Combined Region- and Motion-based 3D Tracking of Rigid and Articulated Objects

Thomas Brox, Bodo Rosenhahn, Juergen Gall, and Daniel Cremers

***Abstract*— In this paper, we propose the combined use of complementary concepts for 3D tracking: region fitting on one side, and dense optical flow as well as tracked SIFT features on the other. Both concepts are chosen such that they can compensate for the shortcomings of each other. While tracking by the object region can prevent the accumulation of errors, optical flow and SIFT can handle larger transformations. Whereas segmentation works best in case of homogeneous objects, optical flow computation and SIFT tracking rely on sufficiently structured objects. We show that a sensible combination yields a general tracking system that can be applied in a large variety of scenarios without the need to manually adjust weighting parameters.**

***Index Terms*— Tracking, segmentation, motion.**

I. Introduction

LOCATING objects in 3D space given 2D images has a long tradition in computer vision research [32], [18], [19], [17] with many applications, such as robot navigation, camera calibration, and human motion analysis. Usually, the intrinsic camera parameters and a 3D object model are assumed to be given. The latter can consist of, e.g., a set of points, lines, or patches. The goal is to find the six parameters of a rigid body motion, i.e., the extrinsic camera parameters relative to the object. For the special case of tracking, the pose of the object is assumed to be known in the first frame of an image sequence. One is then interested in capturing the pose in successive frames of the sequence while the camera or the object are moving.

The task can be extended by assuming no longer rigid objects, but object models that allow for some restricted change in their structure. One application, which has become very popular in recent time, is human motion estimation [16], [3], [38], [25]. Here, the model consists of a number of rigid limbs connected by predefined joints. Additionally to the global rigid body motion, one is interested in the joint angles. There are many recent works on human tracking, most of them making use of learning techniques to constrain the space of solutions and to avoid ambiguities [39], [41], [44], [7]. Others interpret tracking as a recognition task [40], [26], [31], which has many advantages compared to classical tracking, but also requires to solve the much more difficult problem of object localization involving a detailed pose.

T. Brox is with the Computer Science Department at U.C. Berkeley, USA. E-mail: brox@eecs.berkeley.edu. B. Rosenhahn is with the Leibniz-University of Hannover, Germany. E-mail: rosenhahn@tnt.uni-hannover.de. J. Gall is with the Max-Planck Center for Computer Science, Saarbrücken, Germany. E-mail: jgall@mpi-inf.mpg.de. D. Cremers is with the University of Bonn, Germany. E-Mail: dcremers@cs.uni-bonn.de.

This project was partially funded by the German Research Foundation (DFG) and the Max-Planck Center for Visual Computing and Communication. The authors thank the anonymous reviewers for comments leading to improvements of the manuscript.

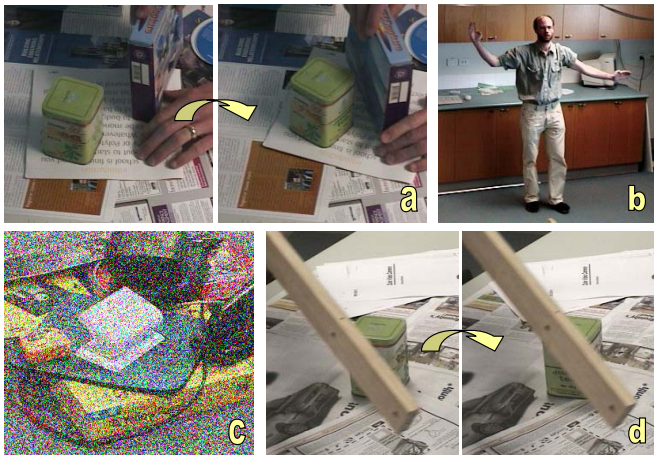


Fig. 1. Sample images from some challenging sequences. In some cases large transformations must be handled (a), in others articulated, textured objects are to be tracked in front of a cluttered background (b). Other objects are homogenous and substantially degraded by noise (c), or they can be partially occluded by another moving object (d). The challenge is to reliably track the object in all these scenarios with a single tracking system, ideally without adapting the parameters.

In this paper, we cover classical tracking of rigid as well as articulated objects focusing on the image-driven part, i.e., we will not cover recognition or learning techniques here. In particular, we deal with the challenge to establish correspondences between image points and model points. Such point correspondences are the fundamental requirement for 3D tracking, and the quality of the correspondences mainly decides on the quality of the estimated pose parameters. Figure 1 shows some images from tracking scenarios highlighting different challenges. While there are numerous specialized methods that can successfully track the object in one or two of the scenes, such methods have their inherent weaknesses that likely make them fail in a complementary scenario. In the present paper, we propose to integrate multiple complementary concepts to establish point correspondences. The ultimate goal of this cue integration is to be able to have a *single* tracking system that can handle *all* scenarios exemplified in Figure 1. The following concepts to establish point correspondences have emerged in the literature:

Edge-based techniques. The classic approach to pose estimation is by means of an edge detector applied to the images. Given a model of the object surface, its silhouette can be matched to the detected edges, seeking to maximize the consistency of both [19]. Though plausible and fast,

the main drawback of this approach are the numerous local minima. They are caused by many spurious edges due to noise, background clutter, or texture on the object itself.

Region-based techniques follow a similar concept as the edge-based approach. Here the overlap error of the projected surface with the object region in the image is sought to be minimized. Unfortunately, extracting the object region from the image is not as easy as edge detection. In principle one is confronted with a segmentation problem. Sometimes background subtraction can be a straightforward solution. More general methods rely on different intensity distributions in the foreground and background region and take the object model as a shape constraint into account [34]. The computational costs are higher than with edge-based approaches. On the other hand segmentation can better deal with low contrast edges and noise. Moreover, texture can be taken into account. Although there are usually fewer local optima than in the edge-based approach, local optima are still a significant problem, as they prohibit tracking in case of large transformations from frame to frame. Another problem are ambiguous solutions. For instance, the pose of a sphere cannot be uniquely determined from its silhouette.

Patch-based techniques. 3D tracking methods very often employ a patch-based tracker that establishes 2D correspondences between successive frames. Knowing the exact pose in the first frame, the 2D points in this frame can be related to 3D points. This effectively yields a set of 2D-3D correspondences. Among the most popular 2D trackers are the KLT tracker [36] and a tracker based on the recently developed SIFT features [20]. Especially the SIFT tracker can deal with small frame rates and fast motion, as it is invariant with respect to scaling, image rotation, and moderate lighting changes. The main drawbacks of patch-based trackers in general are their need for sufficiently textured objects and the accumulation of errors during tracking. The latter is caused by the assumption of knowing the correct pose in the previous frame.

Flow-based techniques. 2D correspondences can also be computed by means of an optical flow method and employed in the same way as correspondences from a patch-based tracker. The success of this approach depends considerably on the chosen optical flow method. Most methods are restricted to small pixel displacements and rely on parametric flow models that might be too restrictive, for instance, in case of human motion estimation. Moreover, optical flow estimation is usually very sensitive even to small brightness changes. These problems are largely avoided by the method in [5], which turns it into an interesting alternative to patch-based trackers. In contrast to those, optical flow provides dense correspondence fields.

Since all these approaches come along with inherent drawbacks, it makes sense to combine complementary concepts. This has been suggested earlier in [12], where optical flow is incorporated as a hard constraint in an edge-based method to face tracking. In this method, the optical flow dominates the tracking. In contrast, the work in [21] uses the optical flow in order to predict the pose parameters in a new frame, which serve as initialization for an edge-based method. The idea in [21] is that a multi-resolution optical flow method captures large displacements of the object and thus helps the edge-based method to hit better local optima. Finally, the authors of [45] propose the combination of a patch-based tracker and an edge-based method. The latter aims at preventing the accumulation of errors of the patch-based tracker. However, they show that the edge-based method tends to degrade results, despite the close initialization by patch-based tracking, since there are still local optima in the vicinity of this initialization. Therefore, the approach in [45] considers multiple hypotheses for edge locations.

These works all propose the use of an edge-based technique in addition to either optical flow or a patch-based tracker for preventing the accumulation of errors. In this paper, we aim at exploiting complementary cues more rigorously in order to investigate the potentials of purely image-driven tracking[1]. In particular, we combine region cues, optical flow, and SIFT features. Whereas region cues are clearly complementary to motion cues, optical flow and SIFT tracking often provide very similar information. However, they are not completely redundant, as we will see in the experimental evaluation.

Besides the selection of the cues to be combined, the main contribution of this paper is their adaptive weighting. Reasonable information fusion is a common challenge in many computer vision tasks. Ideally, the impact of a cue should be large in situations when its extraction is reliable, and small, if the information is likely to be erroneous. While it is rather easy to show advantages of combined cues, if all weights are chosen manually, appropriate fusion mechanisms avoid such a manual parameter tuning. Uncertainty in the cue computation, i.e. optical flow and SIFT, is transferred to the pose estimation stage. This approach has similarities to Kalman filtering and particularly to the work on 2D shape tracking in [47]. In case of the region cues, we propose to couple cue computation and pose estimation by minimizing a joint energy functional. This energy can be interpreted as maximum a-posteriori estimation in a Bayesian setting. It is thus closely related to Bayesian weighting schemes in the context of 2D tracking [43], [37].

Due to its adaptivity, the tracking system is quite generally applicable without the need to tune the parameters for each specific scenario. We demonstrate this by experiments with textured and homogeneous rigid objects, as well as experiments on human motion estimation. The method can deal with considerable amounts of noise, background clutter, and large motion. A further challenge is the presence of partial occlusions. In order to limit the influence of these, we suggest to detect occlusions by means of the object model. They are taken into account when computing the optical flow and when selecting the SIFT keypoints.

The basic idea to combine a region-based tracking technique with point correspondences from dense optical flow has been presented in a preliminary conference paper [6]. The present paper extends this work in several ways. Firstly, the optical flow computation is adapted to the needs of pose tracking including an occlusion detection. Secondly, additional cues from the SIFT tracker are integrated. Thirdly, the paper comprises an

[1]This kind of tracking is also the basis for all methods that further constrain the solution space by means of prior knowledge.

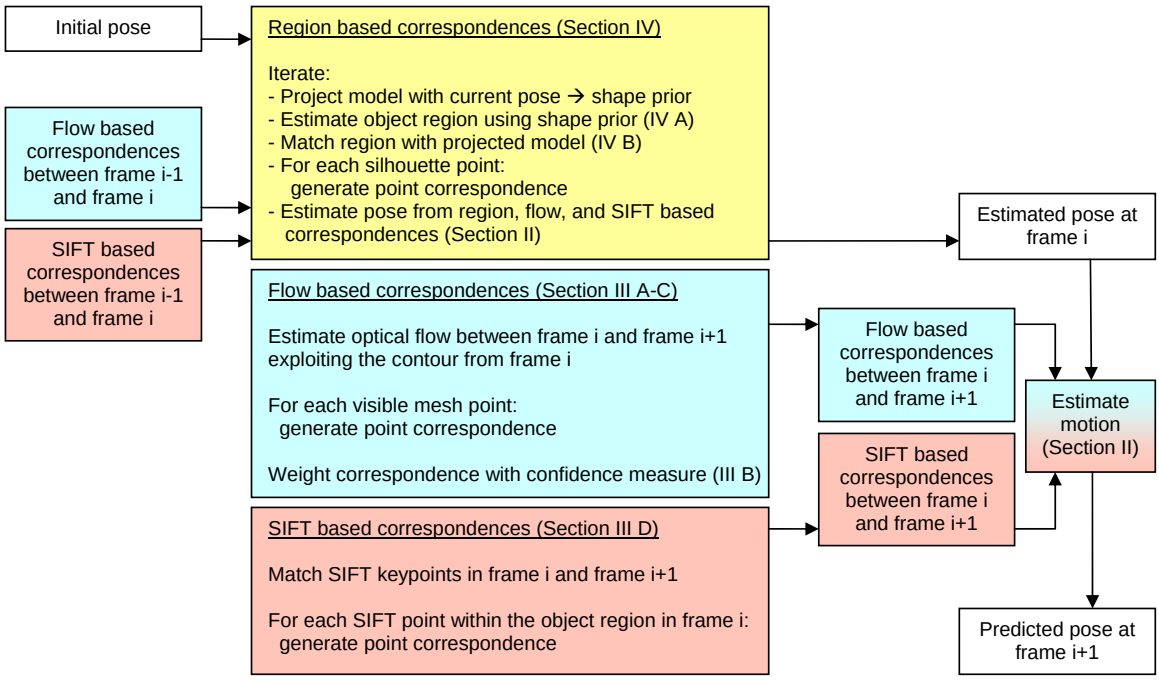


Fig. 2. System overview. Three sources for point correspondences are considered. Motion-based cues allow predicting good initializations, which are iteratively refined by the contour-based estimation.

elaborated way to adaptively weight the different cues. Finally, the method is applied not only to track rigid objects, but also to estimate human motion.

Figure 2 depicts an overview of the presented tracking system. Correspondences between 2D image points and 3D model points are established in three different ways: (a) by matching the projected model to the object region in the image, (b) by matching image points in successive frames via optical flow, and (c) by matching SIFT keypoints of successive frames. Section II clarifies our representation of rigid and articulated objects and explains how pose parameters are estimated from a given set of point correspondences. Section III and Section IV then show how these point correspondences can be derived from the image data. First we show in Section III how a state-of-the-art optical flow estimation technique can be adapted for this task, then we briefly review the concept of the SIFT tracker and explain the region-based part of the tracking system in Section IV. Section V summarizes the system before we experimentally show the effects of the combination and demonstrate the system's general applicability. The paper is concluded with a summary and a discussion on future challenges.

II. 3D Pose Estimation from Point Correspondences

A. Pose representation

In case of tracking rigid bodies, we aim at estimating the six degrees of freedom of a 3D rigid body motion. The corresponding group action can be written as $T(\mathbf{X}) = R\mathbf{X}+\mathbf{t}$, where $\mathbf{t} \in \mathbb{R}^3$ is a translation vector and $R \in SO(3)$ is a rotation matrix. For the purpose of pose estimation, a better representation of rigid body motions is the twist representation

$$\hat{\xi} = \begin{pmatrix} \hat{\omega} & \mathbf{m} \\ 0_{3\times 1} & 0 \end{pmatrix} \text{ with } \hat{\omega} = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}, \quad (1)$$

where the six parameters correspond to the six degrees of freedom. We can write these parameters as a vector $\xi = (m_1, m_2, m_3, \omega_1, \omega_2, \omega_3)$. Each twist can be translated to the corresponding group action by the exponential function $M = \exp(\hat{\xi})$; see [28] for details.

The one-parametric subgroup $M_\xi(\theta) = \exp(\theta\hat{\xi})$ with fixed ξ transforms points along the trajectory of a screw. A degenerate ξ (with no pitch component) can be used to model joints of a kinematic chain [4]. Such a kinematic chain allows modeling articulated objects, e.g., a human body consisting of rigid limbs interconnected by predefined joints like shoulders, elbows, etc. The model can be represented by a tree structure with the main torso as the root of this tree and the limbs as branches.

The motion of a point $(\mathbf{X}, 1)$ behind the jth joint is then described by the consecutive evaluation of exponential functions of all involved twists, including the twist describing the motion of the root:

$$\begin{pmatrix} \mathbf{X}' \\ 1 \end{pmatrix} = \exp(\hat{\xi}) \exp(\theta_1\hat{\xi}_1) \ldots \exp(\theta_j\hat{\xi}_j) \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix}. \quad (2)$$

Consequently, the state of a kinematic chain is defined by a parameter vector $\chi := (\xi, \Theta)$ that consists of the six parameters for the global twist ξ and the joint angles $\Theta := (\theta_1, \ldots, \theta_N)$.

B. Pose estimation

For estimating the parameters χ, a sufficient set of 2D-3D point correspondences is needed. How such correspondences are obtained is subject of later sections. For the moment we assume that a set of correspondences $(\mathbf{x}_i, \mathbf{X}_i)$, with $\mathbf{x}_i \in \mathbb{R}^2$ and $\mathbf{X}_i \in \mathbb{R}^3$ is given.

As the intrinsic camera parameters are known, the projection rays can be reconstructed from the 2D points $\mathbf{x}_i$. 3D lines can be represented implicitly by so-called *Plücker lines* [35], [42]. A Plücker line $\mathbf{L} = (\mathbf{n}, \mathbf{m})$ is described by a unit vector $\mathbf{n}$ and a moment $\mathbf{m}$. This line representation allows to conveniently determine the distance of a 3D point $\mathbf{X}$ to the line

$$d(\mathbf{X}, \mathbf{L}) = \|\mathbf{X} \times \mathbf{n} - \mathbf{m}\|_2, \tag{3}$$

where $\times$ denotes the cross product.

Provided the 2D-3D point correspondences are correct, the transformed 3D points must be on the projection rays reconstructed from their corresponding 2D points. In practice the correspondences are not exact for various reasons, yet we can seek to minimize the above distance. In particular, we seek a transformation $\chi = (\xi, \Theta)$ applied to all points $\mathbf{X}_i$ such that the total distance over all correspondences is minimized in the least squares sense:

$$\underset{\chi}{\operatorname{argmin}} \sum_i \left\| \pi\left(\exp(\hat{\xi}) \prod_{j \in \mathcal{J}(\mathbf{X}_i)} \exp(\theta_j \hat{\xi}_j) \begin{pmatrix} \mathbf{X}_i \\ 1 \end{pmatrix} \right) \times \mathbf{n}_i - \mathbf{m}_i \right\|_2^2, \tag{4}$$

where π denotes the projection of the homogeneous 4D vector to a 3D vector by neglecting the homogeneous component (which is 1), and $\mathcal{J}(\mathbf{X}_i)$ denotes the set of joints that affect the point $\mathbf{X}_i$. It is worth noting that minimizing the distance to the 3D ray and minimizing the 2D re-projection error could be made equivalent by appropriate rescaling of each error vector [34]. In multi-camera set-ups minimizing the 3D error is preferable since it treats all points equally, whereas the reprojection error prefers points closer to a camera.

Equation (4) states a nonlinear least squares problem. To solve for the parameters we use the Gauß-Newton method, i.e., the transformation matrix is linearized and the parameter estimation is iterated. With the identity matrix $\mathbf{I}$ and $\exp(\theta\hat{\xi}) \approx \mathbf{I} + \theta\hat{\xi}$ we can approximate (4) as the linear least squares problem

$$\underset{\chi}{\operatorname{argmin}} \sum_i \left\| \pi\left(\left(\mathbf{I} + \hat{\xi} + \sum_{j \in \mathcal{J}(\mathbf{X}_i)} \theta_j \hat{\xi}_j \right) \begin{pmatrix} \mathbf{X}_i \\ 1 \end{pmatrix} \right) \times \mathbf{n}_i - \mathbf{m}_i \right\|_2^2 \tag{5}$$

which can be solved, e.g., with the Householder method.

Correspondences from different views as well as different cues can be easily combined in the above least-squares framework by considering all of them in the sum of Equation (4). Nonlinear optimization with the Gauß-Newton method yields the optimum pose considering all constraints in the least-squares sense, which is related to the assumption of a Gaussian error distribution.

If there is a way to estimate the expected deviation of the matched points, for instance through a confidence measure, this can be incorporated by means of the variance of the Gaussian distribution. The sums in (4) and (5) are replaced by weighted sums $\sum_i w_i \| \cdot \|_2^2$, where w_i corresponds to the inverse variance of the Gaussian distribution. This leads to the well-known weighted least-squares setting. The detailed choice of the weights is discussed in Section V.

III. Motion-based Tracking

In this section, we consider two methods that compute 2D correspondences between successive frames t and $t+1$: optical flow and SIFT tracking. We assume the pose parameters of the model in frame t to be known. Therefore, it is known, how 3D model points project into this frame. Finding the new positions of the projected points in frame $t+1$ by either optical flow or the SIFT tracker yields 2D-3D point correspondences at $t+1$. From these the new pose of the object can be estimated using the technique described in the preceding section.

Such a procedure obviously accumulates errors over time. This is due to the assumption that the pose in the previous frame is known and is *exact*. As a consequence, even the smallest estimation errors are propagated from frame to frame. Therefore it is crucial to combine motion-based correspondences with region-based ones.

A. Optical flow

Optical flow is the common name for the displacement field $\mathbf{w}(\mathbf{x}) := (u(\mathbf{x}), v(\mathbf{x}), 1)$ between two images of an image sequence $I(\mathbf{x})$, where $\mathbf{x} := (x, y, t)$. Numerous optical flow estimation methods can be found in the literature. Variational methods currently mark the state-of-the-art and yield dense flow fields. Since we are interested in capturing large displacements, we further focus on multi-resolution methods. Building upon the method in [5], [9], we seek the optical flow as the minimizer of

$$\begin{aligned} E(u,v) = &\int_{\Omega_1} r(\mathbf{x}) \cdot \Psi_1\left(|I(\mathbf{x}+\mathbf{w}) - I(\mathbf{x})|^2\right) \mathbf{dx} \\ &+\gamma \int_{\Omega_1} r(\mathbf{x}) \cdot \Psi_1\left(|\nabla I(\mathbf{x}+\mathbf{w}) - \nabla I(\mathbf{x})|^2\right) \mathbf{dx} \\ &+\alpha \int_{\Omega_1} \Psi_2\left(|\nabla u|^2 + |\nabla v|^2\right) \mathbf{dx}. \end{aligned} \tag{6}$$

The energy consists of two parts. The first part states the gray value and the gradient constancy assumption, both weighted relatively to each other by the parameter $\gamma = 5$. This part is usually called data term. It is weighted locally by $r(\mathbf{x})$, which will be explained later. The second term introduces the assumption of a smooth flow field. It is weighted by the parameter $\alpha \geq 0$. $\Psi_1(s^2)$ and $\Psi_2(s^2)$ are so-called robust penalizer functions [2], [23]. In [5], $\Psi_1(s^2) = \Psi_2(s^2) = \sqrt{s^2 + \epsilon^2}$ with $\epsilon = 0.001$. Such a penalizer allows for outliers in the data (e.g. due to noise, specularities, occlusions) and in the smoothness assumptions (due to motion discontinuities). We adopt the same functions for tracking articulated objects and choose $\alpha = 50$.

In case of rigid objects, the model can be simplified by setting $\Psi_2(s^2) = s^2$ and $\alpha = 800$, which leads to a linear term in the Euler-Lagrange equations of the smoothness term. This simplification results in a faster implementation. It

becomes possible because in contrast to [5] the energy is only integrated inside the object region Ω_1. The object region is a byproduct of model-based tracking and beneficial as it already determines most of the relevant motion discontinuities. In case of rigid objects that are far enough from the camera, it even captures *all* relevant motion discontinuities. This is different for articulated objects. One could imagine, e.g., the case of two legs next to each other, one leg partially occluding the other. The legs can move in opposite direction, hence creating a motion discontinuity *within* the object region.

Another difference to the model in [5] is the explicit, local weighting $r(\mathbf{x})$ of the data term. This weighting is for integrating the result of the occlusion detection, which is described in Section III-C. The weights are set to

$$r(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \text{ occluded} \\ 1 & \text{else.} \end{cases} \tag{7}$$

At occluded pixels the data term is ignored and only the smoothness term determines the estimated flow. This yields a smooth interpolation of the flow field in areas, where the data does not reflect the motion of the object.

The minimizer of (6) can be computed with a continuous optimization method in a multi-resolution setting. After discretization of the Euler-Lagrange equations, we obtain a nonlinear system of equations that can be solved via two nested fixed point iteration loops and a solver for sparse linear systems. For details we refer to [5]. With a fast multi-grid solver, the optical flow can be computed in real-time [9]. Further speedups are possible with a GPU implementation [46]

B. Confidence measure for optical flow

Since we are interested in an adaptive weighting of optical flow correspondences versus correspondences from other cues, we need some measure that tells us something about the local confidence of the computed optical flow. A standard confidence measure is the gradient magnitude of the image $c_{\text{grad}}(\mathbf{x}) = |\nabla I(\mathbf{x})|$ or some similar expression [1]. However, this measure does not perform well in case of contemporary, variational optical flow methods, as pointed out in [10]. Instead, it was proposed in [10] to employ the local energy of variational methods as a confidence measure. We adopt this idea and use

$$\begin{aligned} c_{\text{Energy}}(\mathbf{x}) &= \beta \left(1 + e(\mathbf{x})\right)^{-1} \\ e(\mathbf{x}) &:= \Psi_1\big(|I(\mathbf{x}+\mathbf{w}) - I(\mathbf{x})|^2\big) \\ &\quad + \gamma \Psi_1\big(|\nabla I(\mathbf{x}+\mathbf{w}) - \nabla I(\mathbf{x})|^2\big) \\ &\quad + \alpha \Psi_2\big(|\nabla u|^2 + |\nabla v|^2\big) \end{aligned} \tag{8}$$

according to the energy stated in (6). This confidence measure is small in areas, where the assumptions stated in the energy functional cannot be fulfilled. Consequently, it indicates areas where optical flow computation is difficult and not reliable. Point correspondences derived from the optical flow are weighted by this confidence value. The factor β normalizes the confidence, such that $c_{\text{Energy}} = 1$, if the optical flow computation works reasonably well. If the confidence is larger, the correspondence obtains more influence than average, if it is smaller, its relative influence is decreased. Empirical evaluation resulted in $\beta = 12$ for $\Psi_2(s^2) = s^2$ and $\beta = 3$ for $\Psi_2(s^2) = \sqrt{s^2 + \epsilon^2}$.

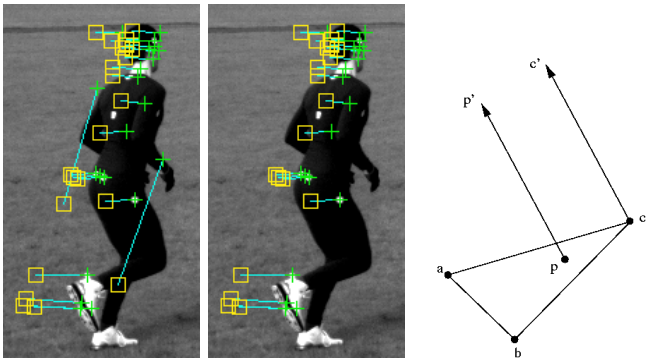


Fig. 3. **Left:** Matches between previous frame (squares) and current frame (crosses). **Center:** The outliers are removed after filtering. **Right:** When a matched SIFT keypoint p does not coincide with a projected mesh vertex, the 2D translation vector $p' - p$ is added to the closest vertex (here c). For the new 2D-2D correspondence c - c', the 2D-3D counterpart is available.

C. Occlusion detection

Occlusions are one of the most severe problems in tracking. In motion-based tracking methods, the motion of the occluding object is erroneously regarded as the motion of the tracked object. For this reason, most 2D trackers imply a monitoring stage, where the appearance of the tracked patch is compared to the patch at some earlier time. Once a patch has changed too much, it is ignored.

For 3D tracking we can make use of the object model in order to refine this concept. Knowing the pose parameters in a (non-occluded) frame, the object region can be mapped onto the model surface. This appearance model $f(\mathbf{X})$ can then be compared with the image in a successive frame by projecting it back to the image. For computing the similarity, we compare the gray value histograms p_a and p_b of the appearance model and the image patch, respectively:

$$d := \frac{1}{2} \int_{\mathbb{R}} |p_a(\zeta) - p_b(\zeta)| d\zeta. \tag{9}$$

It holds $d \in [0, 1]$, and if $d > \frac{1}{4}$ we define the center $\mathbf{x}$ of the patch to be occluded in the new frame and set $r(\mathbf{x}) = 0$. Otherwise the point is not occluded, we set $r(\mathbf{x}) = 1$, and we update the appearance model:

$$f_t(\mathbf{X}) = (1 - \alpha) f_{t-1}(\mathbf{X}) + \alpha I(\mathbf{x}), \tag{10}$$

where $\mathbf{x} = \pi(\mathbf{X})$ is the projection of the surface point $\mathbf{X}$ to the image plane. We set $\alpha = \frac{1}{8}$. Due to the updating step, the appearance model can adapt to changes in lighting. Note that the appearance model is *not* updated, if the point is marked occluded.

D. SIFT

The scale invariant feature transform and its corresponding region descriptor [20] currently belong to the most reliable techniques for sparse matching [24]. Matching is restricted

to keypoints which correspond to local extrema in scale-space. Each keypoint is described by orientation histograms computed in its neighborhood [20]. Correspondences between successive images are then established by nearest neighbor distance ratio matching [24] where conflicting correspondences are deleted. We used the distance ratio threshold of 0.6. Only keypoints that belong to the object region and are not occluded are considered.

As shown in Figure 3, the matching produces reliable point correspondences but also some outliers that need to be eliminated. The rudest mismatches for each pair of images are removed by discarding correspondences with an Euclidean distance that exceeds the average by a multiple as proposed in [15]. When the average is above a threshold, we also delete corresponding features with the same location since the match in frame $t+1$ then usually belongs to a static object in the background. Such pre-selection increases the inlier to outlier ratio, though it does not restrict the applicability to static backgrounds, as demonstrated in Figure 7. After deriving the 2D-3D correspondences, a preliminary pose is estimated and the new 3D correspondences are projected back in order to detect the remaining outliers.

In contrast to dense optical flow, with a point correspondence available for each projected mesh vertex, SIFT keypoints do not necessarily coincide with the projected mesh points. However, if the mesh is fine enough, we can assume the closest projected mesh point to undergo approximately the same 2D translation between two successive images as the SIFT keypoint. This is illustrated on the right hand side of Figure 3.

Thanks to the outlier detection and the high overall robustness of SIFT matching, a separate confidence measure like in case of the optical flow is not needed. The influence of SIFT correspondences automatically increases with the number of successful matches. In case of poorly structured objects, the number of these matches, and thus the influence of SIFT, will be low.

IV. Region-based tracking

In contrast to motion-based methods, region-based tracking does not require the exact pose in previous frames. Given a simplified model of the scene described by a set of parameters, we seek the parameters that best explain the image data. In our case, the scene is described by the object model and the background. They are parameterized by the sought pose parameters χ, the contour C between the object and background region in the image, and intensity distributions p_1 and p_2 in each region.

For convenience, we represent the contour C implicitly by the zero level line of a level set function $\Phi : \Omega \to \mathbb{R}$. It splits the image domain Ω into the object region Ω_1 and the background region Ω_2, where $\Phi(\mathbf{x}) > 0$ if $\mathbf{x} \in \Omega_1$ and $\Phi(\mathbf{x}) < 0$ else. We generally constrain Φ to be the signed distance image of the contour. This means the absolute value of $\Phi(\mathbf{x})$ is the minimum distance of $\mathbf{x}$ to the contour. Seeking the optimum parameters is then described by the following energy minimization problem [34]:

$$\begin{aligned} &E(\Phi, p_1, p_2, \chi) = \\ &\underbrace{-\int_\Omega \big(H(\Phi)\log p_1 + (1-H(\Phi))\log p_2 + \nu|\nabla H(\Phi)|\big)\,\mathbf{dx}}_{\text{segmentation}} \\ &\underbrace{+\lambda\int_\Omega (\Phi - \Phi_0(\chi))^2 \mathbf{dx}}_{\text{shape distance}} \to \min, \end{aligned} \tag{11}$$

where $H(s)$ denotes a regularized version of the step function, Φ_0 denotes the shape of the projected object model, and ν and λ are tuning parameters, which we fix at $\nu = 0.001|\Omega|^{0.7}$ and $\lambda = 0.05$. Obviously, the first part is very similar to segmentation models stated in [27], [11], [30]. The second part couples the segmentation model and the pose parameters as it enforces the projected object model to match the object region. This has two effects: firstly, the pose parameters are adapted such that the projection fits the region extracted by the segmentation part. Secondly, the segmentation is constrained by the shape of the object model and is not allowed to deviate too much from this shape. The tolerated amount of deviation depends on the clarity of the image data and the choice of λ. The pose parameters (yielding Φ_0) and the level set function Φ are optimized in an iterative, alternating scheme. See Figure 2 for the system overview.

In case the image-driven segmentation is not well constrained, e.g. due to heavy clutter or irregular texture, the contour stays close to Φ_0. The solution is then dominated by the optical flow and the SIFT features. In the opposite case, e.g. there is a homogeneous object, the segmentation part is very dominant and yields correspondences that can correct errors of motion-based tracking.

A. Segmentation

The energy in (11) leaves room for various ways to model the probability densities p_1 and p_2. The most simple choice is the approximation of each region by its mean [11]. However, this would restrict the tracking scenarios to homogeneous objects with homogeneous background. Thus in [34] we proposed to model the regions by local Gaussian distributions on a feature space consisting of the gray value and color, as well as some texture descriptors. These can be responses of Gabor filters [14] or, more efficiently, the texture features suggested in [8]. In order to keep the region model manageable, the feature channels are assumed to be independent, so $p_i = \prod_j p_{ij},\ i = 1, 2$. However, due to the variability of the variance in the Gaussian distribution, the relative importance of a channel j is determined automatically by its discriminative properties.

Local distribution models allow to drop the assumption of identically distributed pixels in each region. In contrast, at each spatial position $\mathbf{x}$ we have a separate probability density. For a Gaussian distribution this reads [34]:

$$p_{ij}(s, \mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_{ij}(\mathbf{x})} \exp\left(\frac{(s - \mu_{ij}(\mathbf{x}))^2}{2\sigma_{ij}(\mathbf{x})^2}\right). \tag{12}$$

Estimation of the parameters $\mu_{ij}(\mathbf{x})$ and $\sigma_{ij}(\mathbf{x})$ can be achieved using a Gaussian window with standard deviation

$\rho = 12$ and restricting the estimation only to points within this window.

Minimization of (11) with respect to Φ and p_i is achieved by gradient descent. Having an initialization of Φ by the projected object surface, we can estimate p_i. From these we can update Φ by

$$\Phi^{k+1} = \Phi^k + H'(\Phi^k)\left(\log\frac{p_1^k}{p_2^k} + \nu \mathrm{div}\left(\frac{\nabla\Phi^k}{|\nabla\Phi^k|}\right)\right) + \lambda(\Phi_0 - \Phi^k) \tag{13}$$

with iteration index k. When moving on to a new frame, it makes sense to run a few iterations with the densities from the previous frame before adapting p_i. This allows the contour to capture the new position of the object boundary. We assume the distribution to be sufficiently smooth for being valid also for the displaced regions in the new frame. This is ensured by the large Gaussian window with $\rho = 12$.

We would like to emphasize that it is due to the adaptivity of the variances σ_{ij} that the relative importance of motion- and region-based cues is adapted automatically. If the variances in both regions are large, the first term in (13) will get small and will be dominated by the second term that carries the information of the motion-based tracking. Vice-versa, if the regions are homogeneous, the variances will get small, so the first term in (13) dominates and forces the pose parameters to be adapted for Φ_0 matching Φ.

B. Shape matching

The shape distance between Φ and Φ_0 in (11) relates the pose parameters χ to the region represented by Φ. To estimate the pose parameters for a given Φ, we need 2D-3D point correspondences. Since Φ_0 is the projection of the object model, corresponding 3D points on the model are known. Thus 2D-3D correspondences can be derived by matching the 2D shapes Φ and Φ_0. Towards this end, we seek the displacement vector field $(u(\mathbf{x}), v(\mathbf{x}))$ that minimizes

$$\int_\Omega (\Phi(x,y) - \Phi_0(x+u, y+v, \chi))^2 \mathbf{dx}. \tag{14}$$

In practice, we are only interested in correspondences for points along the contours.

Numerous methods on 2D shape matching can be found in the literature. We are interested in a method that can deal with shape deformations in order to handle projective distortion and articulated objects. Moreover, we can assume that the transformation between the shapes is limited. A suitable and simple method is closest point search. It can be computed efficiently, if the two contours Φ and Φ_0 are represented by distance images, i.e., the value of $\Phi(\mathbf{x})$ is the minimum distance of $\mathbf{x}$ to the contour. A very efficient method for computing the minimum Euclidean distance in linear time is provided in [13].

The estimated region Φ may contain estimation errors, for instance due to partial occlusion or background clutter. For the shape matching to be more robust in such cases, a robust function Ψ can be applied together with a regularity term that penalizes shape deformations:

$$\int_\Omega \Psi\left((\Phi(x,y) - \Phi_0(x+u, y+v, \chi))^2\right)\mathbf{dx} + \alpha\int_\Omega (|\nabla u|^2 + |\nabla v|^2)\mathbf{dx}, \tag{15}$$

where $\Psi(s^2) = \sqrt{s^2 + \epsilon^2}$ with $\epsilon = 0.001$ like in the case of optical flow computation. Indeed this kind of shape registration can be computed by the same numerical scheme as used for optical flow estimation [33]. This can be done very fast. Since Φ and Φ_0 are distance images and very smooth, only few iterations are needed. Figure 4 shows a comparison of standard closest point matching and the regularized matching. Clearly, the correspondences obtained with the regularized matching are more regular and tend to ignore noise in one of the contours.

Fig. 4. **Left:** Projected object surface in blue and the extracted object contour in yellow. One seeks corresponding points between the silhouette of the blue area and the yellow contour. **Center:** Closest point correspondences. **Right:** Regularized closest point computed with optical flow numerics and $\alpha = 10$.

V. Fusion of Point Correspondences and Adaptivity of the System

The previous sections introduced the details on how point correspondences can be established using different matching strategies and how such correspondences can be employed to estimate the pose parameters. The present section summarizes the whole system, particularly the fusion of point correspondences, the way how this fusion can exploit uncertainties, and how such uncertainties are estimated in our system.

Information fusion. There are two places in our system where information is combined to improve the robustness of tracking. The first is the fusion of point correspondences estimated with different matching methodologies. These correspondences are combined in the energy in (4), which states a least-squares problem. This formulation assumes that errors in the correspondences are Gaussian distributed. Since a matching strategy may fail completely at some points, which renders the global Gaussian noise assumption inappropriate, we seek to detect such situations and reflect the uncertainty (or expected error) by means of a confidence measure. Correspondences with a large uncertainty are assigned smaller weights in a weighted least squares setting.

The second place where information is combined is the prediction step; see also Fig. 2. Here the uncertainty of the contour based matches is reduced by means of motion based point correspondences, which can handle larger transformations and yield a better initial contour. This step alleviates one particular shortcoming of the contour based matching, this is its sensitivity to the initial pose.

Estimation of uncertainty. In order to prefer the more reliable correspondences in the weighted least squares setting

in (4), we need an estimate of the reliability of each point correspondence. In case of optical flow, such a confidence measure is defined in (8). In case of SIFT, outliers are detected explicitly and are then removed, so we have a binary confidence estimate here.

In case of the contour matching, there is also a confidence measure, even though it is not explicit. Areas where the segmentation is evident, i.e. the difference of the log-likelihoods of foreground and background is large, the image driven part of the segmentation energy in (11) dominates the shape prior and the contour can deviate much from the projected model. Vice-versa, if the foreground and background distributions fit almost equally well, the segmentation will stay close to the shape prior, i.e., the correspondences reflect the initial pose estimated with flow and SIFT based correspondences. In this case, the correspondence vectors have zero length, i.e. their confidence is zero. If the log-likelihood ratio is large, on the other hand, the vectors have larger length and thus their weight in (4) is larger.

Apart from their confidence also the number of correspondences determines the influence of a matching strategy on the overall system. Therefore, we suggest to normalize the weights such that if all cues can be extracted in an equally reliable manner, they are more or less equally weighted. Let n_{C} and n_{OF} denote the number of contour- and flow-based correspondences, respectively. We take the contour-based correspondences as reference and assign all of them $w_{\mathrm{C}} = 1$. SIFT correspondences are all assigned the weight $w_{\mathrm{SIFT}} = 0.002 \cdot n_{\mathrm{C}}$. Optical flow correspondences are weighted individually by means of the confidence measure described in Section III-B. For a correspondence i with confidence c_i, we assign the weight $w_i = c_i \frac{n_{\mathrm{C}}}{n_{\mathrm{OF}}}$. For the pose prediction, where no contour-based cues are available, the factor n_{C} is replaced by n_{OF}, respectively.

Theoretical gain of the combination. The different matching strategies have different shortcomings and fail in different situations. Thanks to the fusion, if one matching strategy fails and this failure is detected by its confidence measure, other matching strategies can take over and may ensure a good pose estimate. This allows to run the method on different data preferring different cues. In the coming experimental section, we will see how far this theoretical gain can be observed in practical experiments.

VI. Experiments

In order to demonstrate the ability of the tracking system to deal with a number of challenges, we applied it to numerous tracking scenes. These scenes contain homogeneous as well as textured objects, large transformations, noisy images, partial occlusions, and articulated human motion. With the experiments we aim at showing that, due to the combination of complementary cues and their adaptive weighting, the tracking system can handle all these scenes without the need of manual adaptations.

A. Rigid objects

Figure 5 depicts an experiment where a tea box has been moved by about 30 pixels between two frames including a rotation. See Figure 1 for the input images. As the transformation is quite large, the computed optical flow vectors contain errors. This can be seen from the pose prediction in Figure 5b,d. However, thanks to the additional region-based correspondences, the final pose result is good (Figure 5c). Conversely, the pose estimation also fails if only the region-based correspondences are used. This is shown in Figure 5e. Figure 5f reveals that in this scene there are not enough SIFT keypoints on the object (only one, to be precise) for tracking the tea box. This experiment demonstrates two things. Firstly, there are scenes where none of the cues alone is able to correctly track the object. Taking region- and motion-based cues together, on the other hand, leads to a successful tracking. Secondly, there is clearly a difference between the usage of correspondences from optical flow and SIFT. While the estimated flow might not be exact in difficult situations, it provides at least enough correspondences for a unique approximate solution. SIFT correspondences are usually more reliable, but their number is sometimes not sufficient to estimate the pose.

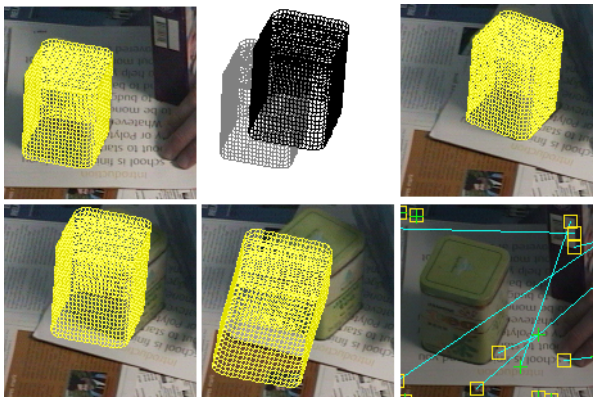


Fig. 5. Combining motion- and region-based tracking allows to capture the large motion of a tea box. **Top row, from left to right: (a)** Object pose at frame 1. **(b)** Object motion due to the estimated optical flow between frame 1 and frame 2. Gray: pose from frame 1. Black: pose prediction for frame 2. **(c)** Estimated pose at frame 2 using combined motion- and region-based tracking. **Bottom row, from left to right: (d)** Bad pose just using motion-based tracking. **(e)** Bad pose just using region-based tracking. **(f)** Not enough distinctive SIFT features are located to allow for a proper prediction. Motion-, region-, or SIFT-based tracking alone cannot handle this situation.

In order to evaluate the sensitivity of the region based pose estimation on the initialization, we added increasing perturbations to the correct pose. This kind of experiment is also commonly used in the scope of active appearance models [22]. The perturbing twists were $0.01\theta(10, 10, 10, 0.5, 0.5, 0.5)^\top$ for increasing θ. The remaining average deviation of all mesh points is depicted in Fig. 6 together with the initial poses for three θ. Clearly, the method can deal very well with small perturbations, and the pose estimates are still quite good with medium perturbations. The reason for some smaller perturbation leading to inferior results than a larger perturbation is due to different ways from the initialization to the next optimum. Already very small structures can be the reason for a local minimum. Initializations that are too far away lead to local minima that correspond to very bad poses. For this reason, motion based cues are needed to handle fast motion at low frame rates.

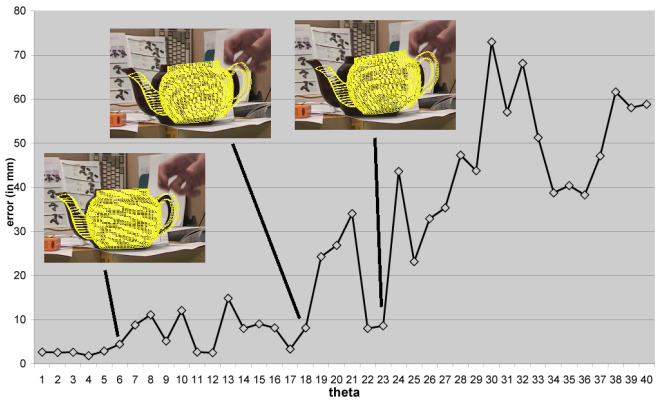


Fig. 6. Sensitivity of the region based method on the initial pose. The diagram shows the average error of the mesh points depending on the amount of disturbance from the correct pose. Three key initial poses are depicted in the images.

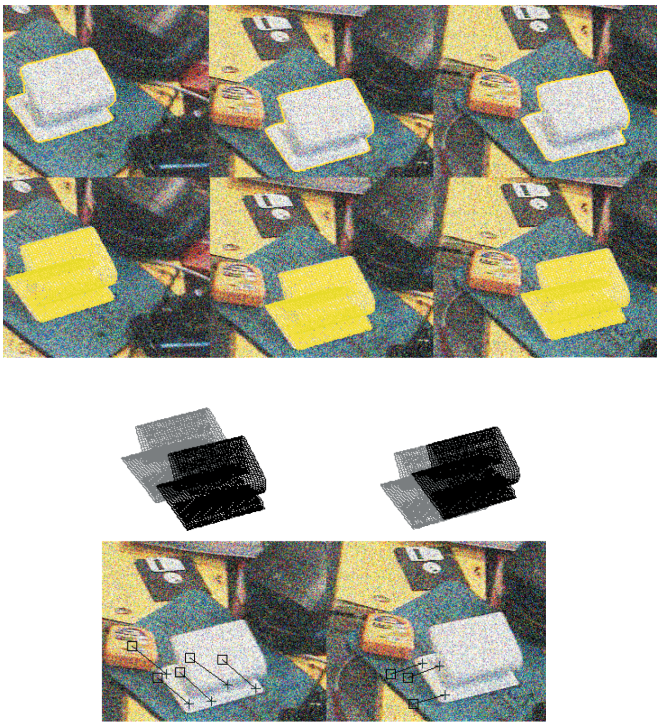


Fig. 7. Four successive frames from a sequence with the camera moving and Gaussian noise with standard deviation 60 added (140 frames, 8fps). **First row:** Extracted contour. **Second row:** Estimated pose. **Third row:** Object motion due to optical flow correspondences. Gray: pose from previous frame. Black: pose prediction at current frame. **Last row:** A very similar result is obtained with the SIFT tracker.

In Figure 7 displacements between successive frames are almost of the size of the object. Without a motion based prediction, region based pose estimation will fail to track this object. Surprisingly, although the object is homogeneous in large parts and there is a very high amount of noise added to the input images, multi-resolution optical flow is still able to capture its motion by means of its coarse-scale structure. The SIFT descriptor works fine as well, though there are only few SIFT regions on the puncher. When further decreasing the frame rate by skipping every second image, optical flow fails as the motion is larger than the tracked structure itself. For

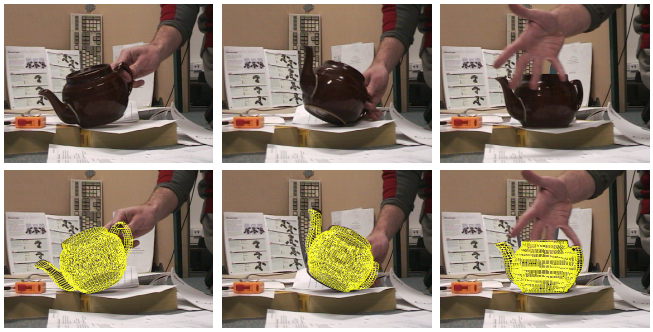


Fig. 8. **Top row:** Frames 97, 116, and 188 of a stereo sequence used for the experiments in Table I. **Bottom row:** Tracking results. See also the supplementary material for a video.

noise level	0	20	40	60	80
region	124	115	95	85	5
region+flow	tracked	115	115	75	5
region+SIFT	tracked	110	100	25	5
region+flow+SIFT	tracked	115	115	85	5

TABLE I

SENSITIVITY TO NOISE IN THE INPUT IMAGES. THE TABLE INDICATES THE FRAME NUMBER WHERE TRACKING FAILED. THE SEQUENCE CONTAINS 196 IMAGES, SOME OF THEM ARE SHOWN IN FIG. 8.

the SIFT tracker, the larger transformation is not a problem. The accumulation of inaccuracies is prevented by the region-based matching. Once the projected object model covers larger parts of the object region, the segmentation can robustly determine the exact location of the object contour, thanks to the homogeneity of the object region. As a consequence, it can correct errors of the motion-based prediction. This experiment shows that the system can deal with homogeneous objects, even if there are large displacements and substantial degradation by noise.

Figure 8 shows a slightly more difficult sequence, which we used to quantitatively determine the sensitivity to noise in the

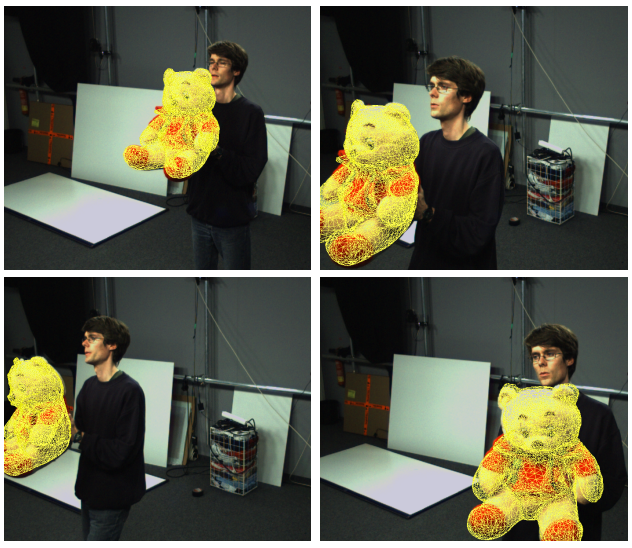


Fig. 9. Tracking result for another rigid object. One out of three camera views is shown. See also the supplementary material for a video.

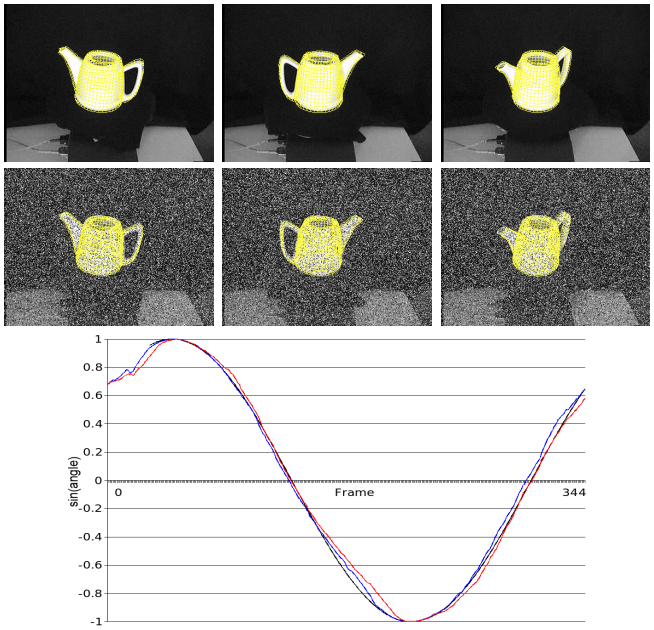


Fig. 10. **Top row:** Tracking results of a tea pot on a turntable. **Center row:** Tracking results with 50% of the pixels in the input image replaced by a uniformly distributed random value. **Bottom:** Comparison of the estimated pose (blue) versus the true motion (black). The red curve shows the result on the noisy input images.

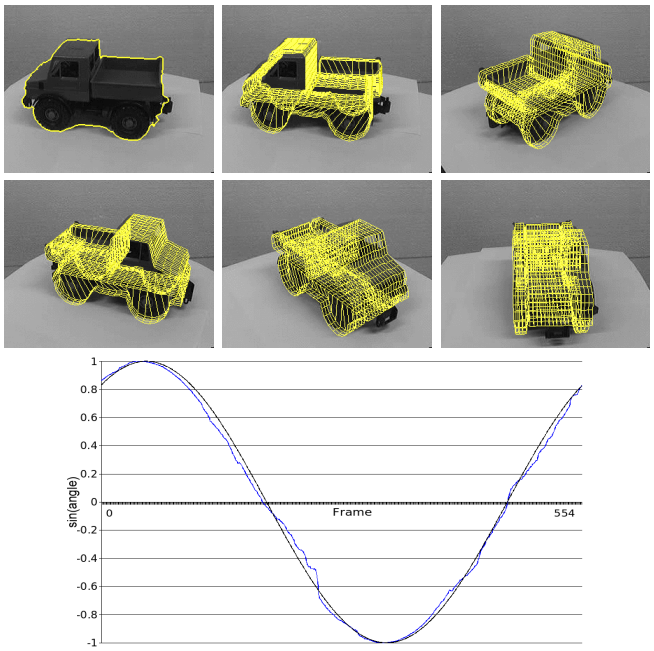


Fig. 11. **Top:** Input image with estimated contour and tracking results of a toy car on a turntable. **Bottom:** Comparison of the estimated pose (blue) versus the true motion (black).

input images. We added increasing amounts of noise to the images and observed the frame number when tracking failed. The results are shown in Table I. Without additional noise, the combined system can track the sequence completely. With increasing noise tracking fails earlier in the sequence. With the combined system successful tracking is possible for a larger number of frames.

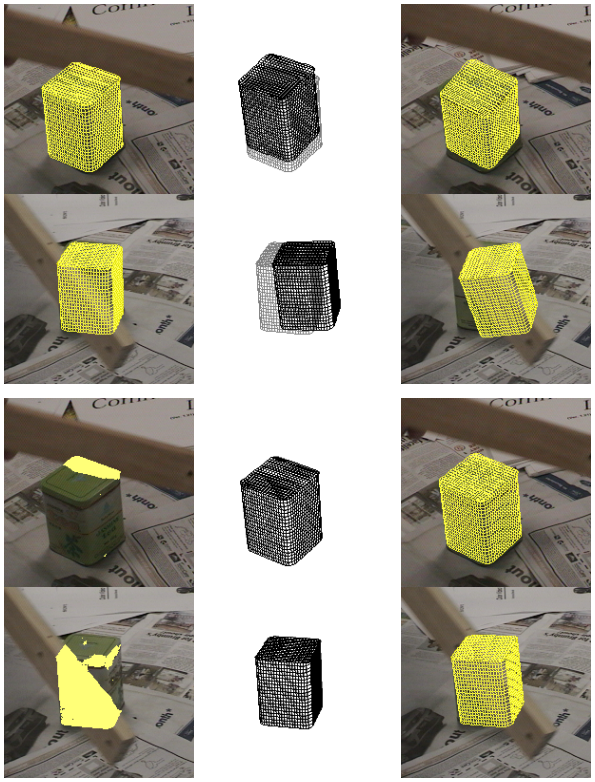


Fig. 12. Relevance of occlusion detection. **Top:** Result without occlusion detection. **From left to right:** (a) Initial pose in a stereo frame. In both views the object is partially occluded. (b) The estimated optical flow in the occluded area reflects the motion of the occluding object. (c) The object pose is disturbed by the bad motion-based correspondences. The region-based tracking cannot compensate the error, because it also suffers from the occlusion. **Bottom:** For comparison the method *with* detection of areas that are occluded. **From left to right:** (d) Areas that deviate from the model appearance are marked as occluded. (e) Ignoring these areas in optical flow estimation, the motion of the tracked object, in this case zero motion, is estimated correctly. (f) Finally estimated pose based on motion- and region-based correspondences.

We performed two further experiments with quantitative results, as depicted in Figure 10 and Figure 11. Ground truth has been provided by placing the tracked objects on a turntable and reading the true pose from the turntable controller[2]. The tracking curves reveal a very accurate tracking of the objects. In case of the tea pot, the average error is only 2.3 degree. The error increased to 4.6 degree replacing 50% of the pixel in the input images by uniform noise. In case of the car, the average error is 2 degree.

Figure 12 demonstrates the occlusion detection. Tracking without occlusion detection leads to large errors since the estimated motion reflects in large part the motion of the occluding stick instead of the tea box to be tracked. Clearly, the proposed appearance based method is able to detect these parts. Estimating the optical flow based on data from the non-occluded areas only, avoids bad pose estimations caused by the motion-based component of the system.

Another demonstration of the occlusion detection is shown in Figure 13 where a tea pot is swayed. Two occluding boxes have been added to the images. They continuously move across

[2]The sequences and ground truth data are provided at `www.tnt.uni-hannover.de/project/TPAMI09Benchmark/`.

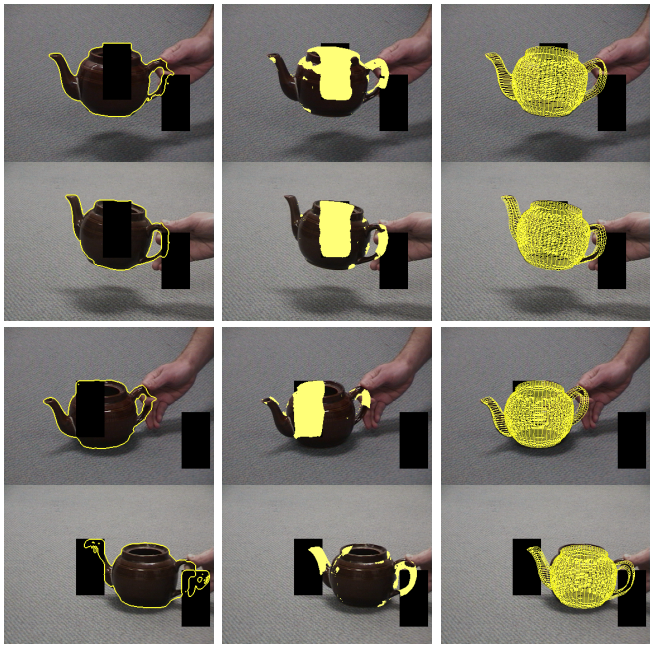


Fig. 13. Three frames from a stereo sequence with a moving object and two moving, occluding boxes. **Left:** Extracted contour. **Center:** Detected occlusions marked in yellow. **Right:** Estimated pose. Thanks to robust shape matching and consideration of occlusions in optical flow estimation, occlusions that are not too large can be handled.

the image (see also the video in the supplementary online material). The occluded areas are quite well detected, which ensures the successful tracking of the tea pot.

B. Human motion tracking

In another set of experiments, we applied the system to the tracking of articulated objects, in particular to human motion tracking. Besides the global rigid motion, the joint angles of the body model represent further degrees of freedom that have to be estimated.

Due to the relatively small size and fast motion of limbs, it is very likely that region-based tracking gets stuck in local optima and tracking fails. Hence, the predicted pose due to optical flow and SIFT matches is particularly important for human motion tracking. This is demonstrated by the experiment in Figure 14 where the upper body of a person waving their arms is tracked. Without a good pose prediction, the arm movement is clearly underestimated, as the contour extraction gets stuck in a local optimum. Optical flow and SIFT together allow for good predictions. SIFT alone is not sufficient, since the number of keypoints is often too small for a unique estimate. Provided a good prediction, the region-based cues ensure a precise final pose estimate without accumulating the errors from motion-based tracking.

The experiment in Figure 15 shows the outcome of a full-body outdoor running sequence. The body model has 26 degrees of freedom and the image data was captured with four Basler gray-scale cameras and 120 frames per second. Ground-truth data was obtained for this sequence through parallel tracking of the person with a marker-based system.

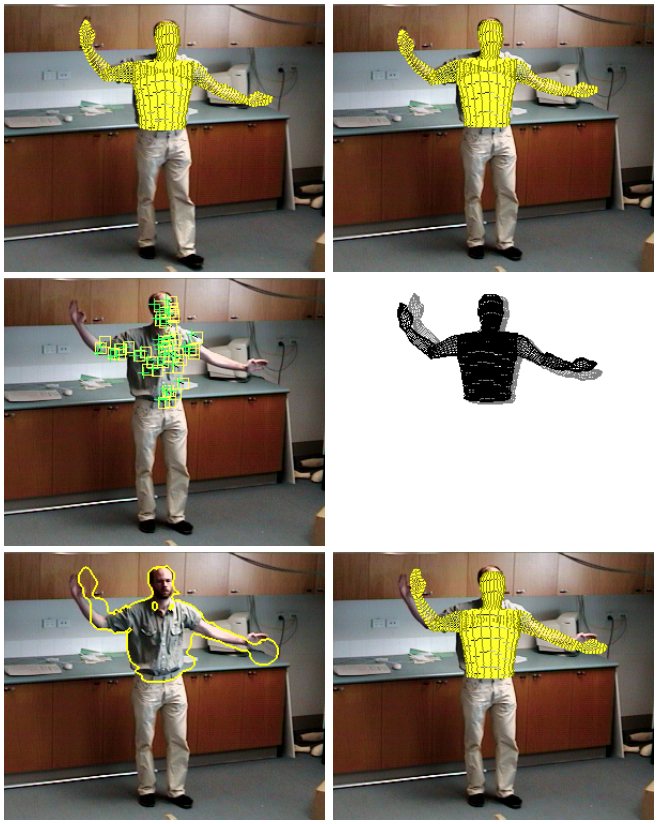


Fig. 14. Combining motion- and region-based tracking allows to capture fast upper body motion. **Top row:** Initialization with the pose from the previous frame (left), and the estimated pose in the new frame when combining all available cues (right). **Middle row:** Matched SIFT keypoints (left). Yellow rectangles indicate keypoints in the previous frame, green crosses keypoints in the new frame. In this frame, successfully matched keypoints are available at the main body but missing at the hands. Right: motion prediction by optical flow and SIFT. **Bottom row:** The same situation with region-based cues only. Lacking a sufficiently close initialization, contour extraction fails (left) and leads to an inaccurate pose estimation (right).

Fig. 15. Full body tracking in a sequence with ground truth data. **Top row:** Input frames from one out of four camera views. **Bottom row:** Synthesized images from the tracked 3D pose. A different viewpoint than in the input images is depicted. Further results are shown in Figures 16, 18, 17, and Table II.

Bad marker correspondences have been corrected manually[3].

Thanks to combined cues, even fast motion can be tracked, as illustrated in Figure 16. The image in the top left corner depicts the start pose. The second image shows the predicted

[3]The sequence and ground truth data are provided at `www.tnt.uni-hannover.de/project/TPAMI09Benchmark/`.

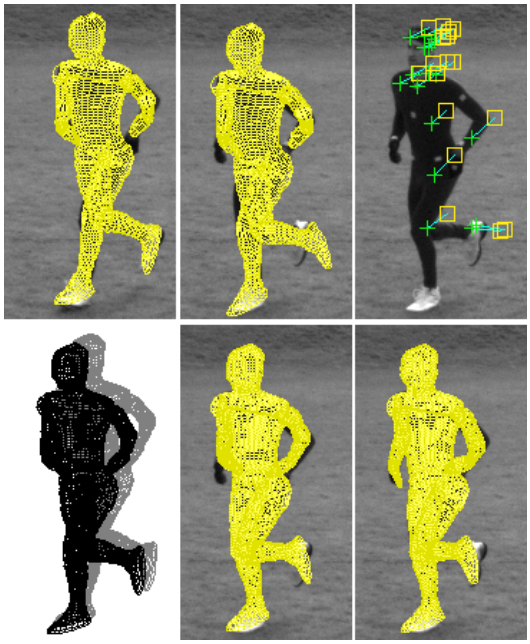


Fig. 16. Combining motion- and region-based tracking allows to capture the fast motion of a jogging person. **Top row, from left to right: (a)** Object pose at frame 1. **(b)** Pose at frame 2 estimated from optical flow correspondences only. **(c)** Tracked SIFT features: not enough features are located to ensure a proper prediction. **Bottom row, from left to right: (d)** Estimated prediction at frame 2 using combined optical flow and SIFT information. Gray: pose in frame 1. Black: prediction for frame 2. **(e)** Prediction from (d) overlaid with the image. The outcome is much better than the result in (b). **(f)** Final outcome for motion- and region-based tracking.

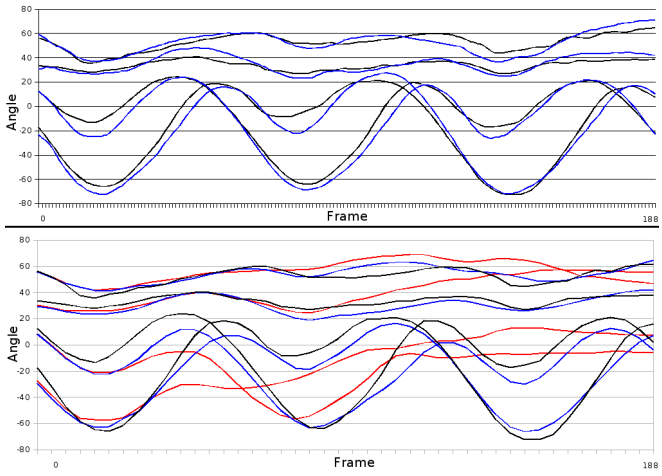


Fig. 17. Tracking diagram for the sequence in Figure 15. The curves show the angles of the two elbow and the two knee joints. **Top:** Comparison of the proposed system (blue) to the ground truth (black). **Bottom:** Comparison of the combined system (blue) to the purely region-based system (red) for a reduced frame rate of 24fps. The black curve shows again the ground truth. The tracking failure of the single-cue system is clearly visible. See Table II for average errors.

motion in the next frame using optical flow. The third image shows the tracked SIFT-features. Due to the black body suit, not enough features are detected to allow for a proper prediction using SIFT tracking alone. Tracking fails even with regularized equations since limb movements are not properly predicted. The left and center image in the bottom row depict

Fig. 18. Illustration of the drift when only flow-based correspondences are used for tracking. **From left to right:** Result at frames 1, 10, 30, and 150. The optical flow yields good results for the first frames, which indicates its suitability for predicting the pose in successive frames. However, errors accumulate over time and are the reason for tracking failures of the limbs. Successful tracking of the main torso even after 150 frames indicates the generally high precision of the flow-based correspondences.

the outcome of the combined optical flow and SIFT tracker. It is superior to the results of the separate motion predictors. The estimate is further refined if also region-based tracking is involved. Compare, e.g., the right hand of the person.

Table II shows quantitative results for the most interesting cue combinations. Clearly, the combination of correspondences improves the robustness of tracking when the frame rate is reduced. When tracking is successful, the results are very precise with average errors of about 5 degrees. Figure 17 depicts corresponding tracking curves for the elbow and knee angles. The system with the combined cues is close to the ground truth even when the frame rate is small, whereas tracking with the purely region-based system fails (red curves).

Tracking with purely motion-based cues always fails due to accumulation of errors. Figure 18 illustrates the corresponding drift. Although the estimated optical flow is extremely precise, as indicated by the successful tracking of the torso over 150 frames, even smallest errors accumulate over time especially at limbs with few correspondences. Such drift can only be avoided by region-based correspondences, which are based on matching the image directly to the model and do not suffer from small errors in previous frames.

The computation time for tracking the full body model with four camera views was around 4 minutes per frame. Tracking the upper body model with two camera views took approximately 80 seconds per frame. While this is clearly not realtime performance, the focus of this paper is on a general and robust system, not on a fast one. The rather large computation time is mainly due to the iterative region-based tracking and the involved local region statistics including a texture feature space. Using less sophisticated components here would substantially reduce the computation time.

VII. Discussion

We have proposed the combination of surface-region matching, optical flow, and SIFT tracking for 3D motion capture of rigid and articulated objects. The system is designed in a way that all involved cues can incorporate their strong aspects, while weaknesses are sought to be suppressed. This is achieved as the system adaptively weights cues according

	flow only	region only	region+SIFT	region+flow	region+SIFT+flow
120 fps	- (30)	4.29 ± 3.42	4.35 ± 3.31	4.42 ± 3.38	4.46 ± 3.38
40 fps	- (30)	- (165)	4.35 ± 3.34	4.31 ± 3.43	4.29 ± 3.38
30 fps	- (33)	- (118)	4.86 ± 4.29	4.47 ± 3.94	4.73 ± 3.99
24 fps	- (21)	- (33)	- (33)	- (25)	5.83 ± 4.91

TABLE II

COMPARISON OF CUE COMBINATIONS AT VARIOUS FRAME RATES CORRESPONDING TO THE EXPERIMENT IN FIGURE 15. THE TABLE SHOWS THE AVERAGE ERROR OF THE KNEE AND ELBOW JOINT ANGLES OVER ALL 180 FRAMES IN DEGREES. THE SECOND VALUE INDICATES THE STANDARD DEVIATION. TRACKING FAILURES ARE MARKED BY '-' AND A NUMBER THAT INDICATES THE FRAME WHERE TRACKING FAILED (ONE BAD LIMB).

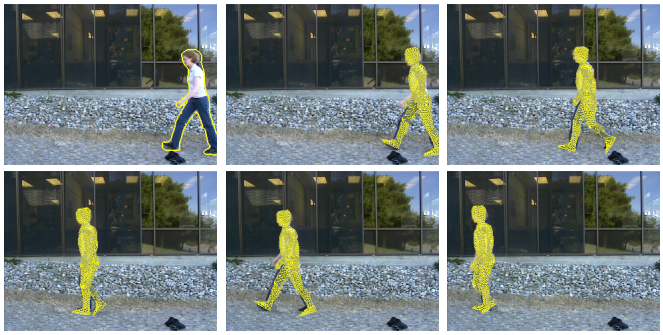


Fig. 19. Monocular standard test sequence with self-occlusions. Input image with estimated contour and tracking results.

to their reliability. This results in a very generally applicable tracking system. We demonstrated this by a number of experiments in very different scenarios, where we obtained stable tracking results although the parameters of the system were not manually adapted when changing the scene. In particular, the system is able to capture large transformations, it can track textured as well as homogeneous objects, and it can deal with noise and partial occlusions. Furthermore, we have demonstrated that the system can be applied to human motion tracking, even when prior knowledge about typical human movements is missing.

Obviously, priors becomes necessary in monocular scenes, which lack image cues for some body parts. Figure 19 shows an example where the right arm of the person is fully occluded. While in this work we focused on *image* cues for tracking and deliberately ignored all priors on the pose or the dynamics of articulated objects, even smoothness priors, it is easy to supplement these priors in our system. Figure 19 shows tracking results where the proposed system has been supported by a kernel density estimate on a set of walking motions [7].

Rather than tracking an object from frame to frame, the object can also be consecutively detected in each frame [40], [26], [29], [31]. This approach has become feasible due to recent advances in object recognition and reveals many appealing properties, among them auto-initialization, no problems with large motion, and re-initialization after full occlusions. On the other hand is is obvious that detection solves a harder problem than tracking. Particularly, it remains the problem to interpolate between recognized views. We believe that the way how to combine good detection results with the temporal consistency of classical tracking methods is an important issue of future research.

REFERENCES

[1] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, Feb. 1994.

[2] M. J. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, Jan. 1996.

[3] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8–15, Santa Barbara, California, 1998.

[4] C. Bregler, J. Malik, and K. Pullen. Twist based acquisition and tracking of animal and human kinematics. *International Journal of Computer Vision*, 56(3):179–194, 2004.

[5] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In T. Pajdla and J. Matas, editors, *Proc. 8th European Conference on Computer Vision*, volume 3024 of *LNCS*, pages 25–36. Springer, May 2004.

[6] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel. High accuracy optical flow serves 3-D pose tracking: exploiting contour and flow based constraints. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Proc. European Conference on Computer Vision*, volume 3952 of *LNCS*, pages 98–111. Springer, 2006.

[7] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel. Nonparametric density estimation with adaptive anisotropic kernels for human motion tracking. In *Proc. 2nd International Workshop on Human Motion*, volume 4814 of *LNCS*, pages 152–165. Springer, 2007.

[8] T. Brox and J. Weickert. A TV flow based local scale estimate and its application to texture discrimination. *Journal of Visual Communication and Image Representation*, 17(5):1053–1073, Oct. 2006.

[9] A. Bruhn and J. Weickert. Towards ultimate motion estimation: Combining highest accuracy with real-time performance. In *Proc. 10th International Conference on Computer Vision*, pages 749–755. IEEE Computer Society Press, Beijing, China, Oct. 2005.

[10] A. Bruhn and J. Weickert. Confidence measures for variational optic flow methods. In R. Klette, R. Kozera, L. Noakes, and J. Weickert, editors, *Geometric Properties from Incomplete Data*, volume 31 of *Computational Imaging and Vision*, pages 283–297. Springer, 2006.

[11] T. Chan and L. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, Feb. 2001.

[12] D. DeCarlo and D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *International Journal of Computer Vision*, 38(2):99–127, July 2000.

[13] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. Technical Report TR2004-1963, Computer Science Department, Cornell University, Sept. 2004.

[14] D. Gabor. Theory of communication. *The Journal of the Institution of Electrical Engineers*, 93:429–457, 1946.

[15] J. Gall, B. Rosenhahn, and H.-P. Seidel. Robust pose estimation with 3D textured models. In *IEEE Pacific-Rim Symposium on Image and Video Technology*, volume 4319 of *LNCS*, pages 84–95, 2006.

[16] D. Gavrila and L. Davis. 3D model based tracking of humans in action: a multiview approach. In *ARPA Image Understanding Workshop*, pages 73–80, Palm Springs, 1996.

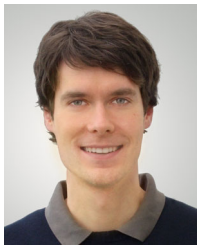
[17] W. E. L. Grimson. *Object Recognition by Computer*. MIT Press, Cambridge, MA, 1990.

[18] D. Lowe. Solving for the parameters of object models from image descriptions. In *Proc. ARPA Image Understanding Workshop*, pages 121–127, 1980.

[19] D. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987.
[20] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
[21] E. Marchand, P. Bouthemy, and F. Chaumette. A 2D-3D model-based approach to real-time visual tracking. *Image and Vision Computing*, 19(13):941–955, Nov. 2001.
[22] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.
[23] E. Mémin and P. Pérez. Robust discontinuity-preserving model for estimating optical flow. In *Proc. 13th International Conference on Pattern Recognition*, pages 920–924, Vienna, Austria, Aug. 1996.
[24] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.
[25] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, 2006.
[26] G. Mori and J. Malik. Recovering 3D human body configurations using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1052–1062, 2006.
[27] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42:577–685, 1989.
[28] R. M. Murray, Z. Li, and S. S. Sastry. *Mathematical Introduction to Robotic Manipulation*. CRC Press, Baton Rouge, 1994.
[29] M. Özuysal, V. Lepetit, F. Fleuret, and P. Fua. Feature harvesting for tracking-by-detection. In *Proc. European Conference on Computer Vision*, volume 3953 of *LNCS*, pages 592–605. Springer, Graz, Austria, 2006.
[30] N. Paragios and R. Deriche. Geodesic active regions: A new paradigm to deal with frame partition problems in computer vision. *Journal of Visual Communication and Image Representation*, 13(1/2):249–268, 2002.
[31] D. Ramanan, D. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):65–81, 2007.
[32] L. G. Roberts. Machine perception of three-dimensional solids. In J. Tippet, editor, *Optical and Electro-optical Information Processing*, pages 159–197. MIT Press, Cambridge, MA, 1966.
[33] B. Rosenhahn, T. Brox, D. Cremers, and H.-P. Seidel. A comparison of shape matching methods for contour based pose estimation. In R. Reulke, U. Eckhardt, B. Flach, U. Knauer, and K. Polthier, editors, *Proc. International Workshop on Combinatorial Image Analysis*, volume 4040 of *LNCS*, pages 263–276, Berlin, Germany, June 2006. Springer.
[34] B. Rosenhahn, T. Brox, and J. Weickert. Three-dimensional shape knowledge for joint image segmentation and pose tracking. *International Journal of Computer Vision*, 73(3):243–262, July 2007.
[35] F. Shevlin. Analysis of orientation problems using Plücker lines. In *International Conference on Pattern Recognition (ICPR)*, volume 1, pages 685–689, Brisbane, 1998.
[36] J. Shi and C. Tomasi. Good features to track. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.
[37] H. Sidenbladh and M. J. Black. Learning the statistics of people in images and video. *International Journal of Computer Vision*, 54:183–209, 2003.
[38] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In D. Vernon, editor, *Proc. European Conference on Computer Vision*, volume 1843 of *LNCS*, pages 702–718. Springer, 2000.
[39] H. Sidenbladh, M. J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proc. European Conference on Computer Vision*, volume 2353 of *LNCS*, pages 784–800. Springer, 2002.
[40] L. Sigal, B. Sidharth, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *Proc. 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 421–428, June 2004.
[41] C. Sminchisescu and A. Jepson. Generative modeling for continuous non-linearly embedded visual inference. In *Proc. International Conference on Machine Learning*, 2004.
[42] G. Sommer, editor. *Geometric Computing with Clifford Algebra: Theoretical Foundations and Applications in Computer Vision and Robotics*. Springer, Berlin, 2001.
[43] M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. In *Proc. 9th International Workshop on Computer Vision Systems*, volume 2095 of *LNCS*, pages 93–106, 2001.
[44] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 238–245. IEEE Computer Society Press, 2006.
[45] L. Vacchetti, V. Lepetit, and P. Fua. Combining edge and texture information for real-time accurate 3D camera tracking. In *3rd International Symposium on Mixed and Augmented Reality*, pages 48–57, 2004.
[46] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *Pattern Recognition - Proc. DAGM*, volume 4713 of *LNCS*, pages 214–223. Springer, 2007.
[47] X. S. Zhou, D. Comaniciu, and A. Gupta. An information fusion framework for robust shape tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):115–129, 2005.

Thomas Brox received the Ph.D. degree in computer science from the Saarland University, Germany in 2005. Subsequently, he spent two years as a postdoctoral researcher at the University of Bonn, Germany, and one year as a temporary faculty member at the University of Dresden, Germany. He is currently a postdoctoral researcher at U.C. Berkeley. His research interests include image segmentation, motion estimation, texture analysis, pose estimation, and detection. Dr. Brox received the Longuet-Higgins Best Paper Award at ECCV 2004.

Bodo Rosenhahn gained his Ph.D. 2003 at the University of Kiel, Germany. From 2003-2005 he was (DFG) PostDoc at the University of Auckland, New Zealand. From 2005 to 2008 he worked as senior researcher at the Max-Planck Insitute for Computer Science in Saarbrücken, Germany. Since 2008 he is a full professor at the Leibniz-University of Hannover. His research focus is on markerless motion capture, human model generation and animation, pose estimation, and image segmentation. His works received several awards.



Juergen Gall obtained his BSc in mathematics from the University of Wales Swansea in 2004 and his Master's degree in mathematics from the University of Mannheim, Germany, in 2005. Since 2006, he is Ph.D. student in computer science at the Saarland University and the Max-Planck-Institut für Informatik. In 2008, he worked for the Machine Learning and Perception group at Microsoft Research Cambridge. His research interests include textured model based tracking, interacting particle systems, and markerless human motion capture.

Daniel Cremers received Bachelor degrees in Mathematics (1994) and Physics (1994), and a Master's degree in Theoretical Physics (1997) from the University of Heidelberg. In 2002 he obtained a Ph.D. in Computer Science from the University of Mannheim, Germany. Subsequently he spent two years as a postdoctoral researcher at the UCLA and one year as a permanent researcher at Siemens Corporate Research in Princeton, NJ. Since 2005 he heads the Computer Vision group at the University of Bonn, Germany. His publications received several awards, and he has given more than 60 talks and invited speeches.