Ministry of Science and Higher Education of the Republic of Kazakhstan
L.N. Gumilyov Eurasian National University

Faculty of Information Technology
Department of Information Systems

Comparison of object (face) detection algorithms with Viola-Jones

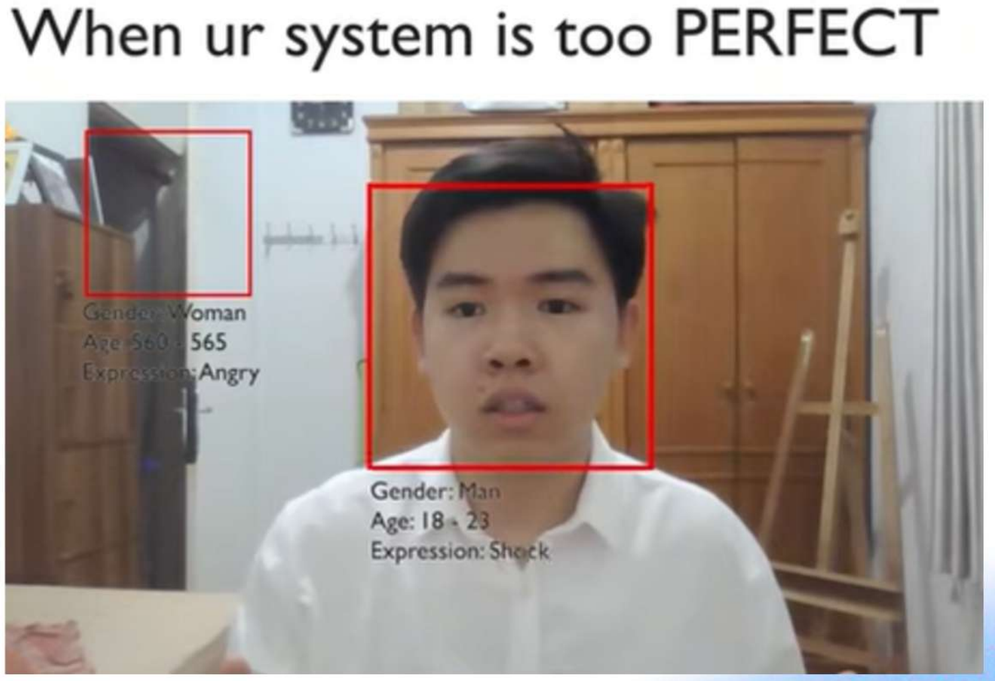
Done by: Iskakov Yerassyl

CHECKED BY: PROF. T.K. ZHUKABAYEVA

INTRODUCTION

The Face Detection task is easily done in the perspective of human visual task but when it comes in the view of computer it is little bit difficult. An image is given in which the faces are detected leaving the illumination, pose variation and lighting factors.

Over the years, face detection techniques have evolved from traditional methods to advanced deep learning-based models.



INTRODUCTION

I have chosed:

MTCNN is a deep learning-based face detection and alignment algorithm. It locates faces and detects key facial landmarks (eyes, nose, mouth, chin).

Faster R-CNN is an advanced object detection model that uses a Region Proposal Network (RPN) to efficiently generate face candidates before applying a convolutional network for precise face localization.

Comparison methodology indicators:

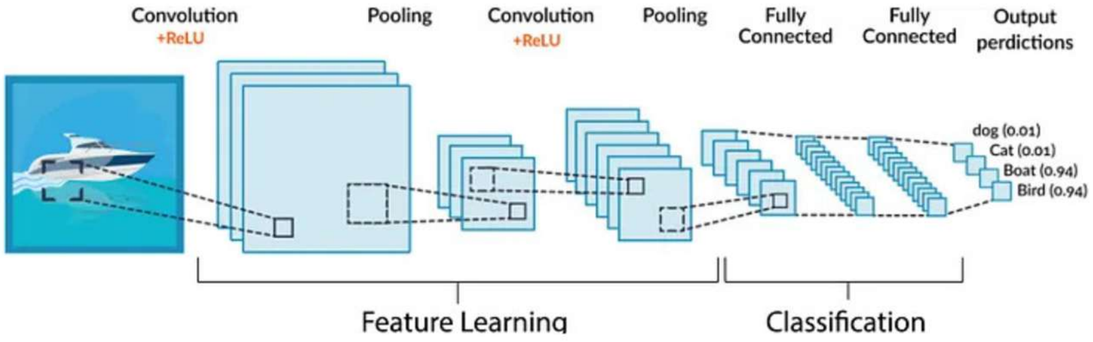
- detection accuracy;
- speed;
- pose invariance;
- lighting adaptability;
- computational complexity.

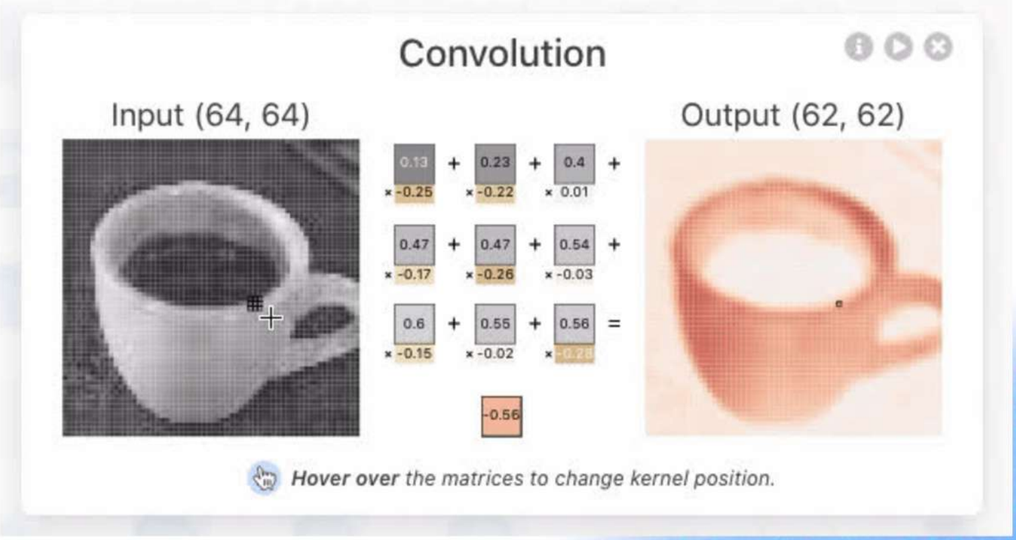
MTCNN THEORY: RECAP OF CNN

Since detection task involves images/video feeds, a neural network could be a potential candidate for solving the problem at hand.

1. Feature Learning: In reference to face detection, the feature learning task of CNN would involve learning how different parts of the face look depending on height, width, and other features.

1. Classification: The classification task assigns a probability for an entity in the image depending on the object to be predicted; In this case, the human face

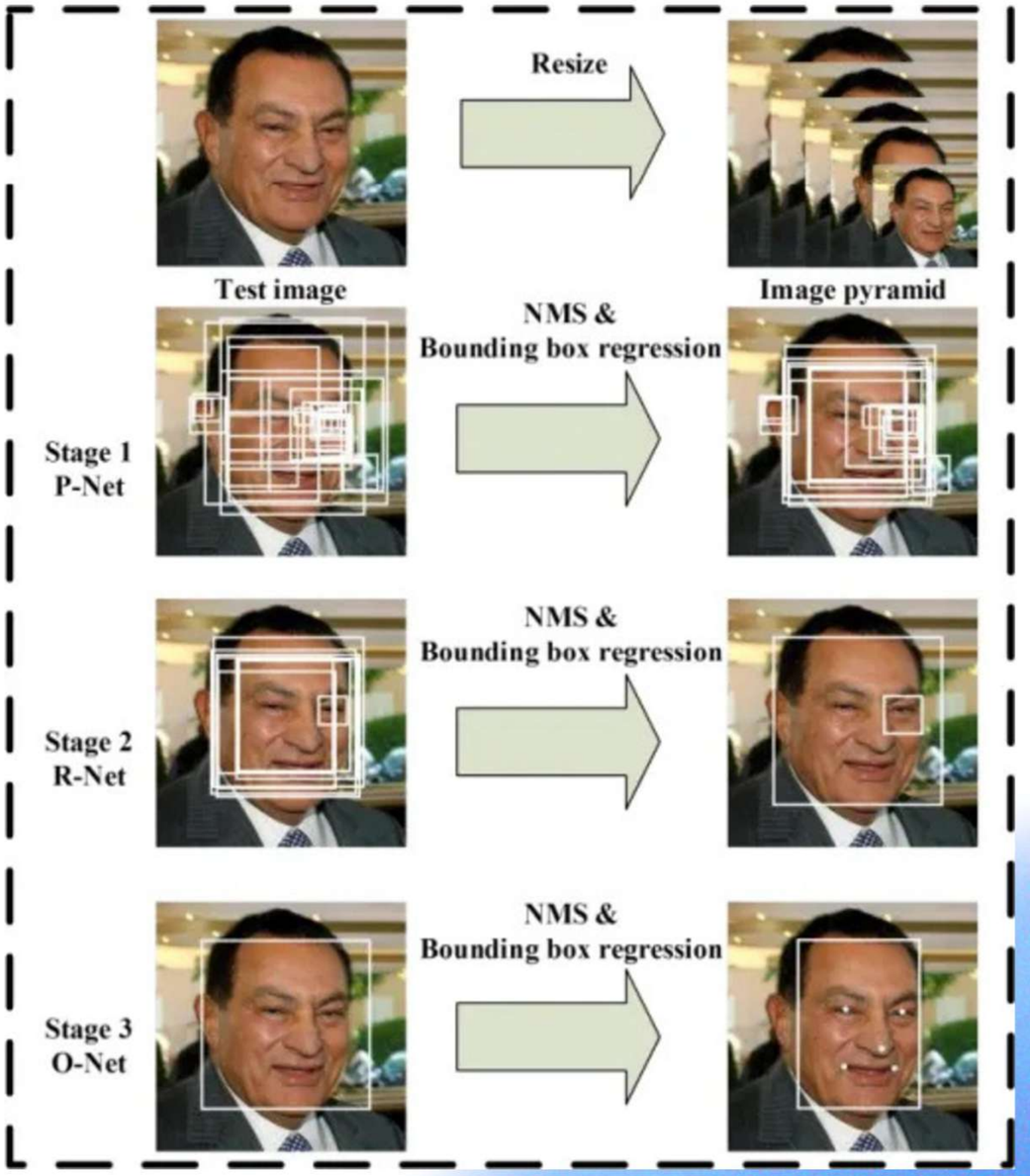




MTCNN THEORY

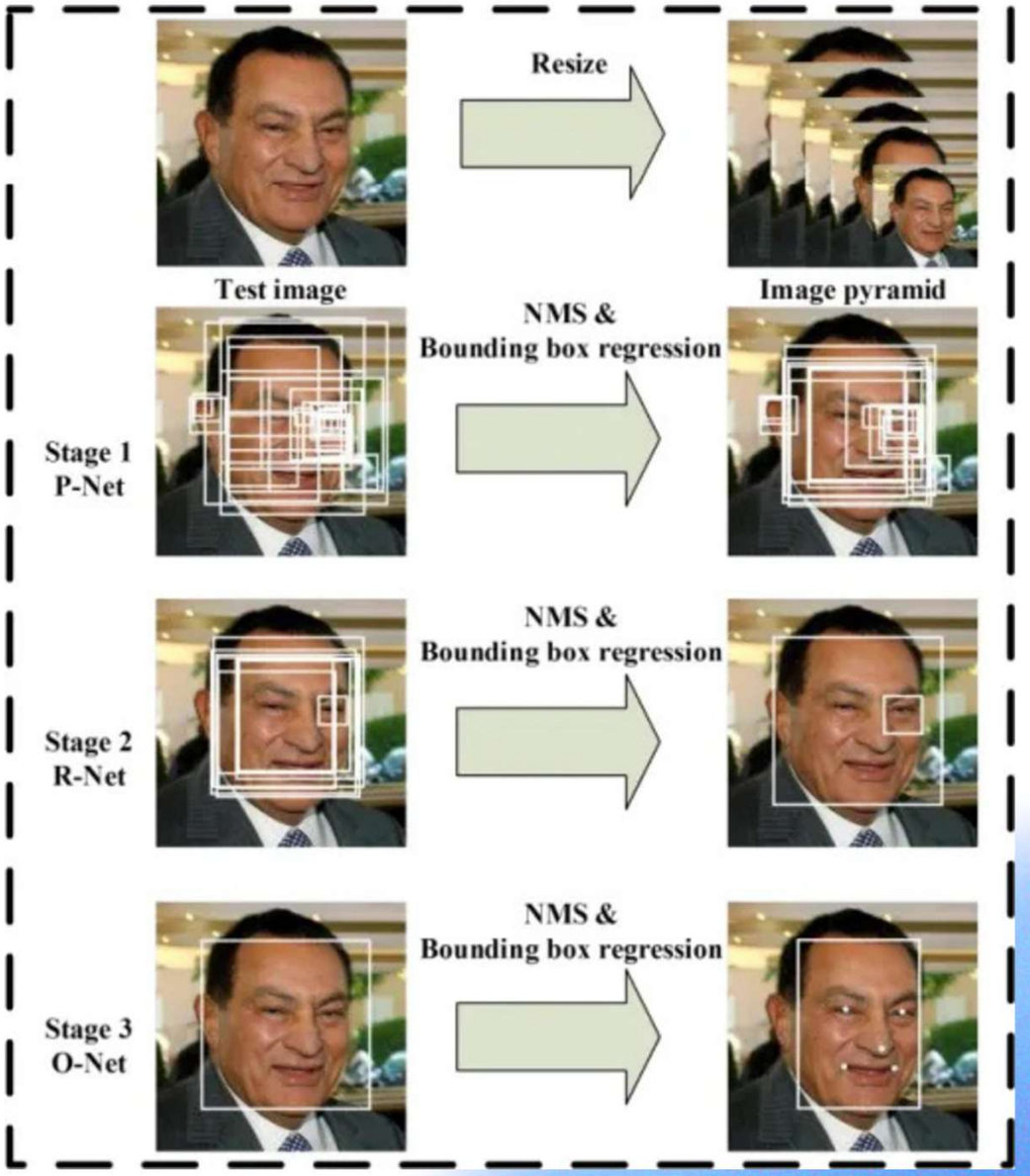
MTCNN (Multi-task Cascaded Neural Network) detects faces and facial landmarks on images/videos. This method was proposed by Kaipeng Zhang et al. in their paper 'Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks', IEEE Signal Processing Letters, Volume: 23 Issue: 10.

The whole concept of MTCNN can be explained in three stages out of which, in the third stage, facial detection and facial landmarks are performed simultaneously. These stages consists of various CNN's with varying complexities.



MTCNN THEORY

1. In the first stage the MTCNN creates multiple frames which scans through the entire image starting from the top left corner and eventually progressing towards the bottom right corner. The information retrieval process is called P-Net(Proposal Net) which is a shallow, fully connected CNN.
2. In the second stage all the information from P-Net is used as an input for the next layer of CNN called as R-Net(Refinement Network), a fully connected, complex CNN which rejects a majority of the frames which do not contain faces.
3. In the third and final stage, a more powerful and complex CNN, known as O-Net(Output Network), which as the name suggests, outputs the facial landmark position detecting a face from the given image/video.

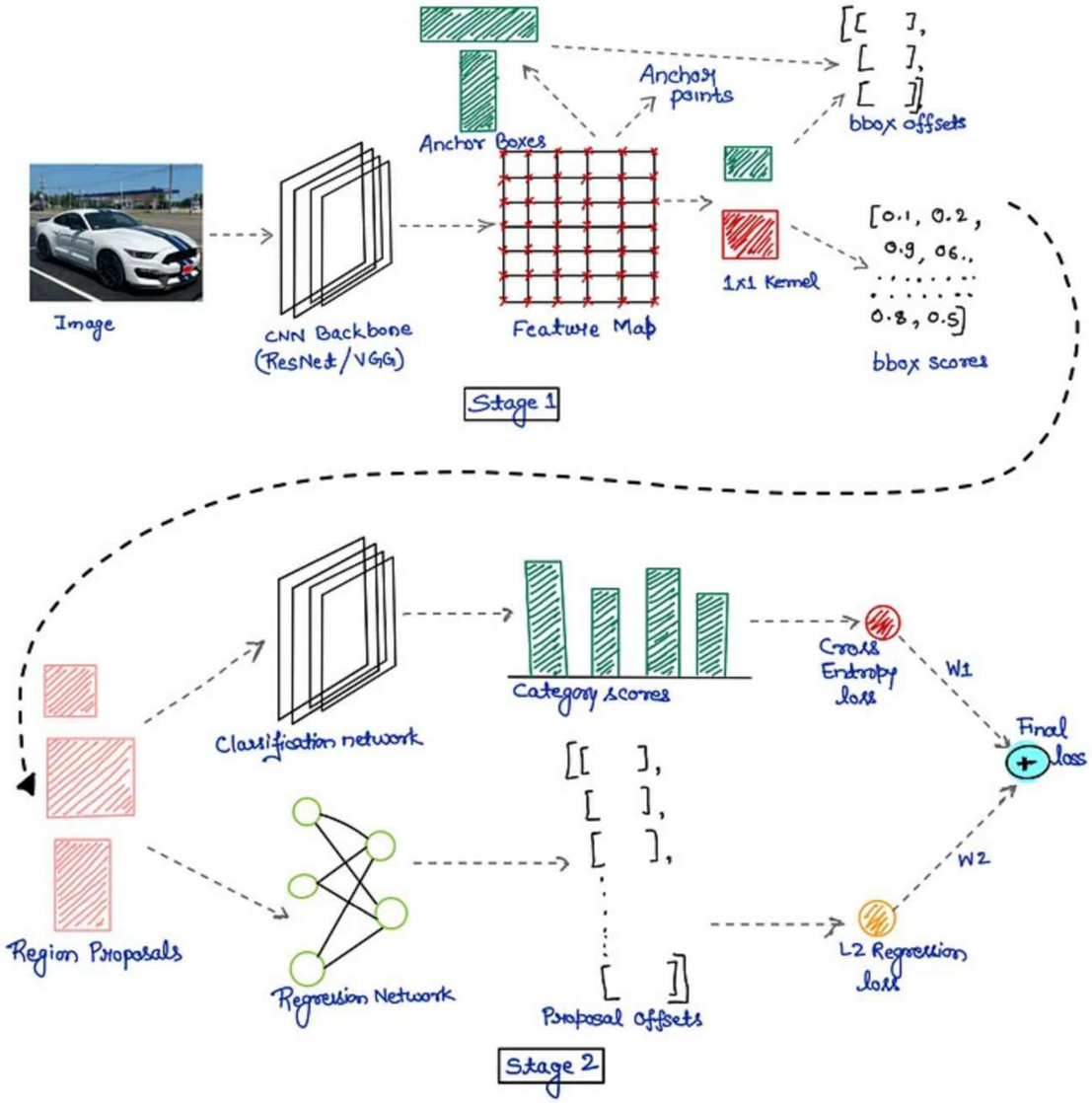


FASTER R-CNN THEORY

Fast R-CNN was introduced in April 2015. It was faster than R-CNN. Faster R-CNN was introduced by Ross Girshick et al. in the June of same year and its much faster that Fast R-CNN.

Faster R-CNN is an object detection model that identifies objects in an image and draws bounding boxes around them, while also classifying what those objects are. It's a two-stage detector:

1. Stage 1: Proposes potential regions in the image that might contain objects. This is handled by the Region Proposal Network (RPN).
2. Stage 2: Uses these proposed regions to predict the class of the object and refines the bounding box to better match the object.



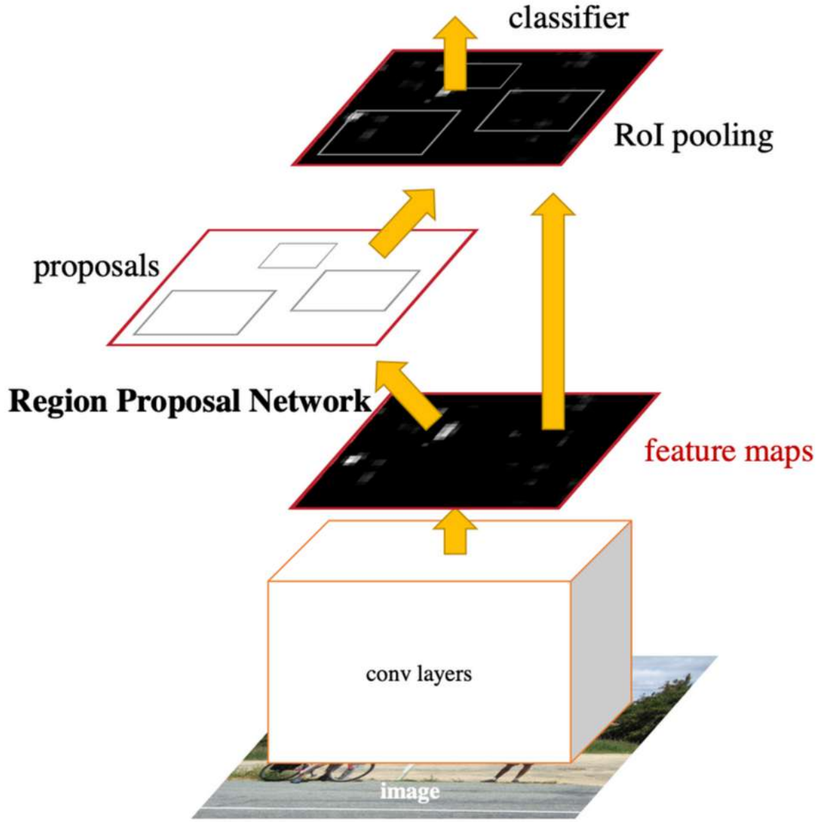
FASTER R-CNN THEORY

Stage 1: Region Proposal Network (RPN)

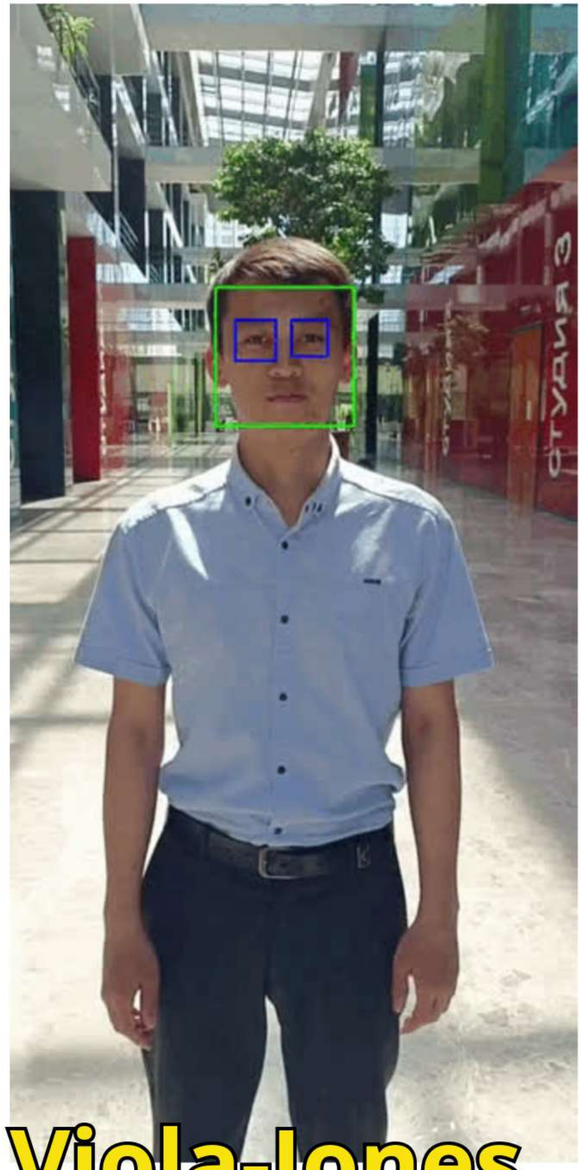
- Backbone Network: Extracts features from the image.
- Anchors: Predefined boxes predict possible object locations.
- Classification: Identifies foreground (object) and background.
- Bounding Box Refinement: Adjusts anchor boxes for better alignment.

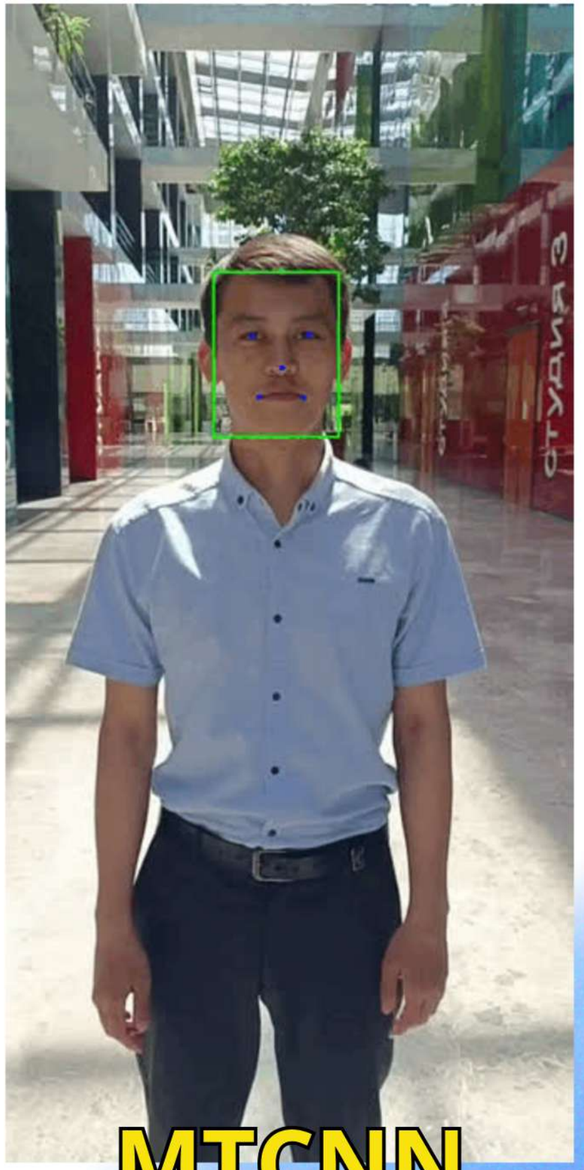
Stage 2: Object Classification & Box Refinement

- Region Proposals: Selected from RPN for further processing.
- ROI Pooling: Resizes proposals to a fixed size.
- Object Classification: Predicts object category.
- Final Box Refinement: Further adjusts box positions.



EXPERIMENTAL RESULTS





COMPARISON TABLE

Indicator	MTCNN (Multi-task Cascaded CNN)	Faster R-CNN	Viola-Jones (Haar Cascades)
Detection Accuracy	Medium	High	Low
Speed	Medium (faster than Faster R-CNN but slower than Viola-Jones)	Slow	Very Fast
Pose Invariance	Good	High	Bad
Lighting Adaptability	Good	High (handles lighting variations well)	Bad
Computational Complexity	High	High	Low

CONCLUSION

In comparing MTCNN, Faster R-CNN, and Viola-Jones for face detection, each method has distinct advantages. Faster R-CNN provides the highest accuracy but requires computational resources.

MTCNN balances accuracy and speed, making it suitable for real-time applications.

Viola-Jones, though outdated, remains the fastest on low-end hardware but lacks robustness. Ultimately, the choice depends on the application's need for accuracy, speed, and computational efficiency.

REFERENCES

1. https://paperswithcode.com/method/faster-r-cnn
2. https://medium.com/dummykoders/face-detection-using-mtcnn-part-1-c35c4ad9c542
3. https://poloclub.github.io/cnn-explainer/
4. https://medium.com/@RobuRishabh/understanding-and-implementing-faster-r-cnn-248f7b25ff96