Proceedings of the 2006 IEEE/RSJ
International Conference on Intelligent Robots and Systems
October 9 - 15, 2006, Beijing, China

Hierarchical Featureless Tracking for Position-Based 6-DoF Visual Servoing

Wolfgang Sepp, Stefan Fuchs and Gerd Hirzinger
Institute of Robotics and Mechatronics
German Aerospace Center (DLR)
82234 Wessling, Germany
Email: {wolfgang.sepp,stefan.fuchs,gerd.hirzinger}@dlr.de

***Abstract*— Classical position-based visual servoing approaches rely on the presence of distinctive features in the image such as corners and edges. In this contribution we exploit a hierarchical approach for object detection, initial-pose estimation, and real-time tracking based first on colour distribution and subsequently on the shape and texture information. The shape model of the object is not limited to surface primitives but allow for any free-form surface not subject to self-occlusion. We evaluate the approach as part of a handshake scenario where a 7-DoF robot takes a free moving object over from a human.**

I. Introduction

Present robotic research efforts are directed towards intelligent service-robots operating side-by-side and also closely together with humans. The tactile contact plays an important role in human-machine interaction, additionally to the visual and auditive information. This contact is realized whenever a person hands an object over to the robot for manipulation.

In order to perform this handshake the robot has to fulfill autonomously different tasks. First, the object has to be detected and classified. Then, the pose of the object has to be estimated to enable tracking of the object motion in 3-D. The latter task has to be performed with 6-DoF to ensure a precise grasp and accurate subsequent manipulation.

The problem has been addressed in the past either with respect to robot control or with respect to computer vision. Grasping of moving objects has been shown for simple geometric objects (e.g. [1], [2]) or for the interception of objects moving on a plane [3] whereas the DoFs are usually restricted.

Image-based visual servoing approaches benefit from a tight coupling of vision and control in that exact knowledge about the object and about hand-eye calibration become dispensable. However, these approaches show view dependency especially with respect to the object distance. Position-based servoing on the other hand uses vision algorithms to directly estimate the relative pose between robot end-effector and the object. Malis et al. [4] combined the benefits of image and position-based visual servoing in their $2\frac{1}{2}$-D approach. The approach however is designed for co-planar features points which severely restricts the objects shape.

The majority of 3-D tracking approaches for position-based servoing rely on the localisation of a-priori known artificial or natural landmarks, e.g. [2], [5]. Often, these approaches depend on the correspondence between brightness steps and edges of the 3-D surface, which are not present for general free-form surfaces.

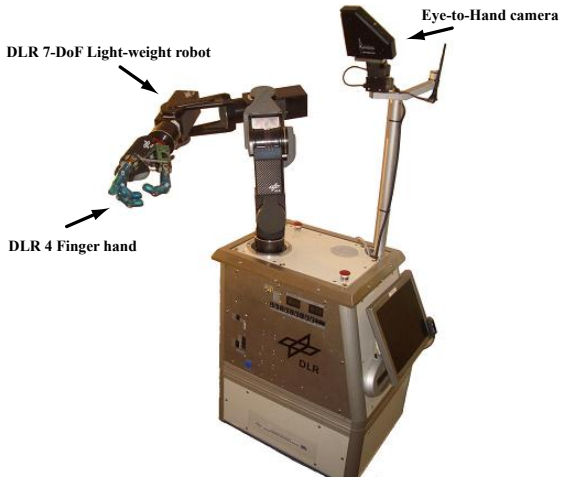

Fig. 1. The DLR-Robutler: A mobile service robot with a 7-DoF light-weight robotic arm and an eye-to-hand camera system.

Direct methods on the other hand, deduce the object motion on a pixel basis. Many approaches have been developed since the pioneering work of Lucas and Kanade [6] but concentrate on the efficient tracking of planes by means of the 8-DoF homography, e.g. Diehl et al. [7], Buenaposada et al. [8] and Baker et al. [9].

Some efforts have been made to upgrade tracking of planar surfaces primitives to more general surfaces. Cernuschi-Frias et al. [10] presented an estimation model for simple parametrised surfaces while Ramey et al. [11] described the target surface with B-splines. Lately, Sepp [12] derived an analytic description of the tracking problem for general surfaces modeled as 3-D point clouds.

In the following we present position-based visual servoing for a mobile robot with a 7-DoF robotic arm. A camera is rigidly mounted on the mobile platform and overlooks the scenery. Here, we don't focus only on a tracking approach but solve also the detection and re-initialisation problems leading to a robust visual servoing application.

First, the process of hierarchical detection and tracking is outlined in Sect. II. The first stages rely on matching by means

1-4244-0259-X/06/$20.00 ©2006 IEEE

of a colour histogram as described in Sect. III while the latter stages are build on shape and texture as similarity measure (see Sect. IV). We evaluate and compare the single stages in Sect. V and draw our conclusions in the final Sect. VI.

II. Hierarchical Approach

Solid visual servoing applications rely on a robust re-initialisation after the target gets lost. In the following a hierarchical approach for object detection, localisation, and tracking is presented which resumes the idea of incremental focus of attention of Toyama and Hager [13]. The localisation property increases step by step from a 2-DoF object detection to an accurate 6-DoF pose representation. For this purpose, the object is first localized and tracked based on a colour histogram as similarity measure which is generally not pose discriminative. Thereafter, the pose is refined to 6-DoF by a similarity measure based on shape and texture. The levels of localisation and tracking are:

1) 2-DoF object detection based on colour histogram
2) 3-DoF object tracking with Mean-Shift
3) initial 6-DoF object tracking with particle filter
4) accurate 6-DoF object tracking with IC-R

Conversely, when the object is lost on a specific level, the next lower level of localisation or tracking is invoked (see Fig. 2). The levels are changed following individual error thresholds.

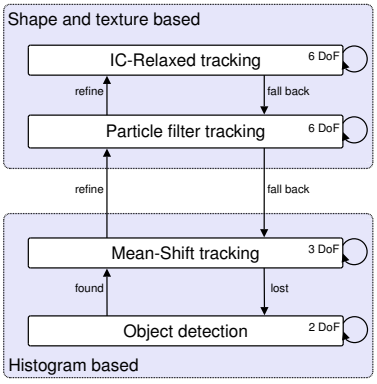

Fig. 2. Hierarchical featureless tracking scheme.

III. Histogram-based Localisation and Tracking

Histograms of colour or gray-level gradient distributions are frequently used because they are robust with respect to illumination changes and to pose variations. Here, we focus on the colour distributions of Comaniciu et al. [14] as target models.

Let therefore $I(\mathbf{v}) \in \mathbb{N}^3$ denote the colour value at the image position $\mathbf{v} \in \mathbb{N}^2$. The function $h(\mathbf{c}) \in \{1, 2, \ldots, m\}$ maps each colour value $\mathbf{c} \in \mathbb{N}^3$ to a colour-bin index. Then, the local colour probability for the specific colour-bin b at the position $\mathbf{u}$ in the image is defined by

$$p(\mathbf{u}, b) = \frac{\sum_{\mathbf{v}} k_\sigma(\mathbf{u}-\mathbf{v})\, \delta(h(I(\mathbf{v})) - b)}{\sum_{\mathbf{v}} k_\sigma(\mathbf{u}-\mathbf{v})} \tag{1}$$

where δ is the Kronecker delta function and where the colour occurrence is weighted by an anisotropic kernel function

$$k_\sigma(\mathbf{u}) = \exp\left(-\frac{1}{2}\mathbf{u}^{\mathrm{T}} \mathrm{diag}\left(\sigma^{-2}\right) \mathbf{u}\right) \tag{2}$$

according to the spatial distance to the center. Here, the diagonal 2×2 matrix $\mathrm{diag}\left(\sigma^{-2}\right)$ ensures individual scaling along the axis.

Now, the object to be tracked is recognized by the prior colour probability $q(b)$ obtained from a reference image. The similarity between a local colour distribution p and the prior distribution q is measured by the Bhattacharyya coefficient

$$\rho(p, q) = \sum_b \sqrt{p(\mathbf{u}, b)\, q(b)} \tag{3}$$

which results in 1 for identical distributions.

A. Localisation

The object can be detected evaluating the Bhattacharyya coefficients in the image plane. In order to reduce the occurrence of local minima and to decrease the computational costs, the coefficients are computed at selected positions only.

In this sense, let σ_x^2 and σ_y^2 be the variance of the kernel k which determine the spatial decay of the function. The Bhattacharyya coefficients are sampled at intervals $2\sigma_x$, $2\sigma_y$ which assures good coverage in the image. Let $S = \{(2i\sigma_x, 2j\sigma_y) | i, j \in \mathbb{N}\}$ be the set of sampling coordinates, then two dimensional object localisation is performed by

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{u} \in S} \sum_b \sqrt{p(\mathbf{u}, b)\, q(b)} \tag{4}$$

where the detection is rejected whenever the corresponding Bhattacharyya coefficient is below the detection threshold θ_{D}.

B. Tracking

Object tracking starts immediately after localisation in the image plane. At this stage the initial two-dimensional position is augmented by the additional scale parameter.

Tracking starts by computing the Mean-Shift of Comaniciu et al. [14] in planar coordinates. Here, every position in the neigbourhood of the current estimate is weighted according to the relevance of the corresponding colour-histogram value. The relevance is determined by the ratio between the probability of the colour value in the reference pattern to the corresponding probability in the current pattern, which reads

$$w(\mathbf{u}, \mathbf{v}) = \sum_b \sqrt{\frac{q(b)}{p(\mathbf{u}, b)}} \delta(h(I(\mathbf{v})) - b) \tag{5}$$

for an image coordinate $\mathbf{v}$ with respect to the location $\mathbf{u}$. Again, the kernel function of (2) is used and finally the new location $\hat{\mathbf{u}}$ of the object is estimated by

$$\hat{\mathbf{u}} = \frac{\sum_{\mathbf{v}} k_\sigma(\mathbf{u}-\mathbf{v})\, w(\mathbf{u}, \mathbf{v})\, \mathbf{u}}{\sum_{\mathbf{v}} k_\sigma(\mathbf{u}-\mathbf{v})\, w(\mathbf{u}, \mathbf{v})} \quad . \tag{6}$$

Subsequently, the scale of the object is identified by the evaluation of its relevance at $2n+1$ scales

$$\sigma_s = \sigma \cdot a^s \; ; \quad -n \leq s \leq n \tag{7}$$

based on the current scale σ and base $a > 1$ of logarithmic coordinate scale. The new scale index $\hat{s}$ and scale $\hat{\sigma}$ are estimated according to

$$\hat{s} = \frac{\sum_s \sum_{\mathbf{v}} k_{\sigma_s}(\hat{\mathbf{u}} - \mathbf{v})\, w(\hat{\mathbf{u}}, \mathbf{v})\, s}{\sum_s \sum_{\mathbf{v}} k_{\sigma_s}(\hat{\mathbf{u}} - \mathbf{v})\, w(\hat{\mathbf{u}}, \mathbf{v})} \; ; \quad \hat{\sigma} = \sigma \cdot a^{\hat{s}} \; . \tag{8}$$

Note that here we simply use the same kernel as in (6) instead of a Laplacian of Gaussian or a Difference of Gaussian as proposed by Collins [15]. The steps of shift and scale computation are alternated with continuous update of the location $\mathbf{u}$ and the scale σ leading to a 3-DoF tracking process.

The parameters shift $\mathbf{u} = (u_x, u_y)$ and scale σ are mapped to position information in 3-D for a full-perspective projection, given the knowledge about the intrinsic camera parameters

$$K = \begin{pmatrix} \alpha & 0 & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{9}$$

and the physical extent d of the object. The latter is related to the position t_z in direction of the z-axis by the formula $\alpha \cdot d \propto t_z \cdot \sigma$. Therefore, the translation parameters of the object are computed according to

$$\begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} = \frac{\alpha d}{\sigma} \begin{pmatrix} (u_x - u_0)/\alpha \\ (u_y - v_0)/\beta \\ 1 \end{pmatrix} \quad . \tag{10}$$

Tracking with the extended Mean-Shift approach is performed for a constant amount of time and switches back to the previous detection stage when the Bhattacharyya coefficient falls below the threshold θ_{D}. The next tracking stage is invoked as soon as the minimization process converges, that is the Euclidean distance $\|\hat{\mathbf{u}} - \mathbf{u}\|$ and $|\hat{\sigma} - \sigma|$ falls below a convergence threshold θ_{M}.

IV. Shape-Texture-based Tracking

In order to get exact information about the object pose in 6-DoF, *brightness* information is linked to *s*hape information. Here, the object surface is modeled by an unordered set of sample points $X = \{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N\} \subset \mathbb{R}^3$. No constraint other than visibility is imposed on the surface. The surface points $\mathbf{x} \in X$ are subject to rigid-body motion described by

$$m(\mathbf{x}, \mu) = R(\mu)\, \mathbf{x} + t(\mu) \tag{11}$$

for a pose $\mu \in \mathbb{R}^6$ and the associated 3-D rotation $R(\mu)$ and translation $t(\mu)$. The so transformed point cloud is finally mapped to the image under full perspective projection

$$p(\mathbf{x}) = \left(\frac{\mathbf{k}_1^{\mathrm{T}} \cdot \mathbf{x}}{\mathbf{k}_3^{\mathrm{T}} \cdot \mathbf{x}}, \frac{\mathbf{k}_2^{\mathrm{T}} \cdot \mathbf{x}}{\mathbf{k}_3^{\mathrm{T}} \cdot \mathbf{x}} \right)^{\mathrm{T}} , \quad K = \begin{pmatrix} \mathbf{k}_1^{\mathrm{T}} \\ \mathbf{k}_2^{\mathrm{T}} \\ \mathbf{k}_3^{\mathrm{T}} \end{pmatrix} \tag{12}$$

where $K \in \mathbb{R}^{3\times3}$ is the matrix of intrinsic camera parameters. Let $I(\mathbf{v}) \in \mathbb{R}$ be the brightness value of the *current* image at position $\mathbf{v}$ then the texture value for a single model point $\mathbf{x}_i \in X$ is determined by

$$I_i(\mu) = I(p(m(\mathbf{x}_i, \mu))) \tag{13}$$

for the pose μ. The surface texture is assumed to remain constant over time. The right pose estimation maximizes the above similarity between the current texture $I_i(\mu)$ and the reference texture $T_i(\mu^0)$ of the surface for the image T and pose μ^0. The relationship between similarity and pose estimation μ is hereafter expressed by the probability density function (p.d.f.)

$$p(I|\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{t}}} \exp\left(-\frac{\left(I_i(\mu) - T_i(\mu^0)\right)^2}{2\sigma_{\mathrm{t}}^2} \right) \quad , \tag{14}$$

given conditionally independent observations $\mathbf{x} \in X$. Thus, the objective is to maximize the probability with respect to the pose μ.

However, the illumination on the surface changes as the object moves and generates biased measurements. This effect can be alleviated by normalizing the brightness value distribution in I_i to that in T_i. Let $\mathrm{E}(.)$ be the expected value of a distribution and let $\mathrm{Var}(.)$ be the variance, then the current texture I_i in (14) is replaced by

$$\bar{I}_i(\mu) = \frac{\mathrm{Var}(T_i)}{\mathrm{Var}(I_i)} \left(I_i - \mathrm{E}(I_i)\right) + \mathrm{E}(T_i) \quad . \tag{15}$$

The task of finding the most probable pose estimation is a classical minimisation problem solve here in two stages. A first estimation is determined by Monte-Carlo sampling which features a large range of convergence. A second stage involves local Gauss-Newton minimisation showing fast convergence.

A. Particle Filter Tracking

Particle filtering is employed in order to refine the 3-DoF pose estimation from the previous Mean-Shift tracking. It serves also as the first stage of 6-DoF tracking. The method approximates the posterior density $p(\mu|I)$ through M pose samples $\mu^{(i)}$ and weights $\pi^{(i)}$ (see [16] for detailed description of particle filtering). Initially, the distribution of samples is set to the single pose estimation $\hat{\mu} = (0, 0, 0, t_x, t_y, t_z)$ fed from Mean-Shift tracking. The pose samples are propagated by a simple stochastic process

$$\mu_t^{(i)} = \mu_{t-1}^{(i)} + w_{t-1}^{(i)} \tag{16}$$

where $w_{t-1}^{(i)}$ represents white normal process noise $p\left(w_{t-1}^{(i)}\right) \sim N(0, Q)$ with covariance matrix Q. The p.d.f. (14) is sampled at $\mu_t^{(i)}$ and assigned to the weights $\pi_t^{(i)}$.

Potentially many different informations can be extracted from the posterior density $p(\mu|I)$ but here we are interested only in a single mode which is the mean pose of the object

$$\mathrm{E}(\mu) = \frac{1}{\sum_i \pi_t^{(i)}} \sum_i \pi_t^{(i)} \mu_t^{(i)} \quad . \tag{17}$$

At the next iteration the samples $\mu_t^{(i)}$ are resampled with a probability proportional to the associated weights $\pi_t^{(i)}$. Despite the loss in frame-rate, we perform several iteration steps of the particle filter. We consider particle filtering here as a method for annealed maximisation which is reflected by the absence

of a deterministic component in the dynamic model (16). Tracking at this stage is performed as long as the p.d.f. at the mean pose $\mathrm{E}(\mu)$ is within the interval $[\theta_{\mathrm{P}^-}, \theta_{\mathrm{P}^+}]$.

B. IC-R Tracking

The likelihood of a pose estimation equals p.d.f. (14) as a function of μ, that is $L(\mu) = p(I|\mu)$. Maximisation of the logarithmic likelihood is equivalent to the minimisation $\hat{\mu}^\star = \arg\min_{\hat{\mu}} O(\hat{\mu})$ of

$$O(\mu) = \sum_{i=1}^{N} \left[I_i(\mu) - T_i\big(\mu^0\big) \right]^2 \quad , \tag{18}$$

which is a sum-of-squared differences dissimilarity measure. Sepp [12] considers rigid-body motion as a composition of two motions, a global motion estimate $\hat{\mu}$ and a differential motion $\hat{\delta\mu}$. The texture (13) is thus redefined to the function

$$I_i(\delta\mu) = I(p(m(m(\mathbf{x}_i, \delta\mu), \hat{\mu}))) \tag{19}$$

of the pose variation $\delta\mu$. The accordingly modified objective function (18) is minimized with a Gauss-Newton approximation to the Hessian by iteratively solving the linear equation system

$$\sum_{i=1}^{N} \left[\partial_{\delta\mu} I_i\right]^{\mathrm{T}} \left[\partial_{\delta\mu} I_i\right]\Big|_{\delta\mu=0,\hat{\mu}} \hat{\delta\mu} = \sum_{i=1}^{N} \left[\partial_{\delta\mu} I_i\right]^{\mathrm{T}}\Big|_{\delta\mu=0,\hat{\mu}} \left[I_i|_{\delta\mu=0,\hat{\mu}} - T_i|_{\mu^0} \right] \tag{20}$$

for the pose variation $\hat{\delta\mu}$ at the initial estimate $\hat{\mu}$ and $\delta\mu = 0$. The estimate $\hat{\mu}^\star$ of global motion is updated according to

$$\hat{\mu}^\star = \hat{\mu} \circ \hat{\delta\mu} \quad \text{with} \quad m(\mathbf{x}, \hat{\mu}^\star) = m\Big(m\Big(\mathbf{x}, \hat{\delta\mu}\Big), \hat{\mu}\Big) \quad . \tag{21}$$

In practice, (20) is expensive since image derivatives have to be computed for every iteration. By taking advantage of a relaxed image-constancy assumption in 3-D which reads

$$I(p(m(m(\mathbf{x}, \delta\mu), \hat{\mu}))) = T\big(p\big(m\big(\mathbf{x}, \mu^0\big)\big)\big) \tag{22}$$

for $\mathbf{x} \in X$, the texture Jacobian in (20) simplifies at $\delta\mu = 0$ to

$$\partial_{\delta\mu} I_i\Big|_{\delta\mu=0,\hat{\mu}} = \partial_{\mathbf{x}} T_i \cdot \partial_{\mu} m\Big|_{\mu=0} \quad . \tag{23}$$

Hence, the image Jacobian and Hessian are constant in this framework, which is a valid approximation for reasonable variations from the reference pose (cf. [12]). IC-R tracking is performed as long as the error $O(\hat{\mu}^\star)$ is below θ_{ICR}.

V. Evaluation

We evaluate the hierarchical detection and tracking approach on two objects: a 1.5l soda bottle with a radius of $4.615\mathrm{cm}$ and a textured box of size $17.7\mathrm{cm} \times 12.5\mathrm{cm} \times 6.5\mathrm{cm}$ (see Fig. 8). The digital Fire-Wire camera used in the setup has a resolution of 780×580 pixels and a lens aperture of $55^\circ \times 40^\circ$. Two distinct kinds of model information are build for each object, that is a colour histogram and a textured 3-D point cloud.

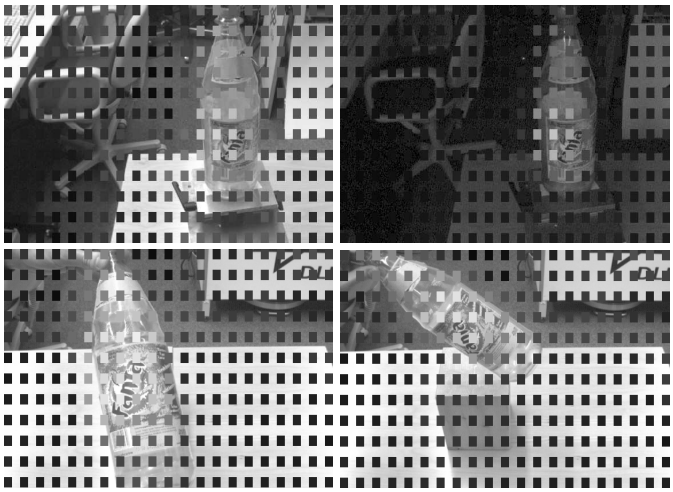

Fig. 3. Subsampled detection of the bottle by means of the Bhattacharya coefficients for colour histograms. Top: Detection under different scene lightness. Bottom: Detection for different poses.

The colour histograms are computed for a selected view on each object by manually selecting an appropriate image region. We choose a $8 \times 8 \times 2$ discretisation of the hue-saturation-value (HSV) colour space to build the histogram. Thus the function h of Sect. III maps to $m = 128$ colour bins.

The 3-D point clouds for the shape-texture based tracking approaches are generated by sampling a 83.17° segment of a cylindrical body and a planar $14.7\mathrm{cm} \times 10.5\mathrm{cm}$ patch of the box at two resolutions. Low resolution models with 100 points for the bottle and respectively 192 for the box are used for particle filtering with 50 samples while high resolution models with 4624 and respectively 4368 points are generated for IC-R tracking. The corresponding reference textures are aquired prior to tracking and manually registered to the geometric models.

In the following, constant computation time is assessed to each tracking method where not stated otherwise. In terms of framrate, Mean-Shift tracking is performed at $8\mathrm{Hz}$, particle filtering at $8\mathrm{Hz}$ and IC-R tracking at $25\mathrm{Hz}$ which corresponds to $1-12$, 75 and 25 iterations respectively for each frame.

The first evaluation concerns the capability of the first stage in detecting an a-priori known object under variations of lightness as well as the pose of the object (see Fig. 3). Detection by means of a colour histogram reveals to be robust with respect to these variations.

Next, we evaluate the tracking performance of the three tracking approaches of Sect. III and IV with Monte-Carlo methods. For the comparison, we off-line registered the camera to an external optical tracking system based on infrared retro-reflecting markers with a precision of $1.5\mathrm{mm}$. Herafter, the externally measured pose of the object is assumed to be the ground-truth. Minimisation is analysed with distinct combinations of initial rotational and translational offsets in 50 images with 30 trials per combination. The direction as well as the rotational axis are randomly chosen.

The average distance to ground-truth after minimisation

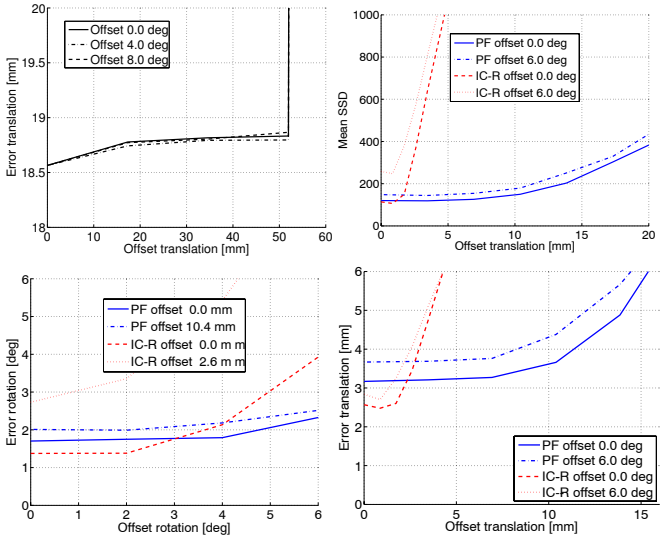

Fig. 4. Accuracy of minimization with erroneous initial estimates. Upper left: Translational error for the extended Mean-Shift tracking. Upper right: Mean squared error for particle filtering (PF) opposed to the IC-R tracking. Lower: Translational and rotational error for PF opposed to IC-R tracking.

is plotted in Fig. 4. The graphs show a broad range of convergence for the extended Mean-Shift tracking but also a relative high error in translation. The particle filter outbeats the IC-R algorithm in range of convergence but not in terms of accuracy. It can be deduced that gain in accuracy is affected by a substantially decreased range of convergence.

The probability of convergence and the speed of convergence of the methods are confronted to each other in Fig. 5 and 6. IC-R tracking shows a narrow range of convergence with respect to translation while it performs more robustly for additional rotational error compared to tracking with the particle filter. The dependence on the number of iterations until convergence from the translational and the rotational offset is shown in Fig. 6. Obviously narrow convergence is faster for the IC-R tracking compared to particle filtering.

Finally, the individual stages are combined to a single visual servoing application on top of the service-robot of picture

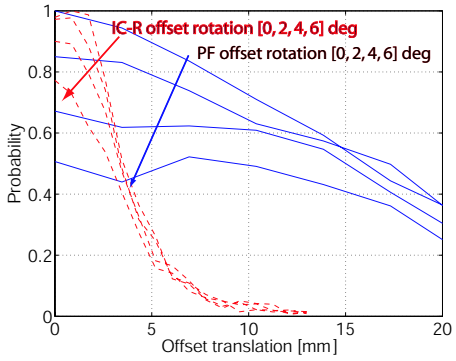

Fig. 5. Probability of convergence with respect to a controlled translational and rotational offset for tracking with the particle filter opposed to the IC-R approach after 75 iterations each.

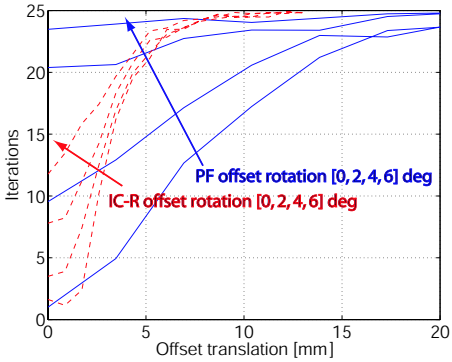

Fig. 6. Speed of convergence with respect to a controlled translational and rotational offset for tracking with the particle filter opposed to the IC-R approach with maximal 25 iterations.

1. In this scenario, a human carries one of the two objects towards the robot. The a-priori known object is detected by the vision system as soon as it becomes visible. Tracking starts with histogram tracking and as soon as control is passed to the shape-texture based methods target tool-center-point (tcp) positions are send to the robotic arm. When the tcp is at a specific distance then the object is catched by the robotic hand with a pre-defined grasp.

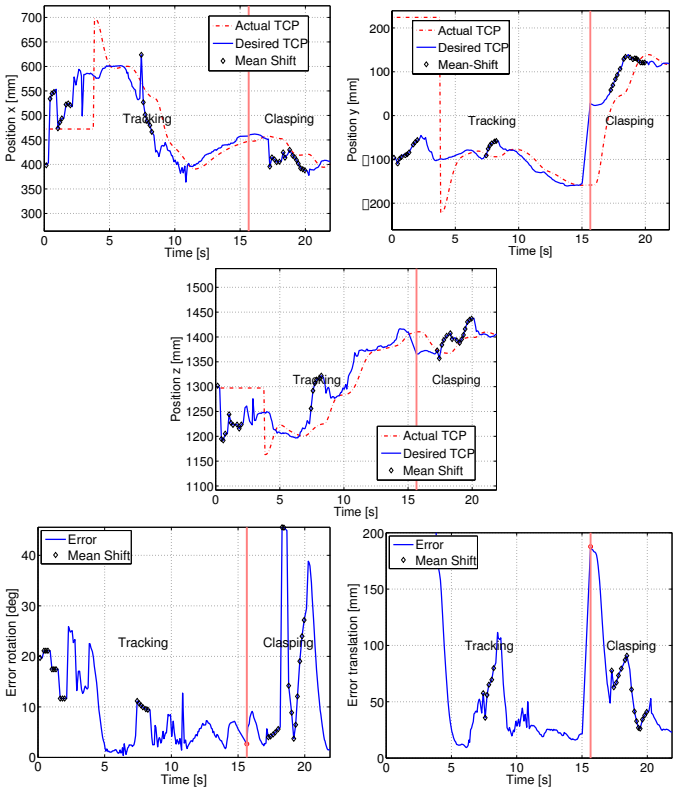

Fig. 7. Trajectories of an exemplified visual servoing session. Upper: Desired and actual absolut position in x,y and z. Bottom: Rotational and translational error between the desired and actual tool-center-point pose.

The pictures of Fig. 8 show tracking of the box and successful grasping of the bottle. The trajectories for a single

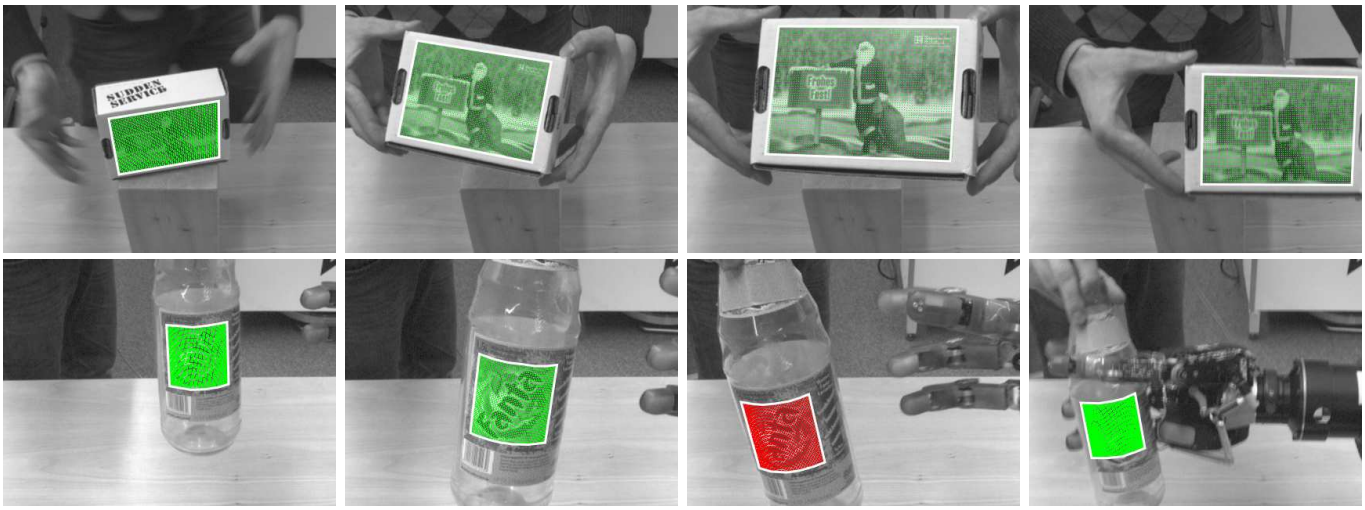

Fig. 8. Screenshots. Top: Successfully tracked box. Bottom: Successful visual-servoing and grasping of a bottle. The tracked point cloud is outlined.

session are documented in Fig. 7, showing the translation parameters of the tracked object as well as the distance and rotational error in position-based servoing by the robot.

VI. Summary and Conclusion

Visual servoing in 6-DoF strongly depends on robustness and accuracy of object tracking. Previous work addressed this problem generally by using artificial landmarks. However, as soon as unaltered real-world objects are used and motion is not constrained to less than 6-DoF the task becomes more difficult.

We address the problem employing a hierarchy of tracking algorithms. In contrast to feature-based algorithms we here track on a per pixel basis. That is, pixel colour histogram values drive localization at the beginning while texture values steer tracking at the final levels.

We assess the probability and range of convergence for each of the tracking methods on real-data and show successfully the integration into position-based visual servoing for grasping moving objects. We attest satisfactory tracking speed and robust behaviour of object-pose re-initialization.

There is still space for increasing the speed of tracking. Currently, the trajectory of the moving object is not extrapolated which would give a better estimate for the pose at the following time instant. Moreover, the number of 3-d model points involved at the highest level can be reduced enabling tracking at a framerate of 50Hz.

Acknowledgment

The work was partly supported by the EU's sixth framework programme (FP6), IP no. 011838-2 *SMErobot*.

References

[1] P. K. Allen, A. Timcenko, B. Yoshimi, and P. Michelman, "Automated tracking and grasping of a moving object with a robotic hand-eye system," *IEEE Trans. on Robotics and Automation*, vol. 9, no. 2, pp. 152–165, 1993.

[2] S. Jörg, J. Langwald, J. Stelter, C. Natale, and G. Hirzinger, "Flexible robot-assembly using a multi-sensory approach," in *IEEE Int. Conf. on Robotics and Automation*, vol. 28, no. 3, July 2000, pp. 3687–3694.

[3] M. Mehrandezh, N. M. Sela, R. G. Fenton, and B. Benhabib, "Robotic interception of moving objects using an augmented ideal proportional navigation guidance technique," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 30, no. 3, pp. 152–165, May 2000.

[4] S. B. Ezio Malis, Franois Chaumette, "2 1/2 d visual servoing," *IEEE Trans. on Robotics and Automation*, vol. 15, no. 2, pp. 234–246, April 1999.

[5] T. Drummond and R. Cipolla, "Real-time visual tracking of complex structures," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 932–946, 2002.

[6] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of the Int. Joint Conf. on Artificial Intelligence*, 1981, pp. 674–679.

[7] N. Diehl and H. Burkhardt, "Planar motion estimation with a fast converging algorithm," in *Proc. of the 8th Int. Conf. on Pattern Recognition*, 1986, pp. 1099–1102.

[8] J. M. Buenaposada and L. Baumela, "Real-time tracking and estimation of plane pose," in *Proc. of the 16th Int. Conf. on Pattern Recognition*, vol. II, August 2002, pp. 697–700.

[9] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *Int. Journal of Computer Vision*, vol. 56, no. 3, pp. 221 – 255, March 2004.

[10] B. Cernuschi-Frias, D. B. Cooper, Y.-P. Hung, and P. N. Belhumer, "Toward a model-based bayesian theory for estimating and recognizing parameterized 3-d objects using two or more images taken from different positions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, no. 10, pp. 1028–1052, 1989.

[11] N. A. Ramey, J. J. Corso, W. W. Lau, D. Burschka, and G. D. Hager, "Real time 3d surface tracking and its applications," in *Proc. of Workshop on Real-time 3D Sensors and Their Use (at CVPR)*, 2004.

[12] W. Sepp, "A direct method for real-time tracking in 3-d under variable illumination," in *Proc. of the 27th Pattern Recognition Symp., DAGM'05*, Aug.-Sept. 2005.

[13] K. Toyama and G. D. Hager, "Incremental focus of attention for robust visual tracking," *Int. Journal of Computer Vision*, vol. 35, no. 1, pp. 45–63, 1999.

[14] V. R. Dorin Comaniciu and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2000, pp. 142–149.

[15] R. Collins, "Mean-shift blob tracking through scale space," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. 2, June 2003, pp. 234–240.

[16] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *Int. Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.