GEOMETRIC OBJECT RECOGNITION IN RANGE IMAGE FOR GRASPING

Mateo C. M., Gil P., Torres F. - University of Alicante
{cm.mateo, pablo.gil, fernando.torres}@ua.es

Universitat d'Alacant
Universidad de Alicante

GRUPO AUROVA
AUTOMÁTICA, ROBÓTICA Y VISIÓN ARTIFICIAL

Abstract

Nowadays, there is a strong interest in the use of 3D feature descriptors for grasping tasks. The advances on 3D computer vision and 3D sensors allow us to make object recognition, geometric categorization and shape/pose retrieval grasping tasks. Therefore, this work describes a study of two recognition pipelines using 3D normal-based descriptors. On the one hand, descriptors behaviour is evaluated in the recognition process using scenes from Kinect sensors. On the other hand, nowadays we are making an analysis of pose and orientation precision of 3d normal-based descriptors.

Objective

The aim of this work is to know which is the best descriptor for grasping problem. The best descriptor has the best relation among computing and matching runtimes and accuracy.

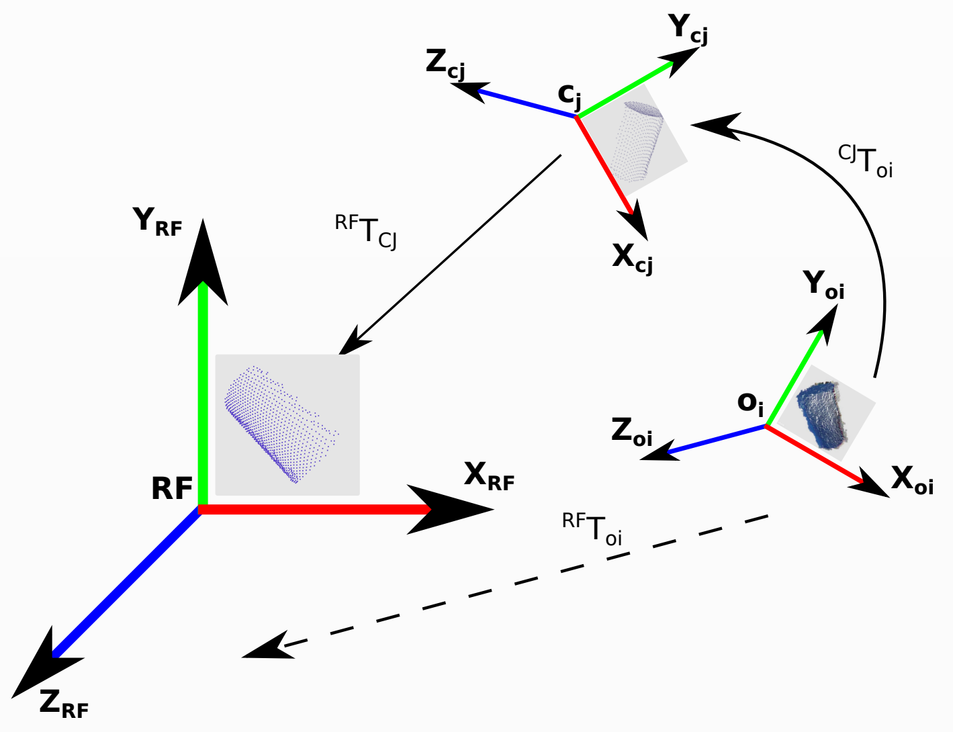
Method

Testing runtime and accuracy

Two stages are done. Firstly, we ealuate the descriptor extraction process with regards to runtime and accuracy. Secondly the matching process is analized between model and scenes. **1) In training stage** the virtual views are created from a set of CAD-Models (cone, cube, cylinder, prism and sphere) from a virtual camera, and their 3D features are computed and saved in a look-up table. **2) In test stage** we makes matching among a test views (captured from Kinect) and the look-up table.

Testing pose estimation

In order to obtain the object pose, the view set of each object are sorted. In addition, its transformation matrix are saved for each view. Regarding the order, each first view is considered the reference frame. Therefore, the transformation, ${}^{RF}T_{oi}$, (6DoF) is done by ${}^{RF}T_{oi} = {}^{RF}T_{cj}\,{}^{cj}T_{oi}$, where c_j is the best matching views (look-up table) with the object view (o_i), ${}^{RF}T_{cj}$ is the transformation save before in the look-up table, and ${}^{cj}T_{oi}$ is obtained in aling step.



References

[1] A., Aldoma, Z.-C., Marton, F., and et al., Tutorial: Point cloud library: Three-dimensional object recognition and 6 dof pose estimation, *IEEE Robot. Automat. Mag.*, 2012

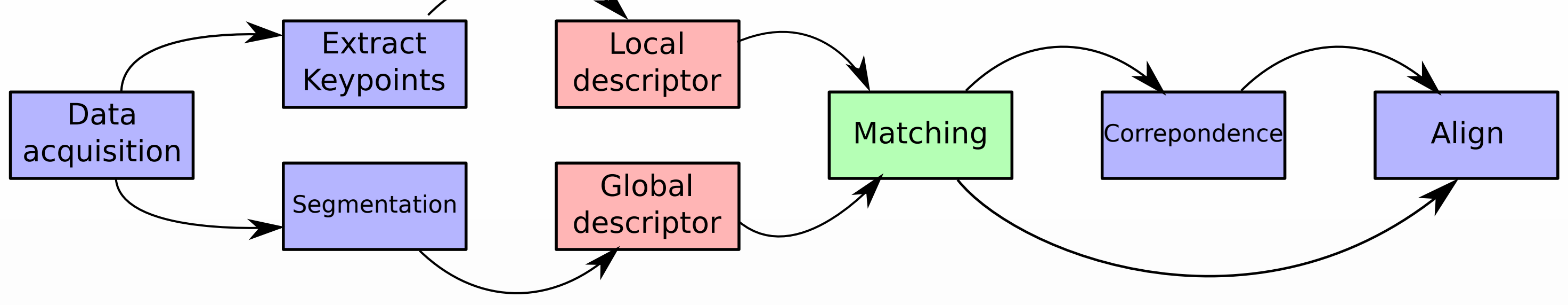
[2] F., Tombari, S., Salti, and L.D. Stefano, Unique sigatures of histograms for local surface description, *Proceedings of the 11th European Conference on Computer Vision: Part II, ECCV'10*, 2010

[3] A., Aldoma, M., Vicze, N., and et al., CAD-Model recognition and 6dof pose estimation using 3d cues, *IEEE International Conference on Computer Vision Workshops, ICCV'11*, 2011

[4] C.M. Mateo, P. Gil, and F. Torres, A Performance Evaluation of Surface Normals-Based Descriptors for Recognition of Objects using CAD-Models, *ICINCO'14*, 2014

Descriptor framework

Typically, recognition process is categorized such as local and global recognition pipelines [1]. Both have two critical steps, descriptor extraction step (red) and matching step (green). Firstly, two descriptors highlight within both set, *SHOT*[2] (local descriptor) and *CVFH*[3] (global descriptor). Both have different level of complexity [4]



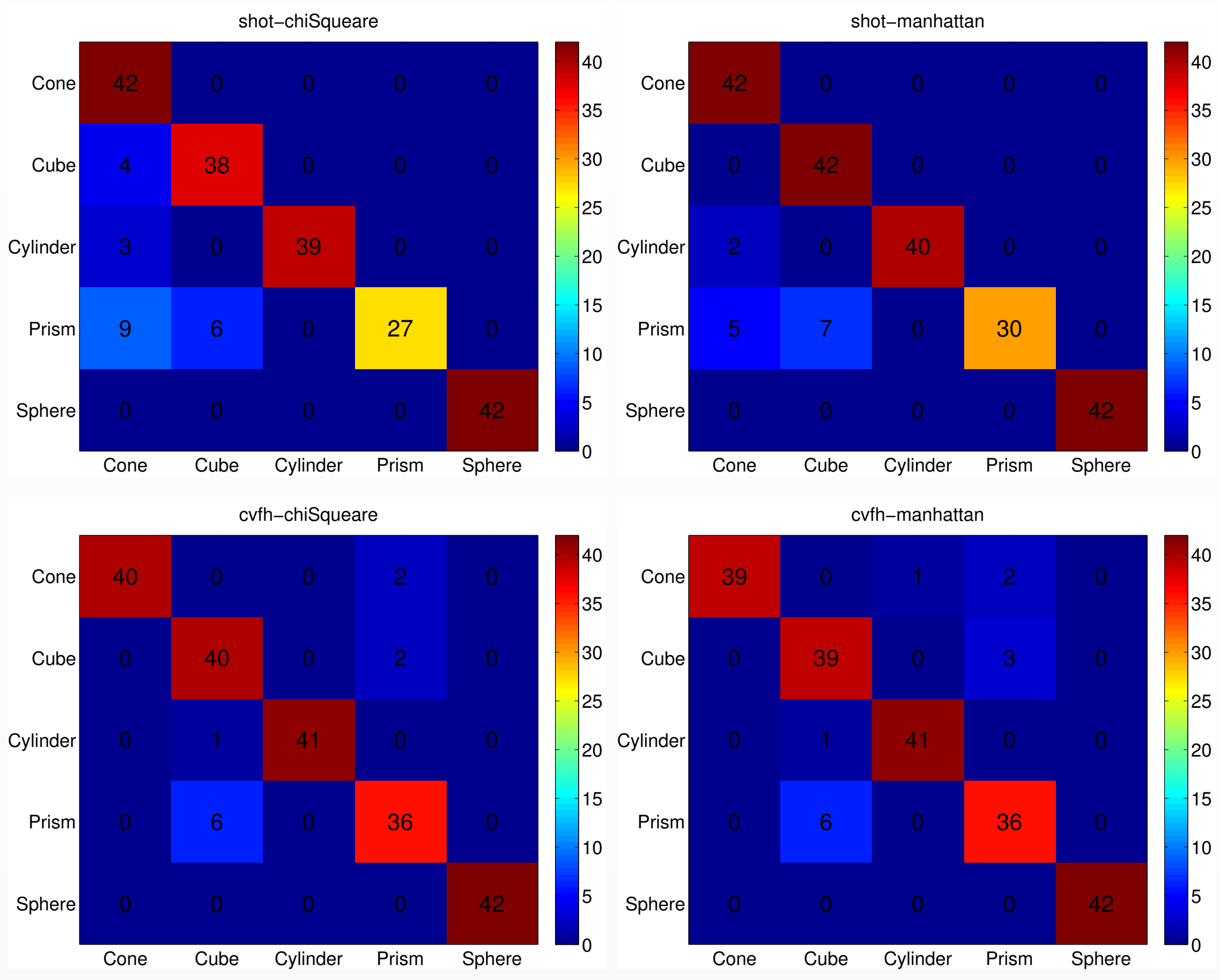
SHOT A partitioned spherical grid is used as local reference frame. For each volume of the partitioned grid, a signature of the amount of $\cos(\theta_i)$ between the normal at each point of surface and the normal at the query feature point is computed.

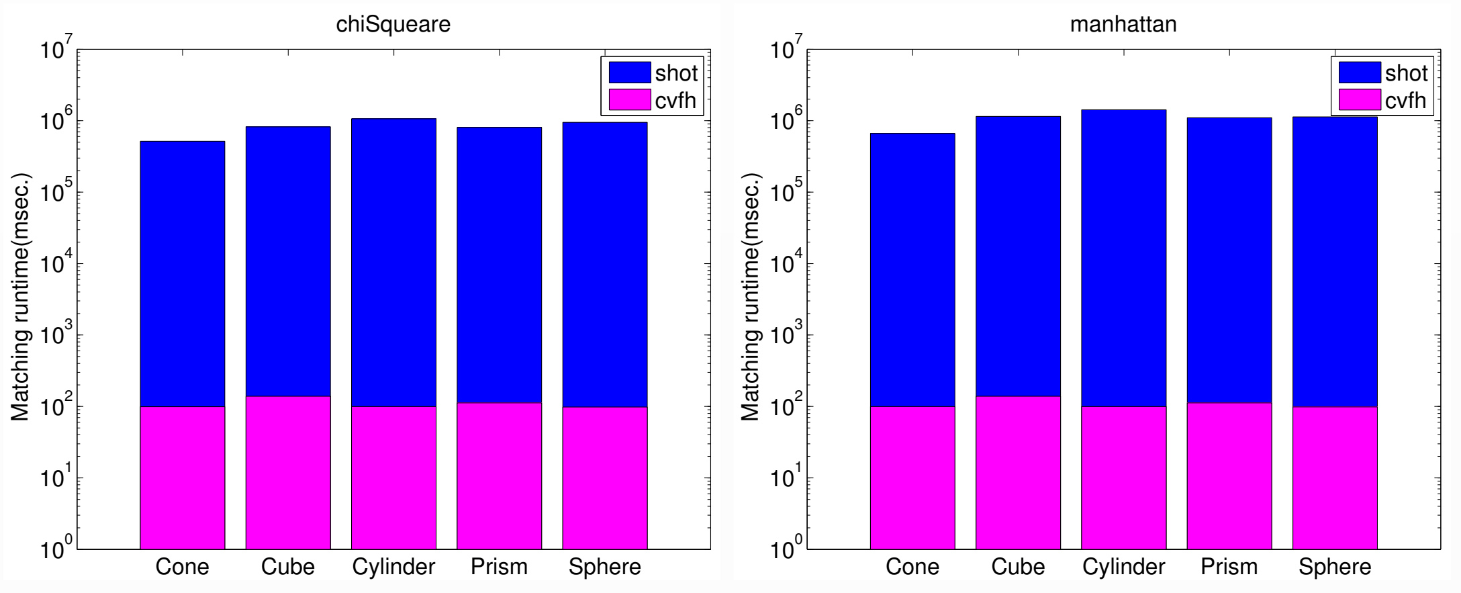
CVFH The basic idea is to identify an object from splitting it in a set of smooth and continuous regions. For each of these regions is computed its *VFH*[1] descriptor. *CVFH* describes a surface as a histogram in which each histogram item represents the centroid to surface and the average of the normals among all points of surface

Secondly, the matching step accuracy depending on the distance metric used. In computer vision there are many common metrics to compare similarity. Two distance metrics stand out among others. **1)** $d_{L1}(p,q) = \sum_{i=1}^{n} p_i - q_i$ **2)** $d_{\chi^2}(p,q) = \sum_{i=1}^{n} \frac{(p_i-q_i)^2}{p_i+q_i}$

Results

A total of 42 views/model and 32 scenes/object were used to make the experiments. With the aim of evaluate the robustness and precision of the descriptors, the recognition pipeline is avoid to the keypoint extractor step. A step of segmentation was done in both recognition pipelines. Bottom is shown 4 confusion matrix with matching results, they have a low dependency of the two metrics used, and we can see how CVFH has the best results in our experiments. In addition, we can see the great differences between matching runtime depending on the descriptor used. Finally, it is show a table with the runtime descriptors.





	Cone	*Cube*	*Cylinder*	*Sphere*	*Prism*
SHOT	11.6±0.15(ms)	20.3±0.21(ms)	24.6±0.31(ms)	16.9±0.15(ms)	18.9±0.36(ms)
CVFH	0.41±0.06(ms)	3.82±.0.05(ms)	1.77±0.05(ms)	1.06±0.07(ms)	3.63±0.11(ms)

Acknowledge

The research leading to these result has received funding from the Spanish Government and European FEDER funds (DPI2012-32390) and the Valencia Regional Government (PROMETEO/2013/085).