Technischen Universität München
Winter Semester 2015/2016

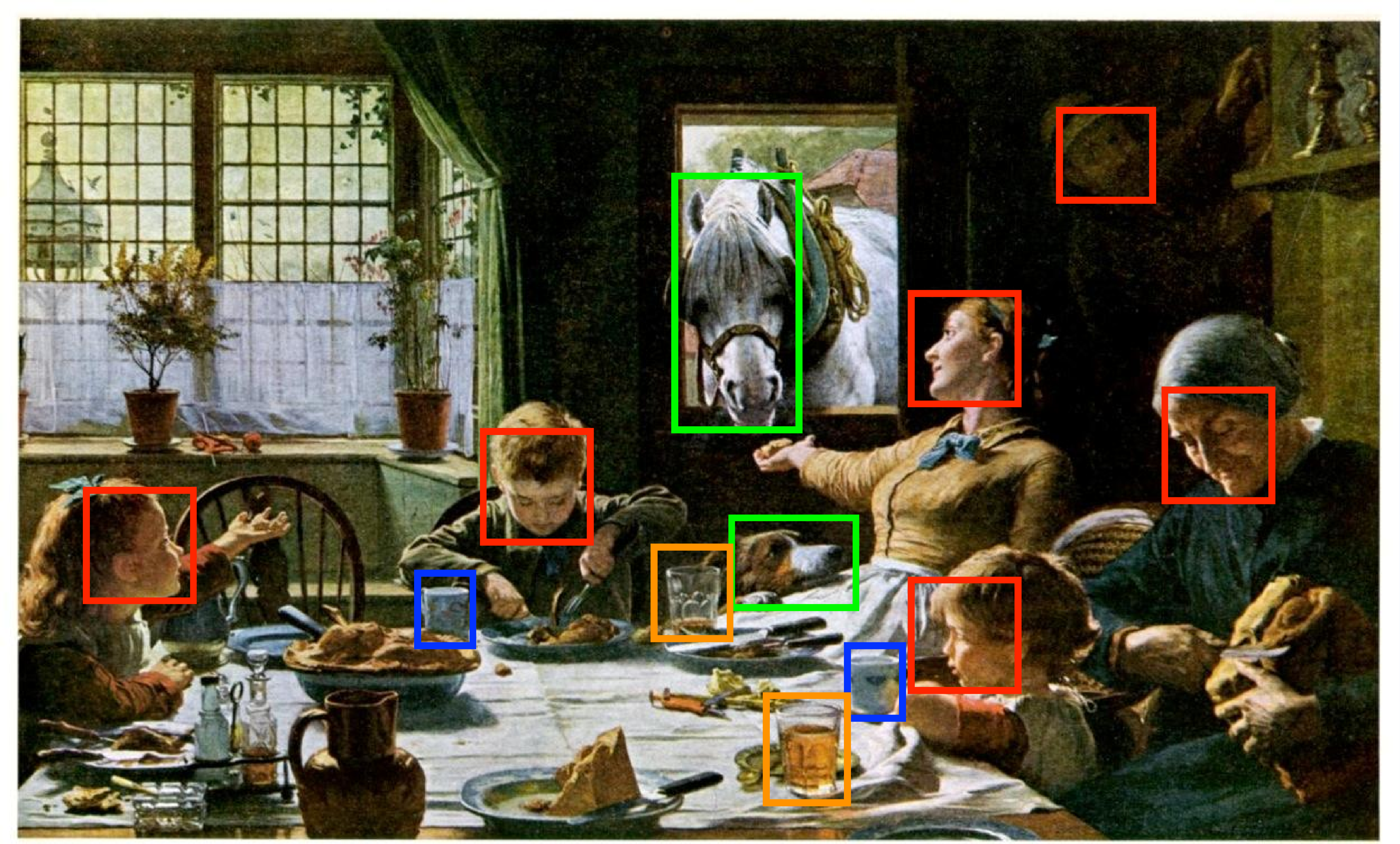
TRACKING and DETECTION in COMPUTER VISION

Slobodan Ilić

Motivation

“One of the Family”, Frederick Cotman, 1880

Computer can recognize objects in images



“One of the Family”, Frederick Cotman, 1880

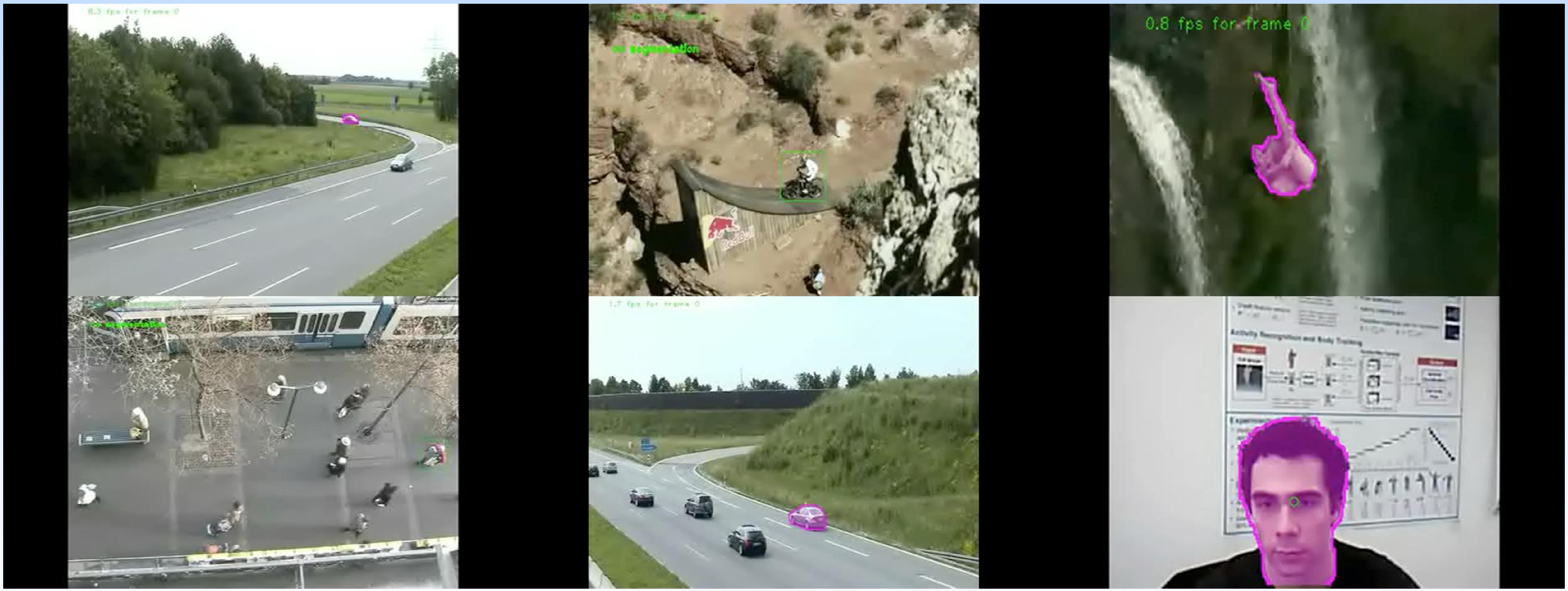
Computer can potentially recognize activities

"One of the Family", Frederick Cotman, 1880

Motivation

What about videos?

Computer can track objects in images



What is Computer Vision?

Human Vision (eyes and the visual cortex in the brain) discovers from images what object are present in the scene, where they are, how they move and what is their shape.

Computer Vision (using cameras attached to the computers) automatically interprets images trying to understand their content similar to the human vision.

What is **not** Computer Vision?

Image Processing - Takes an image and process is to produce new, more desirable image. Image enhancement, image compression, image restoration.

Pattern Recognition - Takes a pattern and classifies it into one of predefined, finite set of classes.

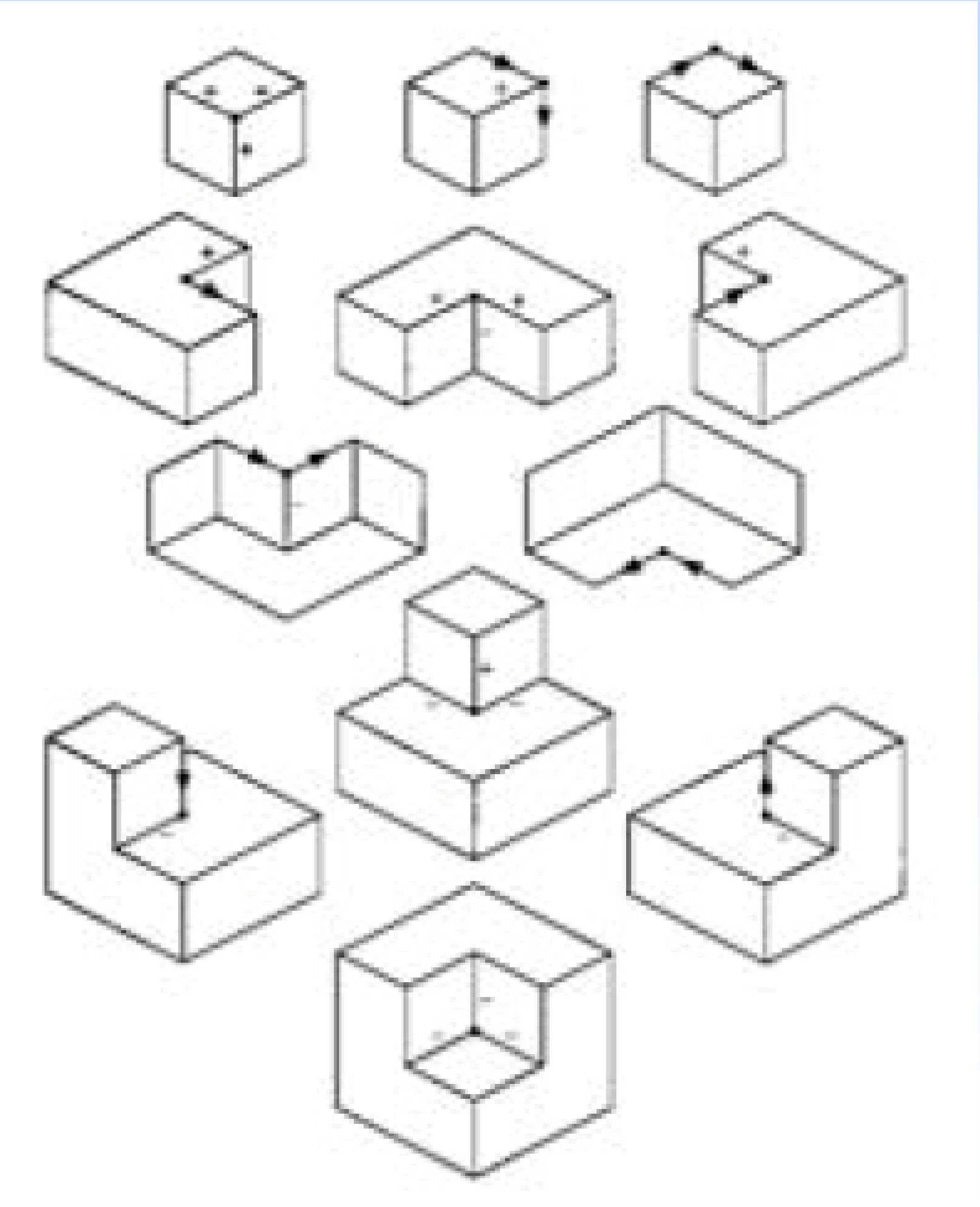
Computer Graphics - Synthesize images using powerful algorithms so that they correspond as close as possible to the real images.

Machine Learning - Hmm... Nowadays more and more learning is part of Computer Vision algorithms.

How did everything start?

Computer Vision started as a semester project at MIT in 1965.

The assumptions were very strong (block world) and the data were perfect, so it seemed to researchers to be an easy task.

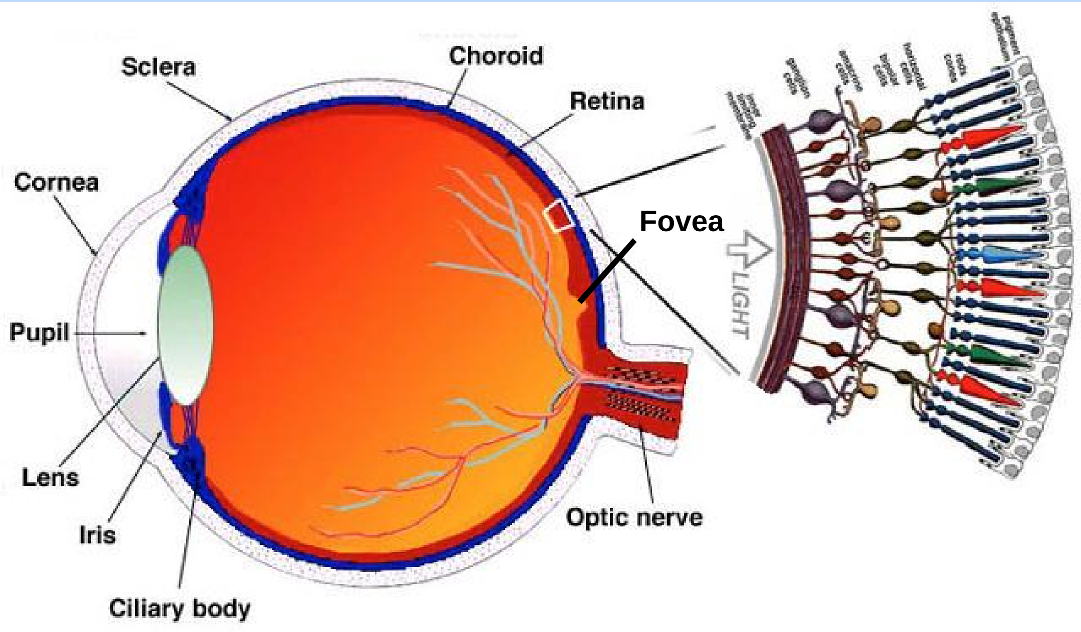


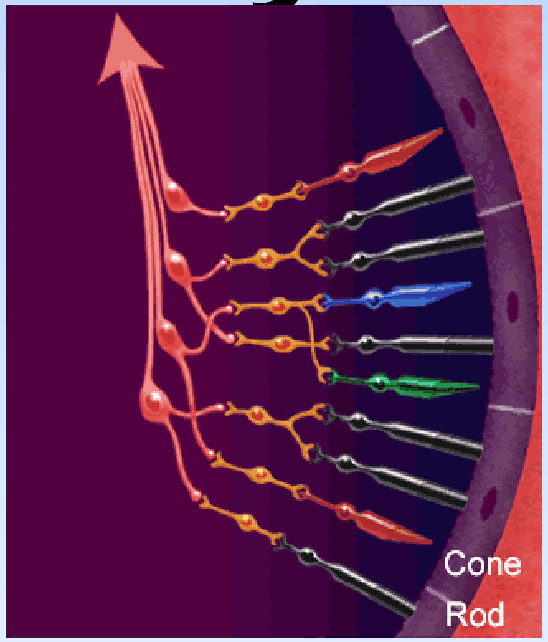
However the wold is not perfect !

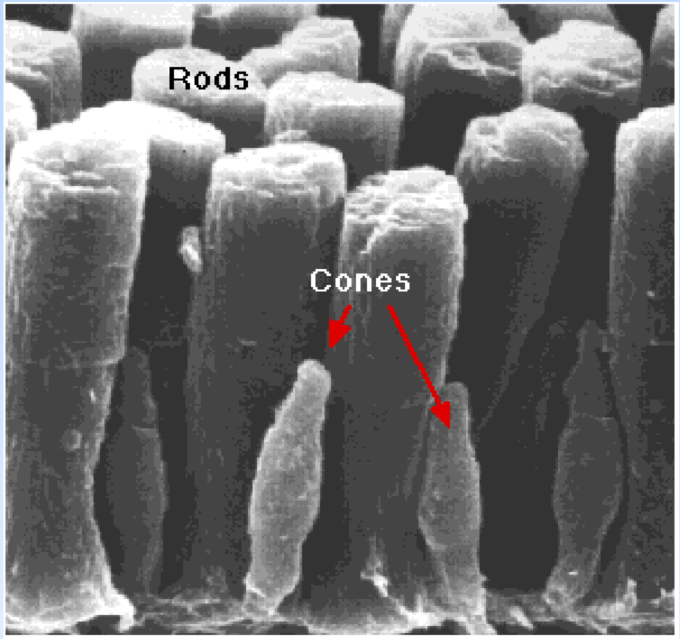
Why to study Computer Vision?

- Intellectual curiosity -- try to mimic the most powerful human sense
- Nowadays a number of industrial applications exist:
  - automation of industrial processes
  - medicine, diagnostics
  - entertainment: film and video games
  - security and surveillance
  - visualization and augmented reality
  - communication
  - human computer interactions (HCI)
  - military and space research

The eye

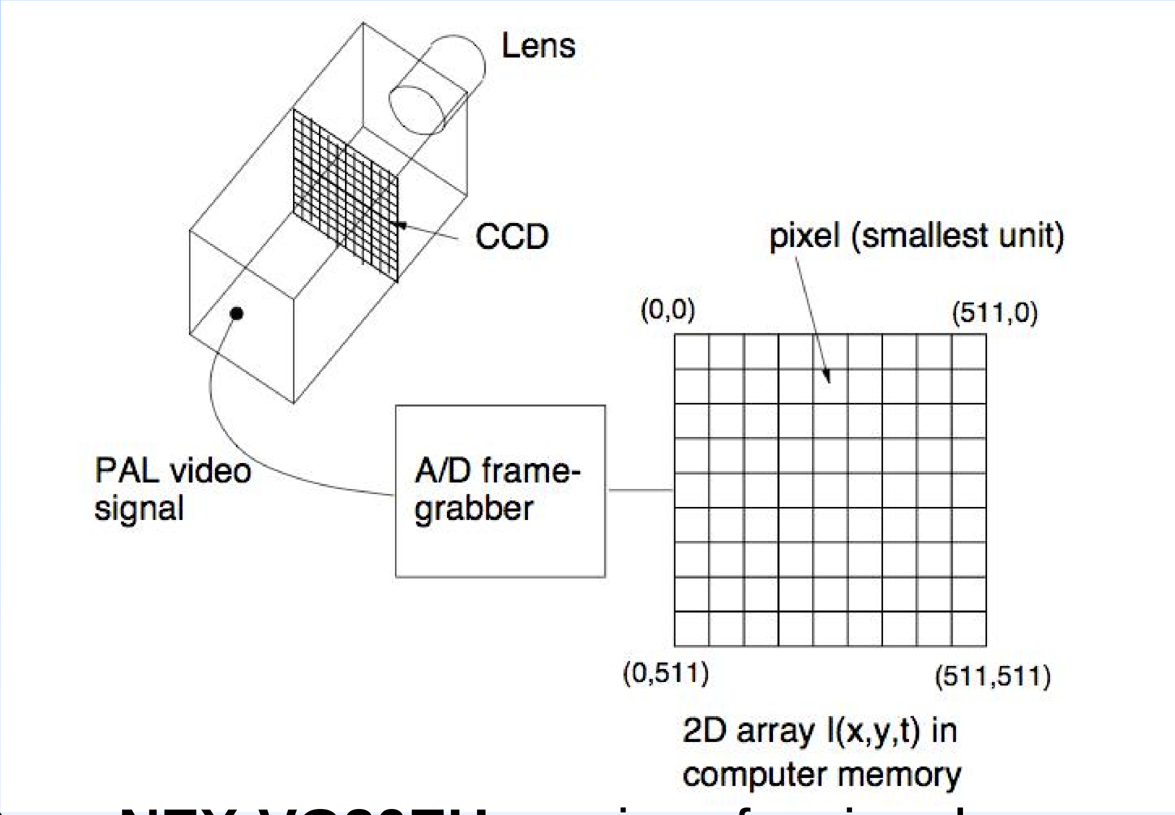






- Retina measures about 5 × 5 cm and contains 10^8 sampling elements (rods: sense brightness, low intensity, e.g night vision and cones: sense color with higher intensity light).
- The eye’s spatial resolution is about 0.01degree over a 150 degree field of view (not evenly spaced, there is a fovea and a peripheral region).
- Intensity resolution is about 11 bits/element, spectral resolution is about 2 bits/element (400–700nm).
- Temporal resolution is about 100 ms (10 Hz).
- Two eyes give a data rate of about **3 GBytes/s**!
- A large chunk of our brain is dedicated to processing the signals from our eyes.

The camera



- For example, Sony **NEX-VG20EH** semi-professional camera has HD resolution of around
~2 Mpixels.
- Intensity resolution is 24bits/pixel (RGB).
- Most computer vision applications work with monochrome images.
- Temporal resolution is about 40-20ms (25-50 Hz), SNR is about 50dB(Pulnix camera spec.).
- One HD camera at 50Hz gives a raw data rate of about **300MBytes/s** (color), i.e **100MBytes/s** (mono)

Should we copy biology?

- No! Human vision is a product of millions of years of the evolution created under different constraints.
- It consists of 60 billion neurons heavily interconnected.
- Computes we have today cannot perform like a human brain.

We really do not understand how the brain works!

We need to try understand underlying principles rather then the particular implementation.

Is the Deep Learning answer?

- Deep Learning is reincarnation of Neural Networks

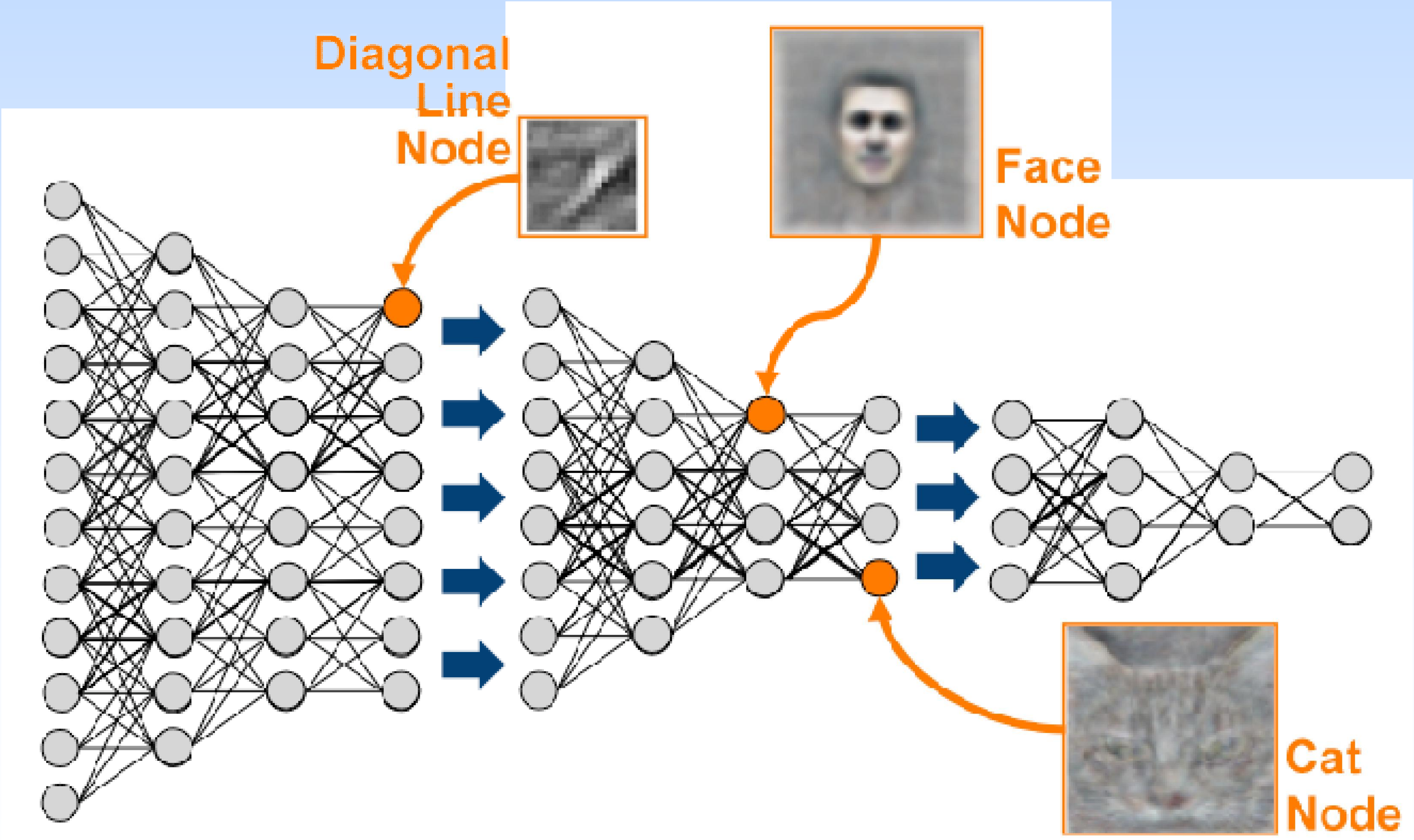
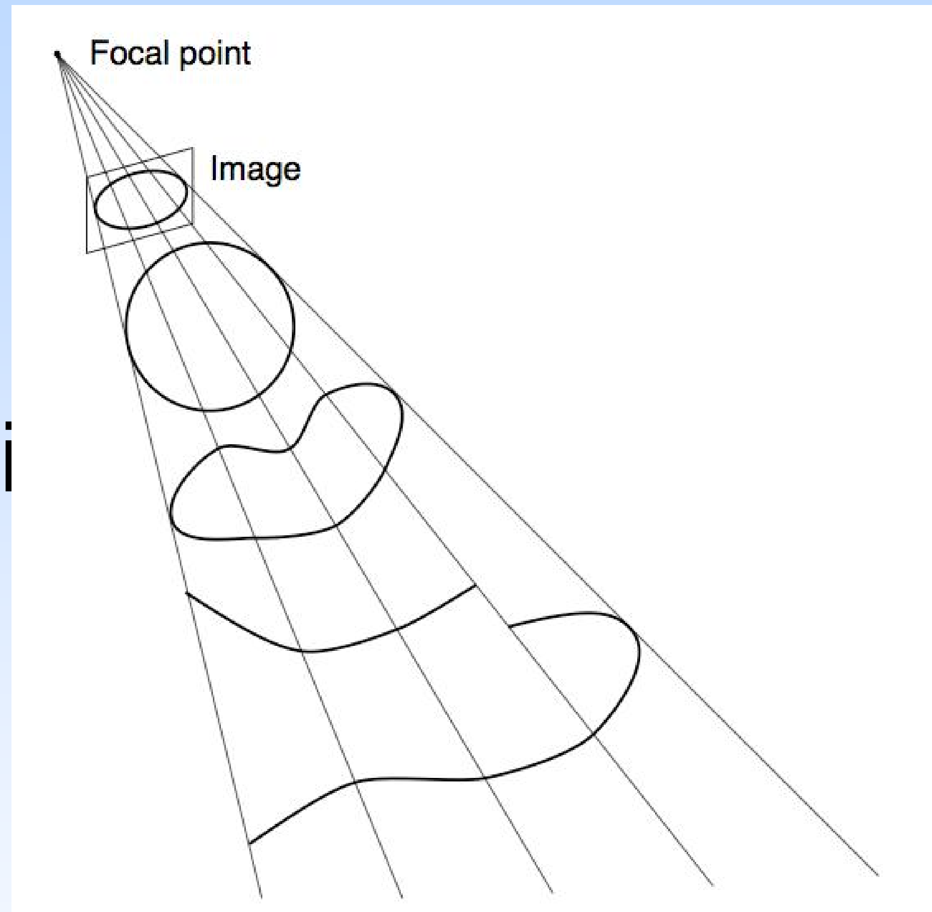


Image ambiguities

Gi



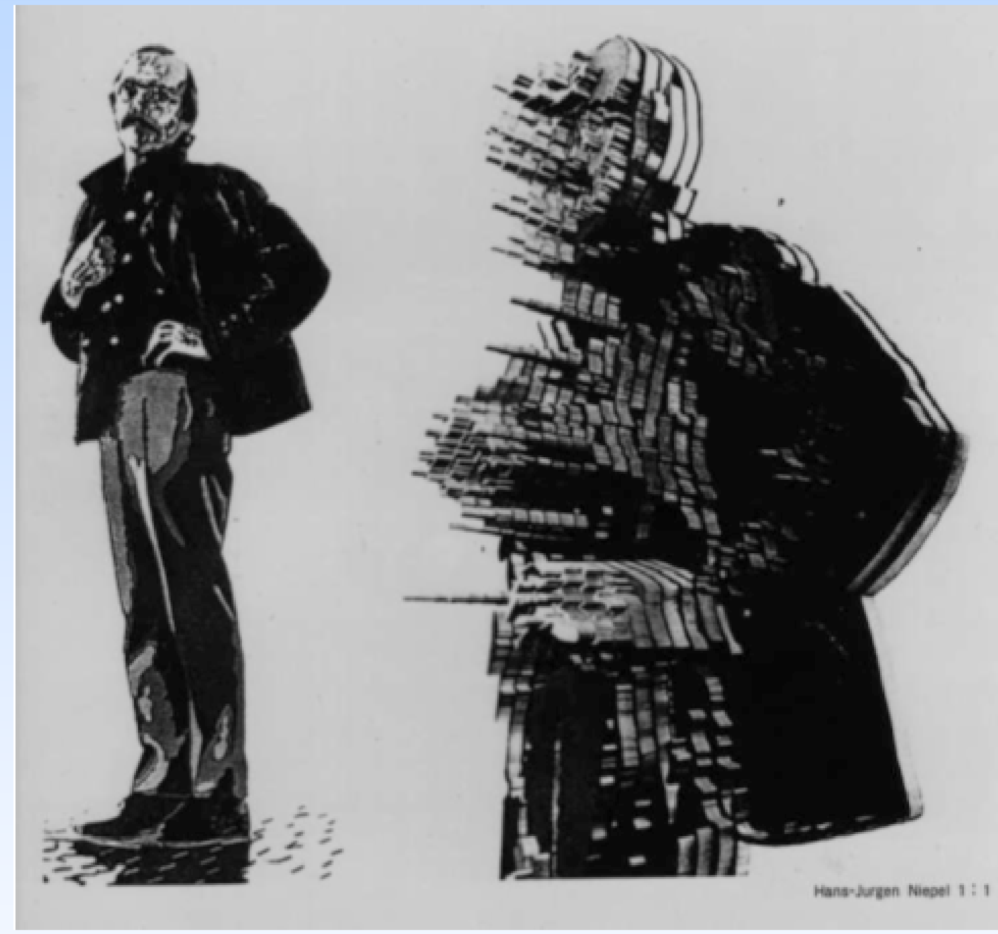


Image formation is many to one mapping. It is simple projection of the 3D object representation and does not say anything about the depth.

What should we do?

“One cannot understand what seeing is and how it works unless one understands the underlying information processing task being solved”

David Marr

The imaging process is ambiguous and we should try to resolve the ambiguities by introducing constraints to our problem like:

- use more then one image of the scene
- make assumptions about the world in the scene
- introduce knowledge about the observed problem

Computer Vision Tasks

Are equivalent to those of Human Vision:

- discover from images what object are present in the scene ⟶ **object recognition/classification**
- where they are ⟶ **object detection**
- how they move and ⟶ **object tracking**
- what is their shape ⟶ **object reconstruction**

Object Recognition/Classification

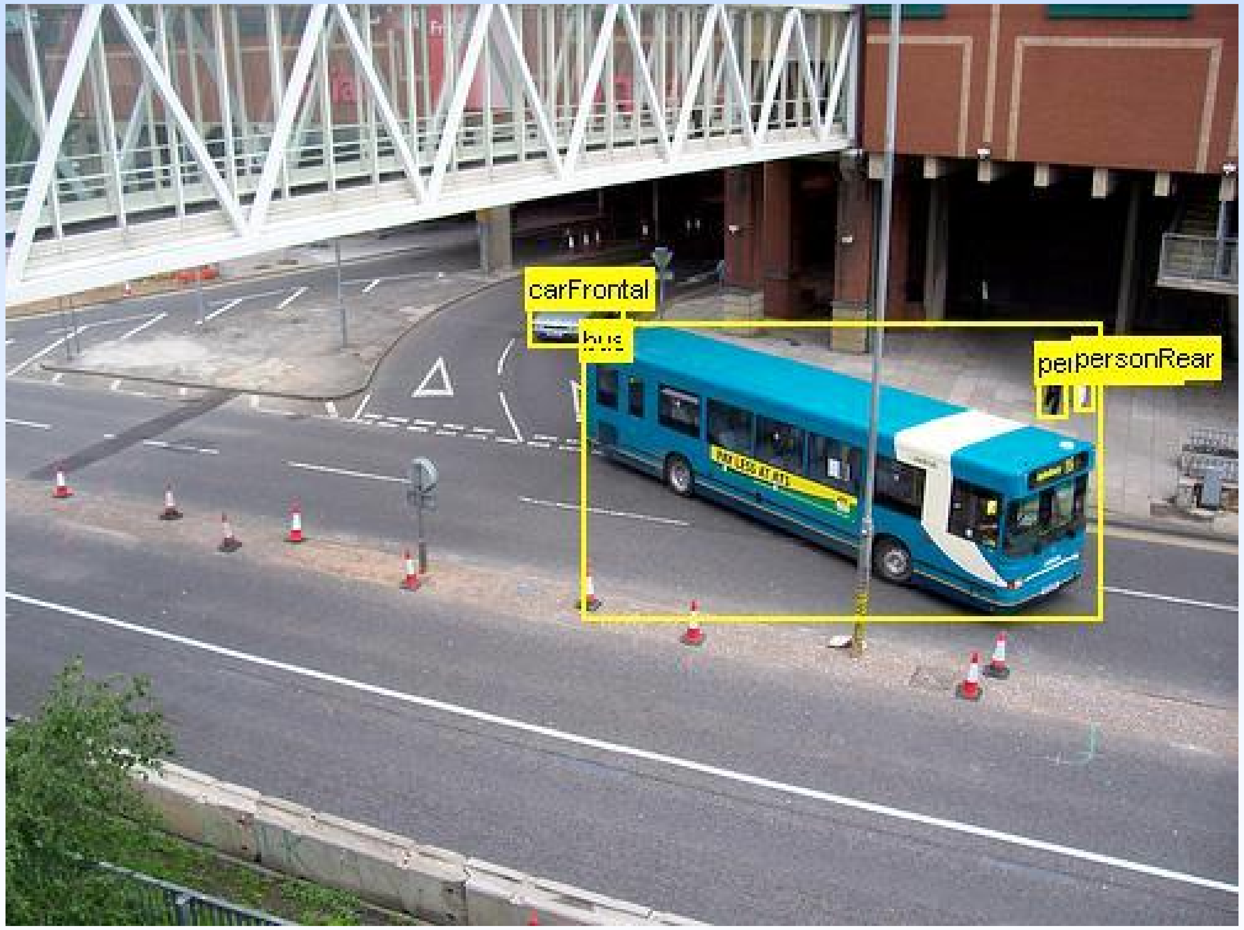
Discover from images what object are present in the scene.

Bicycles

Horse

Object Detection

Where are the objects in images?



From Pascal VOC2012 Challenge

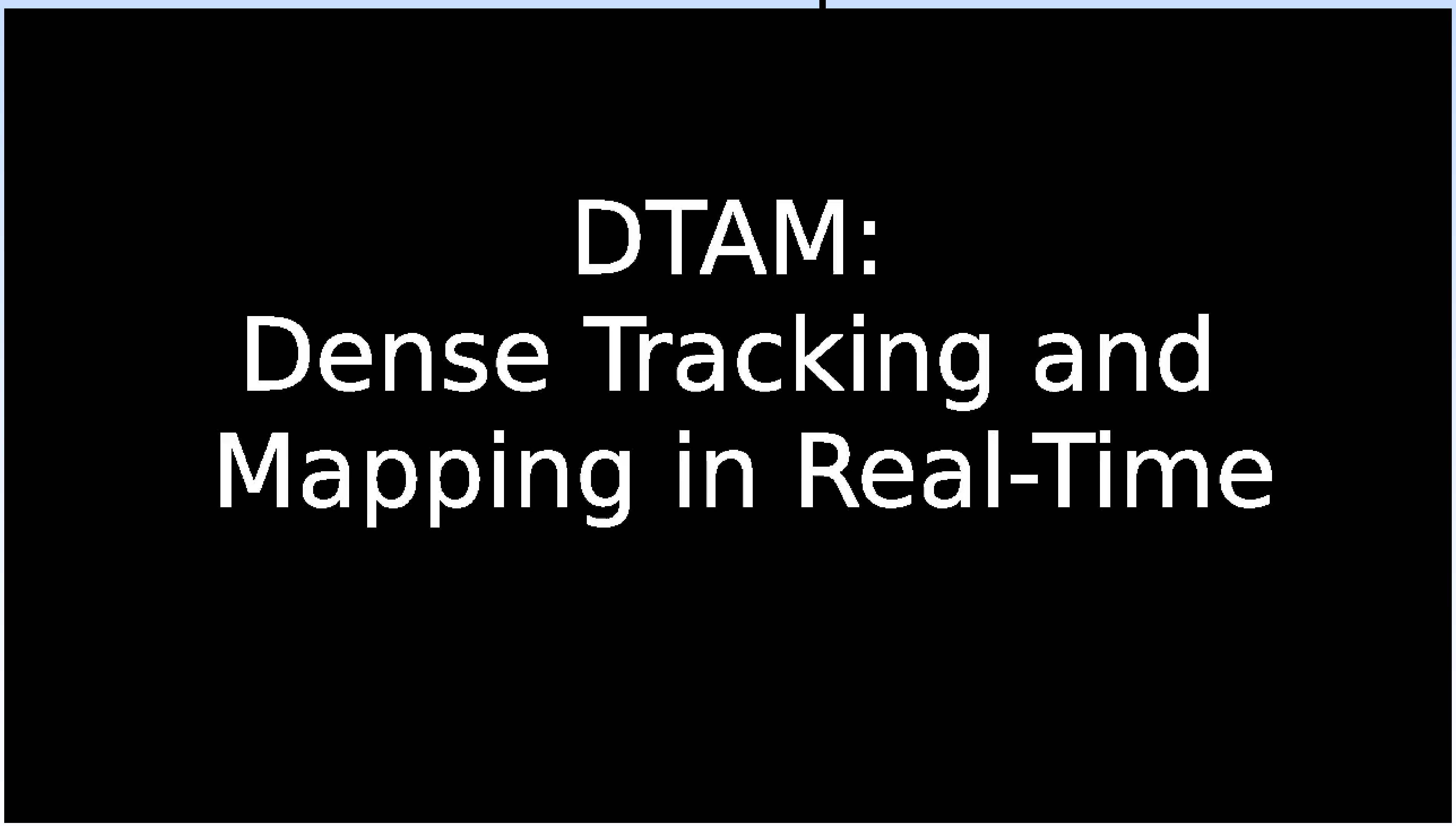
Object Tracking

How the object moves in the images/video?

From V. Belagianis et al. ECCV 2012

Object Reconstruction

How the objects or the scene look like, e.g what is their 3D shape?



From Newcombe et al. ICCV201

What is tracking?

- Tracking means following one or multiple objects or of interest in the scene providing continuously their position.
- Tracking algorithm estimates parameters of the dynamic system, e.g. feature point positions, object position, human joint angles etc.
- The source of information is video from one or multiple cameras

Tracking algorithms

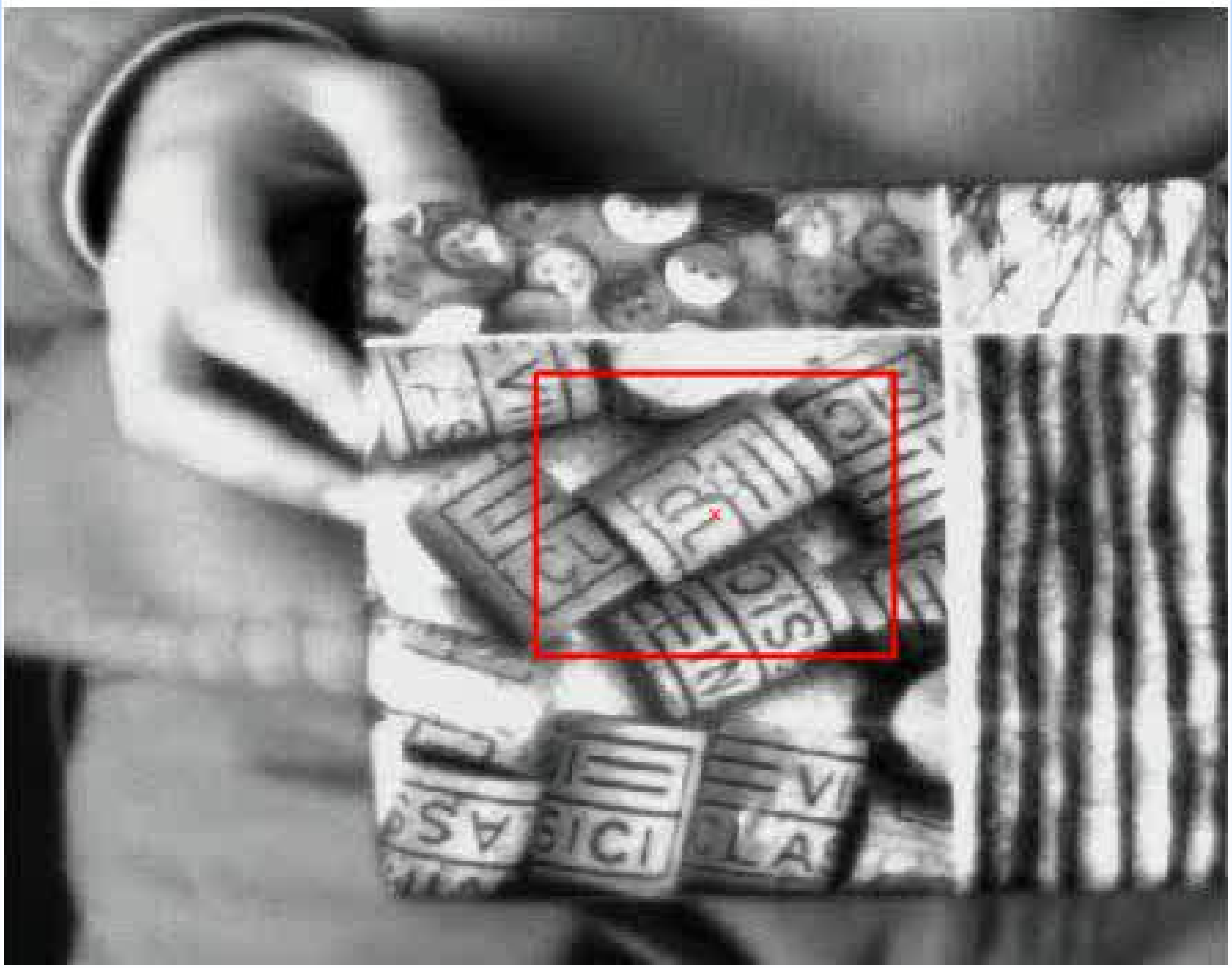
- 2D object tracking
  - follow the object in images and return an object trajectory in image coordinates
- 3D object tracking
  - follow the object in 3D and return its pose (position[translation] and orientation[rotation])
  - this is equivalent to the camera tracking

2D Object Tracking

- Template tracking
  - Lucas-Kanade;
  - Compositional Alg., Inverse Compositional;
  - ESM;
  - Learning a linear predictor;
  - Active Shape and Active Appearance Models;

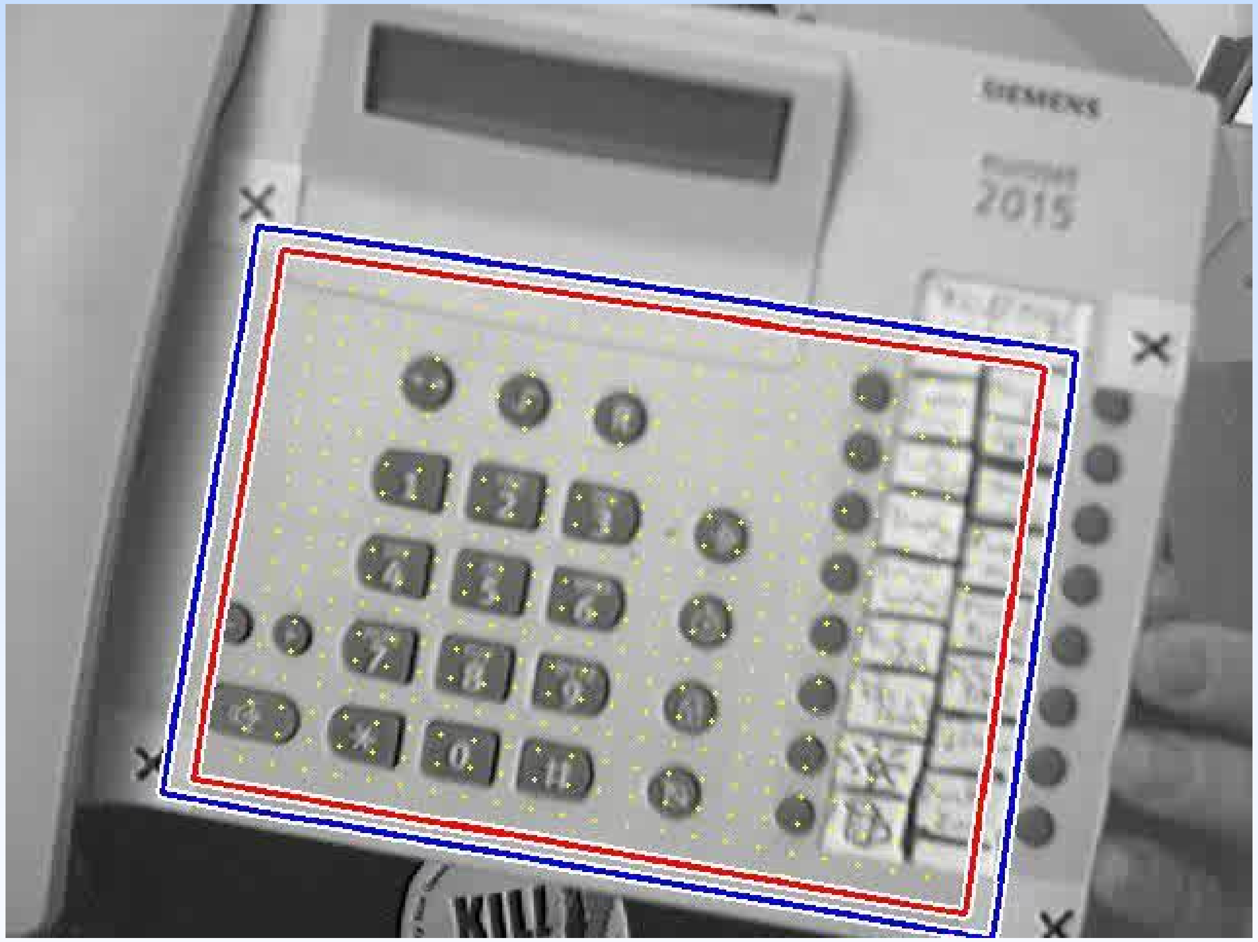
- Mean-Shift tracking
- Kalman filtering
- Particle filtering

Template Tracking



S. Benhimane, E. Malis, Real-time image-based tracking of planes using efficient second-order minimization, IEEE/RSJ International Conference on Intelligent Robots Systems, Sendai, Japan, 2004.

Need for templates of varying size



Template extension Template reduction

S. Holzer, **S. Ilic**, N.Navab, **Adaptive Linear Predictors for Real-Time Tracking**, **CVPR 2010**

AAM based face tracking

2D face tracking using AAM, courtesy of Robotics Institute CMU

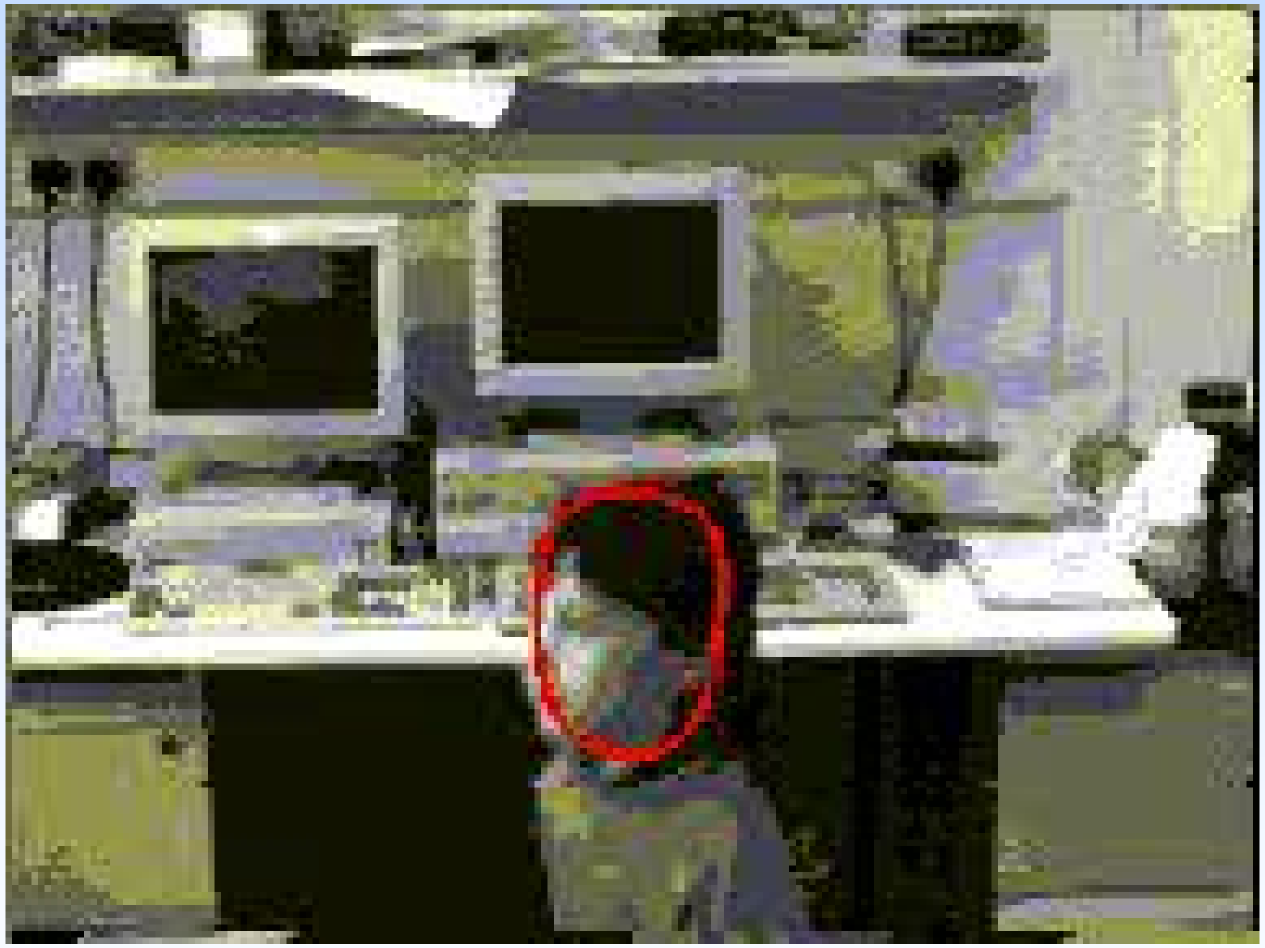
Mean-Shift Tracking

Dorin Comaniciu and Peter Meer, Mean Shift : A Robust approach towards feature space analysis, PAMI, 2002.

Mean-Shift Tracking

V. Nedović, "Tracking moving video objects using mean-shift algorithm", *Unpublished report, University of Amsterdam*, 2004.

Tracking 2D




Tracking in 2D, CONDENSATION alg., M. Isard, A. Blake

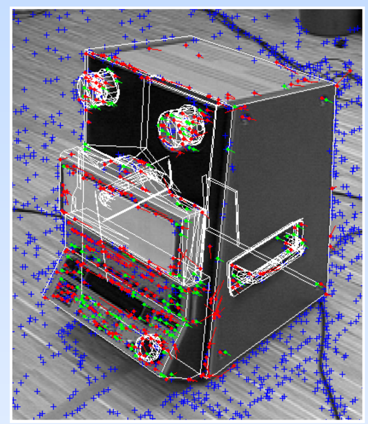
Approaches to 3D Object/Camera Pose Tracking

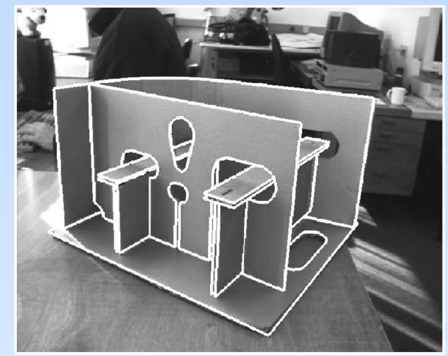
SLAM/SFM

Consider natural features

No *a priori* 3D knowledge

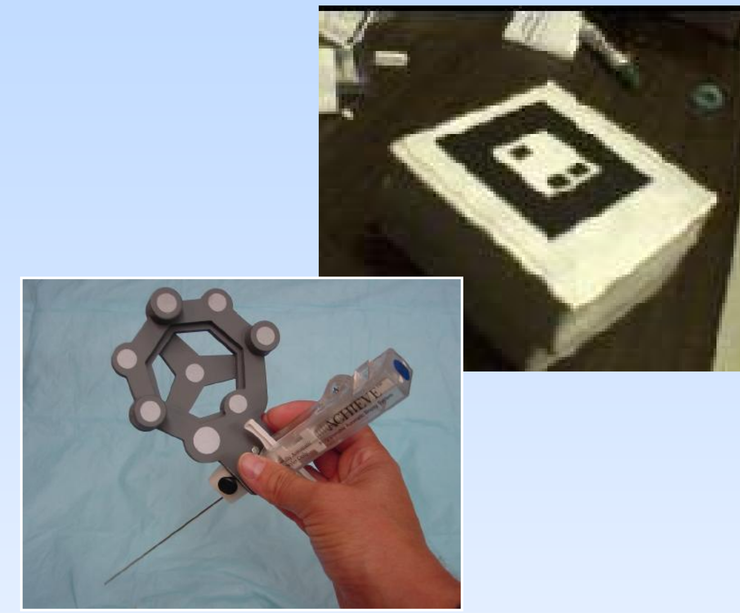
Model-based





Consider natural features

Use some 3D knowledge



Make use of visual markers

Use some 3D knowledge

What will we learn?

3D Object/Camera Tracking

Feature extraction/description

Matching/Tracking features

Object/Camera Pose

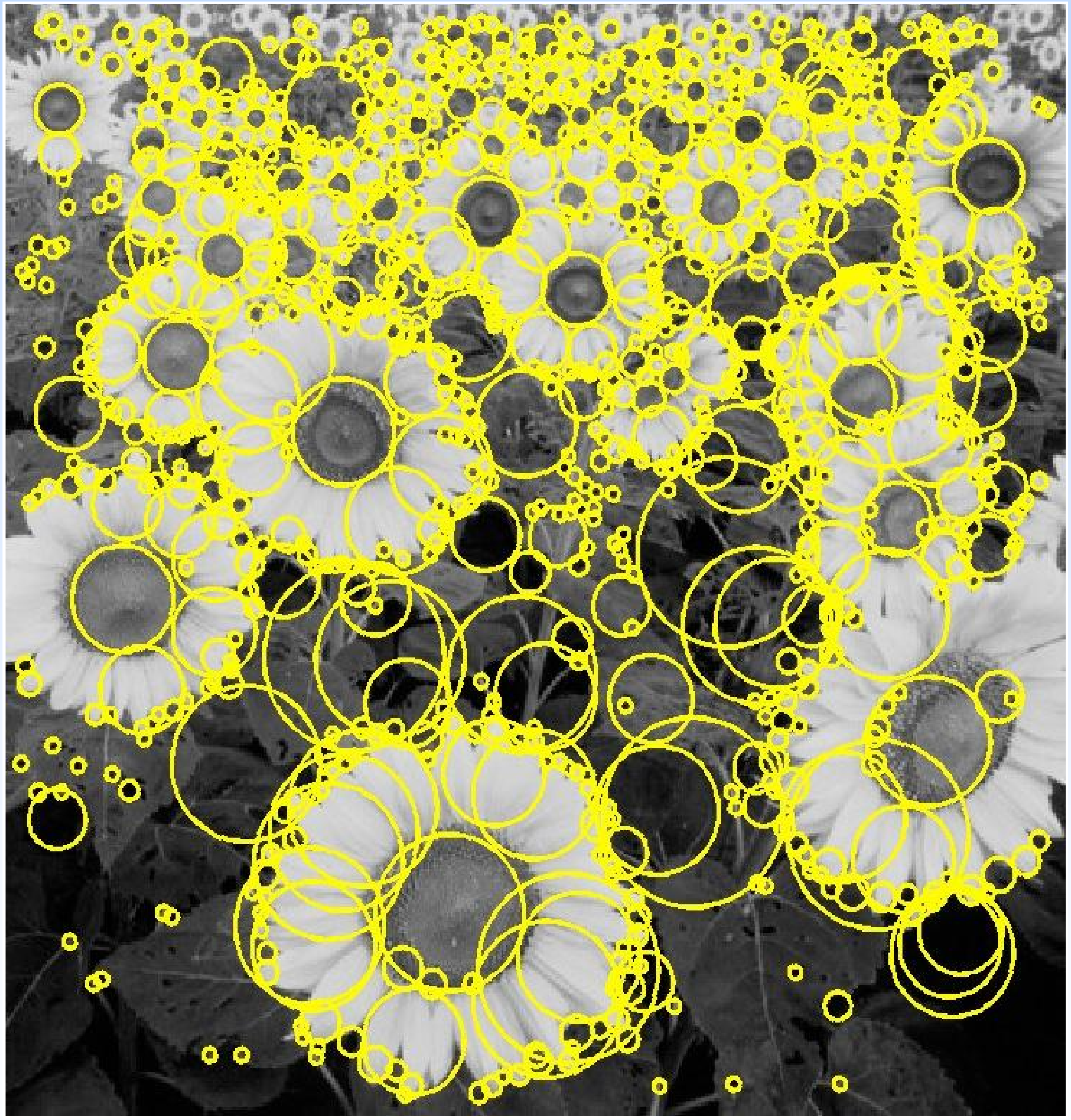
- gradients, edges, corners, regions, blobs
- SIFT, SURF, HoG, ...

- N.N. search, randomized trees, FERNS, signatures

- 3D-2D, 2D-2D corespond.
- obj. func. (non-lin. optim.)
- RANSAC
- robust estimators

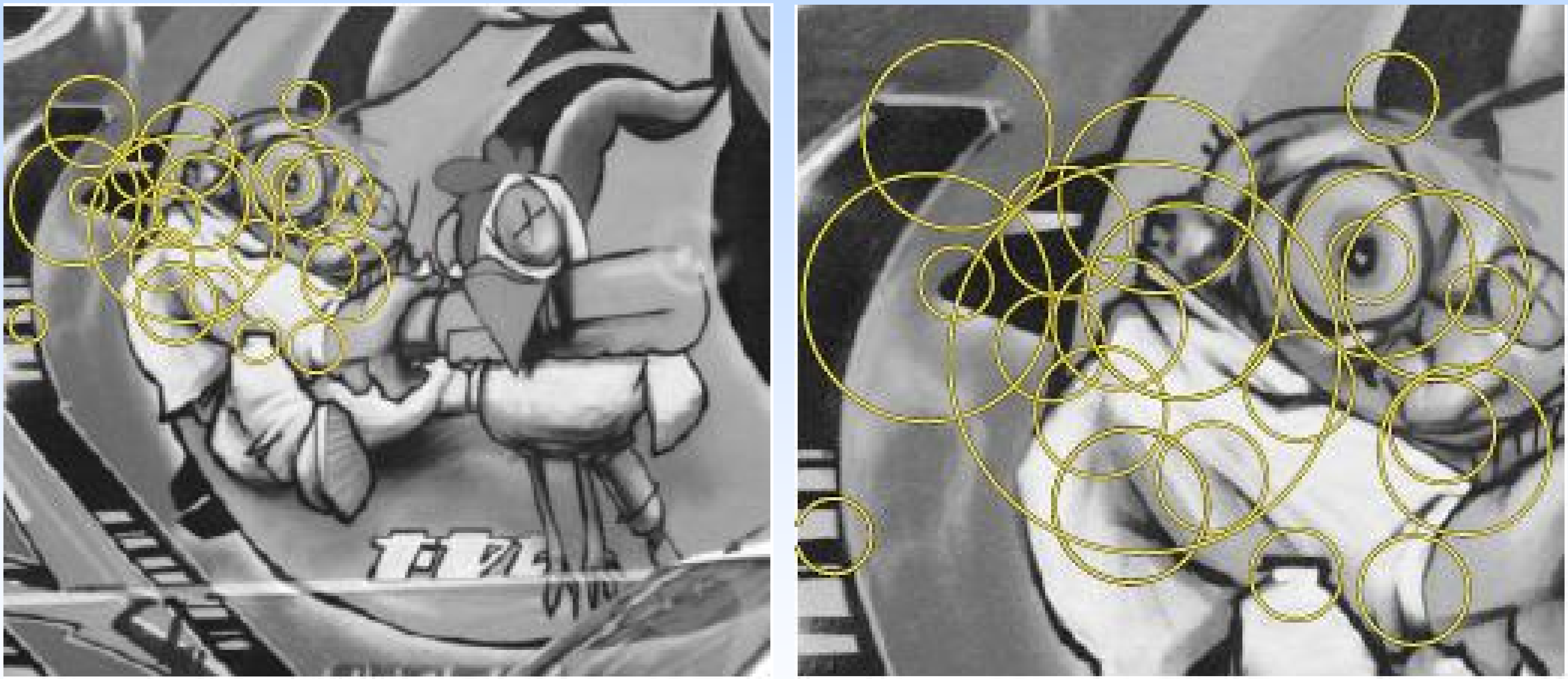
Feature extraction

Corners



Blobs

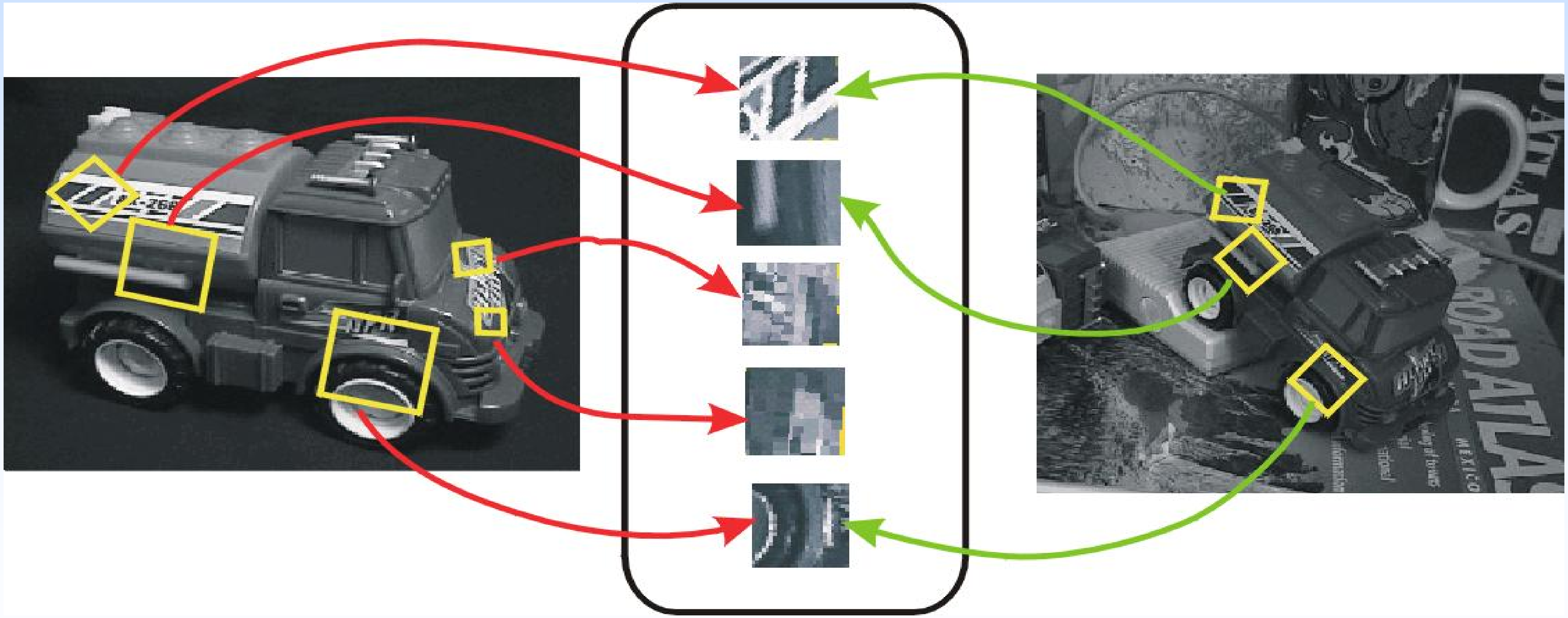
Corners at different scales



Corresponding features found using Harris-Laplace corner detector

Matching

Image content is transformed into local feature coordinates that are invariant to: **scale**, **rotation**, **viewpoint changes** and **illumination changes.**

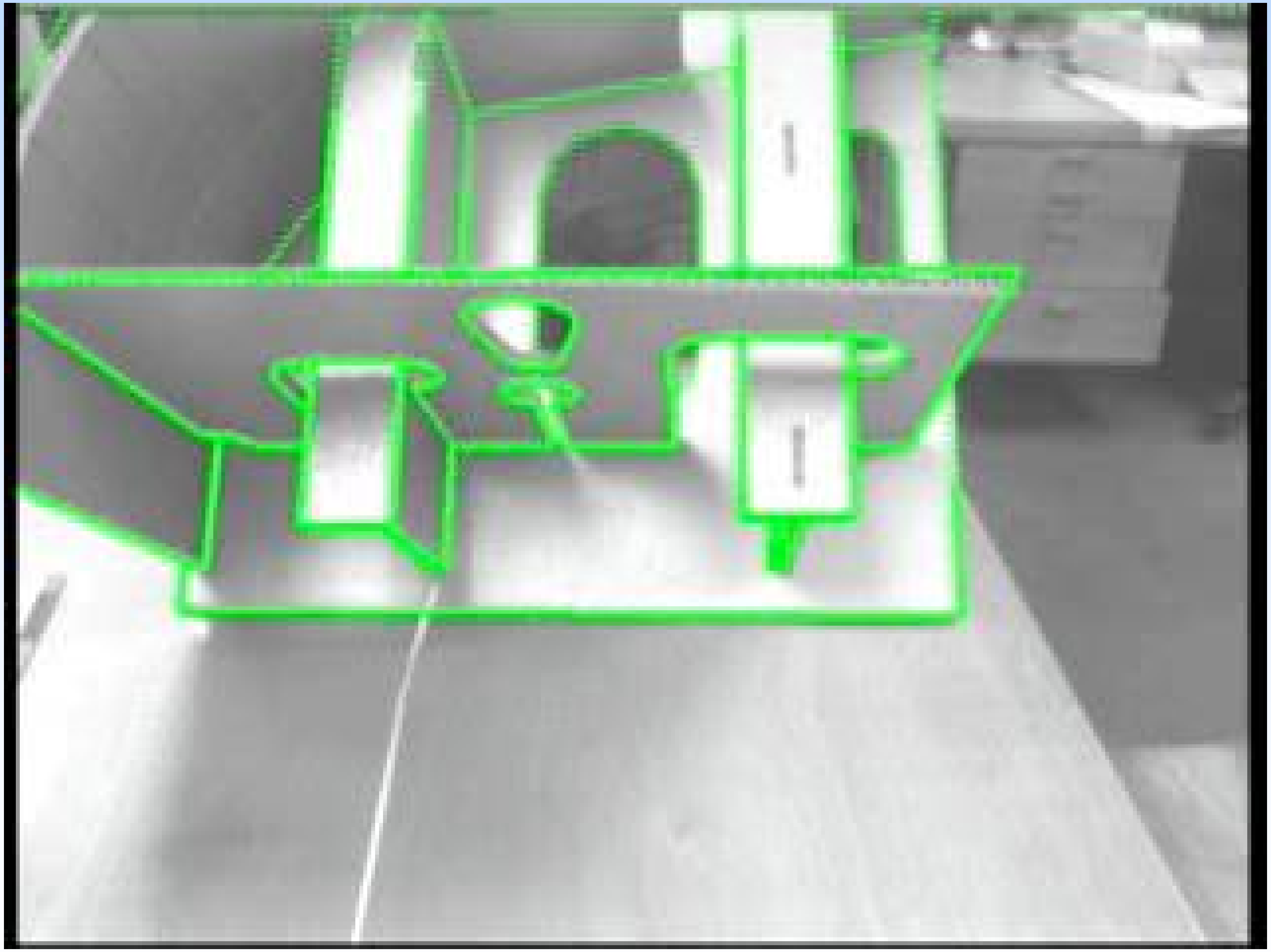


SIFT Features

Model-based face tracking

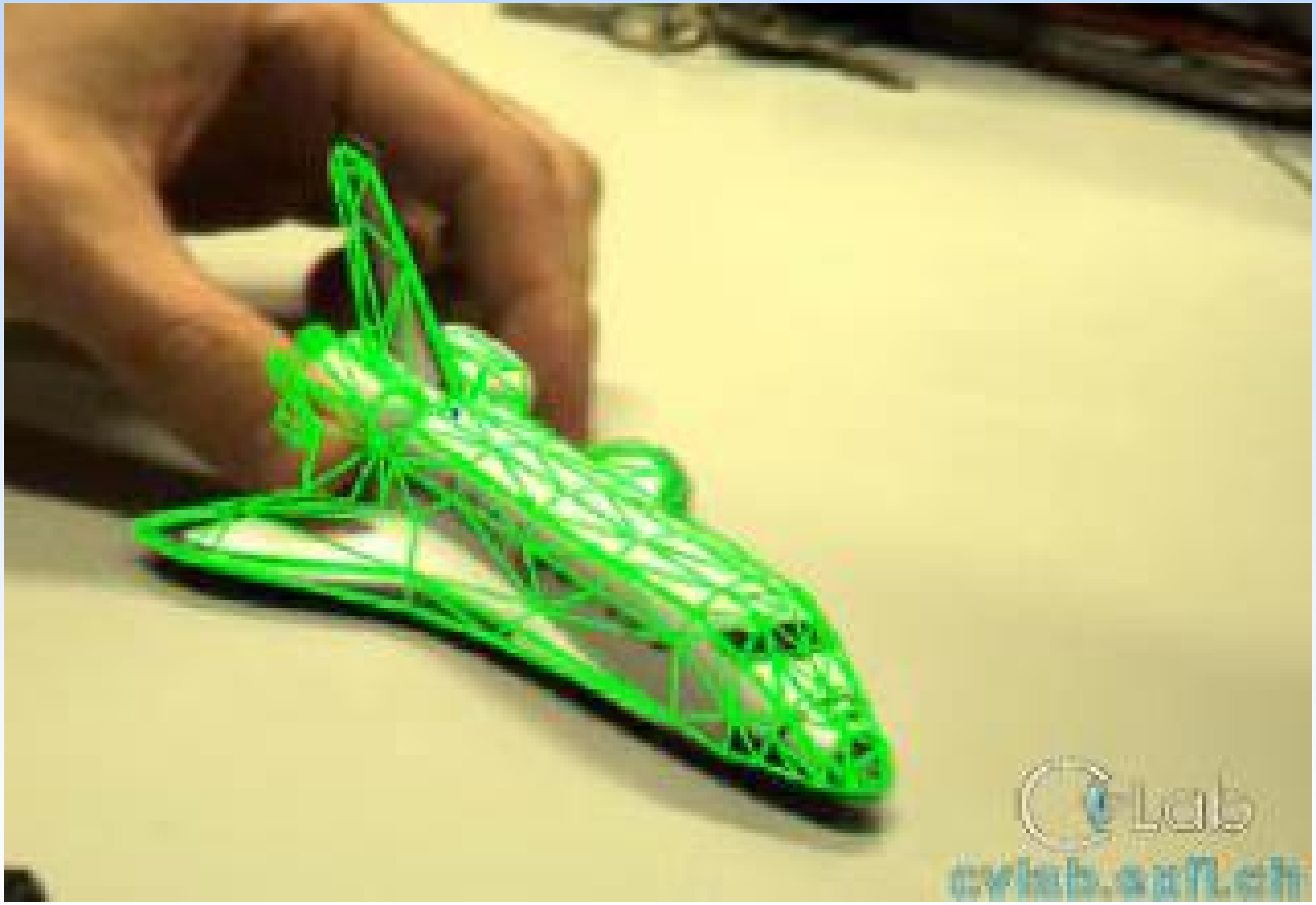
3D face tracking, courtesy of CVLAB, EPFL

Model based object tracking



Video is courtesy of University of Cambridge

Tracking 3D

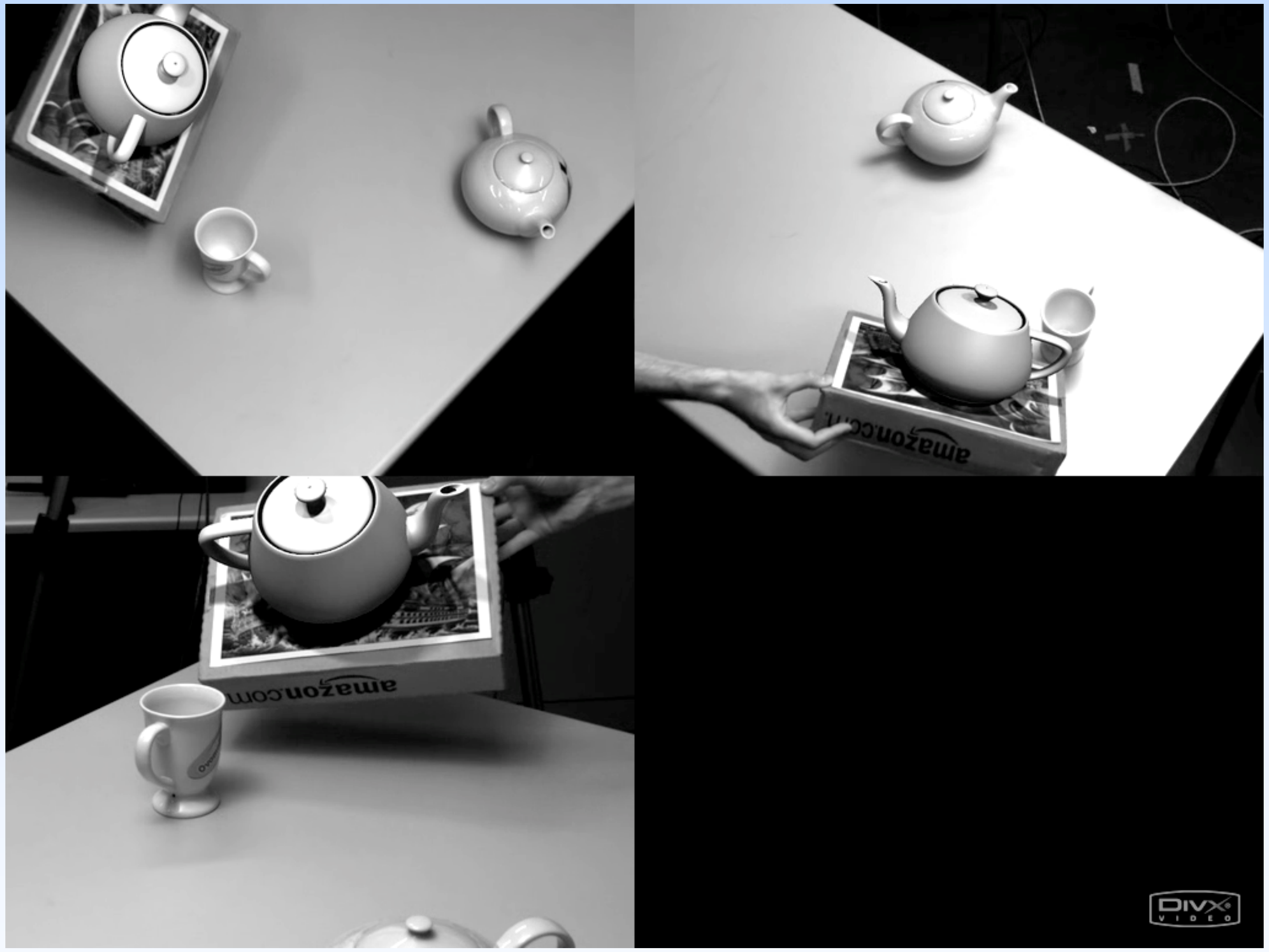


Tracking 3D objects, CVLAB, EPFL

Model-based face tracking

3D object tracking, courtesy of CVLAB, EPFL

Augmented reality

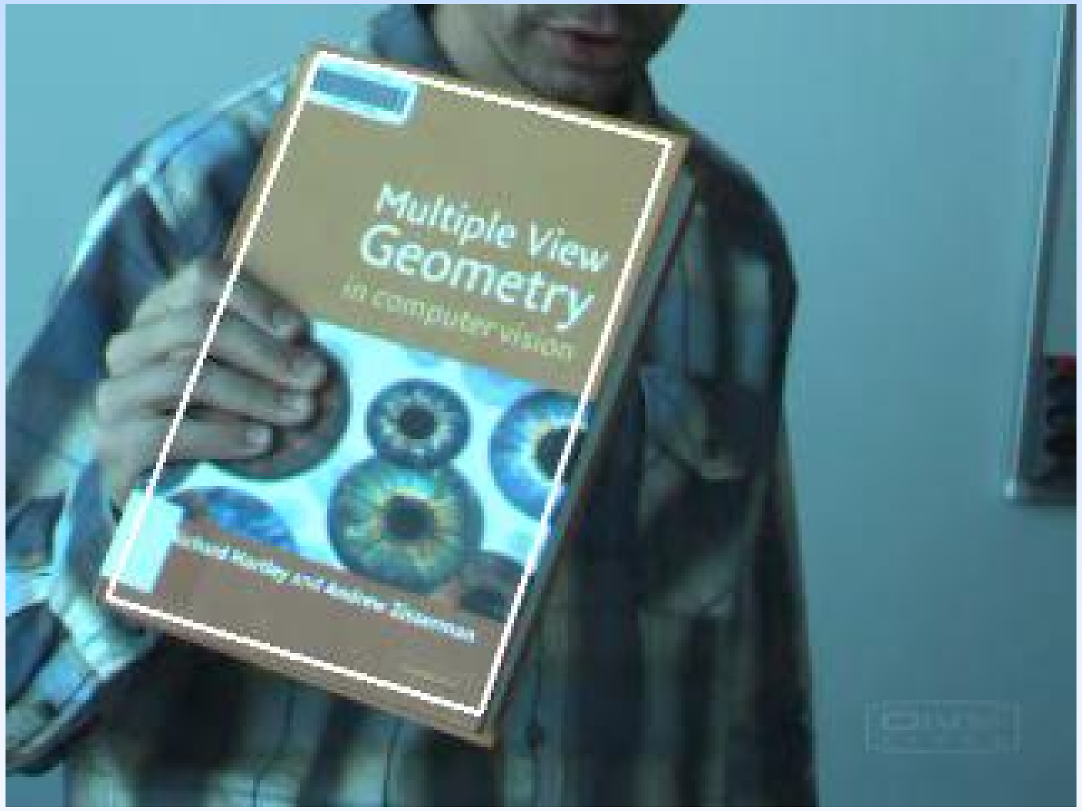


J. Pilet, A. Geiger, P. Lagger, V. Lepetit and P. Fua, An all-in-one solution to geometric and photometric calibration, International Symposium on Mixed and Augmented Reality, October 2006

M. Salzmann, J.Pilet, S.Ilic, P.Fua, Surface Deformation Models for Non-Rigid 3--D Shape Recovery, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, Nr. 8, pp. 1481 - 1487, August 2007

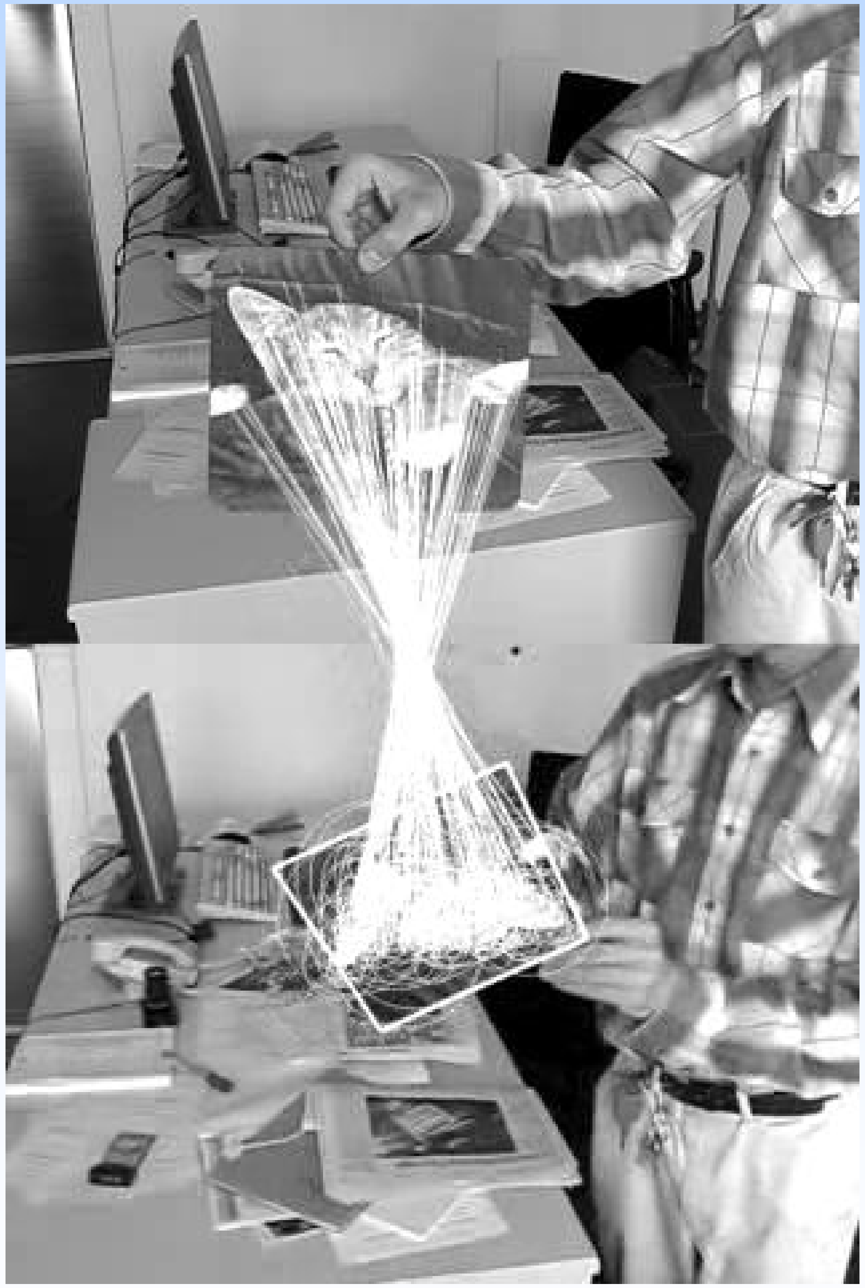
What is detection?

- Detection means finding the object of interest and providing its position in the image.
  - no assumption of the system dynamics
  - the response is not based on temporal consistency
- The applications of detection are various: machine vision and quality control, surveillance, robotics etc.
- The source of information is a single image



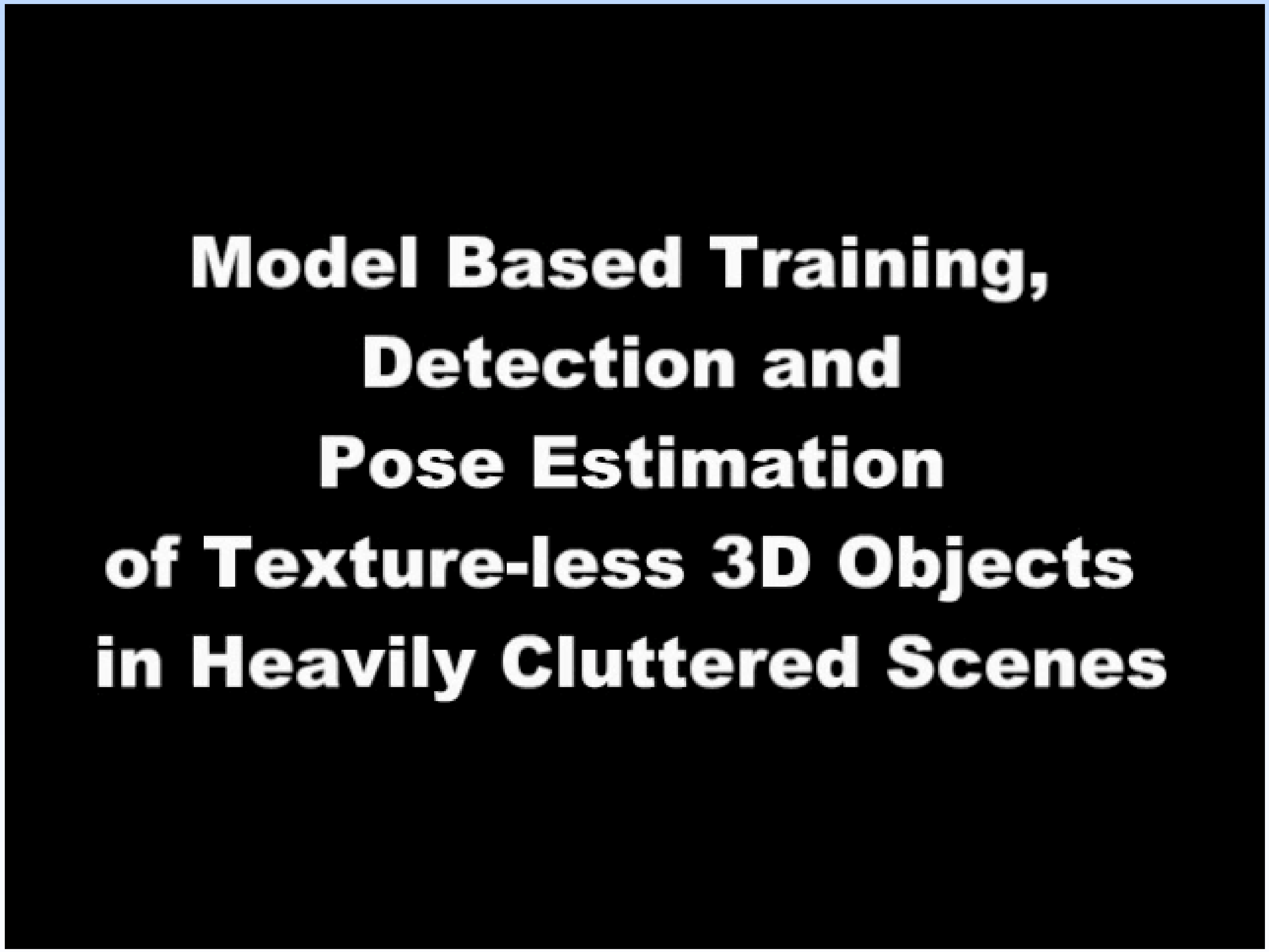
What is tracking by detection

- Tracking-by-Detection means following the object of interest in video by detecting it in every frame separately.
  - no assumption of the system dynamics
  - the response is not based on temporal consistency
- The source of information is a video



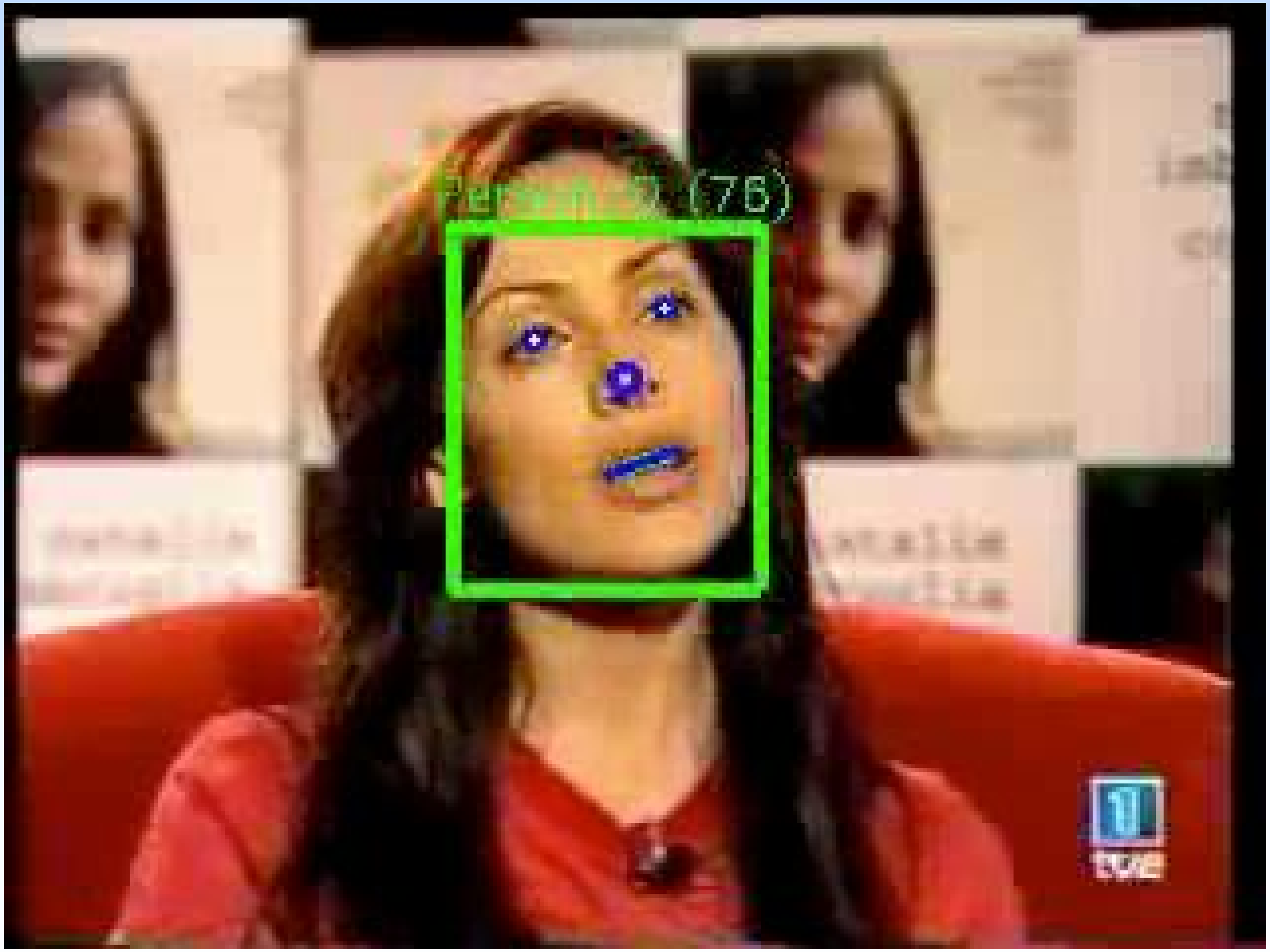
M. Ozuysal, M. Calonder, V. Lepetit and P. Fua, Fast Keypoint Recognition using Random Ferns, accepted to IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009.

3D object detection

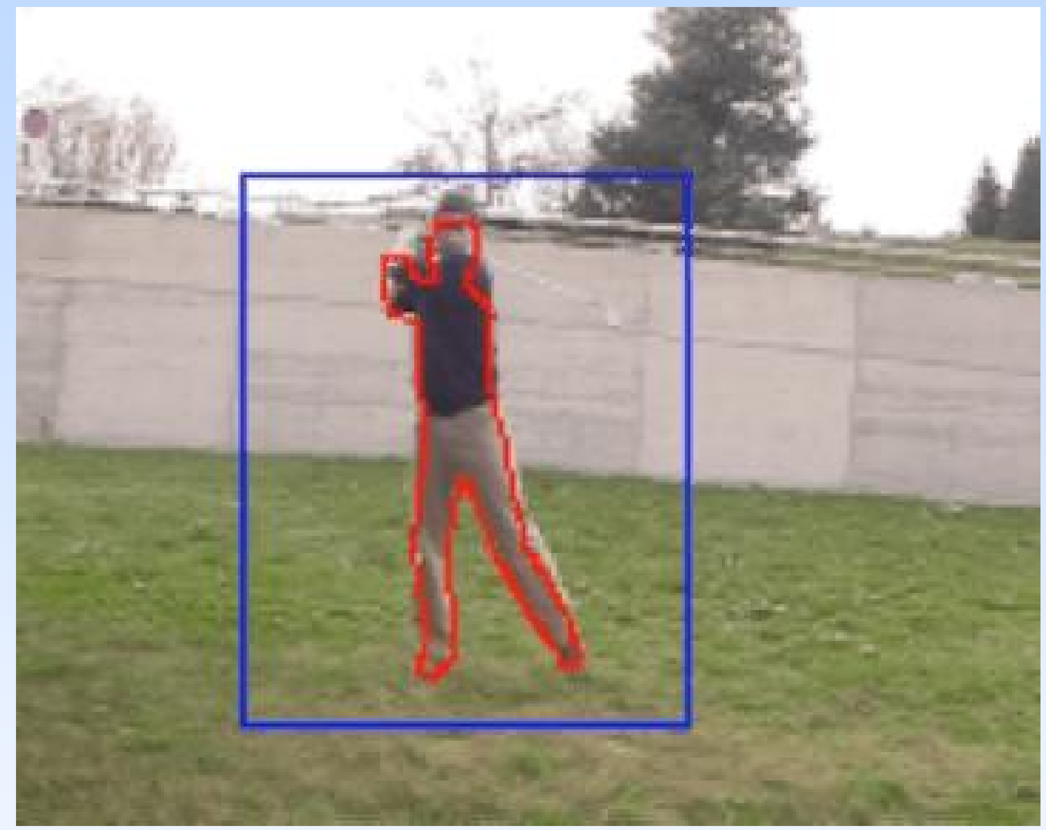


S. Hinterstoisser et al. , Model Based Training, Detection and Pose Estimation of Texture-less 3D Objects in Heavily Cluttered Scenes, ACCV2012

Face detection



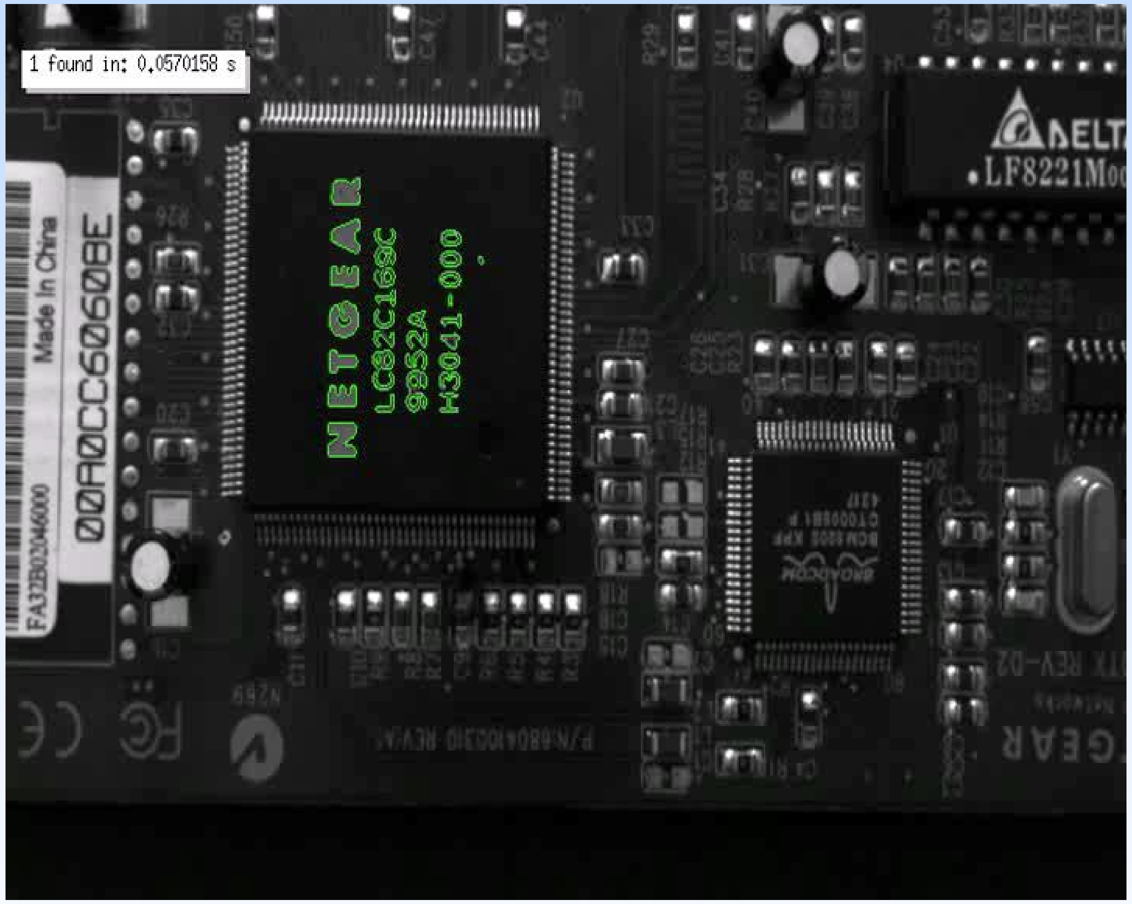
People detection

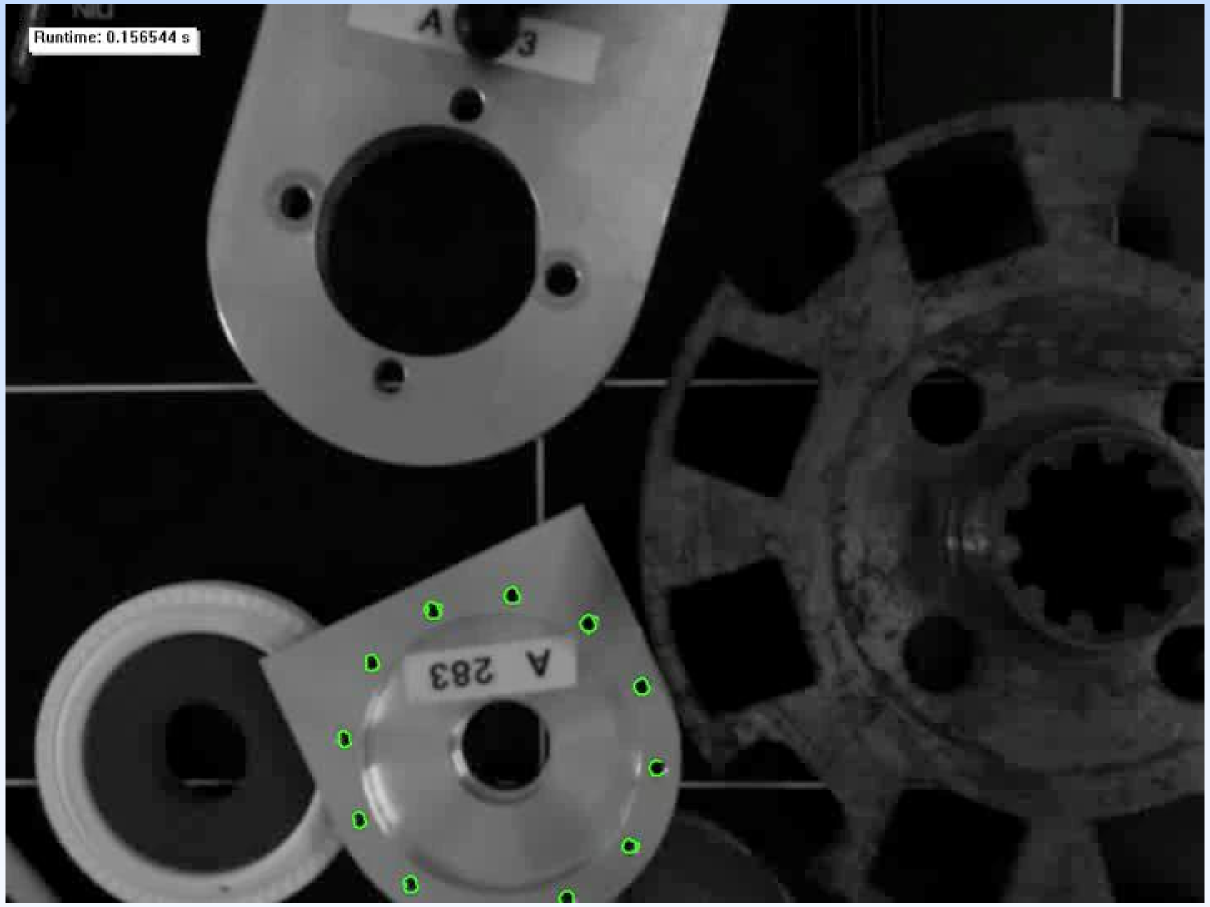


M. Dimitrijevic, V. Lepetit and P. Fua, Human Body Pose Detection Using Bayesian Spatio-Temporal Templates, Computer Vision and Image Understanding, Vol. 104, Nr. 2, pp. 127 - 139, December 2006

A. Fossati, M. Dimitrijevic, V. Lepetit and P. Fua, Bridging the Gap between Detection and Tracking for 3D Monocular Video-Based Motion Capture, Conference on Computer Vision and Pattern Recognition, Minneapolis, MI, June 2007

Detection in Machine Vision





A. Hofhauser, Carsten Steger, N. Navab, **Harmonic deformation model for edge based template matching**, *International Conference on Computer Vision Theory and Applications, Funchal, Portugal, January 2008.*

Overview of the course

- Introduction
- Filtering and edge detection: Convolution; Gaussians; Image derivatives; Edge detection; Canny edge detector
- Local invariant feature detectors:
  - Corner detection: Harris corner detector; Scale space; Harris-Laplace; Harris-Affine; FAST
  - Blob detectors: Hessian; Hessian Laplace/Affine;
  - Feature descriptors: SIFT, SURF, HoG, ORB
- Feature point recognition: Randomized and Regression Forests, FERNS,BRIEF
  - Haar features; Integral images; Ada-Boost; Viola-Jones face detection.

Overview of the course

- Camera models and projections; Model based tracking; Pose estimation form 2D-3D correspondences(“DLT,”POSIT,Soft Posit, RANSAC); Rotation parameterization;
- Non-linear optimization; Robust estimators;
- Template tracking: Lucas-Kanade; Compositional Alg., Inverse Compositional Alg., ESM; Juri-Dhome algorithm(linear predictor); Active appearance and active shape models.
- Mean-Shift Tracking
- Template matching approaches
- Random Forest Based Tracking

Exam, exercises and homeworks

- **Final exam** 100pts (50pts to pass excluding bonus points)
- **Mid-term exam** (max10pts) + Home work projects (max 10-11pts)
- **Total:** 120pts (100pts = 1.0!!!)
- **Lectures:** Monday from 12:15-14:00h at MI 00.13.009A
- **Exercises:** Thursdays 12:00-14:00h at MI 00.13.009A
  - mainly practical on the computer and will serve to explain you given homework tasks from theoretical and practical point of views.
  - check your previous home works(depending on the schedule)
  - you can start doing your home work on the exercises class and ask questions
- **Home works:** will be check individually during the exercises!