Aalto University
School of Electrical Engineering

ELEC-E8126: Robotic Manipulation Perception

Ville Kyrki

1.2.2021

Learning goals

- Know types of perception problems that relate to manipulation.
- Understand basic principles of solutions to the above.

What things would you want to perceive for manipulation?

Perception problems in manipulation

- Measure properties of target objects
  - 3-D shape, location
  - dynamics properties, e.g. mass, surface friction
- Measure properties of environment (e.g. obstacles)
  - Similar to above
- Measure properties of interaction
  - For example, contact force, slipping
- Measure properties of robot
  - For example, pose, dynamics

How those could be measured?

Today

- Visual perception
  - non-contact, primarily 3-D shape/position

- Tactile and force perception
  - in-contact, measurement of interaction

How is 3-D perception possible using light?

How is 3-D perception possible using light?

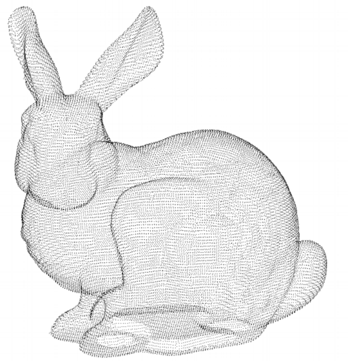
- Direct 3-D measurement (e.g. time-of-flight camera, laser scanner)
- 3-D reconstruction from multiple views
- 3-D measurement from projected light
- 3-D measurement from 2-D image of known target

Direct 3-D sensing: methods and typical characteristics

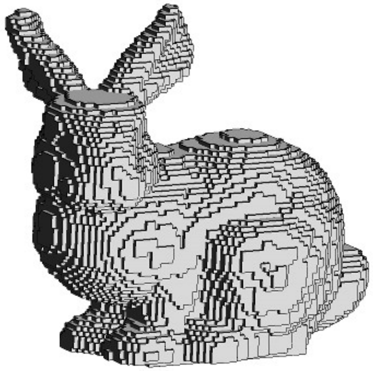
- 3-D laser scanners (LiDARs)
  - limited resolution (for manipulation), not fast
- 2-D laser scanner + mechanical scanning
  - slow
- Time-of-light cameras
  - Lower resolution than normal cameras, often fast
  - Principle: direct tof or phase of rf modulated light

Note: Types of 3-D models

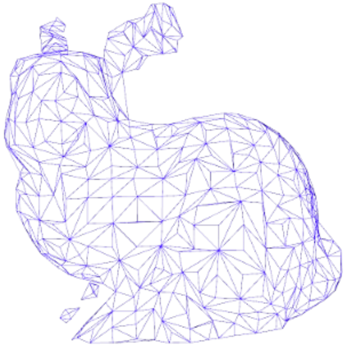
- Point clouds
  - possibly with color information



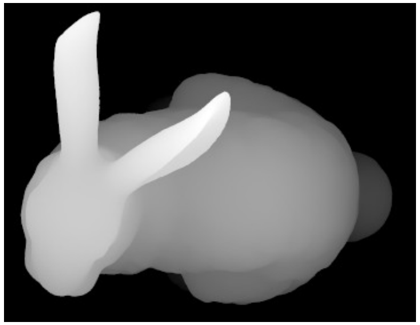
- Voxel grid



- Polygonal/mesh models
  - possibly with texture



- Depth maps
  - possibly with registered (color) image



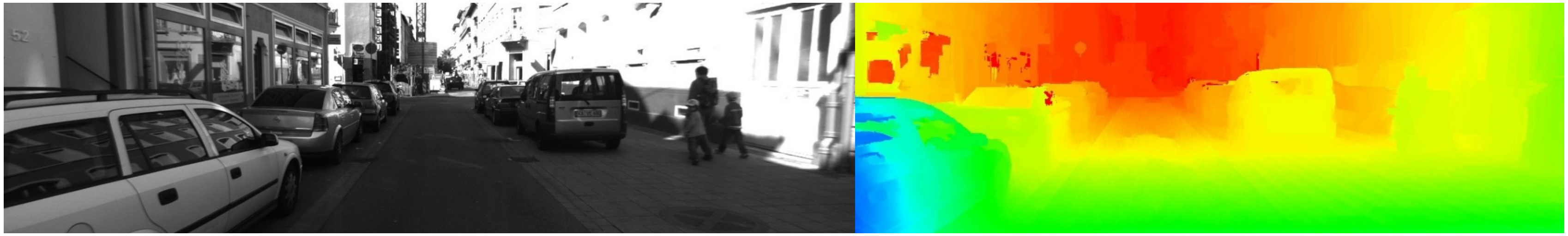
- Sparse feature graphs
  - Each feature represents a neighborhood of a point
  - Does not represent entire shape

3-D reconstruction from two simultaneous views (stereo)

- Experiment: Thumbs up!

3-D reconstruction from two simultaneous views (stereo)

- Experiment: Thumbs up!
- Calibrated stereo cameras detect correspondence and calculate disparity (position difference) map.
- Disparity inversely proportional to distance.
- Relatively high spatial resolution.
- Distance resolution decreases with distance, increases with baseline increase (distance between sensors)



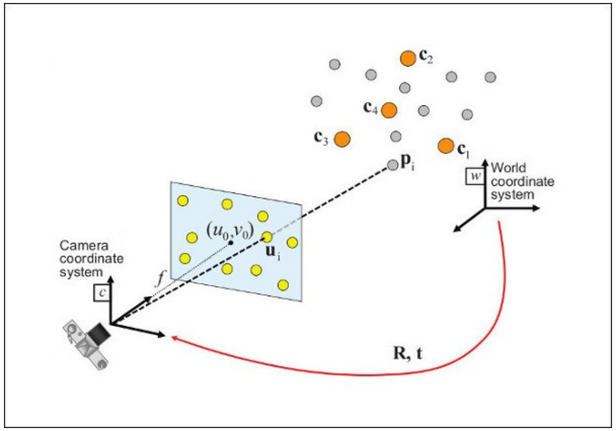
Best quality stereo requires heavier processing (outside integrated stereo camera).

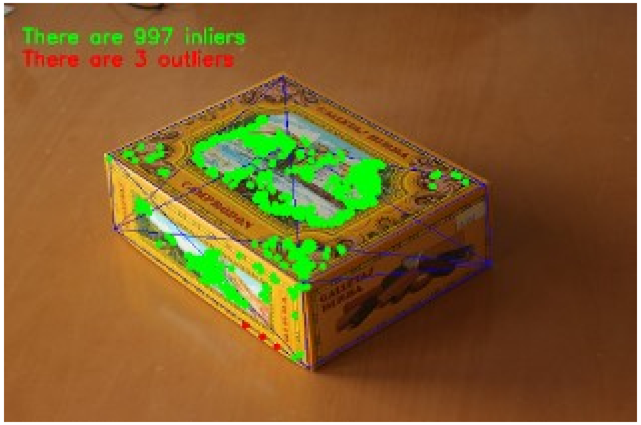
3-D reconstruction from projected light (active stereo)

- Idea: replace one camera with light projector.
  - Project random or deterministic pattern.
  - Geometrically identical problem to stereo, points will move with distance.
  - Characteristics similar to stereo.
- Can find correspondence in uniformly colored scenes (without texture).
- If several sensors are used in same scene, interference may be a problem

Understanding 3-D from 2-D

- Knowing a 3-D model of the object, the 3-D pose of it can be found in a 2-D image (pose estimation).
  - Assumes calibrated camera.
  - Typically based on matching known unique points on the object.





Primary uses of vision in manipulation

- Localize pose of known objects
- Reconstruct objects and/or environment
  - Target objects or obstacles
- Hand-eye coordination

Require understanding of 3-D.

- Detect objects
  - e.g. bounding box
- Recognize objects
  - Often by category

Outside scope of this course, often performed in 2-D. Computer vision course tells about these.

Primary uses of vision in manipulation

- **Localize pose of known objects**
- Reconstruct objects and/or environment
  - Target objects or obstacles
- Hand-eye coordination

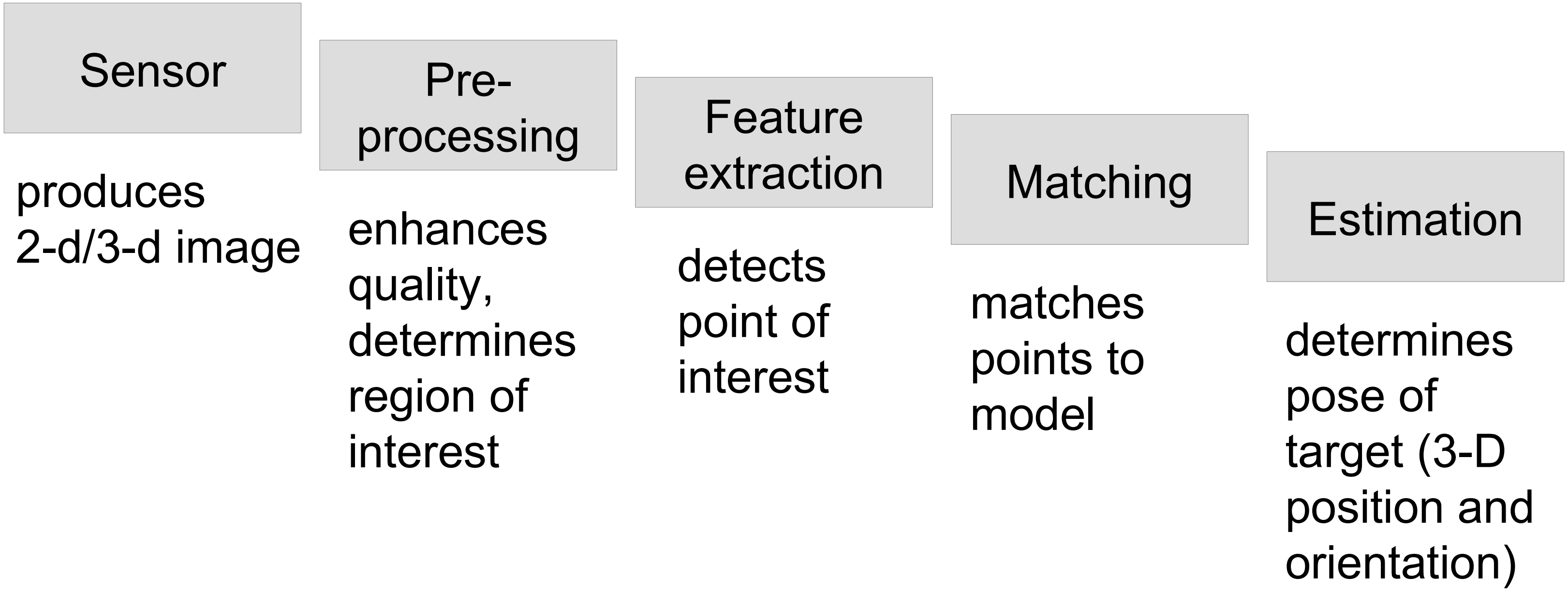
Require understanding of 3-D.

- Detect objects
  - e.g. bounding box
- Recognize objects
  - Often by category

Outside scope of this course, often performed in 2-D. Computer vision course tells about these.

Typical vision pipeline for object localization

- Traditional pipeline



Sensor

produces 2-d/3-d image

Pre-processing

enhances quality, determines region of interest

Feature extraction

detects point of interest

Matching

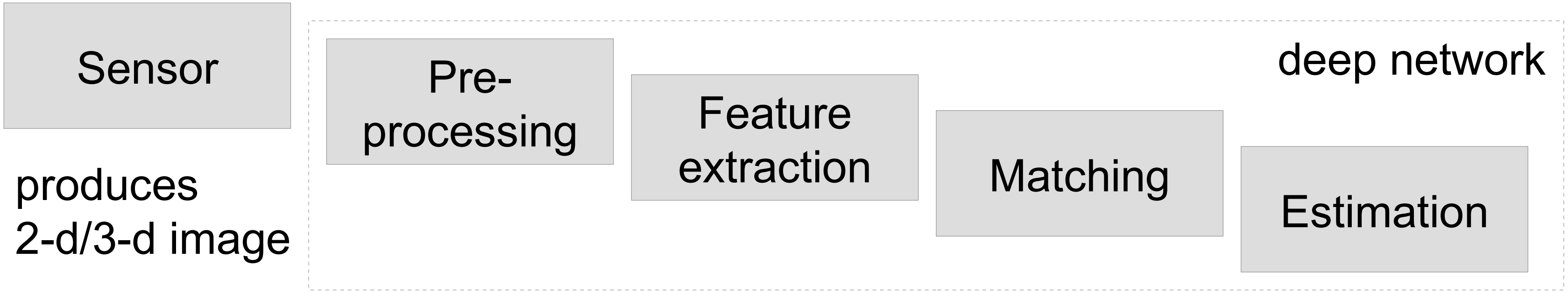
matches points to model

Estimation

determines pose of target (3-D position and orientation)

Typical vision pipeline for object localization

- Deep learning pipeline



Sensor

Pre-processing

Feature extraction

Matching

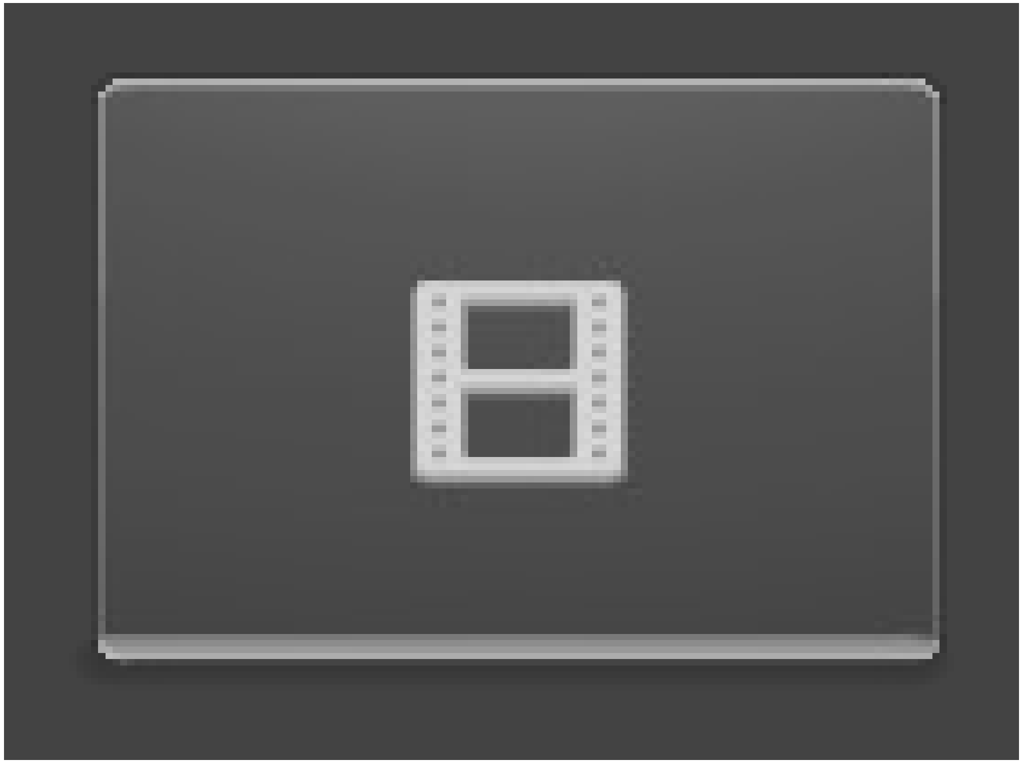
Estimation

deep network

produces
2-d/3-d image

At the moment, deep learning not (yet?) much used in 3-D object localization for robots because of high cost of training data. However, for common targets such as human faces suitable (data general). Also computer graphics may be used for data generation.

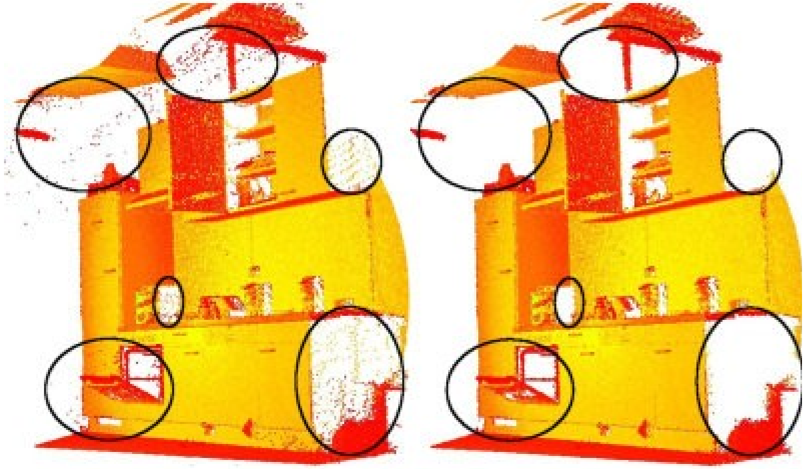
Example of real-time pose tracking

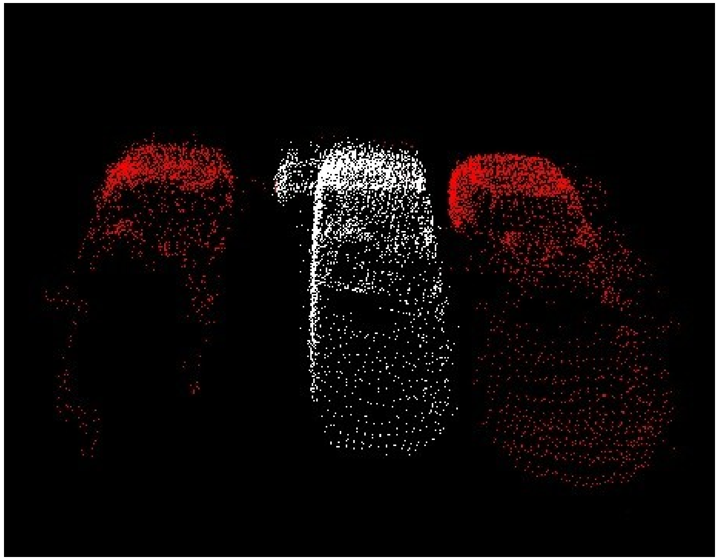


Pre-processing

- Image/data enhancement
  - Removing noise/outliers.
  - Resampling (up/down).
  - Enhancing useful structure.

- Segmentation
  - Dividing data to regions.
  - Detecting region of interest.





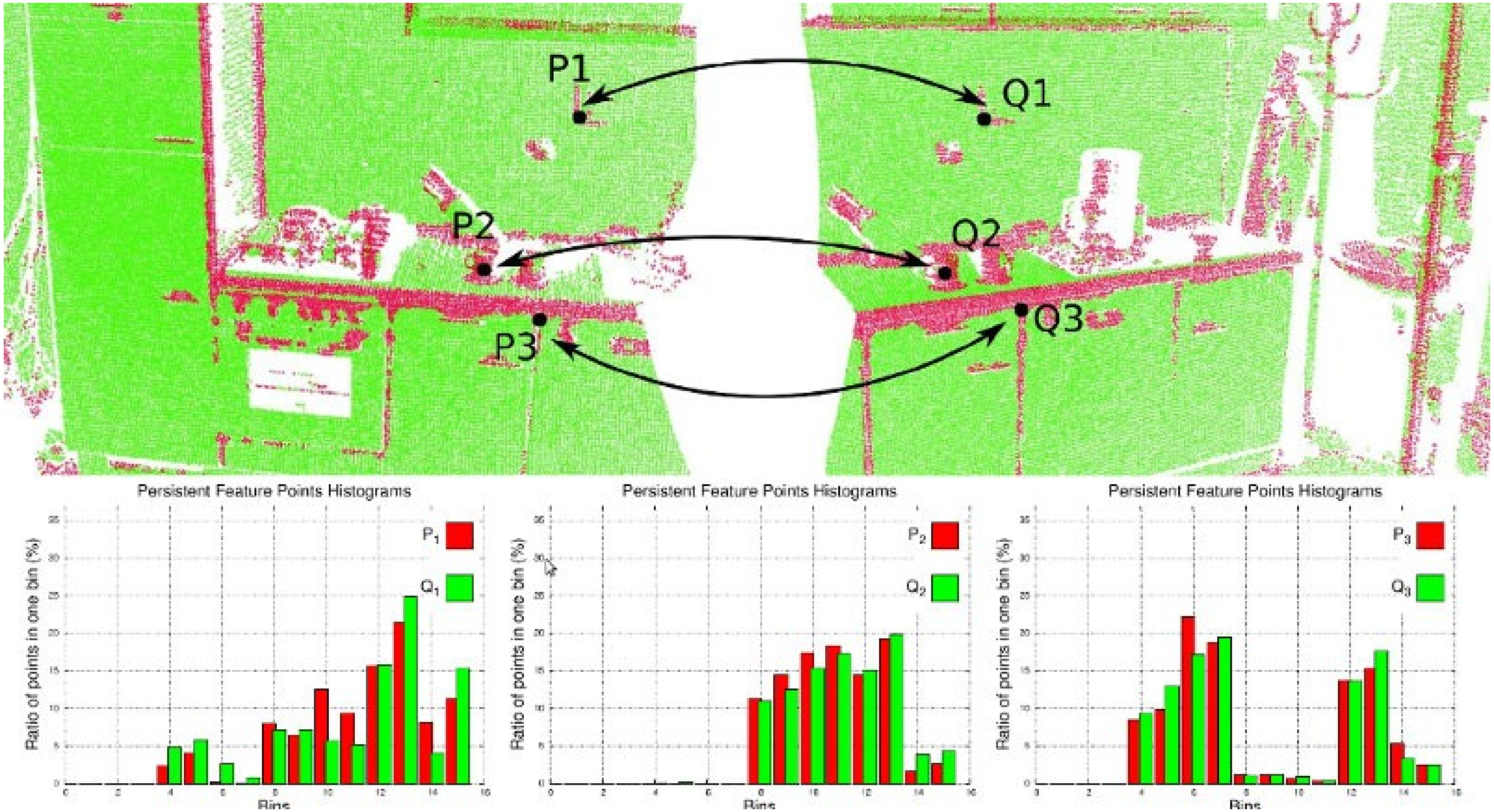
Feature extraction

- Detection
  - Choose which points to describe
  - Good points somehow unique
    - to reduce points to match
    - In 2-D image e.g. corners

- Representation
  - Build a feature vector, numeric descriptor for each point.
  - Good feature is discriminative, robust to noise, invariant to geometric transforms

- Many methods
  - 3-D (point cloud): PFH, FPFH, VFH, NARF, …
  - 2-D (image): SIFT, SURF, MSER, FAST, ...

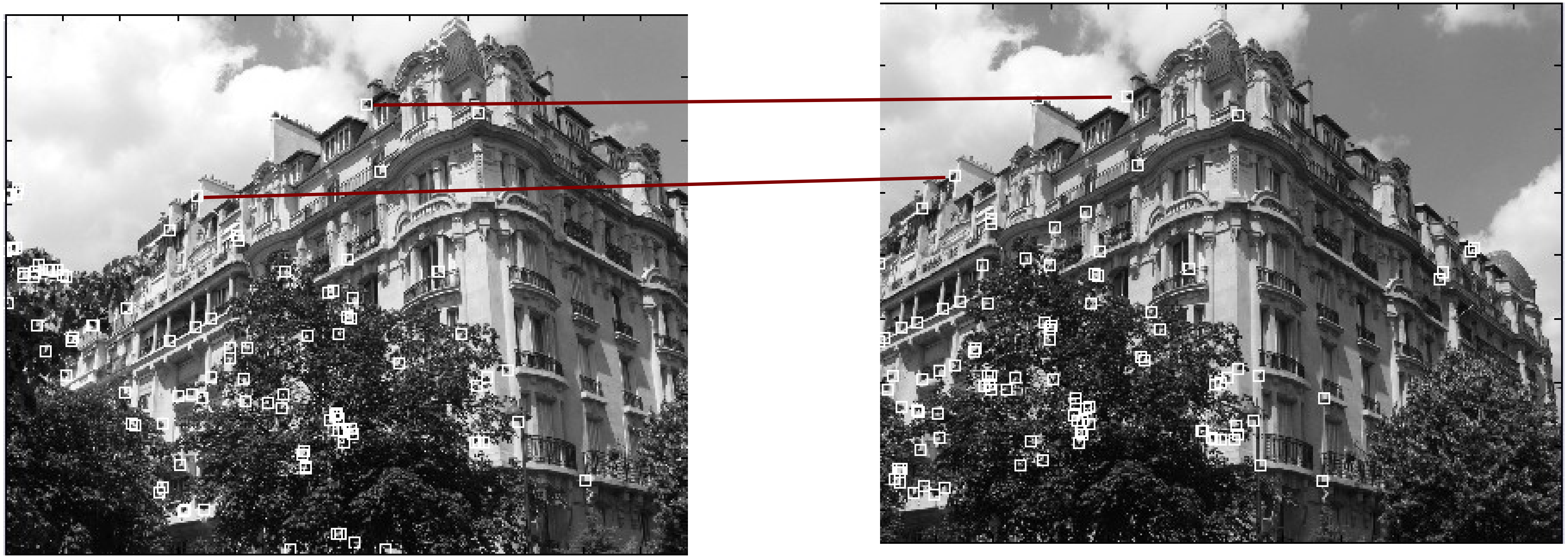
Feature extraction – 3-D example
Point feature histograms



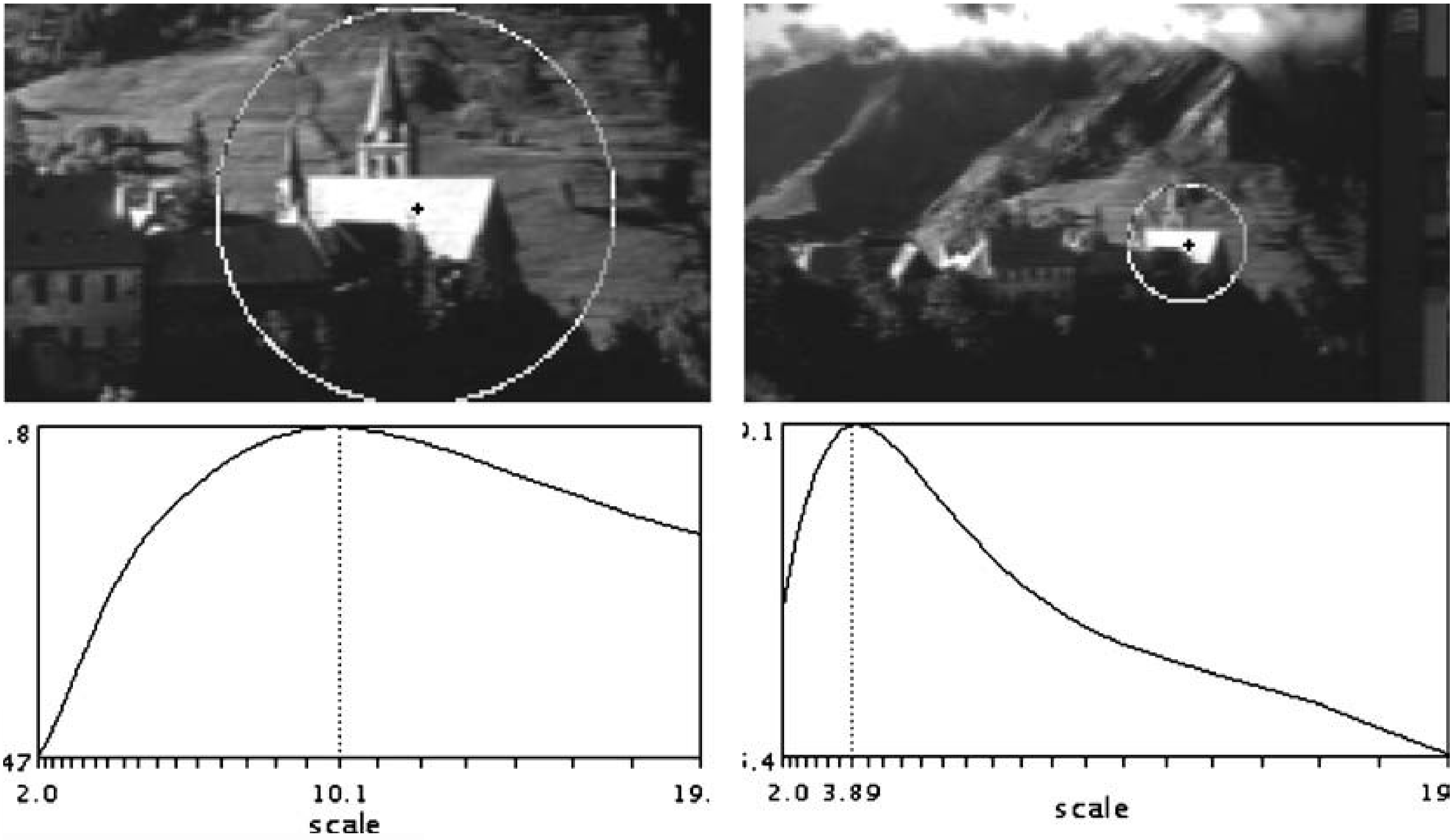
Aalto University
School of Electrical
Engineering

PFH describes difference in normals between points

Feature extraction – 2-D example
Corner matching

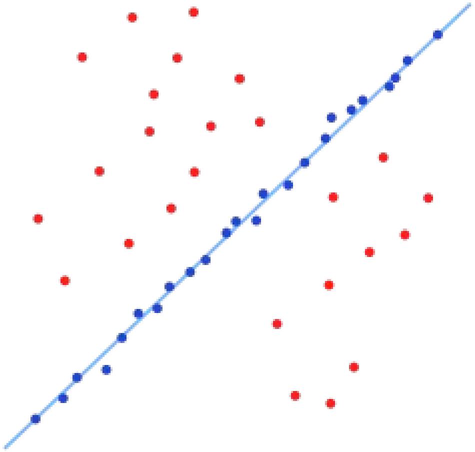


Feature extraction – 2-D example
Scale invariant features



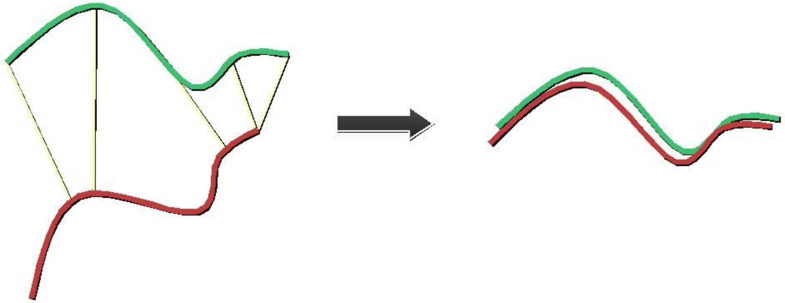
Matching robustly

- Some matches are usually invalid because descriptors are not unique.
- Random Sample Consensus (RANSAC) is a popular algorithm to find largest consistent inlier set for an estimation (fitting) problem.
  - Repeat: Fit to few points and check how many other points match with the fit
  - Choose fit that has the largest number of matching points and re-estimate model with those.



Estimation: 3-D with 3-D

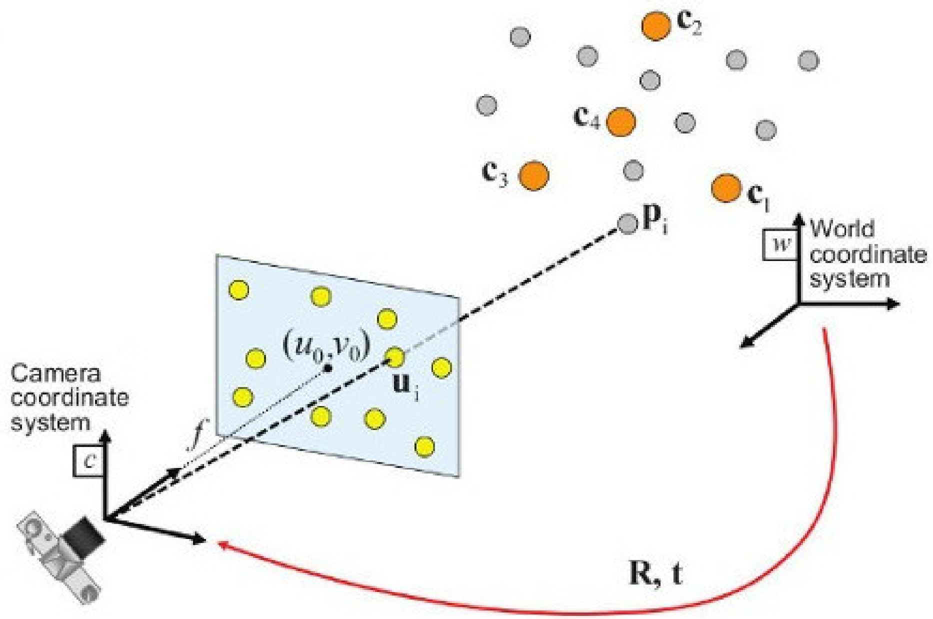
- Unique least squares solution available for 3+ correspondences.
  - Deterministic algorithm available, iterative solutions for other error functions.
- Correspondences may be imperfect: Often followed by iterative dense fitting (iterative closest point, ICP).
  - For each point in source pointcloud,
    choose closest point in target,
    fit using those points.
  - Repeat until convergence.
  - Requires good starting point.



Estimation: 2-D with 3-D

- Minimize reprojection error.
- Unique LS solution for 4+ non co-planar points (6+ co-planar).

- Non-iterative solutions available, but do not minimize geometric error (do not tolerate noise very well).
- Iterative optimization (e.g. Gauss-Newton, Levenberg-Marquardt) usually used.



Primary uses of vision in manipulation

- Localize pose of known objects
- **Reconstruct objects and/or environment**
  - Target objects or obstacles
- Hand-eye coordination

Require understanding of 3-D.

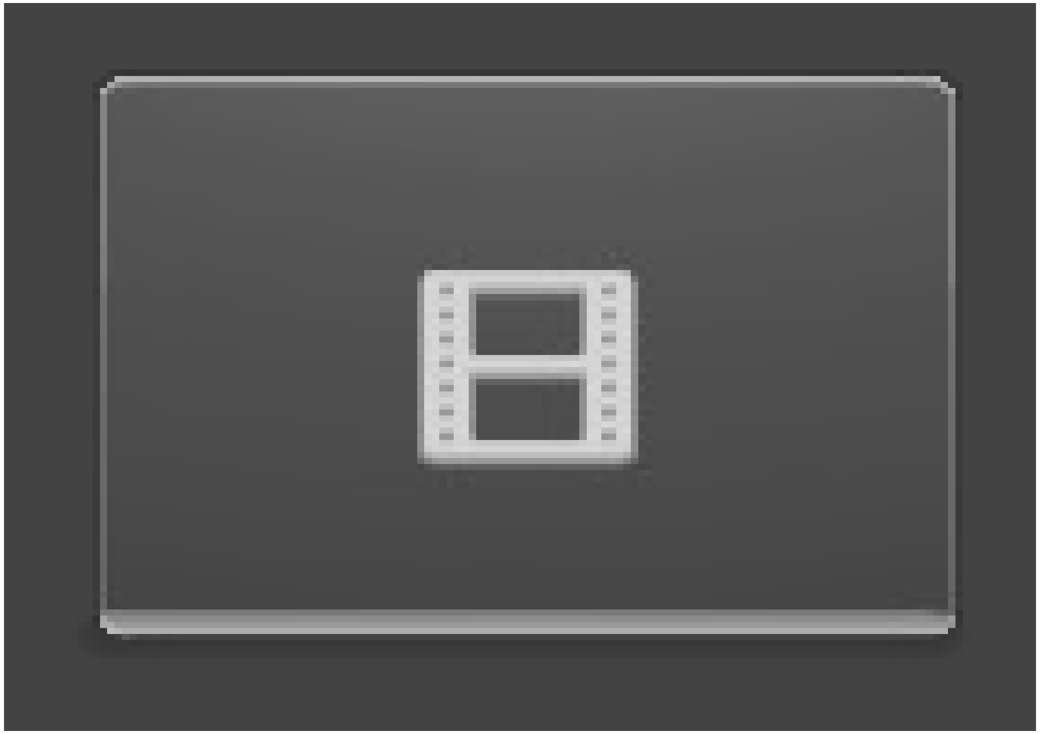
- Detect objects
- Recognize objects

Outside scope of this course, often performed in 2-D.

Scene reconstruction

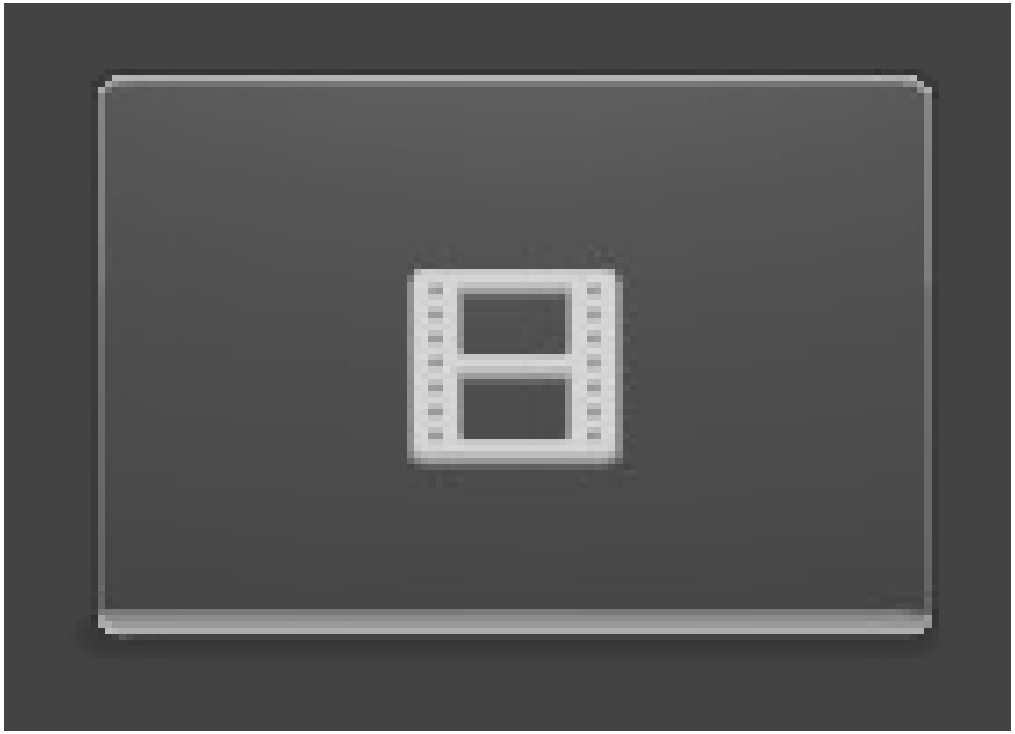
- Aim: Produce a 3-D model of a scene (/object).
- Problem: combine several 3-D models into one.
- Basic approach:
  - Detect relative pose between captured models (align).
  - Combine into one model (merge).
  - Smoothen (post-process, often using voxels).
- Similar approaches for different 3-D sources.

Example



Live Dense Reconstruction with a Single Moving Camera, CVPR 2010

Example with Kinect



Kinect Fusion, SIGGRAPH 2011

Reconstruction from dynamic scene

Video Pop-up: Monocular 3D Reconstruction of Dynamic Scenes, ECCV 2014

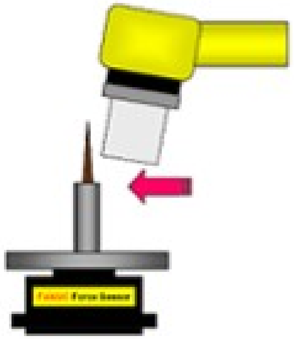
Interaction sensors

- In manipulation, measurement of interaction (e.g. forces) between objects (e.g. robot and environment) may be needed.
  - Difficult to measure with visual sensors.
- In contrast to visual sensors, measure either global properties (sum of interaction forces), or local properties in small area (force distribution in a small area)
- Direct measurement of interaction may be needed for highly controllable interaction such as handling delicate objects or tool use.

Force/torque sensing

- Force/torque sensors can be mounted (typically on robot wrist) to measure interaction forces.
  - Often built of 6 single-axis force sensors (strain gauges) to measure 3-dof force and 3-dof torque.
  - Different F/T ranges available.
  - Often good resolution but tricky to calibrate (creep, tool weight / dynamics).
  - Another option to measure force at each joint.
    - Miniature sensors for robot fingers.

Wrist mounted sensor.

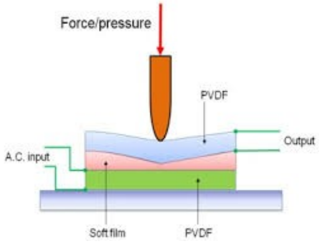


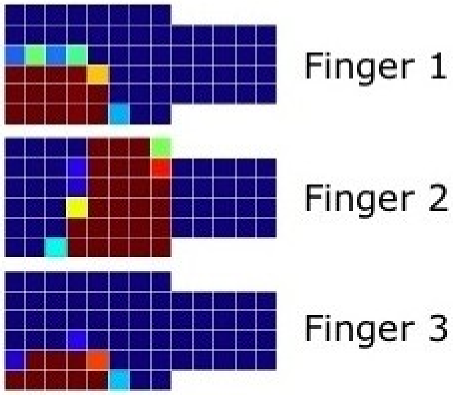
Sensor mounted at fixture.

Tactile sensing

- Tactile sensors may measure
  - Pressure/force (often in grid → tactels), often based on measurement of sensor deformation.
    - Elastoresistive, capacitive, piezoresistive, optical
  - Vibration
    - May be used for measurement of surface texture
  - Heat
  - ...

Most common





Software

- Point cloud library (PCL)
  - ROS integrates with PCL

- OpenCV (computer vision)

Summary

- Visual sensors are often used for estimation of pose and shape of environment, including objects.

- Force/torque and tactile sensors can be used to measure interaction.

Next time: Motion control

- Readings:
  - Lynch & Park, Chapter 11-11.3.2