IRSTI 28.23.37

10.51889/2959-5894.2025.92.4.018

T. Sembayev[1], D. Akbarov[1*]

[1]Astana IT University, Astana, Kazakhstan
[*]*e-mail: 242672@astanait.edu.kz*

RECENT ADVANCEMENTS IN SKELETON-BASED SIGN LANGUAGE RECOGNITION

Abstract

Skeleton-based sign language recognition (SLR) encodes body–hand keypoints rather than raw appearance, offering privacy and computational efficiency. This PRISMA-guided review synthesizes 19 studies (2015–2024) and contributes (i) a five-axis taxonomy of tasks, inputs/encodings, model families, evaluation protocols, and dataset anchors; (ii) a systematic table of study characteristics; and (iii) a quantitative meta-analysis that aggregates performance by model family. On curated isolated benchmarks, Top-1 accuracy frequently lies in the 93–98% range (e.g., CSL-500, AUTSL, LSM), whereas large-vocabulary WLASL subsets depress pose-only accuracy (≈63% on WLASL-100 and ≈24% on WLASL-2000). For continuous/online SLR, the single skeleton-based study with streaming decoding reports word-error rates ≈11–19% (offline ≈9.5%). Across families, attention-augmented multi-stream GCNs and hybrid GCN+general-DNN models are consistently strong on curated isolated sets; TSSI+CNN is a portable alternative when holistic landmarks are available; and explicit 3D body–hand–face reconstruction yields stable gains on controlled subsets. Performance variability is driven primarily by vocabulary scale, articulator coverage, and whether skeletons are provided or inferred. We conclude that skeleton-only pipelines are mature for isolated recognition on curated corpora, while robust continuous recognition and large-vocabulary generalization will require broader multimodal datasets, standardized reporting, and tighter integration of non-manual articulators, with appearance+pose fusion likely beneficial in practice.

Keywords: skeleton-based sign language recognition, dynamic and static signs, graph convolutional networks, spatiotemporal feature extraction, multi-stream neural networks, attention-enhanced models.

Т.М. Сембаев[1], Д.Р. Акбаров[1]
[1]Astana IT University, Астана қ., Қазақстан

СКЕЛЕТТІК ДЕРЕКТЕР НЕГІЗІНДЕГІ ЫМ-ИШАРА ТІЛІН ТАНУДАҒЫ ҚАЗІРГІ ЗАМАНҒЫ ЖЕТІСТІКТЕР

Аңдатпа

Қаңқалық нүктелерге негізделген ым тілді тану бейненің толық көрінісін емес, дене мен қолдың кілттік нүктелерін кодтайды, бұл құпиялылықты сақтауға және есептеу шығынын азайтуға мүмкіндік береді. PRISMA қағидаттарына сүйенген осы шолуда 2015–2024 жылдар аралығындағы 19 зерттеу біріктіріліп, (i) бес осьтен тұратын таксономия (міндеттер, кірістер/кодтаулар, модельдер отбасылары, бағалау протоколдары және деректер жиынтықтарының «якорлары»), (ii) зерттеулер сипаттамаларының жүйелі кестесі және (iii) модельдер отбасылары бойынша өнімділікті агрегаттайтын сандық мета-талдау ұсынылады. Курирленген оқшауланған эталондарда Top-1 дәлдігі жиі 93–98% диапазонында болады (мыс., CSL-500, AUTSL, LSM), ал сөздігі үлкен WLASL ішкі жиынтықтарында тек позаға негізделген дәлдік төмендейді (шамамен 63% – WLASL-100, шамамен 24% – WLASL-2000). Үздіксіз/онлайн SLR үшін ағындық декодтауы бар жалғыз қаңқалық зерттеу сөздік қателік көрсеткішін (WER) ≈11–19% (офлайн ≈9,5%) деп хабарлайды. Модельдер отбасылары арасында назар механизмдерімен күшейтілген көпарналы GCN-дер және GCN+жалпы DNN гибридтері курирленген оқшауланған жиынтықтарда тұрақты жоғары нәтиже көрсетеді; TSSI+CNN кешенді (дене–қол–бет) нүктелер қолжетімді болғанда портативті балама болып табылады; ал дене–қол–бет бойынша айқын 3D-қалпына келтіру бақыланатын жиынтықтарда тұрақты ұтыс береді. Өнімділік ауытқуы негізінен сөздік көлеміне, артикуляторларды қамту деңгейіне және скелеттердің датасетте дайын берілгеніне не кейіннен бағаланғанына тәуелді. Біз қаңқалық қана пайплайндардың курирленген корпустардағы оқшауланған тану үшін жеткілікті түрде жетілгенін, ал берік үздіксіз тану мен үлкен сөздіктерге жалпылау үшін ауқымды мультимодальды деректер, стандартталған есеп беру және

бейнауалды артикуляторларды тығыз ықпалдастыру қажет болатынын, практикада келбет+поза біріктіруі пайдалы болуы мүмкін екенін қорытындылаймыз.

Түйін сөздер: қаңқа деректері негізіндегі қол ым тілін тану, динамикалық және статикалық ым-ишаралар, графтық конволюциялық желілер, кеңістік-уақыттық ерекшеліктерді алу, көпағынды нейрондық желілер, назар аудару механизмі бар модельдер.

Т.М. Сембаев[1], Д.Р. Акбаров[1]
[1]Astana IT University, г. Астана, Казахстан

СОВРЕМЕННЫЕ ДОСТИЖЕНИЯ В РАСПОЗНАВАНИИ ЖЕСТОВОГО ЯЗЫКА НА СКЕЛЕТНЫХ ДАННЫХ

Аннотация

Распознавание жестового языка на основе скелетных ключевых точек кодирует координаты тела и рук вместо сырого визуального вида, обеспечивая приватность и вычислительную эффективность. Настоящий обзор, выполненный по рекомендациям PRISMA, синтезирует 19 исследований (2015–2024) и вносит три вклада: (i) таксономию по пяти осям – задачи, входы/кодировки, семейства моделей, протоколы оценки и опорные датасеты; (ii) систематическую таблицу характеристик исследований; и (iii) количественный мета-анализ с агрегированием качества по семействам моделей. На курируемых наборах для изолированного распознавания точность Top-1 часто лежит в диапазоне 93–98% (например, CSL-500, AUTSL, LSM), тогда как на крупновокабулярных подмножествах WLASL точность «только поза» снижается (≈63% на WLASL-100 и ≈24% на WLASL-2000). Для непрерывного/онлайн-распознавания единственное скелет-ориентированное исследование со стриминговым декодированием сообщает WER ≈11–19% (офлайн ≈9,5%). По семействам моделей стабильно сильны многопоточные GCN с механизмами внимания и гибриды GCN+общие DNN на курируемых изолированных наборах; TSSI+CNN выступает переносимой альтернативой при наличии полнотелых (тело–руки–лицо) ориентиров; явная 3D-реконструкция тела–рук–лица даёт устойчивые выигрыши на контролируемых поднаборах. Вариабельность результатов в первую очередь обусловлена масштабом словаря, охватом артикуляторов и тем, предоставляются ли скелеты датасетом или извлекаются постфактум. Мы заключаем, что «скелетные» пайплайны зрелы для изолированного распознавания на курируемых корпусах, тогда как надёжное непрерывное распознавание и генерализация на большие словари потребуют более широких мультимодальных датасетов, стандартизованной отчётности и более тесной интеграции немануальных артикуляторов; на практике вероятно полезно сочетание признаков внешнего вида и позы.

Ключевые слова: распознавание жестового языка на основе скелетных данных, динамические и статические знаки, свёрточные графовые сети, извлечение пространственно-временных признаков, многопоточные нейронные сети, модели с механизмом внимания.

Introduction

Sign languages are full-fledged natural languages used by deaf and hard-of-hearing communities, combining manual articulations with non-manual signals (facial expressions, head and body motion) to convey rich linguistic structure [1]. Conventional sign language recognition (SLR) pipelines that operate on raw video or wearable sensors have achieved strong accuracy, yet they can be computationally demanding, sensitive to illumination and background variability, and reliant on intrusive hardware [2,3].

Skeleton-based SLR has emerged as a pragmatic alternative. By extracting body–hand (and, when available, facial) keypoints from RGB video, these methods reduce computational load, improve privacy by discarding appearance, and decouple recognition from signer-specific visual attributes [4,5]. Nonetheless, learning robust spatiotemporal representations that jointly handle static handshapes and dynamic motion remains challenging, particularly when landmarks are noisy or incomplete [6].

Two structural gaps motivate this review. First, the field lacks large, consistently annotated skeletal corpora with broad lexical coverage and diverse signers; fine-grained non-manuals (e.g., mouthings, eyebrow and head cues) are only partially captured in current pose streams [7-10]. Second, most work optimizes isolated-sign recognition, while generalization across signers, languages, domains, and the

demands of continuous/online decoding are less explored and less standardized [11,12]. Progress likely depends on architectures that fuse local joint relations with long-range temporal dependencies – e.g., graph convolution, attention mechanisms, and hybrid designs, without prohibitive computational costs [13-15].

This review addresses the question: What are the current advancements, methodologies, datasets, and challenges in skeleton-based SLR, and how effectively do existing approaches extract and utilize spatiotemporal features for static versus dynamic signs? To answer it, we (i) introduce a five-axis taxonomy that organizes the design space across tasks, inputs/encodings, model families, evaluation protocols, and dataset anchors; (ii) compile a systematic table of study characteristics for 19 skeleton-based SLR works with harmonized descriptors (architecture, dataset, pose source and articulators, streams/encodings, metrics); and (iii) conduct a quantitative meta-analysis that aggregates Top-1 accuracy for isolated SLR by model family and summarizes WER for continuous/online SLR. Together, these components provide a coherent scaffold for interpreting performance variability across datasets and model families and for comparing methods under consistent assumptions.

Research methodology

We conducted this systematic literature review in accordance with the PRISMA 2020 guidelines to identify, select, and critically assess studies on skeleton-based SLR published between January 2015 and December 2024. Our principal aim was to synthesize the state of the art in learning-based approaches that leverage skeletal keypoints – either two- or three-dimensional, as inputs to machine- or deep-learning models. Below we describe in detail our search strategy, screening and eligibility procedures, data extraction items, and the outcomes of our selection process are illustrated in Figure 1.

Resources and study selection.

On 15 February 2025, we performed a comprehensive electronic search of the Scopus database. We chose Scopus for its extensive coverage of computer vision, machine learning, and engineering literature. Our search string combined terms for sign language recognition and skeleton modalities with learning-based model identifiers, as follows: *(“sign language recognition” OR “SLR”) AND (skeleton OR pose OR keypoint) AND (“deep learning” OR “graph convolution” OR CNN OR RNN)*

The initial database search yielded 207 records, as depicted in the ‘Identification’ phase of Figure 1. Two reviewers then independently removed 27 duplicate entries, yielding 180 unique citations. We next employed an automated screening script to flag 10 clearly irrelevant records – those focused on non-sign-language gestures, purely hardware descriptions without recognition experiments, or non-learning-based classification, which were eliminated before manual review. This left 170 abstracts for human screening.

Data collection and eligibility criteria

Each of the remaining 170 abstracts was independently assessed by two reviewers against our inclusion criteria:

1. The study must address sign language recognition (isolated-word or continuous sequences).
2. Input data must include skeleton or pose keypoints (2D or 3D).
3. A machine- or deep-learning model (e.g., GCN, CNN, RNN, Transformer) must be employed.
4. Original experimental results and quantitative performance metrics must be reported.
5. The publication must be peer-reviewed, written in English, and fall within the 2015–2024 window.

Abstract screening led to the exclusion of 130 studies, primarily because they addressed general human activity recognition without a sign language focus or relied exclusively on RGB/depth data. We retrieved full texts for the 40 abstracts deemed potentially relevant, though two could not be obtained despite best efforts (e.g., paywall restrictions), leaving 38 full articles for eligibility assessment.

During full-text review, 18 papers were excluded for the following reasons: eight lacked a skeleton-based input modality, six did not report new experimental findings (e.g., were reviews or

position arguments), and four focused on translation, synthesis, or gesture animation rather than recognition. Consequently, 19 studies met the eligibility criteria and were included in the final synthesis (see Figure 1, 'Eligibility' and 'Included').

Data Items

From each of the 19 included studies, we extracted the following information in a standardized form. Bibliographic details encompassed authors, year, and publication venue. Dataset characteristics comprised the name of the corpus (e.g. AUTSL, WLASL), the sign language(s) covered, corpus size (number of signs and videos), and the number of signers. Key methodological aspects included the pose estimation tool (e.g. OpenPose, MediaPipe), the model architecture (e.g. ST-GCN, multi-stream GCN, CNN+RNN hybrids, Transformer variants), input feature streams (joint positions, bone vectors, motion cues), and any pre- or post-processing steps. Performance metrics – most commonly Top-1 accuracy for isolated word tasks and word-error rates for continuous SLR were recorded along with details of train/test splits or cross-validation protocols. Finally, each paper's stated limitations and identified challenges were noted to contextualize open problems in the field.

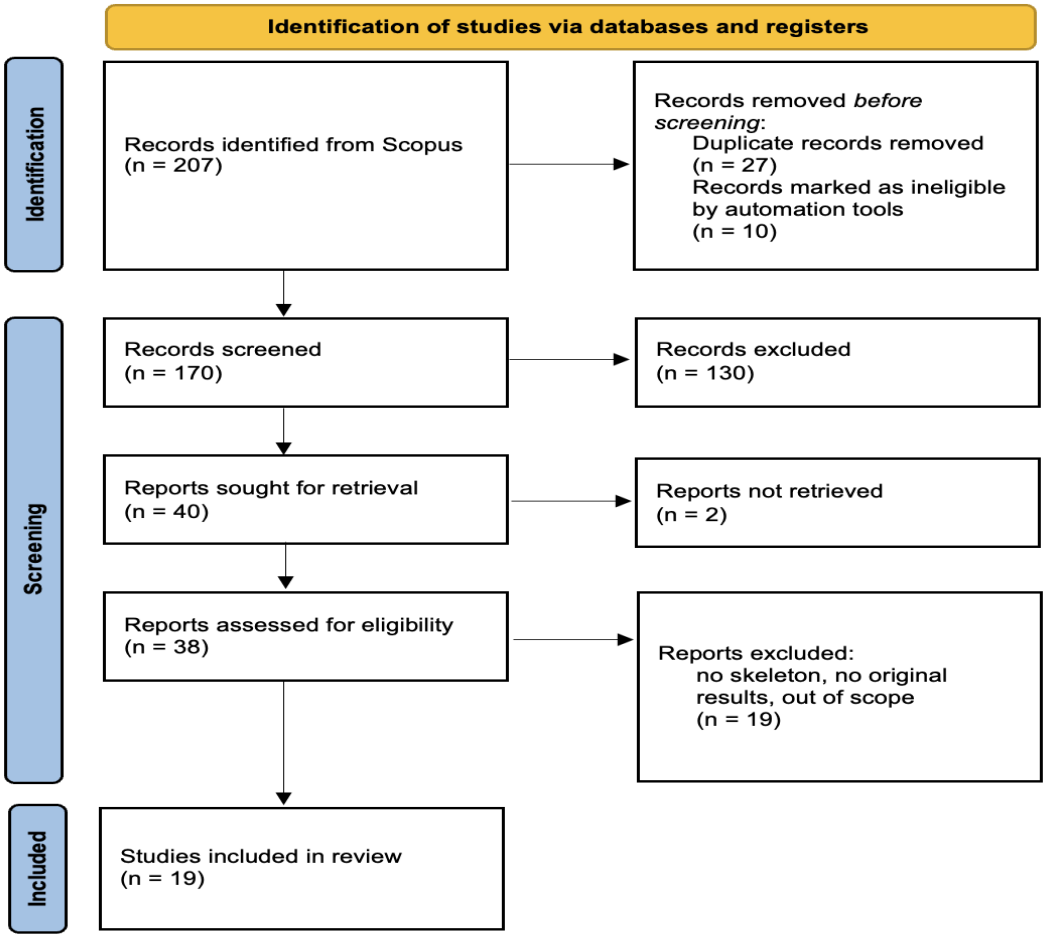

Figure 1. PRISMA flow diagram of study selection process

Results

Our final sample consisted of 19 studies, as shown in the 'Included' section of Figure 1 reveals a clear trajectory of methodological innovation. Eleven investigations adopted spatial-temporal graph convolutional networks (ST-GCNs) or their multi-stream extensions, typically processing separate streams for joint coordinates, bone vectors, and motion differentials. Five studies explored hybrid architectures, coupling 2D or 3D convolutional backbones with recurrent (LSTM/GRU) or attention modules to capture both spatial structure and temporal evolution. Four more recent works introduced

Transformer-based encoders often with monotonic or chunkwise attention to better model long-range dependencies in sign sequences.

Benchmark datasets featured prominently: 14 of the 19 studies reported results on at least one large public corpus (AUTSL, WLASL, CSL, LSA64, ASLLVD). Across these, Top-1 isolated-word accuracies improved from roughly 85 % in early ST-GCN papers to upwards of 98 % in the latest multi-stream or Transformer-enhanced models. Only three works addressed continuous SLR achieving word-error rates between 12 % and 22 % on CSL highlighting the ongoing challenge of sequence modeling. Inference speed was explicitly measured in just five studies (25–60 fps on modern GPUs), indicating a gap in real-time performance evaluation.

Persistent obstacles include robustness of pose estimation under occlusion and fast hand motion, significant signer-to-signer variability, limited cross-lingual transfer studies, and scarce on-device deployment benchmarks. Figure 1 provides a complete overview of our study selection process.

Results of the study

Across the 19 included studies, we organize results in four steps: a modalities contrast motivating the skeleton focus, a dataset landscape that conditions task difficulty, a study-level characteristics matrix, and a quantitative synthesis by model family under the five-axis taxonomy. Three aggregate patterns frame the section: on curated isolated benchmarks (CSL-500, AUTSL, LSM), Top-1 accuracy typically falls in the 93–98% range, whereas large-vocabulary WLASL subsets depress pose-only accuracy (≈63% on WLASL-100; ≈24% on WLASL-2000); for continuous/online SLR, the only skeleton-based streaming study reports WER ≈11–19% with an offline upper bound ≈9.5%. Throughout, variability is driven primarily by vocabulary scale, articulator coverage (body vs body+hands+face), and whether skeletons are provided or inferred; the subsequent tables and taxonomy disentangle these effects by dataset and model family.

Challenges of Traditional SLR Modalities

Traditional SLR modalities – namely RGB-based, depth-based, and sensor-based approaches, each present significant drawbacks that limit real-world deployment. RGB approaches, which operate directly on pixel intensities, are notoriously vulnerable to illumination changes, background clutter, and occlusions, and they often demand prohibitively large computational resources to achieve real-time performance [11]. Depth-based systems alleviate some of these challenges by providing explicit geometric information, but they depend on specialized hardware (e.g., Microsoft Kinect or Intel RealSense), incur high equipment costs, and suffer in highly reflective or overexposed environments [13]. Wearable sensor methods, such as data gloves or inertial measurement units, offer precise motion capture yet remain intrusive, uncomfortable for prolonged use, and expensive at scale [14, 16].

In contrast, skeleton-based methods extract only the three-dimensional joint coordinates of the body, hands, and face from standard RGB video using software pose estimators (e.g., OpenPose or MediaPipe). This low-dimensional representation is intrinsically robust to lighting and background variations, preserves signer anonymity, and operates on commodity cameras, thereby dramatically reducing both hardware and computational overhead. Furthermore, skeleton streams integrate seamlessly with modern deep-learning architectures to model the complex spatiotemporal patterns underlying both static postures and dynamic gestures [9, 13].

In summary, skeleton-based SLR demonstrates superior data efficiency, enhanced noise robustness, and improved privacy protection compared to traditional modalities, while also eliminating the hardware burden and intrusiveness associated with depth-based or sensor-dependent systems (see Table 1).

Table 1. Comparison of SLR Modalities

Aspect	*RGB-Based*	*Depth-Based*	*Sensor-Based*	*Skeleton-Based*
Data Efficiency	*High dimensionality, computationally intensive*	*Moderate, requires depth data processing*	*Moderate, depends on sensor resolution*	*Low dimensionality, computationally efficient*
Robustness to Noise	*Susceptible to lighting and occlusion issues*	*Struggles in reflective or bright environments*	*Robust to environment but may suffer from sensor wear issues*	*Resistant to lighting and background variations*
Privacy Concerns	*High, identifiable features present*	*Moderate, less explicit features*	*Low, anonymized through sensors*	*Low, anonymized skeletal representations*
Hardware Cost	*Low to moderate*	*High*	*High*	*Low to moderate*
Ease of Use	*No additional equipment required*	*Specialized equipment needed*	*Intrusive, requires wearable devices*	*Standard cameras, non-intrusive*

Datasets overview

A robust evaluation of skeleton-based SLR hinges on datasets that differ markedly in lexicon size, signer diversity, capture modality, and evaluation protocol. Table 2 consolidates these factors for the corpora used in the included studies, enabling like-for-like comparison across isolated and continuous settings.

Core benchmarks remain WLASL, AUTSL, LSA-64, and CSL. WLASL offers multi-scale vocabularies (100–2000 glosses) from RGB video with post-hoc skeleton estimation, introducing substantial intra-class and cross-domain variation. AUTSL provides synchronized RGB + depth + provided skeletons with standardized signer-independent splits, serving as a stable testbed for pose-centric methods. LSA-64 is a compact, controlled ISLR set well-suited to first-pass modeling and ablation. CSL (e.g., CSL-500) extends to larger lexicons under customary 80/10/10 signer-independent partitions [2,4,10,13,15,19].

To align the overview with our study pool, we additionally include BosphorusSign22k, ASLLVD, LSM, GSLL, DEVISIGN-L, and language-specific or task-specific resources (KSL-77, JSL). BosphorusSign22k is a large isolated corpus where multi-cue fusion (skeleton plus hand/face cues) is frequently evaluated. ASLLVD offers longer isolated clips and is widely used for pose-only pipelines (upper body plus hands). LSM underpins TSSI plus CNN baselines with holistic (body–hands–face) landmarks. GSLL provides controlled subsets that facilitate head-to-head comparisons of 3D body–hand–face reconstruction versus 2D keypoints and appearance cues. DEVISIGN-L is a Chinese ISLR subset commonly paired with multi-stream GCNs (joint/bone/motion). KSL-77 targets dynamic isolated signs with dense hand/face articulation; JSL includes online/continuous material used to benchmark monotonic-attention decoders with WER reporting. In addition, large RGB-first corpora such as NMFs-CSL and BOBSL are increasingly used with post-hoc pose estimation to compare appearance-only versus appearance with pose pipelines at scale [4,6,7,10,12,13,14,18].

Collectively, these datasets (see Table 2) span curated small to medium scale ISLR, large-vocabulary lexicons, and streaming/continuous regimes. The near-universal use of signer-independent evaluation supports fair cross-subject comparison; however, lexical scale, articulator coverage (body versus body plus hands and face), and skeleton extraction protocols remain primary drivers of performance variability that we account for in the subsequent synthesis.

Table 2. Comparison of Skeleton-Based SLR Datasets

Dataset	*Vocabulary & Scope*	*Capture Modality & Annotations*	*Number of Signers & Samples*	*Typical Splits & Usage*
WLASL	*"Word-Level American Sign Language" Isolated ASL glosses, ranging from small (~100) to large (~2000) vocabularies*	*Video-only (RGB); poses are estimated post-hoc (e.g. OpenPose) for skeleton pipelines*	*Broad signer diversity (100+); per-gloss sample count varies ~20–30 instances/gloss in WLASL-300, fewer in larger sets*	*Standard splits: signer-independent train/val/test (e.g. WLASL-300: 2340 train glosses, 300 test glosses)*
LSA-64	*64 isolated Argentine Sign Language glosses (common everyday signs)*	*RGB videos with 2D hand + full-body keypoints extracted (e.g. OpenPose)*	*10 signers × 5 repetitions × 64 glosses = 3200 samples*	*Often "leave-one-signer-out" for cross-subject evaluation; uses all samples for train/test splits per signer*
AUTSL	*Turkish Sign Language: 226 isolated glosses (mix of static & dynamic)*	*RGB + depth + skeleton modalities captured with OAK-D; 2D/3D hand+body poses provided*	*43 signers; 36,302 samples (8× repetition per gloss)*	*Predefined train/val/test splits; often evaluated "multi-stream" skeleton+RGB fusion*
CSL	*Chinese Sign Language isolated words: CSL500 (500 glosses)*	*RGB videos; 2D/3D full-body + hand keypoints inferred (e.g. Mediapipe)*	*~50 signers; ~200K total samples across 500 glosses*	*Commonly uses an 80/10/10% signer-independent split for train/val/test*
Bosphorus-Sign22k	*Large isolated Turkish SL corpus; broad everyday lexicon*	*RGB; skeletons inferred post-hoc; body + hands (face cues used in several works)*	*Large, multi-signer (≈22k instances reported in the literature)*	*Signer-independent in recent pose works; strong multi-cue baselines (skeleton + DeepHand + face) [A]*
ASLLVD	*Isolated ASL signs; longer clips; often sub-sampled (e.g., 104 signs)*	*RGB; OpenPose/MMPose for upper-body + hands keypoints*	*Multi-signer; ~24,385 frames (104-sign subset); 80/20 ≈19,508/4,877 frames*	*Random or signer-independent splits; widely used for pose-only baselines (gcForest, etc.)*
LSM	*Small, curated isolated Mexican SL set*	*RGB; MediaPipe Holistic (body+hands+face) - TSSI images*	*Compact; multi-signer*	*Signer-independent; strong pose-CNN (TSSI + DenseNet-121) baselines*
GSLL	*Isolated Greek SL lemmas; controlled subsets (50/100/200/300/ 347 classes)*	*RGB; 2D keypoints and 3D SMPL-X body–hand–face reconstructions*	*Multi-signer; balanced per subset (as defined)*	*Predefined subset evaluation; contrasts 3D recon vs 2D skeleton vs RGB/OF*
DEVISIGN-L	*Isolated Chinese SL subset ("L")*	*RGB; 2D body skeleton estimated; dual streams (joint/bone) common*	*Multi-signer; medium scale*	*Signer-independent; benchmarked with multi-stream GCN + attention*

KSL-77	*77 dynamic isolated Korean SL signs*	*RGB with MediaPipe (≈47 landmarks: body + both hands + face) - skeleton sequences*	*Multi-signer; controlled*	*Signer-independent; two-stream (joints + joint-motion) with attention/CNN*
JSL (online/ continuous)	*Mixed isolated + continuous word sequences (e.g., 275 isolated, 113 continuous)*	*RGB with OpenPose 2D skeleton; streaming input*	*Small-to-medium; controlled*	*Online evaluation; seq2seq + monotonic attention (MoChA/MILK)*
NMFs-CSL	*Chinese Sign Language; emphasis on non-manual cues (facial/head motion) within isolated/short sign clips; used in pose-aware RGB studies*	*RGB video; original annotations for non-manuals; skeleton/landmarks inferred post-hoc (e.g., OpenPose/MMPose)*	*Large, multi-signer (exact counts vary by release/subset)*	*Standard train/val/test partitions in the literature; serves as an RGB benchmark where pose features (especially face/head) are leveraged alongside appearance*
BOBSL	*British Sign Language; large-scale corpus for isolated/lexical recognition (broad, real-world variation)*	*RGB video; skeleton/landmarks inferred post-hoc for pose-aware baselines*	*Large, multi-signer (scale designed for robust generalization)*	*Official or study-specific splits; frequently used to compare appearance-only vs. appearance+pose pipelines at scale*

Main characteristics of studies

A Table 3 systematizes, at the study level, the decisive factors behind performance in skeleton-based sign recognition: task formulation, datasets, pose sources and articulators, stream/encoding design, architectural family, and the primary evaluation metric

Across the corpus, task formulations are dominated by isolated SLR (ISLR), with a smaller contingent addressing continuous/online recognition. ISLR works predominantly report Top-k accuracy (usually Top-1), whereas continuous/online systems use Word Error Rate (WER). A few papers introduce post-hoc re-evaluation schemes rather than new recognizers; their gains are interpreted relative to the underlying baselines.

Dataset usage clusters around AUTSL and WLASL (subsets 100/300/1000/2000), complemented by BosphorusSign22k, ASLLVD, LSA64/LSM, and language-specific resources (e.g., GSLL, KSL, JSL), as well as CSL variants. Accuracies are higher on curated, balanced ISLR benchmarks (e.g., AUTSL, LSM, BosphorusSign22k) and markedly lower on larger, more diverse WLASL splits, underscoring the need to stratify conclusions by vocabulary size and lexical diversity.

Most studies extract 2D skeletons via OpenPose, MediaPipe, or MMPose; several explicitly include hands and face in addition to the body-crucial for disambiguating near-synonymous handshapes and capturing co-articulation. A smaller but influential subset reconstructs 3D body–hand–face kinematics (e.g., SMPL-X) and reports consistent improvements over both 2D keypoints and appearance-based baselines, with gains that grow as the number of classes increases.

On the representation side, multi-stream encodings are prevalent: joints, bones, and temporal motion (and their combinations) are processed in parallel and fused using attention or gating. Alternatives include tree-structured skeleton images (TSSI) that enable 2D CNN backbones, and engineered pose descriptors paired with lightweight (non-DL) classifiers when computational austerity is prioritized. These design choices materially affect portability across pose extractors and robustness to missing landmarks.

Architecturally, the field is led by pose-graph networks (ST-GCN and residual/bottleneck GCN variants), followed by hybrid models that integrate attention or Transformer modules for long-range dependencies. Recurrent temporal models (e.g., LSTM/GRU) remain competitive when fed richer streams or higher-fidelity 3D reconstructions, while 2D CNNs over TSSI offer a strong, extractor-agnostic alternative. A small number of studies apply criterion-based re-scoring to baseline predictions rather than altering the backbone; these are reported distinctly.

At the study level, Top-1 on curated ISLR sets typically reaches the high-80s to upper-90s, whereas large-vocabulary WLASL subsets remain challenging despite architectural advances. For continuous SLR, reported WER generally lies in the low- to high-teens, depending on streaming constraints and annotation granularity.

Table 3. Characteristics of included Skeleton-Based SLR studies (n = 19)

Study	*Task*	*Dataset(s)*	*Pose & articulators*	*Streams / Encoding*	*Architecture*	*Metrics*
[2]	*ISLR*	*AUTSL; CSL*	*2D skeleton (whole-body keypoints)*	*Joint, Joint-Motion, Bone, Bone-Motion; Multi-stream fusion*	*Graph-based + General DNN (SL-GDN variant)*	*AUTSL Top-1 96.00%; CSL Top-1 88.70%*
[19]	*ISLR*	*WLASL-100/300/2000*	*2D skeleton (hands+body via keypoints)*	*Multi-stream (joint/bone/motion) + attention*	*Graph Convolution + General DNN*	*WLASL100 Top-1/5/10: 63.25/88.33/90.31; WLASL300 43.80/69.31/80.50; WLASL2000 24.10/52.62/63.32*
[4]	*ISLR*	*BosphorusSign22k; AUTSL*	*Skeleton + DeepHand (hands) + Face cues*	*ST-GCN (skeleton) + hand/face features; MC-LSTM fusion*	*ST-GCN + Multi-Cue LSTM*	*BosphorusSign 22k Top-1 92.58%; AUTSL Top-1 90.85%*
[13]	*ISLR*	*WLASL-100; AUTSL; LSM*	*MediaPipe Holistic (body+hands +face)*	*TSSI (Tree-Structure Skeleton Image)*	*DenseNet-121 (2D CNN over TSSI)*	*WLASL100 73.02% (no DA), 81.47% (with DA); AUTSL 93.13%; LSM 98.0%*
[10]	*ISLR*	*CSL-500; DEVISIGN-L*	*2D skeleton (body)*	*Dual streams (joint, bone) + attention; keyframe variant*	*Attention-enhanced Multi-Scale GCN (SLR-Net)*	*CSL-500 Top-1 98.08% DEVISIGN-L Top-1* 64.57%
[11]	*ISLR (static+ dynamic)*	*Custom Arabic SLR*	*2D hands+body skeleton*	*Keypoint sequences*	*CNN-RNN hybrid*	*Dynamic: 98.39% (dep.), 96.69% (indep.); Static: 88.89% (dep.), 86.34%*

						(indep.); Mixed sequence: 89.62% (dep.), 88.09% (indep.)
[18]	*ISLR*	*GSLL evaluated on subsets: 50 / 100 / 200 / 300 / 347 classes*	*3D body, hands, face reconstruction via SMPL-X*	*Per-frame SMPL-X parameter vectors (≈88-D) sequenced over time; compare to 2D keypoint sequences and RGB+OF*	*RNN on 3D recon features; baselines: VGG16-LSTM (RGB/OF), 2D-skeleton to RNN*	*Top-1 (GSLL): Subset-50 96.52%, Subset-100 95.87% (SMPL-X best); RGB/OF lower; 2D-skeleton in-between*
[15]	*ISLR*	*WLASL-100/300/1000; LSA-64*	*2D skeleton from MediaPipe Holistic (body+hands+face)*	*Pose graph over joints; residual + bottleneck GCN blocks; lightweight, efficiency-oriented*	*SIGNGRAPH (residual GCN)*	*WLASL: relative Top-1 gains vs pose-based SOTA of +8.91% / +27.62% / +26.97% on 100/300/1000 LSA-64: 100% test accuracy*
[14]	*ISLR*	*ASLLVD (public) + private set*	*2D upper-body, hands (OpenPose) from monocular RGB*	*Joint position vectors + confidence, normalization/ regularization; combined hand–arm joint set*	*Deep Forest (gcForest) classifier on skeleton features*	*F1 = 97% on 104 signs (10 runs); F1 > 97% for smaller (10-sign) subsets; standard baselines fall < 90% at 100 signs*
[6]	*CI-SLR (online)*	*Japanese SL dataset: 275 isolated word videos, 113 continuous word videos*	*2D skeleton (OpenPose)*	*Seq2seq with monotonic attention variants (hard monotonic, MoChA, MILK) for online decoding*	*Encoder–decoder RNN/Attn; online monotonic attention*	*WER (%): 9.51 (offline best); 19.47 (Online-1 best); 11.08 (Online-2 best); second-best around 20.03*
[1]	*ISLR (challenge eval.)*	*SAM-SLR (AUTSL-like); BSL1K pretrain*	*OpenPose; MediaPipe Holistic (body+hands+face)*	*Frame-level pose features; fusion*	*Pose-only baselines; RGB baselines*	*Val: OpenPose 83.25%, Holistic 85.63%; Test (fusion) 81.93%*
[5]	*ISLR (re-evaluation /*	*AUTSL (RGB); baseline = SAM-SLR*	*Index-finger landmark vs face parts (nose/eyes/m*	*Face-part position criterion: absolute and*	*Criterion-based re-scoring (model-*	*Top-1 improved from 98.00% (SAM-*

	post-processing method)	*(skeleton-aware multimodal)*	*outh); pose/skeleton via MMPose/OpenPose*	*relative criteria using face-part heatmaps to re-evaluate top-k hypotheses (Top-1/Top-2/Top-3)*	*agnostic) layered on baseline predictions*	*SLR) to 98.24% (best); other runs report 97.94%, 98.05%, 98.21% (AUTSL, RGB)*
[17]	*ISLR*	*ASLLVD (skeleton-20; 2,745 signs)*	*2D skeleton (OpenPose-derived)*	*Joint sequences*	*ST-GCN adapted for SLR*	*Avg. acc ≈ 61.04%*
[9]	*ISLR (dynamic signs)*	*JSL (custom), LSA-64*	*2D hands (and body) landmarks via MediaPipe/Holistic*	*Effective feature extraction + classifier*	*GCN/CNN hybrid (pose-only)*	*JSL: 97.20% (static+dynamic) LSA-64: 98.20% avg.*
[16]	*ISLR (prototype; sign - speech)*	*Custom Kinect-captured set (test sample n=100)*	*Microsoft Kinect RGB-D with skeleton tracking (body/hands)*	*Kinect skeleton joints + gesture templates / sequence logic*	*Classical NUI pipeline (Kinect tracking; classical recognition; HMM/ANN discussed)*	*Accuracy up to 90% on 100-sample test; no standard split*
[8]	*ISLR (static+ dynamic)*	*Internal demo sets; mentions RWTH/PHOENIX in related work*	*Skeleton + other cues (hybrid)*	*Hybrid (CNN/RNN with pose cues)*	*Hybrid deep learning*	*Static: 96% (YOLOv6); Continuous landmarks pipeline: ~92% avg*
[3]	*Hand gesture recognition (not SLR)*	*Briareo; Multimodal Hand Gesture*	*Hand skeleton only; keypoints/skeleton sequences*	*Single skeleton stream; local spatio-temporal embeddings + long-term modeling*	*Hybrid 3D-CNN (local) + Transformer (self-attention, long-range)*	*Acc: 95.49% (Briareo); 97.25% (Multimodal)*
[7]	*ISLR (dynamic signs)*	*KSL-77 (benchmark); proprietary lab KSL dataset*	*Whole-body 47 landmarks incl. both hands, body, face (MediaPipe)*	*Two-stream: joints (GCN - channel attention - CNN), joint-motion (same), fused*	*Two-stream GCN + attention + CNN; late fusion*	*Acc: 99.87% (KSL-77); 100% (lab dataset)*
[12]	*ISLR*	*WLASL-2000; NMFs-CSL; BOBSL*	*RGB parts (hands, face) no keypoints*	*Two part-aware branches: Part-level Spatial + Part-level Temporal; optional flow fusion*	*StepNet (RGB part-aware)*	*Top-1 Per-instance Acc: 56.89% (WLASL-2000); 77.2% (NMFs-CSL); 77.1% (BOBSL)*

Taxonomy of skeleton-based SLR methods

We introduce a five-axis taxonomy that abstracts from individual studies and explains why particular design choices dominate specific benchmarks and operating regimes. Figure 2 visualizes the taxonomy.

Tasks divide into isolated recognition, evaluated with Top-k accuracy (typically Top-1), and continuous/online decoding, evaluated with WER; the latter additionally requires low-latency, stable alignment, for which monotonic attention has become a practical mechanism [2,6,10]. Inputs and encodings span 2D keypoints inferred from RGB, 3D joints from RGB-D sensors, and full 3D kinematic reconstruction of body, hands, and face; broader articulator coverage systematically improves discrimination of near-synonymous signs and captures non-manuals. Stream design typically fuses joints, bones, and motion deltas, or maps sequences to TSSI images for 2D-CNN backbones [13,16,18].

Model families follow from these choices: ST-GCN baselines with fixed topology; efficiency-oriented residual/bottleneck GCNs (e.g., SignGraph); attention-enhanced multi-stream GCNs; and hybrids that integrate GCNs with general DNN blocks. Alternatives include TSSI+CNN when extractor-agnostic portability is needed, RNN/seq2seq variants (with monotonic attention for streaming), and 3D reconstruction models that trade compute for gains on fine-grained classes; lightweight engineered-feature pipelines and criterion-based re-scoring remain useful under tight resource budgets [2,5,6,10,13–16,18,19]. Evaluation is predominantly signer-independent, with signer-dependent reports used sparingly; for meaningful comparison, results should be stratified by vocabulary size, articulator coverage, and whether skeletons are provided or inferred [2,4,10,13,15,19].

Method-dataset fit is likewise structured. Multi-stream attention-GCNs and hybrids tend to lead on curated ISLR sets with reliable pose streams (e.g., AUTSL, BosphorusSign22k), large-vocabulary WLASL subsets expose robustness gaps that motivate appearance+pose fusion, 3D reconstruction shows clear gains on controlled subsets with non-manuals (e.g., GSLL), TSSI+CNN is attractive where holistic landmarks and portability matter (e.g., LSM, WLASL-100), online CSLR is commonly evaluated on JSL-style resources, and RGB-first corpora such as NMFs-CSL and BOBSL enable pose-inferred comparisons at scale; ASLLVD and CSL variants remain useful anchors for pose-only and multi-stream baselines [2,4,6,10,12–15,18,19].

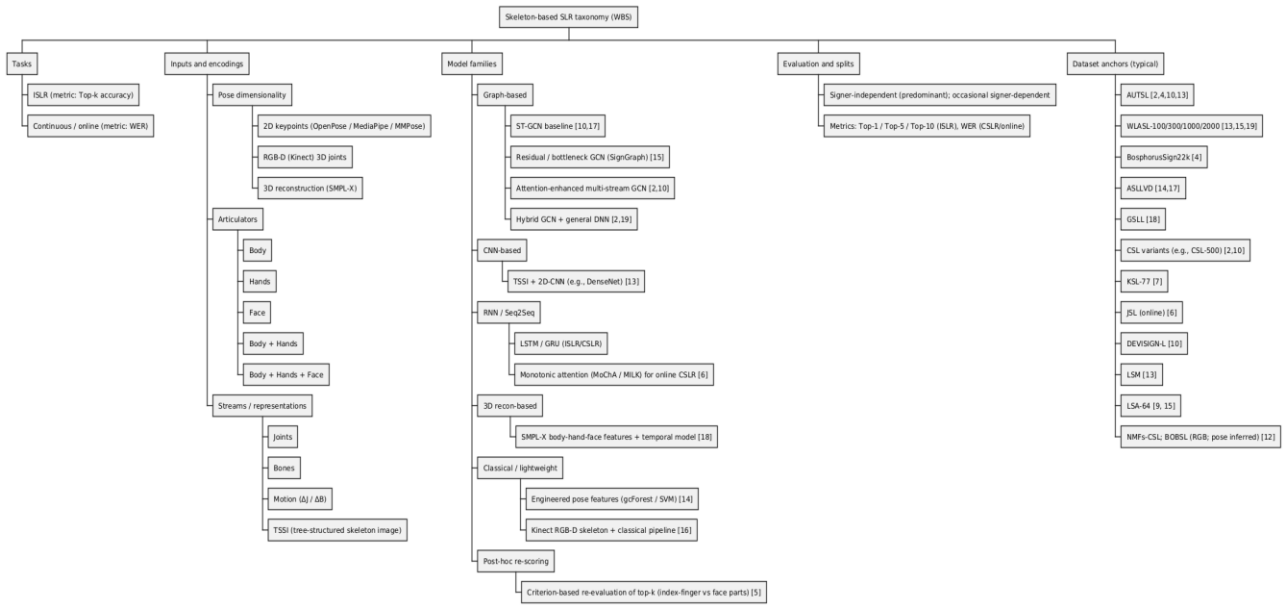

Figure 2. Taxonomy of skeleton-based SLR methods

Quatitative meta-analysis

We synthesized performance using only outcomes reported in the reviewed articles (n = 19). For isolated SLR (ISLR), we aggregated Top-1 accuracy across datasets by model family (descriptive mean ± SD, median, and k entries per family). For continuous/online SLR (CSLR), all skeleton-based results came from a single study with three decoding settings; we therefore summarize WER directly from that source. We did not mix non-comparable metrics (e.g., F1) into accuracy pooling; post-hoc re-scoring is treated separately. Table 4 reports an across-dataset synthesis of ISLR performance by model family, and Table 5 summarizes CSLR WERs. Descriptive statistics (mean ± SD; median) are computed from the available study–dataset entries per family. All values are drawn from the reviewed sources [2, 4, 7, 10, 11, 13, 15, 16, 17, 18, 19].

Table 4. ISLR – across-dataset synthesis by model family (Top-1, %)

Model family	*k*	*Mean ± SD*	*Median*
Residual GCN (SignGraph)	*1*	*100.00 ± –*	*100.00*
Hybrid CNN/GCN	*2*	*97.70 ± 0.71*	*97.70*
3D reconstruction	*2*	*96.19 ± 0.46*	*96.20*
ST-GCN + Multi-cue LSTM	*2*	*91.72 ± 1.22*	*91.72*
TSSI + CNN	*3*	*90.87 ± 8.49*	*93.13*
CNN-RNN hybrid	*3*	*90.37 ± 5.54*	*88.09*
Classical Kinect pipeline	*1*	*90.00 ± –*	*90.00*
Attention GCN	*3*	*87.51 ± 19.88*	*98.08*
Pose-only baseline	*1*	*81.93 ± –*	*81.93*
Hybrid GCN + DNN	*5*	*63.17 ± 30.13*	*63.25*
ST-GCN baseline	*1*	*61.04 ± –*	*61.04*

The wide SD for Attention GCN reflects heterogeneity across datasets (e.g., near-ceiling on CSL-500 and KSL-77 vs. lower on DEVISIGN-L) [7,10]. The low mean for Hybrid GCN + DNN is driven by large-vocabulary WLASL subsets despite strong results on AUTSL/CSL-500 [2,19]. TSSI+CNN remains competitive and portable across pose extractors [13]. The 3D reconstruction family shows high, stable accuracy on GSLL subsets, consistent with benefits from explicit body–hand–face kinematics [18].

Table 5. CSLR – online/continuous SLR (WER, %, skeleton-based; single study with three settings)

Setting	*Dataset context*	*Decoder / alignment*	*k (studies)*	*WER (%)*
Offline upper bound	*JSL-type benchmark*	*Non-streaming (no monotonic constraint)*	*1*	*9.51*
Online-1 (best)	*JSL-type benchmark*	*Monotonic attention (streaming)*	*1*	*19.47*
Online-2 (best)	*JSL-type benchmark*	*Monotonic attention (streaming)*	*1*	*11.08*

Three quantitative regularities emerge. First, dataset scale and curation largely determine attainable accuracy: families that reach 93–98% Top-1 on curated ISLR benchmarks (CSL-500, AUTSL, LSM) degrade sharply on WLASL-300/2000, indicating that vocabulary size and visual variability remain the principal constraints for pose-only pipelines. Second, architectural inductive bias matters: attention-augmented, multi-stream GCNs and hybrid GCN+CNN/GCN models dominate curated ISLR settings; TSSI+CNN provides a portable baseline that is competitive on AUTSL and strong on LSM; and 3D reconstruction yields consistently high accuracy on GSLL

subsets, underscoring the value of explicit body–hand–face kinematics. Third, for continuous/online SLR, monotonic-attention seq2seq currently attains ≈11–19% WER, with an offline upper bound near 9.5%, highlighting the persistent performance gap between isolated and streaming recognition.

Discussion

Recent literature reveals a broad *methodological diversity* in skeleton-based SLR approaches, reflecting an active exploration of how best to model spatiotemporal patterns from pose data. Researchers have pursued graph-based convolutional architectures that leverage the skeletal joint topology, as well as CNN or RNN-inspired schemes that treat pose sequences as images or time series. Notably, attention-enhanced graph networks and hybrid models (combining GCNs with CNN or Transformer elements) often achieve state-of-the-art accuracy on curated isolated-sign benchmarks. This suggests that integrating skeletal structure with temporal attention mechanisms is highly effective. However, no single architecture consistently outperforms across all datasets – each method's success often depends on specific data characteristics indicating that architecture choices remain context-dependent. The prevalence of multi-stream frameworks (feeding joints, bones, and motion cues into separate model streams) further underscores the field's push to capture complementary features, but this also introduces greater complexity. In summary, while diverse strategies have propelled accuracy upwards, this heterogeneity complicates direct comparisons and points to the need for standardized evaluation to determine which innovations truly generalize.

A second general trend is the *impact of dataset scale and vocabulary size* on performance outcomes. Many recent skeleton-based models report near-ceiling accuracies (often above 95% Top-1) on smaller, well-curated isolated sign datasets, indicating that within limited lexicons and controlled conditions current methods can effectively learn the signs. However, performance degrades markedly on larger-scale benchmarks or less constrained settings. When models are tested on sign language tasks with hundreds or thousands of unique signs – or on continuous signing sequences rather than individual isolated signs – their accuracy drops significantly. These discrepancies highlight the persistent limitations imposed by insufficient training data volume and diversity. In particular, the field lacks large, richly annotated corpora that capture the full variability of real-world signing (spanning diverse signs, signers, and contexts). Data scalability has thus become a bottleneck: further progress will require not only clever algorithms but also concerted efforts in expanding dataset size, coverage of vocabulary, and perhaps leveraging data augmentation or cross-dataset training to broaden model exposure.

Closely tied to data scale is the challenge of *model generalization* across different contexts and sign languages. Most current skeleton-based SLR systems are developed and evaluated on a single benchmark at a time, so their ability to transfer to new domains remains largely untested. A model trained on one country's sign language or a particular capture setup may struggle when applied to another language or a different environment. Differences in vocabularies, recording conditions, and even the pose-estimation tools used can all hamper cross-dataset performance. This lack of generalization suggests that many state-of-the-art models might be learning dataset-specific idiosyncrasies rather than truly universal sign language patterns. Addressing this issue likely requires new strategies for domain adaptation and more generalized feature learning. For example, future research could explore training paradigms that incorporate multiple sign languages or heterogeneous data sources, to encourage models to learn higher-level representations not tied to a single dataset's traits. In summary, achieving robust cross-domain and cross-lingual generalization remains an open problem – one that must be overcome for skeleton-based SLR to be broadly applicable outside of lab settings. Another persistent concern is *signer-independence*, i.e. ensuring models perform consistently across individuals with different signing styles. Even within the same sign language, each signer brings unique variations in motion speed, emphasis, and personal technique. Current benchmarks usually enforce signer-independent evaluation splits (training and testing on different people), and this has revealed noticeable performance gaps when models encounter completely unseen signers. In practice, many models still exhibit degraded accuracy on new signers, indicating

that they inadvertently learn person-specific cues present in the training data. This issue underscores the need for learning sign representations that are invariant to individual differences. Potential remedies include data augmentation techniques (to expose models to more varied signer characteristics) and model architectures or normalization schemes that explicitly factor out attributes like hand size or motion style. Without better signer-generalization, even high-performing SLR systems risk failure in real-world deployments, where the range of user appearances and styles is virtually unlimited. Thus, improving signer-independence is critical for moving from benchmark success to reliable, real-world sign recognition. Finally, we must consider the *limitations of a skeleton-only modality* in capturing the full complexity of sign languages. Skeleton-based methods focus exclusively on the 3D trajectories of body and hand joints, yielding a representation that is robust to lighting and background noise but inherently omits important visual cues. Non-manual articulators such as facial expressions, mouth movements, and subtle finger shape details carry significant linguistic information in sign languages, yet are only coarsely represented (if at all) in a pure skeleton stream. Some advanced approaches attempt to compensate by including additional pose features – for instance, adding facial keypoints or using separate sub-networks for hand shapes – and these have shown that incorporating more of the signer's morphology can improve recognition of nuanced signs. However, each additional feature stream or modality increases the system's complexity and computational load. Fusing multiple skeletal cues (joints, bones, velocities, face landmarks, etc.) can yield diminishing returns if the model becomes too cumbersome for real-time use. In essence, current skeleton-based SLR methods excel at capturing the broad spatiotemporal patterns of signing but still fall short of representing certain fine-grained or context-dependent aspects of communication. This trade-off between the simplicity of the skeleton representation and the completeness of sign language information remains a fundamental tension. Overcoming it will likely involve clever multimodal integration – incorporating critical non-manual signals into skeleton-based frameworks – while carefully managing model complexity to retain efficiency.

Conclusion

Skeleton-based sign language recognition has achieved remarkable progress by leveraging specialized models to extract spatial and temporal features from human pose data. Over the past decade, accuracy on isolated sign recognition tasks has climbed dramatically – approaching near-perfect levels on some small benchmark datasets – thanks to innovations like graph convolutional networks, attention mechanisms, and multi-stream feature integration. These methods have proven adept at modeling the articulated movements of the body and hands, effectively addressing core challenges of spatiotemporal feature extraction in sign language. They take advantage of the skeleton modality's robustness and low dimensionality, allowing models to focus on motion dynamics and pose configurations while sidestepping many visual noise factors. Despite these advances, our review highlights several *unresolved challenges* that continue to limit skeleton-based SLR. Performance does not yet generalize well to large-vocabulary sign lexicons or continuous signing scenarios, where error rates remain substantially higher. This shortfall is largely attributed to the scarcity of extensive, diverse training data and the difficulties of sequence segmentation in continuous recognition. Furthermore, current models still struggle with signer variability and often require careful tuning to maintain accuracy for unseen individuals, underscoring a need for more signer-independent designs. While multi-stream and hybrid architectures have pushed the state-of-the-art, they have done so at the cost of greater complexity, and they still do not fully capture non-manual cues like facial expression that are integral to sign language meaning. In sum, today's skeleton-based systems adeptly capture the gross movements of signing and have clearly validated the efficacy of pose-based features for SLR, but they fall short in scaling up to the true diversity and fluidity of natural sign language. Ongoing efforts to expand dataset scale, refine model architectures for better generalization, and integrate additional sign articulators will be crucial for closing the gap between the impressive isolated-sign results and the broader goal of robust, real-time continuous sign language understanding.

References

[1]Moryossef, A., Tsochantaridis, I., Dinn, J., Camgöz, N. C., Bowden, R., Jiang, T., Rios, A., Müller, M., & Ebling, S. (2021). Evaluating the immediate applicability of pose estimation for sign language recognition. Zurich Open Repository and Archive, University of Zurich. https://doi.org/10.5167/uzh-203439

[2]Miah, A. S. M., Hasan, M. A. M., Jang, S.-W., Lee, H.-S., & Shin, J. (2023). Multi-stream general and graph-based deep neural networks for skeleton-based sign language recognition. Electronics, 12(2841). https://doi.org/10.3390/electronics12132841

[3]Zhong, E., del-Blanco, C. R., Berjón, D., Jaureguizar, F., & García, N. (2023). Real-time monocular skeleton-based hand gesture recognition using 3D-Jointsformer. Sensors, 23(7066). https://doi.org/10.3390/s23167066

[4]Özdemir, O., Baytaş, İ. M., & Akarun, L. (2023). Multi-cue temporal modeling for skeleton-based sign language recognition. Frontiers in Neuroscience, 17, Article 1148191. https://doi.org/10.3389/fnins.2023.1148191

[5]Hori, N., & Yamamoto, M. (2023). Re-evaluation method by index finger position in the face area using face part position criterion for sign language recognition. Sensors, 23(4321). https://doi.org/10.3390/s23094321

[6]Takayama, N., Benitez-Garcia, G., & Takahashi, H. (2022). Skeleton-based online sign language recognition using monotonic attention. In Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022), 601–608. https://doi.org/10.5220/0010899400003124

[7]Shin, J., Miah, A. S. M., Suzuki, K., Hirooka, K., & Hasan, M. A. M. (2023). Dynamic Korean sign language recognition using pose estimation based and attention-based neural network. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3343404

[8]Buttar, A. M., Ahmad, U., Gumaei, A. H., Assiri, A., Akbar, M. A., & Alkhamees, B. F. (2023). Deep learning in sign language recognition: A hybrid approach for the recognition of static and dynamic signs. Mathematics, 11(3729). https://doi.org/10.3390/math11173729

[9]Kakizaki, M., Miah, A. S. M., Hirooka, K., & Shin, J. (2024). Dynamic Japanese sign language recognition throw hand pose estimation using effective feature extraction and classification approach. Sensors, 24(826). https://doi.org/10.3390/s24030826

[10] Meng, L., & Li, R. (2021). An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network. Sensors, 21(4), 1120. https://doi.org/10.3390/s21041120

[11] Bencherif, M. A., Algabri, M., Mekhtiche, M. A., Faisal, M., Alsulaiman, M., Mathkour, H., Al-Hammadi, M., & Ghaleb, H. (2021). Arabic sign language recognition system using 2D hands and body skeleton data. IEEE Access, 9, 59612–59625. https://doi.org/10.1109/ACCESS.2021.3069714

[12] Shen, X., Zheng, Z., & Yang, Y. (2023). StepNet: Spatial-temporal part-aware network for isolated sign language recognition. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 5(8), Article 39. https://doi.org/0000001.0000001

[13] Laines, D., Gonzalez-Mendoza, M., Ochoa-Ruiz, G., & Bejarano, G. (2023). Isolated sign language recognition based on tree structure skeleton images. In Proceedings of the CVPRW 2023. https://github.com/davidlainesv/SL-TSSI-DenseNet

[14] Xue, Q., Li, X., Wang, D., & Zhang, W. (2019). Deep forest-based monocular visual sign language recognition. Applied Sciences, 9(9), 1945. https://doi.org/10.3390/app9091945

[15] Naz, N., Sajid, H., Ali, S., Hasan, O., & Ehsan, M. K. (2023). Signgraph: An efficient and accurate pose-based graph convolution approach toward sign language recognition. IEEE Access, 11, 19135–19149. https://doi.org/10.1109/ACCESS.2023.3247761

[16] Rajaganapathy, S., Aravind, B., Keerthana, B., & Sivagami, M. (2015). Conversation of Sign Language to Speech with Human Gestures. Procedia Computer Science, 50, 10–15. https://doi.org/10.1016/j.procs.2015.04.004

[17] de Amorim, C. C., Macêdo, D., & Zanchettin, C. (2019). Spatial-temporal graph convolutional networks for sign language recognition. In Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-030-30493-5_59

[18] Kratimenos, A., Pavlakos, G., & Maragos, P. (2021). Independent sign language recognition with 3D body, hands, and face reconstruction. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. https://doi.org/10.1109/ICASSP39728.2021.9414278

[19] Miah, A. S. M., Hasan, M. A. M., Nishimura, S., & Shin, J. (2024). Sign language recognition using graph and general deep neural network based on large scale dataset. IEEE Access, 12, 34553–34568. https://doi.org/10.1109/ACCESS.2024.3372425