

Robotic 3D Vision

Lecture 16: 3D Object Detection 2 – 3D Keypoints, ICP, Surfel Pair Matching

Prof. Dr. Jörg Stückler

Computer Vision Group, TU Munich

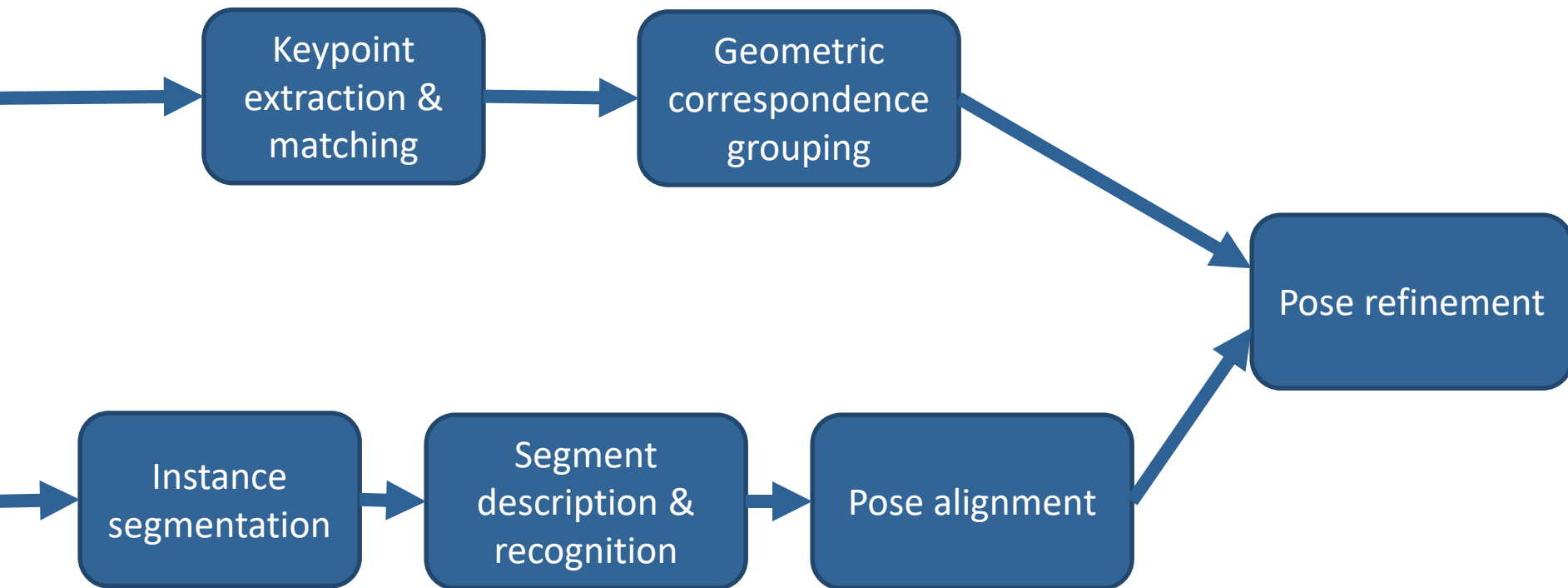
<http://vision.in.tum.de>

What We Will Cover Today

- 3D keypoint detectors and descriptors
- Global 3D object descriptors
- Surfel-pair matching
- Iterative closest points algorithm

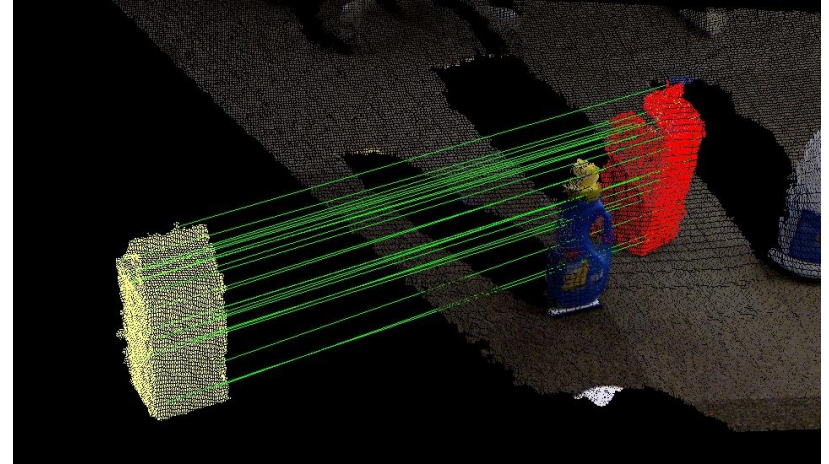
Recap: 3D Object Detection Pipelines

- Local vs. global object description



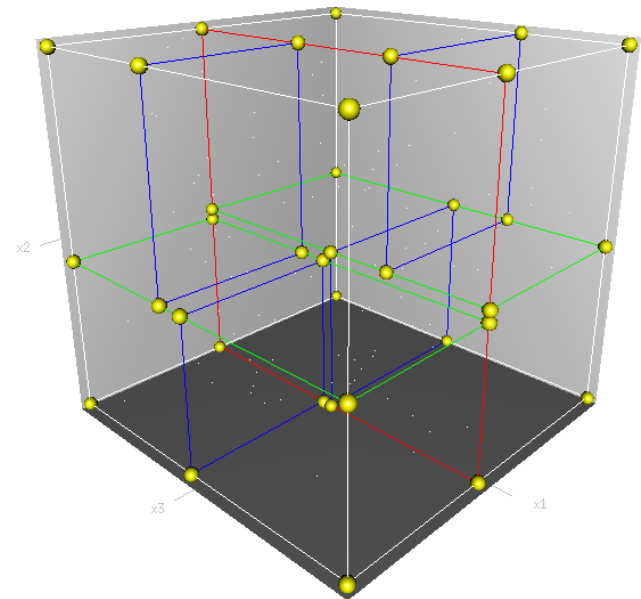
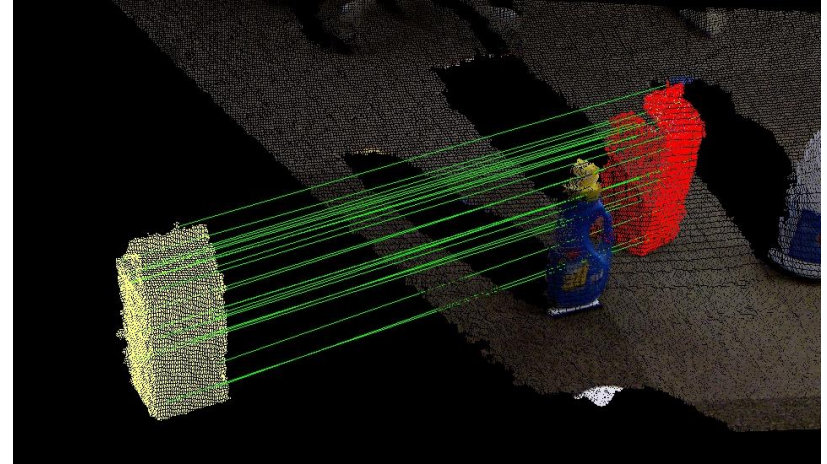
3D Object Detection with Local Keypoints

- Detect and match a set of local keypoints between model and scene
- Locality of keypoints provides robustness against occlusions
- Local keypoints should be distinctive and repeatable, combined properties of detector and descriptor!
- Alignment for pose estimation:
 - 3D-to-3D alignment
 - Pose voting from keypoint match through local reference frames



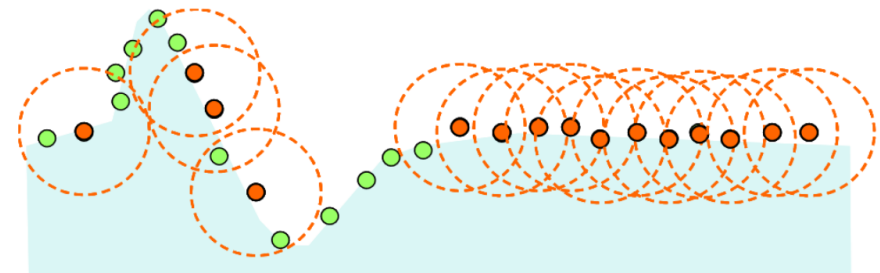
Training Objects with Local Features

- Either
 - Extract keypoints on 3D object models (f.e. CAD or scanned), or
 - Render views on CAD models and extract keypoints for rendered views
- For each object/view, store keypoints in an efficient search data structure for the descriptor metric (e.g. kd-tree)



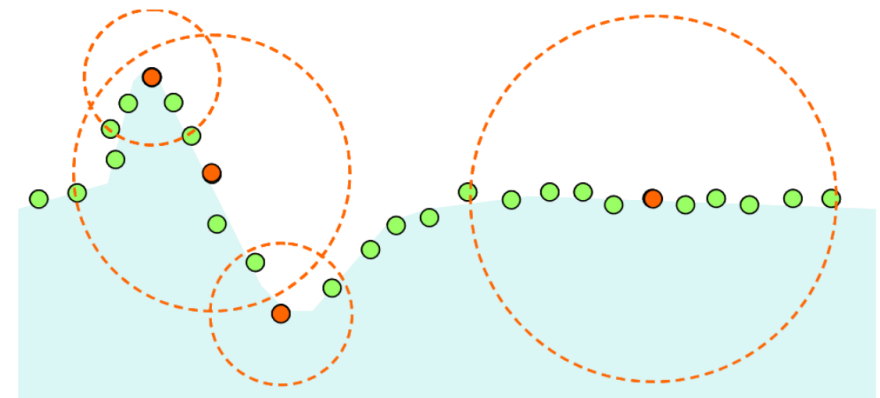
3D Keypoint Detectors

- Strategy 1: Uniform spatial sampling
- Strategy 2: Detection of keypoints at maxima of 3D interest measures
 - Intrinsic Shape Signatures (ISS) Detector, Zhong 2009
 - Harris3D
 - ...
- Multi-scale vs. characteristic scale
- Extraction of a local reference frame



Sparse
but not representative

Exhaustive
but redundant



Data-driven selection
of both locations and neighborhoods

Intrinsic Shape Signatures (ISS) Detector

- Interest measure based on covariance of local point distribution

$$\Sigma(\mathbf{p}_i) = \frac{1}{\sum_{j:|\mathbf{p}_j-\mathbf{p}_i|<r} w_j} \sum_{j:|\mathbf{p}_j-\mathbf{p}_i|<r} w_j (\mathbf{p}_i - \mathbf{p}_j)(\mathbf{p}_i - \mathbf{p}_j)^\top$$

- Weights account for varying point density

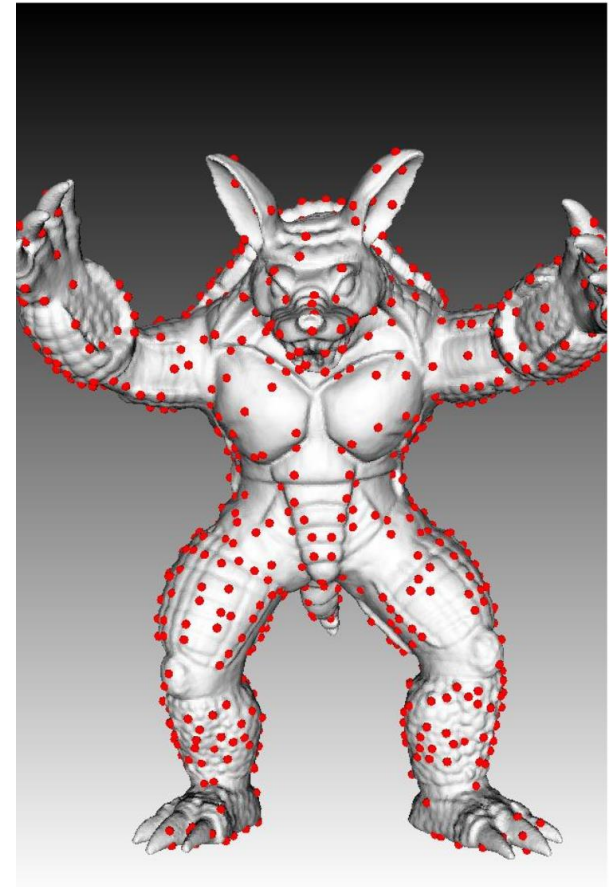
$$w_i := \frac{1}{|j:|\mathbf{p}_j-\mathbf{p}_i|<r|}$$

- Compute eigenvalues of local covariance

$$\lambda_1 > \lambda_2 > \lambda_3$$

- Find local maxima of smallest eigenvalue λ_3

- Constrain by thresholds on λ_2/λ_1 and λ_3/λ_2 to find points with well conditioned eigen vector directions



Local Reference Frame

- Extract local reference frames from eigen vectors to align rotation-variant descriptor
- 4 possible cases for right-handed frame

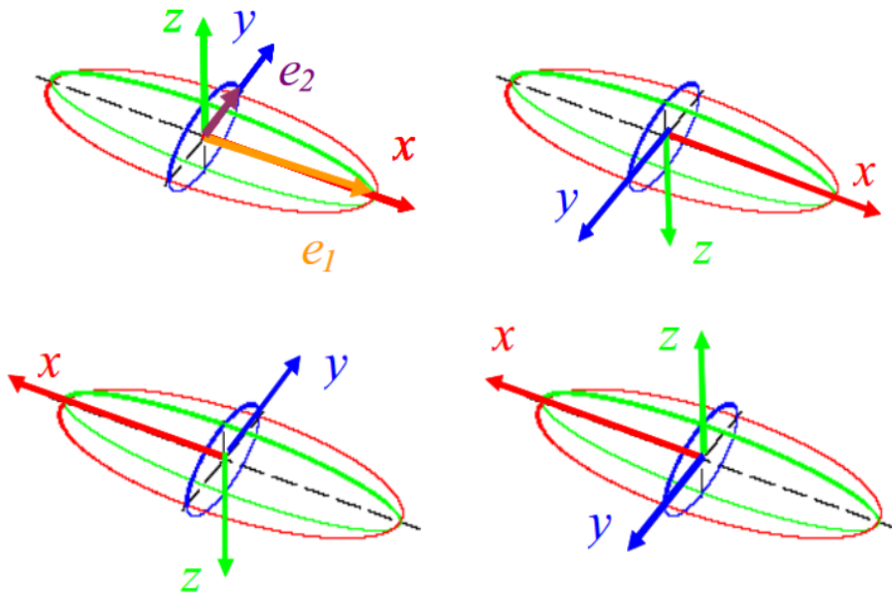


Image from Y. Zhong, 2009

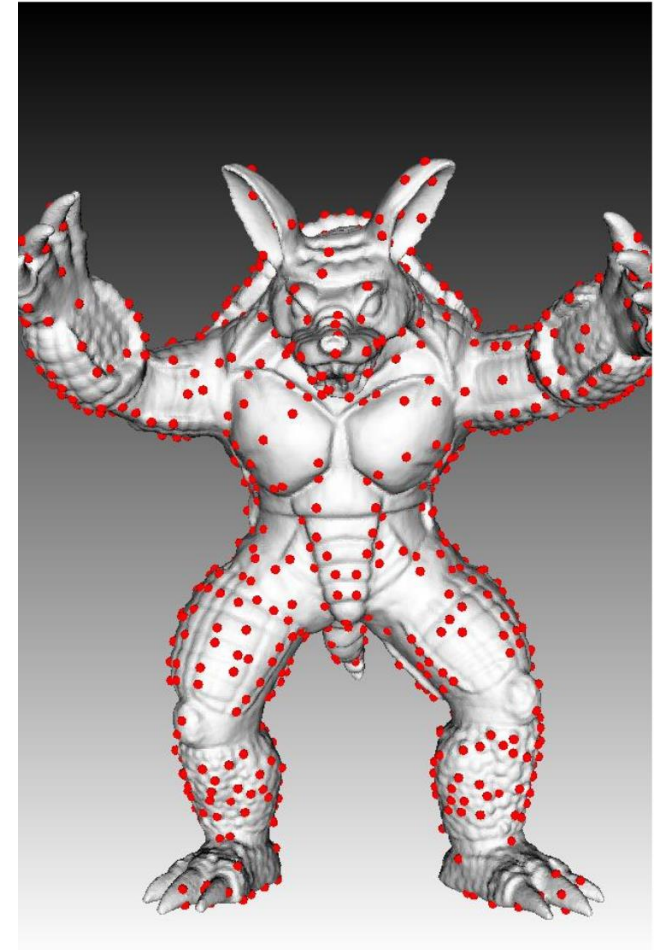


Image from F. Tombari

Local Reference Frame: Disambiguation

- Disambiguate the 4 possible cases by quantifying the support of the directions
- Directions and opposite directions of eigenvectors:

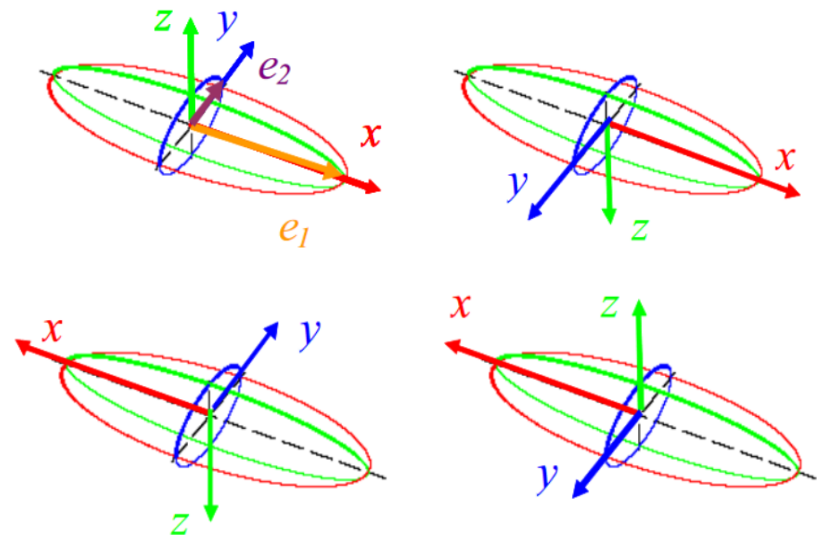
$$\mathbf{x}^+, \mathbf{y}^+, \mathbf{z}^+ \qquad \mathbf{x}^-, \mathbf{y}^-, \mathbf{z}^-$$

- Choose x-axis according to strongest support

$$S_x^+ \doteq \{i : d_i \leq R \wedge (\mathbf{p}_i - \mathbf{p}) \cdot \mathbf{x}^+ \geq 0\}$$

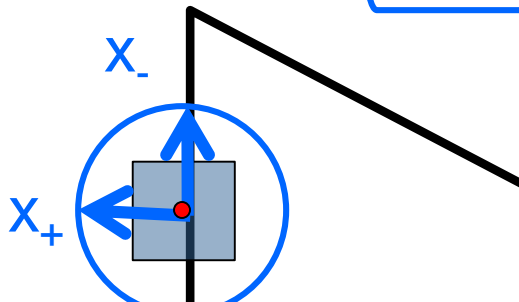
$$S_x^- \doteq \{i : d_i \leq R \wedge (\mathbf{p}_i - \mathbf{p}) \cdot \mathbf{x}^- > 0\}$$

$$\mathbf{x} = \begin{cases} \mathbf{x}^+, & |S_x^+| \geq |S_x^-| \\ \mathbf{x}^-, & \text{otherwise} \end{cases}$$



- z-direction analogously, y through $\mathbf{z} \times \mathbf{x}$

Recap: Structure Tensor

$$E(u, v) = [u \ v] \underbrace{\left(\sum_{(x,y) \in W} \begin{bmatrix} I_x^2 & I_x I_y \\ I_y I_x & I_y^2 \end{bmatrix} \right)}_H \begin{bmatrix} u \\ v \end{bmatrix}$$


Eigenvalues and eigenvectors of H

- Define shifts with the smallest and largest change (E value)
- x_+ = direction of largest increase in E.
- λ_+ = amount of increase in direction x_+
- x_- = direction of smallest increase in E.
- λ_- = amount of increase in direction x_-

$$H x_+ = \lambda_+ x_+$$

$$H x_- = \lambda_- x_-$$

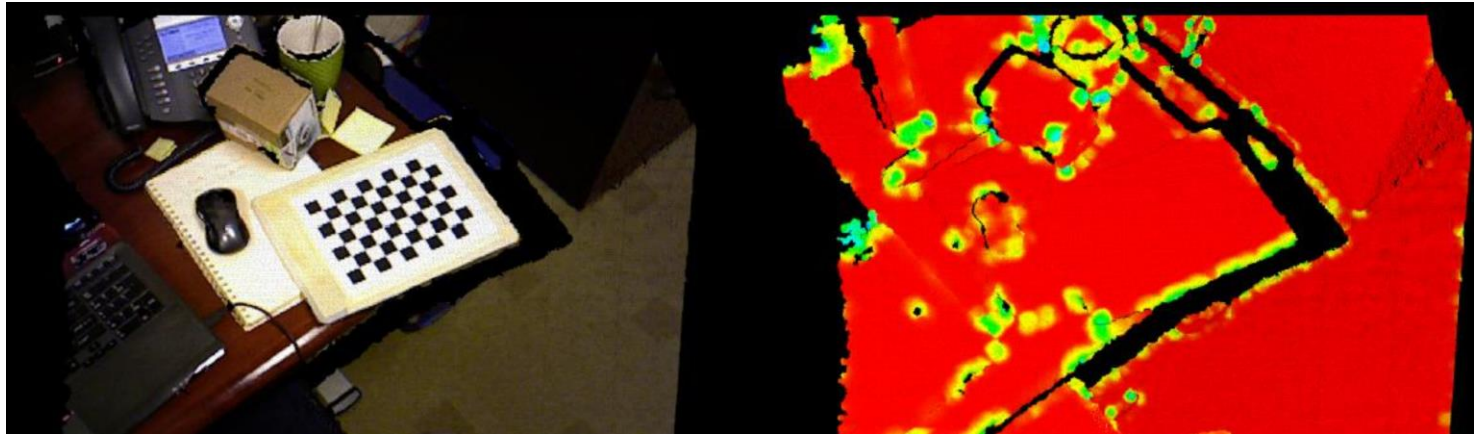
Recap: Harris Operator

- “Harris operator” for corner detection

$$f = \frac{\lambda_- \lambda_+}{\lambda_- + \lambda_+}$$
$$= \frac{\text{determinant}(H)}{\text{trace}(H)}$$

- The trace is the sum of the diagonals, i.e., $\text{trace}(H) = h_{11} + h_{22}$
- Very similar to λ_- but less expensive (no square root)
- Called the “Harris Corner Detector” or “Harris Operator”
- Lots of other detectors, this is one of the most popular

Harris3D

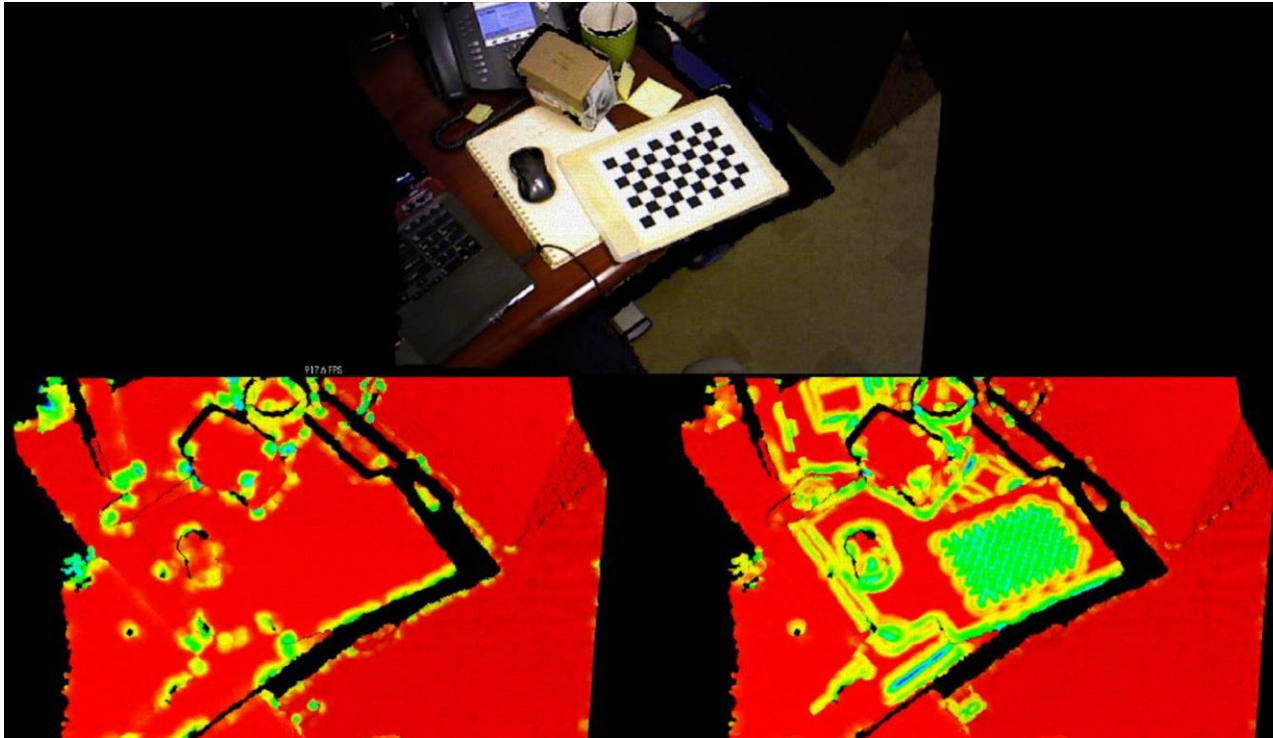


- Replace image gradients with surface normals

$$H = \sum_{i: \mathbf{p}_i \in W} \mathbf{n}_i \mathbf{n}_i^T \quad W: \text{3D window, f.e. sphere}$$

- Harris response: $f = \det(H) - 0.04 \text{trace}(H)^2$
- Lowe response: $f = \det(H) / \text{trace}(H)$
- Noble response: $f = \det(H) / \text{trace}(H)^2$
- Tomasi response: $f = \lambda_{\min}$

Harris5D

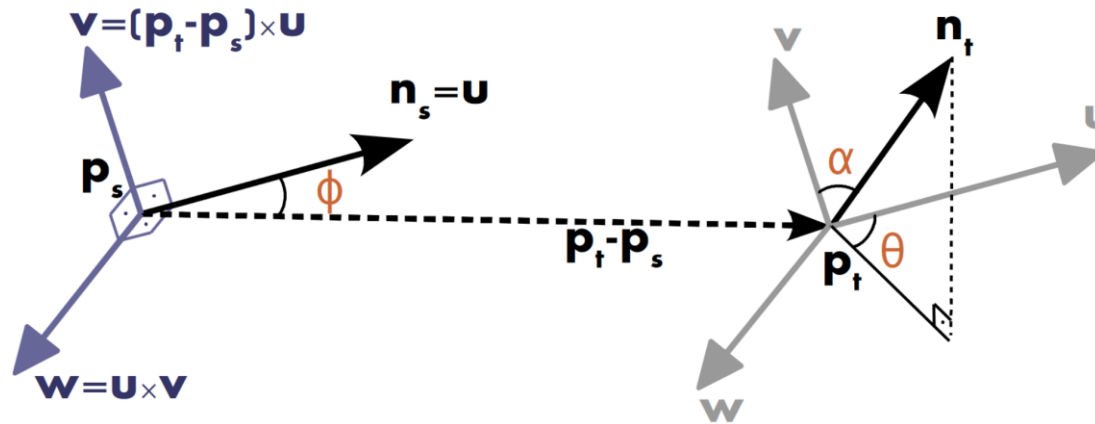


- Can be extended to combined cornerness measure on color and geometry by stacking image gradients and normals

3D Keypoint Descriptors

- Typical approach: Describe local distribution of points and/or surface normals
- Key questions:
 - How to achieve rotation invariance?
 - Description scale of local region?
- Popular descriptors:
 - Fast Point Feature Histograms (FPFH)
 - Signature of Histograms of Orientations (SHOT)
 - 3D Shape Context (3DSC)
 - ...

Surfel-Pair Relations

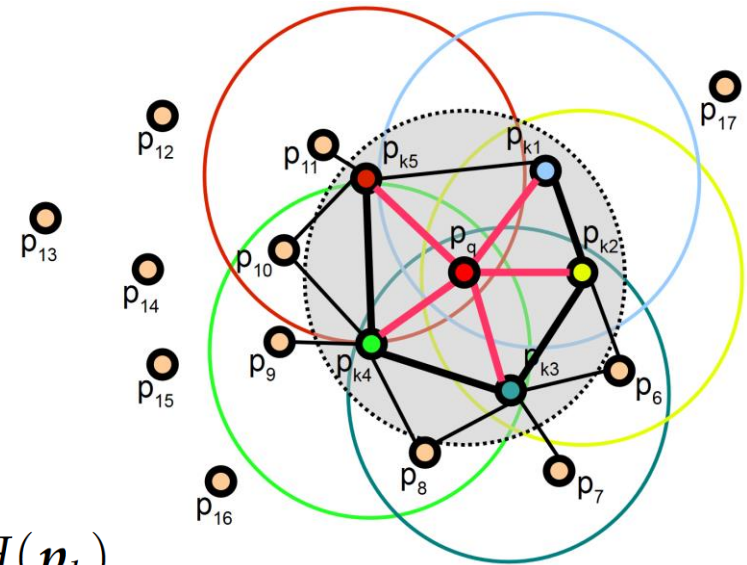
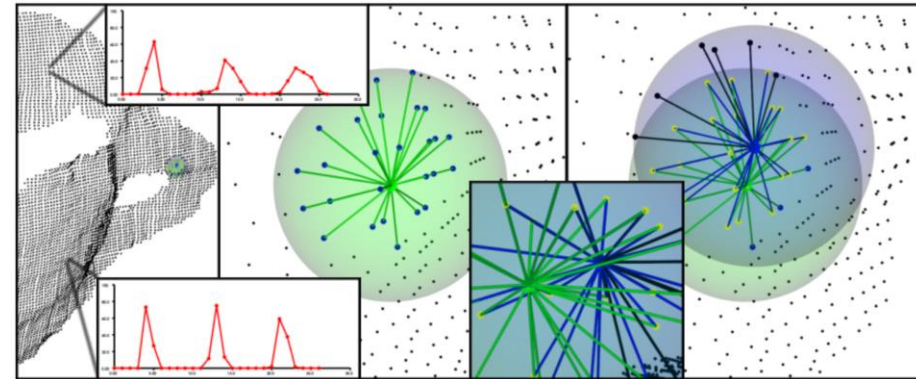


- Surfel $(\mathbf{p}, \mathbf{n}) \in \mathbb{R}^3 \times \mathbb{R}^3$: point \mathbf{p} with normal \mathbf{n}
- Features: geometric relations between two surfels

$f_1 = \mathbf{v}^\top \mathbf{n}_t$	$f_3 = \mathbf{u}^\top (\mathbf{p}_t - \mathbf{p}_s) / f_2$
$f_2 = \ \mathbf{p}_t - \mathbf{p}_s\ _2$	$f_4 = \text{atan2}(\mathbf{w}^\top \mathbf{n}_t, \mathbf{u}^\top \mathbf{n}_t)$
- Construct repeatable local coordinate frame between surfels
- Compute 4 features from constructed frame, normal and point coordinates
- Rotation-invariant features!

Fast Point Feature Histogram (FPFH)

- Describe local neighborhood of a point by histogram of surfel-pair relations
- Fast Point Feature Histogram
 - Compute Simplified Point Feature Histogram for each point from surfel-pair relations between point and local neighbors
 - Accumulate SPFHs in local point neighborhood



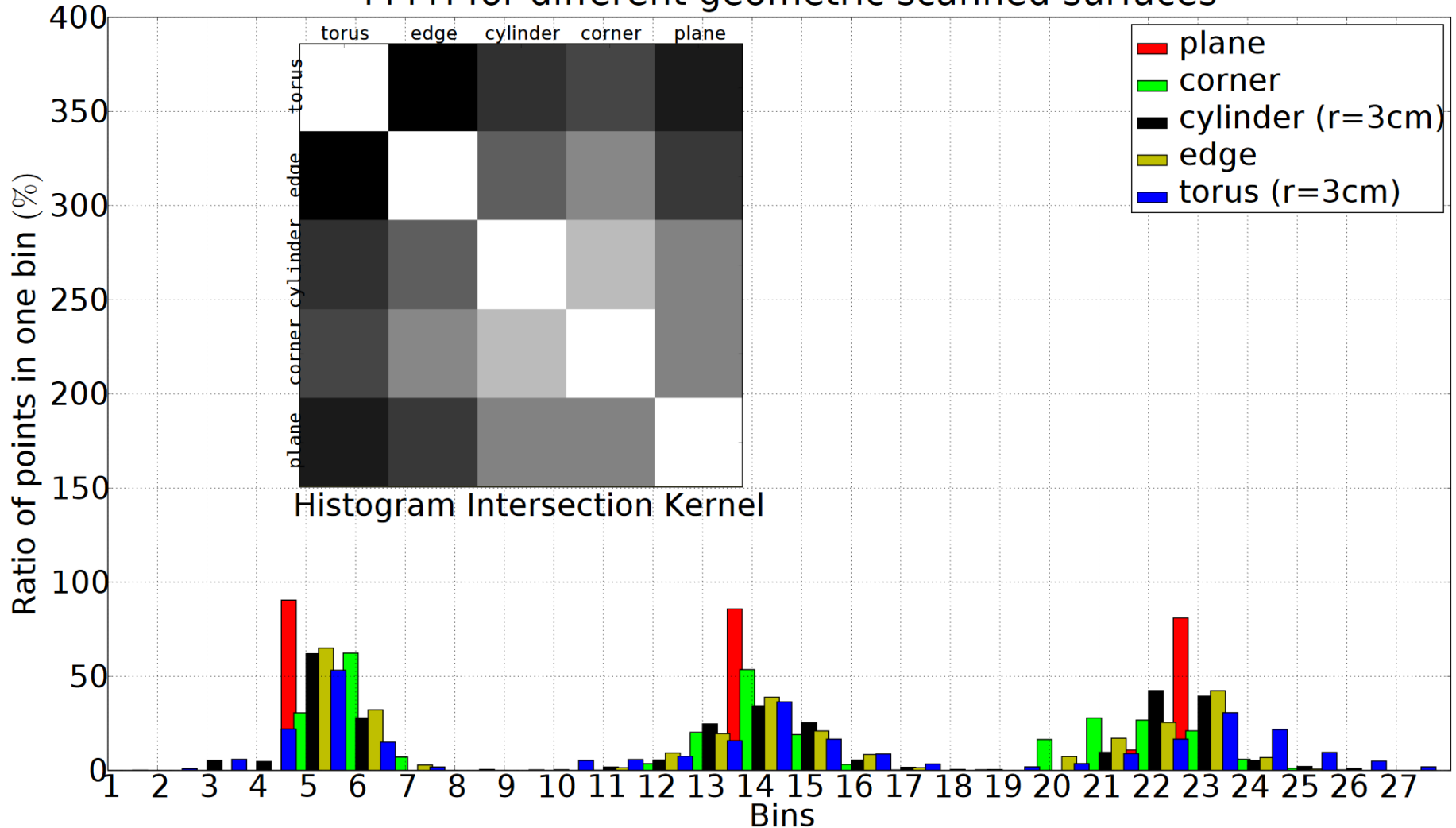
$$FPFH(\mathbf{p}_q) = SPFH(\mathbf{p}_q) + \frac{1}{k} \sum_{i=1}^k \frac{1}{\omega_k} \cdot SPFH(\mathbf{p}_k)$$

- Rotation-invariant

Distance between points

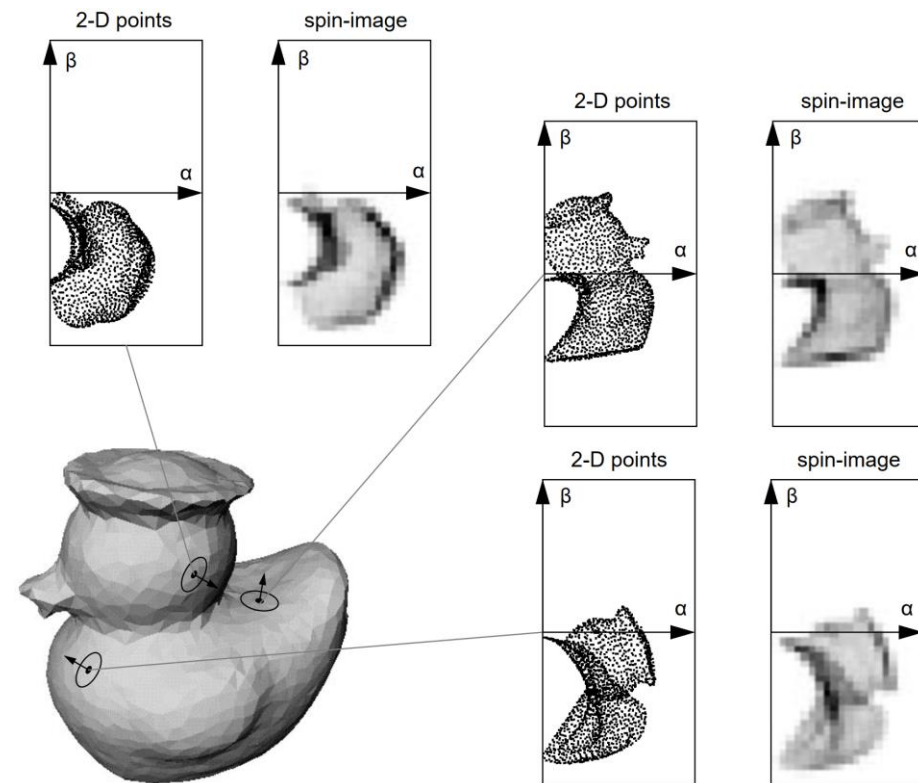
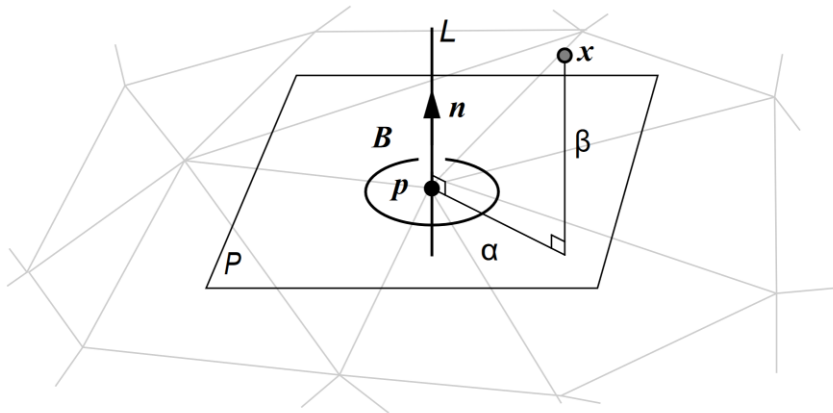
Fast Point Feature Histogram (FPFH)

FPFH for different geometric scanned surfaces



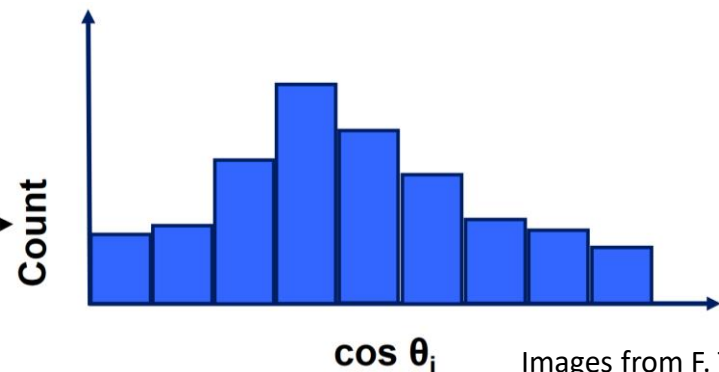
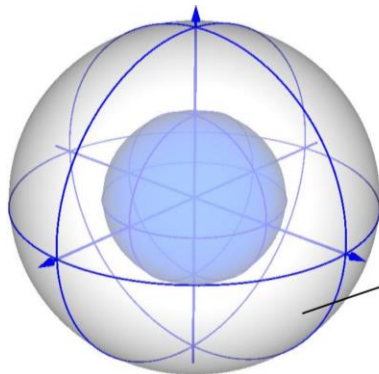
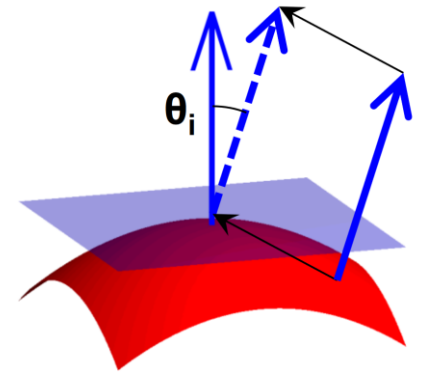
Spin Images

- Describe local point distribution at a keypoint by 2D projection of points on half-plane parallel to surface normal through keypoint
- Also rotation-invariant, but requires a stable normal



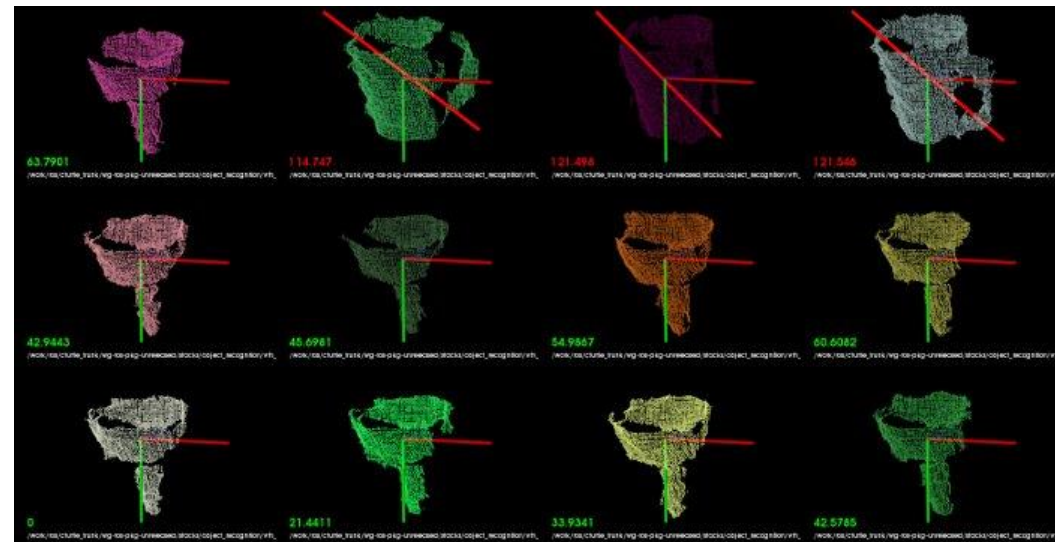
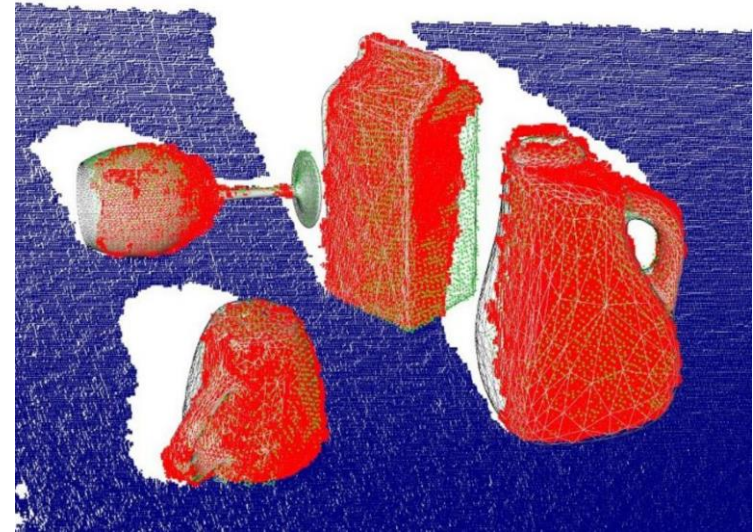
Signature of Histograms of Orientations (SHOT)

- Describe spatial distribution of relative surface orientation around a keypoint
 - Discretize spherical volume around keypoint
 - Discretize spatial bins into angular bins
 - For each neighboring point, determine spatial bin and the angular bin for the angle between its surface normal and the normal of the keypoint
 - Align spherical grid with local reference frame to obtain rotation-invariance



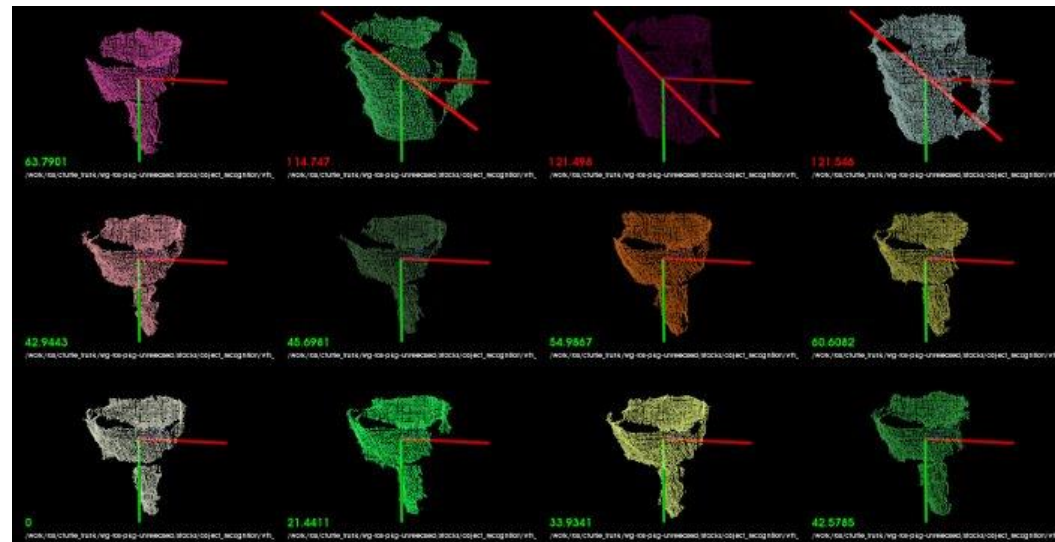
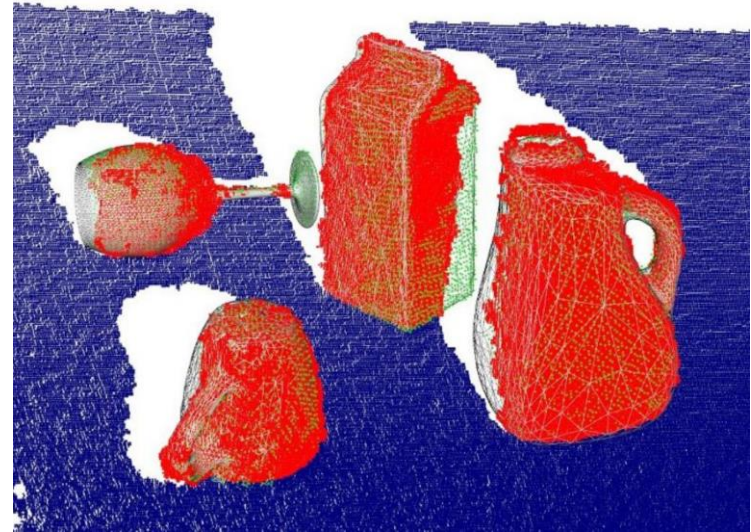
3D Object Detection with Global Features

- Segment RGB-D image into object candidates
- Classify segments using global features extracted on the segments
- Pose estimation by classifying view-point specific features + refinement
 - Simple approach: look-up of N nearest neighbors in kd-tree



Training Objects with Global Features

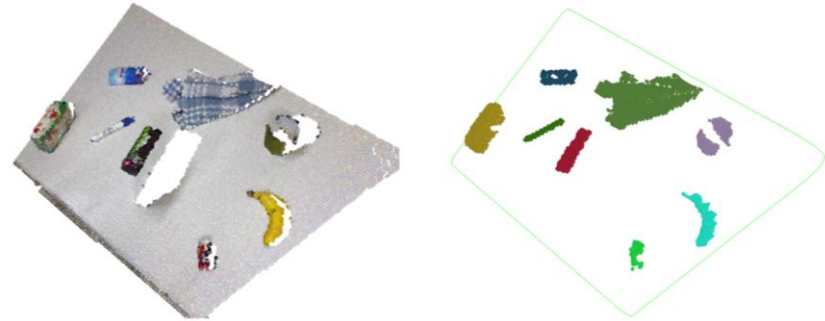
- Render views on CAD models
- Extract global features for rendered views
- Store object indices with global features in an efficient search data structure for the descriptor metric (e.g. kd-tree)



3D Object Segmentation

- Strategy 1: Plane-object (e.g. table-top) segmentation

- Find plane segments in 3D point cloud using RANSAC or Generalized Hough Transform
- Cluster remaining points into object candidates



- Strategy 2: Region-growing

- Grow regions until borders with high curvature (e.g. large smallest eigenvalue of local point covariance) or depth discontinuities



- Many more strategies such as supervoxels, RGB-D superpixels...

Region-Growing Segmentation

- Goal: segment scene into “smooth” regions

- What is smooth?

- Close-by points should have similar normal

$$\exists \mathbf{p}_j \in C_k : \|\mathbf{p}_i - \mathbf{p}_j\|_2 < t_d \wedge \mathbf{n}_i^\top \mathbf{n}_j > t_n \leftarrow \text{e.g. } \cos 10^\circ$$

e.g. 3x point sampling rate

- Prefiltering step: remove points with high curvature above a threshold

- Curvature measure: smallest eigenvalue of the covariance of a local point neighborhood



Image from D. Holz

Region-Growing Segmentation Algorithm

Input: point cloud $P = \{\mathbf{p}_i\}_{i=1}^N$

Output: regions $C_k \subseteq P, \forall k \neq k' : C_k \cap C_{k'} = \emptyset$

1. remove high curvature points from P

2. $k = 0, C_k = \emptyset$

3. while $P \neq \emptyset$

 if $C_k = \emptyset$

 choose random seed point $\mathbf{p}_0 \in P$,
 add it to C_k and remove it from P

 else

 repeat

 for each $\mathbf{p}_i \in P$ if

$\exists \mathbf{p}_j \in C_k : \|\mathbf{p}_i - \mathbf{p}_j\|_2 < t_d \wedge \mathbf{n}_i^\top \mathbf{n}_j > t_n$
 add \mathbf{p}_i to C_k and remove it from P

 until no new point could be added to C_k

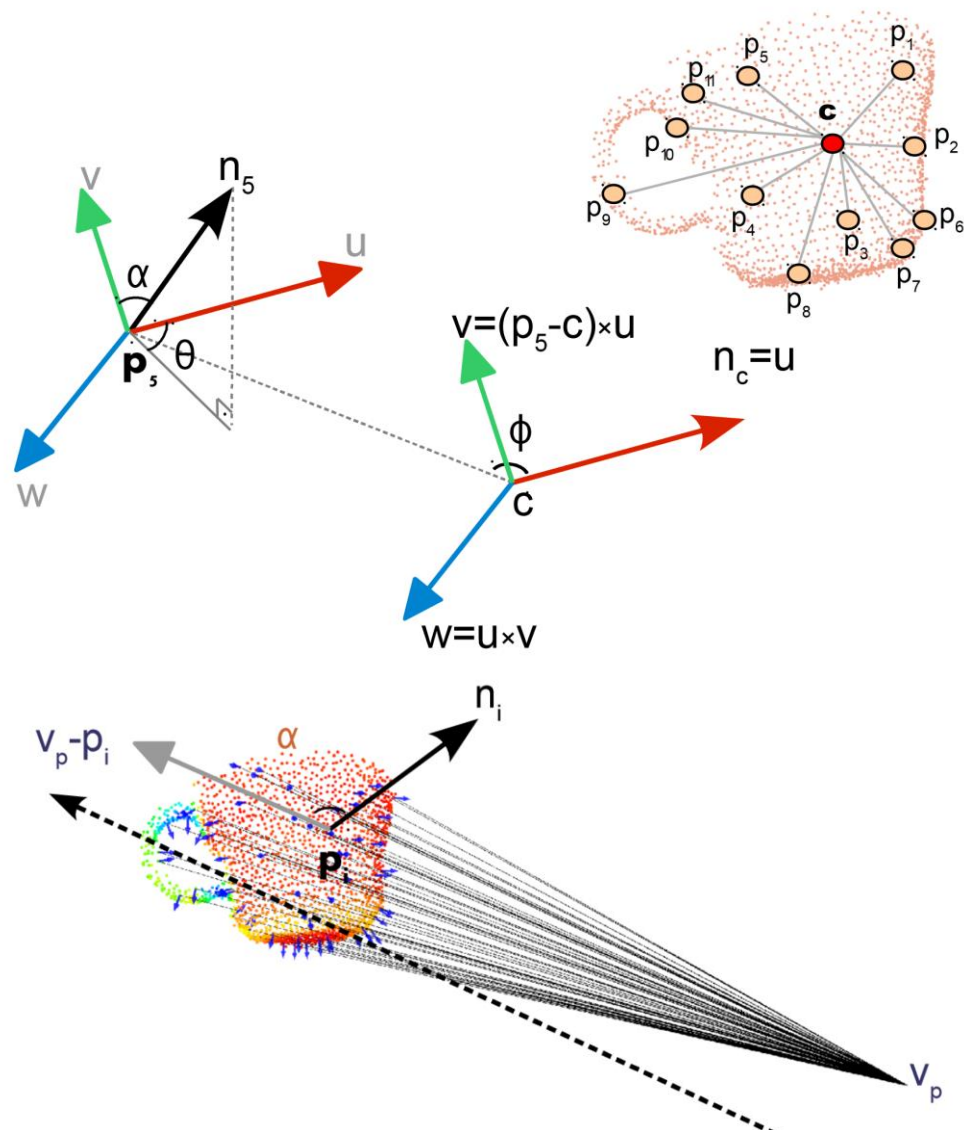
$k = k + 1, C_k = \emptyset$

Global Object Description

- Describe object shape by point/surface normal distribution within segment
- Examples:
 - Viewpoint Feature Histograms (VFH)
 - Cluster Viewpoint Feature Histograms (CVFH)
 - ...

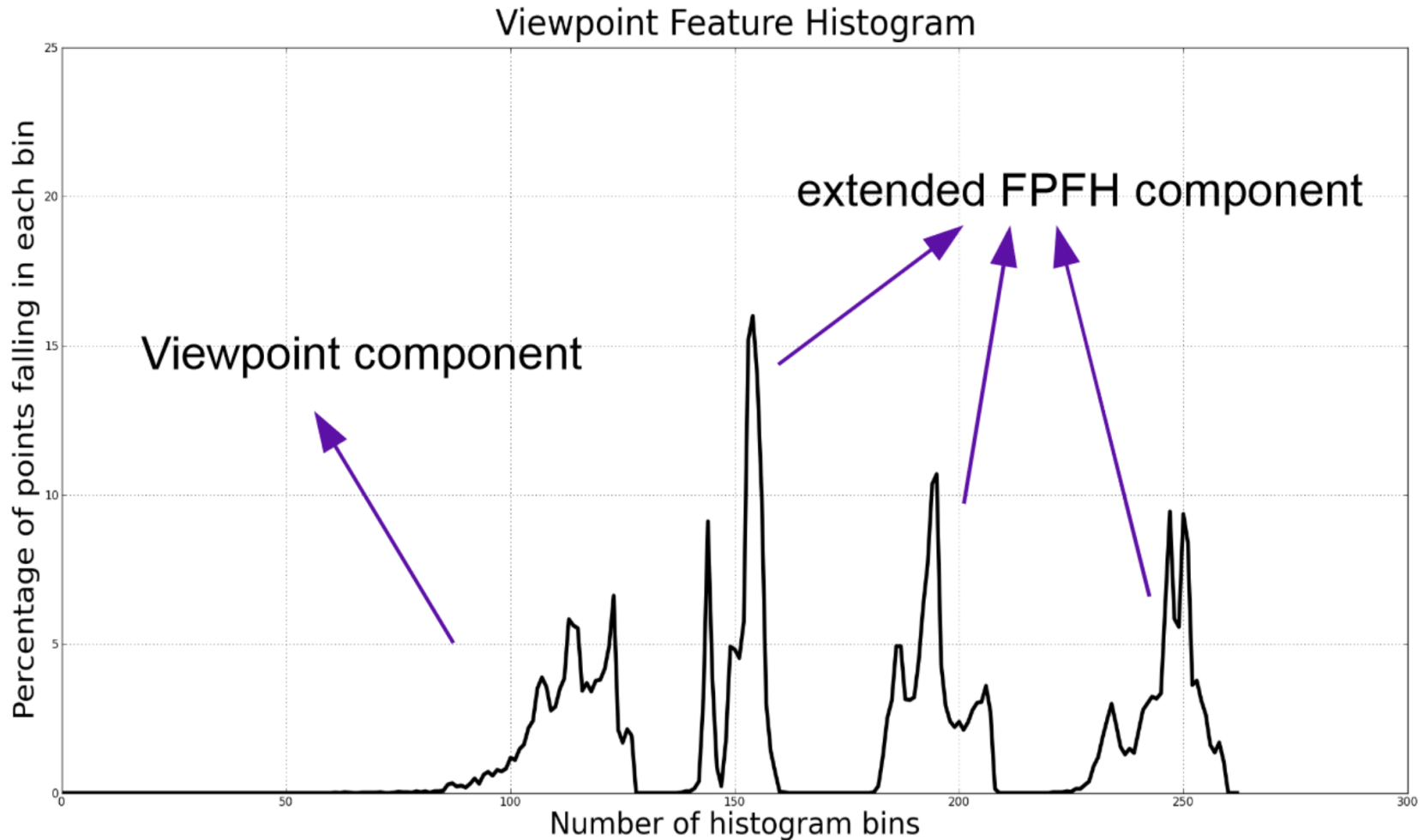
Viewpoint Feature Histogram (VFH)

- Extended FPFH for object segment: histogram over surfel-pair relations to segment centroid \mathbf{c} and average normal \mathbf{n}_c
- Add statistics on angles between point normals in segment and “central” view direction
- Central view direction: view direction to segment centroid, $\mathbf{v}_p - \mathbf{c}$



Images from R. Rusu 2009

Viewpoint Feature Histogram (VFH)



Pose Estimation and Camera-Roll Histogram

- Alignment of the centroids of model and scene segment yields 5-DoF object pose up to rotation around view direction
- Remaining degree of freedom obtained from camera-roll histogram:
 - Project normals in scene cluster to plane orthogonal to central view direction
 - Create camera-roll histogram of angles of projected normal in the plane
 - Find rotation angles as peaks above some threshold in the cross power spectrum between scene and model CRH using the discrete Fourier transform

Clustered Viewpoint Feature Histogram (CVFH)

- Extension 1 (CVFH)
 - Problem of VFH: segment centroid and average normal not robust to partial occlusions
 - CVFH:
 - describes an object view by VFHs \mathcal{H} of multiple part segments \mathcal{S} obtained through region growing
 - The CVFH histogram $h_i \in \mathcal{H}$ of the i -th segment is formed from (45,45,45,45,128)-bin histograms over the variables

$$(\alpha, \phi, \theta, SDC, \beta)$$

- Additional shape distribution component (SDC) describes distance relations between segment centroids \mathbf{p}_c and points in the segment, i.e. for a point $\mathbf{p}_i \in \mathcal{S}$:

$$SDC = \frac{(\mathbf{p}_c - \mathbf{p}_i)^2}{\max((\mathbf{p}_c - \mathbf{p}_i)^2)}$$

Object Recognition with CVFH

- For each object segment independently
 - Perform region growing in the object segment and describe the part segments with a set of CVFHs \mathcal{H}
 - For each $h \in \mathcal{H}$ find the N closest CVFH descriptors in the training set
 - Among the matches select the N best matches according to the metric

$$d(A, B) = 1 - \frac{1 + \sum_{i=1}^{308} \min(A_i, B_i)}{1 + \sum_{i=1}^{308} \max(A_i, B_i)}$$

- Determine 6-DoF pose using CRH matching for the N best matches
- Refine pose estimates using Iterative Closest Points (ICP, later)
- Find best pose estimate by counting matching inliers based on distance threshold

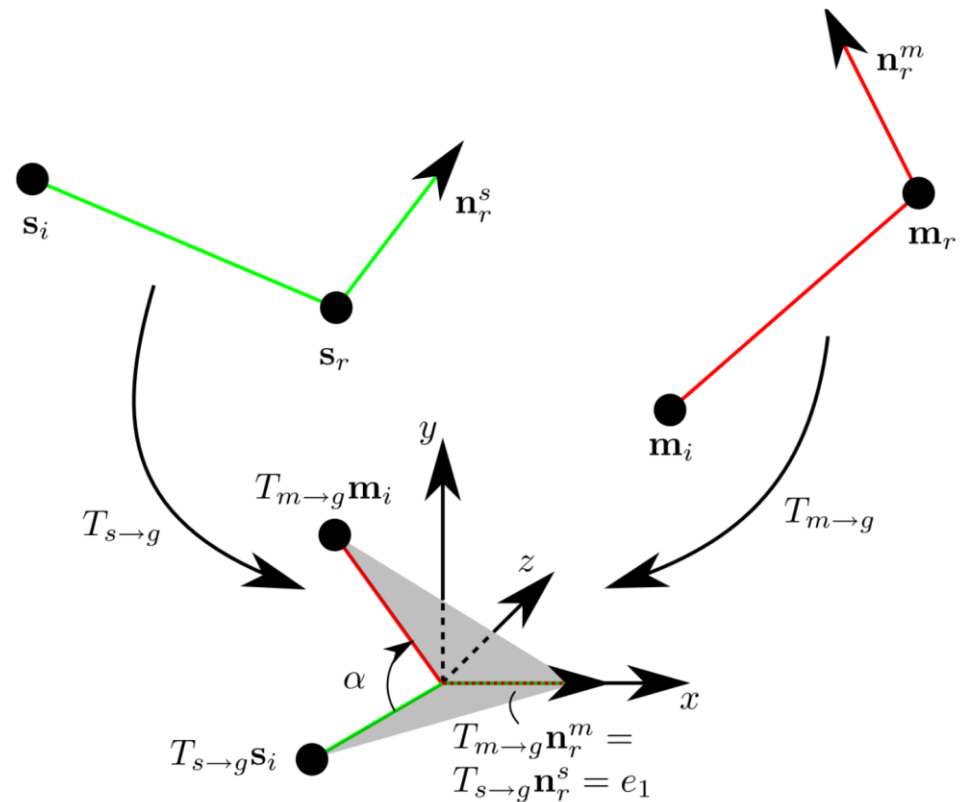
Clustered Viewpoint Feature Histogram (CVFH)

- Extension 2 (OUR-CVFH)
 - Determine local reference frame at each centroid
 - Describe spatial distribution of points in each segment by distance from centroid in the 8 octants of the local reference frame

Surfel-Pair Matching

- If we could identify a corresponding pair of surfels between scene and model, we could directly and uniquely infer the 6-DoF pose

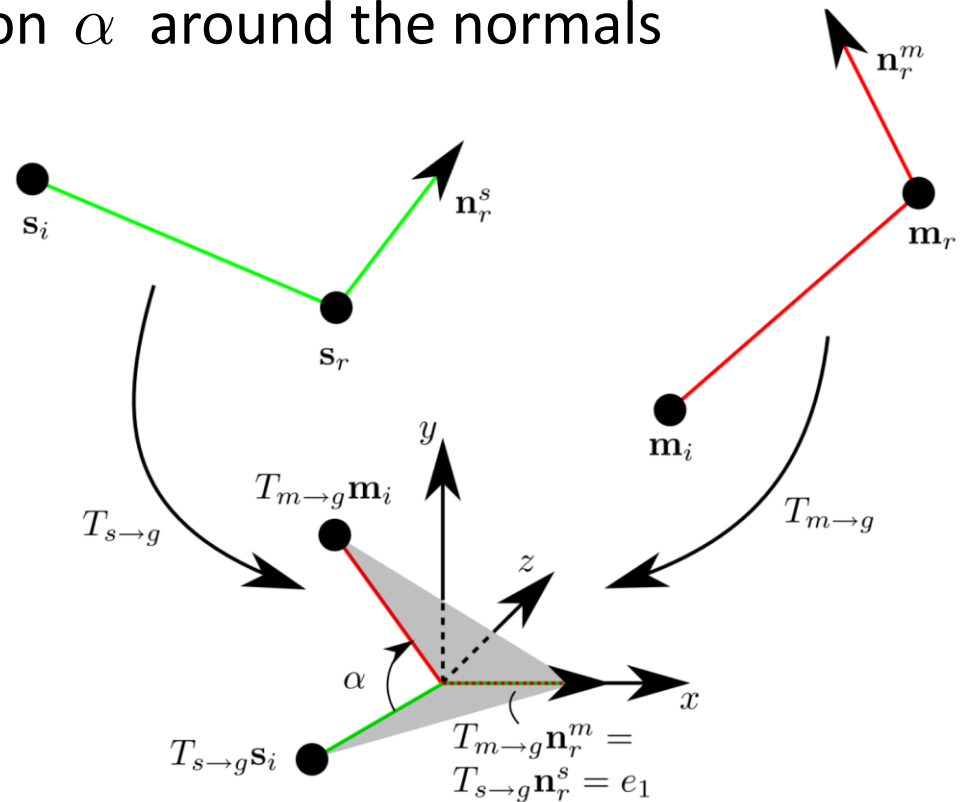
$$\mathbf{s}_i = \underbrace{T_{s \rightarrow g}^{-1} R_{\mathbf{x}}(\alpha) T_{m \rightarrow g}}_{\mathbf{T}(\xi)} \mathbf{m}_i$$



Surfel-Pair Matching

- Align the reference points $\mathbf{s}_r, \mathbf{m}_r$ and their normal $\mathbf{n}_r^s, \mathbf{n}_r^m$ with the x-axis in the world frame using $T_{s \rightarrow g}, T_{m \rightarrow g}$
- Align the secondary points $\mathbf{s}_i, \mathbf{m}_i$ in a common plane parallel to the normal by a 1D rotation α around the normals

$$\mathbf{s}_i = \underbrace{T_{s \rightarrow g}^{-1} R_{\mathbf{x}}(\alpha) T_{m \rightarrow g}}_{\mathbf{T}(\xi)} \mathbf{m}_i$$

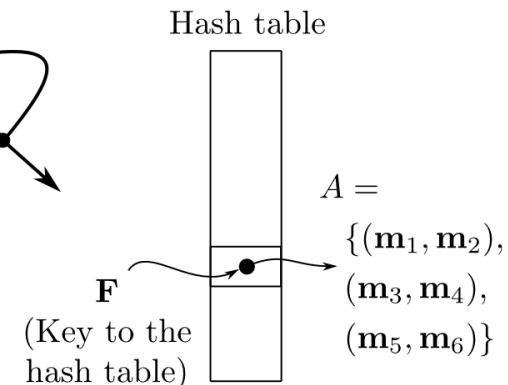
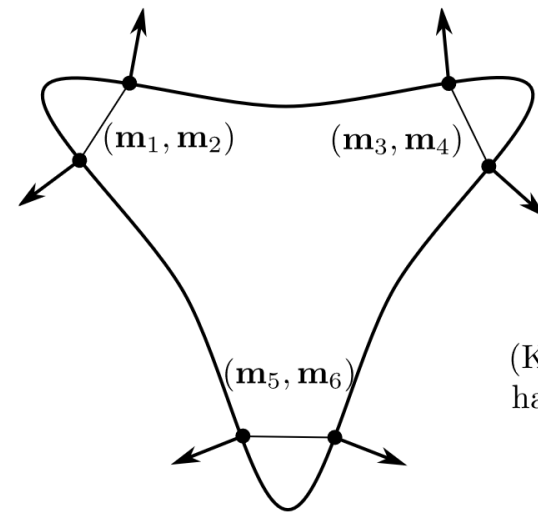
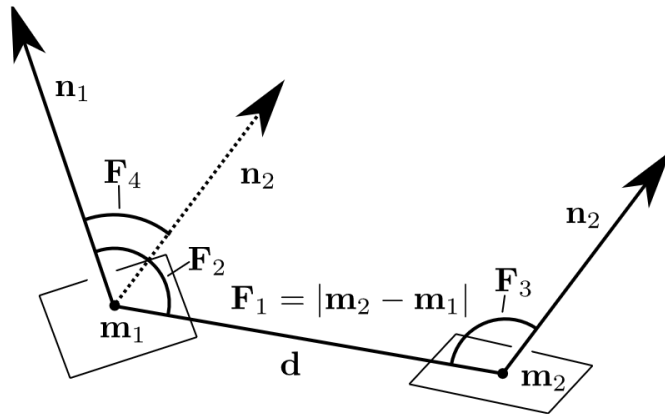


Surfel-Pair Pose Voting

- Match a scene surfel-pair (s_r, s_i) to model pairs (m_r, m_i) according to quantized surfel-pair-relations in a hash look-up table
- Pair all surfels for scene and model description, respectively

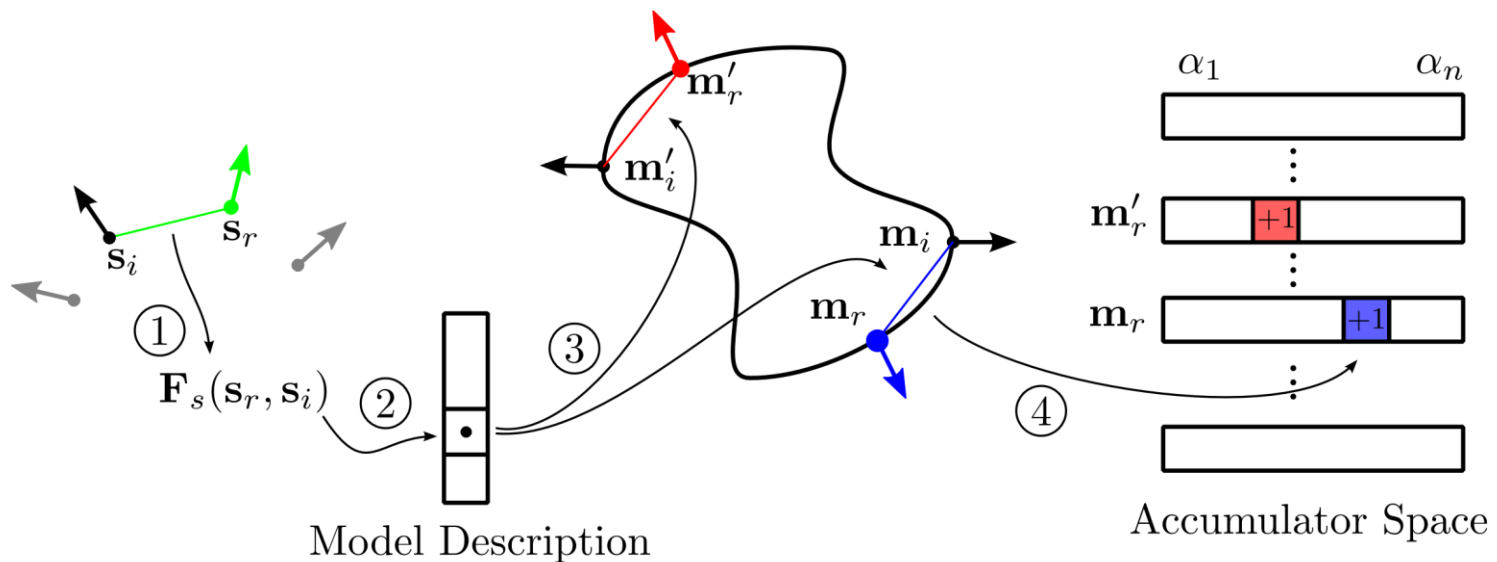
$$\mathbf{d} = \mathbf{m}_2 - \mathbf{m}_1$$

$$\mathbf{F}(\mathbf{m}_1, \mathbf{m}_2) = (\|\mathbf{d}\|_2, \angle(\mathbf{n}_1, \mathbf{d}), \angle(\mathbf{n}_2, \mathbf{d}), \angle(\mathbf{n}_1, \mathbf{n}_2))$$



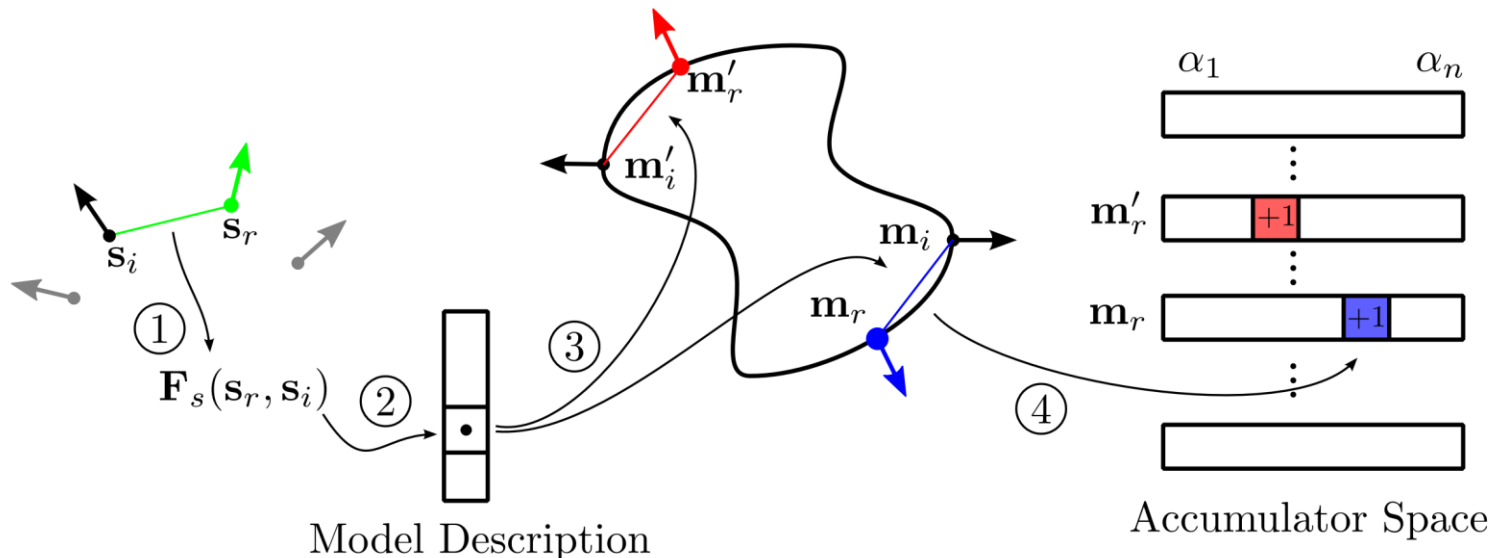
Surfel-Pair Pose Voting

- Cast a vote for each model-pair and corresponding rotation angle α that aligns scene and model pair



Extracting Object Pose from Surfel-Pair Votes

- Object pose could be extracted from maximum peak in Hough voting space for a single reference surfel s_r
- Due to noise and occlusions pose from single surfel is not reliable
- Strategy: Find pose cluster with most votes from multiple reference surfels



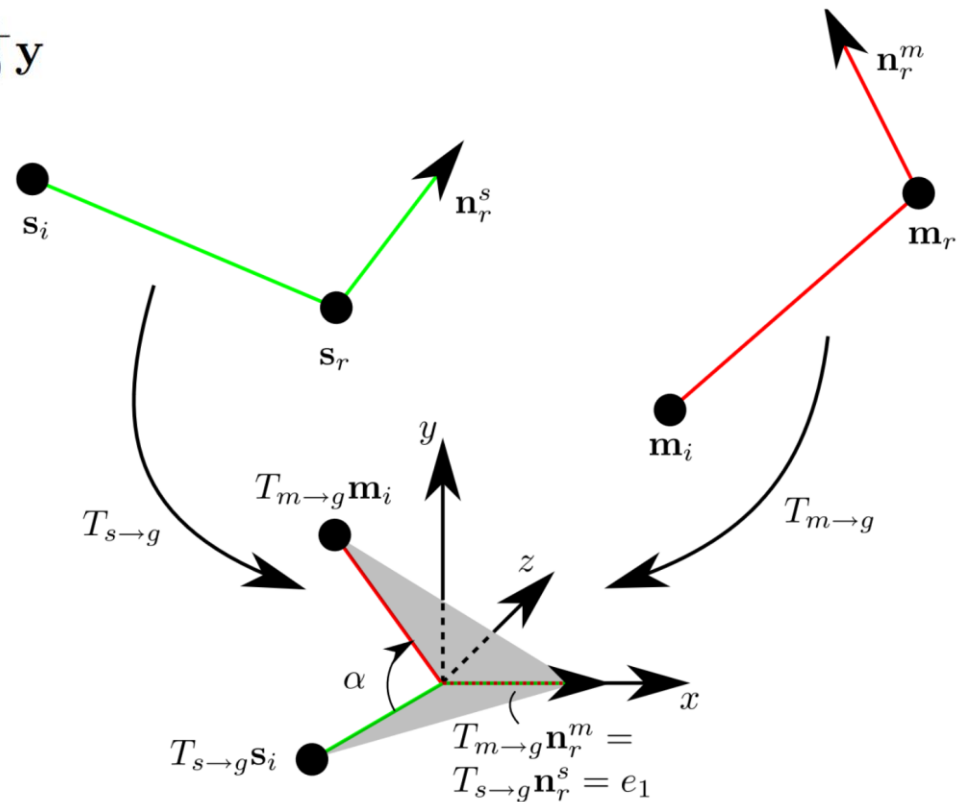
Efficient Surfel-Pair Voting

- α can be split into 2 rotations that align the secondary points in the world x-y plane

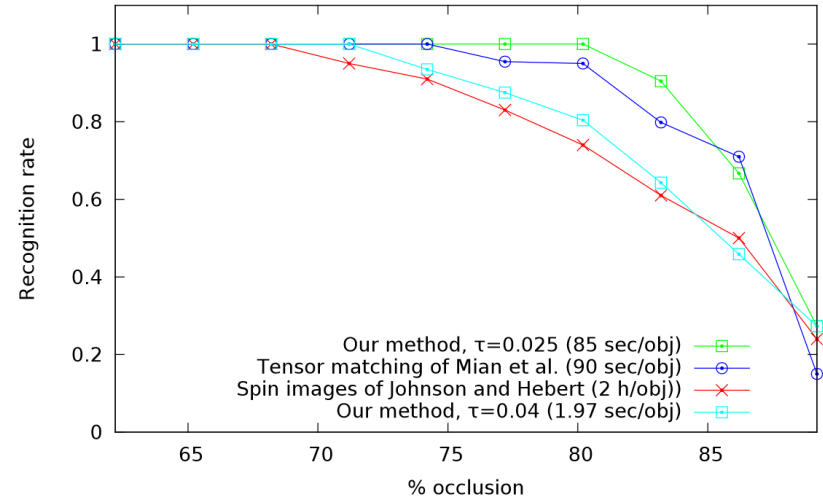
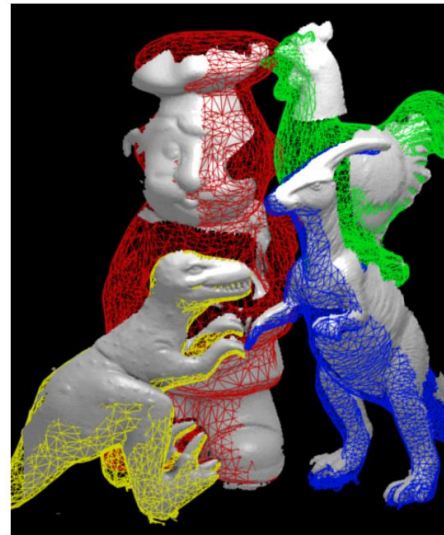
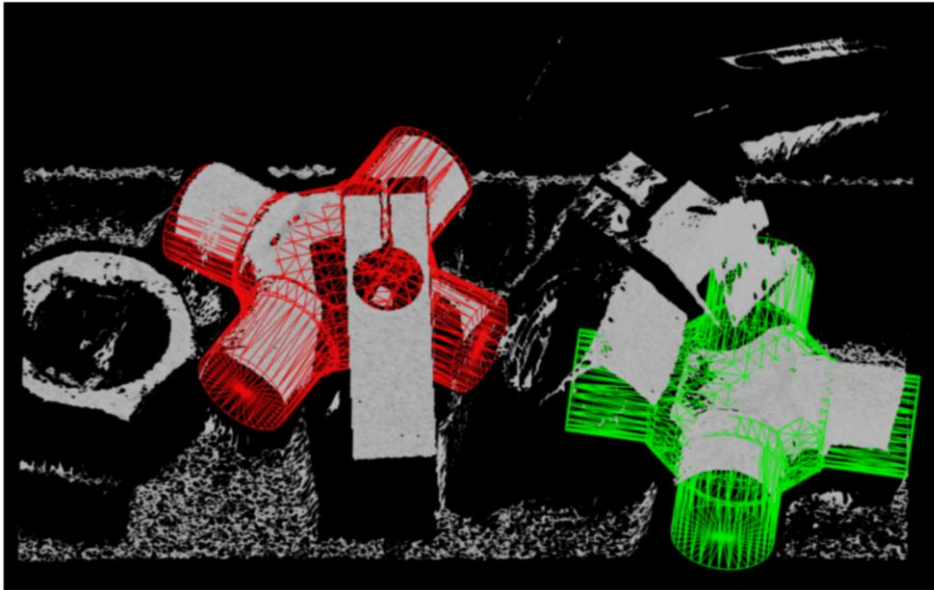
$$R_{\mathbf{x}}(\alpha) = R_{\mathbf{x}}(-\alpha_s)R_{\mathbf{x}}(\alpha_m)$$

$$\alpha = \alpha_m - \alpha_s$$

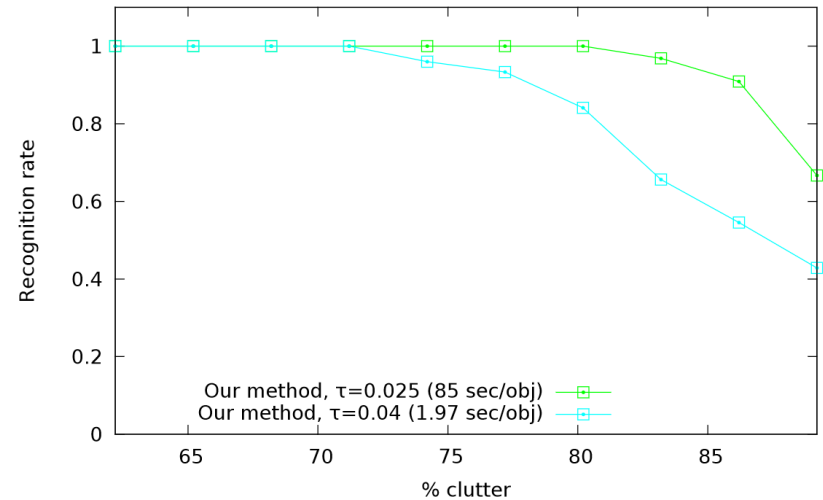
$$\begin{aligned} \mathbf{t} &= R_{\mathbf{x}}(\alpha_s)T_{s \rightarrow g}\mathbf{s}_i = \\ &= R_{\mathbf{x}}(\alpha_m)T_{m \rightarrow g}\mathbf{m}_i \in \mathbb{R}\mathbf{x} + \mathbb{R}_0^+\mathbf{y} \end{aligned}$$



Surfel-Pair Matching Example



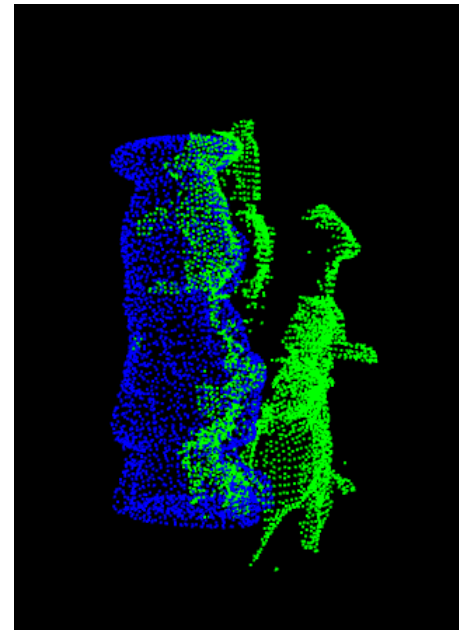
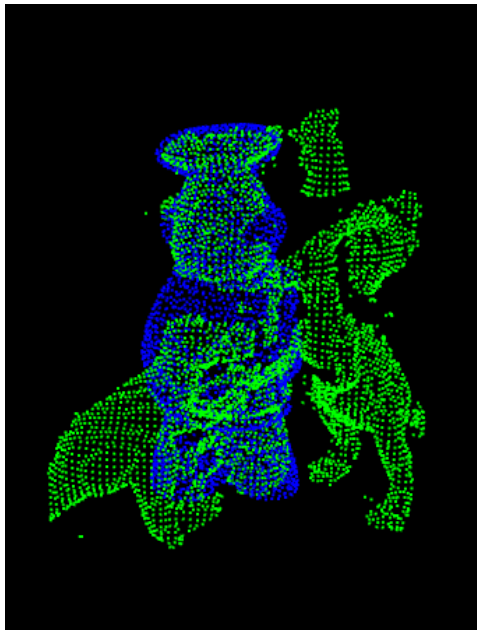
(c)



sampling rate τ for reference surfels in scene and model

Pose Refinement

- So far, local as well as global detection strategies provide only a coarse pose estimate
- Popular strategy for pose refinement: Iterative Closest Points
- Align scene measurements with model point cloud

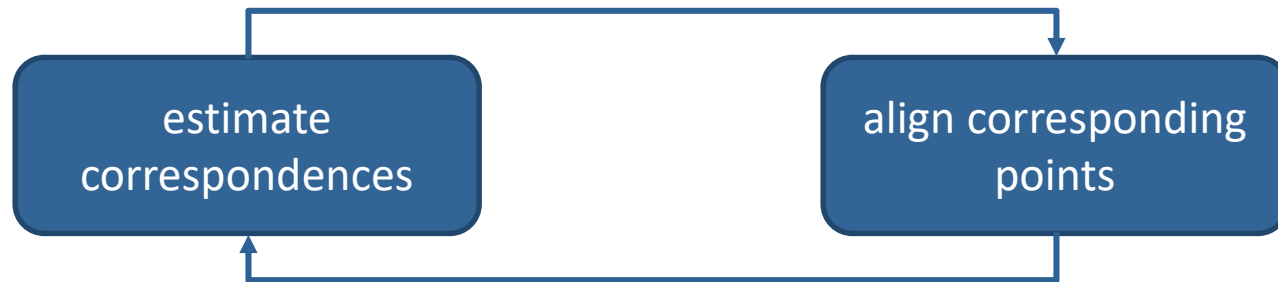


Scene

Model

Iterative Closest Points (ICP)

- Ideally, if we knew the correspondences of points between scene and model, we could directly solve for the 3D-to-3D motion estimate
 - How ?
- ICP: Iteratively and alternately estimate correspondences and pose alignment between point sets $P = \{\mathbf{p}_i\}_{i=1}^N$ and $Q = \{\mathbf{q}_j\}_{j=1}^M$



$$\operatorname{argmin}_c p(P \mid Q, \xi, c)$$

$$\operatorname{argmin}_\xi p(P \mid Q, \xi, c)$$

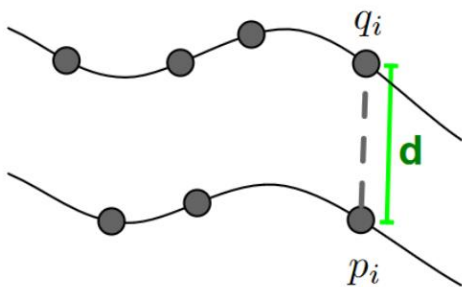
ICP Alignment Objectives

- Alignment objectives: point-point, point-plane, GICP

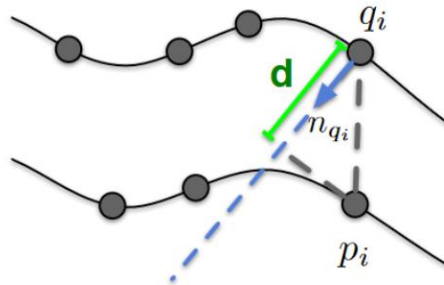
$$E_{\text{point-to-point}}(\mathbf{T}) = \sum_{k=1}^N w_k \|\mathbf{T} \mathbf{p}_k - \mathbf{q}_k\|^2, \text{ and}$$

$$E_{\text{point-to-plane}}(\mathbf{T}) = \sum_{k=1}^N w_k \left((\mathbf{T} \mathbf{p}_k - \mathbf{q}_k) \cdot \mathbf{n}_{\mathbf{q}_k} \right)^2$$

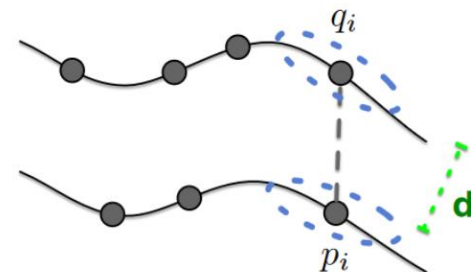
$$E_{\text{Generalized-ICP}}(\mathbf{T}) = \sum_{k=1}^N \mathbf{d}_k^{(\mathbf{T})T} \left(\Sigma_k^Q + \mathbf{T} \Sigma_k^P \mathbf{T}^T \right)^{-1} \mathbf{d}_k^{(\mathbf{T})}$$



(a) Point to point error



(b) Point to plane error



(c) Generalized-ICP

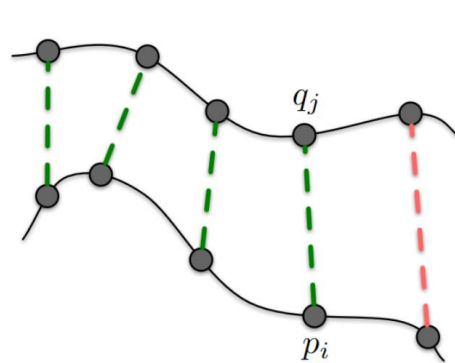
Data Association for ICP

- Closest-points data association
- Use efficient spatial search data structure such as kd-trees

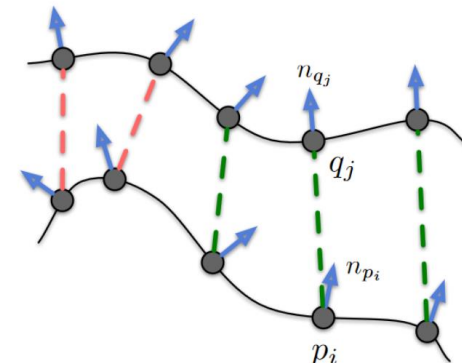
- Popular metric: Euclidean distance

$$\operatorname{argmin}_j \|\mathbf{p}_i - \mathbf{q}_j\|_2$$

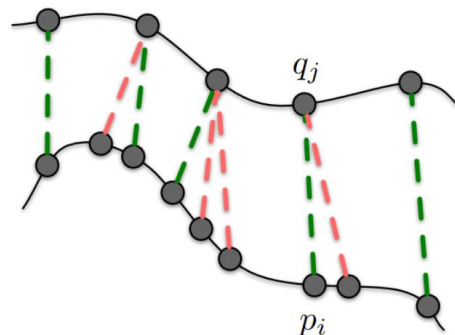
- Determine nearest neighbors and reject outliers



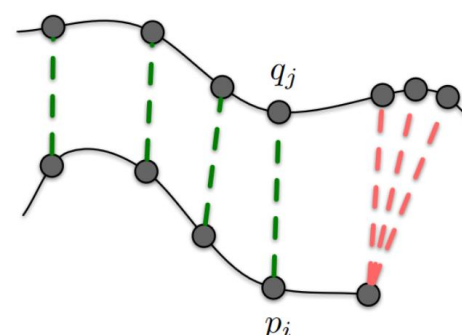
(a) Rejection based on the distance between the points.



(b) Rejection based on normal compatibility.



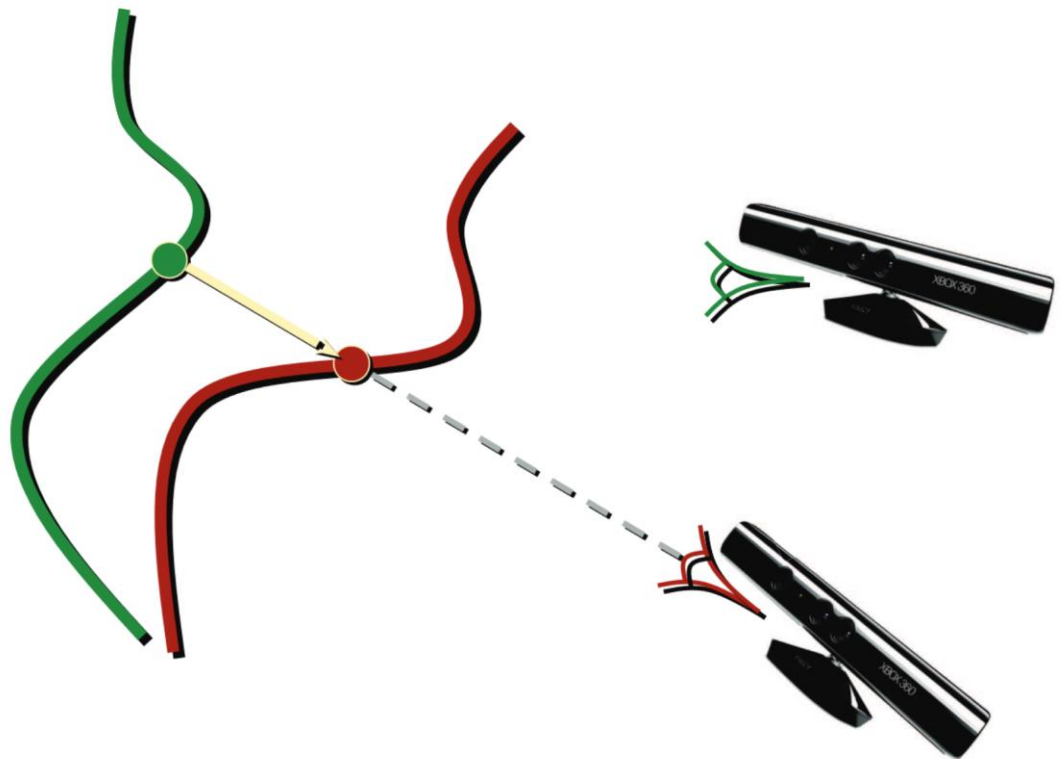
(c) Rejection of pairs with duplicate target matches.



(d) Rejection of pairs that contain boundary points.

Projective Data Association

- For aligning depth or point measurements from a sensor, we can use projective data association
- Warping of measured 3D point
- Analogous association as in direct image alignment!



Keypoint Alignment and ICP Example



Lessons Learned Today

- 3D object detection with local 3D keypoints
 - 3D keypoint detector derived from 2D detector, e.g. Harris3D
 - Intrinsic Shape Signatures detector: points at strong surface curvature
 - 3D keypoint description
 - Extraction of local 3D reference frame from point distribution
 - FPFH, Spin Images, SHOT descriptors
- 3D object detection with global features
 - Object candidate segmentation, f.e. using region growing
 - Object segment classification based on features (VFH, CVFH, ...)
- Object detection based on surfel-pair matching and Hough voting
- Iterative Closest Points algorithm for point cloud alignment

Further Reading

- A. Aldoma et al., Point Cloud Library: Three-Dimensional Object Recognition and 6 DoF Pose Estimation, IEEE RAM 2012
- Holz et al., Registration with the Point Cloud Library, IEEE RAM 2015.
- R. Rusu, Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments, Dissertation thesis 2009
- A. Aldoma et al., OUR-CVFH - Oriented, Unique and Repeatable Clustered Viewpoint Feature Histogram for Object Recognition and 6DOF Pose Estimation, DAGM/OAGM 2012
- Drost et al., Model Globally, Match Locally: Efficient and Robust 3D Object Recognition, CVPR 2010
- A. Johnson, Spin-Images: A Representation for 3-D Surface Matching, Dissertation thesis, 1998

Thanks for your attention!