Detecting Letters and Words from Bangladeshi Sign Language in Real-time

Ahsanullah University of Science and Technology

Presented By :

Oishee Bintey Hoque	15-01-04-005
Al-Farabi Akash	15-02-04-010
Md. Saiful Islam	15-01-04-027
Alvin Sachie Paulson	15-01-04-035

Supervised By :

Mr. Mohammad Imrul Jubair
Assistant Professor,
Department of Computer Science and Engineering,
Ahsanullah University of Science and Technology.

Signers Communication

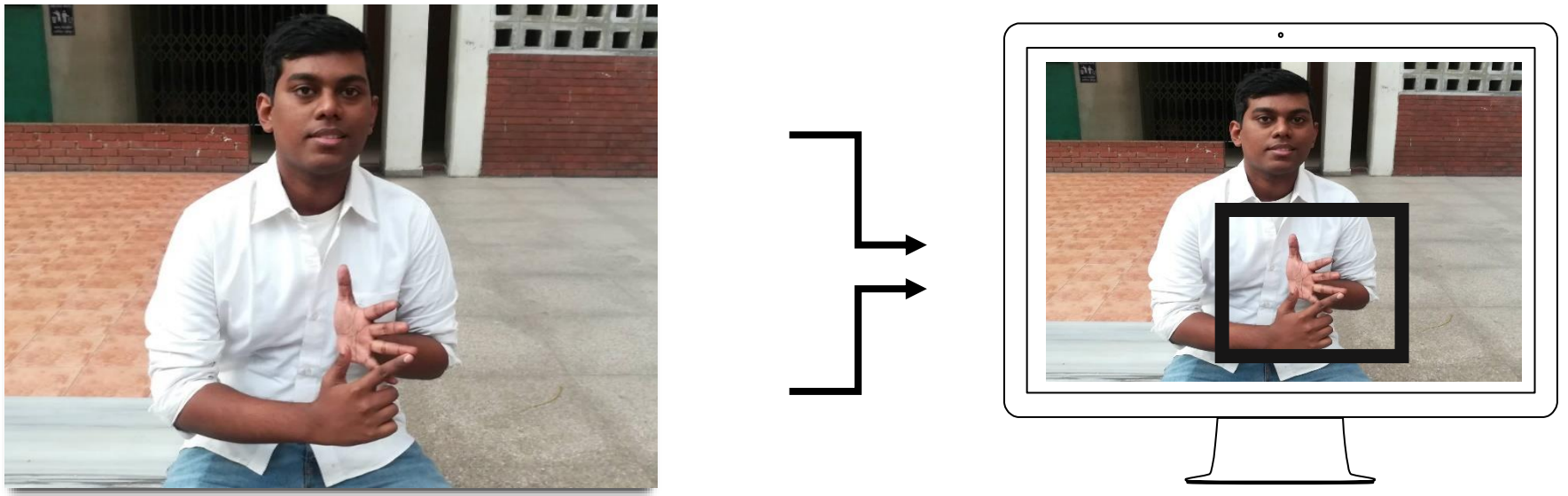
How can a **non-signer** communicate with a **signer**?

- A real-time interpreter might be a possible solution

Signers Communication

How can a **non-signer** communicate with a **signer**?

- A real-time interpreter might be a possible solution



Research Problem

- **Bangladeshi Sign Language Detection**
  - Letters
  - Words
  - In real-time

- **How can we implement that?**
  - Using **Deep Learning** based
    - Object Detection and localization method
    - Action Recognition method

Existing Works

- Letter based
- Word based

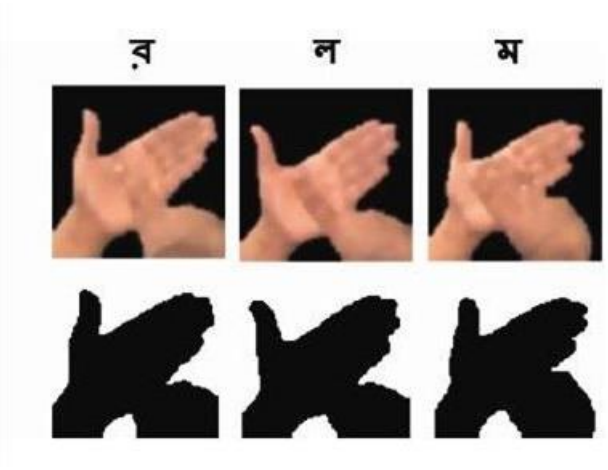
Letter Based Works

To best of our knowledge -

- **Rahman et al. 2018**
  - “Bangla Language Modeling Algorithm For Automatic Recognition of Hand-Sign spelled Bangla Sign Language”
- **Yasir et al. 2018**
  - “Bangla Sign Language Recognition Using Convolutional Neural Network”
- **Ahmed et al. 2016**
  - “Bangladeshi Sign Language Recognition Using Fingertip Position”

Reviews

- Additional device for input
- Dataset
  - Pre-processing
  - Background & angle variation limitation
- Methodology
  - Traditional machine learning method
  - Manual feature extraction
- Output
  - Similarities among signs give faulty recognition
  - All of them are not real-time



**Images are collected from Internet and Rahman et al.*

Word Based Works

To best of our knowledge -

- **Yangho Ji , Sunmok Kim, Ki-Baek Lee 2017**
  - “Sign Language Learning System With Image Sampling And Convolutional Neural Network”
- **Brandon Garcia, Sigberto Alarcon Viesca 2016**
  - “Real-time American Sign Language Recognition With Convolutional Neural Network”
- **Sarfaraz Masood, Adhyan Srivastava, Musheer Ahmad 2018**
  - “Real-Time Sign Language Gesture (Word) Recognition From Video Sequences Using CNN and RNN”

Reviews

- Dataset on the same background
- Correct classification is not possible without high number of features
- Used Dee

Challenges

- Developing dataset for letters and words
  - No dataset on BdSL sign letter or word is available
  - No existing work on BdSL word

- Making the dataset from scratch
  - Gather subjects
  - Ensure lighting conditions
  - Different backgrounds

Challenges

- Very few work done in this area using Deep Learning
- Needs Deep Learning based recognition methods
  - Dependent on Robust Dataset
    - Large number of images/class
    - Enough variation in input data in terms of
      - Background, Gesture Angle, Age and Gender

Contribution

- BdSL-letters dataset
- BdSL-words dataset
- Both datasets are available for further research

Contribution

- System for BdSL letters recognition
  - Using Faster-RCNN

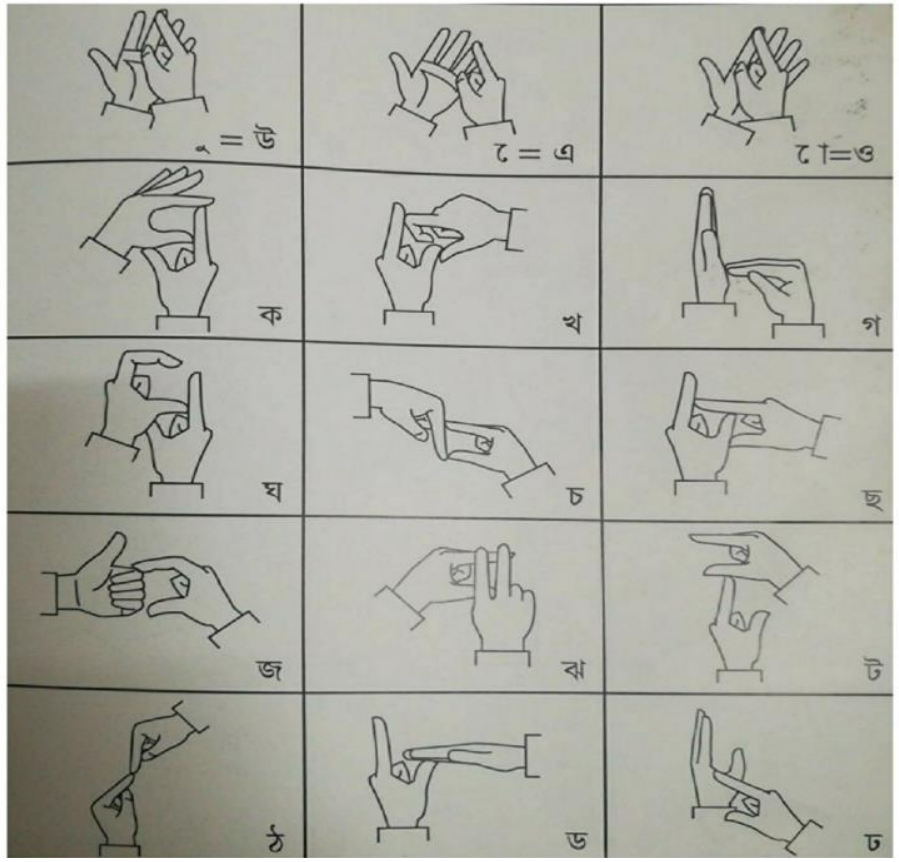
- System for BdSL words recognition
  - Using LSTM

Dataset

- BdSL dataset is not available anywhere.
- Dataset required are two types:
  - BdSL Letters- BdSLImset
  - BdSL Words- BdSLVidset

Dataset

- Our datasets have been verified by –
  - **DHAKA BADHIR HIGH SCHOOL , Paltan**
    (ঢাকা বধির হাই স্কুল)



BdSLImSet

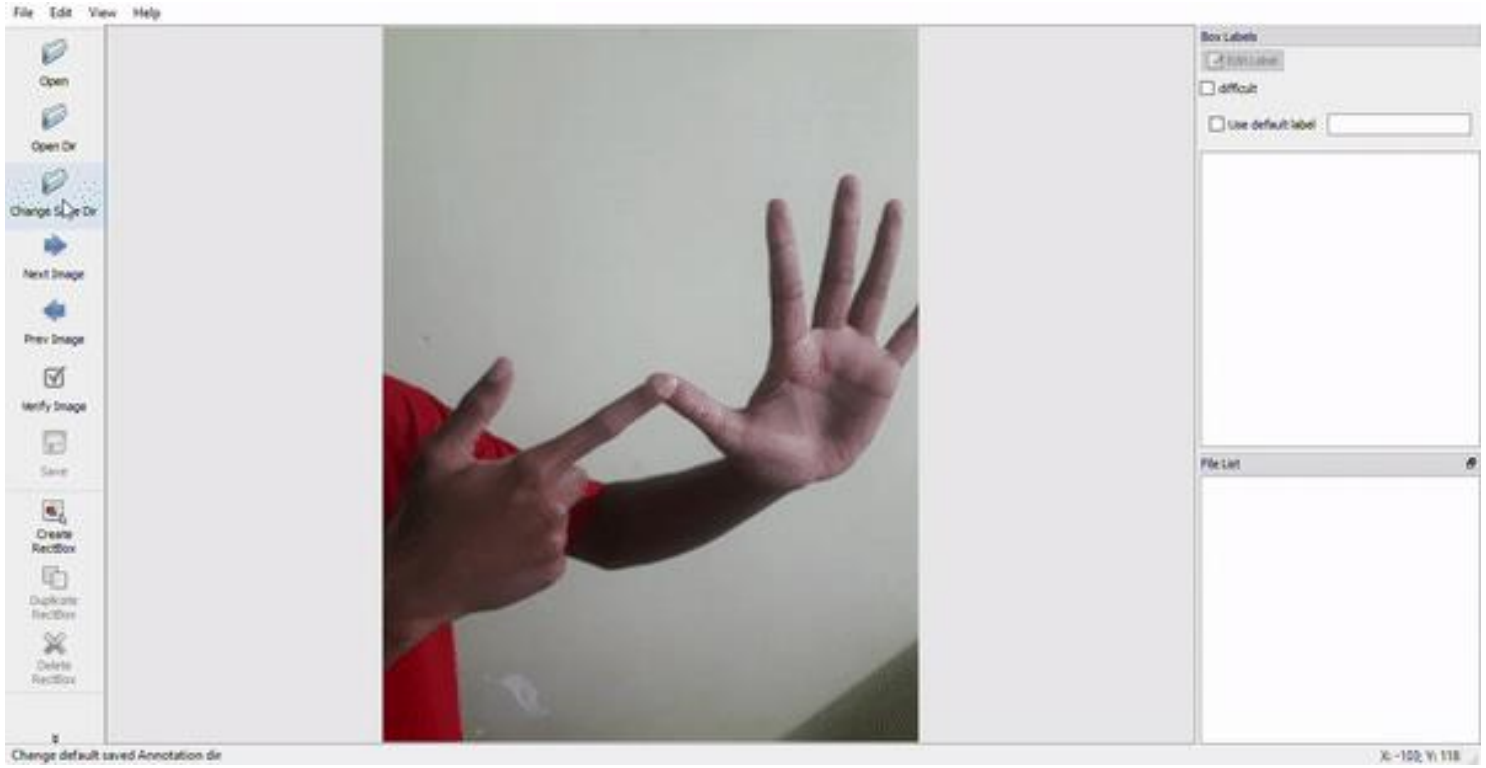
- Background variation
- Different non-signer subjects
- Different angles
- Lighting conditions

Fig : Samples from our Dataset

Fig : Variation in a single letter from our Dataset

BdSLImSet: Labelling

- Individually labelled all 2000 images



BdSLImSet: Labelling

- Converted into **XML** files
- File name
- Image size (height and width)
- Class name
- Bounding box dimensions (x_{min}, y_{min}, x_{max}, y_{max})

Fig: A labelled dataset image.

BdSLImSet: Dataset specifics

Total Images	Total Class	Images/ Class	Image Size	Resolution	Number of Participants	Training Set: Testing Set
2000	**10**	**200**	**≤200kb**	**≤700*1280**	**10**	**8:2**

*Available in **Github** (*https://github.com/imruljubair/bdslimset*)

BdSLVidSet

- Different non-signer subjects
- Even Lighting condition
- Maintain computational simplicity and less time consumption

BdSLVidSet

- White background, even lighting
- For hand tracking simple red colored glove used.
- No external electrical component needed
- So the system can be operated with simple image processing

BdSLVidSet

- Total of 200 video needs to be converted to 40 frames
- So we are working with a total of 8000 frames

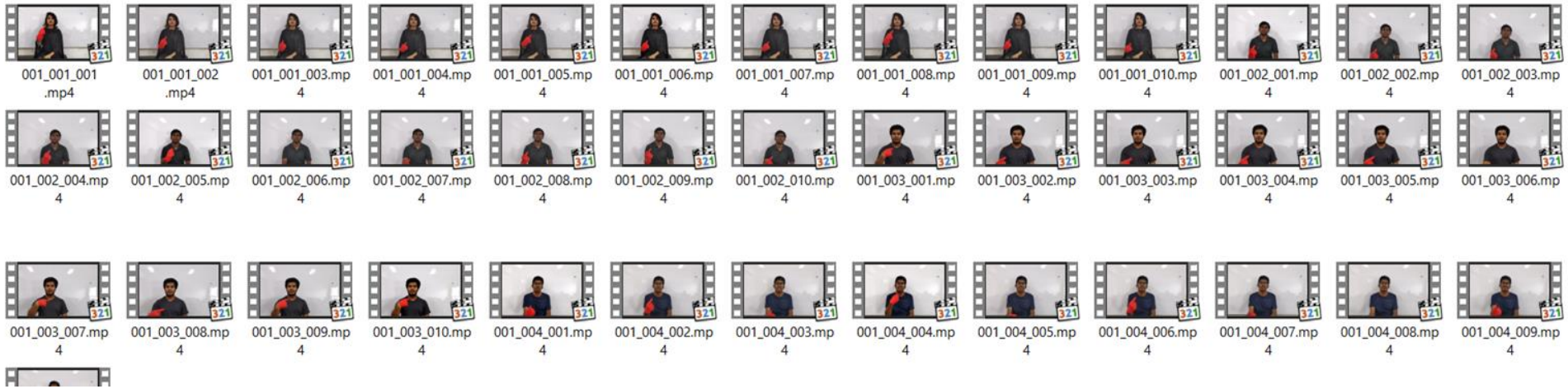


Fig : The sequence of images of a video gesture belonging to class ‘Tumi (You)’.

BdSLVidSet

- BdSL word is separated as a sequence of images

Fig : The sequence of images of a video gesture belonging to class ‘Tumi (You)’.

BdSLVidSet

- Position of the hand recorded, we can simply remove all the background and only track the palm movement.
- With the variation of palm movement the words are recognizable.

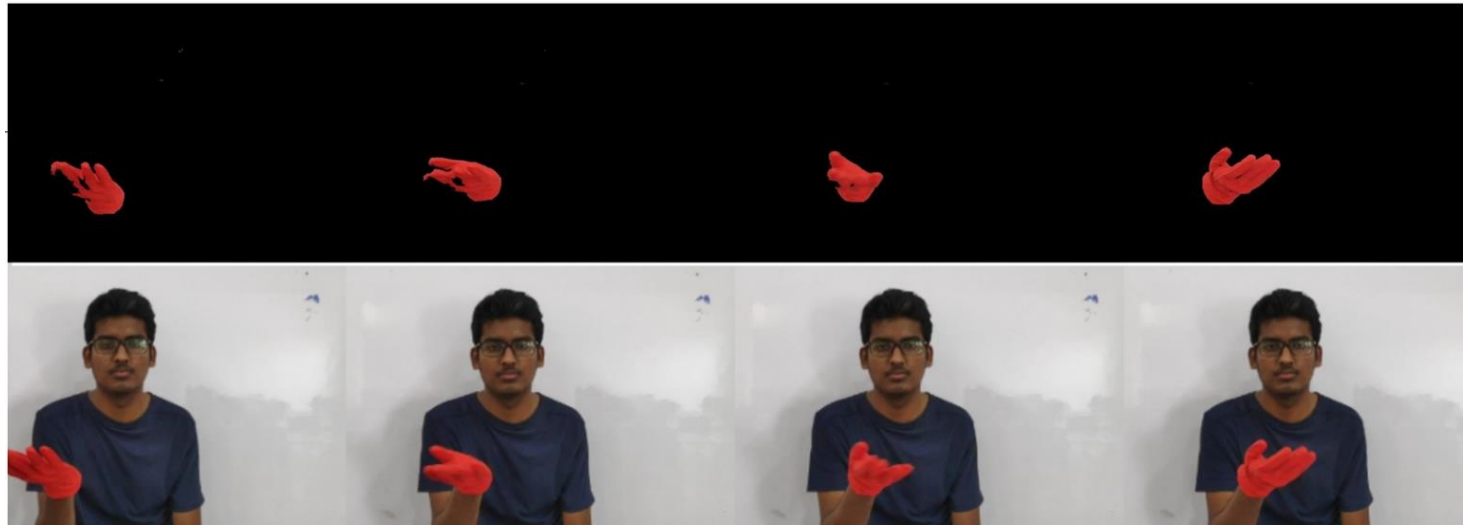


Fig : Sample before and after background removal of class ‘Kemon’.

BdSLVidSet: Gray Scale Frames

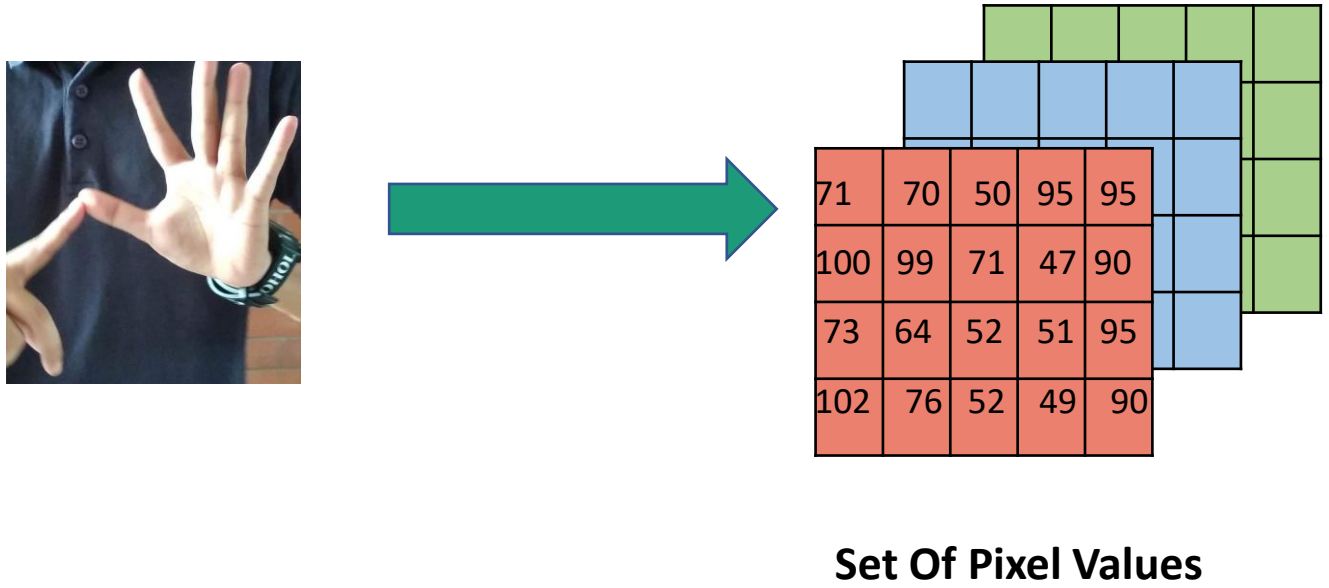
Fig: Some frames belonging to a video sample after processing and background removal.

BdSLVidSet: Current specifics

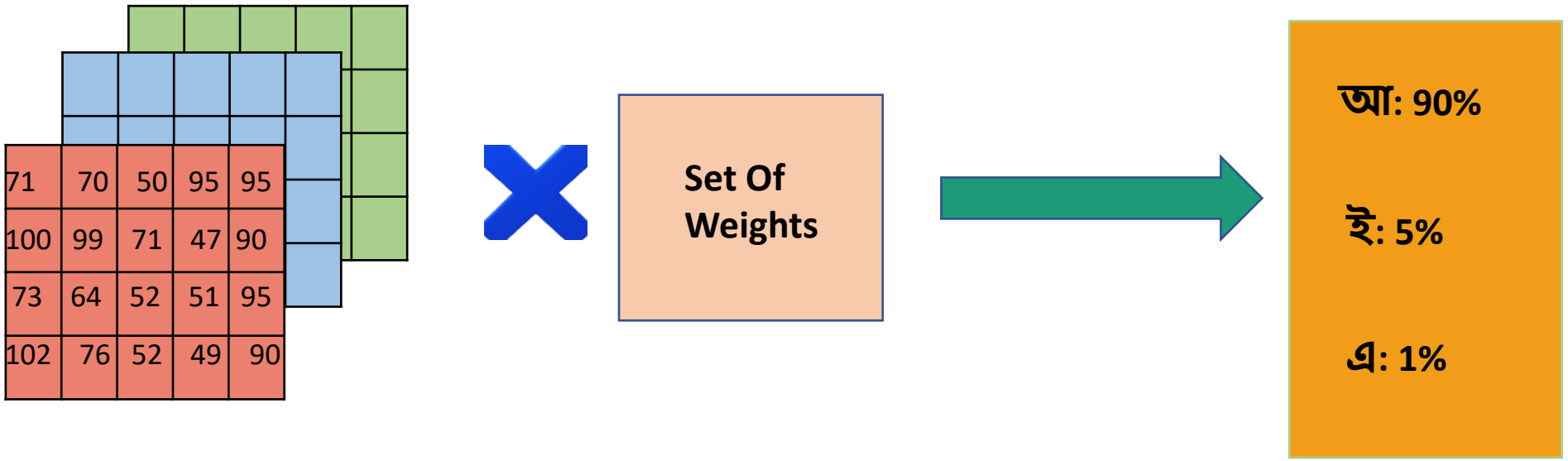
Total Videos	Total Class	Videos/ Class	Video Size	Video Background	No. Of Person Participated	Training Set: Testing Set
200	4	50	≤800kb	White with lighting variation	5	6:4

*Available in **Github (**https://github.com/Oishee30/BdSLVidSet)

Images to Computer

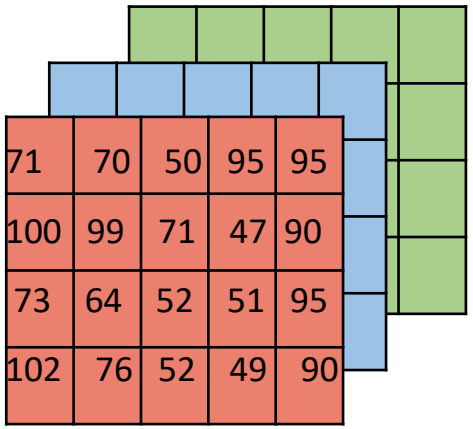


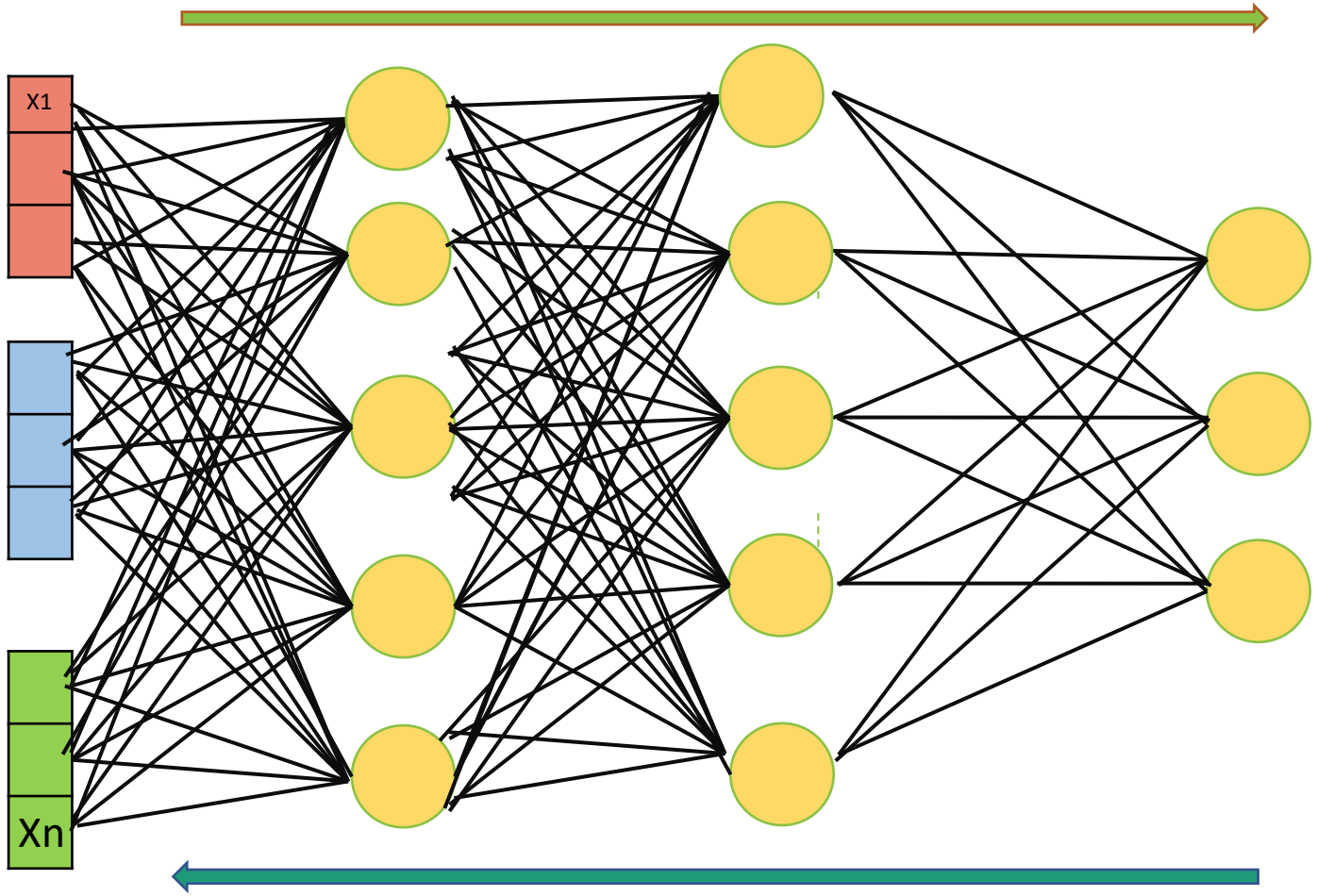
How computers detect classes from images?



Weight Generation

Neural Networks: Weights Update





$Z=W^T X + b$

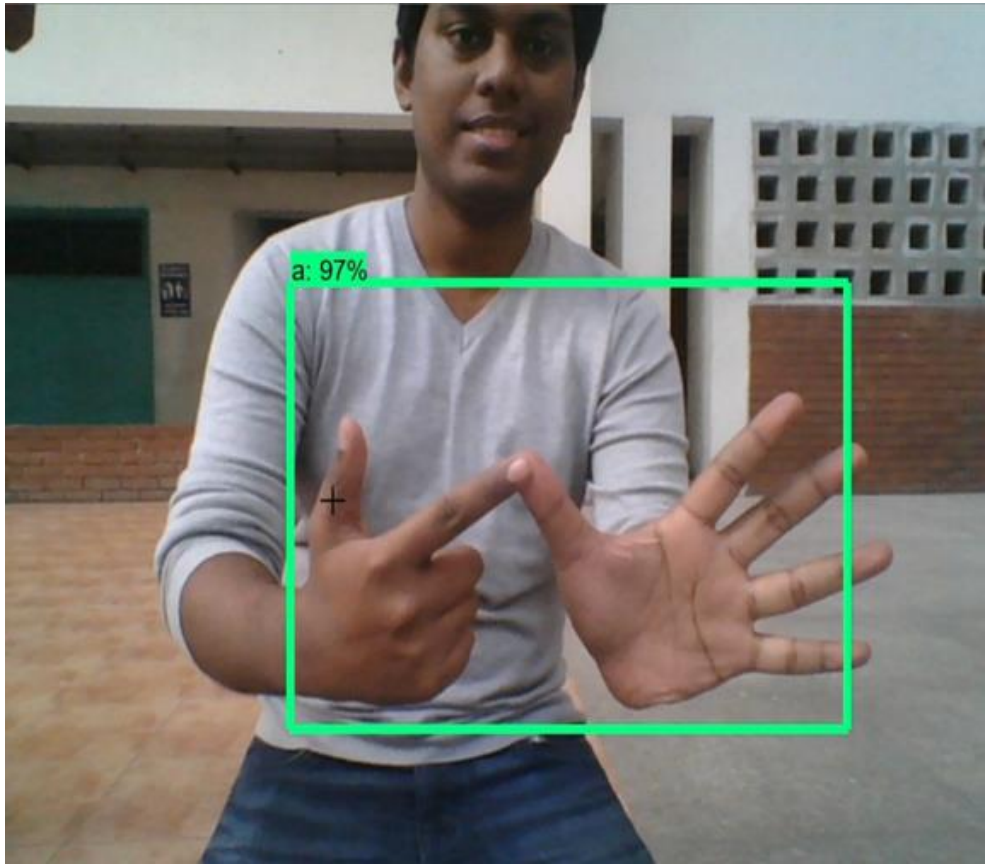
A= activation(Z)

Loss Calculation

Cost Calculation

Image Detection In Dynamic Background

- Image Classification
- Localization



Convolutional Neural Network

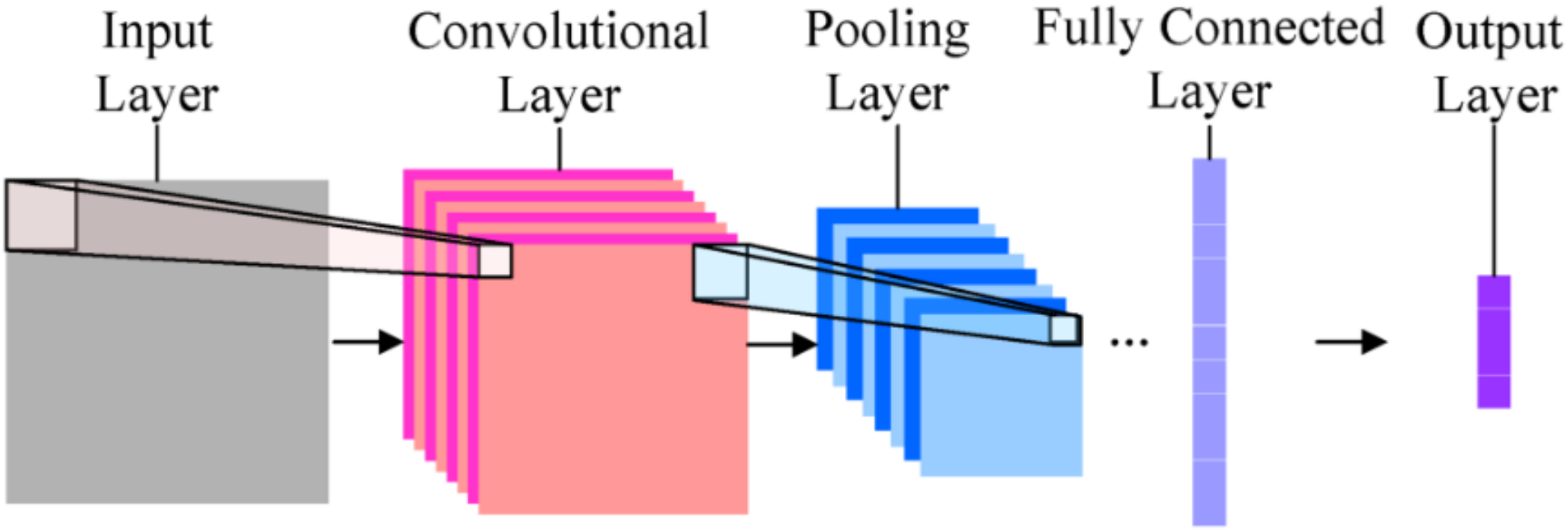


Fig: CNN Architecture (Source : Internet)

Different Layers Of CNN

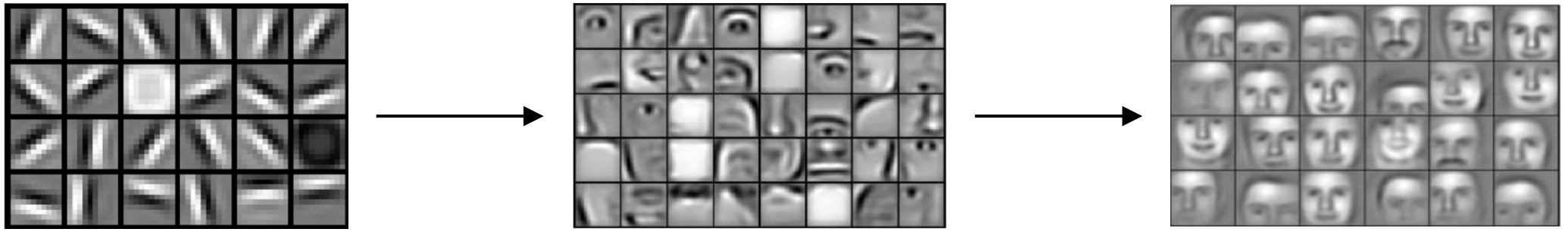


Fig: Different Layers of CNN (Source : Internet)

Transfer Learning

- Pre-trained weights of an already trained model
  - On millions of images belonging to 1000’s of classes
  - On several high-power GPU’s for several days)
- There is no need of an extremely large training dataset.
- Not much computational power is required.

Pre-Trained Models

- Fast R-CNN
- Faster R-CNN
- YOLO
- Mask R-CNN
- SSD Mobilnet

Faster R-CNN Architecture

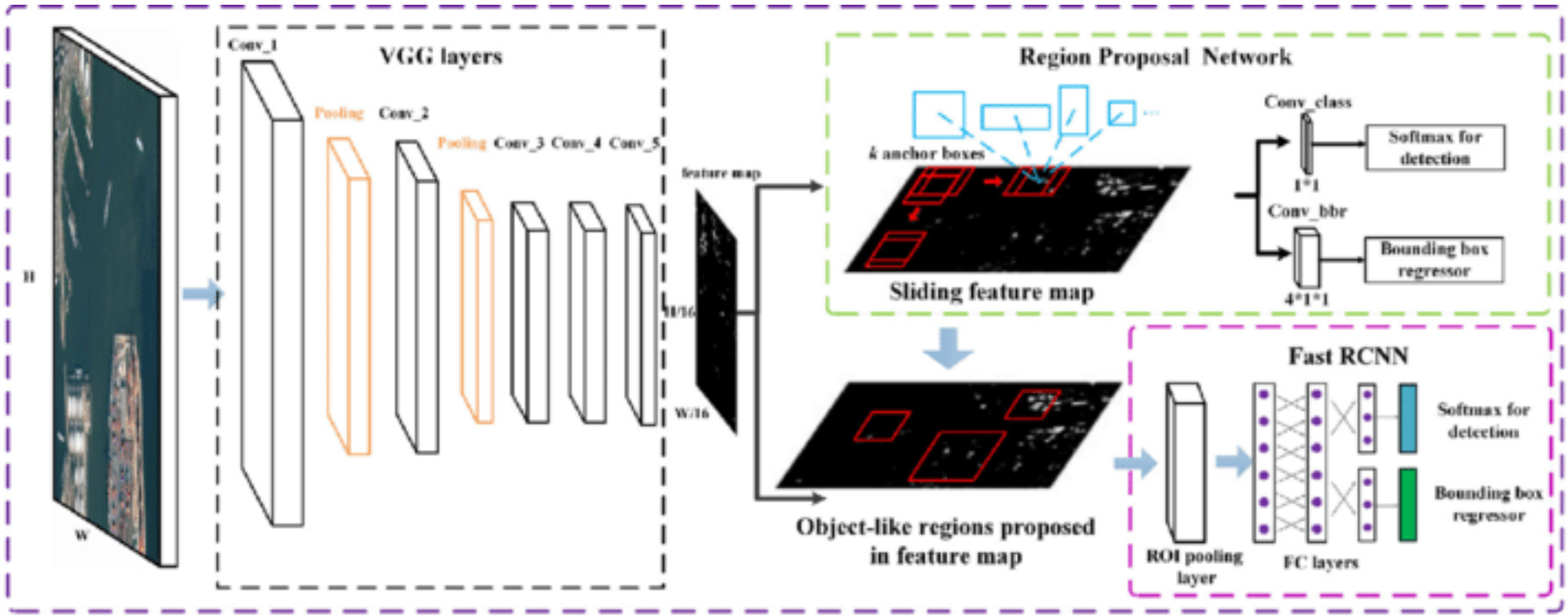


Fig: Faster R-CNN Architecture (Source : Internet)

RPN - Anchor Boxes

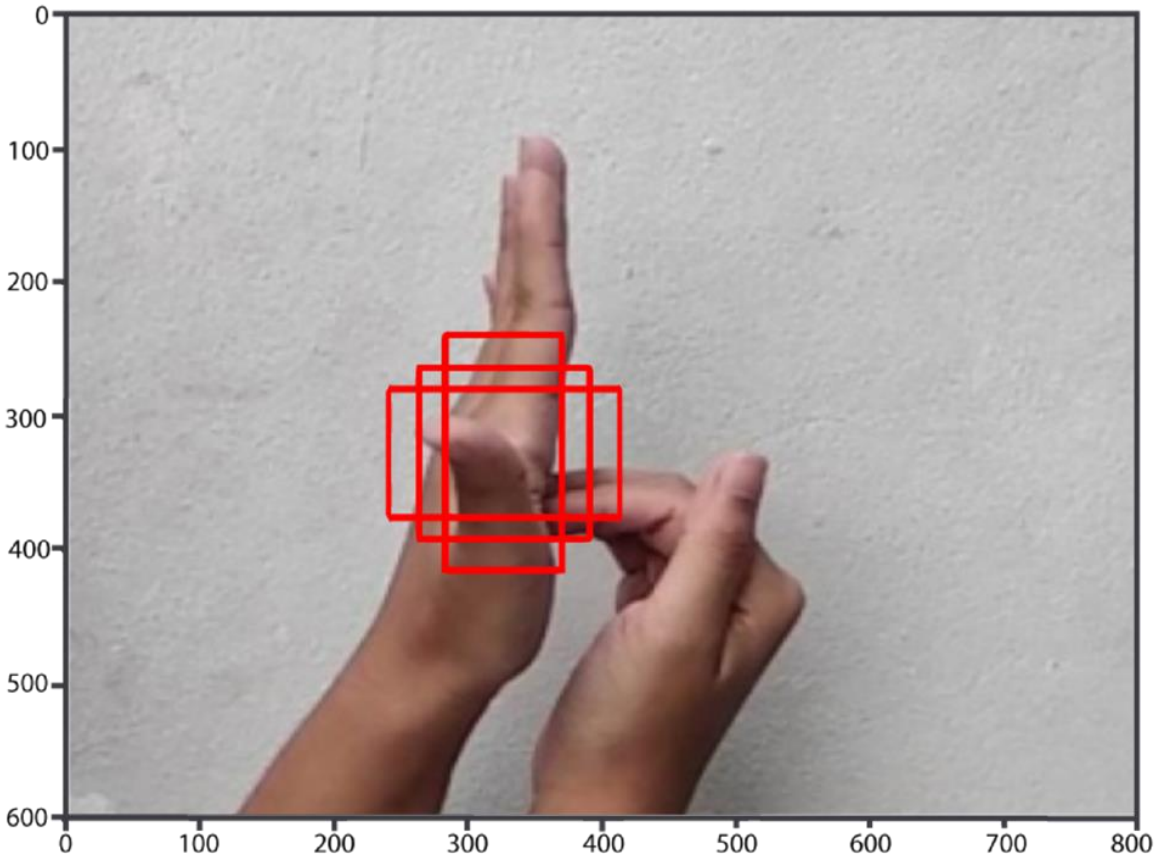


Fig: Anchor Boxes aspect_ratio 0.5, 1,2 with scale size of 0.25

RPN - Anchor Boxes (cont.)

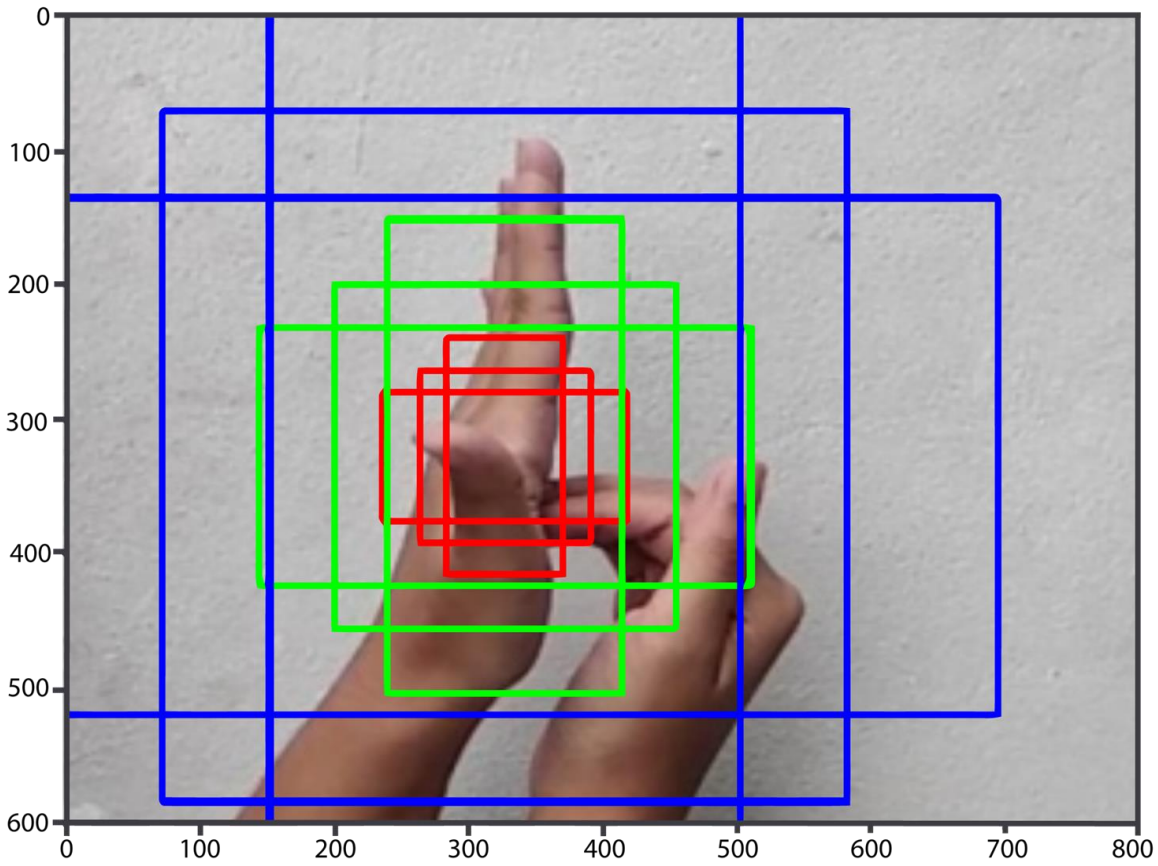


Fig: Anchor Boxes aspect_ratio 0.5, 1,2 with scale size of 0.25, 0.5, 1

RPN - Anchor Boxes (cont.)

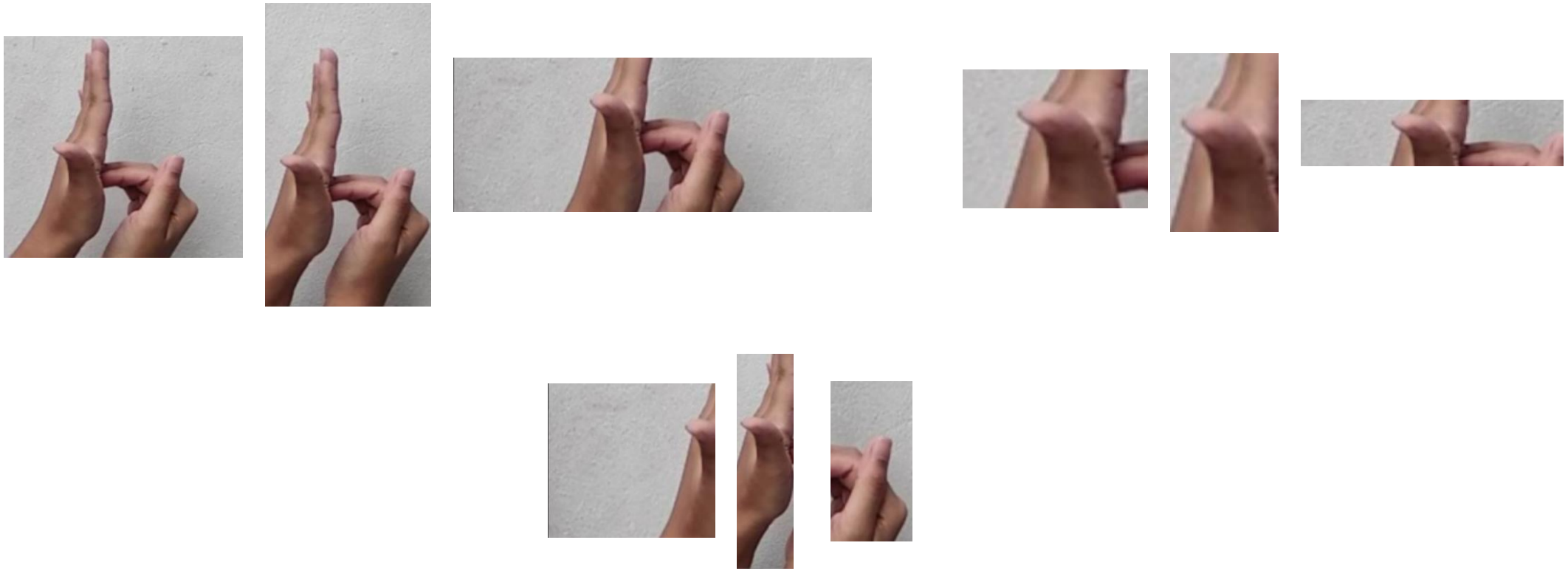


Fig: Sample Regions From an Image

RPN - Anchor Boxes (cont.)

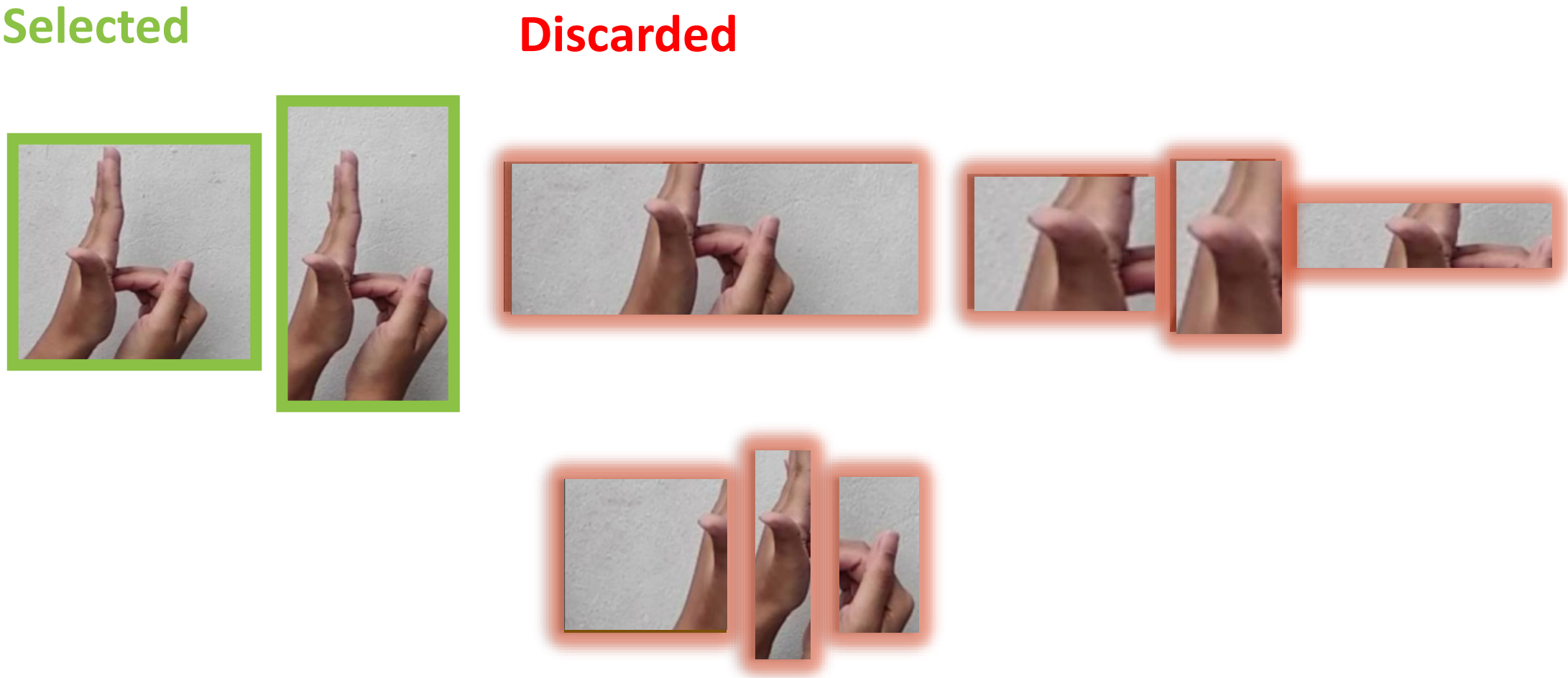


Fig: Green portions are selected as foreground and others are selected as background.

Training BdSLImset With Faster R-CNN Model

- Took about 12 hours
- 28000 iterations to train the model.
- Started with loss of 3.00, quickly dropped to 0.8. Stopped at 0.03.

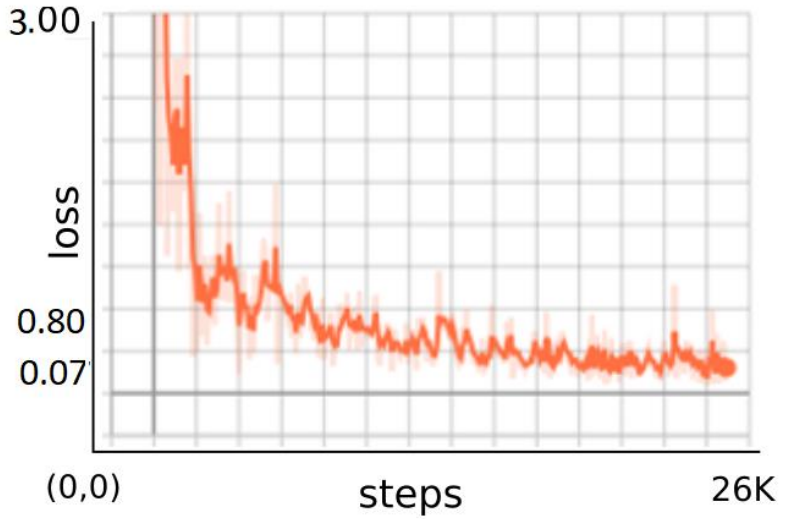


Fig: Loss Graph

Experimental Results

Models	Faster - RCNN	YOLO – Inception V2 Model
Training Time	12 hours	8 hours
Average Accuracy (On Test Set)	0.9415	0.6
Detection Accuracy (On Applcation Level)	Accurate with average of 90percent confidence rate	Faulty

Experimental Results - On BdSLImset(Test Set)

Id	Gesture	No. Of Image In Test Set	No. Of Correct Classification	Accuracy (%)
1	oo (অ)	40	38	95
2	a (আ)	40	40	100
3	I (ই)	40	35	87
4	e (এ)	40	35	87
5	u (উ)	40	30	75
6	k (ক)	40	40	100
7	kh (খ)	40	40	100
8	ga (গ)	40	40	100
9	dh (ঢ)	40	40	100
10	o (ও)	40	39	97.5

Experimental Results - On Real-time System

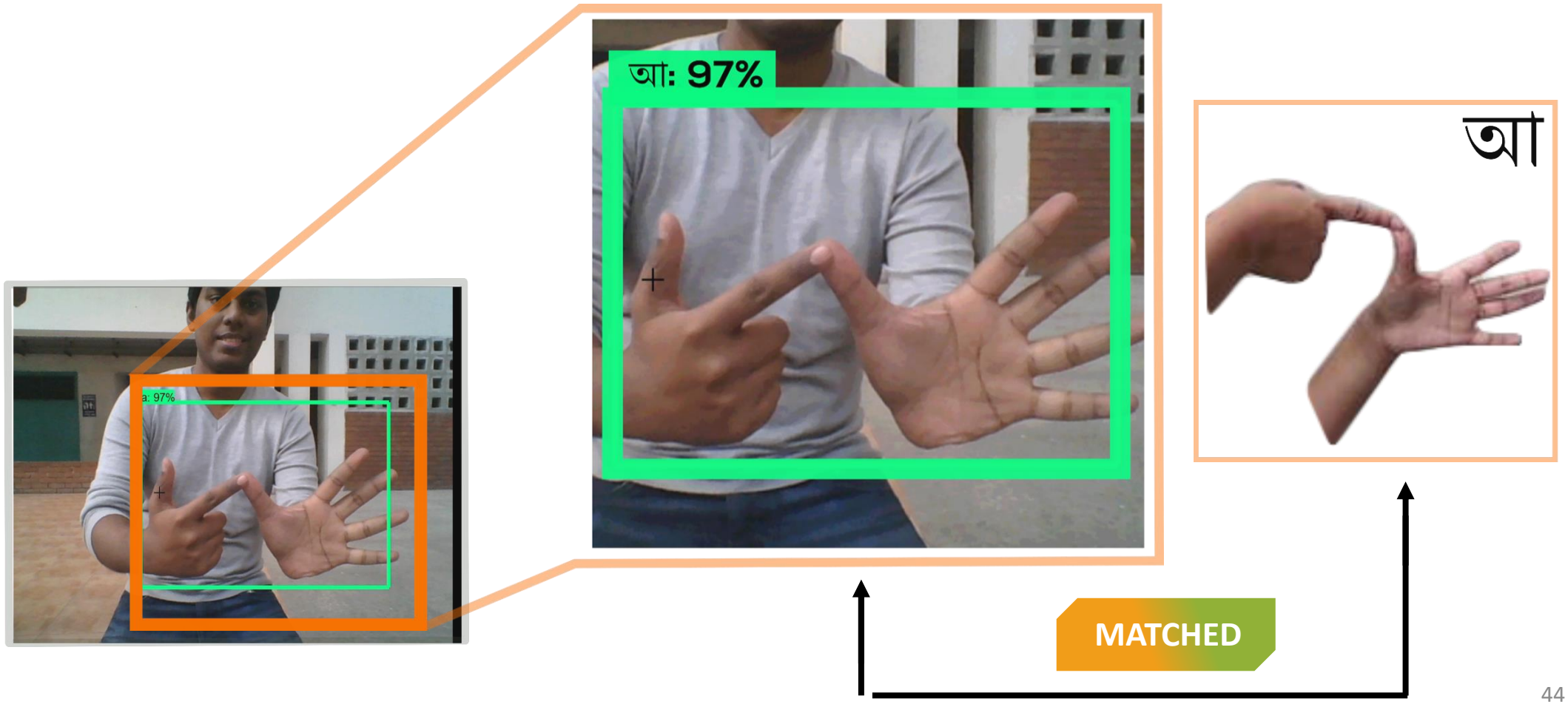


Fig: Detection in Real-time System

Experimental Results - On Real-time System (cont.)

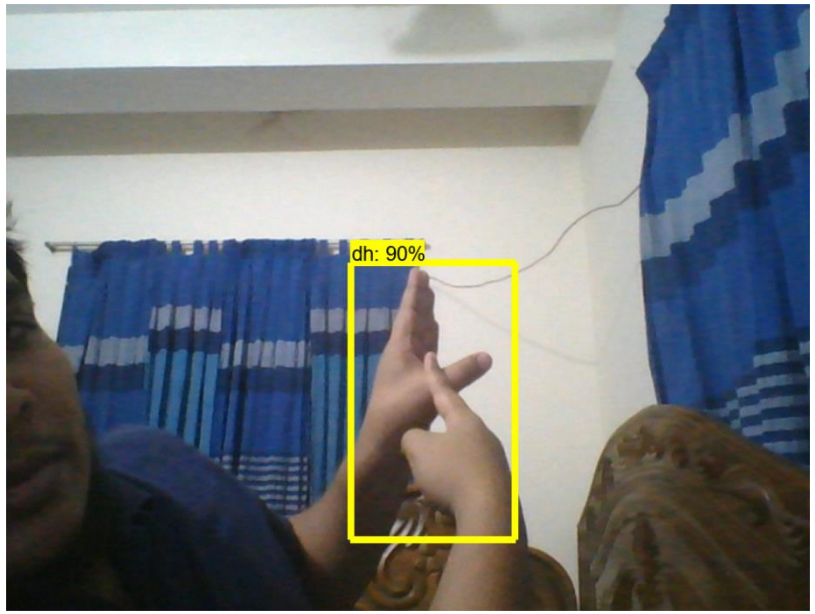


Fig: Some other examples of detection in Real-time System

Experimental Results - On Real-time System (cont.)

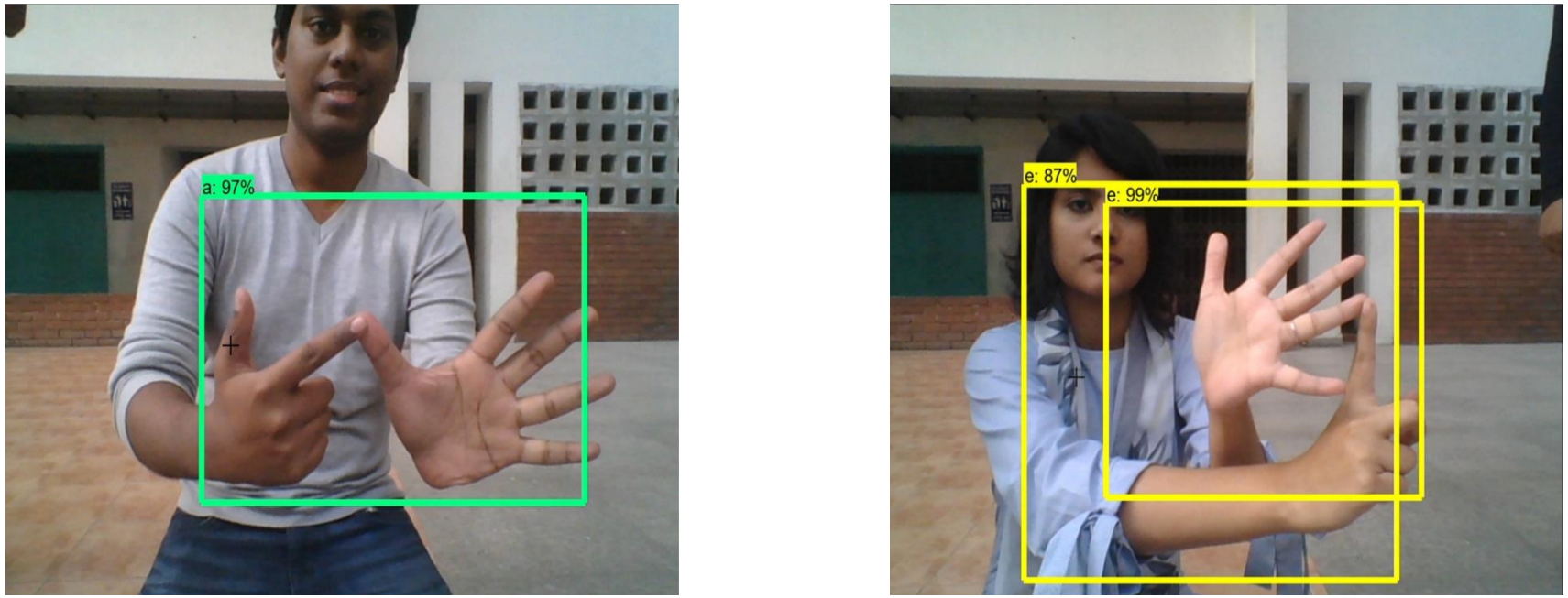


Fig: Some other examples of detection in Real-time System

What About Words?

- Sequence of images combines a word

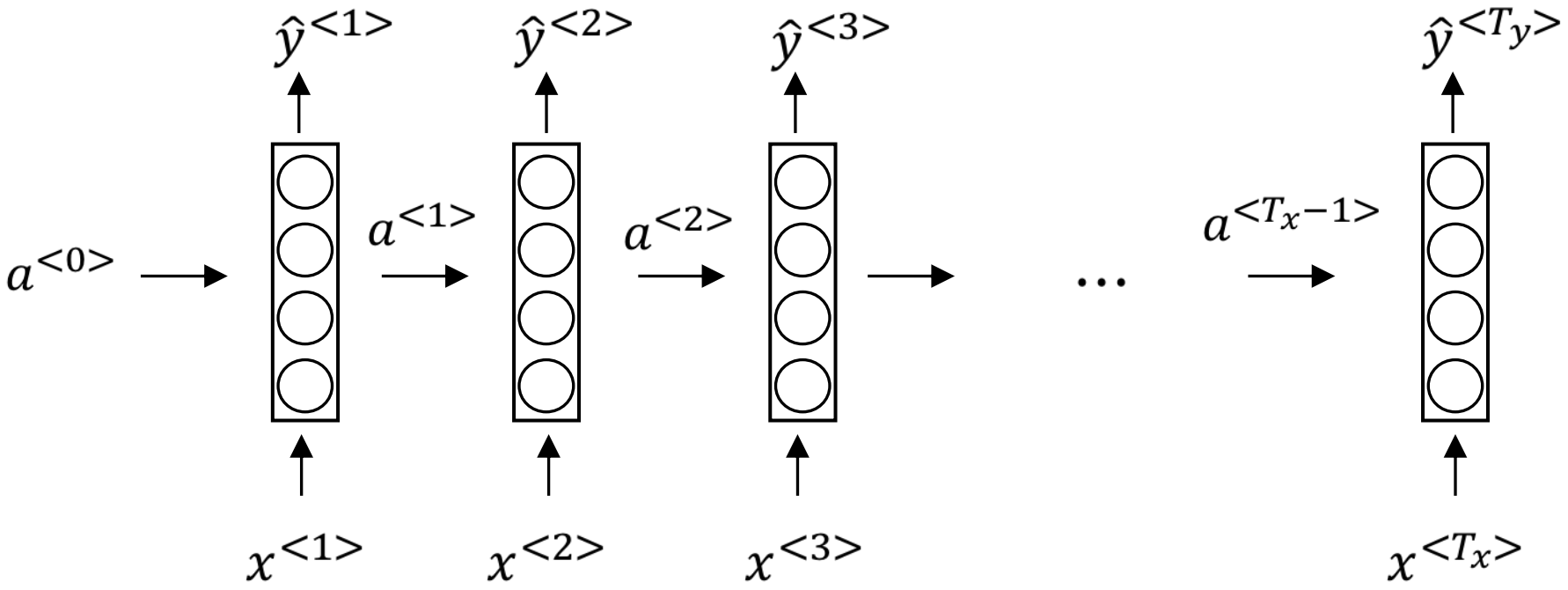
TUMI

How to solve this?

- Extracting features from each of the frame with CNN
- Combines the features through an RNN based model

** Real-Time Argentine Sign Language Gesture (Word) Recognition from video sequences using CNN and RNN (2018)

Recurrent Neural Network



$$a^{<t>} = g(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a) \qquad \hat{y}^{<t>} = g(W_{ya}a^{<t>} + b_y)$$

Fig: RNN Architecture

Long Short-Term Memory

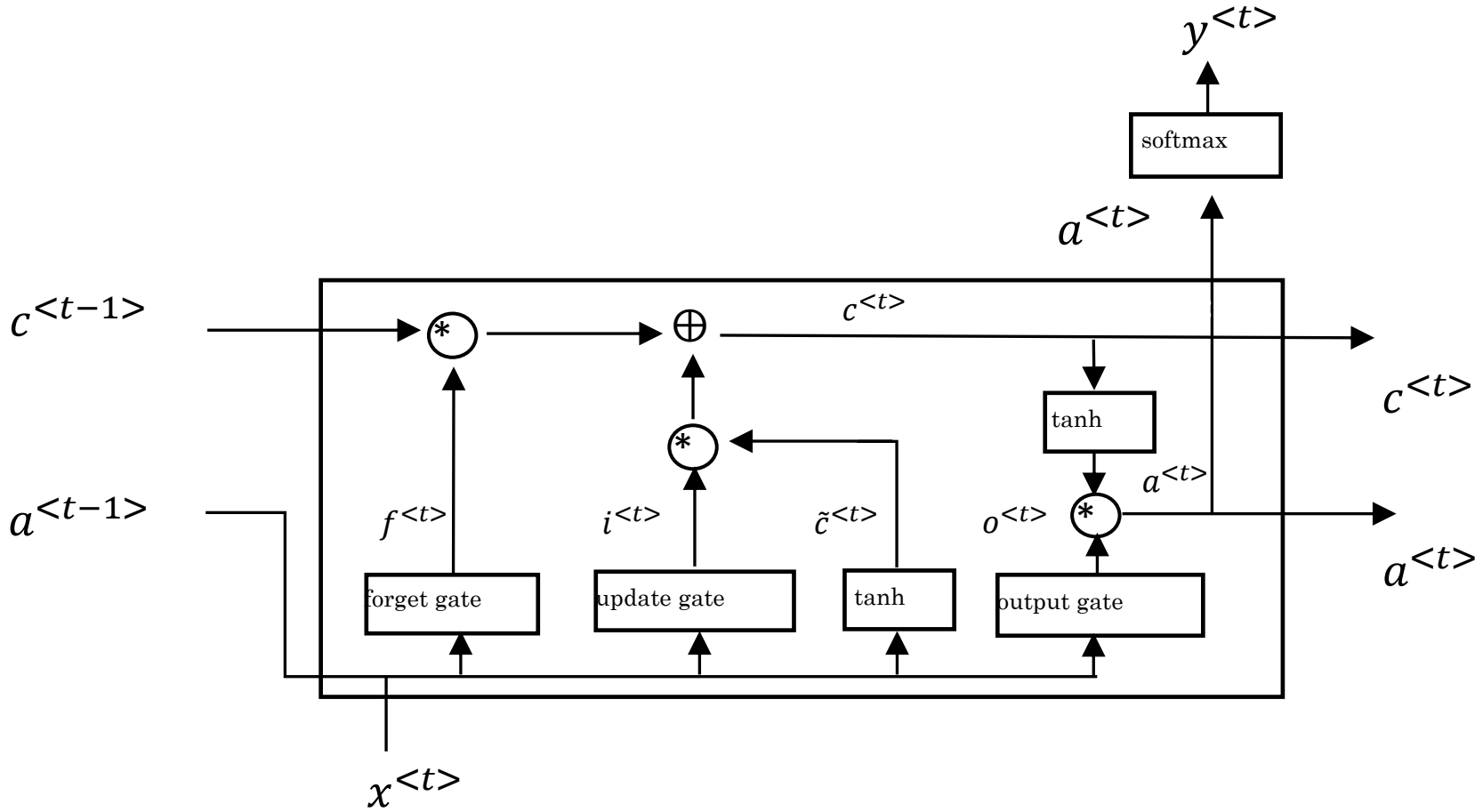


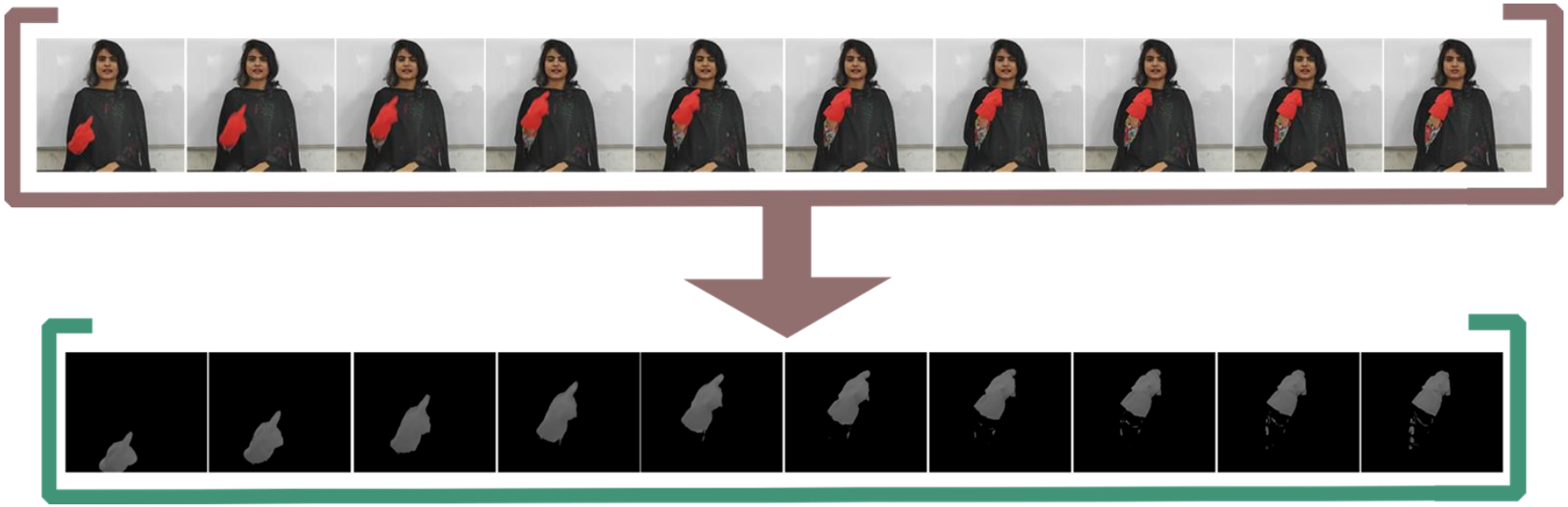
Fig: LSTM Architecture

Word Recognition System – Extracting Frames

- Extracting frames from the video sequences & go through necessary processing.

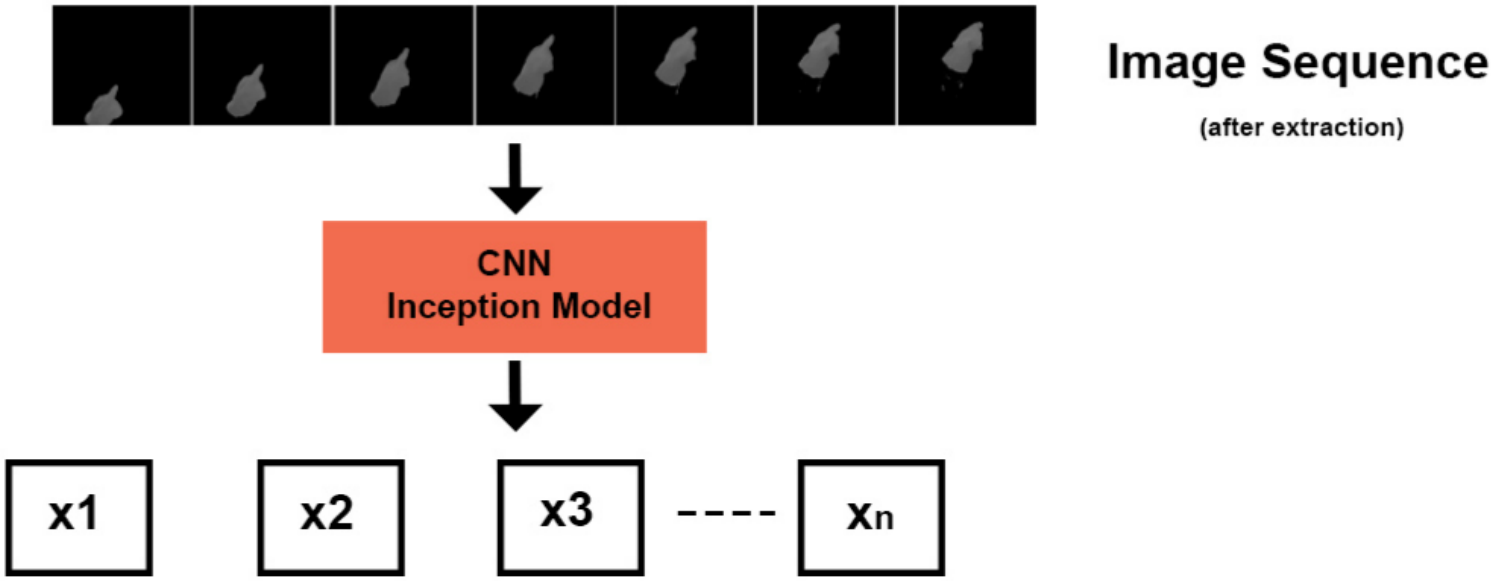
Word Recognition System – Processing Frames

- Extracting frames from the video sequences & go through necessary processing.



Word Recognition System – CNN

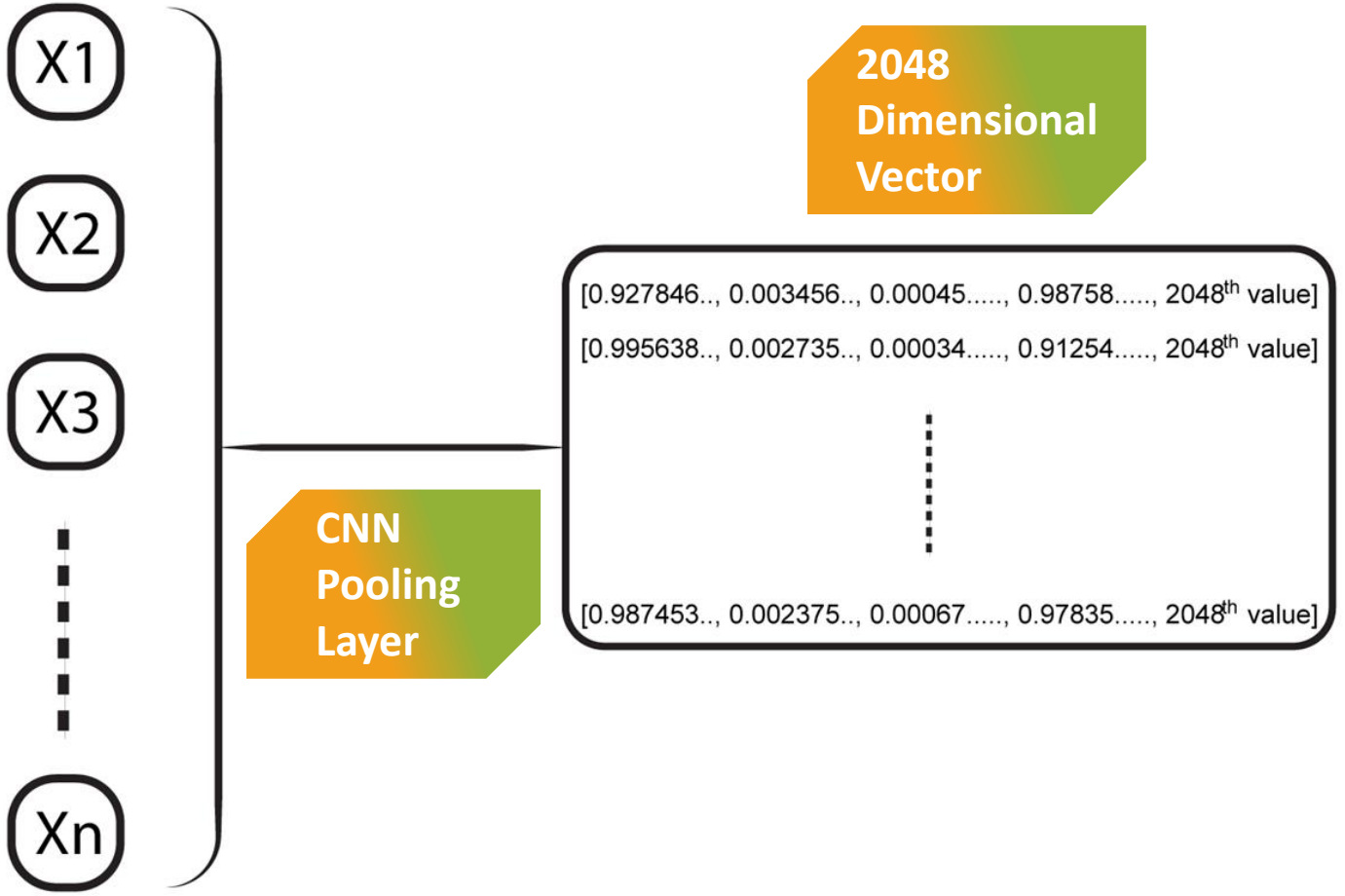
- Extracted frames fed into CNN.
- CNN generates spatial features of each image.



Word Recognition System – LSTM Method I

- Extracted features used in two approaches
  - Method I
    - CNN output layer passed to RNN
    - Returns a list of probability values from frames
    - Belonging to each class
    - 4 class probability for our system

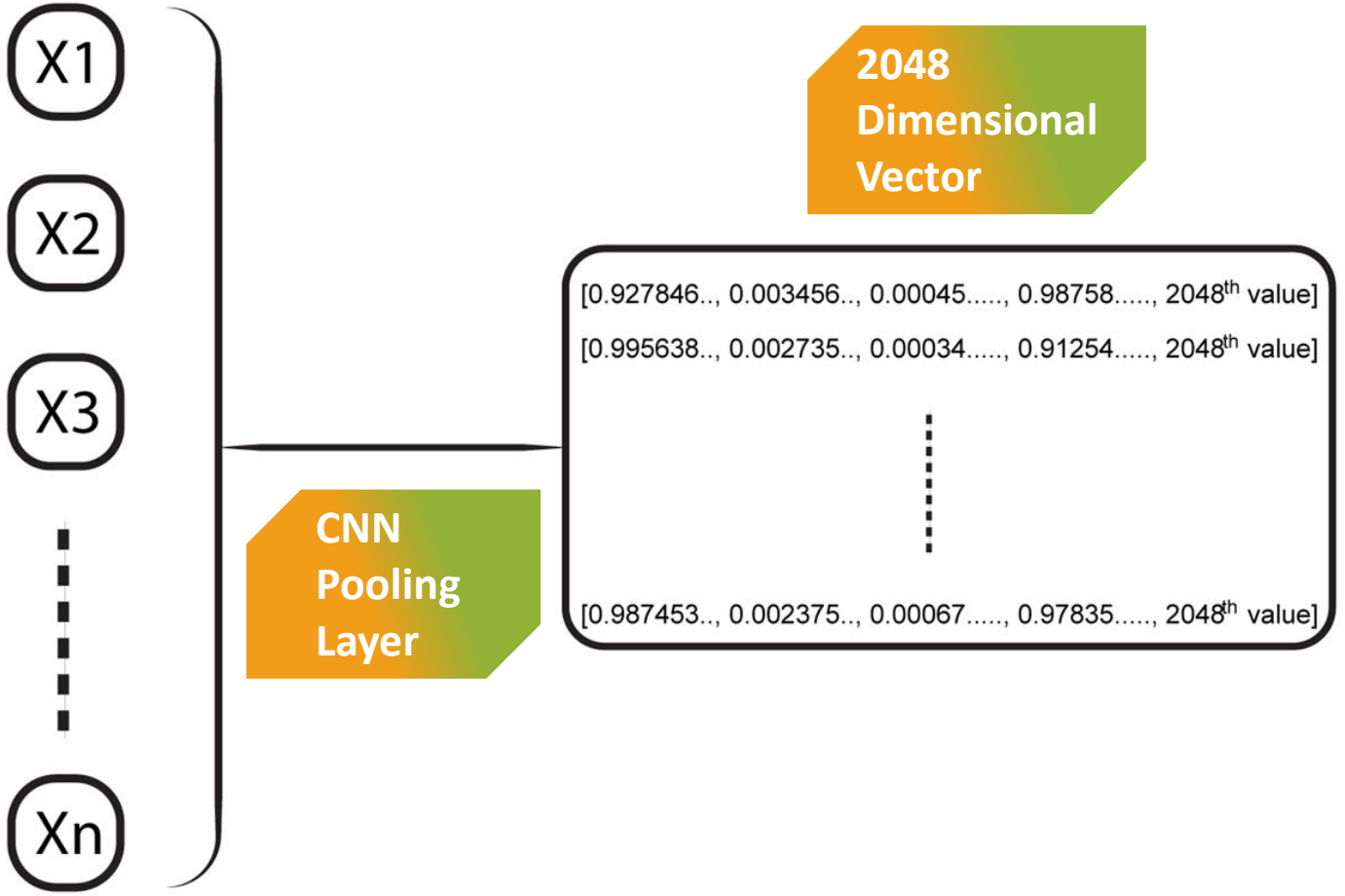
Word Recognition System



Word Recognition System - LSTM Method II

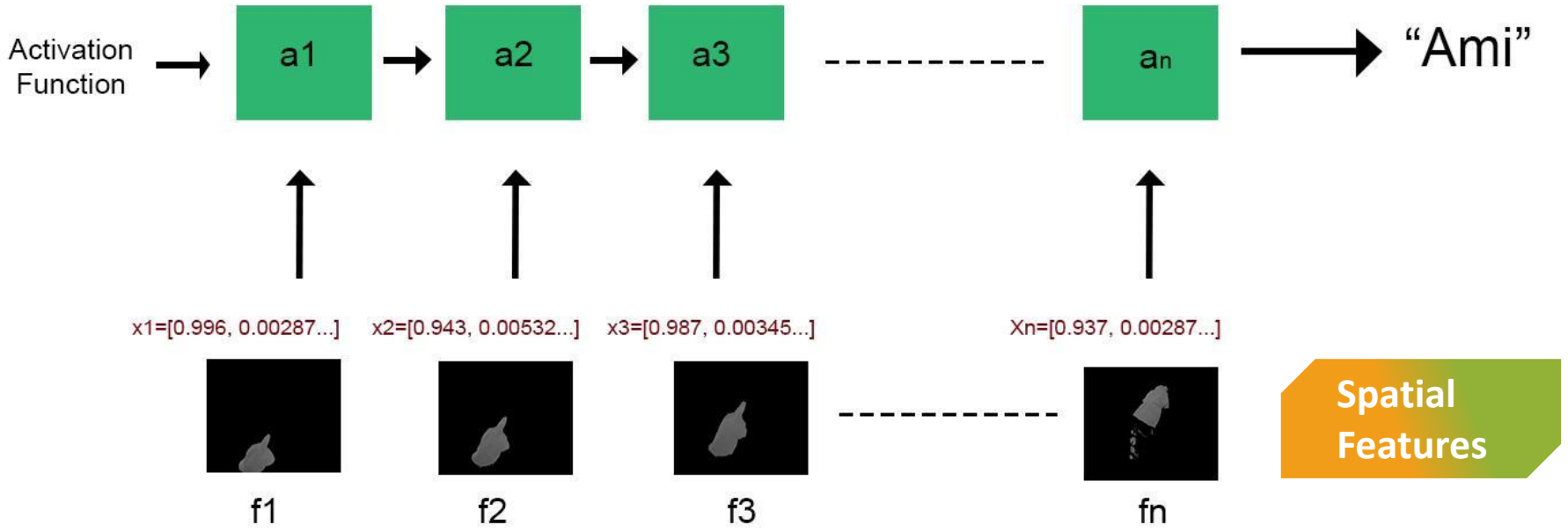
- Method II
    - CNN pooling layer passed to RNN
    - Returns a list of convoluted features from frames
    - 2048 dimensional vector for each frame

Word Recognition System - LSTM Method II



Word Recognition System - LSTM

- Single RNN layer consisting of 256 LSTM units
- Previous layer value works as input for next layer (RNN)



Result

ID	Gesture	Video	Classification		Accuracy	
			Approach 1	Approach 2	Approach 1	Approach 2
1	Ami	20	20	20	100	100
2	Tumi	20	20	20	100	100
3	Kemon_Acho	20	20	20	100	100
4	Valo_Achi	20	20	20	100	100

Approach Comparison (cont.)

- Approach 2 based system outperforms the Approach 1

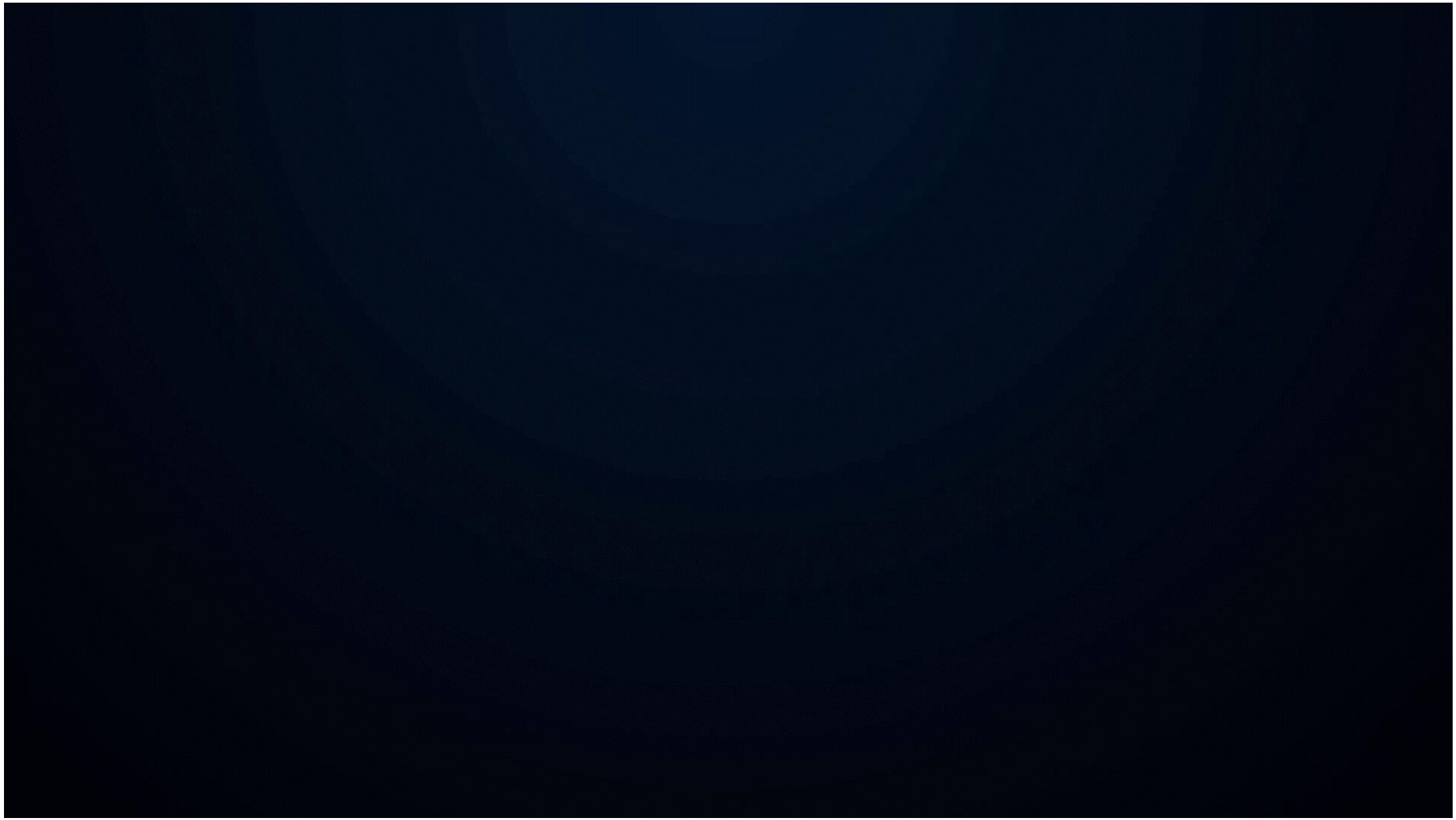
Approaches	Time Needed To Detect Each Gesture	Multiple Gesture at a time
Approach 1	About 1.30 minutes	NO
Approach 2	About 30 seconds	YES

Limitations

- While recognizing the letters with similarities among their patterns gives faulty recognition sometimes.
- While recognizing words with the system, there’s background limitation.
- Less number of class in video dataset.

Future Plan

- Implement a mobile based system
- Increase number of data and classes in both of our datasets
- Overcome the background limitation in word recognition system



Thank You.