ASL HAND GESTURE DETECTION AND SPEECH CONVERSION USING VGG16 ARCHITECTURE

ABSTRACT

This study develops an efficient system for American Sign Language (ASL) hand gesture recognition and speech conversion utilizing VGG16 architecture. The ASL dataset from Kaggle comprises 36 distinct subfolders representing alphabets (A-Z) and digits (0-9). Transfer learning with a pre-trained VGG16 model enhanced the learning process and improved model accuracy. The training phase achieved a remarkable 97% accuracy with precision and recall reaching 98% during testing. Real-time implementation demonstrated superior performance in recognizing hand gestures accurately and consistently. The system extends its functionality by converting recognized gestures into corresponding speech outputs, bridging communication gaps for individuals relying on ASL.

However, challenges such as managing variations in hand orientation, lighting conditions, and user-specific differences during real-time detection were encountered. Future directions include expanding the model's functionality to recognize dynamic gestures and words, optimizing for low-computation contexts, and strengthening robustness through larger, more diverse datasets.

---

1. INTRODUCTION

American Sign Language (ASL) serves as an essential bridge connecting hard-of-hearing and deaf communities with the broader world. Despite its importance, communication gaps frequently arise between ASL users and non-users, creating significant challenges in social settings, educational institutions, and workplaces. This research addresses these barriers by integrating ASL hand gesture recognition with speech conversion, leveraging recent advances in computer vision and machine learning to enable natural real-time communication.

ASL hand gesture detection and speech conversion systems capture ASL signs using computer vision techniques, interpreting and translating them into spoken language. The system identifies hand gestures representing letters or words and immediately converts them into text and audio outputs, enabling seamless two-way communication. This approach offers a new pathway for ASL users to become more interactive with non-signing individuals.

Previous research has tackled hand gesture recognition for ASL, but challenges remain in achieving accurate and consistent recognition across diverse hand shapes, lighting conditions, and gesture speeds. This study employs transfer learning techniques, specifically utilizing the VGG16 architecture. Transfer learning involves fine-tuning a previously trained model—one trained on a large dataset—for a related task. This approach leverages pre-existing knowledge while requiring less training data and computational resources.

VGG16 Architecture

VGG16 is a deep convolutional neural network model highly effective for image classification tasks. Originally designed for object recognition in large-scale image datasets like ImageNet, VGG16 consists of 16 layers: 13 convolutional layers and 3 fully connected layers. These layers enable the model to learn intricate details in images, such as edges, textures, and shapes—essential for accurate hand gesture recognition.

VGG16's architecture uses small 3x3 convolutional filters throughout the network, capturing fine-grained details while maintaining manageable computational complexity. This makes VGG16 particularly effective at detecting subtle variations in hand gestures, even under varying lighting conditions or with different hand shapes. Using a pre-trained VGG16 model enables the system to rapidly adapt to hand gesture recognition without requiring full retraining, significantly reducing time and computing requirements.

The integration of speech synthesis enhances system utility, translating detected gestures into vocalized words and increasing interactivity and accessibility. This transforms a simple visual recognition tool into a complete communication solution between ASL users and non-users of sign language—a critical feature for real-time applications where immediate feedback is necessary.

The development of ASL hand gesture detection and speech conversion represents a revolutionary step in enhancing accessibility and inclusivity for hard-of-hearing and deaf communities. By converting hand gestures into spoken language, this research addresses significant communication barriers and showcases the capability of integrating computer vision and artificial intelligence to solve real-world problems.

---

2. LITERATURE REVIEW

Recent research has made significant strides in sign language recognition systems. Shashidhar et al. (2022) developed a model converting Indian Sign Language (ISL) gestures into speech using CNNs, achieving 85% accuracy. Binwant Kaur et al. (2023) utilized a pre-trained InceptionResNetV2 model for real-time ASL conversion, achieving training and validation accuracies of 98.23% and 97.07% respectively, with implementation on Jetson Nano technology for portable translation.

Tano et al. (2024) compared four models—Faster R-CNN with ResNet50, Faster R-CNN with MobileNetV3, SSDLite, and YOLOv8—finding Faster R-CNN with ResNet50 most accurate for ASL sign detection. Chwesiuk et al. (2024) emphasized the importance of automated ASL recognition technologies, noting that CNNs excel in gesture recognition when combined with data augmentation and transfer learning to improve accuracy and reduce overfitting.

Dabwan et al. (2024) developed a sign language recognition system using DenseNet121, achieving 97% training accuracy and 96% validation accuracy on ASL gestures (excluding dynamic gestures for 'J' and 'Z'). Bayrak et al. (2024) proposed a real-time ASL recognition system using complex Zernike moments for feature extraction and complex-valued deep neural networks, achieving 90-95% accuracy despite challenges with visually similar signs and varying backgrounds.

Singla et al. (2024) developed a real-time sign recognition system using CNNs trained on ASL datasets, achieving 96.3% accuracy with a multi-headed CNN model reaching 98.981% test accuracy. Kaushik et al. (2024) demonstrated the feasibility of VGG16 and ResNet50 for Sign Language Recognition systems, with VGG16 achieving 99.92% accuracy and ResNet50 achieving 99.95% accuracy.

Despite these advances, critical gaps remain. Previous works have not fully addressed the demand for accurate and efficient real-time sign language translation suitable for real-world applications. Challenges persist in handling environmental variations such as lighting conditions, background noise, and signer-specific variations. Most systems exclude dynamic gestures like letters 'J' and 'Z', and few investigate full sentence recognition or variety across different sign languages.

This research addresses these gaps by utilizing VGG16 CNN architecture for ASL gesture detection from a dataset including letters A-Z and digits 0-9. The project

integrates speech conversion capability, translating recognized gestures in real-time to spoken language. This dual functionality improves accessibility, bridging communication gaps for the hard-of-hearing and deaf community in a scalable and efficient manner.

---

3. METHODOLOGY

This research creates an efficient system for recognizing American Sign Language (ASL) hand gestures using a deep learning approach. The system uses transfer learning as its core, leveraging a pre-trained VGG16 model. The pre-trained VGG16 model was trained on the ImageNet dataset containing millions of photos in 1,000 categories. By using pre-trained weights, the model already recognizes general features like edges, textures, and patterns common in most image types, drastically reducing the time and data needed to train for ASL hand gesture detection across 36 different classes (A-Z, 0-9).

The system integrates a speech conversion mechanism for enhanced accessibility and user-friendliness. Once gestures are identified and matched to corresponding letters, the predicted output is vocalized using a text-to-speech engine. This provides real-time interaction with both textual and auditory output, broadening usability for users with communication challenges.

Data Collection and Preprocessing

The dataset, asl_dataset, was downloaded from Kaggle and specifically designed for recognizing ASL gestures. It comprises 36 classes representing 26 letters (A-Z) and 10 digits (0-9), divided into subfolders with each corresponding to a unique label. The dataset contains 2,515 images, with most subfolders containing 70 images and one containing 65 images, stored in standard formats like JPEG.

Data preprocessing included:

- **Resizing**: Images resized to 224x224 to comply with VGG16 input requirements
- **Normalization**: Pixel values scaled to 0-1 range by dividing by 255.0
- **Data Augmentation**: Techniques including rotation (up to 20 degrees), shifting (up to 20% of image size), shear and zoom effects, and horizontal flipping to increase effective dataset size
- **Train-Test Split**: 80% training, 20% testing

- **Label Mapping**: Association of class labels with numerical indices

Model Architecture

The model architecture uses the pre-trained VGG16 network as its base, frozen to retain robust feature extraction capabilities from ImageNet. Input images are processed through VGG16 layers, producing a 3D feature map with dimensions (7, 7, 512). This is flattened into a 1D vector using a Flatten layer, followed by a Dense layer with 512 neurons and ReLU activation to learn task-specific features. A Dropout layer with 50% rate prevents overfitting. Finally, a Dense output layer with softmax activation and 36 neurons maps features to ASL gesture classes (A-Z, 0-9).

The model was compiled with Adam optimizer for efficient learning, sparse categorical cross-entropy as the loss function for multi-class classification, and accuracy as the performance metric. The total model contains approximately 27.6 million parameters, with 12.8 million trainable in new layers, combining efficiency with versatility for gesture detection.

Hyperparameter Tuning

Key hyperparameters include:

- **Learning Rate**: Determined dynamically by Adam optimizer
- **Batch Size**: 32 samples per weight update
- **Epochs**: 10 complete iterations over training dataset
- **Optimizer**: Adam for adaptive learning rate adjustment
- **Loss Function**: Sparse Categorical Cross-Entropy for multi-class classification
- **Dropout Rate**: 0.5 to prevent overfitting

---

4. EVALUATION AND RESULTS

The evaluation demonstrates the VGG16-based model's effectiveness in recognizing ASL gestures and converting them to text and speech. Training history plots show the model improved rapidly, with training accuracy rising from 50% to over 90% within three epochs. Validation accuracy also improved steadily, indicating strong generalization to unseen data. Training loss decreased

rapidly and stabilized at low values, while validation loss decreased at a slower rate but remained acceptable.

Performance Metrics

The model was tested on 503 samples with impressive results:

- **Accuracy**: 97%
- **Precision (macro)**: 98%
- **Recall (macro)**: 98%
- **F1-Score (macro)**: 97%
- **Weighted Avg F1-Score**: 97%

These metrics demonstrate the model's effectiveness in ASL gesture recognition, with high precision and recall indicating accurate identification and categorization with minimal errors.

Testing and Real-Time Performance

From 503 test samples, the model achieved 97% overall accuracy, correctly identifying individual gestures such as 'C,' 'A,' and 'T'. The final word "CAT" was formed and vocalized clearly using the text-to-speech engine, demonstrating reliable real-time performance for practical applications.

Real-time testing utilized a camera to capture live hand gestures processed by the trained VGG16 model. Python implementation completed tasks including image processing, gesture prediction, and text-to-speech conversion using the pyttsx3 library. The system successfully recognized various gestures and simultaneously vocalized recognized words, making the setup efficient and user-friendly.

Challenges encountered during real-time testing included variations in lighting affecting accuracy, background clutter affecting gesture clarity, and the need for precise positioning within the camera's view. Despite these challenges, the model showed consistent performance for clear gestures, and TTS integration significantly improved usability for ASL users through combined visual and auditory output.

---

5. CONCLUSION

The VGG16 model proved highly efficient in identifying ASL hand gestures. Accuracy and loss plots indicated fast learning and strong performance on both training and new data. Performance metrics of 97% accuracy and 98% precision and recall establish the model's capability for accurate gesture classification. Testing demonstrated successful identification of gestures to form words like "CAT," which were clearly vocalized using text-to-speech.

While small challenges existed—including lighting changes and background distractions—the model worked reliably with clear gestures. Future enhancements should include continuous gesture recognition for dynamic ASL sentences, improved robustness against diverse lighting and background conditions through dataset augmentation, and optimization for deployment on portable devices like mobile phones or embedded systems. Overall, the results demonstrate this system as a strong and practical solution for ASL recognition, with significant potential to bridge communication gaps and enhance accessibility for the deaf and hard-of-hearing community.