Pattern Recognition Letters 32 (2011) 2254–2260

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Towards stratified model-based environmental visual perception for humanoid robots

D. Gonzalez-Aguirre *, T. Asfour, R. Dillmann

Institute for Anthropomatics, Humanoids and Intelligence Systems Lab, Karlsruhe Institute of Technology, Adenauerring 2, Karlsruhe, Germany

ARTICLE INFO

Article history:
Available online 13 October 2010

Keywords:
Model-based vision
Object recognition and detection
Cognitive vision
Humanoid robots

ABSTRACT

An autonomous environmental visual perception approach for humanoid robots is presented. The proposed framework exploits the available model information and the context acquired during global localization by establishing a vision-model coupling in order to overcome the limitations of purely data-driven approaches in object recognition and surrounding status assertion. The exploitation of the model-vision coupling through the *properceptive* components is the key element to solve complex visual assertion-queries with proficient performance. An experimental evaluation with the humanoid robot ARMAR-IIIa is presented.

1. Introduction

The emerging research field of humanoid robots for human daily environments is an exciting multidisciplinary challenge. It embodies multiple aspects and disciplines from mechanical engineering up to artificial intelligence. The physical composition and appearance of humanoid robots differentiate them from other robots according to their application domain. This composition will ultimately allow the robots to noninvasively and effectively operate in human-centered environments. In order to properly and efficaciously interact in those environments it is indispensable to equip the humanoid robots with autonomous perception capabilities.

Recently, considerable results (Okada et al., 2006, 2007, 2008a,b) in this field have been achieved and several humanoid robots expose various knowledge-driven capabilities.

However, these approaches mainly concentrate on knowledge processing for grasping with fixed object-centered attention zones, e.g. a kettle's tip for pouring tea, a water faucet for washing a cup, etc.

These approaches assume a fixed pose of the robots in order to perceive and manipulate unattached objects and environmental elements within a kitchen. In addition, the very narrow field-of-view with no objects in the background and the fully saturated colors of the auxiliary localization props constrains their applicability in real daily scenarios.

These perception limitations can be overcome through a *properceptive*[1] stratified sensing approach. It allows an enhanced exploitation of the available model information by including compact but concise cue-extraction from the model and reasoning sublayers within the visual perception system.

There are works on humanoid robots reasoning for task planning and situation interpretation, see (Okada et al., 2008a,b). These approaches focus on atomic operations and discrete transitions between states of the modeled scenario for behavior generation and verification.

This high-level scenario reasoning is not the focus of the present work, but the inclusion of the essential properceptive and 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 robots and the objects.

This article focuses 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 model-based stratified visual perception for humanoid robots and its implementation are introduced. It comes with the experimental results of a demonstration scenario, where the concepts were evaluated providing remarkable

[1] Properception is the counterpart of perception. The properception deals with the external world by internal means through models and knowledge mechanisms, whereas the perception captures the world through external sensory stimuli. In contrast to the properception, the proprioception deals with the sense related to limbs position, self-posture, awareness of equilibrium, and other internal conditions. Both properception and proprioception provide awareness of the outside (models) and inside (interoceptive senses) respectively, see Fig. 1.

* Corresponding author. Tel.: +49 721 608 8489; fax: +49 721 608 4077.
E-mail addresses: gonzalez@ira.uka.de (D. Gonzalez-Aguirre), asfour@kit.edu (T. Asfour), ruediger.dillmann@kit.edu (R. Dillmann).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.09.028

real-time results which purely data-driven algorithms would hardly provide.

2. Stratified visual perception

The Fig. 1 shows the strata or spaces of abstraction involved in this approach. By dividing the whole approach into these container spaces it is possible to establish the bridge (see Figs. 1e and f) between the reality and the models. The *vision-model* coupling is composed by the confluence of the stimuli-novelty (percepts) and inference-prediction (symbols) respectively provided by the perception and properception processes.

In order to make this coupling mechanism tractable and its implementation plausible, it is necessary to profit from both the *vision-to-model* association acquired during the global localization by our previous work (Gonzalez-Aguirre et al., 2006, 2008, 2010, 2009; Wieland et al., 2009) and the *model-to-vision* resulting from the inference rules in the model-based approach.

3. Visual perception framework

The processing of low-level sensor data and higher-level world-model information for segmentation, recognition, and association constitutes the visual perception. It bridges the gap between the image processing and the object recognition components through a cognitive perception framework (Patnaik, 2007). This framework actively extracts valuable information from the real world through stereo color images and the kinematic configuration of the humanoid robots active vision head (Asfour et al., 2008).

The adequate representation, efficient unified storage, automatic recall, and task-driven processing of this information takes place within different states of cognition. These cognition states are categorically organized according to their function as *sensing*, *attention*, *reasoning*, *recognition*, *planning*, *coordination*, and *learning*, see Fig. 2.

3.1. Memory: model and ego spaces

The formal representation of real objects within the application domain and the relationships between them constitute the *long term memory*. Particularly in the environmental perception, the world-model and the defined transformations compose this non-volatile memory. In this approach, an appropriate description has been done by separating the geometric composition from the pose. The attributes are the configuration state of the instances, e.g. name, identifier, type, size, parametric transformation, etc. This persistent graph structure together with the implemented mechanism (see Sec. 2.3-4 in Gonzalez-Aguirre et al. (2008)) for pruning and inexact matching constitute the *spatial query solver*, see Fig. 2.

On the other hand, the *mental imagery* (see Section 3.4) and the acquired percepts are contained within an ego-space which corresponds to the *short term memory*. By attaching the base platform pose (the frame B in Fig. 3a) to the registration pose of the contained percepts, it is possible to have a short term registration frame for the fusion and model exploitation. Obviously, when the humanoid robot moves its platform, the temporal registration frame and the contained percepts of the ego-space are automatically discarded.

3.2. Sensing

The noise-tolerant *vision-to-model* (see Fig. 1e) coupling arises from the full configuration of the active vision system including the calibrated internal joint configuration (Welke et al., 2008), the external position and the orientation of the camera centers as well as all required mechanisms (Azad and Dillmann, 2009) to obtain Euclidean metric from stereo images (Hartley and Zisserman, 2004), see Figs. 3a and b.

3.3. Planning

Visual planning determines two fundamental aspects for the robust perception;

3.3.1. Target subspace

When the target-node has been established (usually by the coordination cycle), the visual-planner provides a frame and the definition of a subspace Ψ where the robot has to be located, so the target-node can be robustly recognized, see Figs. 3c and d. Notice that the subspace Ψ is not a single pose as in Okada et al.

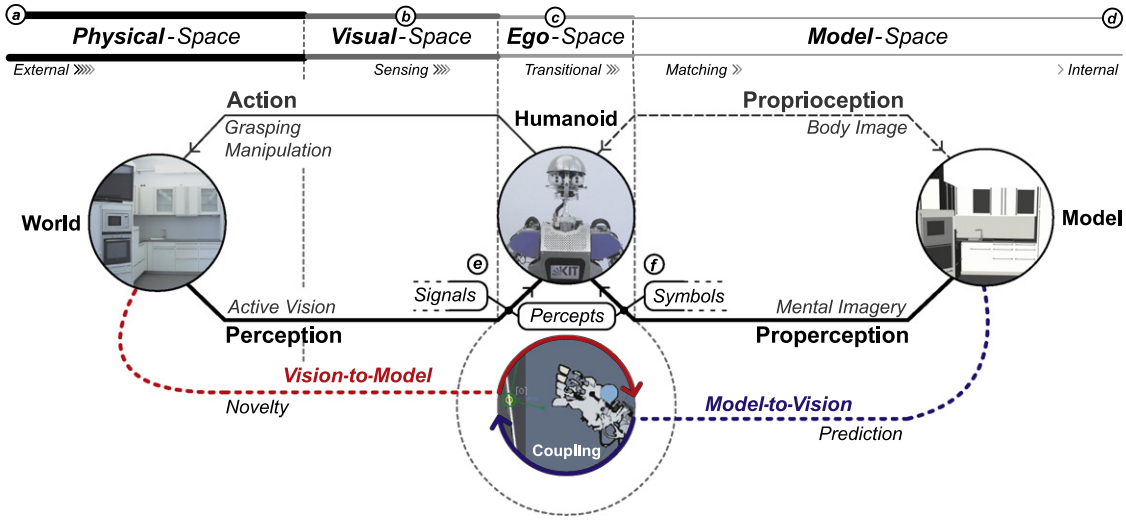

Fig. 1. The model-based stratified visual perception for humanoid robots: a) the *physical-space* embraces the tangible reality, b) the *visual-space* embodies the image projection from the reality to percepts by means of sensor devices and active recognition components, c) the *ego-space* is the short term registration storage for percepts and self-localization, d) the *model-space* contains the geometrical and topological description of the entities of the physical-space, e) the *signals-to-percepts* is the transducer process from visual-space to ego-space. It converts incoming signals from the visual-space to outgoing percepts corresponding to abstracted entities of the model-space, and f) the *symbols-to-percepts* process fuses the percepts corresponding to abstracted entity in the model-space.

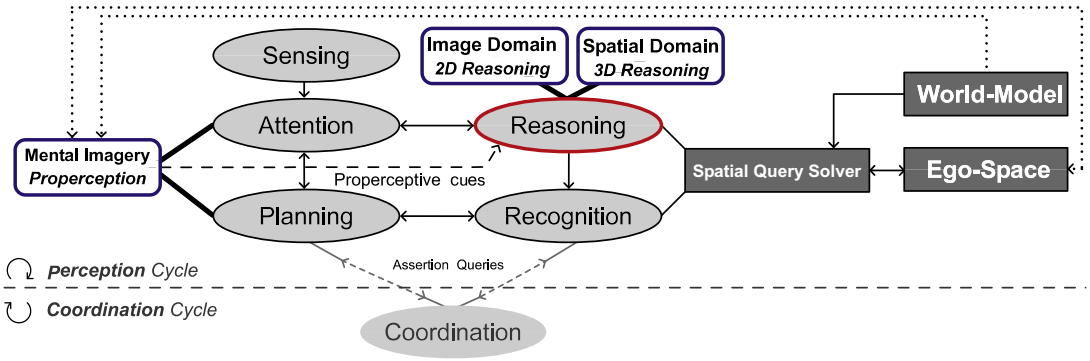

Fig. 2. The states of cognition and cycles involved in the model-based visual perception approach. The properception is implemented by the mental imagery module which fuses information from the long- with the short-term memory, i.e. world- and ego-spaces respectively. The mental imagery provides the properceptive cues for rule-based reasoning and attention planning. The integration of these cues takes place in the image and space domain reasoning submodules, see Section 4. The communication interface with the coordination cycle is done through assertion queries, see an example in Section 5.

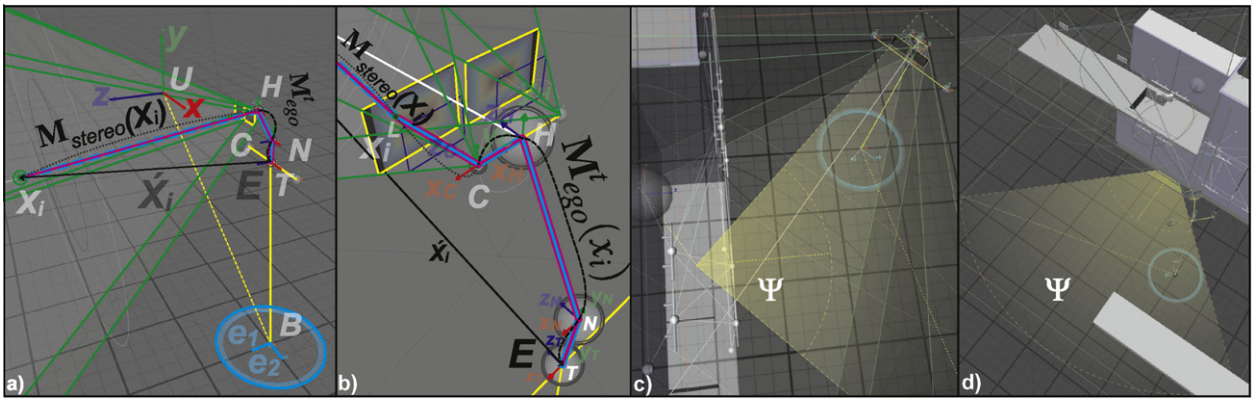

Fig. 3. a) Mapping of percepts from physical-space to the ego-space by the composed transformation $\acute{X}_i = \boldsymbol{M}^t_{ego}[\boldsymbol{M}_{stereo}(X_i)]$, b) the partial transformation from visual-space to the ego-space $\boldsymbol{M}^t_{ego} = [T_{(t)}N_{(t)}HC_L]^{-1}$, see also (Gonzalez-Aguirre et al., 2010), c) the restriction subspace Ψ where the target-node can be robustly recognized, and d) alternative view of the subspace Ψ.

(2008a,b). It is a rather wide range of reachable poses allowing more flexible applicability and more robustness through wider tolerance for uncertainties in the navigation and self-localization.

3.3.2. Appearance context

Once the robot has reached a valid position within Ψ, the visual-planner uses the CAD geometric composition of the node to predict parametric transformations and appearance properties, i.e. how the modeled image content looks like and how the spatial distribution of environmental elements is related to the current pose. Notice that this is not a set of stored image-node associations as in the appearance graph approaches (Koenig et al., 2008) but a novel generative/extractive continuous technique implemented by the following properception mechanism.

3.4. Properception: towards visual mental imagery

The properception skills *cue-extraction* and *prediction* allow the humanoid robot to capture the world by internal means by exploiting the world-model (scene-graph) through the *hybrid* cameras, see Fig. 4. These hybrid devices use the full stereoscopic calibration of the real stereo rig in order to set the projection volume and the projection matrix within the virtual visualization. This half virtual/half physical device is inspired by the inverted concept of augmented reality approaches (Koenig et al., 2008) for overlay image composition. In contrast to augmented reality, this hybrid stereo rig is used to predict and analyze the image content in the world-model, e.g. rigid parametric transformations, extraction of position and orientation cues either for static or dynamic configurations, i.e. rotation or translation over time as in Fig. 4.

4. Visual reasoning for recognition

The reasoning process for perception is decomposed into two interdependent domains;

Visual domain: The 2D signals-to-percepts process (see Fig. 1e) deals with the image content and includes all the chromatic-radiometric sensor calibration and signal processing components required for segmentation, saliency estimation, and geometric primitive extraction. Furthermore, these components are capable of incorporating additional information for purpose-driven extraction, i.e. model-based segmentation. For a detailed example see Section 4.1.

Spatial domain: The 3D percepts-to-symbols matching and reasoning process (see Fig. 1f) manages the geometric entities from both the ego-space and the model-space. This management includes the coupling and the inference through (until now) simple geometric rules, for a more detailed illustration see Section 4.2.

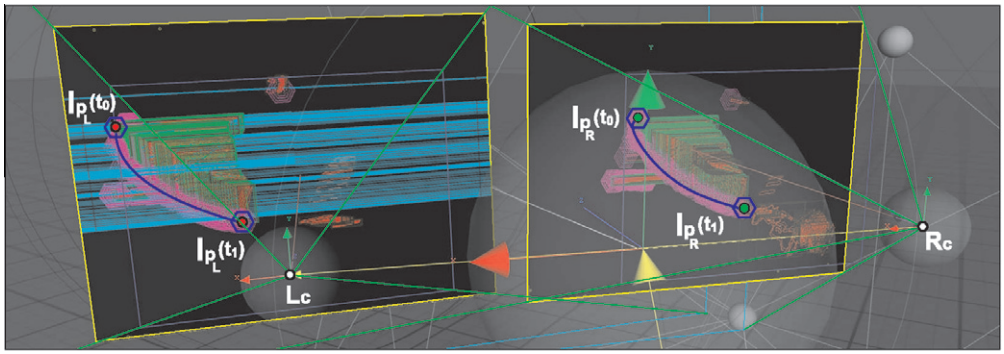

Fig. 4. The properceptive mental imagery for prediction of dynamic configuration trajectories. The blue lines in the left $I_{p_L(t)}$ and right $I_{p_R(t)}$ image planes of the hybrid cameras show the ideal trajectory of the interest point I_p (door handle end-point) during a door opening task. The predicted subspace reduces the size of the region of interest. In addition, the predicted orientation helps to reject complex outliers, see example in Section 4.1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.1. Signals-to-percepts

The pose estimation of a partially occluded door handle, when the robot has already grasped it, turns out to be difficult because of many perturbation factors:

- No size rejection criterion can be assumed, because the robot's hand is partially occluding the handle surface and the hand slides during task execution, producing variation of the apparent size.
- No assumption about color or edges in the background of the handle could be made. This happens when the door is partially open and the perspective view overlaps handles from lower doors similar chromatic distributions appear. This image content avoids the edge tracking (Azad et al., 2009).
- In addition, the glittering of the metal surfaces on both the robot's hand and door's handle produce very confusing phenomena when using standard segmentation techniques (Comaniciu et al., 2002; Hui Zhang and Fritts, 2008).

In this context, we propose an environment-dependent but very robust and fast technique (25–50 ms) to simultaneously segment the regions and erode the borders, producing non-connected regions.

First, the raw *RGB*-color image $I_{rgb}(\boldsymbol{x}) \in \mathbb{N}^3$, $\boldsymbol{x} \in \mathbb{N}^2$ is split per channel and used to compute the *power image* $I_\phi \in \mathbb{R}$, namely

$$I_\phi(\boldsymbol{x}, p) = \left[I_r(\boldsymbol{x}) \cdot I_g(\boldsymbol{x}) \cdot I_b(\boldsymbol{x})\right]^p,$$

where $p \in \mathbb{R}$ and $p > 1$, see Fig. 5.

After linear normalization and locally adaptive threshold (Chang et al., 2006), a binary image $I_B(x)$ is produced. It is used to extract the pixel-connected components (*blobs*) $B_k := \{\boldsymbol{x}_i\}_{i=1}^n$ and build the corresponding feature vectors $F(B_k)$ for rejection and/or matching purposes (see Fig. 5b), namely

$$F(B_k) := [n, \delta(B_k), E_{\sigma_1}(B_k)/E_{\sigma_2}(B_k)]^T.$$

This feature vector is formed by the blob's discrete area $|B_k| = n$, its power density

$$\delta(B_k) := \frac{1}{n}\sum_{i=1}^{n} I_\phi(\boldsymbol{x}_i, p),$$

and the elongation descriptor, i.e. the ratio of the blob's eigenvalues $E_{\sigma_1} : E_{\sigma_2}$ computed by the singular value decomposition

$$\left[\overrightarrow{E_{\sigma_{1,2}}}, E_{\sigma_{1,2}}\right] = SVD(M_{B_k})$$

of the $\bar{\boldsymbol{x}}$-centered and λ-weighted covariance matrix M_{B_k}, namely

$$\bar{\boldsymbol{x}} := \frac{1}{n \cdot \delta(B_k)}\sum_{i=1}^{n} I_\phi(\boldsymbol{x}_i, p) \cdot \boldsymbol{x}_i,$$

$$\lambda(\boldsymbol{x}_i) := \left[I_\phi(\boldsymbol{x}_i, p) - \delta(B_k)\right]^2$$

$$M_{B_k} := \frac{\sum_{i=1}^{n} \lambda(\boldsymbol{x}_i)[\boldsymbol{x}_i - \bar{\boldsymbol{x}}][\boldsymbol{x}_i - \bar{\boldsymbol{x}}]^T}{\sum_{i=1}^{n} \lambda(\boldsymbol{x}_i)}.$$

This characterization allows a powerful rejection of blobs when verifying the right-left cross matching by allowing only candidates in pairs (B_k, B_m) which fulfill the coherence criterion $K(B_k, B_m) > K_{min}$, i.e. the orientation of their main axes shows an angular discrepancy less than $\arccos(K_{min})$ radians.

Until this point, the image feature extraction methods proceed without any model information or knowledge rule, i.e. data-driven. In the next step, the properceptive cue selection and usage is introduced.

The interest point I_p in both images is selected as the furthest pixel along the blob's main axis in the opposed direction of the vector Γ, i.e. unitary vector from the door center to the center of the line segment where the rotation axis is located, see Figs. 6 and 7. This vector is obtained from the mental imagery as stated in Section 3.4. Moreover, the projected edges of a door within the kitchen improves the segmentation results while extracting the door pose. It improves the overall precision by avoiding to consider edge pixels close to the handle, see Fig. 6a.

The key factor of this model-vision coupling relies on the fact that very general information is applied. In other words, from the projected lines and blobs extracted employing mental imagery, only their direction is used (e.g. injected through a noise-tolerant criterion K_{min}) and not the position itself, which normally differs from the real one. These deviations are produced due to the discretization, quantization, sensor noise, actuator deviations, and model uncertainties.

4.2. Percepts-to-symbols

One interesting feature of this approach is the usage of the vision-model coupling to deal with limited visibility. For instance, because of the size of both the door and the 3D field-of-view (3DFOV, see Figs. 3c and d and Fig. 7), it can be easily corroborated that the minimal distance where the robot must be located for the complete door to be contained inside the robot's 3DFOV, lies outside of the reachable space. Therefore, triangulation techniques cannot be used. In this situation, the reasoning for perception uses a simple geometric rule for the recognition and pose estimation of the door.

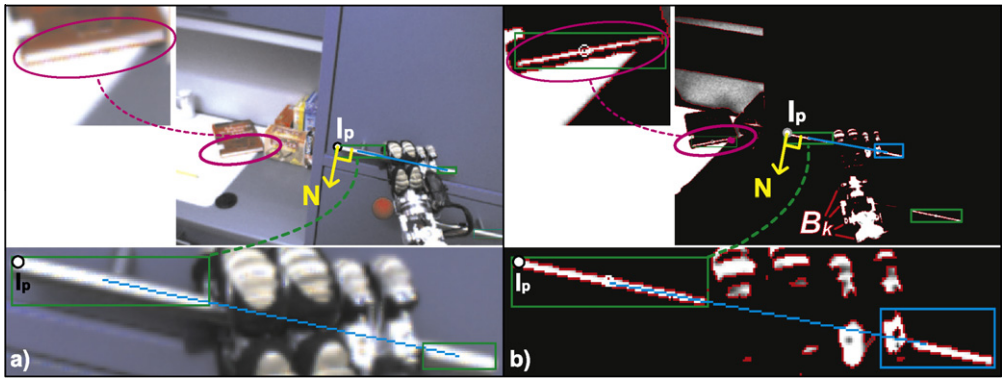

Fig. 5. a) Input image I_{rgb}. Notice that the book (particularly the white paper side) in the background. It shows not only a similar color distribution, but is has almost the same size as the door handle and b) the power image I_{ϕ} and the blobs B_k. Based only on the feature vector B_k (data-driven) recognition it will be hardly possible to reject the presence of a door handle in the location of the book.

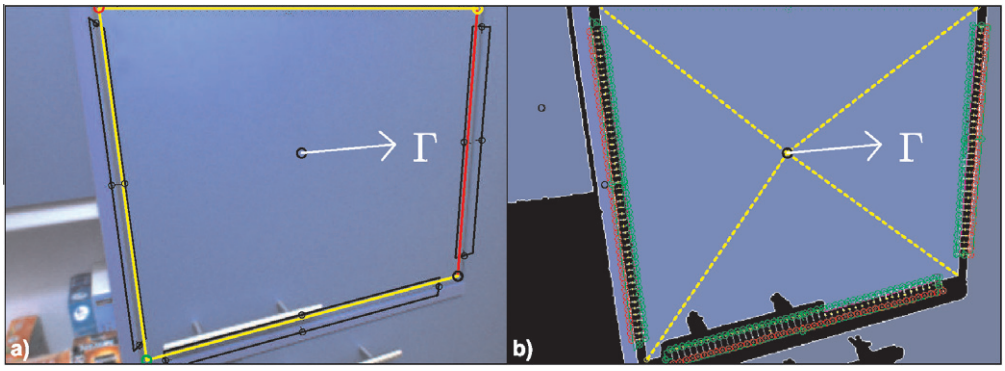

Fig. 6. a) The input image I_{rgb} with recognized edges, projected model edges and the properceptive cue (unitary vector Γ) and b) segmentation results for block and edge analysis.

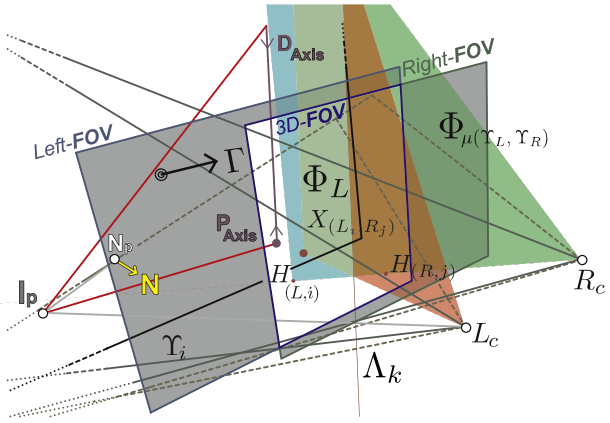

Fig. 7. Geometric elements involved during the spatial reasoning for perception. The 3D field-of-view is the subspace resulting from intersecting the left and right field-of-views of the stereo-rig (Gonzalez-Aguirre et al., 2006).

The preconditions of the rule are: three partially visible edges of the door, the context (robot joint configuration and pose) and the model to assert. The post conditions are: the pose of the model and the geometric derived information, e.g. door's normal vector and the door's angle of aperture.

The rule is computed as follows: First, a 2D-line Υ_i on an image and the center of its capturing camera L_c or R_c define a 3D-space plane Φ, see Fig. 7. Hence, two such planes Φ_L and $\Phi_{\mu(\Upsilon_L,\Upsilon_R)}$, resulting from the matching $\mu(\Upsilon_L,\Upsilon_R)$ of two lines in the left and right images in a stereo system define an intersection subspace, i.e. a 3D-line

$$\Lambda_k = \Phi_L \wedge \Phi_{\mu(\Upsilon_L,\Upsilon_R)}.$$

These 3D-lines Λ_k are subject to noise and calibration artifacts. Thus, they are not suitable for direct computation of 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 $X_{(L_i,R_j)}$ by means of least square triangulation.

Finally, it is possible to acquire corners of the door and directions of the lines connecting them, even when only partial edges are visible. The direction of the vector Γ is the long-term memory cue used to select the 3D edge line by its direction D_{Axis} and the closest point $X_{(L_i,R_j)}$, namely P_{Axis} in Fig. 7.

5. Experimental evaluation

In order to demonstrate the advantages of the presented approach for visual perception and to verify these methods, we accomplished the task of door opening in a regular kitchen with the humanoid robot ARMAR-IIIa (Asfour et al., 2006).

In this scenario (see Fig. 8), the estimation of the relative pose of the furniture not only allows to grasp the door's handle but it also helps to reduce the external forces on the hand during operation. This results from the adaption of the task frame while the manipulation changes the handle's orientation.

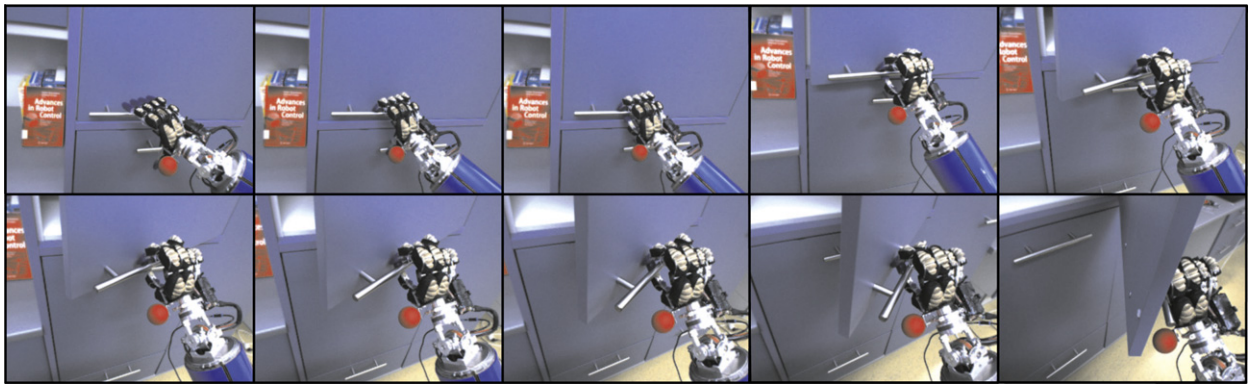

Fig. 8. Experimental evaluation of the framework with the humanoid robot ARMAR-IIIa interacting with a cupboard in the kitchen environment.

In our previous approach (Prats et al., 2008), the results using only one sensory channel (force-torque sensor) were acceptable but not satisfactory because the estimation of the task frame solely depends on the accuracy of the robot kinematics.

In this experimental evaluation, the framework estimates the interest point I_p and normal vector N_p of the door to build the task frame. During task execution this frame is estimated by the previously mentioned methods and the impedance control balances the external forces and torques at the hand. For details on the sensor fusion strategy see Wieland et al. (2009).

Robustness and reliability of the handle tracker are the key to reduce the force stress in the robot's wrist as it has been shown in Wieland et al. (2009).

Combining stereo vision and force control provides the advantage of real-time task frame estimation by vision, which avoids the errors of the robot's kinematics and adjustment of actions by force control.

6. Conclusions

The world-model and the available context acquired during self-localization do not only make it plausible to solve otherwise hardly possible, complex visual assertion queries, but they also help to dispatch them with a proficient performance. This is achieved by the presented framework which implements the basic reasoning skills by extracting simple but compelling geometrical cues from the properception component. These cues are applied as clue-filters for the association of percepts either for tracking (by optimization of the region of interest in terms of size) or handling incomplete visual information.

The novelty of our approach is the coupling of vision-model by means of the properceptive cues generated with the mental imagery and the visual extraction for spatial reasoning.

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. The authors also thank the support from the DAAD-Conacyt Scholarship Reg.180868 funded by the German academic exchange service (DAAD: Deutscher Akademischer Austausch Dienst) and the Mexican National Council of Science and Technology (Conacyt: Consejo Nacional de Ciencia y Tecnología).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2010.09.028.

References

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

Asfour, T., Welke, K., Azad, P., Ude, A., Dillmann, R. 2008. The karlsruhe humanoid head, in: Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on, pp. 447–453.

Azad, P., Asfour, T., Dillmann, R. 2009. Accurate shape-based 6-dof pose estimation of single-colored objects, in: Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on, pp. 2690–2695.

Azad, P., Dillmann, R. 2009. The integrating vision toolkit, http://ivt.sourceforge.net/.

Chang, C.-I., Du, Y., Wang, J., Guo, S.-M., Thouin, P., 2006. Survey and comparative analysis of entropy and relative entropy thresholding techniques, vision, image and signal processing. IEEE Proceedings.

Comaniciu, D., Meer, P., Member, S., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal Machine Intell 24, 603–619.

Gonzalez-Aguirre, D., Bayro-Corrochano, E. 2006. Intelligent autonomous systems 9, proceedings of the 9th international conference on intelligent autonomous systems, university of tokyo, in: Arai, T., Pfeifer, R., Balch, T.R., Yokoi, H. (Eds.), IAS, IOS Press, ISBN:1-58603-595-9.

Gonzalez-Aguirre, D., Asfour, T., Bayro-Corrochano, E., Dillmann, R. 2008. Model-based visual self-localization using geometry and graphs, in: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pp. 1–5.

Gonzalez-Aguirre, D., Wieland, S., Asfour, T., Dillmann, R., 2009. On environmental model-based visual perception for humanoids. In: The 15th Iberoamerican Congress on Pattern Recognition. CIARP 2010, vol. 5856. Springer. ISBN 978-3-642-10267-7.

Gonzalez-Aguirre, D., Asfour, T., Bayro-Corrochano, E., Dillmann, R., 2010. Improving model-based visual self-localization using gaussian spheres. In: Bayro-Corrochano, E., Scheuermann, G. (Eds.), Geometric Algebra Computing. Springer. ISBN 978-1-84996-107-3.

Hartley, R., Zisserman, A., 2004. Multiple view geometry in computer vision, 2nd ed. Cambridge University Press. ISBN 0521540518.

Hui Zhang, S.A.G., Fritts, Jason E., 2008. Image segmentation evaluation: A survey of unsupervised methods. Comput. Vision and Image Understanding 110 (2), 260–280.

Koenig, A., Kessler, J., Gross, H.-M. 2008. A graph matching technique for an appearance-based, visual slam-approach using rao-blackwellized particle filters, in: Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on, pp. 1576–1581.

Okada, K., Kojima, M., Sagawa, Y., Ichino, T., Sato, K., Inaba, M. 2006. Vision based behavior verification system of humanoid robot for daily environment tasks, in: Humanoid Robots, 2006 6th IEEE-RAS International Conference on, pp. 7–12.

Okada, K., Kojima, M., Tokutsu, S., Maki, T., Mori, Y., Inaba, M. 2007. Multi-cue 3D object recognition in knowledge-based vision-guided humanoid robot system, in: Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on, pp. 3217–3222.

Okada, K., Tokutsu, S., Ogura, T., Kojima, M., Mori, Y., Maki, T., Inaba, M. 2008. Scenario controller for daily assistive humanoid using visual verification, in: Intelligent Autonomous Systems 10, 2008. IAS-10. International Conference on, pp. 398–405.

Okada, K., Kojima, M., Tokutsu, S., Mori, Y., Maki, T., Inaba, M. 2008. Task guided attention control and visual verification in tea serving by the daily assistive humanoid HRP2-JSK, in: Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on, pp. 1551–1557.

Patnaik, S. 2007. Robot Cognition and Navigation: An Experiment with Mobile Robots, Springer-Verlag, Berlin Heidelberg, ISBN:978-3-540-23446-3.

Prats, M., Wieland, S., Asfour, T., del Pobil, A., Dillmann, R. 2008. Compliant interaction in household environments by the Armar-III humanoid robot, in: Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on, pp. 475–480.

Welke, K., Przybylski, M., Asfour, T., Dillmann, R. 2008. Kinematic calibration for saccadic eye movements,, Tech. rep., Institute for Anthropomatics, University of Karlsruhe, http://digbib.ubka.uni-karlsruhe.de/volltexte/1000012977.

Wieland, S., Gonzalez-Aguirre, D., Vahrenkamp, N., Asfour, T., Dillmann, R. 2009. Combining force and visual feedback for physical interaction tasks in humanoid robots, in: Humanoid Robots, 2009. Humanoids 2009. 9th IEEE-RAS International Conference on, pp. 439–446.