CROSS-LANGUAGE TRANSFER LEARNING USING VISUAL INFORMATION FOR AUTOMATIC SIGN GESTURE RECOGNITION

D. Ryumin *, D. Ivanko, A. Axyonov

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russian Federation –
ryumin.d@iias.spb.su, denis.ivanko11@gmail.com, axyonov.a@iias.spb.su

Commission II, WG II/8

KEY WORDS: Cross-Language, Transfer Learning, Sign Language Recognition, Assistive Technologies, Intelligent Video Analytics, Computer Vision, Human-Computer Interaction.

ABSTRACT:

Automatic sign gesture recognition (GR) plays a critical role in facilitating communication between hearing-impaired individuals and the rest of society. However, recognizing sign gestures accurately and efficiently remains a challenging task due to the diversity of sign languages (SLs) and their limited availability of labeled data. This scientific paper proposes a new approach to improving the accuracy of automatic sign GR using cross-language transfer learning with visual information. Two large-scale multimodal SL corpora are utilized as the basic SLs for this study: the Ankara University Turkish Sign Language Dataset (AUTSL) and the Thesaurus Russian Sign Language (TheRusLan). Experimental studies were conducted, resulting in an accuracy of 93.33% for 18 different gestures, including the Russian target SL gestures. This result exceeds the previous state-of-the-art accuracy by 2.19%, demonstrating the effectiveness of the proposed approach. The study highlights the potential of the proposed approach to enhance the accuracy and robustness of machine SL translation, improve the naturalness of human-computer interaction, and facilitate the social adaptation of people with hearing impairments. This paper proposes a promising direction for future research to explore the application of the proposed approach to other SLs and to investigate the impact of individual and cultural differences on GR.

1. INTRODUCTION

Human-computer interaction (HCI) has been an area of growing scientific attention in the last few decades (Guo et al., 2021). Recent advances in artificial intelligence (AI), information technologies, and cognitive sciences have been the main drivers behind the growth of HCI (Ahmed et al., 2022). This interdisciplinary interaction allows for the design of complex information and technical spaces, such as digital systems and platforms, which can process information of various modalities more efficiently (Axyonov et al., 2021, Dresvyanskiy et al., 2022).

It can be asserted that the current global trend in modern society is the development of machine learning (ML) and AI technologies to enable effective, natural, and universal HCI. For instance, through the use of visual communication modalities, such as hand gestures (HGs), individuals can interact with intelligent information systems at a distance and in noisy environments where acoustic speech may be ineffective (Ryumin et al., 2020).

In addition, the total number of people suffering from complete deafness or hearing problems is increasing every year. That is why full-fledged automatic machine sign language (SL) translation systems are needed, which currently do not exist. This is due to both a number of technical factors (e.g., presence of visual noise, occlusions, illumination variations), insufficient syntax and semantic description of sign languages (SLs), the absence of a sufficient number of large-scale SLs corpora suitable for model training, the imbalance of available corpora by subject areas (Ryumina and Karpov, 2020), as well as a number of other factors, that are related directly to humans. Interdisciplinary scientific research in the fields of gender linguistics (Carli et al., 1995), non-verbal semiotics (Iriskhanova and Cienki, 2018) and psychology (Carro et al., 2015) indicate that the gender and age characteristics of a single person can affect the size of the palms, the distance of the hands from the body, the distance between active and passive hand, and the speed of representation of various gestures. Along with this, it is well-known that deaf people often accompany HGs with silent articulation of the lips (Rajalakshmi et al., 2023, Ryumin et al., 2023). Therefore, automatic recognition of human gestures is a very actual and complex fundamental and technical task.

In this paper presents a cross-language transfer learning approach that uses visual information to improve the recognition accuracy of a target SL. The main idea of the approach is to combine data from different basic SLs to train preliminary neural network (NN) models, which are then adapted to the target SL.

The paper is structured as follows: Section 2. provides a review of related scientific work and focuses on modern methods for automatic gesture recognition (GR) in SLs. Section 3. discusses the corpora used for training, validation, and testing. Section 4. presents a cross-language transfer learning SL approach. Section 5. includes a comparison of the proposed approach with other state-of-the-art (SOTA) solutions. Section 6. presents the conclusions, highlighting the main findings and outlining future research directions.

2. RELATED WORK

Significant advances have been made in the field of GR in recent years, with the use of deep learning architectures such as convolutional neural networks (CNNs), deep belief networks (DBNs), long short-term memory networks (LSTMs), gated recurrent units (GRUs), graph NNs (GNNs), attention-based networks, capsule networks, and transformer-based networks showing promising results in analyzing spatio-temporal features of gestures. These architectures are particularly effective in analyzing the complex

*Corresponding author

Publisher's note: Copernicus Publications has not received any payments from Russian or Belarusian institutions for this paper.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W3-2023-209-2023 | © Author(s) 2023. CC BY 4.0 License.

spatio-temporal features of gestures, as they can capture both the static and dynamic features of gestures. In addition to deep learning techniques, hybrid approaches, such as combining deep learning with rule-based approaches and fuzzy logic, have been proposed to improve the accuracy and interpretability of GR systems.

The papers discussed in this section introduce various SOTA approaches for enhancing the performance measure of SL recognition.

In a series of papers by the authors, two computer systems were proposed for recognizing manual gestures. The first paper (Ryumin and Karpov, 2017) proposed a prototype system for recognizing continuous fingerspelling gestures and digit sequences in Russian and Kazakh SLs. The system used the Kinect version 2.0 sensor to capture visual information and had a gesture vocabulary of 52 fingerspelling gestures. The authors collected a visual corpus of SL gestures recorded with Kinect version 2.0 to train and test the recognition system. In contrast, the second paper (Ryumin et al., 2019b) proposed an approach for detecting and recognizing 3D one-handed gestures for HCI. The paper described the logical structure for recording a gestural corpus and presented models of deep convolutional networks for detecting faces and hand shapes. Additionally, the paper provided results of automatic detection of the regions with the face and the shape of the hand and suggested that this approach could be used in tasks such as biometrics, computer vision (CV), ML, automatic systems of face recognition, and SLs.

The paper (Camgöz et al., 2020) focused on a transformer-based architecture for joint continuous SL recognition and translation without the need for ground-truth timing information. The paper achieved SOTA results on the RWTH-PHOENIX-Weather-2014T (Koller et al., 2015) corpus and new baseline results for several text-to-text SL translation tasks using transformer networks. In another paper (Hu et al., 2021), the authors proposed pre-training the bidirectional encoder representations from Transformers (BERT) architecture on a large-scale SL video corpus to improve the performance of SL recognition models. The authors introduced a pre-trained model, SignBERT, and fine-tuned it on smaller SL corpora. The paper emphasizes the significance of pre-training and transfer learning for enhancing the accuracy of SL recognition models. Another paper (Jiang et al., 2021) proposed a skeleton aware multimodal SL recognition framework (SAM-SLR) that combined RGB, depth, and skeleton information to achieve SOTA performance in SL recognition. The framework included an SL graph convolution network (SL-GCN) and a separable spatio-temporal convolution network (SSTCN) to model the embedded dynamics and exploit skeleton features, respectively. Another study (Selvaraj et al., 2021) explored a pose-based pre-trained model for cross-lingual SL recognition was proposed, which employed a pose estimation model to extract skeletal features from SL videos and then used a pre-trained transformer model for sequence modeling. The paper achieved SOTA results on various SL corpora, including Indian, American, and Brazilian SLs. In (Coster et al., 2021), a multimodal approach to isolated sign recognition using the video transformer network (VTN) architecture was proposed, which involved pre-extracting information from SL videos to capture body movement and hand shapes. The authors evaluated their approach on an unnamed corpus and achieved significantly higher accuracy compared to the VTN architecture without hand crops and pose flow. In (Saunders et al., 2021), a progressive transformer architecture for end-to-end translation from spoken language sentences to continuous 3D multi-channel sign pose sequences was presented. The authors introduced counter decoding, data augmentation techniques, and an adversarial training regime to produce realistic and expressive sign pose sequences. Also in this paper presented benchmark quantitative results on the PHOENIX14T (Camgoz et al., 2018) corpus and a user evaluation of the SL production model.

In paper (Song and Xiang, 2022), introduced the SL graph time transformer (SLGTformer), which is an approach that utilizes decoupled graph and temporal self-attention with graph relative positional encodings to guide spatial self-attention. This approach achieved SOTA performance on the WLASL (Li et al., 2019) corpus. In another paper (Boháek and Hrúz, 2022), the authors proposed a word-level SL recognition system based on the transformer model, using 2D landmark locations to estimate human body pose. The authors of the paper also introduced a robust pose normalization scheme and several augmentations of the body pose to improve accuracy. Another paper (Amangeldy et al., 2022) described an automatic SL interpretation system based on GR technology, emphasizing the importance of such systems for people with hearing impairments. The authors introduced a palm definition model and linear models to recognize the shapes of numbers and letters in Kazakh SL, achieving an advantage in fully recognizing letters. In yet another paper (Rajalakshmi et al., 2022), the authors proposed a hybrid NN approach for both static and dynamic isolated SL recognition in Indian and Russian SLs, extracting spatio-temporal features and combining them using a NN architecture. In a different paper (Ryumin et al., 2023), the authors introduced a deep NN-based model architecture for GR that included a unique set of spatio-temporal features, including lip articulation information, achieving high accuracy on the Ankara University Turkish Sign Language (AUTSL) (Sincan and Keles, 2020) corpus. In another paper (Boháek and Hrúz, 2023), the authors proposed a few-shot learning approach that used online text-to-video dictionaries to train NN models and achieved SOTA results on multiple SL corpus. Additionally, in another paper (Novopoltsev et al., 2023), the authors investigated fine-tuning on corpora from other SLs to improve recognition quality and whether real-time sign recognition without graphics processing unit was possible, achieving promising results on three different language corpora. Finally, in a separate paper (Rajalakshmi et al., 2023), the authors proposed a novel vision-based hybrid deep NN methodology for recognizing Indian and Russian sign gestures, which used a 3D deep NN with atrous convolutions for spatial feature extraction, attention-based Bi-LSTM for temporal and sequential feature extraction, modified autoencoders for distinguished abstract feature extraction, and a hybrid attention module for discriminative feature extraction. This methodology yielded better results than other SOTA frameworks on a novel multi-signer Indo-Russian SL corpus.

All of the research studies mentioned in this section share the common goal of developing effective approaches for the automatic recognition of SL by analyzing human body movements. However, it is worth noting that separating the digital scene (visual information) from the dynamic behavior of an individual, including SL, remains a challenging task. Currently, there are no fully automatic NN models or ML approaches for SL recognition systems. Developing such comprehensive NN models requires extensive intellectual analysis and improvements in feature extraction techniques, not only for spatial but also for temporal features, from localized regions of an individual.

3. RESEARCH CORPORA

In the field of GR, having a vast and diverse amount of data is crucial to training NN models that can perform well in cross-language transfer learning. However, obtaining such corpora can be challenging, especially for SL recognition. This is because

SLs are complex and require capturing both spatial and temporal information about the signer’s movements. Moreover, SLs have a significant variety of dialects and accents (Karpov and Železnỳ, 2012, Li et al., 2020, Kagirov et al., 2020b), making it challenging to capture the full range of SL gestures, even within a specific SL.

Therefore, collecting and annotating large-scale SL corpora that provide a broad range of gestures, expressions, and dialects is necessary. Additionally, gathering data from signers of different ages, genders, and cultural backgrounds can contribute to creating a more comprehensive corpus. Technological advancements in recent years have made it possible to capture SL data more efficiently, using tools such as motion capture sensors, depth cameras, and smartphones. However, annotating SL data is still a time-consuming process that requires specialized knowledge and skills.

Presently, only a few large-scale corpora (Li et al., 2019, Sincan and Keles, 2020, Kagirov et al., 2020a) are available that contain a significant amount of gesture data, particularly SL data, which is crucial for pre-training NN models for cross-language transfer learning.

In this study, we used two large-scale multimodal SL corpora as the basic SLs, namely: Turkish SL (the Ankara University Turkish Sign Language Dataset (AUTSL (Sincan and Keles, 2020), publicly available) and the Russian SL (the Thesaurus Russian Sign Language (TheRusLan (Kagirov et al., 2020a), available upon request). The hull data was chosen because it contains multimodal data recorded using the Kinect version 2.0 sensor (although the version 2.0 of the Kinect sensor is no longer available at the moment). The target SL is Russian SL, which includes various HGs, as well as some gestures (Ryumin et al., 2019a) specifically designed for interaction with the prototype of an auxiliary mobile information robot in a noisy acoustic environment or for use by people with certain speech disorders or deaf individuals. The prototype of the mobile information robot-assistant is a mobile autonomous robotic platform that consists of a basic device for moving small-sized cargoes of compact dimensions and additional actuators for intelligent visual analysis (Ryumin et al., 2020). Also it is important to note that there can be multiple basic and target languages.

3.1 AUTSL corpus

The AUTSL (Sincan and Keles, 2020) is a large-scale multimodal corpus of Turkish SL recordings, created by researchers at Ankara University in Turkey to facilitate research in SL recognition and understanding.

The corpus comprises video recordings and depth data captured using a Kinect version 2.0 sensor. The video recordings show the signer’s entire body or upper body and hands, while the depth data provides 3D spatial information about their hand and body movements. It includes 226 different gestures demonstrated by 43 signers, with a total of 38 336 video examples of gestures. As can be seen from Figure 1, the videos feature various dynamic backgrounds, indicating recordings in the wild conditions.

The corpus features a variety of signers of different ages (from 19 to 50 years), genders (10 male and 33 female), and signing experience levels, resulting in a diverse range of SL gestures and expressions. Furthermore, the corpus is annotated using glosses and gloss-aligned translations, making it suitable for SL recognition and machine translation research.

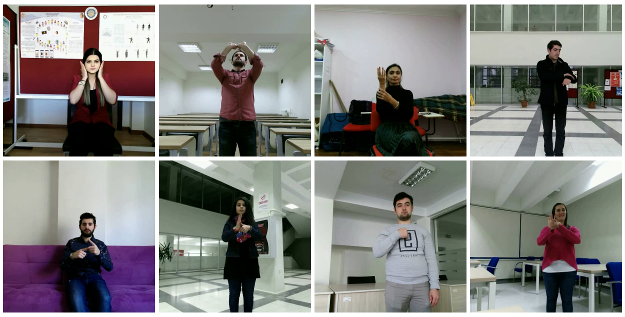

Figure 1: Examples of video frames demonstrating Turkish SL gestures in RGB format from the multimodal corpus AUTSL.

3.2 TheRusLan corpus

The TheRusLan (Kagirov et al., 2020a) is a multimodal corpus of Russian SL elements created by researchers from the Speech and Multimodal Interfaces Laboratory at the St. Petersburg Federal Research Center of the Russian Academy of Sciences. As depicted in Figure 2, the TheRusLan (Kagirov et al., 2020a) corpus comprises video recordings of Russian SL gestures in RGB format, depth map mode, and infrared, making it the only large-scale multimodal resource of its kind for Russian SL.

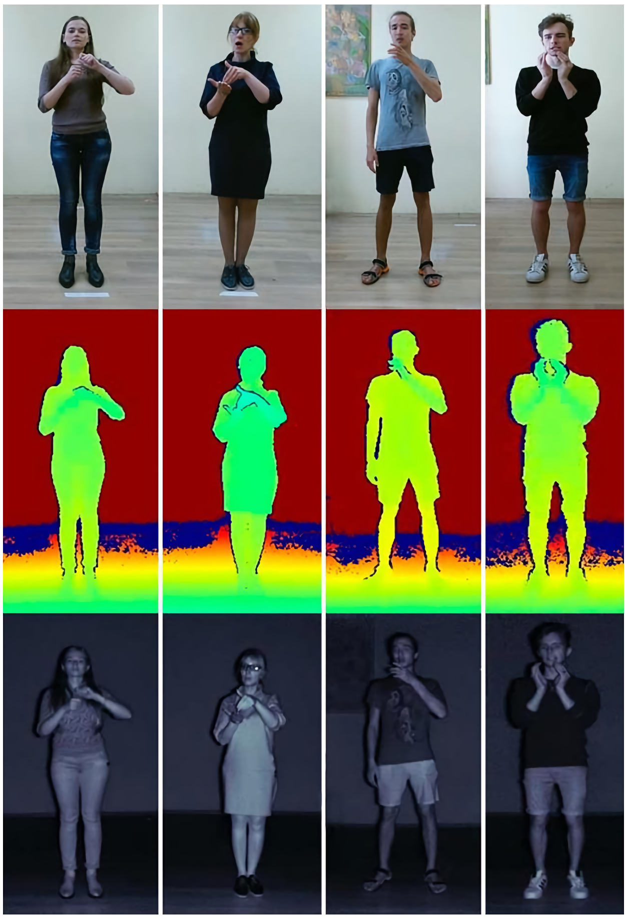

Figure 2: Examples of video frames demonstrating Russian SL gestures in RGB format (top row), depth map mode (middle row), and infrared range (bottom row) from the multimodal corpus TheRuSLan.

The TheRusLan is a multimodal corpus of 13 signers recorded using a Kinect sensor version 2.0 during their studies at a spe-

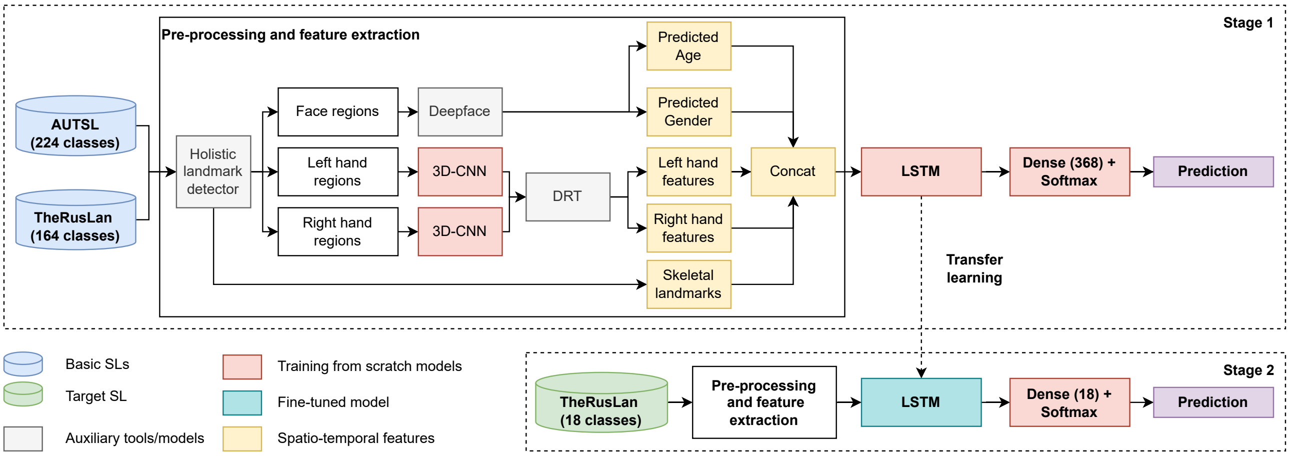

Figure 3: Pipeline of the proposed a cross-language transfer learning approach using visual information for automatic sign gesture recognition.

cial center for people with hearing impairments. The participants included 11 deaf students (two men and nine women) and two teachers. Most of the SL students had only passive knowledge of the standard norm of Russian SL, so translations were standardized by experts from the center, including teachers. The corpus focuses on gestures related to “interaction with a robotic cart while shopping in a supermarket”, and contains recordings of 164 individual words and phrases from Russian SL. Each signer repeated each gesture at least five times, resulting in 10 660 video examples of gestures. The recordings were made at a distance of 1.5-2 meters from the signers, with an average video length of approximately 36 minutes per signer.

4. PROPOSED APPROACH

The pipeline of the proposed a cross-language transfer learning approach using visual information for automatic sign GR is shown in Figure 3.

Our approach divides into two main stages, as shown in Figure 3. In the first stage, we train a model to recognize gestures on the basic corpora: Turkish (AUTSL) and Russian (TheRusLan) sign languages. In the second stage, we fine-tune the pre-trained model on the target corpus, specifically the shorted the TheRusLan corpus, by transferring knowledge from the weights of the pre-trained model. Next, we provide a detailed description of the data pre-processing and feature extraction steps used in our approach.

The initial step in our proposed approach is to precisely locate the skeletal and facial landmarks in the input 2D video. This step is critical, as it forms the basis for subsequent processes such as gender and age determination, feature extraction, and GR. To achieve this, we utilize the SOTA holistic landmark detector from the MediaPipe open-source framework (Bazarevsky et al., 2020). This detector employs multiple NN models that work in real-time to determine 543 2D landmarks of a person’s face, hands, and body. Once the landmarks are accurately localized, we extract the 2D graphic regions of the signer’s hands and face (as shown in Figure 4) to isolate regions of interest and extract the most informative features for the GR task. Thus, the accuracy and reliability of this initial step is critical for the success of our proposed approach.

After localizing the skeletal and facial landmarks in the input 2D frames, the next step is to determine the gender and age of the signer. Gender and age are important characteristics that can provide valuable visual information for GR and interpretation. To accomplish this task, we use a pre-trained NN from the Deepface open-source software library (Serengil and Ozpinar, 2021). This network is specifically designed to extract facial features that can accurately predict the gender and age of a person. It analyzes the localized regions of the face obtained in the previous step and extracts the appropriate visual features needed for gender and age determination. Estimating a person’s age from a single video frame is a complex task. However, the Deepface NN utilizes SOTA techniques such as hyperparameter tuning and deep learning to accurately estimate the age of a person. This NN has been trained on a large corpus of facial images, allowing it to generalize well to new and unfamiliar faces. The estimated age of the signer, ranging from 1 to 100 years, provides valuable visual information for understanding their behavior and interpreting their gestures. For example, different age groups may use different gestures or have different meanings associated with their gestures.

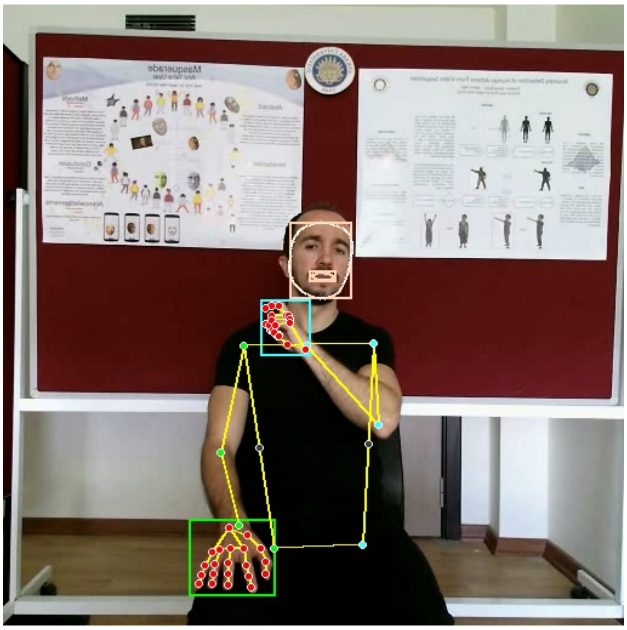

Figure 4: Example of a video frame showing the 2D graphics regions of the signer’s hands and face.

The next step in the proposed approach involves extracting NN features from both hands using a trained 3D-CNN model. One advantage of using 3D-CNNs is that they can process video data directly, making them well-suited for analyzing temporal features. In contrast, 2D-CNNs require multiple frames to be processed independently, which can lead to a loss of temporal information. Another advantage of 3D-CNNs is that they can learn more complex spatio-temporal features in the visual data, which is especially important for tasks such as GR or SL recognition. While the Video Swin Transformer (Liu et al., 2022) has shown promise in recognizing HGs, we found that the sizes of our selected research corpora is insufficient to convergence of this transformer. Therefore, we use the 3D-CNN model (Axyonov et al., 2022a), which has already demonstrated promising results in recognizing and interpreting SL gestures. By leveraging the temporal and spatial features of gestures, our approach can accurately recognize a wide range of SL gestures, even in challenging settings.

The next important step in our proposed approach is the use of a dimensionality reduction technique (DRT), which allows extracting the most informative features from the hand regions. This technique reduces the number of deep features of the hands while retaining the most important information, thus improving the accuracy and efficiency of our approach. There are several DRTs available, such as principal component analysis (PCA) (Abdi and Williams, 2010), linear discriminant analysis (LDA) (Izenman, 2013), and t-distributed stochastic neighbor embedding (t-SNE) (Belkina et al., 2019), among others. In our previous research (Ryumin et al., 2023), we demonstrated that LDA is more effective than other techniques for this task. Therefore, we also use it into our novel approach for this paper.

Thus, five types of spatio-temporal features are formed, including (1) gender, (2) age, (3) features of the left hand, (4) features of the right hand, and (5) normalized skeletal landmarks. These features are combined into one vector, which is normalized using Z-normalization and fed to the GR model. In this study, an LSTM model with an attention layer is used as the GR model. This is because attention mechanisms have been shown to be effective in improving the performance of sequence-based models by allowing them to focus on relevant parts of the input sequence (Ryumina et al., 2022). Our LSTM model consists of two LSTM layers of 64 and 32 neurons with an attention layer (Yang et al., 2016) in between. By using an attention layer in our LSTM model, we can enhance the model's ability to distinguish between different gestures and extract more informative features from the spatio-temporal data. The resulting model can better capture the nuances and complexities of SL gestures, leading to more accurate recognition. The basic NN model is designed to recognize a total of 388 gestures, including 224 Turkish and 164 Russian gestures, and is completed by a fully connected layer.

All these described steps allow combining data from the basic and target SLs with a total of 388 gestures for preliminary training of the basic NN model. Then, the weights of pre-trained NN model are used to re-train the target SL, which includes only 18 Russian gestures from the TheRusLan corpus.

Unlike the first stage, in the second stage training models can lead to their re-training. Therefore, we have applied several techniques to prevent this. Firstly, we utilized the MixUp (Zhang et al., 2017, Liang et al., 2018) data augmentation process to minimize the possible risk of re-training the NN model to extract the most informative features from the hand regions. MixUp is a powerful data augmentation technique that creates virtual training examples by interpolating between pairs of real examples, which encourages the model to learn more robust and generalizable features. Secondly, we used warm restarts with cosine annealing as the learning rate scheduler for all NN models (Axyonov et al., 2022b). Warm restarts are an adaptive learning rate schedule that periodically resets the learning rate to a higher value and then decays it gradually, which helps the model to escape from local minima and find better solutions. Cosine annealing further improves this approach by introducing a cosine-shaped decay pattern, which smoothens the learning rate schedule and improves the model's convergence. By using these techniques, we can effectively train the NN models to recognize the target SL gestures with high accuracy.

5. EXPERIMENTAL RESULTS

The basic NN model was trained on 200 epochs, with the SGD optimizer at a learning rate of 0.001. We used recognition rate r as a performance measure of model for SLs recognition. Recognition rate is calculated as:

$$r = \frac{1}{N}\sum_{i=1}^{N} f(p_i, t_i), \tag{1}$$

$$f(p_i, t_i) = \begin{cases} 1, & if\ p_i = t_i, \\ 0, & else, \end{cases} \tag{2}$$

where N is the total number of samples, p_i is the predicted label for the i^{th} sample, t_i is the true label for the i^{th} sample.

The recognition rates of the basic model trained on the research corpora in two SLs are presented in Table 1.

Corpus	Number of gestures	Recognition rate, %
AUTSL	224	93.38
TheRusLan	164	66.34

Table 1: Recognition rate results of the first stage of the proposed approach in the context of the research corpora.

As shown in Table 1, the recognition rate for gesture recognition on the AUTSL corpus outperforms the recognition rate for gesture recognition on the TheRusLan corpus by 27.04%. This can be explained by the fact that the training set of the AUTSL corpus contains more examples for each gesture demonstrated by signers compared to the TheRusLan corpus. The results of comparing the recognition rates of our approach with the SOTA for the AUTSL corpus are presented in Table 2.

Approach	Recognition rate, %
(Sincan and Keles, 2020)	49.22
(Coster et al., 2021)	92.92
	96.15
(Sincan et al., 2021)	96.55
	97.62
(Jiang et al., 2021)	98.42
(Ryumin et al., 2023)	**98.56**
Our	93.38

Table 2: Comparison of the recognition rate results of our approach with SOTA on the AUTSL corpus.

Table 2 shows that, to date, we achieved results using our approach on-par with SOTA approaches on the AUTSL corpus. For the TheRusLan corpus with 164 gestures, we provide the first baseline result.

The target NN model was trained for 200 epochs using the Adam optimizer and a learning rate that decreased from 0.0001 to 0.00001 with 5 restart cycles. The learning rate scheduler used was cosine annealing. The recognition rates of the target model trained on the TheRusLan corpus with 18 gestures are presented in Table 3. The experimental results demonstrate that our approach using transfer learning achieved an absolute increase of 2.59% in recognition rate (93.33% vs. 90.74%) compared to our approach without the first stage.

Transfer leaning	Recognition rate, %
Without	90.74
With	93.33

Table 3: Recognition rate results of the second stage of our proposed approach on the TheRusLan corpus with 18 gestures.

The results of comparing the recognition rates of our approach with the SOTA for the TheRusLan corpus with 18 gestures are presented in Table 4.

Approach	Recognition rate, %
(Axyonov et al., 2021)	53.07
	68.23
	69.74
	73.54
	74.28
	77.43
	79.98
	84.67
	87.38
	88.92
(Axyonov et al., 2022a)	91.14
Our	**93.33**

Table 4: Comparison of the recognition rate results of our approach with SOTA on the TheRusLan corpus with 18 gestures.

The conducted experiments indicate that the proposed approach is highly promising for improving GR in SL, as demonstrated by achieving an accuracy of 93.33% for 18 different gestures, including Russian SL gestures from the TheRusLan corpus, which is a significant improvement over the previous SOTA accuracy results (exceeding them by an absolute value of 2.19%). This indicates that the proposed approach is highly effective in recognizing SL gestures, particularly for the target SL.

The proposed approach of cross-language transfer learning using visual information to improve the recognition accuracy of the target SL has significant innovative potential. It allows for increasing the accuracy and robustness of machine SL translation, as well as improving the naturalness of HCI in general. Additionally, it has the potential to enhance the social adaptation of people with hearing impairments.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel approach for SL recognition that utilizes transfer learning to improve recognition rates on a target SL corpus. Our approach consists of two stages: (1) training model for simultaneous recognition of gestures on basic corpora of Turkish (AUTSL) and Russian (TheRusLan) SLs, and (2) fine-tuning pre-trained model on the target corpus using transfer learning. We showed that our approach achieved SOTA results on the AUTSL corpus and provided the first baseline result for the TheRusLan corpus with 164 individual words and phrases from Russian SL.

Experimental results on the TheRusLan corpus with 18 gestures demonstrated an absolute increase in recognition rate of 2.59% using our approach with transfer learning compared to our approach without transfer learning. Additionally, we compared our approach with SOTA approaches on the TheRusLan corpus with 18 gestures, and the results showed that our approach achieved competitive recognition rates.

In the future, the proposed approach will be extended to other SLs, and the effectiveness of other data augmentation techniques and learning rate scheduling methods will be explored. The possibility of incorporating additional sensory modalities will also be investigated, and the proposed approach will be deployed on mobile devices to enable real-time GR and translation, which will greatly benefit people with hearing impairments in their daily lives.

Moreover, the impact of cultural and individual differences on the accuracy of GR in different SLs will be investigated, and the proposed approach will be extended to the domain of SL generation. The system could be trained to generate SL gestures from text or speech input, which will be beneficial for people who are learning SL. Finally, the integration of the proposed approach with other assistive technologies will be explored to provide a more comprehensive and effective solution for people with hearing impairments.

ACKNOWLEDGEMENTS

This research is financially supported by the Russian Science Foundation (https://rscf.ru/en/project/21-71-00141/, No. 21-71-00141).

REFERENCES

Abdi, H., Williams, L. J., 2010. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459. doi.org/10.1002/wics.101.

Ahmed, I., Jeon, G., Piccialli, F., 2022. From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: a Survey on What, How, and Where. *IEEE Transactions on Industrial Informatics*, 18(8), 5031–5042. doi.org/10.1109/TII.2022.3146552.

Amangeldy, N., Kudubayeva, S., Kassymova, A., Karipzhanova, A. Z., Razakhova, B., Kuralov, S., 2022. Sign Language Recognition Method Based on Palm Definition Model and Multiple Classification. *Sensors*, 22(3). doi.org/10.3390/s22176621.

Axyonov, A. A., Kagirov, I. A., Ryumin, D. A., 2022a. A Method of Multimodal Machine Sign Language Translation for Natural Human-Computer Interaction. *Journal Scientific and Technical Of Information Technologies, Mechanics and Optics*, 139(3), 585. doi.org/10.17586/2226-1494-2022-22-3-585-593.

Axyonov, A., Ryumin, D., Kagirov, I., 2021. Method of Multi-Modal Video Analysis of Hand Movements for Automatic Recognition of Isolated Signs of Russian Sign Language. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIV-2/W1-2021, 7–13. doi.org/10.5194/ISPRS-ARCHIVES-XLIV-2-W1-2021-7-2021.

Axyonov, A., Ryumin, D., Kashevnik, A., Ivanko, D., Karpov, A., 2022b. Method for Visual Analysis of Driver's Face for Automatic Lip-Reading in the Wild. *Computer Optics*, 46, 955–962. doi.org/10.18287/2412-6179-CO-1092.

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., Grundmann, M., 2020. Blazepose: On-Device Real-Time Body Pose Tracking. *arXiv preprint arXiv:2006.10204*, 1–4. doi.org/10.48550/arXiv.2006.10204.

Belkina, A. C., Ciccolella, C. O., Anno, R., Halpert, R., Spidlen, J., Snyder-Cappione, J. E., 2019. Automated Optimized Parameters for T-Distributed Stochastic Neighbor Embedding Improve Visualization and Analysis of Large Datasets. *Nature Communications*, 10(1), 1–12. doi.org/10.1038/s41467-019-13055-y.

Boháek, M., Hrúz, M., 2022. Sign Pose-based Transformer for Word-level Sign Language Recognition. *Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 182–191. doi.org/10.1109/WACVW54805.2022.00024.

Boháek, M., Hrúz, M., 2023. Learning from What is Already Out There: Few-shot Sign Language Recognition with Online Dictionaries. *International Conference on Automatic Face and Gesture Recognition (FG)*, 1–6. doi.org/10.1109/FG57933.2023.10042544.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., Bowden, R., 2018. Neural Sign Language Rranslation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7784–7793. doi.org/10.1109/CVPR.2018.00812.

Camgöz, N. C., Koller, O., Hadfield, S., Bowden, R., 2020. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10020–10030. doi.org/10.1109/CVPR42600.2020.01004.

Carli, L. L., LaFleur, S. J., Loeber, C. C., 1995. Nonverbal Behavior, Gender, and Influence. *Journal of Personality and Social Psychology*, 68, 1030–1041.

Carro, I. M., Goudbeek, M., Krahmer, E. J., 2015. Coming of Age in Gesture: A Comparative Study of Gesturing and Pantomiming in Older Children and Adults. *GESPIN - Gesture Speech in Interaction Conference*, 1–7.

Coster, M. D., Herreweghe, M. V., Dambre, J., 2021. Isolated Sign Recognition from RGB Video using Pose Flow and Self-Attention. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3436–3445. doi.org/10.1109/CVPRW53098.2021.00383.

Dresvyanskiy, D., Ryumina, E. V., Kaya, H., Markitantov, M., Karpov, A., Minker, W., 2022. End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition in-the-Wild. *Multimodal Technologies and Interaction*, 6(2), 11. doi.org/10.3390/mti6020011.

Guo, L., Lu, Z., Yao, L., 2021. Human-Machine Interaction Sensing Technology Based on Hand Gesture Recognition: A Review. *IEEE Transactions on Human-Machine Systems*, 51, 300–309. doi.org/10.1109/THMS.2021.3086003.

Hu, H., Zhao, W., gang Zhou, W., Wang, Y., Li, H., 2021. SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 11067–11076. doi.org/10.1109/ICCV48922.2021.01090.

Iriskhanova, O., Cienki, A., 2018. The Semiotics of Gestures in Cognitive Linguistics: Contribution and Challenges. *Voprosy Kognitivnoy Lingvistiki*, 2018, 25–36. doi.org/10.20916/1812-3228-2018-4-25-36.

Izenman, A. J., 2013. Linear Discriminant Analysis. *Modern Multivariate Statistical Techniques*, 237–280. doi.org/10.1007/978-0-387-78189-1_8.

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y. R., 2021. Skeleton Aware Multi-modal Sign Language Recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3408–3418. doi.org/10.1109/CVPRW53098.2021.00380.

Kagirov, I., Ivanko, D., Ryumin, D., Axyonov, A., Karpov, A., 2020a. TheRuSLan: Database of Russian sign language. *Language Resources and Evaluation Conference (LREC)*, 6079–6085.

Kagirov, I., Ryumin, D. A., Axyonov, A. A., Karpov, A. A., 2020b. Multimedia Database of Russian Sign Language Items in 3D. *Voprosy Jazykoznanija*, 1, 104–123. doi.org/10.31857/S0373658X0008302-1.

Karpov, A., Železnỳ, M., 2012. Towards Russian Sign Language Synthesizer: Lexical Level. *International Workshop on Representation and Processing of Sign Languages at the LREC*, 83–86. doi.org/10.1109/CVPR.2018.00812.

Koller, O., Forster, J., Ney, H., 2015. Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. *Computer Vision and Image Understanding*, 141, 108–125. doi.org/10.1016/j.cviu.2015.09.013.

Li, D., Rodriguez, C., Yu, X., Li, H., 2020. Word-Level Deep Sign Language Recognition from Video: A New Large-Scale Dataset and Methods Comparison. *Winter Conference on Applications of Computer Vision (WACV)*, 1459–1469. doi.org/10.1109/WACV45572.2020.9093512.

Li, D., Rodriguez-Opazo, C., Yu, X., Li, H., 2019. 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. doi.org/10.1109/WACV45572.2020.9093512.

Liang, D., Yang, F., Zhang, T., Yang, P., 2018. Understanding MixUp Training Methods. *IEEE access*, 6, 58774–58783.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H., 2022. Video Swin Transformer. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3202–3211. doi.org/10.1109/CVPR52688.2022.00320.

Novopoltsev, M., Verkhovtsev, L., Murtazin, R., Milevich, D., Zemtsova, I., 2023. Fine-tuning of Sign Language Recognition Models: a Technical Report. *ArXiv*. doi.org/10.48550/arXiv.2302.07693.

Rajalakshmi, E., Elakkiya, R., Prikhodko, A. L., Grif, M., Bakaev, M. A., Saini, J. R., Kotecha, K., Subramaniyaswamy, V., 2022. Static and Dynamic Isolated Indian and Russian Sign Language Recognition with Spatial and Temporal Feature Detection using Hybrid Neural Network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1), 1–23. doi.org/10.1145/3530989.

Rajalakshmi, E., Elakkiya, R., Subramaniyaswamy, V., Alexey, L. P., Mikhail, G., Bakaev, M., Kotecha, K., Gabralla, L. A., Abraham, A., 2023. Multi-Semantic Discriminative Feature Learning for Sign Gesture Recognition Using Hybrid Deep Neural Architecture. *IEEE Access*, 11, 2226–2238. doi.org/10.1109/ACCESS.2022.3233671.

Ryumin, D., Ivanko, D., Axyonov, A., Kagirov, I., Karpov, A., Zelezny, M., 2019a. Human-Robot Interaction with Smart Shopping Trolley using Sign Language: Data Collection. *International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 949–954. doi.org/10.1109/PERCOMW.2019.8730886.

Ryumin, D., Ivanko, D., Ryumina, E. V., 2023. Audio-Visual Speech and Gesture Recognition by Sensors of Mobile Devices. *Sensors*, 23(4), 1–29. doi.org/10.3390/s23042284.

Ryumin, D., Kagirov, I., Axyonov, A., Pavlyuk, N., Saveliev, A. I., Kipyatkova, I. S., Zelezný, M., Mporas, I., Karpov, A., 2020. A Multimodal User Interface for an Assistive Robotic Shopping Cart. *Electronics*, 9, 2093. doi.org/10.3390/electronics9122093.

Ryumin, D., Kagirov, I., Ivanko, D., Axyonov, A., Karpov, A., 2019b. Automatic Detection and Recognition of 3D Manual Gestures for Human-Machine Interaction. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W12, 179–183. doi.org/10.5194/isprs-archives-XLII-2-W12-179-2019.

Ryumin, D., Karpov, A., 2017. Towards Automatic Recognition of Sign Language Gestures using Kinect 2.0. *Universal Access in Human-Computer Interaction. Designing Novel Interactions: International Conference (UAHCI)*, 89–101. doi.org/10.1007/978-3-319-58703-5_7.

Ryumina, E., Dresvyanskiy, D., Karpov, A., 2022. In Search of a Robust Facial Expressions Recognition Model: A Large-Scale Visual Sross-Corpus Study. *Neurocomputing*, 514, 435–450. doi.org/10.1016/j.neucom.2022.10.013.

Ryumina, E. V., Karpov, A. A., 2020. Comparative Analysis of Methods for Imbalance Elimination of Emotion Classes in Video Data of Facial Expressions. *Journal Scientific and Technical of Information Technologies, Mechanics and Optics*, 129(5), 683–691. doi.org/10.17586/2226-1494-2020-20-5-683-691.

Saunders, B., Camgoz, N. C., Bowden, R., 2021. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. *International Journal of Computer Vision*, 129(7), 2113–2135. doi.org/10.1007/s11263-021-01457-9.

Selvaraj, P., GokulN., C., Kumar, P., Khapra, M. M., 2021. OpenHands: Making Sign Language Recognition Accessible with Pose-based Pretrained Models across Languages. *Annual Meeting of the Association for Computational Linguistics*, 2114–2133. 10.18653/v1/2022.acl-long.150.

Serengil, S. I., Ozpinar, A., 2021. Hyperextended Lightface: A Facial Attribute Analysis Framework. *International Conference on Engineering and Emerging Technologies (ICEET)*, 1–4. doi.org/10.1109/ICEET53442.2021.9659697.

Sincan, O. M., Junior, J., Jacques, C., Escalera, S., Keles, H. Y., 2021. ChaLearn LAP Large Scale Signer Independent Isolated Sign Language Recognition Challenge: Design, Results and Future Research. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3472–3481. doi.org/10.1109/CVPRW53098.2021.00386.

Sincan, O. M., Keles, H. Y., 2020. AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods. *IEEE Access*, 8, 181340–181355. doi.org/10.1109/ACCESS.2020.3028072.

Song, N., Xiang, Y., 2022. SLGTformer: An Attention-Based Approach to Sign Language Recognition. *ArXiv*, 1–12. doi.org/10.48550/arXiv.2212.10746.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E., 2016. Hierarchical Attention Networks for Document Classification. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. doi.org/10.18653/v1/N16-1174.

Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D., 2017. MixUp: Beyond Empirical Risk Minimization. *arXiv preprint arXiv:1710.09412*, 1–13. doi.org/10.48550/arXiv.1710.09412.