DOI 10.54596/2958-0048-2026-1-263-273
UDK 004.85
IRSTI 28.23.15

KSL-EMO MULTIMODAL DATASET FOR EMOTION-AWARE KAZAKH SIGN LANGUAGE RECOGNITION

M. Kabykenov[1], M. Niyazbek[2], A. Zhumadillayeva[1,3*]
[1]Astana IT University, Astana, Kazakhstan*
[2]Xinjiang University, Urumqi, China
[3]L.N. Gumilyov Eurasian national university, Astana, Kazakhstan*
[*]Corresponding author: zhumadillayeva_ak@enu.kz

Abstract

Sign Language Recognition (SLR) is a main technology for bridging the communication gap between the deaf community and the hearing majority. While deep learning has advanced SLR significantly, low resource languages like Kazakh Sign Language (KSL) remain under explored due to the deficit of labeled data. In this paper, we address this limitation by establishing a novel benchmark for KSL, focusing on two distinct tasks: Isolated Sign Language Recognition (ISLR) and Emotion Recognition. We evaluate the performance of three state-of-the-art Vision Transformer architectures ViViT, VideoMAE V2, and TimeSformer on a custom collected dataset comprising 20 lexical gestures and 4 emotional states. Our experiments reveal that TimeSformer achieves superior performance, attaining a Top-1 Accuracy of 96.63% on lexical gestures and 80.87% on emotion recognition. Comparative analysis indicates that TimeSformer's "Divided Space-Time Attention" mechanism captures fine-grained spatiotemporal dynamics more effectively than the factorised encoder of ViViT or the masked modeling approach of VideoMAE.

Keywords: Sign Language Recognition, Kazakh Sign Language, Vision Transformers, Emotion Recognition.

ЭМОЦИЯҒА БЕЙІМДЕЛГЕН ҚАЗАҚ ЫМ-ИШАРА ТІЛІН ТАНУҒА АРНАЛҒАН KSL-EMO МУЛЬТИМОДАЛЬДЫ ДЕРЕКТЕР ЖИЫНТЫҒЫ

Кабыкенов М.[1], Ниязбек М.[2], Жумадиллаева А.К.[1,3*]
[1]Astana IT University, Астана, Қазақстан*
[2]Синьцзян университеті, Үрімші, Қытай
[3]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан*
[*]Хат-хабар үшін автор: zhumadillayeva_ak@enu.kz

Аңдатпа

Ым-ишара тілін тану (SLR) – саңырау қауымдастығы мен есітетін көпшілік арасындағы коммуникациялық алшақтықты азайтуға бағытталған маңызды технология. Терең оқыту әдістері SLR саласын айтарлықтай дамытқанына қарамастан, қазақ ым-ишара тілі (KSL) сияқты ресурсы аз тілдер таңбаланған деректердің жетіспеушілігіне байланысты жеткілікті зерттелмеген. Бұл жұмыста біз осы шектеуді еңсеру үшін KSL бойынша жаңа эталондық деректер жиынтығын ұсынамыз, ол екі негізгі міндетке бағытталған: оқшауланған ымдарды тану (ISLR) және эмоцияны тану. Біз арнайы жиналған, 20 лексикалық ым мен 4 эмоциялық күйден тұратын деректер жиынтығында Vision Transformer негізіндегі үш заманауи архитектураның – ViViT, VideoMAE V2 және TimeSformer – өнімділігін бағаладық. Эксперимент нәтижелері TimeSformer ең жоғары көрсеткіштерге қол жеткізгенін көрсетті: лексикалық ымдар үшін Top-1 дәлдігі 96,63%, ал эмоцияны тануда 80,87%. Салыстырмалы талдау TimeSformer-дің «кеңістік-уақыт бойынша бөлінген назар» механизмі ViViT-тің факторланған энкодеріне немесе VideoMAE-нің маскаланған модельдеу тәсіліне қарағанда кеңістік-уақыттық динамиканы дәлірек бейнелейтінін көрсетеді.

Кілт сөздер: ым-ишара тілін тану, қазақ ым-ишара тілі, Vision Transformer, эмоцияны тану.

МУЛЬТИМОДАЛЬНЫЙ ДАТАСЕТ KSL-EMO ДЛЯ РАСПОЗНАВАНИЯ КАЗАХСКОГО ЖЕСТОВОГО ЯЗЫКА С УЧЁТОМ ЭМОЦИЙ

Кабыкенов М.[1], Ниязбек М.[2], Жумадиллаева А.К.[1,3*]

[1]Astana IT University, Астана, Казахстан*

[2]Синьцзянский университет, Урумчи, Китай

[3]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан*

[*]Автор для корреспонденции: zhumadillayeva_ak@enu.kz

Аннотация

Распознавание жестового языка (SLR) является ключевой технологией для преодоления коммуникационного разрыва между сообществом глухих и слышащим большинством. Несмотря на значительный прогресс в области SLR благодаря глубокому обучению, малоресурсные языки, такие как казахский жестовый язык (KSL), остаются недостаточно изученными из-за нехватки размеченных данных. В данной работе мы решаем эту проблему, создавая новый эталонный набор данных для KSL, сосредоточенный на двух различных задачах: распознавание изолированных жестов (ISLR) и распознавание эмоций. Мы оцениваем производительность трёх современных архитектур Vision Transformer – ViViT, VideoMAE V2 и TimeSformer – на специально собранном наборе данных, включающем 20 лексических жестов и 4 эмоциональных состояния. Наши эксперименты показывают, что TimeSformer демонстрирует наилучшие результаты, достигая точности Top-1 96,63% для лексических жестов и 80,87% для распознавания эмоций. Сравнительный анализ показывает, что механизм «разделённого пространственно-временного внимания» TimeSformer более эффективно улавливает тонкую пространственно-временную динамику по сравнению с факторизованным энкодером ViViT или подходом маскированного моделирования VideoMAE.

Ключевые слова: распознавание жестового языка, казахский жестовый язык, Vision Transformer, распознавание эмоций.

Introduction

Sign Languages are main type of communication for people with hearing disorder. While Sign Language Recognition (SLR) has seen significant advancement, a big gap still exist, the majority of existing systems focused on manual gestures, often neglecting the non manual markers, specifically facial expressions and emotions. Ignoring these cues results in translations that lack nuance and intent. Moreover, current SLR research is heavily focused towards high resource languages such as American, Indian or Chinese Sign Languages. Kazakh Sign Language (KSL) is a low resource language, that lacks robust benchmarks and annotated datasets, hindering development of assistive technologies for the region. Training deep learning models on such limited data presents challenges related to overfitting and generalization. Traditionally, SLR relied on 3D Convolutional Neural Networks (3D-CNNs) and Recurrent Neural Networks (RNNs). However, recent breakthroughs in Vision Transformers (ViTs) offer superior capabilities in modeling long-range spatiotemporal dependencies. Architectures like VideoMAE (Masked Autoencoders) have demonstrated remarkable data efficiency, making them promising candidates for low-resource scenarios.

The main contributions of this work can be summarized as follows:

– Introducing a custom KSL dataset, manually annotated for two distinct tasks: Isolated Sign Language Recognition (ISLR) with 20 lexical gestures and Emotion Recognition with 4

affective states. To the best of our knowledge, this is one of the first datasets to address KSL in a dual stream context.

– We offer an extensive comparative examination of cutting-edge Vision Transformer architectures, such as ViViT, VideoMAE V2, and TimeSformer. We assess their compromises regarding accuracy, training efficiency, and computational expenses specifically for low resource sign language contexts

– We demonstrate through rigorous error analysis that while masked autoencoders are data efficient for macroscopic gestures, they struggle with fine grained emotion recognition. We show that TimeSformer divided space time attention mechanism offers superior robustness for capturing the subtle micro expressions required for affective computing.

Related works

A. Selecting a Template

Sign languages are multi channel communication systems. While manual components like hand shape and movement convey lexical meaning, non-manual markers such as facial expressions, head tilt and body posture provide critical grammatical and affective context (Ong & Ranganath, 2005). For instance, a raised eyebrow can transform a statement into a question, a phenomenon explicitly documented in Kazakh Russian Sign Language (KRSL) linguistics (Bertasius et al., 2021).

Despite their importance, the majority of SLR datasets and benchmarks, such as the widely used WLASL (Li et al., 2020), focus exclusively on manual gestures. Existing emotion recognition models are typically designed for hearing populations and may not generalize well to the exaggerated and distinct facial expressions found in sign language. This work addresses this gap by treating emotion recognition as a complementary stream to gesture recognition. Early approaches to SLR relied on handcrafted features and Hidden Markov Models (HMMs) (Starner et al., 1998). With the approach of deep learning, Convolutional Neural Networks (CNNs) became the standard. Researchers initially utilized 2D-CNNs for spatial feature extraction combined with RNNs or Long Short term Memories (LSTMs) for temporal modeling (Camgoz et al., 2020). Later, 3D-CNNs, such as I3D and C3D, emerged as a dominant method, capable of capturing spatiotemporal features in one (Tran et al., 2015; Carreira & Zisserman, 2017).

However, CNN-based methods are inherently limited by their local receptive fields, often struggling to model long-range temporal dependencies which are crucial for interpreting complex sign sentences (X. Wang et al., 2018). Recently, there has been a paradigm shift towards ViTs (Dosovitskiy et al., 2020), which leverage self-attention mechanisms to effectively capture global dependencies across space and time (Bertasius et al., 2021). While Transformer based architectures have demonstrated remarkable success in high resource languages like ASL or CSL (Camgoz et al., 2020; Hu et al., 2023), their application to low resource languages, such as KSL, remains significantly under explored.

The introduction of the ViT (Dosovitskiy et al., 2020) adapted the transformer architecture from NLP to image recognition. This success was rapidly extended to the video domain, leading to the three state-of-the-art architectures evaluated in this study. Bertasius et al. (2021) introduced TimeSformer, which addresses the high computational cost of 3D self attention. It proposes a Divided Space-Time Attention mechanism, where temporal attention and spatial attention are applied separately within each block. This factorization allows the model to learn spatiotemporal features efficiently, making it a strong alternative to traditional 3D-CNNs. Arnab et al. (2021) proposed ViViT, a pure-transformer architecture. It extracts

spatiotemporal tokens from input video "tubes" and processes them through a series of transformer layers. Specifically, its Factorised Encoder variant models spatial and temporal interactions sequentially. While highly accurate, ViViT is computationally intensive, serving as a robust baseline for complex gesture analysis in our study. Inspired by Devlin et al. (2019), Song et al. (2022) introduced VideoMAE, a self-supervised learner. By masking a high ratio of random video cubes and reconstructing the missing pixels, VideoMAE forces the encoder to learn profound semantic representations. This method is particularly effective for data efficient learning, making it ideally suited for low resource tasks like KSL recognition where labeled data is scarce.

Despite their theoretical strengths, these architectures exhibit distinct trade-offs when applied to SLR. In terms of performance on standard action recognition benchmarks like Kinetics-400, VideoMAE achieves results approximately 87% (L. Wang et al., 2023), generally outperforming TimeSformer (Bertasius et al., 2021) and ViViT (Arnab et al., 2021), which hover in the 80-82% range. However, this accuracy comes at a cost. ViViT offers high model capacity but incurs significant computational overhead, making it less ideal for real time deployment (Arnab et al., 2021). TimeSformer provides a balanced trade-off between speed and accuracy via its divided attention mechanism, though it may lose some fine-grained spatial correlation compared to full 3D attention (Bertasius et al., 2021). VideoMAE excels in data efficiency due to its masked pre training a crucial advantage for low-resource languages-but its reconstruction-based objective may not always align with discriminative tasks requiring fine-grained facial analysis (Song et al., 2022; L. Wang et al., 2023).

While these transformers have been extensively tested on high resource languages (Adaloglou et al., 2021) or general action recognition, their comparative effectiveness on low resource sign languages like KSL remains unexplored. Furthermore, existing benchmarks typically isolate manual gestures from non-manual markers (Rastgoo et al., 2020). There is currently no comprehensive study evaluating how these specific architectures handle the dual challenge of recognizing macroscopic lexical gestures and microscopic emotional expressions simultaneously. This work aims to bridge this gap by benchmarking these Vision Transformers on a novel, dual-stream KSL dataset.

Methodology

In this section, we detail the construction of our custom KSL dataset, the preprocessing pipeline, the Transformer architectures employed, and the experimental implementation details.

B. Dataset Collection and Splitting

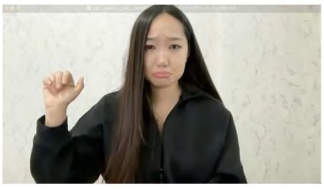
Due to the deficit of publicly available benchmarks for KSL, we constructed a novel, large scale dataset specifically for this study. The dataset was manually collected and annotated for ISLR, focusing on discrete lexical gestures and emotional states.

Crucially, the lexical vocabulary and execution standards for the gestures were strictly derived from the official educational guidelines provided by the National Scientific and Practical Center for the Development of Special and Inclusive Education (NSPCRSIE, 2024). This ensures that the recognized signs correspond to the standardized academic curriculum used in Kazakhstan.

The dataset consists of a total of 7,247 video samples, amounting to approximately 20,000 seconds of footage. It covers 20 distinct lexical gesture classes and 4 emotional states (Neutral, Happy, Sad, Angry). Visual examples of these pairings are presented in Figure 1. Specifically, Figure 1a displays the sign `mysyq' (cat) performed with a neutral expression, while Figure 1b depicts `saubol' (goodbye) combined with a sad emotion. High arousal emotions also represented, such as `apke' (sister) with happiness (Figure 1c) and `adam'

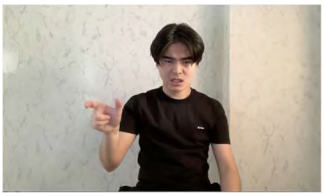
(human) with anger (Figure 1d). Data was collected from 9 distinct participants (7 male and 2 female) to ensure demographic diversity and robustness to signer variations.

(a) Sign: `mysyq' (cat) / Emotion: Neutral



(b) Sign: `saubol' (goodbye) / Emotion: Sad

(c) Sign: `apke' (sister) / Emotion: Happy



(d) Sign: `adam' (human) / Emotion: Angry

Figure 1. Representative samples from the collected KSL dataset illustrating the dual-stream nature of the data.

To carefully evaluate the generalization capability of the models, we employed a subject-independent splitting strategy. Instead of random shuffling, we separated participants into mutually exclusive sets. Training set contain 4,367 videos and testing set contain 2,880 videos. This strict separation ensures that the models are evaluated on "unseen" signers, simulating real world deployment scenarios where the system must recognize users it has never encountered during training.

C. Data Preprocessing

To make consistent input for the Transformer models, all video clips passed a unified preprocessing pipeline. All frames were resized to a spatial resolution of 224 x 224 pixels using bilinear interpolation. To establish a rigorous baseline, no data augmentation techniques, for example random cropping, flipping or color jittering were applied during training or inference. We adopted specific sampling strategies tailored to the architectural requirements of each model to balance performance and computational cost. For ViViT it's 32 frames, for both VideoMAE it's 16 frames and for TimeSformer it's 8 frames.

D. Model Architecture

We benchmark three state-of-the-art Vision Transformer architectures. To ensure robust feature initialization and facilitate faster convergence on our limited dataset, all models are initialized with weights pre-trained on the Kinetics-400 dataset.

We use a variant of ViViT, which uses a factorized encoder architecture. Unlike architectures that process space time markers together, this architecture divides model into spatial and temporal encoders. This is consistent with the principles of axial attention (Ho et al., 2019), which theoretically demonstrate that dividing multidimensional attention

significantly reduces computational complexity from quadratic to linear, allowing the model to process longer sequences with high accuracy while maintaining manageable memory usage.

To examine the trade-off between model capacity and computational efficiency, two distinct scales of the VideoMAE V2 architecture was evaluated. We utilize VideoMAE-Small as a lightweight variant suitable for resource limited edge deployment, and VideoMAE-Base to capture deeper semantic features. The masked modeling paradigm implemented in these architectures (Song et al., 2022) is particularly critical in our low resource setting, as it forces the network to learn robust spatiotemporal structures from limited data.

Finally, we benchmark the TimeSformer architecture, which represents a paradigm shift from traditional volumetric convolutions. By implementing a 'Divided Space-Time Attention' mechanism, the model decouples temporal and spatial feature extraction within each block. As highlighted in recent comparative surveys (Selva et al., 2023), this factorization significantly facilitate computational overhead linked with full 3D attention, offering a scalable solution for modeling long-range dependencies in video sequences.

E. Implementation Details

All models were implemented using the PyTorch framework and the Hugging Face Transformers library. The training process was runned on a dual-GPU setup consisting of two NVIDIA Tesla T4 GPUs.

We fine-tuned all architectures for a total of 10 epochs. Due to the significant differences in computational memory footprint among the models, we adjusted the batch sizes to maximize hardware utilization: a batch size of 8 was used for VideoMAE, 4 for TimeSformer, and 2 for ViViT. Optimization was performed using the AdamW optimizer with an initial learning rate of 5×10^{-5}, employing a linear learning rate scheduler to provide stable convergence.

Results

To provide a comprehensive analysis of the Vision Transformers capabilities in the context of KSL, we structured our evaluation into two distinct experimental streams. First, we benchmark the models on ISLR, focusing on the hand gestures which convey lexical meaning. Second, we evaluate the models on Emotion Recognition, focusing on facial expressions which provide essential affective context.

F. Isolated Sign Language Recognition

According Table I for gestures TimeSformer achieved highest performance across all metrics, reaching top-1 accuracy of 96.63% and macro f1-score of 0.9659. This suggests that the "Divided Space-Time Attention" mechanism is highly effective at capturing distinct spatiotemporal features of KSL. Since many KSL signs share similar hand configurations but differ significantly in their motion trajectories, the decoupled temporal attention likely allows the model to focus more granularly on these dynamic cues compared to the joint tokenization approaches.

Table 1. Comparsion of models on gesture.

Model	Accuracy	F1-Score	Training time
Vivit	91.60%	0.9173	18h 37m
Timesformer	96.63%	0.9659	6h 15m
VideoMAE-Small	91.60%	0.9153	8h 18m
VideoMAE-Base	94.51%	0.9447	9h 01m

VideoMAE-Base followed closely with an accuracy of 94.51%, proving the robustness of masked modeling. A particularly notable observation is that the lightweight VideoMAE-Small variant achieved parity with the significantly larger and computationally heavier ViViT model, both shows accuracy 91.6%. This highlights two critical findings, masked autoencoder pre-training strategy acts as a strong regularizer, allowing smaller models to learn robust representations even on limited datasets where larger "pure" Transformers like ViViT may struggle with optimization or overfitting. Increasing model parameters does not linearly translate to better accuracy for ISLR tasks. ViViT lower performance suggests that its complex Factorised Encoder might require fundamentally more training data to totally converge compared to the more efficient architectures of TimeSformer or VideoMAE.

Finally, the minimal difference between Accuracy and Macro F1-Scores across all architectures indicates that models maintain a high balance between precision and recall, effectively handling multi-class classification without significant bias toward specific frequent gestures.

G. Emotion Recognition

Recognizing emotions from facial expressions and body language proved to be a significantly more challenging task than lexical gesture recognition. As shown in Table II performance dropped across all models compared to the ISLR, with the best model TimeSformer dropping from 96.6% to 80.9%. TimeSformer again demonstrated superior performance with an accuracy of 80.87%, outperforming the VideoMAE-Base by over 6%. We attribute this performance gap to two primary factors. Lexical gestures rely on macroscopic body movements, which are robust to downsampling. In contrast, emotion recognition relies on microscopic facial cues. TimeSformer divided attention mechanism preserves these subtle temporal details more effectively than the competing architectures. VideoMAE employs a high masking ratio during pre-training. While effective for reconstructing global motion, this strategy appears detrimental for fine-grained emotion recognition. If the random mask occludes critical facial regions during training, the model struggles to learn the necessary discriminative features for affect recognition.

Table 2. Comparsion of models on emotion recogntion

Model	Accuracy	F1-Score	Training time
Timesformer	80.87%	0.8089	6h 52m
VideoMAE-Small	71.91%	0.7189	7h 34m
VideoMAE-Base	74.58%	0.7463	9h 07m

Furthermore, unlike the gesture task where VideoMAE-Small and Base performed identically, here we observe a clear degradation in the smaller model, 71.91% against 74.58%. This suggests that while lightweight models are sufficient for classifying distinct hand shapes, the nuances of affective computing require the deeper semantic capacity provided by the Base architecture.

H. Computational Efficiency

A crucial factor for real world deployment is the balance between accuracy and computational cost. ViViT was the most computationally expensive model, requiring 18.5 hours to train. Despite this massive resource consumption, it failed to outperform the much

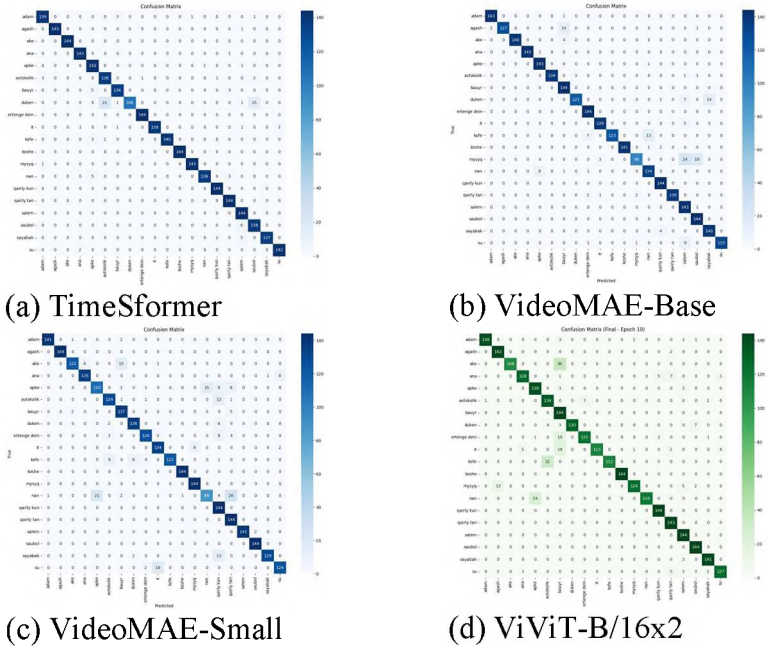
lighter VideoMAE-Small. This indicates that pure Transformer architectures with factorised encoders may be prone to overfitting or slow convergence on low-resource datasets like KSL.

TimeSformer was not only the most accurate but also the fastest to train, approximately 6 hours for both gesture and emotions. It achieved convergence 3 times faster than ViViT, making it the most viable candidate for practical applications.

I. Error Analysis

To analyze the distinctive failure modes of each architecture, we visualize the confusion matrices for each models in Figure 2 and Figure 3.

Despite achieving identical accuracy of ViViT and VideoMAE as 91.6%, the error patterns differ. ViViT (Figure 2d) struggled significantly with structurally similar signs, misclassifying 'ake' (father) as 'bauyr' (brother) in 36 instances and 'kofe' (coffee) as 'avtokolik' (car) 32 times. This indicates that the factorised encoder loses fine grained motion cues. VideoMAE-Small (Figure 2c) partially mitigated this, reducing the 'ake'/'bauyr' confusion to 15 instances, suggesting more robust spatial feature learning.



(a) TimeSformer (b) VideoMAE-Base

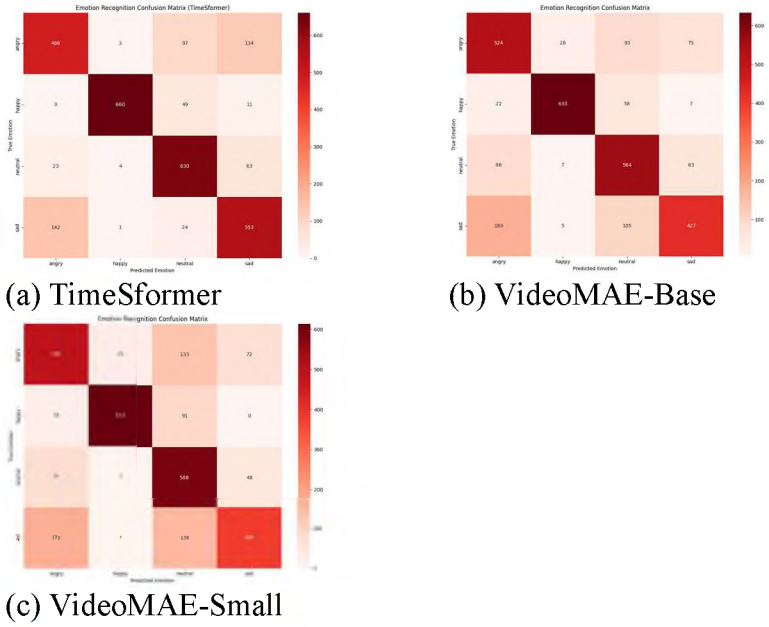
(c) VideoMAE-Small (d) ViViT-B/16x2

Figure 2. Confusion Matrices for ISLR (Gestures). TimeSformer (a) shows the cleanest diagonal, indicating superior class separation compared to ViViT (d).

TimeSformer demonstrated superior class separation. Notably, it completely resolved the specific 'ake'/'bauyr' ambiguity (0 errors) that plagued the other models. The only remaining confusion cluster was between 'duken' (shop) and 'su' (water) (15 errors). This confirms that the Divided Space-Time Attention mechanism effectively captures the subtle dynamic trajectories required to disambiguate complex lexical gestures.

All models achieved their highest precision on the 'Happy' class, for example TimeSformer (Figure 3a) correctly classified 660 samples, confirming that distinct facial deformations like smiling are robust to architectural differences. A significant ambiguity was observed between high-arousal (Angry) and low-arousal (Sad) negative emotions. VideoMAE-

Base (Figure 3b) struggled severely, misclassifying 'Sad' samples as 'Angry' in 183 instances. VideoMAE-Small (Figure 3c) showed the weakest performance, correctly identifying only 380 'Sad' samples. TimeSformer (Fig. 3a) minimized this error, reducing the 'Sad' to 'Angry' misclassification to 142 instances and achieving the highest true positive count for 'Sad' (553). This reinforces that temporal attention is critical for capturing the subtle micro-expressions, like brow lowering vs. lip tightening, that distinguish these complex negative affects.



(a) TimeSformer (b) VideoMAE-Base

(c) VideoMAE-Small

Figure 3. Confusion Matrices for Emotion Recognition.

Conclusion

In summary, this work established a robust benchmark for Kazakh Sign Language recognition by evaluating distinct Vision Transformer architectures on a novel dual-stream dataset. Our experiments demonstrated that the TimeSformer architecture shows the superior balance of accuracy and computational efficiency, achieving a Top-1 Accuracy of 96.63% on lexical gestures and 80.87% on emotion recognition. While masked autoencoders like VideoMAE proved highly data-efficient for macroscopic sign detection, our error analysis revealed their limitations in capturing fine-grained micro-expressions, particularly when distinguishing between high-arousal and low-arousal negative affects compared to the Divided Space-Time Attention mechanism.

Future work will focus on scaling the complexity and variety of the dataset to bridge the gap toward real world deployment. We plan to significantly expand the lexicon size beyond the current set of classes and increase the number of participants to capture wider range of inter signer variability. These enhancements will ensure a more rigorous testing ground for developing deeper, more generalized Transformer models capable of handling the nuances of natural communication in KSL.

Acknowledgements

This research has been funded by the SR-LAB-202504 Collaborative Innovation Seed Fund of Silk Road Multilingual Cognitive Computing International Cooperation Joint

Laboratory, "Key Technologies for Terminology Extraction and Multidirectorial Translation in Multilingual Information Technology and Biomedical Fields in Central Asia", 2025-2027.

References:

1. Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G. T., Zacharopoulou, V., Xydopoulos, G. J., Atzakas, K., Papazachariou, D., & Daras, P. (2021). A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition. *IEEE Transactions on Multimedia*, 24, 1750-1762. https://doi.org/10.1109/tmm.2021.3070438
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. *In 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 6816-6826). IEEE. https://doi.org/10.1109/ICCV48922.2021.00676
3. Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention all you need for video understanding? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2102.05095
4. Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020, March 30). Sign Language Transformers: joint end-to-end sign language recognition and translation. *arXiv.org*. https://arxiv.org/abs/2003.13830
5. Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4724-4733. https://doi.org/10.1109/cvpr.2017.502
6. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv.org*. https://arxiv.org/abs/2010.11929
8. Ho, J., Kalchbrenner, N., Weissenborn, D., & Salimans, T. (2019). Axial attention in multidimensional transformers. *arXiv.org*. https://arxiv.org/abs/1912.12180
9. Hu, H., Zhao, W., Zhou, W., & Li, H. (2023). SignBERT+: Hand-Model-Aware Self-Supervised Pre-Training for Sign Language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9)*, 11221-11239. https://doi.org/10.1109/tpami.2023.3269220
10. Kimmelman V, Imashev A, Mukushev M, Sandygulova A (2020) Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study. PLOS ONE 15(6): e0233731. https://doi.org/10.1371/journal.pone.0233731
11. Koller, O., Zargaran, S., & Ney, H. (2017). Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3416-3424. https://doi.org/10.1109/cvpr.2017.364
12. Li, D., Opazo, C. R., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: a new large-scale dataset and methods comparison. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1448-1458. https://doi.org/10.1109/wacv45572.2020.9093512
13. National Scientific and Practical Center for the Development of Special and Inclusive Education. (2024). *Methodological guidelines for Kazakh sign language*. https://special-edu.kz/kz/news/6/single/961
14. Ong, E. J., & Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(6), 873-891.
15. Rastgoo, R., Kiani, K., & Escalera, S. (2020). Sign Language Recognition: A deep survey. *Expert Systems With Applications*, 164, 113794. https://doi.org/10.1016/j.eswa.2020.113794
16. Selva, J., Johansen, A. S., Escalera, S., Nasrollahi, K., Moeslund, T. B., & Clapés, A. (2023). Video Transformers: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11), 12922-12943. https://doi.org/10.1109/tpami.2023.3243465
17. Starner, T., Weaver, J., & Pentland, A. (1998). Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12)*, 1371-1375. https://doi.org/10.1109/34.735811

18. Song, Y., Tong, Z., Wang, J., & Wang, L. (2022). VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. *Neural Information Processing Systems Foundation, Inc. (NeurIPS)*, 10078-10093. https://doi.org/10.52202/068431-0732
19. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. *IEEE International Conference on Computer Vision (ICCV)*, 4489-4497. https://doi.org/10.1109/iccv.2015.510
20. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., & Qiao, Y. (2023). VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14549-14560. https://doi.org/10.1109/cvpr52729.2023.01398
21. Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7794-7803).

Information about the authors

M.F. Kabykenov – Master student of Astana IT University, Astana, Kazakhstan; e-mail: 242890@astanait.edu.kz;
M. Niyazbek – Associate professor, PhD, Xinjiang University, Urumqi, China; e-mail: muheyatn@xju.edu.cn;
A.K. Zhumadillayeva – corresponding author, Associate professor of the Department of Computer and Software engineering, candidate technical sciences, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: zhumadillayeva_ak@enu.kz.