Model-Based Visual Self-Localization using Geometry and Graphs

D. Gonzalez-Aguirre, T. Asfour, *E. Bayro-Corrochano, R. Dillmann
Institute of Computer Science and Engineering, University of Karlsruhe-Germany
**CINVESTAV, Guadalajara-Mexico*
{*gonzalez,asfour,dillmann*}*@ira.uka.de, *edu@gdl.cinvestav.mx*

Abstract

In this paper, a geometric approach for global self-localization based on a world-model and active stereo vision is introduced. The method uses class specific object recognition algorithms to obtain the location of entities within the surroundings. The perceived entities in recognition trials are simultaneously filtered and fused to provide a robust set of class features. These classified perceptions which simultaneously satisfy geometric and topological constraints are employed for pruning purposes upon the world-model generating the location hypotheses set. Finally, the hypotheses are validated and disambiguated by applying visual recognition algorithms to selected entities of the world-model. The proposed approach has been successfully used with a humanoid robot.

1. Introduction

The self-localization capability is essential for autonomous systems, like humanoid robots, operating in *built-for-humans* environments, where the use of vision is the only natural approach. In those structured environments the geometrical and topological interrelations of the elements provide substantial advantages for feature extraction, object recognition and self-localization.

Self-localization can be categorically divided into *global* and *fine* localization [8]. The first one considered in this paper determines the position and orientation of the robot (*pose*) within a world coordinate system U, see Fig.6. The second one deals with the continuous state (*dynamic-pose*) of the robot.

So far, vision-based self-localization approaches ([4],[8]) have been commonly conceptualized as the extraction and processing of image features, which by means of recursive state estimations provide the continuous location of the camera(s). However, partially significant visual landmarks (stored scale invariant features) and assessed poses (only linked to the *unknown* initial *pose*) provide insufficient useful information for real applications. In these cases, the robot requires plenary environmental information (vision-to-model coupling) to actively interact with its environment, i.e. solving assertions concerning the status of its world, visual planning, grasping, etc. These limitations can be overcome with a proper mechanism which provides fast and reliable global localization by systematically exploiting the intrinsic natural constraints accessible through an effective and consistent world-model representation.

2. Outline of Visual Self-Localization

This approach[1] consists of a collection of active visual *perception-recognition* components, a *world-model* and a hypotheses *generation-validation* apparatus, see Fig.1.

2.1. Object Recognition

The basic inputs are perceived-recognized objects, i.e. *Percepts*, see Fig.1-a. For instance, but not limited to handles, doors or windows in a building, see Fig.2. The advantage of using *class-based* object recognition schemas has been previously exploited [10]. In this way not only fast and robust methods are applied but also the data association between features and model entities is partially[2] solved. In contrast, general feature approaches [8] lack of model association while offering poor reliability compared to those specific ones.

In this approach, doors and handles are robustly recognized by means of gaussian classification over characteristics feature spaces extracted from class specific descriptors of the eigenvectors[3] of corresponding color-segmented regions from stereo images, see Fig.2-b.

[1] Additional material at *http://i61www.ira.uka.de/users/gonzalez/*
[2] Up to the class-instance association level.
[3] From the covariance matrix of the color-clustered regions.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

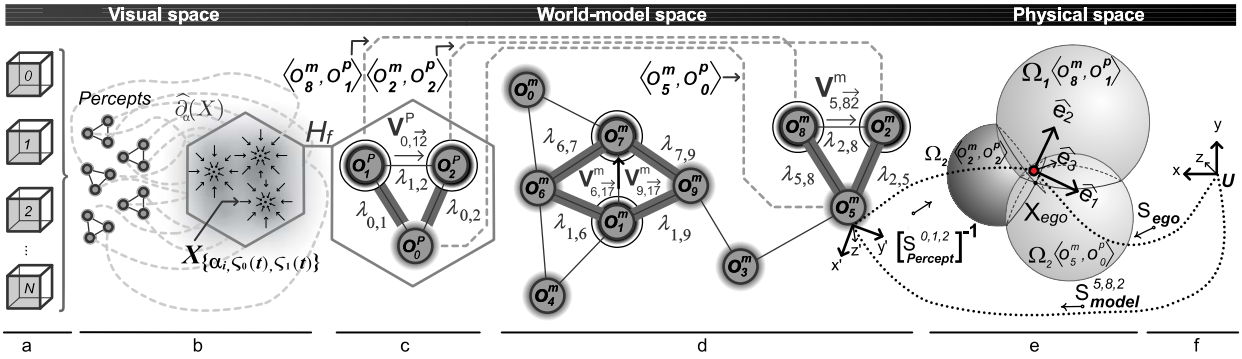

Figure 1. The model-based visual self-localization approach. a) Visual perception-recognition components. b) Recognition fusion. c) Percept subgraphs. d) World-model at graph pruning. e) Hypotheses generation-validation. f) Pose estimation.

Subsequently, the left-right cross match using size, position, orientation, perpendicular distance to the epipolar line and standard disparity constraints [7] allows powerful rejection of the remaining outliers.

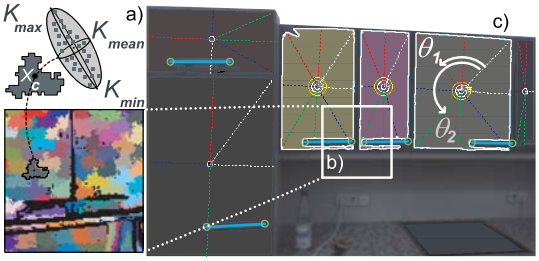

Figure 2. Class specific object recognition.

The used growing region generation criteria admit pixels considering the size of the region ($S_i < S_{max}$) and the volume ($V_i < V_{max}$) of the ellipsoid expanded by points K_{mean}, K_{min} and K_{max}, see Fig.2-a. Afterwards, 5D feature vectors

$$F_i := [X_c, K_{mean}^T]^T, \tag{1}$$

are used to compute the mean shift algorithm [3] for clustering regions, which present chromatic and spatial proximity, i.e. *Blobs*. In this manner, the shifting method benefits from processing less (but still coherent) features. Experimental results show a reduction from 2000-3000 *ms* to 100-200 *ms*.

Subsequently, the blobs representing doors were recognized by means of the descriptor

$$V_d := [\frac{d_{max}^{\rho_1}}{d_{min}^{\rho_1}}, \frac{d_{max}^{\rho_2}}{d_{min}^{\rho_2}}, \frac{S_i}{L_i}, \theta_1, \theta_2]^T, \tag{2}$$

which components involve the ratios of the maximal-minimal lengths (the projection from pixels along each blob axis), the elasticity (ratio area-perimeter $\frac{S_i}{L_i}$) and the angles θ_1 and θ_2 (see Fig.2-c), which restrict the convexity and skewness of the blob. Afterwards, the descriptor of the handle is given by

$$V_h := [\frac{\rho_1}{\rho_2}, \arccos(\widehat{Q} \cdot \widehat{K}_{mean}), ||Q - K_{mean}||]^T, \tag{3}$$

where ρ_1 and ρ_2 denote the eigenvalues of the blob. Here, the components represent the axis-compactness and the angle-length discrepancy between the mean color of the blob and the ideal color of the handle. The remaining outliers are discarded using the *Shi-Tomasi* response function [9] to verify the existence of two parallel edges meeting at clear corners on both ends.

In addition, many specific recognition components may be added to improve the performance of the system at graph filtering by increasing the amount of partitions of the graph, i.e. reinforcing constraints and increasing pruning.

2.1.1 Ego Perception and Recognition Fusion. Despite the robustness of the class specific algorithms and due to certain *phenomena* (varying illumination, singularities of the field of view, etc.) false positives might sporadically occur. In order to anticipate these situations, all recognized objects are related to the ego center X_{ego} of the robot by considering its kinematic configuration while executing the *scanning-strategy*, i.e. the planned trajectory for capturing stereo images using the head of the robot. The resulting registration data structures include frame identifier, type of the percept and its 3D location.

In the next phase, fusion begins by calculating distances between percepts of same type in frame t and those found within the frame range $[\varsigma_0(t),\varsigma_1(t)]$, herein the functions $\varsigma_0(t)$ and $\varsigma_1(t)$ provide the first and last frame that share visual space with the frame t. As a result, percepts which closest distance to other percepts exceeds the threshold[4] ϖ are ignored.

Within this phase, there is a tacit underlying 3D multimodal spatial-density function $\widehat{\partial_\alpha}(X) : \Re^3 \mapsto \Re$ of the percepts type α (see Fig.1-b), which implies that the stationary points $X_{\{\alpha_i,\varsigma_0(t),\varsigma_1(t)\}}$ (the lo-

[4]The size of the 3D field of view.

cations of the α-modes) describe the fused locations of the α-elements of the set. Elements converging to $X_{\{\alpha_i,\varsigma_0(t),\varsigma_1(t)\}}$ constitute the fusion set (cluster delineation in [3]) with a cardinality $C_{\{\alpha_i,\varsigma_0(t),\varsigma_1(t)\}}$. Finally, the point $X_{\{\alpha_i,\varsigma_0(t),\varsigma_1(t)\}}$ allows to determine the amount of frames T_s in which the corresponding percept has to be found. Consequently, subsets carrying through the minimal confidence criterion

$$\frac{C_{\{\alpha_i,\varsigma_0(t),\varsigma_1(t)\}}}{T_s(X_{\{\alpha_i,\varsigma_0(t),\varsigma_1(t),\iota\}})} > E_{min} \tag{4}$$

are merged into a *fused percept* $O_i^{p_f}$, where the parameter ι relaxes the calculation contemplating errors introduced by noisy percepts and ego-mapping. Conclusively, the collection of all $O_i^{p_f}$ forms the set H_f, see Fig.1-b,c.

2.2. World-Model

The world-model has two levels of abstraction. On the first level, the 3D vertices and their composition describing geometric primitives are stored. On the second level, these structures compose instances of *object models* O_i^m with attributes, e.g. identifier, type, size and pose. The collection of object models instances constitutes the *node set*

$$\Xi := \{O_i^m\}_{i=1,\ldots,n}, \tag{5}$$

whereas the *link set*

$$\Lambda \subset \{O_i^m \times O_j^m \ : \ i > j, \ |X_i - X_j| < \varpi\} \tag{6}$$

depicts the connections $\lambda_{i,j}$ formed by all object model instances which relative distances fall below the threshold ϖ .

The considered world-model[5] kitchen consists of 611 rectangular prisms, 124 cylinders, 18 general polyhedra with 846 faces all arranged by 1,524 general transformations (rotation, translation and scaling) with a total of 13,853 vertices and 25,628 normal vectors composed in the scene-graph[6] from the construction CAD model and verified against real furniture with laser devices, see Fig.6.

2.3. Graph Pruning

The previous world representation has been enriched with schemas, which not only integrate and filter the ideal graph model with the noisy percepts, but also create constraints $\Omega_\xi \left\langle O_i^m, O_j^p \right\rangle$ yielding to the hypotheses set Δ, see Fig.1-c,e.

2.3.1 Proximity Filtering. When our algorithm filters links in the world graph, noise is taken into account as deviation

$$\epsilon_i \cong \frac{1}{\zeta}(||X_i^f - C_L||)^2, \tag{7}$$

this is a function describing the distance between the perceived-recognized objects $O_i^{p_f}$ with locations X_i and the center of the left camera C_L [6].

The result of the filter is a set of links $\psi_{\{\alpha,\beta,\phi,\tau\}} \subset \Lambda$ connecting nodes of type α to nodes type β which are separated by a distance ϕ with an error-tolerance $\tau = max_{i\in\Theta}(\epsilon_i)$, where Θ denotes the subset of objects of both types

$$\psi_{\{\alpha,\beta,\phi,\tau\}} \subset \{O_{(i,\alpha)}^m \times O_{(j,\beta)}^m : (\phi - ||X_i - Xj||) < \tau\}. \tag{8}$$

The *active link set* ψ_{act} consists of nodes from the intersection of m proximity filtering results

$$\psi_{act} := \bigcap_i^m \psi_{\{\alpha_i,\beta_i,\phi_i,\tau_i\}}. \tag{9}$$

Each filtering stage $O(n)$ produces a remarkable reduction of the cardinality of the set ψ_{act}, i.e. the remaining nodes should have neighbours with restricted types at constrained distance ranges. A high performance was accomplished by means of dynamic programming techniques (distances-lookup table $O(n^2)$) filtering only previously selected nodes.

2.3.2 Orientation Filtering. A more powerful technique to reduce the nodes cardinality in the set ψ_{act} consists of accepting only those elements which incidence-neighbour nodes have a relative pose, i.e. a displacement vector $V_{i,\vec{jk}}^{p_f}$ from the neighbour node $O_j^{p_f}$ to a third linked node $O_k^{p_f}$ in terms of the created reference frame $S_{Percept}^{i,j,k}$, which is linked to the ego-perception frame, see Fig.3.

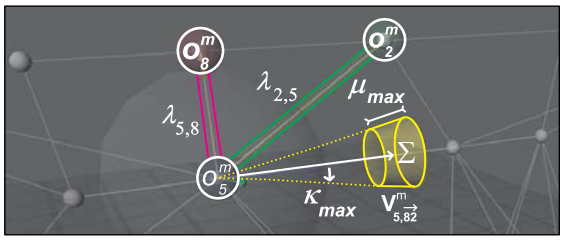

Figure 3. The world-model graph being pruned by means of orientation filtering, with a complexity $O(m^2)$, where m is the degree of the node being filtered.

In this sense, the definition of the frame has to be consistent while considering the noisy nature of the percepts, as follows: first, three non-collinear elements $O_i^{p_f}$, $O_j^{p_f}$ and $O_k^{p_f}$ are selected from H_f specifying the frame $S_{Percept}^{i,j,k} := [R_{Percept}^{i,j,k}, X_i^f]$ relative to the ego-perception frame[7]

$$\delta_1 = X_j^f - X_i^f, \ \ \delta_2 = \delta_1 \times (X_k^f - X_i^f),$$

[5] Human-centered environment [1].

[6] Extending *http://www.coin3d.org/*

[7] Which orthonormal basis vectors are $\{\widehat{e_1}, \widehat{e_2}, \widehat{e_3}\}$.

$$\delta_3 = \delta_1 \times \delta_2, \;\; R^{i,j,k}_{Percept} = \left[\widehat{\delta^n} \cdot \widehat{e_n}\right]_{n=1,\ldots,3}.$$

Next, the relative displacement is computed

$$V^{p_f}_{i,j\vec{k}} = S^{i,j,k}_{Percept}(X^{p_f}_j - X^{p_f}_k). \quad (10)$$

Therefore, it is possible to reject nodes which do not have a "*similar*" displacement vector among two of their neighbours with corresponding type and proximity. This *noisy-similarity* is made quantifiable as the discrepancy length μ and the angle κ between the percepts $V^p_{i,j\vec{k}}$ and those from the model $V^m_{u,u\vec{w}}$, expressed on the world-model frame $S^{u,w,v}_{model}$. Fig.3 shows the subspace Σ bounded by $||V^p_{i,j\vec{k}} - V^m_{u,v\vec{w}}|| < \mu_{max}$ and $\arccos(\widehat{V^p_{i,j\vec{k}}} \cdot \widehat{V^m_{u,v\vec{w}}}) < \kappa_{max}$. When filtering nodes, the combinational burden is reduced by computing only subgraphs which link lengths fall into the range $||V^p_{i,j\vec{k}}|| \pm \mu_{max}$.

2.4. Hypotheses Generation and Validation

The sequence of previous stages extracts model subgraphs, which simultaneously match the *typed-incidences* and the *relative-poses* of acquired percept subgraphs. Latter associations establish the coupling between the *visual space*, *world-model* and *physical world*, see Fig.1-c,d. In fact they impose restraints which are the *geometric-compelling* keys to deduct the 6D pose of the robot.

Each association $\left\langle O^{p_f}_i, O^m_j \right\rangle$ constrains the position of the robot X_{ego} to a subspace of all points which are $\left|\left|X^{p_f}_i\right|\right|$ units away from X^m_j. This subspace is in fact a sphere $\Omega\left\langle O^{p_f}_i, O^m_j \right\rangle$ centered at X^m_j (the position of the matched world-model node) with a radius $\left|\left|X^{p_f}_i\right|\right|$, i.e. the distance from the fused percept to the ego-center, see Fig.4-a. Now considering $\Omega_1\left\langle O^{p_f}_i, O^m_j \right\rangle$ and $\Omega_2\left\langle O^{p_f}_k, O^m_l \right\rangle$, two *restriction spheres* (see Fig.4-b), they implicate that the position of the robot belongs to both subspaces. Hence, the restricted subspace is a circle, i.e. the intersection of spheres

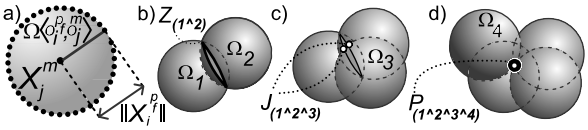

Figure 4. The Ω_i subspaces-intersections constraining the position of the robot. a) Sphere. b) Circle. c) Pair of points. d) Point.

$$Z_{(1\wedge2)} = \Omega_1\left\langle O^{p_f}_i, O^m_j \right\rangle \wedge \Omega_2\left\langle O^{p_f}_k, O^m_l \right\rangle. \quad (11)$$

Following the same pattern, a third sphere $\Omega_3\left\langle O^{p_f}_r, O^m_s \right\rangle$ enforces the restriction to a pair of points, i.e. circle-sphere intersection, see Fig.4-c

$$J_{(1\wedge2\wedge3)} = Z_{(1\wedge2)} \wedge \Omega_3\left\langle O^{p_f}_r, O^m_s \right\rangle. \quad (12)$$

Finally, a fourth sphere $\Omega_4\left\langle O^{p_f}_t, O^m_h \right\rangle$ uniquely determines the position of the robot, i.e. the intersection of the latter pair of points with $\Omega_4\left\langle O^{p_f}_t, O^m_h \right\rangle$, see Fig.4-d

$$P_{(1\wedge2\wedge3\wedge4)} = J_{(1\wedge2\wedge3)} \wedge \Omega_4\left\langle O^{p_f}_t, O^m_h \right\rangle. \quad (13)$$

When more than one matched percept subgraph was extracted, it implicates different plausible positions of the robot. In order to generate and disambiguate those location hypotheses the conformal geometric algebra [6] is used by expressing spheres as computational primitives as well as computing general intersections among them. Fig.5 shows the latter concepts in an efficient location hypotheses generation mechanism.

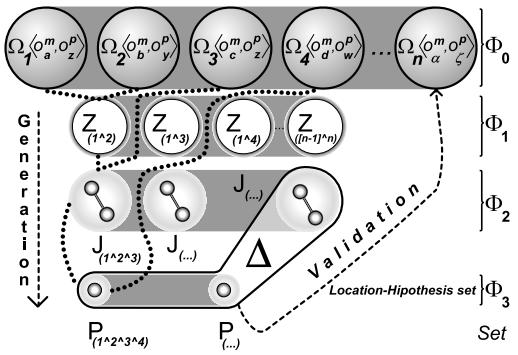

Figure 5. Location hypotheses generation.

Due to uncertainty in the fused percepts, the intersection between restriction spheres is likely to fall in degenerated states (e.g. spheres may not meet, uncertainty from distant percept could diminish the system precision, etc.), which could compromise the quality and existence of the pose-solution. In order to contemplate these facts with their side effects, a statistical method (for a complete description including detailed experimental results see our approach [5]) has been introduced, which in a closed-form simultaneously ensures the solution existence (i.e. maximal density position) and improves the precision of the localization by considering the uncertainties from Eq.7 and the mapping process in section 2.1.1.

2.4.1 Generation. Percept subgraphs are used to produce the *zero-level set*, composed of spheres

$$\Phi_0 := \{\Omega_\varsigma\left\langle O^m_i, O^p_j \right\rangle\}_{\varsigma=1,\ldots,n}. \quad (14)$$

These spheres are intersected by means of the *wedge* operator $\wedge$ in an *upper triangular* fashion producing the *first-level set* Φ_1 comprising circles. The *second-level set* Φ_2 is computed by intersecting those circles with spheres from Φ_0. The latter resulting pair of points are intersected in the same way creating the *third-level set* Φ_3. Here, points from the intersection of four spheres

are contained. Elements of Φ_2 without descendants in Φ_3 and all elements in Φ_3 represent location hypotheses

$$\Delta := \bigwedge_{\xi} \Omega_{\xi} \left\langle O_i^m, O_j^p \right\rangle . \tag{15}$$

The total computational complexity $O(n^4)$ is feasible because in the practice $n \leq 4$.

2.4.2 Validation. Hypotheses are checked by selecting associations $\left\langle O_i^{pf}, O_j^m \right\rangle$ which were not considered when the current validating hypothesis was generated. In case there is more than one prevailing hypothesis, an *active validation* needs to take place by selecting objects from the model and localizing them in the visual space. The criterion to select the discriminator percept is the maximal pose difference between pairs of hypotheses.

2.4.4 Pose Estimation. Once the location hypothesis has revealed the position of the robot X_{ego}, the 6D pose is expressed as

$$S_{ego} = S^{u,w,v}_{model} [S^{i,j,k}_{Percept}]^{-1}. \tag{16}$$

This is the transformation of the kinematic chain coupling the world-model frame S_{model} with the perception frame $S_{Percept}$, see Fig.6.

3. Experimental Results and Conclusions

The global self-localization of the humanoid robot ARMAR-III [2] within the modeled environment was successfully performed using this approach, see Fig.6. The scanning strategy takes 15-20 seconds processing 20 real stereo images, graph model pruning takes 100-150 ms. Finally, the hypotheses generation-validation takes 200-500 ms.

The proposed approach solves the global localization by using the conformal geometric framework and an efficient graph representation of interrelated geometric object features. The resulting pose and those *vision-to-model* subgraphs associations provide very substantial information which is fundamental for autonomous systems, where the visual coupling is needed for higher planning, strategic or semantic abstraction levels.

4 Acknowledgment

The work described in this paper was conducted within the German Humanoid Research project SFB588 funded by the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft).

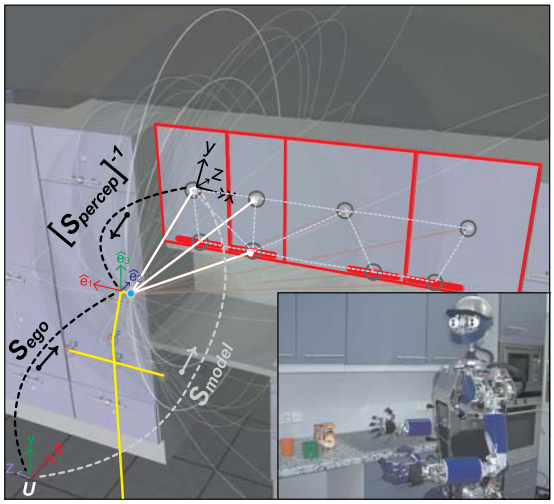

Figure 6. ARMAR-III self-localization.

References

[1] T. Asfour, P. Azad, N. Vahrenkamp, K. Regenstein, A. Bierbaum, K. Welke, J. Schröder, and R. Dillmann. Toward humanoid manipulation in human-centred environments. *Robot. Auton. Syst.*, 56(1):54–65, 2008.

[2] T. Asfour, K. Regenstein, P. Azad, J. Schroder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann. Armar-III: An integrated humanoid platform for sensory-motor control. *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, pages 169–175, Dec. 2006.

[3] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, May 2002.

[4] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1052–1067, June 2007.

[5] D. Gonzalez-Aguirre, T. Asfour, E. Bayro-Corrochano, and R. Dillmann. Improving model-based visual self-localization using gaussian spheres. *To Appear, International Conference AGACSE 2008.*

[6] D. Gonzalez-Aguirre and E. Bayro-Corrochano. A geometric approach for an intuitive perception system of humanoids. In *IAS*, pages 399–407, 2006.

[7] M. Pollefeys, L. V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *Int. J. Comput. Vision*, 59(3):207–232, 2004.

[8] S. Se, D. Lowe, and J. Little. Vision-based global localization and mapping for mobile robots. *Robotics, IEEE Transactions on*, 21(3):364–375, June 2005.

[9] J. Shi and C. Tomasi. Good features to track. *1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pages 593 – 600, 1994.

[10] S. Ullman. *High-Level Vision*. MIT press, Massachusetts, MA, 1996.