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 the context acquired during global localization in order to overcome the limitations of pure data-driven approaches. The reasoning for perception and the *properceptive*¹ components are the key elements to solve complex visual assertion queries with a proficient performance. Experimental evaluation with the humanoid robot ARMAR-IIIa is presented.

Keywords: 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 to properly and effectively interact and operate within those environments it is indispensable to equip the humanoid robots with autonomous perception capabilities.

Recently, considerable results in this field have been achieved (see [1],[2]) and several humanoid robots exposed various knowledge-driven capabilities and skills. However, those approaches mainly concentrate on knowledge processing for graspable objects with fixed object-centered attention zones, e.g. kettle tip while pouring tea or water faucet while washing a cup.

These approaches assume a fixed pose of the robot in its 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 in real daily scenarios.

These perception limitations can be overcome through an enhanced exploitation of the available model and knowledge information by including a reasoning sublayer within the visual perception system. There exist works on humanoids reasoning for task planning and situation interpretation, see [3], [4]. However, they focus on atomic operations and discrete transitions between states of the

¹ Capturing the world through internal means, e.g. models and knowledge mechanisms. It is the counterpart of perception which captures the world through external means, sensory stimuli.

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 occlusions and different poses of both, the humanoid and the environmental objects.

The processing of low-level sensor data and higher-level knowledge model for segmentation, rejection and recognition constitutes the reasoning for visual perception. It bridges the gap between the image processing and object recognition components through a cognitive perception framework [5].

In order to make this reasoning mechanism tractable and its implementation plausible it is necessary to profit from both the *model-to-vision* coupling resulting from the model-based approach and the *vision-to-model* association acquired during the global localization by means of our previous work, see [6], [7].

The focus herein is placed on rigid elements of the environment which could be transformed through rigid-parametric transformations, e.g. furniture, kitchen appliances, etc.

In the following sections the visual perception framework is 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 Visual Perception Framework

The visual perception framework extracts valuable information from the real world through stereo color images and kinematic configuration of the humanoids active vision head. The adequate representation, unified and efficient storage, automatic recall and task-driven processing of this information take place within different layers (so called *states of cognition*) of the framework.

Latter cognition states are categorically organized according to [5] 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 Model and Ego Spaces. The formal representation of real objects within the application domain and the relationships between them constitutes the *long term memory*, i.e. the world-model. In our approach 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 graph pruning and inexact 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*.

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

3 Visual Sensing and Planning

Sensing. The noise tolerant vision-to-model coupling arise from 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. Note that this subspace Ψ is not a single pose as in [3] and [4], but a wide range of reachable poses allowing more flexible applicability and more robustness through wider tolerance for uncertainties in the navigation and self-localization.

Subsequently, the visual-planner uses the restricted subspace and target node frame to generate a transformation from the current pose to a set of valid poses. These poses are used as input of the navigation layer [10] to be unfolded and executed.

Finally, once the robot has reached the desired position, 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. Note that this is not a set of stored image-node associations (as in the appearance graph) but a novel generative-extractive continuous approach implemented by the following properception mechanism.

3.1 Mental Imagery

The properception skills *prediction* and *cue extraction* allow the humanoid to capture the world through internal means by synergistically exploiting both, the world-model (full scene-graph) and the *hybrid virtual cameras*, see Fig.3. These virtual devices use the plenary stereoscopic calibration of the real stereo rig in

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

Fig. 3. The properceptive mental imagery for trajectory prediction. Note that the blue lines in the left and right image planes of the hybrid virtual cameras show the ideal trajectory of the interest point (door handle end-point) during the door opening. This predicted subspace reduce region of interest and helps to reject complex outliers, see example in Fig.4.

order to set the projection volume and matrix within the virtual visualization, a common practice in the augmented reality [11] for overlay image composition. However, this hybrid virtual stereo rig is used to predict and analyze the image content within the world-model, including the previously described parametric transformations, extraction of position and orientations cues either for static or dynamic configurations.

4 Visual Reasoning for Recognition

The reasoning process for perception is decomposed in two domains; the visual domain (2D reasoning) which concerns with the image content and the spatial domain (3D reasoning), which manages the geometric content.

Visual Domain. The pose estimation of a partial occluded door handle, when the robot has already grasped it, turns out to be difficult because of many perturbation factors. No size rejection criteria may be assumed, because the robot hand is partially occluding the handle surface and 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 handles from lower doors similar chromatic distribution appear. In addition, the glittering of the metal surfaces on both, the robots hand and doors handle, produce very confusing phenomena, when using standard segmentation techniques [12].

In this context, we propose an environment 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

$$I_\phi(x, y, n) = [I_r(x, y) \cdot I_g(x, y) \cdot I_b(x, y)]^n, \text{ where } n \text{ and } I_\phi(x, y, n) \in \mathbb{R}.$$

After linear normalization and adaptive thresholding, a binary image $I_B(x, y)$ 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 $E_{\sigma_i}(B_k)$ of the energy covariance matrix² 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 is selected as the furthest pixel along the blobs main axis in opposed direction of the vector Γ_R , i.e. unitary vector from the door center to the center of the line segment where the rotation axis is located, see Fig.2-c. This vector is obtained from the mental imagery as stated in Sec.3.1-c. Moreover, the projected edges of a door within the kitchen aid the segmentation phase to extract the door pose and improves the overall precision by avoiding to consider edges pixels close to the handle.

² From the power image by selecting masked blobs $\omega(B_k)$ in the binary image $I_B(x, y)$.

Fig. 4. a) Input image I_{rgb} . Note that the book (particularly the white paper side) in the background shows not only similar color distribution, but almost the same size of the door handle. b) The power image I_ϕ . Based only on the pure data-driven classification it will be hardly possible to reject the presence of a handle within the location of the book.

The key factor of this model-to-vision coupling relies on the fact that very general information is applied, i.e. from the projected lines and blobs employing mental imagery, only their direction is used (e.g. noise-tolerant criterion K_{min}) and not the position itself, which differs from the real one, due to the discretization, quantization, noise and uncertainties.

Spatial Domain. One of the most interesting features of our approach is the usage of the vision-to-model coupling to deal 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 size of both, the door and the 3D field of view (3DFOV, see Fig.2-a,b), it can be easily corroborated that the minimal distance, where the robot must be located for the complete door to be contained inside the robots 3DFOV, lies outside of the reachable space, therefore common triangulation techniques may not be used. In this situation, the module reasoning for 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 (robot pose) and model to assert the orientation of the door's normal vector and the door's angle of aperture.

First, a 2D-line $\mathcal{I}_i$ on an image and the center of its capturing camera L_c or R_c define a 3D-space plane $\Phi_{(i,j)}$, see Fig.2-c. Hence, two such planes $\Phi_{(L,L)}$ and $\Phi_{(\mu(\mathcal{I}_L, \mathcal{I}_R), R)}$, resulting from the matching $\mu(\mathcal{I}_L, \mathcal{I}_R)$ of two lines in the left and right images in a stereo system define an intersection subspace, i.e. a 3D-line

$$\Lambda_i = \Phi_{(L,L)} \wedge \Phi_{(\mu(\mathcal{I}_L, \mathcal{I}_R), R)}.$$

These 3D-lines A_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 $\mathcal{Y}_i$ 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 fashion.

Finally, 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 Γ_R is the long-term memory cue used to select the 3D line edge direction D_{Axis} and its point P_{Axis} .

5 Experimental Evaluation

In order to demonstrate the advantages of the presented framework for visual perception and to verify our methods, we accomplished the task of door opening in a regular kitchen with the humanoid robot ARMAR-IIIa [10], see Fig.5.

In this scenario, the estimation of the normal vector N_p , and therewith the minimization of the external forces on the hand, because the door changes its orientation during manipulation. In our previous approach [13] 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.

In this experimental evaluation the framework estimates the interest point and normal vector of the door, therewith the task frame. During task execution this frame is estimated by the before mentioned methods and the impedance control is balancing the external forces and torques at the hand, caused by vision artifacts. 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.6.

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.

Fig. 5. Experimental evaluation of the perception framework

Fig. 6. **Left;** Cartesian position of the handle midpoint. Smooth movement in the three cartesian dimensions, until iteration 144 when the handle is completely occluded. **Right;** Total stress forces at the task frame. The red plot represents the force in the pulling direction using only force-torque sensor and previous kinematic configuration. The blue plot is the result when applying a vision-based estimation of the task frame in a sensor fusion fashion.

6 Conclusions

The world-model and the available context acquired during self-localization will not only make it possible to solve, otherwise hardly possible, complex visual assertion queries, but it will also dispatch them with a proficient performance. This is possible through the 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 speed) and handling of incomplete visual information, see Fig.5. The coupling of model-to-vision by means of the properceptive cues generated by the mental imagery along with the visual and spatial reasoning mechanism for perception are the main novel contributions of the framework. A more general formulation, exploitation and exploration of these ideas are the main axis of our current work.

Acknowledgments

The work described in this article was partially conducted within the German Humanoid Research project SFB588 funded by the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft) and the EU Cognitive Systems projects GRASP (FP7-215821) and PACO-PLUS (FP6-027657) funded by the European Commission.

References

1. Okada, K., Kojima, M., Sagawa, Y., Ichino, T., Sato, K., Inaba, M.: Vision based behavior verification system of humanoid robot for daily environment tasks. In: IEEE-RAS Int. Conference on Humanoid Robots (2006)
2. 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. In: IEEE/RSJ International Conference on Intelligent Robots and Systems 2007 (2007)
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. Okada, K., Kojima, M., Tokutsu, S., Mori, Y., Maki, T., Inaba, M.: Task guided attention control and visual verification in tea serving by the daily assistive humanoid HRP2JSK. In: IROS 2008. IEEE/RSJ International Conference on Intelligent Robots and Systems (2008)
5. Patnaik, S.: *Robot Cognition and Navigation: An Experiment with Mobile Robots*. Springer, Heidelberg (2007)
6. Gonzalez-Aguirre, D., Asfour, T., Bayro-Corrochano, E., Dillmann, R.: Model-based visual self-localization using geometry and graphs. In: 19th International Conference on Pattern Recognition. ICPR 2008 (2008)
7. Gonzalez-Aguirre, D., Asfour, T., Bayro-Corrochano, E., Dillmann, R.: Improving Model-Based Visual Self-Localization using Gaussian Spheres. In: 3rd International Conference on Applications of Geometric Algebras in Computer Science and Engineering. AGACSE 2008 (2008)
8. Hartley, R., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, 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. In: IEEE-RAS Int. Conference on Humanoid Robots (2006)
11. Gordon, G., Billingham, M., Bell, M., Woodfill, J., Kowalik, B., Erendi, A., Tilander, J.: The use of dense stereo range data in augmented reality. In: Proceedings of International Symposium on Mixed and Augmented Reality, 2002. ISMAR 2002, pp. 14–23 (2002)
12. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(5), 603–619 (2002)
13. Prats, M., Wieland, S., Asfour, T., del Pobil, A.P., Dillmann, R.: Compliant interaction in household environments by the Armar-III humanoid robot. In: IEEE-RAS Int. Conference on Humanoid Robots (2008)