
# Vision and Cognitive systems Course Final Project

Hand detection and segmentation by implementing R-CNN and Meanshift

Prof. Lamberto Ballan

Physics of data

Asal Rangrasiasl - 2046832

03/03/2023

# CONTENTS

- INTRODUCTION
- AIM OF THE PROJECT
- MATERIAL AND METHODS
- DATASET
- METHODOLOGY
- RESULTS
- CONCLUSION & QUESTIONS

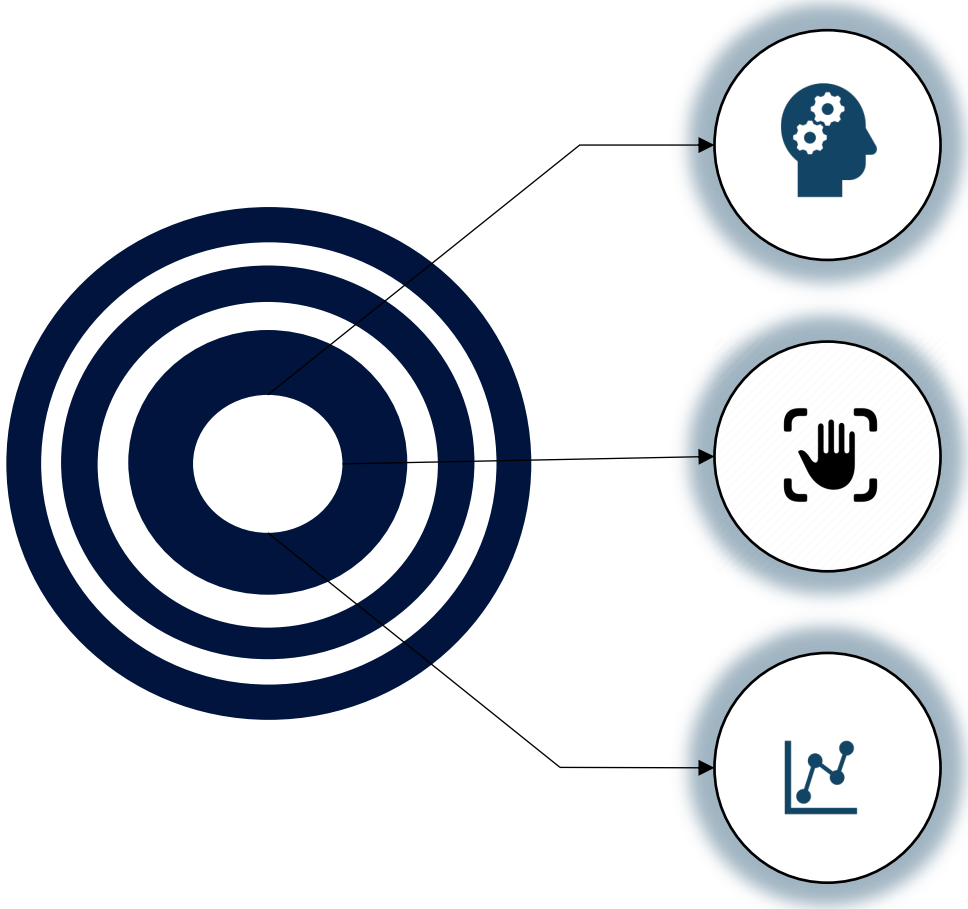
# INTRODUCTION

Object detection is an important task in computer vision and has a wide range of applications. Hand detection is an important technology with many practical applications in various industries, including gaming, healthcare, education, entertainment

In this work, we propose a combined approach for hand detection and segmentation using Region-based Convolutional Neural Networks (RCNN) and Mean Shift Segmentation.

The pre-processing step involves applying bilateral Filter and erosion on the images to reduce noise. In addition, Mean Shift Segmentation is applied to segment the objects in the image.

# AIM OF THE PROJECT



Proposed Method :

The proposed methods structure is to use the Ego-Hands dataset to train a Region-based Convolutional Neural Network (R-CNN) for hand detection

R-CNN:

Combine fine-tune a classification CNN with Selective Search, we'll be able to build our R-CNN object detector.

**Evaluate:**

Training and Validation accuracy of R-CNN for classification (Hand, Non-hand), showing the outputs of different filters

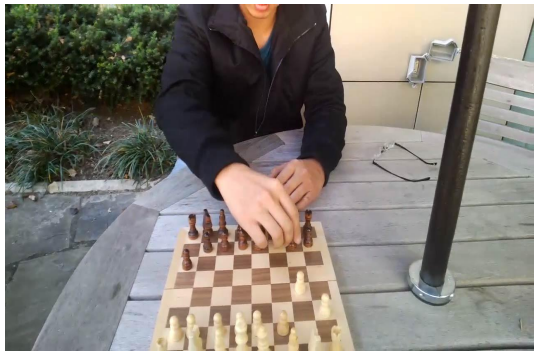
# MATERIAL AND METHODS

| Understanding The Context | Retrieving The Data | Dataset Preparation | Building Models and Training | Evaluation of Models and Reporting |
|---|---|---|---|---|
| study about different methods for hand detection | The Ego- Hands dataset contains 48 Google Glass videos of complex, first-person interactions between two people. | Build an object detection dataset using Selective Search | Fine-tune a classification network 'VGG16 ' for R-CNN with pre-trained weights | Combining Mean Shift Segmentation & R-CNN to detect hands in an image |

# DATASET Prepration

EgoHands ( A dataset for Hands in Complex Egocentric Interactions )

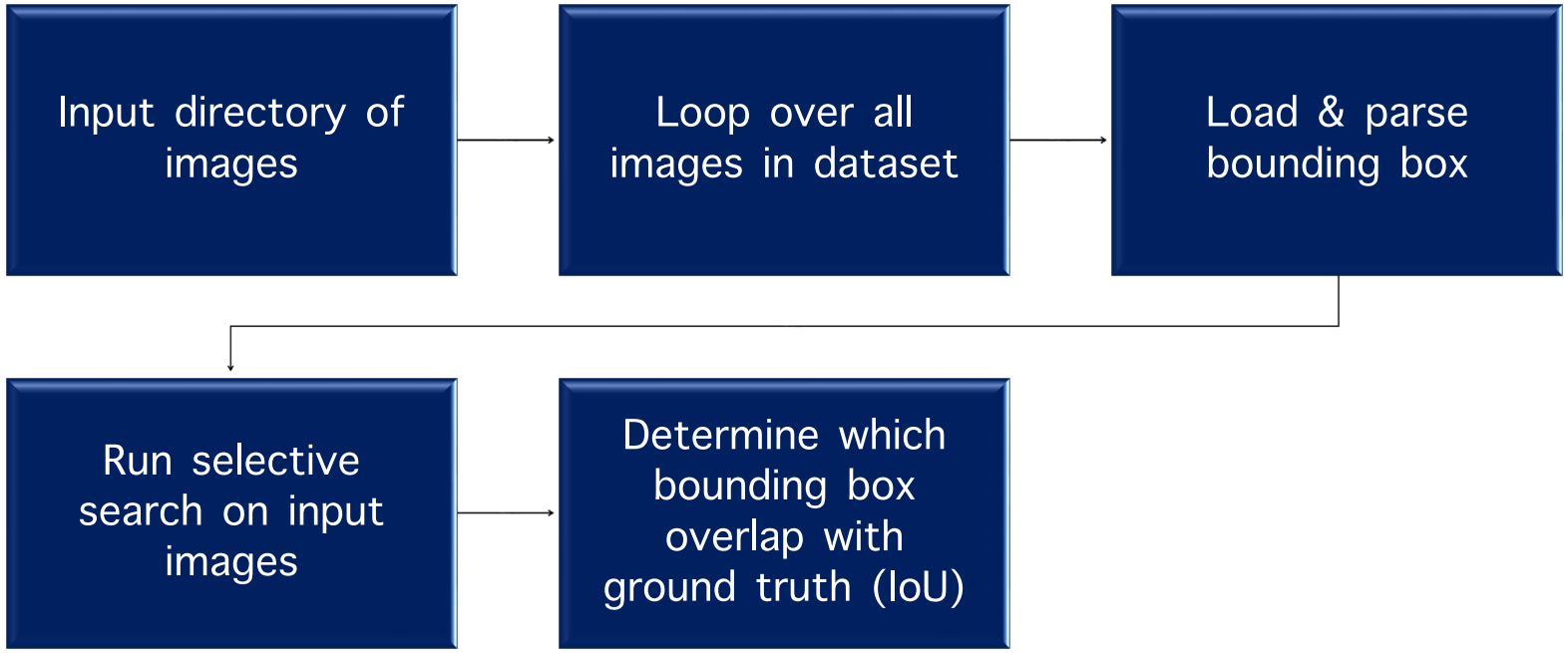
- The Ego- Hands dataset contains 48 Google Glass videos of complex( interactions between two people )
- 41 of the images are chosen from all kinds of hands
- ground truth (XML files containing the coordinates of the hands in each image )
- selective search algorithm is applied to each image

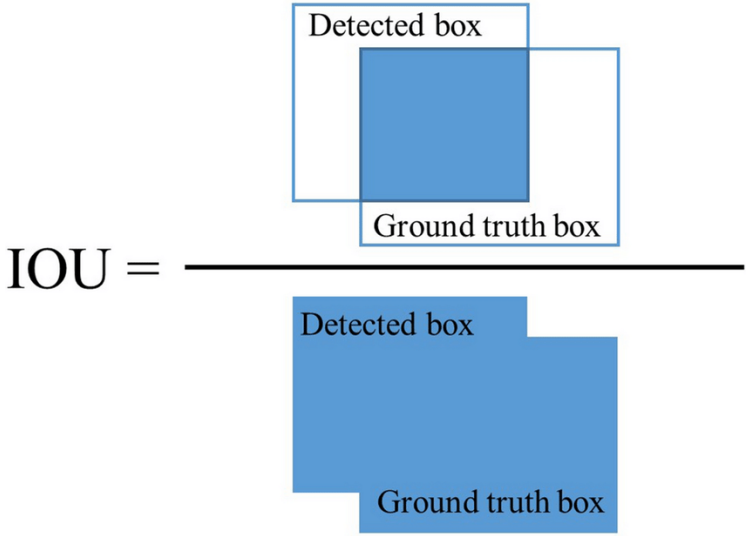



Examples of dataset

# Object detection dataset builder

Build an object detection dataset using Selective Search



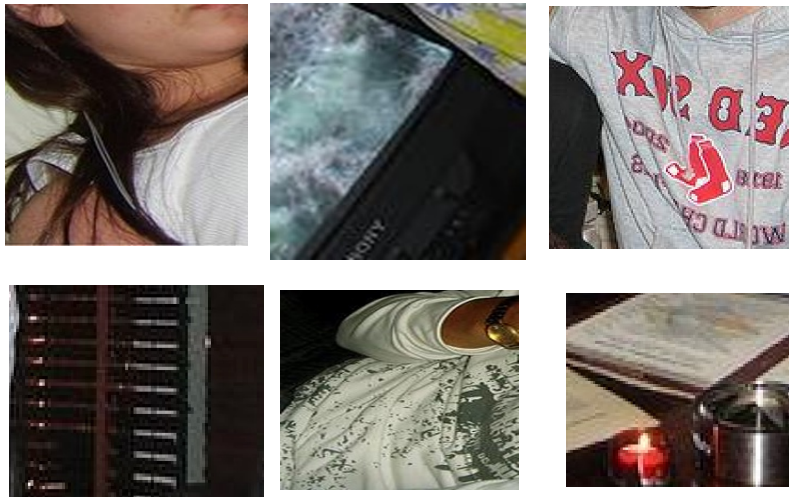
# DATASET Prepration

**Preparing our image dataset for object detection**

- IoU > 70 % : HAND
- Number of samples = 441
- INPUT_DIMS = (224, 224)

HAND dataset

- Not full Overlap and IoU < 5% : NON-HAND
- Number of samples = 190
- INPUT_DIMS = (224, 224)

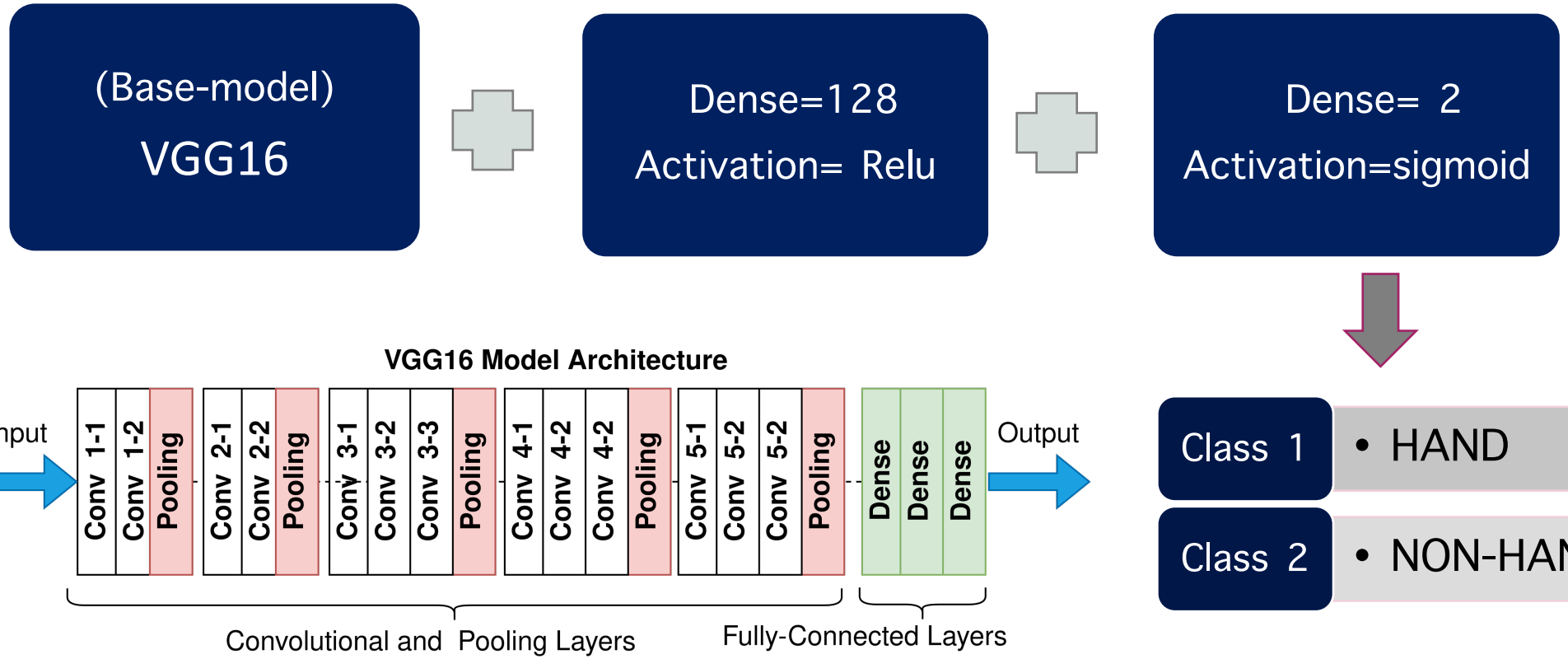


NON-HAND dataset

# Training: Fine-tuning a network for object detection

With our dataset created via the previous section, we're now ready to fine-tune a classification CNN to recognize both of these classes(hand, non-hand)

- Data augmentation (rotation, width shift, height shift)
- Splitting: 70% training, 30% testing



# RESULTS (classification of HAND, NON-HAND)

**Training & validation accuracy :** In 11th epoch, our model achieved 93.23% and 91.05% train and test accuracy, respectively

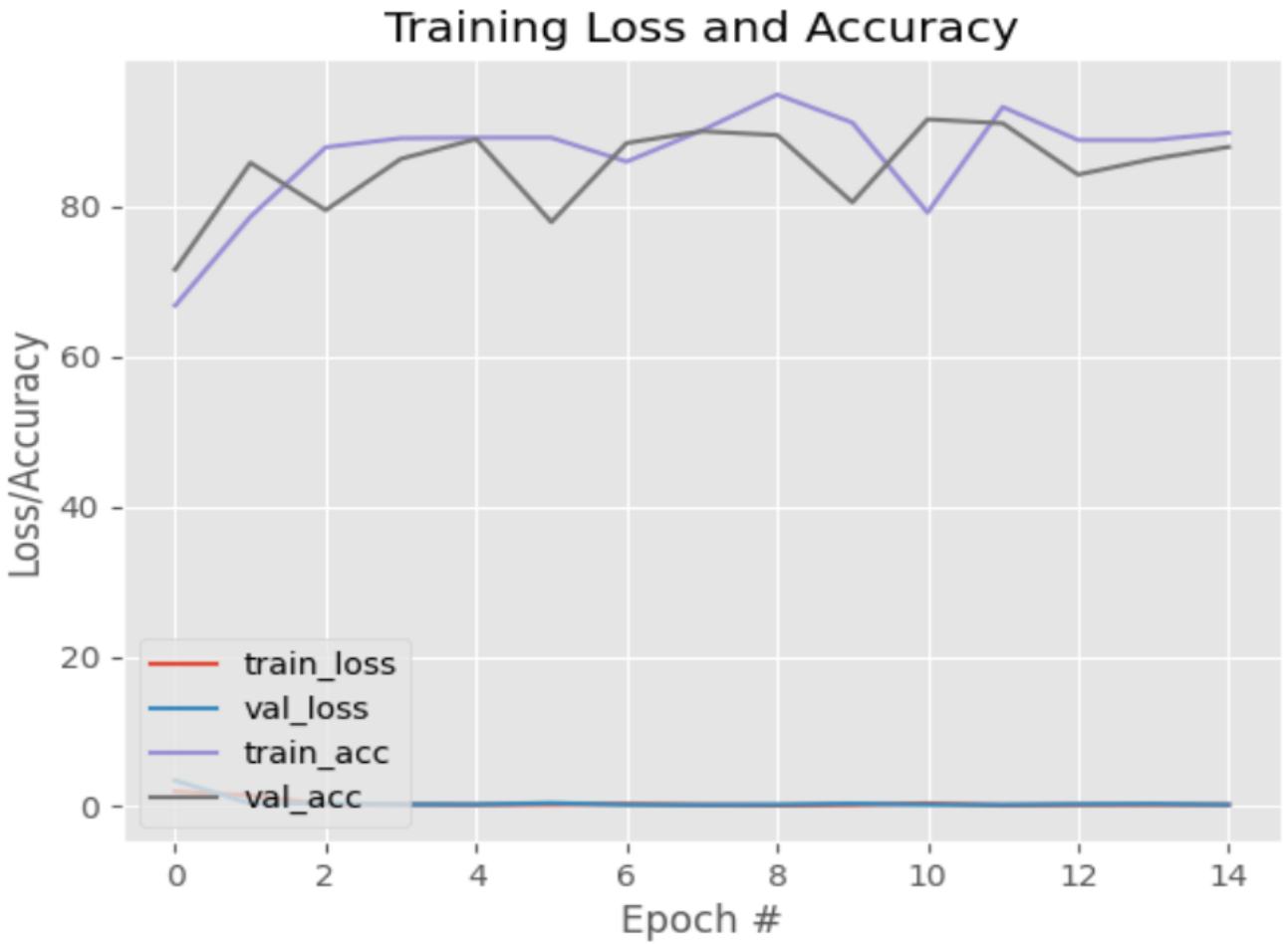
- Optimizer = RMSprop (learning rate =0.0006)
- Batch size = 128
- Epochs =15

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| hand | 0.71 | 0.96 | 0.82 | 54 |
| no_hand | 0.98 | 0.85 | 0.91 | 136 |
| accuracy | | | 0.88 | 190 |
| macro avg | 0.85 | 0.90 | 0.86 | 190 |
| weighted avg | 0.91 | 0.88 | 0.88 | 190 |

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

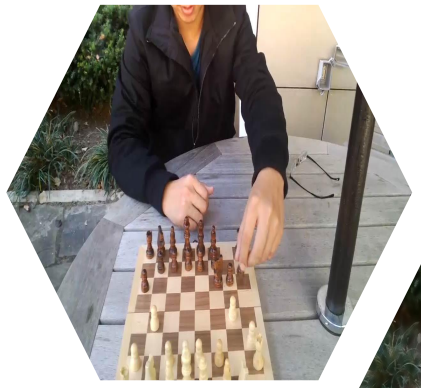
$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

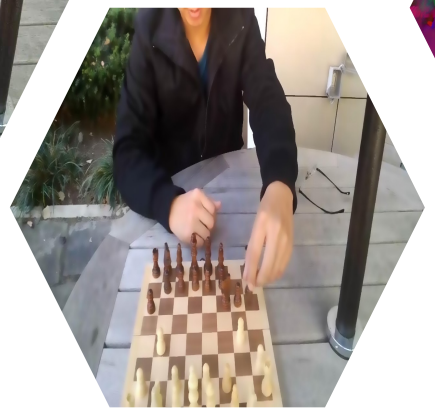


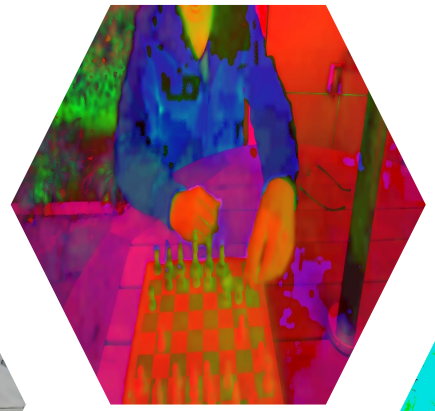
# METHODOLOGY: Meanshift segmentation

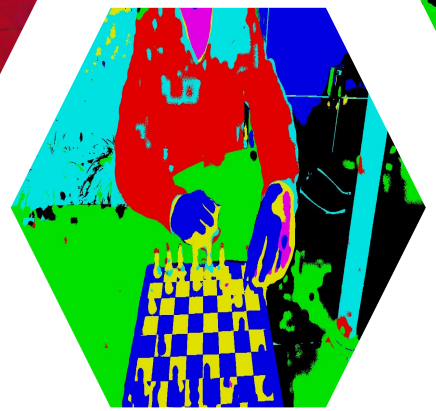
Put our trained model to work to perform object detection inference on new images.

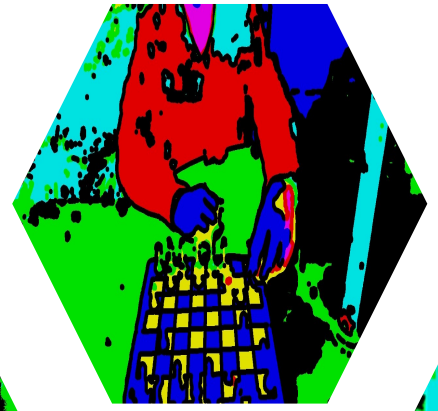
Input image → Bilateral → RGB to HSV → Mean shift → erosion → find Contours

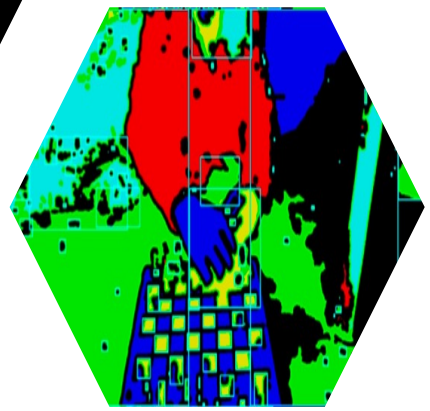










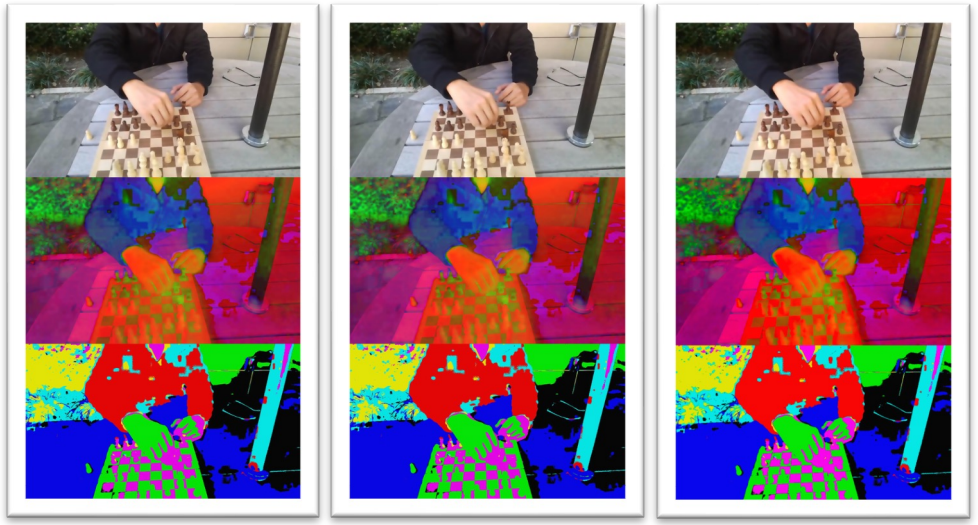


# METHODOLOGY: Meanshift segmentation

Mean Shift is a non-parametric clustering algorithm that can be used for image segmentation. It works by iteratively shifting each pixel in an image towards the direction of the highest density of nearby pixels until it converges to a stable state.

1. Choose a window size ("bandwidth") & starting pixel location
2. Define a kernel function(to weight the contribution of each pixel )
3. Shift the pixel to the new location

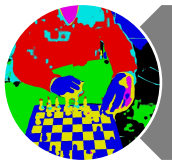
Repeat steps 2 and 3 until convergence



Kernel = 5 Kernel = 45 Kernel = 65

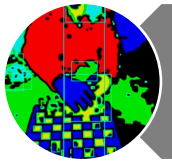
# METHODOLOGY: R-CNN (meanshift+cnn)

Put our trained model to work to perform object detection inference on new images.



Mean shift
Kernel size=(5,45,65)

erosion
Kernel size= 5

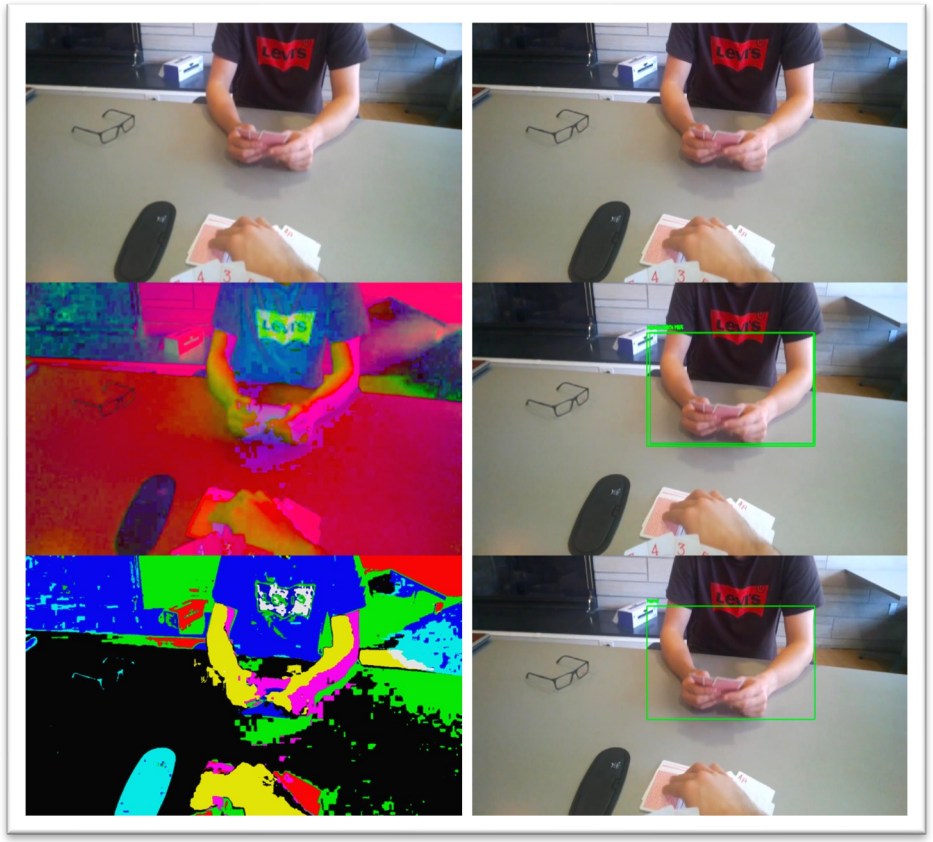


find Contours
Kernel size=(5,45,65)

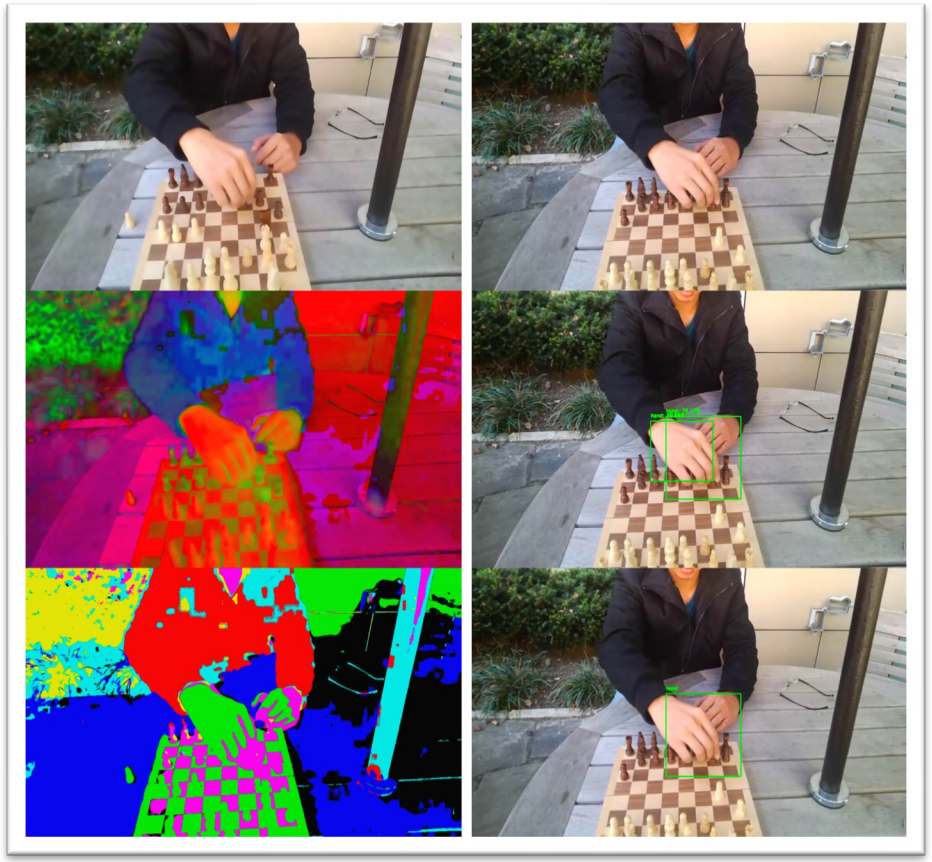
Apply mean shift segmentation and other filters on dataset → apply find Contours → make prediction on each proposal using fine-tuned model → Apply (non max suppression) → Return final object detection results

# RESULTS

## Demo: Images from a batch of test dataset to show the hand detection



Final result of R-CNN by mean shift segmentation for hand detection
kernel size= 65



Final result of R-CNN by mean shift segmentation for hand detection
kernel size= 5


THANK YOU

QUESTIONS ARE WELCOME