Literature Review: Sign Language Production and Recognition Systems

Shifa Shah

Habib University

Abstract

This literature review examines four key research papers that address fundamental challenges in sign language technology, covering both production and recognition systems. The review analyzes Progressive Transformers for end-to-end sign language production [Saunders et al., 2020], Text2Sign's approach using neural machine translation and generative adversarial networks [Stoll et al., 2020], comparative analysis of CNN-based architectures for sign language recognition [Louison et al., 2024], and real-time ASL gesture recognition using MediaPipe and CNNs [Kumar et al., 2023]. These works collectively represent significant advances in bridging communication gaps between hearing and Deaf communities through automated sign language systems. The review identifies key technical contributions, limitations, and future research directions in this rapidly evolving field.

Contents

1 Progressive Transformers for End-to-End Sign Language Production

1.1 Introduction

This review examines "Progressive Transformers for End-to-End Sign Language Production" by Saunders et al. [2020]. This paper is crucial for my Text to Sign Language project. The core problem addressed is Sign Language Production (SLP), converting spoken language into continuous sign sequences. This problem is highly significant as automatic SLP is essential for better involving the Deaf community in the wider, predominantly spoken world. Previous work often produced only concatenated isolated signs, lacking architectures for continuous sign language. This paper proposes a novel end-to-end solution.

1.2 Critical Analysis

The paper introduces Progressive Transformers, presented as the first SLP model to translate from discrete spoken language sentences to continuous 3D sign pose sequences in an end-to-end manner. The aim is to achieve human-level SLP to revolutionize Deaf-hearing communication.

The Progressive Transformer adapts the standard Transformer architecture for continuous outputs. It translates symbolic input (text or gloss) to continuous 3D sign pose sequences. A novel counter decoding technique is introduced to enable continuous sequence generation at training and inference, tracking progress with a value between 0 and 1, concluding generation when the counter reaches 1. This allows prediction of variable-length sequences and drives timing. Two configurations are evaluated:

- **Text to Gloss to Pose (T2G2P):** Uses a Symbolic Transformer for text-to-gloss, then a Progressive Transformer for gloss-to-pose.
- **Text to Pose (T2P):** A single Progressive Transformer is used for end-to-end text-to-pose translation. To mitigate drift in continuous sequence generation, Conditional Future Prediction and Gaussian Noise augmentations are applied, achieving the best performance.

Models are trained and evaluated on the PHOENIX14T dataset using normalized 3D skeleton pose data.

1.3 Results

The Symbolic Transformer achieved top performance on the PHOENIX14T Text-to-Gloss task. Progressive Transformers, evaluated via back-translation using BLEU and ROUGE scores, showed strong results. Data augmentation—especially future prediction combined

with Gaussian noise—reduced drift and improved accuracy. The direct Text-to-Pose (T2P) method outperformed the Text-to-Gloss-to-Pose (T2G2P) approach, proving effective when gloss data is unavailable. Qualitative outputs showed smooth, realistic sign language motion matching the ground truth.

1.4 Strengths and Weaknesses

This work stands out as the first end-to-end model for translating text to continuous 3D sign pose, introducing counter decoding and effective data augmentation to achieve strong back-translation results. It demonstrates the feasibility of direct text-to-pose translation, addressing gaps in prior work that relied on isolated signs or rule-based systems unsuited for continuous sign language. While body motion is smooth and realistic, limitations remain in accurately generating proper nouns, specific entities, and fully expressive hand shapes. Future improvements should focus on incorporating non-manual features like facial expressions and mouthings.

1.5 Relevance and Conclusion

This paper is highly relevant to my Text-to-Sign Language project, as it presents a state-of-the-art deep learning approach using Progressive Transformers for translating text directly into continuous 3D sign language poses. The paper introduces novel techniques like counter decoding and effective data augmentation, which significantly improve performance and offer practical methods that I can apply to my own project. It also highlights the importance of moving beyond traditional isolated sign generation and instead focusing on continuous, variable-length outputs. The model's ability to generate smooth body movements makes it a strong foundation for real-world applications. Overall, the paper provides a solid framework, innovative techniques, and valuable insights that strongly support and guide the development of my own end-to-end Text-to-Sign Language system.

2 Text2Sign: Neural Machine Translation and Generative Adversarial Networks

2.1 Introduction

This review focuses on the paper "Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks" by Stoll et al. [2020]. This paper addresses the significant problem of Sign Language Production (SLP), specifically converting spoken language text into continuous sign language video. Automatic SLP is crucial for facilitating communication between the hearing and Deaf communi-

ties. While previous work often relied on limited avatar-based systems or focused on isolated signs, this paper proposes a novel end-to-end approach to generate continuous sign language video without needing a classical graphical avatar.

2.2 Critical Analysis

The paper presents Text2Sign, a novel system for automatic Sign Language Production that generates sign videos directly from spoken language sentences. A key contribution is achieving continuous sign video generation without relying on traditional animated avatars. It requires minimal gloss and skeletal annotations for training compared to heavily annotated approaches.

The system consists of two main stages: Text to Pose (text2pose) and Pose to Video (pose2video). In the text2pose stage, a Neural Machine Translation (NMT) model translates text into gloss probabilities, which guide transitions through a Motion Graph built from sign language pose data. This process generates smooth 2D skeletal pose sequences. In the pose2video stage, a Generative Adversarial Network (GAN) converts these pose sequences into realistic video frames, conditioned on signer appearance. The system is trained across multiple datasets, showing strong generalization and effective cross-dataset learning.

2.3 Results

The system achieved strong performance across stages: the Text2Gloss model performed well on standard metrics, and the Text2Pose output aligned closely with glosses. Multi-Signer video generation produced realistic images, preserving key details like hands and faces. The full Text-to-Video pipeline showed smooth, consistent signing, though expressiveness was slightly limited. HD generation notably improved realism, especially in hand and facial features, by using more detailed keypoints.

2.4 Strengths and Weaknesses

This work presents a major advancement in continuous sign language video generation by moving beyond traditional avatar-based systems. It integrates Neural Machine Translation, Motion Graphs, and GANs into a unified, data-driven pipeline that requires minimal specialized annotations and supports both multi-signer and high-definition video generation. Key strengths include its scalability, ability to generate continuous sign sequences without avatars, and successful demonstration of signer diversity and improved realism. However, limitations remain—output resolution and expressiveness still fall short of motion capture or avatar-based methods. Challenges include generating accurate hand and arm movements, likely due to averaged pose sequences and timing inaccuracies in

training data. Critically, non-manual features like facial expressions, body posture, and mouthings are not yet fully captured, marking a key area for future improvement. The paper effectively addresses gaps in prior work, which largely focused on isolated signs or robotic avatars, by offering a more natural and scalable approach to continuous text-to-sign translation.

2.5 Relevance and Conclusion

This paper combines Neural Machine Translation and Motion Graphs for pose generation, followed by GANs for video synthesis. The approach supports multi-signer and high-definition outputs, offering flexibility and realism. While it highlights challenges like limited expressiveness, hand accuracy, and missing non-manual features, it provides a strong foundation and clear direction for improving continuous sign language generation in my research.

3 Comparative Analysis of CNN Architectures for Sign Language Recognition

3.1 Introduction

This review focuses on the paper "Learning Sign Language Representation using CNN-LSTM, 3DCNN, CNN–RNN–LSTM and CCN-TD" by Louison et al. [2024]. This paper is relevant to my project in Sign Language Recognition (SLR) as it explores and evaluates different neural network algorithms for real-time video sign translation and grading of sign language accuracy for new users. The ability to automatically recognize and process sign language from video is critical for bridging communication gaps. The paper specifically examines the performance of these algorithms on datasets, including the previously unexplored Trinidad and Tobago Sign Language (TTSL).

3.2 Critical Analysis

The paper aims to evaluate and identify the best neural network algorithm among CNN-LSTM, 3DCNN, CNN-RNN-LSTM, and CNN-TD for facilitating a sign language tuition system that allows for real-time, video sign translation and grading of sign language accuracy. It requires algorithms capable of recognizing and processing spatial and temporal features within sign language videos. The study tests these algorithms on a dataset containing both American Sign Language (ASL) and Trinidad and Tobago Sign Language (TTSL). The researchers implement and compare four main neural network architectures:

- **CNN-LSTM:** An architecture primarily for sequence prediction tasks using spatial inputs like videos. The implementation uses a prebuilt Keras model, Convnet LSTM Model (based on ConvLSTM2D), followed by a decision network.
- **3DCNN:** Utilizes 3D convolution layers with learnable 3D filters to produce spatiotemporal feature maps from stacking frames. It's expected to thrive on spatio-temporal datasets. The model used is based on prior work and executed using TensorFlow/Keras' Conv3D model. It has the highest number of parameters among the models tested.
- **CNN-RNN-LSTM:** This architecture defines layers individually, differing from the prebuilt approach of CNN-LSTM. The model used is based on prior work.
- **CNN-TD (Time Distributed):** An addition to the CNN approach where a Time Distributed layer is wrapped around convolutional and pooling layers. This layer applies the same layer to multiple batch inputs over time, allowing CNNs to work with image sequences.

The dataset frames were divided into training (80%), validation (10%), and testing (10%) sets. Videos were limited to a sequence length of 35 frames for consistency. The models were implemented in Python 3 using Keras and TensorFlow backend. Training consistently used the Categorical Cross-Entropy (Softmax Loss) function for loss calculation and the Adam optimizer. Hyper parameters included epochs (15-20 with early stopping) and batch numbers. Model performance was evaluated using a confusion matrix and a classification report, including metrics like accuracy, precision, recall, F1 scores, and support. Training and loss accuracy progression graphs were also analyzed.

3.3 Results

- **Overall Comparison:** The 3DCNN algorithm was found to be the best performing neural network algorithm among those tested. It demonstrated the best performance in terms of accuracy and overall results for both TTSL and ASL, followed by the CNN-LSTM and CNN-TD models.
- **Accuracy Scores:** 3DCNN achieved 91% accuracy on the TTSL dataset and 83% accuracy on the ASL dataset. CNN-LSTM had average F1 scores of 83% (ASL) and 74% (TTSL). CNN-TD was a surprisingly high performer, showing better results than CNN-RNN-LSTM.
- **Training Quality:** CNN-LSTM and 3DCNN showed favorable training quality with smooth accuracy ascent and low loss descent. CNN-RNN-LSTM had low training quality with significantly higher error values than accuracy. CNN-TD showed some symptoms of overfitting.
- **New Sign Assessment Simulation:**

- The 3DCNN model correctly identified 3 out of 4 TTSL signs, making one incorrect translation with 78% certainty.
- The CNN-TD model only correctly identified 1 out of 4 TTSL signs translating others incorrectly with 100% certainty.
- The 3DCNN model was significantly faster in execution time for translating new signs compared to CNN-TD.

3.4 Strengths and Weaknesses

Strengths include evaluating modern neural network architectures on a new, previously unexplored dataset (TTSL). The study provides a direct comparison of four different model types for video-based SLR. It successfully identifies 3DCNN as the top-performing model for this task and the specific datasets used. Weaknesses and areas for improvement include the small dataset size, which negatively impacted training accuracy, particularly for signs with fewer examples. Issues with model setup for certain architectures (like CNN-RNN-LSTM) led to poor performance. The "real-time"/new sign assessment simulation highlighted that the models were sensitive to variations from the training data (same signer, same environment), suggesting the need for a larger, and more varied dataset for better generalizability. The focus is on recognizing individual signs rather than the continuous flow and grammar of sign language sentences. The paper points out the need for more studies focusing on Sign Language Recognition using neural networks with accessible technologies in the Caribbean region.

3.5 Relevance and Conclusion

This paper is highly relevant to my research on Sign Language Recognition and the development of sign language tuition systems. It provides a thorough comparison of several neural network architectures for recognizing signs from video, specifically focusing on the TTSL and ASL datasets. The study identifies 3DCNN as the most effective model for sign recognition, which offers a strong foundation for selecting a base model in my Sign Language Recognition tasks. Additionally, the paper emphasizes the importance of dataset size and variability, highlighting key challenges that can inform strategies for data collection and preparation in my own work. While this paper centers on recognition (understanding signs from video), and the previously reviewed Text2Sign paper focuses on sign production (generating signs from text), both papers complement each other and address crucial aspects of the broader text-to-sign challenge. The findings from this paper will help in developing practical applications, such as real-time feedback systems or sign language grading tools, to assist learners in improving their sign language skills, particularly for underrepresented languages like TTSL.

4 Real-Time ASL Recognition with MediaPipe and CNNs

4.1 Introduction

This review examines the paper "Mediapipe and CNNs for Real-Time ASL Gesture Recognition" by Kumar et al. [2023]. The study leverages modern computer vision and machine learning approaches to facilitate communication for people with hearing impairments through real-time American Sign Language alphabet recognition.

4.2 Critical Analysis

The paper presents a system for real-time American Sign Language (ASL) gesture recognition. The primary goal is to identify ASL alphabets, with potential applications in communication devices for people with hearing impairments and applicability to other sign languages. The study focuses on computer vision-based ASL gesture recognition using a dataset of 26 classes (A–Z), with 4500 images per class. Mediapipe is used for real-time hand tracking, extracting 21 landmarks per hand. These landmarks are processed by adjusting coordinates relative to the hand center, normalizing for scale, and flattening into a 1D feature vector. A CNN classifies the gestures from the 42-coordinate input. The model is trained on 80% of the dataset and tested on 20%, with performance evaluated using classification metrics and a confusion matrix.

4.3 Results

- The proposed system achieved remarkable performance on the test set.
- It obtained a high accuracy score of 99.95% on the test set.
- The classification report showed 100% precision, recall, and F1-score values for all classes, indicating that the model properly identified every sample in the test set without errors.

4.4 Strengths and Weaknesses

The paper presents a highly accurate, real-time ASL alphabet recognition system using MediaPipe for hand tracking and a CNN for classification. It stands out for its lightweight, sensor-free design and potential use in communication aids. The system effectively demonstrates MediaPipe's capability when combined with CNNs, achieving near-perfect accuracy on a large static ASL dataset. However, it is limited to isolated static gestures and lacks integration of facial expressions and body language, which are vital for full sign language understanding. The study builds on prior work involving various techniques (e.g., Inception V3, Naive Bayes, SVM, LSTM), but uniquely highlights

MediaPipe with CNN for static gesture recognition. Future improvements include expanding to dynamic gestures, incorporating richer sign features, and enhancing robustness to real-world variability.

4.5 Relevance and Conclusion

The paper by Kumar et al. presents a highly effective, real-time ASL alphabet recognition system that combines MediaPipe for hand tracking with a CNN for classification, achieving 99.95% accuracy on a large static gesture dataset. Its relevance lies in addressing the static sign recognition subproblem within the broader Sign Language Recognition (SLR) field, providing a strong benchmark and a practical, sensor-free solution for communication tools. The feature extraction method using hand landmarks is particularly valuable and adaptable for more complex, dynamic sign tasks with additional temporal processing. While limited to isolated static gestures, the work lays a solid foundation for expanding into dynamic and continuous sign language recognition and integrating into larger communication or educational applications.

5 Synthesis and Future Directions

The reviewed literature demonstrates significant progress in both sign language production and recognition systems, yet reveals several key areas for future development. The Progressive Transformers approach [Saunders et al., 2020] establishes a strong foundation for continuous sign generation, while Text2Sign [Stoll et al., 2020] advances video-based production without traditional avatars. The comparative analysis by Louison et al. [2024] identifies 3DCNN as the most effective architecture for recognition tasks, and the MediaPipe-based system [Kumar et al., 2023] achieves near-perfect accuracy for static gesture recognition.

Key challenges across all approaches include: (1) incorporation of non-manual features such as facial expressions and body language, (2) handling of proper nouns and domain-specific vocabulary, (3) scalability to larger, more diverse datasets, and (4) real-time performance optimization. Future research should focus on multimodal integration, cross-lingual sign language systems, and development of more robust evaluation metrics that better capture sign language fluency and naturalness.

References

Kumar, R., Singh, S. K., Bajpai, A., and Sinha, A. (2023). Mediapipe and cnns for real-time asl gesture recognition. *arXiv preprint arXiv:2305.05296*.

Louison, N., Goodrige, W., and Khan, K. (2024). Learning sign language representation using cnn-lstm, 3dcnn, cnn–rnn–lstm and ccn-td. *arXiv preprint arXiv:2412.18187*.

Saunders, B., Camgoz, N. C., and Bowden, R. (2020). Progressive transformers for end-to-end sign language production. In *Lecture Notes in Computer Science*, pages 687–705. Springer.

Stoll, S., Camgoz, N. C., Hadfield, S., and Bowden, R. (2020). Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.