

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

NGUYỄN QUỐC KHÁNH – 20521452

KHÓA LUẬN TỐT NGHIỆP
XÂY DỰNG HỆ THỐNG CHUYỂN ĐỔI NGÔN NGỮ
KÝ HIỆU CHO NGƯỜI KHIẾM THÍNH THÔNG QUA
CÔNG NGHỆ GENERATIVE AI

**BUILDING A SIGN LANGUAGE TRANSLATION SYSTEM
FOR THE HEARING IMPAIRED USING GENERATIVE AI
TECHNOLOGY**

CỬ NHÂN NGÀNH HỆ THỐNG THÔNG TIN

TP. HỒ CHÍ MINH, 2024

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

NGUYỄN QUỐC KHÁNH – 20521452

KHÓA LUẬN TỐT NGHIỆP

XÂY DỰNG HỆ THỐNG CHUYỂN ĐỔI NGÔN NGỮ
KÝ HIỆU CHO NGƯỜI KHIẾM THÍNH THÔNG QUA
CÔNG NGHỆ GENERATIVE AI

**BUILDING A SIGN LANGUAGE TRANSLATION SYSTEM
FOR THE HEARING IMPAIRED USING GENERATIVE AI
TECHNOLOGY**

CỬ NHÂN NGÀNH HỆ THỐNG THÔNG TIN

GIÁO VIÊN HƯỚNG DẪN

TS. NGUYỄN THANH BÌNH

TP. HỒ CHÍ MINH, 2024

THÔNG TIN HỘI ĐỒNG CHẤM KHÓA LUẬN TỐT NGHIỆP

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số

ngày của Hiệu trưởng Trường Đại học Công nghệ

Thông tin.

1. – Chủ tịch.
2. – Thư ký.
3. – Ủy viên.
4. – Ủy viên.

ĐẠI HỌC QUỐC GIA

TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC

CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA

VIỆT NAM

Độc lập - Tự do - Hạnh phúc

TP. HCM, ngày... tháng... năm 2024

NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

CỦA CÁN BỘ HƯỚNG DẪN

Tên khóa luận:

**XÂY DỰNG HỆ THỐNG CHUYỂN ĐỔI NGÔN NGỮ KÝ HIỆU CHO
NGƯỜI KHIẾM THÍNH THÔNG QUA CÔNG NGHỆ GENERATIVE AI**

Nhóm SV thực hiện:

Nguyễn Quốc Khánh - 20521452

Cán bộ hướng dẫn

TS Nguyễn Thanh Bình

Đánh giá Khóa luận

1. Về cuốn báo cáo:

Số trang _____

Số chương _____

Số bảng số liệu _____

Số hình vẽ _____

Số tài liệu tham khảo _____

Sản phẩm _____

Một số nhận xét về hình thức cuốn báo cáo:

.....
.....
.....

2. Về nội dung nghiên cứu

.....
.....
.....

3. Về chương trình ứng dụng

.....
.....
.....

4. Về thái độ làm việc của sinh viên

.....
.....
.....

Đánh giá chung:

.....
.....
.....

Điểm từng sinh viên:

Nguyễn Quốc Khánh: __ /10

Người nhận xét

(Ký và ghi rõ họ tên)

LỜI CẢM ƠN

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Trong suốt học kỳ này khi bắt đầu đăng ký học phần Khóa Luận Tốt Nghiệp, em đã nhận được rất nhiều sự quan tâm, giúp đỡ của Thầy Cô, các anh chị khóa trên và bạn bè trong và ngoài lớp.

Với lòng biết ơn sâu sắc nhất, về phía Giảng viên phụ trách học phần em xin chân thành cảm ơn Thầy Nguyễn Thành Bình đã tận tâm hướng dẫn, giải đáp kịp thời các thắc mắc về khóa luận. Nếu không có những lời hướng dẫn của Thầy và các anh chị khóa trên thì em nghĩ bài báo cáo của em rất khó để hoàn thành được.

Trong thời gian một học kỳ tham gia và thực hiện khóa luận, em đã cố gắng vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới từ Thầy Cô, Anh Chị, bạn bè cũng như nhiều nguồn tài liệu tham khảo. Từ đó, em vận dụng tối đa những gì đã thu thập được để hoàn thành một báo cáo khóa luận một cách tốt nhất. Tuy nhiên, vì kiến thức chuyên môn còn hạn chế và bản thân còn thiếu nhiều kinh nghiệm thực tiễn nên nội dung của báo cáo không tránh khỏi những thiếu sót, em rất mong nhận được sự góp ý, chỉ bảo thêm của các Thầy/Cô nhằm để hoàn thiện những kiến thức của mình để em có thể dùng làm hành trang thực hiện tiếp các đề tài khác trong tương lai cũng như là trong việc học tập và làm việc sau này.

Sau cùng, kính chúc các Thầy và Cô thật nhiều sức khỏe, niềm tin để tiếp tục thực hiện sứ mệnh cao đẹp của mình là truyền đạt kiến thức cho thế hệ mai sau.

TP. Hồ Chí Minh, tháng 07 năm 2024

Sinh viên thực hiện
Nguyễn Quốc Khánh

MỤC LỤC

TÓM TẮT KHÓA LUẬN	1
CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI	3
1.1 Đặt vấn đề	3
1.2 Mục tiêu đề tài.....	4
1.3 Phạm vi nghiên cứu.....	5
1.4 Phương pháp thực hiện.....	5
1.5 Công cụ và môi trường phát triển	5
CHƯƠNG 2. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN.....	8
2.1 Các công trình nghiên cứu liên quan về nhận diện và dịch ngôn ngữ ký hiệu .8	
2.1.1 Sign Language Recognition and Translation Systems for Enhanced Communication for the Hearing Impaired	8
2.1.2 Real-Time Sign Language Recognition using Deep Learning Techniques	8
2.1.3 Sign Language Translation Across Multiple Languages	9
2.1.4 A two-stage sign language recognition method focusing on the semantic features of label text	10
2.1.5 Sign Language to Text Translation with Computer Vision: Bridging the Communication Gap	11
2.2 Các công trình nghiên cứu liên quan về hệ thống chuyển đổi ngôn ngữ ký hiệu	12
2.2.1 A Comprehensive Application for Sign Language Alphabet and World Recognition, Text-to-Action Conversion for Learners, Multi-Language Support and Integrated Voice Output Functionality	12
2.2.2 Software based sign language converter	14
2.2.3 Real-Time Gesture Based Sign Language Recognition System	15
2.2.4 Deaf talk using 3D animated sign language: A sign language interpreter using Microsoft's kinect v2	16
2.2.5 Design and Development of Teaching and Learning Tool Using Sign Language Translator to Enhance the Learning Skills for Students With Hearing and Verbal Impairment.....	18
2.3 Các công trình nghiên cứu liên quan tại Việt Nam	19
2.3.1. SOS - Máy phiên dịch ngôn ngữ ký hiệu dành cho người khiếm thính .19	

2.3.2. Thiết bị giao tiếp thông minh dành cho người khiếm thanh, khiếm thính – “Speak your mind” (SYM)	20
2.3.3. SIGNTEGRATE - Ứng dụng dịch thuật Ngôn ngữ kí hiệu sử dụng trí tuệ nhân tạo Việt Nam.....	21
2.3 Phương pháp đề xuất.....	22
CHƯƠNG 3. CƠ SỞ LÝ THUYẾT	24
3.1 Giới thiệu ngôn ngữ ký hiệu	24
3.2 Lịch sử của ngôn ngữ ký hiệu và văn hóa người khiếm thính.....	26
3.3 Tổng quan về ngôn ngữ kí hiệu.....	27
3.4 Biểu diễn ngôn ngữ kí hiệu	30
3.5 Các bài toán của ngôn ngữ kí hiệu	34
3.5.1 Phát hiện ngôn ngữ ký hiệu.....	34
3.5.2 Nhận dạng ngôn ngữ kí hiệu	36
3.5.3 Phân đoạn ngôn ngữ ký hiệu.....	36
3.5.4 Nhận dạng, dịch, và tạo ngôn ngữ ký hiệu.....	38
3.5.5 Video-To-Pose	40
3.5.6 Pose-To-Video	43
3.6 Sign Language Avatars.....	44
3.7 Mô hình tạo ảnh và video.....	47
3.8 Phương pháp đánh giá – Evaluation Metricsc	62
3.9 Truy xuất ngôn ngữ ký hiệu	64
3.9 Fingerspelling - Đánh vần bằng ngón tay	64
3.11 Pretraining and Representation - Learning	67
CHƯƠNG 4. THỰC NGHIỆM PHƯƠNG PHÁP CHUYỂN ĐỔI NGÔN NGỮ KÝ HIỆU	69
4.1 Giới thiệu về Công nghệ Nhận diện Ngôn ngữ Ký hiệu.....	69
4.1.1 Định nghĩa.....	69
4.1.2 Ứng dụng.....	70
4.2 Các thành phần của hệ thống chuyển đổi.....	70
4.2.1 Hệ thống Chuyển đổi Ngôn ngữ Ký hiệu	70
4.2.2 Kiến trúc hệ thống.....	71
4.2.3 Công nghệ và công cụ sử dụng	71

4.2.4 Quy trình hoạt động của hệ thống.....	72
4.3 Nhận diện ngôn ngữ.....	80
4.4 Chuẩn hóa văn bản.....	81
4.5 Ngôn ngữ kí hiệu – SignWriting.....	81
4.5.1 Sign Writing – Illustration.....	82
4.5.2 SignWriting Description.....	82
4.6 Pose Anonymization.....	83
CHƯƠNG 5. THIẾT KẾ VÀ TRIỂN KHAI ỨNG DỤNG CHUYỂN ĐỔI NGÔN	
NGỮ KÝ HIỆU.....	86
5.1 Thiết kế và triển khai giao diện người dùng.....	87
5.2 Normalized Text.....	89
5.3 Dịch Text sang Gloss.....	89
5.4. Chuyển đổi các Gloss thành Pose.....	92
5.5 Chuyển đổi Pose thành Video.....	92
5.6 Hỗ trợ quốc tế hóa nhiều ngôn ngữ.....	93
CHƯƠNG 6. KẾT QUẢ ĐẠT ĐƯỢC VÀ HƯỚNG PHÁT TRIỂN.....	94
6.1 Kết quả đạt được.....	94
6.2 Hướng phát triển.....	94
TÀI LIỆU THAM KHẢO.....	96

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Dịch sang Tiếng Anh	Ý nghĩa Tiếng Việt
CV	Computer Vision	Thị Giác Máy Tính
LTR	Left to Right	Trái sang phải
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
RTL	Right to Left	Phải sang trái
SLP	Sign Language Processing	Xử lý ngôn ngữ ký hiệu
SRT	Sign Language Recognition	Nhận diện ngôn ngữ ký hiệu
SLT	Sign Language Translation	Dịch ngôn ngữ ký hiệu

DANH MỤC HÌNH ẢNH

Hình 1 Hình minh họa từng biểu diễn ngôn ngữ ký hiệu trong báo cáo.....	30
Hình 2 Hình minh họa các cách biểu diễn khác nhau cho các dấu hiệu	34
Hình 3 Tổng quát các nhiệm vụ liên quan tới ngôn ngữ ký hiệu	40
Hình 4 Thành công mới của AI: Chuyển lời nói sang ngôn ngữ ký hiệu	69
Hình 5 Sơ đồ kiến trúc hệ thống pipeline 1 Spoken to Signed	73
Hình 6 Sơ đồ kiến trúc hệ thống pipeline 2 Signed to Spoken	78
Hình 7 Giao diện đơn giản hệ thống đề tài	80
Hình 8 Mô tả các lớp (Layer) của mô hình mạng CLD3 của Google.....	81
Hình 9 Bảng miêu tả sự diễn giải thành tượng hình và ngôn ngữ viết của SignWriting	82
Hình 10 Bảng giải mã một số SignWriting thành ngôn ngữ nói.....	83
Hình 11 Các dạng biến đổi khác nhau của khung xương (Skeleton Viewer)	84
Hình 12 Minh họa từ Gloss -> HamNoSys -> Video.....	85
Hình 13 Kiến trúc toàn bộ đề tài	86
Hình 14 Giao diện của hệ thống dịch ngôn ngữ ký hiệu (Web application).....	87
Hình 15 Giao diện dịch từ text sang của khung xương của hệ thống dịch	88
Hình 16 Giao diện người dùng khi dịch từ video/camera sang text	89
Hình 17 Hình minh họa quy trình text-to-gloss của phương pháp chuyển văn bản sang Gloss	89

TÓM TẮT KHÓA LUẬN

Hiện nay, người khiếm thính chủ yếu sử dụng thủ ngữ¹ để giao tiếp trong cuộc sống hàng ngày. Tuy nhiên, phương pháp này tồn tại nhiều hạn chế. Để giao tiếp, người tương tác buộc phải biết thủ ngữ, dẫn đến việc tương tác trở nên khó khăn và mất nhiều thời gian.

Mục tiêu của đề tài nghiên cứu là tạo ra một công cụ hỗ trợ giao tiếp hiệu quả, giúp người khiếm thính dễ dàng trao đổi thông tin với mọi người. Hệ thống này không chỉ giúp cải thiện chất lượng cuộc sống của người khiếm thính mà còn góp phần tạo ra một xã hội hòa nhập và không còn rào cản giao tiếp.

Đề tài thực hiện tiếp cận giải quyết vấn đề thành các bài toán sau

- **Nhận dạng giọng nói và xác định ngôn ngữ:** Hệ thống bắt đầu bằng việc sử dụng các mô hình nhận dạng giọng nói tiên tiến như Whisper hoặc các giải pháp của Google để chuyển đổi giọng nói thành văn bản. Tiếp theo, hệ thống sử dụng các mô hình nhận diện ngôn ngữ tự động như Google's cld3 hoặc MediaPipe Solutions để xác định ngôn ngữ của văn bản đầu vào. Việc xác định ngôn ngữ này giúp hệ thống có thể xử lý nhiều ngôn ngữ khác nhau, cải thiện trải nghiệm người dùng và độ chính xác của dịch. Sau đó, văn bản đầu vào sẽ được chuẩn hóa thông qua mô hình LLM (Large Language Models) để đảm bảo chất lượng dịch thuật cao hơn.
- **Dịch văn bản sang ngôn ngữ ký hiệu:** Văn bản chuẩn hóa sẽ được dịch sang SignWriting bằng các mô hình dịch máy như Conditional Variational Autoencoder (CVAE). Sau đó, các mô hình như SMPL-X sẽ được sử dụng để chuyển đổi SignWriting thành chuỗi Pose, tạo ra các hình ảnh động 3D cho ngôn ngữ ký hiệu. Điều này giúp biến văn bản thành các động tác ký hiệu, giúp người dùng khiếm thính dễ dàng hiểu nội dung.

¹ **Thủ ngữ** là từ được gọi ở Việt Nam thay thế cho Ngôn ngữ ký hiệu, tên Tiếng Anh là Sign Language.

- **Tích hợp và triển khai hệ thống:** Cuối cùng, hệ thống sẽ được tích hợp và triển khai trên nền tảng web. Sau đó, thực hiện các thử nghiệm và tinh chỉnh để đạt được hiệu suất tối ưu và phản hồi tích cực từ người dùng. Điều này đảm bảo hệ thống hoạt động hiệu quả và đáp ứng nhu cầu thực tế của người sử dụng.

Nội dung khóa luận bao gồm 7 chương chính:

- CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI
- CHƯƠNG 2: CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN
- CHƯƠNG 3: CƠ SỞ LÝ THUYẾT
- CHƯƠNG 4: THỰC NGHIỆM PHƯƠNG PHÁP CHUYỂN ĐỔI NGÔN NGỮ KÝ HIỆU
- CHƯƠNG 5: THIẾT KẾ VÀ TRIỂN KHAI ỨNG DỤNG CHUYỂN ĐỔI NGÔN NGỮ KÝ HIỆU
- CHƯƠNG 6: KẾT QUẢ ĐẠT ĐƯỢC VÀ HƯỚNG PHÁT TRIỂN
- CHƯƠNG 7: TÀI LIỆU THAM KHẢO

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề

Hiện trạng và vấn đề

Hiện có hơn 1,5 tỷ người (gần 20% dân số toàn cầu) sống chung với tình trạng nghe kém, trong đó có 430 triệu người bị điếc tai. Dự kiến đến năm 2050, con số 430 triệu người bị điếc tai có thể tăng đến 700 triệu người.

Nhưng 50% các trường hợp điếc tai có thể phòng tránh được. Riêng ở trẻ em, 60% trường hợp điếc tai có thể ngăn ngừa được bằng các biện pháp như tiêm chủng ngừa rubella và viêm màng não; cải thiện chăm sóc sức khỏe bà mẹ và trẻ sơ sinh; tầm soát và xử trí sớm bệnh viêm tai giữa. Còn với người lớn, biết cách nghe “an toàn”, cần tránh tiếng ồn, không dùng bừa bãi các loại thuốc gây độc cho tai và vệ sinh tai tốt sẽ giúp giảm nguy cơ mất thính lực. Ngoài ra, cũng cần khám sàng lọc để phát hiện bệnh và điều trị kịp thời. (Nguồn: Ngày Thính lực Thế Giới - World Hearing Day, Tổ chức Y tế Thế giới WHO, 2023).

Mặc dù tỉ lệ người khiếm thính cao, nhưng họ lại là một trong những nhóm đối tượng thiệt thòi nhất trong việc tiếp cận thông tin. Nguyên nhân chủ yếu là do rất ít các phương tiện truyền thông sử dụng ngôn ngữ ký hiệu dành cho người khiếm thính. Điều này dẫn đến sự cách biệt lớn trong việc tiếp nhận thông tin giữa người khiếm thính và cộng đồng.

Sự khan hiếm phiên dịch viên ngôn ngữ ký hiệu

Một trong những khó khăn lớn nhất đối với người khiếm thính là sự thiếu hụt trầm trọng của các phiên dịch viên ngôn ngữ ký hiệu. Tại Hà Nội, chỉ có 6 người đạt trình độ có thể dịch các lĩnh vực cho người điếc, trong khi cả nước chỉ có khoảng 10 người. Điều này khiến cho người khiếm thính gặp rất nhiều khó khăn trong giao tiếp và tiếp cận thông tin trong các lĩnh vực quan trọng như y tế, giáo dục, và dịch vụ công cộng.

Hạn chế của ngôn ngữ ký hiệu

Hiện nay, người khiếm thính chủ yếu sử dụng thủ ngữ (ngôn ngữ ký hiệu Việt Nam) để giao tiếp trong cuộc sống hàng ngày. Tuy nhiên, phương pháp này tồn tại nhiều hạn chế. Để giao tiếp, người tương tác buộc phải biết thủ ngữ, dẫn đến việc tương tác trở nên khó khăn và mất nhiều thời gian. Hơn nữa, việc giao tiếp bằng thủ ngữ dễ dẫn đến hiểu lầm và nhầm lẫn. Dù nhu cầu học và sử dụng ngôn ngữ ký hiệu rất lớn, nhưng hiện nay Việt Nam có rất ít trung tâm dạy thủ ngữ, làm hạn chế khả năng tiếp cận của người khiếm thính.

Khó khăn trong tiếp cận dịch vụ và cảm nhận kỳ thị

Người khuyết tật nói chung và người khiếm thính nói riêng gặp nhiều khó khăn trong việc tiếp cận các dịch vụ y tế, giáo dục và có tỷ lệ thất nghiệp cao. Cảm nhận bị kỳ thị và phân biệt đối xử cũng ảnh hưởng nghiêm trọng đến đời sống vật chất và tinh thần của họ. Việc thiếu các công cụ hỗ trợ giao tiếp hiệu quả càng làm tăng thêm khó khăn này.

Đề xuất giải pháp

Đây chính là những lý do quan trọng để nhóm chọn đề tài xây dựng hệ thống chuyển ngôn ngữ ký hiệu sang. Mục tiêu của đề tài này là tạo ra một công cụ hỗ trợ giao tiếp hiệu quả, giúp người khiếm thính dễ dàng trao đổi thông tin với mọi người. Hệ thống này không chỉ giúp cải thiện chất lượng cuộc sống của người khiếm thính mà còn góp phần tạo ra một xã hội hòa nhập và không còn rào cản giao tiếp.

1.2 Mục tiêu đề tài

- Phát triển một hệ thống sử dụng công nghệ AI để chuyển đổi ngôn ngữ ký hiệu thành văn bản và phát âm.
- Tạo điều kiện giao tiếp mượt mà hơn cho người khiếm thính với những người không biết ngôn ngữ ký hiệu.

- Tăng cường khả năng hòa nhập và tương tác xã hội cho cộng đồng người khiếm thính.

1.3 Phạm vi nghiên cứu

- Nghiên cứu và thực hiện trên bộ dữ liệu có sẵn cho ngôn ngữ ký hiệu, với đối tượng là người khiếm thính
- Phát triển hệ thống có khả năng mở rộng cho các ngôn ngữ ký hiệu Việt.

1.4 Phương pháp thực hiện

- Thu thập dữ liệu video ngôn ngữ ký hiệu và tương ứng văn bản/phát âm từ nguồn mở và cộng đồng người khiếm thính.
- Sử dụng các mô hình AI tạo sinh như Generative Adversarial Networks (GANs) hoặc Variational Autoencoders (VAEs) để phân tích và học hỏi từ dữ liệu.
- Phát triển thuật toán để nhận dạng chính xác các ký hiệu và chuyển đổi chúng thành văn bản/phát âm.
- Kiểm thử và tối ưu hệ thống thông qua các bước lặp để đạt được hiệu suất tối ưu.

1.5 Công cụ và môi trường phát triển

Công nghệ phát triển

Frontend:

- **Framework:** Angular (TypeScript)
 - Angular là một framework phát triển ứng dụng web được viết bằng ngôn ngữ TypeScript, phát triển bởi Google. Nó cung cấp các công cụ và khung làm việc cho các lập trình viên để tạo ra các ứng dụng web động và đa dạng. Angular cho phép phát triển ứng dụng theo mô hình MVVM (Model-View-ViewModel), với các tính năng như bộ lọc, các

hàm giao tiếp với API, routing, dependency injection và nhiều tính năng khác giúp đơn giản hóa quá trình phát triển ứng dụng.

- **Ngôn ngữ lập trình:** TypeScript, HTML, SCSS, JavaScript, Tex.

Backend:

- **Ngôn ngữ lập trình:** Node.js
 - Node.js là môi trường chạy JavaScript phía server, cho phép xây dựng các ứng dụng server hiệu suất cao.
- **Framework:** Express.js
 - Express.js là một framework web dành cho Node.js, được sử dụng để xây dựng các API backend hiệu quả.

Framework và Công cụ phát triển

- **Framework:**
 - **Angular:** Framework phát triển ứng dụng web, sử dụng TypeScript.
 - **Express.js:** Framework web dành cho Node.js, được sử dụng để xây dựng các API backend.
- **Công cụ phát triển:**
 - **Visual Studio Code:** Trình soạn thảo mã nguồn đa năng, hỗ trợ nhiều ngôn ngữ lập trình.
 - **Postman:** Công cụ kiểm tra API.
 - **Git:** Hệ thống quản lý mã nguồn phân tán.
 - **GitHub:** Nền tảng lưu trữ mã nguồn và cộng tác phát triển phần mềm.
- **Công cụ thiết kế:**
 - draw.io: Công cụ thiết kế sơ đồ, flowchart.

- Figma: Công cụ thiết kế giao diện người dùng (UI/UX).
- Visio: Công cụ thiết kế sơ đồ chuyên nghiệp.
- **Các bộ công cụ phát triển khác:**
 - **TensorFlow và Keras:** Thư viện học máy và deep learning.
 - **OpenPose:** Thư viện phát hiện và nhận diện tư thế của con người.
 - **Google Cloud Speech-to-Text:** Công cụ chuyển đổi giọng nói thành văn bản.
 - **Google Cloud Text-to-Speech:** Công cụ chuyển đổi văn bản thành giọng nói.
- **Công cụ quản lý mã nguồn (source code):**
 - GitHub: Nền tảng lưu trữ và quản lý mã nguồn.

CHƯƠNG 2. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

2.1 Các công trình nghiên cứu liên quan về nhận diện và dịch ngôn ngữ ký hiệu

2.1.1 Sign Language Recognition and Translation Systems for Enhanced Communication for the Hearing Impaired

Nhóm tác giả: Kambhampati Sai Sindhu cùng các cộng sự xuất bản 2024 [1] tại 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU).

Bài báo này đề cập đến những thách thức trong Hệ thống nhận dạng ngôn ngữ ký hiệu (SLR) và Biên dịch ngôn ngữ ký hiệu (SLT), tập trung vào việc biên dịch ngôn ngữ ký hiệu sang văn bản/lời nói và quá trình ngược lại. Cấu trúc ngữ pháp độc đáo của ngôn ngữ ký hiệu đặt ra một vấn đề cốt lõi, thúc đẩy sự phát triển của các mô hình tính toán. Mô-đun Biên dịch ngôn ngữ ký hiệu (SLT), biên dịch thuật ngữ được nhận dạng thành văn bản ngôn ngữ nói, đặt ra một thách thức to lớn do sự phức tạp về ngữ pháp và ngữ nghĩa. Ngoài ra, nhu cầu về các tập dữ liệu đa dạng, bao gồm nhiều phương ngữ ngôn ngữ ký hiệu khác nhau ở Ấn Độ, làm phức tạp thêm quá trình phát triển các hệ thống SLR và SLT mạnh mẽ. Bài báo khám phá những tiến bộ công nghệ, thách thức và giải pháp trong cả hai mô-đun, góp phần tạo nên các công cụ giao tiếp toàn diện.

2.1.2 Real-Time Sign Language Recognition using Deep Learning Techniques

Nhóm tác giả: Abhishek Wahane cùng các cộng sự xuất bản 2022 [2] tại International Conference on Advanced Computer Science and Information Systems (ICACSIS)

Công nghệ sử dụng:

Nhận diện cử chỉ: Hệ thống nhận diện cử chỉ sử dụng mô hình kép của Single Shot Multibox Detector (SSD) và mô hình Machine Learning dựa trên tọa độ 2D-Pose của người dùng tại thời gian thực.

Nhận diện bảng chữ cái ASL: Module này sử dụng Inception v3 của Google để học chuyển giao (transfer learning), đạt độ chính xác 89.91%.

Nghiên cứu có thể hoạt động hiệu quả trong thời gian thực, hỗ trợ đáng kể cho người khiếm thính và người có khiếm khuyết ngôn ngữ trong giao tiếp hàng ngày.

2.1.3 Sign Language Translation Across Multiple Languages

Nhóm tác giả: Sonali M. Antad và các cộng sự xuất bản 2024 [3] tại International Conference on Emerging Systems and Intelligent Computing (ESIC)

Công trình nghiên cứu này của nhóm tác giả đã phát triển thành công một nền tảng hỗ trợ dịch ngôn ngữ ký hiệu qua nhiều ngôn ngữ khác nhau. Bằng cách sử dụng Mạng nơ-ron tích chập (CNN) và các thuật toán học sâu, nền tảng này cung cấp khả năng lựa chọn giữa Ngôn ngữ Ký hiệu Ấn Độ và Ngôn ngữ Ký hiệu Mỹ, và dịch sang nhiều ngôn ngữ khu vực Ấn Độ. Điều này giúp người sử dụng ngôn ngữ ký hiệu khác nhau giao tiếp hiệu quả và cung cấp văn bản viết cho những người không hiểu ngôn ngữ ký hiệu, tạo điều kiện cho sự hòa nhập và tiếp cận thông tin trong cộng đồng người khiếm thính trên toàn cầu.

Công nghệ sử dụng:

- Mạng nơ-ron tích chập (Convolutional Neural Networks - CNN): Sử dụng CNN và các thuật toán học sâu khác để nhận diện và dịch ngôn ngữ ký hiệu.

- Dữ liệu huấn luyện: Nền tảng được huấn luyện trên bộ dữ liệu bao gồm các ký hiệu từ Ngôn ngữ Ký hiệu Ấn Độ và Ngôn ngữ Ký hiệu Mỹ, cũng như các ngôn ngữ khu vực Ấn Độ.

Kết quả đạt được:

- Nền tảng cho phép người dùng lựa chọn giữa Ngôn ngữ Ký hiệu Ấn Độ và Ngôn ngữ Ký hiệu Mỹ.
- Cung cấp bản dịch sang nhiều ngôn ngữ khu vực Ấn Độ, giúp cải thiện giao tiếp giữa người khiếm thính và cộng đồng rộng lớn hơn.

2.1.4 A two-stage sign language recognition method focusing on the semantic features of label text

Nhóm tác giả: Xuebin Xu và các cộng sự xuất bản 2024 [4] tại 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP).

Công trình nghiên cứu này của nhóm tác giả đã phát triển thành công một phương pháp nhận diện ngôn ngữ ký hiệu hai giai đoạn, tập trung vào các đặc trưng ngữ nghĩa của văn bản nhãn trong quá trình chuyển đổi cử chỉ thành GLOSS. Bằng cách triển khai mô-đun sửa lỗi văn bản và sử dụng các mô-đun nhận diện ngôn ngữ ký hiệu tiên tiến, phương pháp này đã cải thiện đáng kể độ chính xác của mô hình nhận diện. Các thử nghiệm trên các tập dữ liệu lớn như RWTHPHOENIX-Weather-2014-T và CSL đã chứng minh hiệu quả của phương pháp, mở ra các hướng đi mới cho việc phát triển các công cụ hỗ trợ giao tiếp cho cộng đồng người khiếm thính và câm.

Công nghệ sử dụng:

- Nhận diện ngôn ngữ ký hiệu: Sử dụng một mô-đun nhận diện ngôn ngữ ký hiệu để đưa ra các dự đoán ban đầu về các cử chỉ.

- Mô-đun sửa lỗi văn bản: Triển khai mô-đun sửa lỗi văn bản để sửa chữa các chuỗi GLOSS dự đoán ban đầu, nhằm đạt được kết quả nhận diện cuối cùng với độ chính xác cao hơn.
- Tập dữ liệu: Phương pháp được kiểm tra trên các tập dữ liệu RWTHPHOENIX-Weather-2014-T và CSL để đánh giá hiệu quả nhận diện ngôn ngữ ký hiệu trên quy mô lớn.

Kết quả đạt được:

- Phương pháp đề xuất đã cải thiện đáng kể độ chính xác của mô hình nhận diện ngôn ngữ ký hiệu.
- Các kết quả thực nghiệm chứng minh rằng phương pháp hai giai đoạn, với trọng tâm vào các đặc trưng ngữ nghĩa của văn bản nhàn, mang lại hiệu quả cao trong nhận diện ngôn ngữ ký hiệu.

2.1.5 Sign Language to Text Translation with Computer Vision: Bridging the Communication Gap

Nhóm tác giả: So Xue Thong và các công sự xuất bản 2024 [5] tại 3rd International Conference on Digital Transformation and Applications (ICDXA).

Công trình nghiên cứu này của nhóm tác giả đã phát triển thành công một hệ thống dịch ngôn ngữ ký hiệu sang văn bản thời gian thực bằng công nghệ thị giác máy tính. Sử dụng các mô hình CNN và LSTM, hệ thống có thể nhận diện các ký hiệu tĩnh và động với độ chính xác cao. Bên cạnh đó, việc sử dụng 'word-ninja' và Mô hình ngôn ngữ lớn (LLM) giúp phân đoạn từ và tạo câu chính xác. Hệ thống còn tích hợp chức năng dịch máy và chuyển văn bản thành giọng nói, góp phần cải thiện khả năng tiếp cận và giao tiếp của người khiếm thính. Mặc dù vẫn còn một số thách thức như điều kiện môi trường và nhận diện nhầm, hệ thống này vẫn đóng góp tích cực vào việc giảm bớt rào cản giao tiếp và thúc đẩy sự hòa nhập trong xã hội.

Công nghệ sử dụng:

- Nhận diện ký hiệu tĩnh và động: Sử dụng Mạng nơ-ron tích chập (CNN) để nhận diện các ký hiệu tĩnh và Mạng nơ-ron LSTM (Long Short-Term Memory) để nhận diện các ký hiệu động.
- Phân đoạn từ và tạo câu: Sử dụng 'word-ninja' và Mô hình ngôn ngữ lớn (LLM) để phân đoạn từ và tạo câu chính xác.
- Dịch máy và chuyển văn bản thành giọng nói: Tích hợp chức năng dịch máy và chuyển đổi văn bản thành giọng nói để cải thiện khả năng tiếp cận của hệ thống.

Kết quả đạt được:

- Hệ thống đạt được độ chính xác 99.20% cho nhận diện ký hiệu tĩnh và 90.08% cho nhận diện ký hiệu động.
- Tạo câu có độ chính xác khá cao, đạt 90% từ đầu ra của mô hình LLM.
- Mặc dù có một số vấn đề như điều kiện môi trường ảnh hưởng đến độ chính xác và nhận diện nhầm các ký hiệu tương tự, hệ thống vẫn góp phần giảm bớt rào cản giao tiếp và thúc đẩy sự hòa nhập trong xã hội.

2.2 Các công trình nghiên cứu liên quan về hệ thống chuyển đổi ngôn ngữ ký hiệu

2.2.1 A Comprehensive Application for Sign Language Alphabet and World Recognition, Text-to-Action Conversion for Learners, Multi-Language Support and Integrated Voice Output Functionality

Nhóm tác giả: D. Shofia Priyadharshini và các cộng sự xuất bản 2024 [6] tại 2024 International Conference on Science Technology Engineering and Management (ICSTEM)

Công trình nghiên cứu này của nhóm tác giả đã phát triển thành công một ứng dụng toàn diện nhằm hỗ trợ người sử dụng ngôn ngữ ký hiệu trong việc học và giao tiếp. Ứng dụng này giải quyết các thách thức trong việc giao tiếp hiệu quả với người không sử dụng ngôn ngữ ký hiệu và cải thiện trải nghiệm học tập của người dùng. Các tính năng như nhận diện bảng chữ cái và từ ngữ ngôn ngữ ký hiệu, chuyển đổi văn bản thành hành động, hỗ trợ đa ngôn ngữ và chức năng đầu ra giọng nói tích hợp giúp tạo điều kiện cho sự hòa nhập và tiếp cận thông tin, góp phần tạo nên một xã hội công bằng và bao dung hơn cho cộng đồng người khiếm thính và câm.

Công nghệ sử dụng:

- Nhận diện bảng chữ cái và từ ngữ ngôn ngữ ký hiệu: Ứng dụng tích hợp các thuật toán tiên tiến để nhận diện bảng chữ cái và từ ngữ trong ngôn ngữ ký hiệu.
- Chuyển đổi văn bản thành hành động: Tính năng này giúp người học hiểu và thực hành các cử chỉ ngôn ngữ ký hiệu dựa trên văn bản.
- Hỗ trợ đa ngôn ngữ: Ứng dụng hỗ trợ nhiều ngôn ngữ, giúp người dùng có thể giao tiếp với người không biết ngôn ngữ ký hiệu bằng ngôn ngữ của họ.
- Chức năng đầu ra giọng nói tích hợp: Ứng dụng cung cấp chức năng đầu ra giọng nói để hỗ trợ người dùng giao tiếp hiệu quả hơn.

Kết quả đạt được:

- Ứng dụng đã phát triển các thuật toán tiên tiến để nhận diện chính xác bảng chữ cái và từ ngữ ngôn ngữ ký hiệu.
- Chuyển đổi văn bản thành hành động giúp người học hiểu và thực hành ngôn ngữ ký hiệu một cách hiệu quả.
- Hỗ trợ đa ngôn ngữ giúp người dùng giao tiếp dễ dàng với người không biết ngôn ngữ ký hiệu.

- Chức năng đầu ra giọng nói tích hợp giúp tăng cường khả năng giao tiếp của người dùng.

2.2.2 Software based sign language converter

Nhóm tác giả: Keerthi S Warriar và các cộng sự tại 2016 [7] International Conference on Communication and Signal Processing (ICCSPP)

Nhóm tác giả đã phát triển thành công một công cụ chuyển đổi âm thanh thành ngôn ngữ ký hiệu dựa trên Python, nhằm cải thiện khả năng giao tiếp cho người khiếm thính và người câm. Bằng cách sử dụng các kỹ thuật Xử lý Ngôn ngữ Tự nhiên (NLP) tiên tiến, công cụ này có thể chuyển đổi đầu vào giọng nói sống thành các cử chỉ ngôn ngữ ký hiệu, tạo điều kiện cho sự hòa nhập và giao tiếp hiệu quả giữa người khiếm thính và cộng đồng không khiếm thính. Công cụ này không chỉ hỗ trợ giao tiếp mà còn là một phương tiện học tập ngôn ngữ ký hiệu hữu ích, góp phần nâng cao khả năng tiếp cận và tạo ra các kết nối ý nghĩa trong xã hội.

Công nghệ sử dụng:

- Nhận diện cử chỉ tay: Hệ thống nhận diện cử chỉ tay của người khiếm thính sử dụng phần mềm LabVIEW để phân tích và nhận dạng các cử chỉ ký hiệu ngôn ngữ (ASL Gestures).
- Chuyển đổi cử chỉ thành văn bản và giọng nói: Khi nhận diện được cử chỉ, hệ thống sẽ chuyển đổi thành văn bản tương ứng và sau đó chuyển đổi văn bản này thành giọng nói. Ví dụ, khi người khiếm thính thực hiện cử chỉ cho số 5, hệ thống sẽ hiển thị văn bản "FIVE" và phát âm từ này qua loa.
- Lập trình đồ họa: Công cụ này sử dụng lập trình đồ họa để tạo giao diện người dùng thân thiện, giúp dễ dàng sử dụng và thao tác.

Kết quả đạt được:

- Cải thiện giao tiếp: Công cụ này giúp người khiếm thính có thể giao tiếp hiệu quả hơn với người không biết ngôn ngữ ký hiệu mà không cần sự hỗ trợ của phiên dịch viên.

- Tiềm năng ứng dụng rộng rãi: Với khả năng chuyển đổi cử chỉ thành giọng nói, công cụ này có thể được tích hợp trên các nền tảng web và di động, mở rộng phạm vi sử dụng và hỗ trợ giao tiếp trong nhiều tình huống khác nhau.
- Giảm khoảng cách giao tiếp: Việc sử dụng công cụ này giúp giảm khoảng cách giao tiếp giữa người khiếm thính và người không khiếm thính, tạo điều kiện cho sự hòa nhập và giao tiếp hiệu quả hơn trong xã hội.

2.2.3 Real-Time Gesture Based Sign Language Recognition System

Nhóm tác giả: Jeet Debnath và các cộng sự xuất bản 2024 [8] tại International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS).

Công trình nghiên cứu này của nhóm tác giả đã phát triển thành công một hệ thống nhận diện ngôn ngữ ký hiệu dựa trên cử chỉ trong thời gian thực, sử dụng kỹ thuật thị giác máy tính và học sâu. Bằng cách sử dụng Python, OpenCV và MediaPipe Holistic, hệ thống có thể ước lượng chính xác tư thế tay và cơ thể, sau đó sử dụng mạng nơ-ron LSTM để nhận diện các cử chỉ ngôn ngữ ký hiệu. Hệ thống đạt được độ chính xác cao và độ trễ thấp trong nhận diện ngôn ngữ ký hiệu, giúp cải thiện giao tiếp giữa cộng đồng người điếc và người không điếc. Nghiên cứu này mở ra các hướng đi mới cho việc phát triển các công cụ giao tiếp hòa nhập và tiếp cận, đồng thời thúc đẩy sự phát triển của các ứng dụng trong giáo dục, tiếp cận và tương tác xã hội.

Công nghệ sử dụng:

- Xử lý ảnh và Thị giác máy tính: Sử dụng Python, thư viện OpenCV (Open-Source Computer Vision Library) và MediaPipe Holistic để ước lượng tư thế tay và cơ thể trong thời gian thực.
- Mạng nơ-ron LSTM (Long Short-Term Memory): Hệ thống sử dụng LSTM để xử lý các dữ liệu về cử chỉ và tư thế, nhờ khả năng mô hình hóa các chuỗi và ngữ cảnh trong ngôn ngữ ký hiệu.

- Học chuyển giao (Transfer Learning): Áp dụng kỹ thuật học chuyển giao để tinh chỉnh mô hình, cải thiện hiệu suất nhận diện ngôn ngữ ký hiệu.

Kết quả đạt được:

- Hệ thống có thể nhận diện chính xác các cử chỉ của Ngôn ngữ Ký hiệu Mỹ (ASL) trong thời gian thực, với độ chính xác cao và độ trễ thấp.
- Mô hình có khả năng học và nhận diện các cử chỉ trong các câu hoặc cụm từ, nhờ khả năng học các phụ thuộc thời gian và ngữ cảnh trong ngôn ngữ ký hiệu.
- Hệ thống đã được đánh giá và chứng minh hiệu quả trong các kịch bản thực tế, hỗ trợ tốt cho giao tiếp giữa người điếc và người không điếc.

2.2.4 Deaf talk using 3D animated sign language: A sign language interpreter using Microsoft's kinect v2

Nhóm tác giả: Mateen Ahmed và các cộng sự xuất bản 2016 [9] tại SAI Computing Conference (SAI).

Công trình nghiên cứu này của nhóm tác giả đã phát triển thành công hệ thống Deaf Talk sử dụng công nghệ Kinect for Windows V2 của Microsoft để tạo ra một công cụ phiên dịch ngôn ngữ ký hiệu và dịch ngôn ngữ nói thành ngôn ngữ ký hiệu. Hệ thống này cung cấp chế độ giao tiếp hai chiều giữa người sử dụng ngôn ngữ ký hiệu và người nói ngôn ngữ tự nhiên, giúp giảm bớt rào cản giao tiếp và tạo điều kiện cho sự hòa nhập trong xã hội. Với độ chính xác cao trong việc nhận diện và chuyển đổi, hệ thống Deaf Talk đóng góp quan trọng vào việc cải thiện giao tiếp cho cộng đồng người điếc và người có khó khăn về thính giác.

Công nghệ sử dụng:

- Kinect for Windows V2: Sử dụng cảm biến Kinect để nhận diện các cử chỉ ngôn ngữ ký hiệu và chuyển đổi chúng thành giọng nói, đồng thời chuyển đổi ngôn ngữ nói thành ngôn ngữ ký hiệu.
- Chuyển đổi cử chỉ thành giọng nói: Người sử dụng ngôn ngữ ký hiệu thực hiện các cử chỉ trong phạm vi nhìn của Kinect. Hệ thống nhận diện các cử chỉ này, so sánh với cơ sở dữ liệu đã được huấn luyện, và chuyển đổi thành từ khóa tương ứng. Sau đó, các từ khóa này được chuyển đổi thành giọng nói thông qua mô-đun chuyển đổi văn bản thành giọng nói.
- Chuyển đổi giọng nói thành ngôn ngữ ký hiệu: Người nói ngôn ngữ tự nhiên đứng trong phạm vi nhìn của Kinect và nói bằng ngôn ngữ mẹ đẻ (trong trường hợp này là tiếng Anh). Hệ thống chuyển đổi giọng nói thành văn bản, sau đó ánh xạ các từ khóa thành các cử chỉ ký hiệu 3D đã được lưu trữ trước và hiển thị các hoạt hình này trên màn hình.

Kết quả đạt được:

- Hệ thống Deaf Talk đạt được độ chính xác 87% cho việc chuyển đổi giọng nói thành ngôn ngữ ký hiệu và 84% cho việc chuyển đổi ngôn ngữ ký hiệu thành giọng nói.
- Hệ thống cung cấp chế độ giao tiếp hai chiều hiệu quả, giúp giảm bớt rào cản giao tiếp và thúc đẩy sự hòa nhập trong cộng đồng.

2.2.5 Design and Development of Teaching and Learning Tool Using Sign Language Translator to Enhance the Learning Skills for Students With Hearing and Verbal Impairment

Nhóm tác giả: Mehwish Sultana và các cộng sự xuất bản 2024 [10] tại Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE).

Công trình nghiên cứu này của nhóm tác giả đã phát triển thành công một hệ thống hỗ trợ học tập và giao tiếp thời gian thực cho học sinh khiếm thính và câm. Sử dụng công nghệ MediaPipe và kiến trúc LSTM, hệ thống cho phép chuyển đổi chính xác giữa Ngôn ngữ Ký hiệu Ấn Độ (ISL) và văn bản. Ứng dụng web thân thiện với người dùng giúp tăng cường khả năng truy cập và sử dụng, góp phần thúc đẩy sự hòa nhập và khả năng giao tiếp. Nghiên cứu này đại diện cho một bước tiến quan trọng trong công nghệ trợ giúp, thúc đẩy sự hòa nhập và kết nối trong xã hội, bất kể khả năng thính giác của họ.

Công nghệ sử dụng:

- MediaPipe: Sử dụng MediaPipe để trích xuất các điểm chính toàn diện, bao gồm các chuyển động của tay và biểu cảm khuôn mặt.
- Kiến trúc LSTM (Long Short-Term Memory): Sử dụng kiến trúc LSTM kết hợp với TensorFlow và Keras để diễn giải chính xác ngôn ngữ ký hiệu.
- Chuyển đổi Ngôn ngữ Ký hiệu sang văn bản và ngược lại: Hệ thống cho phép người dùng nhập văn bản và chuyển đổi thành hoạt hình ngôn ngữ ký hiệu, và ngược lại, chuyển đổi ngôn ngữ ký hiệu thành văn bản để đảm bảo giao tiếp mượt mà.
- Ứng dụng web thân thiện với người dùng: Phát triển ứng dụng web sử dụng HTML, CSS, và JavaScript để tăng cường khả năng truy cập và sử dụng dễ dàng cho giao tiếp thời gian thực.

Kết quả đạt được:

- Hệ thống cho phép chuyển đổi thời gian thực giữa Ngôn ngữ Ký hiệu Ấn Độ (ISL) và văn bản, cải thiện khả năng học tập và giao tiếp cho học sinh khiếm thính và câm.
- Tính năng chuyển đổi giữa ngôn ngữ ký hiệu và văn bản giúp người không biết ngôn ngữ ký hiệu có thể giao tiếp tự nhiên với người sử dụng ngôn ngữ ký hiệu.
- Ứng dụng web thân thiện với người dùng giúp tăng cường khả năng truy cập và sử dụng, thúc đẩy giao tiếp hiệu quả.

2.3 Các công trình nghiên cứu liên quan tại Việt Nam

2.3.1. SOS - Máy phiên dịch ngôn ngữ ký hiệu dành cho người khiếm thính

Nhóm tác giả: Nhóm sinh viên Trường Đại học Bách Khoa - ĐHQG-HCM (HCMUT) – Vòng bình chọn SV.STARTUP 2021

“Thiết bị giao tiếp thông minh dành cho người khiếm thanh, khiếm thính - Sound of Silence” (thiết bị SOS) là thiết bị giao tiếp chuyển ngôn ngữ ký hiệu sang văn bản và giọng nói để những người khiếm thanh, khiếm thính dễ dàng trao đổi thông tin với mọi người.

05 sự khác biệt tạo nên SOS đó là:

- Tính tiên phong: Sản phẩm chưa có trên thị trường Việt Nam.
- Tính cấp thiết: Những người khuyết tật nói chung, người khiếm thanh, khiếm thính nói riêng rất ít được tiếp cận các dịch vụ y tế, giáo dục, tỷ lệ thất nghiệp cao và trên cả nước ta số lượng phiên dịch ngôn ngữ ký hiệu chuyên nghiệp lại chỉ có khoảng hơn 10 người. Vì vậy sản phẩm chuyển thủ ngữ (ngôn ngữ ký hiệu) sang văn bản và chữ viết sẽ mang lại sự hỗ trợ giao tiếp đáng kể cho người khiếm thanh, khiếm thính.
- Phương pháp xử lý: Áp dụng các thuật toán học sâu và thị giác máy tính để nhận diện và phiên dịch ngôn ngữ ký hiệu trên phần cứng nhỏ gọn với độ chính xác cao.
- Thiết kế: Mẫu sản phẩm được nghiên cứu phát triển nhằm tối ưu hoá tính năng mà không cản trở hoạt động sinh hoạt thường ngày.
- Chi phí: Giá thành sản phẩm thấp phù hợp với mức thu nhập người khuyết tật nghe, nói.

Kết quả đạt được:

- Dự án là giải pháp áp dụng công nghệ trí tuệ nhân tạo AI để chuyển đổi ngôn ngữ ký hiệu sang văn bản và giọng nói tiếng Việt, góp phần giúp người khiếm thanh khiếm thính dễ dàng giao tiếp với mọi người xung quanh. Từ mã nguồn đến thiết kế sản phẩm đều do nhóm tự nghiên cứu chế tạo. SOS mong muốn đem đến một sản phẩm có tính thiết thực, nhân văn sâu sắc, giúp cộng đồng người khiếm thanh khiếm thính ở Việt Nam dễ dàng hòa nhập với cộng đồng, tiếp cận các dịch vụ xã hội như giáo dục, y tế và cơ hội phát triển bản thân toàn diện như những người bình thường khác.
- Kế hoạch sản xuất về quy trình, công nghệ là hoàn toàn khả thi và sẵn sàng triển khai với số lượng phù hợp nhu cầu của thị trường.
- Cơ cấu chi phí, giá thành sản phẩm là hoàn toàn khả thi và có thể mang lại lợi nhuận cao.
- Sản phẩm hiện tại chưa có trên thị trường Việt Nam nên chưa có đối thủ cạnh tranh. Sản phẩm có giá trị vì cộng đồng, đóng vai trò tiên phong trong việc xóa bỏ rào cản giao tiếp với người khiếm thanh, khiếm thính.
- Sản phẩm hướng đến đối tượng là những người khiếm thanh, khiếm thính, đáp ứng được nhu cầu của khách hàng thông qua các khảo sát thị trường nhu cầu tại Việt Nam.

2.3.2. Thiết bị giao tiếp thông minh dành cho người khiếm thanh, khiếm thính – “Speak your mind” (SYM)

Nhóm tác giả: Nhóm sinh viên trường ĐH Bách khoa (ĐHQG TP. HCM)

“Thiết bị giao tiếp thông minh dành cho người khiếm thanh, khiếm thính – Speak your mind” (SYM) là thiết bị giao tiếp sử dụng công nghệ AI để chuyển ngôn ngữ ký hiệu sang văn bản và giọng nói phát ra từ các thiết bị điện thoại thông minh để những người khiếm thanh, khiếm thính dễ dàng trao đổi thông tin với mọi người. Sản phẩm là một thiết bị tích hợp phần mềm thông minh nhằm hỗ trợ người khiếm thanh có thể tương tác với mọi người một cách dễ dàng và thuận tiện hơn bằng cách chuyển đổi thủ ngữ thành văn bản và giọng nói. Thiết bị bao gồm hai thành phần: mô-đun camera được gắn trên nón và ứng dụng trên điện thoại Android.

Kết quả đạt được:

- Làm chủ công nghệ sản xuất, toàn bộ mã nguồn do nhóm tự xây dựng, thiết kế, hoàn toàn chưa có trên thị trường Việt Nam cũng như trên thế giới.
- Toàn bộ sản phẩm được mã hóa bằng tiếng Việt, dễ dàng sử dụng đối với khách hàng. Bên cạnh đó, thiết bị cũng tích hợp các ngôn ngữ khác nhau đáp ứng các đối tượng khách hàng trên toàn thế giới.

- Là sản phẩm với những tính năng đáp ứng được những yêu cầu của khách hàng thông qua các khảo sát thị trường nhu cầu tại Việt Nam.
- Chất lượng sản phẩm tốt, giá cả hợp lý phù hợp cho người thu nhập thấp, thích hợp cho mọi độ tuổi, thiết kế nhỏ gọn, dễ lắp đặt, dễ mang theo bên người mọi lúc mọi nơi.

2.3.3. SIGNTEGRATE - Ứng dụng dịch thuật Ngôn ngữ kí hiệu sử dụng trí tuệ nhân tạo Việt Nam

Nhóm tác giả: Nhóm học sinh Sở Giáo dục và Đào tạo Hà Nội – Vòng bình chọn SV.STARTUP 2021

Signtegrate là ứng dụng phiên dịch Ngôn ngữ kí hiệu Hà Nội thành ngôn ngữ Đọc-Viết tiếng Việt đầu tiên tại Việt Nam.

Sử dụng công nghệ nhận diện Trích chọn Khung xương với độ chính xác cao khi nhận diện trong thời gian thực do Google sáng chế nhằm chức năng dịch thuật video thao tác Ngôn ngữ kí hiệu Hà Nội sang văn bản Tiếng Việt.

Nhóm dự án Signtegrate đưa ra những tính mới và ưu việt như:

- Là ứng dụng dịch thuật Ngôn ngữ kí hiệu Hà Nội có chức năng xử lý tiếng Việt đầu tiên được phát triển và đưa vào thử nghiệm dành cho người Khiếm Thính tại Việt Nam.
- Là ứng dụng có khả năng xử lý nhận dạng video động trong thời gian thực đầu tiên của Việt Nam.
- Là phần mềm đầu tiên nhận dạng dịch thuật Ngôn ngữ kí hiệu theo câu, đồng thời độ chính xác lên đến 90%, không bị ảnh hưởng bởi yếu tố gây nhiễu nền như ánh sáng, môi trường hay góc nghiêng không đáng kể khi quay.
- Giao diện bắt mắt, thân thiện với người dùng.
- Phục vụ nhu cầu giao tiếp lớn còn đang gặp khó khăn của cộng đồng người sử dụng Ngôn ngữ kí hiệu tại Việt Nam, đặc biệt trong hoàn cảnh dịch bệnh Covid-19 gây khó khăn cho giao tiếp và hòa nhập trực tuyến trên toàn cầu.

Sản phẩm dịch vụ là hoàn toàn mới chưa có trên thị trường. Là giải pháp ý nghĩa và thiết thực đầu tiên cho bài toán không còn mới mẻ nhưng vẫn vô cùng nan giải của xã hội suốt hàng thập kỉ qua, vấn đề giao tiếp còn gặp khó khăn của cộng đồng người Khiếm Thính tại Việt Nam.

Kết quả đạt được:

- Xây dựng thành công ứng dụng demo Signtegrate.
- Phần mềm đã chạy trên thời gian thực, dịch thành công 10 câu thoại phổ biến trong xã giao thường ngày và có độ chính xác tốt (90%).
- Quá trình nhận dạng trên thời gian thực nhanh: hiển thị nội dung của hành động sau 1 giây kể từ khi mô hình đã thu đủ số lượng khung hình (120) ở camera.
- Nhóm đang trong quá trình xây dựng kế hoạch khởi nghiệp, xúc tiến thương mại.

2.3 Phương pháp đề xuất

Dựa vào các công trình nghiên cứu liên quan, nhóm đề xuất xây dựng một hệ thống dịch ngôn ngữ nói sang ngôn ngữ ký hiệu (Spoken to Signed Language Translation) theo các bước sau đây:

1. Nhận dạng giọng nói và xác định ngôn ngữ:

- Trước hết, hệ thống sẽ sử dụng các mô hình nhận dạng giọng nói tiên tiến như Whisper hoặc các giải pháp của Google để chuyển đổi giọng nói thành văn bản.
- Tiếp theo, sử dụng các mô hình nhận diện ngôn ngữ tự động như Google's cld3 hoặc MediaPipe Solutions để xác định ngôn ngữ của văn bản đầu vào. Việc này đảm bảo hệ thống có thể xử lý nhiều ngôn ngữ khác nhau, cải thiện trải nghiệm người dùng và độ chính xác của dịch.
- Văn bản đầu vào sẽ được chuẩn hóa thông qua mô hình LLM (Large Language Models) để đảm bảo chất lượng dịch thuật cao hơn.

2. Dịch văn bản sang ngôn ngữ ký hiệu:

- Văn bản chuẩn hóa sẽ được dịch sang SignWriting bằng các mô hình dịch máy như Conditional Variational Autoencoder (CVAE).
- Sau đó, các mô hình như SMPL-X sẽ được sử dụng để chuyển đổi SignWriting thành chuỗi Pose, tạo ra các hình ảnh động 3D cho ngôn ngữ ký hiệu.

3. Tạo hình ảnh động 3D và avatar sống động:

- Các mô hình Mạng thần kinh tích chập (Convolutional Neural Network) và Generative Adversarial Networks (GAN) sẽ giúp tạo ra các avatar 3D sống động, hỗ trợ người dùng dễ dàng theo dõi và hiểu ngôn ngữ ký hiệu hơn.

4. Tích hợp và triển khai hệ thống:

- Cuối cùng, hệ thống sẽ được tích hợp và triển khai trên nền tảng web.
- Thực hiện thử nghiệm và tinh chỉnh để đạt được hiệu suất tối ưu và phản hồi tích cực từ người dùng.

CHƯƠNG 3. CƠ SỞ LÝ THUYẾT

3.1 Giới thiệu ngôn ngữ ký hiệu

Ngôn ngữ ký hiệu (Signed Language - còn gọi là Sign Language) là ngôn ngữ sử dụng phương thức cử chỉ - hình ảnh để truyền đạt ý nghĩa thông qua cách phát âm thủ công kết hợp với các yếu tố không thủ công như khuôn mặt và cơ thể. Chúng đóng vai trò là phương tiện liên lạc chính của nhiều người khiếm thính và khiếm thính. Tương tự như ngôn ngữ nói, ngôn ngữ ký hiệu là ngôn ngữ tự nhiên được điều chỉnh bởi một bộ quy tắc ngôn ngữ [4], cả hai đều xuất hiện thông qua một quá trình lão hóa kéo dài, trù tượng và phát triển tự nhiên theo thời gian. Các ngôn ngữ ký hiệu không phổ biến hoặc không thể hiểu được lẫn nhau, mặc dù chúng thường có những điểm tương đồng nổi bật giữa chúng. Chúng cũng khác biệt với các ngôn ngữ nói, tức là Ngôn ngữ ký hiệu Mỹ (ASL) không phải là một dạng tiếng Anh trực quan mà là ngôn ngữ độc đáo của riêng nó.

Xử lý ngôn ngữ ký hiệu (Bragg và cộng sự, 2019 [5] ; Yin và cộng sự, 2021 [6]) là một lĩnh vực mới nổi của trí tuệ nhân tạo liên quan đến việc xử lý và phân tích tự động nội dung ngôn ngữ ký hiệu. Mặc dù nghiên cứu tập trung nhiều hơn vào các khía cạnh trực quan của ngôn ngữ ký hiệu, nhưng đây là một lĩnh vực phụ của cả Xử lý ngôn ngữ tự nhiên (NLP) và Thị giác máy tính (CV). Những thách thức trong xử lý ngôn ngữ ký hiệu thường bao gồm dịch máy các video ngôn ngữ ký hiệu thành văn bản ngôn ngữ nói (dịch ngôn ngữ ký hiệu), từ văn bản ngôn ngữ nói (sản xuất ngôn ngữ ký hiệu) hoặc nhận dạng ngôn ngữ ký hiệu để hiểu ngôn ngữ ký hiệu.

Thật không may, những tiến bộ mới nhất trong trí tuệ nhân tạo dựa trên ngôn ngữ, như dịch máy và trợ lý cá nhân, chỉ yêu cầu đầu vào là ngôn ngữ nói (văn bản hoặc lời nói đã phiên âm), loại trừ khoảng 200 đến 300 ngôn ngữ ký hiệu khác nhau (Liên hợp quốc 2022 [7]) và tới 70 triệu người khiếm thính (Tổ

chức Y tế Thế giới 2021 [8] ; Liên đoàn Người khiếm thính Thế giới 2022 [9]) .

Trong suốt chiều dài lịch sử, cộng đồng người khiếm thính đã đấu tranh cho quyền được học và sử dụng ngôn ngữ ký hiệu và cho sự công nhận công khai của ngôn ngữ ký hiệu là ngôn ngữ hợp pháp. Ngôn ngữ ký hiệu là phương thức giao tiếp tinh vi, ít nhất là có khả năng như ngôn ngữ nói ở mọi khía cạnh, cả về ngôn ngữ và xã hội. Tuy nhiên, trong một xã hội chủ yếu là truyền miệng, người khiếm thính liên tục được khuyến khích sử dụng ngôn ngữ nói thông qua việc đọc khẩu hình hoặc giao tiếp dựa trên văn bản. Việc loại trừ ngôn ngữ ký hiệu khỏi các công nghệ ngôn ngữ hiện đại càng làm hạn chế việc ký hiệu để ủng hộ ngôn ngữ nói. Việc loại trừ này bỏ qua sở thích của cộng đồng người khiếm thính, những người rất thích giao tiếp bằng ngôn ngữ ký hiệu cả trực tuyến và trong các tương tác hàng ngày trực tiếp, giữa họ với nhau và khi tương tác với các cộng đồng ngôn ngữ nói (CA Padden và Humphries 1988 [10]; Glickman và Hall 2018 [11]) . Do đó, việc làm cho ngôn ngữ ký hiệu dễ tiếp cận là điều cần thiết.

Cho đến nay, một lượng lớn nghiên cứu về Xử lý ngôn ngữ ký hiệu (Sign language Processing - SLP) đã tập trung vào khía cạnh trực quan của ngôn ngữ ký hiệu, do cộng đồng Thị giác máy tính (CV) dẫn đầu, với rất ít sự tham gia của NLP. Trọng tâm này không phải là không có lý, vì một thập kỷ trước, chúng ta thiếu các công cụ CV thích hợp để xử lý video nhằm phân tích ngôn ngữ sâu hơn. Tuy nhiên, tương tự như ngôn ngữ nói, ngôn ngữ ký hiệu là hệ thống hoàn chỉnh thể hiện tất cả các đặc điểm cơ bản của ngôn ngữ tự nhiên và các kỹ thuật SLP hiện tại không giải quyết hoặc tận dụng đầy đủ cấu trúc ngôn ngữ của ngôn ngữ ký hiệu. Các ngôn ngữ ký hiệu đưa ra những thách thức mới cho NLP do phương thức cử chỉ thị giác, tính đồng thời, sự gắn kết về không gian và thiếu hình thức viết. Việc thiếu dạng văn bản khiến cho quy trình xử lý ngôn ngữ nói - thường bắt đầu bằng phiên âm âm thanh trước khi

xử lý - không tương thích với ngôn ngữ ký hiệu, buộc các nhà nghiên cứu phải làm việc trực tiếp trên tín hiệu video thô.

Hơn nữa, SLP không chỉ hấp dẫn về mặt trí tuệ mà còn là một lĩnh vực nghiên cứu quan trọng có tiềm năng to lớn mang lại lợi ích cho cộng đồng ngôn ngữ ký hiệu. Các ứng dụng có lợi được hỗ trợ bởi công nghệ ngôn ngữ ký hiệu bao gồm cải thiện tài liệu về các ngôn ngữ ký hiệu đang bị đe dọa; các công cụ giáo dục dành cho người học ngôn ngữ ký hiệu; các công cụ để truy vấn và truy xuất thông tin từ các video ngôn ngữ ký hiệu; trợ lý cá nhân phản ứng với ngôn ngữ ký hiệu; phiên dịch ngôn ngữ ký hiệu tự động theo thời gian thực; và nhiều hơn nữa. Điều cần làm là khi giải quyết lĩnh vực nghiên cứu này, các nhà nghiên cứu nên làm việc cùng và dưới sự chỉ đạo của cộng đồng người khiếm thính, và trên hết là mang lại lợi ích cho cộng đồng ngôn ngữ ký hiệu (Harris, Holmes và Mertens 2009 [12]).

Trong báo cáo nghiên cứu này, nhóm này mô tả các cách biểu diễn khác nhau được sử dụng để xử lý ngôn ngữ ký hiệu, cũng như khảo sát các nhiệm vụ khác nhau và những tiến bộ gần đây về chúng.

3.2 Lịch sử của ngôn ngữ ký hiệu và văn hóa người khiếm thính

Trong suốt lịch sử hiện đại, ngôn ngữ nói chiếm ưu thế, đến mức ngôn ngữ ký hiệu phải vật lộn để được công nhận là ngôn ngữ riêng và các nhà giáo dục đã phát triển những quan niệm sai lầm rằng việc học ngôn ngữ ký hiệu có thể cản trở sự phát triển của các kỹ năng nói. Ví dụ, vào năm 1880, một hội nghị quốc tế lớn của các nhà giáo dục khiếm thính có tên là "Đại hội quốc tế lần thứ hai về giáo dục người khiếm thính" ("Second International Congress on Education of the Deaf") đã cấm giảng dạy ngôn ngữ ký hiệu, thay vào đó ủng hộ liệu pháp ngôn ngữ. Phải đến khi công trình quan trọng về Ngôn ngữ ký hiệu Hoa Kỳ (ASL) của Stokoe Jr (1960 [13]) thì ngôn ngữ ký hiệu mới bắt đầu được công nhận là ngôn ngữ tự nhiên, độc lập và được định nghĩa rõ ràng, điều này đã truyền cảm hứng cho các nhà nghiên cứu khác khám phá thêm ngôn ngữ

ký hiệu như một lĩnh vực nghiên cứu. Tuy nhiên, thái độ lỗi thời coi nhẹ ngôn ngữ ký hiệu vẫn tiếp tục gây hại và khiến nhiều người bị bỏ bê ngôn ngữ (Humphries và cộng sự 2016 [14]). Một số nghiên cứu đã chỉ ra rằng trẻ em khiếm thính chỉ lớn lên bằng ngôn ngữ nói, không được tiếp cận đủ với ngôn ngữ đầu tiên trong giai đoạn quan trọng của quá trình học ngôn ngữ (Murray, Hall và Snoddon 2020 [15]). Sự thiếu hụt ngôn ngữ này có thể dẫn đến hậu quả suốt đời đối với sự phát triển nhận thức, ngôn ngữ, xã hội, cảm xúc và học tập của người khiếm thính (Hall, Levin và Anderson 2017 [16]).

Ngôn ngữ ký hiệu là ngôn ngữ giao tiếp chính của người điếc và là trung tâm của cộng đồng người điếc. Trước đây, việc không công nhận ngôn ngữ ký hiệu là hệ thống ngôn ngữ tự nhiên hoàn chỉnh theo đúng nghĩa của chúng đã gây ra những tác động bất lợi và trong một thế giới ngày càng số hóa, nghiên cứu NLP nên cố gắng tạo ra một thế giới mà tất cả mọi người, bao gồm cả người điếc, đều có thể tiếp cận với các ngôn ngữ phù hợp với trải nghiệm sống của họ.

3.3 Tổng quan về ngôn ngữ kí hiệu

Ngôn ngữ ký hiệu bao gồm các cấp độ cấu trúc âm vị, hình thái, cú pháp và ngữ nghĩa đáp ứng các mục đích xã hội, nhận thức và giao tiếp giống như các ngôn ngữ tự nhiên khác. Trong khi ngôn ngữ nói chủ yếu truyền tải phương thức thính giác, miệng, thì ngôn ngữ ký hiệu lại sử dụng phương thức thị giác - cử chỉ, dựa vào khuôn mặt, bàn tay, cơ thể và không gian xung quanh của người ký để tạo ra sự khác biệt về ý nghĩa. Báo cáo trình bày các đặc điểm ngôn ngữ của ngôn ngữ ký hiệu mà các nhà nghiên cứu phải xem xét trong quá trình lập mô hình của họ.

Ngữ âm học

Các dấu hiệu bao gồm các đơn vị tối thiểu kết hợp các đặc điểm thủ công như cấu hình bàn tay, hướng lòng bàn tay, vị trí, tiếp xúc, chuyển động đường đi,

chuyển động cục bộ cũng như các đặc điểm không thủ công bao gồm khẩu độ mắt, chuyển động đầu và định vị thân (Liddell và Johnson 1989 [17]; Johnson và Liddell 2011 [17]; Sandler 2012 [18] ; Không phải tất cả các âm vị có thể có đều được thể hiện bằng cả ngôn ngữ ký hiệu và ngôn ngữ nói, đồng thời việc kiểm kê các âm vị/đặc điểm của hai ngôn ngữ có thể không trùng lặp hoàn toàn. Các ngôn ngữ khác nhau cũng phải tuân theo các quy tắc về sự kết hợp các tính năng được phép.

Tính đồng thời

Mặc dù một ký hiệu ASL mất khoảng gấp đôi thời gian để tạo ra một từ tiếng Anh, nhưng tốc độ truyền tải thông tin giữa hai ngôn ngữ là tương tự nhau (Bellugi và Fischer 1972 [19]) . Một cách ngôn ngữ ký hiệu bù đắp cho tốc độ tạo ra ký hiệu chậm hơn là thông qua tính đồng thời: Ngôn ngữ ký hiệu sử dụng nhiều tín hiệu thị giác để truyền đạt thông tin khác nhau cùng một lúc (Sandler 2012 [20]) . Ví dụ, người ký hiệu có thể tạo ra ký hiệu cho "cốc" bằng một tay trong khi đồng thời chỉ vào cốc thực tế bằng tay kia để diễn đạt "cái cốc đó". Tương tự như giọng điệu trong ngôn ngữ nói, khuôn mặt và thân mình có thể truyền đạt thông tin tình cảm bổ sung (Liddell 2003 [21]; Johnston và Schembri [22]) . Biểu cảm khuôn mặt có thể sửa đổi tính từ, trạng từ và động từ; một cái lắc đầu có thể phủ định một cụm từ hoặc câu; hướng mắt có thể giúp chỉ ra người tham chiếu.

Tham chiếu

Người nói ngôn ngữ ký hiệu có thể giới thiệu người tham chiếu trong bài phát biểu bằng cách chỉ vào vị trí thực tế của họ trong không gian hoặc bằng cách chỉ định một vùng trong không gian ký hiệu cho một người tham chiếu không có mặt và bằng cách chỉ vào vùng này để tham chiếu đến người đó (Rathmann và Mathur 2011 [23] ; Schembri, Cormier và Fenlon 2018 [24]) . Người nói ngôn ngữ ký hiệu cũng có thể thiết lập mối quan hệ giữa những người tham chiếu dựa trên không gian ký hiệu bằng cách sử dụng các dấu hiệu chỉ hướng

hoặc thể hiện người tham chiếu bằng cách dịch chuyển cơ thể hoặc nhìn chăm chăm (Dudis 2004 [25] ; Liddell và Metzger 1998 [26]) . Tham chiếu không gian cũng tác động đến hình thái khi tính hướng của động từ phụ thuộc vào vị trí của người tham chiếu đến chủ ngữ và/hoặc tân ngữ của nó (Beuzeville 2008 [27] ; Fenlon, Schembri và Cormier 2018 [28]) : Ví dụ, một động từ chỉ hướng có thể di chuyển từ vị trí chủ ngữ của nó và kết thúc ở vị trí tân ngữ của nó. Trong khi mỗi quan hệ giữa tham chiếu và động từ trong ngôn ngữ nói mang tính tùy ý hơn, mỗi quan hệ tham chiếu thường dựa trên ngôn ngữ ký hiệu.

Một cách khác mà các thực thể ẩn dụ được tham chiếu trong ngôn ngữ ký hiệu là sử dụng các bộ phân loại hoặc mô tả các dấu hiệu (Supalla 1986 [29] ; Wilcox và Hafer 2004 [30]; Roy 2011 [31]) giúp mô tả các đặc điểm của vật được ám chỉ. Các bộ phân loại thường là các dấu hiệu bằng một tay không có vị trí hoặc chuyển động cụ thể được gán cho chúng hoặc rút ra các đặc điểm từ diễn ngôn có ý nghĩa (Liddell và những người khác 2003 [21]) , vì vậy chúng có thể được sử dụng để truyền đạt mối liên hệ của vật được đề cập với các thực thể khác, mô tả nó như thế nào. chuyển động và cung cấp thêm thông tin chi tiết. Ví dụ: để kể về một chiếc ô tô chuyển hướng và đâm vào, người ta có thể sử dụng bộ phân loại bằng tay cho một phương tiện, di chuyển nó để biểu thị sự chuyển hướng và đâm nó với một thực thể khác trong không gian.

Để trích dẫn một người nào đó không phải chính mình, người ký thực hiện chuyển đổi vai trò (Cormier, Smith và Sevcikova-Sehyr 2015 [32]) , nơi họ có thể dịch chuyển vật lý trong không gian để đánh dấu sự khác biệt và thể hiện một số đặc điểm của những người mà họ đại diện. Ví dụ, để kể lại đoạn hội thoại giữa người cao hơn và người thấp hơn, người ra dấu có thể dịch sang một bên và nhìn lên khi đảm nhận vai người thấp hơn, chuyển sang phía bên kia và nhìn xuống khi đảm nhận vai người cao hơn.

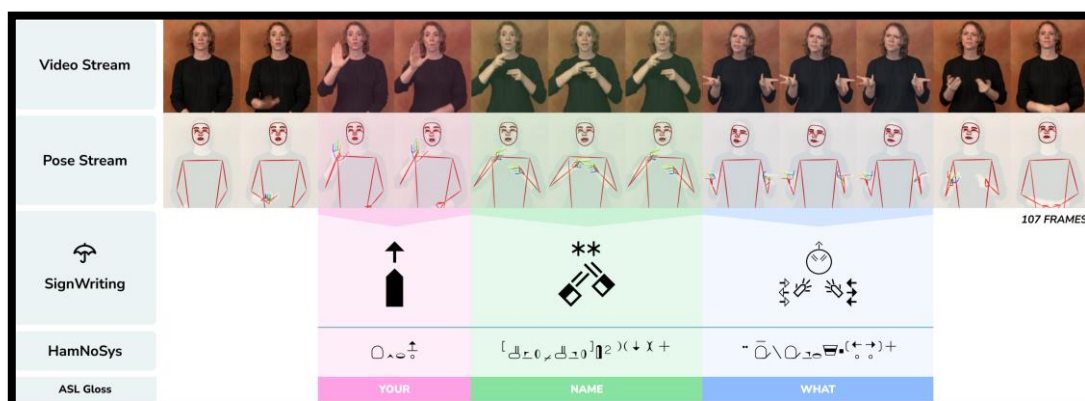
Đánh vần bằng ngón tay

Fingerspelling là kết quả của sự tiếp xúc ngôn ngữ giữa ngôn ngữ ký hiệu và ngôn ngữ nói xung quanh dạng viết (Battison 1978 [33]; Wilcox 1992 [34]; Brentari và Padden 2001 [35]; Patrie và Johnson 2011 [36]). Một tập hợp các cử chỉ bằng tay tương ứng với một hệ thống chính tả hoặc ngữ âm viết. Hiện tượng này, được tìm thấy trong hầu hết các ngôn ngữ ký hiệu, thường được sử dụng để chỉ tên hoặc địa điểm hoặc khái niệm mới từ ngôn ngữ nói nhưng thường được tích hợp vào ngôn ngữ ký hiệu như một chiến lược ngôn ngữ khác (Padden 1998 [37]; Montemurro và Brentari 2018 [38]).

3.4 Biểu diễn ngôn ngữ ký hiệu

Biểu diễn là một thách thức đáng kể đối với SLP. Không giống như ngôn ngữ nói, ngôn ngữ ký hiệu không có dạng viết được chấp nhận rộng rãi. Vì ngôn ngữ ký hiệu được truyền đạt thông qua phương thức thị giác-cử chỉ, nên ghi hình video là cách trực tiếp nhất để ghi lại chúng. Tuy nhiên, vì video bao gồm nhiều thông tin hơn mức cần thiết để mô hình hóa và tốn kém để ghi hình, lưu trữ và truyền tải, nên người ta đã tìm kiếm một biểu diễn có chiều thấp hơn.

Hình dưới đây minh họa từng biểu diễn ngôn ngữ ký hiệu mà báo cáo sẽ mô tả bên dưới. Trong phần trình diễn này, báo cáo giải cấu trúc video thành các khung riêng lẻ để minh họa sự liên kết của các chú thích giữa video và các phần trình bày.



Hình 1 Hình minh họa từng biểu diễn ngôn ngữ ký hiệu trong báo cáo

Video

Là cách thể hiện trực tiếp nhất của ngôn ngữ ký hiệu và có thể kết hợp đầy đủ thông tin được truyền tải thông qua ký hiệu. Một nhược điểm lớn của việc sử dụng video là tính đa chiều cao của chúng: Chúng thường bao gồm nhiều thông tin hơn mức cần thiết để mô hình hóa và tốn kém để lưu trữ, truyền tải và mã hóa. Vì các đặc điểm khuôn mặt là yếu tố thiết yếu trong ký hiệu, việc ẩn danh video thô vẫn là một vấn đề chưa có lời giải, hạn chế khả năng công khai các video này (Isard 2020).

Tư thế bộ xương

Giảm các tín hiệu thị giác trong video thành các khung lưới hoặc lưới giống như bộ xương thể hiện vị trí của các khớp. Kỹ thuật này đã được sử dụng rộng rãi trong lĩnh vực thị giác máy tính để ước tính tư thế của con người từ dữ liệu video, trong đó mục tiêu là xác định cấu hình không gian của cơ thể tại mỗi thời điểm. Mặc dù có thể đạt được ước tính tư thế chất lượng cao bằng cách sử dụng thiết bị ghi lại chuyển động, nhưng những phương pháp như vậy thường tốn kém và tốn nhiều công sức. Do đó, việc ước tính tư thế từ video đã trở thành phương pháp được ưa chuộng trong những năm gần đây (Pishchulin và cộng sự 2012; Chen và cộng sự 2017; Cao và cộng sự 2019; Güler, Neverova và Kokkinos 2018) . So với các hình ảnh thể hiện bằng video, các tư thế bộ xương chính xác có độ phức tạp thấp hơn và cung cấp hình ảnh thể hiện bản ẩn danh của cơ thể con người, đồng thời quan sát thấy mức độ mất thông tin tương đối thấp. Tuy nhiên, chúng vẫn là một biểu diễn đa chiều, liên tục và không phù hợp với hầu hết các mô hình NLP.

Hệ thống ký hiệu chữ viết

Biểu diễn các dấu hiệu như các đặc điểm trực quan rời rạc. Một số hệ thống được viết theo dạng tuyến tính, và một số khác sử dụng các ký tự chữ cái trong

hai chiều. Trong khi nhiều hệ thống ký hiệu phổ quát (Sutton 1990; Prillwitz và Zienert 1990) và ký hiệu dành riêng cho ngôn ngữ (Stokoe Jr [1060](#) ; Kakumasu 1968 ; Bergman 1977) đã được đề xuất, không có hệ thống chữ viết nào được bất kỳ cộng đồng ngôn ngữ ký hiệu nào áp dụng rộng rãi và việc thiếu các tiêu chuẩn cản trở việc trao đổi và thống nhất các nguồn lực và ứng dụng giữa các dự án. Hình trên mô tả hai hệ thống ký hiệu phổ quát: SignWriting (Sutton 1990) , một hệ thống tượng hình hai chiều và HamNoSys (Prillwitz và Zienert 1990) , một luồng ký tự chữ cái tuyến tính được thiết kế để máy có thể đọc được.

Bóng (Glosses)


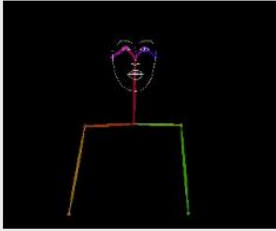


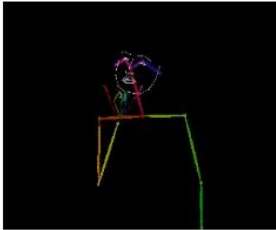




Là phiên âm của các ngôn ngữ ký hiệu theo từng ký hiệu, với mỗi ký hiệu có một mã định danh ngữ nghĩa duy nhất. Trong khi các dự án kho ngữ liệu ngôn ngữ ký hiệu khác nhau đã cung cấp hướng dẫn về chú thích độ bóng (Mesch và Wallin 2015; Johnston và De Beuzeville 2016; Konrad và cộng sự 2018) , thì một giao thức chú thích độ bóng được tiêu chuẩn hóa vẫn chưa được thiết lập. Chú thích độ bóng tuyến tính đã bị chỉ trích vì thể hiện ngôn ngữ ký hiệu không chính xác. Các chú thích này không nắm bắt được tất cả thông tin được thể hiện đồng thời thông qua các tín hiệu khác nhau, chẳng hạn như tư thế cơ thể, ánh mắt hoặc quan hệ không gian, dẫn đến mất thông tin có thể ảnh hưởng đáng kể đến hiệu suất ở các nhiệm vụ SLP sau này (Yin and Read 2002; Müller et al . 2013) .

Müller và cộng sự. (2023) tiến hành đánh giá sâu rộng việc sử dụng thuật ngữ trong nghiên cứu dịch thuật ngôn ngữ ký hiệu và đưa ra những khuyến nghị sau đây cho nghiên cứu sử dụng thuật ngữ:

- Thể hiện nhận thức về những hạn chế của các phương pháp tiếp cận độ bóng và thảo luận rõ ràng về chúng.

- Tập trung vào các tập dữ liệu ngoài RWTH-PHOENIX-Weather-2014T (Camgöz và cộng sự, 2018) . Thảo luận cởi mở về quy mô hạn chế và phạm vi ngôn ngữ của tập dữ liệu này.
- Sử dụng các số liệu được thiết lập tốt trong MT. Nếu sử dụng BLEU (Papineni et al. 2002) , hãy tính toán bằng SacreBLEU (Bài đăng 2018) , báo cáo chữ ký số liệu và vô hiệu hóa mã thông báo nội bộ cho đầu ra gloss. Không so sánh với điểm số được tạo ra bằng quy trình đánh giá khác hoặc không xác định.
- Do chú thích là đặc thù của từng ngữ liệu nên cần xử lý chú thích theo cách đặc thù của từng ngữ liệu, dựa trên các quy ước phiên âm.
- Tối ưu hóa các đường cơ sở dịch bóng bằng các phương pháp được chứng minh là có hiệu quả đối với MT có nguồn lực thấp.

Bảng sau đây minh họa thêm các cách biểu diễn khác nhau cho các dấu hiệu biệt lập hơn. Trong ví dụ này, sử dụng SignWriting làm hệ thống ký hiệu. Lưu ý rằng cùng một dấu hiệu có thể có hai cách chú giải không liên quan và cùng một cách chú giải có thể có nhiều bản dịch ngôn ngữ nói hợp lệ.

Video	Pose Estimation	Notation	Gloss	English Translation
			HOUSE	House
			WRONG-WHAT	What's the matter? What's wrong?
			DIFFERENT BUT	Different But

Hình 2 Hình minh họa các cách biểu diễn khác nhau cho các dấu hiệu

3.5 Các bài toán của ngôn ngữ ký hiệu

Cho đến nay, cộng đồng thị giác máy tính chủ yếu dẫn dắt nghiên cứu SLP tập trung vào việc xử lý các đặc điểm trực quan trong video ngôn ngữ ký hiệu. Do đó, các phương pháp SLP hiện tại không giải quyết đầy đủ sự phức tạp về mặt ngôn ngữ của ngôn ngữ ký hiệu. Báo cáo khảo sát các tác vụ SLP phổ biến và những hạn chế của các phương pháp hiện tại, dựa trên các lý thuyết ngôn ngữ của ngôn ngữ ký hiệu.

3.5.1 Phát hiện ngôn ngữ ký hiệu

Phát hiện ngôn ngữ ký hiệu (Borg và Camilleri 2019 ; Moryossef và cộng sự 2020 ; Pal và cộng sự 2023) là nhiệm vụ phân loại nhị phân để xác định xem hoạt động ký hiệu có xuất hiện trong một khung video nhất định hay không. Nhiệm vụ tương tự trong ngôn ngữ nói là phát hiện hoạt động giọng nói (VAD) (Sohn, Kim và Sung 1999 ; Ramirez và cộng sự 2004) , phát hiện khi nào giọng nói của con người được sử dụng trong tín hiệu âm thanh. Vì các phương

pháp VAD thường dựa vào các biểu diễn dành riêng cho giọng nói như biểu đồ phổ nên chúng không nhất thiết phải áp dụng cho video.

Borg và Camilleri (2019) đã giới thiệu cách phân loại khung hình lấy từ video YouTube là ký hoặc không ký. Họ đã áp dụng cách tiếp cận không gian và thời gian dựa trên VGG-16 (Simonyan và Zisserman 2015) CNN để mã hóa từng khung hình và sử dụng Đơn vị tái phát có cổng (GRU) (Cho và cộng sự 2014) để mã hóa chuỗi khung hình trong cửa sổ 20 khung hình ở tốc độ 5 khung hình / giây. Ngoài khung thô, chúng còn mã hóa lịch sử luồng quang, lịch sử chuyển động tổng hợp hoặc chênh lệch khung.

Moryossef và cộng sự. (2020) đã cải tiến phương pháp của họ bằng cách thực hiện phát hiện ngôn ngữ ký hiệu trong thời gian thực. Họ xác định rằng việc sử dụng ngôn ngữ ký hiệu liên quan đến chuyển động của cơ thể và do đó, họ đã thiết kế một mô hình hoạt động dựa trên các tư thế ước tính của con người thay vì trực tiếp trên tín hiệu video. Họ đã tính toán định mức dòng quang của mọi khớp được phát hiện trên cơ thể và áp dụng mô hình bối cảnh hóa nông nhưng hiệu quả để dự đoán cho mọi khung hình xem người đó có đang ký hay không.

Mặc dù các mô hình phát hiện gần đây này đạt được hiệu suất cao nhưng chúng tôi cần dữ liệu được chú thích rõ ràng, bao gồm cả sự can thiệp và gây phiền nhiễu với các trường hợp không ký để đánh giá chính xác trong thế giới thực. Pal và cộng sự. (2023) đã tiến hành phân tích chi tiết về tác động của sự chồng chéo người ký giữa tập huấn luyện và tập kiểm tra trên hai bộ dữ liệu chuẩn phát hiện dấu hiệu (Signing in the Wild (Borg và Camilleri 2019) và DGS Corpus (Hanke et al. 2020)) được sử dụng bởi Borg và Camilleri (2019) và Moryossef và cộng sự. (2020) . Bằng cách so sánh độ chính xác có và không có sự trùng lặp, họ nhận thấy hiệu suất giảm tương đối đối với những người ký không có mặt trong quá trình đào tạo. Do đó, họ đã đề xuất các phân

vùng tập dữ liệu mới giúp loại bỏ sự chòng chéo giữa tập huấn luyện và tập kiểm tra, đồng thời tạo điều kiện đánh giá hiệu suất chính xác hơn.

3.5.2 Nhận dạng ngôn ngữ ký hiệu

Nhận dạng ngôn ngữ ký hiệu (Gebre, Wittenburg và Heskes 2013 ; Monteiro và cộng sự 2016) phân loại ngôn ngữ ký hiệu nào được sử dụng trong một video nhất định.

Gebre, Wittenburg và Heskes (2013) đã phát hiện ra rằng một công cụ phân loại rừng ngẫu nhiên đơn giản sử dụng phân bố âm vị có thể phân biệt giữa Ngôn ngữ ký hiệu Anh (BSL) và Ngôn ngữ ký hiệu Hy Lạp (ENN) với điểm F1 95%. Phát hiện này được hỗ trợ thêm bởi Monteiro et al. (2016) , dựa trên bản đồ hoạt động trong không gian ký hiệu, quản lý để phân biệt giữa Ngôn ngữ ký hiệu của Anh và Ngôn ngữ ký hiệu của Pháp (Langue des Signes Française, LSF) với điểm F1 98% trong các video có nền tĩnh và giữa Ngôn ngữ ký hiệu của Mỹ và Ngôn ngữ ký hiệu của Anh, với 70% điểm F1 cho các video được khai thác từ các trang chia sẻ video phổ biến. Các tác giả cho rằng thành công của họ chủ yếu là nhờ các hệ thống đánh vần bằng ngón tay khác nhau, đó là sử dụng hai tay trong trường hợp BSL và một tay trong trường hợp ASL và LSF.

Mặc dù những kết quả phân loại theo cặp này có vẻ hứa hẹn, nhưng cần có những mô hình tốt hơn để phân loại từ một tập hợp lớn các ngôn ngữ ký hiệu. Các phương pháp này chỉ dựa vào các đặc điểm trực quan cấp thấp, trong khi ngôn ngữ ký hiệu có một số đặc điểm riêng biệt ở cấp độ ngôn ngữ, chẳng hạn như sự khác biệt về từ vựng hoặc cấu trúc (McKee và Kennedy 2000 ; Kimmelman 2014 ; Ferreira-Brito 1984 ; Shroyer và Shroyer 1984) chưa được khám phá cho nhiệm vụ này.

3.5.3 Phân đoạn ngôn ngữ ký hiệu

Phân đoạn bao gồm việc phát hiện ranh giới khung cho các ký hiệu hoặc cụm từ trong video để chia chúng thành các đơn vị có ý nghĩa. Trong khi cách chuẩn

mục nhất để chia văn bản ngôn ngữ nói thành một chuỗi từ tuyến tính, do tính đồng thời của ngôn ngữ ký hiệu, khái niệm "từ" ngôn ngữ ký hiệu không được định nghĩa rõ ràng và ngôn ngữ ký hiệu không thể được mô hình hóa tuyến tính hoàn toàn.

Các phương pháp hiện tại sử dụng các đơn vị phân đoạn được ánh xạ lỏng lẻo tới các đơn vị ngôn ngữ ký hiệu (Santemiz và cộng sự 2009 ; Farag và Brock 2019 ; Bull, Gouiffès và Braffort 2020 ; Renz, Stache và cộng sự 2021 a , 2021 b ; Bull và cộng sự 2021) và không tận dụng một cách rõ ràng các yếu tố dự đoán ngôn ngữ đáng tin cậy về ranh giới câu như ngữ điệu trong ngôn ngữ ký hiệu (tức là tạm dừng, thời lượng ký hiệu kéo dài, nét mặt) (Sandler 2010 ; Ormel và Crasborn 2012) . De Sisto và cộng sự. (2021) kêu gọi hiểu rõ hơn về cấu trúc ngôn ngữ ký hiệu mà họ tin rằng đó là nền tảng cần thiết cho việc thiết kế và phát triển các phương pháp phân đoạn và nhận dạng ngôn ngữ ký hiệu.

Santemiz và cộng sự. (2009) đã tự động trích xuất các dấu hiệu riêng biệt từ việc ký liên tục bằng cách căn chỉnh các chuỗi thu được thông qua nhận dạng giọng nói, được mô hình hóa bằng phương pháp Dynamic Time Warping (DTW) và Hidden Markov Models (HMM).

Farag và Brock (2019) đã sử dụng bộ phân loại rừng ngẫu nhiên để phân biệt các khung chứa ký hiệu trong Ngôn ngữ ký hiệu Nhật Bản dựa trên thành phần của các đặc điểm góc không gian-thời gian và khoảng cách giữa các cặp phân đoạn khớp cụ thể theo miền.

Bull, Gouiffès và Braffort (2020) đã phân đoạn Ngôn ngữ ký hiệu tiếng Pháp thành các phân đoạn tương ứng với các đơn vị phụ đề bằng cách dựa vào sự liên kết giữa phụ đề và video ngôn ngữ ký hiệu, tận dụng mạng tích chập đồ

thị không gian-thời gian (ST-GCN; Yu, Yin và Zhu (2018)) với BiLSTM trên dữ liệu khung 2D.

Renz, Stache, et al. (2021 a) đã xác định ranh giới thời gian giữa các dấu hiệu trong video ngôn ngữ ký hiệu liên tục bằng cách sử dụng biểu diễn mạng nơ-ron tích chập 3D với tinh chỉnh phân đoạn thời gian lặp lại để giải quyết sự mơ hồ giữa các tín hiệu ranh giới dấu hiệu. Renz, Stache, et al. (2021 b) đã đề xuất thêm thuật toán Nhãn giả điều chế điểm thay đổi (CMPL) để giải quyết vấn đề thích ứng miền không có nguồn.

Bull và cộng sự. (2021) đã trình bày cách tiếp cận dựa trên Transformer để phân đoạn video ngôn ngữ ký hiệu và căn chỉnh chúng đồng thời với phụ đề, mã hóa phụ đề bằng BERT (Devlin et al. 2019) và video bằng cách trình bày video CNN.

Moryossef, Jiang, et al. (2023) đã trình bày một phương pháp được thúc đẩy bởi các tín hiệu ngôn ngữ được quan sát thấy trong các tập hợp ngôn ngữ ký hiệu, chẳng hạn như ngữ điệu (tạm dừng, nhịp độ, v.v.) và thay đổi hình dạng bàn tay. Họ cũng thấy rằng việc sử dụng BIO, một lược đồ chú thích ghi chú phân đầu, bên trong và bên ngoài, tạo ra sự khác biệt đáng kể so với các lược đồ trước đây chỉ ghi chú IO (bên trong hoặc bên ngoài). Họ thấy rằng việc bao gồm luồng quang học và chuẩn hóa bàn tay 3D cũng giúp khái quát hóa ngoài miền và các ngôn ngữ ký hiệu khác.

3.5.4 Nhận dạng, dịch, và tạo ngôn ngữ ký hiệu

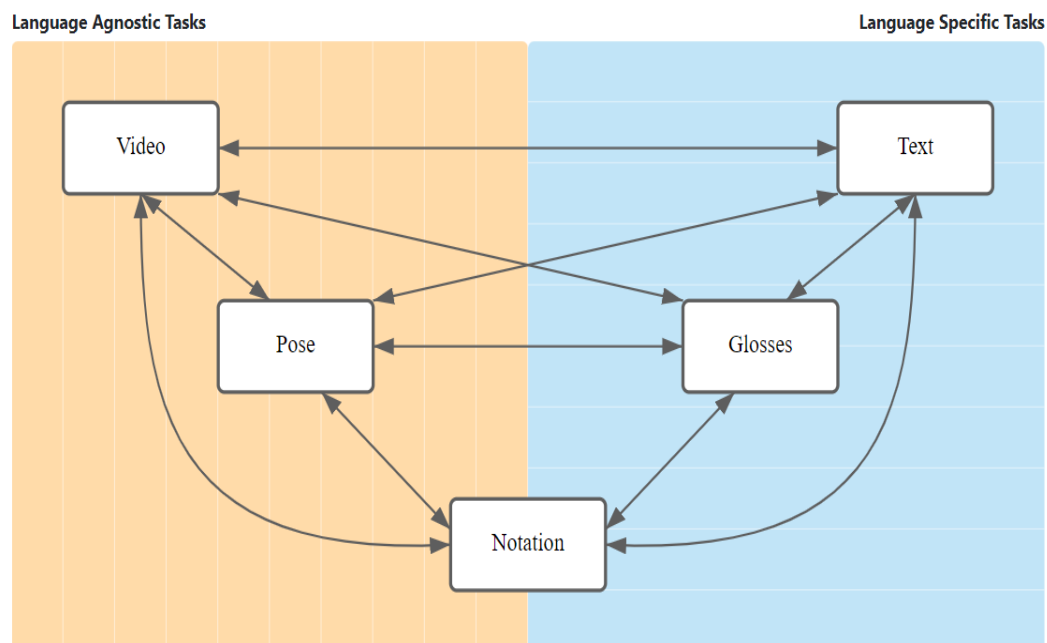
Dịch ngôn ngữ ký hiệu (SLT) thường đề cập đến việc dịch ngôn ngữ ký hiệu sang ngôn ngữ nói (De Coster và cộng sự, 2022 ; Müller và cộng sự, 2022) . Tạo (sản xuất) ngôn ngữ ký hiệu là quá trình ngược lại của việc sản xuất video ngôn ngữ ký hiệu từ văn bản ngôn ngữ nói. Nhận dạng ngôn ngữ ký hiệu (SLR) (Adaloglou và cộng sự, 2020) phát hiện và dán nhãn các ký hiệu từ video, trên các ký hiệu riêng lẻ (Imashev và cộng sự, 2020 ; Sincan và Keles, 2020) hoặc

liên tục (Cui, Liu và Zhang, 2017 ; Camgöz và cộng sự, 2018 ; NC Camgöz và cộng sự, 2020 b) .

Trong biểu đồ sau, chúng ta có thể thấy một hình ngũ giác được kết nối đầy đủ trong đó mỗi nút là một biểu diễn dữ liệu duy nhất và mỗi cạnh có hướng biểu thị nhiệm vụ chuyển đổi một biểu diễn dữ liệu này sang một biểu diễn dữ liệu khác.

Chúng tôi chia biểu đồ thành hai phần:

- Mỗi cạnh bên trái, trên nền màu cam, biểu thị một tác vụ trong tầm nhìn máy tính. Các tác vụ này vốn không phụ thuộc vào ngôn ngữ; do đó, chúng khái quát giữa các ngôn ngữ ký hiệu.
- Mỗi cạnh bên phải, trên nền xanh, thể hiện một nhiệm vụ trong xử lý ngôn ngữ tự nhiên. Các nhiệm vụ này dành riêng cho ngôn ngữ ký hiệu, yêu cầu từ vựng ngôn ngữ ký hiệu cụ thể hoặc mã thông báo ngôn ngữ nói.
- Mỗi cạnh trên cả hai nền đều biểu thị một nhiệm vụ đòi hỏi sự kết hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên.



Hình 3 Tổng quát các nhiệm vụ liên quan tới ngôn ngữ ký hiệu

Có 20 nhiệm vụ được định nghĩa theo khái niệm trong biểu đồ này, với số lượng nghiên cứu trước đó khác nhau. Mỗi đường dẫn giữa hai nút có thể hoặc không thể hợp lệ, tùy thuộc vào mức độ mất mát của các nhiệm vụ trong quy trình.

3.5.5 Video-To-Pose

Video-to-Pose—thường được gọi là ước tính tư thế—là nhiệm vụ phát hiện hình người trong hình ảnh và video, để người ta có thể xác định, ví dụ, khuỷu tay của ai đó xuất hiện ở đâu trong hình ảnh. Người ta đã chỉ ra rằng tư thế khuôn mặt tương quan với các đặc điểm không phải bằng tay trên khuôn mặt như hướng đầu (Vogler và Goldenstein 2005) .

Lĩnh vực này đã được nghiên cứu kỹ lưỡng (Pishchulin và cộng sự, 2012 ; Chen và cộng sự, 2017 ; Cao và cộng sự, 2019 ; Güler, Neverova và Kokkinos , 2018) với các mục tiêu khác nhau, từ dự đoán tư thế 2D/3D đến lựa chọn một tập hợp nhỏ các điểm mốc cụ thể hoặc một mạng lưới dày đặc của một người.

OpenPose (Cao et al. 2019 ; Simon et al. 2017 ; Cao et al. 2017 ; Wei et al. 2016) là hệ thống nhiều người đầu tiên cùng phát hiện các điểm chính trên cơ thể con người, bàn tay, khuôn mặt và bàn chân (tổng cộng 135 điểm chính điểm chính) ở dạng 2D trên các hình ảnh đơn lẻ. Mặc dù mô hình của họ có thể ước tính tư thế đầy đủ trực tiếp từ một hình ảnh trong một suy luận duy nhất, nhưng họ cũng đề xuất một phương pháp tiếp cận đường ống trong đó trước tiên họ ước tính tư thế cơ thể và sau đó ước tính độc lập tư thế bàn tay và khuôn mặt bằng cách thu được các loại cắt có độ phân giải cao hơn xung quanh các khu vực đó. Dựa trên cách tiếp cận đường dẫn chậm, một mô hình OpenPose toàn bộ mạng duy nhất đã được đề xuất (Martinez và cộng sự 2019) , mô hình này nhanh hơn và chính xác hơn trong trường hợp lấy được tất cả các điểm chính. Với nhiều góc ghi, OpenPose cũng cung cấp tính năng tam giác điểm chính để tái tạo lại tư thế ở dạng 3D.

DensePose (Güler, Neverova và Kokkinos 2018) có cách tiếp cận khác. Thay vì phân loại cho mọi điểm chính mà pixel có nhiều khả năng xảy ra nhất, họ đề xuất một phương pháp tương tự như phân đoạn ngữ nghĩa, để mỗi pixel phân loại nó thuộc về phần cơ thể nào. Sau đó, đối với mỗi pixel, khi biết bộ phận cơ thể, họ dự đoán vị trí của pixel đó trên bộ phận cơ thể so với hình chiếu 2D của mô hình cơ thể đại diện. Cách tiếp cận này dẫn đến việc tái cấu trúc lưới toàn thân và cho phép lấy mẫu để tìm các điểm chính cụ thể tương tự như OpenPose.

Tuy nhiên, tư thế của con người 2D có thể không đủ để hiểu đầy đủ vị trí và hướng của các điểm mốc trong không gian và việc áp dụng ước tính tư thế trên mỗi khung hình sẽ bỏ qua thông tin chuyển động theo thời gian của video, đặc biệt là trong các trường hợp chuyển động nhanh, có chuyển động mờ.

Pavlo và cộng sự (2019) đã phát triển hai phương pháp để chuyển đổi giữa các tư thế 2D sang tư thế 3D. Phương pháp đầu tiên, một phương pháp có giám sát, được đào tạo để sử dụng thông tin thời gian giữa các khung hình để dự

đoán trục Z bị thiếu. Phương pháp thứ hai là một phương pháp không giám sát, tận dụng thực tế là các tư thế 2D chỉ đơn thuần là phép chiếu của một tư thế 3D chưa biết và đào tạo một mô hình để ước tính tư thế 3D và chiếu ngược lại các tư thế 2D đầu vào. Phép chiếu ngược này là một quá trình xác định, áp dụng các ràng buộc vào bộ mã hóa tư thế 3D. Zelinka và Kanis (2020) đã làm theo một quy trình tương tự và thêm một ràng buộc để xương duy trì độ dài cố định giữa các khung hình.

Panteleris, Oikonomidis và Argyros (2018) đề xuất chuyển đổi các tư thế 2D thành 3D bằng cách sử dụng động học nghịch đảo (IK), một quy trình lấy từ hoạt hình máy tính và rô bốt để tính toán các tham số khớp biến đổi cần thiết để đặt phần cuối của chuỗi động học, chẳng hạn như bộ xương của người máy hoặc nhân vật hoạt hình, ở một vị trí và hướng nhất định so với phần đầu của chuỗi. Để chứng minh cách tiếp cận của mình đối với ước tính tư thế bàn tay, họ mã hóa thủ công các ràng buộc và giới hạn của từng khớp, tạo ra 26 bậc tự do. Sau đó, quá trình giảm thiểu bình phương nhỏ phi tuyến tính phù hợp với mô hình 3D của bàn tay với các vị trí khớp 2D ước tính, khôi phục lại tư thế bàn tay 3D. Quy trình này tương tự như phép chiếu ngược được Pavllo và cộng sự (2019) sử dụng , ngoại trừ ở đây, không có thông tin thời gian nào được sử dụng.

MediaPipe Holistic (Grishchenko và Bazarevsky 2020) cố gắng giải quyết ước tính tư thế 3D bằng cách áp dụng cách tiếp cận tương tự như OpenPose, có hệ thống đường ống để ước tính cơ thể, sau đó là khuôn mặt và bàn tay. Không giống như OpenPose, các tư thế ước tính ở dạng 3D và trình ước tính tư thế chạy theo thời gian thực trên CPU, cho phép tạo mô hình ngôn ngữ ký hiệu dựa trên tư thế trên các thiết bị di động có công suất thấp. Công cụ ước tính tư thế này có sẵn rộng rãi và được xây dựng cho Android, iOS, C++, Python và Web bằng JavaScript.

3.5.6 Pose-To-Video

Pose-to-Video, còn được gọi là chuyển động hoặc hoạt hình xương trong lĩnh vực robot và hoạt hình, là việc chuyển đổi một chuỗi các tư thế thành video. Nhiệm vụ này là "kết xuất" cuối cùng của ngôn ngữ ký hiệu theo phương thức trực quan.

Chan và cộng sự. (2019) đã trình diễn cách tiếp cận bán giám sát trong đó họ lấy một tập hợp video, chạy ước tính tư thế bằng OpenPose (Cao và cộng sự 2019) và học cách dịch từ hình ảnh sang hình ảnh (Isola và cộng sự 2017) giữa bộ xương được kết xuất và video gốc. Họ đã thể hiện cách tiếp cận của mình đối với điệu nhảy của con người, trích xuất các tư thế từ vũ đạo và khiến bất kỳ người nào cũng như thể họ đang khiêu vũ. Họ dự đoán hai khung hình liên tiếp sẽ cho kết quả video nhất quán theo thời gian và giới thiệu một quy trình riêng để tổng hợp khuôn mặt chân thực hơn, mặc dù vẫn còn thiếu sót.

Vương và cộng sự. (2018) đã đề xuất một phương pháp tương tự bằng cách sử dụng các biểu diễn DensePose (Güler, Neverova và Kokkinos 2018) bên cạnh các biểu diễn OpenPose (Cao et al. 2019) . Họ đã chính thức hóa một mô hình khác, với nhiều mục tiêu khác nhau để tối ưu hóa, chẳng hạn như tách nền-tiền cảnh và sự gắn kết thời gian bằng cách sử dụng hai đầu thời gian trước đó trong dữ liệu đầu vào.

Sử dụng phương pháp của Chan et al. (2019) trên “Mọi người nhảy ngay bây giờ”, Giró-i-Nieto (2020) hỏi, “Mọi người có thể ký ngay bây giờ không?” và điều tra xem liệu mọi người có thể hiểu ngôn ngữ ký hiệu từ các video được tạo tự động hay không. Họ đã tiến hành một nghiên cứu trong đó những người tham gia xem ba loại video: video ký hiệu ban đầu, video chỉ hiển thị các tư thế (bộ xương) và video được tái tạo lại với ký hiệu thực tế. Các nhà nghiên cứu đánh giá mức độ hiểu biết của người tham gia sau khi xem từng loại video. Kết quả cho thấy họ ưa thích video được xây dựng lại hơn là video về bộ xương. Tuy nhiên, các phương pháp tổng hợp video tiêu chuẩn được sử dụng

trong nghiên cứu này không đủ hiệu quả để dịch ngôn ngữ ký hiệu rõ ràng. Những người tham gia gặp khó khăn trong việc hiểu các video được xây dựng lại, cho thấy rằng cần phải cải thiện để dịch ngôn ngữ ký hiệu tốt hơn trong tương lai.

Như một câu trả lời trực tiếp, Saunders, Camgöz và Bowden (2020 a) đã chỉ ra điều đó giống như trong Chan et al. (2019) , trong đó một tổn thất đối nghịch được thêm vào để tạo khuôn mặt cụ thể, việc thêm một tổn thất tương tự vào quá trình tạo bàn tay sẽ mang lại các video ngôn ngữ ký hiệu liên tục có độ phân giải cao, chân thực hơn. Để cải thiện hơn nữa chất lượng tổng hợp hình ảnh bàn tay, họ đã giới thiệu chức năng mất dựa trên điểm chính để tránh các vấn đề do nhòe chuyển động gây ra.

Trong một bài báo tiếp theo, Saunders, Camgöz và Bowden (2021) đã giới thiệu nhiệm vụ Ẩn danh video ngôn ngữ ký hiệu (SLVA) như một phương pháp tự động để ẩn danh hình thức trực quan của video ngôn ngữ ký hiệu trong khi vẫn giữ lại nội dung ngôn ngữ ký hiệu gốc. Bằng cách sử dụng khung mã hóa tự động biến đổi có điều kiện, trước tiên, họ trích xuất thông tin tư thế từ video nguồn để loại bỏ hình dáng người ký ban đầu, sau đó tạo ra một video ngôn ngữ ký hiệu chân thực bằng hình ảnh về hình dáng mới lạ từ trình tự tư thế. Các tác giả đã đề xuất một phương pháp mất đi phong cách mới để đảm bảo tính nhất quán về phong cách trong các video ngôn ngữ ký hiệu ẩn danh.

3.6 Sign Language Avatars

JASigning

Là một hệ thống ký hiệu ảo tạo ra các màn trình diễn ngôn ngữ ký hiệu bằng cách sử dụng các ký tự con người ảo. Hệ thống này phát triển từ hệ thống SiGMLSigning trước đó, được phát triển trong các dự án ViSiCAST (Bangham và cộng sự, 2000 ; Elliott và cộng sự, 2000) và eSIGN (Zwitserslood

và cộng sự, 2004) , và sau đó được phát triển thêm như một phần của dự án Dicta-Sign (Matthes và cộng sự, 2012 ; Efthimiou và cộng sự, 2012) .

Ban đầu, JASigning dựa vào các ứng dụng Java JNLP để sử dụng độc lập và tích hợp vào các trang web. Tuy nhiên, cách tiếp cận này đã trở nên lỗi thời do thiếu sự hỗ trợ cho Java trong các trình duyệt hiện đại. Do đó, hệ thống Avatar ký CWA (CWASA) gần đây hơn đã được phát triển, dựa trên HTML5, sử dụng công nghệ JavaScript và WebGL.

SiGML (Ngôn ngữ đánh dấu cử chỉ ký hiệu) (Elliott và cộng sự, 2004) là một ứng dụng XML cho phép phiên âm các cử chỉ ngôn ngữ ký hiệu. SiGML được xây dựng trên HamNoSys và thực tế, một biến thể của SiGML về cơ bản là mã hóa các tính năng thủ công của HamNoSys, kèm theo biểu diễn các khía cạnh không thủ công. SiGML là ký hiệu đầu vào được sử dụng bởi các ứng dụng JASigning và các ứng dụng web. Một số công cụ chỉnh sửa cho SiGML hiện có, chủ yếu do Đại học Hamburg sản xuất.

Hệ thống phân tích văn bản tiếng Anh thành SiGML trước khi ánh xạ nó vào hình đại diện ký tên 3D có thể tạo ra ký hiệu. Sau đó, CWASA sử dụng một cơ sở dữ liệu lớn về hình ảnh động avatar ký tên 3D được xác định trước, có thể được kết hợp để tạo thành các câu mới. Hệ thống bao gồm trình chỉnh sửa 3D, cho phép người dùng tạo hình đại diện và hoạt ảnh ký tùy chỉnh.

PAULA (Davidson 2006)

Là hình đại diện ngôn ngữ ký hiệu dựa trên máy tính, ban đầu được phát triển để dạy ngôn ngữ ký hiệu cho người lớn khiếm thính. Hình đại diện là mô hình 3D của một người với từ vựng về ký hiệu được tạo hình động theo cách thủ công. Nó lấy cách phát âm ASL dưới dạng một luồng chú giải, thực hiện các sửa đổi cú pháp và hình thái, quyết định âm vị và thời gian thích hợp, đồng thời kết hợp các kết quả thành hoạt ảnh 3D của hình đại diện. Qua nhiều năm, một số kỹ thuật đã được sử dụng để làm cho hình đại diện trông chân thực hơn.

Trong những năm qua, một số tiến bộ đã được thực hiện để tăng cường tính chân thực và khả năng biểu cảm của hình đại diện PAULA, chẳng hạn như tinh chỉnh chuyển động của lông mày để trông tự nhiên hơn (Wolfe và cộng sự, 2011), kết hợp cảm xúc và các tín hiệu không bằng tay trên khuôn mặt đồng thời xảy ra (Schnepf và cộng sự, 2012, 2013), cải thiện độ mượt mà đồng thời tránh các chuyển động giống như rô bốt (McDonald và cộng sự, 2016) và tạo điều kiện cho tính đồng thời (McDonald và cộng sự, 2017). Những phát triển khác bao gồm giao diện với các hệ thống ký hiệu ngôn ngữ ký hiệu như AZee (Filhol, McDonald và Wolfe 2017), tăng cường hoạt ảnh nói bằng miệng (Johnson, Brumm và Wolfe 2018; Wolfe và cộng sự 2022), nhiều lớp kết cấu khuôn mặt và trang điểm (Wolfe và cộng sự 2019) và áp dụng các từ bỏ nghĩa trạng từ (Moncrief 2020, 2021).

Những cải tiến bổ sung cho PAULA tập trung vào việc làm cho hình đại diện giống thật hơn bằng cách nói lỏng hướng cổ tay và các góc "toán học" cực đoan khác (Filhol và McDonald 2020), tinh chỉnh quá trình chuyển đổi hình dạng bàn tay, thư giãn và va chạm (Baowidan 2021), triển khai các quá trình chuyển đổi cụm từ theo thứ bậc (McDonald và Filhol 2021), tạo khả năng kiểm soát cơ mặt thực tế hơn (McDonald, Johnson và Wolfe 2022) và hỗ trợ các chuyển vị hình học (Filhol và McDonald 2022).

SiMAX (“SiMAX - the Sign Language Avatar SiMAX Project Fact Sheet H2020”)

Là một ứng dụng phần mềm được phát triển để chuyển đổi văn bản đầu vào thành các biểu diễn ngôn ngữ ký hiệu hoạt hình 3D. Bằng cách sử dụng cơ sở dữ liệu toàn diện và kiến thức chuyên môn của các chuyên gia ngôn ngữ ký hiệu dành cho người khiếm thính, SiMAX đảm bảo bản dịch chính xác cả nội dung viết và nói. Quá trình này bắt đầu bằng việc tạo ra một đề xuất dịch thuật, sau đó sẽ được người dịch khiếm thính xem xét và sửa đổi nếu cần thiết để đảm bảo tính chính xác và phù hợp về mặt văn hóa. Những bản dịch này được

thực hiện bởi một hình đại diện kỹ thuật số có thể tùy chỉnh có thể được điều chỉnh để phản ánh danh tính công ty hoặc đối tượng mục tiêu của người dùng. Cách tiếp cận này mang lại giải pháp thay thế hiệu quả về mặt chi phí cho việc sản xuất video bằng ngôn ngữ ký hiệu truyền thống vì nó loại bỏ nhu cầu về xưởng phim đắt tiền và công nghệ video phức tạp thường gắn liền với những sản phẩm như vậy.

3.7 Mô hình tạo ảnh và video

Gần đây nhất trong lĩnh vực tạo ảnh và video, đã có những tiến bộ đáng chú ý trong các phương pháp như Kiến trúc tạo dựa trên phong cách cho mạng đối nghịch tạo sinh (Karras, Laine và Aila 2019 ; Karras và cộng sự 2020 , 2021) , Mô hình khuếch tán biến thiên (Kingma và cộng sự 2021) , Tổng hợp ảnh độ phân giải cao với mô hình khuếch tán tiềm ẩn (Rombach và cộng sự 2021) , Tạo video độ nét cao với mô hình khuếch tán (Ho và cộng sự 2022) và Tổng hợp video độ phân giải cao với mô hình khuếch tán tiềm ẩn (Blattmann và cộng sự 2023) . Các phương pháp này đã cải thiện đáng kể chất lượng tổng hợp ảnh và video, mang lại kết quả cực kỳ chân thực và hấp dẫn về mặt thị giác.

Tuy nhiên, bất chấp sự tiến bộ đáng kể trong việc tạo ra hình ảnh và video chất lượng cao, các mô hình này lại đánh đổi hiệu quả tính toán. Độ phức tạp của các thuật toán này thường dẫn đến thời gian suy luận chậm hơn, khiến các ứng dụng thời gian thực trở nên khó khăn. Việc triển khai các mô hình này trên thiết bị mang lại những lợi ích như chi phí máy chủ thấp hơn, chức năng ngoại tuyến và quyền riêng tư của người dùng được cải thiện. Trong khi các tối ưu hóa nhận thức về tính toán, đặc biệt nhắm mục tiêu vào khả năng phần cứng của các thiết bị khác nhau, có thể cải thiện độ trễ suy luận của các mô hình

này, Chen và cộng sự (2023) nhận thấy rằng việc tối ưu hóa các mô hình như vậy trên các thiết bị di động hàng đầu như Samsung S23 Ultra hoặc iPhone 14 Pro Max có thể giảm độ trễ suy luận trên mỗi khung hình từ khoảng 23 giây xuống còn khoảng 12.

ControlNet (L. Zhang và Agrawala 2023) gần đây đã trình bày một cấu trúc mạng nơ-ron để kiểm soát các mô hình khuếch tán lớn được đào tạo trước với các điều kiện đầu vào bổ sung. Phương pháp này cho phép học từ đầu đến cuối các điều kiện cụ thể của tác vụ, ngay cả với một tập dữ liệu đào tạo nhỏ. Đào tạo ControlNet nhanh như tinh chỉnh mô hình khuếch tán và có thể được thực hiện trên các thiết bị cá nhân hoặc mở rộng thành lượng dữ liệu lớn bằng cách sử dụng các cụm tính toán mạnh mẽ. ControlNet đã được chứng minh là tăng cường các mô hình khuếch tán lớn như Stable Diffusion với các đầu vào có điều kiện như bản đồ cạnh, bản đồ phân đoạn và điểm chính. Một trong những ứng dụng của ControlNet là điều khiển dịch chuyển từ thế sang hình ảnh, cho phép tạo hình ảnh dựa trên thông tin tư thế. Mặc dù phương pháp này đã cho thấy kết quả khả quan, nhưng nó vẫn yêu cầu đào tạo lại mô hình và không hỗ trợ tính nhất quán về mặt thời gian, điều này rất quan trọng đối với các tác vụ như dịch ngôn ngữ ký hiệu.

Trong tương lai gần, chúng ta có thể mong đợi nhiều công trình điều khiển mô hình khuếch tán video trực tiếp từ văn bản để dịch ngôn ngữ ký hiệu. Những mô hình này có thể sẽ tạo ra các video chân thực và hấp dẫn về mặt hình ảnh. Tuy nhiên, họ vẫn có thể mắc lỗi và bị giới hạn trong các tình huống có sẵn nhiều dữ liệu đào tạo hơn. Việc phát triển các mô hình có thể tạo chính xác video ngôn ngữ ký hiệu từ văn bản hoặc thông tin tư thế trong khi vẫn duy trì chất lượng hình ảnh và tính mạch lạc về thời gian sẽ là điều cần thiết để thúc đẩy lĩnh vực sản xuất ngôn ngữ ký hiệu.

Gloss-To-Pose

Gloss-to-Pose, được xếp vào nhiệm vụ sản xuất ngôn ngữ ký hiệu, là nhiệm vụ tạo ra một chuỗi các tư thế thể hiện đầy đủ một chuỗi các ký hiệu được viết dưới dạng bóng.

Để sản xuất video ngôn ngữ ký hiệu, Stoll và cộng sự (2018) đã xây dựng một bảng tra cứu giữa các chú thích và chuỗi tư thế 2D. Họ căn chỉnh tất cả các chuỗi tư thế tại khớp cổ của một bộ xương tham chiếu và nhóm tất cả các chuỗi thuộc cùng một chú thích. Sau đó, đối với mỗi nhóm, họ áp dụng phương pháp bề cong thời gian động và tính trung bình tất cả các chuỗi trong nhóm để xây dựng chuỗi tư thế trung bình. Cách tiếp cận này gặp phải vấn đề là không có một tập hợp các tư thế chính xác được căn chỉnh theo chú thích và có các chuyển động chuyển tiếp không tự nhiên giữa các chú thích.

Để giảm bớt những nhược điểm của công việc trước đó, Stoll et al. (2020) đã xây dựng một bảng tra cứu độ bóng cho một nhóm các chuỗi tư thế thay vì tạo ra một chuỗi tư thế trung bình. Họ đã xây dựng Đồ thị chuyển động (Min và Chai 2012) , đây là một quy trình Markov được sử dụng để tạo ra các chuỗi chuyển động mới đại diện cho chuyển động tự nhiên và chọn các nguyên mẫu chuyển động (chuỗi các tư thế) trên mỗi bóng có xác suất chuyển tiếp cao nhất. Để làm trơn chuỗi đó và giảm chuyển động không tự nhiên, họ đã sử dụng bộ lọc làm mịn chuyển động chuyển động Savitzky–Golay (Savitzky và Golay 1964) . Moryossef, Müller và cộng sự. (2023) đã triển khai lại cách tiếp cận của họ và biến nó thành nguồn mở.

Hoàng và cộng sự. (2021) đã sử dụng một mô hình không tự hồi quy mới để tạo ra một chuỗi các tư thế cho một chuỗi các độ bóng. Họ lập luận rằng các mô hình hiện có như Saunders, Bowden và Camgöz (2020) dễ bị tích lũy lỗi và độ trễ suy luận cao do tính chất tự hồi quy của chúng. Mô hình của họ thực hiện việc lấy mẫu dần dần các tư thế, bằng cách bắt đầu với một tư thế chỉ bao gồm hai khớp ở lớp đầu tiên và dần dần giới thiệu nhiều điểm chính hơn. Họ

đã đánh giá mô hình của mình trên tập dữ liệu RWTH-PHOENIX-WEATHER 2014T (Camgöz và cộng sự 2018) bằng cách sử dụng Dynamic Time Warping (DTW) (Berndt và Clifford 1994) để căn chỉnh các tư thế trước khi tính toán Lỗi khớp trung bình (DTW-MJE). Họ đã chứng minh rằng mô hình của họ vượt trội hơn các phương pháp hiện có về độ chính xác và tốc độ, khiến nó trở thành một phương pháp đầy hứa hẹn để sản xuất ngôn ngữ ký hiệu nhanh và chất lượng cao.

Video-To-Gloss

Video-to-Gloss, còn được gọi là nhận dạng ngôn ngữ ký hiệu, là nhiệm vụ nhận dạng một chuỗi ký hiệu từ một video.

Để nhận ra điều này, Cui, Liu và Zhang (2017) đã xây dựng mô hình tối ưu hóa ba bước. Đầu tiên, họ huấn luyện một mô hình từ đầu đến cuối video-to-gloss, trong đó họ mã hóa video bằng bộ mã hóa CNN không gian-thời gian và dự đoán độ bóng bằng cách sử dụng Phân loại thời gian kết nối (CTC) (Graves et al. 2006) . Sau đó, từ đề xuất danh mục và căn chỉnh CTC, họ mã hóa từng phân đoạn cấp độ bóng một cách độc lập, được đào tạo để dự đoán danh mục độ bóng và sử dụng mã hóa phân đoạn video độ bóng này để tối ưu hóa mô hình học theo trình tự.

Camgöz et al. (2018) về cơ bản khác với cách tiếp cận đó và xây dựng vấn đề này như thể nó là một vấn đề dịch ngôn ngữ tự nhiên. Họ mã hóa từng khung hình video bằng AlexNet (Krizhevsky, Sutskever và Hinton 2012) , được khởi tạo bằng trọng số được đào tạo trên ImageNet (Deng et al. 2009) . Sau đó, họ áp dụng kiến trúc bộ mã hóa-giải mã GRU với Luong Attention (Luong, Pham và Manning 2015) để tạo độ bóng. Trong công việc tiếp theo, NC Camgöz et al. (2020 b) sử dụng bộ mã hóa biến áp (Vaswani et al. 2017) để thay thế GRU và sử dụng CTC để giải mã độ bóng. Họ cho thấy một cải tiến nhỏ với cách tiếp cận này đối với tác vụ chuyển đổi video thành độ bóng.

Adaloglou và cộng sự (2020) thực hiện đánh giá thử nghiệm so sánh các phương pháp dựa trên thị giác máy tính cho tác vụ chuyển đổi video thành bóng. Họ triển khai nhiều phương pháp tiếp cận khác nhau từ nghiên cứu trước đây (Camgöz và cộng sự, 2017 ; Cui, Liu và Zhang , 2019 ; Joze và Koller, 2019) và thử nghiệm chúng trên nhiều tập dữ liệu (Huang và cộng sự, 2018 ; Camgöz và cộng sự, 2018 ; Von Agris và Kraiss , 2007 ; Joze và Koller, 2019) cho cả nhận dạng dấu hiệu riêng lẻ và nhận dạng dấu hiệu liên tục. Họ kết luận rằng các mô hình tích chập 3D hoạt động tốt hơn các mô hình chỉ sử dụng mạng hồi quy để nắm bắt thông tin thời gian và các mô hình này có khả năng mở rộng hơn do trường tiếp nhận bị hạn chế, xuất phát từ kỹ thuật "cửa sổ trượt" CNN.

Momeni, Bull, Prajwal, và những người khác. (2022 a) đã phát triển một quy trình toàn diện kết hợp nhiều mô hình khác nhau để chú thích dày đặc cho các video bằng ngôn ngữ ký hiệu. Bằng cách tận dụng việc sử dụng từ đồng nghĩa và căn chỉnh ký hiệu phụ đề, cách tiếp cận của họ thể hiện giá trị của việc gắn nhãn giả từ mô hình nhận dạng dấu hiệu để phát hiện dấu hiệu. Họ đề xuất một phương pháp mới để tăng chú thích cho cả lớp đã biết và lớp chưa biết, dựa vào các mẫu trong miền. Do đó, khuôn khổ của họ đã mở rộng đáng kể số lượng chú thích tự động đáng tin cậy trên kho ngữ liệu ngôn ngữ ký hiệu BOBSL BSL (Albanie et al. 2021) từ 670K lên 5 triệu và họ hào phóng công bố những chú thích này.

Gloss-To-Video

Gloss-to-Video, còn được gọi là sản xuất ngôn ngữ ký hiệu, là nhiệm vụ sản xuất video thể hiện đầy đủ một chuỗi các ký hiệu được viết dưới dạng bóng.

Tính đến năm 2020, không có nghiên cứu nào thảo luận về nhiệm vụ dịch trực tiếp giữa gloss và video. Việc thiếu thảo luận này là kết quả của tính không thực tế về mặt tính toán của mô hình mong muốn, khiến các nhà nghiên cứu

không thực hiện nhiệm vụ này trực tiếp mà thay vào đó dựa vào các phương pháp tiếp cận đường ống sử dụng các biểu diễn tư thế trung gian.

Gloss-To-Text

Gloss-to-Text hay còn gọi là dịch ngôn ngữ ký hiệu là nhiệm vụ xử lý ngôn ngữ tự nhiên dịch giữa văn bản bóng đại diện cho các ký hiệu ngôn ngữ ký hiệu và văn bản ngôn ngữ nói. Những văn bản này thường khác nhau về thuật ngữ, cách viết hoa và cấu trúc câu.

Camgoz và cộng sự. (2018) đã thử nghiệm nhiều kiến trúc dịch máy khác nhau và so sánh việc sử dụng LSTM (Hochreiter và Schmidhuber 1997) với GRU cho mô hình lặp lại, cũng như Luong, Pham, và Manning 2015) với Bahdanau chú ý (Bahdanau, Cho và Bengio 2015) và nhiều kích cỡ lô khác nhau. Họ kết luận rằng trên tập dữ liệu RWTH-PHOENIX-Weather-2014T, cũng được trình bày trong tác phẩm này, sử dụng GRU, Luong sự chú ý và kích thước lô 1 vượt trội hơn tất cả các cấu hình khác.

Song song với những tiến bộ trong dịch máy ngôn ngữ nói, Yin and Read (2020) đã đề xuất thay thế RNN bằng mô hình bộ mã hóa-giải mã Transformer (Vaswani et al. 2017) , cho thấy những cải tiến trên cả RWTH-PHOENIX-Weather-2014T (DGS) và bộ dữ liệu ASLG-PC12 (ASL) đều sử dụng một mô hình duy nhất và tập hợp các mô hình. Điều thú vị là, trong tính năng chuyển độ bóng sang văn bản, họ cho thấy rằng việc sử dụng đầu ra của hệ thống nhận dạng ngôn ngữ ký hiệu (video-to-gloss) sẽ hoạt động tốt hơn khi sử dụng các độ bóng có chú thích bằng vàng.

Dựa trên mã do Yin và Read (2020) công bố , Moryossef, Yin và cộng sự (2021) cho thấy việc đào tạo trước các mô hình dịch này bằng cách sử dụng các ngữ liệu ngôn ngữ nói đơn ngữ tăng cường là có lợi. Họ thử ba cách tiếp cận khác nhau để tăng cường dữ liệu: (1) Dịch ngược; (2) Các quy tắc chung

từ văn bản sang chú thích, bao gồm lemmatization, sắp xếp lại từ và loại bỏ từ; (3) Các quy tắc cụ thể cho từng cặp ngôn ngữ tăng cường cú pháp ngôn ngữ nói thành cú pháp ngôn ngữ ký hiệu tương ứng. Khi đào tạo trước, tất cả các phép tăng cường đều cho thấy sự cải thiện so với đường cơ sở đối với RWTH-PHOENIX-Weather-2014T (DGS) và NCSLGR (ASL).

Text-To-Gloss

Text-to-gloss, một ví dụ về bản dịch ngôn ngữ ký hiệu, là nhiệm vụ dịch giữa văn bản ngôn ngữ nói và chú thích ngôn ngữ ký hiệu. Đây là một lĩnh vực nghiên cứu hấp dẫn vì tính đơn giản của nó khi tích hợp vào các đường ống NMT hiện có, mặc dù các công trình gần đây như Yin và Read (2020) và Müller và cộng sự (2023) tuyên bố rằng chú thích là một cách thể hiện không hiệu quả của ngôn ngữ ký hiệu và chú thích không phải là cách thể hiện đầy đủ các ký hiệu (Pizzuto, Rossini và Russo 2006) .

Zhao và cộng sự (2000) đã sử dụng hệ thống dựa trên Ngữ pháp cây liên kết (TAG) để dịch các câu tiếng Anh sang các chuỗi chú thích Ngôn ngữ ký hiệu Hoa Kỳ (ASL). Họ phân tích cú pháp văn bản tiếng Anh và đồng thời lắp ráp một cây chú thích ASL, sử dụng TAG đồng bộ (Shieber và Schabes 1990 ; Shieber 1994) , bằng cách liên kết các cây cơ bản ASL với các cây cơ bản tiếng Anh và liên kết các nút mà tại đó có thể xảy ra các phép thay thế hoặc phép cộng tiếp theo. TAG đồng bộ đã được sử dụng để dịch máy giữa các ngôn ngữ nói (Abeille, Schabes và Joshi 1990) , nhưng đây là ứng dụng đầu tiên cho ngôn ngữ ký hiệu.

Đối với bản dịch tự động của gloss-to-text, Othman và Jemni (2012) đã xác định nhu cầu về một ngữ liệu gloss ngôn ngữ ký hiệu song song lớn và ngữ liệu văn bản ngôn ngữ nói. Họ đã phát triển một ngữ pháp dựa trên loại lời nói để chuyển đổi các câu tiếng Anh từ bộ sưu tập sách điện tử của Dự án

Gutenberg (Lebert 2008) thành gloss Ngôn ngữ ký hiệu Hoa Kỳ. Ngữ liệu cuối cùng của họ chứa hơn 100 triệu câu tổng hợp và 800 triệu từ và là ngữ liệu gloss Anh-ASL rộng lớn nhất mà chúng tôi biết. Thật không may, rất khó để chứng thực chất lượng của ngữ liệu, vì các tác giả đã không đánh giá phương pháp của họ trên các cặp gloss Anh-ASL thực tế.

Egea Gómez, McGill và Saggion (2021) đã trình bày một bộ chuyển đổi nhận biết cú pháp cho nhiệm vụ này, bằng cách chèn các thẻ phụ thuộc từ để tăng cường các nhúng được nhập vào bộ mã hóa. Điều này liên quan đến các sửa đổi nhỏ trong kiến trúc nơ-ron dẫn đến tác động không đáng kể đến độ phức tạp tính toán của mô hình. Khi thử nghiệm mô hình của họ trên RWTH-PHOENIX-Weather-2014T (Camgöz và cộng sự, 2018) , họ đã chứng minh rằng việc chèn thông tin bổ sung này dẫn đến chất lượng bản dịch tốt hơn.

Video-To-Text

Chuyển video thành văn bản, còn được gọi là dịch ngôn ngữ ký hiệu, là nhiệm vụ dịch video thô thành văn bản ngôn ngữ nói.

NC Camgöz và cộng sự (2020 b) đã đề xuất một kiến trúc duy nhất để thực hiện nhiệm vụ này có thể sử dụng cả chú thích ngôn ngữ ký hiệu và văn bản ngôn ngữ nói trong giám sát chung. Họ sử dụng nhúng không gian được đào tạo trước từ Koller và cộng sự (2019) để mã hóa từng khung độc lập và mã hóa các khung bằng bộ chuyển đổi. Trong mã hóa này, họ sử dụng Phân loại thời gian kết nối (CTC) (Graves và cộng sự 2006) để phân loại chú thích ngôn ngữ ký hiệu. Sử dụng cùng một mã hóa, họ sử dụng bộ giải mã bộ chuyển đổi để giải mã văn bản ngôn ngữ nói từng mã thông báo một. Họ chỉ ra rằng việc thêm giám sát chú thích cải thiện mô hình hơn là không sử dụng nó và nó vượt trội hơn các phương pháp tiếp cận đường ống video-sang-chú thích-sang-văn bản trước đây (Camgöz và cộng sự 2018) .

Tiếp theo, NC Camgöz và cộng sự (2020 a) đề xuất một kiến trúc mới không yêu cầu giám sát chú thích, có tên là "Bộ chuyển đổi đa kênh để dịch ngôn ngữ ký hiệu đa khớp". Trong phương pháp này, họ cắt tay ký hiệu và khuôn mặt và thực hiện ước tính tư thế 3D để có được ba kênh dữ liệu riêng biệt. Họ mã hóa từng kênh dữ liệu riêng biệt bằng bộ chuyển đổi, sau đó mã hóa tất cả các kênh lại với nhau và nối các kênh riêng biệt cho mỗi khung. Giống như công trình trước đây của họ, họ sử dụng bộ giải mã bộ chuyển đổi để giải mã văn bản ngôn ngữ nói, nhưng không giống như công trình trước đây của họ, họ không sử dụng chú thích làm giám sát bổ sung. Thay vào đó, họ thêm hai tổn thất "neo" để dự đoán hình dạng bàn tay và hình dạng miệng từ mỗi khung một cách độc lập, vì chú thích bạc có sẵn cho họ bằng cách sử dụng mô hình được đề xuất trong Koller và cộng sự (2019) . Họ kết luận rằng phương pháp này ngang bằng với các phương pháp trước đây yêu cầu chú thích, và do đó, họ đã phá vỡ sự phụ thuộc vào thông tin chú thích chú thích tồn kém trong tác vụ chuyển đổi video thành văn bản.

Shi et al. (2022) giới thiệu OpenASL, một tập dữ liệu Ngôn ngữ ký hiệu Hoa Kỳ (ASL) - tiếng Anh quy mô lớn được thu thập từ các trang web video trực tuyến (ví dụ: YouTube), sau đó đề xuất một bộ kỹ thuật bao gồm tìm kiếm ký hiệu như một nhiệm vụ tiền đề để đào tạo trước và hợp nhất các đặc điểm về miệng và hình dạng bàn tay để cải thiện chất lượng bản dịch khi không có chú thích và khi có dữ liệu khó hiểu về mặt hình ảnh.

Yutong Chen, Zuo và cộng sự. (2022) trình bày mạng hai luồng để nhận dạng ngôn ngữ ký hiệu (SLR) và dịch thuật (SLT), sử dụng kiến trúc bộ mã hóa hình ảnh kép để mã hóa các khung hình video RGB và đặt các điểm chính trong các luồng riêng biệt. Các luồng này tương tác thông qua các kết nối hai chiều. Đối với SLT, bộ mã hóa hình ảnh dựa trên xương sống S3D (Xie và cộng sự 2018

) xuất ra mạng dịch đa ngôn ngữ bằng mBART (Liu và cộng sự 2020) . Mô hình này đạt được hiệu suất tiên tiến trên RWTH-PHOENIX-Weather-2014 (Koller, Forster và Ney 2015) , RWTH-PHOENIX-Weather-2014T (Camgöz et al. 2018) và CSL-Daily (Zhou và cộng sự 2021) .

B. Zhang, Müller và Sennrich (2023) đề xuất một phương pháp học đa phương thức, đa tác vụ cho bản dịch ngôn ngữ ký hiệu đầu cuối. Mô hình có các biểu diễn được chia sẻ cho các phương thức khác nhau như văn bản và video và được đào tạo chung trên một số tác vụ như video-to-gloss, gloss-to-text và video-to-text. Phương pháp này cho phép tận dụng dữ liệu bên ngoài như dữ liệu song song cho bản dịch máy ngôn ngữ nói.

Zhao và cộng sự. (2024) giới thiệu CV-SLT, sử dụng bộ mã hóa tự động biến thiên có điều kiện để giải quyết khoảng cách về phương thức giữa video và văn bản. Cách tiếp cận của họ liên quan đến việc hướng dẫn mô hình mã hóa dữ liệu hình ảnh và văn bản tương tự thông qua hai đường dẫn: một đường chỉ có dữ liệu hình ảnh và đường kia có cả hai phương thức. Bằng cách sử dụng phân kỳ KL, họ điều khiển mô hình theo hướng tạo ra các phần nhúng nhất quán và đầu ra chính xác bất kể đường dẫn. Khi mô hình đạt được hiệu suất nhất quán trên các đường dẫn, nó có thể được sử dụng để dịch mà không cần giám sát độ bóng. Đánh giá trên bộ dữ liệu RWTH-PHOENIX-Weather-2014T (Camgöz và cộng sự 2018) và CSL-Daily (Zhou và cộng sự 2021) cho thấy tính hiệu quả của nó. Họ cung cấp cách triển khai mã chủ yếu dựa trên Yutong Chen, Wei và cộng sự. (2022) .

Gong và cộng sự (2024) giới thiệu SignLLM, một khuôn khổ cho bản dịch ngôn ngữ ký hiệu không bóng bẩy tận dụng thế mạnh của Mô hình ngôn ngữ lớn (LLM). SignLLM chuyển đổi video ký hiệu thành các biểu diễn rời rạc và phân cấp tương thích với LLM thông qua hai mô-đun: (1) Mô-đun Ký hiệu trực quan lượng tử hóa vectơ (VQ-Sign), dịch video ký hiệu thành các mã thông báo "cấp ký tự" rời rạc và (2) Mô-đun Tái cấu trúc và Căn chỉnh số mã

(CRA), tái cấu trúc các mã thông báo này thành các biểu diễn "cấp từ". Trong quá trình suy luận, các mã thông báo "cấp từ" được chiếu vào không gian nhúng của LLM, sau đó được nhắc dịch. Bản thân LLM có thể được "lấy ra khỏi kệ" và không cần phải đào tạo. Trong quá trình đào tạo, mô-đun "cấp ký tự" VQ-Sign được đào tạo bằng tác vụ dự đoán ngữ cảnh, mô-đun "cấp từ" CRA với kỹ thuật vận chuyển tối ưu và mất căn chỉnh ký hiệu-văn bản giúp tăng cường hơn nữa sự căn chỉnh ngữ nghĩa giữa các mã thông báo ký hiệu và văn bản. Khung này đạt được kết quả tiên tiến nhất trên các tập dữ liệu RWTH-PHOENIX-Weather-2014T (Camgöz và cộng sự, 2018) và CSL-Daily (Zhou và cộng sự, 2021) mà không cần dựa vào chú thích chú thích.

Rust và cộng sự. (2024) giới thiệu một phương pháp nhận thức về quyền riêng tư gồm hai giai đoạn để dịch ngôn ngữ ký hiệu (SLT) trên quy mô lớn, được gọi là Đào tạo trước video tự giám sát để dịch ngôn ngữ ký hiệu (SSVP-SLT). Giai đoạn đầu tiên liên quan đến quá trình đào tạo trước tự giám sát của máy biến áp thị giác Hiera (Ryali và cộng sự 2023) trên các tập dữ liệu video lớn không được quản lý (Duarte và cộng sự 2021 ; Uthus, Tanzer và Georg 2023) . Trong giai đoạn thứ hai, đầu ra của mô hình tầm nhìn được đưa vào mô hình ngôn ngữ đa ngôn ngữ (Raffel và cộng sự 2020) để tinh chỉnh trên tập dữ liệu How2Sign (Duarte và cộng sự 2021) . Để giảm thiểu rủi ro về quyền riêng tư, hệ thống này sử dụng tính năng làm mờ khuôn mặt trong quá trình đào tạo trước. Họ nhận thấy rằng trong khi huấn luyện trước bằng cách làm mờ ảnh hưởng đến hiệu suất, một số có thể được phục hồi khi tinh chỉnh với dữ liệu không bị mờ. SSVP-SLT đạt được hiệu suất tiên tiến trên How2Sign (Duarte và cộng sự 2021) . Họ kết luận rằng các mô hình SLT có thể được đào tạo trước theo cách nhận thức được quyền riêng tư mà không phải hy sinh quá nhiều hiệu suất. Ngoài ra, các tác giả còn phát hành DailyMoth-70h, bộ dữ liệu ASL 70 giờ mới từ The Daily Moth .

Text-To-Video

Chuyển văn bản thành video, còn được gọi là sản xuất ngôn ngữ ký hiệu, là nhiệm vụ sản xuất video thể hiện đầy đủ văn bản ngôn ngữ nói bằng ngôn ngữ ký hiệu.

Tính đến năm 2020, không có nghiên cứu nào thảo luận về nhiệm vụ dịch trực tiếp giữa văn bản và video. Việc thiếu thảo luận này là kết quả của tính không thực tế về mặt tính toán của mô hình mong muốn, khiến các nhà nghiên cứu không thực hiện nhiệm vụ này trực tiếp mà thay vào đó dựa vào các phương pháp tiếp cận đường ống sử dụng các biểu diễn tư thế trung gian.

Pose-To-Text

Pose-To-Text, còn được gọi là dịch ngôn ngữ ký hiệu, là nhiệm vụ dịch chuỗi tư thế được chụp hoặc ước tính sang văn bản ngôn ngữ nói.'

Ko et al. (2019) đã chứng minh hiệu suất ấn tượng trong nhiệm vụ chuyển tư thế thành văn bản bằng cách nhập chuỗi tư thế vào mạng dịch mã hóa-giải mã chuẩn. Họ đã thử nghiệm với cả GRU và nhiều loại chú ý khác nhau (Luong, Pham và Manning 2015 ; Bahdanau, Cho và Bengio 2015) và với Transformer (Vaswani et al. 2017) và cho thấy hiệu suất tương tự, với transformer hoạt động kém hơn trên tập xác thực và hoạt động tốt hơn trên tập kiểm tra, bao gồm những người ký tên không nhìn thấy. Họ đã thử nghiệm với nhiều lược đồ chuẩn hóa khác nhau, chủ yếu là trừ đi giá trị trung bình và chia cho độ lệch chuẩn của từng điểm chính riêng lẻ liên quan đến toàn bộ khung hoặc "đối tượng" có liên quan (Cơ thể, Khuôn mặt và Bàn tay).

Text-To-Pose

Text-to-Pose, còn được gọi là sản xuất ngôn ngữ ký hiệu, là nhiệm vụ tạo ra một chuỗi các tư thế thể hiện đầy đủ một văn bản ngôn ngữ nói bằng ngôn ngữ ký hiệu, như một biểu diễn trung gian để vượt qua các thách thức trong hoạt hình. Hầu hết các nỗ lực đều sử dụng các tư thế như một biểu diễn trung gian để vượt qua các thách thức trong việc tạo video trực tiếp, với mục tiêu sử dụng

hoạt hình máy tính hoặc các mô hình pose-to-video để thực hiện sản xuất video.

Saunders, Camgöz và Bowden (2020 b) đã đề xuất Progressive Transformers, một mô hình để dịch từ các câu ngôn ngữ nói rời rạc sang các chuỗi tư thế ký hiệu 3D liên tục theo cách tự hồi quy. Không giống như các biến đổi tương trung (Vaswani và cộng sự, 2017) , sử dụng một vốn từ vựng rời rạc và do đó có thể dự đoán một EOS mã thông báo kết thúc chuỗi () ở mọi bước, biến đổi tiến bộ dự đoán một counter $\in [0 , 1]$ ngoài tư thế. Trong thời gian suy luận, counter = 1 được coi là kết thúc của chuỗi . Họ đã thử nghiệm cách tiếp cận của mình trên tập dữ liệu RWTH-PHOENIX-Weather-2014T (Camgöz et al. 2018) bằng cách sử dụng ước tính tư thế OpenPose 2D, nâng cấp lên 3D (Zelinka và Kanis 2020) và cho thấy kết quả thuận lợi khi đánh giá bằng cách sử dụng dịch ngược từ các tư thế được tạo ra sang ngôn ngữ nói. Họ tiếp tục chỉ ra (Saunders, Bowden và Camgöz 2020) rằng việc sử dụng bộ phân biệt đối nghịch giữa các tư thế thực tế và các tư thế được tạo ra, có điều kiện trên văn bản ngôn ngữ nói đầu vào, cải thiện chất lượng sản xuất khi đo bằng cách sử dụng dịch ngược.

Để khắc phục các vấn đề về phát âm không chuẩn xác được thấy trong các tác phẩm trên, Saunders, Camgöz và Bowden (2020 a) đã mở rộng mô hình máy biến áp tiến bộ bằng cách sử dụng Mạng mật độ hỗn hợp (MDN) (Bishop 1994) để mô hình hóa biến thể được tìm thấy trong ngôn ngữ ký hiệu. Mặc dù mô hình này hoạt động kém hiệu quả trên tập xác thực, so với các tác phẩm trước đó, nhưng nó hoạt động tốt hơn trên tập kiểm tra.

Zelinka và Kanis (2020) đã trình bày một phương pháp giải mã hồi quy tự động tương tự, với việc thêm vào tính năng làm cong thời gian động (DTW) và sự chú ý nhẹ nhàng. Họ đã thử nghiệm phương pháp của mình trên dữ liệu thời tiết Ngôn ngữ ký hiệu Séc được trích xuất từ tin tức, không được chú thích

thủ công hoặc căn chỉnh theo chú thích ngôn ngữ nói và cho thấy DTW của họ có lợi thế cho loại nhiệm vụ này.

Xiao, Qin và Yin (2020) đã khép lại vòng lặp bằng cách đề xuất mô hình chuyển văn bản thành văn bản cho trường hợp nhận dạng ngôn ngữ ký hiệu biệt lập. Đầu tiên, họ huấn luyện một bộ phân loại để lấy một chuỗi các từ thể được mã hóa bởi BiLSTM và phân loại dấu hiệu liên quan, sau đó đề xuất một hệ thống sản xuất lấy một dấu hiệu duy nhất và lấy mẫu chuỗi có độ dài không đổi gồm 50 từ thể từ Mô hình hỗn hợp Gaussian. Các thành phần này được kết hợp sao cho có một lớp dấu hiệu y , một chuỗi từ thể được tạo ra, sau đó được phân loại lại thành một lớp dấu hiệu \hat{y} , và sự mất mát được áp dụng giữa y và \hat{y} chứ không phải trực tiếp trên chuỗi từ thể được tạo ra. Họ đánh giá cách tiếp cận của họ trên tập dữ liệu CSL (Huang và cộng sự 2018) và chỉ ra rằng các chuỗi từ thể được tạo của họ gần như đạt được hiệu suất phân loại tương tự như các chuỗi tham chiếu.

Do nhu cầu về các phương pháp đánh giá tự động phù hợp hơn cho các ký hiệu được tạo ra, các tác phẩm hiện tại phải dùng đến phương pháp đo chất lượng dịch ngược, không thể nắm bắt chính xác chất lượng của các ký hiệu được tạo ra cũng như khả năng sử dụng của chúng trong bối cảnh thực tế. Hiểu được cách tạo ra sự khác biệt về ý nghĩa trong ngôn ngữ ký hiệu có thể giúp phát triển một phương pháp đánh giá tốt hơn.

Notation-To-Text

Giang và cộng sự. (2023) khám phá việc chuyển văn bản thành văn bản bằng ký hiệu sang ngôn ngữ nói, với SignWriting là hệ thống ký hiệu ngôn ngữ ký hiệu được chọn. Mặc dù SignWriting thường được biểu diễn dưới dạng 2D, nhưng họ sử dụng đặc tả SignWriting chính thức 1D và đề xuất phương pháp dịch máy dựa trên hệ số thần kinh để mã hóa các chuỗi biểu đồ SignWriting cũng như vị trí của chúng trong không gian 2D. Họ xác minh cách tiếp cận được đề xuất trên bộ dữ liệu SignBank bằng cả thiết lập song ngữ (Ngôn ngữ

ký hiệu Mỹ sang tiếng Anh) và hai thiết lập đa ngôn ngữ (lần lượt là 4 và 21 cặp ngôn ngữ). Họ áp dụng một số kỹ thuật dịch máy có nguồn tài nguyên thấp được sử dụng để cải thiện bản dịch ngôn ngữ nói nhằm cải thiện hiệu suất dịch ngôn ngữ ký hiệu một cách tương tự. Phát hiện của họ xác nhận việc sử dụng cách trình bày văn bản trung gian để dịch ngôn ngữ ký hiệu và mở đường cho việc đưa dịch ngôn ngữ ký hiệu vào nghiên cứu xử lý ngôn ngữ tự nhiên.

Text-To-Notation

Jiang và cộng sự (2023) cũng khám phá hướng dịch ngược, tức là dịch văn bản sang SignWriting. Họ tiến hành các thí nghiệm trong cùng điều kiện với thí nghiệm SignWriting đa ngôn ngữ sang văn bản (4 cặp ngôn ngữ) của họ và một lần nữa đề xuất phương pháp dịch máy có phân tích thần kinh để giải mã các ký tự và vị trí của chúng riêng biệt. Họ mượn BLEU từ bản dịch ngôn ngữ nói để đánh giá các ký tự được dự đoán và sai số tuyệt đối trung bình để đánh giá các số vị trí.

Walsh, Saunders và Bowden (2022) khám phá bản dịch Văn bản sang HamNoSys (T2H), với HamNoSys là hệ thống ký hiệu ngôn ngữ ký hiệu mục tiêu. Họ thử nghiệm với T2H trực tiếp và Văn bản sang chú thích sang HamNoSys (T2G2H) trên một tập hợp con dữ liệu từ tập dữ liệu MEINE DGS (Hanke và cộng sự, 2020) , trong đó tất cả các chú thích đều được ánh xạ sang HamNoSys bằng cách tra cứu từ điển. Họ nhận thấy rằng bản dịch T2H trực tiếp tạo ra BLEU cao hơn (mặc dù vẫn cần làm rõ mức độ BLEU thể hiện chất lượng bản dịch HamNoSys tốt như thế nào). Họ mã hóa HamNoSys bằng BPE (Sennrich, Haddow và Birch, 2016) , vượt trội hơn mã hóa cấp độ ký tự và cấp độ từ. Họ cũng tận dụng BERT để tạo những cấp độ câu tốt hơn và sử dụng HamNoSys để trích xuất hình dạng bàn tay của một ký hiệu như một sự giám sát bổ sung trong quá trình đào tạo.

Notation-To-Pose

Arkushin, Moryossef và Fried (2023) đã đề xuất Ham2Pose, một mô hình để hoạt hình hóa HamNoSys thành một chuỗi các tư thế. Đầu tiên, họ mã hóa HamNoSys thành một biểu diễn "bối cảnh" có ý nghĩa bằng cách sử dụng bộ mã hóa biến đổi và sử dụng nó để dự đoán độ dài của chuỗi tư thế sẽ được tạo ra. Sau đó, bắt đầu từ một khung hình tĩnh, họ đã sử dụng bộ giải mã lặp lại không tự hồi quy để dần dần tinh chỉnh dấu hiệu qua T bước. Trong mỗi bước thời gian t từ T đến 1 , mô hình dự đoán sự thay đổi cần thiết từ bước t đến bước $t - 1$. Sau T bước, trình tạo tư thế đưa ra chuỗi tư thế cuối cùng. Mô hình của họ vượt trội hơn các phương pháp trước đây như Saunders, Camgöz và Bowden (2020), hoạt hình hóa HamNoSys thành các chuỗi ngôn ngữ ký hiệu thực tế hơn.

3.8 Phương pháp đánh giá – Evaluation Metrics

Các phương pháp đánh giá tự động quá trình xử lý ngôn ngữ ký hiệu thường chỉ phụ thuộc vào đầu ra và không phụ thuộc vào đầu vào.

Đầu ra văn bản

Đối với các tác vụ xuất văn bản bằng ngôn ngữ nói, các phương pháp đánh giá dịch máy tiêu chuẩn như BLEU, chrF hoặc COMET thường được sử dụng.

Đầu ra Gloss

Đầu ra Gloss cũng có thể được tự động chấm điểm, mặc dù không phải không có vấn đề. Đặc biệt, Müller và cộng sự. (2023) đã phân tích điều này và đưa ra một loạt khuyến nghị.

Đầu ra Pose

Đối với việc dịch từ ngôn ngữ nói sang ngôn ngữ ký hiệu, các số liệu đánh giá tự động là một hướng nghiên cứu mở, mặc dù một số số liệu liên quan đến dịch ngược đã được phát triển

Ở mức đơn giản, các tác phẩm trong miền này đã sử dụng các số liệu như Lỗi bình phương trung bình (MSE) hoặc Lỗi vị trí trung bình (APE) cho kết quả đầu ra về tư thế (ahuja 2019 Language2Pose Natural Language ;ghosh 2021 Synthesis Compositional Animations; petrovich 2022 TEMOS Generated Diverse). Tuy nhiên, những thước đo này có những hạn chế đáng kể đối với việc Sản xuất Ngôn ngữ Ký hiệu.

Ví dụ: MSE và APE không tính đến sự thay đổi độ dài chuỗi. Trong thực tế, cùng một dấu hiệu không phải lúc nào cũng mất cùng một khoảng thời gian để tạo ra, thậm chí bởi cùng một người ký. Để giải quyết sự thay đổi theo thời gian, Huang et al. (2021) đã giới thiệu một thước đo cho đầu ra của chuỗi tư thế dựa trên việc đo khoảng cách giữa các chuỗi tư thế được tạo và tham chiếu ở cấp độ khớp bằng cách sử dụng độ cong thời gian động, được gọi là DTW-MJE (Độ cong thời gian động - Lỗi khớp trung bình). Tuy nhiên, số liệu này không đề cập rõ ràng cách xử lý các điểm chính bị thiếu. Arkushin, Moryossef và Fried (2023) đã thử nghiệm nhiều phương pháp đánh giá và đề xuất thêm hàm khoảng cách để giải quyết các điểm chính bị thiếu này. Họ đã áp dụng chức năng này bằng cách chuẩn hóa các điểm chính, đặt tên số liệu của họ là nDTW-MJE.

Multi-Channel Block Output

Để thay thế cho trình tự Gloss, Kim et al. (2024) đã đề xuất cách trình bày đầu ra đa kênh cho ngôn ngữ ký hiệu và giới thiệu SignBLEU, một phương pháp tính điểm giống BLEU cho các đầu ra này. Thay vì một chuỗi các chú giải tuyến tính duy nhất, các phân đoạn biểu diễn đầu ra ngôn ngữ ký hiệu thành nhiều kênh tuyến tính, mỗi kênh chứa các “khối” rời rạc. Các khối này đại diện cho cả tín hiệu thủ công và không thủ công, ví dụ: một khối cho mỗi bàn tay và các khối khác cho các tín hiệu không thủ công khác nhau như chuyển động của lông mày. Sau đó, các khối này được chuyển đổi thành n-gram: các chuỗi thời gian ghi lại các chuỗi trong một kênh và các gram kênh

ghi lại các lần xuất hiện đồng thời trên các kênh. Sau đó, điểm SignBLEU được tính cho n-gram của các đơn hàng khác nhau. Họ đã đánh giá SignBLEU trên các bộ dữ liệu DGS Corpus v3.0 (Konrad và cộng sự 2020 ; Prillwitz và cộng sự 2008) , NIASL2021 (Huerta-Enochian et al. 2022) và NCSLGR (Neidle và Sclaroff 2012 ; Vogler và Neidle 2012) , so sánh nó với các số liệu đơn kênh (độ bóng) như BLEU, TER, chrF và METEOR, cũng như đánh giá của con người bởi những người ký bản địa. Các tác giả nhận thấy rằng SignBLEU có mối tương quan nhất quán với đánh giá của con người tốt hơn so với các lựa chọn thay thế này. Tuy nhiên, một hạn chế của phương pháp này là thiếu bộ dữ liệu phù hợp. Các tác giả đã xem xét một số ngữ liệu ngôn ngữ ký hiệu, lưu ý rằng sự khan hiếm tương đối của các chú thích đa kênh. Mã nguồn của SignBLEU có sẵn. Như với SacreBLEU (Bài đăng năm 2018) , mã có thể tạo ra các chuỗi “chữ ký phiên bản” tóm tắt các tham số chính để nâng cao khả năng tái tạo.

3.9 Truy xuất ngôn ngữ ký hiệu

3.9 Fingerspelling - Đánh vần bằng ngón tay

Fingerspelling là đánh vần một từ theo từng chữ cái, mượn từ bảng chữ cái ngôn ngữ nói (Battison 1978 ; Wilcox 1992 ; Brentari và Padden 2001 ; Patrie và Johnson 2011) . Hiện tượng này, được tìm thấy trong hầu hết các ngôn ngữ ký hiệu, thường xảy ra khi không có ký hiệu nào được thống nhất trước đó cho một khái niệm, như trong ngôn ngữ kỹ thuật, các cuộc trò chuyện thông tục liên quan đến tên, các cuộc trò chuyện liên quan đến các sự kiện hiện tại, các hình thức nhấn mạnh và bối cảnh chuyển đổi mã giữa ngôn ngữ ký hiệu và ngôn ngữ nói tương ứng (Padden 1998 ; Montemurro và Brentari 2018) . Lượng tương đối của fingerspelling khác nhau giữa các ngôn ngữ ký hiệu và đối với Ngôn ngữ ký hiệu Hoa Kỳ (ASL), chiếm 12-35% nội dung được ký hiệu (Padden và Gunsauls 2003).

Patrie và Johnson (2011) đã mô tả thuật ngữ sau đây để mô tả ba dạng đánh vần bằng ngón tay khác nhau:

- Cẩn thận - đánh vần chậm hơn để mỗi chữ cái được hình thành rõ ràng.
- Đánh vần nhanh - đánh vần nhanh trong đó các chữ cái thường không được hoàn thành và chứa phần còn lại của các chữ cái khác trong từ.
- Từ vựng hóa - một dấu hiệu được tạo ra bằng cách thường sử dụng không quá hai hình dạng chữ viết tay (Battison 1978). Ví dụ, ALL sử dụng từ vựng hóa A và L, BUZZ sử dụng từ vựng hóa B và Z, v.v.

Nhận dạng

Nhận dạng chữ viết bằng ngón tay, một nhiệm vụ phụ của nhận dạng ngôn ngữ ký hiệu, là nhiệm vụ nhận dạng các từ được đánh vần bằng ngón tay trong video ngôn ngữ ký hiệu.

Shi và cộng sự. (2018) đã giới thiệu một tập dữ liệu lớn có sẵn để nhận dạng chính tả bằng ngôn ngữ ký hiệu của Mỹ. Tập dữ liệu này bao gồm cả hình thức đánh vần ngón tay “cẩn thận” và “nhanh chóng” được thu thập từ các video diễn ra tự nhiên “trong tự nhiên”, vốn khó khăn hơn so với điều kiện studio. Họ đã đào tạo một mô hình cơ sở để lấy một chuỗi hình ảnh được cắt xung quanh bàn tay ký và sử dụng bộ giải mã tự hồi quy hoặc CTC. Họ phát hiện ra rằng CTC hoạt động tốt hơn mô hình bộ giải mã tự hồi quy, nhưng cả hai đều đạt tỷ lệ nhận dạng kém (độ chính xác ở mức ký tự 35-41%) so với hiệu suất của con người (khoảng 82%).

Trong công việc tiếp theo, Shi et al. (2019) đã thu thập gần như một tập dữ liệu lớn hơn có quy mô lớn hơn và thiết kế một mô hình nhận dạng mới. Thay vì phát hiện bàn tay đang ký, họ phát hiện khuôn mặt và cắt một vùng rộng lớn xung quanh nó. Sau đó, họ thực hiện một quá trình lặp đi lặp lại là phóng to bàn tay bằng cách sử dụng sự chú ý trực quan để giữ lại đủ thông tin ở độ

phân giải cao của bàn tay. Cuối cùng, giống như công việc trước đây, họ đã mã hóa chuỗi hình ảnh được cắt thủ công và sử dụng CTC để lấy nhãn khung. Họ đã chỉ ra rằng phương pháp này vượt trội hơn 4% so với phương pháp “cắt thủ công” ban đầu của họ và họ có thể đạt được độ chính xác ở cấp độ ký tự lên tới 62,3% bằng cách sử dụng dữ liệu bổ sung được thu thập. Xem qua tập dữ liệu này, chúng tôi nhận thấy rằng các video trong tập dữ liệu được lấy từ các video dài hơn và khi bị cắt, chúng không giữ lại chữ ký trước khi gõ ngón tay. Bối cảnh này liên quan đến mô hình ngôn ngữ, trong đó lúc đầu, một ngón tay đánh vần một từ một cách cẩn thận và khi lặp lại nó, có thể đánh vần nó nhanh chóng, nhưng người đối thoại có thể suy ra rằng họ đang đánh vần cùng một từ.

Sản xuất

Sản xuất chính tả bằng ngón tay, một nhiệm vụ phụ của sản xuất ngôn ngữ ký hiệu, là nhiệm vụ sản xuất video đánh vần bằng ngón tay cho các từ.

Ở dạng cơ bản, việc sản xuất đánh vần bằng ngón tay “cẩn thận” có thể được giải quyết một cách dễ dàng bằng cách sử dụng nội suy hình dạng bàn tay chữ cái được xác định trước. Adeline (2013) đã chứng minh cách tiếp cận này cho Ngôn ngữ ký hiệu Hoa Kỳ và đánh vần bằng ngón tay tiếng Anh. Họ đã lắp một khung tay cho mỗi chữ cái trong bảng chữ cái tiếng Anh ($N = 26$) và tạo ra tất cả ($N^2 = 676$) chuyển đổi giữa mỗi hai chữ cái bằng cách sử dụng nội suy hoặc hoạt ảnh thủ công. Sau đó, để đánh vần bằng ngón tay toàn bộ các từ, họ nói các cặp chuyển đổi chữ cái. Ví dụ, đối với từ “CHLOE”, họ sẽ nói các chuyển đổi sau theo trình tự: #C CH HL LO OE E#.

Tuy nhiên, để tạo ra các hình ảnh động giống như thật, người ta cũng phải xem xét nhịp điệu và tốc độ giữ các chữ cái, và chuyển đổi giữa các chữ cái, vì những yếu tố này có thể ảnh hưởng đến mức độ dễ hiểu của các chuyển động đánh vần bằng ngón tay đối với người đối thoại (Wilcox (1992)). Wheatland và cộng sự (2016) đã phân tích cả video đánh vần bằng ngón tay "cẩn thận"

và "nhanh" để tìm các tính năng này. Họ phát hiện ra rằng đối với cả hai hình thức đánh vần bằng ngón tay, trung bình, từ càng dài thì thời gian chuyển tiếp và giữ càng ngắn. Hơn nữa, họ phát hiện ra rằng trung bình thời gian dành cho các chữ cái ở giữa ít hơn và chữ cái cuối cùng được giữ trung bình lâu hơn các chữ cái

3.11 Pretraining and Representation - Learning

Trong mô hình này, thay vì nhắm vào một nhiệm vụ cụ thể (ví dụ như chuyển từ thể thành văn bản), mục tiêu là tìm hiểu một mô hình hoặc cách biểu diễn Hiểu ngôn ngữ ký hiệu hữu ích nói chung, có thể được áp dụng hoặc tinh chỉnh cho các nhiệm vụ cụ thể tiếp theo.

Hu et al. (2023) giới thiệu SignBERT+, một phương pháp tiền đào tạo tự giám sát để hiểu ngôn ngữ ký hiệu (SLU) dựa trên mô hình hóa mặt nạ của chuỗi từ thể. Đây là phần mở rộng của SignBERT trước đó của họ (H. Hu, Zhao, et al. 2021), với một số cải tiến. Để tiền đào tạo, họ trích xuất chuỗi từ thể từ hơn 230 nghìn video bằng MMPose (Những người đóng góp 2020) . Sau đó, họ thực hiện mô hình hóa mặt nạ đa cấp (khớp, khung, clip) trên các chuỗi này, tích hợp mô hình bàn tay thống kê (Romero, Tzionas và Black 2017) để hạn chế các dự đoán của bộ giải mã về tính hiện thực giải phẫu và độ chính xác được nâng cao. Xác thực trên SLR bị cô lập (MS-ASL (Joze và Koller 2019) , WLASL (Li và cộng sự 2020) , SLR500 (Huang và cộng sự 2019)) , SLR liên tục (RWTH-PHOENIX-Weather 2014 (Koller, Forster và Ney 2015)) và SLT (RWTH-PHOENIX-Weather 2014T (Camgöz và cộng sự 2018)) chứng minh hiệu suất tiên tiến.

Zhao và cộng sự (2023) giới thiệu BEST (BERT Pre-training for Sign Language Recognition with Coupling Tokenization), một phương pháp tiền đào tạo dựa trên mô hình hóa mặt nạ của các chuỗi từ thể sử dụng một lược đồ mã hóa được ghép nối. Phương pháp này lấy các đơn vị bộ ba từ thể (tay trái, tay phải và thân trên có cánh tay) làm đầu vào, mỗi đơn vị được mã hóa thành

các mã rời rạc (Oord, Vinyals và Kavukcuoglu 2017) sau đó được ghép nối với nhau. Sau đó, mô hình hóa mặt nạ được áp dụng, trong đó bất kỳ hoặc tất cả các thành phần của bộ ba (tay trái, tay phải hoặc thân trên) có thể được che giấu, để tìm hiểu các mối tương quan phân cấp giữa chúng. Không giống như Hu và cộng sự (2023) , BEST không che giấu các chuỗi tư thế nhiều khung hoặc các khớp riêng lẻ. Các tác giả xác thực phương pháp đào tạo trước của họ về các tác vụ nhận dạng dấu hiệu bị cô lập (ISR) bằng cách sử dụng MS-ASL (Joze và Koller 2019) , WLASL (Li và cộng sự 2020) , SLR500 (Huang và cộng sự 2019) và NMFs-CSL (H. Hu, Zhou và cộng sự 2021) . Bên cạnh tác vụ pose-to-gloss, họ cũng thử nghiệm các tác vụ video-to-gloss thông qua hợp nhất với I3D (Carreira và Zisserman 2017) . Kết quả trên các tập dữ liệu này chứng minh hiệu suất tiên tiến so với các phương pháp trước đây và tương đương với SignBERT+ (Hu và cộng sự 2023) .

CHƯƠNG 4. THỰC NGHIỆM PHƯƠNG PHÁP CHUYỂN ĐỔI NGÔN NGỮ KÝ HIỆU

Công nghệ nhận diện ngôn ngữ ký hiệu và hệ thống chuyên đổi.

4.1 Giới thiệu về Công nghệ Nhận diện Ngôn ngữ Ký hiệu

4.1.1 Định nghĩa

Công nghệ Nhận diện Ngôn ngữ Ký hiệu là một lĩnh vực nghiên cứu thuộc ngành trí tuệ nhân tạo và thị giác máy tính, tập trung vào việc phát triển các hệ thống và ứng dụng có khả năng nhận diện, hiểu và dịch các cử chỉ tay và biểu cảm khuôn mặt của người sử dụng ngôn ngữ ký hiệu thành văn bản hoặc lời nói. Mục tiêu chính của công nghệ này là tạo ra các công cụ hỗ trợ giao tiếp hiệu quả giữa người khiếm thính và người không biết ngôn ngữ ký hiệu, từ đó giúp họ hòa nhập tốt hơn vào cộng đồng.



Hình 4 Thành công mới của AI: Chuyển lời nói sang ngôn ngữ ký hiệu

Các hệ thống nhận diện ngôn ngữ ký hiệu thường sử dụng các cảm biến như camera, găng tay có cảm biến, hoặc các thiết bị theo dõi chuyển động để thu thập dữ liệu về các cử chỉ và biểu cảm của người dùng. Sau đó, các thuật toán xử lý hình ảnh và học sâu (deep learning) sẽ phân tích và nhận diện các mẫu hình học, hình thái và động học của các cử chỉ này để chuyển chúng thành thông tin có thể hiểu được.

Công nghệ nhận diện ngôn ngữ ký hiệu hiện đang được ứng dụng trong nhiều lĩnh vực khác nhau, bao gồm giáo dục, y tế, và các dịch vụ công cộng. Những tiến bộ trong lĩnh vực này không chỉ giúp cải thiện chất lượng cuộc sống của người khiếm thính mà còn mở ra nhiều cơ hội mới cho việc nghiên cứu và phát triển các ứng dụng thông minh hơn trong tương lai.

4.1.2 Ứng dụng

Công nghệ nhận diện ngôn ngữ ký hiệu đã và đang tạo ra những thay đổi đáng kể trong nhiều lĩnh vực của đời sống xã hội. Trong giáo dục, công nghệ này hỗ trợ người khiếm thính bằng cách chuyển đổi các bài giảng thành ngôn ngữ ký hiệu, giúp họ tiếp cận kiến thức một cách dễ dàng hơn. Ngoài ra, trong giao tiếp hàng ngày, các ứng dụng di động và thiết bị thông minh sử dụng công nghệ này giúp người khiếm thính giao tiếp hiệu quả với cộng đồng.

Không chỉ dừng lại ở đó, công nghệ này còn được áp dụng rộng rãi trong các dịch vụ công cộng. Tại các bệnh viện, trung tâm dịch vụ khách hàng, và các cơ sở công cộng khác, các hệ thống hỗ trợ nhận diện ngôn ngữ ký hiệu giúp người khiếm thính có thể sử dụng dịch vụ một cách bình đẳng. Trong lĩnh vực giải trí, việc tạo ra phụ đề ngôn ngữ ký hiệu cho các chương trình truyền hình và phim ảnh giúp người khiếm thính có thể thưởng thức nội dung giải trí một cách trọn vẹn.

Các ứng dụng này không chỉ cải thiện chất lượng cuộc sống của người khiếm thính mà còn mở ra nhiều cơ hội nghiên cứu và phát triển cho các nhà khoa học và kỹ sư trong lĩnh vực trí tuệ nhân tạo và học máy.

4.2 Các thành phần của hệ thống chuyển đổi

4.2.1 Hệ thống Chuyển đổi Ngôn ngữ Ký hiệu

Hệ thống chuyển đổi ngôn ngữ ký hiệu là một giải pháp công nghệ tích hợp, được thiết kế để chuyển đổi các động tác ký hiệu thành văn bản hoặc âm thanh

ngôn ngữ nói. Hệ thống này bao gồm nhiều bước và thành phần, mỗi bước đóng vai trò quan trọng trong quá trình xử lý và chuyển đổi.

4.2.2 Kiến trúc hệ thống

Kiến trúc của hệ thống chuyển đổi ngôn ngữ ký hiệu được xây dựng dựa trên các công nghệ tiên tiến như nhận diện tư thế, phân đoạn video, nhận dạng ngôn ngữ ký hiệu, tokenization, và dịch ký hiệu sang ngôn ngữ nói.

Quá trình bắt đầu với việc sử dụng các công cụ như OpenPose hoặc MediaPipe để nhận diện các tư thế của người thực hiện ký hiệu trong video. Những công cụ này giúp xác định các điểm đặc trưng trên cơ thể người, từ đó phân tích các động tác ký hiệu. Sau khi nhận diện tư thế, hệ thống tiến hành phân đoạn video thành các phần nhỏ hơn để xác định và phân tách từng động tác ký hiệu.

Tiếp theo, hệ thống sử dụng các mô hình học máy để nhận dạng ngôn ngữ ký hiệu dựa trên các đặc điểm tư thế và động tác. Quá trình này giúp xác định ngôn ngữ ký hiệu được sử dụng trong video, từ đó chuyển đổi các động tác ký hiệu thành các token để dễ dàng xử lý và dịch thuật. Cuối cùng, các ký hiệu được dịch từ SignWriting thành văn bản ngôn ngữ nói và chuyển đổi văn bản này thành âm thanh ngôn ngữ nói thông qua các công cụ Text-to-Speech.

4.2.3 Công nghệ và công cụ sử dụng

Để xây dựng và triển khai hệ thống này, chúng tôi sử dụng nhiều công nghệ và công cụ tiên tiến. OpenPose và MediaPipe là hai công cụ quan trọng trong việc nhận diện tư thế và phân tích các động tác ký hiệu. TensorFlow và Keras được sử dụng để phát triển và huấn luyện các mô hình học máy, giúp cải thiện độ chính xác của quá trình nhận diện và dịch thuật.

Google Cloud Speech-to-Text và Text-to-Speech cung cấp các dịch vụ chuyển đổi giọng nói thành văn bản và ngược lại, đảm bảo rằng hệ thống có thể tạo ra các bản dịch chính xác và dễ hiểu. Node.js và Express.js được sử dụng để xây dựng backend của hệ thống, hỗ trợ việc xử lý dữ liệu và giao tiếp với các dịch

vụ khác, trong khi Firebase được sử dụng để lưu trữ và quản lý dữ liệu người dùng, cung cấp các tính năng bảo mật và xác thực người dùng.

Angular được sử dụng để phát triển frontend của hệ thống, cung cấp giao diện người dùng tương tác và dễ sử dụng. Các công cụ này kết hợp với nhau tạo thành một hệ thống chuyển đổi ngôn ngữ ký hiệu hoàn chỉnh, không chỉ mang lại lợi ích lớn cho người khiếm thính mà còn mở ra nhiều cơ hội phát triển và nghiên cứu trong tương lai.

- OpenPose
- MediaPipe
- TensorFlow
- Keras

4.2.4 Quy trình hoạt động của hệ thống

Pipeline 1. Tạo Ngôn Ngữ Ký Hiệu (Sign Language Production)

Input:

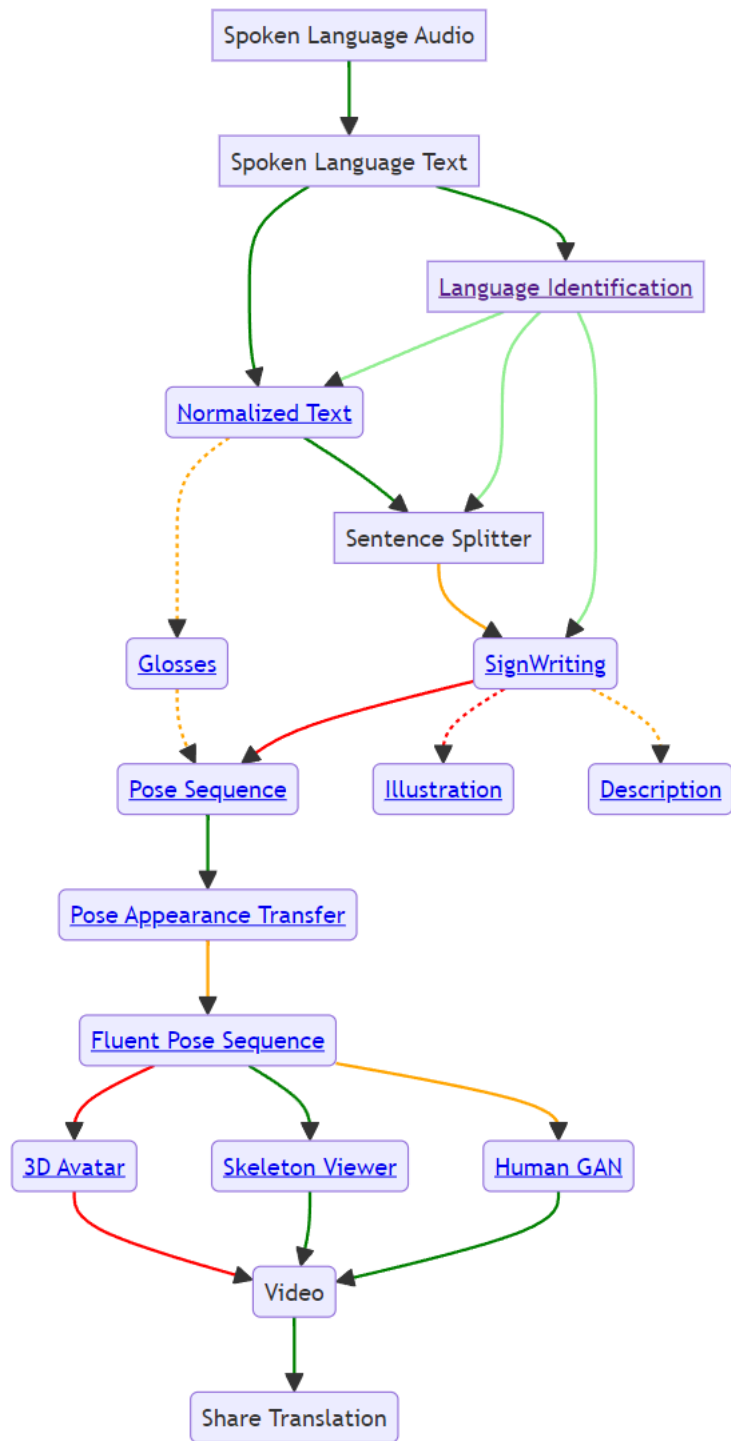
Âm thanh Ngôn Ngữ Nói (Spoken Language Audio).

Quy Trình:

Âm thanh được chuyển đổi thành văn bản (Spoken Language Text). Văn bản này sau đó được chuẩn hóa (Normalized Text) và chuyển thành SignWriting (một hệ thống ký hiệu ngôn ngữ ký). SignWriting được dùng để tạo ra chuỗi động tác (Pose Sequence).

Output:

Chuỗi động tác này được hiển thị qua một chương trình xem khung xương (Skeleton Viewer) hoặc một hình đại diện 3D (3D Avatar), được sinh ra thông qua một mô hình GAN (Generative Adversarial Network) cho hình ảnh con người.



Hình 5 Sơ đồ kiến trúc hệ thống pipeline 1 Spoken to Signed

Mô tả chi tiết pipeline 1 gồm các giai đoạn chính như sau:

Giai đoạn 1: Xây dựng giao diện thiết kế trang web để người dùng có thể sử dụng trên 2 nền tảng là ứng dụng website và ứng dụng di động được xây dựng bằng

Thiết kế và triển khai giao diện người dùng:

- Xây dựng giao diện ứng dụng trên nền tảng web và di động bằng framework Angular.
- Sử dụng TypeScript, HTML, và SCSS để phát triển các thành phần giao diện.
- Tạo Progressive Web App (PWA) để cung cấp trải nghiệm liền mạch trên cả thiết bị di động và máy tính.

Phát triển backend:

- Sử dụng Node.js để xây dựng server backend.
- Quản lý gói với npm và triển khai Firebase để lưu trữ và xử lý dữ liệu.

Giai đoạn 2: Chuyển đổi âm thanh ngôn ngữ nói thành văn bản

Tích hợp công cụ chuyển đổi giọng nói thành văn bản (Speech-to-Text):

- Sử dụng API của Google Cloud Speech-to-Text hoặc Microsoft Azure Cognitive Services để nhận dạng giọng nói và chuyển đổi thành văn bản.

Nhận dạng ngôn ngữ tự động:

- Hỗ trợ lựa chọn ngôn ngữ thủ công bởi người dùng (107 ngôn ngữ).
 - Nhận dạng ngôn ngữ tự động sử dụng Google's cld3 và MediaPipe Solutions.
- Sử dụng ngôn ngữ trình duyệt ban đầu và ghi nhớ cặp ngôn ngữ ưa thích của người dùng.

Chuẩn hóa văn bản:

- Sử dụng mô hình chuẩn hóa văn bản đa ngôn ngữ trên máy chủ để xử lý văn bản.

Giai đoạn 3: Dịch ngôn ngữ nói sang SignWriting

Dịch máy đa ngôn ngữ:

- Sử dụng mô hình dịch máy trên máy chủ để dịch văn bản ngôn ngữ nói sang SignWriting (chất lượng thấp~).

Triển khai dịch Client/Server:

- Sử dụng dịch vụ dịch Client/Server với Bergamot để cải thiện chất lượng dịch.

Phục vụ các mô hình dịch:

- Đảm bảo triển khai các mô hình dịch trên máy chủ để hỗ trợ việc dịch từ văn bản sang SignWriting.

Giai đoạn 4: Chuyển đổi SignWriting thành Pose Sequence

Chuyển đổi từ SignWriting sang Pose Sequence:

- Triển khai server-side sign-stitching, sử dụng OpenPose để chuyển đổi văn bản SignWriting thành các chuỗi động tác (chất lượng thấp~).
- Tạo hoạt hình trực tiếp từ các chuỗi SignWriting/HamNoSys.

Hỗ trợ suy luận ngoại tuyến:

- Triển khai mô hình hoạt hình trên client để hỗ trợ hoạt động ngoại tuyến.

Giai đoạn 5: Tạo hình ảnh từ Pose Sequence

Tạo hình ảnh từ Pose Sequence:

- Sử dụng Pose Viewer để xem khung xương (Skeleton Viewer), giúp gỡ lỗi nhanh chóng và tiết kiệm năng lượng.
- Sử dụng Human GAN để tạo hình ảnh giống con người từ các chuỗi động tác.
- Hoạt hình hóa avatar 3D sử dụng học máy và hỗ trợ AR.

Giai đoạn 6: Tạo video từ Pose Sequence

Chuyển đổi Pose Sequence thành video:

- Chuyển đổi các chuỗi động tác thành video để tiết kiệm năng lượng và bộ nhớ thiết bị.
- Hỗ trợ các thao tác Copy, Download, và Share sau khi video sẵn sàng.

Giai đoạn 7: Hỗ trợ quốc tế hóa

Quốc tế hóa:

- Hỗ trợ 104 ngôn ngữ và cả bố cục LTR và RTL.
- Sử dụng ngôn ngữ của trình duyệt/điện thoại người dùng và cho phép chọn ngôn ngữ khác thông qua tham số URL.

Các tính năng bổ sung:

Tiết kiệm năng lượng và bộ nhớ:

- Tạo video từ các chuỗi động tác để tiết kiệm năng lượng và bộ nhớ thiết bị.
- Hỗ trợ các thao tác Copy, Download, và Share sau khi video sẵn sàng.

Pipeline 2 Dịch Ngôn Ngữ Ký Hiệu (Sign Language Translation)

Input:

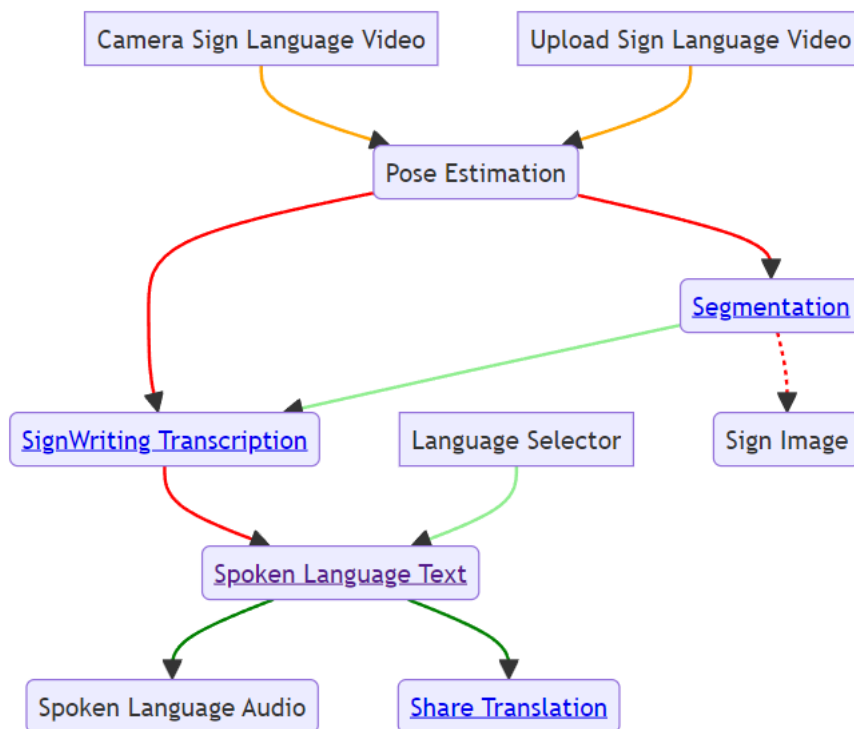
Video Ngôn Ngữ Ký Hiệu được tải lên (Upload Sign Language Video) hoặc được ghi hình trực tiếp (Camera Sign Language Video).

Quy Trình:

Video được phân đoạn (Segmentation) để nhận diện và phân tách từng động tác ký hiệu. Các động tác này được chuyển thành SignWriting.

Output:

SignWriting được chuyển đổi thành văn bản Ngôn Ngữ Nói (Spoken Language Text). Văn bản này cuối cùng được chuyển thành Âm thanh Ngôn Ngữ Nói (Spoken Language Audio).



Hình 6 Sơ đồ kiến trúc hệ thống pipeline 2 Signed to Spoken

Giai đoạn 1: Xây dựng giao diện người dùng

Thiết kế và triển khai giao diện người dùng:

- Xây dựng giao diện ứng dụng trên nền tảng web và di động bằng framework Angular.
- Sử dụng TypeScript, HTML, và SCSS để phát triển các thành phần giao diện.
- Tạo Progressive Web App (PWA) để cung cấp trải nghiệm liền mạch trên cả thiết bị di động và máy tính.

Phát triển backend:

- Sử dụng Node.js để xây dựng server backend.
- Quản lý gói với npm và triển khai Firebase để lưu trữ và xử lý dữ liệu.

Giai đoạn 2: Chuyển đổi video ngôn ngữ ký hiệu thành SignWriting

Dự đoán tư thế (Pose Estimation):

- Sử dụng các mô hình học máy như OpenPose hoặc MediaPipe để nhận diện tư thế của người thực hiện ký hiệu trong video.

Phân đoạn video (Segmentation):

- Sử dụng các thuật toán xử lý hình ảnh và học máy để phân đoạn video thành các phần nhỏ hơn.

Giai đoạn 3: Nhận dạng ngôn ngữ ký hiệu và chuyển đổi thành ngôn ngữ nói

Nhận dạng ngôn ngữ ký hiệu (Signed Language Identification):

- Áp dụng các mô hình nhận dạng ngôn ngữ để tự động phát hiện ngôn ngữ ký hiệu.

Tokenization:

- Phân tách các chuỗi động tác ký hiệu thành các đơn vị nhỏ hơn (token).

Chuyển đổi SignWriting thành ngôn ngữ nói (SignWriting to Spoken Language Translation):

- Sử dụng các mô hình dịch máy để chuyển đổi các ký hiệu từ SignWriting thành văn bản ngôn ngữ nói.

Chuyển đổi văn bản thành giọng nói (Text-to-Speech):

- Sử dụng các công cụ chuyển đổi văn bản thành giọng nói (Text-to-Speech) để tạo ra âm thanh từ văn bản ngôn ngữ nói.

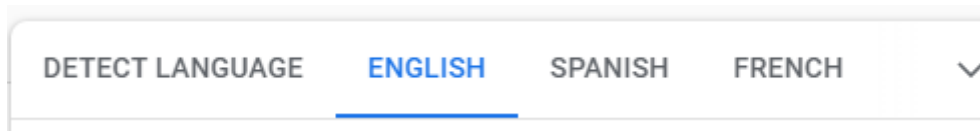
Giai đoạn 4: Chia sẻ bản dịch

Chia sẻ bản dịch (Copy/Share Translation):

- Cung cấp tùy chọn sao chép hoặc chia sẻ bản dịch văn bản và âm thanh ngôn ngữ nói.

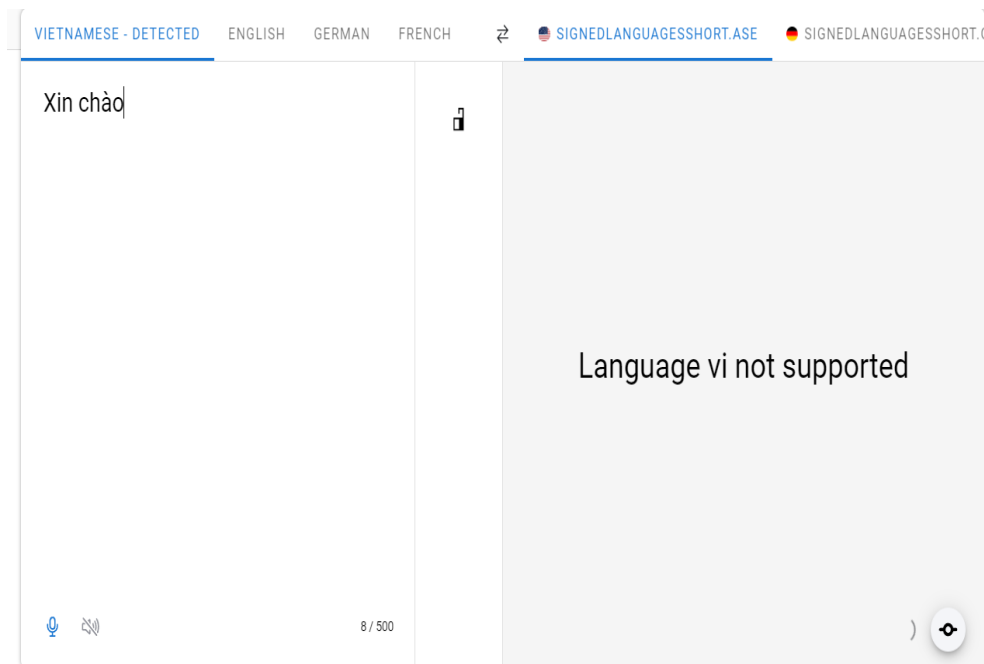
4.3 Nhận diện ngôn ngữ

Phát hiện ngôn ngữ để dịch ngôn ngữ nói sang ngôn ngữ ký hiệu phải xác định ngôn ngữ của văn bản ngôn ngữ nói đầu vào.



Sử dụng nhận dạng tự động bằng cld3 của Google hỗ trợ 107 ngôn ngữ

CLD3 là một mô hình mạng lưới thần kinh để nhận dạng ngôn ngữ. Gói này chứa mã suy luận và mô hình được đào tạo. Mã suy luận trích xuất các ngram ký tự từ văn bản đầu vào và tính toán tỷ lệ số lần mỗi ký tự đó xuất hiện.



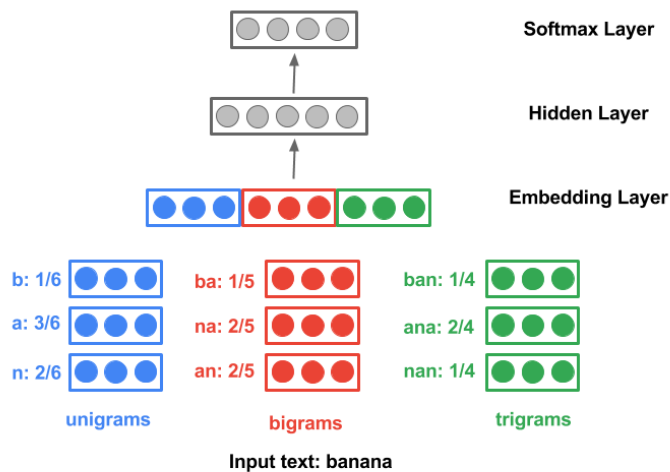
Hình 7 Giao diện đơn giản hệ thống đề tài

Mã suy luận sẽ trích xuất các ngram ký tự từ văn bản đầu vào và tính toán tỷ lệ số lần xuất hiện của từng ký tự. Ví dụ, như thể hiện trong hình bên dưới, nếu văn bản đầu vào là "banana", thì một trong các trigram được trích xuất là "ana" và tỷ lệ tương ứng là 2/4. Các ngram được băm xuống thành một id trong

một phạm vi nhỏ và mỗi id được biểu diễn bằng một vector nhưng dày đặc được ước tính trong quá trình đào tạo.

Mô hình tính trung bình các phần nhúng tương ứng với từng loại ngram theo các phân số và các phần nhúng trung bình được nối với nhau để tạo ra lớp nhúng. Các thành phần còn lại của mạng là lớp ẩn (Tuyến tính được chỉnh sửa) và lớp softmax.

Để có được dự đoán ngôn ngữ cho văn bản đầu vào, chúng tôi chỉ cần thực hiện chuyển tiếp qua mạng.



Hình 8 Mô tả các lớp (Layer) của mô hình mạng CLD3 của Google

4.4 Chuẩn hóa văn bản

Chuẩn hóa văn bản đã cho để dịch ngôn ngữ ký hiệu.










Đưa vào văn bản có thể chứa lỗi chính tả, viết hoa sai, thiếu dấu gạch nối, số, đơn vị hoặc hiện tượng khác yêu cầu cách phát âm đặc biệt. Trả về văn bản được chuẩn hóa cho ngôn ngữ ký hiệu, với các sửa đổi về cách viết hoa, chính tả và gạch nối cũng như cách phát âm đặc biệt cho số và đơn vị.

4.5 Ngôn ngữ kí hiệu – SignWriting

4.5.1 Sign Writing – Illustration

Những người không có kinh nghiệm về SignWriting trước đó sẽ khó hiểu được ký hiệu SignWriting.

Phần này nhằm mục đích cung cấp một cái nhìn thay thế cho SignWriting, sử dụng các hình minh họa do máy tính tạo ra cho các ký hiệu.










	00004	00007	00015
Băng hình			
Đã ký viết			
Hình minh họa			
Lời nhắc	Hình minh họa một người có mái tóc gắn với mũi tên màu đen.	Hình minh họa một người phụ nữ có mái tóc gắn với những mũi tên màu đen.	Hình ảnh minh họa của một người đàn ông có mái tóc gắn. Các mũi tên có màu đen.

Hình 9 Bảng miêu tả sự diễn giải thành tượng hình và ngôn ngữ viết của SignWriting

4.5.2 SignWriting Description

Việc thể hiện bằng văn bản của các ngôn ngữ ký hiệu là một thách thức do sự phức tạp của các hệ thống chữ viết khác nhau. SignWriting, mặc dù có lợi, nhưng vẫn đòi hỏi kiến thức chuyên môn về ngôn ngữ để sử dụng hiệu quả.

Phần này nhằm mục đích cung cấp giải pháp mô tả tự động SignWriting bằng ngôn ngữ nói. Giải pháp này có thể được sử dụng để dạy SignWriting cho người mới học, tinh chỉnh mô hình dịch hoặc suy luận không chính xác trên các mô hình tạo chuyển động.

	Xin chào	Với bàn tay thuận của bạn mở ra, chạm vào trán và đưa tay ra xa, lòng bàn tay hướng ra ngoài.
	Cảm ơn	Chạm bàn tay thuận đang mở của bạn vào mũi, sau đó đưa tay về phía trước, lòng bàn tay hướng lên.
	Giúp anh ấy cô ấy)	Đặt nắm tay của bàn tay thuận (ngón cái hướng lên) lên lòng bàn tay không thuận đang mở của bạn. Cùng nhau di chuyển cả hai tay hướng lên trên.
	KHÔNG	Với bàn tay thuận của bạn, mở rộng ngón trỏ và ngón giữa trong khi vẫn giữ các ngón khác của bạn nhét vào trong. Chạm các ngón tay này vào ngón cái của bạn.
	KHÔNG	Lắc đầu theo chiều ngang trong khi nhú mào.
	Lấy làm tiếc	Nắm tay thuận bằng tay thuận, lòng bàn tay hướng vào trong. Vòng tròn lên trái tim của bạn.
	Bạn bè	Liên kết các ngón trỏ của cả hai tay với nhau, xen kẽ vị trí của chúng.
	Yêu	Khoanh tay trước ngực như thể đang ôm mình, hai tay tạo thành nắm đấm.
	Tên	Với bàn tay thuận của bạn, mở rộng ngón trỏ và ngón giữa. Chạm các ngón tay này hai lần vào ngón trỏ mở rộng của bàn tay không thuận, ngón này được giữ theo chiều ngang.

Hình 10 Bảng giải mã một số SignWriting thành ngôn ngữ nói

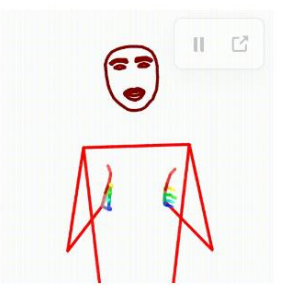
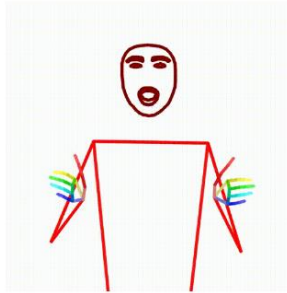
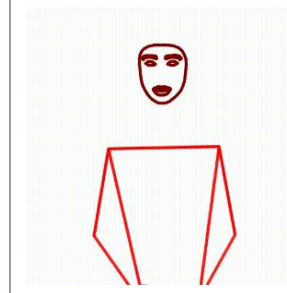
4.6 Pose Anonymization

Xóa thông tin nhận dạng khỏi tư thế ngôn ngữ ký hiệu

Có tồn tại thông tin nhận dạng trong mọi phát ngôn ngôn ngữ ký hiệu. Từ ngoại hình, đến nhịp điệu, kiểu chuyển động và lựa chọn ký hiệu.

Vì vậy, nếu muốn loại bỏ toàn bộ thông tin nhận dạng thì chúng ta cần loại bỏ toàn bộ thông tin.

Bằng cách sử dụng giá trị trung bình của các tư thế ngôn ngữ ký hiệu được tính toán bởi [sign-language-processing/sign-vq](https://www.sign-language-processing.com/sign-vq/), chúng ta có thể ẩn danh chuỗi tư thế bằng cách giả sử khung đầu tiên chỉ là hình dáng bên ngoài của người đó và xóa nó khỏi các khung còn lại, sau đó thêm giá trị trung bình.

Sign	Original__	Anonymized_	Transferred
Kleine			

Hình 11 Các dạng biến đổi khác nhau của khung xương (Skeleton Viewer)

4.7 Fluent (Sign Language) Pose Synthesis

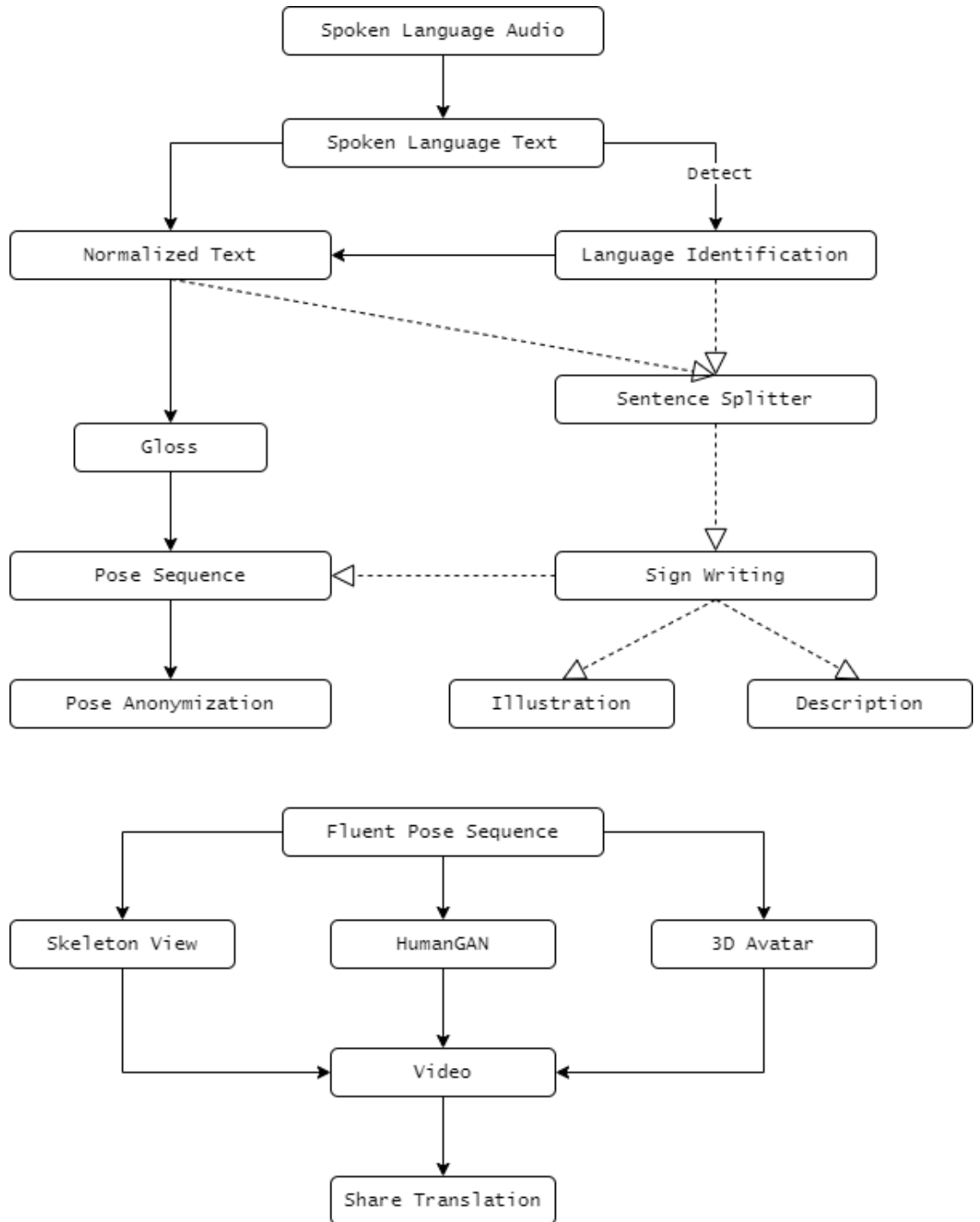
Làm cho các tư thế ngôn ngữ ký hiệu không thành thạo trở nên trôi chảy bằng cách chỉnh sửa các trình tự tư thế

Ví dụ: Dịch text "We were expecting something simple, like a youth hostel" dịch sang glosses DIFFERENT1 IMAGINATION1A LIKE3B* EASY1 YOUNG1* HOME1A. Sau đó, bằng cách sử dụng dịch sang ngôn ngữ kí hiệu, các video được tìm thấy cho từng phần chú giải rồi ghép lại với nhau. Hoặc bằng cách sử dụng [Ham2Pose](#), mỗi HamNoSys được tạo hoạt hình theo trình tự tư thế.

Gloss	HamNoSys	Video
DIFFERENT1^	□□□□□□□□□□	
IMAGINATION1A^	□□□□□□□□	
LIKE3B*	□□□□□□□□□□□□□□□□□□□□	
EASY1	□□□□□□□□□□□□	
YOUNG1*	□□□□□□□□□□□□	
HOME1A	□□□□□□□□□□□□	

Hình 12 Minh họa từ Gloss -> HamNoSys -> Video

CHƯƠNG 5. THIẾT KẾ VÀ TRIỂN KHAI ỨNG DỤNG CHUYỂN ĐỔI NGÔN NGỮ KÝ HIỆU



Hình 13 Kiến trúc toàn bộ đề tài

Pipeline dịch từ text/audio sang ngôn ngữ ký hiệu: nét liền – đã triển khai, nét đứt – chưa triển khai

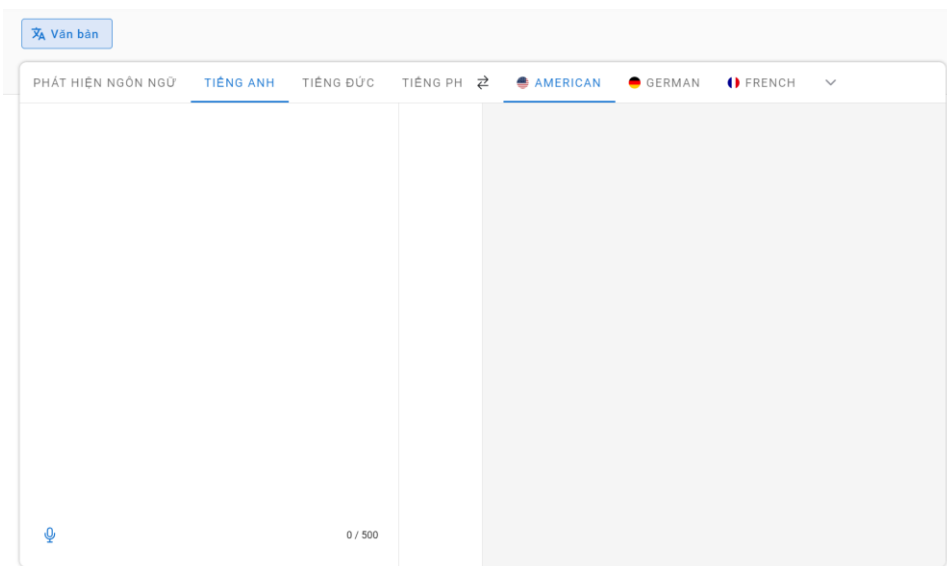
5.1 Thiết kế và triển khai giao diện người dùng

Thiết kế và triển khai giao diện người dùng:

- Xây dựng giao diện ứng dụng trên nền tảng web và di động bằng framework Angular.
- Sử dụng TypeScript, HTML, và SCSS để phát triển các thành phần giao diện.
- Tạo Progressive Web App (PWA) để cung cấp trải nghiệm liền mạch trên cả thiết bị di động và máy tính.

Phát triển backend:

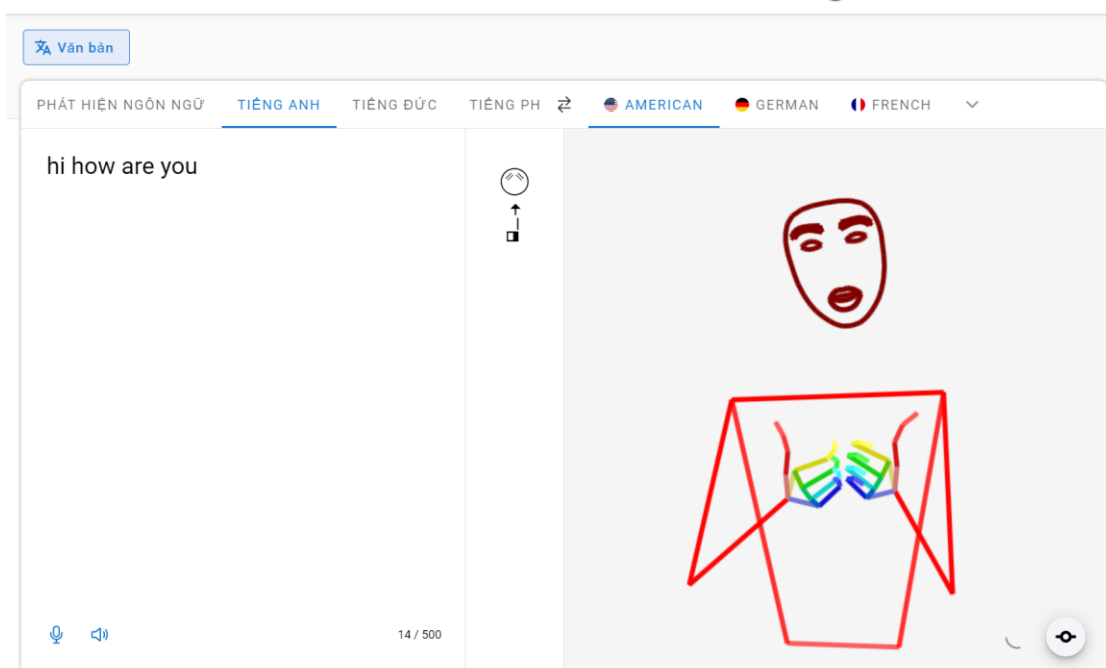
- Sử dụng Node.js để xây dựng server backend.
- Quản lý gói với npm và triển khai Firebase để lưu trữ và xử lý dữ liệu.



Hình 14 Giao diện của hệ thống dịch ngôn ngữ ký hiệu (Web application)

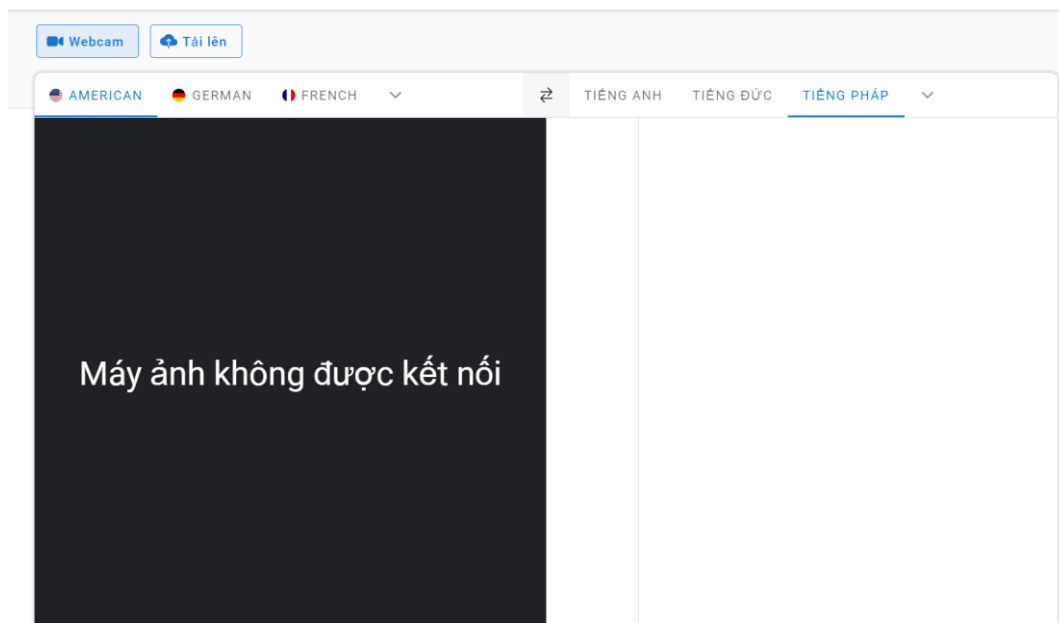
Giao diện người dùng – Web application

Đối với phần dịch từ text sang ngôn ngữ kí hiệu, giao diện gồm 1 text box bên trái để người dùng điền text vào và sẽ convert qua ngôn ngữ kí hiệu



Hình 15 Giao diện dịch từ text sang của khung xương của hệ thống dịch

Giao diện người dùng khi nhập text vô, skeleton view sẽ được hiển thị bên phải, và ở giữa sẽ là SignWriting (ngôn ngữ viết của ngôn ngữ ký hiệu)



Hình 16 Giao diện người dùng khi dịch từ video/camera sang text

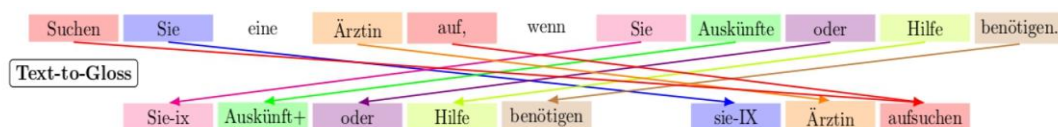
5.2 Normalized Text

Sau khi nhận diện ngôn ngữ xong, Text sẽ được

Chuẩn hóa văn bản đã cho để dịch ngôn ngữ ký hiệu.

Đưa vào văn bản có thể chứa lỗi chính tả, viết hoa sai, thiếu dấu gạch nối, số, đơn vị hoặc hiện tượng khác yêu cầu cách phát âm đặc biệt. Trả về văn bản được chuẩn hóa cho ngôn ngữ ký hiệu, với các sửa đổi về cách viết hoa, chính tả và gạch nối cũng như cách phát âm đặc biệt cho số và đơn vị.

5.3 Dịch Text sang Gloss



Hình 17 Hình minh họa quy trình text-to-gloss của phương pháp chuyển văn bản sang Gloss

Bắt đầu bằng một câu tiếng Đức, hệ thống sẽ áp dụng bản dịch từ văn bản sang Gloss sử dụng một từ dựa trên quy tắc sắp xếp lại và loại bỏ các thành phần không liên quan. Tập hợp các Gloss thu được được sử dụng để tìm kiếm các video có liên quan từ điển ngôn ngữ ký hiệu tiếng Đức Thụy Sĩ (DSGS).

Text-to-Gloss Translation: Văn bản đầu vào (ngôn ngữ nói) trước tiên được xử lý bởi thành phần dịch văn bản sang Gloss, chuyển đổi nó thành một chuỗi các Gloss

Gloss là cách thể hiện ngôn ngữ ký hiệu trong văn bản mà qua đó giữ lại cấu trúc ngữ pháp và các yếu tố không ngôn từ đặc trưng của ngôn ngữ ký hiệu.

Example:

- **English:** I'm happy to meet you.
- **ASL Gloss:** ME HAPPY MEET YOU.

Chúng tôi khám phá ba thành phần khác nhau như một phần của dịch văn bản sang Gloss, bao gồm cả từ vựng, sắp xếp lại và loại bỏ các từ dựa trên quy tắc thành phần và hệ thống dịch máy thần kinh (Neural Machine Translation - NMT).

Lemmatizer

Chúng tôi sử dụng Simplemma, một công cụ đơn giản để chuyển đổi từ về dạng gốc (lemma) trong nhiều ngôn ngữ lập trình Python (Barbaresi, 2023). Việc chuyển đổi này giúp chúng tôi giữ nguyên ý nghĩa của từ đồng thời giảm bớt độ phức tạp của dữ liệu đầu vào. Tuy nhiên, phương pháp này còn hạn chế do chỉ sử dụng kỹ thuật đơn giản không phụ thuộc vào ngữ cảnh, không nắm bắt được thông tin về nghĩa của từ, dẫn đến sự mơ hồ.

Sắp xếp và loại bỏ từ

Chúng tôi tạo ra các bản gần đúng của ngôn ngữ ký hiệu từ văn bản ngôn ngữ nói bằng cách sử dụng phương pháp dựa trên quy tắc. Quá trình chuyển đổi từ câu nói thành chuỗi ký hiệu ngôn ngữ ký hiệu có thể được tóm tắt một cách đơn giản bằng cách loại bỏ biến tố của từ, bỏ qua dấu câu và các từ cụ thể, và sắp xếp lại từ. Để giải

quyết những khác biệt này, chúng tôi áp dụng phương pháp dựa trên quy tắc từ Moryossef et al. (2021) để tạo ra các bản gần đúng từ ngôn ngữ nói: chuyển đổi từ về dạng gốc, xóa từ phụ thuộc vào từ loại và hoán vị thứ tự từ.

Cụ thể, chúng tôi sử dụng spaCy (Montani et al., 2023) để chuyển đổi từ về dạng gốc, gán nhãn từ loại và phân tích cú pháp phụ thuộc. Không giống như Simplelemma, trình chuyển đổi từ về dạng gốc của spaCy là cụ thể cho từng ngôn ngữ và dựa trên ngữ cảnh. Chúng tôi loại bỏ các từ không phải là từ nội dung (ví dụ: mạo từ, giới từ), vì chúng phần lớn không được sử dụng trong ngôn ngữ ký hiệu, nhưng giữ lại đại từ sở hữu và đại từ nhân xưng cũng như danh từ, động từ, tính từ, trạng từ và số. Chúng tôi đưa ra một danh sách ngắn các quy tắc chuyển đổi cú pháp dựa trên ngữ pháp của ngôn ngữ ký hiệu và ngôn ngữ nói tương ứng. Chúng tôi xác định chủ ngữ, động từ và tân ngữ trong văn bản đầu vào và sắp xếp lại chúng để phù hợp với thứ tự được sử dụng trong ngôn ngữ ký hiệu. Ví dụ: đối với tiếng Đức sang Ngôn ngữ Ký hiệu Đức (Deutsche Gebardensprache " , DGS), chúng tôi sắp xếp lại các câu SVO thành SOV, di chuyển các trạng từ bổ nghĩa cho động từ và các từ chỉ vị trí lên đầu câu (một dạng của topicalization), chuyển các từ phủ định về cuối

Trước tiên, chúng tôi chia mỗi câu thành các mệnh đề riêng biệt và sắp xếp lại chúng trước khi áp dụng các quy tắc này cho từng mệnh đề. Việc sắp xếp lại các mệnh đề có thể cần thiết cho các câu điều kiện trong đó mệnh đề phụ điều kiện phải đứng trước mệnh đề chính, như trong "if. . . then. . .". Các quy tắc này cho phép chúng tôi chuyển đổi văn bản ngôn ngữ nói thành các bản gần đúng phù hợp hơn với thứ tự từ và cấu trúc của ngôn ngữ ký hiệu. Nhìn chung, phương pháp dựa trên quy tắc của chúng tôi cung cấp một cách linh hoạt và hiệu quả để tạo ra các bản gần đúng cho ngôn ngữ ký hiệu từ văn bản ngôn ngữ nói, với khả năng kết hợp các quy tắc cụ thể theo ngôn ngữ để nắm bắt các sắc thái của các ngôn ngữ ký hiệu khác nhau. Cách tiếp cận này sử dụng một trình chuyển đổi từ về dạng gốc chính xác hơn, tuy nhiên, nó vẫn gặp phải sự mơ hồ về nghĩa của từ.

Neural Machine Translation

Chúng tôi sử dụng một hệ thống dịch máy thần kinh pretrained model (NMT). Dịch máy thần kinh (Neural Machine Translation - NMT) là một phương pháp thay thế cho việc chuyển đổi văn bản thành các bản gán đúng dựa trên quy tắc.

5.4. Chuyển đổi các Gloss thành Pose

Dịch gloss sang pose là việc chuyển đổi các gloss ngôn ngữ ký hiệu thành một chuỗi các tư thế đại diện đầy đủ cho một chuỗi các ký hiệu.

Khi thu thập được các bộ dữ liệu về ngôn ngữ kí hiệu cho các ngôn ngữ khác nhau. Chúng tôi trích xuất các tư thế xương từ các video này bằng cách sử dụng Mediapipe Holistic (Grishchenko và Bazarevsky, 2020), một khung ước tính tư thế tiên tiến ước tính tọa độ 3D của các điểm mốc khác nhau trên cơ thể người, bao gồm cả khuôn mặt, bàn tay và cơ thể. Chúng tôi tiền xử lý các tư thế bằng cách đảm bảo rằng cổ tay của cơ thể ở cùng một vị trí với cổ tay, loại bỏ chân, tay và mặt khỏi tư thế cơ thể và cắt các video ở đầu và cuối để tránh quay lại vị trí trung lập của cơ thể.

Chúng tôi nối các tư thế cho mỗi gloss bằng cách tìm điểm ‘khâu’ tốt nhất giảm thiểu khoảng cách L2. Sau đó, chúng tôi nối các tư thế này, thêm 0,2 giây ‘đệm’ ở giữa, trước khi áp dụng làm mịn khối trên mỗi khớp để đảm bảo chuyển tiếp mượt mà giữa các ký hiệu và điền vào các điểm chính còn thiếu. Cuối cùng, chúng tôi áp dụng bộ lọc làm mịn chuyển tiếp chuyển động Savitzky-Golay (Savitzky và Golay, 1964), tương tự như Stoll et al. (2020), để giảm chuyển động không tự nhiên.

5.5 Chuyển đổi Pose thành Video

Chúng tôi sử dụng một hệ thống avatar giống người bán thực tế để làm sống động các tư thế được tạo ra. Hệ thống avatar là một mô hình Pix2Pix (Isola et al., 2016) được điều chỉnh để hoạt động trên chuỗi tư thế, không phải hình ảnh riêng lẻ. Đồng thời sử dụng OpenCV (Bradski, 2000) để kết xuất các tư thế thành hình ảnh và đưa chúng vào mô hình Pix2Pix để tạo khung hình video trông giống thật. Hệ thống avatar có thể chạy trong thời gian thực trên các thiết bị được hỗ trợ.

Skeleton Viewer: Trình xem Barebone sử dụng [Pose Viewer](#) (Nhanh, tiêu thụ điện năng thấp, dễ debug)

- Chuỗi tư thế được chuyển thành video, tiết kiệm pin và bộ nhớ
- Khi video đã sẵn sàng, hỗ trợ các thao tác **Sao chép, Tải xuống và Chia sẻ**

Human GAN: Sử dụng mô hình học máy để tạo ra tư thế giống như con người. Dựa vào mô hình (nặng) để tạo hình ảnh có độ phân giải thấp (256x256) và mô hình (nhẹ) để nâng cao phân giải hình ảnh (768x768)

- Các model GANs hiện tại đang không tốt và nhanh và vừa tốt và lâu (Diffusion model)

5.6 Hỗ trợ quốc tế hóa nhiều ngôn ngữ

- Mở rộng trên nhiều bộ dữ liệu
- Hỗ trợ 104 ngôn ngữ và cả bố cục LTR và RTL.
- Sử dụng ngôn ngữ trình duyệt/điện thoại của người dùng và các ngôn ngữ khác nhau thông qua tham số URL.

CHƯƠNG 6. KẾT QUẢ ĐẠT ĐƯỢC VÀ HƯỚNG PHÁT TRIỂN

6.1 Kết quả đạt được

- Hệ thống hỗ trợ chuyển đổi toàn diện từ giọng nói sang văn bản (và ngược lại) cho tất cả các ngôn ngữ được hỗ trợ cục bộ.
- Hệ thống cho phép người dùng lựa chọn ngôn ngữ thủ công (107 ngôn ngữ) và tự động xác định ngôn ngữ bằng công cụ cld3 của Google (107 ngôn ngữ) hoặc giải pháp MediaPipe của Google. Hệ thống cũng sử dụng ngôn ngữ trình duyệt ban đầu và ghi nhớ tùy chọn cặp ngôn ngữ.
- Hệ thống có mô hình chuẩn hóa văn bản đa ngôn ngữ phía máy chủ LLM và mô hình dịch máy đa ngôn ngữ phía máy chủ (chất lượng thấp) để chuyển đổi văn bản ngôn ngữ nói sang SignWriting. Ngoài ra, hệ thống còn hỗ trợ triển khai dịch phía máy khách/máy chủ với Bergamot và phục vụ các mô hình dịch.
- Hệ thống thực hiện chuyển đổi SignWriting thành chuỗi tư thế phía máy chủ (chất lượng thấp), sử dụng tư thế OpenPose, dựa vào văn bản ngôn ngữ nói. Hệ thống cũng có triển khai phía máy chủ mới, tạo hoạt ảnh trực tiếp từ chuỗi SignWriting/HamNoSys và hỗ trợ suy luận ngoại tuyến phía máy khách cho mô hình hoạt ảnh.
- Hệ thống cung cấp trình xem xương cơ bản bằng công cụ Pose Viewer nội bộ (nhanh, tiết kiệm năng lượng và hữu ích cho việc gỡ lỗi), mô hình GAN của con người để tạo dáng cho tư thế giống như con người và mô hình 3D Avatar để tạo hoạt ảnh cho avatar 3D giống người bằng máy học, bao gồm cả hỗ trợ AR.
- Các tính năng bổ sung bao gồm chuyển đổi chuỗi tư thế thành video, hỗ trợ sao chép, tải xuống và chia sẻ video, cũng như hỗ trợ 104 ngôn ngữ và cả bố cục LTR và RTL. Hệ thống sử dụng ngôn ngữ trình duyệt/điện thoại của người dùng và các ngôn ngữ khác nhau thông qua tham số URL.

6.2 Hướng phát triển

- Cải tiến và thu thập bộ dữ liệu Tiếng Việt (Hiện chưa có bộ dữ liệu ngôn ngữ tiên hiệu nào cho ngôn ngữ)
- Cải thiện mô hình HumanGAN

TÀI LIỆU THAM KHẢO

- [1] F. W. Z. C. B. M. J. Y. X. T. Ronglai Zuo, "A Simple Baseline for Spoken Language to Sign Language Translation with 3D Avatars," *Computer Vision and Pattern Recognition*, 2023.
- [2] A. M. M. M. S. E. Zifan Jiang, "Machine Translation between Spoken Languages and Signed Languages Represented in SignWriting," *EACL*, 2023.
- [3] Z. J. Amit Moryossef, "SignBank+: Preparing a Multilingual Sign Language Dataset for Machine Translation Using Large Language Models," September 2023.
- [4] W. a. D. L.-M. Sandler, "Sign Language and Linguistic Universals," *Cambridge University Press*, 2006.
- [5] D. O. K. M. B. L. B. P. B. A. B. N. C. Bragg, "Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective," *The 21st International Acm Sigaccess Conference on Computers and Accessibility*, pp. 16-31, 2019.
- [6] K. A. M. J. H. Y. G. a. M. A. Yin, "Including Signed Languages in Natural Language Processing," *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1, p. 7347–60, 2021.
- [7] U. Nations, "International Day of Sign Languages," 2022.
- [8] W. H. Organization, "Deafness and Hearing Loss.," 2021.
- [9] W. F. o. t. Deaf, "World Federation of the Deaf - Our Work," 2022.
- [10] C. A. a. T. H. Padden, "Deaf in America. Harvard University Press," 1988.

- [11] N. S. a. W. C. H. Glickman, "Language Deprivation and Deaf Mental Health," 2018.
- [12] R. H. M. H. a. D. M. M. Harris, "Research Ethics in Sign Language Communities." *Sign Language Studies* 9 (2), pp. 104-31.
- [13] W. C. Stokoe Jr, "Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf.," *The Journal of Deaf Studies and Deaf Education* 10 (1), p. 3–37, 1960.
- [14] T. P. K. G. M. D. J. N. C. P. C. R. a. S. S. Humphries, "Avoiding Linguistic Neglect of Deaf Children.," *Social Service Review* 90 (4), p. 589–619.
- [15] J. J. W. C. H. a. K. S. Murray, "The Importance of Signed Languages for Deaf Children and Their Families," *The Hearing Journal* 73 (3), pp. 30-32, 2020.
- [16] W. C. L. L. L. a. M. L. A. Hall, "Language Deprivation Syndrome: A Possible Neurodevelopmental Disorder with Sociocultural Origins.," *Social Psychiatry and Psychiatric Epidemiology* 52 (6), p. 761–76, 2017.
- [17] S. K. v. R. E. J. Liddell, "American Sign Language: The Phonological Base.," *Sign Language Studies* 64 (1), p. 195–277, 1989.
- [18] "The Phonological Organization of Sign Languages," *Language and Linguistics Compass* 6 (3);, p. 162–82.
- [19] U. a. S. F. Bellugi, "A Comparison of Sign Language and Spoken Language," *Cognition* 1 (2-3), p. 173–200, 1972.
- [20] "The Phonological Organization of Sign Languages.," *Language and Linguistics Compass* 6 (3), p. 162–82..

- [21] S. K. a. o. Liddell, "Grammar, Gesture, and Meaning in American Sign Language," *Cambridge University Press.*, 2003.
- [22] T. a. A. S. Johnston, "Australian Sign Language (Auslan): An Introduction to Sign Language Linguistics," *Cambridge University Press*, 2007.
- [23] C. a. G. M. Rathmann, "A Featural Approach to Verb Agreement in Signed Languages," *Theoretical Linguistics* 37 (3-4), p. 197–208, 2011.
- [24] J. A. S. a. K. C. Fenlon, "Modification of Indicating Verbs in British Sign Language: A Corpus-Based Study.," *Language* 94 (1), p. 84–118, 2018.
- [25] P. G. Dudis, "Body Partitioning and Real-Space Blends," *Cognitive Linguistics* 15 (2), p. 223–38, 2004.
- [26] S. K. a. M. M. Liddell, "Gesture in Sign Language Discourse," *ournal of Pragmatics* 30 (6), p. 657–97, 1998.
- [27] L. d. Beuzeville, "Pointing and Verb Modification: The Expression of Semantic Roles in the Auslan Corpus.," *In Workshop Programme*, vol. 13, 2008.
- [28] J. A. S. a. K. C. enlon, "Modification of Indicating Verbs in British Sign Language: A Corpus-Based Study," *Language* 94 (1), p. 84–118, 2018.
- [29] T. Supalla, "The Classifier System in American Sign Language.," *Noun Classes and Categorization* 7, p. 181–214, 1986.
- [30] S. a. S. H. Wilcox, "Rethinking Classifiers. Emmorey, K.(Ed.).(2003). Perspectives on Classifier Constructions in Sign Languages. Mahwah, Nj: Lawrence Erlbaum Associates. 332 Pages. Hardcover," *Oxford University Press*, 2004.
- [31] C. B. Roy, "Discourse in Signed Languages," *Gallaudet University Press*, 2011.

- [32] K. S. S. a. Z. S.-S. Cormier, "Rethinking Constructed Action," *Sign Language & Linguistics 18 (2)*, p. 167–204, 2015.
- [33] R. Battison, "Lexical Borrowing in American Sign Language," 1978.
- [34] S. Wilcox, "The Phonetics of Fingerspelling.," *John Benjamins Publishing*, vol. 4, 1992.
- [35] D. a. C. P. Brentari, "A Language with Multiple Origins: Native and Foreign Vocabulary in American Sign Language.," *Foreign Vocabulary in Sign Language: A Cross-Linguistic Investigation of Word Formation*, p. 87–119, 2001.
- [36] C. J. a. R. E. J. Patrie, "Fingerspelled Word Recognition Through Rapid Serial Visual Presentation," 2011.
- [37] C. A. Padden, "1998," *The ASL Lexicon.*" *Sign Language & Linguistics 1 (1)*, p. 39–60.
- [38] K. a. D. B. Montemurro, "Emphatic Fingerspelling as Code-Mixing in American Sign Language.," *Proceedings of the Linguistic Society of America 3 (1)*, p. 61–61.
- [39] T. T. News, "Thành công mới của AI: Chuyển lời nói sang ngôn ngữ ký hiệu".
- [40] R. E. a. S. K. L. Johnson, "Toward a Phonetic Representation of Signs: Sequentiality and Contrast," *Sign Language Studies 11 (2)*, p. 241–74, 2011.
- [41] D. Brentari, "Sign Language Phonology," *The Handbook of Phonological Theory*, p. 691–721, 2011.