Interpreting Sign Language Using CNN-based Models

Maximilian Du

Introduction

Sign language is a well-established method of communication among the hearing impaired. However, such language can be very difficult to understand to the outsider. For spoken languages, simple text-to-text mapping services help to bridge the communication barrier, but because sign language is neither spoken nor written, a new method of translation is needed. To address this problem, this project uses computer vision models in the form of convolution neural networks to visually identify sign language gestures from video.

Previous Research and Dataset

Dataset and Motion Energy Usage

- Data obtained from the open-source *American Sign Language Lexicon Video Dataset (ASLLVD)*
- Hundreds of labeled videos with 4714 unique signs made up of 82 gesture types
- Makers of ASLLVD used non-binary Motion Energy diagrams, but only to correlate with known samples
- Makers of ASLLVD also used word-level encodings instead of gesture-level encodings, which resulted in hindered scalability

MotionSavvy—Leap Motion Sign Language Translation

- Used Leap Motion tracking hardware to track hand position and translate sign language
- Requires specialized hardware

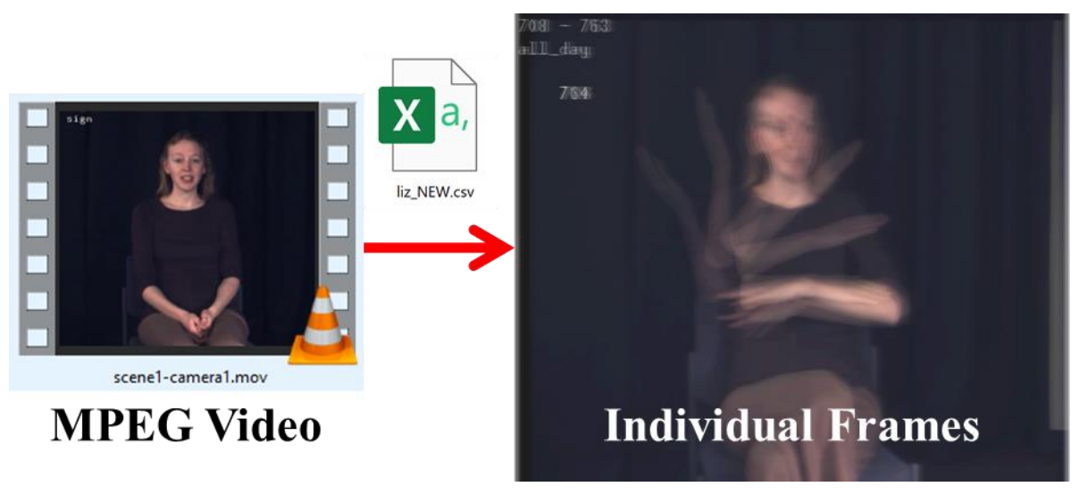
Objectives

1. Represent video frames of a gesture as a single frame while retaining crucial gesture information
2. Use a Convolutional Neural Network (CNN) to classify each processed frame's dominant and non-dominant gestures
3. Use classified gestures and gesture-to-word dictionary to identify signed word

Data Processing

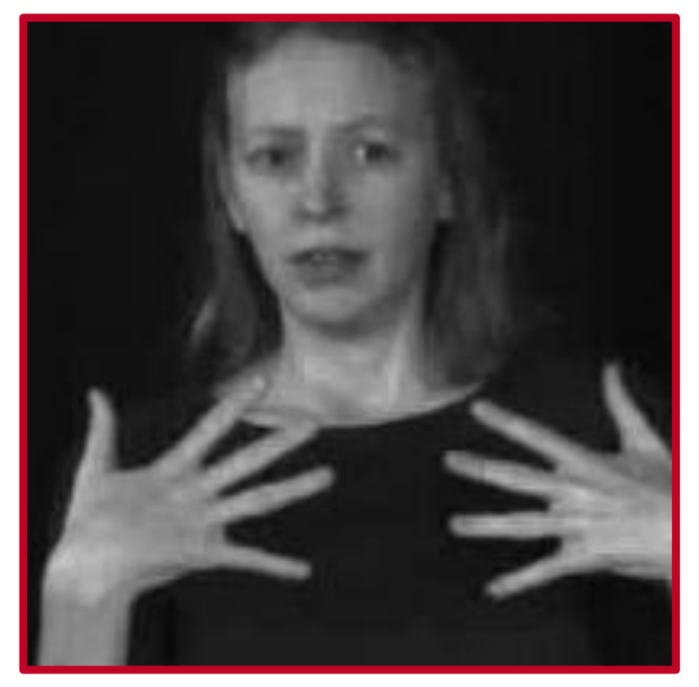
Video Parsing

- Labeled video (640 × 480 @ 60 fps) was split into individual frames
- Gestures ranged from 20 to 180 frames in length
- Cropped and downsized to 100 × 100 greyscale frames
- Only single gesture words were used to reduce complexity



Center Sampling

- Primitive feature extraction: the frame from the middle of the video was used to represent the gesture
- Advantages: image processing is minimal, hands are clear
- Disadvantages: loses temporal information

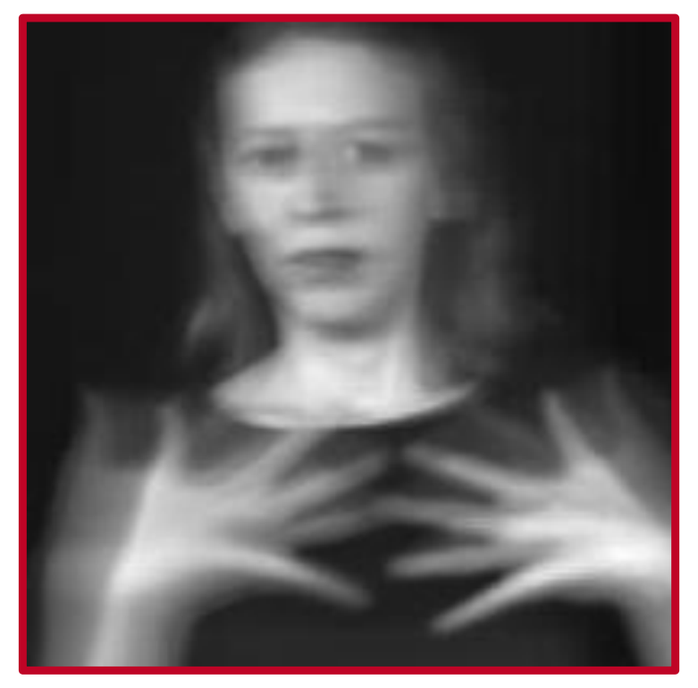


Frame Overlap

- Frames are averaged by pixel, meaning that each frame has an equal contribution to the final image

$$Z(x,y) = \frac{1}{N}\sum_{t=0}^{N} P_t(x,y)$$

- Advantages: captures motion and some detail
- Disadvantages: noisy image, faster movements blurred

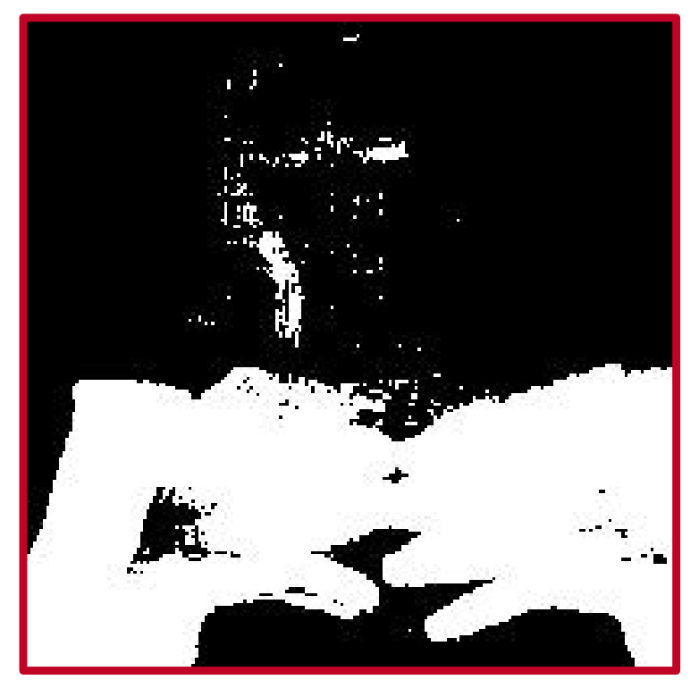


Binary Motion Energy Diagrams

- Marks changing pixels

$$Z(x,y) = \bigcup_{t=1}^{N} \min(1, |P_t(x,y) - P_{t-1}(x,y)|)$$

- Advantages: simple, effective segmentation with only two possible pixel states
- Disadvantages: loses temporal order and detail



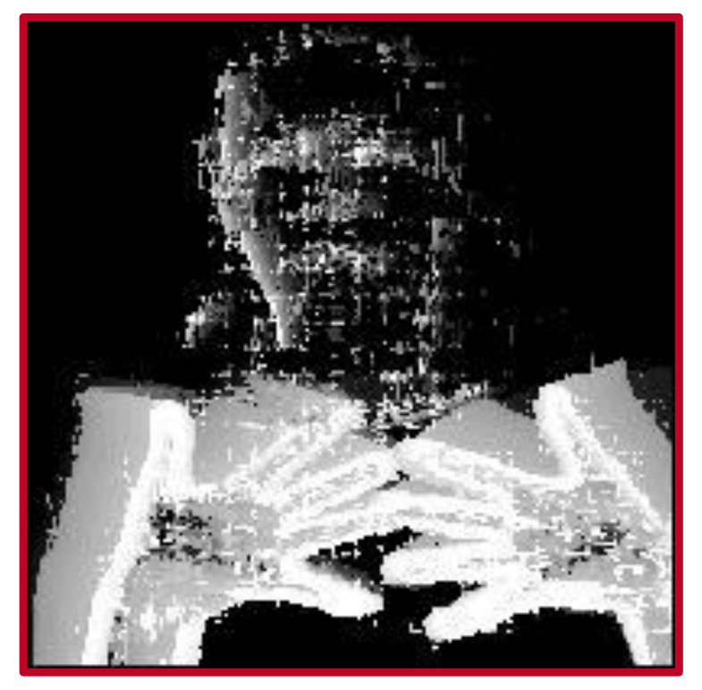
Motion History Diagrams

- Marks changing pixels with decay inversely proportional to video length

$$D_t = \min(1, |P_t(x,y) - P_{t-1}(x,y)|)$$

$$Z_t(x,y) = \begin{cases} t, D_t = 1 \\ \max\left(0, Z_{t-1}(x,y) - \frac{K}{N}\right), D_t \neq 1 \end{cases}$$

- Advantages: segments area of motion while showing motion direction and emphasizing recent motion
- Disadvantages: even with activation thresholds, output is noisy



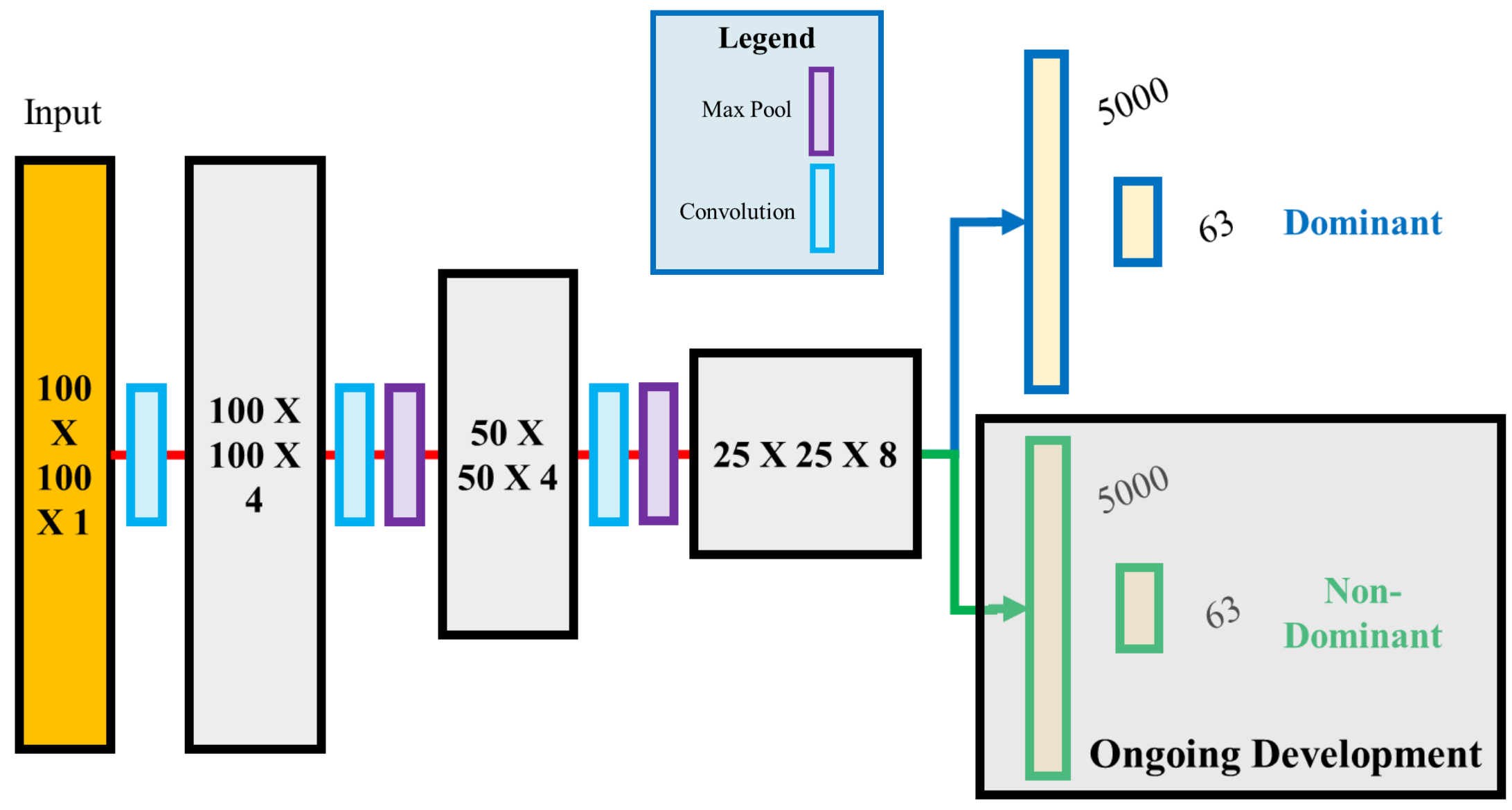
Model Architectures

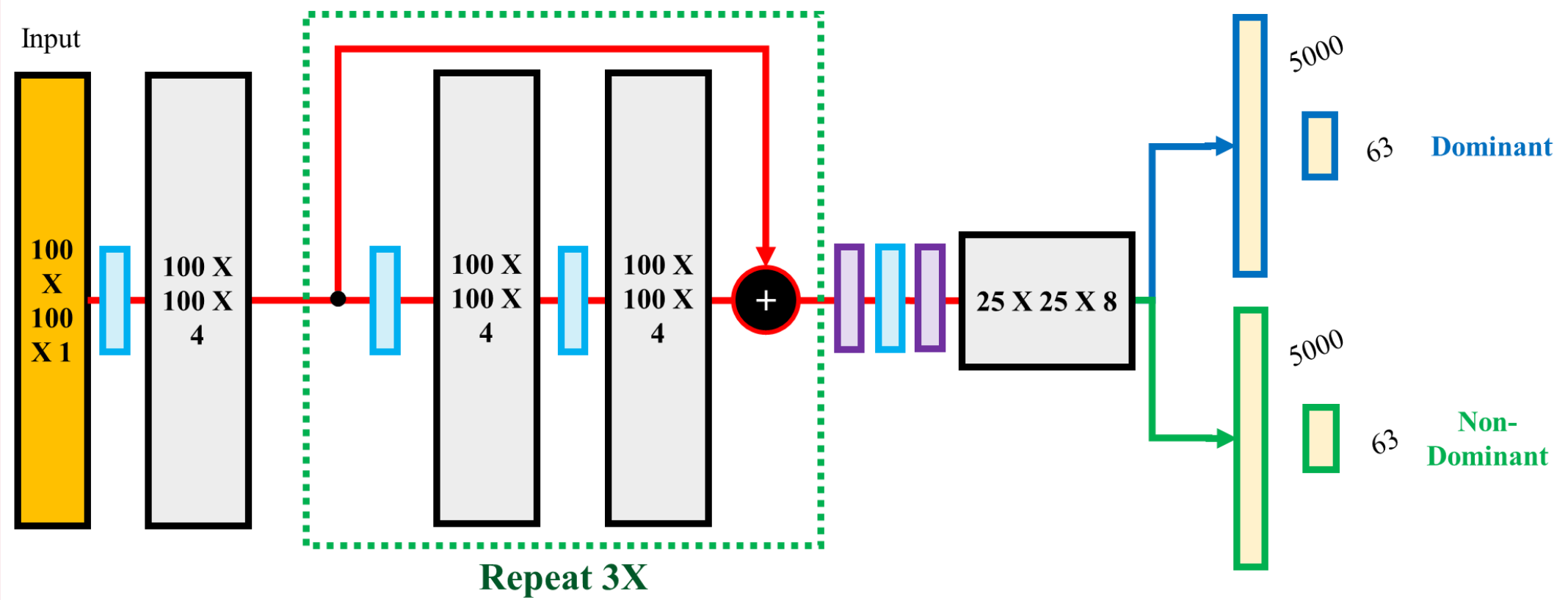
Goals

- Make models that take in images and output a probability distribution of possible gestures for both hands
- Use the same feature extraction layers for both dominant and non-dominant hand signs
- Combine benefits of two different video processing techniques by creating a dual-input CNN model with separate feature extraction

Standard CNN

- 3x3 filter
- Feature encoding through the convolutional section
- Gesture extraction with independent fully-connected layers
- Increased number of layers decrease accuracy
- Not pooled on the first layer to increase layer count



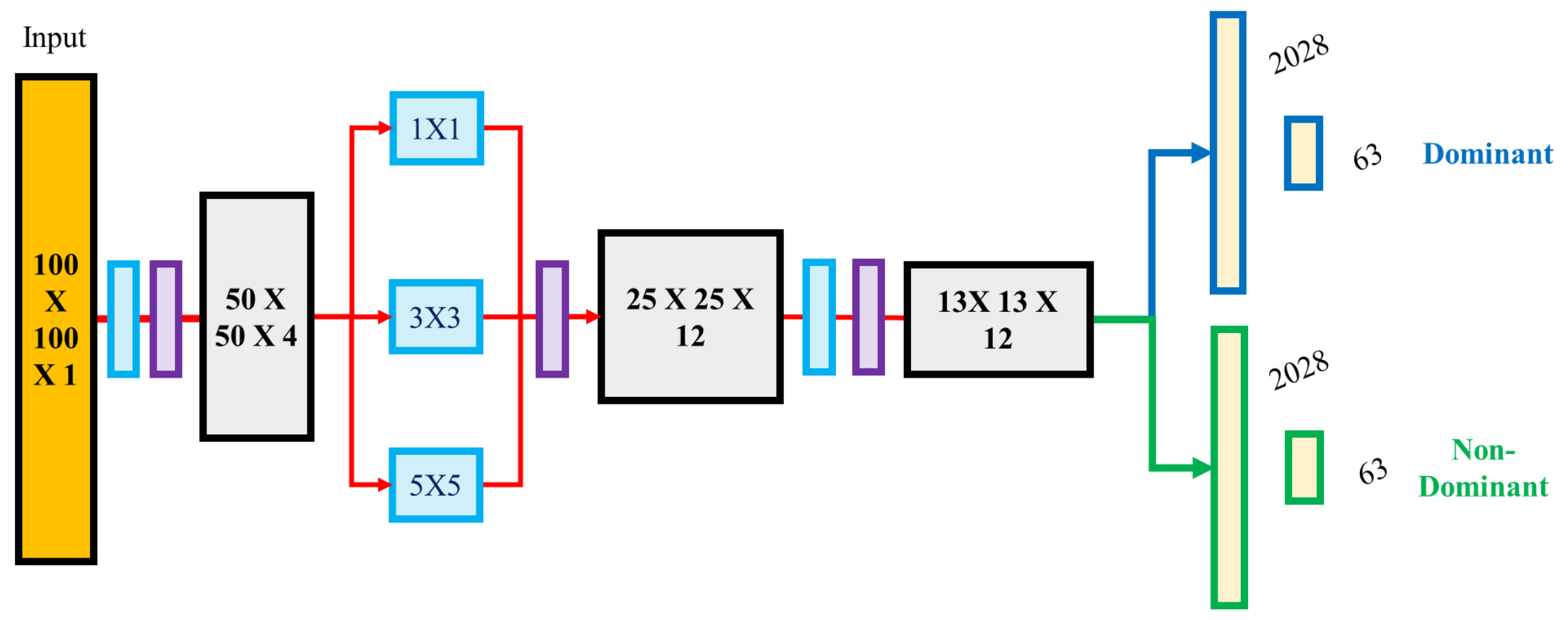


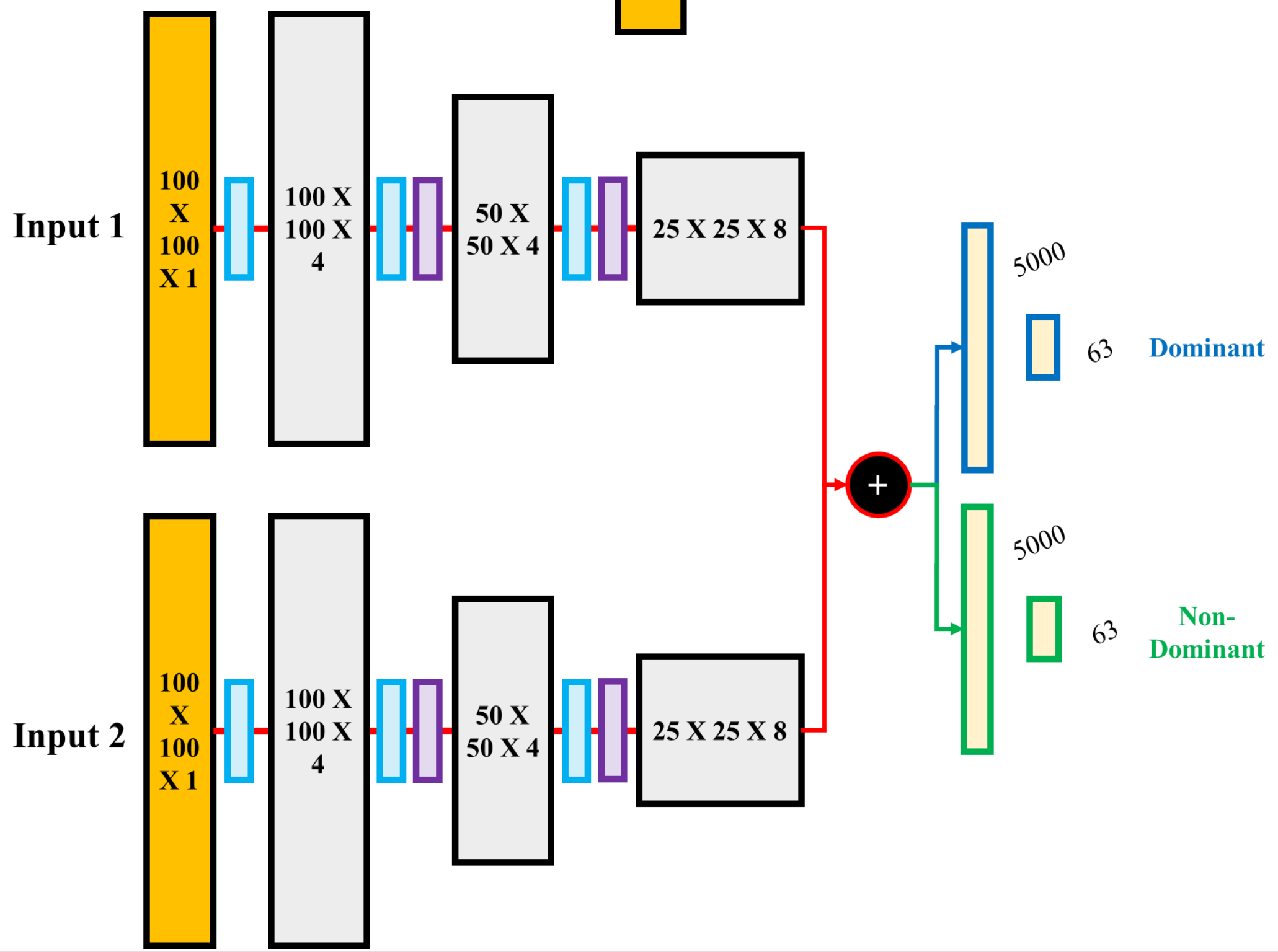
Residual CNN

- Allows layer bypass, reducing the vanishing gradient problem
- More layers can be added
- At risk of overfitting due to high layer count

Inception CNN

- Reduces dependence on consistent image scale
- Trades depth for breadth and avoids vanishing gradient





Proposed: Dual-Input CNN

- Allows more than one type of processed image to be used
- *Motion History* and *Center Sample* were used as the two inputs
  - History provides whole-video snapshot
  - Center sample provides detailed mid-gesture snapshot

Data Feeding, Training, and Evaluation

Representing Gestures

- One-hot (single high value) vector label representation
- Future expansion to multiple-gesture videos can use a multiple-hot vector encoding

Dataset Balancing

- Imbalanced number of gestures (200:1 ratio) that persisted even as more data was added
- Selective augmentation was a naïve solution
  - Augmented rarer gestures more than abundant ones
  - Enabled proper convergence, but led to some overfitting

Augmentation

- Random horizontal/vertical translation (±5 px)
- Addition of uniform noise

Training

- Cross entropy loss function
- Batch gradient descent to reduce optimization noise
- Used 1200 non-augmented (12000 augmented) signs
- L2 regularization losses

Testing

- 100 untrained gestures from augmented pool as test set
- Evaluated prediction accuracy

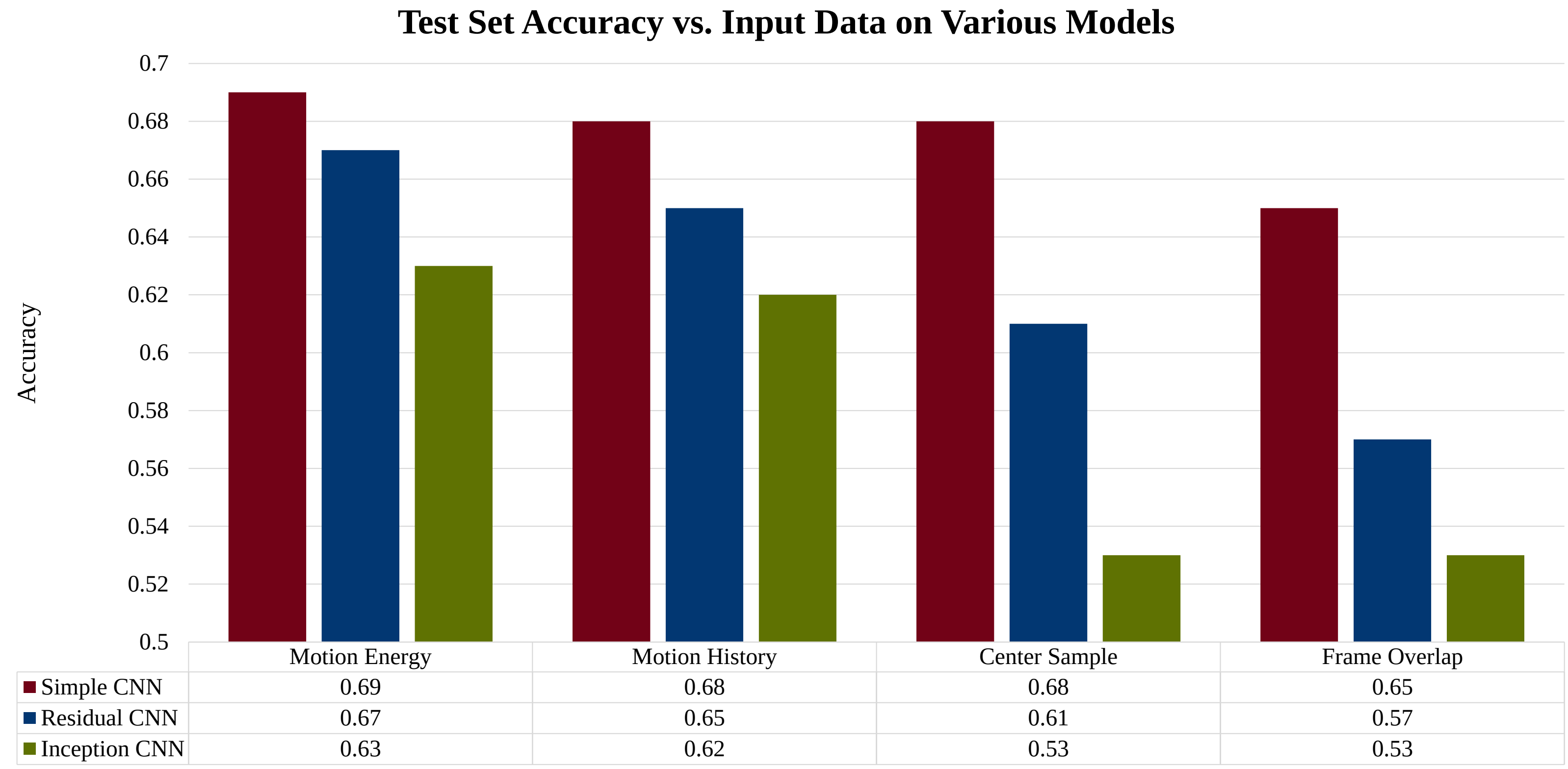
Results and Analysis

Hyperparameters			
Epochs	Batch Size	Learning Rate	L2 Weights
1000	150	0.0025	0.05

Evaluation Objectives

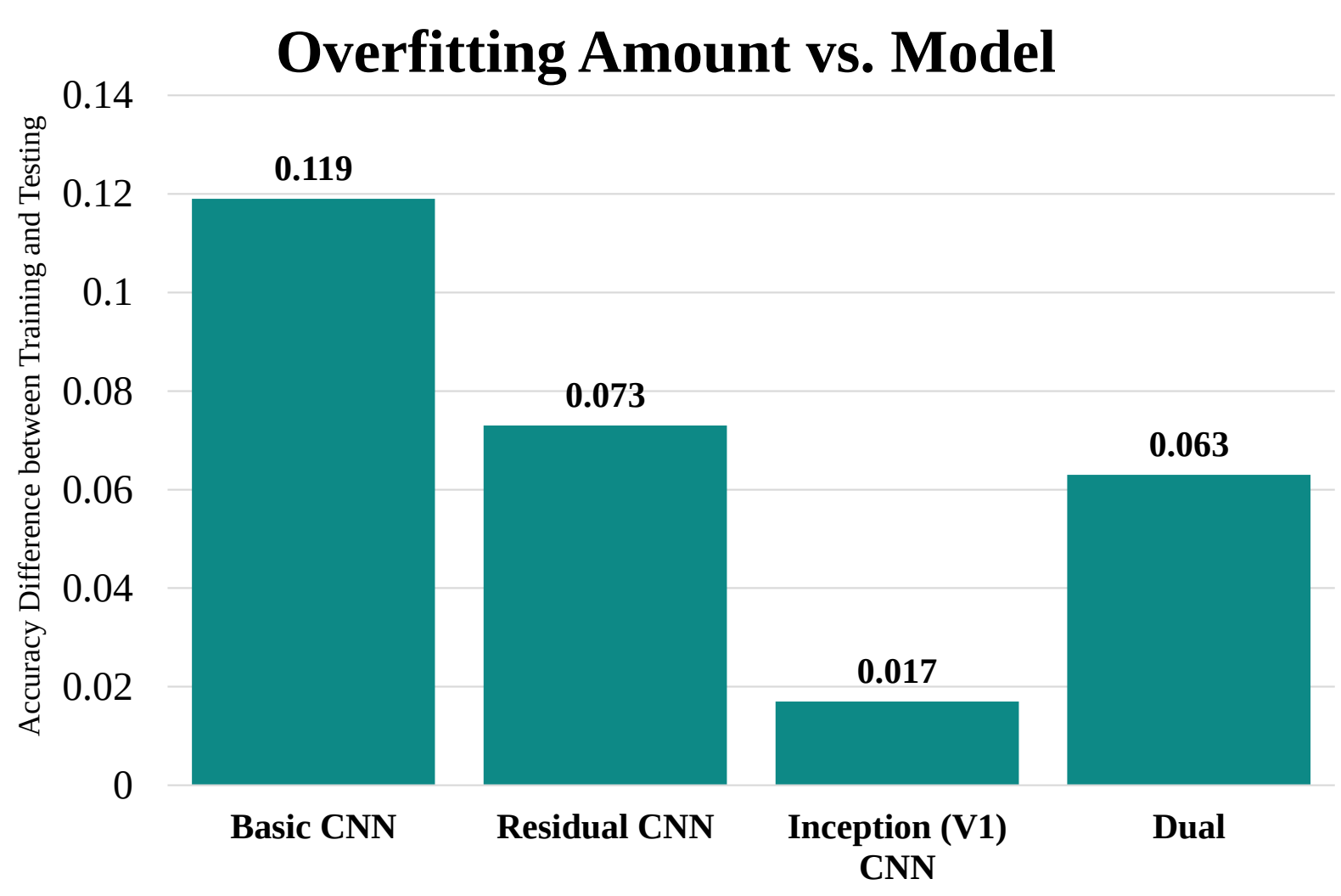
- Search for models that have high accuracies with a low difference between training and testing performance
- Find the optimal video processing algorithm by comparing accuracies across all single-fed models

Test Set Accuracy vs. Input Data on Various Models

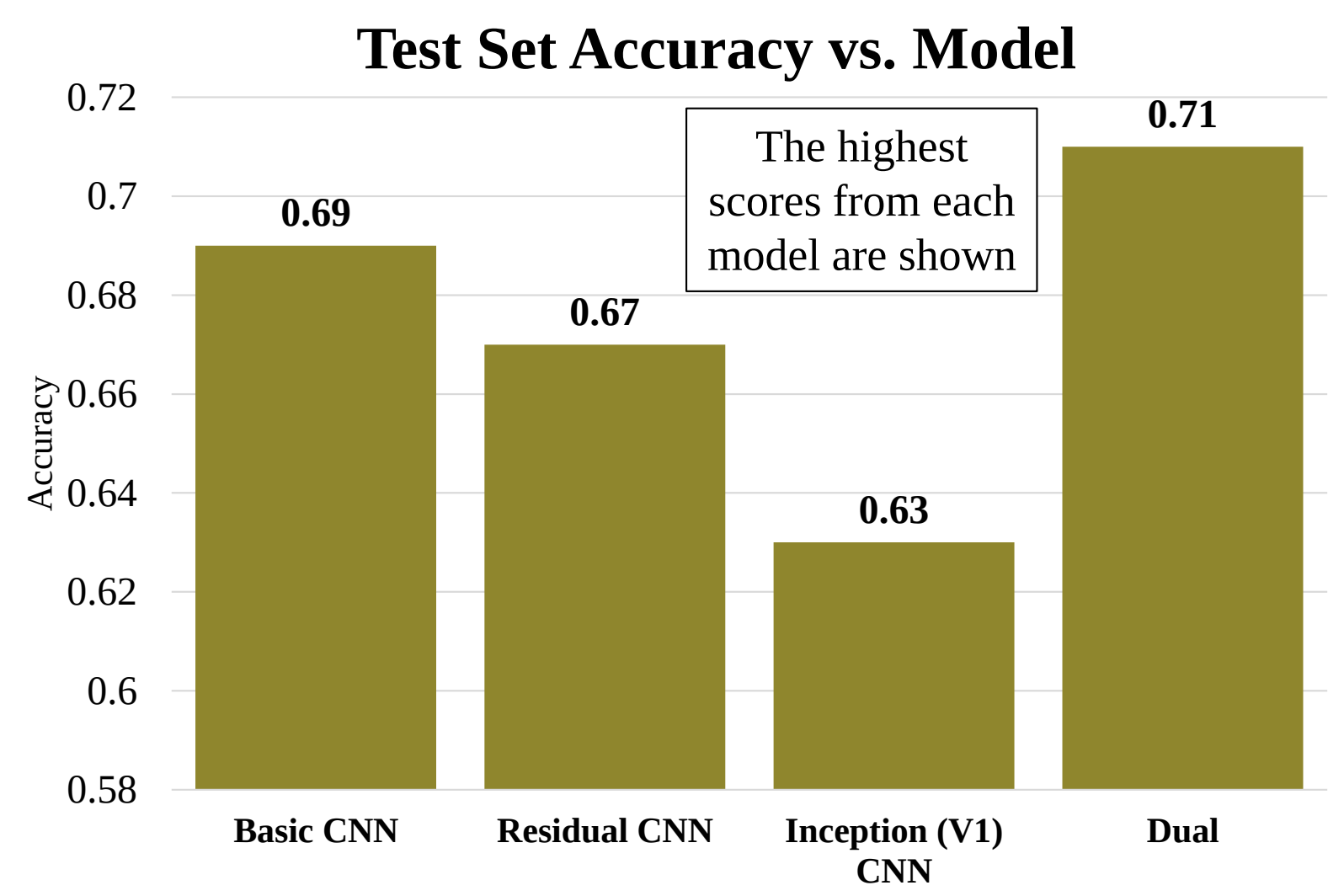


	Motion Energy	Motion History	Center Sample	Frame Overlap
■ Simple CNN	0.69	0.68	0.68	0.65
■ Residual CNN	0.67	0.65	0.61	0.57
■ Inception CNN	0.63	0.62	0.53	0.53

Overfitting Amount vs. Model

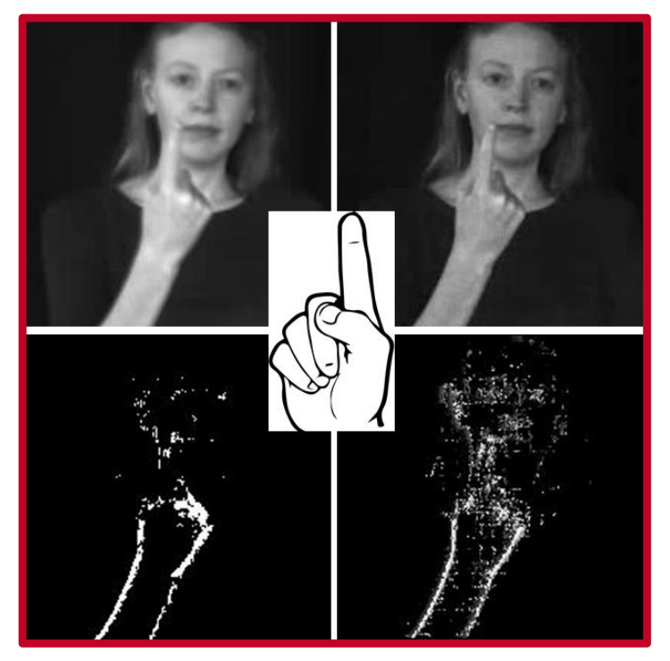


Test Set Accuracy vs. Model



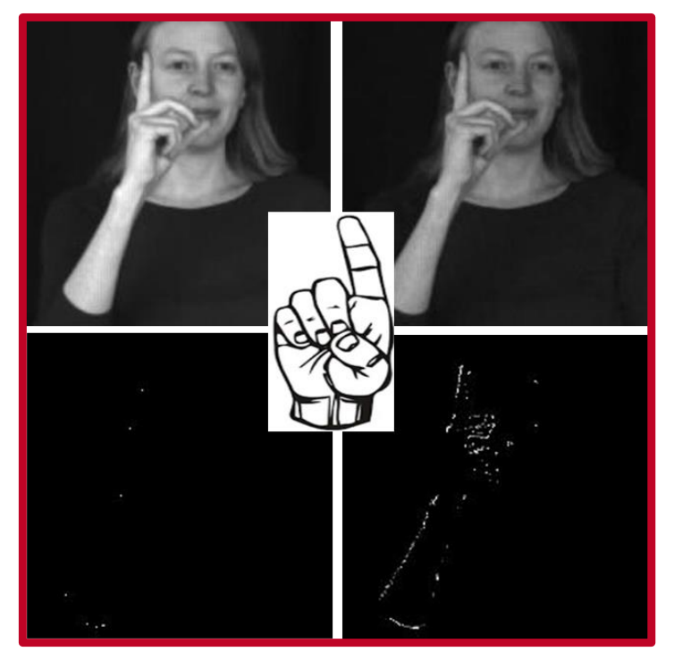
The highest scores from each model are shown

Top Four Missed Gestures



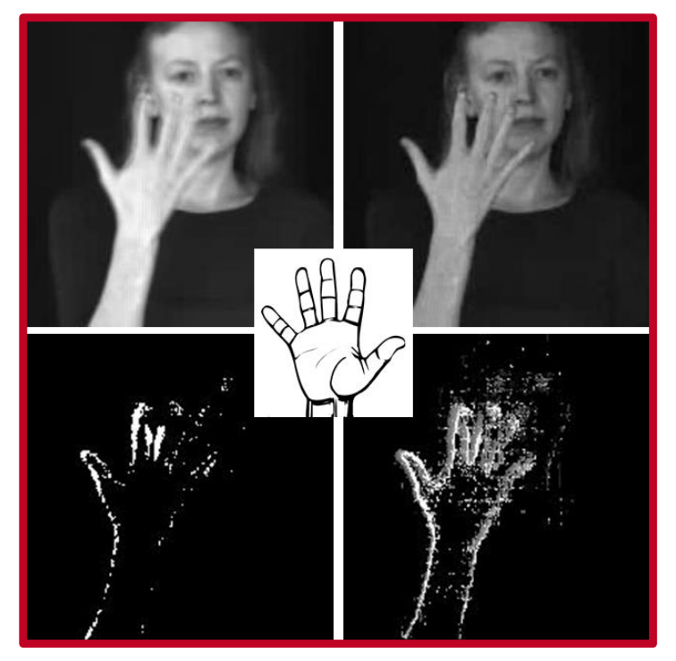
One

- Very static
- Results in low quality motion energy diagram
- Similar to other hand signs



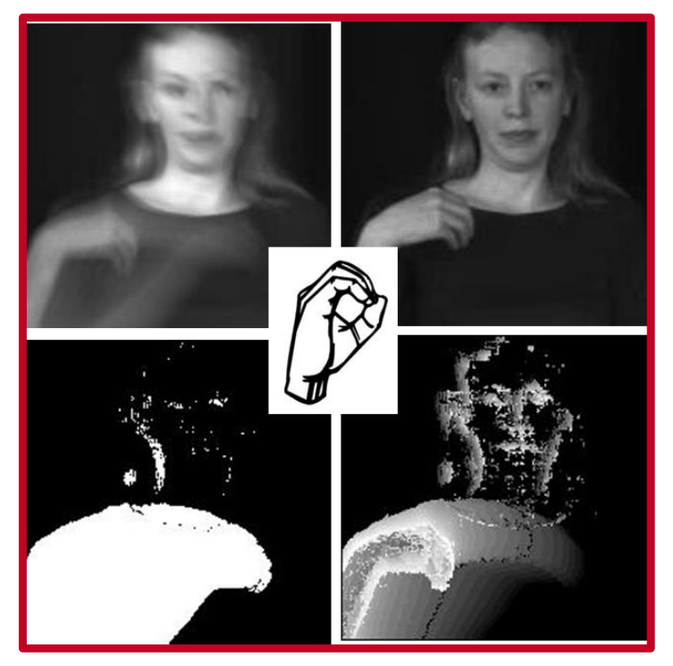
"D"

- Lack of motion resulted in no contours in motion energy image
- Similar to "One"



Five

- Also lack of motion
- Caused higher noise levels and unclosed contours on motion energy



Flat "O"

- Entire body moves, causing excessive noise
- Gesture is at a sub-optimal angle of view

Images Source: https://www.yescoloring.com/learn-sign-language.html

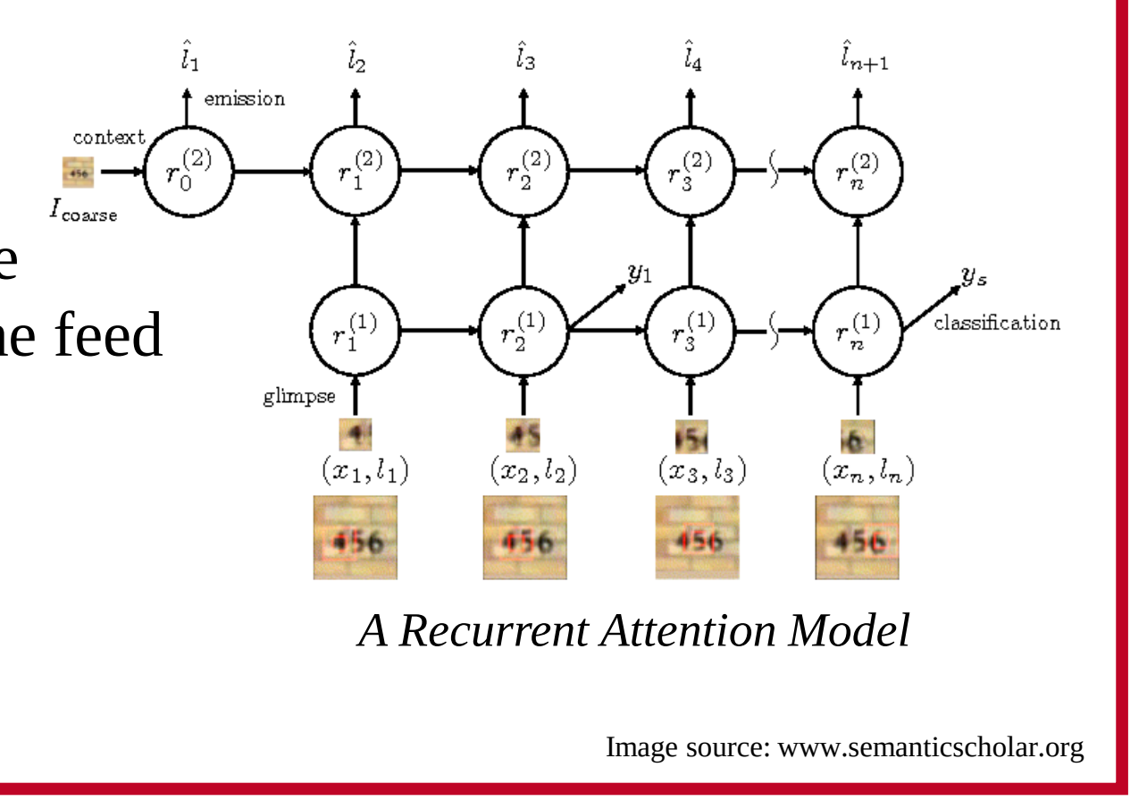
Conclusion and Future Directions

Conclusion

- The final model is affected more by the complexity of the input image than the information it contains
  - Motion Energy was the best video representation algorithm
- Inception CNNs are harder to train but they are less prone to overfitting
- Dual models have higher accuracy than Inception CNNs but are more prone to overfitting

Future Directions

- Deep reinforcement learning—recurrent attention model
  - Can be trained on an imbalanced dataset and is more adaptive
- Experimentations with CNN-LSTM combination and direct frame feed
- Interpretation of facial expressions
- Real-time implementation
  - Rolling motion history/motion energy generation
- Expansion of dataset to further reduce overfitting



A Recurrent Attention Model

Image source: www.semanticscholar.org