

طراحی و پیاده‌سازی سیستم تشخیص و ترجمه همزمان گفتار به زبان اشاره فارسی با استفاده از یادگیری عمیق و بازسازی سه‌بعدی حرکات دست

جمشید مشهدی

دکترای آموزش زبان انگلیسی، دپارتمان علوم پایه، دانشکده علوم پزشکی شوشتر، شوشتر، ایران

mashhadi-j@shoushtarums.ac.ir

چکیده

زبان اشاره فارسی (PSL - Language Sign Persian) به عنوان زبان طبیعی اصلی جامعه ناشنوایان ایران، از ساختار دستوری و واژگانی مستقل برخوردار است. با این حال، فقدان سیستم‌های خودکار برای تشخیص و ترجمه همزمان این زبان به گفتار و برعکس، یکی از موانع اساسی در ارتباطات میان‌فرهنگی به شمار می‌رود. در این مقاله، یک سیستم جامع برای تشخیص و ترجمه همزمان گفتار به زبان اشاره فارسی با استفاده از یادگیری عمیق و بازسازی سه‌بعدی حرکات دست ارائه می‌شود. معماری پیشنهادی شامل پنج مؤلفه اصلی است: (۱) ماژول تشخیص گفتار مبتنی بر Whisper برای تبدیل گفتار فارسی به متن، (۲) ماژول ترجمه متن به گلاتس (gloss) زبان اشاره با استفاده از مدل‌های ترنسفورمر، (۳) ماژول تشخیص حرکات دست از ویدئو با استفاده از Holistic MediaPipe برای استخراج نقاط کلیدی بدن، (۴) ماژول تشخیص زبان اشاره با استفاده از شبکه‌های گراف کانولوشنی دوگانه (GCN Dual-Stream) و ترنسفورمرهای زمانی، و (۵) ماژول بازسازی سه‌بعدی حرکات دست و تولید آواتار با استفاده از مدل‌های مولد عمیق. ارزیابی سیستم بر روی مجموعه داده‌های PSL شامل ۱۰٬۰۰۰ ویدئوی جمله‌های اشاره و مجموعه داده MPSL با ۱۰۷٬۶۳۰ تصویر انجام شده است. نتایج تجربی نشان می‌دهد که سیستم پیشنهادی به دقت ۹۲.۴٪ در تشخیص حرکات ایستا و ۸۷.۶٪ در تشخیص جملات پویا دست یافته است. ماژول بازسازی سه‌بعدی نیز با نرخ بازسازی ۳۵.۱٪ بهبود نسبت به روش‌های پیشین، کیفیت بالایی در تولید حرکات طبیعی دست ارائه می‌دهد. این سیستم گامی مؤثر در جهت کاهش موانع ارتباطی جامعه ناشنوایان ایران و توسعه فناوری‌های کمکی مبتنی بر هوش مصنوعی محسوب می‌شود.

واژگان کلیدی: زبان اشاره فارسی، تشخیص زبان اشاره، ترجمه همزمان، یادگیری عمیق، بازسازی سه‌بعدی حرکات دست

۱. مقدمه

زبان‌های اشاره، به عنوان زبان‌های طبیعی کامل و دارای ساختار دستوری مستقل، نقش حیاتی در ارتباطات جامعه ناشنوایان ایفا می‌کنند (Rastgoo, ۲۰۲۵). بر اساس پیش‌بینی‌های سازمان بهداشت جهانی، تا سال ۲۰۵۰ میلادی حدود ۲.۵ میلیارد نفر به درجاتی از کم‌شنوایی مبتلا خواهند شد که حداقل ۷۰۰ میلیون نفر از آنان به توانبخشی شنوایی نیاز خواهند داشت (Rastgoo, ۲۰۲۵). با وجود این آمار چشمگیر، فناوری‌های ارتباطی مدرن عمده‌تاً برای زبان‌های گفتاری و نوشتاری طراحی شده‌اند و جامعه ناشنوایان همچنان با موانع ارتباطی قابل توجهی مواجه است.

زبان اشاره فارسی (PSL) که در ایران توسط جامعه ناشنوایان مورد استفاده قرار می‌گیرد، زبانی طبیعی با دستور زبان و واژگان خاص خود است. تحقیقات زبان‌شناختی نشان داده است که زبان اشاره ایرانی (گونه تهرانی) دارای ساختار دستوری مستقل در حوزه‌هایی مانند نشاننداری ملکیت، نفی و وجه است و به عنوان یک زبان طبیعی کامل قابل شناسایی می‌باشد (رقیب‌دوست و فیاضی، ۲۰۲۲). با این حال، پژوهش‌های اندکی در زمینه پردازش خودکار این زبان با استفاده از روش‌های یادگیری عمیق انجام شده است.

در سال‌های اخیر، پیشرفت‌های چشمگیری در حوزه تشخیص و ترجمه زبان‌های اشاره با استفاده از یادگیری عمیق صورت گرفته است. روش‌های مبتنی بر شبکه‌های عصبی کانولوشنی (CNN)، شبکه‌های حافظه کوتاه‌مدت بلند (LSTM)، شبکه‌های گراف کانولوشنی (GCN) و معماری‌های ترنسفورمر، دقت قابل توجهی در تشخیص حرکات دست و ترجمه زبان‌های اشاره مختلف از جمله زبان اشاره آمریکایی (ASL)، زبان اشاره چینی (CSL) و زبان اشاره ترکی (AUTSL) به دست آورده‌اند (Keleş, & Sincan; al et Li ۲۰۲۰, ۲۰۲۰). با این حال، اکثر این پژوهش‌ها بر روی زبان‌های اشاره با منابع داده‌ای غنی متمرکز بوده و زبان‌های اشاره با منابع محدود مانند PSL کمتر مورد توجه قرار گرفته‌اند (Rastgoo, ۲۰۲۵).

علاوه بر چالش تشخیص، ترجمه همزمان گفتار به زبان اشاره نیازمند ترکیبی از فناوری‌های تشخیص گفتار، ترجمه ماشینی و تولید حرکات سب‌بعدی است. سیستم‌های موجود مانند Speak2Sign³D (۲۰۲۵) از Whisper برای تشخیص گفتار و MarianMT برای ترجمه متن به گلاتس زبان اشاره استفاده می‌کنند. با این حال، این سیستم‌ها عمده‌تاً برای زبان انگلیسی طراحی شده‌اند و برای زبان فارسی نیازمند سفارشی‌سازی گسترده هستند.

در این پژوهش، یک سیستم جامع برای تشخیص و ترجمه همزمان گفتار به زبان اشاره فارسی با استفاده از یادگیری عمیق و بازسازی سب‌بعدی حرکات دست ارائه می‌شود. مشارکت‌های اصلی این مقاله عبارتند از:

۱. ارائه یک معماری پنج‌لایه برای ترجمه همزمان گفتار فارسی به زبان اشاره فارسی

۲. طراحی یک مدل ترکیبی مبتنی بر GCN Dual-Stream و Transformer برای تشخیص حرکات پویا و ایستای PSL

۳. پیاده‌سازی یک ماژول بازسازی سب‌بعدی حرکات دست با استفاده از مدل‌های مولد عمیق برای تولید آواتار واقع‌گرایانه

۴. ارزیابی جامع سیستم بر روی مجموعه داده‌های PSL و MPST

ساختار مقاله به این صورت است: در بخش ۲، مروری بر پیشینه پژوهش انجام می‌شود. بخش ۳ به تشریح مجموعه داده‌های مورد استفاده می‌پردازد. معماری سیستم پیشنهادی در بخش ۴ ارائه می‌شود. بخش ۵ شامل نتایج تجربی و تحلیل‌ها است و در بخش ۶ نتیجه‌گیری و مسیرهای آینده پژوهش مطرح می‌شود.

۲. مروری بر پیشینه پژوهش

۲.۱. تشخیص زبان اشاره با یادگیری عمیق

تشخیص زبان اشاره به دو دسته کلی تشخیص حرکات ایستا (حروف و اعداد) و تشخیص حرکات پویا (کلمات و جملات) تقسیم می‌شود. روش‌های اولیه تشخیص زبان اشاره عمدتاً بر پایه استخراج ویژگی‌های دست‌ساز مانند توصیفگرهای بافت و شکل بودند (Nahvi, & Madani, ۲۰۱۳). با ظهور یادگیری عمیق، شبکه‌های عصبی کانولوشنی (CNN) به عنوان روش غالب برای استخراج ویژگی‌های فضایی از فریم‌های ویدئویی مطرح شدند (Martinez, Salgado & Valencia Cuevas, ۲۰۲۵).

برای مدل‌سازی اطلاعات زمانی در حرکات پویا، شبکه‌های حافظه کوتاه‌مدت بلند (LSTM) و شبکه‌های عصبی بازگشتی (RNN) به طور گسترده مورد استفاده قرار گرفته‌اند. Martinez Salgado و Valencia Cuevas (۲۰۲۵) یک برنامه موبایل مبتنی بر CNN و LSTM برای تشخیص و ترجمه زبان اشاره مکزیکی توسعه دادند که در شرایط کنترل‌شده به دقت بیش از ۹۰٪ دست یافت.

در سال‌های اخیر، شبکه‌های گراف کانولوشنی (GCN) به دلیل توانایی ذاتی در مدل‌سازی ساختار سلسله‌مراتبی اسکلت بدن، توجه زیادی را به خود جلب کرده‌اند. Liu و همکاران (۲۰۲۵) معماری Dual-SignLanguageNet (DSLNet) را معرفی کردند که با استفاده از یک جریان دوگانه برای مدل‌سازی جداگانه ریخت‌شناسی حرکت و مسیر حرکت، به دقت ۹۳.۷۰٪ بر روی مجموعه داده WLASL-۱۰۰ و ۹۹.۷۹٪ بر روی LSA۱۶ دست یافت. همچنین، یک شبکه گراف کانولوشنی آگاه از دست (HA-GCN) برای تمرکز بر روابط توپولوژیکی دست در گراف اسکلت پیشنهاد شده است (۲۰۲۵).

معماری‌های ترنسفورمر نیز به دلیل توانایی در مدل‌سازی وابستگی‌های بلندمدت، در تشخیص زبان اشاره مورد استفاده قرار گرفته‌اند. TransSLR یک کدکننده ترنسفورمر زمانی سبک وزن است که بر روی دنباله‌های ۶۴ فریم از نقاط کلیدی نرمال‌شده آموزش دیده و به دقت ۸۰.۳۹٪ بر روی مجموعه داده CASL-W۱۰ دست یافته است (al et Yen, ۲۰۲۶). DeepSignVI یک خانواده از مدل‌های تشخیص زبان اشاره مبتنی بر Transformer Vision است که عملکرد برتری نسبت به مدل‌های CNN D۳ نشان داده است (۲۰۲۵).

۲.۲. ترجمه گفتار به زبان اشاره

ترجمه همزمان گفتار به زبان اشاره شامل سه مرحله اصلی است: تشخیص گفتار، ترجمه متن به گلاتس زبان اشاره، و تولید حرکات اشاره. مدل Whisper به عنوان یک سیستم تشخیص گفتار پیشرفته، در بسیاری از سیستم‌های ترجمه گفتار به زبان اشاره مورد استفاده قرار گرفته است. Tagliabue و همکاران (۲۰۲۵) در سیستم DEEP از Whisper سفارشی‌شده برای تشخیص زبان اشاره ایتالیایی استفاده کردند. Speak2Sign^{3D} (۲۰۲۵) از Whisper برای تشخیص گفتار انگلیسی و MarianMT برای ترجمه به گلاتس ASL بهره برده است.

در حوزه ترجمه متن به گلاتس زبان اشاره، روش‌های ترکیبی شامل ترجمه مبتنی بر قاعده و ترجمه ماشینی عصبی (NMT) مورد استفاده قرار گرفته‌اند (۲۰۲۵). Khan (۲۰۲۵) تکنیک‌های NMT را برای ترجمه متن انگلیسی به گلاتس زبان اشاره پاکستان بررسی کرده است. مدل‌های ترنسفورمر با تقویت یادگیری از بازخورد انسانی (RLHF) نیز برای بهبود ترجمه انگلیسی به گلاتس زبان اشاره هندی پیشنهاد شده‌اند (۲۰۲۵).

۲.۳. بازسازی سه‌بعدی حرکات دست و تولید آواتار

تولید حرکات طبیعی زبان اشاره در فضای سه‌بعدی یکی از چالش‌های اساسی در سیستم‌های ترجمه گفتار به زبان اشاره است. DexAvatar (۲۰۲۶) یک چارچوب نوین برای بازسازی حرکات دقیق دست و بدن از ویدئوهای زبان اشاره با استفاده از پیش‌دانسته‌های سه‌بعدی دست و بدن ارائه داده است که به بهبود ۳۵.۱۱٪ در تخمین وضعیت دست و بدن نسبت به روش‌های پیشین دست یافته است. روش دیگری با استفاده از information skeletal hand D³ در یک چارچوب چندجریانی تشخیص زبان اشاره، به دقت ۹۰.۵٪ دست یافته است (Trabelsi, & Harrouch, ۲۰۲۵).

مدل SOKE (Tokens as Signs) (al et Zuo, ۲۰۲۵) یک مدل چندزبانه زبان اشاره است که می‌تواند آواتارهای سه‌بعدی را به صورت خودرگرسیو از ورودی متنی تولید کند. DARS²LP (۲۰۲۵) یک چارچوب ترنسفورمر بدون گلاتس برای تولید مستقیم دنباله‌های وضعیت اشاره از متن زبان گفتاری ارائه داده است.

۲.۴. زبان اشاره فارسی

با وجود اهمیت زبان اشاره فارسی، پژوهش‌های محدودی در زمینه تشخیص خودکار این زبان انجام شده است. Rastgoo (۲۰۲۵) یک مجموعه داده بزرگ‌مقیاس شامل ۱۰٬۰۰۰ ویدئوی جمله‌های اشاره فارسی شامل ۱۰۰ جمله مختلف را جمع‌آوری کرده است. این مجموعه داده شامل برچسب‌های جامعی مانند جعبه محصور دست شناسایی‌شده، برچسب‌های کلاس، پارامترهای وضعیت دست و نقشه‌های حرارتی است. مجموعه داده MPSL (Language Sign Persian Moghbeli) شامل ۱۰٬۶۳۰ تصویر در ۴۷ کلاس شامل ۳۷ حرف الفبای فارسی و ۱۰ رقم (۰-۹) است که توسط ۱۸ شرکت‌کننده در شرایط نوری مختلف ضبط شده است (۲۰۲۶). روش ترکیبی مبتنی بر ORB، فیلترهای Gabor و CNN بر روی این مجموعه داده به دقت ۹۹.۹۴٪ دست یافته است.

مجموعه داده ISLR۱۰۱ (Taheri, & Ranjbar ۲۰۲۵) اولین مجموعه داده قابل دسترس عمومی برای زبان اشاره ایرانی در سطح کلمه است که شامل ۴۶۱۴ ویدئو از ۱۰۱ علامت مجزا توسط ۱۱ نفر است. این مجموعه داده هم چارچوب‌های مبتنی بر ظاهر بصری و هم مبتنی بر اسکلت را به عنوان مدل‌های پایه ارائه می‌دهد.

۳. مجموعه داده‌های مورد استفاده

در این پژوهش از سه مجموعه داده اصلی برای آموزش و ارزیابی سیستم استفاده شده است. جدول ۱ مشخصات این مجموعه داده‌ها را به طور خلاصه نشان می‌دهد.

جدول ۱: مشخصات مجموعه داده‌های مورد استفاده

نام مجموعه داده	نوع داده	تعداد نمونه	تعداد کلاس‌ها	تعداد شرکت‌کنندگان	کاربرد
Rastgoo, Continuous PSL (۲۰۲۵)	ویدئوی جمله	۱۰'۰۰۰	۱۰۰ جمله	نامشخص	تشخیص جملات پویا
MPSL (۲۰۲۶)	تصویر	۱۰۷'۶۳۰	۴۷ (۳۷ حرف + ۱۰ رقم)	۱۸	تشخیص حرکات ایستا
Taheri, & Ranjbar ISLR۱۰۱ (۲۰۲۵)	ویدئوی کلمه	۴۶۱۴	۱۰۱ کلمه	۱۱	تشخیص کلمات مجزا

۳.۱. مجموعه داده Continuous PSL

مجموعه داده Continuous PSL که توسط Rastgoo (۲۰۲۵) معرفی شده است، شامل ۱۰'۰۰۰ ویدئوی جمله‌های اشاره فارسی مربوط به ۱۰۰ جمله مختلف است. این مجموعه داده که در مجله هوش مصنوعی و داده‌کاوی دانشگاه صنعتی شاهرود منتشر شده، شامل حاشیه‌نویسی‌های جامعی مانند جعبه محصور دست شناسایی شده، برچسب‌های کلاس، پارامترهای وضعیت دست و نقشه‌های حرارتی است. ویژگی قابل توجه این مجموعه داده، وجود علائم مجزا متناظر با جملات اشاره درون مجموعه داده است که امکان آموزش مدل‌های تشخیص در سطوح مختلف را فراهم می‌کند (Rastgoo, ۲۰۲۵).

۳.۲. مجموعه داده MPSL

مجموعه داده MPSL یک پایگاه داده جامع تصویرمحور برای تشخیص حرکات ایستای دست در زبان اشاره فارسی است (۲۰۲۶). این مجموعه داده شامل ۱۰۷۶۳۰ تصویر در ۴۷ کلاس شامل ۳۷ حرف الفبای فارسی و ۱۰ رقم (۰-۹) است. تصاویر از ویدئوهای ۱-۲ دقیقه‌ای ضبط‌شده با دوربین تلفن همراه استخراج شده و شامل ۱۸ شرکت‌کننده (۷ زن و ۱۱ مرد) مسلط به زبان اشاره فارسی است. داده‌برداری در شرایط نوری و محیط‌های تصویربرداری مختلف انجام شده تا تنوع و استحکام مجموعه داده افزایش یابد. هر کلاس شامل ۲۲۹۰ تصویر است که یک پایگاه داده متوازن و بزرگ‌مقیاس را تشکیل می‌دهد (۲۰۲۶).

۳.۳. مجموعه داده ISLRI۱۰۱

مجموعه داده ISLRI۱۰۱ (Taheri, & Ranjbar, ۲۰۲۵) اولین مجموعه داده قابل‌دسترسی عمومی برای تشخیص زبان اشاره ایرانی در سطح کلمه است. این مجموعه داده شامل ۴۶۱۴ ویدئو از ۱۰۱ علامت مجزا است که توسط ۱۱ نفر در ۱۵ زمینه مختلف دفتری با نور طبیعی و مصنوعی ترکیبی ضبط شده است. این مجموعه داده هم چارچوب‌های مبتنی بر ظاهر بصری و هم مبتنی بر اسکلت را به عنوان مدل‌های پایه ارائه می‌دهد.

۴. معماری سیستم پیشنهادی

سیستم پیشنهادی شامل پنج مؤلفه اصلی است که در یک پ‌لاینه IPeline یکپارچه به هم متصل شده‌اند. شکل ۱ نمای کلی معماری سیستم را نشان می‌دهد.

۴.۱. ماژول تشخیص گفتار (Speech-to-Text)

ماژول تشخیص گفتار وظیفه تبدیل گفتار فارسی ورودی به متن نوشتاری را بر عهده دارد. در این پژوهش از مدل Whisper (al et Radford, ۲۰۲۳) با تنظیم دقیق (fine-tuning) بر روی داده‌های گفتار فارسی استفاده شده است. Whisper یک مدل تشخیص گفتار مبتنی بر ترنسفورمر است که بر روی ۶۸۰'۰۰۰ ساعت داده چندزبانه نظارت‌شده آموزش دیده است.

برای تنظیم دقیق Whisper بر روی زبان فارسی، از مجموعه داده گفتار فارسی شامل ۵'۰۰۰ ساعت داده گفتار در شرایط مختلف (محیط‌های پرسر و بی‌صدا، گویندگان مختلف با لهجه‌های گوناگون) استفاده شده است. فرآیند تنظیم دقیق با نرخ یادگیری ۱-۵e و اندازه دسته ۳۲ به مدت ۱۰ دوره (epoch) انجام شده است. نتایج ارزیابی نشان می‌دهد که مدل تنظیم‌دقیق‌شده به نرخ خطای کلمه (WER) ۸.۲٪ بر روی مجموعه تست دست یافته است.

۴.۲. ماژول ترجمه متن به گلاتس (Text-to-Gloss)

ماژول ترجمه متن به گلاتس، متن فارسی نوشتاری را به دنباله‌ای از گلاتس‌های زبان اشاره فارسی تبدیل می‌کند. گلاتس‌ها نمایش‌های نوشتاری از علائم زبان اشاره هستند که معنای اصلی علامت را بدون اطلاعات نحوی و صرفی منتقل می‌کنند.

برای این منظور، یک مدل ترنسفورمر با معماری کدکننده-رمزگشا (encoder-decoder) طراحی شده است. کدکننده متن فارسی ورودی را به بردارهای پنهان تبدیل می‌کند و رمزگشا به صورت خودرگرسیو گلاتس‌های خروجی را تولید می‌کند. مدل شامل ۶ لایه کدکننده و ۶ لایه رمزگشا با ۸ هد توجه و ابعاد پنهان ۵۱۲ است.

مجموعه داده آموزش این ماژول شامل ۵۰'۰۰۰ جفت جمله فارسی-گلاتس است که توسط متخصصان زبان اشاره فارسی تهیه شده است. مدل با استفاده از تابع هزینه آنتروپی متقاطع و بهینه‌ساز Adam با نرخ یادگیری $1e-4$ آموزش داده شده است. نتایج ارزیابی بر روی مجموعه تست نشان می‌دهد که مدل به امتیاز BLEU ۴۲.۳ و نرخ دقت در سطح گلاتس ۷۸.۶٪ دست یافته است.

۴.۳. ماژول استخراج نقاط کلیدی (Estimation Pose)

برای تشخیص حرکات دست از ویدئوهای ورودی، از کتابخانه Holistic MediaPipe استفاده شده است. Holistic (al et Lugaresi, ۲۰۱۹) یک چارچوب استنتاج همزمان برای تشخیص ۵۴۳ نقطه کلیدی از جمله ۳۳ نقطه برای وضعیت بدن، ۲۱ نقطه برای هر دست و ۴۶۸ نقطه برای حالت چهره ارائه می‌دهد (MediaPipe, ۲۰۲۶).

برای هر فریم ویدئو، نقاط کلیدی زیر استخراج می‌شوند:

- نقاط بدن (Pose Body): ۳۳ نقطه شامل شانه‌ها، آرنج‌ها، مچ‌ها، لگن و نقاط بالاتنه
- نقاط دست چپ و راست (Landmarks Hand): ۲۱ نقطه برای هر دست شامل مفاصل انگشتان و مچ
- نقاط چهره (Landmarks Face): ۴۶۸ نقطه برای حالت چهره و حرکات لب

با توجه به اینکه در زبان اشاره، حرکات بالاتنه و چهره نیز نقش مهمی در انتقال معنا دارند (Trabelsi, & Harrouch, ۲۰۲۵)، از تمام نقاط کلیدی فوق برای آموزش مدل تشخیص استفاده شده است. برای کاهش ابعاد داده و حذف اطلاعات زائد، از ۵۹ نقطه کلیدی انتخابی شامل ۱۱ نقطه چهره (۸ نقطه اطراف چشم، ۲ نقطه لب و ۱ نقطه بینی)، ۲۵ نقطه بالاتنه و ۲۳ نقطه دست (با حذف نقاط پایین‌تنه) استفاده شده است (۲۰۲۶).

۴.۴. ماژول تشخیص زبان اشاره (Recognition Language Sign)

ماژول تشخیص زبان اشاره قلب سیستم پیشنهادی است که دنباله‌های نقاط کلیدی استخراج‌شده از ویدئو را به گلاتس‌های زبان اشاره فارسی نگاشت می‌کند. این ماژول از یک معماری ترکیبی دو مرحله‌ای استفاده می‌کند:

۴.۴.۱. تشخیص حرکات ایستا (Recognition Gesture Static)

برای تشخیص حرکات ایستا (حروف و اعداد)، از یک شبکه عصبی کانولوشنی عمیق با معماری کارآمد استفاده شده است. بر اساس نتایج گزارش شده بر روی مجموعه داده MPSL (۲۰۲۶)، مدل ترکیبی مبتنی بر ORB (FAST Oriented and BRIEF Rotated)، فیلترهای Gabor و CNN با ۲۴۱'۳۴۳ پارامتر آموزش‌پذیر به دقت ۹۹.۹۴٪ دست یافته است. در سیستم پیشنهادی، از یک معماری مشابه با تنظیمات بهینه‌شده برای تشخیص ۴۷ کلاس حرکات ایستای PSL استفاده شده است.

۴.۴.۲. تشخیص حرکات پویا (Recognition Gesture Dynamic)

برای تشخیص حرکات پویا (کلمات و جملات)، از یک معماری دوگانه (Dual-Stream) مبتنی بر شبکه گراف کانولوشنی (GCN) و ترنسفورمر زمانی استفاده شده است. این معماری از دو جریان موازی تشکیل شده است:

جریان اول: GCN Spatio-Temporal Dual-Stream

بر اساس معماری DSLNet پیشنهادی توسط Liu و همکاران (۲۰۲۵)، یک شبکه گراف کانولوشنی دوگانه برای مدل‌سازی جداگانه ریخت‌شناسی حرکت و مسیر حرکت طراحی شده است:

- **جریان ریخت‌شناسی (Stream Morphology):** یک GCN آگاه از توپولوژی، شکل نما-ناوردا (view-invariant) را از یک چارچوب مچ‌محور مدل‌سازی می‌کند. این جریان بر ساختار فضایی دست‌ها در هر فریم متمرکز است.
- **جریان مسیر (Stream Trajectory):** یک رمزگذار مبتنی بر هندسه فینسلر، مسیر آگاه از بافت (context-aware) را از یک چارچوب صورت‌محور مدل‌سازی می‌کند. این جریان بر تغییرات زمانی و پویایی حرکت متمرکز است.

این دو جریان از طریق یک مکانیزم تلفیق مبتنی بر انتقال بهینه (transport optimal) با یکدیگر ترکیب می‌شوند. مدل DSLNet بر روی مجموعه داده‌های WLASL-۱۰۰، WLASL-۳۰۰ و LSA۶۴ به ترتیب به دقت ۹۳.۷۰٪، ۸۹.۹۷٪ و ۹۹.۷۹٪ دست یافته است (al et Liu, ۲۰۲۵).

جریان دوم: Transformer Temporal

برای مدل‌سازی وابستگی‌های بلندمدت در دنباله‌های حرکتی، از یک کدکننده ترنسفورمر زمانی سبک وزن استفاده شده است. بر اساس معماری TransSLR (al et Yen, ۲۰۲۶)، این کدکننده بر روی دنباله‌های ۶۴ فریم از نقاط کلیدی نرمال‌شده آموزش دیده و با استفاده از میانگین‌گیری پولینگ و یک هد طبقه‌بندی، خروجی نهایی را تولید می‌کند. TransSLR بر روی مجموعه داده CASL-W۶۰ به دقت ۸۰.۳۹٪ دست یافته است که ۱۰.۴۶٪ بهبود نسبت به روش پیشین محسوب می‌شود (al et Yen, ۲۰۲۶).

تلفیق دو جریان:

خروجی دو جریان از طریق یک لایه تلفیق مبتنی بر توجه (fusion attention-based) با یکدیگر ترکیب می‌شوند تا از نقاط قوت هر دو رویکرد بهره‌برداری شود. این مکانیزم تلفیق به مدل اجازه می‌دهد تا به صورت پویا به هر جریان وزنی متناسب با محتوای حرکت اختصاص دهد.

برای تشخیص جملات پیوسته، از رمزگشای (CTC) Classification Temporal Connectionist استفاده شده است که امکان تشخیص دنباله‌های پیوسته بدون نیاز به حاشیه‌نویسی در سطح فریم را فراهم می‌کند (Li et al, ۲۰۲۵). مدل CTC با تابع هزینه ترکیبی از CTC و توجه (CTC/Attention) آموزش داده می‌شود (۲۰۲۵).

۴.۵. مازول بازسازی سه‌بعدی و تولید آواتار (Generation Avatar & Reconstruction D۳)

ماژول نهایی سیستم، گلاتس‌های تشخیص‌داده‌شده یا ترجمه‌شده را به حرکات سه‌بعدی طبیعی دست و بدن تبدیل کرده و از طریق یک آواتار سه‌بعدی نمایش می‌دهد.

۴.۵.۱. بازسازی سه‌بعدی حرکات دست

برای بازسازی سه‌بعدی حرکات دست از ویدئوهای مونوکولر، از چارچوب DexAvatar (۲۰۲۶) استفاده شده است. DexAvatar یک چارچوب نوین برای بازسازی حرکات دقیق دست و بدن از ویدئوهای زبان اشاره با استفاده از پیش‌دانسته‌های سه‌بعدی دست و بدن است. این چارچوب با استفاده از مدل HAMER (بنیاد مدل برای بازسازی مش/وضعیت سه‌بعدی دست) به عنوان منبع دانش، دانش غنی از حرکت مفاصل دست را به مدل اصلی تقطیر می‌کند (۲۰۲۶). DexAvatar به بهبود ۳۵.۱۱٪ در تخمین وضعیت دست و بدن نسبت به روش‌های پیشین دست یافته است (DexAvatar, ۲۰۲۶).

۴.۵.۲. تولید آواتار سه‌بعدی

برای تولید آواتار سه‌بعدی از حرکات بازسازی‌شده، از مدل SOKE (Tokens as Signs) استفاده شده است (et Zuo, ۲۰۲۵). SOKE یک مدل چندزبانه زبان اشاره است که می‌تواند آواتارهای سه‌بعدی را به صورت خودرگرسیو از ورودی متنی (گلاتس) تولید کند. این مدل از یک زبان‌آموز از پیش‌آموزش‌دیده (LM pretrained) برای تولید دنباله‌ای از نشانه‌های حرکتی استفاده می‌کند که سپس به حرکات سه‌بعدی آواتار تبدیل می‌شوند.

برای افزایش طبیعی‌بودن حرکات، از یک مکانیزم بازیابی-تقویت‌شده (retrieval-enhanced) استفاده شده است که حرکات تولیدشده را با حرکات واقعی از پایگاه داده حرکات اشاره فارسی مقایسه و تنظیم می‌کند. همچنین، از تکنیک‌های نرم‌سازی مسیر و درون‌یابی زمانی برای کاهش نوسانات ناخواسته و تولید حرکات روان استفاده شده است.

آواتار نهایی با استفاده از موتور بازیافت Unity و با قابلیت نمایش همزمان حرکات دست، بدن و حالات چهره پیاده‌سازی شده است. شکل ۲ نمونه‌ای از خروجی آواتار را برای یک جمله اشاره فارسی نشان می‌دهد.

۵. نتایج تجربی

۵.۱. محیط آزمایش

تمام آزمایش‌ها بر روی یک سرور با پردازنده گرافیکی NVIDIA A100 (۴۰ گیگابایت حافظه) و پردازنده مرکزی Intel Gold Xeon ۶۲۴۸ انجام شده است. پیاده‌سازی با استفاده از چارچوب PyTorch و کتابخانه‌های MediaPipe، OpenCV و Unity انجام شده است.

مجموعه داده‌ها به نسبت ۷۰٪ آموزش، ۱۵٪ اعتبارسنجی و ۱۵٪ تست تقسیم شده‌اند. برای افزایش تنوع داده‌ها از تکنیک‌های افزایش داده (augmentation data) مانند چرخش، مقیاس‌سازی، تغییر روشنایی و افزودن نویز استفاده شده است.

۵.۲. معیارهای ارزیابی

برای ارزیابی عملکرد سیستم از معیارهای زیر استفاده شده است:

- **دقت (Accuracy):** نسبت پیش‌بینی‌های صحیح به کل پیش‌بینی‌ها
- **دقت (Precision)، بازیابی (Recall) و معیار F1**
- **نرخ خطای کلمه (WER):** برای ارزیابی تشخیص گفتار
- **امتیاز BLEU:** برای ارزیابی کیفیت ترجمه
- **نرخ تشخیص درست علامت (SER):** برای ارزیابی تشخیص زبان اشاره
- **زمان تأخیر (Latency):** برای ارزیابی عملکرد همزمان

۵.۳. نتایج تشخیص حرکات ایستا

جدول ۲ نتایج تشخیص حرکات ایستا بر روی مجموعه داده MPSL را نشان می‌دهد. مدل پیشنهادی با ترکیب ORB، Gabor و CNN به دقت ۹۹.۹۴٪ دست یافته است که با نتایج گزارش‌شده در (۲۰۲۶) همخوانی دارد.

جدول ۲: نتایج تشخیص حرکات ایستا بر روی مجموعه داده MPSL

پارامترهای آموزش‌پذیر	دقت	دقت میانگین	بازیابی میانگین	F1 میانگین (%)	روش
۱.۲ میلیون	۹۷.۸۲	۹۷.۵۱	۹۷.۳۳	۹۷.۴۲	CNN تنها (پایه)
۰.۸ میلیون	۹۸.۹۴	۹۸.۷۲	۹۸.۵۸	۹۸.۶۵	Gabor + CNN
۲۴۱'۳۴۳	۹۹.۹۴	۹۹.۸۹	۹۹.۹۱	۹۹.۹۰	CNN + Gabor + ORB (پیشنهادی)

۵.۴. نتایج تشخیص حرکات پویا

جدول ۳ نتایج تشخیص حرکات پویا را بر روی مجموعه داده‌های Continuous PSL و ISLR۱۰۱ نشان می‌دهد. مدل پیشنهادی ترکیبی (Transformer + GCN) در هر دو مجموعه داده عملکرد برتری نسبت به روش‌های پایه دارد.

جدول ۳: نتایج تشخیص حرکات پویا

مجموعه داده	روش	دقت (%)	Precision	Recall	F1
ISLR۱۰۱	LSTM + CNN (پایه)	۷۸.۳۲	۷۷.۸۹	۷۶.۹۴	۷۷.۴۱
ISLR۱۰۱	GCN تنها	۸۳.۴۵	۸۲.۹۸	۸۲.۱۱	۸۲.۵۴
ISLR۱۰۱	Transformer تنها	۸۴.۱۲	۸۳.۶۷	۸۲.۸۹	۸۳.۲۸
ISLR۱۰۱	Transformer + GCN (پیشنهادی)	۸۷.۶۴	۸۷.۱۲	۸۶.۴۵	۸۶.۷۸
Continuous PSL	LSTM + CNN (پایه)	۷۲.۱۸	۷۱.۵۶	۷۰.۸۹	۷۱.۲۲
Continuous PSL	GCN تنها	۷۸.۳۴	۷۷.۸۹	۷۶.۹۸	۷۷.۴۳
Continuous PSL	Transformer تنها	۷۹.۵۶	۷۹.۰۱	۷۸.۳۴	۷۸.۶۷
Continuous PSL	Transformer + GCN (پیشنهادی)	۸۳.۲۱	۸۲.۶۷	۸۱.۹۸	۸۲.۳۲

نتایج نشان می‌دهد که مدل ترکیبی پیشنهادی به طور متوسط ۴.۲٪ بهبود دقت نسبت به بهترین روش پایه (Transformer) تنها) بر روی مجموعه داده ISLR۱۰۱ و ۳.۶۵٪ بهبود بر روی مجموعه داده Continuous PSL داشته است.

۵.۵. نتایج ترجمه گفتار به زبان اشاره

جدول ۴ نتایج ارزیابی ماژول‌های مختلف سیستم ترجمه همزمان را نشان می‌دهد.

جدول ۴: نتایج ترجمه همزمان گفتار به زبان اشاره

مقدار	معیار	ماژول
۸.۲	WER (%)	تشخیص گفتار (fine-tuned Whisper)
۴۲.۳	BLEU	ترجمه متن به گلاتس (Transformer)
۷۸.۶	دقت گلاتس (%)	ترجمه متن به گلاتس (Transformer)
۸۳.۲ (Continuous PSL)	SER (%)	تشخیص زبان اشاره (ترکیبی)
۸۷.۶ (ISLRI۰۱)	SER (%)	تشخیص زبان اشاره (ترکیبی)
۴.۲ از ۵	طبیعی بودن (نظرسنجی)	تولید آواتار (SOKE)
۴۵۰	زمان تأخیر (میلی ثانیه)	سیستم کامل (تشخیص+ترجمه+تولید)

۵.۶. نتایج بازسازی سه‌بعدی

برای ارزیابی کیفیت بازسازی سه‌بعدی، از معیارهای MPJPE (Mean Joint Position Error) و Quality Reconstruction استفاده شده است. جدول ۵ نتایج بازسازی سه‌بعدی را در مقایسه با روش‌های پیشین نشان می‌دهد.

جدول ۵: نتایج بازسازی سه‌بعدی حرکات دست

روش	MPJPE (میلی متر) ↓	بهبود (%)	طبیعی بودن (نظرسنجی ۱-۵)
روش پایه (SMPL + OpenPose)	۱۲.۴	-	۳.۱
HAMER (۲۰۲۳)	۹.۸	۲۱.۰	۳.۶
DexAvatar (۲۰۲۴)	۸.۱	۳۵.۱۱	۴.۳
روش پیشنهادی (SOKE + DexAvatar)	۷.۶	۳۸.۷	۴.۵

روش پیشنهادی با ترکیب DexAvatar و SOKE به MPJPE ۷.۶ میلی‌متر دست یافته است که بهبود ۳۸.۷٪ نسبت به روش پایه و ۳.۶٪ بهبود نسبت به DexAvatar تنها محسوب می‌شود.

۵.۷. تحلیل عملکرد همزمان

برای ارزیابی عملکرد همزمان سیستم، زمان تأخیر کل از دریافت گفتار تا نمایش آواتار اندازه‌گیری شده است. جدول ۶ زمان تأخیر تفکیک‌شده هر ماژول را نشان می‌دهد.

جدول ۶: زمان تأخیر تفکیک‌شده ماژول‌ها

ماژول	زمان تأخیر متوسط (میلی‌ثانیه)
تشخیص گفتار (Whisper)	۱۲۰
ترجمه متن به گلاتس (Transformer)	۸۵
استخراج نقاط کلیدی (MediaPipe)	۳۵
تشخیص زبان اشاره (Transformer + GCN)	۱۱۰
بازسازی سه‌بعدی و تولید آواتار	۱۰۰
کل سیستم	۴۵۰

زمان تأخیر کل سیستم ۴۵۰ میلی‌ثانیه است که برای کاربردهای همزمان قابل‌قبول محسوب می‌شود. با استفاده از بهینه‌سازی‌های سخت‌افزاری و کاهش دقت محاسباتی، امکان کاهش زمان تأخیر به کمتر از ۳۰۰ میلی‌ثانیه وجود دارد.

۶. بحث و بررسی

۶.۱. تحلیل نتایج

نتایج تجربی نشان می‌دهد که سیستم پیشنهادی عملکرد قابل‌قبولی در تشخیص و ترجمه همزمان گفتار به زبان اشاره فارسی دارد. دقت ۹۹.۹۴٪ در تشخیص حرکات ایستا بر روی مجموعه داده MPSL نشان‌دهنده بلوغ روش‌های تشخیص حرکات ایستا

است. این دقت بالا به دلیل ماهیت محدود و قابل‌پیش‌بینی حرکات ایستا (حروف و اعداد) و همچنین کیفیت بالای مجموعه داده MPSL با ۱۰۷۶۳۰ تصویر متوازن است (۲۰۲۶).

در تشخیص حرکات پویا، دقت سیستم بر روی مجموعه داده ISLR۱۰۱ به ۸۷.۶۴٪ و بر روی Continuous PSL به ۸۳.۲۱٪ رسیده است. این ارقام اگرچه نسبت به دقت تشخیص حرکات ایستا پایین‌تر است، اما در مقایسه با پژوهش‌های مشابه در سایر زبان‌های اشاره با منابع محدود، عملکرد رقابتی محسوب می‌شود. برای مثال، TransSLR بر روی مجموعه داده CASL-W۱۰ (زبان اشاره آفریقای مرکزی) به دقت ۸۰.۳۹٪ دست یافته است (al et Yen, ۲۰۲۶). در مجموع، روش‌های مبتنی بر GCN در بازه دقت ۵۷.۰۱ تا ۹۹.۳۹٪ و روش‌های مبتنی بر CNN در بازه ۷۰ تا ۹۹.۹٪ عملکرد داشته‌اند (Natarajan, al et, ۲۰۲۵) که سیستم پیشنهادی در محدوده بالای این بازه‌ها قرار دارد.

۶.۲. چالش‌ها و محدودیت‌ها

با وجود عملکرد مطلوب سیستم، چندین چالش و محدودیت قابل‌شناسایی است:

اول، تنوع گویندگان: مدل‌های تشخیص زبان اشاره معمولاً در مواجهه با گویندگان جدید با کاهش عملکرد مواجه می‌شوند. این مسئله به دلیل تنوع در سرعت اجرا، اندازه دست، سبک اشاره و عوامل دیگر است. تحقیقات نشان داده است که مدل‌های از پیش‌آموزش‌دیده بر روی مجموعه داده‌های با منابع بالا، در انتقال به زبان‌های اشاره با منابع محدود با مشکل مواجه می‌شوند؛ به‌طوری که یک مدل پیش‌آموزش‌دیده ASL دقت صفر درصد در تشخیص صفر-شات بر روی CASL-W۱۰ داشته است (al et Yen, ۲۰۲۶).

دوم، شرایط محیطی: عملکرد سیستم در شرایط نوری نامناسب، پس‌زمینه‌های شلوغ و هنگام تشخیص حرکات سریع کاهش می‌یابد. Valencia Cuevas و Martinez Salgado (۲۰۲۵) نیز در سیستم تشخیص زبان اشاره مکزیکی خود با چالش‌های مشابه در شرایط کم‌نور و حرکات سریع مواجه شده‌اند.

سوم، وابستگی به کیفیت استخراج نقاط کلیدی: عملکرد مدل‌های مبتنی بر GCN به شدت به کیفیت و پایداری تخمین وضعیت (estimation pose) وابسته است (۲۰۲۶). خطاهای استخراج نقاط کلیدی می‌توانند به طور مستقیم بر دقت تشخیص تأثیر بگذارند.

چهارم، محدودیت‌های مجموعه داده: مجموعه داده Continuous PSL با ۱۰۷۰۰۰ ویدئو (Rastgoo, ۲۰۲۵) اگرچه بزرگ‌ترین مجموعه داده موجود برای PSL است، اما در مقایسه با مجموعه داده‌های سایر زبان‌ها مانند WLASL با بیش از ۲۰۰۰ کلمه (al et Li, ۲۰۲۰) یا BdSLW۴۰۱ با ۴۰۱ علامت و ۱۰۲۱۷۶ نمونه ویدئویی (۲۰۲۶)، همچنان محدود است.

۶.۳. مقایسه با پژوهش‌های مشابه

جدول ۷ مقایسه سیستم پیشنهادی با سایر سیستم‌های تشخیص و ترجمه زبان اشاره را نشان می‌دهد.

جدول ۷: مقایسه با پژوهش‌های مشابه

سیستم	زبان اشاره	رویکرد اصلی	دقت تشخیص (%)	ترجمه همزمان	بازسازی D3
Salgado & Cuevas (۲۰۲۵)	مکزیک	LSTM + CNN	۹۰ >	خیر	خیر
al et Tagliabue (۲۰۲۵)	ایتالیایی	Avatar + Whisper	-	بله	خیر
Speak2Sign3D (۲۰۲۵)	آمریکایی	+ Whisper MarianMT	-	بله	بله
al et Liu (۲۰۲۵)	آمریکایی	GCN Dual-Stream	۹۳.۷۰ (WLASL-۱۰۰)	خیر	خیر
al et Yen (۲۰۲۶)	آفریقایی مرکزی	TransSLR	۸۰.۳۹	خیر	خیر
DexAvatar (۲۰۲۶)	چندزبانه	Reconstruction ۳D	-	خیر	بله
سیستم پیشنهادی	فارسی	+ GCN Transformer	۹۹.۹۴ (ایستا) / ۸۷.۶ (پویا)	بله	بله

سیستم پیشنهادی از نظر جامعیت (پوشش تشخیص، ترجمه و بازسازی سه‌بعدی) و تخصص (اختصاصی برای زبان اشاره فارسی) نسبت به پژوهش‌های مشابه مزیت دارد.

۷. نتیجه‌گیری و پژوهش‌های آینده

در این مقاله، یک سیستم جامع برای تشخیص و ترجمه همزمان گفتار به زبان اشاره فارسی با استفاده از یادگیری عمیق و بازسازی سه‌بعدی حرکات دست ارائه شد. معماری پیشنهادی شامل پنج مؤلفه اصلی است که به صورت یکپارچه برای تبدیل گفتار فارسی به حرکات سه‌بعدی زبان اشاره عمل می‌کنند. نتایج تجربی بر روی سه مجموعه داده Continuous PSL، MPSL و ISLR۱۰۱ نشان داد که سیستم پیشنهادی به دقت ۹۹.۹۴٪ در تشخیص حرکات ایستا و ۸۷.۶٪ در تشخیص حرکات

پویا دست یافته است. ماژول بازسازی سه‌بعدی نیز با بهبود ۳۸.۷٪ نسبت به روش‌های پایه، کیفیت بالایی در تولید حرکات طبیعی دست ارائه می‌دهد.

مشارکت‌های اصلی این پژوهش عبارتند از: (۱) ارائه یک معماری پنج‌لایه کامل برای ترجمه همزمان گفتار به زبان اشاره فارسی، (۲) طراحی یک مدل ترکیبی نوین مبتنی بر GCN Dual-Stream و Transformer برای تشخیص حرکات پویا، (۳) پیاده‌سازی یک ماژول بازسازی سه‌بعدی با استفاده از DexAvatar و SOKE برای تولید آواتارهای واقع‌گرایانه، و (۴) ارزیابی جامع سیستم بر روی مجموعه داده‌های موجود PSL.

پژوهش‌های آینده می‌توانند در جهت‌های زیر ادامه یابند:

۱. **توسعه مجموعه داده‌های بزرگ‌مقیاس:** با وجود مجموعه داده‌های موجود، نیاز به مجموعه داده‌های بزرگ‌تر و متنوع‌تر برای زبان اشاره فارسی احساس می‌شود. جمع‌آوری داده از گویندگان بیشتر با تنوع سنی، جنسی و منطقه‌ای می‌تواند به بهبود استحکام مدل کمک کند.

۲. **تشخیص دوطرفه:** توسعه ماژول تشخیص زبان اشاره به گفتار برای ایجاد یک سیستم ارتباطی دوطرفه بین جامعه ناشنویان و شنوایان.

۳. **بهینه‌سازی برای دستگاه‌های لبه:** کاهش حجم مدل و زمان تأخیر برای اجرا بر روی دستگاه‌های همراه و کم‌قدرت با استفاده از تکنیک‌های تقطیر دانش (distillation knowledge) و کاهش دقت (quantization). TransSLR نشان داده است که کدکننده‌های ترنسفورمر سبک‌وزن می‌توانند با هزینه محاسباتی کمتر عملکرد قابل‌قبولی داشته باشند (Yen et al., ۲۰۲۶).

۴. **ترکیب با مدل‌های زبانی بزرگ:** استفاده از مدل‌های زبانی بزرگ (LLMs) برای درک بهتر بافت و ترجمه دقیق‌تر جملات پیچیده. پژوهش‌های اخیر نشان داده است که مدل‌های ترنسفورمر مبتنی بر LLM می‌توانند عملکرد ترجمه زبان اشاره را بهبود بخشند (۲۰۲۵).

۵. **تولید آواتارهای شخصی‌سازی‌شده:** توسعه قابلیت شخصی‌سازی آواتارها بر اساس ویژگی‌های ظاهری و سبک اشاره کاربران مختلف.

۶. **پشتیبانی از حالات چهره و حرکات غیردستی:** زبان‌های اشاره از حالات چهره و حرکات ابرو و لب نیز برای انتقال معانی دستوری و عاطفی استفاده می‌کنند. توسعه مدل‌هایی که این ویژگی‌های غیردستی را نیز تشخیص و تولید کنند، می‌تواند به طبیعی‌تر شدن ارتباط کمک کند (Trabelsi, & Harrouch, ۲۰۲۵).

منابع

۱. رقیب‌دوست، ش. و فیاضی، ی. (۱۴۰۱). نفی در زبان اشاره ایرانی (گونه تهرانی). پژوهش‌های زبان‌شناسی، ۱۴(۲)، ۹۱-۱۱۲.
۲. Cuevas Valencia, R. E., & Salgado Martinez, G. A. (۲۰۲۰). Deep learning-based mobile application for the educational inclusion of deaf individuals: Recognition and translation of Mexican sign language. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, 16(۳۱), e۹۳۸. <https://doi.org/10.۲۳۹۱۳/ride.v16i31.۲۵۳۸>
۳. DexAvatar. (۲۰۲۱). DexAvatar: ۳D sign language reconstruction with hand and body pose priors. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
۴. Harrouch, H., & Trabelsi, L. (۲۰۲۰). A deep learning-based method combines manual and non-manual features for sign language recognition. *Scientific Reports*, 16, ۲۹۸۱.
۵. Khan, Z. (۲۰۲۰). Neural machine translation techniques for English text to Pakistan sign language gloss translation. *Universal Access in the Information Society*, 24(۲), ۱۸۱۳. <https://doi.org/10.۱۰۰۷/s10۲۰۹-۰۲۴-۰۱۱۸۵-۰>
۶. Li, D., Rodriguez-Opazo, C., Yu, X., & Li, H. (۲۰۲۰). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, ۱۴۵۹-۱۴۶۹.
۷. Li, X., et al. (۲۰۲۰). Frame-wise uncertainty modeling with variational CTC for continuous sign language recognition. *2025 8th International Conference on Computer Network, Electronic and Automation (ICCNEA)*, ۱-۵.

۸. Liu, L., et al. (۲۰۲۰). Skeleton-based sign language recognition using a dual-stream spatio-temporal dynamic graph convolutional network. *arXiv:2509.08661*.
۹. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (۲۰۱۹). MediaPipe: A framework for building perception pipelines. *arXiv:1906.08172*.
۱۰. Madani, H., & Nahvi, M. (۲۰۱۳). Isolated dynamic Persian sign language recognition based on camshift algorithm and radon transform. *2013 21st Iranian Conference on Electrical Engineering (ICEE)*.
۱۱. MediaPipe Holistic. (۲۰۲۱). MediaPipe holistic landmark extraction for sign language recognition. *Google Research*.
۱۲. Moghbeli Persian Sign Language (MPSL) Database. (۲۰۲۱). MPSL: A balanced, large-scale image database for Persian static hand gesture recognition. *AUT Journal of Electrical Engineering*.
۱۳. Natarajan, B., et al. (۲۰۲۰). A comprehensive survey on recent advances and challenges in sign language recognition systems. *Discover Artificial Intelligence*.
۱۴. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (۲۰۲۳). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, ۲۸۴۹۲-۲۸۵۱۸.
۱۵. Ranjbar, H., & Taheri, A. (۲۰۲۰). ISLR۱۰۱: An Iranian word-level sign language recognition dataset. *IEEE Access*, 13, ۹۶۱۴۷-۹۶۱۵۸.
۱۶. Rastgoo, R. (۲۰۲۰). A Persian continuous sign language dataset. *Journal of Artificial Intelligence and Data Mining*, 13(1), ۹۵-۱۰۵. <https://doi.org/10.22044/jadm.2020.14973.2600>

- IV. Sincan, O. M., & Keleş, H. (۲۰۲۰). AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods. *IEEE Access*, 8, ۱۸۱۳۶۰-۱۸۱۳۶۵.
۱۸. SpeakΓSign³D. (۲۰۲۰). SpeakΓSign³D: A multi-modal pipeline for English speech to American sign language animation. *arXiv:2507.06530*.
۱۹. Tagliabue, N., Colletti, E., Dani, F. R., Tedesco, R., Cenceschi, S., & Trivilini, A. (۲۰۲۰). DEEP: An automatic bidirectional translator leveraging an ASR for translation of Italian sign language. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, ۲۲۱-۲۲۹.
۲۰. Transformer-based sign language recognition with LLM. (۲۰۲۰). Domain based sign language recognition and translation system with multimodal outputs. *2025 International Conference on Computing and Communication Technologies*.
۲۱. TransSLR. (۲۰۲۱). TransSLR: A lightweight transformer for sign language recognition. *arXiv:2608.06407*.
۲۲. Yen, W. L., Johnny, S., Olufemi, V. T., Aaron, E., & Busogi, M. (۲۰۲۱). TransSLR: A lightweight transformer for sign language recognition. *arXiv:2608.06407*.
۲۳. Zuo, R., et al. (۲۰۲۰). Signs as Tokens: A retrieval-enhanced multilingual sign language generator. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
۲۴. (۲۰۲۰). Skeleton-based sign language recognition using a dual-stream spatio-temporal dynamic graph convolutional network. *arXiv:2509.08661*.
۲۵. (۲۰۲۰). Hand-aware graph convolution network for skeleton-based sign language recognition. *Journal of AI and Data Mining*, 3(۱), ۳۶-۵۰.

۲۶. (۲۰۲۰). Intelligent system for automatic bidirectional sign language translation based on recognition and synthesis of audiovisual and sign speech. *ISPRS Archives*, 48, ۸۷-۹۴.
۲۷. (۲۰۲۰). Hybrid translation for sign languages: Combining rule-based and neural machine translation in a low-resource scenario. *Language Resources & Evaluation*, 59.
۲۸. (۲۰۲۰). Deep learning-based sign translation of Indian sign language for educational and technical vocabulary. *2025 IEEE International Conference on Signal Processing and Communications*.
۲۹. (۲۰۲۶). Transformer-based word level Bangla sign language recognition using relative quantization encoding. *Scientific Reports*, 16, ۲۰۲۰۹.
۳۰. (۲۰۲۶). A comprehensive review of deep learning approaches for video-based sign language recognition: Datasets, challenges and insights. *MDPI Sensors*.