệt. Các mô phỏng được tải về và phát theo đúng trình tự ký hiệu nhằm tạo trải nghiệm quan sát liền mạch và dễ hiểu cho người dùng.

Ở bước kế tiếp, bộ xử lý hiển thị (Rendering/Playback Engine) phụ trách kết hợp các video mô phỏng đơn lẻ thành một đoạn trình bày hoàn chỉnh. Toàn bộ kết quả được phát trực tiếp trên giao diện dưới dạng video liên tục, hỗ trợ người xem nắm bắt ý nghĩa nội dung văn bản sau khi chuyển sang dạng ngôn ngữ ký hiệu.

### *2.4.4 Lợi ích và ý nghĩa*

Việc xây dựng phần giao diện kết hợp với cơ chế kết nối đến bộ mô phỏng ba chiều giúp toàn bộ hệ thống vận hành gọn gàng hơn, đồng thời cung cấp cho người dùng một cách thức tương tác rõ ràng và thân thiện. Giao diện đóng vai trò như lớp công cụ trung gian, giúp người sử dụng dễ dàng tiếp cận chức năng của hệ thống và tạo nên giá trị thực tiễn trong nhiều bối cảnh làm việc.

Đối với người khiếm thính, giải pháp này tạo ra một phương tiện tiếp nhận nội dung dựa trên cử chỉ trực quan. Thông qua việc chuyển văn bản sang các động tác mô phỏng, họ có thể nắm bắt ý chính một cách nhanh chóng mà không phải phụ thuộc vào hình thức hỗ trợ bổ sung. Cách tiếp cận này giúp giảm bớt trở ngại trong trao đổi thông tin, đồng thời hỗ trợ nâng cao mức độ hòa nhập của nhóm người dùng này.

Không chỉ hỗ trợ việc tiếp nhận thông tin, hệ thống còn phù hợp cho mục đích giảng dạy và tự rèn luyện. Người học hoặc người dạy có thể xem đây như một công cụ giúp minh họa, làm quen hoặc củng cố kiến thức về ký hiệu. Nhờ khả năng thay đổi tốc độ hiển thị và tùy chỉnh cách xem, người học có thể điều chỉnh quá trình luyện tập cho phù hợp với khả năng cá nhân, tạo ra trải nghiệm học linh hoạt hơn.

Escaneado con CamScanner

Bên cạnh đó, hướng tiếp cận này có tiềm năng được triển khai trong các cơ sở giáo dục, nền tảng truyền thông hoặc những dịch vụ hỗ trợ cộng đồng. Khi nhu cầu tiếp cận thông tin của người khiếm thính ngày càng được quan tâm, các công nghệ chuyển đổi dựa trên mô phỏng ký hiệu sẽ có cơ hội ứng dụng rộng rãi hơn trong những lĩnh vực liên quan đến xã hội, giảng dạy và truyền tải nội dung.

Escaneado con CamScanner

# CHƯƠNG 3: THỬ NGHIỆM MÔ HÌNH CHUYỂN ĐỔI VĂN BẢN SANG KÝ HIỆU 3D

## 3.1. Quy trình xử lý đầu vào và xây dựng mô phỏng chuyển động 3D

Quy trình này nhằm biến văn bản tiếng Việt đầu vào thành chuỗi ký hiệu ngôn ngữ được mô phỏng bằng hình ảnh 3D. Đây là bước quan trọng vì nó tạo ra cầu nối giữa ngôn ngữ viết và hình thức biểu đạt trực quan mà người khiếm thính có thể tiếp nhận. Việc xây dựng mô phỏng chuyển động 3D đòi hỏi nhiều giai đoạn liên tiếp, mỗi giai đoạn xử lý một dạng dữ liệu khác nhau (từ chữ sang ký hiệu sang chuyển động sang video 3D).

### *3.1.1. Các bước xử lý chính*

Giai đoạn đầu tiên của quy trình tập trung đưa văn bản đầu vào về trạng thái dễ xử lý. Khi người dùng nhập câu tiếng Việt, hệ thống tiến hành tách nội dung thành từng lớp nhỏ như câu, cụm ý và đơn vị từ, giúp việc phân tích cú pháp và ngữ nghĩa được thực hiện có trật tự.

Sau khi tách lớp, các thành phần này được đối chiếu với bộ từ điển ký hiệu đã được hệ thống xây dựng trước đó. Mục đích là tìm ra những phần có thể chuyển sang ký hiệu ngay lập tức. Với những mục không khớp dữ liệu, hệ thống chọn hướng thay thế như dùng biểu đạt gần nghĩa hoặc lựa chọn ký hiệu tương đương để giữ nguyên thông điệp gốc mà không làm sai lệch nội dung.

Kết quả của giai đoạn chuẩn hóa là một tập hợp đơn vị ngôn ngữ đã được xử lý đồng nhất, đóng vai trò lớp trung gian trước khi chuyển sang các bước tạo ký hiệu. Chính lớp dữ liệu này giúp các mô đun tiếp theo hoạt động ổn định và giảm rủi ro sai khác.

Bước kế tiếp là gán từng từ hoặc nhóm từ vào các ký hiệu tương ứng dựa trên bộ dữ liệu chuẩn như QIPEDC. Vì ký hiệu không đi theo cấu trúc ngữ pháp của tiếng

Escaneado con CamScanner

Việt, đây là bước đặc biệt quan trọng để đảm bảo cách thể hiện phù hợp với quy tắc của ngôn ngữ ký hiệu.

Sau khi việc gán hoàn thành, hệ thống sẽ điều chỉnh lại toàn bộ chuỗi ký hiệu để phù hợp với thứ tự diễn đạt của ký hiệu, tạo ra mạch trình bày tự nhiên, cho dù vị trí xuất hiện của các ký hiệu có thể không giống câu văn ban đầu.

Kết quả thu được là một chuỗi ký hiệu ở dạng trung gian, đóng vai trò nguồn đầu vào cho quá trình dựng mô phỏng 3D nhằm tái hiện lại nội dung bằng chuyển động.

Trong bước tạo hoạt ảnh, hệ thống chọn chuyển động 3D tương ứng cho từng ký hiệu. Các mẫu động tác được lấy từ tập video huấn luyện, nơi lưu trữ chuyển động ký hiệu do người thật thực hiện.

Quá trình tái dựng bao gồm nhiều thao tác liên hoàn. Đầu tiên hệ thống trích khung xương từ video nguồn bằng các công cụ như OpenPose, thu được tọa độ các khớp chính. Sau đó, những mô hình cơ thể như SMPLifyX hoặc các dạng mesh khác được dùng để sinh ra hoạt ảnh 3D hoàn chỉnh với hình dạng nhân vật và cử động chính xác.

Khi mỗi ký hiệu đã có một đoạn chuyển động riêng, chúng được lưu lại dưới dạng các clip ngắn. Điều này cho phép tái sử dụng khi ký hiệu đó xuất hiện lần nữa trong những câu khác.

Giai đoạn tiếp theo là ghép toàn bộ các đoạn chuyển động thành một chuỗi liền mạch, đại diện cho nội dung câu văn gốc. Hệ thống cũng thêm các khung trung gian để chuyển động giữa hai ký hiệu không bị giật, giúp quá trình chuyển tư thế trở nên mượt.

Khi các bước xử lý hoàn tất, hệ thống có thể tạo video 3D hoàn chỉnh hoặc trình chiếu trực tiếp trên giao diện. Thành phẩm cuối cùng là chuỗi ký hiệu 3D dễ quan sát, hỗ trợ người điếc và người học ký hiệu tiếp cận nội dung thuận lợi hơn.

### *3.1.2. Ý nghĩa của quy trình*

Escaneado con CamScanner

Việc chuyển đổi nội dung văn bản thành hình thức ký hiệu 3D đóng góp đáng kể cho khả năng tiếp cận thông tin của người điếc. Khi dữ liệu chữ được tái hiện bằng chuyển động trực quan, người xem có thể tiếp nhận nội dung nhanh hơn và dễ hiểu hơn, đặc biệt là trong các hoạt động giao tiếp hằng ngày, nơi tính tức thời và rõ ràng rất quan trọng.

Bên cạnh mục tiêu truyền tải thông điệp, quy trình này còn tạo ra nền móng để phát triển những hệ thống tiên tiến hơn trong tương lai. Kho ký hiệu có thể được mở rộng theo thời gian, chuyển động mô phỏng có thể được cải thiện để đạt độ tự nhiên cao hơn, và hiệu năng của toàn bộ pipeline cũng có thể được tối ưu nhằm phục vụ tốt hơn cho những tình huống sử dụng thực tế. Nhờ vậy, nền tảng này có tiềm năng được mở rộng sang nhiều lĩnh vực như giảng dạy, hỗ trợ hành chính hoặc tương tác trực tuyến.

Tổng quan lại, quy trình xử lý không chỉ hỗ trợ cộng đồng người điếc mà còn góp phần hình thành các công cụ phục vụ học tập và trao đổi thông tin hiệu quả, phù hợp cho nhiều môi trường và mục đích sử dụng khác nhau.

## 3.2. Thử nghiệm chuyển đổi văn bản sang trình tự ký hiệu 3D

Mục tiêu của phần thử nghiệm là kiểm tra và đánh giá khả năng hoạt động của hệ thống trong ba khía cạnh chính. Đầu tiên, hệ thống cần chứng minh khả năng nhận đầu vào là văn bản tiếng Việt và xử lý văn bản này một cách chính xác. Tiếp theo, hệ thống phải chuyển đổi văn bản tiếng Việt thành chuỗi ký hiệu ngôn ngữ ký hiệu tương ứng, đảm bảo sự chính xác và hợp lý trong ánh xạ từ ngữ nghĩa sang cử chỉ ký hiệu.

Cuối cùng, hệ thống sẽ được đánh giá qua khả năng hiển thị chuỗi ký hiệu đó dưới dạng mô phỏng 3D liền mạch, tức là quá trình chuyển đổi văn bản thành hình ảnh động phải diễn ra mượt mà và không có sự gián đoạn. Thông qua thử nghiệm, nghiên cứu viên sẽ tập trung vào việc đánh giá tính đúng đắn của hệ thống, chất lượng mượt mà của hoạt ảnh, cũng như mức độ tự nhiên và chính xác trong trình tự ký hiệu.

Escaneado con CamScanner

### *3.2.1 Chuẩn bị thử nghiệm*

Để tiến hành thử nghiệm, nghiên cứu viên sẽ sử dụng một tập hợp các câu tiếng Việt phổ biến làm dữ liệu đầu vào, ví dụ như "Chúng tôi không cần địa chỉ". Những câu này được chọn vì chứa các từ hoặc cụm từ có sẵn ký hiệu trong cơ sở dữ liệu, giúp hệ thống dễ dàng thực hiện ánh xạ sang ngôn ngữ ký hiệu.

Về công cụ và môi trường, hệ thống sẽ sử dụng bộ từ điển ký hiệu để chuyển đổi các từ thành chuyển động 3D, được trích xuất từ dữ liệu huấn luyện như OpenPose hoặc SMPLifyX. Mô-đun xử lý văn bản sẽ đảm nhiệm việc phân tách câu và ánh xạ từ điển, trong khi mô-đun dựng nhân vật 3D (avatar) sẽ tạo ra các mô hình hình ảnh động. Các mô phỏng này sẽ được chạy trên các máy tính có GPU, hoặc sử dụng Google Colab để tối ưu hóa hiệu suất tính toán.

Để đánh giá kết quả thử nghiệm, ba tiêu chí chính sẽ được áp dụng. Đầu tiên là độ chính xác ánh xạ, tức là kiểm tra xem văn bản đầu vào có được chuyển sang đúng ký hiệu không. Tiếp theo là tính mượt mà của các ký hiệu khi ghép lại với nhau, đảm bảo rằng không có hiện tượng giật, khựng. Cuối cùng, tiêu chí tính dễ hiểu sẽ đánh giá mức độ người dùng quen thuộc với ngôn ngữ ký hiệu có thể hiểu đúng nội dung của các chuỗi ký hiệu đã được mô phỏng.

Trong thử nghiệm này, ta sẽ sử dụng 150 video từ bộ dữ liệu QIPEDC, mỗi video có thời gian dài khoảng 3-4 giây và chứa từ 30-40 frame. Các video này sẽ được sử dụng để huấn luyện và đánh giá khả năng chuyển đổi video ký hiệu tiếng Việt thành các chuyển động ký hiệu 3D.

### *3.2.2 Quy trình thực hiện*

Quy trình thực hiện bắt đầu khi người dùng nhập câu tiếng Việt vào hệ thống thông qua giao diện web hoặc ứng dụng. Đây là bước đầu tiên trong quá trình chuyển đổi văn bản thành ngôn ngữ ký hiệu 3D.

Escaneado con CamScanner

Sau khi văn bản được nhập, hệ thống sẽ thực hiện bước tiếp theo là tách từ và ánh xạ từng từ, cụm từ sang chuỗi ký hiệu tương ứng. Câu văn sẽ được chuyển đổi theo cú pháp của ngôn ngữ ký hiệu Việt Nam, nếu có sự khác biệt so với cấu trúc ngữ pháp tiếng Việt chuẩn. Ví dụ, câu "Địa chỉ ở đâu?" có thể được ánh xạ thành chuỗi ký hiệu ["địa chỉ"] + ["ở đâu"].

Bước tiếp theo là sinh mô phỏng 3D cho từng ký hiệu. Mỗi ký hiệu sẽ được hệ thống truy xuất mô phỏng chuyển động 3D đã được dựng sẵn từ cơ sở dữ liệu. Dữ liệu khung xương (skeleton data) sẽ điều khiển avatar 3D thực hiện các cử chỉ tương ứng với ký hiệu đó.

Sau khi các mô phỏng 3D được tạo ra, bước tiếp theo là ghép nối các chuỗi ký hiệu lại thành một đoạn hoạt ảnh liên tục. Các mô phỏng 3D này sẽ được nối tiếp nhau theo đúng trình tự, và khung chuyển tiếp (transition frames) sẽ được thêm vào để đảm bảo rằng avatar di chuyển một cách mượt mà, tránh hiện tượng giật cục.

Kết quả sẽ được hiển thị trực tiếp trên giao diện 3D, cho phép người dùng theo dõi trực quan quá trình chuyển đổi. Nếu cần, kết quả cũng có thể được kết xuất thành một file video để người dùng tải về và sử dụng

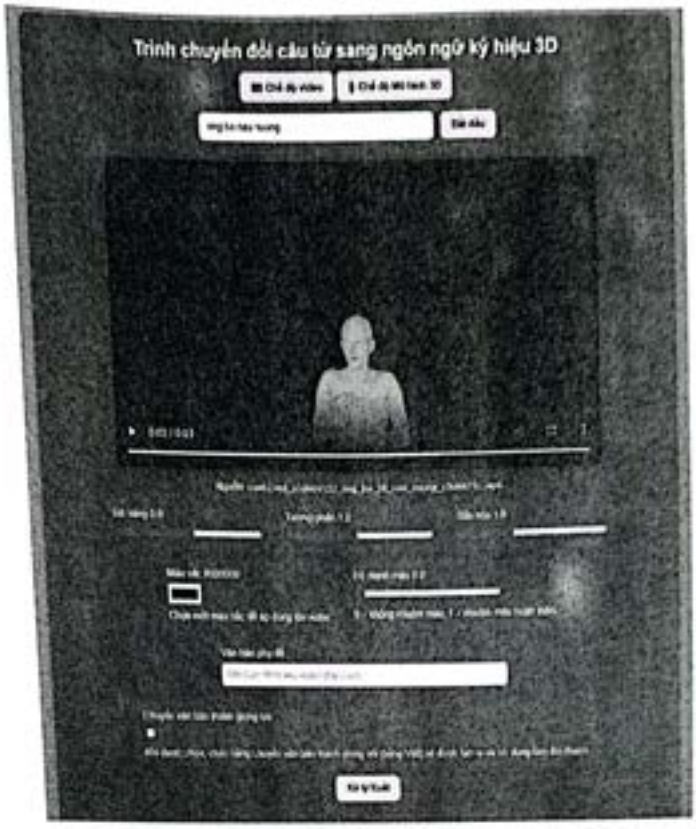
### *3.2.3 Kết quả thử nghiệm và quan sát thực tế*

Escaneado con CamScanner




**Hình 3.1 Kết quả render mô hình 3D**

Escaneado con CamScanner




**Hình 3.2 Kết quả chế độ xem video mô hình 3D**

Kết quả thử nghiệm cho thấy hệ thống có khả năng chuyển đổi chính xác các câu ngắn và từ vựng phổ biến thành chuỗi ký hiệu. Các ký hiệu được hiển thị rõ ràng và đúng với nghĩa của chúng trong ngữ cảnh văn bản gốc. Hoạt ảnh 3D thể hiện đúng từng ký hiệu riêng lẻ, và các chuyển động của avatar 3D mô phỏng khá chính xác các cử chỉ ký hiệu. Tuy nhiên, khi ghép chuỗi các ký hiệu lại với nhau, vẫn xuất hiện một số chuyển động chưa mượt, đôi khi có hiện tượng giật nhẹ khi chuyển từ ký hiệu này sang ký hiệu khác. Mặc dù vậy, các đoạn mô phỏng nhìn chung vẫn dễ hiểu và truyền đạt đúng nội dung.

Escaneado con CamScanner

**Bảng 3.1 Số liệu từ quá trình huấn luyện dữ liệu dataset**

| Số video sử dụng | Số ký hiệu 3D đạt được | Số từ bị hỏng | Mức độ từ chính xác |
|---|---|---|---|
| 150 | 100 | 50 (33%) | 80-90% |

Trong quá trình thử nghiệm biến đổi video 2D sang model 3D, hệ thống đã sử dụng 150 tệp video .mp4, huấn luyện chuyển đổi thành công tổng cộng 100 từ. Tuy nhiên, 50 tệp trong số đó bị lỗi do sự phức tạp của các tư thế và hạn chế của mô hình. Mặc dù vậy, các tệp còn lại vẫn cho kết quả khả quan, với độ chính xác (mức độ tương đương với video ký hiệu gốc) của từ, khả năng hiểu đúng nội dung của người dùng thử nghiệm, qua quan sát, đạt xấp xỉ từ 80–90% đối với các câu đơn giản. Điều này cho thấy hệ thống đang hoạt động hiệu quả, mặc dù vẫn còn một số cải thiện cần thực hiện để hoàn thiện hơn nữa độ mượt mà trong quá trình ghép chuỗi ký hiệu.

**Bảng 3.2 Thống kê nguyên nhân lỗi hỏng dữ liệu huấn luyện**

| Nguyên nhân Thất bại | Số lượng Từ | Tỉ lệ trong số Thất bại |
|---|---|---|
| Lỗi OpenPose (Phát hiện sai Keypoint) | 4 | 8% |
| Lỗi SMPLify-X (Thao tác khớp mô hình cơ thể) | 46 | 92% |
| Tổng số Thất bại | 50 | 100% |

Kết quả trên cho thấy phần lớn các lỗi trong quá trình tạo dữ liệu huấn luyện xuất phát từ bước xử lý bằng SMPLify-X, chiếm tới 92% tổng số thất bại. Trong khi đó, lỗi từ OpenPose (phát hiện sai keypoint) chỉ chiếm 8%. Tổng cộng có 50 trường hợp thất bại được ghi nhận, phản ánh rằng khâu khớp mô hình cơ thể 3D là nguyên nhân chính gây hỏng dữ liệu.

Escaneado con CamScanner

**Bảng 3.3 So sánh các nguyên nhân lỗi trong SMPLify-X**

| Loại lỗi quan sát (diễn đạt lại) | Số lượng mẫu | Tỷ lệ (%) | Mô tả lỗi |
|---|---|---|---|
| Ngón tay chạm sát gây khó nhận diện | 16 | 35.0 | Các ngón tay quá gần nhau khiến mô hình khó tách biệt, dẫn đến sai nghĩa. |
| Chuyển động tay theo chiều sâu dễ gây nhầm lẫn | 14 | 30.0 | Khi tay đưa ra trước hoặc ra sau, thuật toán dễ nhầm hướng hoặc khoảng cách. |
| Tư thế cánh hoặc bàn tay tương tự dễ bị nhận nhầm | 16 | 35.0 | Các tư thế gần giống nhau (úp/ngửa) dễ gây sai lệch ý nghĩa ký hiệu. |
| Tổng cộng | 46 | 100.0 | |

Trong thống kê này, ba nhóm lỗi chính được ghi nhận khi sử dụng SMPLify-X. Các trường hợp ngón tay chạm sát nhau và các tư thế bàn tay tương tự chiếm tỷ lệ ngang nhau, mỗi loại khoảng 35% tổng số lỗi. Lỗi do chuyển động tay theo chiều sâu cũng xuất hiện đáng kể với 30%. Tổng cộng có 46 mẫu bị ảnh hưởng, cho thấy những yếu tố hình thái và chuyển động đều góp phần gây sai lệch trong quá trình mô hình hóa.

Escaneado con CamScanner

**Bảng 3.4 Kết quả thống kê so sánh độ rung với huấn luyện FPS khác nhau**

| FPS | Mức độ rung lắc quan sát | Diễn giải ngắn |
|---|---|---|
| 15 | Không đến tối thiểu | Chuyển động mượt, hầu như không thấy rung lắc. |
| 30 | Dễ nhận thấy | Rung lắc bắt đầu rõ ràng, ảnh hưởng đến độ ổn định. |
| 60 | Rất mạnh | Rung lắc nổi bật, gây khó khăn trong việc giữ nguyên hình dáng chính xác. |

Trong quá trình thử nghiệm, ta đã tiến hành huấn luyện và đánh giá trên 10 đoạn video mẫu với ba mức tốc độ khung hình khác nhau (15, 30 và 60 FPS). Kết quả cho thấy 15 FPS mang lại chuyển động mượt mà nhất, gần như không xuất hiện rung lắc, đồng thời tiết kiệm đáng kể thời gian xử lý và huấn luyện lượng dữ liệu lớn. Ngược lại, ở mức 30 FPS hiện tượng rung lắc bắt đầu trở nên rõ ràng và ảnh hưởng đến độ ổn định, còn ở 60 FPS rung lắc rất mạnh, gây khó khăn trong việc duy trì hình dáng chính xác. Những quan sát này cho thấy việc lựa chọn 15 FPS vừa đảm bảo chất lượng hình ảnh ổn định vừa tối ưu hiệu quả thời gian.

Điều này có thể được giải thích bởi sự khác biệt nhỏ trong ước tính của mô hình 3D, ví dụ như một số khung hình có kích thước nhỏ hơn một chút, trong khi những khung hình khác có kích thước lớn hơn. Việc sử dụng nhiều khung hình hơn hay FPS cao sẽ phóng đại sự khác biệt này. Tuy nhiên, hiện tượng này cũng có thể xuất phát từ sự không hoàn hảo của bộ dữ liệu dataset QIPEDC, chỉ hiển thị nửa trên cơ thể (từ hông trở lên), điều này có thể ảnh hưởng đến độ chính xác của các chuyển động được mô phỏng.

## 3.3. Đánh giá kết quả chuyển đổi và khả năng biểu đạt của mô hình

Escaneado con CamScanner

Phần này tập trung vào việc phân tích hiệu quả của hệ thống khi thực hiện chuyển đổi văn bản sang mô phỏng ký hiệu 3D. Các khía cạnh được quan tâm bao gồm:

- Độ chính xác ngữ nghĩa: mức độ đúng đắn khi văn bản được chuyển thành chuỗi ký hiệu.
- Tính mượt mà của hoạt ảnh: sự liền mạch giữa các động tác ký hiệu.
- Khả năng biểu đạt: mức độ dễ hiểu, tự nhiên, và gần gũi với người dùng, đặc biệt là người khiếm thính.

### *3.3.1 Kết quả quan sát*

Hệ thống thể hiện khả năng ánh xạ khá chính xác các từ vựng phổ biến trong bộ dữ liệu sang ký hiệu 3D. Đối với các câu ngắn, hệ thống truyền tải được ý nghĩa cơ bản của nội dung văn bản, giúp người dùng dễ dàng tiếp nhận thông tin. Tuy nhiên, với các câu phức tạp hoặc những từ ngữ chưa có trong bộ dữ liệu, hệ thống gặp khó khăn trong việc tìm ra ký hiệu tương ứng. Trong những trường hợp này, hệ thống đôi khi bỏ qua hoặc thay thế các từ này bằng những ký hiệu có nghĩa tương tự, làm giảm độ chính xác trong việc thể hiện toàn bộ ý nghĩa của câu.

Khi trình chiếu các ký hiệu riêng lẻ, chuyển động của avatar rõ ràng và dễ nhận diện, giúp người dùng theo dõi và hiểu được từng động tác ký hiệu. Tuy nhiên, khi ghép nhiều ký hiệu thành một chuỗi, vẫn xuất hiện độ trễ và hiện tượng giật nhẹ ở các điểm nối giữa các động tác. Một số ký hiệu cũng có tư thế kết thúc không hoàn toàn khớp với ký hiệu tiếp theo, dẫn đến sự thiếu tự nhiên trong toàn bộ chuỗi, làm giảm tính mượt mà của quá trình mô phỏng.

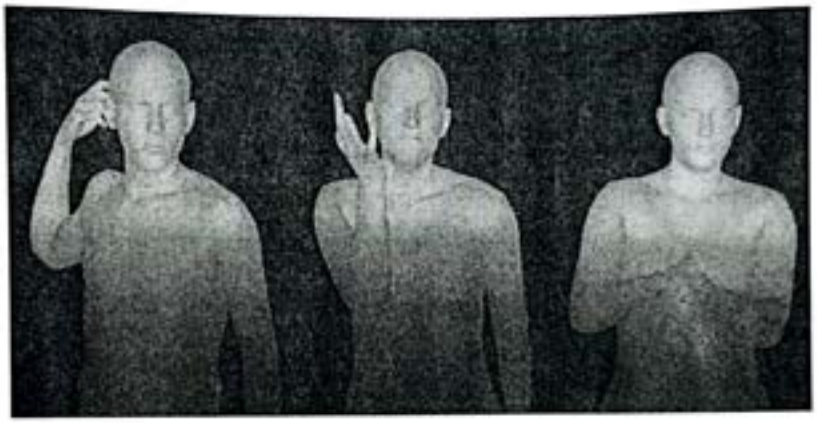
Avatar 3D có thể truyền tải được ngữ nghĩa cốt lõi của từng ký hiệu, giúp người dùng nhận diện được ý nghĩa cơ bản của các động tác ký hiệu. Tuy nhiên, hệ thống

Escaneado con CamScanner

thiếu biểu cảm khuôn mặt, ánh mắt và ngữ điệu cơ thể, những yếu tố quan trọng trong ngôn ngữ ký hiệu. Những thiếu sót này làm giảm tính tự nhiên và sinh động của mô phỏng. Dù vậy, người dùng quen thuộc với ngôn ngữ ký hiệu đánh giá hệ thống là dễ hiểu ở mức cơ bản. Tuy nhiên, để có thể thay thế giao tiếp trực tiếp bằng ngôn ngữ ký hiệu, hệ thống vẫn cần cải thiện để mang lại trải nghiệm tự nhiên và sống động hơn.

### *3.3.2 Các hạn chế còn tồn tại*

Kho từ còn ít: Một vấn đề đáng chú ý của hệ thống là phạm vi từ vựng hiện vẫn còn rất hẹp. Bộ dữ liệu mới chỉ bao phủ một lượng nhỏ các đơn vị ký hiệu trong tiếng ký hiệu Việt Nam, khiến việc xử lý các câu dài, câu mang tính học thuật hoặc những từ ít gặp trở nên khó khăn. Điều này làm giảm khả năng đáp ứng nhu cầu giao tiếp đa dạng của người dùng khiếm thính.



**Hình 3.3 Một số ví dụ về hạn chế kết quả huấn luyện**

Nhiều ký hiệu bị lỗi trong quá trình huấn luyện: Khi xây dựng tập dữ liệu, một số lượng đáng kể video hoặc mô phỏng ký hiệu bị sai lệch do mô hình không tái tạo đúng hình dạng bàn tay hoặc chuyển động cần thiết. Tình trạng này xuất hiện nhiều ở những ký hiệu có cấu trúc ngón tay phức tạp hoặc đòi hỏi sự khác biệt rất nhỏ giữa các tư thế.

Escaneado con CamScanner

Sai lệch về lưới 3D (mesh) và hình dạng bàn tay: Trong giai đoạn dựng hoạt ảnh, mô hình 3D thường gặp lỗi như méo tay, biến dạng lưới hoặc chồng chéo ngón tay. Một số ví dụ phổ biến gồm:

- Ngón tay áp sát nhau khiến hệ thống khó phân tách, dẫn đến ký hiệu bị diễn đạt sai.
- Chuyển động tay đưa ra phía trước hoặc phía sau dễ làm thuật toán nhầm lẫn hướng hoặc độ sâu.
- Những tư thế có ngoại hình gần giống nhau (chẳng hạn bàn tay xoay úp so với xoay ngửa) rất dễ bị nhầm, làm thay đổi ý nghĩa ban đầu.

Chưa xử lý tốt đặc điểm cú pháp riêng của ngôn ngữ ký hiệu: Tiếng ký hiệu không tuân theo trật tự câu giống tiếng Việt, nên nếu chuyển đổi trực tiếp từng từ theo thứ tự văn bản, câu ký hiệu thường trở nên rời rạc, khó hiểu. Hệ thống hiện tại chưa có cơ chế biến đổi cấu trúc câu theo đúng quy tắc ngữ pháp đặc thù trong ngôn ngữ ký hiệu.

Hoạt ảnh thiếu yếu tố phi ngôn ngữ: Hệ thống chủ yếu tái tạo chuyển động tay, trong khi tiếng ký hiệu còn phụ thuộc nhiều vào nét mặt, hướng nhìn, độ nghiêng cơ thể và các tín hiệu phi ngôn ngữ khác. Sự thiếu vắng các yếu tố này khiến hoạt ảnh chưa thể tái hiện đầy đủ sắc thái biểu đạt, làm giảm độ tự nhiên và độ chính xác về mặt ý nghĩa.

Thiếu kiểm nghiệm trên phạm vi rộng: Dù đã có thử nghiệm ban đầu trong một vài tình huống, hệ thống vẫn chưa được đánh giá với số lượng lớn người dùng. Việc không có dữ liệu đánh giá rộng rãi làm khó xác định mức độ dễ hiểu, mức độ hài lòng và sự hữu ích thực tế của hệ thống, đặc biệt đối với những người không quen sử dụng công nghệ.

Escaneado con CamScanner

### *3.3.3 Đánh giá và định hướng cải tiến*

Qua quá trình thử nghiệm, hệ thống cho thấy khả năng ánh xạ văn bản sang ký hiệu tương đối chính xác, giúp người dùng có thể tiếp nhận nội dung một cách trực quan và rõ ràng. Mô phỏng 3D tạo ra các động tác dễ hiểu, hỗ trợ tốt cho người học hoặc người khiếm thính trong việc nắm bắt thông tin.

Tuy nhiên, hệ thống vẫn tồn tại một số hạn chế kỹ thuật cần cải thiện. Trước hết, hệ thống chưa hỗ trợ tốt cho các câu dài hoặc các từ/cụm từ chưa có trong từ điển ký hiệu, gây khó khăn trong xử lý văn bản đa dạng. Thứ hai, các chuỗi ký hiệu khi ghép nối vẫn có hiện tượng giật, thiếu mượt, do thiếu cơ chế nội suy hoặc chuyển tiếp động tác phù hợp. Cuối cùng, biểu cảm khuôn mặt và ánh mắt chưa được mô phỏng, làm giảm độ tự nhiên của chuyển động, ảnh hưởng đến khả năng truyền đạt cảm xúc và ngữ pháp trong ngôn ngữ ký hiệu.

Để nâng cao hiệu quả, có một số định hướng cải tiến quan trọng như sau:

- Mở rộng dữ liệu ký hiệu: Cần bổ sung thêm các bộ dữ liệu ngôn ngữ ký hiệu chuẩn hóa từ các dự án như QIPEDC hoặc các nguồn mở quốc tế, giúp hệ thống hỗ trợ nhiều từ vựng hơn, bao gồm các cụm từ phức tạp hoặc ít gặp.
- Tối ưu ghép chuyển động (motion blending): Cải thiện thuật toán nội suy để làm mượt các điểm nối giữa các động tác, giảm hiện tượng giật và tạo ra chuyển động liền mạch, tự nhiên hơn.
- Tích hợp biểu cảm khuôn mặt và ánh mắt: Sử dụng kỹ thuật motion capture hoặc animation điều khiển bằng AI để tái hiện chính xác các yếu tố phi ngôn ngữ, giúp avatar sống động và dễ hiểu hơn.
- Xử lý ngữ pháp ký hiệu chuyên biệt: Ngôn ngữ ký hiệu có cấu trúc cú pháp khác với tiếng Việt. Việc xây dựng mô-đun xử lý ngữ pháp riêng sẽ giúp hệ

Escaneado con CamScanner

thống chuyển đổi văn bản thành câu ký hiệu đúng trật tự ngôn ngữ ký hiệu, dễ hiểu và phù hợp với người dùng.

Những cải tiến này sẽ đóng vai trò quan trọng trong việc nâng cao chất lượng hệ thống, hướng tới khả năng ứng dụng trong môi trường thực tế như lớp học, trung tâm hỗ trợ người khiếm thính hoặc các nền tảng giao tiếp trực tuyến.

Ngoài các hướng cải tiến nêu trên, trong lĩnh vực tái tạo mô hình cơ thể 3D từ dữ liệu động, hiện tồn tại nhiều phương pháp khác nhau bên cạnh SMPLify-X. Một số mô hình tiêu biểu như HMR (Human Mesh Recovery) hoặc VIBE (Video Inference for Human Body Pose and Shape Estimation) sử dụng mạng học sâu để hồi quy trực tiếp tham số mô hình 3D từ ảnh hoặc chuỗi video, giúp rút ngắn thời gian xử lý và giảm phụ thuộc vào quá trình tối ưu lặp. Tuy nhiên, các phương pháp này thường yêu cầu tập dữ liệu huấn luyện lớn, khó kiểm soát chi tiết hình thái bàn tay và chưa thực sự ổn định đối với các tư thế ký hiệu phức tạp.

Một hướng tiếp cận khác là sử dụng hệ thống motion capture hoặc dữ liệu cảm biến chiều sâu để thu trực tiếp chuyển động 3D. Các phương pháp này cho độ chính xác cao và chuyển động mượt, nhưng đòi hỏi thiết bị chuyên dụng và điều kiện thu thập dữ liệu đặc thù, không phù hợp với mục tiêu xây dựng hệ thống dựa trên video 2D sẵn có như trong đề tài.

So với các hướng tiếp cận trên, SMPLify-X được lựa chọn do khả năng kết hợp tốt giữa keypoint 2D và mô hình cơ thể toàn diện, bao gồm cả thân người và bàn tay. Mặc dù còn tồn tại một số hạn chế về độ ổn định và lỗi khớp hình trong các tư thế phức tạp, phương pháp này cho phép kiểm soát tốt quá trình dựng hình và dễ phân tích nguyên nhân sai lệch. Trong tương lai, hệ thống có thể được mở rộng bằng cách kết hợp SMPLify-X với các mô hình hồi quy học sâu hoặc lựa chọn linh hoạt phương pháp theo từng loại ký hiệu, nhằm cải thiện chất lượng và độ mượt của mô phỏng 3D.

Escaneado con CamScanner

# KẾT LUẬN

Trong quá trình triển khai đề tài, nhóm đã xây dựng được một mô hình thử nghiệm có khả năng chuyển văn bản tiếng Việt thành chuỗi động tác ký hiệu dưới dạng hoạt họa 3D. Phiên bản hiện tại cho thấy hệ thống đã hình thành được một quy trình xử lý tương đối rõ ràng: tiếp nhận nội dung đầu vào, phân tích để tạo ra chuỗi ký hiệu tương ứng và trình bày chúng bằng chuyển động 3D. Đây là bước khởi động quan trọng, cho thấy hướng nghiên cứu này hoàn toàn khả thi và có thể tiếp tục phát triển để hỗ trợ giao tiếp cho người khiếm thính.

Dựa trên những thử nghiệm ban đầu, mô hình đã thể hiện được khả năng truyền tải thông tin ở mức nền tảng thông qua các động tác ký hiệu. Tuy nhiên, các hạn chế vẫn còn đáng kể: lượng ký hiệu chưa phong phú, tiết tấu chuyển động còn thiếu tự nhiên và phần biểu cảm chưa được thể hiện đầy đủ. Những vấn đề này chỉ ra rằng hệ thống cần được bổ sung thêm dữ liệu, cải thiện kỹ thuật dựng chuyển động và đào sâu nghiên cứu về cấu trúc của ngôn ngữ ký hiệu. Đây là nền tảng quan trọng để định hướng phát triển cho các giai đoạn tiếp theo, hướng đến một hệ thống có tính ứng dụng cao hơn trong thực tế.

Trong tương lai, đề tài có thể được mở rộng theo nhiều hướng. Việc gia tăng số lượng và sự đa dạng của bộ dữ liệu sẽ giúp hệ thống xử lý được nhiều đơn vị ngữ nghĩa hơn. Đồng thời, việc tích hợp mô phỏng chuyển động toàn thân cùng các biểu cảm trên gương mặt sẽ giúp nhân vật ảo thể hiện sinh động và thuyết phục hơn. Bên cạnh đó, hệ thống cũng có thể được liên kết với các nền tảng giáo dục, dịch vụ công hoặc công cụ giao tiếp trực tuyến. Cuối cùng, việc đánh giá thử nghiệm với người khiếm thính và chuyên gia trong lĩnh vực ký hiệu sẽ giúp tinh chỉnh hệ thống, tiến gần hơn đến một giải pháp hoàn thiện, ổn định và có giá trị sử dụng lâu dài.

Escaneado con CamScanner

# DANH MỤC CÁC TÀI LIỆU THAM KHẢO

[1] B. D. Khuat, D. T. Phung, H. T. T. Pham, A. N. Bui, S. Tung (2021). "Vietnamese sign language detection using Mediapipe" *2021 10th International Conference on Software and Computer Applications* , pp. 162 – 165.

[2] D. Q. Nguyen, A. T. Nguyen (2020), "PhoBERT: Pre-trained language models for Vietnamese" *Findings of the Association for Computational Linguistics: EMNLP 2020* , pp 1037–1042.

[3] G. V. Martins, J. Magalhães, A. Quinaz, C. Viegas, S. Cavaco (2024). "SLVideo: A Sign Language Video Moment Retrieval Framework" *arXiv*, 2407(15668), pp 1-4.

[4] H. Brock, F. Law, K. Nakadai, Y. Nagashima (2020), "Learning Three-dimensional Skeleton Data from Sign Language Video" *ACM Transactions on Intelligent Systems and Technology (TIST)*. 11(3), pp 1 – 2.

[5] CMU-Perceptual-Computing-Lab (2018). *GitHub:Openpose.* Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose

[6] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, M. J. Black (2019), "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image" *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 10967-10977

[7] QIPEDC. *Danh mục ngôn ngữ ký hiệu.* Available: https://qipedc.moet.gov.vn/slang2

[8] D. Li, C. R. Opazo, X. Yu, H. Li (2020), "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison" *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. pp 1448-1458

Escaneado con CamScanner

[9] Z. Yu, S. Huang, Y. Cheng, T. Birdal (2024), "SignAvatars: A Large-scale 3D Sign Language Holistic Motion Dataset and Benchmark" *ECCV2024*, pp 1-19

[10] R. Rastgoo1, K. Kiani, S. Escalera (2025), "Diffusion-Based Continuous Sign Language Generation with Cluster-Specific Fine-Tuning and Motion-Adapted Transformer" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4127-4136

[11] J. Moon, J. Park, J. Kim, J. Bae, H. Jeon, H. Y. Kim (2024), "DiffSLT: Enhancing Diversity in Sign Language Translation via Diffusion Model" *Pattern Recognition Letters*, 196(C), pp 117 - 125

[12] D. D. Chakladar, P. Kumar, S. Mandal, P. Roy, M. Iwamura, B. Kim (2021), "3D Avatar Approach for Continuous Sign Movement Using Speech/Text" *Applied Sciences* 11, pp 3439

[13] P. Jedlička, Z. Krňoul, M. Zelezny, L. Muller (2022), "MC-TRISLAN: ALarge 3DMotion Capture Sign Language Data-set" *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pp 88–93

Escaneado con CamScanner

# BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung đề án/luận văn/luận án qua phần mềm Turnitin một cách trung thực và đạt kết quả mức độ tương đồng ..4.3.. % toàn bộ nội dung đề án/luận văn/luận án. Bản đề án/luận văn/luận án kiểm tra qua phần mềm là bản cứng đề án/luận văn/luận án đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học viện.

*Hà Nội, ngày ...5. tháng 12... năm 2025*

HỌC VIÊN CAO HỌC/NCS
(Ký và ghi rõ họ tên)

Trần Quang Hưng

Escaneado con CamScanner

turnitin Trang 2/75 - Tổng quan về tính toàn vẹn

ID bài nộp: trn:oid:::3434374741

# 13% Tính tương đồng nói chung

Tổng cộng của tất cả các kết quả trùng khớp, bao gồm cả các nguồn trùng lặp, cho mỗi c...

## Đã lọc khỏi Báo cáo

- Mục lục tham khảo
- Văn bản được trích dẫn
- Văn bản được trích dẫn
- Kết quả trùng khớp nhỏ (ít hơn 8 từ)

## Nguồn hàng đầu

11% Nguồn Internet

3% Ấn bản

2% Bài tập được nộp (bài của học sinh)

**Học viên**

Trần Quang Hưng

**Người hướng dẫn khoa học**

Lê Hoài Nam

ID bài nộp: trn:oid:::3434374741

turnitin Trang 2/75 - Tổng quan về tính toàn vẹn

Escaneado con CamScanner

BỘ THÔNG TIN VÀ TRUYỀN THÔNG
**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**
**Độc lập - Tự do - Hạnh phúc**

# BÁO CÁO GIẢI TRÌNH
# SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Họ và tên học viên: Trần Quang Hưng

Chuyên ngành: Hệ thống thông tin

Khóa: 2024 đợt 1

Tên đề tài: Nghiên cứu, thiết kế và thử nghiệm hệ thống chuyển văn bản tiếng Việt thành ngôn ngữ ký hiệu sử dụng mô hình 3D hỗ trợ người khiếm thính

Người hướng dẫn khoa học: TS. Vũ Hoài Nam

Ngày bảo vệ: 24/01/2025

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

| TT | Ý kiến hội đồng | Sửa chữa của học viên |
|---|---|---|
| 1 | Thực hiện khảo sát thêm một số thuật toán ở pha chuyển đổi DL động sang mô hình 3D | Học viên đã tiếp thu và bổ sung phần khảo sát về thuật toán khác vào mục 3.3.3 trang 62 dòng 20. |
| 2 | Giảm bớt tài liệu tham khảo và chỉnh sửa cho đúng định dạng | Học viên đã lược bớt đi nhiều tài liệu tham khảo và chỉnh sửa định dạng. |
| 3 | Chuẩn hóa các hình vẽ từ tiếng Anh chuyển sang tiếng Việt và hình vẽ cần rõ ràng hơn | Học viên đã tiếp thu và chỉnh sửa các hình ảnh sang Tiếng Việt: Hình 1.2, 1.4, 2.6, 3.1, 3.2 |
| | | |
| | | |

*Hà Nội, ngày 10 tháng 2 năm 2025*

**Ký xác nhận của**

| CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN | THƯ KÝ HỘI ĐỒNG | NGƯỜI HƯỚNG DẪN KHOA HỌC | HỌC VIÊN |
|---|---|---|---|
| PGS.TS. Trần Đình Quế | TS. Đào Thị Thúy Quỳnh | TS. Vũ Hoài Nam | Trần Quang Hưng |

Escaneado con CamScanner

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

# BIÊN BẢN
# HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 3299/QĐ-HV ngày 26 tháng 12 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi ...8... giờ ...00 phút, ngày 24 tháng 01 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Trần Quang Hưng**

Tên đề án tốt nghiệp: **Nghiên cứu, thiết kế và thử nghiệm hệ thống chuyển văn bản tiếng Việt thành ngôn ngữ ký hiệu sử dụng mô hình 3D hỗ trợ người khiếm thính**

Ngành: **Hệ thống thông tin** Mã số: **8.48.01.04**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: ...5.../ 05

| TT | HỌ VÀ TÊN | TRÁCH NHIỆM TRONG HĐ | GHI CHÚ |
|---|---|---|---|
| 1 | **PGS.TS. Trần Đình Quế** | Chủ tịch | |
| 2 | **TS. Đào Thị Thúy Quỳnh** | Thư ký | |
| 3 | **TS. Nguyễn Quang Hưng** | Phản biện 1 | |
| 4 | **PGS.TS. Trần Đăng Hưng** | Phản biện 2 | |
| 5 | **PGS.TS. Nguyễn Hà Nam** | Uỷ viên | |

**Các nội dung thực hiện:**

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên (Có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

1) Pha chuyển đổi dữ liệu chuyển động sang mô hình có thể 3D có tốn thêm thời gian MRP không? Tại sao tỷ lệ lỗi lại cao như vậy?

2) Nếu vốn cao thì khả dụng, liệu có phù hợp với thực tế ứng dụng không?

Escaneado con CamScanner

7. Trả lời của học viên:

1) Học viên cũng đã khảo sát một số thuật toán thuộc nhóm độ lệ rất cao nên em quyết định chọn thuật toán S.M.Lijey-X.

2) Hiện tại mới dừng ở mô phỏng, trong tương lai sẽ thực hiện phát triển thêm.

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: PGS. TS. Trần Đình Quế
  2. Uỷ viên Ban kiểm phiếu: TS. Nguyễn Quang Hưng
  3. Uỷ viên Ban kiểm phiếu: TS. Đào Thị Thúy Quỳnh
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 7,4

**Kết luận:**

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:
   - Thực hiện khảo sát thêm một số thuật toán ở pha chuyển đổi D.L động sang mô hình 3D.
   - Cần bổ sung tài liệu tham khảo và chỉnh sửa cho đúng định dạng.
   - Chuẩn hoá các hình vẽ từ tiếng anh chuyển sang tiếng Việt và hình vẽ cần rõ ràng hơn.
2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên: Đề nghị công nhận
3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS.....

Buổi làm việc kết thúc vào...9h00..... cùng ngày.

| **Chủ tịch** | **Thư ký** |
|---|---|
| **PGS.TS. Trần Đình Quế** | **TS. Đào Thị Thúy Quỳnh** |

Escaneado con CamScanner

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

# BẢN NHẬN ĐỀ ÁN THẠC SĨ CỦA CÁN BỘ PHẢN BIỆN

Về đề tài: Nghiên cứu, thiết kế và thử nghiệm hệ thống chuyển văn bản tiếng Việt thành ngôn ngữ kí hiệu sử dụng mô hình 3D hỗ trợ người khiếm thính.

Chuyên ngành: Hệ thống thông tin

Của học viên: Trần Quang Hưng

Họ và tên cán bộ phản biện: PGS.TS Trần Đăng Hưng

Cơ quan công tác: Trường CNTT&TT, Trường Đại học Công nghiệp Hà Nội

## NỘI DUNG NHẬN XÉT

Việc xây dựng hệ thống chuyển từ văn bản sang ngôn ngữ ký hiệu sử dụng mô hình 3D xuất phát từ nhu cầu cấp thiết trong việc hỗ trợ người khiếm thính tiếp cận thông tin. Trong khi phần lớn nội dung hiện nay được trình bày dưới dạng văn bản hoặc lời nói, người khiếm thính lại chủ yếu sử dụng ngôn ngữ ký hiệu để giao tiếp và học tập. Sự thiếu tương thích này tạo ra rào cản lớn, hạn chế khả năng tiếp cận tri thức, dịch vụ công và các hoạt động xã hội của họ.

Bên cạnh đó, số lượng phiên dịch viên ngôn ngữ ký hiệu hiện nay còn rất hạn chế, không thể đáp ứng đầy đủ nhu cầu trong các lĩnh vực như giáo dục, y tế hay hành chính. Việc phụ thuộc hoàn toàn vào con người khiến chi phí cao và khó triển khai rộng rãi, đặc biệt trong các tình huống cần hỗ trợ kịp thời. Do đó, một hệ thống tự động chuyển đổi văn bản sang ngôn ngữ ký hiệu là giải pháp cần thiết nhằm giảm bớt sự phụ thuộc vào nguồn nhân lực phiên dịch.

Ngoài giá trị thực tiễn, bài toán này còn mang ý nghĩa xã hội và nhân văn sâu sắc. Việc phát triển các công nghệ hỗ trợ người khuyết tật thể hiện sự quan tâm đến quyền tiếp cận thông tin và bình đẳng xã hội, phù hợp với các chủ trương, chính sách an sinh

Escaneado con CamScanner

và các công ước quốc tế về quyền của người khuyết tật. Đồng thời, đây cũng là một hướng nghiên cứu có tính thời sự cao, tận dụng sự phát triển mạnh mẽ của trí tuệ nhân tạo, xử lý ngôn ngữ tự nhiên và công nghệ đồ họa 3D để giải quyết một vấn đề thực tiễn quan trọng. Vì vậy đề tài đề án có tính cấp thiết và ý nghĩa thực tiễn cao.

Đề án được tổ chức trong 3 chương:

Chương 1 trình bày tổng quan về hệ thống chuyển văn bản thành ngôn ngữ kí hiệu. Trong đó có khái quát về ngô ngữ kí hiệu và các tiếp cận công nghệ trong chuyển đổi văn bản sang ngôn ngữ kí hiệu.

Chương 2 xây dựng hệ thống chuyển văn bản sang mô hình 3D, trong đó có nêu quy trình xử lý, thiết kế kiến trúc hệ thống và thiết kế giao diện người dùng.

Chương 3 thử nghiệm mô hình chuyển đổi văn bản sang ký hiệu 3D, trong đó có thử nghiệm và dánh giá kết quả chuyển đổi và khả năng biểu đạt của mô hình.

**Nhận xét và góp ý:**

- Đề tài có ý nghĩa thực tiễn và có tính nhân văn.

Một số góp ý:

- Nội dung đề án trình bày dưới dạng giới thiệu, chưa rõ quá trình thực hiện xây dựng hệ thống chuyển đổi như thế nào, nội dung của Chương 2 chủ yếu là mô tả cách làm và quy trình. Cần làm rõ hơn nội dung và đóng góp của đề án.
- Khá nhiều hình vẽ đang để bằng tiếng Anh, cần chuyển sang tiếng Việt.
- Khá nhiều sơ đồ, hình vẽ chữ bị vỡ và khó đọc. Cần chỉnh sửa.

**Kết luận:**

- Đồng ý cho học viên bảo vệ kết quả đề án trước hội đồng.

*Hà Nội, ngày 18 tháng 1 năm 2026*

CÁN BỘ PHẢN BIỆN

PGS.TS Trần Đăng Hưng

Escaneado con CamScanner

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

# BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

*(Dùng cho người phản biện)*

Tên đề tài đề án tốt nghiệp: Nghiên cứu, thiết kế và thử nghiệm hệ thống chuyển văn bản tiếng Việt thành ngôn ngữ ký hiệu sử dụng mô hình 3D hỗ trợ người khiếm thính
Chuyên ngành: Hệ thống thông tin
Mã chuyên ngành: 8.48.01.04
Họ và tên học viên: Trần Quang Hưng
Họ và tên người nhận xét: Nguyễn Quang Hưng
Học hàm, học vị: Tiến sỹ
Chuyên ngành: Mạng và kênh thông tin liên lạc
Cơ quan công tác: Học viện CN Bưu chính viễn thông
Số điện thoại: .....E-mail: hungnq1@ptit.edu.vn

## NỘI DUNG NHẬN XÉT

**I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:**

- Đề tài có tính nhân văn và đáp ứng nhu cầu thực tế của cộng đồng người khiếm thính tại Việt Nam.
- Định hướng ứng dụng rõ ràng, phù hợp với xu hướng công nghệ hỗ trợ người khuyết tật.

**II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:**

- Có tổng quan tốt về ngôn ngữ ký hiệu, các tiếp cận công nghệ hiện có.
- Trình bày khá chi tiết quy trình xử lý từ văn bản đến mô hình 3D.
- Có công thức toán học minh họa các bước xử lý.
- Có thử nghiệm thực tế với dataset QIPEDC (150 video).
- Cấu trúc đề án logic, đầy đủ các phần: tổng quan, xây dựng hệ thống, thử nghiệm.

Về kỹ thuật:

- Sử dụng các công cụ phù hợp: OpenPose, SMPLify-X.
- Pipeline xử lý được mô tả tương đối rõ ràng.

**III/ Những vấn đề cần giải thích thêm:**

Hạn chế về phương pháp:

- Hệ thống chủ yếu dựa vào ánh xạ trực tiếp (lookup) từ dictionary, chưa có mô hình AI/ML học sâu
- Chưa xử lý được ngữ pháp đặc thù của ngôn ngữ ký hiệu
- Thiếu các yếu tố phi-thủ (biểu cảm, ánh mắt)

Hạn chế về dữ liệu:

- Dataset nhỏ (chỉ 150 video, thành công 100 từ, tỷ lệ lỗi 33%)
- Chưa có đánh giá định lượng với người dùng thực tế

Escaneado con CamScanner

Hạn chế về kết quả:

- Chất lượng chuyển động chưa mượt (có hiện tượng giật)
- Độ chính xác chỉ 80-90% cho câu đơn giản (chưa nói rõ thang đo lựa chọn/quy trình thực hiện để tính toán kết quả - hay chỉ là định tính của học viên?).

**IV/ Kết luận:**

Tuy còn một số hạn chế về phương pháp, dữ liệu, và đánh giá khoa học, đề án đạt được mục tiêu xây dựng prototype chuyển văn bản tiếng Việt sang ký hiệu 3D, có ý nghĩa thực tiễn, có thể là một công trình định hướng ứng dụng ở mức khởi đầu.

Đồng ý cho học việc được bảo vệ trước Hội đồng chấm đề án.

**CÂU HỎI:**

**Câu 1**: Tại sao học viên lại chọn phương pháp ánh xạ trực tiếp (dictionary-based lookup) thay vì sử dụng các mô hình học máy hiện đại như Transformer, seq2seq hoặc diffusion models mà chính đề án đã đề cập trong phần tổng quan? Liệu phương pháp này có đủ khả năng mở rộng khi cần xử lý ngữ pháp phức tạp?

**Câu 2**: Tỷ lệ lỗi khi huấn luyện lên đến 33% (50/150 video), trong đó 92% lỗi từ SMPLify-X (Bảng 3.2, 3.3). Học viên đã thử các phương pháp thay thế hoặc cải tiến SMPLify-X chưa? Có kế hoạch nào để giảm tỷ lệ lỗi này xuống mức chấp nhận được cho ứng dụng thực tế không?

**Câu 3**: Với chỉ 100 từ ký hiệu khả dụng, liệu hệ thống có đủ để ứng dụng trong các tình huống thực tế như giáo dục, y tế, dịch vụ công mà đề án đề cập không?

Hà Nội, ngày 23 tháng 1 năm 2026

NGƯỜI NHẬN XÉT

TS. Nguyễn Quang Hưng

Escaneado con CamScanner