DRIVES

Development and Research on Innovative Vocational Education Skills

COMPUTER VISION FUNDAMENTALS

U2.E10. OBJECT DETECTION AND TRACKING, MOTION ESTIMATION, FACIAL RECOGNITION, SCENE UNDERSTANDING AND 3D RECONSTRUCTION

Computer Vision Expert

May 2021, Version 1

Co-funded by the Erasmus+ Programme of the European Union

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

LEARNING OBJECTIVES

The student is able to …

CVE.U2.E10.PC1	The student is able to define object detection and tracking as well as understand their differences.
CVE.U2.E10.PC2	The student understands the machine learning and deep learning approaches for object detection and is able to select the best approach according to a specific problem or situation.
CVE.U2.E10.PC3	The student understands the process of object tracking and knows the classification of traditional methods for tracking objects.
CVE.U2.E10.PC4	The student is able to define target representation and localization algorithms and list some examples.
CVE.U2.E10.PC5	The student is able to define filtering and data association algorithms and list some examples
CVE.U2.E10.PC6	The student understands the differences between target representation and localization algorithms and filtering and data association algorithms.

LEARNING OBJECTIVES

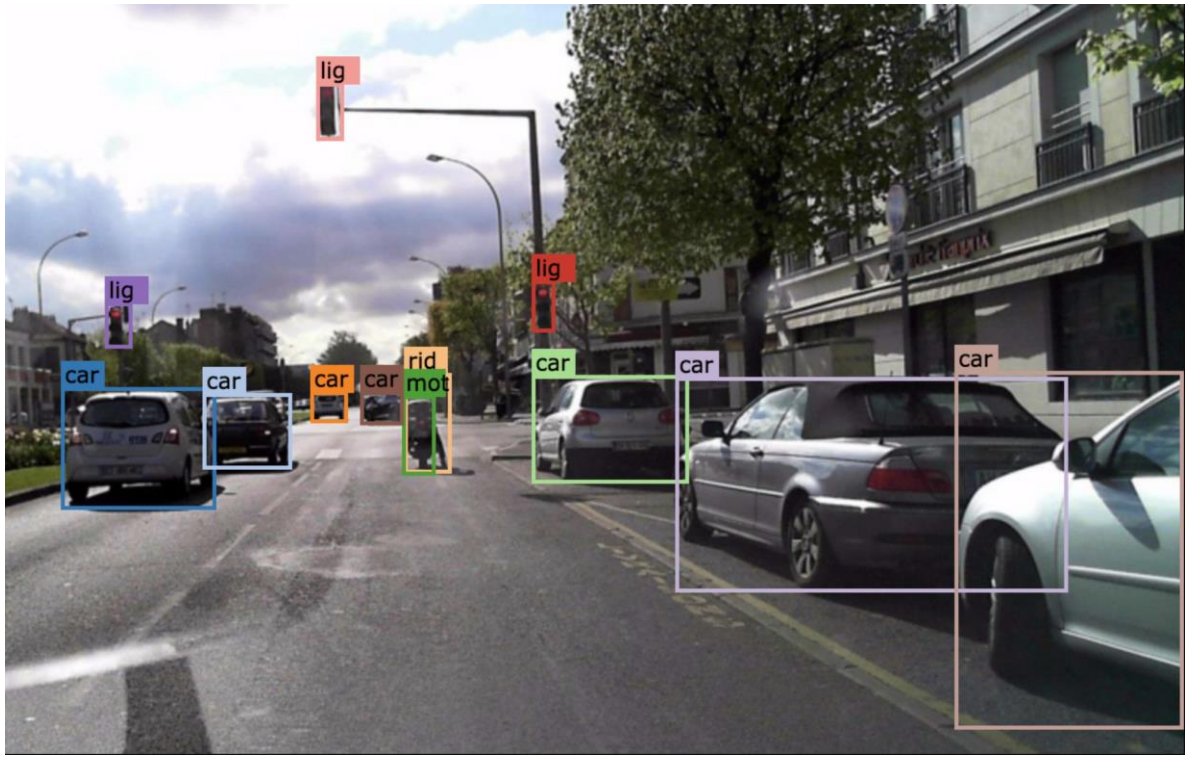
The student is able to …

CVE.U2.E10.PC7	The student is able to define and understand what facial recognition systems are as well as identify the applications of these systems.
CVE.U2.E10.PC8	The student knows the different techniques for face acquisition.
CVE.U2.E10.PC9	The student understands the motion estimation problem and knows the different methods and algorithms for finding motion vectors.
CVE.U2.E10.PC10	The student is able to define the need and purposes of scene understanding in typical computer vision problems.
CVE.U2.E10.PC11	The student understands the need, definition, and applications of 3D reconstruction.
CVE.U2.E10.PC12	The student knows the active and passive methods for 3D reconstruction and understands their differences.

OBJECT DETECTION

The focus of **object detection** is finding all objects of certain classes in an image. These detections are commonly represented with bounding boxes.

To identify a certain class of image and then detect and tabulate their appearance in an image or video, it is used **image classification**. For example detecting damages on an assembly line or identifying machinery that requires maintenance.



OBJECT DETECTION

A detection pipeline contain the following steps:

- Preprocessing, region of interest extraction (ROI);
- Object classication, and verification.

In the preprocessing step, tasks such as exposure and gain adjustment, as well as camera calibration and image rectification, are usually performed.

OBJECT TRACKING

Object tracking consists of follows or tracks an object once it is detected.

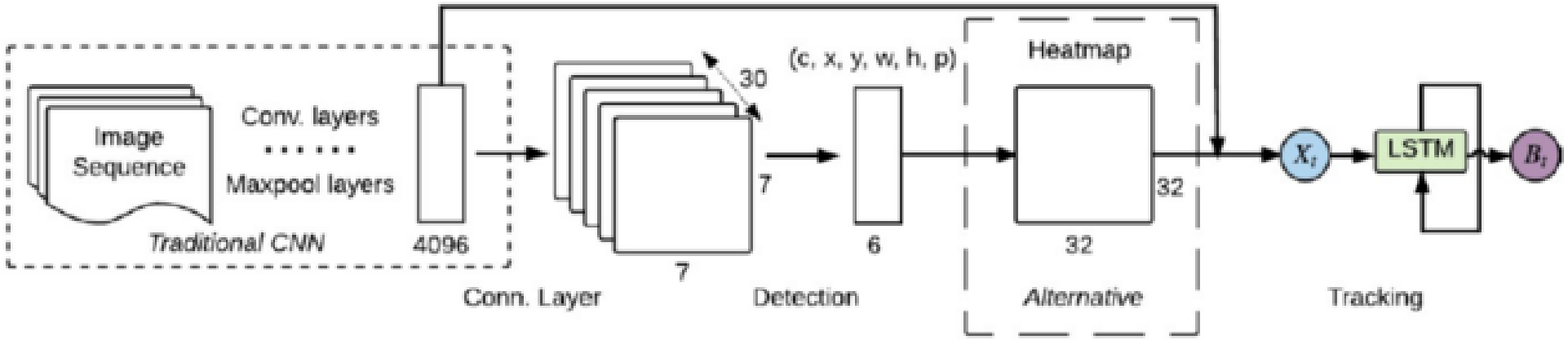
This task is often performed with images captured in sequence or real-time video feeds. For example, autonomous vehicles need to classify, track and detect objects such as pedestrians, other cars and road infrastructure to avoid collisions and obey traffic laws.

There are mainly two levels of object tracking:

- Single Object Tracking (SOT)
- Multiple Object Tracking (MOT)

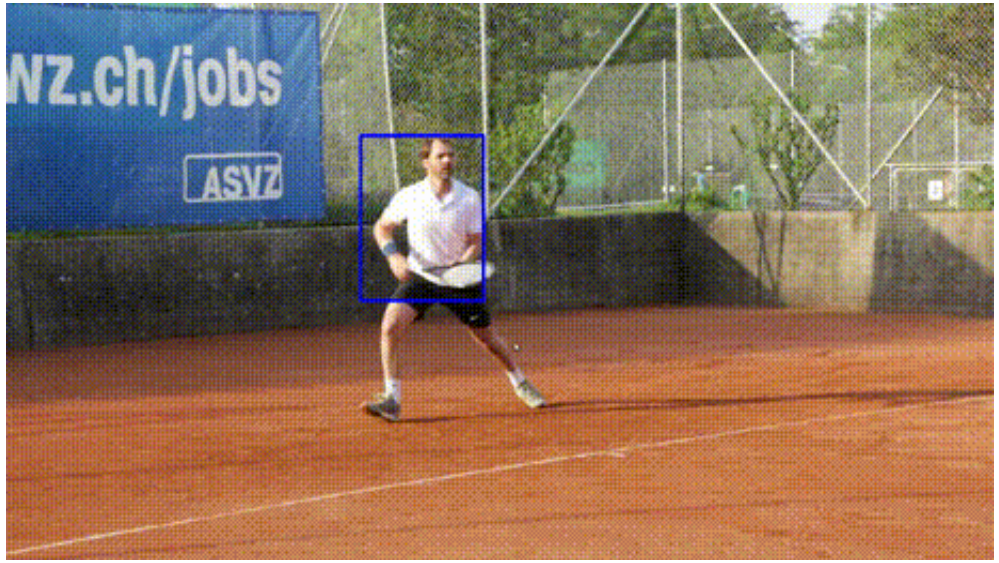
OBJECT TRACKING METHODS: **ROLO**

Recurrent YOLO (ROLO) is a single object tracking method that combines object detection and recurrent neural networks. It is a combination of YOLO and LSTM. The object detection module uses YOLO to collect visual features, along with location inference priors. At each time-step (frame), the LSTM receives an input feature vector of length 4096, and returns the location of the tracked object.



OBJECT TRACKING METHODS: **SIAMMASK**

SiamMask is a good choice when it comes to single object tracking, and it is based on the charming **siamese neural network**, which rose in popularity with Google's Facenet. This provides object segmentation masks, besides producing rotated bounding boxes at 55 frames per second. In order to track the desired object, SiamMask requires to be initialized with a single bounding box. Multiple object tracking (**MOT**) is not viable with SiamMask, and changing the model to support that will leave a significantly slower object detector.



OBJECT TRACKING METHODS: **DEEP SORT**

SORT is an algorithmic approach to object tracking. **Deep SORT** is improving SORT by replacing the associating metric with a novel **cosine metric learning**, a method for learning a feature space where the cosine similarity is successfully optimized through reparametrization of the softmax regime.

The track handling and Kalman filtering framework is almost identical to the original SORT, except the bounding boxes are computed using a pre-trained convolutional neural network, trained on a large-scale person re-identification dataset. It is a simple method to implement with a great starting point for multiple object detection, and offers solid accuracy, running in real-time.

OBJECT TRACKING METHODS: **TRACKR-CNN**

TrackR-CNN can be introduced as a baseline for the Multi Object Tracking and Segmentation (MOTS) challenge. The object detection module uses Mask R-CNN on top of a ResNet-101 backbone. The tracker is created by integrating 3D convolutions which are applied to the backbone features, incorporating temporal context of the video. Convolutional LSTM is considered as an alternative but the latter method does not yield any gains compared with the baseline.

TrackR-CNN also extends Mask R-CNN by an **association head**, to be capable to associate detections over time.

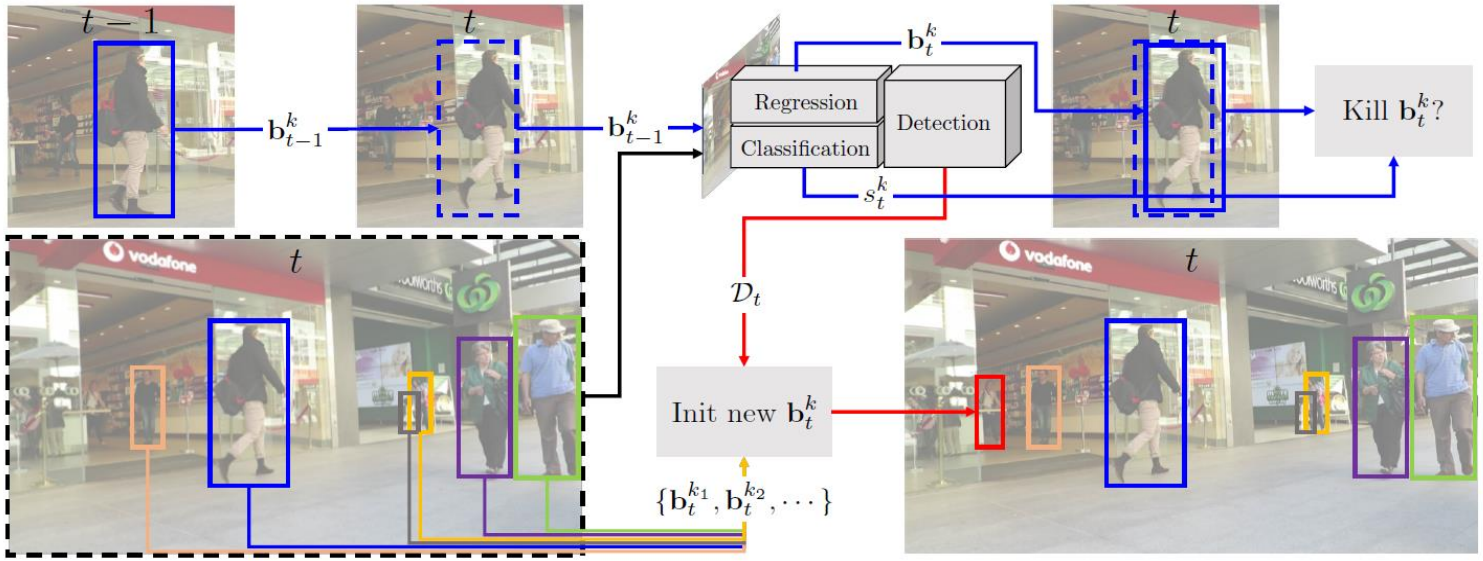
OBJECT TRACKING METHODS: **TRACKR-CNN**

The association head draws inspiration from siamese networks and the embedding vectors used in person re-identification. It is trained using a video sequence adaptation of **batch hard triplet loss**, which is a more efficient method than the original triplet loss.

In the final result production, the system must decide which detections should be reported. The matching between the previous frame detections and current proposals is done using the Hungarian algorithm, while only enabling pairs of detections with association vectors smaller than some threshold.

OBJECT TRACKING METHODS: **TRACKTOR++**

The Multiple Object Tracking Benchmark, thanks to its public leaderboard, makes it easier to discover the most recent breakthroughs in MOT. **Tracktor++** dominated the leaderboard with a simple and effective approach. This model predicts the position of an object in the next frame by calculating the bounding box regression, without needing to train or optimize on tracking data whatsoever. The usual Faster R-CNN with 101-layer ResNet and FPN is the object detector for Tracktor++, trained on the MOT17Det pedestrian detection dataset.



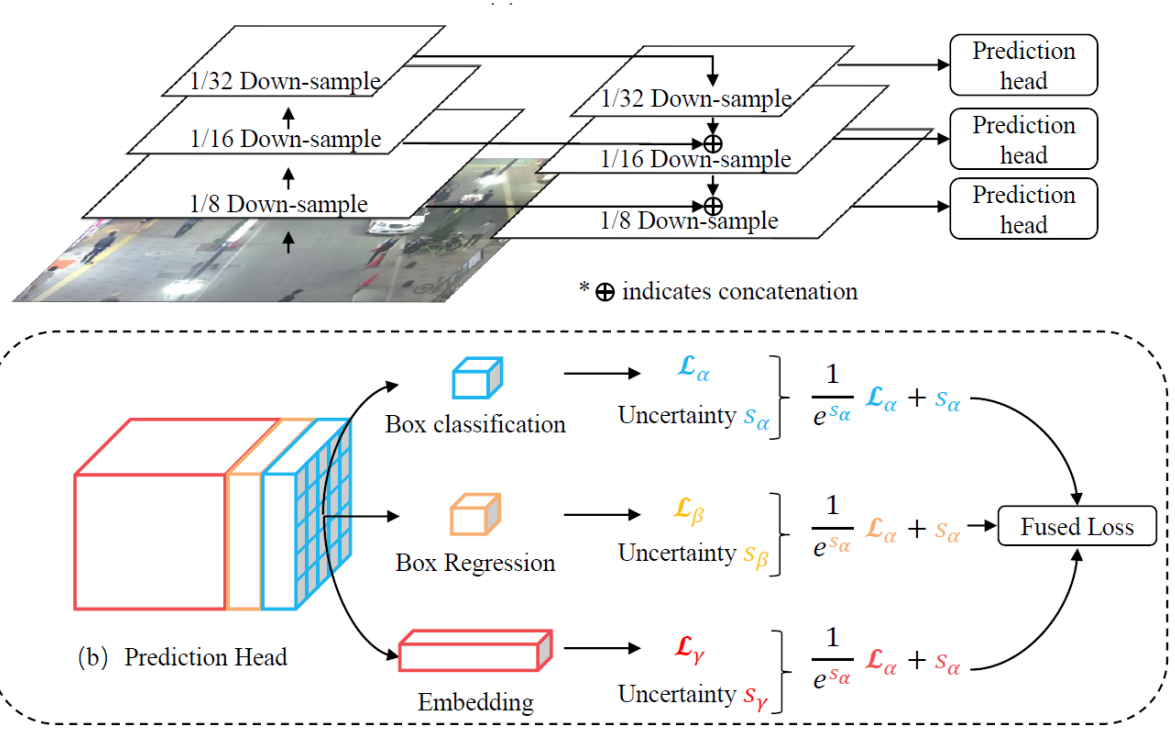
OBJECT TRACKING METHODS: **TRACKTOR++**

The principal idea of Tracktor++ is to use the regression branch of Faster R-CNN for frame-to-frame tracking by extracting features from the current frame, and then using object locations from the previous frame as input for the RoI pooling process to regress their locations into the current frame.

It also uses several motion models like the camera motion compensation based on image registration, and short-term **re-identification**. The re-identification method store deactivated tracks for a set number of frames, and compares the newly detected tracks to them for possible re-identification. The siamese neural network measures the distance between tracks.

OBJECT TRACKING METHODS: **JDE**

Joint Detection and Embedding (JDE) is identical to RetinaNet that deviates from the two-stage paradigm, and it is a recent proposal. This single-shot detector is designed to solve a multi-task learning problem, such as anchor classification, bounding box regression and embedding learning. JDE uses Darknet-53 as the backbone to get feature maps of the input at three scales. Then, the feature maps are merged together using up-sampling and residual connections. Lastly, predictions heads are attached on top of the merged feature maps.



OBJECT TRACKING METHODS: **JDE**

To reach object tracking, the JDE model outputs appearance embedding vectors when processing the frames.

The appearance embeddings are compared to embeddings of previously detected objects using an affinity matrix. The Hungarian algorithm and Kalman filter are used for smoothing out the trajectories and predicting the locations of previously detected objects in the current frame.

- A Hungarian algorithm **inform if an object in current frame is the same as the one in previous frame.** It will be used for association and id attribution.
- A Kalman Filter is an algorithm that can **predict future positions based on current position**. Also, **estimate current position better** than what the sensor inform. It will be used to have better association.

THE MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR OBJECT DETECTION

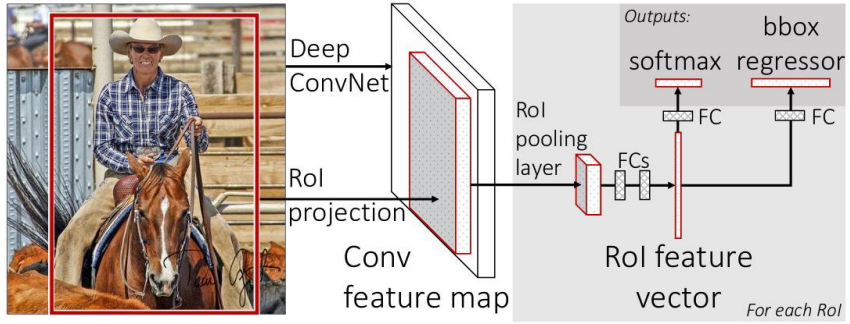
Machine learning uses algorithmic models that allow the computer to teach itself about the context of visual data. The computer will “look” at the data and teach itself to tell one image from another,
if enough data is provided through the model,. Algorithms allow the machine to learn by itself, rather than someone programming it to recognize an image.

A CNN aids a machine learning or deep learning model “look” by breaking images down into pixels that are given tags or labels. It uses the labels to perform convolutions (a mathematical operation on two functions to generate a third function) and do predictions about what it is “seeing.” The neural network runs convolutions and checks the accuracy of its predictions in a series of iterations until the predictions start to come true.
Then, it is recognizing or seeing images similar to humans.

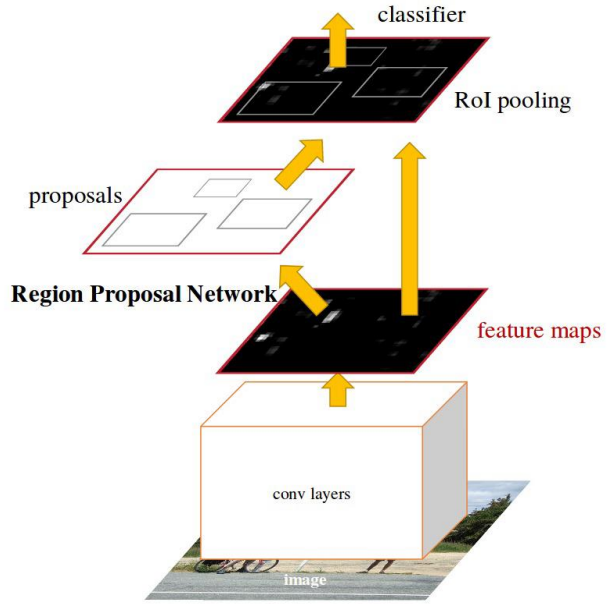
THE MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR OBJECT DETECTION

Convolutional Neural Networks (CNN) first recognize hard edges and simple shapes, then fills in information as it runs iterations of its predictions. It is used to understand single images. CNN have been applied to the object detection problem, resulting in increased performance. The three most popular architectures are in Figure below.

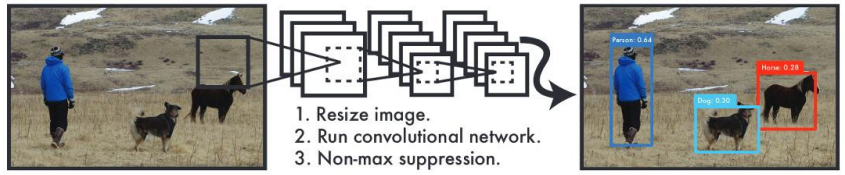
Region Network (Fast-RCNN)



Region Proposal Network (Faster-RCNN)



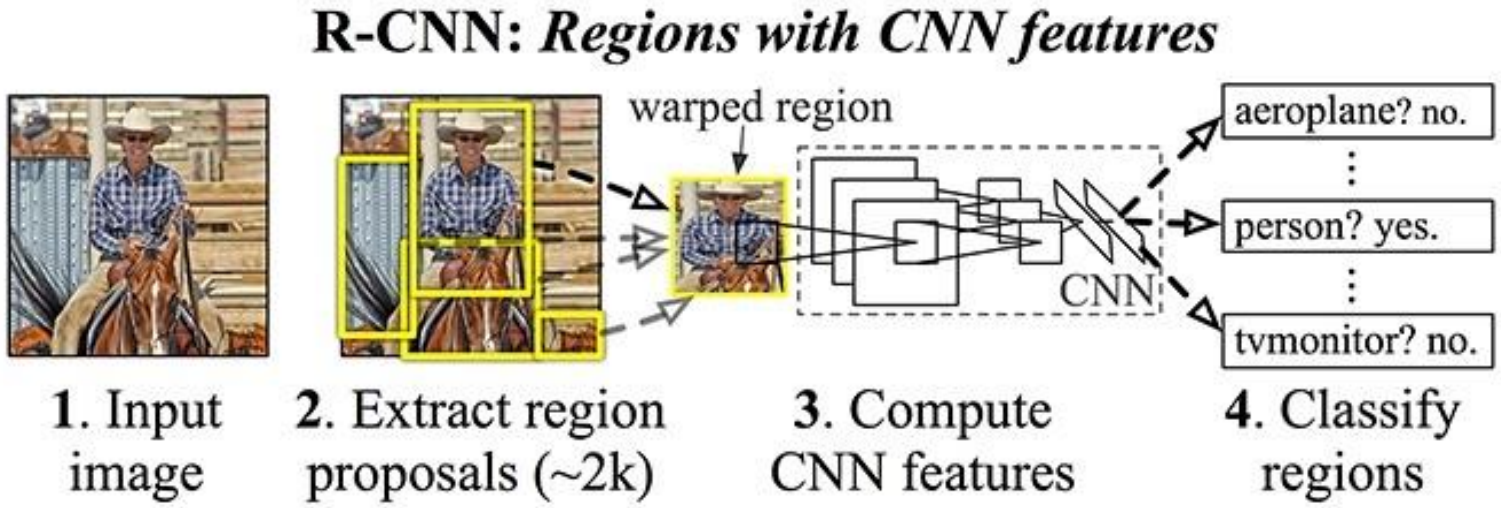
One-Stage Detector (YOLO)



THE MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR OBJECT DETECTION

A recurrent neural network (RNN) is used in a equal way for video applications to help computers understand how pictures in a series of frames are related to one another. It is simple to understand. There are three stage process:

- Extract possible objects using a region proposal method.
- Identify features in each region using a CNN.
- Classify each region utilizing **SVMs**.



THE MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR OBJECT DETECTION

Step 3 is very important as it decreases the number of object candidates, which makes the method less computationally expensive.

The features extracted here are less intuitive. A CNN is used to extract a 4096-dimensional feature vector from each region proposal. Given the nature of this, it is necessary that the input ever have the same dimension. This is usually one of the CNN's weak points and the various approaches address this in different ways. The trained CNN architecture requires inputs of a fixed area of 227 × 227 pixels.

THE MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR OBJECT DETECTION

The operations used in computer vision based on a Deep Learning perspective are:

- **Convolution**: is an operation in which a learnable kernel is “convolved” with the image.The kernel is slided across the image pixel by pixel, and an element-wise multiplication is performed between the kernel and the image at every pixel group.
- **Pooling:** is an operation to reduce the dimensions of an image by performing operations at a pixel level. The kernel slides across the image, and only one pixel from the corresponding pixel group is chosen for processing, reducing the image size. For example: Max Pooling, Average Pooling.
- **Non-Linear Activations:** introduce non-linearity to the neural network, as a result of enabling the stacking of multiple convolutions and pooling blocks to rise model depth.

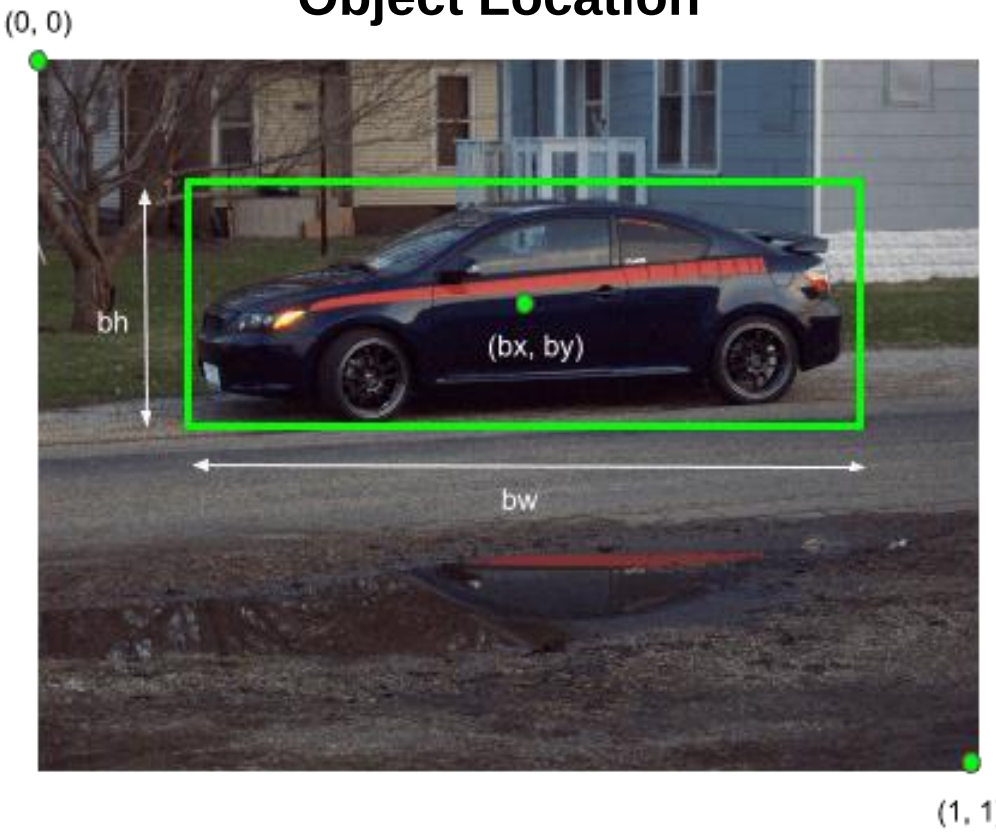
TARGET REPRESENTATION AND LOCALIZATION ALGORITHMS

An image classification or image recognition model detect the probability of an object in an image. **Object localization** is about identifying the location of an object in the image. The algorithm will output the coordinates of the location of an object with respect to the image. The most common way to localize an object in an image is to represent its location with the help of bounding boxes like in figure below that shows an example of a bounding box.

A bounding box use the following parameters:

- bx, by : coordinates of the center of the bounding box
- bw : width of the bounding box w.r.t the image width
- bh : height of the bounding box w.r.t the image height

Object Location



Defining the target variable

The target variable for a multi-class image classification problem is defined as:

$$\hat{y} = c_i$$

where,

c_i = Probability of the i_{th} class.

For example, if there are four classes, the target variable is defined as

$$y = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}$$

We can extend this approach to define the target variable for object localization. The target variable is defined as

$$y = \begin{bmatrix} p_c \\ b_x \\ b_y \\ b_h \\ b_w \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}$$

where,

p_c = Probability/confidence of an object (i.e the four classes) being present in the bounding box.

b_x,b_y,b_h,b_w = Bounding box coordinates.

c_i = Probability of the ith class the object belongs to.

For example, the four classes be ‘truck’, ‘car’, ‘bike’, ‘pedestrian’ and their probabilities are represented as c_1,c_2,c_3,c_4. So,

$$p_c = \begin{cases} 1, & c_i : \{c_1, c_2, c_3, c_4\} \\ 0, & otherwise \end{cases}$$

LOCALIZATION ALGORITHMS

Image localization is a spin-off of regular CNN vision algorithms. In object localization, the algorithm predicts a set of 4 continuous numbers, that is x coordinate, y coordinate, height, and width, to draw a bounding box around an object of interest.

Initial layers are convolutional neural network layers ranging from a couple of layers to 100 layers, depending on the application, amount of data, and computational resources available. There are a pooling layer, after the CNN layers, then one or two fully connected layers. The last layer is the output layer that gives a probability of an object being present in an image. For example, an algorithm identifies 100 different objects, so the last layer gives an array length of 100 and values ranging from 0 to 1 that indicate the probability of an object being present in an image. In an image localization algorithm, everything is the same except the output layer.

LOCALIZATION ALGORITHMS

Other algorithms that are prefered:

- R-CNN (region-based CNN)
- Fast and Faster R-CNN (improved version of R-CNN)
- YOLO (highly efficient object detection framework)
- SSD (single shot detectors)

FILTERING AND DATA ASSOCIATION ALGORITHMS

Thinking in images as functions mapping locations in images to pixel values, **filters** can be just systems that form a new, and preferably enhanced, image from a combination of the original image's pixel values.

Data association can be defined as the process of matching information about newly observed objects with information that was previously observed about them. This information may be about their identities, positions, or trajectories.

Data association algorithms look for matches that optimize certain match criteria and are subject to physical conditions.

DATA ASSOCIATION ALGORITHMS

Data association algorithms can be classified as single-scan, multi-scan, or batch.

- A **single-scan** algorithm just uses measurement or track information from the recent time step,

whereas multi-scan algorithms utilize information from previous and/or future time steps.

- **Batch**, or offline multi-object tracking, is an extreme version of multi-scan where the all sequence

is available.

- **Multi-scan** methods, usually are preferable in situations where the objects of interest are

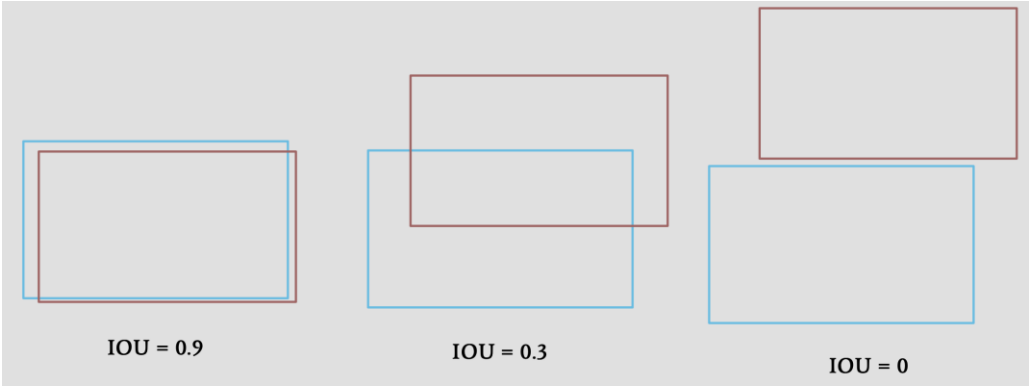
closely spaced and there are a lot of false alarms and missed detections.

THE HUNGARIAN ALGORITHM (KUHN-MUNKRES)

The hungarian algorithm, or Kuhn Munkres algorithm, based on a score, can associate an obstacle from one frame to another.

There are many scores that we can think of :

- **IOU** (Intersection Over Union): define that if the bounding box is overlapping the previous one, it's very likely the same.



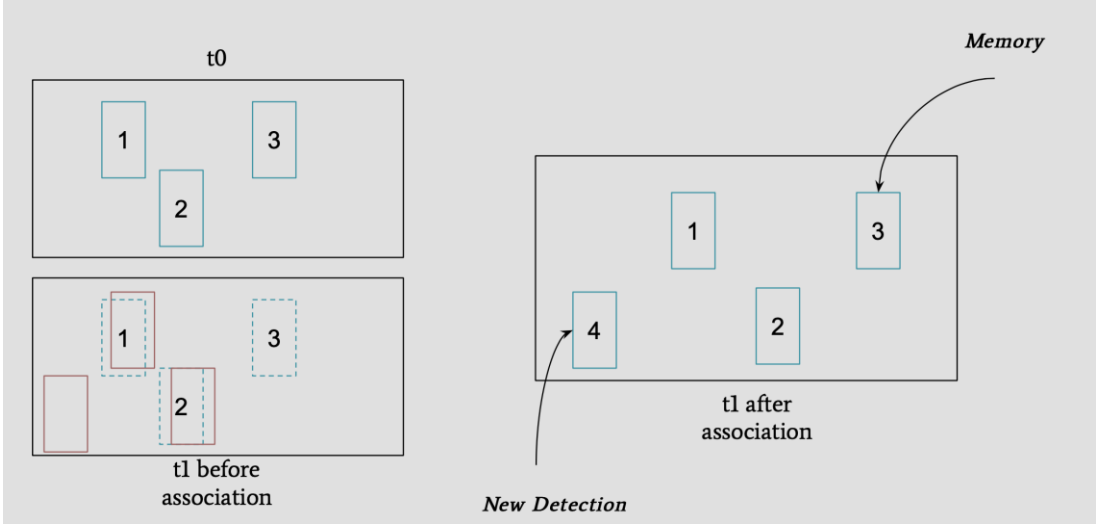
- **Shape Score:** the score increases, if the shape or size didn't vary a lot during two consecutives frames.
- **Convolution Cost:** it is possible to run a CNN (Convolutional Neural Network) on the bounding box and compare the result with the one from the frame b efore. If the convolutional features are the same, it means the objects looks the same. If there is a partial bl ockage, the convolutional features will stay in part the same and the association will remain.

THE HUNGARIAN ALGORITHM (KUHN-MUNKRES)

How to do the association ?

The process is the following :

- There are two lists of boxes from YOLO : a **tracking list (t-1)** and a **detection list (t)**.
- Calculate IOU, shape, convolutional score going through tracking and detection list. For example, consider just IOUs, it is possible to have a cost function giving importance to each score. For convolutions, cosine distance metrics would be used. **Store the IOU scores in a matrix**
- In some cases of overlapping bounding boxes, it is possible to have two or more matches for one candidate.



THE KALMAN FILTER

Kalman Filters are known for tracking obstacles and predicting current and future positions. It is used in all sort of robots, drones, self-flying planes, self-driving cars, multi-sensor fusion.

A Kalman Filter is used on each bounding box, so it comes after a box has been matched. After the association, **predict** and **update** functions are called. These functions implement the math of Kalman Filters composed of formulas for determining state mean and covariance. These are what we want estimate.

Mean is the coordinates of the bounding box, **Covariance** is the uncertainty of the bounding box having such coordinates.

Mean (x) is a state vector. It is composed by coordinates of the center of the bounding box (cx,cy), size of the box (width, height) and the change of each of these parameters, velocities.

Covariance (P) is our uncertainty matrix in the estimation.

FACIAL RECOGNITION

Facial Recognition is a subpart of object detection where the primary object being detected is the human face.

Features are detected and localized as in object detection, but facial recognition not only performs detection as also recognition of the detected face. The system search for usual features and landmarks in faces like nose, eyes, and mouth and classify with the help of these features and the positioning of these landmarks.

Traditional Image Processing based methods for facial recognition include Haar Cascades easily accessible via the OpenCV library while more robust methods including the use of Deep Learning based algorithms are found in works like FaceNet.

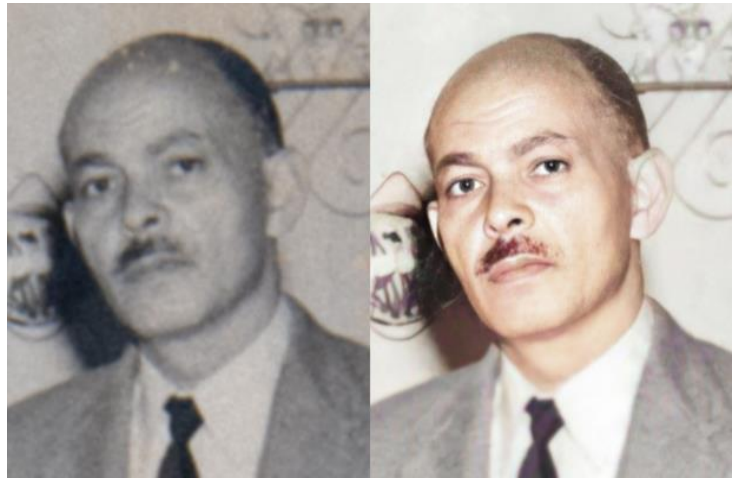
FACE RECOGNITION APPLICATIONS

- Automated surveillance
- Monitoring closed circuit television
- Image database investigations
- Multimedia environments with adaptive human computer interfaces
- Airplane-boarding gate
- Sketch-based face reconstruction
- Forensic applications
- Face spoofing and anti-spoofing, where a photograph or video of an authorised person's face could be used to gain access to services.

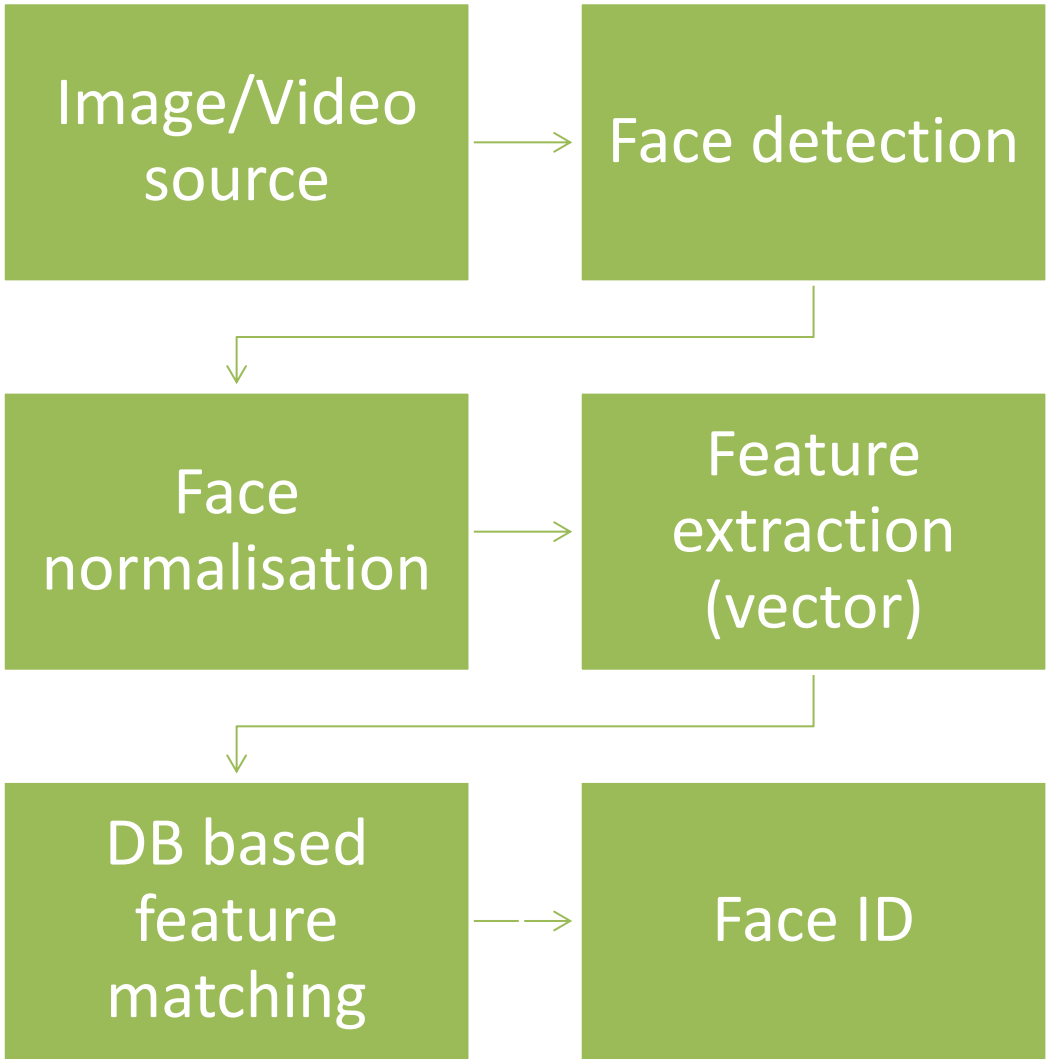
FACE RECONSTRUCTION

Image Restoration is the reconstruction of faded and old image hard copies that have been captured and stored in an improper manner, leading to loss of quality of the image. This process implies the reduction of additive noise via mathematical tools, while reconstruction demand major changes, leading to further analysis and the use of image in painting.

In Image in painting, damaged parts of an image are filled with the help of generative models that make an estimate of what the image is trying to convey. Frequently, the restoration process is followed by a colourization process that colours the subject of the picture (black and white) in the most realistic manner possible.



TECHNIQUES FOR FACE ACQUISITION



MOTION ESTIMATION

Video motion analysis is a task in machine vision that study moving objects or animals and the trajectory of their bodies.

Motion analysis can be defined as a combination of many subtasks, specially object detection, tracking, and segmentation, and pose estimation.

Human motion analysis is used in areas like sports, medicine, surveillance, and physical therapy. Also is used in other areas like manufacturing and to count and track microorganisms such as bacteria and viruses.

Motion estimation examines the movement of objects in an image sequence to try to obtain vectors representing the estimated motion.

MOTION ESTIMATION: TECHNIQUES

Feature-based methods

- Extract visual features (corners, textured areas) and track them
- Sparse motion fields, but possibly robust tracking
- Appropriate when image motion is large (10s of pixels)

Direct-methods

- Recover image motion directly from spatio-temporal image brightness variations
- Global motion parameters directly recovered without an intermediate feature motion calculation
- Dense motion fields, yet more sensitive to appearance variations
- Appropriate for video and when image motion is small (< 10 pixels)

MOTION ESTIMATION: APERTURE PROBLEM

The **aperture problem** has a traditional form that focuses on the motion of opaque objects. It describes the ambiguity of the inferred motions when observing local image structures that only vary along one direction.

The “aperture problem” describes the intrinsic ambiguity of perceiving the motion of an object through a local observation. This ambiguity depends on the complexity of the structure observed through an aperture.

MOTION ESTIMATION ALGORITHMS

The majority of the motion estimation algorithms make some assumptions:

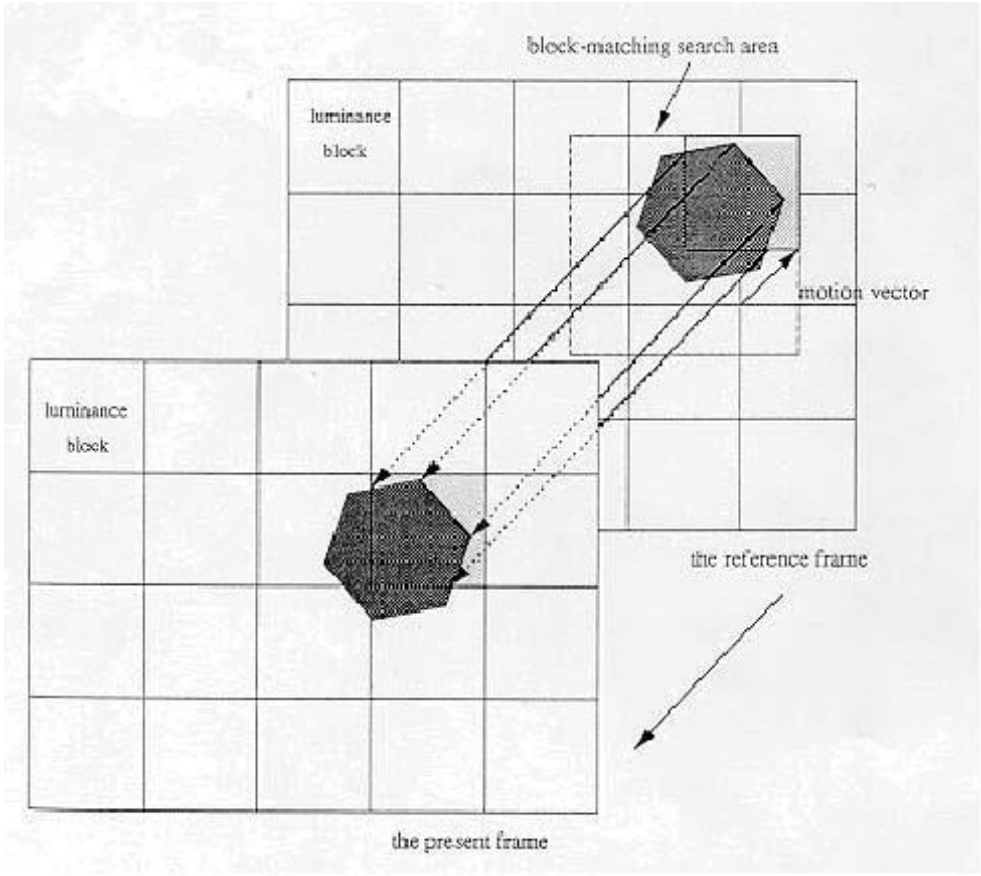
- Objects move in translation in a plane which is parallel to the camera plane. Camera zoom effects, and object rotations are not considered.
- Illumination is spatially and temporally uniform.
- Occlusion of one object by another, and uncovered background are neglected.

There are two mainstream techniques of motion estimation: **pel-recursive algorithm** (PRA) and **block-matching algorithm** (BMA). PRA are iterative refining of motion estimation for individual pels by gradient methods. BMA suppose that all the pels within a block has the same motion activity. This estimate motion on the basis of rectangular blocks and produce one motion vector for each block. It is more suitable for a simple hardware realization due to its regularity and simplicity.

PRA involves computational complexity and less regularity.

MOTION ESTIMATION ALGORITHMS

The figure on the right illustrates a process of block-matching algorithm. Each frame is divided into blocks, each of which consists of luminance and chrominance blocks. Generally, motion estimation is performed only on the luminance block for coding efficiency. Each luminance block is matched against candidate blocks in a search area on the reference frame. These candidate blocks are only the displaced versions of original block. The best (lowest distortion) candidate block is found and its displacement (motion vector) is recorded. The input frame is subtracted from the prediction of the reference frame. So, the motion vector and the resulting error can be transmitted. The decoder builds the frame difference signal from the received data at the receiver end, and adds it to the reconstructed reference frames.



SCENE UNDERSTANDING IN TYPICAL COMPUTER VISION PROBLEM

Scene understanding has the goal of build a machine that can see like humans to automatically interpret the content of the images. One of the basic tasks of basic level scene understanding is to be able to classify a natural image into a limited number of semantic categories.

Comparing with traditional vision problem:

- Study on larger scale
- Human vision related tasks

More image information/Context information.

SCENE UNDERSTANDING IN TYPICAL COMPUTER VISION PROBLEM

Human vision related task

Similar as the way that human understand the image imply more useful information from image.

How do human learn?

- Bayesian Rules:

$$P(A|B) = P(B|A)\cdot P(A)/P(B)$$

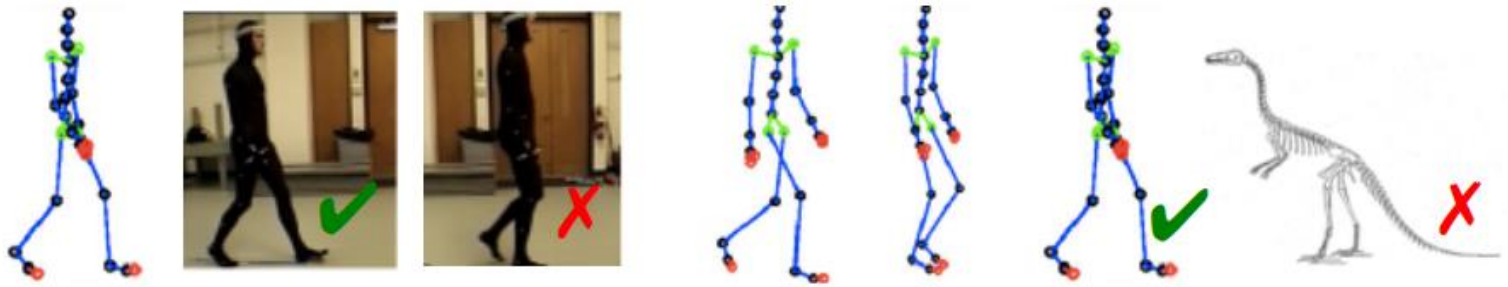
- In practice: Infer abstract knowledge based on observation

$$P(W \mid I) = P(I \mid W)\cdot P(W)/P(I) \propto P(I \mid W)\cdot P(W)$$

Prior: The probability of W w/o seeing any observation

Posterior probability

Likelihood: The probability of getting **I** given model **W**



SCENE UNDERSTANDING IN TYPICAL COMPUTER VISION PROBLEM

How do human learn?

- To teach a human what is “horse” showing 3 pictures and let him learn by himself.
- All the following concepts can explain the images:
  - “horse” = all horse
  - “horse” = all horse but not Clydesdales
  - “horse” = all animal

"horse"

I =

SCENE UNDERSTANDING IN TYPICAL COMPUTER VISION PROBLEM

Scene understanding involves:

- Localization of all instances of foreground objects (“things”)
- Localization of all background classes (“stuff”)
- Pixel-wise segmentation
- 3D reconstruction
- Pose detection
- Action recognition
- Event recognition
- etc

3D RECONSTRUCTION

The **objective** of multi-view 3D reconstruction is to infer 3D geometry from a set of 2D images by inverting the image formation process using appropriate prior assumptions.

Difference from two-view stereo, multi-view reconstruction algorithms retrieve the complete 3D shape of an object by deducing shape from many viewpoints.

3D RECONSTRUCTION APPLICATIONS

3D reconstruction system has various applications in a variety of field such as:

- Medicine
- Film industry
- Robotics
- City planning
- Gaming
- Virtual environment
- Earth observation
- Archaeology augmented reality
- Reverse engineering
- Animation
- Human computer interaction

ACTIVE METHODS

There are two principal methods of creating 3D reconstruction models: **active** methods and **passive** methods.

Active methods are known as range data methods, and numerical approximation approach is used to reconstruct the 3D profile and construct the model. The use of this method requires active interference with the reconstructed object, by using either radiometric or mechanical rangefinders. To measure distances to an object that is rotating on a turntable is used depth gauges by a mechanical method, for example, radiometric methods used include ultrasound, microwaves, and moving light sources.

PASSIVE METHODS

There are two passive methods that are often used:

- monocular cues methods;
- binocular stereo vision.

Shape-from-shading is often used, in the former category. Shade information is gathered from an object to learn about the depth of normal information on the object, and this enables the reconstruction of the object. Other example of a monocular cues method is photometric stereo, and this is a more sophisticated version of shape-from-shading. Other method widely used is shape-from-texture. This is used to discover the depth of normal information on the surface of an object by using hints from the distortion and perspective of 2D images. When binocular stereo vision is used, multiple images are collected to obtain the 3-dimensional geometric information about an object. Two cameras are used at the same time to collect the images from different angles, or one camera is used to take multiple photographs from different perspectives.

DIFFERENCE BETWEEN ACTIVE AND PASSIVE METHODS

The **difference** between active methods and passive methods is that passive methods do not interfere with the object. These methods use sensors to measure the radiance from the surface of the object and then use this for image understanding.

A large number of applications could benefit from apply 3D reconstruction to a greater number of applications.

SUMMARY

- The focus of **object detection** is finding all objects of certain classes in an image.
- **Object tracking** consists of follows or tracks an object once it is detected.
- **Object tracking methods are** Recurrent YOLO, SiamMask, Deep SORT, TrackR-CNN, Tracktor++ and JDE.
- CNN have been applied to the object detection problem, resulting in increased performance.
- **Object localization** is about identifying the location of an object in the image.
- Some **localization algorithms** are R-CNN (region-based CNN), Fast and Faster R-CNN (improved version of R-CNN), YOLO (highly efficient object detection framework), and SSD (single shot detectors).
- **Data association** can be defined as the process of matching information about newly observed objects with information that was previously observed about them.
- Data association algorithms are: The hungarian algorithm and Kalman Filters.

SUMMARY

- **Motion analysis** can be defined as a combination of many subtasks, specially object detection, tracking, and segmentation, and pose estimation.
- The **aperture problem** has a traditional form that focuses on the motion of opaque objects.
- There are two mainstream techniques of motion estimation: **pel-recursive algorithm** (PRA) and **block-matching algorithm** (BMA)
- **Scene understanding** has the goal of build a machine that can see like humans to automatically interpret the content of the images.
- The **objective** of multi-view 3D reconstruction is to infer 3D geometry from a set of 2D images by inverting the image formation process using appropriate prior assumptions.
- There are two methods of creating 3D reconstruction models: **active** methods and **passive** methods.
- The **difference** between active methods and passive methods is that passive methods do not interfere with the object

REFERENCES

Bianco, G., Gallo, A., Bruno, F., & Muzzupappa, M. (2013). A comparative analysis between active and passive techniques for underwater 3D reconstruction of close-range objects. *Sensors (Switzerland)*, *13*(8), 11007–11031. https://doi.org/10.3390/s130811007

Bobick, A. (n.d.). *Motion models CS 4495 Computer Vision-A. Bobick CS 4495 Computer Vision Motion Models*.

Computer Vision for tracking. In Computer Vision, one of the most… | by Jeremy Cohen | Towards Data Science. (n.d.). Retrieved from https://towardsdatascience.com/computer-vision-for-tracking-8220759eee85

Emami, P., & Pardalos, P. M. (2020). *Machine Learning Methods for Data Association in Multi-Object Tracking*. https://doi.org/0000001.0000001

Introduction to Motion Estimation and Compensation. (n.d.). Retrieved from https://www.cmlab.csie.ntu.edu.tw/cml/dsp/training/coding/motion/me1.html

Introduction to Object Detection | Machine Learning | HackerEarth Blog. (n.d.). Retrieved from https://www.hackerearth.com/blog/developers/introduction-to-object-detection/

REFFERENCES

Khilar, R., Chitrakala, S., & Selvamparvathy, S. (2013). 3D image reconstruction: Techniques, applications and challenges. *2013 International Conference on Optical Imaging Sensor and Security, ICOSS 2013*, 1–6. https://doi.org/10.1109/ICOISS.2013.6678395

Liang, R. H., Pan, Z. G., & Chen, C. (2004). New algorithm for 3D facial model reconstruction and its application in virtual reality. *Journal of Computer Science and Technology*, *19*(4), 501–509. https://doi.org/10.1007/BF02944751

Object Detection and Tracking in 2020 | by Borijan Georgievski | Netcetera Tech Blog. (n.d.). Retrieved from https://blog.netcetera.com/object-detection-and-tracking-in-2020-f10fb6ff9af3

Understanding Object Localization with Deep Learning. (n.d.). Retrieved from https://www.einfochips.com/blog/understanding-object-localization-with-deep-learning/

What is Computer Vision? | IBM. (n.d.). Retrieved from https://www.ibm.com/topics/computer-vision

Wójcik, W., Gromaszek, K., & Junisbekov, M. (2016). Face Recognition: Issues, Methods and Alternative Applications. *Face Recognition - Semisupervised Classification, Subspace Projection and Evaluation Methods*. https://doi.org/10.5772/62950

Xue, T., Mobahi, H., Durand, F., & Freeman, W. T. (n.d.). *The Aperture Problem for Refractive Motion*.

Yang, M. (n.d.). *Computer Vision II-Scene Understanding*.

REFERENCE TO AUTHORS

Regina Sousa

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

0000-0002-2988-196X

Diana Ferreira

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

0000-0003-2326-2153

Ana Luísa Sousa

- PhD student in Information System and Tecnologies
- Research Collaborator of the Algoritmi Research Center

0000-0001-5731-3583

REFERENCE TO AUTHORS

António Abelha

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0001-6457-0756

José Machado

- Associate Professor with Habilitation at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0003-4121-6169

Victor Alves

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0003-1819-7051

REFERENCE TO AUTHORS

This Training Material has been certified according to the rules of **ECQA – European Certification and Qualification Association.**

The Training Material was developed within the international job role committee **“Computer Vision Expert”:**

UMINHO – University of Minho (https://www.uminho.pt/PT)

The development of the training material was partly funded by the EU under Blueprint Project DRIVES.

DRIVES

Development and Research on Innovative Vocational Education Skills

Thank you for your attention

DRIVES project is project under **The Blueprint for Sectoral Cooperation on Skills in Automotive Sector**, as part of New Skills Agenda.

The aim of the Blueprint is **to support an overall sectoral strategy and to develop concrete actions to address short and medium term skills needs.**

Follow DRIVES project at:

More information at:

www.project-drives.eu

Co-funded by the Erasmus+ Programme of the European Union

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.