Robust Real-time 6D Active Visual Localization for Humanoid Robots

D. Gonzalez-Aguirre, M. Vollert, T. Asfour and R. Dillmann
Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany
{*david.gonzalez, vollert, asfour, dillmann*}@kit.edu

***Abstract*— Overcoming the perceptual limitations of humanoid robots requires representations exploitable by highly integrable simulation, sensing, planning and acting components. Therefore, a novel active visual localization component for humanoid robots based on particle filtering in CAD environments is introduced. Specifically, two new components are presented: i) A vector-graphics prediction method employing hierarchical CAD environmental representations is presented. ii) A gaze attention method within the prediction-update cycle of the particle filter increases the available amount of visual features for localization while allowing adjustable task coupling. Finally, large and unobstructive ground-truth validation with the humanoid robot ARMAR-IIIb [1] in a made-for-humans environment shows the robustness, accuracy and performance of the proposed methods.**

I. Introduction

In order to appropriately act in the world, a mobile robot should attain its dynamic 6D pose in the environment. This is essential for metric and semantic context acquisition before planning, during execution and verification of actions. It implies environmental representations including semantic attributes, geometric models and functional descriptions. In these object-centered representations, the surrounding elements are spatially and hierarchically registered. Thus, the available environmental visual features contain all the information required to unambiguously estimate the 6D pose of the robot. For the realization of such a dynamic localization, there exist two main feature based paradigms: *bottom-up* and *top-down*. In the case of bottom-up methods, the extracted visual features are matched to the environmental representation to determine the optimal 6D pose. These methods do not assume a priori knowledge of the pose [2]. In contrast, top-down methods assume a priori knowledge of the robot pose. They are formulated as the generation, distribution and validation of hypotheses representing time varying 6D poses as state vectors. Top-down approaches track the state vector in a recursive and probabilistic manner. This models the underling dynamics of both sensor measurements and robot motions. In this domain, Bayesian filtering established as reliable state estimation framework. In particular, the non-parametric and time discrete formulation of particle filters has been widely studied in the context of robot localization. However, less attention has been given to active visual localization for humanoid robots within semantic endowed CAD environments. There, the robot not only tracks its ego-motion but also focuses its visual attention for physical interaction. Up to now, proposed representations are usually manually annotated attention zones [3] or assume bounded sets of preconditions and outcomes.

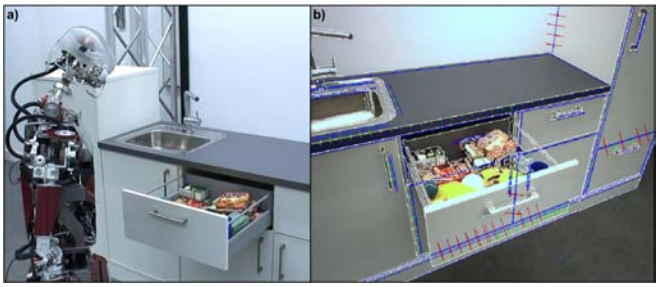

Fig. 1: a) The humanoid robot ARMAR-IIIb localizes itself while is exposed to environmental changes, b) robust visual coupling to appearance and geometric modifications.

Additionally in the literature, the visual localization for humanoid robots regards either static environments or all changes in the surroundings are considered results of previous robot interactions. This restricts the integration of robots in human-centered environment where other actors (humans, robots, etc) interact with the world producing appearance and geometric changes. The contributions in this paper alleviate these limitations by new prediction and attention methods:
Environmental visual prediction: A novel CAD (computer aided modeling [4]) prediction method in the observation model is presented. This method reduces and exploits semantically labeled and functionally described CAD representations in a general, scalable and efficient manner. Based on this prediction, it is possible to efficiently manage complex dynamic environments fully automatically.
Adjustable coupled attention: An active attention method for in-the-loop gaze planning with adjustable task coupling is presented. The gaze planning simultaneously exploits both the visually estimated 6D pose and the proposed *virtual forces* (predicted amount of visual features observable at each particular robot configuration) maximizing the information gain during localization and task execution. An adjustable coupling enables continuous fading between optimal view (for feature extraction during ego-tracking) and focus on the target in the view. This coordinates the perception-planning-action loop while improving localization.

Section II presents related work on visual localization for robots. Subsequently, the environmental representation is described in section III allowing a detailed presentation of the observation model. Based on the prediction method, the virtual forces and their dynamics are formulated into the attention model for gaze planning in section IV. Extensive experimental evaluation using reference ground-truth measurements is presented in section V. Finally, conclusions are provided in section VI.

II. Related Work

In this paper, the focus is placed on dynamic visual localization using CAD environmental representations. This clearly differentiates itself from SLAM/MOT (simultaneous localization and mapping / moving object tracking) where the environmental representation is acquired while pose estimation takes place. There are important contribution in SLAM for humanoid robots [5], [6] or SLAMMOT for other robots [7]. Despite their wide applicability, the asserted 6D poses are only linked to the initially unknown pose of the robot in the environment. Until now, SLAM/MOT representations are neither semantically nor functionally endowed. This limits their use in robots with physical interaction.

In the top-down robot localization paradigm, the application of conditional density propagation [8] for tracking was introduced in [9]. In particular [9] proved long lasting localization based on a vision and odometry in a museum by a robot equipped with a camera oriented towards the ceiling. The authors proposed an appearance-based map representation where the pose estimation is deducted solely form an intensity similarity. Despite of the limited dimensionality of the state space and rather simple observation model, the contribution notably changed the dominating paradigm on sensor-based state estimation for mobile robots.

Another pioneer work is the contribution [10]. The authors exploited dense stereo reconstruction justifying lower data association complexity during matching as well as the necessity of information for footstep planning. Their visual localization was formulated in 3D continuous state space (two DoF for position and one DoF for orientation) assuming neither odometry nor kinematic deviations. The humanoid robot H7 was the experimental platform equipped with low resolution (320x240 pixels) cameras allowing 1-3m depth range. The environmental representations were volumetric descriptions of artificially textured obstacles registered by a marker-based ground-truth system. A critical issue was ignoring the kinematic deviations and odometry drifting during pose estimation (see analysis on this issue in [5]). Despite having ground-truth measurements, authors provided no quantitative assessment of their method. Nevertheless, even with sensor limitations, strong restrictive textured environments and a simplified state space, the authors showed the plausibility of sequential Monte Carlo filtering [11] for ego-motion tracking. More recently in the work of K. Okada et al., perception, planning and action coordination have been successfully realized [3], [12]. The approaches are based on particle filters providing the robot with a full-fledged environment representation. This representation contains all necessary information for allowing the robot to accomplish complicated assignments. The constructed representation accelerates the developmental process compared to other approaches such as those in artificial intelligence and neuroscience. Since this (so-called) knowledge-based representation provides pre-stored solutions to various complex problems during task execution, for example, grasp planning, best view pose, feature selection and focus of attention. The sensor-based adaptation to partially modeled, unknown or dynamic situations was the novel contribution. Despite the enormous advances in visual localization, environmental state estimation and unified task representation, the handcrafted critical information (manual annotations called *knowledge models* as navigation spots, attention zones, etc. [12]) is not suitable for general, scalable and autonomous humanoid robots. These limitations of visual attention and feature selection are managed in this paper by the proposed prediction and attention methods. Moreover, active sensing approaches are being broadly and intensively studied. Remarkable contributions in visual localization using depth cameras [13] show promising results. In our research, the focus lies on passive cameras with active robot joints due to the wider sensing possibilities, namely those unreachable scenarios for active cameras such as large range sensing, frontal robot collaboration and outdoors scenarios just to mention a few.

III. Environmental Representation

Spatial model: Spatial hierarchy is established by the arrangement of environmental objects according to the *spatial enclosing*. The spatial enclosing of an object O_i is defined by the subspace extraction function $\Theta(O_i) : O_i \mapsto \mathbb{S}_i \subset \mathbb{R}^3$, which determines the subspace $\mathbb{S}_i$ occupied by the object. Consequently, an object O_j is denoted to be fully contained $O_j \succ O_i$ in the object O_i if and only if the subspace $\mathbb{S}_j = \Theta(O_j)$ is a proper subspace of $\mathbb{S}_i = \Theta(O_i)$. As consequence, the object O_j is placed in a lower hierarchical level compared to the hierarchy level of O_i producing a directed hierarchy tree $\mathbb{T}(\mathbb{W}, \Theta)$. In this hierarchy, three different types of nodes are distinguished: The root node $O_{\mathbb{W}}$ is an abstract entity embracing the complete spatial domain of the representation $O_i \succ O_{\mathbb{W}} \ : \ \forall O_i \in \mathbb{W}$. There is at least one leaf node O_l enclosing no subordinate object $\nexists\, O_i \in \mathbb{W} \ : \ O_i \succ O_l$. There are nodes O_b (neither leaves nor the root node) enclosing at least one subordinate object $O_i \succ O_b \ : \ \exists\, O_i \in \mathbb{W}$, see Fig. 2. Due to the acyclic structure of $\mathbb{T}(\mathbb{W}, \Theta)$, it is possible to insert, remove or change objects without affecting other spatially unrelated objects. This is the structural key for multiresolution and scalability. Depending on the functional description of an object, a directed link of the hierarchy tree $\mathcal{L}_{ij} := (O_i, O_j) \iff (O_j \succ O_i)$ can contain parametric rotations and/or translations expressed as

$$P(\mathcal{L}_{ij}) = \begin{cases} \mathcal{T}(T, \alpha, \beta, \theta) \in SE^3, & \text{if 6D Transformation} \\ L \in \mathbb{R}^3, \omega \in \mathbb{R}, & \text{if 3D-Axis-Rotation} \\ T \in \mathbb{R}^3, & \text{if 3D-Translation.} \end{cases} \tag{1}$$

Object model: The boundary description [4] is used to formulate objects as graphs composed by $O_i := (V, E, F) \in \mathbb{W}$, where V is the list of vertices, E represents the set of edges, F denotes the set of triangular surfaces. The vertices $v_k \in \mathbb{R}^3$ describe the metric of an object. The list of m vertices is denoted as $V(O_i) := \{v_k\}_{k=1}^{m}$. An edge connecting vertices v_α and v_β is expressed as $e_{\alpha\beta} := (v_\alpha, v_\beta) \mid v_\alpha, v_\beta \in V(O_i) \Rightarrow \alpha, \beta \in \mathbb{N}^+$, where $(1 \leq \alpha < \beta \leq m)$.

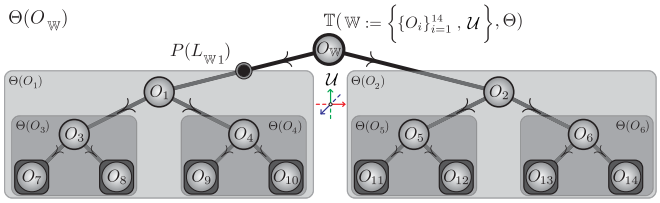

Fig. 2: The hierarchy tree of the CAD environmental representation $\mathbb{T}(\mathbb{W}, \Theta)$ is topologically organized by spatial enclosing. A parametric transformation ($P(L_{\mathbb{W}1})$ Eq. 1) dynamically affects a subtree. The collection of dynamic transformations $P(\mathcal{L}_{ij})$ describes the kinematic tree $\mathbb{T}_P(\mathbb{W}, \Theta)$.

Due to the vertex ordering ($\alpha < \beta$), there is no ambiguity (neither loops nor multigraphs) in this composition. The object O_i contains an edge set $E(O_i)$ expressed as $E(O_i) := \{e_{\alpha\beta}\}^p_{\alpha,\beta \in V(O_i)} \subset \{v_\alpha \otimes v_\beta \mid \alpha < \beta\}$, where $\otimes$ is the Cartesian product. General surfaces are approximated by subdivision into planar polygonal oriented surfaces forming closed sequences. Due to the advantages (convexity, coplanarity, verbosity, etc.), the polygonal surfaces are triangles. A triangular surface denoted as $f_{\alpha\beta\chi}$ is uniquely defined by three non-collinear vertices v_α, v_β and v_χ. Surface orientation is denoted by the normal $\widehat{N}(f_{\alpha\beta\chi})$. Finally, the union of triangular surfaces $F(O_i) := \{f_{\alpha\beta\chi}\}^q_{\alpha,\beta,\chi \in V(O_i)}$, defines the boundary of the object.

IV. Particle Filter

The fully registered body pose of a humanoid robot is determined by the robot base frame $\mathcal{P} \in SE^3$ and the time varying joint configuration $\Theta(t) \in \mathbb{R}^n$ of the multi-limb kinematic tree with n DoFs. Using the modeled direct kinematics of the robot $\mathbb{K}_{\Theta(t)} : (SE^3, \mathbb{R}^n) \mapsto SE^3$ and providing either the robot platform frame $\mathcal{P}$ or the camera frame $\mathcal{C}$, it is possible to bidirectionally determine complementary frames.

Unfortunately, there is accumulated uncertainty collected along the path (encoding resolution, irregularities, construction or wastage deviations and calibration errors) between frames of the robot kinematic chain. In practice, the estimated visual localization transformation

$$T^{\mathcal{E}}_{\mathcal{C}}(t) = T^{\mathcal{E}}_{\mathcal{P}}(t) \cdot \mathbb{K}_{\Theta(t)}(\mathcal{C}, \mathcal{P}) \tag{2}$$

deviates from its real value. Considering these and other external effects during localization can be collectively and stochastically approximated as

$$T^{\mathcal{E}}_{\mathcal{C}}(t) \approx \overbrace{T^{\mathcal{E}}_{\mathcal{P}}(t)}^{\text{estimation}} \underbrace{\mathbb{K}_{\Theta(t)}(\mathcal{H}, \mathcal{P})}_{\text{kinematic model}} \overbrace{T^{\mathcal{H}}_{\mathcal{C}}(t)}^{\text{compensation}}, \tag{3}$$

where the compensation transformation from the camera frame $\mathcal{C}$ to the neck frame $\mathcal{H}$ integrates both neck encoding $\{\alpha^{\mathcal{H}}_t, \beta^{\mathcal{H}}_t, \gamma^{\mathcal{H}}_t\} \in \mathbb{R}^3$ (roll, pitch and yaw angles) and the independent and identically distributed stochastic deviations $\{e^r_t, e^p_t, e^y_t\} \in \mathbb{R}^3$. In omni-wheel humanoid robots, it is possible to represent the transformation $T^{\mathcal{P}}_{\mathcal{E}}(t)$ using the pose $(x^{\mathcal{E}}_t, y^{\mathcal{E}}_t) \in \mathbb{R}^2$ and orientation $\alpha^{\mathcal{E}}_t \in \mathbb{R}$ of the platform.

State space: Dynamic visual localization is a 6D problem in continuous state space. Hence, the state vector $\boldsymbol{x}$ should reflect the dimensionality of the problem. By considering the uncertainty (in Eq. 3), it is plausible to simultaneously integrate joint configuration while compensating uncertainty deviations and reducing the range (in each of dimension e^i_t) of the state space without reducing the intrinsic 6D dimensionality of the process. This has the advantage of a stochastic adaptive compensation while reducing the state hypervolume (compact spreading) of the particles. Thus, the 6D state vector is

$$\boldsymbol{x}_t = (\underbrace{x^{\mathcal{E}}_t\ ,\ y^{\mathcal{E}}_t\ ,\ \alpha^{\mathcal{E}}_t}_{\text{platform pose}}\ ,\ \underbrace{e^{\mathsf{r}}_t\ ,\ e^{\mathsf{p}}_t\ ,\ e^{\mathsf{y}}_t}_{\text{neck deviation}})^{\mathsf{T}}. \tag{4}$$

Motion model: Dynamics of the robot include both platform and neck motion. The state transition function is expressed as two independent models, one for the platform pose $f_{\mathcal{P}}$ and one for the neck frame compensation $f_{\mathcal{H}}$, namely $\boldsymbol{x}_{t+1} = f(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{v}_t) = (f_{\mathcal{P}}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{v}_t), f_{\mathcal{H}}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{v}_t))^{\mathsf{T}}$, where the state vector $\boldsymbol{x}_t$ is affected by the speed vector $\boldsymbol{u}_t = \{v^{\mathcal{P}}_x, v^{\mathcal{P}}_y, v^{\mathcal{P}}_\alpha\}$ of the platform odometry and the uncertainty spread model $\boldsymbol{v}_t$. The motion model of the platform $f_{\mathcal{P}}$ requires the speed measurement in the environment frame $\mathcal{E}$. Thus, the platform orientation $\overline{\alpha}^{\mathcal{P}}$ between two discrete time steps is approximated as $\overline{\alpha}^{\mathcal{P}}(t) = \frac{1}{2}(v^{\mathcal{P}}_\alpha(t+1)\Delta t + \alpha^{\mathcal{P}}(t))$. The platform motion model includes two stochastic contributions: i) proportional speed spreading $\{\mathcal{N}(1, \sigma_{\dot{x}}), \mathcal{N}(1, \sigma_{\dot{y}}), \mathcal{N}(1, \sigma_{\dot{\alpha}})\} \in \boldsymbol{v}_t$ and ii) refinement spreading $\{\mathcal{N}(0, \sigma_x), \mathcal{N}(0, \sigma_y), \mathcal{N}(0, \sigma_\alpha)\} \in \boldsymbol{v}_t$ as

$$\begin{bmatrix} x^{\mathcal{E}}_{t+1} \\ y^{\mathcal{E}}_{t+1} \\ \alpha^{\mathcal{E}}_{t+1} \end{bmatrix} = \begin{bmatrix} x^{\mathcal{E}}_t + \mathcal{N}(0, \sigma_x) \\ y^{\mathcal{E}}_t + \mathcal{N}(0, \sigma_y) \\ \alpha^{\mathcal{E}}_t + \mathcal{N}(0, \sigma_\alpha) \end{bmatrix} + \Delta t \begin{bmatrix} \mathrm{R}(\overline{\alpha}^{\mathcal{P}}_t) & 0_2 \\ 0^{\mathsf{T}}_2 & 1 \end{bmatrix} \begin{bmatrix} v^{\mathcal{P}}_x \cdot \mathcal{N}(1, \sigma_{\dot{x}}) \\ v^{\mathcal{P}}_y \cdot \mathcal{N}(1, \sigma_{\dot{y}}) \\ v^{\mathcal{P}}_\alpha \cdot \mathcal{N}(1, \sigma_{\dot{\alpha}}) \end{bmatrix},$$

where $\mathcal{N}$ represents the normal distribution and $\mathrm{R}(\overline{\alpha}^{\mathcal{P}}_t) \in SO^2$ is the rotation submatrix. The neck compensation expressed as a state transition function is formulated as

$$\begin{bmatrix} e^r_{t+1} \\ e^p_{t+1} \\ e^y_{t+1} \end{bmatrix} = \begin{bmatrix} e^r_t \\ e^p_t \\ e^y_t \end{bmatrix} + \begin{bmatrix} \mathcal{N}(0, \sigma_t(\mathrm{r})) \\ \mathcal{N}(0, \sigma_t(\mathrm{p})) \\ \mathcal{N}(0, \sigma_t(\mathrm{y})) \end{bmatrix}, \text{ with spreading} \tag{5}$$

$$\sigma_t(i) = \begin{cases} \sigma_u & \text{if neck moves} \\ \mathbf{max}(\sigma_l\ ,\ \sigma_{t-1}(i) - \epsilon \cdot (\sigma_u - \sigma_l)) & \text{else,} \end{cases} \tag{6}$$

where the temporal convergence $0 < \epsilon < 1 \in \mathbb{R}$ implies annealing bounded to $[\sigma_l, \sigma_u] \in \boldsymbol{v}_t$.

Observation model: It determines the degree of similarity $\mathrm{P}(\boldsymbol{y}_t | \boldsymbol{x}^i_t)$ between a particle $\boldsymbol{x}^i_t$ (Eq. 4) and the real state of the humanoid robot by visual measurements $\boldsymbol{y}_t$, namely the *prediction* and *similarity assessment* methods.

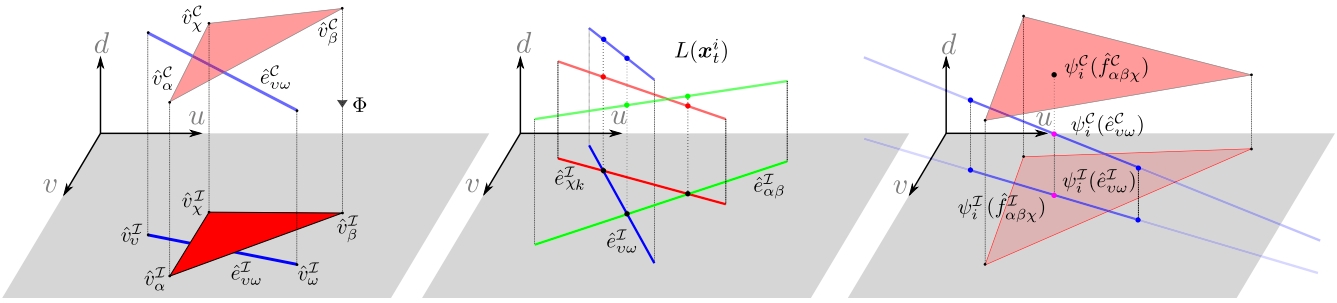

Fig. 3: CAD and vector graphics prediction method. a) During the projection process, the depth of a projected vertex is kept in the homogeneous representation for the hidden line process. b) The intersection of all clipped image edges is efficiently computed by the Sweep-Line algorithm [14]. c) The occlusion test is numerically stable and efficiently computed using barycentric coordinates.

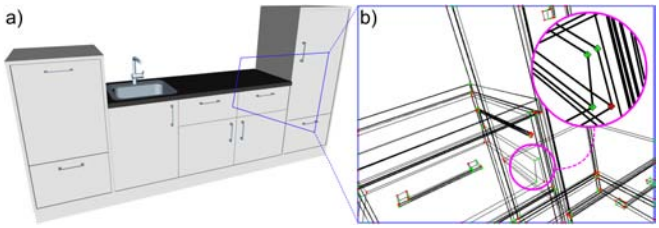

Fig. 4: a) CAD representation. b) Clipping of projected edges with visible end-points including those at image boundary.

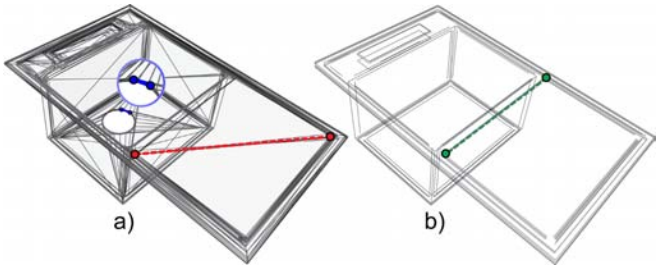

Fig. 5: CAD model reduction. a) Source representation. b) Automatic reduction by auxiliar tessellation removal (red edge), aperture (green edge) and length filtering (in blue), see summary in Tab. I.

Prediction method: The integration of three elements: i) the 6D pose of the humanoid robot (Eq. 3), ii) the CAD environmental representation and iii) the intrinsic camera calibration enables accurate virtual camera simulation within the environmental representation. Straightforward generation of images using this virtual camera to determine $\mathrm{P}(\boldsymbol{y}_t|\boldsymbol{x}_t)$ has widely been done in visual tracking, object detection and pose estimation, see [15], [16], [17]. These visual prediction methods (based on raster graphics) expose many drawbacks: i) Visual feature extraction should be applied for each hypothesis generating huge computational overhead and strongly limiting the amount of tracking particles for real time processing. ii) Even if this process is parallelized in various GPUs or large multi-core systems (usually not available in mobile systems) the rasterizing (pixel discretization and quantization) of visual features limits the reliable estimation even within short depths [18]. iii) The extraction of edge-end-points and/or junctions at arbitrary depths is unreliable even using optimal filters [19]. The proposed prediction model (based on vector graphics) overcomes these limitations by changing the raster-rendering-extraction paradigm to the continuous projection-occlusion-extraction paradigm.

This highly parallelizable method (see performance in Tab. I) simultaneously determines visibility and extracts continuous visual features from the environment represented in compact lists of visible edge segments and junctions. Based on the camera calibration matrix $K \in \mathbb{R}^{3\times3}$, image (width and height $w, h \in \mathbb{R}$) and frustum (near and far clipping depths $n, f \in \mathbb{R}$), the projection matrix [20] is

$$\Phi = \begin{bmatrix} 2k_{11} & -2k_{12}/w & 1-2k_{13}/w & 0 \\ 0 & 2k_{22}/h & 2k_{23}/h-1 & 0 \\ 0 & 0 & (f+n)/(n-f) & (2fn)/(n-f) \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

An environmental vertex $v_k^{\mathcal{E}} \in \mathbb{R}^3$ is mapped (in homogeneous coordinates $\mathbb{P}^n$) $\hat{v}_k^{\mathcal{C}} \in \mathbb{P}^3 \subset \mathbb{R}^4$ relative to the camera frame as $\hat{v}_k^{\mathcal{C}} = T_{\mathcal{E}}^{\mathcal{C}}(t)[v_k^{\mathcal{E}}\ 1]^\mathsf{T}$. Its projection $\hat{v}_k^{\mathcal{I}} \in \mathbb{P}^2 \subset \mathbb{R}^3$ on the image plane (with coordinate system $\mathcal{I}$) is expressed as $\hat{v}_k^{\mathcal{I}} = \Phi\hat{v}_k^{\mathcal{C}}$. Notice that the homogeneous depth component is preserved for the next stages, see Fig. 3-a). For objects whose bounding box $\Theta(O_i)$ is (at least partially) within the camera frustum, all edges $e_{\alpha\beta}$ are vertex-wise projected into the image plane $\hat{e}_{\alpha\beta}^{\mathcal{I}} = (\hat{v}_\alpha^{\mathcal{I}}, \hat{v}_\beta^{\mathcal{I}})$. Projected image edges are clipped inside the image boundary using the Cohen-Sutherland algorithm [21]. Fig. 4 shows this as red and green markers. Next, the intersection of all clipped image edges $\hat{e}_{\alpha\beta}^{\mathcal{I}}$ is computed using [14], see Fig. 3-b). Each of the resulting edge segments has no possible intersections with other edge. Thus, the image midpoint $\psi_i^{\mathcal{I}}(\hat{e}_{v\omega}^{\mathcal{I}})$ of each segment uniquely determines whether it is occluded by a projected triangular surface $\hat{f}_{\alpha\beta\chi}^{\mathcal{I}} = (\hat{v}_\alpha^{\mathcal{I}}, \hat{v}_\beta^{\mathcal{I}}, \hat{v}_\chi^{\mathcal{I}})$. The triangle-to-edge occlusion test is done efficiently and numerically stable using barycentric coordinates $(\lambda_1, \lambda_2) \in \mathbb{R}^2$ of the image midpoint $\psi_i^{\mathcal{I}}(\hat{f}_{\alpha\beta\chi}^{\mathcal{I}}) = \hat{v}_\alpha^{\mathcal{I}} + \lambda_1(\hat{v}_\beta^{\mathcal{I}} - \hat{v}_\alpha^{\mathcal{I}}) + \lambda_2(\hat{v}_\chi^{\mathcal{I}} - \hat{v}_\alpha^{\mathcal{I}})$. Hence, $\hat{e}_{v\omega}^{\mathcal{I}}$ is occluded only if both conditions occur: i) the image midpoint is inside the projected triangle, namely $(\lambda_1 \in [0,1]) \wedge (\lambda_2 \in [0,1]) \wedge ((\lambda_1+\lambda_2) \leq 1)$ and ii) the spatial midpoint on the triangle $\psi_i^{\mathcal{C}}(\hat{f}_{\alpha\beta\chi}^{\mathcal{C}}) = \hat{v}_\alpha^{\mathcal{C}} + \lambda_1(\hat{v}_\beta^{\mathcal{C}} - \hat{v}_\alpha^{\mathcal{C}}) + \lambda_2(\hat{v}_\chi^{\mathcal{C}} - \hat{v}_\alpha^{\mathcal{C}})$ is closer to the camera than the midpoint on the edge $\psi_i^{\mathcal{C}}(\hat{e}_{v\omega}^{\mathcal{C}})$, namely $(\psi_i^{\mathcal{C}}(\hat{f}_{\alpha\beta\chi}^{\mathcal{C}}) \cdot [0,0,1]) < (\psi_i^{\mathcal{C}}(\hat{e}_{v\omega}^{\mathcal{C}}) \cdot [0,0,1])$, see Fig. 3-c). The efficient computation of this visibility test is done by extending the Sweep-Line algorithm with the depth information for triangular faces and edges. This object-space visibility analysis [22], has an input-dependent complexity. Thus, complexity reduction is done offline, see Fig. 5.

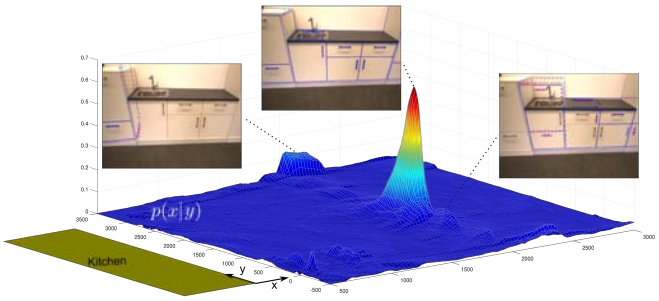

Fig. 6: Observation model $\mathrm{P}(\boldsymbol{x}_t|\boldsymbol{y}_t)$. For the offline computation of Eq. 7, the physical pose of the robot is kept static while the particles are varied in the 6D state space. For each position $(x^{\mathcal{E}}, y^{\mathcal{E}})$ the maximal value (varying all other 4-DoF) was stored. Notice the multimodality produced by symmetries in the environment.

Similarity assessment methods: The observation model asserts the degree of similarity $\mathrm{P}(\boldsymbol{y}_t|\boldsymbol{x}_t^i)$ between the prediction list of visible edges $L(\boldsymbol{x}_t^i) = \{\hat{e}_{v\omega}^{\mathcal{I}}\}$ for each hypothetical state $\boldsymbol{x}_t^i$ and the current visual measurement $\boldsymbol{y}_t$ in terms of the edge map $E : (u, v) \in \mathbb{N}^2 \mapsto \{1, 0\}$ extracted from the real robot camera. This aspect has been previously modeled in diverse manners. For example in [23], the authors established a metric based on fixed length line segments. Generalization of these and other metrics were introduced in [24]. In order to determine the optimal metric (by offline analysis with ground-truth data), the Gaussian similarity (GS) and the inlier/outliers ratio similarity (IORS) were implemented. Additionally, the inclusion of junction points was considered with both metrics producing a total of four different similarity assessment methods. For each predicted edge segment $\hat{e}_{v\omega}^{\mathcal{I}}$, a set of sampling points is distributed using a regular image length. From each of these sampling points, a Bresenham scan [25] is conducted perpendicularly to the line direction within a $\delta_{\mathbf{max}}$ range, see Fig. 7-a). The first occurrence of an active pixel in the edge map E has a distance to the edge denoted as $\delta_{v\omega,l}$. Base on these distances, the Gaussian similarity assessment P_{GS} integrates all predicted edge segments [24] as

$$\mathrm{P}_{\mathrm{GS}}(\boldsymbol{x}_t^i|\boldsymbol{y}_t) = \mathbf{exp}\left[-\frac{\sum_{v\omega}^{L(\boldsymbol{x}_t^i)} \sum_{l}^{\Omega(\hat{e}_{v\omega}^{\mathcal{I}})} \delta_{v\omega,l}^2}{2\sigma^2 \sum_{v\omega}^{L(\boldsymbol{x}_t^i)} \Omega(\hat{e}_{v\omega}^{\mathcal{I}})} \right], \quad (7)$$

where the function $\Omega : \hat{e}_{v\omega}^{\mathcal{I}} \mapsto \mathbb{N}^+$ determines the amount of sampling points for each image segment $\hat{e}_{v\omega}^{\mathcal{I}}$, see Fig. 7-b). The inlier/outliers ratio similarity $\mathrm{P}_{\mathrm{IOR}}$ determines if an active edge is present. When an active edge is found the inlier counter c_i is incremented. Otherwise, the outliers counter c_o is incremented. Its formulation (see Fig. 7-c) is expressed as

$$\mathrm{P}_{\mathrm{IOR}}(\boldsymbol{x}_t^i|\boldsymbol{y}_t) = \mathbf{exp}[-c_o/(2\sigma^2(c_o + c_i))]. \quad (8)$$

The similarity using junction points is computed as in Eq. 7 and Eq. 8 where the prediction list $\check{L}(\boldsymbol{x}_t)$ contains only short visible edges connected to junction points, see Fig. 6.

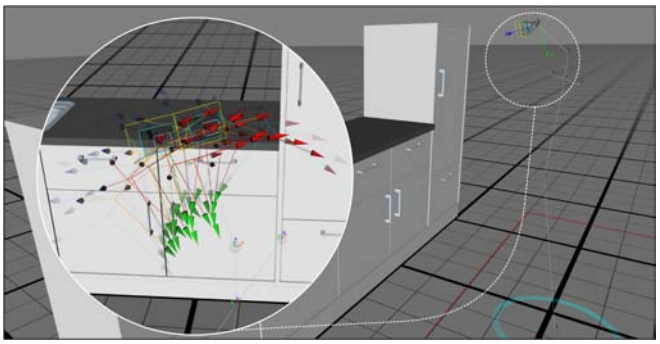

Fig. 8: The proposed attention method for gaze planning is based on information gain by efficiently sampling the amount of visual features (virtual force) available at each particular neck configuration using vector graphics prediction method from Fig. 3. Transparency shows amount of information per configuration.

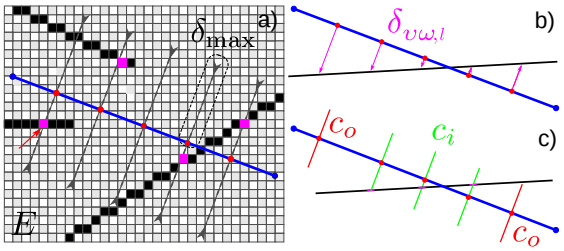

Fig. 7: a) Sampling points (red) on the predicted edge (blue) by Bresenham scan [25] along the normal direction for detection of sensed edges (black). b) The Gaussian similarity is estimated based on a continuous model (Eq. 7) of the distance to the first active (magenta) pixel. c) Inlier (green) / outliers (red) ratio of similarity.

Attention model: Frequently, it is necessary to track the ego-pose of the humanoid robot while concurrently analyzing the scene. For instance, during object exploration or shape based object categorization (see [26] particularly figures 4-6). Ideally, fixating an attention target $\boldsymbol{T} \in \mathbb{R}^3$ while exploring or visually reconstructing the scene from diverse view points is realizable. However, in common situations (for instance, in front of a door), due to the lack of environmental visual features for the ego-tracking, the assessed 6D poses are not adequate for online multiview registration. This is achieved by our attention method which smoothly controls the robot gaze orientation while switching between subtasks.

The gaze planning is formulated as a spring-mass system: The current neck configuration $\boldsymbol{C} = (c_r, c_p, c_y)^\intercal$ has a virtual mass m. The virtual forces w_s^i correspond to the amount of visual features predicted at each sampled neck configuration $\boldsymbol{S}^i$, see Fig. 8. The effect of these virtual forces is expressed as $\boldsymbol{f}_s(\boldsymbol{S}^i) = k_s w_s^i(\boldsymbol{S}^i - \boldsymbol{C})$ where $k_s \in \mathbb{R}$ denotes the Hooke's constant. By discretization of the neck configuration space (see Fig. 9), it is possible to approximate the effects of all n virtual forces to estimate the next configuration with higher amount of visual features.

Hence, the collective effect of all sampled virtual forces is expressed as $\boldsymbol{f}_s = \sum_{i=1}^{n} \boldsymbol{f}_s(\boldsymbol{S}^i)$. Additionally, the target location (in terms of neck configuration by the inverse kinematics $\mathbb{K}_{\Theta(t)}^{-1}(\mathcal{P}, \boldsymbol{T}) = \mathcal{H}_{\boldsymbol{T}}$) also applies a virtual force on the current configuration as $\boldsymbol{f}_T = k_{\boldsymbol{T}}(\boldsymbol{T} - \boldsymbol{C})$.

The target constant $k_{\boldsymbol{T}} \in \mathbb{R}$ serves fading purposes when adjusted relative to k_s. In addition to these external virtual forces, there are two internal virtual forces. First, because the neck configuration should transit smoothly from one configuration to another, the mass effects are considered by the virtual force $\boldsymbol{f}_m = m\ddot{\boldsymbol{C}}$ which forbids abrupt motions producing a more human-like behavior, less joints stress and sharp images for the ego-tracking. Second, in order to avoid undesired oscillations, a damping effect is also considered by the force $\boldsymbol{f}_d = -d\dot{\boldsymbol{C}}$, where $d \in \mathbb{R}$ denotes the viscous damping coefficient serving convergence purposes. Considering discrete time, the dynamic system is

$$\ddot{\boldsymbol{C}}_{t+1} = \frac{1}{m}\left(k_T(\boldsymbol{T} - \boldsymbol{C}_t) - d\dot{\boldsymbol{C}}_t + k_s \sum_{i=1}^{n} w_s^i(\boldsymbol{S}_t^i - \boldsymbol{C}_t)\right).$$

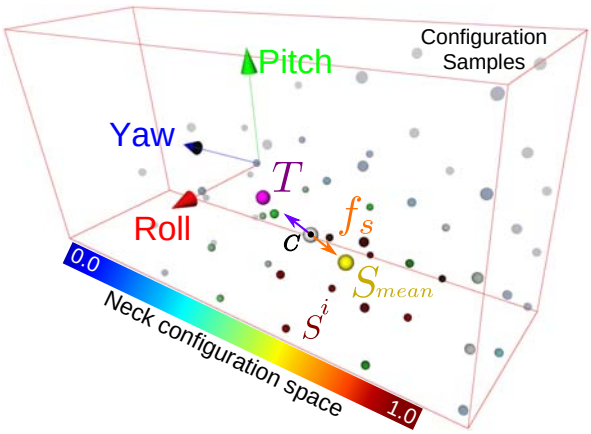

Fig. 9: Spring-mass system with virtual forces for gaze planning in the configuration space of the humanoid neck, see Tab. VI.

The first integral approximation is $\dot{\boldsymbol{C}}_{t+1} = \dot{\boldsymbol{C}}_t + \Delta_t \ddot{\boldsymbol{C}}_{t+1}$ and thereafter the resulting configuration is $\boldsymbol{C}_{t+1} = \boldsymbol{C}_t + \Delta_t \dot{\boldsymbol{C}}_{t+1}$. The fading between the target force and the predicted forces w_s^i is modeled by the variable $0 \leq \tau \leq 1$. Thus, the target gain is $k_T = \tau$ and the prediction gain $k_s = (1-\tau) \cdot (\mathbf{max}(w_s^i)/\sum_i^n w_s^i)$. This normalization holds k_s in cases where the prediction provides no salient cue.

V. Experimental Evaluation

The performance of the implementation is shown in Tab. IV. These results were obtained on a CPU Intel Core i7, 2.93GHz using 200 particles at 15 FPS. The model reduction and prediction methods were evaluated with two environmental models demonstrating the high performance of the approach, see Tab. I. During the evaluation of the localization precision, the registration of the robot platform and environmental objects were done using a precise and high-speed marker-based system, see details on the whole registration process in our previous work [18]. The storage of all tracking data (see Fig. 10) including ground-truth poses, raw camera images and platform speed measurements allow the systematic offline evaluation under the exact same conditions.

Furthermore, since the methods are stochastic, all evaluations were computed five times for all frames to estimate the mean, RMS and maximal errors. This enables the evaluation of various estimation methods from the collection of particles, see Tab. V. Further, four similarity assessment methods were evaluated (including annealing) to determine their accuracy, see Tab. II and Tab. III.

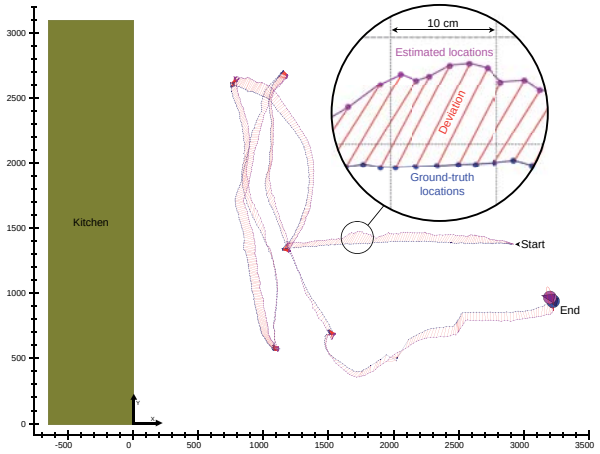

Fig. 10: Visualization of evaluation track.

VI. Conclusions

The contributions of this paper are two methods for improving the flexibility and robustness of the visual dynamic localization of humanoid robots in CAD modeled environments. First, our visual prediction method avoids the rasterization drawbacks while extracting visual features from object models. In contrast to raster graphic predictors, our prediction method is based on vector graphics and it is realized by efficient algorithms and data structures for exploiting the proposed general and extensible CAD environmental representation. Since the method is efficient and fully automatic, a robot moving from one place to another could access the environmental representations to dynamically localize itself for interaction with the environment. Second, the contributed visual attention method enables the humanoid robot to smoothly plan its gaze by proper integration of the environmental representation, the estimated pose and the task target. The transition between these visual processes is adjusted by a robust fading coupling. Finally, an extensive experimental evaluation with ground-truth registration shows that when using the prediction with Gaussian similarity, it is possible to ensure the dynamic pose of the robot within an average deviation of less than five centimeters, see Tab. V. Notice that this motion accuracy is measured while manually and arbitrarily changing the speed and orientation of the robot within the full (4x4m) capture area of the motion caption lab. When the robot remains static, the refinement process reduces the localization deviation up to one order of magnitude. The slightly less accurate localization (see Tab. VI) attained with the attention model is neglectable considering the fact that during the experimentation with active attention the robot 6D pose was never lost.

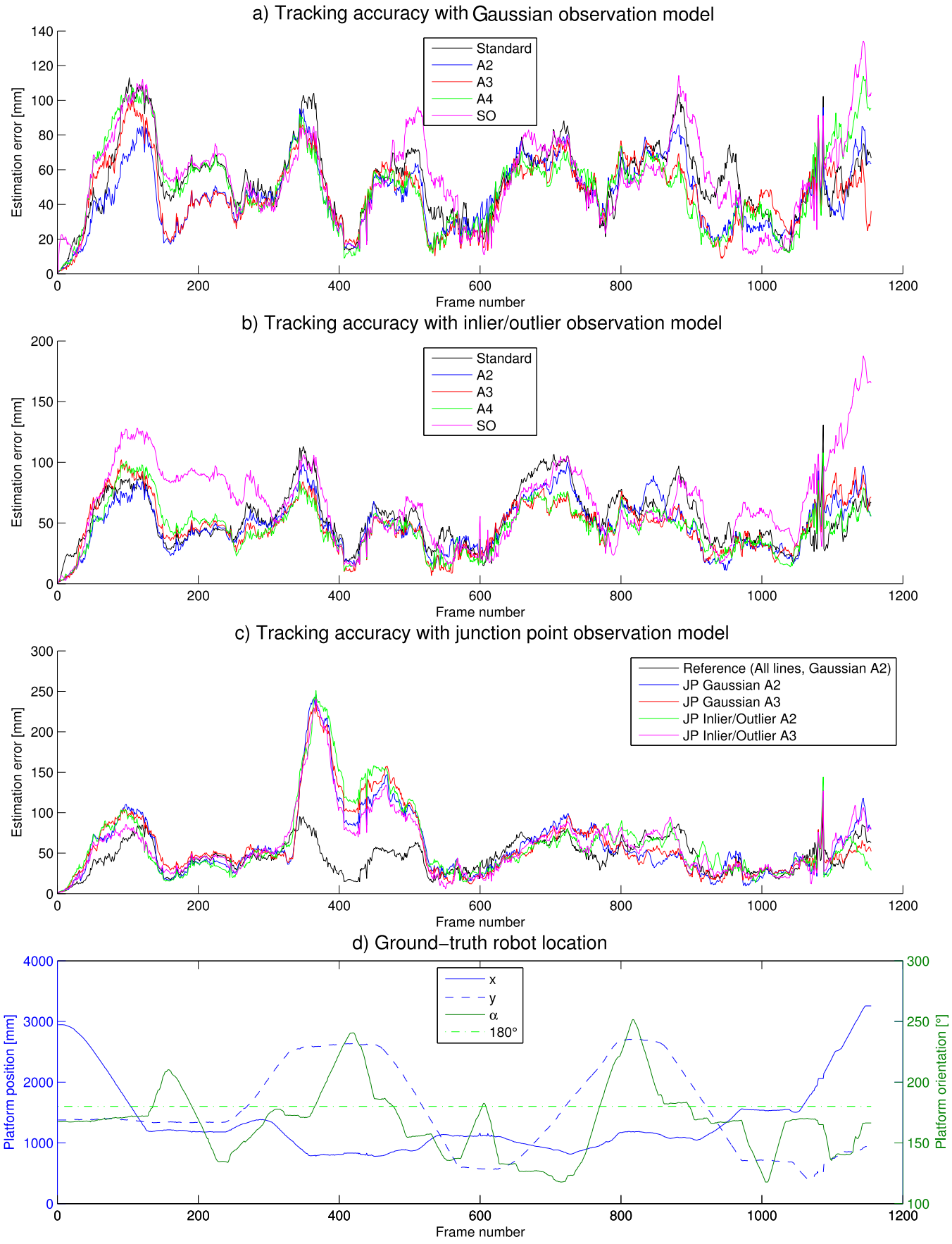

Fig. 11: Plots of the evaluation track.

VII. ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement no 611832 (WALK-MAN) and grant agreement no 611909 (KoroiBot).

Environment Representations	Edges	Junctions
Full Mobile Kitchen (3234 edges, 677 faces)	8	8
Reduced Mobile Model (130 edges, 112 faces)	2325	1720
Full Kitchen Room (3775 edges, 2069 faces)	6	5.7
Reduced Kitchen Room (415 edges, 306 faces)	370	250

TABLE I: Reduction and prediction performance in FPS.

Similarity Assessment Method	None	A2	A3	A4
GS - Mean	53.8	45.8	46.6	50.4
GS - RMSE	60.7	51.2	51.8	57.7
GS - Max	201.2	121.1	136.4	176.9
IORS - Mean	55.3	48.9	47.9	46.3
IORS - RMSE	59.9	54.6	53.8	52.4
IORS - Max	161.1	149.9	141.5	197.5

TABLE II: Accuracy (in mm) with **G**aussian **S**imilarity and **I**nlier / **O**utliers **R**ate **S**imilarity. Notice annealing A2,3,4.

Error Metric	GS			IORS		
Annealing	None	A2	A3	None	A2	A3
Mean	69.8	61.3	61.5	68.4	63.3	60.2
RMSE	87.02	76.6	76.5	87.6	79.6	74.0
Max	291.4	255.6	250.6	283.9	270.3	256.3

TABLE III: Accuracy in mm using junctions.

Threads	1	2	3	4	5	6	7	8
GD	4.96	9.8	14.6	19.4	24.1	27.0	32.2	37.6
DR	4.92	6.0	6.6	7.2	7.3	7.2	7.3	8.07

TABLE IV: Particle filter performance with 200 particles in FPS. **G**lobal **D**istributed. **D**istributed **R**esampling.

Error Metric	Weighted Mean	MAP	Threshold Mean
Mean	45.8	51.0	50.0
RMSE	51.2	56.9	55.9
Max Error	121.1	294.9	138.0

TABLE V: Pose accuracy in mm. **M**aximum **a** **P**osteriori.

Error Metric	GS			IORS		
Annealing	None	A2	A3	None	A2	A3
Mean	64.1	54.6	59.2	70.7	61.3	64.2
RMSE	73.5	64.9	69.3	80.2	72.4	74.3
Max	204.2	179.4	204.6	244.3	213.7	201.7

TABLE VI: Accuracy in mm using the attention model.

REFERENCES

[1] 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," in *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, 2006, pp. 169–175.

[2] D. Gonzalez-Aguirre, T. Asfour, E. Bayro-Corrochano, and R. Dillmann, "Model-Based Visual Self-localization Using Gaussian Spheres," in *Geometric Algebra Computing*, E. Bayro-Corrochano and G. Scheuermann, Eds. Springer London, 2010, pp. 299–324.

[3] K. Okada, M. Kojima, S. Tokutsu, Y. Mori, T. Maki, and M. Inaba, "Task guided attention control and visual verification in tea serving by the daily assistive humanoid hrp2jsk," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, 2008, pp. 1551–1557.

[4] G. Farin, J. Hoschek, M.-S. Kim, J. Hoschek, and M.-S. Ki, *Handbook of Computer Aided Geometric Design*. North-holland, 2002.

[5] O. Stasse, A. J. Davison, R. Sellaouti, and K. Yokoi, "Real-time 3d slam for humanoid robot considering pattern generator information," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, oct. 2006, pp. 348 –355.

[6] S. Yoon, S. Hyung, M. Lee, K. S. Roh, S. Ahn, A. Gee, P. Bunnun, A. Calway, and W. W. Mayol-Cuevas, "Real-time 3D simultaneous localization and map-building for a dynamic walking humanoid robot," *Advanced Robotics*, vol. 27, no. 10, pp. 759–772, 2013.

[7] K.-H. Lin and C.-C. Wang, "Stereo-based simultaneous localization, mapping and moving object tracking," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, Oct 2010, pp. 3975–3980.

[8] M. Isard and A. Blake, "CONDENSATION: Conditional Density Propagation for Visual Tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[9] F. Dellaert, W. Burgard, D. Fox, and S. Thrun, "Using the condensation algorithm for robust, vision-based mobile robot localization," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2, 1999, pp. –594 Vol. 2.

[10] S. Thompson and S. Kagami, "Humanoid robot localisation using stereo vision," in *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, 2005, pp. 19–25.

[11] F. Dellaert, D. Fox, W. Burgard, and S. Thrun, "Monte Carlo Localization for Mobile Robots," in *IEEE International Conference on Robotics and Automation (ICRA99)*, May 1999.

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

[13] J. M. Carranza and W. Mayol-Cuevas, "Real-Time Continuous 6D Relocalisation for Depth Cameras," in *Robotics Science and Systems*. to appear, June 2013.

[14] U. Bartuschka, K. Mehlhorn, and S. Nher, "A robust and efficient implementation of a sweep line algorithm for the straight line segment intersection problem," in *Workshop on Algorithm Engineering*, 1997, pp. 124–135.

[15] W.-H. Chang, C.-H. Hsia, Y.-C. Tai, S.-H. Chang, F. Ye, and J.-S. Chiang, "An efficient object recognition system for humanoid robot vision," in *Pervasive Computing (JCPC), 2009 Joint Conferences on*, dec. 2009, pp. 209 –214.

[16] V. Lepetit, L. Vacchetti, D. Thalmann, and P. Fua, "Fully automated and stable registration for augmented reality applications," in *Mixed and Augmented Reality, 2003. Proceedings. The Second IEEE and ACM International Symposium on*, oct. 2003, pp. 93 – 102.

[17] G. Klein and D. Murray, "Full-3d edge tracking with a particle filter," in *Proc. British Machine Vision Conference (BMVC'06)*, vol. 3. Edinburgh: BMVA, September 2006, pp. 1119–1128.

[18] D. Gonzalez-Aguirre, T. Asfour, and R. Dillmann, "Ground-Truth Uncertainty Model of Visual Depth Perception for Humanoid Robots," in *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on*, 2012, pp. 436–442.

[19] M. A. Oskoei and H. Hu, "A survey on edge detection methods," School of Computer Science & Electronic Engineering University of Essex, Colchester CO4 3SQ, United Kingdom, Tech. Rep., 2010.

[20] A. Straw, "Augmented reality - computing the OpenGL projection matrix from intrinsic camera," [Online; accessed 16-May-2013].

[21] M. Agoston, *Computer Graphics and Geometric Modelling: Mathematics*, ser. Computer Graphics and Geometric Modeling: Implementation and Algorithms. Springer, 2005.

[22] I. E. Sutherland, R. F. Sproull, and R. A. Schumacker, "A characterization of ten hidden-surface algorithms," *ACM Comput. Surv.*, vol. 6, no. 1, pp. 1–55, Mar. 1974.

[23] K. Okada, M. Kojima, S. Tokutsu, T. Maki, Y. Mori, and M. Inaba, "Multi-cue 3D object recognition in knowledge-based vision-guided humanoid robot system," in *IROS 2007*, 2007, pp. 3217 –3222.

[24] M. L. Pupilli, "Particle filtering for real-time camera localisation," Ph.D. dissertation, University of Bristol, 2006.

[25] J. E. Bresenham, "Algorithm for computer control of a digital plotter," *IBM Systems Journal*, vol. 4, no. 1, pp. 25–30, 1965.

[26] D. Gonzalez-Aguirre, J. Hoch, S. Roehl, T. Asfour, E. Bayro-Corrochano, and R. Dillmann, "Towards shape-based visual object categorization for humanoid robots," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011, pp. 5226 – 5232.