

Object Detection and Tracking

DIH & University of Salzburg: *Deep Learning Workshop 2023*

Robert Jöchl and Andreas Uhl

Universität Salzburg
Department of Artificial Intelligence and Human Interfaces
A-5020 Salzburg, Austria

12.10.2023 - 13.10.2023



1 Introduction

- Object Detection and Tracking Literature
- What is planned

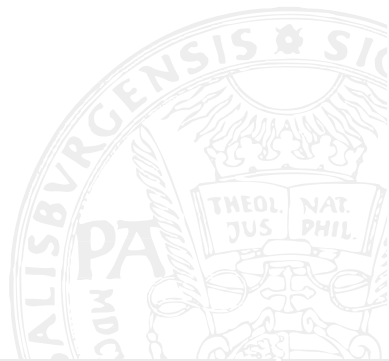
2 Object Detection

- Traditional Object Detection Methods and Strategies
- Deep Learning Based
 - Two-Stage Detectors
 - One-Stage Detectors

3 Object Tracking

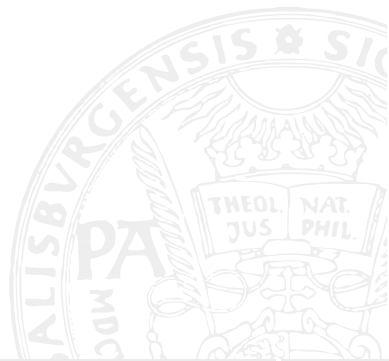
- Multiple Object Tracking
 - Motion Model
 - Appearance Model
 - Possible Extensions
 - Multiple Object Tracker

4 Evaluation Metric

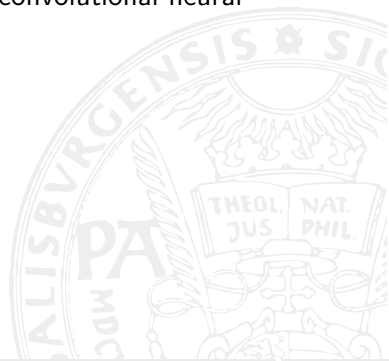


Object detection and **Object tracking** are:

- two of the most challenging and important branches in computer vision.
- necessary for applications in various fields:
 - health care monitoring
 - autonomous driving
 - anomaly detection
 - surveillance
 - ...

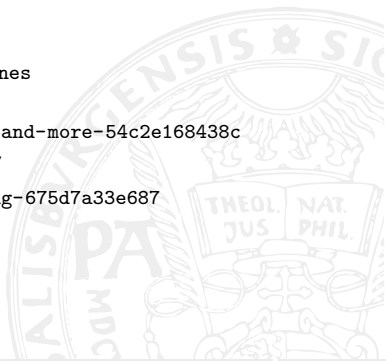


- Jiang, Xiaoyue et al. Deep Learning in object detection and recognition. Springer, 2019
- Michelucci, Umberto. Advanced applied deep learning: convolutional neural networks and object detection. Springer, 2019
- ...



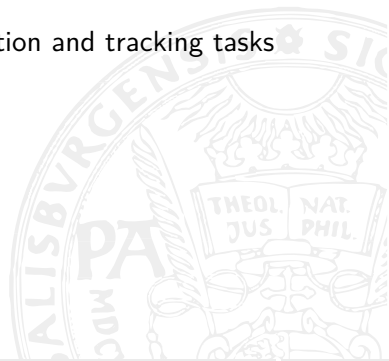
- Pal, Sankar K et al. Deep learning in multi-object detection and tracking: state of the art. Applied Intelligence, 2021
- Roth, Peter M and Winter, Martin. Survey of appearance-based methods for object recognition. Technical report ICGTR0108, 2008
- Chen, Zhihao et al. Real time object detection, tracking, and distance and motion estimation based on deep learning: Application to smart mobility, International Conference on Emerging Security Technologies, 2019
- Jiao, Licheng et al. A survey of deep learning-based object detection. IEEE access, 2019
- Huang, Rachel and Pedoem, Jonathan and Chen, Cuixian. YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. IEEE International Conference on Big Data, 2019
- Fang, Wei and Wang, Lin and Ren, Peiming. Tinier-YOLO: A real-time object detection method for constrained environments. IEEE access, 2019
- He, Zhenwei and Zhang, Lei. Multi-adversarial faster-rcnn for unrestricted object detection. International Conference on Computer Vision, 2019
- Roh, Myung-Cheol and Lee, Ju-young. Refining faster-RCNN for accurate object detection. International conference on machine vision applications, 2017
- Voigtlaender, Paul et al. Mots: Multi-object tracking and segmentation. Conference on Computer Vision and Pattern Recognition, 2019
- Jiao, L. et al. A survey of deep learning-based object detection. IEEE access, 7, 128837-128868, 2019
- ...

- A Beginner's Guide to Object Detection:
<https://www.datacamp.com/community/tutorials/object-detection-guide>
- Object Detection Guide: <https://www.fritz.ai/object-detection/>
- How to implement a YOLO (v3) object detector from scratch in PyTorch:
<https://blog.paperspace.com/how-to-implement-a-yolo-object-detector-in-pytorch/>
- Guide to Object Detection using PyTorch: <https://medium.com/analytics-vidhya/guide-to-object-detection-using-pytorch-3925e29737b9>
- Face Detection using Viola Jones Algorithm:
<https://www.mygreatlearning.com/blog/viola-jones-algorithm/>
- Viola Jones Alogrithm: <https://github.com/Donny-Hikari/Viola-Jones>
- A Guide to two-stage object detection: <https://medium.com/codex/a-guide-to-two-stage-object-detection-r-cnn-fpn-mask-r-cnn-and-more-54c2e168438c>
- Object tracking: <https://viso.ai/deep-learning/object-tracking/>
- Object tracking: <https://medium.com/visionwizard/object-tracking-675d7a33e687>
- ...

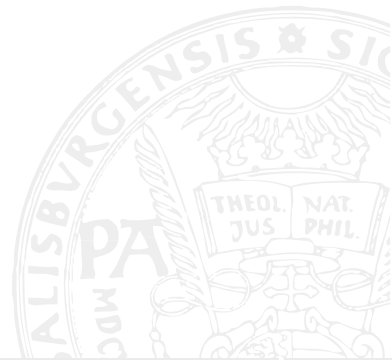


Plan of today's course...

- 1 Overview of basic concepts in object detection and tracking
 - Traditional ideas
 - Deep-Learning ideas
- 2 Discuss several strategies and approaches to fulfil detection and tracking tasks
- 3 Get to know sample implementations of both topics...

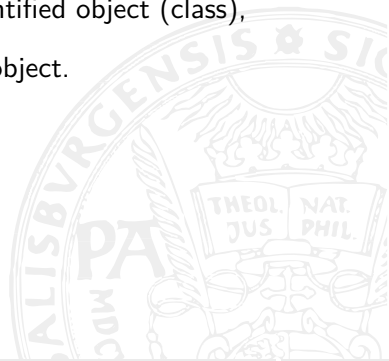


Object Detection



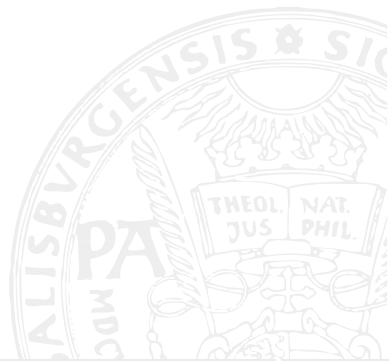
The objectives of an object detector are:

- identify which objects (classes) are present in an image,
- to indicate the associated confidence score for each identified object (class),
- return a bounding box to identify the location of each object.

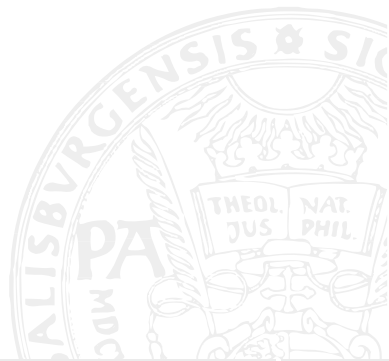


In general object detection is a 2 step process:

- 1 Find the foreground entities which are considered as object hypothesis.
- 2 Verify the candidates using a classifier.



- Variations in scale
- Orientation
- Lighting conditions
- Occlusions
- Complex background
- Too many objects
- Tiny objects
- ...



Example: Face Detection

“The goal of face detection is to determine whether there are any faces, and if any, return the bounding box of each face.¹”

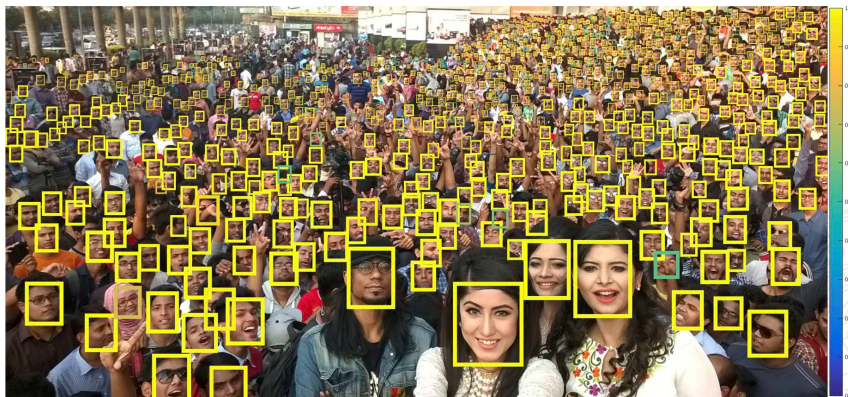


Figure: Source: Peiyun Hu and Deva Ramanan. “Finding tiny faces”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 951–959

¹Shifeng Zhang et al. “Refineface: Refinement neural network for high performance face detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

Object Detection Schematic Description

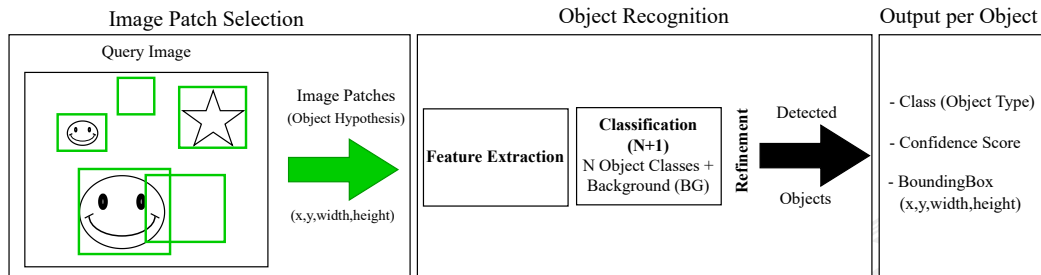
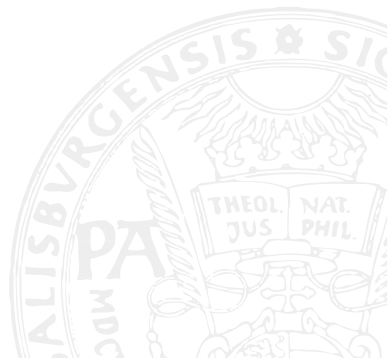
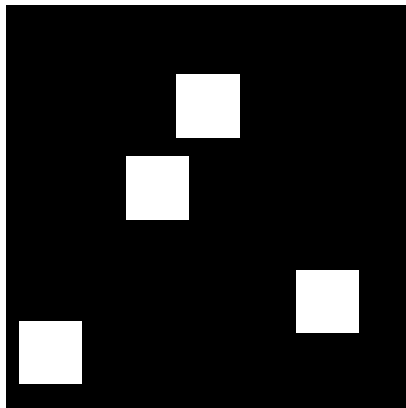


Figure: Simplified schematic description of an object detection system.

Coding Example 01 - Build a Simple Object Detector

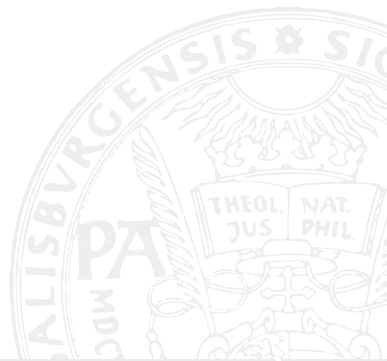
Coding Example: Build a simple object detector that detects all white rectangles in a back image.



A naive approach is to select the image patches using a sliding window (brute force).

Computationally very expensive:

- search all possible locations in the image,
- consider different windows scales,
- and aspect ratios.



Region Proposal Algorithms

Output bounding boxes for all regions of an image that are most likely to be objects.

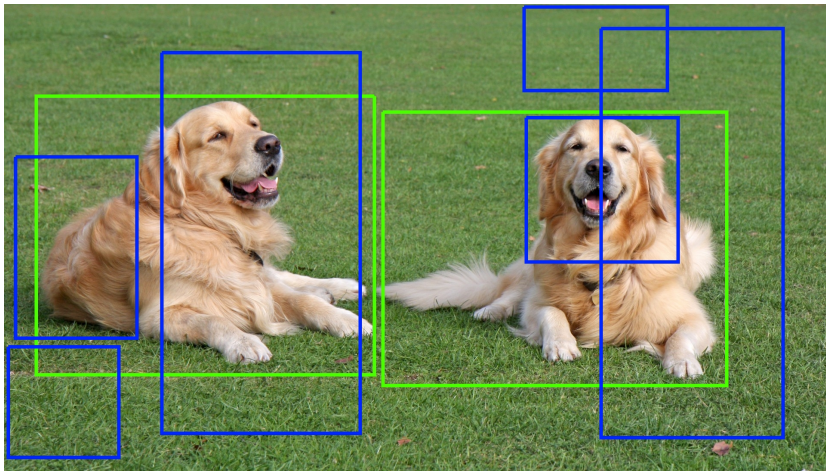


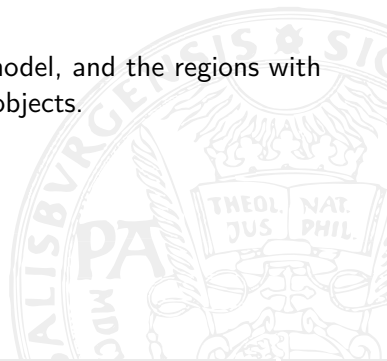
Figure: Output region proposal, green boxes represent True Positives and blue boxes False Positives. Source: Chandel Vaibhaw Singh. *Object Detection vs. Object Recognition*.

<https://learnopencv.com/selective-search-for-object-detection-cpp-python/>. [Online, Accessed 05.10.2023]

The returned region proposals can be:

- noisy,
- overlapping,
- and may not contain the object completely.

All region suggestions are passed to the object recognition model, and the regions with the highest confidence scores represent the locations of the objects.

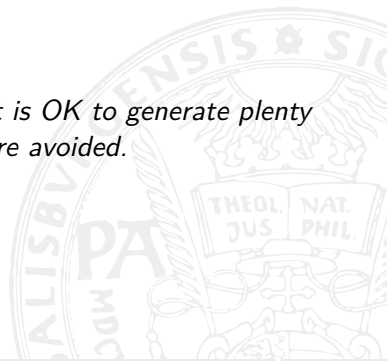


The basic idea is to group adjacent regions that are similar to each other. Criteria for similarity could be for example:

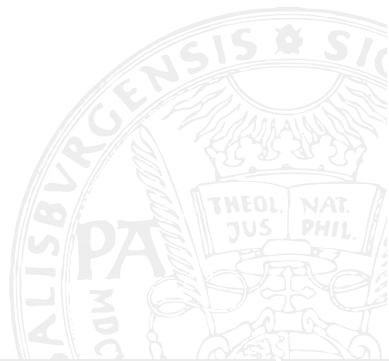
- color,
- or texture.

The resulting groups represent the region proposals.

Region proposal algorithms should have a high recall, i.e., it is OK to generate plenty of false positives as long as false negatives are avoided.



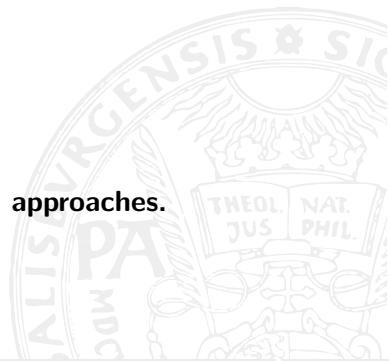
Traditional Object Detection Methods and Strategies



Traditionally used features are:

- Low level features (e.g., skin color, motion based)
- Active shape model (e.g., Snakes)
- Histogram of oriented gradients (HOG)
- Local Binary Pattern (LOB)
- Scale-Invariant Feature Transform (SIFT)
- ...

Traditional methods are non-neural network approaches.



Viola Jones² is a traditional object detection framework proposed 2001 by Paul Viola and Michael Jones.

- Compute low level rectangular features (or low level classifiers) based on simple Haar features.
- These low level classifiers (binary) are combined to a **classifier cascade**.
- Adaptive Boosting (**AdaBoost**) is used for training.
- Viola Jones selects the image patches using a sliding window approach with a fixed window size (square). Different scales are detected using an image pyramid.

²Paul Viola and Michael Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. Ieee. 2001, pp. 1-1.

Cascading Classifiers

The classifier cascade is an ensemble of several weak classifiers leading to a strong classifier. Each image patch (sub-image) is propagated through the cascade. If all classifiers approve the image patch, an object is detected.

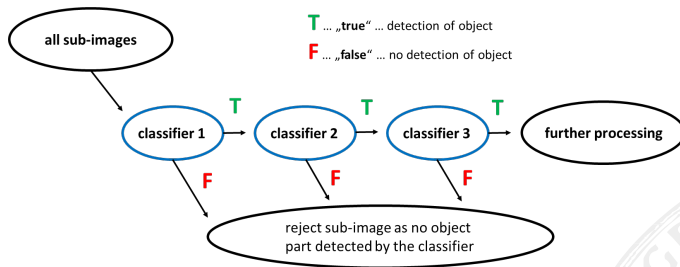


Figure: Overview of the classifier cascade.

The idea of the cascade is to speed up the classification by rejecting as many non-objects as quickly as possible, i.e., the first classifier in the cascade should be trained to already reject the vast majority of non-object image patches.

Classification Features - Haar-Like Features

The value of each feature is the sum of the pixel values in the black area minus the sum of the pixel values in the white area. This can be performed efficiently by using integral images.



Figure: Examples of Haar-Like features.

Integral Image

The value of each point in an integral image is the sum of all pixels above and to the left, including the target pixel.

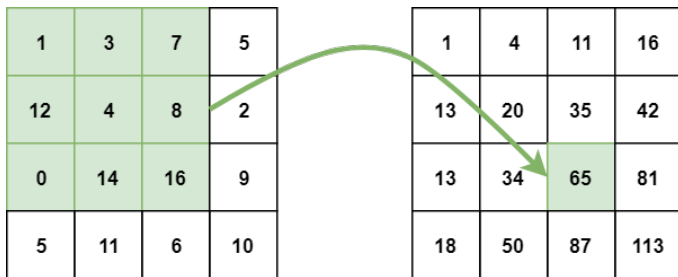


Figure: Source: Kristijan Ivancic. *Traditional Face Detection With Python*.
<https://realpython.com/traditional-face-detection-python/>. [Online, Accessed 05.10.2023]

Integral Image

What is the sum of pixels in the rectangle ABCD?

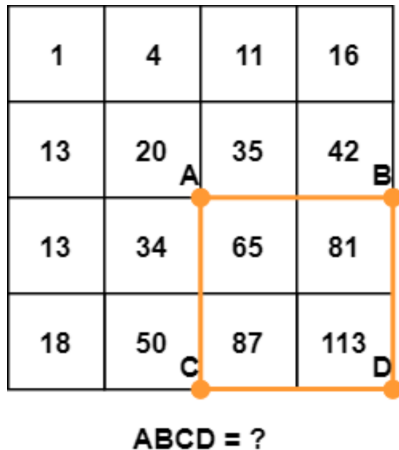


Figure: Source: Kristijan Ivancic. *Traditional Face Detection With Python*.
<https://realpython.com/traditional-face-detection-python/>. [Online, Accessed 05.10.2023]

Integral Image

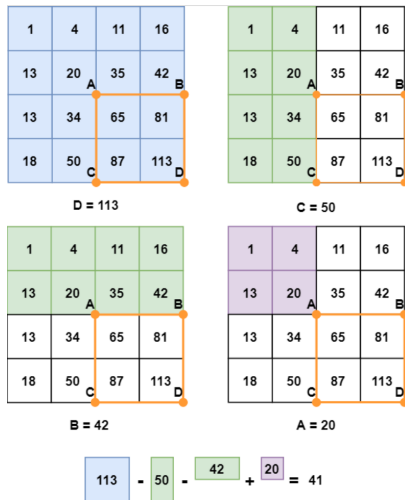
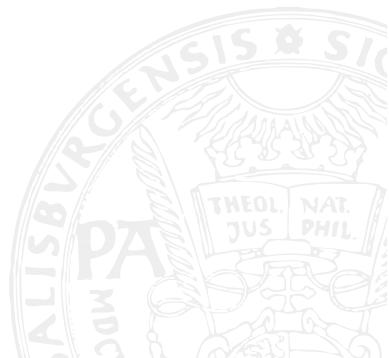
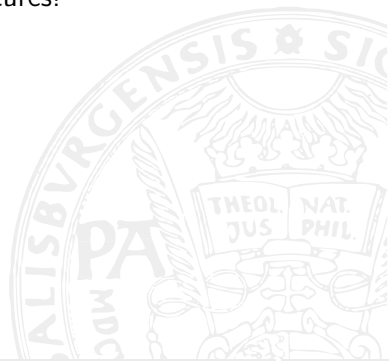


Figure: Source: Kristijan Ivancic. *Traditional Face Detection With Python*.
<https://realpython.com/traditional-face-detection-python/>. [Online, Accessed 05.10.2023]



Integral images provide a simple and fast way to calculate the sum of pixel values of a rectangle. This is ideal for Haar-like features!



Face Detection based on Haar-Like Features

For example, in an image showing a person's face, usually the eye region is darker than the nose and the cheeks are brighter than the eye region.

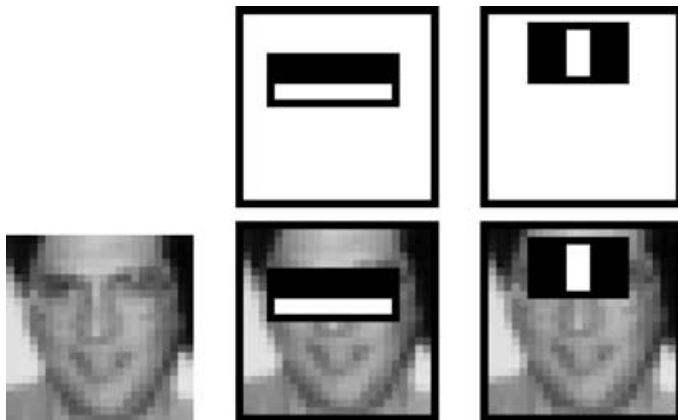


Figure: The first and second low level features selected.

The basic idea of boosting is to combine several weak classifiers (slightly better than random guessing) into a single strong classifier.

For example:

- Each Haar-Like feature represents a weak classifier.
- To decide which features to include in the final classifier, AdaBoost checks the performance of all provided classifiers (at least better than random prediction).
- The end result is a strong classifier containing the best performing weak classifiers.

It is called adaptive because the weak classifiers that perform better on hard examples are weighted more than others.

AdaBoost Example

Build a combined classifier that separates red rectangles from blue triangles.

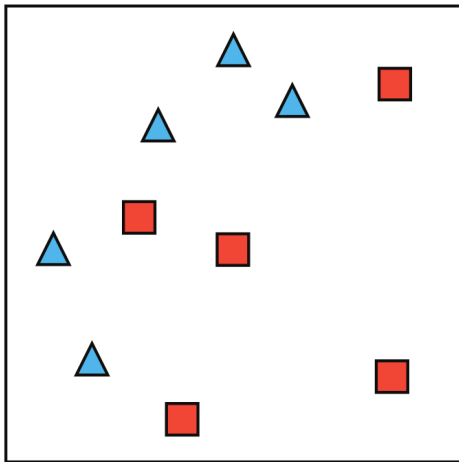


Figure: Source: Luis Serrano. *Grokking Machine Learning*. Simon and Schuster, 2021

AdaBoost Example

First weak classifier.

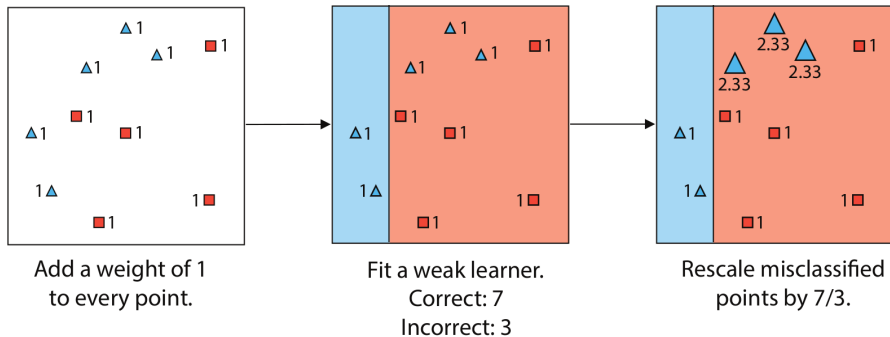


Figure: Source: Luis Serrano. *Grokking Machine Learning*. Simon and Schuster, 2021

AdaBoost Example

Second weak classifier, for which the sum of the weights of the correctly classified points is the highest.

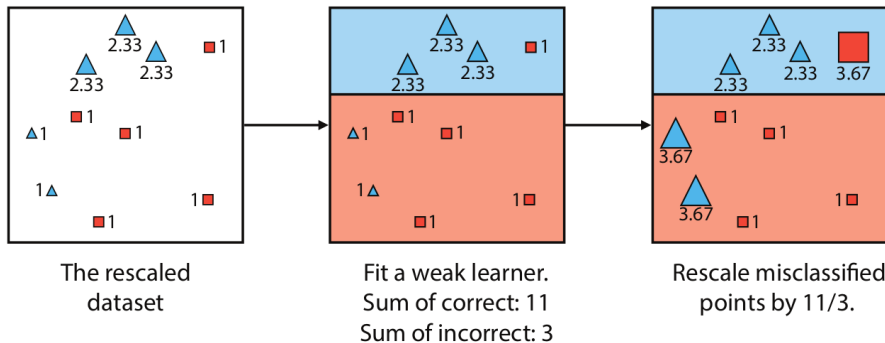


Figure: Source: Luis Serrano. *Grokking Machine Learning*. Simon and Schuster, 2021

AdaBoost Example

Third weak classifier.

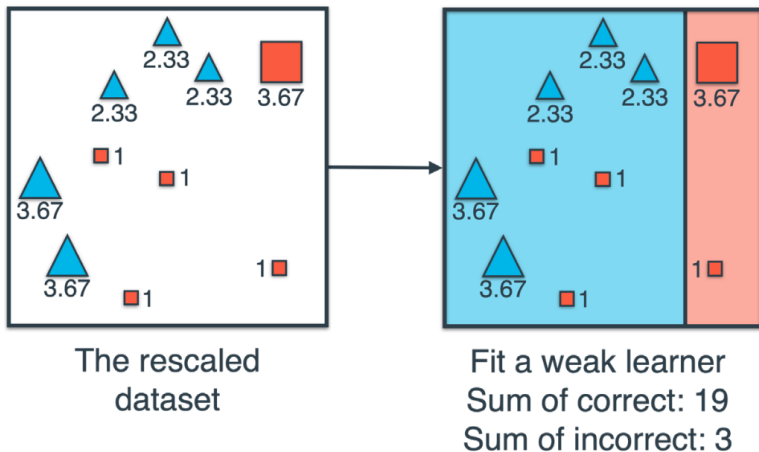


Figure: Source: Luis Serrano. *Grokking Machine Learning*. Simon and Schuster, 2021

AdaBoost Example

Combining the classifiers by the sum of the scores (predictions are $+1$ and -1 , instead of 1 and 0). The scores for each classifier are obtained by the log-odds.

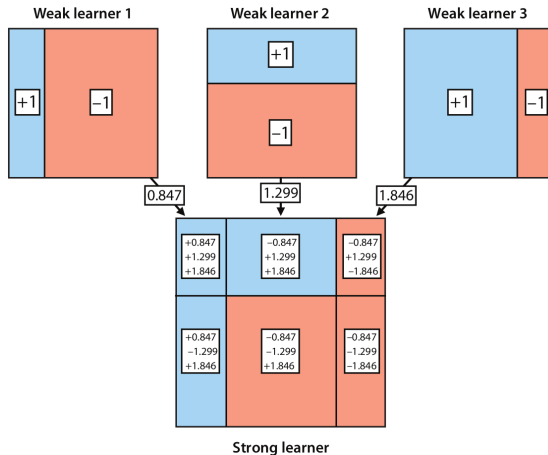


Figure: Source: Luis Serrano. *Grokking Machine Learning*. Simon and Schuster, 2021

AdaBoost Example

Assign a prediction of 1 if the sum of scores is greater than or equal to 0 and a prediction of 0 otherwise.

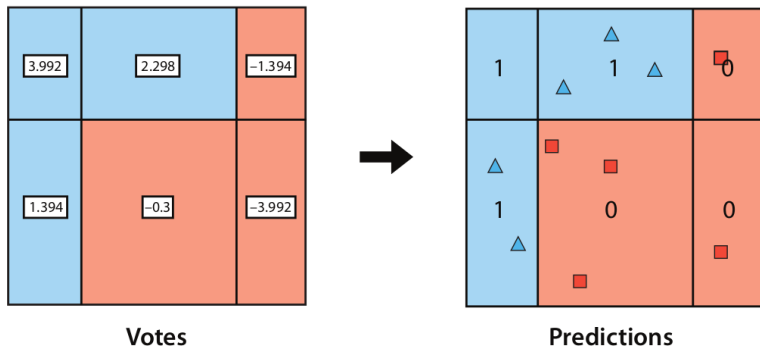
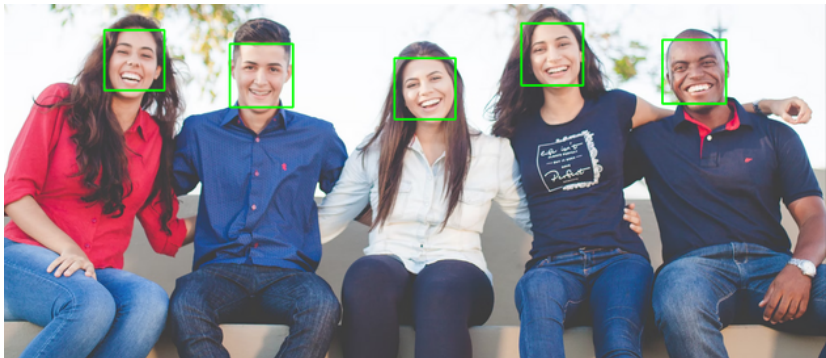


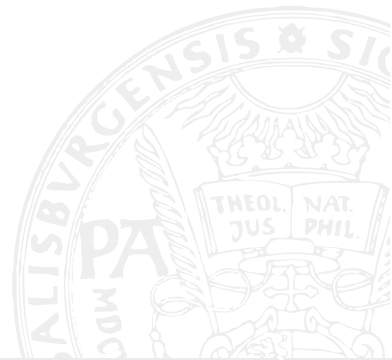
Figure: Source: Luis Serrano. *Grokking Machine Learning*. Simon and Schuster, 2021

Coding Example 02 - Viola Jones Face Detector

Coding Example: Detect all the faces in an image by using the Viola Jones face detector.



Deep Learning Based Object Detection



Can basically be divided into two types:

- Two-Stage Detector:
 - Has excellent performance but suffers from long latency and slower speed.
- One-Stage Detector:
 - A one-stage detector has the potential to be faster and simpler.
 - Problem of one-stage detector is the extreme imbalance of the foreground-background class.
 - Can be addressed by 'Online Hard Example Mining (OHEM)³'

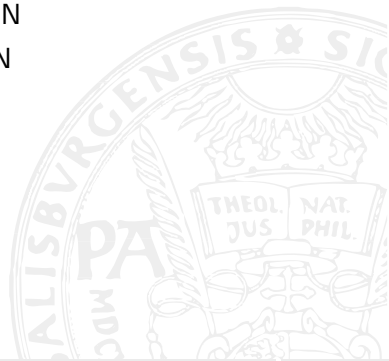
³Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. "Training region-based object detectors with online hard example mining". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 761–769.

One-Stage Detector:

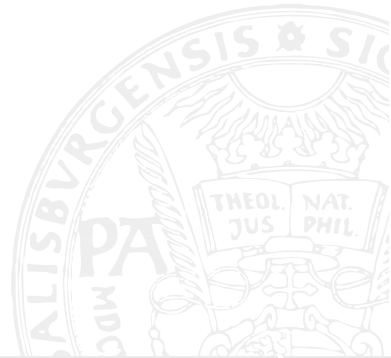
- SSD
- ASFD
- Refineface
- YOLO
- ...

Two-Stage Detector:

- R-CNN
- Fast R-CNN
- Faster R-CNN
- Mask R-CNN
- ...



Two-Stage Object Detectors



Concept of Two-Stage Object Detection

As the name implies, detectors belonging to this class combine two different models to fulfill the task of object detection:

- 1 Model: *extract the regions of the objects*
 - 2 Model: *classify and further refine the localization of the object*
- Compared to one-stage detectors these methods are usually slower in inference, but more accurate in terms of high localization and object recognition accuracy
 - Recent developments in the field try to close the computational costs difference

Two-Stage Detectors

Two stages:

- 1 proposals for object regions (either conventional or deep learning based methods),
- 2 object classification, based on extracted features.

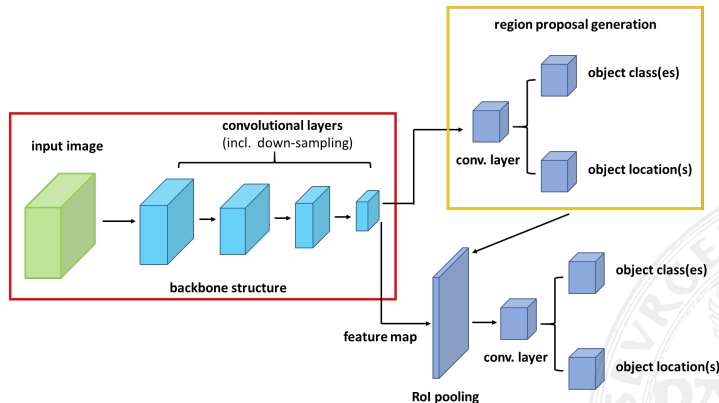
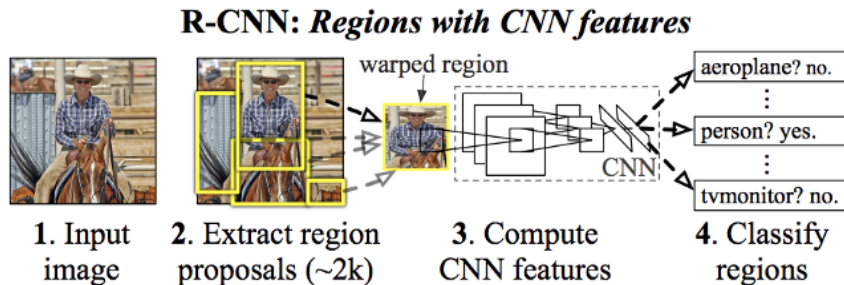


Figure: Schematic description of two-stage detectors.

Regions with CNN features (R-CNN)³



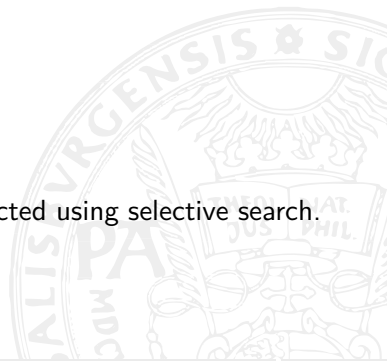
Main idea based on four modules:

- 1 Module: is responsible for selecting candidate regions of objects in the input image, independent from the category.
- 2 Module: extracts a fixed-length feature vector from each candidate region.
- 3 Module: classifies all objects in an image, including the BG class (no object).
- 4 Module: is a bounding-box regressor for precisely bounding-box prediction.

³Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.



- In the first stage, about 2000 region proposals are extracted using selective search.



Selective Search

Selective search starts with an oversegmented image.



Figure: Source: Chandel Vaibhaw Singh. *Object Detection vs. Object Recognition*.
<https://learnopencv.com/selective-search-for-object-detection-cpp-python/>. [Online, Accessed 05.10.2023]

Bounding boxes corresponding to segmented parts are generated and added to the list of regional proposals.

Selective Search

Iteratively, adjacent regions are combined based on their similarity (color, shape, texture, and size).

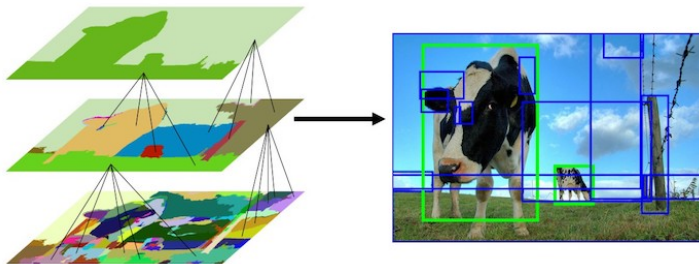
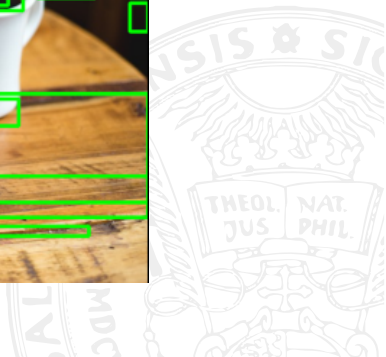
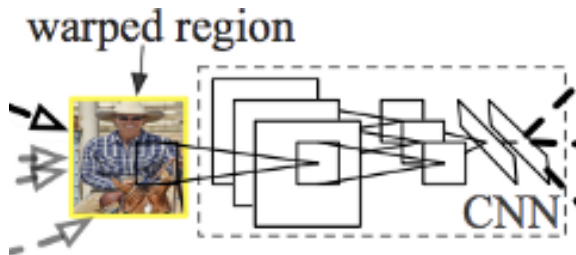


Figure: Source: Chandel Vaibhaw Singh. *Object Detection vs. Object Recognition*.
<https://learnopencv.com/selective-search-for-object-detection-cpp-python/>. [Online, Accessed 05.10.2023]

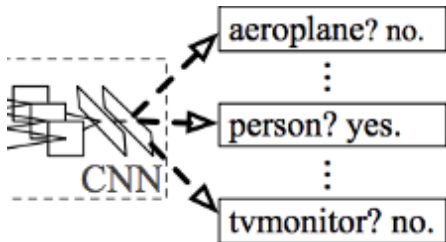
Coding Example 03 - Selective Search

Coding Example: Apply selective search to query image and visualize the top 80 boxes.

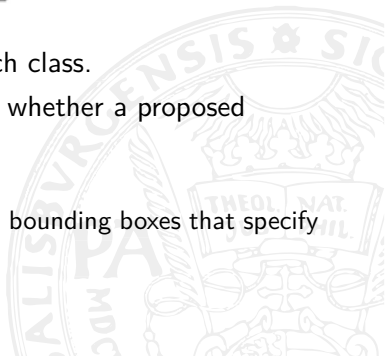




- The objects in an image are usually of different sizes and aspect ratios, so the region proposals extracted by the selective search vary in size as well.
- Regardless of the size or aspect ratio of the candidate region, all pixels are warped and resized such that they fit in a bounding box of fixed size 227×227 pixel.
- The resulting fixed size image patches are propagated through the backbone CNN for feature extraction.



- As classification head, a separate SVM is trained for each class.
- Intersection over Union (IoU) is calculated to determine whether a proposed region belongs to a specific class.
- To calculate the IoU, two bounding boxes are needed:
 - 1 The ground-truth bounding boxes (i.e., the hand labeled bounding boxes that specify where in the image the objects are located)).
 - 2 The predicted bounding boxes.



Intersection over Union

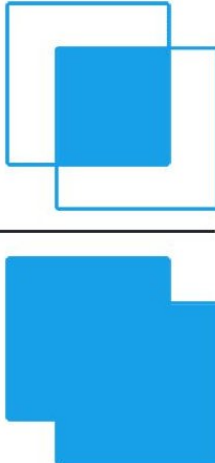
$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$


Figure: Source: Adrian Rosebrock. *Intersection over Union (IoU) for object detection.*

<https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/>. [Online, Accessed 05.10.2023]

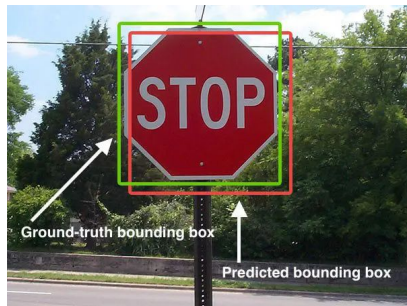
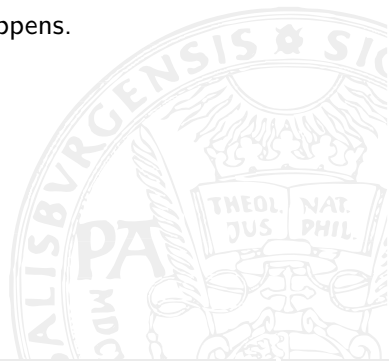


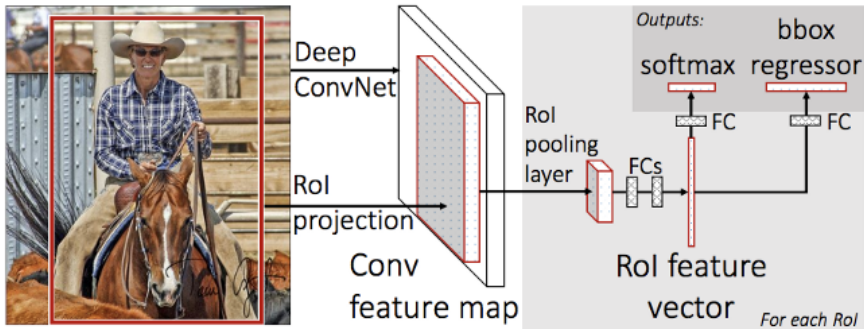
Figure: Source: Adrian Rosebrock. *Intersection over Union (IoU) for object detection.*

<https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/>. [Online, Accessed 05.10.2023]

- While fine-tuning R-CNN a $IoU > 0.5$ allows to decide that detected object belong to a certain class.
- If a region results into $0.3 < IoU < 0.5$ compared to a ground truth area, the region is considered as partially overlapping. For such cases a mixed label of background and ground truth is assigned.

- Very slow, since about 2000 image patches per image have to be passed through the network.
- Selective search is a fixed method where no learning happens.





A faster version of R-CNN.

- As with R-CNN, region proposals are generated from the input image by selective search.
- In Fast R-CMM, the Regions of Interest (RoI) are extracted from the feature map.
→ The entire image is passed through the CNN once, i.e., avoiding feeding ≈ 2000 image patches into the network individually.

⁴Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

Region of Interest (RoI) Extraction

The feature map from which the RoIs are extracted is much smaller than the input image (e.g., $512 \times 512 \times 3 \rightarrow 16 \times 16 \times 512$).

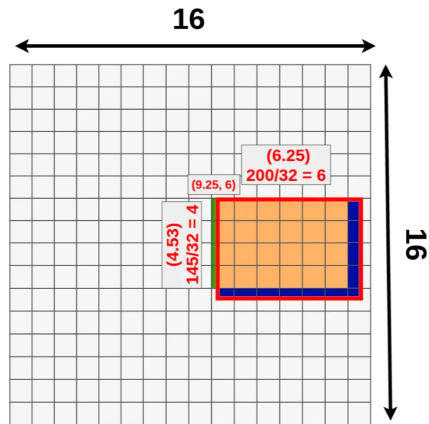


Figure: The red rectangle represents the downscaled RoI on the feature map. Since the pixel raster is discrete, quantization is performed: **information is lost** (blue area) and additional information is gained (green area). Source: Kemal Erdem. *Understanding Region of Interest — (RoI Pooling)*.

<https://towardsdatascience.com/understanding-region-of-interest-part-1-roi-pooling-e4f5dd65bb44>. [Online, Accessed 05.10.2023]

Remember that RoIs vary in size and aspect ratio.

- In R-CNN, the RoIs are cropped out of the image and wrapped in a 227×227 image patch to ensure a fixed size input.
- To ensure a fixed size input (i.e., $3 \times 3 \times 512$) to the FC layer, RoI pooling is introduced in Fast R-CNN.

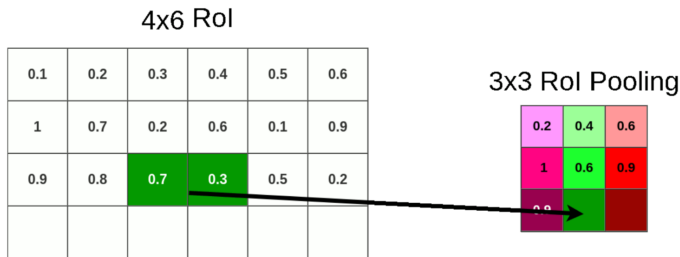


Figure: Source: Kemal Erdem. *Understanding Region of Interest — (RoI Pooling)*.

<https://towardsdatascience.com/understanding-region-of-interest-part-1-roi-pooling-e4f5dd65bb44>. [Online, Accessed 05.10.2023]

Since 4 is not a multiple of 3, information is lost again!

Fast R-CNN: Multi-Task Loss

Replace the multi-stage training protocol of R-CNN (SVM classification and bounding box regression).

A specialized *multi-task loss function*, which combines a softmax classifier and bounding-box regressors is used.

$$L(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geq 1]L_{loc}(t^u, v),$$

where

u ... ground-truth class

v ... ground-truth bounding box

$L_{cls}(p, u) = -\log p_u$ is the log loss for the true class u

$L_{loc}(t^u, v)$ is the loss of a bounding box for a true class u defined by a predicted tuple of offset parameters $t^u = (t_x^u, t_y^u, t_w^u, t_h^u)$

$\lambda[u \geq 1]$ called Iverson bracket indicator, which is 0 in case a background region is found. In this case no bounding box is needed.

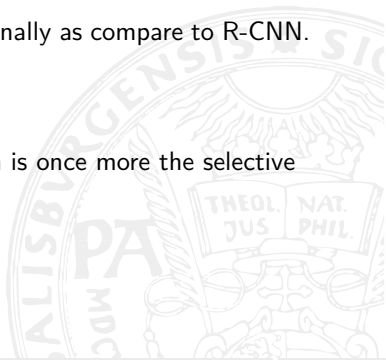
Advantages and Disadvantages of Fast R-CNN

■ Advantages:

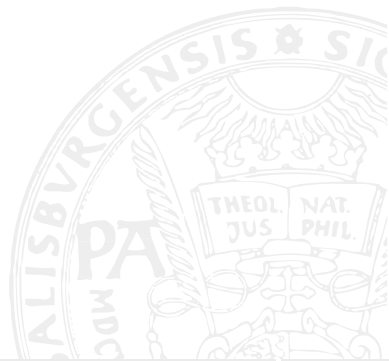
- 1 Drastically improves the training (8.75 hrs vs 84 hrs) and detection time from R-CNN.
- 2 Extraction of the features is not longer done separately for each region proposal, instead once for the entire image.
- 3 Also improves the Mean Average Precision (mAP) marginally as compare to R-CNN.

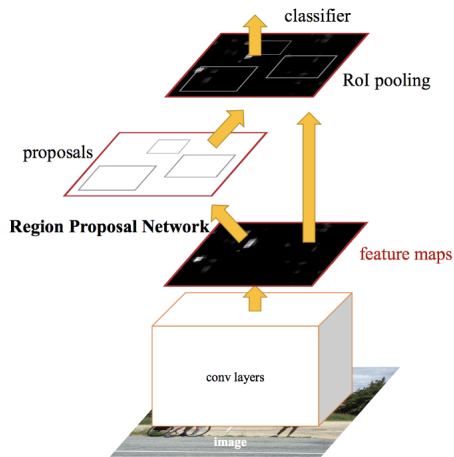
■ Disadvantages:

- Most of the time taken by Fast R-CNN during detection is once more the selective search region proposal generation algorithm.



.... time to get rid of the selective search bottleneck





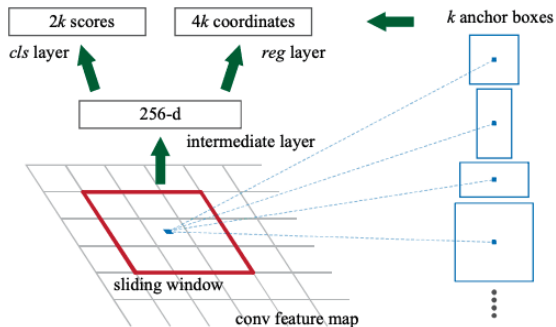
A separate network is used to predict the region proposals:

Region Proposal Network (RPN)

⁵Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015), pp. 91–99.

Region Proposal Network

At each position of a sliding window (e.g., each point in the feature map) over the convolutional feature map, a region is proposed for each of k anchors.



For each sliding window position, two informations are returned for all k anchors:

- 1 $2 \times k$ probabilities that an object/background is classified.
- 2 $4 \times k$ coordinate tuples defining the bounding boxes.

Region Proposal Network - Anchor Boxes

The common configuration is to have $k = 9$ pre-defined anchors.

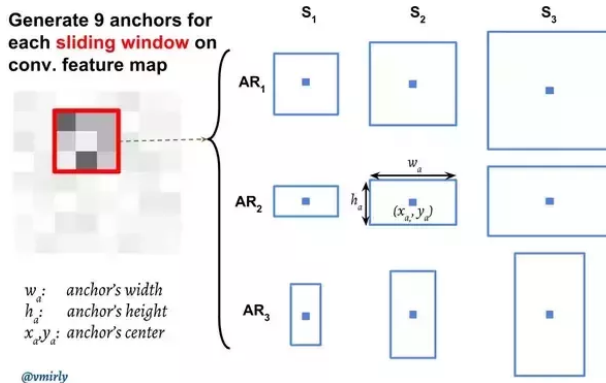
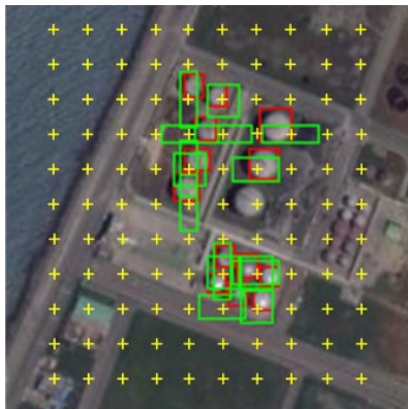


Figure: Source: Vidhya Sagar, Sailesh Jain, and VIDHYASAGAR BS. "Yield Estimation using faster R-CNN". In: (May 2018)

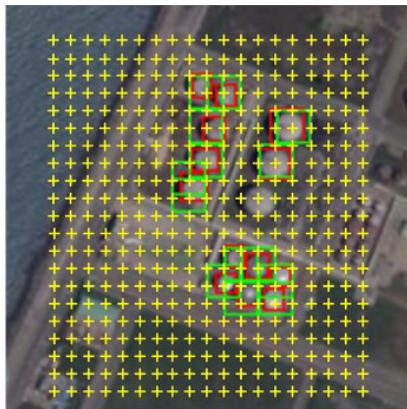
Anchor boxes are reference boxes with different shapes and scales that parameterize the k proposed regions at each point of the feature map.

Region Proposal Network

The corresponding anchor points at image level, 16 and 8 are strides used depending on the backbone network.



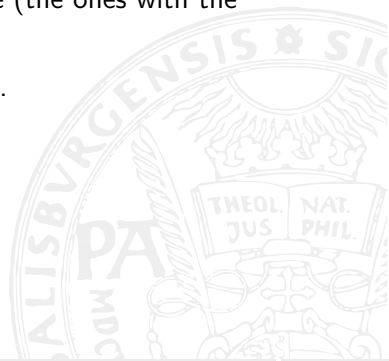
(a) $S_A = 16$

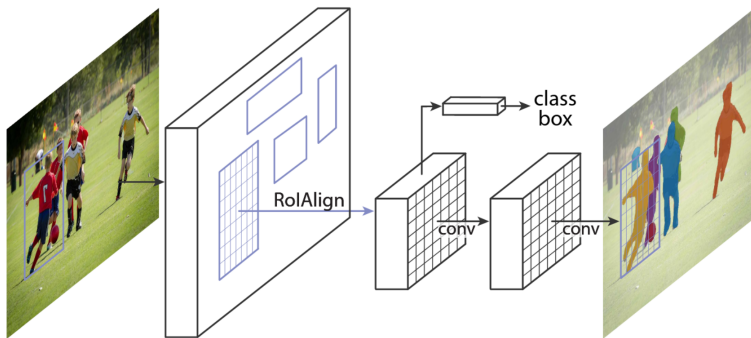


(b) $S_A = 8$

Figure: Source: Jianguo Yan et al. "IoU-Adaptive Deformable R-CNN: Make Full Use of IoU for Multi-Class Object Detection in Remote Sensing Imagery". In: *Remote Sensing* 11.3 (2019). ISSN: 2072-4292. DOI: 10.3390/rs11030286

- IoU is used as loss function to train the RPN.
- In total, $k \times$ (number of points in the feature map) regions are proposed.
 - The proposed regions are filtered by confidence score (the ones with the highest confidence score are kept).
- RoI polling is performed on the filtered region proposals.





- Based on Faster R-CNN and adding an additional branch to output the object mask.
- Introduce RoIAlign to fix the information loss problem of RoIPooling.

⁶Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

Solves the problem of information loss through RoI pooling.

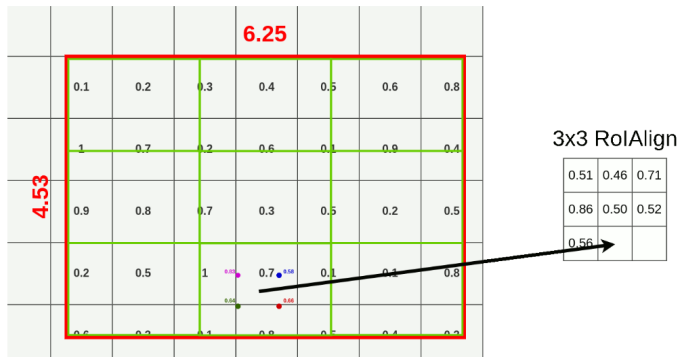
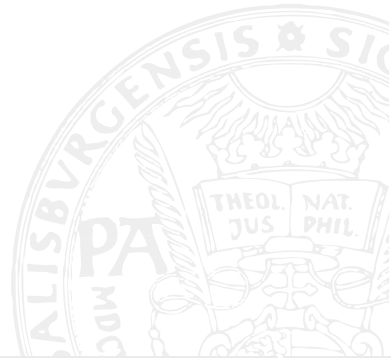


Figure: Source: Kemal Erdem. *Understanding Region of Interest — (RoI Align and RoI Warp)*.

<https://towardsdatascience.com/understanding-region-of-interest-part-2-roi-align-and-roi-warp-f795196fc193>. [Online, Accessed 05.10.2023]

Dividing the original RoI into 9 equal-sized boxes and applying bilinear interpolation in each of these boxes.

One-Stage Object Detectors



One-Stage Detectors

Bounding boxes are predicted directly from the input images.

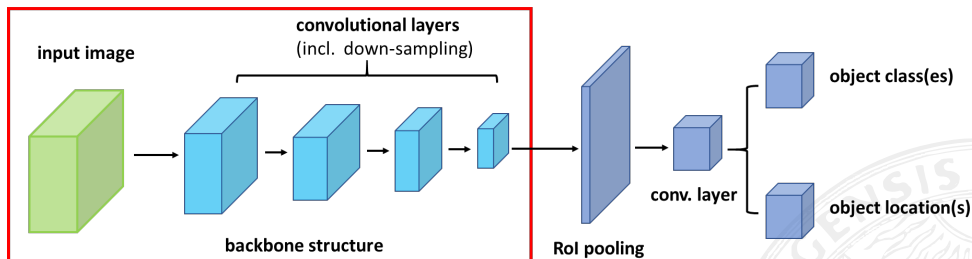


Figure: Schematic description of one-stage detectors. Source: Licheng Jiao et al. "A survey of deep learning-based object detection". In: *IEEE access* 7 (2019), pp. 128837–128868

Single Shot Detector (SSD)⁷

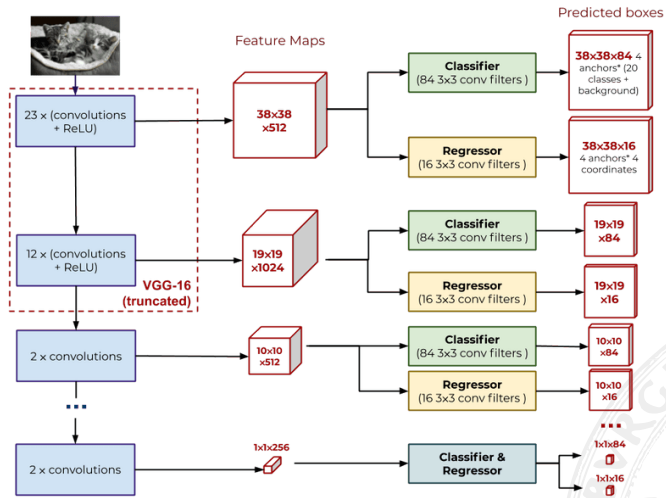


Figure: Source: Bastien Ponchon. *Face Detectors: Understand DSFD and the State-of-the-art Algorithms*.

<https://www.sicara.fr/blog-technique/2019-09-26-face-detectors-dsfd-state-of-the-art-algorithms>. [Online, Accessed 05.10.2023]

⁷Wei Liu et al. "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

Single Shot Detector (SSD) - Keyfacts

- Bounding boxes are directly predicted from the feature map, which results in short prediction times
- Classifiers and regressors are run on several feature maps at different scales of the core network.
 - The feature maps correspond to the different sizes of the receptive field, i.e., the deeper the feature map, the wider the receptive field (larger objects).
- A smaller number of anchors is required just to account for the different possible shapes, since detection is already performed at different scales.
- The core network is composed of VGG-16 network layers
- Finally, the number of positive bounding boxes (not classified as background) is reduced by non-maximum suppression based on the confidence of the classifiers.

Non-Maximum Suppression

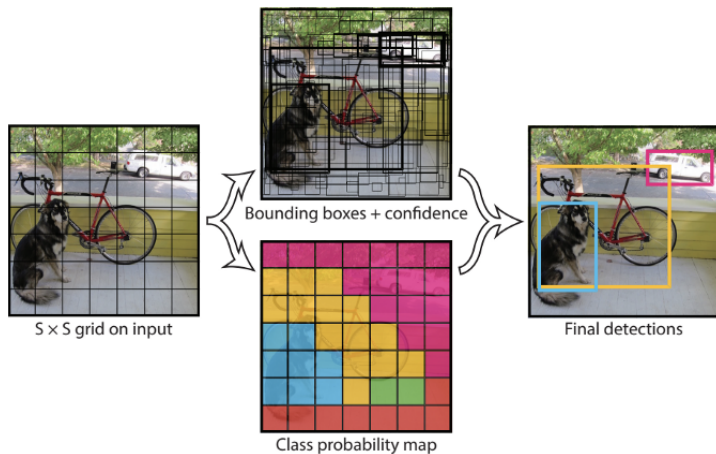
A technique to select one entity (e.g., bounding boxes) out of many overlapping entities.



Figure: Source: Jatin Prakash. *Non Maximum Suppression: Theory and Implementation in PyTorch*.
<https://learnopencv.com/non-maximum-suppression-theory-and-implementation-in-pytorch/>. [Online, Accessed 05.10.2023]

Select the predictions with the **maximum confidence** and **suppress** all the other predictions having overlap (IoU) with the selected predictions greater than a threshold.

You Only Look Once (YOLO)⁸



"We reframe object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities"

⁸Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

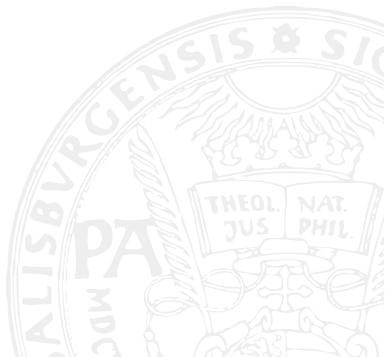
YOLO - Object Detection Steps

The following steps are performed by YOLO:

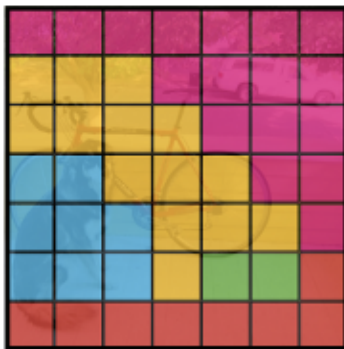
- Divide the input image into an $S \times S$ cell grid.
→ A grid cell is responsible to detect the object whose centre falls into.
- Each grid cell predicts B bounding boxes (x, y, w, h) and a confidence score for each box.
→ The confidence is obtained by a combination (multiplication) of the probability that the box contains an object and the IoU (how accurate the box contains that object), i.e., $P(Object) * IoU_{pred}^{truth}$.



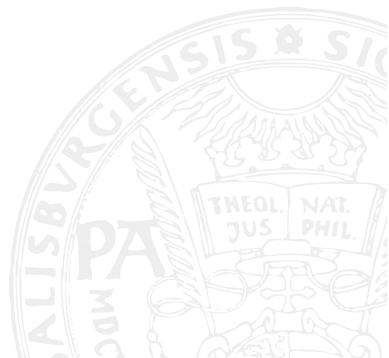
Bounding boxes + confidence



- Each grid also predicts conditional class probabilities, i.e., $P(Class_i|Object)$.

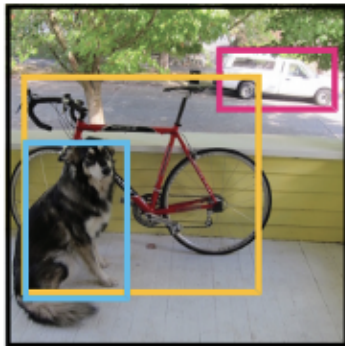


Class probability map

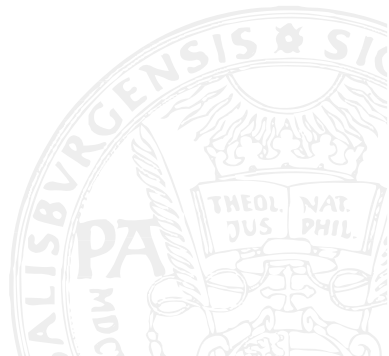


YOLO - Object Detection Steps

- The conditional class probabilities and the box confidence scores are multiplied, i.e., $P(Class_i|Object) * P(Object) * IoU_{pred}^{truth} = P(Class_i) * IoU_{pred}^{truth}$.
→ Class-specific confidence scores for each box.
- NMS is performed to obtain the final detections.



Final detections

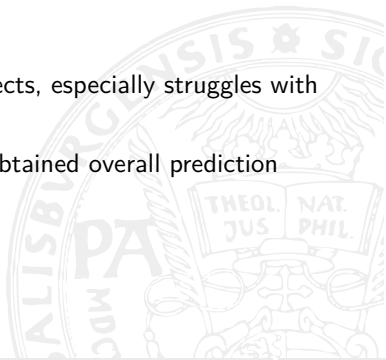


■ Advantages:

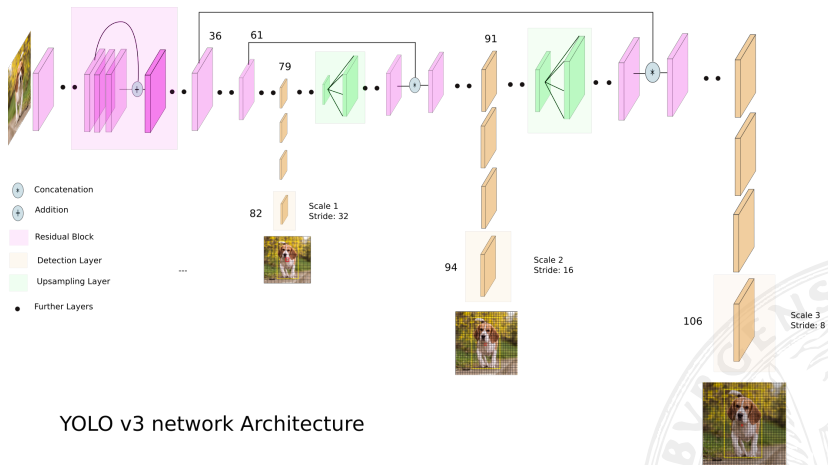
- 1 Very fast compared to all two-stage detectors.
- 2 Fast R-CNN makes many background false positives mistakes while YOLO achieves 3 times less of those.

■ Disadvantages:

- Performs not really good at accurate localization of objects, especially struggles with small objects.
- The localization error was the main component of the obtained overall prediction error.



Perform detection at three different scales.



YOLO v3 network Architecture

Figure: Source: Ayoosh Kathuria. *What's new in YOLO v3?*

<https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b>. [Online, Accessed 05.10.2023]

⁹Joseph Redmon and Ali Farhadi. "Yolov3: An incremental improvement". In: *arXiv preprint arXiv:1804.02767* (2018).

- Uses a different backbone (deeper) network, that is more accurate, however, with trade off in speed.
- Detection at different layers helps address the issue of detecting small objects.
- In total uses 9 anchor boxes, three for each scale.



Feature Pyramid Network (FPN)¹⁰

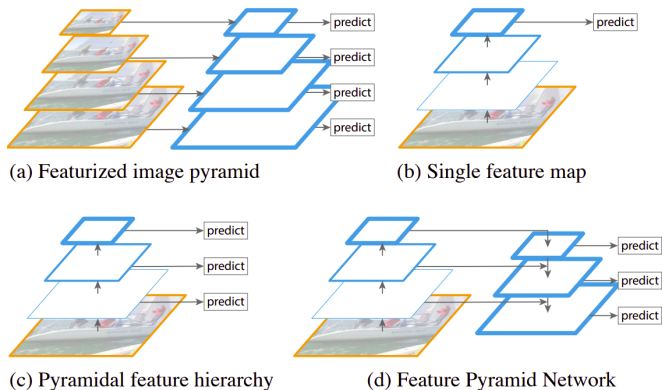
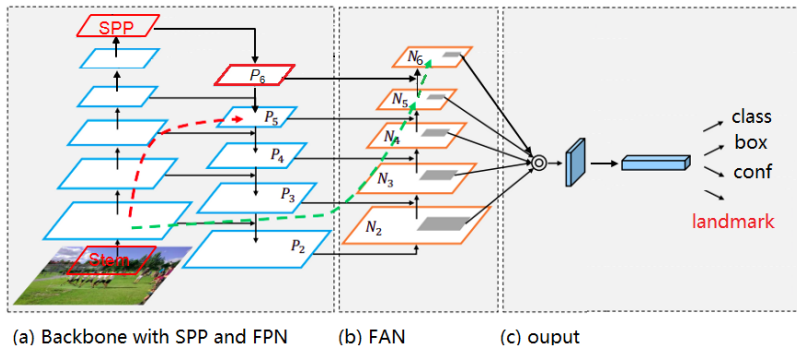


Figure: In this figure, feature maps are indicated by blue outlines and thicker outlines denote semantically stronger features.

¹⁰Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

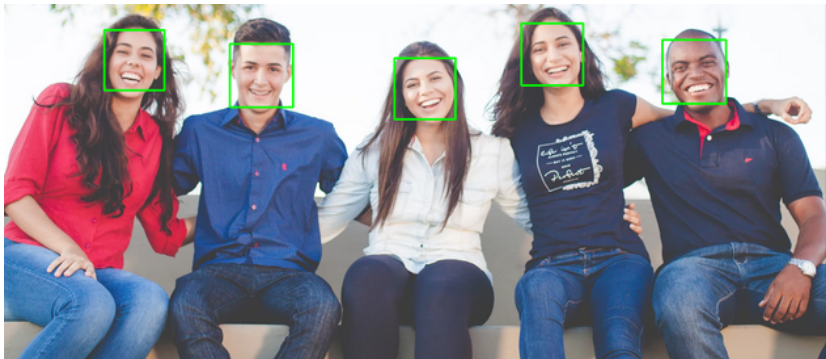


- Add landmark regression head.
- Additional spatial pyramid pooling (SPP) block to increase the receptive field.
- Additional output block for prediction so called P_6 .

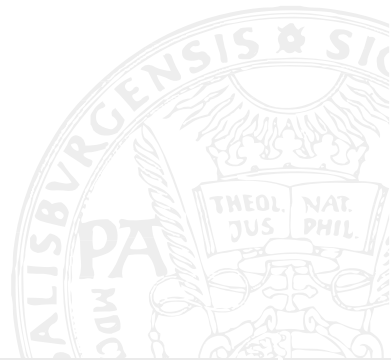
¹¹Delong Qi et al. "YOLO5Face: Why Reinventing a Face Detector". In: *arXiv preprint arXiv:2105.12931* (2021).

Coding Example 04 - YOLOv5 Face

Coding Example: Apply the YOLOv5 Face detector to detect all the faces in the sample images.



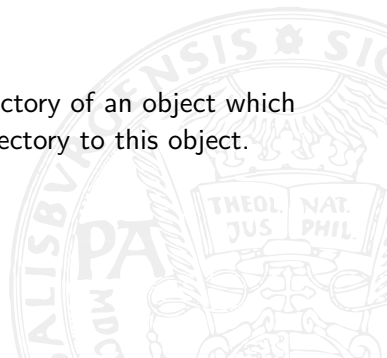
Object Tracking



Object tracking follows detection and is:

- a popular and widely applicable problem in computer vision,
- applied in many fields (autonomous driving, pedestrian tracking, surveillance, ...),
- applied for different use cases and benchmarks.

The aim of object tracking is to localize/describe the trajectory of an object which may be moved to a different position and link this trajectory to this object.



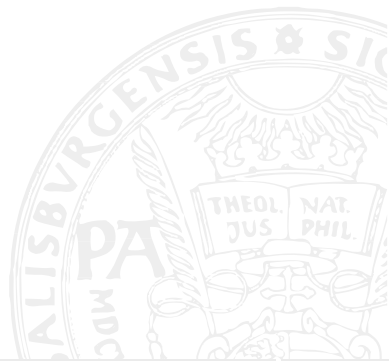
Object Tracking - Challenges

- Occlusions, an object can be occluded across multiple frames.
- Variations due to geometric changes in pose, articulation, scale changes of objects due movement.
- Differences due to photo-metric factors like illumination and appearance.
- Non-linear motion like sudden change in movements, or non-linear blur caused by sharp change in camera quality.
- Limited resolution of a video which is captured by a low-end mobile phone.
- Similar objects in a scene, e.g. people of similar body measures wearing the same coloured clothes and accessories (surveillance).
- Highly crowded scenarios like shopping malls, concert halls, sport stadiums and many more.
- ...

Object tracking: Main categories

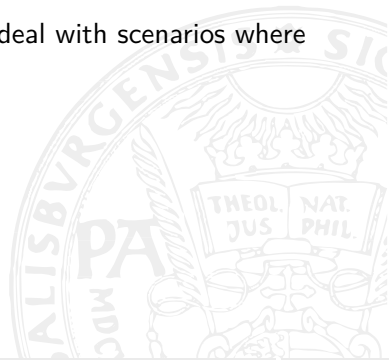
The most common tracking categories are:

- Single Object Tracking (SOT)
- Video Object Segmentation (VOS)
- Multiple Object Tracking (MOT)
- Multiple Object Tracking and Segmentation (MOTS)



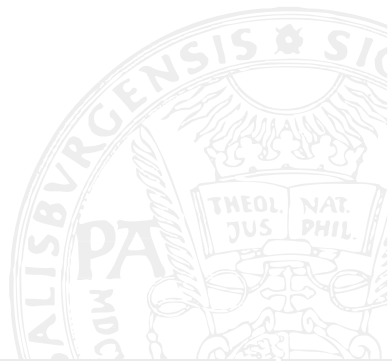
Single Object Tracking (SOT)

- SOT implies that one singular object is tracked, even in environments involving other objects.
- The object of interest is determined in the first frame, which is where the object to be tracked is initialized for the first time.
- Is a type of detection-free tracking category and thus, requires manual initialization of a fixed number of objects in the first frame. These objects are then localized in consequent frames.
- A drawback of detection-free tracking is that it cannot deal with scenarios where new objects appear in the middle frames.



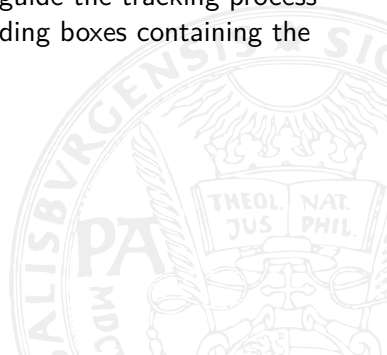
Video Object Segmentation (VOS)

- Tracking in terms of VOS means estimating the location of an arbitrary user-annotated target object throughout a video, where the location of the object is represented by a pixel-wise mask.
- Thus, this task shares several similarities with SOT, where the tracked object is represented by a bounding box.



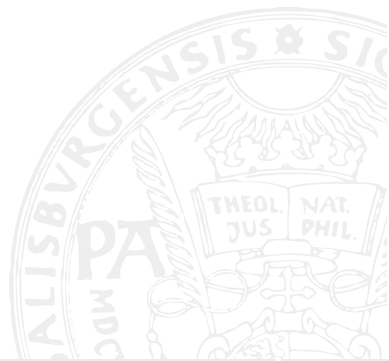
Multiple Object Tracking (MOT)

- Most MOT algorithms incorporate an approach called *tracking-by-detection*.
- This method involves an independent detector that is applied to all image frames to obtain likely detections, and then a tracker, which is run on the set of detections.
- The tracker attempts to perform data association (for example, linking the detections to obtain complete trajectories).
- The detections extracted from video inputs are used to guide the tracking process by connecting them and assigning identical IDs to bounding boxes containing the same target.



Multiple Object Tracking and Segmentation (MOTS)

- Compared to single object tracking tasks in MOT, MOTS and pose estimation and tracking tasks a detection step is necessary to identify the targets leaving or entering the scene.
- The main difficulty in tracking multiple objects simultaneously stems from the various occlusions and interactions between the single objects occurring (sometimes at the same time) in the video scene.

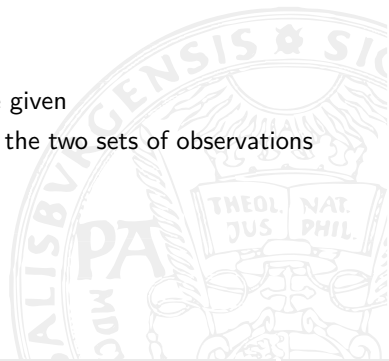


■ Propagation

- localise a target object in the current video frame given its location in the previous one
- in SOT and VOS this is the core problem

■ Association

- target states in both previous and current frames are given
- the goal is to determine the correspondence between the two sets of observations
- in MOT and MOTS this is the core problem



Prediction of the location of an arbitrary object in a video = Propagation

- The information of previous frames serve as input for the prediction
- The output is either a bounding box or a pixel-wise mask

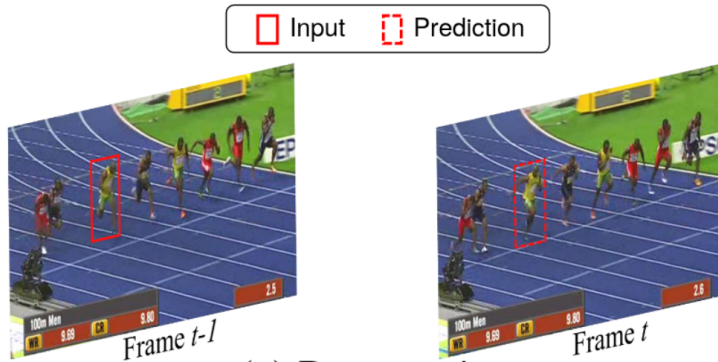


Figure: Source: Zhongdao Wang et al. "Do Different Tracking Tasks Require Different Appearance Models?" In: *Advances in Neural Information Processing Systems* 34 (2021)

Association of (multiple) objects across video frames

- Describes the problem of identity association and forming trajectories
- As input act information of object states from other frames

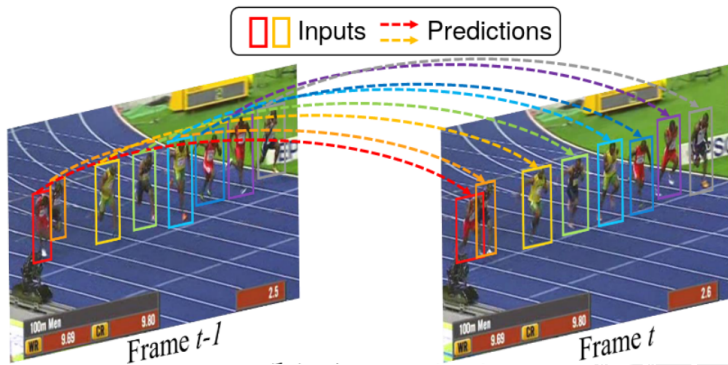
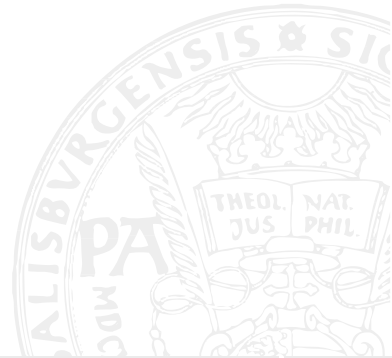


Figure: Source: Zhongdao Wang et al. "Do Different Tracking Tasks Require Different Appearance Models?" In: *Advances in Neural Information Processing Systems* 34 (2021)

Multiple Object Tracking



Generic Tracking Pipeline

Generic Tracking Pipeline:

- handle occlusions: buffer and recover, or terminate track
- initialize new tracking hypothesis (initialization phase)
- ...

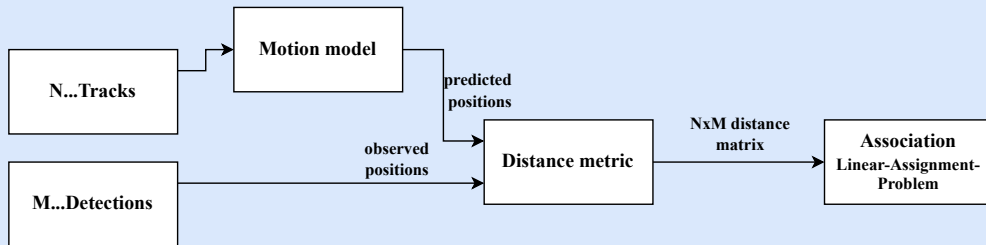


Figure: Generic tracking pipeline, association based on motion prediction.

Generic Tracking Pipeline

Generic Tracking Pipeline:

- handle occlusions: buffer and recover, or terminate track
- initialize new tracking hypothesis (initialization phase)
- ...

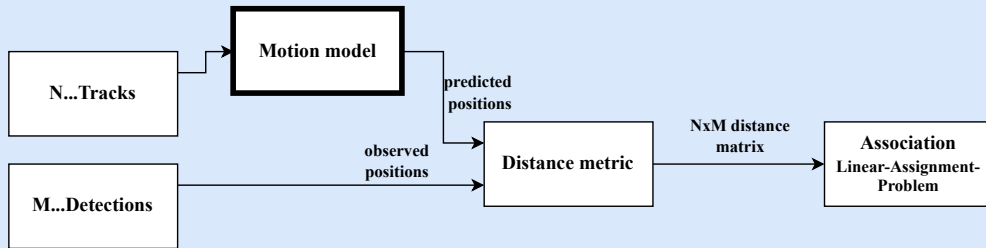
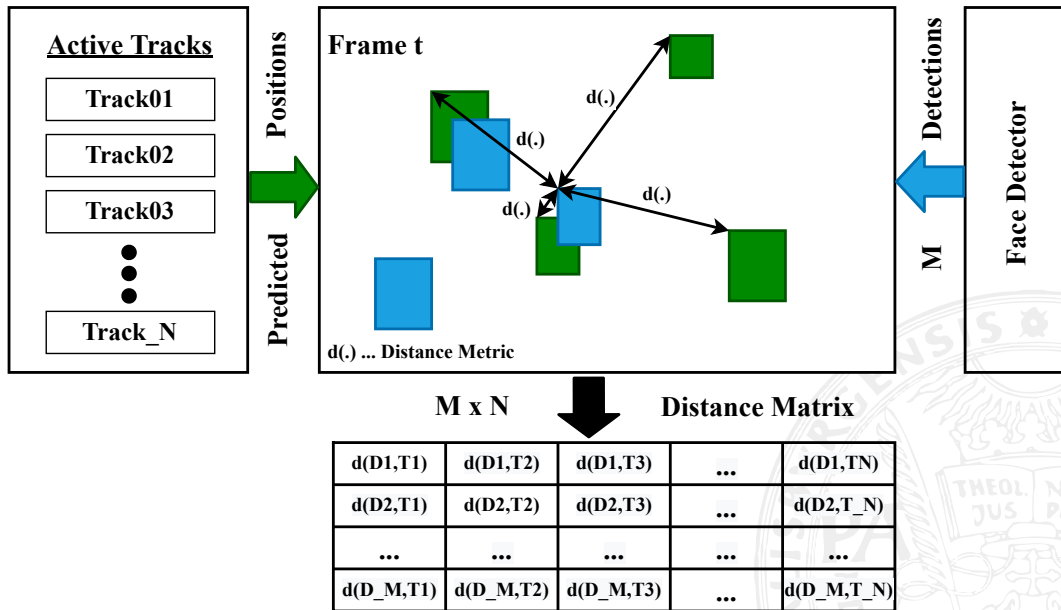


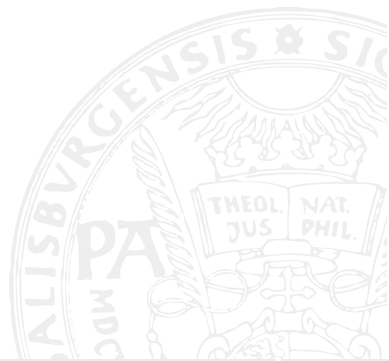
Figure: Generic tracking pipeline, association based on motion prediction.

Motion based Association



Coding Example 05 - Simple Multiple Object Tracker

Coding Example: Build a simple multiple object tracker based on motion prediction.



- Estimates the states of a (linear) dynamic system.
- The estimation is based on:
 - the prediction of the current state (based on prior knowledge) and
 - the measurement of the current state.
- Both (prediction and measurement) are not perfect (assumption, they follow a Gaussian distribution).
- The state can optimally be estimated by the Bayesian inference:

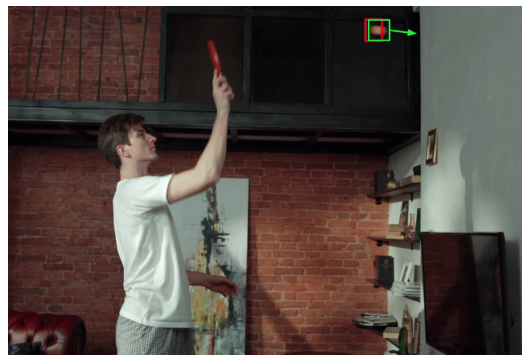
$$P(x|m) = \frac{P(m|x) \cdot P(x)}{P(m)} \quad (1)$$

¹²R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1 (Mar. 1960), pp. 35–45.
ISSN: 0021-9223. DOI: 10.1115/1.3662552.

Motion Prediction: Kalman Filter Example



result object detector (red bbox)



estimated bbox (green)

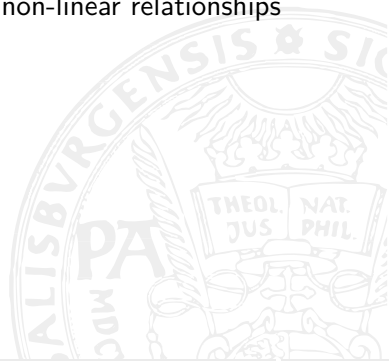
Figure: Example Kalman filter state estimation (green bbox), the estimation is a combination of the current measurement (red bbox, returned from the object detector) and the prediction. The green arrow indicates the direction and magnitude used to create the next prediction.

Source: Daniel Tomer. *What I Was Missing While Using The Kalman Filter For Object Tracking*.

<https://towardsdatascience.com/what-i-was-missing-while-using-the-kalman-filter-for-object-tracking-8e4c29f6b795>. [Online, Accessed 05.10.2023]

Motion Prediction: Kalman Filter Problem

- Kalman filter fails if there are non-linear relationships between the hidden state variable and the observed variable (e.g., fails with non-linear motion during occlusion).
- Variations of the traditional Kalman filter for capturing non-linear relationships are:
 - Extended Kalman Filter,
 - Particle Filter.



Generic Tracking Pipeline

Generic Tracking Pipeline:

- handle occlusions: buffer and recover, or terminate track
- initialize new tracking hypothesis (initialization phase)
- ...

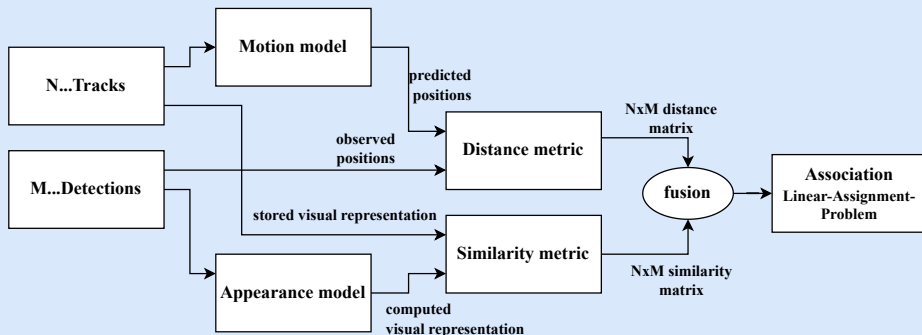


Figure: Generic tracking pipeline, association based on a combination of motion prediction and appearance similarity.

Generic Tracking Pipeline

Generic Tracking Pipeline:

- handle occlusions: buffer and recover, or terminate track
- initialize new tracking hypothesis (initialization phase)
- ...

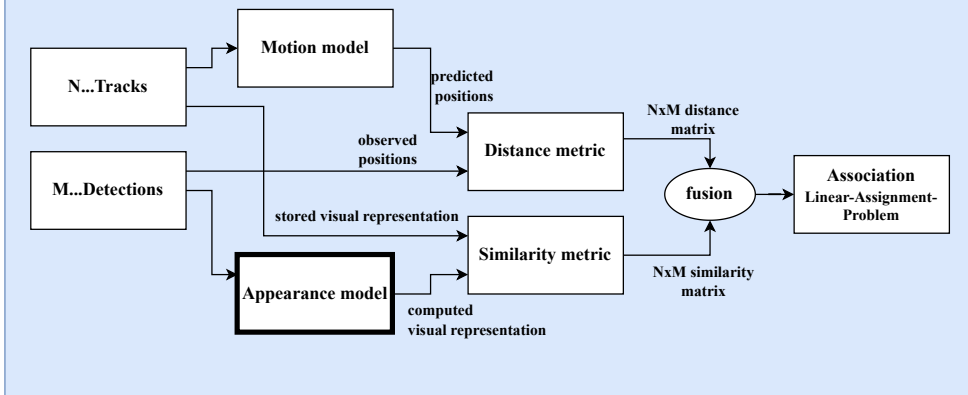
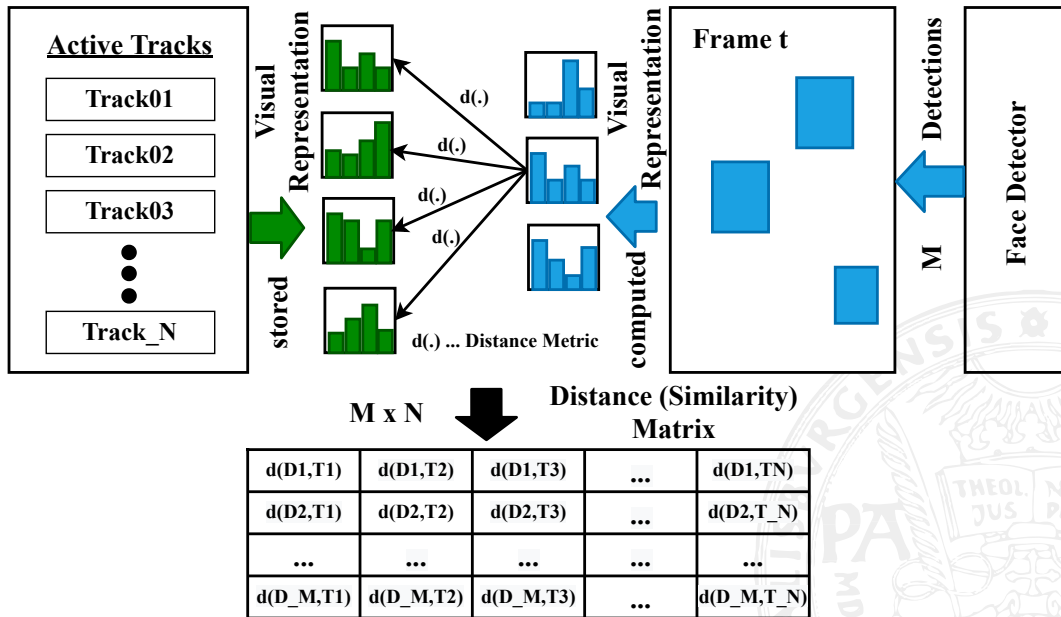


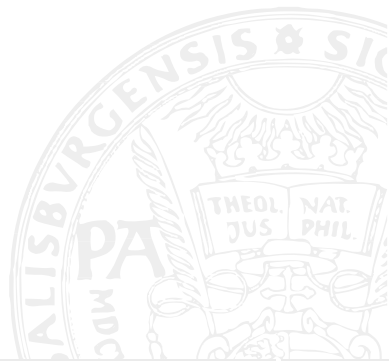
Figure: Generic tracking pipeline, association based on a combination of motion prediction and appearance similarity.

Appearance based Association

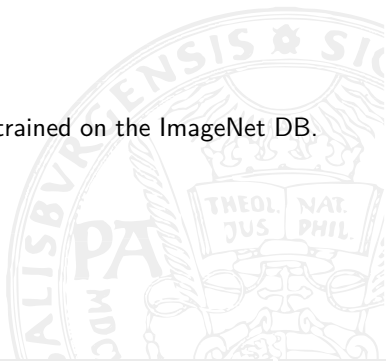


Coding Example 06 - Simple Multiple Object Tracker

Coding Example: Build a simple multiple object tracker based on a combination of motion prediction and appearance.



- Describes an object according to different types of features.
- Traditional handcrafted features:
 - Color histogram,
 - Gradient-based representation (e.g., HOG),
 - Key point-based representation (e.g., SURF, SIFT),
 - ...
- Features generated by a neuronal network.
 - A image classification net, like the ResNet50 or VGG16 trained on the ImageNet DB.



■ Featurebank:

- Let F_{T_j} be a set of features of the j^{th} track extracted from the last N matched detections (i.e., $|F_{T_j}| = N$). The distance to the i^{th} detected object D_i is computed as follows:

$$d(D_i, T_j) = \min_{f \in F_{T_j}} d(D_i, f). \quad (2)$$

■ Update Features:

- Exponential Moving Average (EMA)¹³: updates the feature state e_j^t for the j^{th} track at frame t as follows:

$$e_j^t = \alpha e_j^{t-1} + (1 - \alpha) f_j^t, \quad (3)$$

where f_j^t is the appearance embedding of the current matched detection and α is the momentum term.

¹³Zhongdao Wang et al. "Towards real-time multi-object tracking". In: *European Conference on Computer Vision*. Springer, 2020, pp. 107–122.

Generic Tracking Pipeline

Generic Tracking Pipeline:

- handle occlusions: buffer and recover, or terminate track
- initialize new tracking hypothesis (initialization phase)
- ...

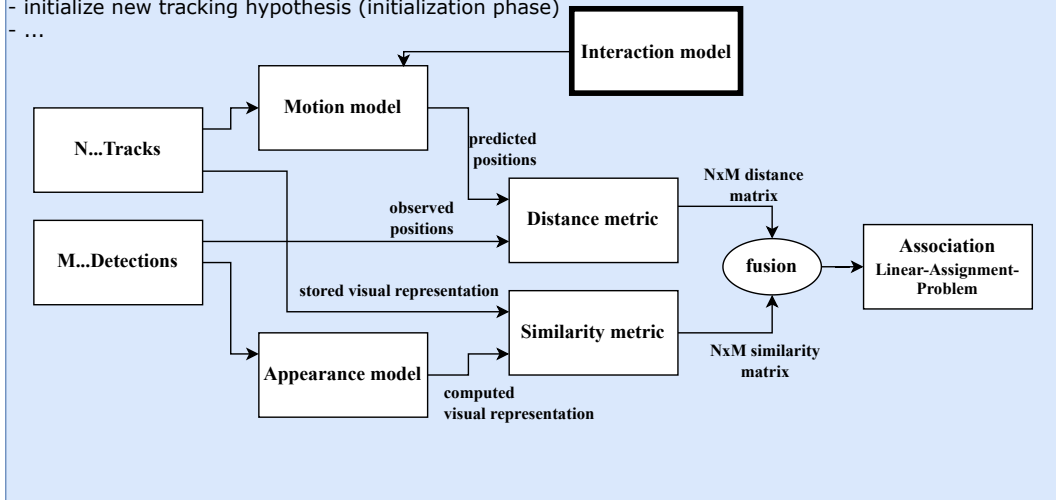


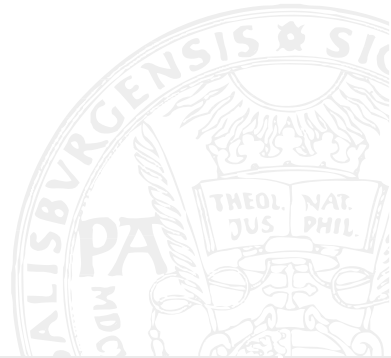
Figure: Generic tracking pipeline, include interaction model.

- Is also known as mutual motion model.
- It captures the influence of an object on other objects.
- *Example:* When people move in a queue, each of them follows other people and guides others at the same time.
- Two typical interaction models are the social forces models¹⁴ and the crowd motion pattern models¹⁵

¹⁴Dirk Helbing and Peter Molnar. "Social force model for pedestrian dynamics". In: *Physical review E* 51.5 (1995), p. 4282.

¹⁵Min Hu, Saad Ali, and Mubarak Shah. "Detecting global motion patterns in complex videos". In: *2008 19th International Conference on Pattern Recognition*. Ieee. 2008, pp. 1–5.

Multiple Object Tracker



Simple Online Realtime Tracking (SORT)¹⁶

Simple Online Realtime Tracking (SORT):

- any detection with an overlap less than IoU_{min} initiates a new track,
- tracks are terminated if they are not detected for T_{Lost} frames.

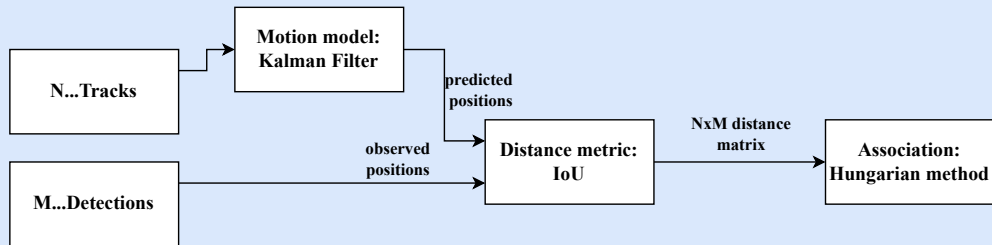


Figure: Overview Simple Online Realtime Tracking (SORT).

¹⁶Alex Bewley et al. "Simple online and realtime tracking". In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 3464–3468.

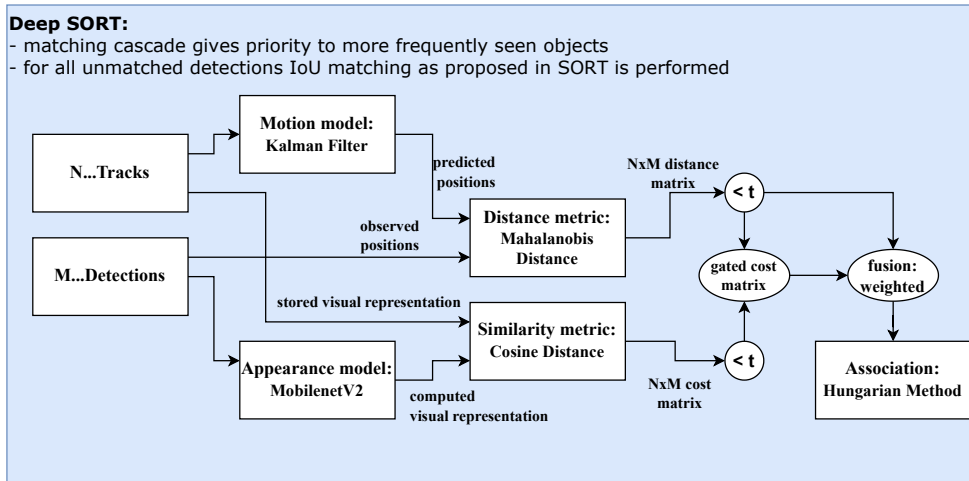


Figure: Overview Deep SORT, as described in the paper.

¹⁷Nicolai Wojke, Alex Bewley, and Dietrich Paulus. "Simple online and realtime tracking with a deep association metric". In: *2017 IEEE International Conference on Image Processing (ICIP)*. 2017, pp. 3645–3649.

Strong SORT:

- replace matching cascade with vanilla global linear assignment
- IoU matching (SORT) fallback routine

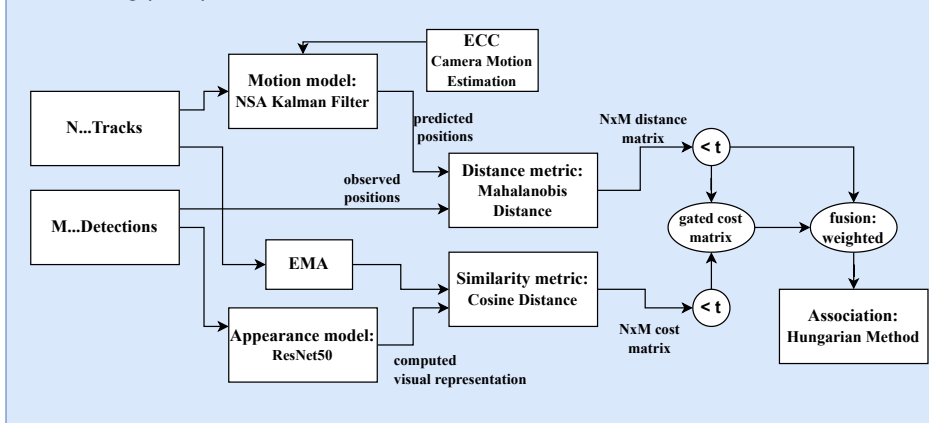


Figure: Overview Strong SORT, including a Noise Scale Adaptive (NSA) Kalman Filter and Exponential Moving Average (EMA) feature updating.

¹⁸Yunhao Du et al. "Strongsort: Make deepsort great again". In: *arXiv preprint arXiv:2202.13514* (2022).

Unitrack MOT part:

- for all unmatched detections IoU matching as proposed in SORT is performed

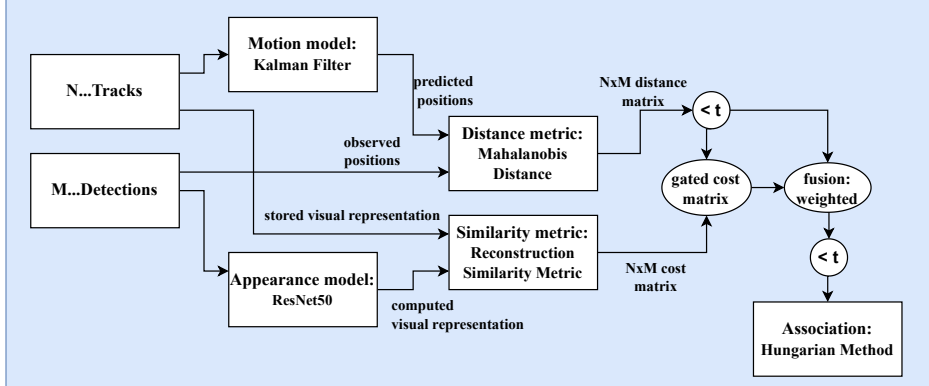
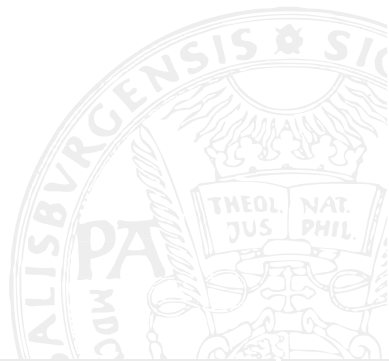


Figure: Overview MOT part of Unitrack.

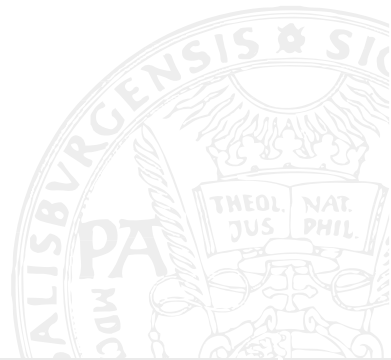
¹⁹Zhongdao Wang et al. "Do Different Tracking Tasks Require Different Appearance Models?" In: *Advances in Neural Information Processing Systems* 34 (2021).

Coding Example 07- Strong SORT

Coding Example: Apply the Strong SORT multiline object tracker.



Evaluation Metric HOTA



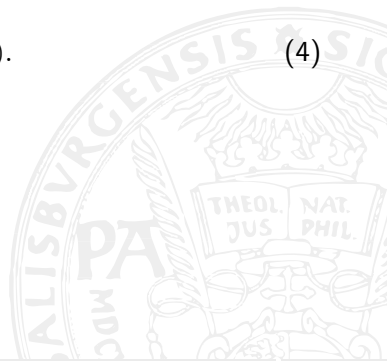
Higher Order Tracking Accuracy (HOTA)²⁰

- Explicitly balances the effect of performing accurate detection, association and localization.
- The HOTA score is a combination of three IoU scores:
 - Localization,
 - Detection,
 - Association.

²⁰Jonathon Luiten et al. "Hota: A higher order metric for evaluating multi-object tracking". In: *International journal of computer vision* 129.2 (2021), pp. 548–578.

- Spatial alignment (IoU) between one predicted and one ground-truth detection.
- The overall Localization Accuracy (LocA) is the average IoU (Loc-IoU) over all matched pairs (predicted and ground-truth) in the whole sequence:

$$LocA = \frac{1}{|TP|} \sum_{c \in TP} \text{Loc-IoU}(c). \quad (4)$$



- Alignment between the set of all predicted and the set of all ground-truth detections.
- Detection and ground-truth intersect (TP) if Loc-IoU $> \alpha$.
- Overlaps will be solved with the Hungarian algorithm to ensure a one-to-one matching.
- Overall Detection Accuracy (DetA):

$$DetA = \frac{|TP|}{|TP| + |FN| + |FP|}, \quad (5)$$

where the TPs, FNs and FPs are counted over the whole sequence.

- Measures how well a tracker links detections over time into the same identities.
- The Association IoU (Ass-IoU) measures the alignment over the whole track:

$$\text{Ass-IoU} = \frac{|TPA|}{|TPA| + |FNA| + |FPA|}. \quad (6)$$

- The overall Association Accuracy (AssA) represents the average Ass-IoU:

$$\text{AssA} = \frac{1}{|TP|} \sum_{c \in TP} \text{Ass-IoU}(c). \quad (7)$$

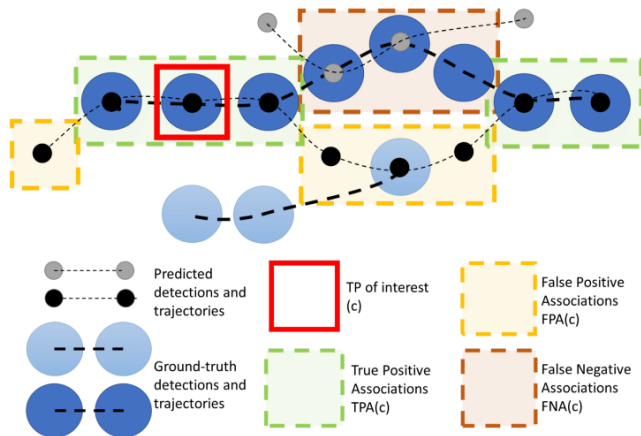


Figure: A visual example of the definitions TPA, FNA and FPA²¹.

²¹Luiten et al., "Hota: A higher order metric for evaluating multi-object tracking".

- All three components are combined in the HOTA metric:
 - $HOTA_\alpha$ represents the geometric mean of the detection and the association score:

$$HOTA_\alpha = \sqrt{DetA_\alpha \cdot AssA_\alpha} = \sqrt{\frac{\sum_{c \in TP_\alpha} AssIoU(c)}{|TP_\alpha| + |FN_\alpha| + |FP_\alpha|}} \quad (6)$$

- By integrating over the different α thresholds, the localization accuracy is integrated into the final score:

$$HOTA = \int_{0 < \alpha \leq 1} HOTA_\alpha \approx \frac{1}{19} \sum_{\alpha=0.05, \alpha+=0.05}^{0.95} HOTA_\alpha \quad (7)$$

Thank you for your attention!

