A Framework for Real-Time Emotion-Aware Sign Language Translation

Motivation

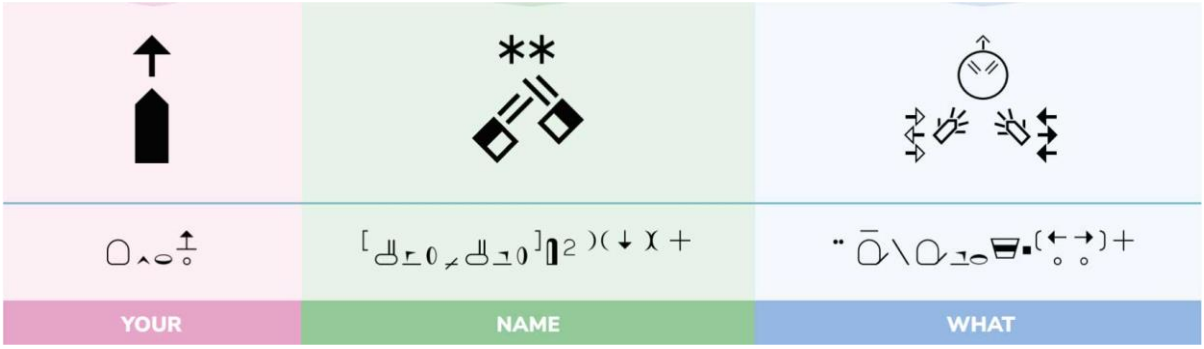
Around 70 million people worldwide use sign language, and there are over 200 different sign languages. However, there is a big gap in communication between signers and people who use spoken languages. Sign language is just as complex as spoken language, but many people struggle to understand it. Often, signers are expected to lip-read or learn written languages, but the wider society should also make space for sign language in daily life. This research aims to create a real-time system that can translate sign language while also capturing the emotions behind it. By including emotions in the translation, the system can make communication between signers and non-signers more natural and clearer, helping to close the communication gap.

Related Studies

Existing research on sign language processing has primarily focused on sign detection, identification, recognition, translation, and production. Detection and recognition methods often use hand, body, and facial landmarks to identify sign movements, while recognition systems leverage deep learning models to understand sign gestures from video data. These models have shown promise but still face challenges such as limited datasets and difficulty distinguishing between gestures and general movements.

For translation, many studies focus on converting signs to text or spoken language, often using gloss notation systems. However, this approach has limitations due to high lexical and syntactical differences between sign language glosses and natural spoken or written languages.

Sign language production research has largely revolved around systems that either generate signs from gloss annotations or produce 3D avatar-based animations. Sign writing systems, like HamNoSys, are widely used to represent signs in text form, enabling translation systems to convert between written and sign languages. However, these systems still depend heavily on manual annotations and are hindered by the limitations of current pose estimation technologies.



Recent research has also explored the inclusion of emotion detection in sign language systems. This addition aims to capture the emotional context of a signer's expression, as facial emotions play a crucial role in conveying the full meaning of a sign. Although some systems use AI techniques to detect emotions, the accuracy of these models is still a challenge, especially in real-time applications.

Research Gaps

Despite the growing interest in sign language processing, current systems are still far from providing an accurate, real-time solution for recognition and translation. Sign languages are inherently complex, involving not only hand gestures but also facial expressions, body movements, and non-manual signals. Existing models struggle to handle this multimodal nature of communication effectively.

One major challenge lies in the lack of large-scale, diverse datasets that can represent the richness and variety of sign languages across different regions and communities. Without sufficient data, it's difficult to train models that can generalize well across various sign languages and signing styles. Furthermore, the transfer of emotions between signers and non-signers remains a significant hurdle. Emotions are critical for effective communication, but capturing and transferring them accurately between different modalities (sign language and spoken language) is challenging due to the different types of multimodal data involved.

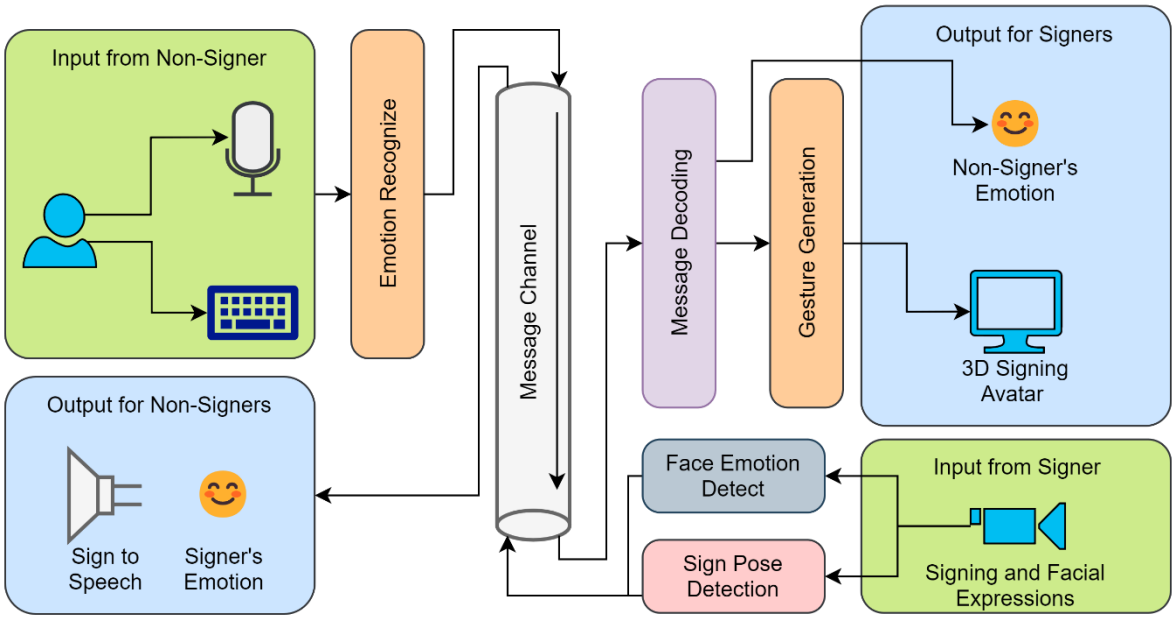
Objectives

The primary goal of this research is to develop an end-to-end framework for real-time, emotion-aware sign language recognition, translation, and production. The system will use advanced AI techniques and Human-Computer Interaction (HCI) approaches to reduce the communication gap between signers and non-signers. Specific objectives include:

1. **Developing an Emotion-Aware System:** The framework will focus on detecting and recognizing sign language, including the non-manual components like facial expressions and body language, in real-time. The system will also capture emotions from both signers and non-signers, enhancing the natural flow of communication.
2. **AI for Emotion Detection:** The system will utilize AI techniques to detect emotions from non-signer speech, and it will apply AI-based facial expression recognition to understand the emotions of signers. Additionally, the system will detect non-manual movements like head tilts and shoulder shifts for deeper emotion analysis.
3. **Sign Language Processing (SLP) and NLP Integration:** Natural Language Processing (NLP) techniques will be incorporated into the system to tokenize and analyze sign language at the word level. The framework will also account for the unique grammatical structures of sign languages and integrate this understanding into the Sign Language Processing (SLP) pipeline.
4. **3D Avatar-Based Sign Language Production:** To enable more intuitive interaction, the system will produce sign language through 3D avatars in a virtual environment. Augmented Reality (AR) and Virtual Reality (VR) technologies will be used to create immersive experiences for signers, while low-cost Internet of Things (IoT) solutions, such as smart gloves, will be explored to facilitate interaction with the virtual world.

Proposed Framework

The proposed architecture presents a comprehensive framework for bidirectional communication between signers and non-signers as given in the following figure leveraging various AI technologies to bridge the communication gap. The system is designed to process inputs from both parties and generate appropriate outputs, facilitating seamless interaction. Here's a detailed breakdown of the architecture:



1. Input from Non-Signer:
    - The system accepts input from non-signers through two primary modes: a) Speech: Captured via a microphone b) Text: Entered through a keyboard
    - This dual input method ensures flexibility for non-signers to communicate effectively.
2. Emotion Recognition:
    - The non-signer's input is processed through an Emotion Recognition module.
    - This module analyzes the speech or text to detect the emotional context of the message.
    - The recognized emotion is then integrated into the subsequent processing steps.
3. Message Channel:
    - Acts as the central conduit for information flow within the system.
    - Receives input from both the Emotion Recognition module and the non-signer's raw input.
    - Manages the routing of information to appropriate processing modules.
4. Output for Non-Signers:
    - The system generates two types of output for non-signers: a) Sign to Speech: Converts sign language to audible speech. b) Signer's Emotion: Displays the detected emotion of the signer, likely through an emoji or similar visual representation.
5. Input from Signer:
    - Captures the signer's input through a camera.
    - Records both signing and facial expressions.

6. Face Emotion Detection:
    - Analyzes the facial expressions of the signer to detect their emotional state.
    - This information is fed into the Message Channel for integration with other data.
7. Sign Pose Detection:
    - Utilizes computer vision and AI models to detect and interpret sign language gestures.
    - Likely employs landmark detection techniques to accurately capture hand and body movements.
8. Message Decoding:
    - Processes the interpreted sign language gestures.
    - Converts the gestures into a format that can be further processed or translated.
9. Gesture Generation:
    - Takes the decoded message and emotion data as inputs.
    - Generates appropriate gestures for the 3D signing avatar.
10. Output for Signers:
    - Provides two main outputs for signers: a) Non-Signer's Emotion: Displayed as an emoji, conveying the emotional context of the non-signer's message. b) 3D Signing Avatar: A visual representation that performs sign language gestures corresponding to the non-signer's input.

This architecture demonstrates a holistic approach to sign language processing, addressing not just the linguistic aspects but also the crucial emotional components of communication. By incorporating emotion recognition and conveyance for both parties, the system aims to provide a more nuanced and effective communication experience.
The use of AI and computer vision technologies throughout the pipeline enables real-time processing and translation, making this a powerful tool for breaking down communication barriers between signers and non-signers. The bidirectional nature of the system ensures that both parties can express themselves naturally in their preferred mode of communication, with the technology bridging the gap between spoken/written language and sign language.

Results

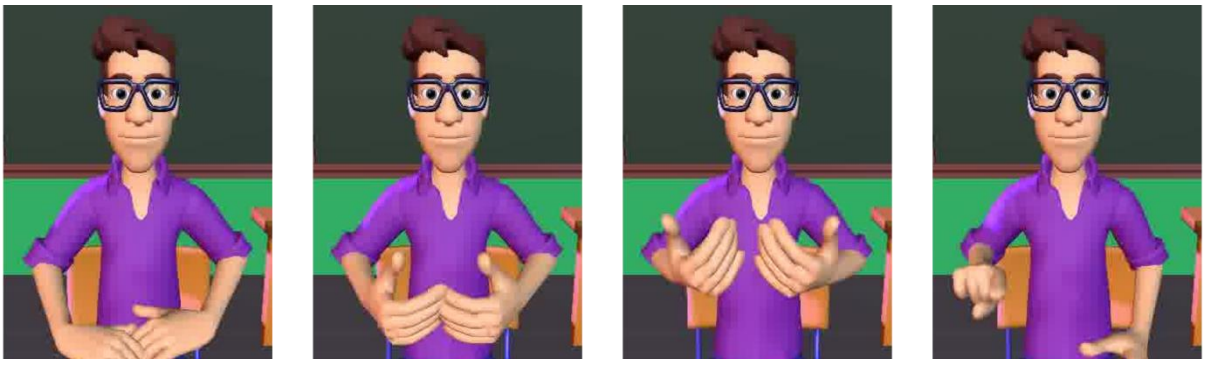
The proposed end-to-end system for sign language processing demonstrates impressive performance across various components, leveraging multiple AI models to achieve high accuracy in different tasks.

Speech Emotion Recognition: For speech emotion recognition, we implemented a hybrid CNN-XGBoost model. This approach begins with feature extraction from audio inputs, followed by classification using the combined power of Convolutional Neural Networks (CNN) and XGBoost. The model has shown exceptional performance, achieving a remarkable 98% accuracy when tested on a collection of several benchmark datasets. This high accuracy ensures that the emotional context of non-signers' speech is reliably captured and conveyed to signers, enhancing the quality of communication.

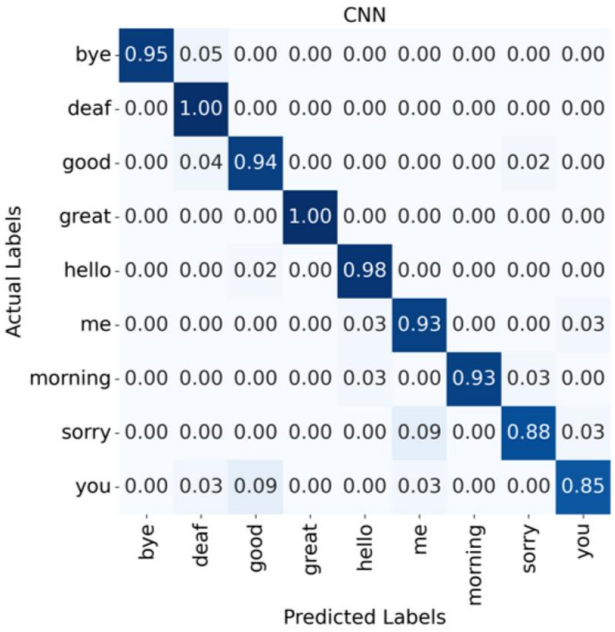
Face Emotion Recognition: In the domain of face emotion recognition, our system utilizes a landmark-based approach coupled with an XGBoost classifier. The process involves extracting facial landmarks from the signer's video input, which are then fed into the XGBoost model for emotion classification. This method has proven to be highly effective, yielding an accuracy of 89%. While slightly lower than the speech emotion recognition accuracy, this performance is still impressive given the complexity of facial expression analysis and the potential for subtle variations in expressions.

ASL to English Text Detection: For the critical task of translating American Sign Language (ASL) to English text, we employed a CNN-based technique. This model focuses on detecting and translating 9 different ASL words, achieving a solid accuracy of 90%. This high performance indicates the system's strong capability in interpreting basic ASL gestures and converting them into text, facilitating effective communication from signers to non-signers. To provide a more detailed understanding of the model's performance, we have included a confusion matrix for the ASL to Word detector, which likely shows the distribution of correct predictions and misclassifications across the 9 words.

3D Avatar-based ASL Generation: The system also includes a component for generating ASL signs through a 3D avatar, which serves as the visual output for signers. While specific accuracy metrics for this component are not provided, the results indicate that several signs are successfully displayed through the avatar. This suggests that the system can effectively convert non-signers' input into visual ASL representations, completing the communication loop.



The high accuracy rates across speech emotion recognition (98%), face emotion recognition (89%), and ASL to English text detection (90%) demonstrate the robustness and effectiveness of the proposed system. These results indicate that the framework can reliably process and translate between spoken language and ASL, while also capturing and conveying emotional context in both directions.



The inclusion of a confusion matrix for the ASL to Word detector provides valuable insights into the specific strengths and potential areas for improvement in sign language recognition. This detailed analysis can guide future refinements of the model, potentially leading to even higher accuracy and a broader vocabulary of recognized signs.

Overall, these results suggest that the proposed architecture offers a promising solution for bridging the communication gap between signers and non-signers, with high accuracy in crucial areas of language processing and emotion recognition. The system's ability to handle multiple aspects of communication, including linguistic content and emotional context, positions it as a comprehensive tool for facilitating more natural and nuanced interactions between diverse user groups.

Future Research

To further enhance the capabilities and accuracy of our sign language processing system, several promising avenues for future research emerge. A primary focus will be on improving ASL detection from videos through advanced sign segmentation techniques. This will enable more precise recognition of individual signs within continuous signing, crucial for handling natural conversations. We also aim to expand the system's versatility by developing models for translation between different sign languages, not just between ASL and spoken language. This cross-sign language translation will significantly broaden the system's applicability across diverse deaf communities. Additionally, we plan to explore the application of more advanced AI models, such as transformer-based architectures or large language models fine-tuned on sign language data, to improve real-time sign language processing and translation accuracy. Another critical area for improvement is the 3D avatar's ability

to express non-manual movements. These subtle facial expressions, head tilts, and body postures are integral to conveying meaning in sign languages. Enhancing our avatar's capability to accurately reproduce these non-manual elements will greatly increase the naturalness and expressiveness of the generated signs. Finally, we intend to conduct extensive user studies with both deaf and hearing participants to gather feedback on the system's usability, accuracy, and overall effectiveness in facilitating communication. These insights will guide further refinements and ensure that our research continues to address the real-world needs of the signing community.