Form 3: Methodology

1. Team No: 20

2. Project Title: Multi-Modal Assistive System for people with disabilities

3. Proposed Method:

SNO	Module	Proposed Method
1	Sign language translation	**Two-Stream Mixed Convolutional Neural Network**
2	Visual Question Answering on Images	**Multimodal Transformers**
3	image-to-speech and speech-to-text	**Optical Character Recognition**

4. Proposed Method illustration

a. Sign language translation:

- **The methodology involves using a Two-Stream Mixed Convolutional Neural Network. Its architecture is detailed below.**

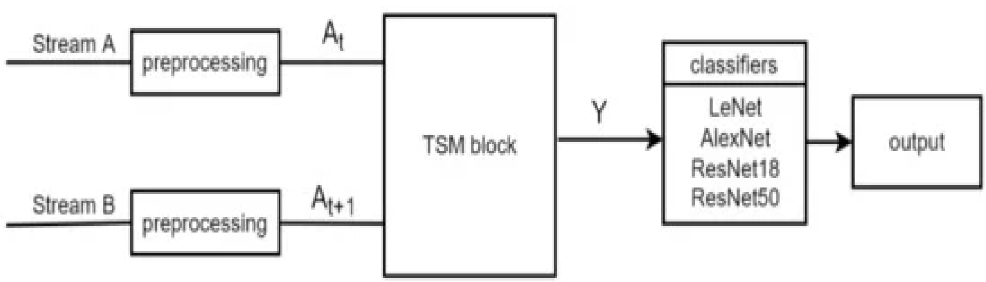


Fig 1. The 2S-CNN structure.

b. Visual Question Answering on images (Multimodal Transformers):

- **The methodology involves the following three phases**

   i. Featurization of Image and Question:

Model	Hugging Face Model Name	Description
ViT	google/vit-base-patch16-224-in21k	Vision Transformer (first image transformer encoder, trained on ImageNet)
DeiT	facebook/deit-base-distilled-patch16-224	Data-Efficient Image Transformer (more efficiently trained transformers for image classification, requiring much less data and computing resources compared to ViT)
BEiT	microsoft/beit-base-patch16-224-pt22k-ft22k	Bidirectional Encoder representation from Image Transformers (regular vision transformer, but pre-trained in a self-supervised way rather than supervised)

Model	Hugging Face Model Name	Description
BERT	bert-base-uncased	Bidirectional Encoder Representations from Transformers (basic BERT)
RoBERTa	roberta-base	Robustly Optimized BERT Pretraining Approach (builds on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with larger mini-batches and learning rates)
ALBERT	albert-base-v2	A Lite BERT (consumes lower memory and increases training speed of BERT by splitting the embedding matrix into two smaller matrices and using repeating layers split among groups)

Fig 2. Pretrained image transformers for experimentation to provide visual features.

Fig 3 Pretrained text transformers for experimentation to provide textual features.

ii. Feature Fusion:

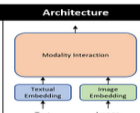
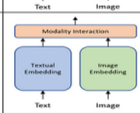
Model Type	Architecture	Single Modality Features	Modality Interaction	Training
Early Fusion		Very simple Example: GLoVe, CNNs	Complex Examples: Cross-Modal Transformers	Slow & difficult to train
Late Fusion		Heavy Example: Pretrained Transformers	Very simple Example: Concatenation + Single hidden layer	Simple to train

Fig 4. Types of multimodal data fusion

iii. Answer Generation:

answer generation involve a simple classifier for one-word/phrase answers within a fixed answer space.

c. OCR- powered image-to-speech and speech-to-text:

i. Image to text :

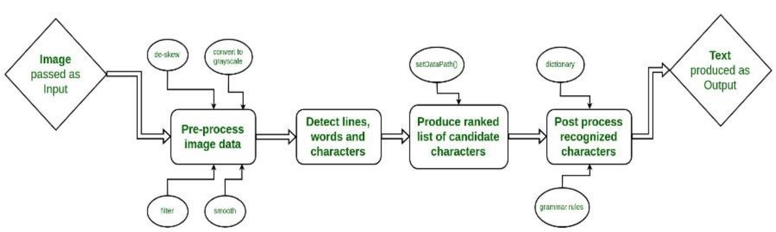


Fig 5. General Working of OCR

ii. Text to speech/ speech to text:

In order to implement this feature, we leverage existing speech recognition and text-to-speech engines through Python.

5. Parameter Formulas

a. Sign language translation:

1. The feature extraction in TSM is calculated with the following equation: $Ht=\sum j\sum kWi[j,k]At[a-j,a-k]$
2. The feature map Z in TSM is calculated with Equation:.
   $Z=Ht+Ht+1=\sum j\sum kWA[j,k]At[a-j,a-k]+WB[j,k]At+1[a-j,a-k]$
3. The feature map Y in TSM is calculated with Equation (3):
   $Y=OutputTSM=\sum l=1c-lZi\&\sum c-lcHt$

b. Visual Question Answering on Images:

1. Learning rate decay**: w_i^(t+1) w_i^(t)-α*∇L(w)/∇w_i^(t)**
2. Adam optimizer: **w_{t+1} = w_t - α * m_t / (√(v_t + ε))**

c. OCR- powered image-to-speech and speech-to-text:

1. Kernel Function : f (x) = sgn(X 1 i=1 αiyiK(xi, x) + b)
2. Finding the probability of nearest sample : p(y|q) = P k∈K Wk .1(ky=y)/ P k∈K Wk

SUPERVISOR:

G.Kiran Kumar

Assitant Professor

TEAM DETAILS:

C.Chandhana 20eg105310

K.Sri pavani 20eg105323

M.Vamshi Krishna 20eg105328

S.Jhansi 20eg105348