Gesture Recognition Using Deep Learning

Malhar Patwari
A20410420

Herick Asmani
A20399752

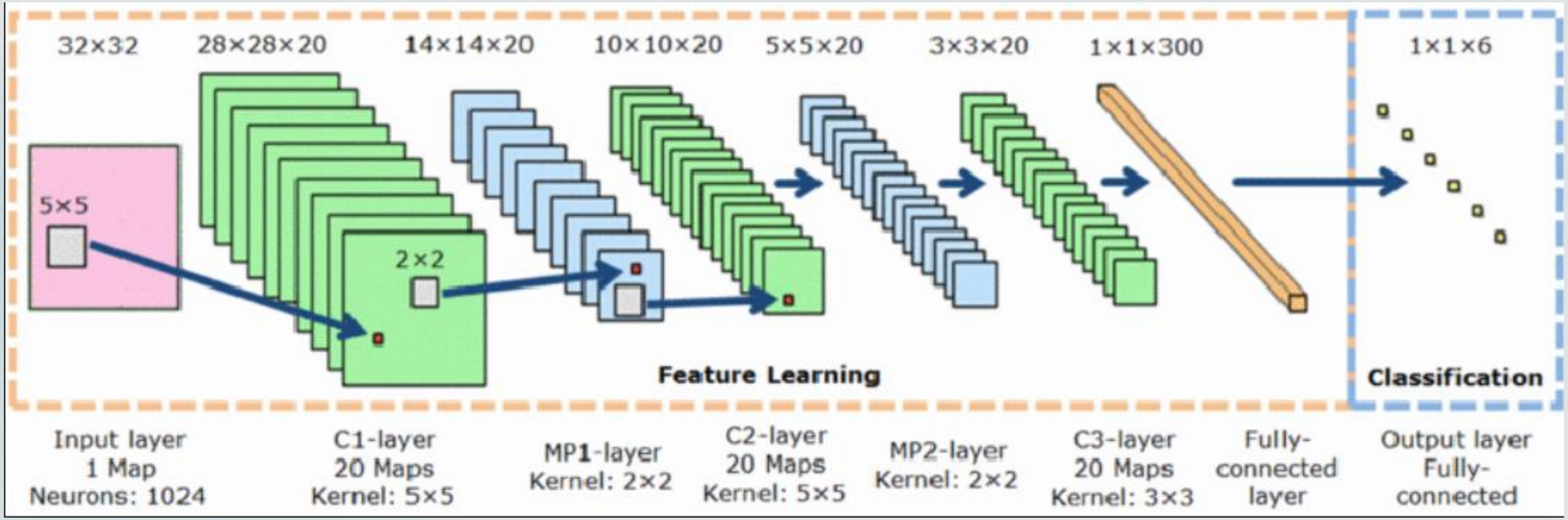
Problem Statement

- Control computer, Tv, other devices just by using hand!
- Application in advanced driver assistance systems (ADASs)
- Advancement in deep learning
- Better approach than feature engineered Machine Learning models
- Challenges:

  1. Intra and inter-persons variations in hand gesture motion
  2. Inter-person variations in the shape and size of the human hand
  3. Illumination variations
  4. Background noise

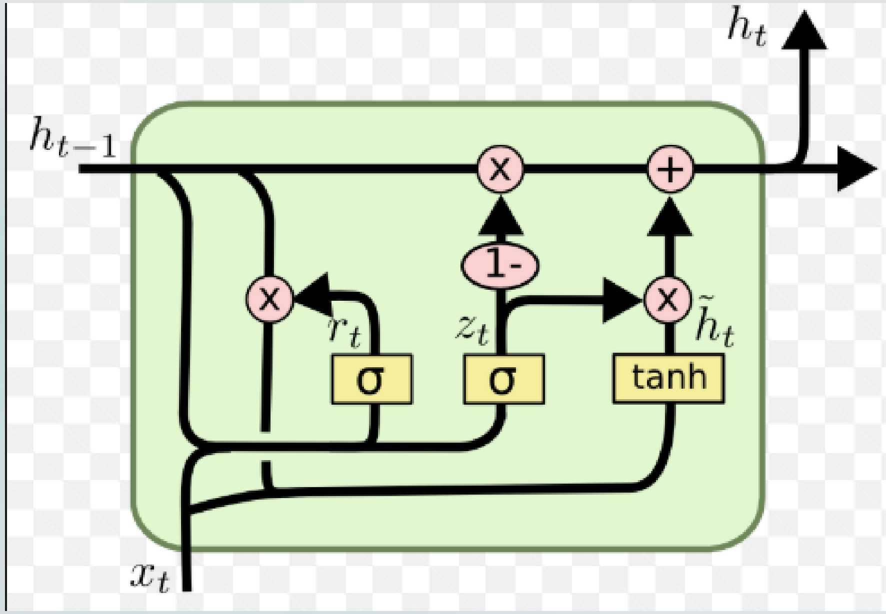
Background

2D Convolutional Neural Network



Background

Long Short Term Memory unit



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

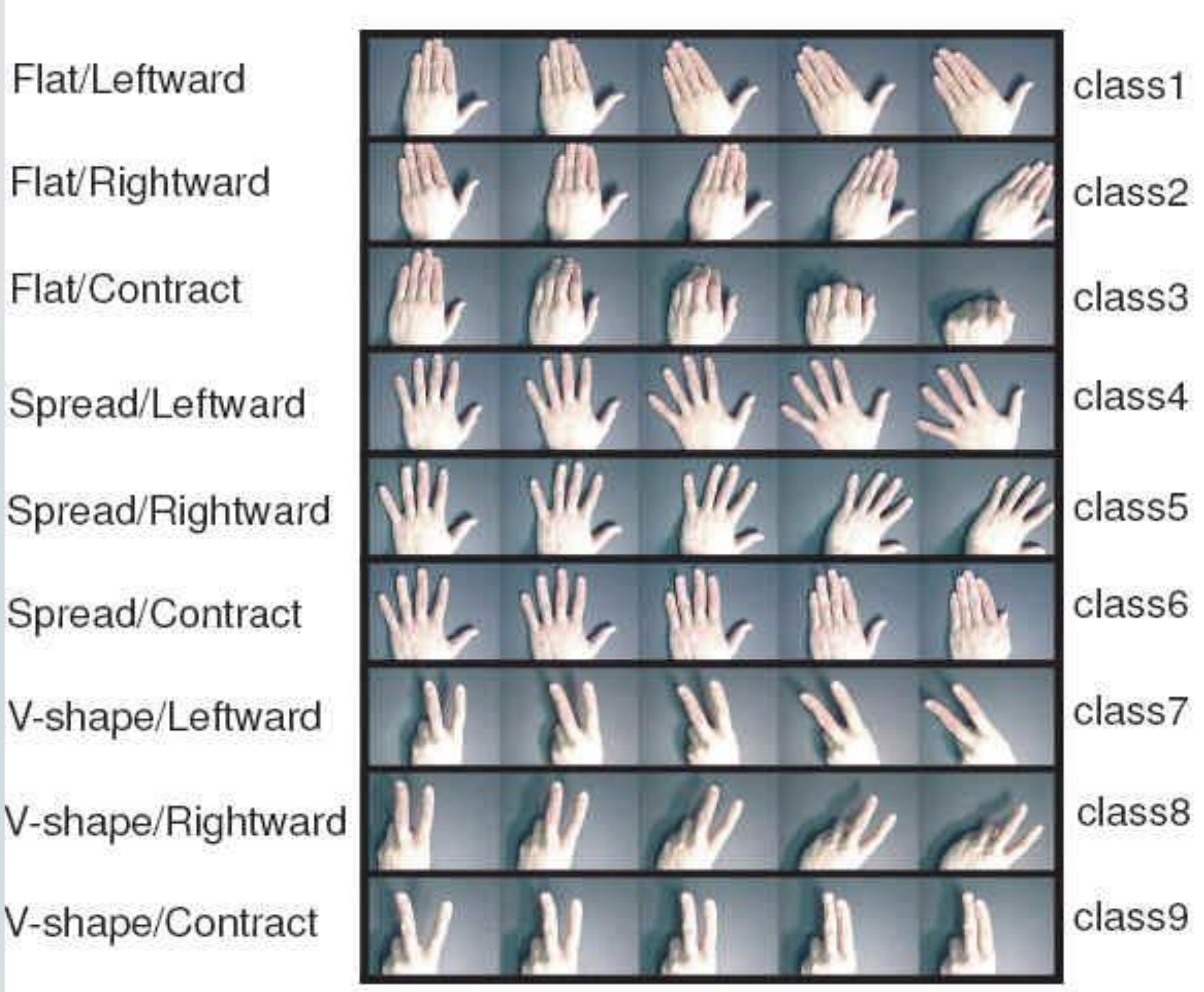
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Data

- The dataset used for this project is Cambridge Hand Gesture Dataset. The data set consists of 900 image sequences of 9 gesture classes, which are defined by 3 primitive hand shapes and 3 primitive motions as shown in figure below.
- Each class contains 100 image sequences (5 different illuminations x 10 arbitrary motions x 2 subjects).

Data


Flat/Leftward
class1
Flat/Rightward
class2
Flat/Contract
class3
Spread/Leftward
class4
Spread/Rightward
class5
Spread/Contract
class6
V-shape/Leftward
class7
V-shape/Rightward
class8
V-shape/Contract
class9

Proposed Solution

Step 1

Sparse Modeling Representative Frames (SMRF)

- Extract representative frames of a video sequence

$$\sum_{i=1}^{T} \|y_i - Yc_i\|_2^2 = \|Y - YC\|_F^2$$

$$\min \|Y - YC\|_F^2 \quad s.t. \|C\|_{1,q} \leq \tau, \quad \mathbf{1}^\top C = \mathbf{1}^\top,$$

- “See all by looking at a few: Sparse modeling for finding representative objects” by Ehsan Elhamifar , Guillermo Sapiro , René Vidal

Proposed Solution

Step 2

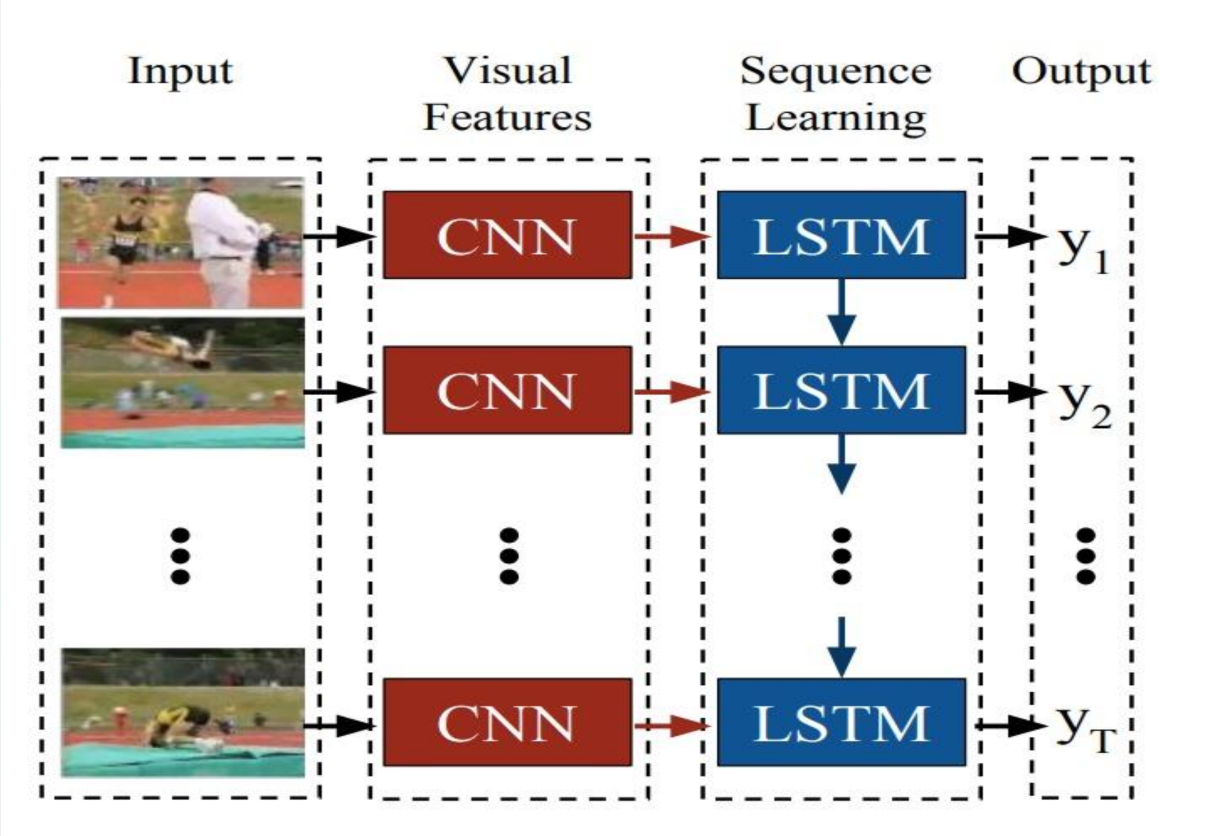
Preprocessing steps

- One hot encoding of class labels
- Normalize data by subtracting mean pixel value and divided by standard deviation
- Shuffling data to reduce bias
- Split data in to 700, 100, 100 videos for train, validation and test respectively

Proposed Solution

Step 3

Long Term Recurrent Convolutional Network (LRCN)



Results

Best Model parameters with using SMRF

Data: Training : 700 videos , each having 5 frames

Validation : 700 videos , each having 5 frames

Testing : 700 videos , each having 5 frames

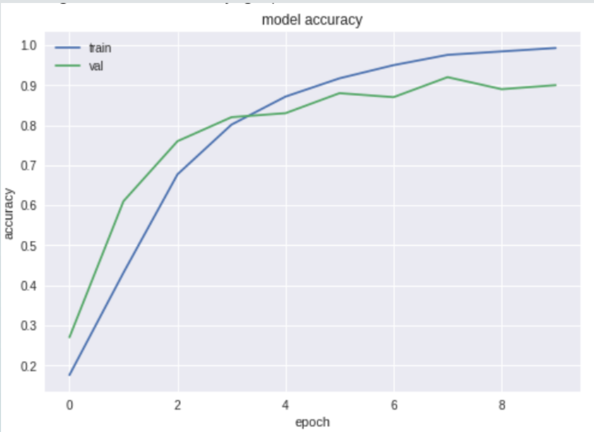
Parameter	**Value**
2D convolution layer	5
Max pooling layer	5
LSTM units	256
Optimizer	Adam
Dropout	0.5
Epoch	10
Batch size	64

Train Accuracy:	99.29%
Validation Accuracy :	90%
Test Accuracy :	98%
Test Precision :	0.97
Test Recall :	0.97
Test F1 Score:	0.97

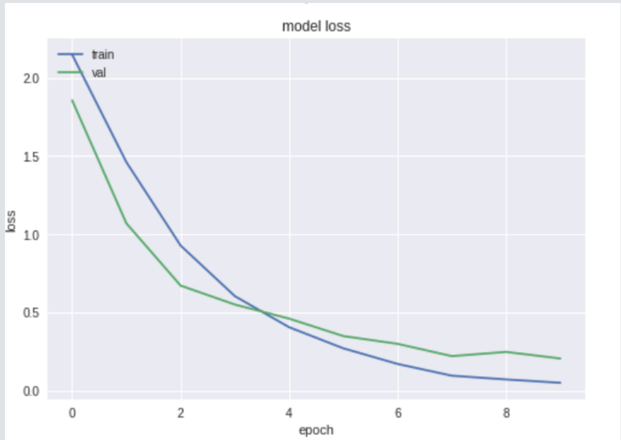
Results

Graphs:

Accuracy vs epoch



loss vs epoch



Precision , Recall, f1 score, support:

```
precision: [0.94117647 1.         1.         1.         1.         1.
 1.         0.85714286 1.        ]
recall: [1.         1.         1.         1.         1.         0.91666667
 1.         1.         0.9       ]
fscore: [0.96969697 1.         1.         1.         1.         0.95652174
 1.         0.92307692 0.94736842]
support: [16 10 13 13  7 12 13  6 10]
```

Results

Best Model parameters without using SMRF

Data: Training : 700 videos , each having 5 frames

Validation : 700 videos , each having 5 frames

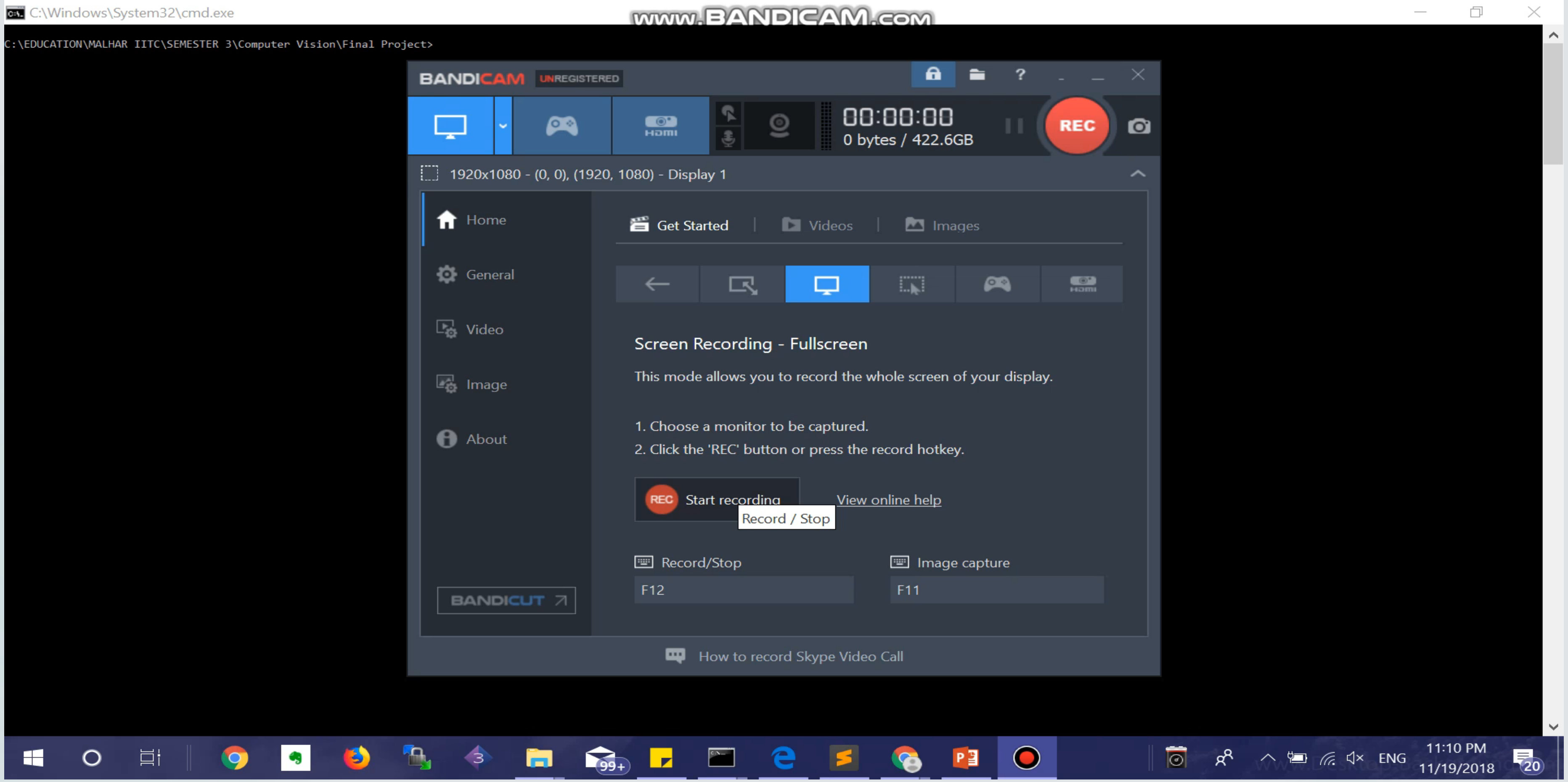
Testing : 700 videos , each having 5 frames

Parameter	**Value**
2D convolution layer	5
Max pooling layer	5
LSTM	256
Optimizer	Adam
Dropout	0.5
Epoch	10
Batch size	64

Train Accuracy:	90.12%
Validation Accuracy :	87.37%
Test Accuracy :	94.78%
Test Precision :	0.92
Test Recall :	0.93
Test F1 Score:	0.92

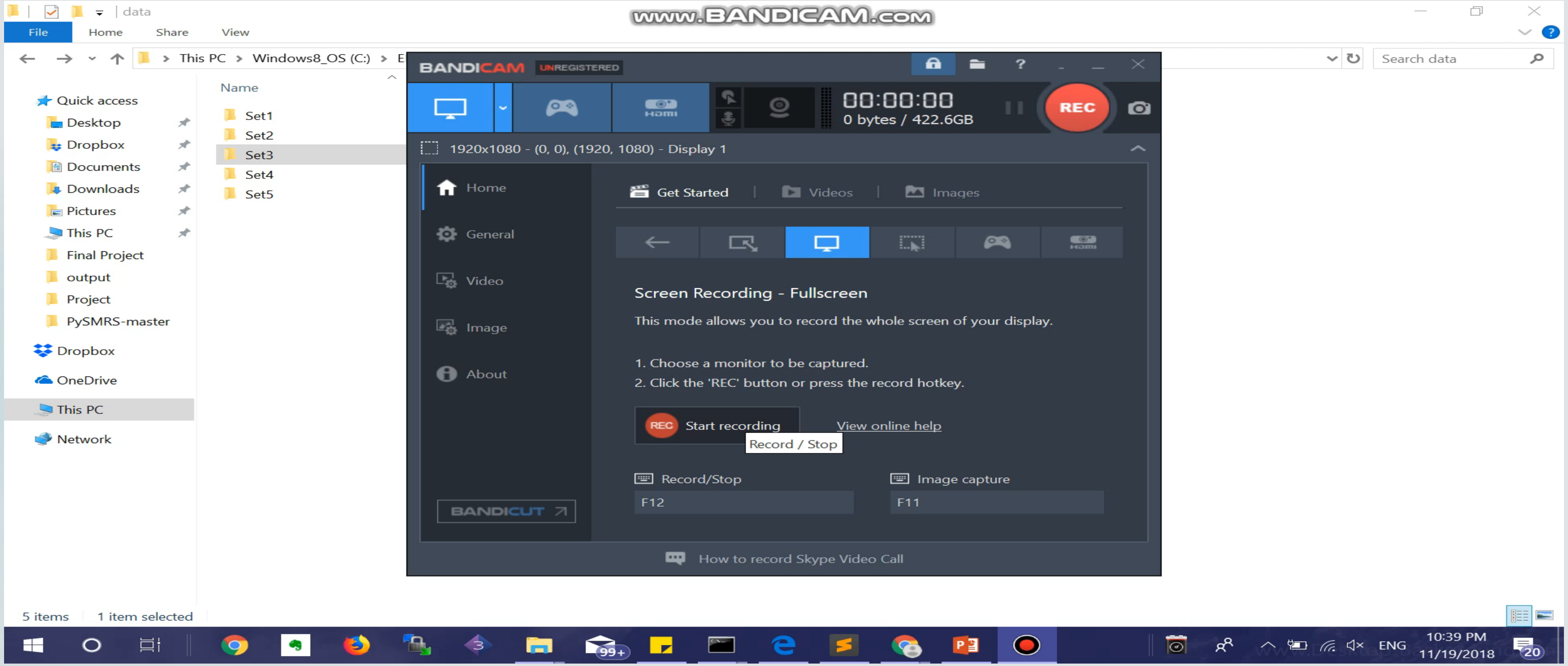
Results

Correct Classification example:



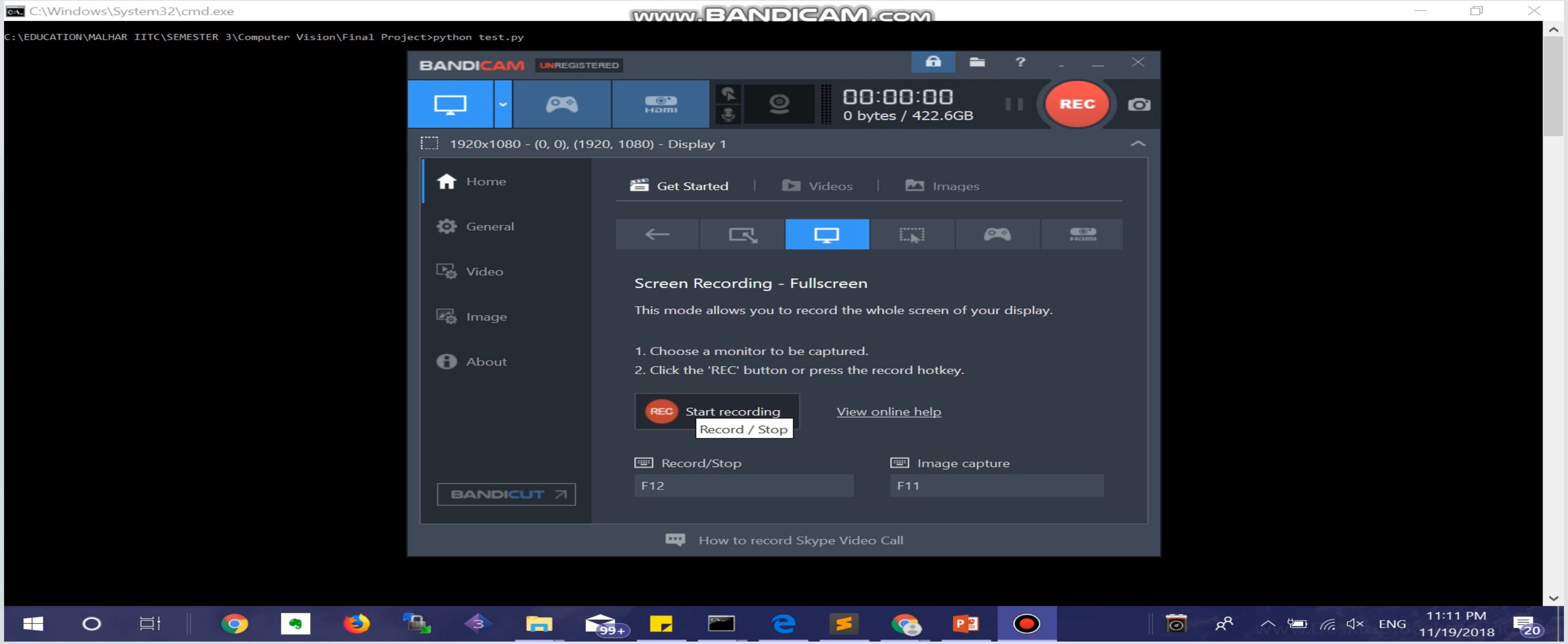
Results

Correct Classification example:



Results

Misclassification example:



Conclusion & Future Work

- In this project, a Deep Learning based Hand Gesture Recognition algorithm is proposed for operating computer applications.
- The Long term Recurrent Convolutional Neural Network is utilized to perform the hand gesture recognition.
- The reasonable classification accuracy and computational efficiency of the long term recurrent convolutional neural network is obtained by extracting 5 representative frames from the video sequence.
- Our proposed algorithm is evaluated on the Cambridge public dataset.
- In our future work, we will evaluate with a larger dataset containing flow images along with RGB images in order to improve the robustness of the system. With larger dataset pre-trained model such as VGG19, INCEPTION V3, etc will be used as CNNs.

References

- Deep Learning-Based Fast Hand Gesture Recognition Using Representative Frames. Vijay John ; Ali Boyali ; Seiichi Mita ; Masayuki Imanishi ; Norio Sanma. 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)
- J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in CVPR, 2015.

Thank You!