On Environmental Model-Based Visual Perception for Humanoids

D. Gonzalez-Aguirre, S. Wieland, T. Asfour, and R. Dillmann

Institute for Anthropomatics, University of Karlsruhe,
Haid-und-Neu-Strasse 7, Karlsruhe-Germany.
{gonzalez,wieland,asfour,dillmann}@ira.uka.de

Abstract. In this article an autonomous visual perception framework for humanoids is presented. This model-based framework exploits the available knowledge and context acquired during global localization in order to overcome the limitations of pure data driven approaches. The reasoning for perception and the properceptive components are the key elements to solve complex visual assertion queries with a proficient performance. Experimental evaluation with the humanoid robot ARMAR-III is presented.

Key words: Model-Based Vision, Object Recognition, Humanoids.

1 Introduction

The emerging research field of humanoid robots for human daily environment is an exciting multidisciplinary challenge. In order for humanoid robots to properly and effectively interact and operate within daily environments it is indispensable to equip them with an autonomous perception framework. Recently, considerable results in this field have been achieved (see [2],[1]) and several humanoid robots exposed various knowledge-driven capabilities and skills. However, those approaches mainly concentrate on manipulation knowledge for graspable objects and fixed object-centered attention zones, e.g. kettle tip while pouring tee or water faucet while washing a cup. These approaches assume fixed pose of the robot in their environment in order to perceive and manipulate objects and environmental elements within a kitchen. In addition, the very narrow field of view with no objects in the background constrains their applicability.

These perception limitations can be overcome through an enhanced exploitation of the available knowledge and model information by including a compact reasoning sublayer within the perception of the humanoid. There exist works on humanoids reasoning for task planning and situations interpretation [3]. However, they focus on atomic operations and discrete transitions between states of the modeled world for behavior generation and verification. This high level reasoning is not the focus of the present work, but the inclusion of the essential reasoning mechanism while perception takes place in order to robustly recognize and interpret complex patterns, i.e. distinguish and track environmental objects

in presence of cluttered backgrounds, grasping occlusion and different poses of both the humanoid and/or the environmental object.

The manipulation of low-level sensor data and higher-level models for segmentation, rejection and recognition constitutes the reasoning for visual perception, which bridges the image processing and object recognition components through a cognitive perception framework [7].

In order to make this reasoning mechanism tractable and implementation plausible it is necessary to profit from both the *vision-to-model* coupling resulting from the model-based approach and the *association-linkage* acquired during the global localization by means of our previous work (see [4],[5]).

The main focus is placed on rigid elements of the environment which could be transformed through parametric (rotational or translational) transformations, e.g. furniture, kitchen appliances, etc.

In the following sections the perception framework and its methods are introduced along experimental results of the demonstration application scenario where these concepts were implemented and evaluated providing remarkable real-time results which pure data driven algorithms would hardly provide.

2 Perception Framework

The aim of the perception framework is to extract valuable information from the real world in the form of stereoscopic color images and joint-encoders values from the humanoids active vision head. The adequate representation, unified storage, automatic recall and task-driven manipulation of this information take place within different layers (*states of cognition*) of the perception framework.

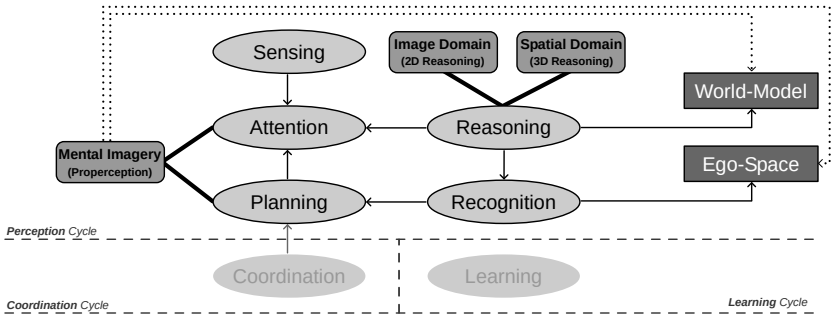

Fig. 1. The perception framework, including states of cognition and principal cycles.

Latter cognition states are categorically organized according to [6] as *sensing*, *attention*, *reasoning*, *recognition*, *planning*, *coordination* and *learning*. In this manner three *principal cycles* arise, namely *perception-cycle*, *coordination-cycle* and *learning-cycle*, see Fig.1.

Memory; World and Ego Spaces. The formal representation of the real objects within the domain and the relationships between them constitutes the *long term memory*, i.e. the world-model. Appropriate description has been done by simultaneously separating the geometric composition from the pose and encapsulating the attributes which correspond to the configuration of the instances, e.g. name, identifier, type, size, parametric transformation, etc. This structure, along the implemented mechanism for pruning and matching lay down the *spatial query solver* used in Sec.4. On the other hand, the *mental imagery* (see Sec.3.1) and the acquired percepts are contained within an ego centered space which corresponds to the *short term memory*.

3 Visual Sensing and Planning

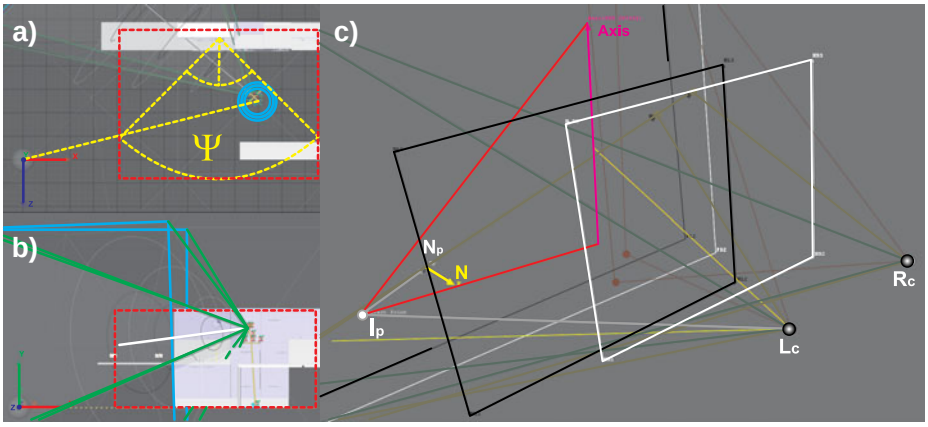

Fig. 2. a) Restriction subspace Ψ where the target node can be robustly recognized, top view. b) Restriction subspace Ψ, side view. c) Geometric elements involved during the spatial reasoning for perception.

Sensing. The noise tolerant vision-to-model coupling consist of the full configuration of the active vision system including the internal joint configuration, external position and orientation of the cameras centers as well as all required mechanisms to obtain euclidean metric from stereo images (see [8], [9]).
Planning. It involves three fundamental aspects. First, once the visual target-node has been established it provides a frame and the definition of a subspace Ψ where the robot has to be located, therewith the target-node can be robustly recognized, see Fig.2-a,b. Notice that this subspace Ψ is not a single pose like in [1], but a wide range of reachable poses allowing a more flexible applicability and more tolerance for uncertainties in the navigation and self-localization.

Subsequently, the visual-planner uses the restriction subspace and target node frame to generate a transformation from the current pose to a set of valid

poses. These poses are submitted to the navigation layer [10] to be unfolded and executed into a safe trajectory. Once the the robot has reached his desired position, once again the visual-planner uses the description of the node to predict parametric transformations and appearance properties, namely, how the image content should look like, and how the spatial distribution of environmental elements is related to the current pose. This is done by the following *properception* mechanism.

3.1 Mental Imagery

The properception skills (prediction, clue extraction, etc.) allow the humanoid to capture the world through internal means by exploiting the full scene-graph of the CAD world-model and the *hybrid virtual cameras*, see Fig.3. These virtual devices use the full-stereoscopic calibration of the real stereo rig in order to set the projection volume and matrix within the virtual visualization, a common practice in the augmented reality [11] for image composition and overlay. However, here this hybrid virtual stereo rig is used to predict and analyze the image content within the world-model, including those previous discussed parametric transformations, extraction of cues like position and orientations even for trajectories of elements, see Fig.3.

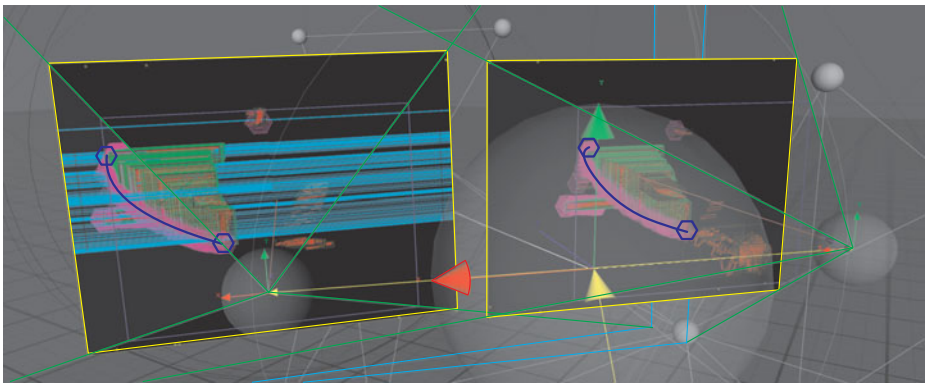

Fig. 3. Properceptive mental imagery for trajectory prediction. Notice the blue lines in the left and right image planes of the hybrid virtual cameras depicting the ideal trajectory of the point of interest (door handle end point) during the opening of the door. This predicted subspace not only allows to reduce region of interest, but it also helps to reject complex outliers.

4 Visual Reasoning for Recognition

The reasoning process for perception could be decomposed in two phases; the visual domain and the spatial domain.

2D Reasoning. The pose estimation of the partial occluded door handle, when the robot has already grasped it, turns out to be a difficult task because there are many perturbations factors. No size rejection criteria may be assumed, because the robot hand is partially occluding the handle surface and also the hand slides during task execution, producing variation of the apparent size. No assumption about the background of the handle could be made, because when the door is partially open and the perspective view overlaps handlers from lower doors same chromatic distribution appear. On the top of that, the glittering of the metal surfaces on both, the robots hand and doors handle, produce very confusing phenomena, when using previously standard segmentation techniques [4].

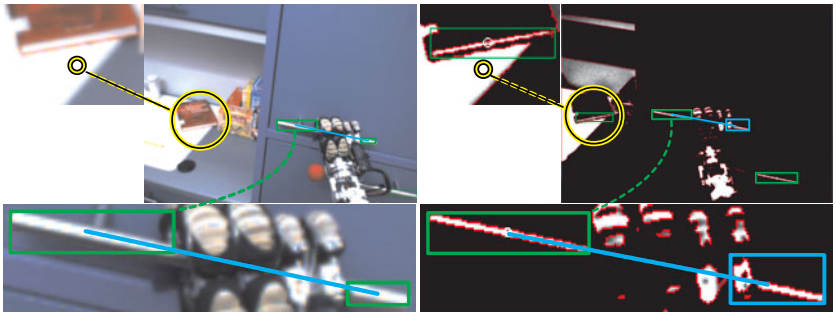

Fig. 4. Left; Input Image. Notice that the book (the white paper side) in the background shows not only similar color distribution, but almost the same size of the door handler. Right; The power image results. Based only on the pure data-driven classification it will be hardly possible to reject the presence of a handler in the current location of the book.

In this context, we propose an application dependent but very robust and fast technique (15-20 ms) to simultaneously segment the regions and erode the borders, producing non-connected regions which suits our desired preprocessing-filtering phase. First, the raw RGB-color image $I_{RGB}(x, y) \in \mathbb{N}^3$ is split per channel and used to compute the *power image* I_ϕ, see Fig.4

$$I_\phi(x, y, n) = [I_R(x, y) \cdot I_G(x, y) \cdot I_B(x, y)]^n \,, \; n \text{ and } I_\phi(x, y, n) \in \mathbb{R}.$$

After a linear normalization and adaptive thresholding a binary image $I_B(x, y) \in \{0, 1\}$ is produced, which is used to extract the blobs B_k and build feature vectors for rejection purposes.

The feature vector $F(B_k)$ is formed by the blobs area $\omega(B_k)$, the energy density $\delta(B_k)$, and the elongation descriptor, i.e. the ratio of the eigen values $E_{\sigma_{i,j}}(B_k)$ of the energy covariance matrix M_{B_k} expressed by $F(B_k) := [\delta(B_k), \omega(B_k), E_{\sigma_1}(B_k)/E_{\sigma_2}(B_k)]$. This characterization enables a powerful rejection of blobs when verifying the right-left cross matching by only allowing candidates in pairs (B_k, B_m) where the criterion is fulfilled, i.e. the orientation of their axis shows a discrepancy less than $\arccos(K_{min})$ radians, i.e.

$K(B_k, B_m) := \|E_{\sigma_1}(B_k) \cdot E_{\sigma_1}(B_m)\| > K_{min}$. The interest point I_p in both images are selected as the furtherest pixel along the blobs main axis in opposed direction of the vector $\Gamma_{R_{Axis}}$, i.e. unitary vector from the door center to the center of the line segment where the rotation axis is located, see Fig.2. This vector is obtained from the mental imagery as stated in Sec.3.1. Moreover, the projected edges of a door within the kitchen aids the segmentation phase to extract the door pose and improves precision by avoiding to consider edges pixels close to the handle. The key factor of this vision-to-model coupling relies on the fact that very general information is used, i.e. from the projected lines and blobs using the mental imagery, only their direction is used (i.e. noise-tolerant criterion K_{min}) and not the position itself, which differs to the real one, due to the discretization, quantization, noise and uncertainty.

3D Reasoning. One of the most interesting rewards of our approach is the usage of the vision-to-model coupling dealing with limited visibility. In order to provide the required information from the global planner or coordinator module it is necessary to estimate the interest point I_p, and the normal vector N_p of the grasping element (see Fig.2-c, e.g. the door handle).

Because of the sizes of both, the door and the 3D field of view (3DFOV, see Fig.2-c), it can be easily corroborated that the minimal distance within the subspace Ψ, where the robot must be located for the complete door to be contained inside the robots 3DFOV, lie outside of the reachable space. In this situation reasoning perception switches from pure data driven algorithm to the following recognition method which only requires three partially visible edges of the door and uses the context (robots pose) and model to assert the orientation of the door's normal vector and the door's angle of aperture. First, a 2D-line Υ_i on an image and the center of its capturing camera C_j define a 3D-space plane $\Phi_{(i,j)}$, hence two such planes $\Phi_{(L,L)}$ and $\Phi_{(\mu(\Upsilon_L,\Upsilon_R),R)}$, resulting from the matching $\mu(\Upsilon_L, \Upsilon_R)$ of two lines in left and right images in a stereo system define an intersection subspace $\Lambda_i = \Phi_{(L,L)} \wedge \Phi_{(\mu(\Upsilon_L,\Upsilon_R),R)}$, i.e. a 3D-line. These 3D-lines Λ_i are subject to noise and calibration artifacts. Thus, they are not suitable to compute 3D intersections. However, their direction is robust enough. Next, the left image 2D points $H_{(L,i)}$ resulting from the intersection of 2D-lines Υ_i are matched against those in the right image $H_{(R,j)}$ producing 3D points $X_{(R,j)}$ by means of triangulation in a minimal square solution. In this way it is possible to acquire corners of the door and directions of the lines connecting them, even when only partial edges are visible. Herein, the direction of the vector $\Gamma_{R_{Axis}}$ (provided by the mental imagery and the spatial reasoning) is the long-term memory clue simultaneously used to select 3D line edge direction D_{Axis} and its point P_{Axis}. In addition, the framework uses proximity knowledge to switch between our components for door and handle recognition and pose estimation.

5 Experiment

Advanced manipulation tasks to perform physical interaction with the environment are important on humanoid robots to be useful in daily life and in cooper-

ation with humans. To demonstrate the advantages of the perception framework and to verify our methods, we accomplished the task of door opening in a regular kitchen environment with the humanoid robot ARMAR-III. In this scenario, the estimation of this normal vector N_p, and therewith the minimization of the external forces at the hand, is the main challenge, because the door changes its orientation during manipulation. In our previous approach [14] the results using only one sensory channel (force-torque sensor) are acceptable but not satisfactory, because the estimation of the task frame depends on the accuracy of the robot kinematics and the tangent is always imprecisely. To decrease the external forces on the hand during the task execution, we use our perception framework to estimate the inters point and normal vector of the door, see Fig.6. The compared results achieved with both methods are shown in Fig. 5.

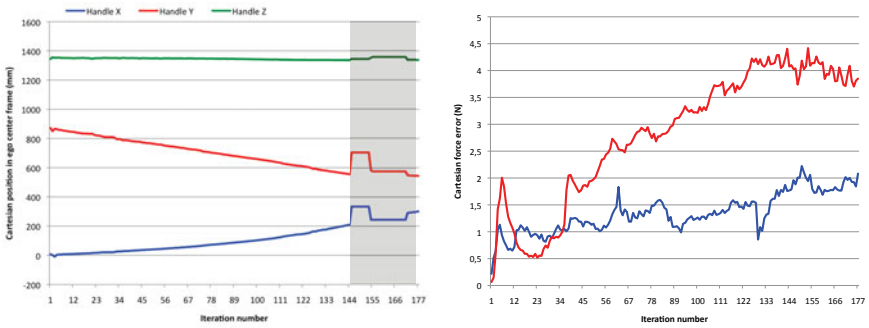

Fig. 5. Left; Cartesian position of the handle midpoint, related to the ego center frame. Smooth movement in the three cartesian dimensions, until iteration 144 when the handle is completely occluded. Right; Comparative plot of the total stress forces at the task frame. The red curve represents the force in the pulling direction using only force-torque sensor and previous kinematic configuration. The blue curve represents our improved results when using the vision estimated task frame in a sensor fusion fashion.

Robustness and reliability of the handle tracker are the key to reduce the force stress in the robots wrist as it can easily be seen in Fig.5. In fact, the sensor fusion in the task space improves the overall performance. Combining stereo vision and force control provides the advantage of real-time task frame estimation by vision, which avoids the errors of the robots kinematics and adjustment of actions by the force control.

6 Conclusions

The world-model and the available context acquired during self-localization (the associations between model elements and visual percepts) will not only make it possible to solve, otherwise hardly possible, complex visual assertion queries,

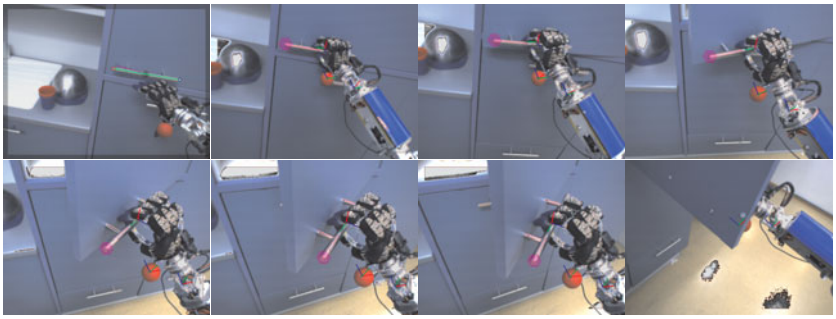

Fig. 6. Experimental evaluation of the perception framework.

but it will also dispatch them with a proficient performance. This is possible through the previous introduced perception framework which implements the basic reasoning skills by extracting simple but compelling geometrical cues from the properception component and then applying them as filters for the classification of percepts, tracking and optimization of the region of interest (in terms of size and trajectory) and finally handling of incomplete visual information. The present work is an on going work which is concern to certain kind of elements in the world. A more general exploitation and exploration of those ideas are the main axis of our future work.

References

1. Okada, K.; Kojima, M.; Tokutsu, S.; Maki, T.; Mori, Y.; Inaba, M., "Multi-cue 3D object recognition in knowledge-based vision-guided humanoid robot system," Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on , vol., no., pp.3217-3222, Oct. 29 2007-Nov. 2 2007
2. Okada, K.; Kojima, M.; Sagawa, Y.; Ichino, T.; Sato, K.; Inaba, M., "Vision based behavior verification system of humanoid robot for daily environment tasks," Humanoid Robots, 2006 6th IEEE-RAS International Conference on , vol., no., pp.7-12, 4-6 Dec. 2006
3. Okada, K.; Tokutsu, S; Ogura, T.;, Kojima, M.;, Mori, Y.; Maki, T.; Inaba,M., "Scenario controller for daily assistive humanoid using visual verification, task planning and situation reasoning," Intelligent Autonomous Systems 10, 2008, ISBN 978-1-58603-887-8
4. Gonzalez-Aguirre, D.; Asfour, T.; Bayro-Corrochano, E.; Dillmann, R., "Model-based visual self-localization using geometry and graphs," Pattern Recognition, 2008. ICPR 2008. 19th International Conference on , vol., no., pp.1-5, 8-11 Dec. 2008
5. Gonzalez-Aguirre, D.; Asfour, T.; Bayro-Corrochano, E.; Dillmann, R., "Improving Model-Based Visual Self-Localization using Gaussian Spheres," Applications of Geometric Algebras in Computer Science and Engineering, 2008. AGACSE 2008. 3rd International Conference on,.
6. Patnaik, S.; Karibasappa K.: Edge, Shade and Mixed Range Detection by Fuzzy Gaussian Filter for an Autonomous Robot. Journal of Intelligent and Robotic Systems, vol. 37, pp. 251–271 (2003)
7. Patnaik, S.: Robot Cognition and Navigation: An Experiment with Mobile Robots. Springer-Verlag, Berlin Heidelberg (2007)
8. Hartley, R.; Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, San Francisco (2004)
9. The Integrating Vision Toolkit (IVT), http://ivt.sourceforge.net/
10. Asfour, T.; Regenstein, K.; Azad, P.; Schroder, J.; Bierbaum, A.; Vahrenkamp, N.; Dillmann, R., "ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control," Humanoid Robots, 2006 6th IEEE-RAS International Conference on , vol., no., pp.169-175, 4-6 Dec. 2006
11. Gordon, G.; Billinghurst, M.; Bell, M.; Woodfill, J.; Kowalik, B.; Erendi, A.; Tilander, J., "The use of dense stereo range data in augmented reality," Mixed and Augmented Reality, 2002. ISMAR 2002. Proceedings. International Symposium on , vol., no., pp. 14-23, 2002
12. Comaniciu, D.; Ramesh, V.; Meer, P., "Real-time tracking of non-rigid objects using mean shift," Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on , vol.2, no., pp.142-149 vol.2, 2000
13. Comaniciu, D.; Meer, P., "Mean shift: a robust approach toward feature space analysis," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.24, no.5, pp.603-619, May 2002
14. Prats, M.; Wieland, S.; Asfour, T.; del Pobil, A.P.; Dillmann, R., "Compliant interaction in household environments by the Armar-III humanoid robot," Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on , vol., no., pp.475-480, 1-3 Dec. 2008