12th Workshop on the Representation and Processing of Sign Languages: Language in Motion

A Video-Based Reverse Dictionary for Sign Language Using Gesture Similarity

Batyrbek Orazumbekov, Daniyal Bayanov, Aruzhan Kaltay, Anara Sandygulova

School of Engineering and Digital Sciences · Nazarbayev University, Astana, Kazakhstan

AI-YM

1. Problem Statement

You see a sign but don't know the word. A text dictionary can't help — there's no word to type. Most sign-language systems are classifiers: they map a video to a fixed label, and they fail when the label is unknown.

- **Low-shot regime** — WLASL provides only 3–4 video instances per gloss, making robust intra-class variation modeling severely constrained
- **Signer variability** — differences in execution speed, body scale, camera perspective, and signing style introduce high intra-class variance
- **No textual supervision** — similarity must be learned purely from visual motion, without gloss annotations or text-video alignment
- **Retrieval vs. classification** — the optimization objective shifts from cross-entropy label prediction to ranking-based metrics (Recall@K, mAP)

CLASSIFICATION
video → label
Fails when the label is unknown

RETRIEVAL (this work)
video → similar videos

2. Contributions

- **Video-to-video retrieval** for sign language — no text supervision required.
- **Pose-based representation** normalized to be invariant to signer scale and signing speed.
- **Metric-learning losses** (SupCon, ArcFace, ProxyNCA) compared head-to-head on the same backbone.
- **Two-Stream ST-GCN vs Transformer** compared in a low-shot regime (3–4 clips/class).
- **Cross-language evaluation** on AUTSL (Turkish SL) — fully zero-shot, no fine-tuning.

3. Datasets

2,000 glosses	11,980 clips	3 - 4 videos	1 - 5s duration

WLASL (American Sign Language) is the training and evaluation corpus. Multiple instances per class enable metric learning; signer variability drives invariance. **AUTSL** 226-class subset (Turkish SL) is held out as a zero-shot transfer test (4 gallery + 1 query per class).

4. Pose Preprocessing

MediaPipe · 33 body + 2×21 hand = 75 joints (225 features/frame).

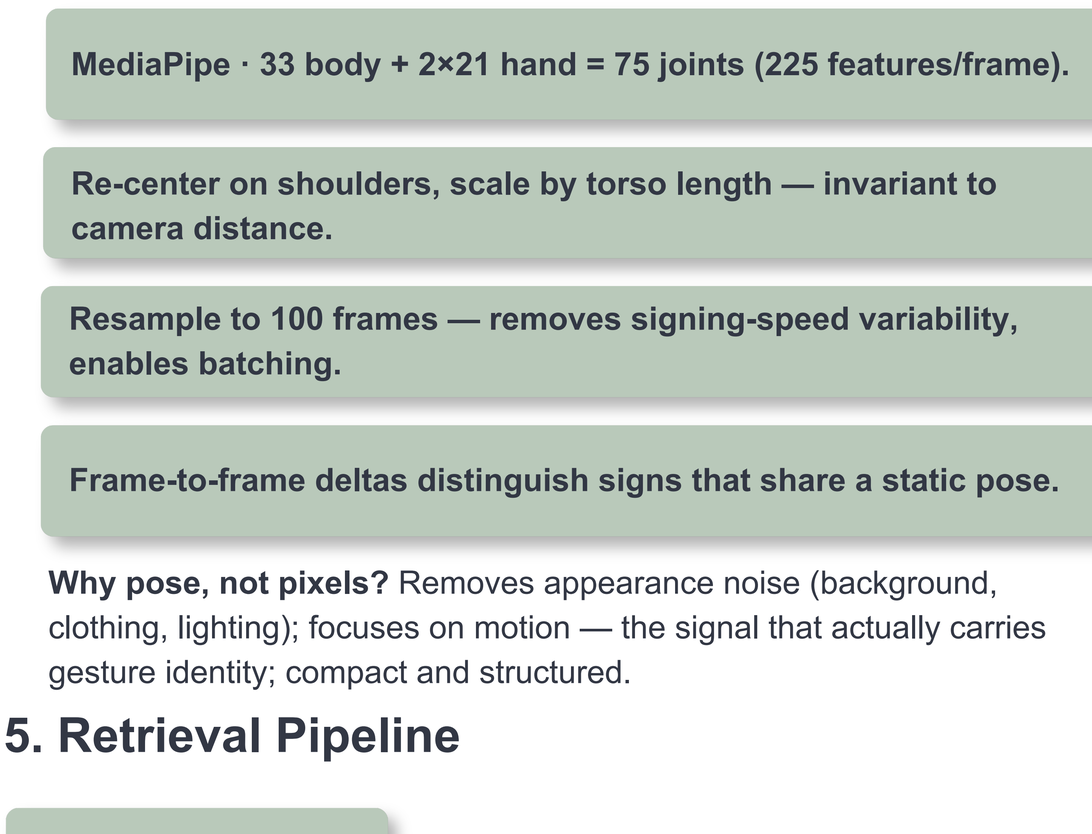
Re-center on shoulders, scale by torso length — invariant to camera distance.

Resample to 100 frames — removes signing-speed variability, enables batching.

Frame-to-frame deltas distinguish signs that share a static pose.

Why pose, not pixels? Removes appearance noise (background, clothing, lighting); focuses on motion — the signal that actually carries gesture identity; compact and structured.

5. Retrieval Pipeline



6. Two Temporal Architectures

TRANSFORMER

Self-attention over frames

- 100 pose vectors + positional encoding
- Multi-head self-attention — global view
- Attention-pooling → fixed-size embedding
- Wins under low-shot conditions
- Best overall: M5 (mAP 0.237)

TWO-STREAM ST-GCN

Graph over the skeleton

- Joints = nodes, bones = edges
- Spatial graph conv. + temporal conv.
- Two-stream: pose + hands fused late
- Strong with abundant data
- Best ST-GCN: M4 (mAP 0.192)

7. Losses & Evaluation

Loss Functions

- **SupCon –** pull same-class together, push others apart.
- **ArcFace** – angular margin between classes.
- **ProxyNCA** – proxy-based baseline; weakest in low-shot.
- **SupCon+ArcFace** – combines clustering with angular margin. (P-K sampling K=3: guarantees positive pairs/batch.)

Evaluation Protocol:
1) Embed every test video
2) Cosine similarity vs gallery
3) Rank gallery by similarity
4) Report Recall@K and mAP. (Retrieval task — ranking, not classification accuracy.)

8. WLASL Validation

Model	Architecture	Loss	R@1	R@5	R@10	R@50	mAP
M1	Transformer	SupCon	0.178	0.455	0.576	0.771	0.212
M2	Two-Stream ST-GCN	ProxyNCA	0.105	0.311	0.424	0.704	0.098
M3	Two-Stream ST-GCN	ArcFace	0.16	0.396	0.51	0.756	0.16
M4	Two-Stream ST-GCN	ArcFace + SupCon	0.177	0.446	0.559	0.769	0.192
M5	Transformer (Attn.)	SupCon	0.183	0.433	0.554	0.732	0.237

9. Cross-Dataset (AUTSL)

R@1: 3.54%	R@10: 29.2%	mAP: 0.091

10. Dicussion and Outlook

Why the Best Model Wins?

- **Global temporal view:** Self-attention compares every pair of frames directly.
- **No rigid skeletal prior:** Flexibility helps under low-shot regimes.
- **Attention pooling:** Weights informative frames over preparation/return.
- **Loss matters as much as model:** SupCon + margin shape retrieval geometry.

11. Future Directions

- **Continuous-sign segmentation:** Extend to sentence-level retrieval via temporal segmentation of continuous signing streams.
- **Self-supervised pre-training:** Pre-train on unlabeled corpora to reduce annotation dependence and improve cross-vocabulary generalization.
- **Re-ranking via DTW Apply Dynamic Time Warping** post-retrieval to compensate for signing speed variability.
- **User study with Deaf learners:** Validate retrieval utility and UX with Deaf and hard-of-hearing participants in practice.

Selected References

[1] Li et al. WLASL: Word-Level Deep Sign Language Recognition. WACV 2020.
[2] Sincan & Keles. AUTSL: Large-Scale Multimodal Turkish SL Dataset. IEEE Access 2020.
[3] Khosla et al. Supervised Contrastive Learning. NeurIPS 2020.
[4] Deng et al. ArcFace: Additive Angular Margin Loss. CVPR 2019.
[5] Vaswani et al. Attention Is All You Need. NeurIPS 2017.
[6] Yan et al. ST-GCN for Skeleton-Based Action Recognition. AAAI 2018.
[7] Duarte et al. Sign Language Video Retrieval with Free-Form Textual Queries. CVPR 2022.
[8] Cheng et al. CiCo: Domain-Aware Sign Language Retrieval. CVPR 2023.

Contact:

batyrbek.orazumbekov@nu.edu.kz
daniyal.bayanov@nu.edu.kz
aruzhan.kaltay@nu.edu.kz
anara.sandygulova@nu.edu.kz

Scan to visit ai-ym.kz