Visual Tracking of an End-Effector by Adaptive Kinematic Prediction

Andreas Ruf, Martin Tonko, Radu Horaud, Hans-Hellmut Nagel

HAL Id: inria-00590080
https://hal.inria.fr/inria-00590080

Submitted on 3 May 2011

Visual Tracking of An End–Effector by Adaptive Kinematic Prediction

A. Ruf, M. Tonko[†], R. Horaud, H.-H. Nagel[†‡]

GRAVIR–IMAG, INRIA Rhône–Alpes, 38330 Montbonnot St. Martin, FRANCE
[†] IAKS, Universität Karlsruhe (TH), 76128 Karlsruhe, GERMANY
[‡] Fraunhofer–Institut für Informations– und Datenverarbeitung, 76131 Karlsruhe, GERMANY

Abstract

This paper presents results of a model–based approach to visual tracking and pose estimation for a moving polyhedral tool in position–based visual servoing. This enables the control of a robot in look-and-move mode to achieve six degree of freedom goal configurations. Robust solutions of the correspondence problem – known as "matching" in the static case and "tracking" in the dynamic one – are crucial to the feasibility of such an approach in real–world environments. The object's motion along an arbitrary trajectory in space is tracked using visual pose estimates through consecutive images. Subsequent positions are predicted from robot joint angle measurements. To deal with inaccurate models and to relax calibration requirements, adaptive on–line calibration of the kinematic chain is proposed. The kinematic predictions enable unambiguous feature matching by a pessimistic algorithm. The performance of the suggested algorithms and the robustness of the proposed system are evaluated on real image sequences of a moving gripper. The results fulfill the requirements of visual–servoing, and the computational demands are sufficiently low to allow for real–time implementation.

1 Introduction

This paper considers the problem of tracking general motions of a robot end–effector in space, using image sequences taken with a mobile camera mounted on a second robot. As the camera tracks the tool's motions, the execution of complex trajectories by position–based control becomes possible, even in real–world environments with limited visibility and restricted accessibility such as engine compartements.

One difficulty is the non–linearity of both the sensor and the actuator, i.e. perspective projection and six–axis robot arms. For simplicity, many existing approaches use either approximate sensor models e.g. [8] or reduce the complexity of task–space [9], [1].

Secondly, by allowing off–the–shelf robot tools that are not ideal convex polyhedra, both the combinatoric complexity of matching and the uncertainty in feature measurements increase significantly. Allowing a six degrees of freedom task-space further raises the search complexity in prediction and optimization algorithms.

To achieve real–time execution with limited computational power, trade–offs are required between task complexity, simplicity of approache, or the performance, and overall capability of the system.

In previous systems, tracking focused on artificial features on a supplementary ring [17] or plate [7] rigidly attached to the end–effector. This allowed the use of simplified algorithms, but the observable or accessible reach of the end–effector was significantly restricted by occlusion of features as well as by the encumbrance and the obstruction caused by the added tops. Moreover, an additional calibration demand, ring–to–hand, was introduced.

Moreover, feature or pose prediction in many current approaches is based on trajectory extrapolation [15]. The required model of object motion usually imposes lower order polynomial constraints on the trajectories. As soon as non–smooth motions are to be tracked, e.g. zigzag or stop–and–go, either short cycle times or slow robot motions need to be assumed. In addition, robustness with respect to stoppage in the image stream or in the computational process is rather low.

Below we demonstrate that adaptive kinematic prediction allows robust and accurate tracking of end–effector motion, even for complex non–smooth trajectories. The computational demands are also within the capabilities of current real-time systems.

The text is divided into two sections, one on the theoretical approach and the other one on the experimental evaluation. After a brief introduction, the four major stages of the approach are further detailed: prediction, matching, pose estimation, and calibration update. The evaluation section describes the experimental environment and reports on the error in image and 3D–space, as well as giving the probability distributions for the adapted parameters.

2 Approach

The tracking loop consists of four stages:
1) prediction, 2) feature matching, 3) pose estimation, and 4) adaptive calibration.

In stage 1, the spatial position of the object relative to the camera is predicted using the online-adapted kinematic chain. In consequence, the image location of model features can be predicted by backprojection.

Stage 2 establishes 2D–2D correspondences between predicted and extracted line segments with a pessimistic matching algorithm. 2D–3D line matches are conjectured from the known object model. This concludes feature tracking and yields the input to the pose algorithm.

During stage 3, the object pose is estimated from corresponding lines (not line segments) by the Levenberg–Marquardt [11] minimization of a quadratic error–function. The error measures the algebraic distance of image points from the corresponding image line, but also has a spatial geometric interpretation. The necessary initialization is provided by the kinematic prediction. The estimated sequence of object poses gives the tracked spatial trajectory.

During stage 4, the difference between the object pose predicted using the nominal kinematic chain and the pose estimated from the image is used to adapt the calibration of the kinematic chain.
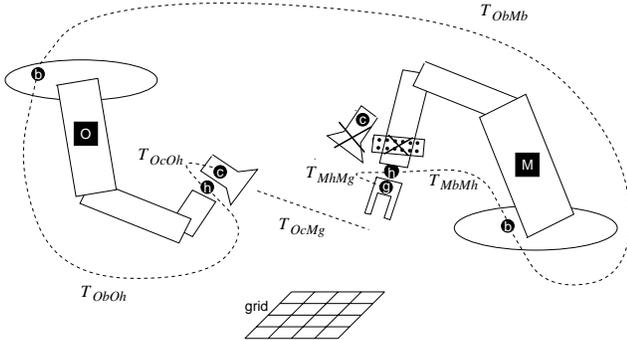

Figure 1: Robotic cell at the FhG–IITB, Karlsruhe–Germany.

2.1 Prediction

The prediction of the gripper position with respect to the camera frame is determined by the product of the homogeneous transforms[1,2] $\mathcal{T}$ that constitute the kinematic chain (see Fig. 1).

$$\mathcal{T}_{OcMg}^{(t_i)} = \mathcal{T}_{OcOh}\mathcal{T}_{OhOb}^{(t_i)}\mathcal{T}_{ObMb}\mathcal{T}_{MbMh}^{(t_i)}\ \mathcal{T}_{MhMg}^{(t_{i-k})}. \quad (1)$$

---

[1]The source and target frames of the transforms are given in right to left order by subscripts. Lowercase letters b,h,c,g denote, respectively, base, hand, camera, and gripper frame of either the observing robot O or the manipulating robot M, e.g. Mg is the frame of the gripper carried by the manipulating robot.

The static transforms[2] are initially determined with powerful calibration techniques. Firstly, $\mathcal{T}_{OcOh}$ and $\mathcal{T}_{McMh}$ result from hand–eye calibration with a variation of [14] from multiple views of a calibration grid common to both robots [5]. Secondly, the positions relative to the grid allow the base–to–base calibration $\mathcal{T}_{ObMb}$ to be calculated. Finally, $\mathcal{T}_{MhMg}^{(t_0)}$ is known for a fixed starting configuration of the robots.

With the factory–given model of robot forward kinematics, the dynamically varying transforms $\mathcal{T}_{OhOb}^{(t_i)}$ and $\mathcal{T}_{MbMh}^{(t_i)}$ follow from the joint–angle measurements at time instant t_i.
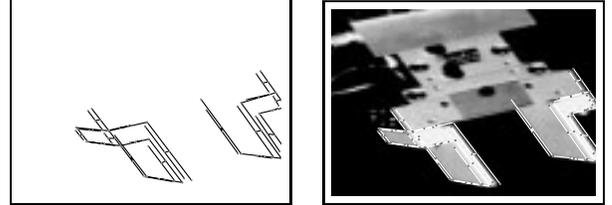

Figure 2: Image of the gripper with superimposed line segments. Only the successfully matched segments are displayed by an interconnection of their mid-points.

2.2 Matching

The term matching is understood as the one–to–one association of extracted image features with the visible features of the object model. Straight edges of the polyhedral model turn out to be the most discriminating features, since they have in comparison with points two additional dimensions, i.e. orientation and length.

Line–segments are represented by their midpoint (mx, my), length l and orientation θ. Noise in the feature extraction process is modeled by the variances of the endpoint coordinates in an aligned frame, i.e with axes parallel and perpendicular to the segment. This leads [3] to a block–diagonal covariance–matrix for the parameter vector $(mx, my, l, \theta)^T$.

This allows for the computation of the Mahalanobis–distance d_{Mah} between the parameter vectors $\boldsymbol{p}$ and $\boldsymbol{e}$ of predicted and extracted segments, i.e. a distance with weights respecting the given uncertainties.

To increase robustness, particularly of pose estimation, feature pairs are associated pessimistically, i.e. possibly ambiguous pairs are considered as mismatches and preference is given to a smaller but unambiguous set of matches. Candidate pairs are ascertained by a symmetric nearest–neighbour search within both feature sets, with a threshold β on the distance.

More formally, let $\mathbf{E}$ be the set of extracted segments $\boldsymbol{e}$ and $\mathbf{P}$ the set of predicted segments $\boldsymbol{p}$. A pair $(\boldsymbol{p}_j, \boldsymbol{e}_i)$

---

[2]Dynamically varying transforms bear time superscripts (t_i), static transforms have no superscription. k is the length of a discrete time-delay.

will be matched if and only if:

$$
\begin{aligned}
d_{Mah}(\boldsymbol{p}_j, \boldsymbol{e}_i) &< \beta, \\
d_{Mah}(\boldsymbol{p}_j, \boldsymbol{e}_i) &= \min_{\boldsymbol{e} \in \boldsymbol{E}} d(\boldsymbol{e}, \boldsymbol{p}_j), \quad (2)\\
d_{Mah}(\boldsymbol{p}_j, \boldsymbol{e}_i) &= \min_{\boldsymbol{p} \in \boldsymbol{P}} d(\boldsymbol{e}_i, \boldsymbol{p}).
\end{aligned}
$$

Thresholding corresponds to a χ^2–test of the hypothesis that the calculated distance is within uncertainty bounds given the assumed distributions and covariances.
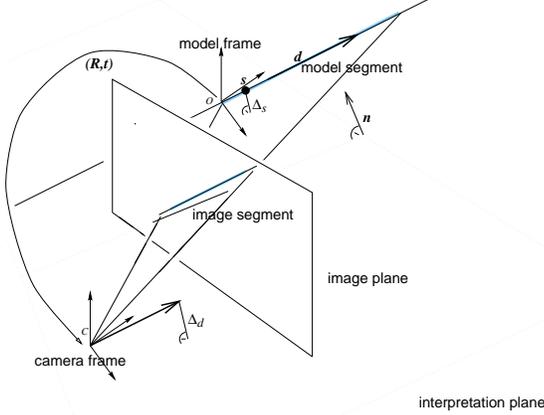

Figure 3: Orthogonal distance between a point in space and the interpretation plane.

2.3 Pose Estimation

The error function is defined as the algebraic distance from the projection p of an arbitrary point P on an object edge to the extracted image line n [13], [4]. More precisely, let $\boldsymbol{n}$ and $\boldsymbol{p}$ be the coordinate vectors of n and p in the projective plane $I\!\!P^2$. They are incident if and only if $\boldsymbol{n} \cdot \boldsymbol{p} = 0$. Furthermore, let O be the object frame, C the camera frame – Mg and Oc in our case (Fig. 1) – and $(R, \boldsymbol{t})$ rotation and translation of the transform from O to C. An arbitrary point P on a model edge with start point $\boldsymbol{s}$ and direction vector $\boldsymbol{d}$ has the form $\boldsymbol{s} + \lambda\boldsymbol{d}$ in O, where λ is a free parameter. Transforming it to C yields $\boldsymbol{p}_\lambda = (x, y, z)^T = \lambda R\boldsymbol{d} + R\boldsymbol{s} + \boldsymbol{t}$. Its projection p with vector $(x/z, y/z, 1)^T$ is on n if and only if $\boldsymbol{n} \cdot \boldsymbol{p_\lambda} = 0$. Specializing to $\lambda \to \infty$ and $\lambda = 0$ gives

$$
\Delta_d \equiv \boldsymbol{n} \cdot (R\boldsymbol{d}) = 0 \quad (3)
$$

$$
\Delta_s \equiv \boldsymbol{n} \cdot (R\boldsymbol{s} + \boldsymbol{t}) = 0 \quad (4)
$$

At first glance, Δ_s and Δ_d are the algebraic distances of the respective points in $I\!\!P^2$ to n. After normalization, $\boldsymbol{n}$ corresponds to the unit normal in C to the

interpretation plane corresponding to n [10]. As it passes through the origin of C, $\boldsymbol{n}$ corresponds to the Hessian form of this plane and equally Δ_s and Δ_d (3), (4) turn out to be the geometric distances from this plane to the points in C with vectors $\boldsymbol{p_0}$ and $\boldsymbol{p_\infty}$. They represent points on the model edge, namely its start–point and the point of direction vector after translation to the origin of C (see Fig. 3). Let $\Delta_{\boldsymbol{s,n}}^{(i)}$ and $\Delta_{\boldsymbol{d,n}}^{(i)}$ now be the functions of R and $\boldsymbol{t}$ corresponding to the distance or error of the match with index i.

Pose-estimation is done by non–linear minimization of the quadratic error–function $\Phi(R, \boldsymbol{t})$

$$
\Phi(R, \boldsymbol{t}) = \sum_i \left[\left(\Delta_{\boldsymbol{s,n}}^{(i)}\right)^2 + \left(\Delta_{\boldsymbol{d,n}}^{(i)}\right)^2 \right] \quad (5)
$$

using the Levenberg–Marquardt algorithm. Φ varies with the coefficients of matrix $R^{3\times3}$ and vector $\boldsymbol{t}^{3\times1}$, representing orientation and location of the object in the camera frame. The nine coefficients of the rotation matrix R are parametrized by a unit quaternion q, $R = R(q)$. The unity of q adds another quadratic constraint, so the minimization becomes:

$$
\min_{q,t} \left[\mu(1 - \|q\|^2) + \Phi(R(q), \boldsymbol{t}) \right]. \quad (6)
$$

Note that we minimize both over rotation and translation, which further increases the accuracy.

Finally $\mathcal{T}_{OcMg}^{(t_i)}$, the homogeneous transform from gripper frame to camera frame, is given by $R(q)$ and $\boldsymbol{t}$ with minimal error.

2.4 Adaption

Inaccuracies in the static calibrations and especially divergence from the modeled kinematics result in deviations of the predicted position $\mathcal{T}_{OcMh}$ of manipulating robot's hand–frame from the real position[3] $\mathcal{T}_{OcMh!}$. As these errors vary systematically during robot motion, an on-line adaption of the kinematic chain, i.e. of its terminal link, is advisible.

$$
\mathcal{T}_{MhMg}^{(t_i)} = \mathcal{T}_{MhMb}^{(t_i)} \mathcal{T}_{MbOb} \mathcal{T}_{ObOh}^{(t_i)} \mathcal{T}_{OhOc} \mathcal{T}_{OcMg}^{(t_i)}. \quad (7)
$$

In contrast, $\mathcal{T}_{MhMg}^{(t_i)}$ is assumed to be locally constant, since for subsequent time instants only minor changes in joint–angles occur. Given this, a time–delay k in the prediction (1),(8) can be tolerated.

$$
\mathcal{T}_{OcMg}^{(t_{i+k})} = \mathcal{T}_{OcMh}^{(t_{i+k})} \cdot \mathcal{T}_{MhMg}^{(t_i)}. \quad (8)
$$

For a fixed cycle time this means; the longer the feasible time–delay k, the higher the velocities that still can

---

[3]The frame which is rigidly attached to the terminal joint of the robot arm and which once coincides with the tool–center–point frame, e.g. in the reference position of the robot, is understood as "real" hand frame $Mh!$.

be tracked. What is more, values of k about 5–10 allow major stoppages to be tolerated, even in the case of a zigzag–movement during failure. This distinguishes adaptive kinematic tracking from extrapolation-based approaches.

Additionally, the adaption $\mathcal{T}_{MhMg}^{(t_i)}$ allows the position error in the current tool–center–point frame Mh to be computed from the error in the gripper frame, e.g. as observed by the camera. This is a significant advantage as the input to the joint–level robot controller is supposed to be given in the Mh frame.

Figure 4: Experimental platform for car dismantling at FhG–IITB, Karlsruhe (Germany).

3 Experimental Evaluation

The experiments were conducted at the platform for automatic car dismantling at FhG–IITB, Karlsruhe (see Fig. 4, Fig. 1). It consists of two 6–DOF industrial robots. The Reis robot in hanging position on the left hand side, serves as observing robot O, carrying a stereo-head. The Cloos robot on the right hand side in upright position serves as manipulating robot M, carrying the gripper. The camera mounted on M is required only for base-to-base calibration at installation time. As a further result of tracking the motion by directly tracking the tool, the system could be liberated from the restricting ring (see Fig. 1).

We will give results for an image sequence which is taken from about 1.2 m distance. It shows the gripper translating along the three perpendicular edges of a 20 cm cube and returning along the diagonal. Twenty images were taken at equally spaced stops on each of the four straight movements. Note that for a pessimistic cycle time of 100 ms and a time-delay of $k = 5$, tracking motions at velocities of 0.5 m/s should be feasible.

3.1 Error in image

The image error is calculated as the average orthogonal distance in *pixel* of the endpoints of a backprojected segment to the image line containing an extracted segment. The total error is then the average distance be-
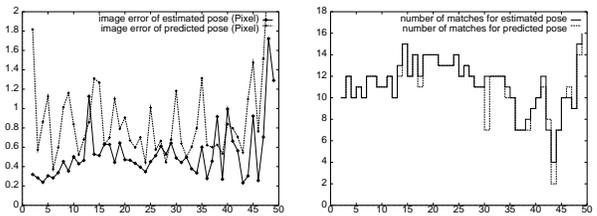

Figure 5: Average image error (left) and number of matches (right) for predicted and estimated pose compared on the tracked sequence.

tween paired segments. The image errors during the sequence, both before and after pose estimation, are depicted in Fig. 5. Even before pose estimation, the error of predicted features is on average by 1 *pixel*. This underlines the high accuracy obtained by the adaptive kinematic predictions. However, the non-linear pose estimation further halves the error to about 0.5 *pixel*. Another indication of near-optimal effectiveness of the prediction is the fact that rerunning the matching algorithm after pose estimation only rarely increases the number of matched pairs (see Fig. 5).

3.2 Error in space

To obtain ground-truth, a mathematical model of the gripper trajectory is defined by the Cartesian output of the joint-level controller during the execution of the motion, e.g. a cube in case of the investigated sequence. The output of visual tracking - the sequence of pose estimates $\mathcal{T}_{OcMg}^{(t_i)}$ - is fitted to this model by minimizing the mean-square error. This yields an instance of the trajectory model which is henceforth considered as ground-truth for the gripper's position (Fig. 6).

The standard-deviation of the tracked motion from ground-truth is given in terms of the six pose parameters x, y, z, ϕ, θ, ψ. The latter three give the gripper orientation in terms of the angles for roll, pitch, yaw (Table 1).

	$x_{[mm]}$	$y_{[mm]}$	$z_{[mm]}$	$\phi_{[°]}$	$\theta_{[°]}$	$\psi_{[°]}$
σ	0.43	0.89	5.88	0.58	0.65	1.40
σ/μ	1.3%	2.1%	0.53 %	$--$	$--$	$--$

Table 1: Deviation of tracked gripper pose from ground truth and relative error in pose parameters.

The much higher standard deviation in z–direction, parallel to the optical axis is a characteristic of monocular pose estimation. The value of $\sigma_y = 0.89$ mm, which is roughly two times $\sigma_x = 0.43$ mm, underlines the fact that half-frame video images – each second scan line only – are processed. This is also observed when regarding the relative errors, whereas now the error in z is less grave.
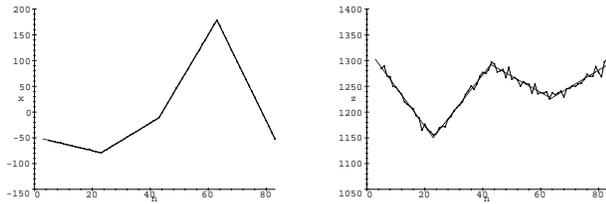
Figure 6: Graph of x- and z-coordinates of the gripper translation during tracking.

Figure 7: Sample density of the parameters ϕ, y (points) and the density function of the fitted normal distribution (curve).

3.3 Distribution of Parameters

A disadvantage of the simple way the kinematic chain is adapted in equation (7) is that noise in pose estimates immediately affects the adapted transform $\mathcal{T}_{MhMg}$, called system state S from now on. Indeed, as soon as the noise in the state parameters exceeds a certain level, tracking becomes unstable and might even break-down.

The distribution of the state estimates obtained during tracking is being investigated in order to identify an appropriate noise model. This should also allow for optimal estimation of the system state by using filtering techniques [2].

If errors in the kinematic models and pose estimates are neglected, $\mathcal{T}_{MhMg}$ or S corresponds to a fixed and static transform from the gripper frame to a frame rigidly attached to the terminal link of the robot arm, the "real" hand frame. Hence, the distribution of the state parameters is assumed to be normal with sample mean and sample variance. This hypothesis is verified by a χ^2-test (see Table 2, Fig. 7). Even though the sample size is comparably small, the hypothesis can not be rejected at $0.09 < \alpha < 0.23$.

	x	y	z	ϕ	θ	ψ
χ^2	6.2	8.1	7.4	8.18	11.09	10.54
α	0.18	0.09	0.12	0.23	0.09	0.10
m	4	4	4	6	6	6

Table 2: Results of a χ^2-test on the hypothesis of normal distribution of the measured state parameters. α gives the significance level with respect to m, the degrees of freedom of the employed partitioning.

4 Conclusions

To sum up, the visual tracking of a non-ideal polyhedral manipulator is feasible by imposing merely robotic kinematic constraints on the object position. Standard hand-eye calibration and factory-given kinematic models are sufficiently accurate to locally predict end-effector positions such that pessimistic matching becomes unambiguous. The resulting pose estimates en-
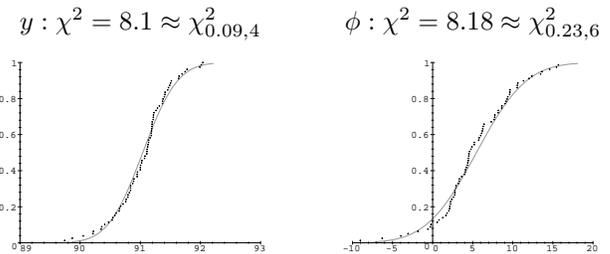
able the on-line corrective adaption of the kinematic chain, that subsequently allows for robust and global tracking of the end-effector motion. Robustness to major delays in the tracking process and the ability to cope with non-smooth trajectories are significant advantages of this approach.

In practice, visual tracking by adaptive kinematic prediction makes a step towards position-based control of a kinematic chain with a large number of degrees of freedom, 12 in our case. Especially in the independent-eye configuration, this allows complex tasks to be performed in environments with clutter and limited accessability.

The computational demands introduced by the stages specific to this approach - prediction and adaption - are negligible compared to those for approaches based solely on filtering. With specialized hardware up to the level of feature processing and recently available linear iterative methods for pose estimation [6], an implementation running at frame-rate is feasible. However, synchronisation of joint-angle measurements with image acquistion still remains a crucial issue.

The evaluation in a real-world environment shows that the approach is both robust and accurate. In particular the results on the distribution of state parameters point out remaining potentials, which can be exploited to further increase robustness and accuracy by filtering the adapted parameters [16].

References

[1] P. K. Allen, A. Timcenko, B. Yoshimi, P. Michelman: *Automated Tracking and Grasping of a Moving Object with a Robotic Hand–Eye System.* IEEE Transactions on Robotics and Automation **RA-9**:2 (1993) 152–165.

[2] Y. Bar-Shalom, T.E. Fortmann, *Tracking and Data-Association*, Academic Press, Orlando/FL, 1988.

[3] R. Deriche, O. Faugeras: *Tracking Line Segments.* Proc. First European Conference on Computer Vi-

sion, Antibes, France, 23–26 April 1990, Lecture Notes in Computer Science **427**, Springer-Verlag, Berlin, Heidelberg, New York/NY u.a., 1990, pp. 259–268.

[4] F. Dornaika: *Contributions à l'intégration vision/robotique: Calibrage, localisation et asservissement.* Thèse, L'Institut National Polytechnique de Grenoble, September 1995.

[5] V. Gengenbach: *Einsatz von Rückkopplungen in der Bildauswertung bei einem Hand-Auge-System zur automatischen Demontage.* Dissertation, Fakultät für Informatik der Universität Karlsruhe (TH), Juli 1994. Erschienen in der Reihe *Dissertationen zur Künstlichen Intelligenz,* Band DISKI 72, infix Verlag, Sankt Augustin, 1994.

[6] R. Horaud, F. Dornaika, B. Lamiroy, and S. Christy: *Object Pose: The Link between Weak Perspective, Para Perspective, and Full Perspective.* International Journal of Computer Vision, **22**:2 (1997), to appear.

[7] R. Horaud, F. Dornaika, B. Espiau: *Visually Guided Object Grasping.* In IEEE Transactions on Robotics and Automation, (1997) to appear.

[8] G. M. T. Cross, R. Cipolla: *Affine Visual Servoing.* In Proceedings of the 7th British Machine Vision Conference, Edinburgh, 9 – 12 September 1996, BMVA, pp. 425–434.

[9] U. Uenohara, T. Kanade: *Geometric Invariants for Verification in 3-D Object Tracking.* Proceedings of the 1996 IEEE/RSJ International Conference on Intelligent Robots and Systems. November 4–8, 1996, Osaka, Japan, pp. 785–789.

[10] K. Kanatani: *Geometric Computation for Machine Vision.* Clarendon Press, Oxford, 1993.

[11] J.J. Moré: *The Levenberg-Marquardt Algorithm: Implementation and Theory.* In G.E. Watson (Hrsg.), *Lecture Notes in Mathematics 630,* Springer-Verlag, Berlin, Heidelberg, New York/NY, Tokyo, 1979, pp. 105–116.

[12] N. P. Papanikolopoulos, P. K. Khosla, T. Kanade: *Visual Tracking of a Moving Target by a Camera Mounted on a Robot: A Combination of Control and Vision.* IEEE Transactions on Robotics and Automation **RA-9**:1 (1993) 14–35.

[13] T. Phong, R. Horaud, A. Yassine, P. Tao: *Object Pose from 2-D to 3-D Point and Line-Correspondences.* International Journal of Computer Vision **15** (1995) 225–243.

[14] R. Y. Tsai, R.K. Lenz: *A New Technique for Fully Autonomous and Efficient 3D Robotics Hand/Eye Calibration.* IEEE Journal of Robotics and Automation **RA-5**:3 (1989) 345–358.

[15] M. Tonko, K. Schäfer, V. Gengenbach, H.–H. Nagel: *Multi-Level 3D-Tracking of Objects Integrating Velocity Estimation based on Optical Flow and Kalman-Filtering.* In O. Khatib, K. Salisbury (Eds.): Proc. Fourth Int. Symp. on Experimental Robotics, Stanford/CA, 30 June–2 July, 1995, Lecture Notes in Control and Information Sciences **223**, Springer-Verlag, Berlin, Heidelberg, New York/NY, 1997, pp. 212–221.

[16] M. Tonko, K. Schäfer, F. Heimes, H.-H. Nagel: *Towards Visually Servoed Manipulation of Car Engine Parts.* In R. W. Harrigan (Ed.): Proc. IEEE Int. Conf. on Robotics and Automation, Albuquerque/NM, 20–25 April 1995, IEEE Computer Society Press, Los Alamitos/CA, 1997, pp. 3166-3171.

[17] K. Schaefer, M. Tonko, V. Gengenbach, H.-H. Nagel, R. Horaud, R. Mohr: *Real-Time Pose Estimation of Workpiece and Actuator Using an Independently Mobile Camera System.* In IAR Jahresbericht 1995, Dt.-frz. Institut für Automation und Robotik an der Universität Karlsruhe (Ed.), pp. 259-263.