Information about the authors

Yerzhan Seitkulov* – PhD, L.N. Gumilyov Eurasian National University, Republic of Kazakhstan; e-mail: yerzhan.seitkulov@gmail.com. ORCID: https://orcid.org/0000-0002-5172-8339.

Ruslan Ospanov – Researcher, Gumilyov Eurasian National University, Republic of Kazakhstan; e-mail: ospanovrm@gmail.com. ORCID: https://orcid.org/0000-0002-0771-575X.

Banu Yergaliyeva – Researcher, Gumilyov Eurasian National University, Republic of Kazakhstan; e-mail: banu.yergalieva@gmail.com. ORCID: https://orcid.org/0000-0002-1252-857X.

Kuat Utebayev – Master student, Almaty Branch, National Research Nuclear University MEPhi, Almaty 050040, Kazakhstan; e-mail: utebayevkuat@gmail.com. ORCID: https://orcid.org/0009-0004-3032-5049.

Авторлар туралы ақпарат

Ержан Сейтқұлов* – PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Қазақстан Республикасы; e-mail: yerzhan.seitkulov@gmail.com. ORCID: https://orcid.org/0000-0002-5172-8339.

Руслан Оспанов – Ғылыми қызметкер, Гумилев атындағы Еуразия ұлттық университеті, Қазақстан Республикасы; e-mail: ospanovrm@gmail.com. ORCID: https://orcid.org/0000-0002-0771-575X.

Бану Ерғалиева – Ғылыми қызметкер, Гумилев атындағы Еуразия ұлттық университеті, Қазақстан Республикасы; e-mail: banu.yergalieva@gmail.com. ORCID: https://orcid.org/0000-0002-1252-857X.

Қуат Өтебаев – магистрант, MEPhi Ұлттық зерттеу ядролық университетінің Алматы филиалы, Алматы 050040, Қазақстан; e-mail: utebayevkuat@gmail.com. ORCID: https://orcid.org/0009-0004-3032-5049.

Информация об авторах

Ержан Сейткулов* – PhD, Л.Н. Евразийский национальный университет имени Гумилева, Республика Казахстан; e-mail: erzhan.seitkulov@gmail.com. ORCID: https://orcid.org/0000-0002-5172-8339.

Руслан Оспанов – научный сотрудник Евразийского национального университета имени Гумилева, Республика Казахстан; e-mail: ospanovrm@gmail.com. ORCID: https://orcid.org/0000-0002-0771-575X.

Бану Ергалиева – научный сотрудник Евразийского национального университета имени Гумилева, Республика Казахстан; e-mail: Banu.yergalieva@gmail.com. ORCID: https://orcid.org/0000-0002-1252-857X.

Куат Утебаев – магистрант, Алматинский филиал Национального исследовательского ядерного университета МИФИ, Алматы Казахстан; e-mail: utebayevkuat@gmail.com. ORCID: https://orcid.org/0009-0004-3032-5049.

Received 04.05.2026
Revised 19.06.2026
Accepted 22.06.2026

https://doi.org/10.53360/2788-7995-2026-3(23)-8

IRSTI: 28.23.25

CC BY 4.0
Check for updates

A. Zhakanova[1], Zh. Oralbekova[1]*, D. Oralbekova[2]
[1]Astana IT University,
010000, Republic of Kazakhstan, Astana, 55/11 Mangilik El avenue, C1 block
[2]Institute of Information and Computational Technologies,
050010, Republic of Kazakhstan, Almaty, 28 Shevchenko street
*e-mail: zh.oralbekova@astanait.edu.kz

DEVELOPMENT OF A MODEL OF KAZAKH SIGN LANGUAGE FOR CLOSED-SET PHRASE RECOGNITION AND TRAINING

Annotation: *Kazakh Sign Language (KSL) remains underrepresented in current sign language processing research due to a lack of large, annotated datasets and specialized computational tools. This paper presents a deep learning-based framework for phrase-level recognition of sign language using visual keypoint sequences, in which each recognized phrase is mapped to its gloss label and a verified reference translation in Russian. The approach employs the FluentSigners-50 dataset as a benchmark corpus and extracts pose*

©A. Zhakanova, Zh. Oralbekova, D. Oralbekova, 2026

and hand landmarks from video data using the MediaPipe Holistic model. A two-layer bidirectional long short-term memory (BiLSTM) network is trained on the resulting spatiotemporal sequences to classify forty distinct KSL phrases. The model takes as input a fixed-length sequence of 150 frames, each containing 225 keypoint features, and outputs a predicted phrase class through a multilayer perceptron. Experimental results demonstrate strong performance in phrase classification, with accurate mappings from video to gloss labels and their corresponding Russian translations, achieving an accuracy of 94.73%. Rather than performing open-vocabulary machine translation, the system addresses closed-set phrase classification with gloss and reference-translation lookup, making it suitable for structured assistive and training applications in sign language learning environments.

Key words: *kazakh sign language, sign language recognition, deep learning, bidirectional LSTM, MediaPipe Holistic, keypoint sequence modeling, closed-set phrase classification.*

Introduction

Sign language plays a critical role in enabling communication for deaf and hard-of-hearing (DHH) individuals across the world. In Kazakhstan, Kazakh Sign Language (KSL) serves as the primary medium of communication for a substantial DHH population. Despite its importance, KSL remains severely underrepresented in sign language recognition (SLR) research. This lack of computational support restricts access to education, healthcare, and public services, and creates a persistent communication gap between hearing and non-hearing communities [1, 2]. The development of automatic translation systems for KSL is thus not only technically relevant but also socially imperative.

Globally, recent advancements in deep learning have enabled significant progress in automatic sign language recognition and translation, particularly for high-resource sign languages such as American Sign Language (ASL), Chinese Sign Language (CSL), and Arabic Sign Language [3], [4]. These developments rely on large annotated datasets and spatiotemporal modeling using Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models [5, 6]. Moreover, the use of pose estimation frameworks such as MediaPipe or OpenPose has enabled the extraction of structured body and hand keypoints for gesture modeling [7]. These advances, however, do not directly transfer to Kazakh Sign Language due to unique grammatical structures, signer variation, and resource scarcity [2].

Existing studies on KSL have explored isolated letter and word recognition using CNNs, SVMs, and LSTMs, reporting high accuracy on alphabet datasets [8, 9]. However, these works primarily focus on classification of static signs and lack support for phrase-level recognition or translation. Furthermore, many proposed systems rely on small scale, internally collected datasets with limited vocabulary coverage. The FluentSigners-50 dataset, introduced in [10], is a recent attempt to address this gap by providing sentence-level video samples of Kazakh Russian Sign Language (KRSL). Still, end-to-end frameworks for automatic KSL translation based on such data remain underexplored.

This paper presents a phrase-level recognition framework for Kazakh Sign Language based on spatiotemporal keypoint modeling and sequence classification. The model uses pose and hand landmarks extracted from the FluentSigners-50 dataset via MediaPipe Holistic and applies a two-layer Bidirectional LSTM for phrase classification. A subset of 40 commonly used KSL phrases was selected, with each input video converted into a uniform 150-frame sequence of 225-dimensional keypoint vectors. The model was trained and evaluated on this data, demonstrating reliable mapping of gestures to gloss labels and supporting future integration into KSL training and communication systems.

The system targets closed-set phrase recognition with gloss and reference-translation lookup: each input video is mapped to a fixed gloss label and a verified reference sentence in Russian. This is a deliberate scope, matched to structured KSL learning and communication scenarios where the target phrase set is known in advance.

Materials and methods

The experimental setup is based on the FluentSigners-50 dataset [10], a large-scale collection of phrase-level videos in Kazakh Russian Sign Language (KRSL). It contains over 43,000 samples from 50 fluent signers, covering 173 unique phrases. For this study, a curated subset of 40 representative KRSL phrases was selected. Each phrase (class) contains 250 video samples, resulting in a total of 10,000 annotated samples.

Each video is labeled with a class index (0-39), the corresponding sign gloss (e.g., “ЗДРАВСТВУЙТЕ”) and a Russian textual equivalent (e.g., “Здравствуйте”).

The dataset is split into training (70%), validation (15%), and test (15%) sets, stratified to ensure uniform class distribution across splits. In absolute terms, this corresponds to 7,000 training, 1,500 validation, and 1,500 test videos (37-38 samples per class in each of the validation and test partitions). The stratified split was performed at the level of individual video samples and balanced only with respect to phrase class.

To extract meaningful motion and posture information, the MediaPipe Holistic model was used, which detects 33 full body pose landmarks and 21 hand landmarks per hand. Each landmark provides three coordinates (x, y, z), resulting in a 225-dimensional feature vector per frame.

Each video is processed frame-by-frame. First, RGB frames are passed through MediaPipe. Then, if a hand is not visible, its keypoints are zero-filled. Finally, each frame’s 225-dimensional vector is stored sequentially.

All sequences are normalized to 150 frames. Shorter sequences are zero-padded, longer ones are truncated to 150. Thus, every input video becomes a tensor of shape (150, 225), ready for time-series processing.

As shown in Figure 1, the MediaPipe Holistic model extracts pose and hand keypoints directly from RGB video frames, yielding a total of 225 features per frame. The overlaid landmarks comprise body pose and both-hand joints, which together form the 225-dimensional feature vector used for gesture classification. These skeletal keypoints serve as a compact and modality-agnostic representation of signer motion and articulation across the entire video sequence.



Figure 1 – Example of pose and hand landmarks extracted from a Kazakh Sign Language video frame using MediaPipe Holistic

The proposed classification model is based on a Bidirectional Long Short-Term Memory (BiLSTM) architecture followed by a fully connected multilayer perceptron (MLP) head. The model is designed to capture both spatial and temporal characteristics of gesture sequences and to map them to phrase-level outputs.

The input to the network consists of sequences with shape (150, 225), where each sequence contains 150-time steps and each time step is represented by a 225-dimensional feature vector corresponding to frame-level keypoint features. These sequences are processed by two stacked BiLSTM layers. Each BiLSTM layer has a hidden size of 128 units per direction, resulting in a combined output dimensionality of 256 at each time step. Consequently, the BiLSTM encoder produces an output sequence of shape (150, 256), effectively capturing bidirectional temporal dependencies within the gesture data.

To transform the variable-length temporal information into a fixed-size representation, mean pooling is applied across the temporal dimension. This operation aggregates the sequence of 150-time steps into a single 256-dimensional feature vector, representing the entire video sequence.

The resulting representation is then passed through a multilayer perceptron classification head. The first dense layer reduces the feature dimension from 256 to 128 and is followed by batch normalization and a Rectified Linear Unit (ReLU) activation function. A dropout layer with a rate of 0.5 is applied to prevent overfitting. The second dense layer further reduces the dimensionality from 128 to 64, followed by another ReLU activation and a dropout layer with a rate of 0.3. Finally, a fully connected layer maps the 64-dimensional feature vector to 40 output classes, corresponding to the target phrase categories. A softmax activation function is applied to produce class probabilities over the 40 output classes.

This architecture effectively captures bidirectional temporal dependencies in gesture sequences while maintaining a compact and efficient representation. The overall design ensures that both spatial and temporal information are preserved and utilized for accurate phrase-level classification.

The complete architecture of the proposed model is illustrated in Figure 2. To improve generalization and ensure stable convergence during training, regularization techniques such as dropout and batch normalization are incorporated throughout the network. These design choices contribute to both spatiotemporal sensitivity and computational efficiency, making the model suitable for phrase-level sign language recognition tasks.

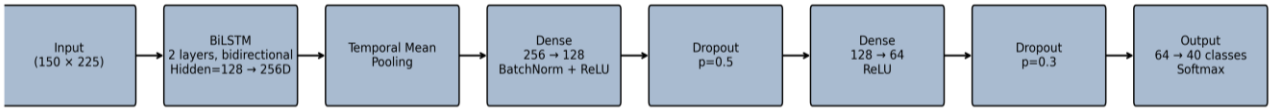


Figure 2 – Architecture of the model

The model was trained using a supervised learning approach with categorical cross-entropy as the loss function. Optimization was performed using the Adam optimizer, which provides adaptive learning rate adjustments and efficient convergence for deep neural networks.

The training process used a fixed schedule of 150 epochs with batch size 32, a ReduceLROnPlateau learning-rate scheduler (factor 0.5, patience 5 epochs on validation loss), and gradient clipping (max norm 1.0) to stabilize convergence. A checkpoint was saved after every epoch. The checkpoint achieving the highest validation accuracy across all 150 epochs was selected as the final model, the test set was held out completely from this selection process and used exactly once.

The final model was selected based on its validation set performance, with the best results obtained at epoch 150. This model achieved a test accuracy of 94.73%, demonstrating strong generalization capability on unseen data.

The BiLSTM-MLP classifier contains 802,792 trainable parameters (around 0.80 million and around 3.1 MB at single precision), substantially smaller than typical image-based or transformer-based sign language recognition models, reflecting the compactness gained by operating on low-dimensional keypoint sequences rather than raw video. All experiments were run on a single Apple M4 machine, the BiLSTM classifier was trained on CPU only, with no dedicated GPU. MediaPipe Holistic keypoint extraction required on average 3 seconds of processing per video on this hardware (a one-time, offline preprocessing cost of 8.3 hours for the full 10,000-video subset used in this study). Training the BiLSTM classifier for 150 epochs on this 40-class subset took approximately 3.4 hours.

Results

A total of 10,000 videos representing 40 distinct phrases were selected from the FluentSigners-50 dataset. Each phrase contains 250 samples performed by multiple fluent signers. The dataset was stratified into 70% training, 15% validation, and 15% test sets, ensuring balanced class representation. The final test set consisted of 1,500 examples (approximately 37-38 samples per class).

The final model, trained over 150 epochs using the Adam optimizer and cross-entropy loss, achieved an overall test accuracy of 94.73%. Table 1 summarizes the per-class precision, recall, and F1-score results across all 40 classes.

The overall accuracy on the test set was 94.73%, with a macro-average F1-score of 94.74%. The weighted average across all metrics remained consistently above 94%.

The normalized confusion matrix in Figure 3 illustrates classification confidence across all 40 classes. Most classes exhibited strong diagonal dominance, confirming high accuracy and low confusion. However, several misclassifications occur in semantically similar classes. In particular, the model confuses class 14 (“Какой ваш адрес”) with class 15 (“Где вы живете”), and class 17 (“Где находится магазин”) with class 16 (“Как вас зовут”), which correspond to contextually related location and identity queries.

Table 1 – per class classification metrics

Class ID	Precision	Recall	F1-score	Support
0	1.0000	1.0000	1.0000	37
1	0.9744	1.0000	0.9870	38
2	0.9500	1.0000	0.9744	38
3	1.0000	0.9737	0.9867	38
4	0.9722	0.9459	0.9589	37
5	0.8293	0.8947	0.8608	38
6	0.9737	1.0000	0.9867	37
7	0.8718	0.8947	0.8831	38
8	1.0000	0.9459	0.9722	37
9	0.9730	0.9730	0.9730	37
10	1.0000	0.9474	0.9730	38
11	0.9500	1.0000	0.9744	38
12	0.9730	0.9730	0.9730	37
13	0.9487	1.0000	0.9737	37
14	1.0000	1.0000	1.0000	38
15	0.6923	0.7297	0.7105	37
16	1.0000	0.9730	0.9863	37
17	0.7059	0.6486	0.6761	37
18	1.0000	0.9737	0.9867	38
19	0.9474	0.9730	0.9600	37
20	1.0000	0.9474	0.9730	38
21	1.0000	0.9474	0.9730	38
22	0.9250	0.9737	0.9487	38
23	1.0000	0.9737	0.9867	38
24	1.0000	0.8649	0.9275	37
25	0.9231	0.9730	0.9474	37
26	0.9474	0.9730	0.9600	37
27	0.9714	0.9189	0.9444	37
28	0.9444	0.8947	0.9189	38
29	0.9189	0.8947	0.9067	38
30	0.9444	0.8947	0.9189	38
31	0.7826	0.9730	0.8675	37
32	1.0000	0.9474	0.9730	38
33	1.0000	1.0000	1.0000	37
34	1.0000	1.0000	1.0000	37
35	0.9730	0.9474	0.9600	38
36	0.9500	1.0000	0.9744	38
37	0.9250	1.0000	0.9610	37
38	1.0000	0.9211	0.9589	38
39	1.0000	1.0000	1.0000	37

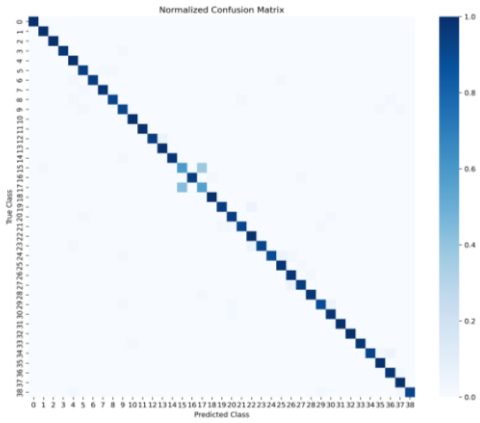


Figure 3 – Confusion matrix on the test set

These confusions suggest that although the model captures the dominant spatiotemporal patterns, gesture overlap and subtle hand trajectory differences may lead to ambiguity across neighboring classes. This indicates a potential need for more training examples or finer-grained motion modeling in future iterations.

To assess model convergence and generalization, we analyzed validation accuracy across training epochs. Table 2 reports performance at selected checkpoints.

Table 2 – validation accuracy across epochs

Experiment	Epochs	Val Accuracy	Val Loss
1	20	0.8020	1.3223
2	50	0.9020	0.9853
3	80	0.9207	0.9208
4	100	0.9293	0.9093
5	135	0.9280	0.9099
6	150	0.9473	0.9118
Final	150	0.9473	0.9118

As shown, performance steadily improved up to epoch 80. The best validation accuracy was recorded at epoch 150, which was selected as the final model checkpoint.

To further examine model learning dynamics, training and validation loss, as well as validation accuracy, were monitored across 150 epochs. The training dynamics illustrated in Figure 4 demonstrate stable convergence over the course of 150 epochs.

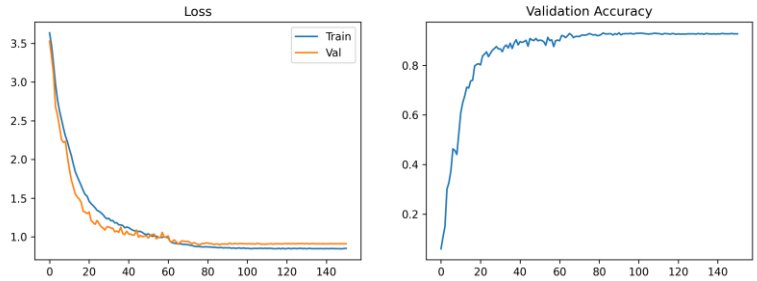


Figure 4 – Training and validation loss and validation accuracy curves over 150 epochs

The training loss exhibits a smooth and monotonic decrease, approaching a minimum below 0.9. The validation loss follows a similar trend, stabilizing after approximately 60 epochs, indicating the model has generalized well without overfitting. The validation accuracy improves rapidly during early epochs, plateauing near 0.91, confirming robust learning capacity of the BiLSTM-based architecture and the effectiveness of the chosen optimization strategy.

Discussion

The results of this study demonstrate the feasibility and practical value of using pose-based temporal modeling for phrase-level recognition in Kazakh Sign Language (KSL). The system achieved over 94.73% classification accuracy across 40 frequently used KRSL phrases, using MediaPipe keypoints and a BiLSTM architecture trained on a curated subset of the FluentSigners-50 dataset.

The use of MediaPipe Holistic for keypoint extraction provides a compact and noise-resistant representation of gesture dynamics. By transforming raw video input into structured skeletal features, the model effectively reduces sensitivity to environmental variations such as background clutter and illumination changes, while preserving essential spatial and temporal information. This representation enables efficient learning and contributes to the stable convergence observed during training. The training and validation loss curves exhibit consistent behavior, with no significant divergence, suggesting that the applied regularization techniques, including dropout and batch normalization, successfully mitigate overfitting.

Some of the model's confusion patterns can be attributed to semantic and gestural similarity across multiple KSL phrases. For example, confusion between phrases related to location ("Где вы живете" and "Какой ваш адрес") and identity-related queries indicates that overlapping motion trajectories and similar temporal patterns remain challenging for the model. Similarly, lower

performance in classes such as 15 and 17 (F1-scores of 0.71 and 0.67, respectively) suggests that subtle differences in gesture execution are not always fully captured by the current representation. These findings are consistent with prior work, where dynamic gesture similarity and coarticulation effects are known to limit classification performance in sign language recognition systems [11].

Furthermore, signer variability, including hand speed, articulation, and background, affects model consistency. While the MediaPipe Holistic model provides robust keypoints, even minor differences in pose estimation can propagate to misclassifications, particularly in boundary cases with short or incomplete signs. These findings highlight the need for more fine-grained attention mechanisms and signer-aware adaptation in future work [7].

One of the distinguishing features of this system is its applicability to KSL learning environments. The proposed model can serve as a foundational component in interactive training tools where learners practice phrase articulation and receive real-time classification feedback. For example, a language learner can perform a KSL phrase on camera, and the model would predict and display the detected phrase in gloss and spoken Kazakh, enabling self-correction and visual reinforcement.

This aligns with recent trends in computer-assisted sign language learning, where models trained on phrase-level gestures act as digital tutors [4]. Compared to prior systems trained on isolated signs or letters, the use of phrase-level data from FluentSigners-50 allows for more contextual and grammatically relevant feedback, which is critical for meaningful conversation training.

The work presented in this study introduces a technically novel contribution to Kazakh Sign Language recognition. It constitutes the first implementation of a phrase-level classification model for KSL that employs bidirectional long short-term memory (BiLSTM) networks operating on MediaPipe-based pose and hand keypoints. This architecture bridges spatially structured gesture input with sequential temporal modeling, enabling accurate phrase-level recognition in a low-resource linguistic setting.

In addition, the system incorporates a fully automated end-to-end pipeline that processes raw video data from the FluentSigners-50 dataset, performs landmark extraction, applies sequence normalization, and classifies complete phrases. Unlike prior work that typically focuses on isolated signs or alphabets, this model targets semantically meaningful phrase recognition, offering a more linguistically relevant output for real-world communication.

A further strength of the proposed system lies in its modular design: the architecture maintains a clear separation between feature extraction (via pose/hand keypoints) and sequence learning (via BiLSTM), ensuring flexibility and scalability. This modularity allows for seamless extension to additional phrase classes or sign languages with minimal structural adjustment. Collectively, these contributions position the model as a viable foundation for research and assistive applications in underrepresented sign languages such as KSL.

From a methodological perspective, the proposed approach can be considered competitive with recent state-of-the-art (SOTA) methods for low-resource sign language recognition. While CNN-based and transformer-based architectures have demonstrated strong performance in large-scale datasets, they typically require significantly higher computational resources and large annotated corpora. In contrast, the proposed BiLSTM model reduces computational overhead by operating on low-dimensional keypoint features rather than high-resolution image data, while maintaining high accuracy. This makes the model suitable for real-time applications and deployment in resource-constrained environments. Compared to approaches using YOLO-based detection combined with temporal modeling, the MediaPipe-based pipeline reduces preprocessing overhead while still capturing essential gesture dynamics.

To substantiate this comparison, Table 3 summarizes recent results on Kazakh Sign Language recognition.

[9] reported 99% accuracy using a Support Vector Machine on MediaPipe hand landmarks, but for a considerably simpler task, classifying 42 static dactyl (finger-spelling) signs from single images, with no temporal modeling. [8] achieved 95.7% accuracy with a ResNet-50/VGG-19 ensemble on the same 42-class static Kazakh dactyl alphabet. The methodologically closest prior work [7], which applies a BiLSTM over YOLO-extracted keypoints and optical flow, reported 92% accuracy on its largest set (47 isolated gesture classes, without optical flow).

Table 3 – comparison with related work

Study	Method	Task / classes	Dataset	Accuracy
Zholshiyeva et al. [9], 2023	MediaPipe + SVM	Static dactyl signs, 42 classes	4,200 images	99%
Amirgaliyev et al. [8], 2024	CNN ensemble (ResNet-50+VGG-19)	Static dactyl signs, 42 classes	57,708 images	95.7%
Buribayev et al. [7], 2025	YOLOv8-Pose + optical flow + BiLSTM	Isolated gestures, 47 classes	166,010 samples	92.0% (no flow) / 54.0% (with flow)
Present study	MediaPipe Holistic + BiLSTM	Isolated phrase classification, 40 classes	FluentSigners-5, 10,000 videos	94.73%

Accuracy fell to 54% once optical-flow features were added, which the authors attributed to additional noise at higher class counts. By comparison, the present model reaches 94.73% accuracy on 40 classes using the larger, signer-diverse FluentSigners-50 corpus [10], without a dedicated keypoint detector or an auxiliary motion stream. It should be noted that all three studies above address static or short isolated-sign classification, while the present work addresses phrase-level (multi-word) classification, a linguistically more complex unit.

Despite the strong performance, several limitations remain. The current framework focuses on isolated phrase classification and does not address continuous sign language recognition, where temporal segmentation and context modeling are required. In addition, the model relies primarily on hand and pose keypoints, without incorporating facial expressions or lip movements, which are important components of sign language communication. These factors limit the system’s ability to fully capture the multimodal nature of KSL.

It is also worth situating this system relative to the broader sign language translation literature. As formalized in [12], Sign Language Translation is an open-vocabulary sequence-to-sequence problem capable of producing previously unseen sentences. The present system addresses a complementary problem: high-precision recognition over a fixed, known phrase set, where each recognized phrase is mapped to its gloss and a verified reference translation. This closed-set design guarantees a correct reference translation for every recognized class, avoiding the mistranslation risk inherent to generative open-vocabulary output, which matters directly for an educational or assistive tool. Broadening coverage toward open-vocabulary sequence-to-sequence translation, as outlined in planned follow-up work, would extend the system's reach beyond this fixed phrase set while preserving the reliability the closed-set design provides for the phrases it already covers.

Overall, the proposed method demonstrates that keypoint-based spatiotemporal modeling provides a reliable and efficient solution for phrase-level sign language recognition, while also highlighting areas for further enhancement.

Conclusion

This paper presented a deep learning-based model for phrase-level recognition of Kazakh Sign Language using MediaPipe keypoints and bidirectional LSTM sequence modeling. The proposed system achieved 94.73% accuracy and a macro-average F1-score of 94.74% across 40 common KSL phrases using a compact architecture and a well-curated subset of the FluentSigners-50 dataset. By converting each video into normalized skeletal sequences, the model efficiently captured temporal gesture dynamics without the need for raw video or manual gloss annotations.

The findings confirm that keypoint-based representations provide a robust and computationally efficient alternative to traditional image-based methods. By reducing input dimensionality while preserving essential motion information, the approach enables stable training, strong generalization, and suitability for real-time applications. The integration of bidirectional temporal modeling further enhances performance by capturing contextual dependencies across gesture sequences.

Analysis of the experimental results reveals that the model performs consistently across most classes, with high precision and recall. Misclassifications are primarily associated with gestures that exhibit similar spatial and temporal trajectories. These observations provide valuable insights into the challenges of phrase-level sign language recognition and highlight directions for future research.

In conclusion, the proposed approach contributes to the advancement of sign language recognition for low-resource languages by demonstrating that efficient keypoint-based sequence modeling can achieve high accuracy in phrase-level tasks. The developed framework provides a foundation for the creation of scalable assistive technologies aimed at improving communication accessibility for the deaf and hard of hearing community.

References

1. Aitim A. A systematic review on translating Kazakh sign language to speech using dynamic gesture recognition / A. Aitim, D. Sattarkhuzhayeva, A. Khairullayeva // The Bulletin of KazATC. – 2025. – T. 138, № 3. – P. 122-133. https://doi.org/10.52167/1609-1817-2025-138-3-122-133.
2. Creating a Parallel Corpus for the Kazakh Sign Language and Learning / A. Yerimbetova et al // Applied Sciences. – 2025. – T. 15, № 5. – P. 2808. https://doi.org/10.3390/app15052808.
3. Advancements in Machine Learning Techniques for Hand Gesture-Based Sign Language Recognition: A Comprehensive Review / U. Rastogi et al // Archives of Computational Methods in Engineering. – 2025. – T. 32, № 7. – P. 4265-4302. https://doi.org/10.1007/s11831-025-10258-z.
4. Almjally A., Almukadi W.S. Deep computer vision with artificial intelligence based sign language recognition to assist hearing and speech-impaired individuals / A. Almjally, W.S. Almukadi // Scientific Reports. – 2025. – T. 15, № 1. – P. 32268. https://doi.org/10.1038/s41598-025-09106-8.
5. Deep learning pathways for automatic sign language processing / M. Toshpulatov et al // Pattern Recognition. – 2025. – T. 164. – P. 111475. https://doi.org/10.1016/j.patcog.2025.111475.
6. Luqman H. SignVLM: a pre-trained large video model for sign language recognition / H. Luqman // PeerJ Computer Science. – 2025. – T. 11. – P. e3112. https://doi.org/10.7717/peerj-cs.3112.
7. Enhancing Kazakh Sign Language Recognition with BiLSTM Using YOLO Keypoints and Optical Flow / Z. Buribayev et al // Applied Sciences. – 2025. – T. 15, № 10. – P. 5685. https://doi.org/10.3390/app15105685.
8. Application of neural networks ensemble method for the Kazakh sign language recognition / Y. Amirgaliyev et al // Bulletin of Electrical Engineering and Informatics. – 2024. – T. 13, № 5. – P. 3275-3287. https://doi.org/10.11591/eei.v13i5.7803.
9. Real-time Kazakh sign language recognition using Mediapipe and SVM / L. Zholshiyeva et al // Academic Scientific Journal of Computer Science. – 2023. – № 1. – P. 82-93. https://doi.org/10.32014/2023.2518-1726.170.
10. FluentSigners-50: A signer independent benchmark dataset for sign language processing / M. Mukushev et al // PloS ONE. – 2022. – T. 17, № 9. – P. e0273649. https://doi.org/10.1371/journal.pone.0273649.
11. Continuous sign language recognition and its translation into intonation-colored speech / N. Amangeldy et al // Sensors. – 2023. – T. 23, № 14. – P. 6383. https://doi.org/10.3390/s23146383.
12. Neural Sign Language Translation / N.C. Camgoz et al // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – Salt Lake City, USA, 2018. – P. 7784-7793. https://doi.org/10.1109/CVPR.2018.00812.

Ә. Жаканова[1], Ж. Оралбекова[1]*, Д. Оралбекова[2]

[1]Astana IT University, 010000, Республика Казахстан, Астана, проспект Мәңгілік Ел, 55/11, блок С1

[2]Институт информационных и вычислительных технологий,
050010, Республика Казахстан, Алматы, ул. Шевченко, 28

*e-mail: zh.oralbekova@astanait.edu.kz

РАЗРАБОТКА МОДЕЛИ КАЗАХСКОГО ЖЕСТОВОГО ЯЗЫКА ДЛЯ РАСПОЗНАВАНИЯ И ОБУЧЕНИЯ ЗАМКНУТЫХ ФРАЗ

Казахский язык жестов (KSL) остается недостаточно представленным в современных исследованиях по обработке языка жестов из-за отсутствия больших аннотированных наборов данных и специализированных вычислительных инструментов. В этой статье представлен основанный на глубоком обучении фреймворк для распознавания языка жестов на уровне фраз с использованием визуальных последовательностей ключевых точек, в котором каждая распознанная фраза сопоставляется с соответствующей и проверенным переводом на русский язык. Этот подход использует набор данных FluentSigners-50 в качестве эталонного корпуса и извлекает позы и ориентиры рук из видеоданных, используя целостную модель MediaPipe. Двухуровневая двунаправленная сеть долговременной кратковременной памяти (BiLSTM) обучается на основе полученных пространственно-временных последовательностей для классификации сорока различных фраз KSL. Модель принимает в качестве входных данных последовательность фиксированной длины из 150 кадров, каждый из которых содержит 225 ключевых объектов, и выводит предсказанный класс фраз через многослойный персептрон. Результаты экспериментов демонстрируют высокую эффективность в классификации фраз, с

точным сопоставлением видео с глосс-метками и соответствующими переводами на русский язык, точность которых достигает 94,73%. Вместо полноценного открытого машинного перевода система решает задачу классификации фраз закрытого множества с отображением gloss-метки и эталонного перевода, что делает её применимой как для вспомогательных, так и для обучающих приложений.

Ключевые слова: *казахский язык жестов, распознавание языка жестов, глубокое обучение, двунаправленный LSTM, MediaPipe Holistic, моделирование последовательности ключевых точек, классификация замкнутых фраз.*

Ә. Жақанова[1], Ж. Оралбекова[1]*, Д. Оралбекова[2]

[1]Astana IT University, 010000, Қазақстан Республикасы, Астана, Мәңгілік Ел даңғылы, 55/11, С1 блогы

[2]Ақпараттық және есептеуіш технологиялар институты,

050010, Қазақстан Республикасы, Алматы, Шевченко көшесі, 28

*e-mail: zh.oralbekova@astanait.edu.kz

ЖАБЫҚ СӨЗ ТІРКЕСТЕРІН ТАНУ ЖӘНЕ ОҚЫТУ ҮШІН ҚАЗАҚ ЫМДАУ ТІЛІНІҢ МОДЕЛІН ӘЗІРЛЕУ

Қазақ ымдау тілі (KSL) қазіргі ымдау тілдерін өңдеу саласындағы зерттеулерде жеткілікті деңгейде ұсынылмаған, бұл ең алдымен ауқымды таңбаланған деректер жиынтықтарының және арнайы есептеу құралдарының жетіспеушілігімен байланысты. Бұл мақалада визуалды кілттік нүктелер тізбегіне негізделген, ымдау тілін фразалық деңгейде тану және әрбір танылған фразаны оның глосс-белгісімен және эталондық аудармасымен сәйкестендіруге арналған терең оқытуға негізделген әдіс ұсынылады. Ұсынылған тәсіл эталондық корпус ретінде FluentSigners-50 деректер жиынтығын пайдаланады және бейнемәліметтерден поза мен қол қимылдарының белгілерін MediaPipe Holistic моделі арқылы бөліп алады. Алынған кеңістіктік-уақыттық тізбектер негізінде екі қабатты екібағытты ұзақ қысқа мерзімді жад (BiLSTM) нейрондық желісі оқытылып, қазақ ымдау тіліндегі қырық түрлі фразаны жіктеуді жүзеге асырады. Модель кіріс ретінде ұзындығы бекітілген 150 кадрдан тұратын тізбекті қабылдайды, әрбір кадр 225 кілттік нүкте белгілерін қамтиды, ал шығыс ретінде көпқабатты перцептрон арқылы болжанған фраза класын береді. Эксперименттік нәтижелер фразаларды жіктеу міндетінде жоғары тиімділікті көрсетеді: бейнеден глосс белгілеріне және олардың сәйкес орыс тіліндегі аудармаларына дәл салыстырулар 94,73% дәлдікке қол жеткізді. Толыққанды ашық сөздікті машиналық аударманың орнына жүйе жабық жиынтықтағы фразаларды жіктеу және глосс-белгі мен эталондық аударманы сәйкестендіру міндетін шешеді, бұл оны көмекші және оқыту қолданбаларында қолдануға жарамды етеді.

Түйін сөздер: *қазақ ымдау тілі; ымдау тілін тану; терең оқыту; екібағытты LSTM; MediaPipe Holistic; кілттік нүктелер тізбегін модельдеу; жабық сөз тіркестерінің жіктелуі.*

Сведения об авторах

Әсем Ерікқызы Жақанова – магистрант 1 курса, Astana IT University, Казахстан; e-mail: 255692@astanait.edu.kz. ORCID: https://orcid.org/0009-0000-2416-645X.

Жанар Орымбаевна Оралбекова* – PhD, ассоциированный профессор, и.о. профессора Школы ИИ и науки о данных, Astana IT University, Казахстан; e-mail: zh.oralbekova@astanait.edu.kz. ORCID: https://orcid.org/0000-0002-3726-8882.

Дина Орымбаевна Оралбекова – PhD, ассоциированный профессор, старший научный сотрудник Института информационных и вычислительных технологий, Казахстан; e-mail: dinaoral@mail.ru. ORCID: https://orcid.org/0000-0003-4975-6493.

Авторлар туралы мәліметтер

Әсем Ерікқызы Жақанова – 1 курс магистранты, Astana IT University, Қазақстан; e-mail: 255692@astanait.edu.kz. ORCID: https://orcid.org/0009-0000-2416-645X.

Жанар Орымбаевна Оралбекова* – PhD, қауымдастырылған профессор, Жасанды интеллект және деректер ғылымы мектебінің профессоры м.а., Astana IT University, Қазақстан; e-mail: zh.oralbekova@astanait.edu.kz. ORCID: https://orcid.org/0000-0002-3726-8882.

Дина Орымбаевна Оралбекова – PhD, қауымдастырылған профессор, Ақпараттық және есептеуіш технологиялар институтының аға ғылыми қызметкері, Қазақстан; e-mail: dinaoral@mail.ru. ORCID: https://orcid.org/0000-0003-4975-6493.

Information about the authors

Assem Zhakanova – first year master's student, Astana IT University, Kazakhstan; e-mail: 255692@astanait.edu.kz. ORCID: https://orcid.org/0009-0000-2416-645X.

Zhanar Oralbekova* – PhD, associate professor, acting Professor of the School of AI and Data Science, Astana IT University, Kazakhstan; e-mail: zh.oralbekova@astanait.edu.kz. ORCID: https://orcid.org/0000-0002-3726-8882.