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

by

Batyrbek Orazumbekov

Submitted to the School of Engineering and Digital Sciences, Nazarbayev University
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

March 2026

© Nazarbayev University 2026. All rights reserved.

Author .....
School of Engineering and Digital Sciences, Nazarbayev University
April 11, 2026

Certified by .....
Anara Sandygulova
Associate Professor
Thesis Supervisor

Accepted by .....
Adnan Yazici
Chair, Department of Computer Science

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

by

Batyrbek Orazumbekov

Submitted to the School of Engineering and Digital Sciences, Nazarbayev University
on April 11, 2026, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Most sign language recognition systems have been developed as classification models that associate gesture videos with pre-defined glosses, but such systems do not facilitate similarity search, where users can make queries without knowing the labels of gestures. This thesis proposes a sign language retrieval system based on pose representation that functions as a reverse gesture dictionary, enabling users to retrieve visually similar gestures directly from video input. The proposed method converts gestures into normalized skeletal joints rather than RGB images to minimize variations in appearance, such as background, lighting, and clothing, and to focus on dynamic motion patterns. The extracted keypoints are temporally normalized and optionally augmented with motion features to better capture gesture dynamics. In order to model the temporal relationships within the data, two models are considered; one being a Transformer model with a self-attention mechanism and another one being a Spatial-Temporal Graph Convolutional Network (ST-GCN). Both of these can be used to compare the capabilities of sequence models in modeling temporal dependencies. The model was evaluated using the WLASL dataset under the few-shot setting and ranking metrics like Recall@K and mean Average Precision (mAP), rather than using classification accuracy as it better suits a retrieval task. According to experimental results, it can be concluded that the Transformer model performs better when it comes to modeling temporal relationships between frames in gesture sequences compared to graph-based models. Additionally, employing attention-driven pooling during temporal aggregation improves the results significantly and achieves an mAP of 0.237 on the validation set. Transferability of the embedding space to novel gestures is tested by applying the trained model to the AUTSL dataset (using only a subset of 226 labels). Finally, the impact of approximate nearest neighbor search on retrieval results is examined.

Keywords: sign language recognition, gesture retrieval, pose estimation, Transformer, ST-GCN, few-shot learning, embedding space, approximate nearest neighbor search, WLASL

Thesis Supervisor: Anara Sandygulova
Title: Associate Professor

Contents

Chapter 1

Introduction

1.1 Background and Motivation

Sign languages are natural languages used by Deaf and hard-of-hearing people, employing the coordinated movement of hands, arms, and body to express meaning. Over the last few years, sign language recognition has advanced significantly thanks to recent developments in computer vision and deep learning. Most of the recent work in sign language recognition has focused on the classification problem, where a gesture video is mapped to a predefined label or gloss [20].

Although the classification-based system is effective in translation and transcription, it is not effective in similarity-based search. In some situations, a gesture may be visible, but the meaning or the gloss associated with the gesture is unknown. Under these conditions, it is not possible to use a text-based search. Instead, the user will benefit from a system that takes a gesture video as input and retrieves the visually similar gestures from the database. This system can be regarded as a reverse gesture dictionary, where the search is done based on the similarity of the gestures [4].

Even though there is some research on cross-modal retrieval between video and text [1, 2], visual video-to-video retrieval in sign language is still a relatively unexplored field. Many of the current methods rely on text supervision or gloss annotations [1]. Hence, there is a need for a model that learns motion-based representations, which can be used for similarity search without text supervision. The current thesis is centered

around developing a retrieval system.

1.2 Problem Statement

The problem that the research in this thesis attempts to solve is the design and implementation of a system that retrieves the most visually similar gestures from a set of isolated sign language videos, given a query video containing an isolated sign language gesture. This differs from the classification setting: in classification, the model predicts the label that best represents the input; in retrieval, the model must learn to represent inputs in an embedding space where visually similar gestures are mapped close to one another and dissimilar gestures are placed far apart [8, 12].

The problem, however, becomes more complicated because there are several sources of variability, including the variability between the signers, the variability in the speeds at which the signers perform the gestures, and the variability in the camera perspectives. Moreover, in most cases, there are limited samples for each gesture, which complicates the problem further, especially in low-shot metric learning scenarios [19, 9].

1.3 Research Objectives

The main objective of this work is to develop and assess the effectiveness of the pose-based gesture similarity learning framework for video-to-video retrieval of isolated signs.

To achieve the main research goal, this study focuses on the following three aspects: first, the structured motion representation in terms of skeletal keypoints is constructed in order to eliminate the background noises and highlight the dynamics of the gestures [3]; second, the effectiveness of different architectures for modeling the temporal information in the videos, aiming at transforming the gesture information into fixed-dimensional embeddings, is examined [7]; and third, the metric learning methods are used for the organization of the embedding spaces for the purpose of the retrieval tasks [8, 21, 22].

To assess the effectiveness of the proposed methods, the ranking-based evaluation metrics, i.e., Recall@K and Mean Average Precision (mAP), are used, which are common in retrieval-oriented metric learning [12]. This thesis focuses on the learning of the embeddings for the gestures. While large-scale indexing is not the primary focus of this work, its impact on retrieval efficiency is also explored through additional experiments [16, 17].

1.4 Proposed Approach

In the proposed system, the embedding-based retrieval approach is adopted [8]. Instead of directly utilizing the RGB frames, the proposed approach utilizes the keypoints of the upper body and hands with the help of MediaPipe. The utilization of keypoints reduces the noise associated with the images and aligns with structured gesture modeling approaches [3].

Figure 1-1: Pose, hand, and face keypoint detection using MediaPipe (adapted from [25]).

After the keypoints are extracted, the keypoints are normalized to ensure that the length of the sequence is fixed. In addition to the keypoints, other features like velocity are also considered to enhance the accuracy of the system. The keypoints are then passed to the temporal models to generate the embeddings, following the paradigm of temporal embedding learning [7].

The proposed pipeline employs two distinct models: the transformer model and the spatial-temporal graph convolutional network (ST-GCN). In the transformer model, self-attention is applied to the input keypoints to capture the temporal dependencies. In the ST-GCN model, the skeletal structure is considered as a graph and the evolution of the graph is captured with the help of the convolutional neural network. In our approach, the metric learning approach is adopted to train the models [8, 21]. In the retrieval phase, the similarity is calculated with the help of the cosine similarity function.

1.5 Contributions

This thesis advances the field of gesture similarity learning by presenting a pose-based video-to-video retrieval system for isolated sign language gestures. A comparative analysis of Transformer-based and graph-based temporal models under low-shot learning conditions [19], as well as metric learning strategies for retrieval [8, 12], is also provided. The experimental findings show that Transformer-based temporal modeling is more effective than graph-based modeling for this task, and that attention-based temporal pooling further improves ranking quality. In particular, the best-performing model, a Transformer with attention pooling trained using supervised contrastive learning, achieves an mAP of 0.237 on the WLASL validation set. In addition, cross-dataset evaluation on AUTSL shows that the learned embedding space retains non-trivial retrieval performance on previously unseen sign vocabularies, which supports the potential of visual gesture retrieval without text supervision.

Chapter 2

Literature Review

2.1 The evolution of sign language technology and the emergence of retrieval systems

Sign language technology has progressed from early gesture recognition approaches to sophisticated systems that focus on retrieval and translation. Conventional recognition systems were designed only to classify sign language gestures corresponding to predefined glosses. While the systems were proficient at the recognition task, they were inefficient for other tasks like exploration and finding matching gestures that the user has observed. This motivated the development of reverse dictionaries for sign language systems whereby the user provides a video of the sign language gesture that has to be searched for in a database.

Recent developments in the field have seen an increased trend towards the use of retrieval techniques for sign language technology. Specifically, Duarte et al. [1] built a retrieval mechanism for sign language which made it possible to bridge the gap between sign language and spoken languages through visual-text embeddings. This method enables two-way video search from either sign or spoken language. In addition to that, Cheng et al. [2] proposed an approach for cross-linguistic retrieval using contrastive learning known as Cross-lingual Contrastive Learning (CiCo). The CiCo algorithm showed promising results on the How2Sign and RWTH-PHOENIX-2014T

datasets.

In addition to these approaches, more research has been done into cross-modal alignment between sign language and text representations. Jiang et al. [5] propose the SignCLIP algorithm which utilizes a contrastive learning technique inspired by CLIP. In the process of cross-modal learning, the SignCLIP model learns embeddings for both sign videos and their textual counterparts, aligning semantically matching video-text pairs. As such, it makes it possible to retrieve signs based on text and vice versa. Although the approaches mentioned above are quite capable of handling cross-modal semantic retrieval tasks, they depend on linguistic supervision.

Previous studies on sign language retrieval predate the advent of embedding approaches based on deep learning models. For instance, the Dicta-Sign Wiki [6] proposed a web-based interface for sign language communication and access to sign language material. At the same time, early works on gesture recognition and early works on retrieval systems experimented with the concept of query by example using sensor inputs, such as the Kinect device, and the classic approach of dynamic time warping [10, 11]. The dynamic time warping technique allows for the alignment of sequences of different speeds, and therefore it can be used to compare gestures. Such systems enabled similarity-based comparisons of motion sequences, but they lacked the use of learned representations and depended on template matching and handcrafted features.

In general, previous efforts at tackling the problem of sign language retrieval can be divided into two broad classes: cross-modal retrieval systems relying on textual information, and traditional video-based retrieval systems that operate exclusively on visual features. However, such systems did not employ learned embeddings of videos.

Paralleling these efforts, other studies have progressed the area of gesture-based retrieval independently of linguistic information. Parian-Scherb et al. [3] proposed a pose-based method of gesture retrieval in 2024 based on the use of body keypoints and the attention mechanism. This particular experiment validated the applicability of pose-based retrieval in practical scenarios. Hassan et al. [4] proposed an AI-assisted video-based dictionary for the American Sign Language in 2025. This experiment

emphasized the viewers' preference for video-based querying systems. It also identified the remaining challenges in gesture-similarity search, including the robustness of the ranking function and the signer variation.

Collectively, the above works demonstrate a clear research thread from recognition to similarity-based retrieval. Nonetheless, the state-of-the-art systems are predominantly designed for cross-modal retrieval (text to video) or semantic alignment. The area of video-to-video retrieval solely based on visual features has been less investigated. This would improve accessibility for users preferring gesture-based querying rather than typing.

2.2 Representation Learning and Temporal Modeling of Gestures

The representation of motion and shape is crucial for retrieving accurate gesture similarity. Most current systems use pose-based feature extraction, capturing and analyzing the key movements of the hands, arms, and body across video frames with the help of frameworks such as MediaPipe or OpenPose. Landmark-based representations focus on the movement patterns rather than the background or a signer's appearance, making them less affected by visual noise or differences in color and surroundings.

Furthermore, the effectiveness of such representations is improved by temporal modeling, which accounts for the evolution of the gestures through time. Temporal embeddings, whereby the frames of videos are embedded into a continuous manifold where the motion order is preserved, were proposed by Ramanathan et al. [7]. Temporal embeddings proved crucial in recognizing similar motions with small deviations because the model created video representations of actions sensitive to motion transitions, not static body configurations. Similarly, Duarte et al. [1] used temporal embeddings to match sign language frames with the corresponding glosses, demonstrating that temporal consistency contributes to better results in frame retrieval.

The development of modelling structures such as Transformers [26], TCNs and

LSTMs has greatly helped the field. Those models could extract complex spatial-temporal patterns from gesture sequences and represent long-term dependencies in data. According to Parian-Scherb et al. [3], the use of attention layers that would concentrate on the most informative parts of a gesture could help keep performance levels stable even if there is partial occlusion or an alternative movement pattern. Overall, Parian-Scherb et al.'s research indicates that the temporal patterns of normalized keypoint trajectories offer a clear depiction of motion over time.

2.3 Metric Learning, Domain Robustness, and Re-Ranking Techniques

The last property that will influence the result of the process is how the gesture embeddings are arranged in the metric space. Metric learning offers a systematic method for this problem since it learns how to place visually similar gestures in proximity and different ones further apart from each other in the embedding space. While traditional classifiers depend on category labels, metric learning deals with relative relations in terms of similarity, making it highly relevant for video-to-video gesture retrieval.

2.3.1 Theoretical foundations of metric learning

In metric learning, the concept is an advancement of generalized Mahalanobis distance, where one learns a data-driven metric. In their study, Ghojogh et al. [8] categorize metric learning algorithms as spectral, probabilistic, and deep. All these have the common denominator of minimizing the distance among classes and maximizing the distance between classes. Deep metric learning takes the idea and enhances it by applying neural networks with loss functions like contrastive loss and triplet loss [31]. Contrastive loss places positive pairs inside the margin and pushes negative pairs outside the margin. Triplet margin loss preserves the distance between anchors and positive pairs while pushing anchors and negative pairs away from each other.

Intuitively, contrastive loss treats pairs of samples as either similar or dissimilar and adjusts their distances accordingly, while triplet loss operates on triples of samples — an anchor, a positive (same class), and a negative (different class) — and enforces a ranking among them: the anchor must be closer to the positive than to the negative by a fixed margin. Both losses shape the embedding space so that gesture similarity corresponds to geometric proximity. The idea was taken forward and improved by Cakir et al. [12], who developed Deep Metric Learning to Rank, where AP is optimized within the batch end-to-end without the use of pair-based constraints. Relevance of this particular type of criterion becomes exceptionally significant when considering the area of retrieval because the objective of optimization matches the set of criteria based on which evaluation takes place (Recall@K and mAP). This approach enables making sure that gesture similarity is taken into account during the process of retrieval.

2.3.2 Robustness and generalization

The real-world challenges associated with gesture retrieval are characterized by significant intra-class variation due to variations between signers, viewpoints, and motion speed. On this note, the work on Deep Adversarial Metric Learning [13] presents the concept of adversarial perturbation that encourages the embedding model to preserve separability despite minor distortions and therefore improves robustness to noise. Another instance of effective contrastive learning for generating visual representations can be observed within the paradigm of self-supervised learning, including the momentum contrast method [30]. Similar arguments exist within the field of re-identification studies, where robustness across heterogeneous visual environments is key. According to Zahra et al. [14], part-level embedding features along with domain adaptation and attention normalization improve robustness to variations in viewpoint and lighting conditions. Similar strategies have been considered in addressing the robustness issue within gesture retrieval, where different instances of the same signs might vary significantly depending on their motion speed, body posture, and hand orientation. On the other hand, Musgrave et al. [9] note that, while most recently developed metric-learning architectures promise dramatic accuracy improvements, the

lion's share of these improvements evaporates when experiments are conducted under comparable settings. Their "reality check" study shows that most loss functions achieve very similar performance when their hyperparameters are properly tuned through cross-validation. The finding underlines the importance of evaluation discipline, especially crucial in gesture retrieval, since dataset bias and uncontrolled variation may misleadingly inflate performance. In this respect, future research in gesture-similarity learning should be performed based on cross-validation and fair comparisons between competing loss functions.

2.3.3 Ranking and post-retrieval refinement

Once the embeddings have been created, then it is time for re-ranking to determine the ranking order for the retrieved gestures. According to Ouyang et al. [15], the researchers proposed collaborative image relevance learning, which is one approach to re-ranking after retrieval, involving re-calibration of the retrieved top results based on pair-wise correlation. The same can be done for gesture retrieval, wherein the temporal approach like DTW and frame-wise correlation may be applied to improve the similarity score even further.

2.3.4 Summary

Overall, recent progress in metric learning for gesture retrieval converges along three key dimensions:

- Metric structure learning: based on generalized distance formulations and deep embedding spaces; spectral, probabilistic and deep metric learning frameworks.
- Robust representation learning borrows from adversarial and domain-invariant modeling to enhance stability across various signers, viewpoints, and recording conditions.
- Re-ranking optimization: this focuses on the refinement of retrieval lists for capturing fine-grained temporal and semantic similarities between gestures.

But as Musgrave et al. [9] pointed out in their work, *A Metric Learning Reality Check*, true progress in this area demands the highest degree of methodological rigor: standardized evaluation and comparison which is fair across models.

This means that for gesture retrieval, the systems have to learn, apart from the discriminative metrics, reproducibility, robustness, and consistency with retrieval-specific performance measures like Recall@K and mAP.

2.4 Large-Scale Indexing and Efficient Similarity Search

Large-scale gesture databases must depend on scalable search algorithms. For existing databases that have thousands of videos, the computation required for extensive comparison between database and video files would be impractical. Approximate nearest-neighbor (ANN) search techniques would instead come into play.

Malkov and Yashunin [16] proposed the Hierarchical Navigable Small World (HNSW) graph. This represents an ANN approach through hierarchical construction of the graph. This approach has sublinear complexity in search. HNSW graphs structure the embeddings in multiple levels so that links between the nodes enable the traversal of the graph from the broader to the finer neighborhood quickly. This approach strikes a balance between high recall and efficient search. This technique can essentially meet the requirement of the dense and high-dimensional embeddings of the gestures.

Further scaling can be gained through hybrid indexing techniques. For billion-scale similarity search, Emanuilov and Dimov [17] proposed a hybrid model based on graph search techniques as well as quantization-based techniques. This hybrid model enables the inclusion of efficient filtering techniques that offer metadata-aware filtering functionalities like signer filtering, dataset filtering, and gesture filtering. The inclusion of such techniques will ensure that the efficiency of gesture search systems can be retained even when the database gets increasingly large.

2.5 Sign and Gesture Retrieval: Current Systems and Trends

In the last years, a range of systems has been developed in the context of sign language and gesture retrieval. Cheng et al. [2] showed that cross-lingual contrastive learning successfully allows sign language retrieval frameworks to process multiple sign languages. The works of Duarte et al. [1] and Martins [18] have developed models for an instance level of video retrieval in sign language, providing evidence that video embeddings' quality is more relevant for retrieval performance compared to the overall classification performance of the model.

This has also been demonstrated for gesture retrieval based on keypoints by Parian-Scherb et al. [3], while Hassan et al. [4] highlighted the user advantages of video-controlled dictionary interfaces, which offer greater engagement and usability benefits. In many of the above studies, however, the systems are dependent either on textual or multimodal inputs, while approaches that are based only on video similarity remain limited. Lack of purely visual retrieval methods offers an immense potential for future research. Recent advances like SignCLIP [5] further support this idea through establishing correspondence between sign videos and textual embeddings in a joint representation space, focusing more on semantics rather than visuals.

Combining findings from the above paragraph with the advances in metric learning [13], large-scale indexing [16, 17], and re-ranking [15], we achieve a robust retrieval pipeline that allows efficient and understandable similarity ranking. This bridges the gap between algorithmic development and practical sign language applications because the pipeline performs retrievals based on motion patterns only.

2.6 Research Gaps and Theoretical Positioning

While the current literature gives a good starting point for visual-based retrieval, there are still key gaps:

- Lack of video-to-video retrieval systems: most current approaches focus on text-

to-video or multimodal, rather than purely video-to-video retrieval based on visual similarity.

- Poor normalization for variability in gesture: most of the works in gesture retrieval are not properly normalized to account for differences in speed, scale, or orientation.

- Restricted scalability in sign retrieval: HNSW and hybrid indexing techniques have proven their efficiency in other fields, but applications to databases of gesture or sign videos remain limited.

- Lack of retrieval-oriented evaluation metrics: sign-language research is still reliant on recognition-based metrics and not ranking-based measures, such as Recall@K and mAP.

This review points out the need for a reverse dictionary that allows video-to-video retrieval using gestures. Integration of pose-based representation, metric-learning methods, and efficient ANN search can address challenges in computational complexity and accessibility. Removing dependence on textual input, such a system would offer a more natural, inclusive, and real-time way of exploring and learning any sign language.

Chapter 3

Datasets

3.1 Introduction

The aim of the current study is to build a system that can learn similarity relationships between isolated sign language gestures using a video-to-video gesture retrieval approach. Although the objective of a sign language recognition system is the classification of individual signs, the objective of a gesture retrieval system is the retrieval of gestures based on similarity relationships. This distinction between recognition and retrieval has been emphasized in recent sign language retrieval works [1, 2].

To achieve the objective of a gesture retrieval system, the system must be based on a dataset that contains a number of instances per class. This requirement aligns with principles from deep metric learning, where multiple samples per class are necessary to construct meaningful positive and negative pairs [8, 9]. Based on the above requirements, the Word Level American Sign Language (WLASL) dataset has been selected as the primary dataset to be used in the system [20]. Apart from the primary dataset, the internal dataset of the isolated sign language gestures in the Russian and Kazakh languages provided by the thesis supervisor has been considered.

3.2 WLASL Dataset

3.2.1 Dataset Characteristics

The primary dataset for training and testing the proposed retrieval models is the Word-Level American Sign Language (WLASL) dataset. The WLASL dataset comprises 2,000 isolated glosses, with an average of three to four video instances per isolated sign. The dataset contains recordings from various signers, making it diverse in terms of signing styles, motion speeds, and recording conditions.

After pre-processing and removing unusable samples, there are 11,980 videos in total, spread across various official splits. The number of training, validation, and test samples are 8,313, 2,253, and 1,414, respectively. Each video contains a single isolated sign, lasting from one to five seconds.

The presence of multiple instances for each isolated sign is vital for metric learning, as supervised contrastive and margin-based loss functions need at least two samples per class to compute meaningful positive pairs [8, 12]. Although there are only a few samples per class, this setting is a reasonable low-shot learning scenario, consistent with few-shot metric learning literature [19].

3.2.2 Justification for Dataset Selection

The primary reason for choosing WLASL is that it meets three essential criteria for gesture similarity learning.

Firstly, there are several instances per gloss, which is a pre-requisite for similarity-based training [8]. Secondly, there is variability across different signers, which enables the embedding model to learn to be invariant to different signers—an important property in retrieval systems and person re-identification problems [14]. Thirdly, it is an isolated sign language dataset, which is a direct match to our goal of creating a reverse dictionary for isolated signs, similar to recent video-based ASL dictionary systems [4].

While there are larger datasets with sentence-level information, they pose a funda-

mentally different problem, involving continuous sign recognition and segmentation. Recent works in sign language video retrieval with free-form queries focus on sentence-level retrieval [1, 2], and video moment retrieval frameworks extend this to temporal localization tasks [18]. However, as this work is focused on isolated gesture retrieval, WLASL is a relevant and appropriate dataset for our work.

3.3 Internal Russian and Kazakh Isolated Sign Dataset

The research was given access to an internal dataset of approximately 1,500 isolated sign videos from the research supervisor. The internal dataset contains sign gestures from the Russian and Kazakh languages, captured from various individuals in frontal view from the upper body. The videos vary in length from a minimum of one second to a maximum of five seconds. The videos were recorded in 720p or 1080p resolution.

The key structural characteristic of the internal dataset is that all the words in the dataset occur only once. As a result, the dataset cannot be used to train a metric learning model, which requires the presence of positive pairs to compute the intra-class similarity constraints [8, 9]. Without multiple instances of the same class, it is not possible to define the anchor-positive relationships required by contrastive or triplet-based objectives [12, 13].

Therefore, the internal dataset was not utilized in the experiment. However, the dataset could be used in future work to evaluate the generalization ability of the model across datasets and the zero-shot retrieval performance of the model. Cross-domain generalization and retrieval robustness are active research topics in visual retrieval systems [14, 15]. Once the embedding model has been trained using the WLASL dataset, it could be used to evaluate the model's ability to generalize to the internal dataset.

3.4 Consideration of Sentence-Level Datasets

In considering sentence-level datasets during the selection of datasets, it was noted that this type of dataset would require additional processing steps to segment the video into individual gestures, as they are continuous signing sequences. Sentence-level sign retrieval tasks introduce cross-modal alignment challenges between sign and text modalities [1, 2].

In this work, as the goal is to produce a space for isolated sign retrieval, additional complexities are introduced if sentence-level datasets are considered, moving into a realm of continuous sign recognition and moment retrieval [18]. As such, sentence-level datasets are not considered in this work to keep focus on method and experiment.

Future work will look to expand this system to a hierarchical model to address continuous sign recognition.

3.5 Preprocessing Consistency Across Datasets

The same preprocessing steps are applied to all datasets to ensure representation consistency. Each video is converted into a sequence of pose and hand keypoints, temporally normalized to a fixed frame length, and spatially normalized.

Learning embeddings from temporal sequences aligns with prior work on temporal embedding learning for video representation [7]. The use of keypoint-based representations is also consistent with gesture retrieval research relying on body keypoints and attention mechanisms [3].

The same preprocessing ensures that the embeddings are learned under the same feature space representation, which is essential for stable metric learning behavior [8].

3.6 Summary

To recap, the WLASL dataset was chosen as the primary training dataset based on its ability to fulfill the structural needs of metric learning and isolated gesture retrieval. The internal dataset of the Russian and Kazakh gestures is not appropriate

for supervised training but offers the possibility of a generalization study. The sentence-based datasets were excluded to ensure consistency with isolated sign retrieval. In addition to WLASL, a controlled subset of the AUTSL dataset with 226 gesture classes is prepared for cross-dataset evaluation by selecting four gallery samples and one query sample per class and processing them with the same pose pipeline. Together, these datasets provide a good foundation for the evaluation of gesture similarity learning under low-shot conditions, as studied in the deep metric learning literature [19].

Chapter 4

Methods

4.1 Overview of the Proposed Framework

The objective of the study is to discover an embedding space for gestures where similar isolated signs are placed near each other while dissimilar gestures are far apart based on their visual characteristics. In the proposed work, the representation is based on poses with the incorporation of temporal information and metric learning [8, 9].

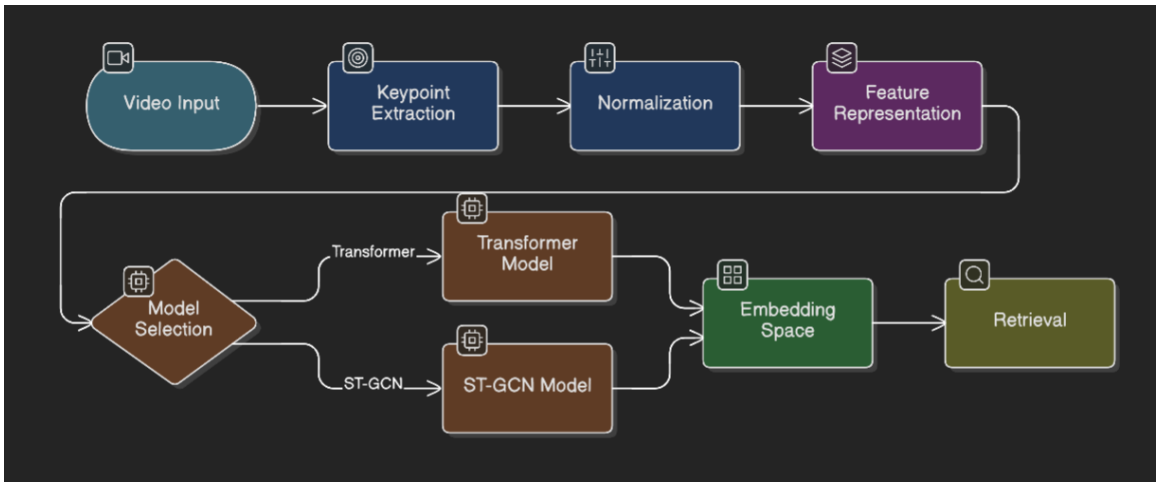

Figure 4-1: Overview of the proposed gesture retrieval pipeline.

The proposed system has several stages: video input, keypoint extraction, preprocessing and normalization, feature building, training the embedding network, and evaluation with the retrieval task. The key difference between the proposed work and

other gesture recognition systems is that our approach is based on optimizing the structure of the embedding space to enable the retrieval task based on ranking [12, 1].

4.2 Pose and Hand Keypoint Extraction

Instead of working with RGB values, the system utilizes the pose-based representation, which helps eliminate background noise and appearance variations. Pose estimation is performed using MediaPipe, which leverages advanced deep learning techniques for robust multi-person pose detection [28]. In the system, 75 keypoints are detected per frame, with 33 keypoints detected for the upper body and 21 keypoints detected for each hand. Each of these keypoints is represented by three-dimensional coordinates, resulting in 225 features per frame. Face keypoints were deliberately excluded from the representation: facial expressions do not carry lexical meaning for the isolated American Sign Language glosses present in WLASL, and including them would introduce unnecessary variation without improving gesture discriminability.

This representation only considers movement and pose, without any consideration of texture and color. The structured skeletal representation helps the system cope with changes in illumination, background, and clothing worn by the signer, which is consistent with gesture retrieval approaches that rely on structured motion representations [3].

4.3 Temporal and Spatial Normalization

Regarding the videos, there are some steps done during preprocessing. The missing landmarks are solved using interpolation from the existing frames. The spatial normalization is achieved by normalizing the coordinates relative to the mid-point of the shoulders and the length of the torso. The temporal normalization of the videos is done by setting the number of frames for all videos to 100. Such a choice was made considering the nature of the data in WLASL: the videos have 25 to 30 fps, a sign lasts 1 to 5 seconds and thus, has around 25 to 150 frames; hence, 100 is a median

number that avoids zero-padding and truncating many frames from the data.

Since the videos are of different lengths, it is important to have the same length so that the videos can be batch-trained and the embeddings can be extracted. Learning embeddings from temporally aligned sequences follows prior work on temporal embedding learning for video analysis [7].

4.4 Feature Representation

Several feature components are built to represent static posture as well as dynamic motion.

The primary feature representation is a set of keypoint coordinates after normalization. To explicitly represent motion, first-order temporal derivatives of the keypoint coordinates, i.e., velocity features, are computed by subtracting consecutive frames. The velocity features help in differentiating gestures with similar static posture but different motion trajectories.

For graph-based experiments, bone vectors are computed by subtracting parent joint coordinates from child joint coordinates based on skeletal structure. Additionally, bone velocities are computed to represent relative joint motion, and this extended feature representation is beneficial for exploiting structural information inherent in the human skeleton using spatial-temporal graph convolutional networks. Such structured modeling aligns with graph-based gesture modeling approaches [3].

4.5 Embedding Architectures

Two types of architectures are considered, namely, the transformer-based approach to modeling temporality and the Spatial-Temporal Graph Convolutional Networks.

For the former, a higher-dimensional embedding of the input feature space is performed, followed by a positional encoding to preserve temporality. Positional encodings are necessary because the Transformer architecture has no inherent sense of temporal order; they inject information about each frame's position in the sequence

so that the model can reason about motion progression from preparation through execution to return. Multiple self-attention layers are then employed to learn long-range temporal dependencies, enabling the model to learn global motion patterns [26, 27]. Intuitively, self-attention enables every frame to attend to all other frames simultaneously, so the model can relate, for example, the preparation phase of a sign to its execution phase without being constrained by the distance between them in time. Rather than using a mean pooling strategy, attention-based pooling is adopted to compute a weighted aggregation of frame-level representations. Intuitively, not all frames contribute equally to gesture identity — the peak of a motion is far more discriminative than the preparation or return phase — and attention pooling learns to assign higher weights to these informative frames automatically, producing a more representative embedding of the whole gesture. Transformer-based temporal embedding learning builds upon prior work on learning temporal embeddings for video representation [7].

For the latter, a skeleton is modeled as a graph, where joints are represented as nodes, and edges are formed by bones. This graph structure mirrors the physical connectivity of the human body, allowing the network to propagate information along anatomically meaningful connections — for instance, relating finger joint motion to wrist and elbow positions — rather than treating all joints as independent features. Spatial graph convolutions [29] are employed to capture these joint dependencies, and temporal convolutions are adopted to capture motion evolution across frames. Two-stream variants are also considered to separately process pose and hand information before feature fusion. Graph-based modeling has been widely used for structured motion representation in visual analysis tasks [14].

4.6 Metric Learning Objectives

Several metric learning objectives are considered to organize the embedding space [8, 9]. In supervised contrastive loss, embeddings of different classes are forced to be far apart, and embeddings of the same class are encouraged to be close [21].

ArcFace adds a penalty term for angular margin, which helps to improve separation for classification-based training [22]. Intuitively, ArcFace enforces a minimum angular gap between class boundaries in the hyperspherical embedding space, making each gesture class more compact and increasing the separation between visually similar but distinct signs. A combination of both objectives is also considered. These objectives directly affect the embedding space, which is essential for retrieval [12, 13].

4.7 Retrieval Evaluation Setup

While training, the metric learning pipeline remains stable using P-K batch sampling, whereby P number of classes is randomly sampled for each batch and K samples from each class are collected. In this case, K=3 was chosen to mimic the average number of classes seen in WLASL, which normally ranges from three to four videos per gloss, making sure that nearly all classes will be able to generate their respective positive pair sets without leaving behind rare classes. This guarantees the generation of adequate positive pairs for supervised contrastive learning [9]. Additionally, mild data augmentation is done in the optimal Transformer model.

After training, all test set samples are mapped to a feature embedding space. The similarity between two gestures is calculated using cosine similarity, which measures the angle between two embedding vectors rather than their absolute distance. This makes it invariant to the magnitude of the embeddings and well-suited to normalized feature spaces, where directional proximity reflects semantic similarity. For each query, its k-nearest neighbors are retrieved and ranked according to their similarity score.

The performance of the system is measured using various ranking-based metrics, such as Recall@K and mean Average Precision (mAP). These metrics are based on the system's ability to retrieve relevant gestures within its top-ranked results, which is in line with the purpose of a reverse gesture dictionary and ranking-based metric learning approaches [12, 1].

To examine generalization beyond the training distribution, the same retrieval protocol is additionally applied to the AUTSL subset described in the previous chapter.

In this setting, the WLASL-trained model is used without fine-tuning, four samples per class form the gallery, and one sample per class is used as the query. This makes it possible to evaluate how well the learned embedding space transfers to unseen sign vocabularies under the same preprocessing pipeline.

4.8 Indexing and Efficient Similarity Search

While the retrieval process presented in the previous sections requires the calculation of all pairwise similarities between embeddings of the query image and images from the gallery dataset, such an operation may be costly when the size of the database increases. In order to make use of indexing and speed up the query process, approximate nearest neighbor (ANN) algorithms have been studied as a possible extension of the proposed retrieval system.

Approximate nearest neighbor algorithms seek to trade off some amount of accuracy for much lower computational cost compared to exact algorithms. One of the most popular ANN algorithms is Hierarchical Navigable Small World (HNSW) graphs, first presented by Malkov and Yashunin [16]. This algorithm constructs a multi-level graph in which search can take place within logarithmic time. Another well-known efficient similarity searching library is FAISS [24] which implements exact algorithms (flat) and approximate search based on HNSW as well as inverted index (IVF) searching.

The following indexing methods have been tested on top of the learned embedding space:

- **Brute-force search**: exact cosine similarity computation between query and all gallery embeddings.
- **FAISS Flat**: exact nearest neighbor search implemented using FAISS.
- **FAISS HNSW**: graph-based approximate nearest neighbor search.
- **FAISS IVF-Flat**: clustering-based approximate search using inverted files.
- **HNSWlib**: a standalone implementation of HNSW for efficient ANN search.

All algorithms work with L2-normalized vectors and calculate inner product similarities, which correspond to cosine similarities in the normalized case. The test procedure still follows the retrieval paradigm outlined in Section 4.7, where the gallery is composed of the train partition and queries are sampled from the validation partition of the WLASL dataset. The accuracy of the algorithm is measured using Recall@K and mean Average Precision (mAP), along with other metrics reflecting agreement with the brute-force solution.

4.9 Summary

The methodological approach to the problem of learning similarity based on pose representations of gestures is given within this chapter. The proposed system includes a keypoint-based representation, temporal representation through the Transformer and ST-GCN models, and metric learning losses designed to optimize retrieval. Besides the performance of the system in terms of retrieval accuracy, an extension of the system in terms of search efficiency is considered by incorporating indexing into the process of similarity computation. The efficiency of search using both exact and approximate nearest neighbors will be analyzed in order to provide the assessment of the proposed method both in terms of accuracy and efficiency in application. The findings from the experiments will be reported in the following chapter.

Chapter 5

Results

5.1 Experimental Overview

In this chapter, we will explore the experimentally achieved results for the proposed gesture retrieval framework. First of all, it should be noted that all considered models were trained on the WLASL training dataset [20]. Additionally, the preprocessing method mentioned in the previous chapter was applied. Furthermore, it should be highlighted that for the experiments, only those gestures that had at least two samples were used because metric learning aims to distinguish between different categories [8, 9].

The presented experiment was performed based on the ranking evaluation criteria. For each sample from the validation set, its cosine similarity value with respect to all embeddings from the gallery set was estimated. Afterward, the values were sorted by their similarity score. The metrics used in the experiment were Recall@1, Recall@5, Recall@10, Recall@50, and mAP, which are common in metric learning for retrieval tasks [12, 1]. Table 5.1 shows the results obtained by all considered architectures.

5.2 Quantitative Results

The results demonstrate clear performance differences across architectural and loss configurations. The Transformer-based architecture's effectiveness aligns with recent

Table 5.1: Retrieval Performance Comparison on the WLASL Validation Set. The table reports Recall@K and mean Average Precision (mAP) for different architectures and metric learning configurations. The best results are highlighted in bold.

Model	Architecture	Loss Function	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	0.160	0.396	0.510	0.756	0.160
M4	Two-Stream ST-GCN	ArcFace + SupCon	0.177	0.446	0.559	0.769	0.192
M5	Transformer (Attention Pooling)	SupCon	**0.183**	0.433	0.554	0.732	**0.237**

findings on the power of self-attention mechanisms in visual representation learning [27]. To the best of our knowledge, no prior work reports retrieval metrics such as Recall@K or mAP on the WLASL dataset under the same few-shot setup, making a direct comparison with existing methods infeasible.

5.3 Qualitative Retrieval Results

In addition to quantitative evaluation, the results of the qualitative retrieval are provided to demonstrate the effectiveness of the system. Figure 5-1 shows the results of the query, along with the top 5 retrieved gestures in each case, which are part of the gallery set.

In each case, the retrieved gestures are of the same gesture category, although they are performed by different signers, under different backgrounds, and in different styles. This shows that the proposed model is focused on gesture semantics, not on low-level features of the gestures.

The results show that the proposed model is successful in retrieving visually similar gestures, which can be used to perform video-to-video retrieval without the need for textual supervision.

Figure 5-1: Example gesture retrieval results.

5.4 Transformer vs. ST-GCN Architectures

In all cases, the performance of the Transformer-based models surpassed that of the ST-GCN-based models, as measured by mean Average Precision. The baseline Transformer, even with supervised contrastive loss [21], achieved a value of 0.212, which was already higher than that of ST-GCN with ProxyNCA and ArcFace [22].

Despite the explicit modeling of skeletal structure using graph convolutions, ST-GCN failed to achieve a higher retrieval accuracy than the Transformer-based model. This may indicate that global temporal modeling using self-attention is a more powerful approach for modeling fine-grained motion similarity in isolated sign sequences, consistent with temporal embedding learning principles [7]. The best ST-GCN architecture, ArcFace + SupCon, achieved a value of 0.192, still lower than that of the baseline Transformer.

5.5 Impact of Loss Functions

The impact of the selection of the loss function is substantial. The performance of the ST-GCN model with ProxyNCA was the lowest among all the configurations. Switching the ProxyNCA loss function to the ArcFace loss function [22] boosted the performance. This shows the importance of angular margin-based objectives, as they improve the inter-class distance in the embedding space [8]. Further performance

improvements were observed when the supervised contrastive loss function [21] was added to the ArcFace loss function. This shows the importance of metric learning objectives for the task of gesture retrieval, particularly ranking-oriented embedding optimization [12].

5.6 Effect of Attention Pooling

In the last experiment, attention-based temporal pooling was integrated with the Transformer model, along with mild data augmentation and P-K batch sampling (K=3), which is commonly used in metric learning batch construction [9]. This resulted in the best performance in terms of mean Average Precision (0.237), as well as Recall@1 (0.183), compared to mean pooling. Compared to mean pooling, attention-based temporal pooling assigns higher weights to the most informative parts of a gesture sequence rather than averaging all frames equally. This is particularly useful for isolated signs, where preparation, execution, and return phases do not contribute equally to gesture identity. As a result, the obtained embeddings offer a more accurate representation of the entire gesture, hence providing higher ranking quality. Though the slight decrease in Recall@5 and Recall@10 is observed when comparing with the baseline transformer, there is an evident increase in ranking quality, measured through mAP, in accordance with retrieval-based metrics [12].

5.7 Cross-Dataset AUTSL Evaluation

To test the generalization ability on other datasets, the best model, which was a Transformer with attention pooling, was tested on the AUTSL dataset with 904 gallery and 226 query samples, and it achieved a Recall@1 of 3.54%, Recall@10 of 29.20%, and mAP of 0.0914. This shows that the model is able to maintain the partial structure in the embedding space despite not being trained on the AUTSL dataset.

5.8 Embedding Space Visualization

To gain more insight into the learned embedding space, a visual representation of the embedding space using the t-SNE method has been given in Figure 5-2. t-SNE is a non-linear dimensional reduction approach which embeds the high dimensional vectors in two dimensions such that similar vectors (points that are closer to each other in the graph) have been mapped from similar vectors in the 256 dimensional embedding space, while vectors that have been mapped farther away have been dissimilar to each other in the high dimensional space. In the following figure, each dot represents one sample of the gesture, while the color represents the class of the gesture.

It has been observed from the figure that the points of the same class are forming clusters, and the clusters of different classes are well separated, which implies that the embedding model is effectively learning a space in which distances between the points of the same class are minimized, and the distances between points of different classes are maximized, which is required for a retrieval system.

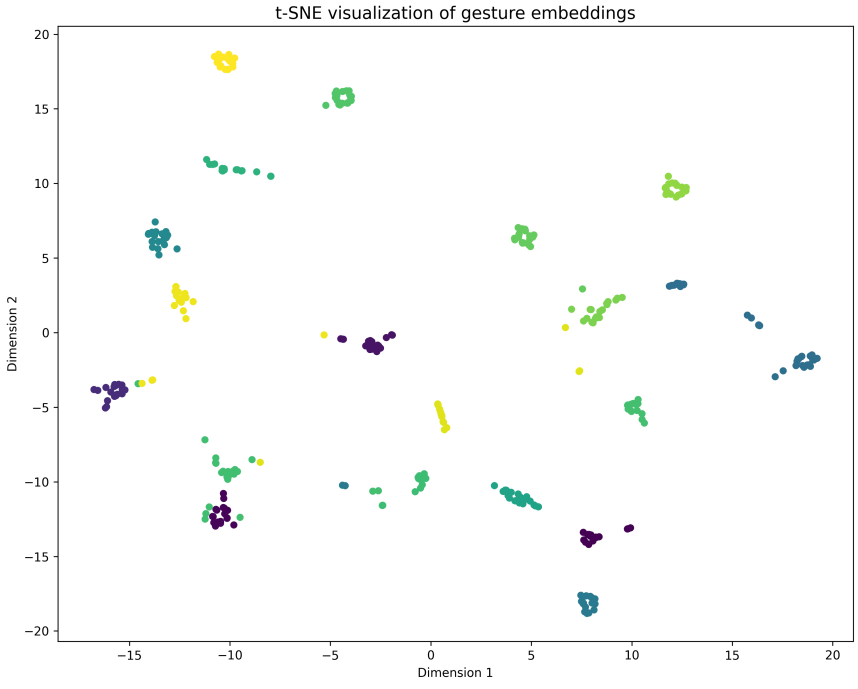

Figure 5-2: t-SNE visualization of gesture embeddings.

The UMAP embedding space can be seen in Figure 5-3. UMAP (Uniform Manifold

Approximation and Projection) is another dimensionality reduction algorithm that, contrary to t-SNE, retains more of the structure of the global nature of the data along with local structure. It comes in handy in determining if the clearly visible cluster separation in the t-SNE embedding plot is due to the global nature or just a local compacting of the data. The UMAP embedding is better than the t-SNE embedding when considering both aspects.

The above results also verify the effectiveness of the learned embedding, which achieves well-separated clusters for different gesture classes, and the effectiveness of the proposed metric learning method for the modeling of gesture similarity.

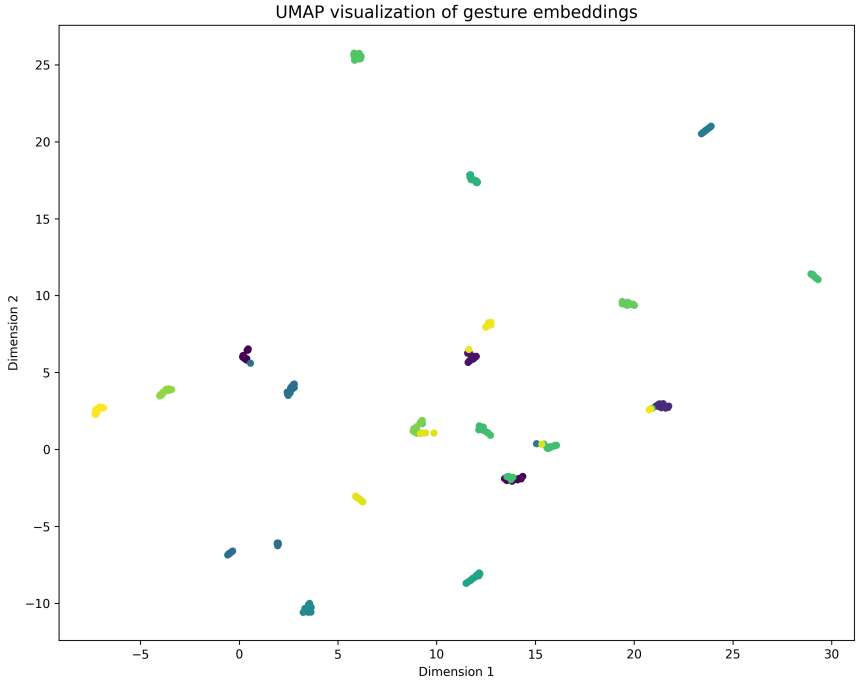

Figure 5-3: UMAP visualization of gesture embeddings.

5.9 Indexing Experiments

For analyzing the performance of indexing approaches for gesture retrieval, experiments were conducted using WLASL dataset. The model is trained on the WLASL train split containing 8,297 instances and tested on the validation split comprising 2,231 queries. The dimensions of the produced embedding space are 256, a widely used

embedding dimensionality in deep metric learning that provides sufficient capacity to encode gesture-level variation while remaining computationally tractable [9].

The comparative performance analysis between brute-force approach and indexing-based retrieval techniques is shown in Table 5.2.

Table 5.2: Comparison of indexing methods on WLASL retrieval

Method	ms/query	R@1	R@5	R@10	R@50	mAP
Brute-force	**0.0037**	0.1793	0.4361	0.5477	0.7333	0.2405
FAISS Flat	0.0503	0.1806	0.4388	0.5477	0.7333	0.2409
FAISS HNSW	0.0799	0.1793	0.4361	0.5477	0.7333	0.2405
FAISS IVF	0.0378	0.1806	0.4388	0.5482	0.7302	0.2412
HNSWlib	0.0442	0.1793	0.4361	0.5477	0.7333	0.2405

As can be seen from the results, each method provides almost identical recall and precision when compared to brute force search. The difference in the values of both Recall@K and mAP is very small (less than approximately 0.001–0.003). It is thus possible to conclude that approximate methods do not alter the structure of the learned embeddings space significantly.

Next, we calculated agreement between the brute-force baseline and each ANN method to investigate their ranking consistency. Both FAISS HNSW and HNSWlib produce an absolute agreement (100%) in both top-1 predictions and top-5 overlap. On the other hand, FAISS Flat and IVF have some discrepancies in their results, namely, 92.6–92.7% of agreement in top-1 results and more than 97.8% overlap for top-5 nearest neighbors.

However, a different picture emerges once we consider query times. In this case, brute-force approach shows the best (the lowest) query times with approximately 0.0037 ms per query. On the contrary, ANN approaches are slower due to extra operations required to build an index and traverse it. Therefore, ANN methods cannot leverage matrix multiplication on GPU. Hence, at the moment, there is no benefit in terms of query time in this particular task due to the dataset at such scale.

5.10 Summary of Findings

Based on the experiments, the following conclusions are made:

- Transformer-based temporal modeling achieves higher retrieval performance than graph-based modeling for isolated gesture retrieval.
- Angular margin losses improve embedding separation compared to proxy-based objectives [22].
- The combination of ArcFace and supervised contrastive loss enhances retrieval accuracy [21].
- The attention-based temporal pooling approach provides better embedding qualities compared to the mean pooling technique.
- The approximate nearest neighbor indexing techniques ensure retrieval quality without offering any latency advantages considering the existing dataset size; in this case, the exhaustive search using GPU becomes a more efficient alternative.

From among all the proposed configurations, the transformer model equipped with the attention-based pooling mechanism and supervised contrastive loss shows better results in terms of the retrieval quality. The experimental findings are analyzed further in the following chapter.

Chapter 6

Discussion

6.1 Overview

This chapter examines the experimental results and discusses their implications for gesture similarity learning. The intent is not to repeat the numerical results, but rather to understand why some architectures and loss functions were better performing and what this means in terms of isolated sign retrieval [12, 1]. The experiments show that the quality of the embedding is affected by the temporal modeling strategy, loss function, and pooling method [8]. The Transformer-based architecture with the attention pooling method yielded the best overall retrieval performance, which indicates the strength of global modeling in this problem.

6.2 Transformer vs. Graph-Based Modeling

One of the key findings from the current study is that the performance of the Transformer-based model surpassed the performance of the Spatial-Temporal Graph Convolutional Networks (ST-GCN) model in the isolated sign retrieval task. The ST-GCN model incorporates the skeletal structure of the hand through the use of a graph convolutional network. This has been a popular approach in action recognition, where the spatial relationships between the joints are crucial. However, in the current low-shot retrieval setting, the ST-GCN model did not achieve superior performance in

comparison to the Transformer-based encoder.

There are a number of reasons why the ST-GCN model did not achieve superior performance in the current study. Firstly, the WLASL dataset contains only three to four instances per class on average [20]. The graph-based approach relies on the ability to learn robust spatial-temporal patterns from multiple instances.

Secondly, the Transformers incorporate self-attention mechanisms that capture global temporal dependencies across the whole sequence [7]. Contrary to the graph convolutional approach that emphasizes local joint relationships, the self-attention mechanisms emphasize all time steps relative to each other. The ability to capture global temporal dependencies could help capture the nuanced differences in motions between visually similar gestures.

Finally, the Transformer architecture is less restricted by the skeletal structure. For the graph models, the structure is heavily influenced by the joint adjacency matrix. In the attention mechanisms, the ability to capture long-range dependencies is flexible and not restricted to the skeletal structure. In a retrieval scenario where the differences matter, the flexibility could offer an advantage.

These results are also in line with the recent trends in sequence modeling where self-attention models have achieved remarkable performance in modeling long-range dependencies across different domains.

6.3 Influence of Metric Learning Objectives

As mentioned earlier, the experiments also demonstrate the significance of the learning objectives used for the metric learning [8]. ProxyNCA yielded the poorest performance among all the losses used for evaluation. While proxy-based approaches ease the optimization process by learning the proxy representation for all classes, they may not capture the intra-class clustering adequately, especially under low-shot conditions [9]. Substituting ProxyNCA with the ArcFace loss significantly boosted the performance [22]. ArcFace incorporates the angular margin, thus improving the inter-class discriminability [22]. This shows the effectiveness of angular discriminability in the gesture

retrieval task.

Further performance gains were observed when the supervised contrastive loss was added with the ArcFace loss [21]. While the angular margin improves the inter-class discriminability, the addition of the supervised contrastive loss strengthens the intra-class clustering. This shows the effectiveness of the retrieval-oriented optimization objectives [12].

6.4 Impact of Attention Pooling

The greatest improvement in the overall quality of the retrieval process was achieved by the attention pooling method. The reason for this is the ability of the model to learn which parts of the video contribute the most to the gesture identity, unlike the mean pooling method. Isolated signs have different phases: preparation, execution, and return. Not all frames of the gesture are equally discriminative. The mean pooling method treats all the frames equally, which may reduce the discriminative information of the gesture. The better mAP values of the attention pooling method imply that the model has learned the importance of the frames, which is a crucial component of the gesture similarity modeling process, consistent with attention-based modeling approaches [3].

6.5 Low-Shot Learning Considerations

WLASL dataset is a low-shot learning scenario, i.e., there are only a few examples per word sense [20]. This also presents a number of challenges for representation learning. Less repetition of classes makes it difficult for the model to learn intra-class variation. The better performance of the Transformer-based model might be due to the robustness of global temporal modeling, as compared to graph-based structural modeling, in low-shot learning scenarios [19].

Additionally, using P-K sampling with K=3 was beneficial for the stability of contrastive learning, as there were enough positive pairs in each batch [9]. This

also emphasizes the role of batch formation in metric learning, especially in low-shot learning.

The cross-dataset experiment on AUTSL further emphasizes that the learned embedding captures transferable motion patterns even when the sign vocabulary and signer distribution are very different from WLASL. While the absolute values of the retrieval scores are not high, the fact that it is able to retrieve the correct matches in a 226-class open-set scenario is an indication that it has learned gesture-level similarities rather than learning dataset-specific information.

6.6 Error Analysis

Although the quantitative analysis proves the validity of the method suggested, there is an urgent need to analyze cases when the system retrieves signs incorrectly in order to see what difficulties the task still poses and how they can be overcome.

The first type of retrieval errors is caused by signs that are visually very close to each other: they use the same hand gesture or have almost identical paths of motion, but vary in their relation to the signer, the palm direction, the number of gestures performed, and so on. In the WLASL database, where each class includes merely three to four examples, the network does not get enough training to recognize differences between near-duplicates, and therefore makes mistakes while retrieving top-1 signs.

The second source of error comes from the quality of MediaPipe detection. As the entire embedding is derived from the MediaPipe keypoints, failure in detection leads directly to bad representations. Unusual hand orientations, rapid movement causing motion blur, or partial self-occlusions — where one hand occludes the other in a signing gesture — may cause MediaPipe to provide faulty detection results. While missing keypoints are linearly interpolated during preprocessing, erroneous or severely erroneous detections cannot be interpolated entirely and therefore affect the final embeddings, which no longer capture the underlying gestures. This can be seen in the low Recall@1 scores of (0.178-0.183), while the Recall@50 scores (0.732-0.771) indicate that the correct class is present but not necessarily ranked first in the neighborhood.

Lastly, when evaluating the model on the AUTSL dataset, we encounter another type of error, known as domain shift. Recall@1 of 3.54% on AUTSL versus 18.3% on WLASL suggests that while the model generalizes to a new sign language vocabulary, the embeddings become less effective when tested on signers with different characteristics and environmental conditions.

6.7 Limitations

Although the results are promising, several aspects are to be noted. Firstly, the study only considers isolated sign retrieval. Continuous sign sequences are not considered. More segmentation mechanisms would be needed to tackle the retrieval of sentences [18].

Secondly, large-scale indexing techniques like FAISS or HNSW were not fully explored [16, 17]. Although the evaluation of the efficiency of the embeddings with the help of the cosine similarity is sufficient for the analysis of the embeddings' quality, the use of the above-mentioned techniques would be necessary for the effective deployment of the system.

Thirdly, the study is mostly based on the WLASL dataset [20]. The cross-lingual and cross-dataset evaluation of the sign language dataset for the Russian and Kazakh languages is an important task to be validated.

6.8 Implications and Future Directions

The results of this thesis demonstrate the viability of video-to-video gesture retrieval based on pose representations, as well as the effectiveness of attention-based Transformer models in low-shot gesture retrieval scenarios. The proposed framework shows that meaningful gesture similarity can be learned without reliance on textual supervision, supporting the development of intuitive, visually driven retrieval systems.

The indexing experiments provide additional insights into the scalability of the system. While approximate nearest neighbor methods such as FAISS and HNSW are

designed to accelerate large-scale search [16, 17], the results show that, at the current dataset scale, GPU-based exhaustive search remains more efficient. Meanwhile, the high correlation between results of approximate techniques and brute-force search demonstrates a good structure and stability of learned embedding space. In particular, HNSW-based approaches produce exact nearest neighbors, which proves that approximate search correctly maintains the semantics of the relation between gestures. Thus, indexing must be considered as a scalable addition to the system, and its relevance increases with an increase in the size of the gesture database.

There are several ways to enhance the presented approach. Firstly, experiments can be conducted using larger and more varied data sets to investigate the potential benefits of indexing approaches. Secondly, evaluation of the system for the domain-specific gesture retrieval task, including testing on internal Russian and Kazakh gesture databases, can demonstrate the applicability of the system for a cross-lingual and cross-cultural scenario. Thirdly, more advanced methods of temporal gesture sequence processing can allow for extending the system to the level of sign language sequence similarity assessment [7].

In conclusion, the discussed approach is a step towards creating effective visual gesture similarity systems and reverse sign language dictionaries [4].

Chapter 7

Conclusion

In this thesis, a video-based reverse dictionary for sign languages has been proposed by reformulating gesture understanding as a video-to-video retrieval problem. The system utilizes pose and hand keypoints, normalized spatiotemporal representations, and a combination of Transformer and ST-GCN architectures, with embedding learning performed using a metric learning approach. The main experiments were conducted on the WLASL dataset, as it is more suitable for the retrieval task due to the presence of multiple samples per gloss and signer variability. The results demonstrate the superiority of the Transformer architecture over the ST-GCN architecture, achieving an mAP of 0.237 and a Recall@1 of 0.183. In addition, the study highlights the importance of loss function design for improving embedding quality, where the combination of SupCon and ArcFace loss functions with attention pooling enables the model to focus on the most relevant temporal segments of a gesture.

Also, evaluation on a subset of the AUTSL dataset revealed limited generalization capability (mAP: 0.0914, Recall@1: 3.54%), indicating that the model learns partially transferable gesture representations rather than overfitting entirely to the training data. The main limitations of the approach are its restriction to isolated sign gestures and its reliance on keypoint-based representations. In addition, the proposed model does not support efficient large-scale retrieval and is not applicable to continuous signing scenarios.

Overall, pose-based embedding learning with Transformer architectures shows

strong potential for gesture retrieval in sign language and can serve as a foundation for reverse dictionary systems.

Bibliography

[1] A. Duarte, S. Albanie, X. Giró-i-Nieto and G. Varol, "Sign Language Video Retrieval with Free-Form Textual Queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14094–14104, 2022, doi: 10.48550/arXiv.2201.02495.

[2] Y. Cheng, F. Wei, J. Bao, D. Chen and W. Zhang, "CiCo: Domain-Aware Sign Language Retrieval via Cross-Lingual Contrastive Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19016–19026, 2023.

[3] M. Parian-Scherb, P. Uhrig, L. Rossetto, S. Dupont and H. Schuldt, "Gesture Retrieval and its Application to the Study of Multimodal Communication," *International Journal on Digital Libraries*, vol. 25, pp. 585–601, 2024, doi: 10.1007/s00799-023-00367-0.

[4] S. Hassan, M. Boháček, C. Kim and D. Crochet, "Towards an AI-Driven Video-Based American Sign Language Dictionary: Exploring Design and Usage Experience with Learners," *arXiv preprint*, 2025, doi: 10.48550/arXiv.2504.05857.

[5] Z. Jiang, G. Sant Muniesa, A. Moryossef, M. Müller, R. Sennrich and S. Ebling, "SignCLIP: Connecting Text and Sign Language by Contrastive Learning," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9171–9193, 2024, doi: 10.18653/v1/2024.emnlp-main.518.

[6] E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos and F. Lefebvre-Albaret, "The Dicta-Sign Wiki: Enabling Web Communication for the Deaf," in *International Conference on Computers Helping People with Special Needs*, pp. 205–212, Springer, 2012.

[7] V. Ramanathan, K. Tang, G. Mori and L. Fei-Fei, "Learning Temporal Embeddings for Complex Video Analysis," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4471–4479, 2015, doi: 10.1109/ICCV.2015.508.

[8] B. Ghojogh, A. Ghodsi, F. Karray and M. Crowley, "Spectral, Probabilistic, and Deep Metric Learning: Tutorial and Survey," *arXiv preprint*, 2022, doi: 10.48550/arXiv.2201.09267.

[9] K. Musgrave, S. Belongie and S.-N. Lim, "A Metric Learning Reality Check," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 681–699, 2020, doi: 10.1007/978-3-030-58595-2_41.

[10] M. Müller, "Dynamic Time Warping," in *Information Retrieval for Music and Motion*, Springer, 2007.

[11] S. Celebi, A. S. Aydin, T. T. Temiz and T. Arici, "Gesture Recognition using Skeleton Data with Weighted Dynamic Time Warping," in *Proceedings of VISAPP*, pp. 620–625, 2013.

[12] F. Cakir, K. He, X. Xia, B. Kulis and S. Sclaroff, "Deep Metric Learning to Rank," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1861–1870, 2019.

[13] Y. Duan, J. Lu, W. Zheng and J. Zhou, "Deep Adversarial Metric Learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 2037–2051, 2020, doi: 10.1109/TIP.2019.2948472.

[14] A. Zahra, N. Perwaiz, M. Shahzad and M. M. Fraz, "Person Re-Identification: A Retrospective on Domain-Specific Open Challenges and Future Trends," *Pattern Recognition*, vol. 142, p. 109669, 2023, doi: 10.1016/j.patcog.2023.109669.

[15] J. Ouyang, W. Zhou, M. Wang, Q. Tian and H. Li, "Collaborative Image Relevance Learning for Visual Re-Ranking," *IEEE Transactions on Multimedia*, vol. 23, pp. 3646–3656, 2021, doi: 10.1109/TMM.2020.3029886.

[16] Y. A. Malkov and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 824–836, 2018, doi: 10.1109/TPAMI.2018.2889473.

[17] S. Emanuilov and A. Dimov, "Billion-Scale Similarity Search Using a Hybrid Indexing Approach with Advanced Filtering," *Cybernetics and Information Technologies*, vol. 24, no. 4, pp. 45–58, 2024, doi: 10.2478/cait-2024-0035.

[18] G. V. Martins, "SLVideo: A Sign Language Video Moment Retrieval Framework," Master's thesis, Universidade NOVA de Lisboa, 2024.

[19] X. Li, X. Yang, Z. Ma and J.-H. Xue, "Deep Metric Learning for Few-Shot Image Classification: A Review of Recent Developments," *Pattern Recognition*, vol. 138, p. 109381, 2023.

[20] D. Li, C. R. Opazo, X. Yu and H. Li, "Word-Level Deep Sign Language Recognition from Video: A New Large-Scale Dataset and Methods Comparison," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1459–1469, 2020.

[21] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu and D. Krishnan, "Supervised Contrastive Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 18661–18673, 2020.

[22] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019, doi: 10.1109/CVPR.2019.00482.

[23] O. M. Sincan and H. Y. Keles, "AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods," *IEEE Access*, vol. 8, pp. 181340–181355, 2020.

[24] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021, doi: 10.1109/TBDATA.2019.2921572.

[25] MediaPipe Holistic Demo, "MediaPipe 0.8 Colab Holistic," 2021. [Online]. Available: https://tensorflow.classcat.com/2021/03/10/mediapipe-0-8-colab-holistic/

[26] A. Vaswani, N. Shazeer, P. N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[28] Z. Cao, G. Hidalgo, T. Simon, S. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[29] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[30] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

[31] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.