Introduction to Computer Vision

Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr

Version 3.2

Artificial Intelligence & Information Analysis Lab

Computer vision overview

- **Image and video acquisition**
- Camera geometry
- Stereo and Multiview imaging
- Shape from X
- 3D Robot Localization and Mapping
- Semantic 3D world mapping
- Object detection and tracking
- 3D object localization
- Object pose estimation
- Computational cinematography

Images $f(x, y)$ and videos signal $f(x, y, t)$

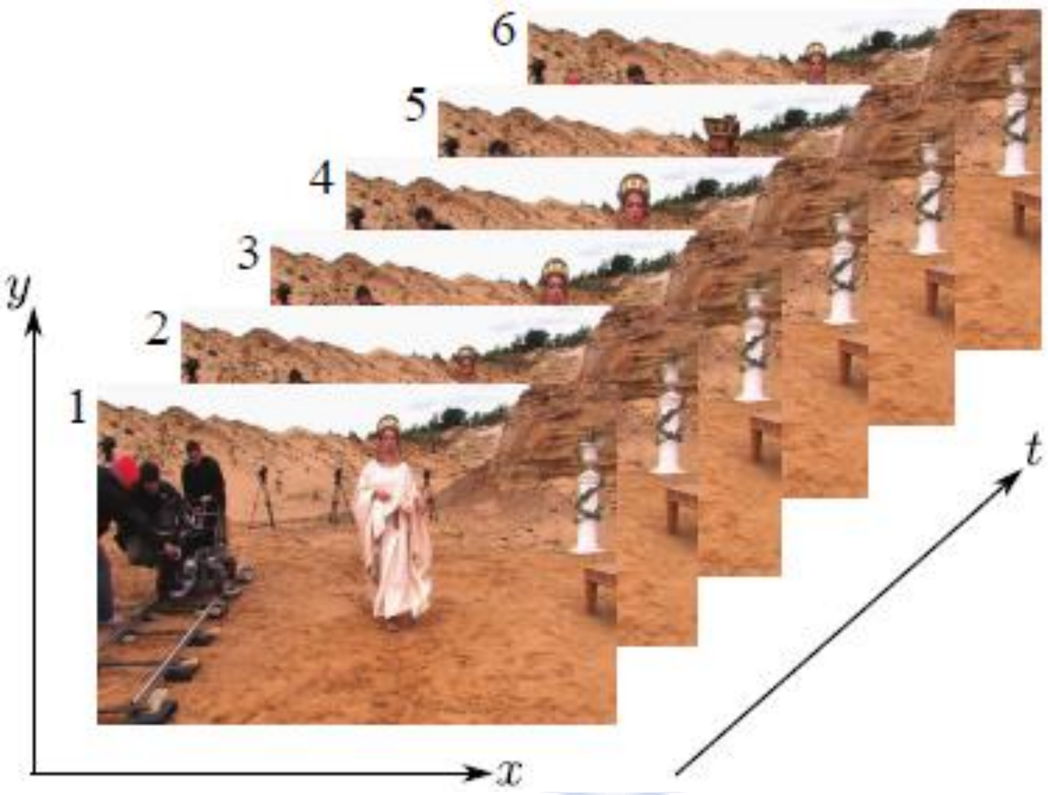
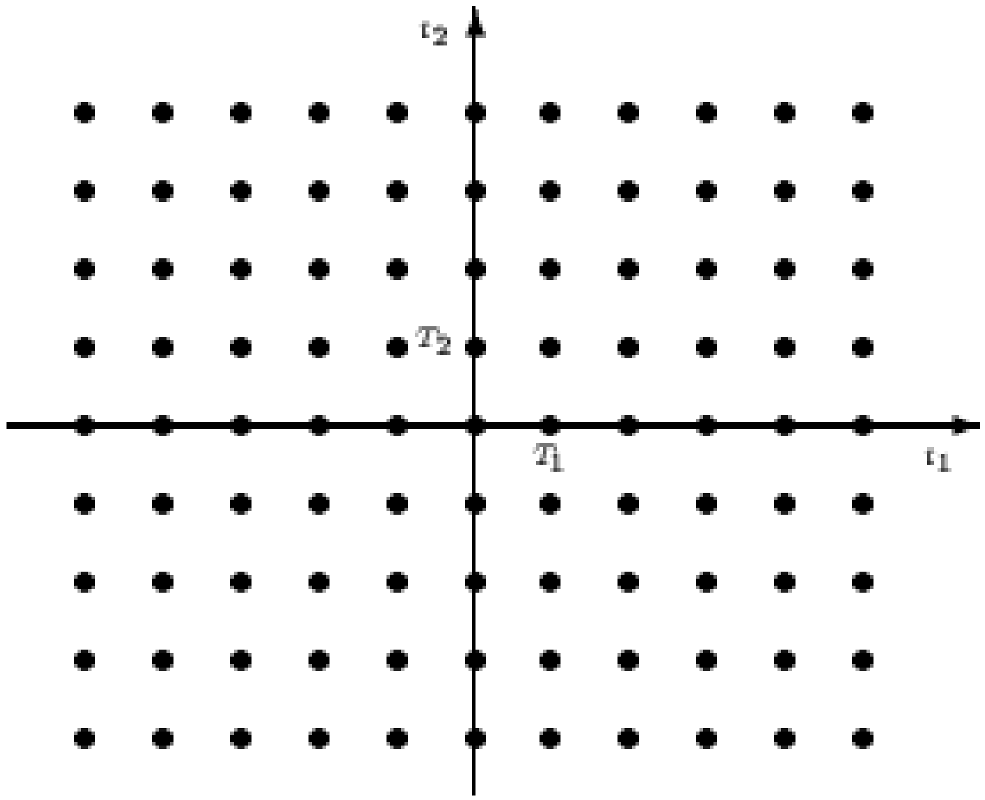


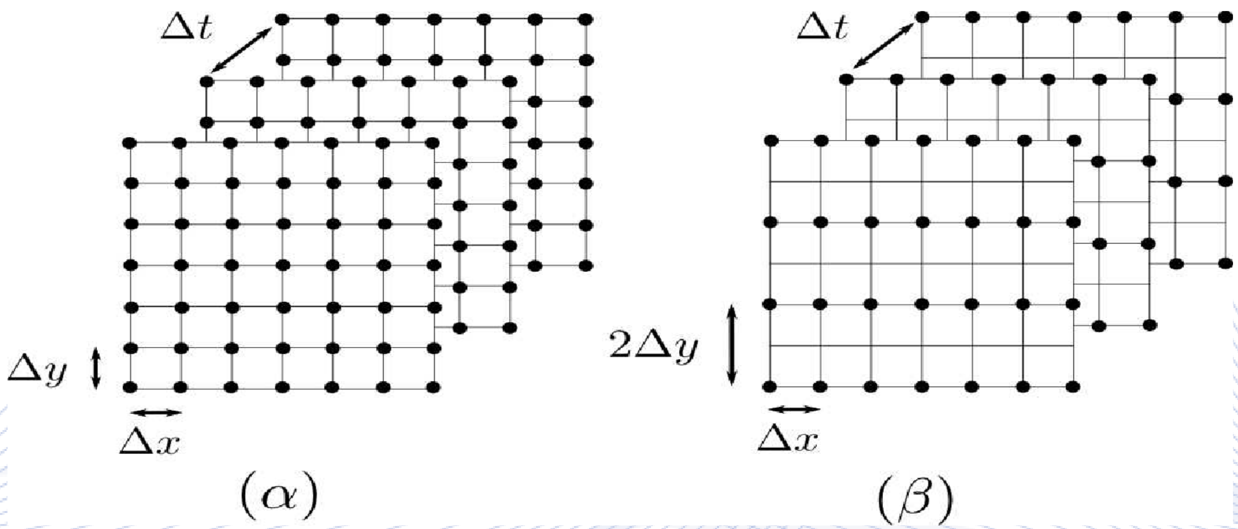
Image sampling



Rectangular sampling grid

Artificial Intelligence & Information Analysis Lab

Video sampling

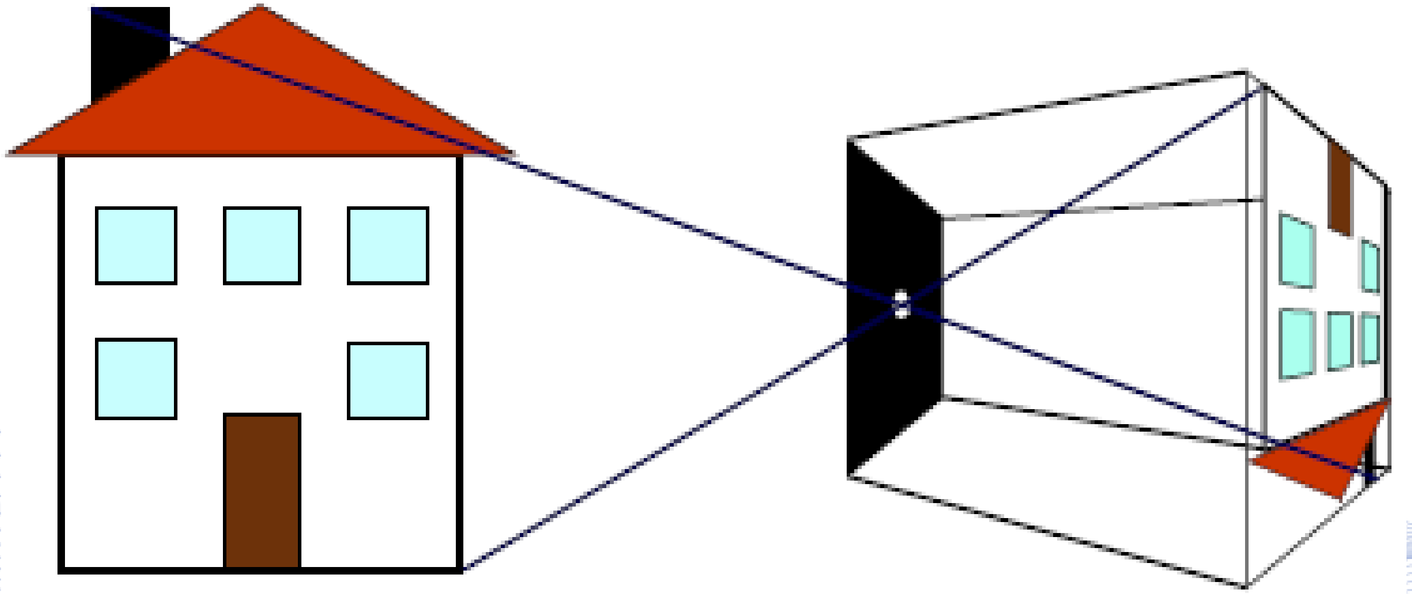


Sampling grids for: a) progressive and b) 2:1 interlaced video

Computer vision overview

- Image and video acquisition
- **Camera geometry**
- Stereo and Multiview imaging
- Shape from X
- 3D Robot Localization and Mapping
- Semantic 3D world mapping
- Object detection and tracking
- 3D object localization
- Object pose estimation
- Computational cinematography

Pinhole Camera and Perspective Projection



Pinhole camera geometry.

Artificial Intelligence & Information Analysis Lab

Camera Parameters and Projection Matrix

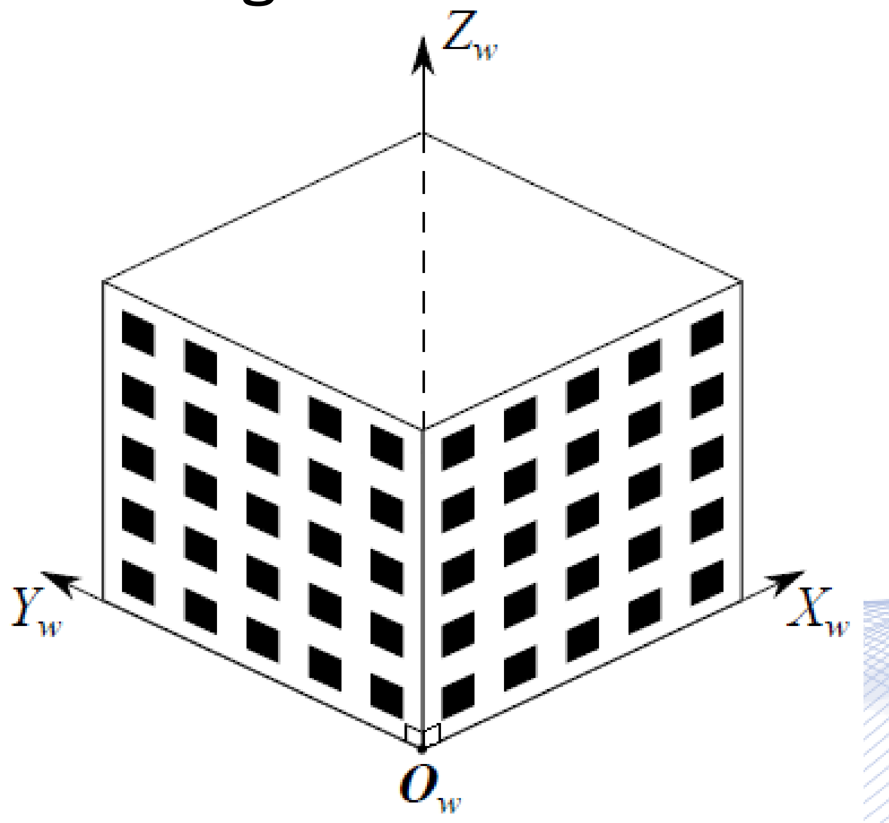
Mathematical camera description:

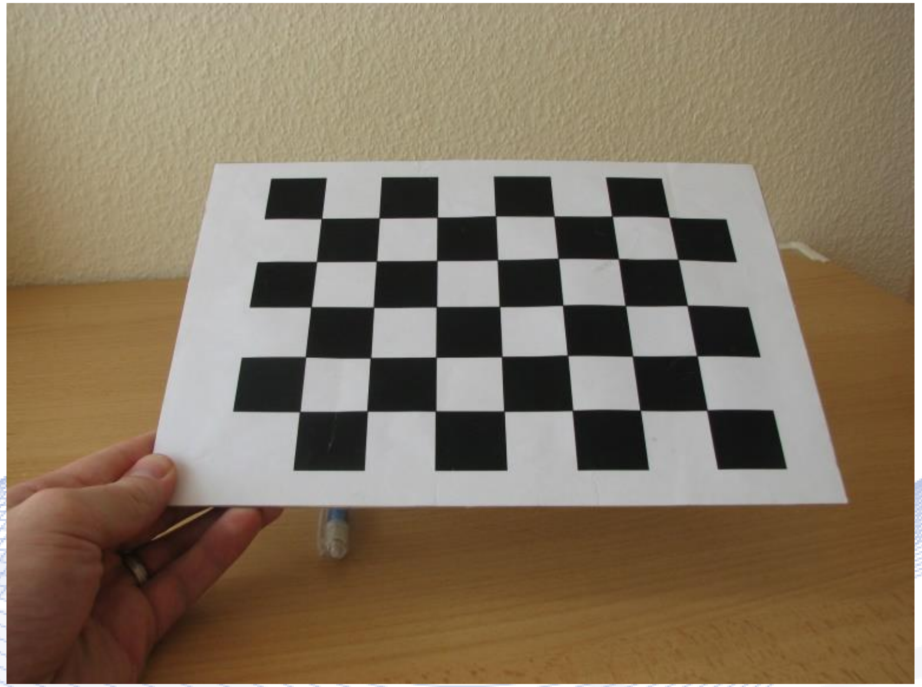
- $\boldsymbol{P} = \mathbf{P}_I \, \mathbf{P}_E$ is the 3×4 *camera projection matrix*:

$$\boldsymbol{P} = \begin{bmatrix} -\frac{f}{s_x} & 0 & o_x \\ 0 & -\frac{f}{s_y} & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & -\mathbf{R}_1^T\mathbf{T} \\ r_{21} & r_{22} & r_{23} & -\mathbf{R}_2^T\mathbf{T} \\ r_{31} & r_{32} & r_{33} & -\mathbf{R}_3^T\mathbf{T} \end{bmatrix}$$

Camera Calibration

Determining the extrinsic and intrinsic camera parameters:





Calibration patterns.

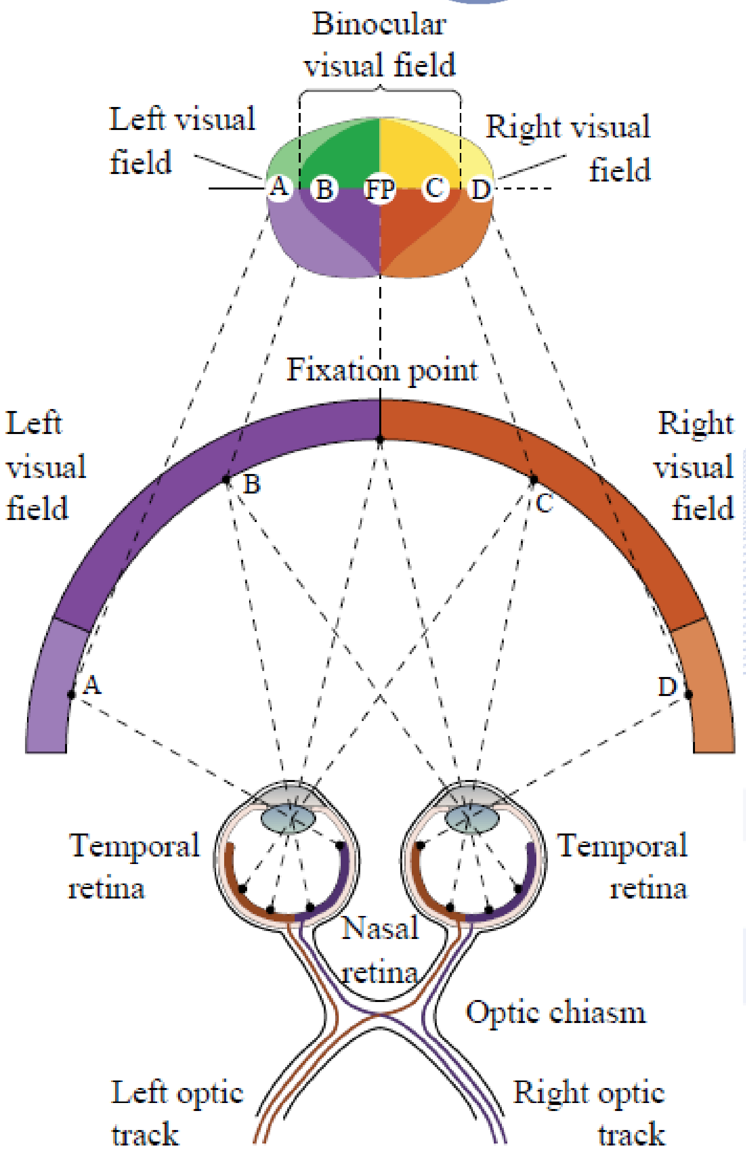
Computer vision overview

- Image and video acquisition
- Camera geometry
- **Stereo and Multiview imaging**
- Shape from X
- 3D Robot Localization and Mapping
- Semantic 3D world mapping
- Object detection and tracking
- 3D object localization
- Object pose estimation
- Computational cinematography

Stereopsis

- The horizontal separation of the eyes leads to a difference, *stereo parallax,* in image location and appearance of an object between the two eyes, called *stereo disparity.*

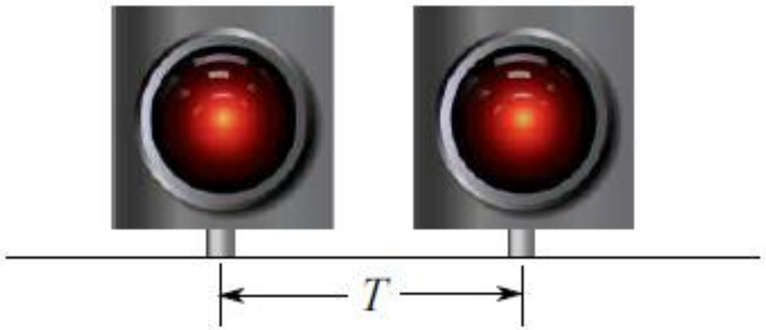
- Stereo parallax is utilized by the brain in order to extract depth information.

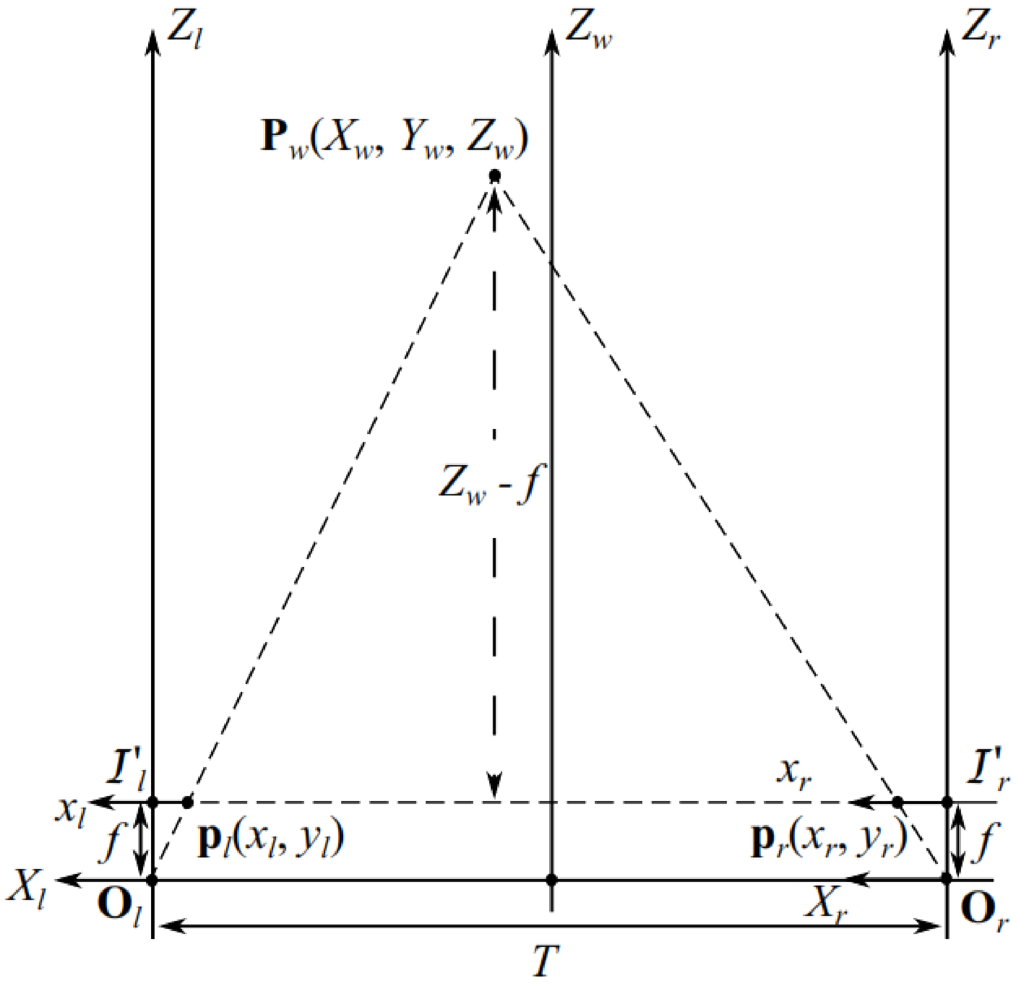


Stereo vision

Parallel Stereo vision Geometry

T: baseline
f: focal length





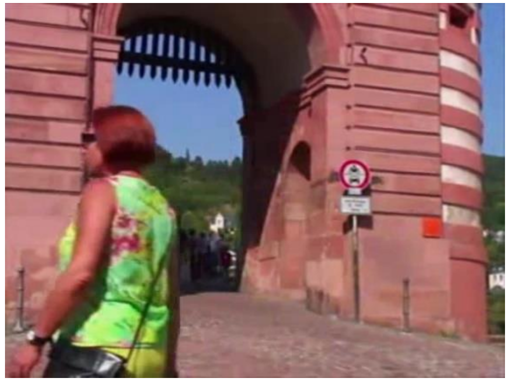
Basics of Stereopsis

Left image

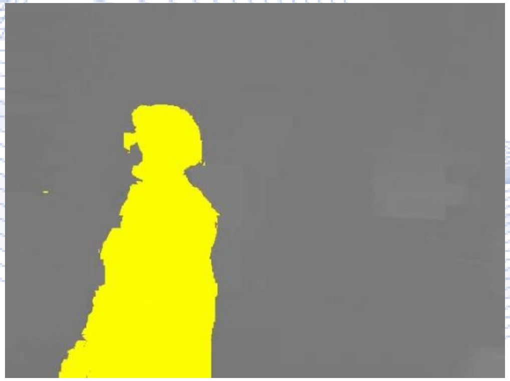
Right image

Dense disparity map

Stereo vision



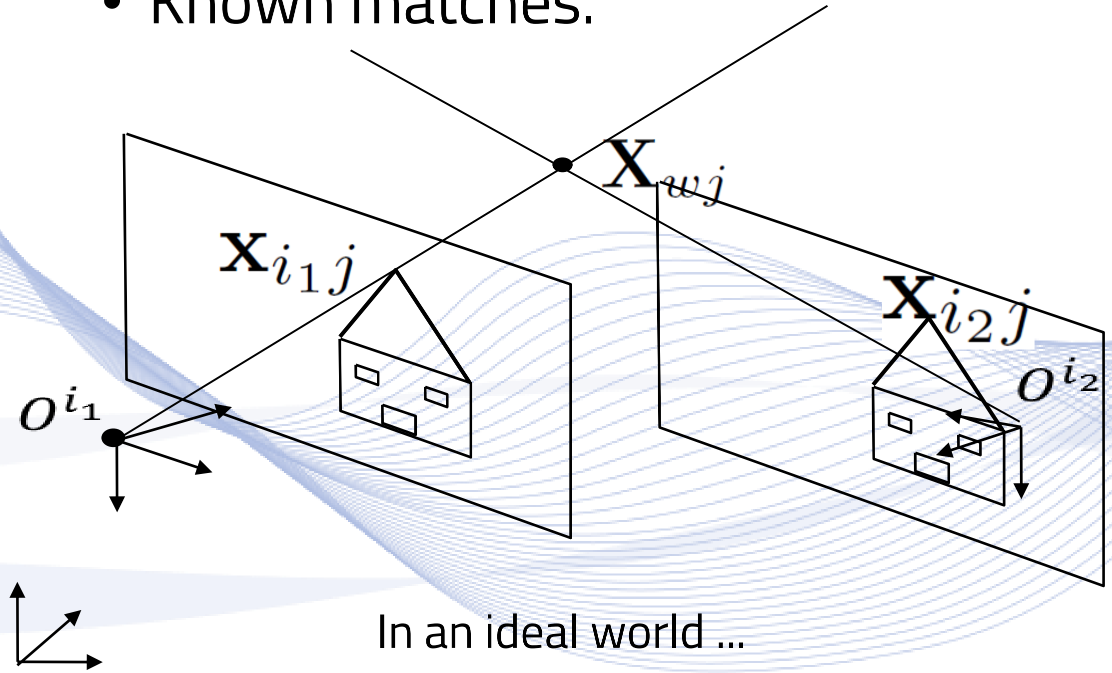
Segmented dense disparity map.



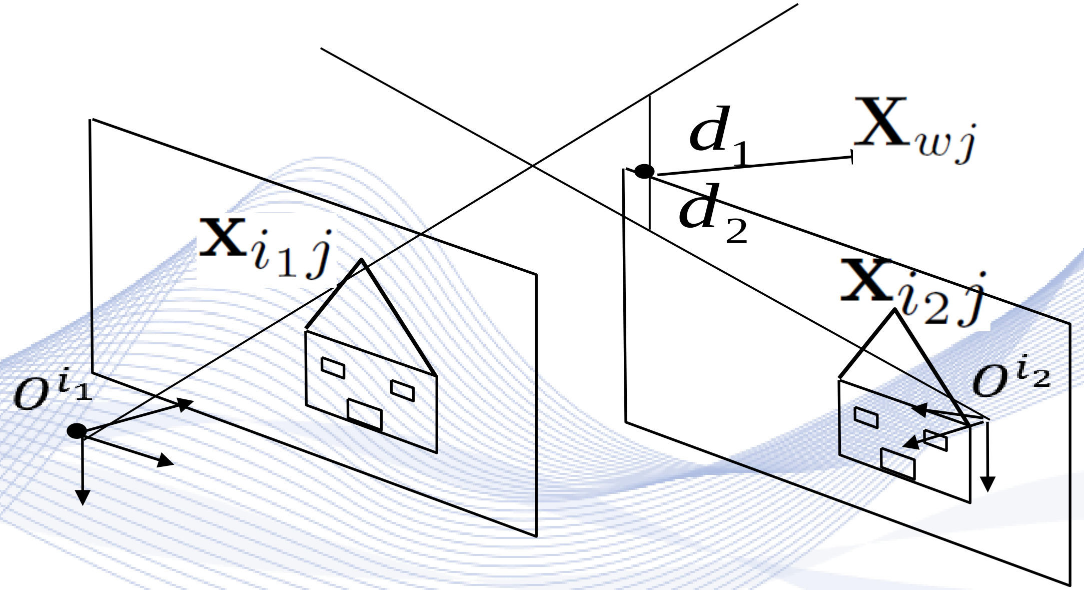
Feature Correspondence

3D perception (at least two views)

- Two cameras in known locations.
- Calibrated cameras.
- Known matches.

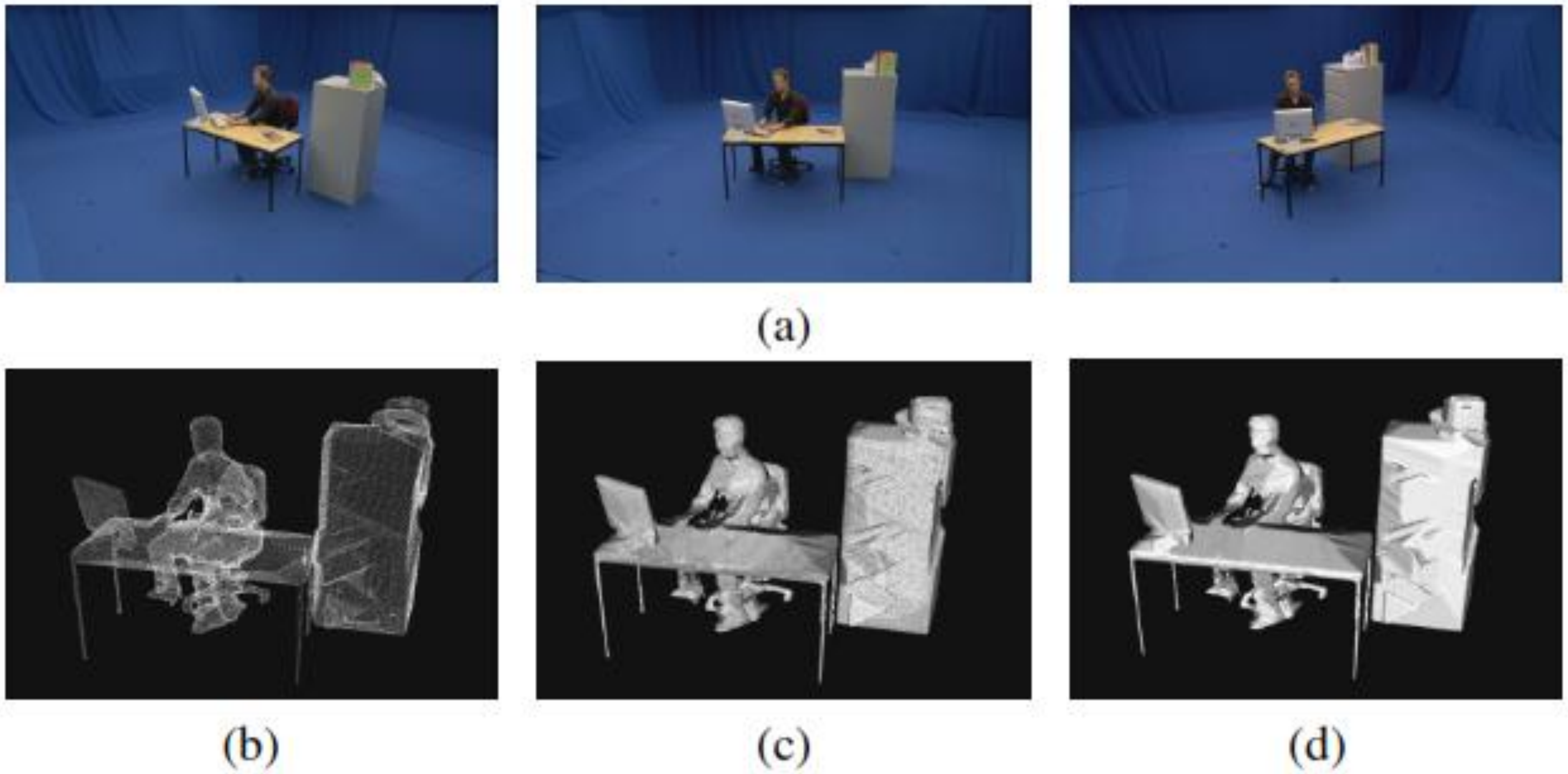


In an ideal world …



In this real world …

3D geometry reconstruction



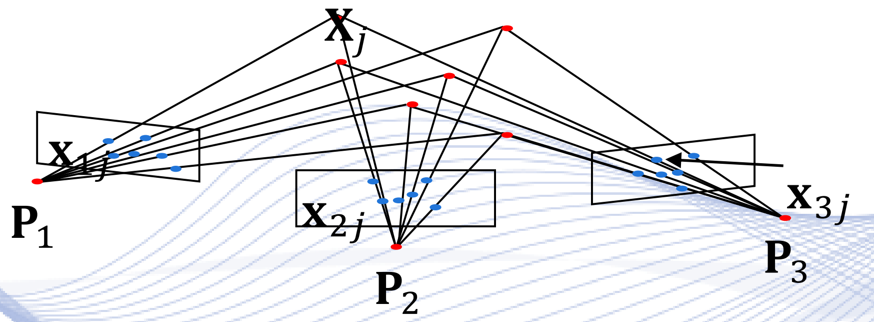
(a) (b) (c) (d)

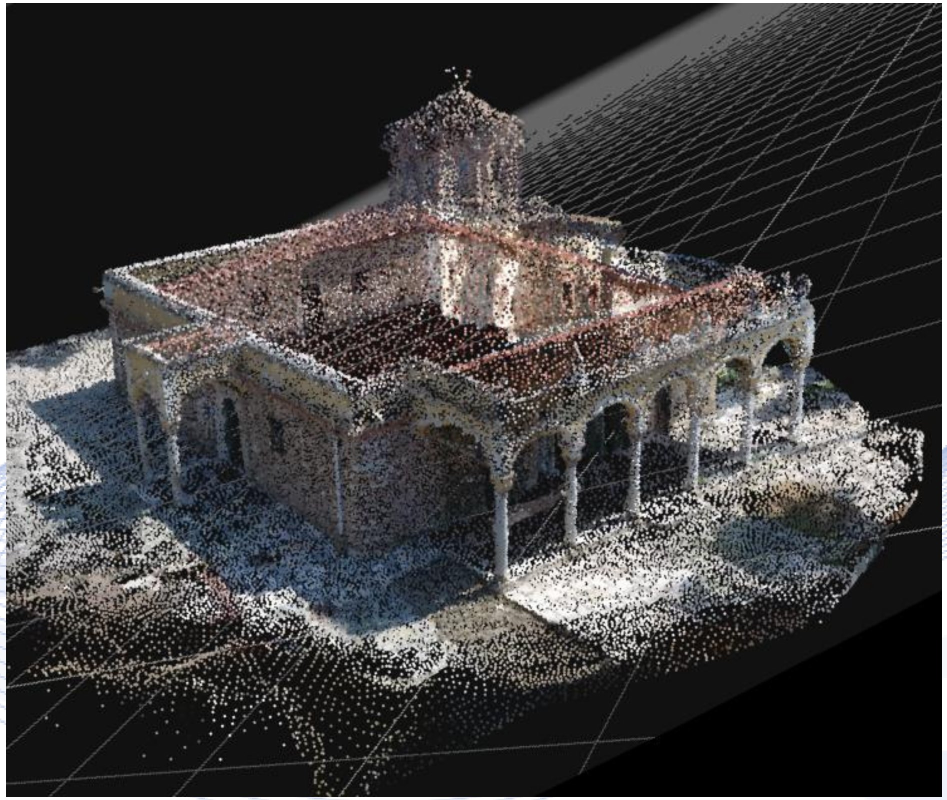
Computer vision overview

- Image and video acquisition
- Camera geometry
- Stereo and Multiview imaging
- **Shape from X**
- 3D Robot Localization and Mapping
- Semantic 3D world mapping
- Object detection and tracking
- 3D object localization
- Object pose estimation
- Computational cinematography

Structure from Motion

- Feature point correspondence
- Feature point matching
- Bundle adjustment and triangulation





3D building reconstruction

- Vladaton monastery

3D building reconstruction

3D building reconstruction

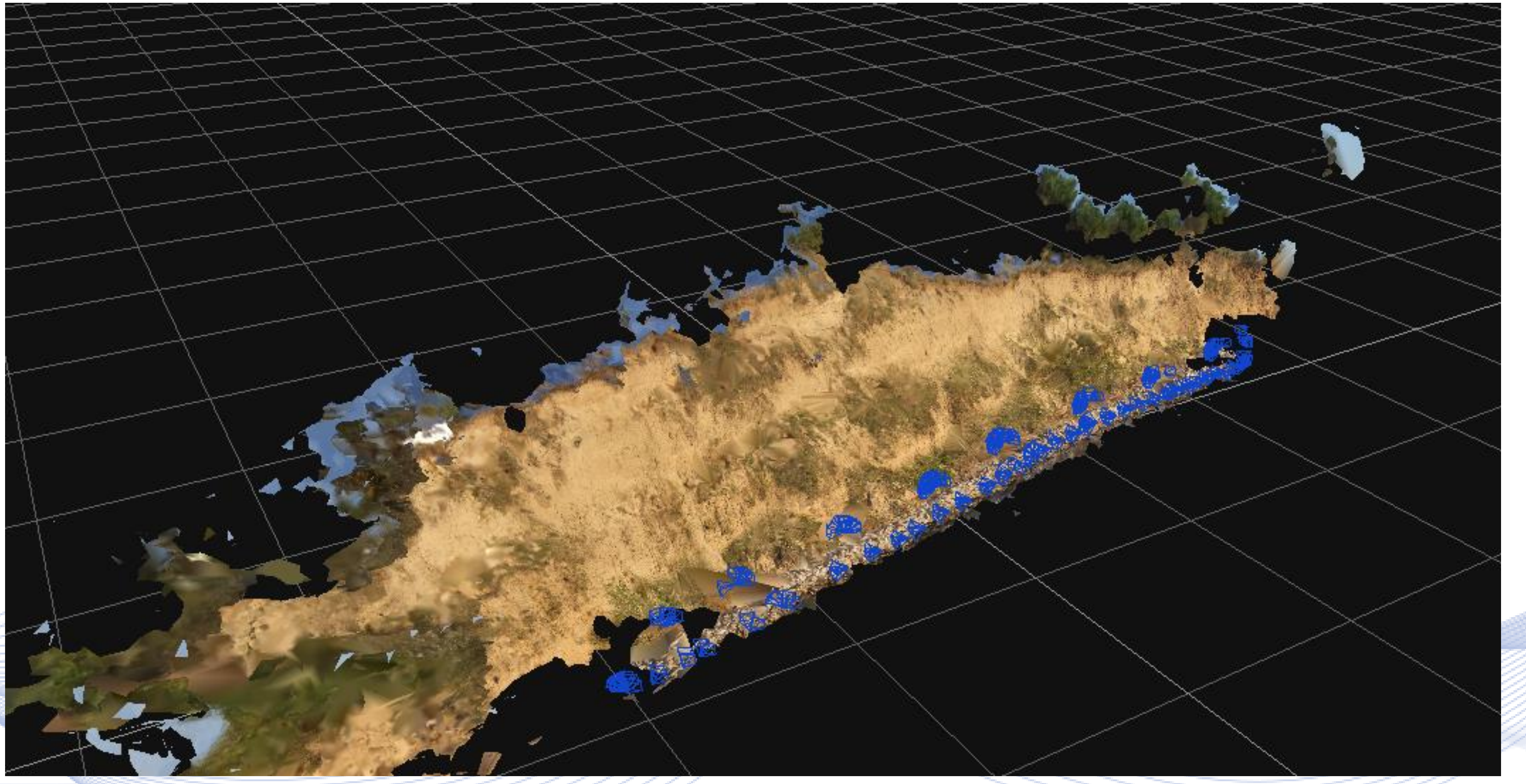
Copyright Hellenic Ministry of Culture and Sports (L. 3028/2002)

Artificial Intelligence & Information Analysis Lab

SfM in 3D landscape reconstruction

- Cliff images

SfM in 3D landscape reconstruction



3D Cliff surface reconstruction.

3D painting reconstruction

3D painting reconstruction and flattening.

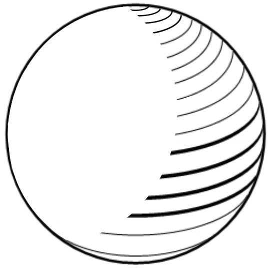
3D landscape modeling

RGB-Depth image acquired from monocular video [APOLLO].

Artificial Intelligence &
Information Analysis Lab

Shape from X

- Shape from shade.
- Shape from focus.



Computer vision overview

- Image and video acquisition
- Camera geometry
- Stereo and Multiview imaging
- Shape from X
- **3D Robot Localization and Mapping**
- Semantic 3D world mapping
- Object detection and tracking
- 3D object localization
- Object pose estimation
- Computational cinematography

3D Localization and Mapping

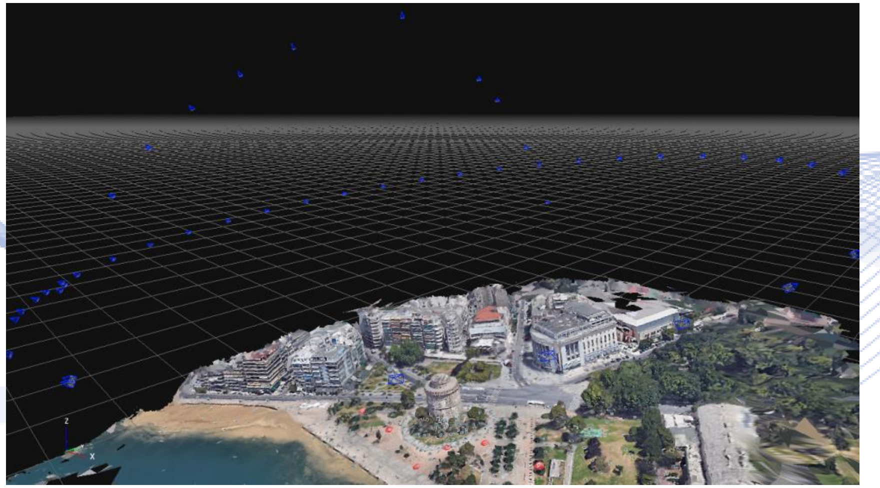
3D scene mapping and vehicle/sensor (primarily camera) localization:

- Mapping: create or get 2D and/or 3D maps.
- Localization: find the 3D location based on sensors.
- Simultaneous Localization and Mapping (SLAM).
- Information fusion in localization and mapping.

3D Robot Localization and Mapping

- 3D scene point mapping+Camera calibration


Images obtained from Google Earth

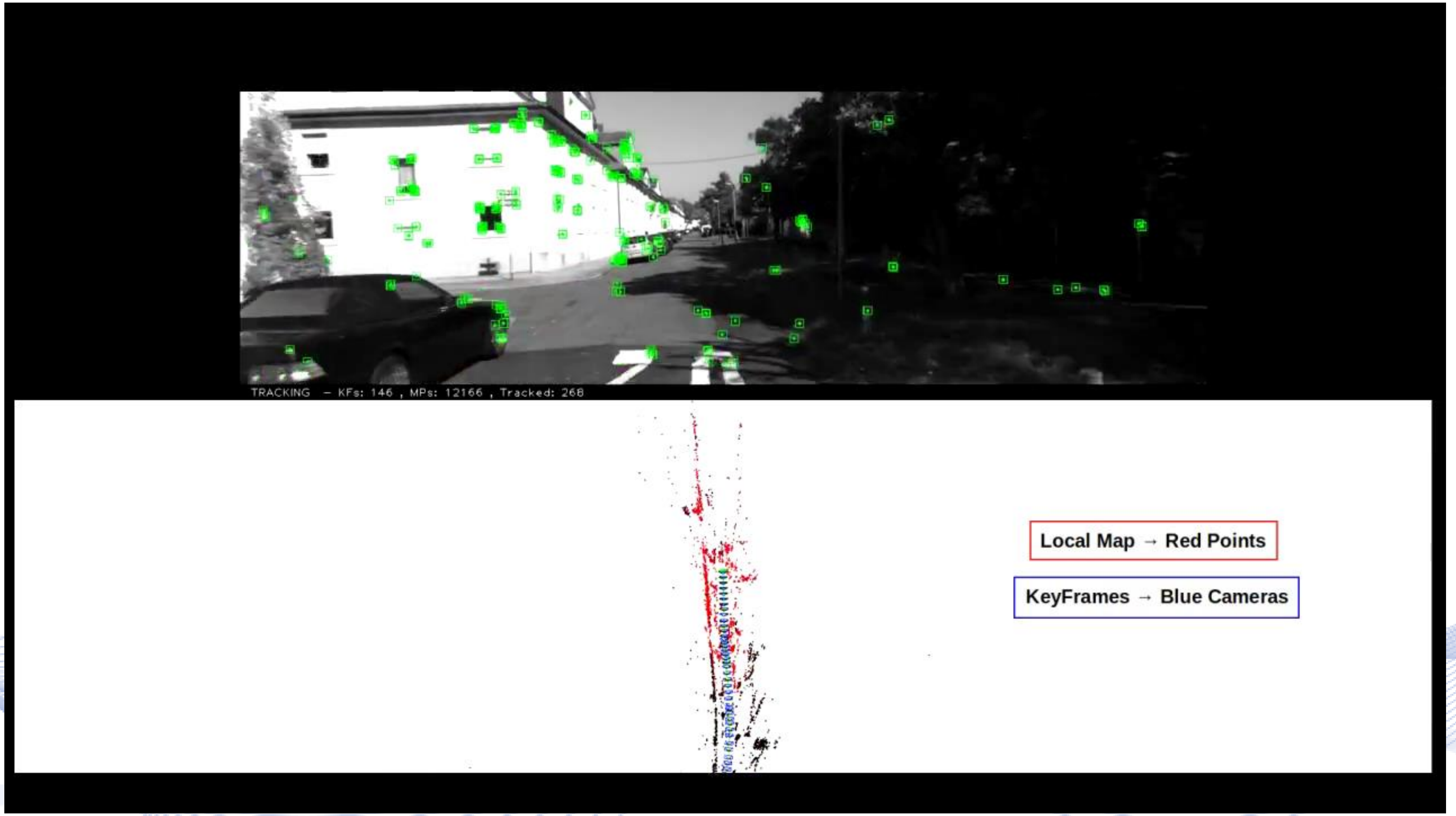



3D models reconstructed in 3DF Zephyr Free using 50 images from Google Earth

Visual SLAM

- From the sole input of the video stream:
  - Simultaneous estimation of the camera motion and the 3D scene.
  - Real-time at frame rate.
  - Sequential processing.
  - The field of view of the camera $\ll$ than the map size.
- Pivotal piece of information in automated scene interaction:
  - Sensor/robot pose with respect to the scene.
  - Localization for robots, cars, drones, autonomous navigation.
  - AR/VR user/sensor positional tracking.

Visual SLAM



https://youtu.be/sr9H3ZsZCzc

Why is place recognition difficult

Likely algorithm answer:

NO	**NO** ✗	**TRUE NEGATIVE** ☺
NO	**YES** ✓	**FALSE POSITIVE** ☹

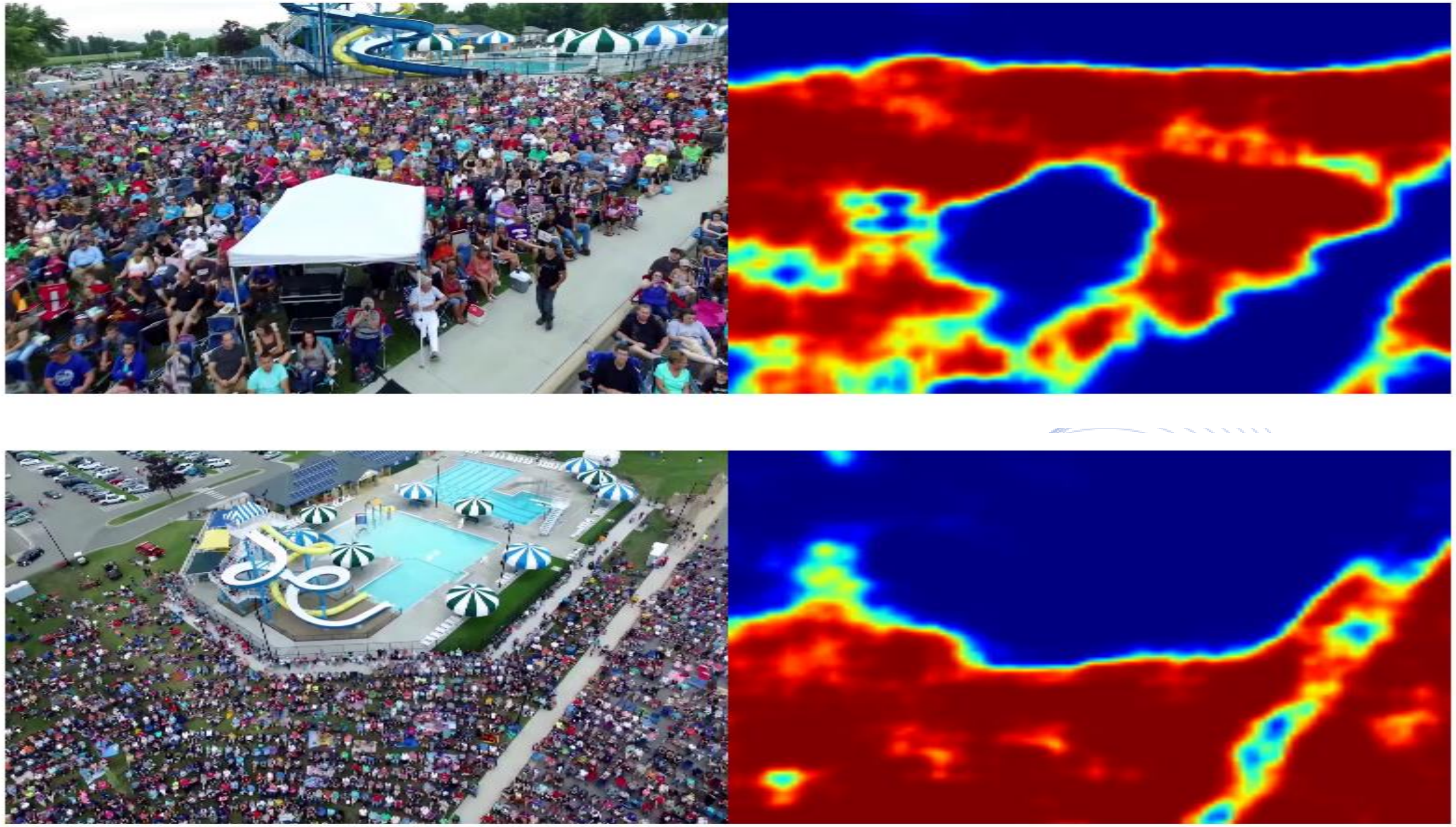
Computer vision overview

- Image and video acquisition
- Camera geometry
- Stereo and Multiview imaging
- Shape from X
- 3D Robot Localization and Mapping
- **Semantic 3D world mapping**
- Object detection and tracking
- 3D object localization
- Object pose estimation
- Computational cinematography

Semantic 3D World Mapping

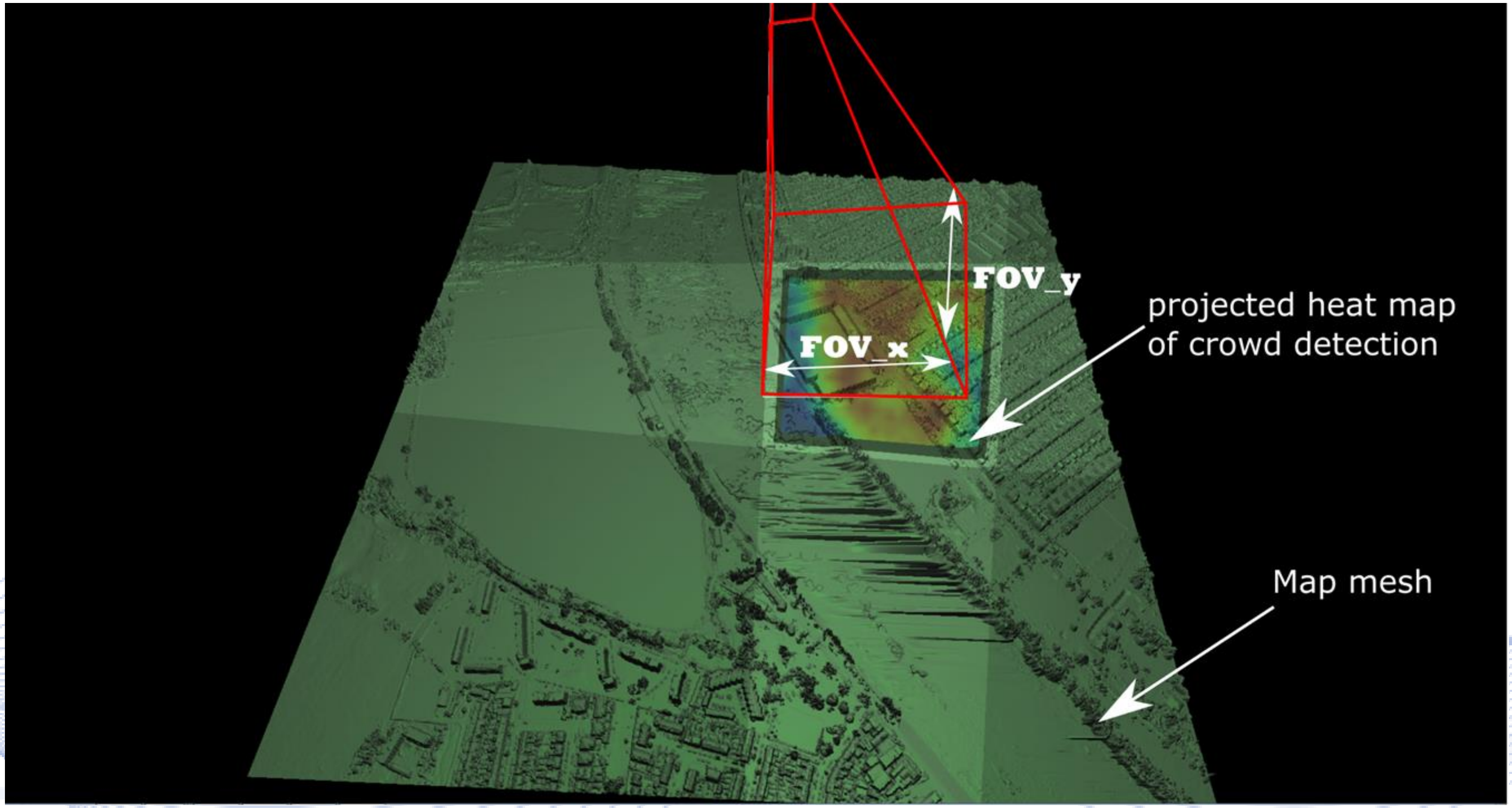
- Semantic mapping overlays semantic information on 2D or 3D scene maps.
  - These semantic entities are assigned specific spatial coordinates in a consistent manner and overlay a geometric 3D scene map.
  - The goal is cognitive comprehension of the outdoors environment where a robot moves and operates.

Semantic 3D Map Annotation for crowd localization



Artificial Intelligence & Information Analysis Lab

Semantic 3D Map Annotation for crowd localization

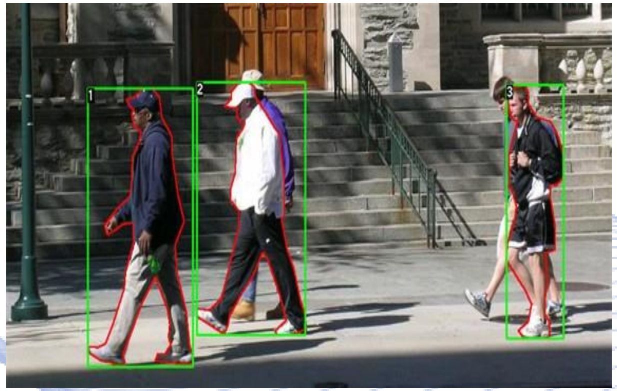


Computer vision overview

- Image and video acquisition
- Camera geometry
- Stereo and Multiview imaging
- Shape from X
- 3D Robot Localization and Mapping
- Semantic 3D world mapping
- **Object detection and tracking**
- 3D object localization
- Object pose estimation
- Computational cinematography

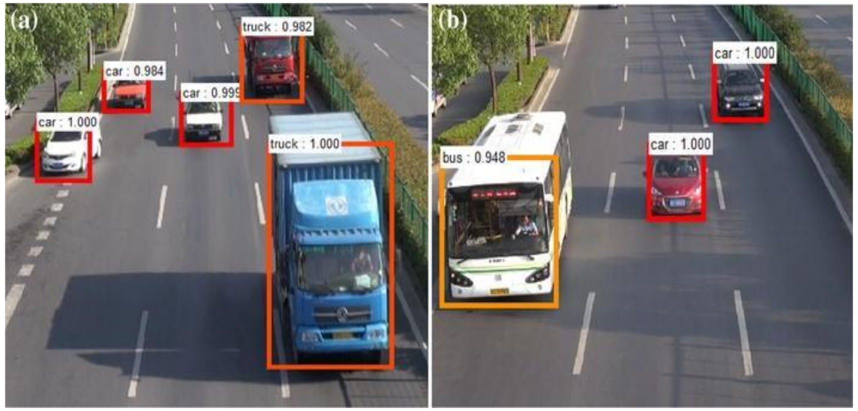
Object detection

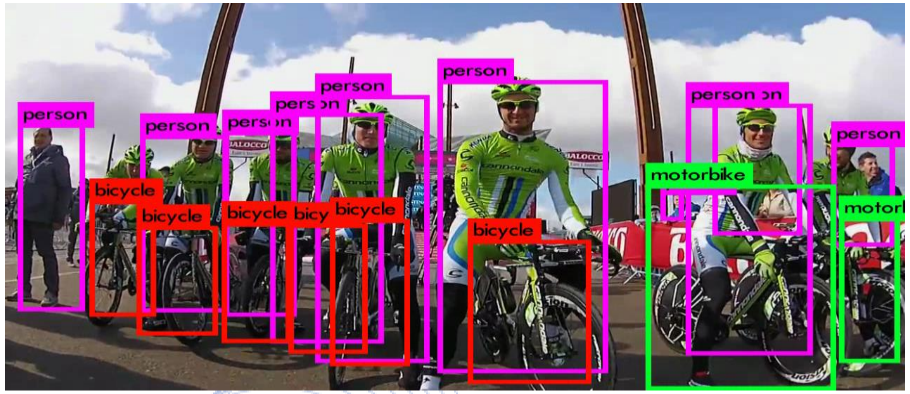
- Pedestrian, cars/vans/cyclist, road sign detection
- Current neural detectors are very capable of accurately detecting objects
- SSD, YOLO



Object detection

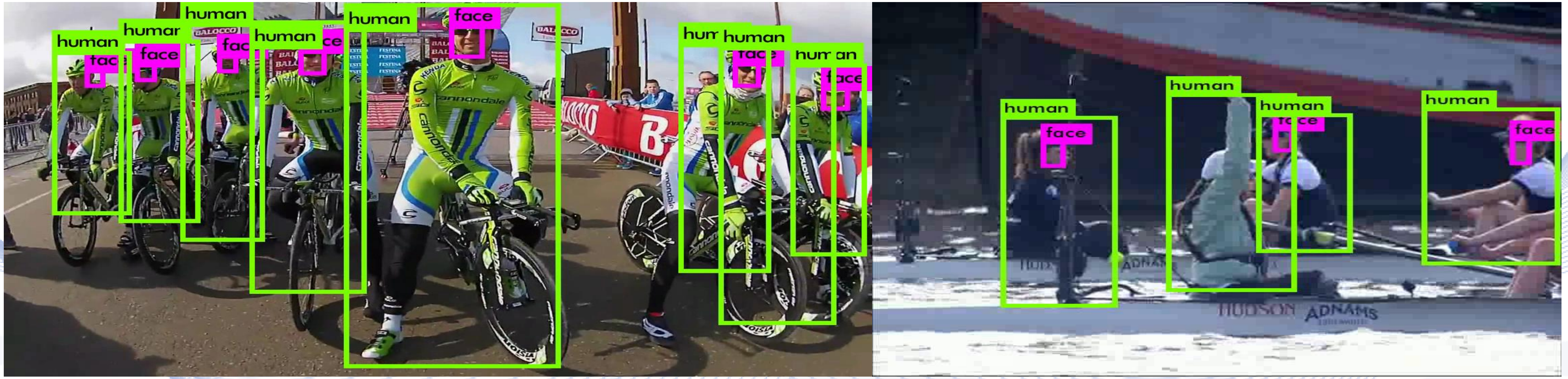
- **But** require domain-specific training or fine-tuning





Object detection

- Both can be trained when suitable annotations are available,
  - e.g., YOLO for face and human detection, trained on WIDER dataset



Object detection acceleration

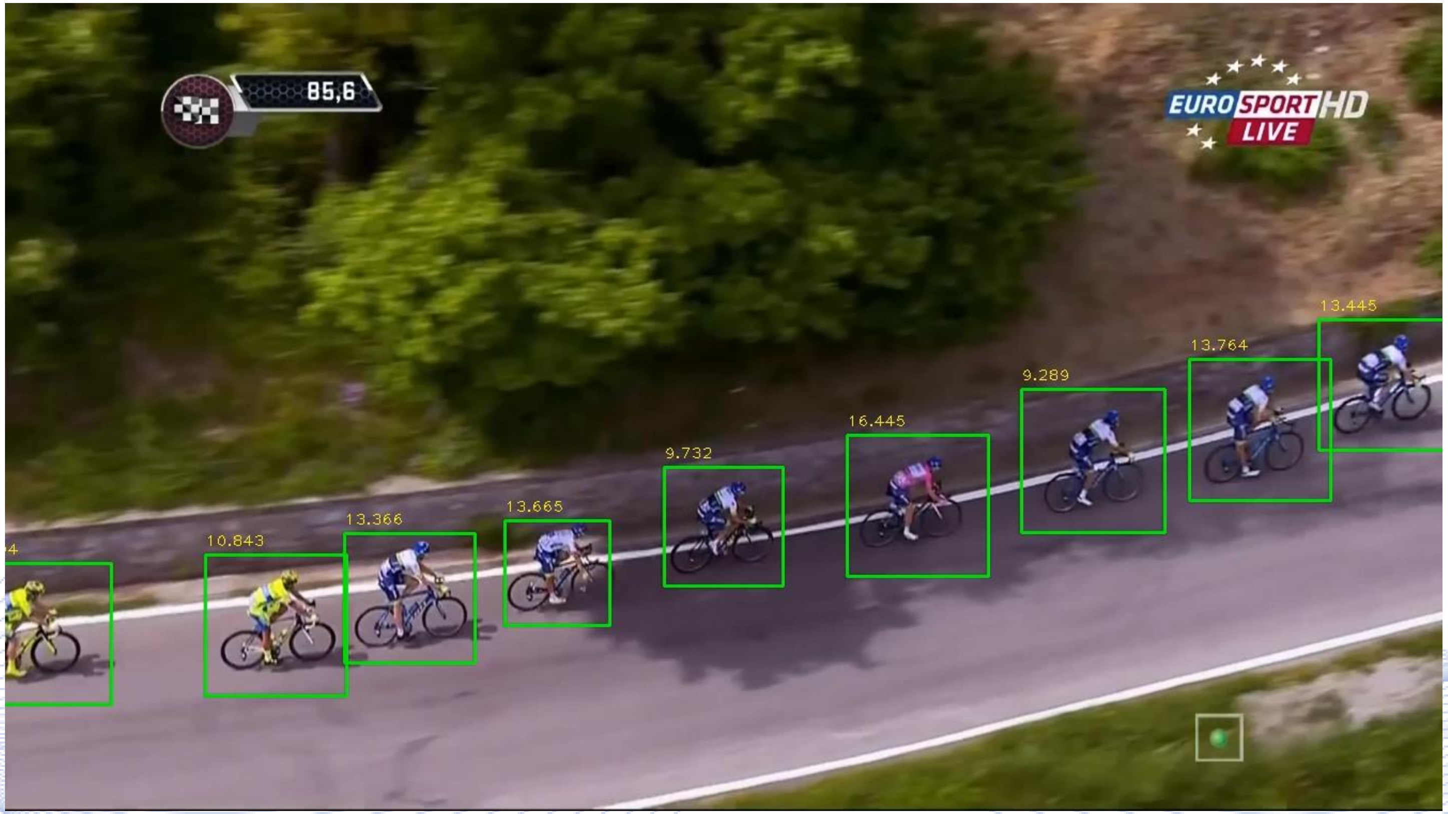
- Examples of acceleration techniques:
  - Input size reduction.
  - Specific object detection instead of multi-object detection.
  - Parameter reduction.
  - Post-training optimizations with TensorRT, including FP16 computations.

Object detection acceleration

- YOLO: good precision in general, but too heavyweight
  - small objects are more challenging to detect.
- Evaluation on VOC (Mean average precision, time):

Input Size	FPS	mAP	Forward time (ms) No TensorRT	Forward time (ms) TensorRT	Forward time (ms) FP16
608	2.9	71.26	241.5	128.8	69.3
544	3.2	73.64	214.4	121.2	64.3
480	5.4	74.50	155.4	62.3	35.7
416	6.4	73.38	155.3	56.5	32.5
352	7.8	71.33	111.0	45.0	24.3
320	8.5	70.02	103.0	40.4	22.8

UAV Object detection & tracking



Object Tracking specs for car vision

- 2D visual tracking will be employed for target following.
- Satisfactory performance in road footage is required.
- Target tracking should be performed in real-time, i.e., $> 25\ fps$.
- Embedded implementation is required and low computational complexity is preferred.
- Parallel or parallelizable methods (e.g., with CUDA implementations) should be preferred as well.
- Assuming 2D target tracking methods operate faster than combining target detection and recognition methods, long-term object tracking is also preferred.

Joint Detection & Tracking

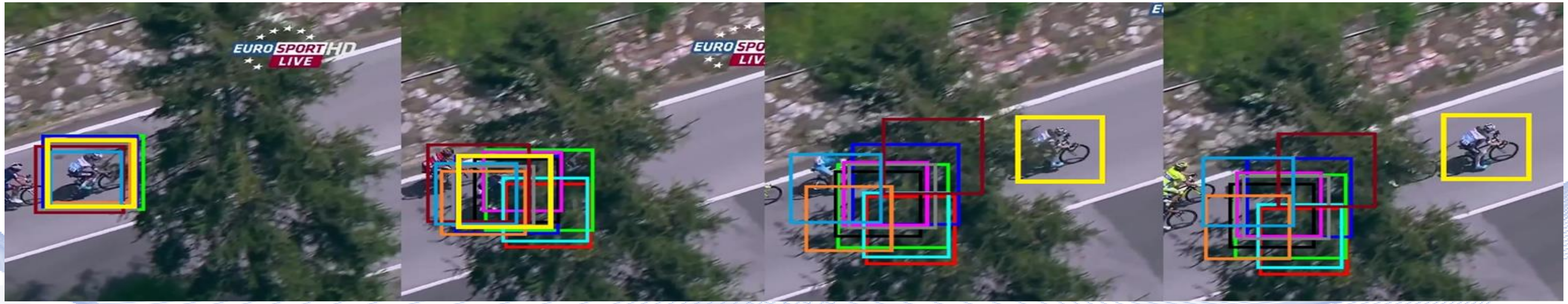
- **Tracker:** Given the initialized position of a target, the tracker T is responsible for estimating the bounding box of the target in the subsequent frames.

- **Detector/Verifier:** Given a bounding box defining the target in a specific frame produced by the tracker, the detector D is responsible for verifying this result, and then provide the appropriate feedback to the system. If the verification fails this module is responsible for detecting the target in a local search area and provide the correct bounding box to the master node M

- **Master**: M is responsible for the coordination of the two aforementioned modules. The node provides the necessary services to control the verification, the detection and the tracking tasks and controls the communication between the different parts of the system.

Joint Detection & Tracking

- Target re-initialization by the detector in hard tracking cases when tracking algorithms fail



Joint Detection & Tracking

- Target re-initialization by the detector in hard tracking cases when tracking algorithms fail

Multi-Target Tracking

- The implementation is extended to support the tracking of multiple targets while maintaining real-time performance

Multiview Object Detection and Tracking

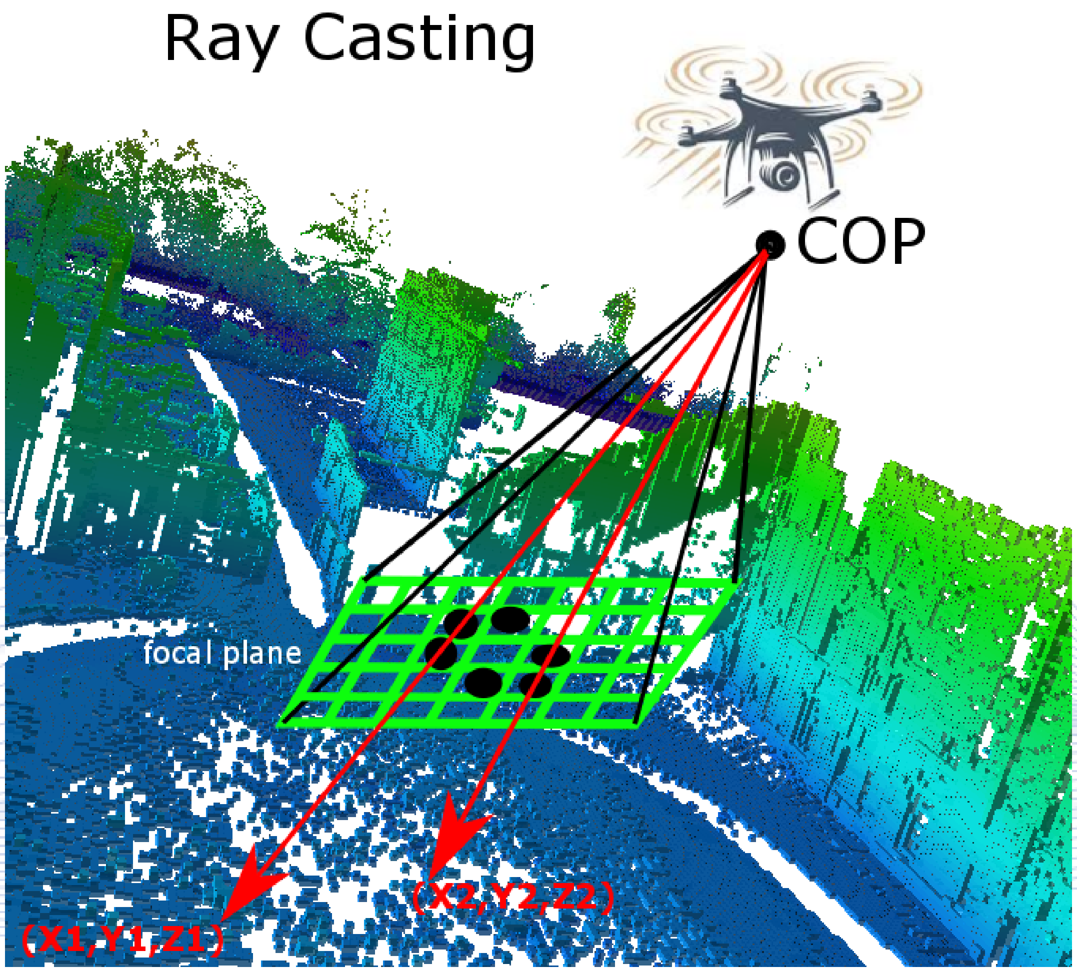
Multiview 3-UAV ORBIT

(a) Video frame from UAV 0. (b) Video frame from UAV 1. (c) Video frame from UAV 2.

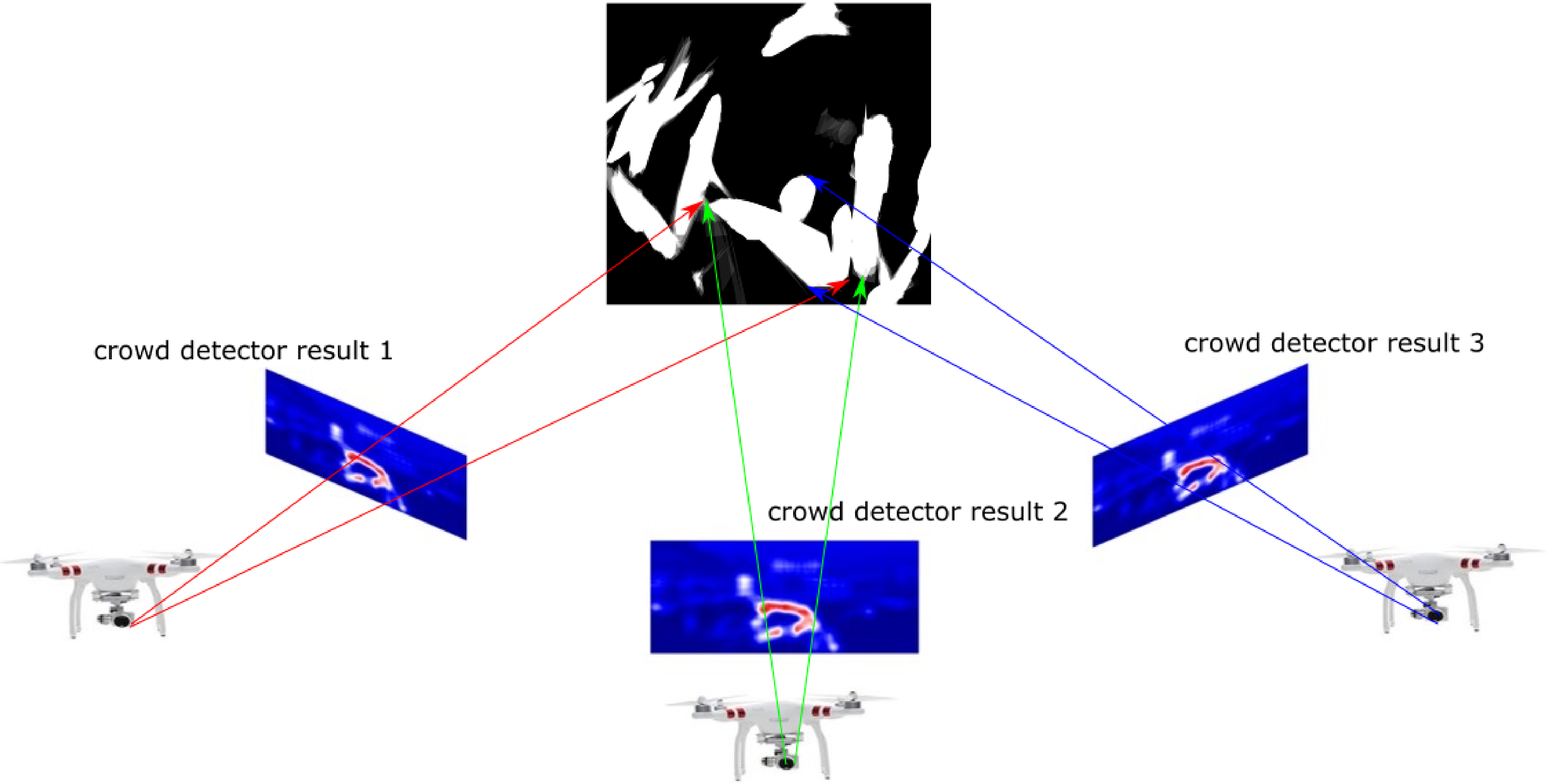
Computer vision overview

- Image and video acquisition
- Camera geometry
- Stereo and Multiview imaging
- Structure from X
- 3D Robot Localization and Mapping
- Semantic 3D world mapping
- Object detection and tracking
- **3D object localization**
- Object pose estimation
- Computational cinematography

3D object localization using 3D maps

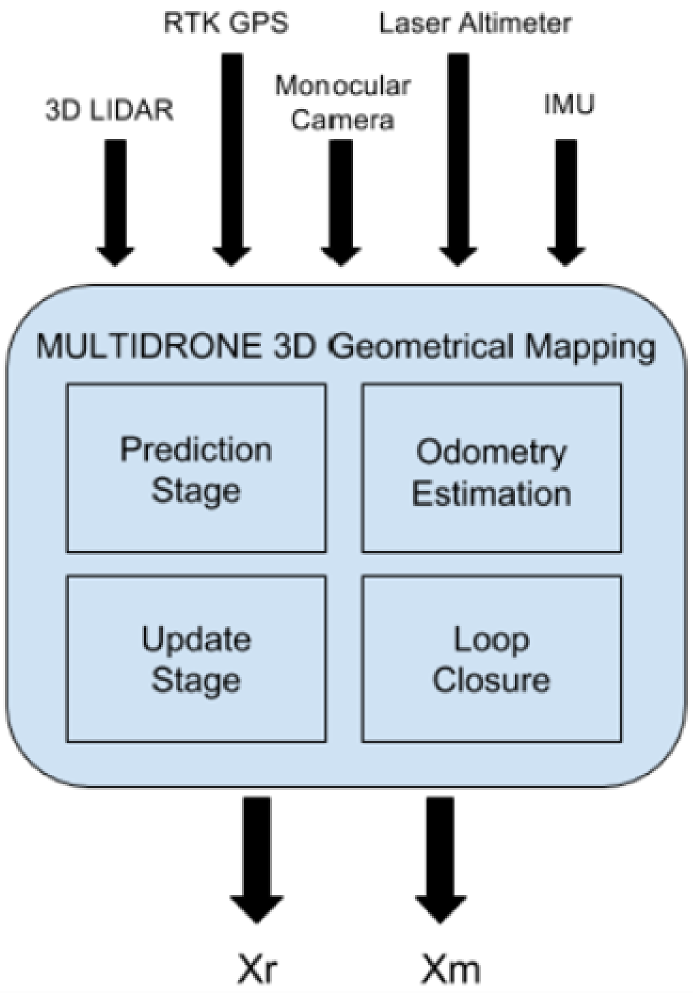


Multiview 3D Object localization



Sensor fusion

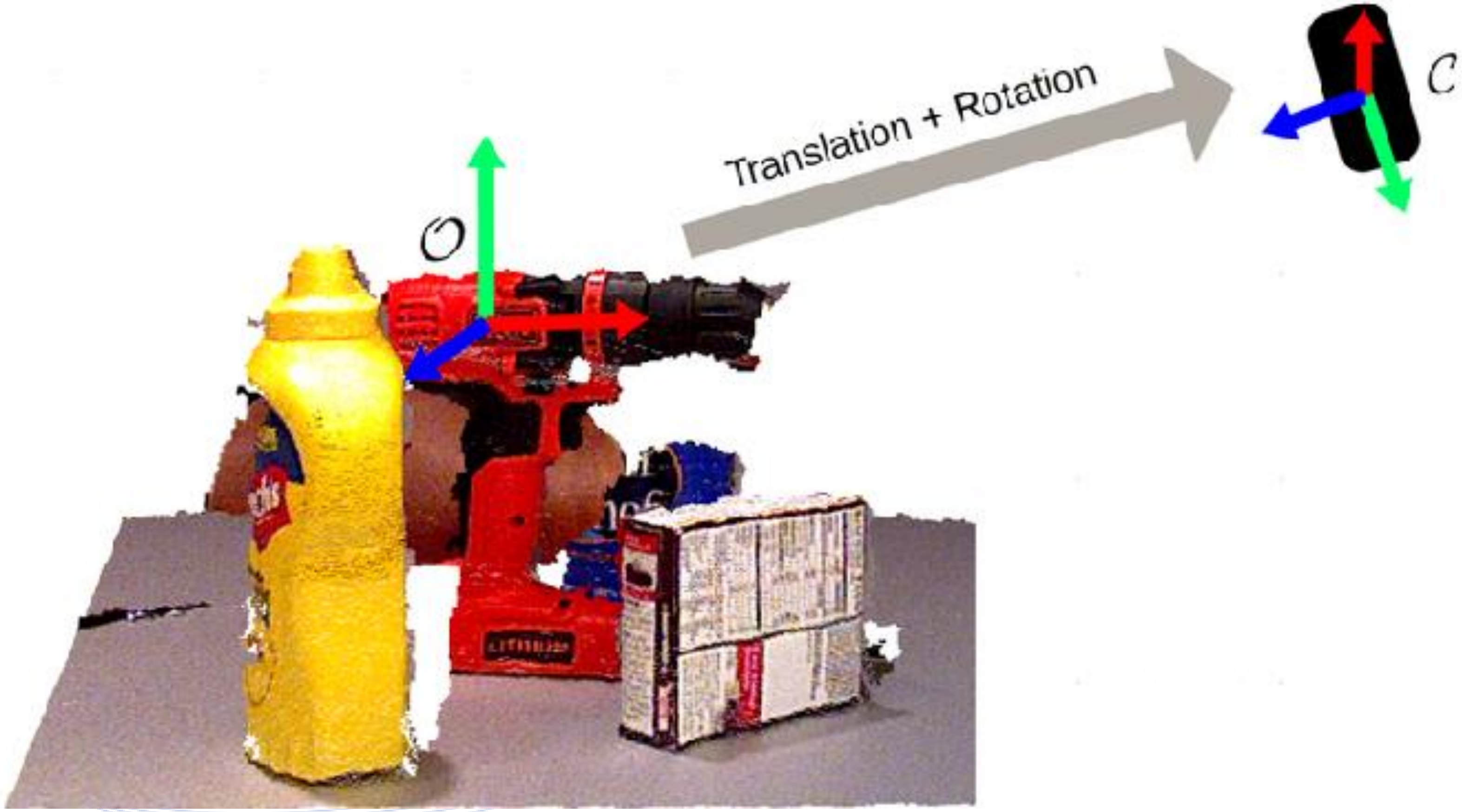
- On vehicle Sensors:
  - Lidar
  - Monocular camera
  - IMU
  - laser altimeter
  - RTK D-GPS
- Embedded processing:
  - Intel NUC NUC6i7KYK2 i7-6770HQ
  - Jetson TX2



Computer vision overview

- Image and video acquisition
- Camera geometry
- Stereo and Multiview imaging
- Structure from X
- 3D Robot Localization and Mapping
- Semantic 3D world mapping
- Object detection and tracking
- 3D object localization
- **Object pose estimation**
- Computational cinematography

6D object pose estimation



Target Pose Estimation

- **Computer Vision Approach**
  - Relies on detecting a set of *predefined points* (e.g., facial landmarks) and then using a method for solving the respective *Perspective-n-Point* (PnP) *problem*, i.e., estimation of the camera position with respect to the object.
- **Limitations:**
  - The 3-D coordinates for the landmark points must be known, i.e., a 3-D model of the object is needed
  - The landmarks points must be precisely tracked, i.e., the texture of the object must allow for setting enough discriminative landmarks

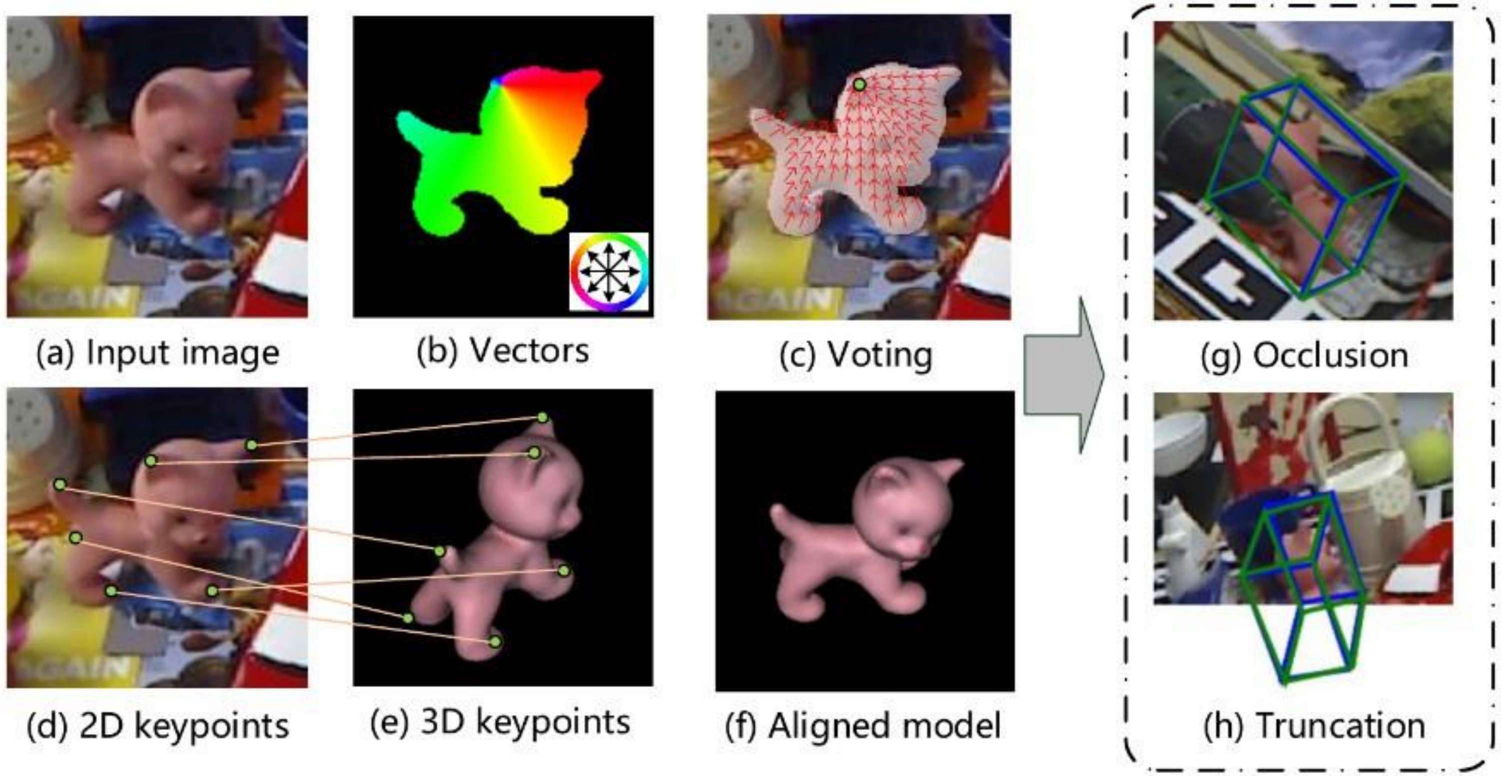
Target Pose Estimation

- **Machine Learning Approach**
  - A neural network receives the object and directly *regresses* its pose
  - Only a set of pose-annotated object pictures are needed
    - There is no need to manually develop 3-D models
    - The models are more robust to variations of the object for which we want to estimate its pose
    - The pose estimation can run entirely on GPU and (possibly) incorporated into a unified detection+pose estimation neural network
  - Very few pre-trained models are available
    - Models must be trained for the objects of interest (faces, bicycles, boats, etc.)

Target Pose Estimation

- **Machine Learning Approach**
  - We integrated a pre-trained yaw estimation model of facial pose (DeepGaze library) into the SSD-300 object detector (trained to detect human faces)
  - Varying illumination conditions seem to affect the estimation.

6D object pose estimation using Deep Learning



6D object pose estimation with 2D keypoint detection.

Posture estimation (Openpose)

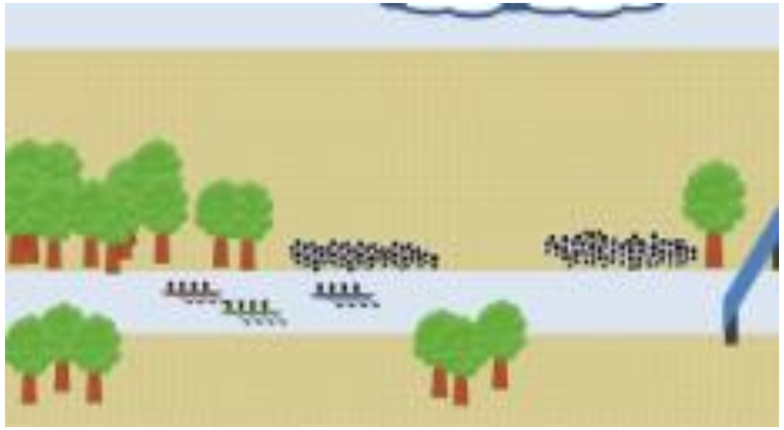
Posture estimation

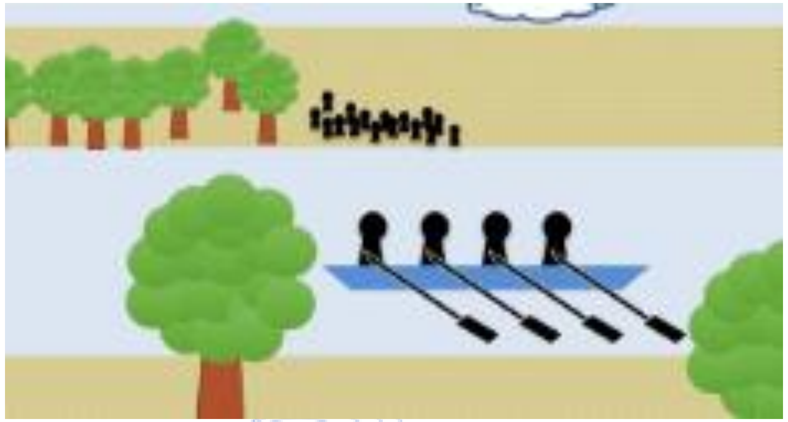
Computer vision overview

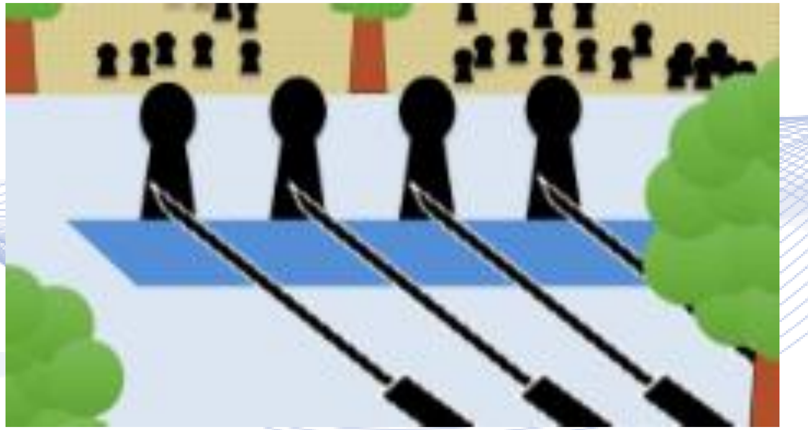
- Image and video acquisition
- Camera geometry
- Stereo and Multiview imaging
- Structure from X
- 3D Robot Localization and Mapping
- Semantic 3D world mapping
- Object detection and tracking
- 3D object localization
- Object pose estimation
- **Computational cinematography**

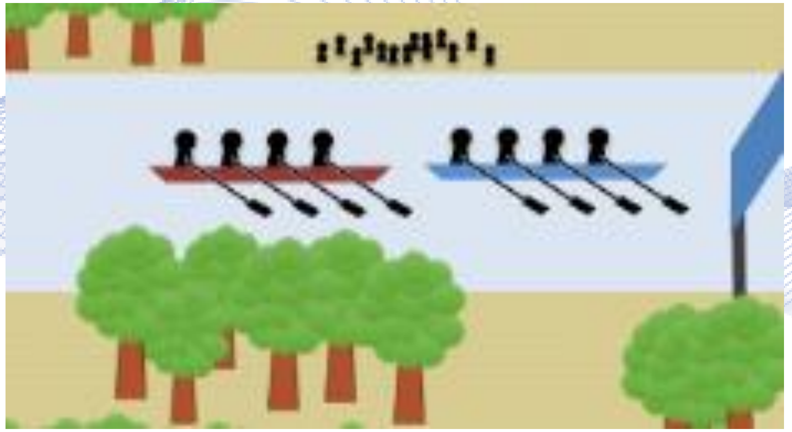
Framing Shot Types

- Example UAV shot types when shooting boat targets from the side.


Extreme Long Shot


Long Shot


Medium Close Up


Two Shot

Shot type constraints for intelligent UAV AV shooting

Determining the desired focal length to achieve specific shot types (constant distance between UAV and target)

Motion type	$\min f_{max}$	f_s when $c_s = 50\ \%$ (medium shot)	f_s when $c_s = 80\ \%$ (closeup)
LTS	$77.8\ mm$	$103.4\ mm$, not feasible	$150.7\ mm$, not feasible
CHASE	$162.9\ mm$	$103.4\ mm$, feasible	$150.7\ mm$, feasible
ORBIT	$128.8\ mm$	$103.4\ mm$, feasible	$150.7\ mm$, not feasible

A shot type is feasible if the $\mathbf{f}_{max} > \mathbf{f}_s$

Bibliography

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.
[SZE2011] R.Szelinski, “ Computer Vision ” , Springer 2011
[HAR2003] Hartley R, Zisserman A. , “ Multiple view geometry in computer vision” . Cambridge University Press; 2003.
[DAV2017] Davies, E. Roy. “Computer vision: principles, algorithms, applications, learning ”. Academic Press, 2017
[TRU1998] Trucco E, Verri A. “Introductory techniques for 3-D computer vision”, Prentice Hall, 1998.
[PIT2017] I. Pitas, “Digital video processing and analysis ” , China Machine Press, 2017 (in Chinese).
[PIT2013] I. Pitas, “Digital Video and Television ” , Createspace/Amazon, 2013.
[PIT2000] I. Pitas, Digital Image Processing Algorithms and Applications, J. Wiley, 2000.
[NIK2000] N. Nikolaidis and I. Pitas, 3D Image Processing Algorithms, J. Wiley, 2000.
[APOLLO] http://apolloscape.auto/

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr