Probabilistic Object Tracking Using Multiple Features

David Serby, Esther-Koller-Meier, Luc Van Gool
Computer Vision Laboratory (BIWI), ETH Zürich, Switzerland
{dserby,ebmeier,vangool}@vision.ee.ethz.ch

Abstract

We present a generic tracker which can handle a variety of different objects. For this purpose, groups of low-level features like interest points, edges, homogeneous and textured regions, are combined on a flexible and opportunistic basis. They sufficiently characterize an object and allow robust tracking as they are complementary sources of information which describe both the shape and the appearance of an object. These low-level features are integrated into a particle filter framework as this has proven very successful for non-linear and non-Gaussian estimation problems. In this paper we concentrate on rigid objects under affine transformations. Results on real-world scenes demonstrate the performance of the proposed tracker.

1 Introduction

Object tracking in monocular image sequences still suffers from a lack of robustness due to temporary occlusions, objects crossing, changing lighting conditions, specularities and out-of-plane rotations. In general, trackers can be subdivided into two categories. First, there are *generic* trackers which use only a minimum amount of a priori information as e.g. the mean-shift approach by Comaniciu *et al.* [3] and the color-based particle filter developed by Perez *et al.* [9]. Secondly, there are trackers that use a very *specific* model of the object, like e.g. the spline representation of the contour by Isard *et al.* [7, 8]. The goal of this paper is to develop a tracker that is generic in terms of handling different objects, but includes many different features that together build a good representation of the object. By using local and complementary information the robustness against appearance changes and distractors can be greatly increased. As an initialization, the user delineates the object to be tracked. Within this region of interest (ROI) the features are extracted. We focus on low-level features like interest points, edges and color distributions of regions and combine them in a flexible way. Each feature alone has certain drawbacks, for example color distributions are not very robust against occlusion and lighting changes while interest points are not discriminative enough and unstable under unexpected transformations. But these features can be seen to complement each other. Moreover, they can be extracted automatically. We integrate these features into the observation process of a particle filter [7] which has been proven very successful for non-linear and non-Gaussian estimation problems and handles clutter and temporary occlusions well.

The integration of different cues into tracking frameworks has already been discussed in the literature [8, 10, 14, 15]. The authors in [8, 15] integrate multiple features into a particle filter. Isard *et al.* [8] combine color and contour information using importance sampling. Wu *et al.* [15] present an approach to combine visual cues by including them into the state, but then decouple the prediction and observation of the different cues. Triesch *et al.* [14] propose an adaptive scheme that they call “Democratic Integration”, for incorporating visual cues. The cues agree on a result which serves as the basis for the adaptation of the individual cues. Finally, Rasmussen *et al.* [10] define a target as a conjunction of parts, and introduce a constrained joint likelihood filter as a data association method to generate the measurement for a Kalman filter.

To the best of our knowledge, the proposed tracker has a unique combination of properties. It automatically combines multiple, low-level cues as equivalent partners in the tracking process. As the tracking performance does not depend on a single cue, the system can behave opportunistically. It has a good chance to find some features that allow to track an object robustly. The tracker is simultaneously tolerant to a wide gamut of object and camera motions and allows to cope with affine deformations. Also, a particle filter is used, with the resulting broad scope of allowed motions and noise models. The features all contribute to a joint likelihood of the object state given the observations.

2 Feature Extraction

From the whole wealth of possible features we restrict ourselves to local, low-level features. They are extracted from the image in a fast bottom-up way without any a priori knowledge besides the object outline / region of interest that the user (or an object detector) indicates in the first frame. The locality of the features helps to overcome problems of

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

partial occlusion and image clutter. A variety of different feature types (points, edges, regions) are considered which are described in the sequel. An example output of the feature extraction is shown in figure 1.

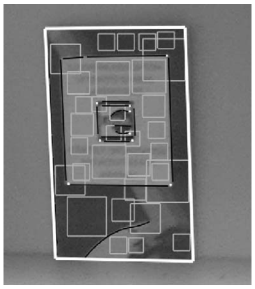


Figure 1. Features extracted in the white framed region of interest. White dots indicate the interest points, the edges are shown in black and the regions in gray.

Interest Points Robust tracking requires first and foremost that the interest points are repeatable. In their comparison Schmid *et al.* [11] concluded that the Harris detector has the highest repeatability rate. The "good features to track" of Shi and Tomasi [12] are, similarly to the Harris detector, based on the second moment matrix. Another important criterion is speed. Here, the SUSAN[1] detector by Smith *et al.* [13] is especially appealing. It is on average ten times faster than the Harris detector. Other advantages of the SUSAN detector are its better localization accuracy and the fact that fewer parameters are involved. However, our experiments on real data have shown that the SUSAN detector is less repeatable, but this can be remedied by renouncing on the contiguity test that is used to suppress false positives. Without this test, the SUSAN detector is also very repeatable, the disadvantage being that quite a lot of false positives are detected, particularly at strong edges. In order to get the best of both worlds, we have combined the two detectors. SUSAN (without contiguity test) is used to quickly extract candidate points. To suppress false positives, the Harris response is computed only at the candidate points which strongly reduces computational complexity. Finally, a non-maximum suppression is applied to find the local maxima; these are the interest points. As we deal with color images, this additional information should not be neglected. The use of color information (RGB) in the context of the Harris detector has been reported to improve the repeatability of the detector [4]. The SUSAN detector can also be straight-forwardly generalized to color images. Again, there is an improvement in comparison with the gray-level version with respect to the repeatability.

Straight and Curved Edges For edge extraction we use the well-known Canny edge detector enhanced for color images by replacing the norm of the intensity gradient from the single-band case with the maximum of the norm of the gradients in the R-, G-, and B-channel. This way some more edges than in the single-band case are detected. After edge detection and edge linking, the edges are broken at points of high curvature and straight line segments are fitted to those segments; these are the straight edges. Then remaining short broken edges are linked together and approximated by B-splines; these are the curved edges.

Textured Regions A simple and efficient way to extract textured regions is to detect interest points that have an associated scale which fixes the size of a square region around the point. We use the approach introduced by Baumberg [1], extended to make use of color information. First, the color Harris detector [4] is run at multiple integration scales using a geometric progression of fixed scale settings. Then the features are ordered based on the scale-normalized Harris response and stored in a list. Finally, starting from the region with the strongest response, every region in the list is checked, whether it overlaps with any of the previous regions and if yes, it is abandoned. This way the regions are spread over the object and their number can be kept at a computationally tractable level. The photometrical information content of the textured regions is represented by a weighted color histogram, as in [3, 9].

Homogeneous Regions The idea to extract the homogeneous regions is very similar to the approach used for the textured regions. First, at different scales the average of the variances in the three color channels is computed over a local window with a size given by the current scale. Then, the features are ordered based on the scale-normalized variance and stored in a list. Finally, starting from the region with the lowest response, every region in the list is checked, whether it overlaps with any of the previous regions and if yes, it is abandoned. Again, a limited number of regions is created, distributed over the object. The photometrical information is given by the average color of the region. Again we work in the HSV space to be more invariant against illumination changes. The saturation and the brightness value of each pixel of such a region are compared against two thresholds 0.1 and 0.2, respectively, to decide if the region is a "hue region" or a "value region". If more than 3/4 of the pixels have a saturation and a brightness bigger than the thresholds, the region is taken as a "hue region". In the converse case, the region is considered as a "value region", and if less than 3/4 of the pixels are of one kind, then the region is abandoned.

3 Tracking

Particle filters [7] offer a probabilistic framework for recursive dynamic state estimation. They are based on a sampling approach, where the posterior density function is approximated by a weighted particle set $\{(\mathbf{x}_t^{(n)}, \pi_t^{(n)})\}_{n=1}^{N}$. Each particle $\mathbf{x}_t^{(n)}$ represents one hypothetical state of the object with a corresponding dis-

[1] **S**mallest **U**nivalue **S**egment **A**ssimilating **N**ucleus

crete sampling probability $\pi_t^{(n)}$. In our case, an object configuration vector $\mathbf{x}_t^{(n)}$ is denoted as $\mathbf{x}_t^{(n)} = \{x^{(n)}, y^{(n)}, v_x^{(n)}, v_y^{(n)}, s_x^{(n)}, s_y^{(n)}, \theta^{(n)}, h^{(n)}\}$, where the first four values specify the kinematic parameters, that is the position and the velocity of the modeled object. The remaining values indicate the affine transformation parameters which describe the scaling in x- and y-direction, orientation and shearing. The dynamics are currently represented as a first order model. Because the low-level features $(\mathbf{x}_t^1, \ldots, \mathbf{x}_t^M)$ are rigidly connected to each other, the object state is completely described by mapping the features according to the transformation parameters, under the assumption that they are coplanar. The samples are weighted using the multi-feature observation density described next. Assuming conditional independence of the observations from the different cues given $\mathbf{x}_t$ [5, 10], the multi-feature observation density $p(\mathbf{z}_t|\mathbf{x}_t)$ is the product of the individual likelihoods of the different features described below:

$$\begin{aligned} p(\mathbf{z}_t|\mathbf{x}_t) &= \prod_{i=1}^{I} p_I(\mathbf{z}_t|\mathbf{x}_t^i) \prod_{j=1}^{J} p_E(\mathbf{z}_t|\mathbf{x}_t^j) \\ & \prod_{k=1}^{K} p_T(\mathbf{z}_t|\mathbf{x}_t^k) \prod_{l=1}^{L} p_H(\mathbf{z}_t|\mathbf{x}_t^l) \qquad (1) \end{aligned}$$

where I is the number of interest points, J the number of edges, K the number of textured regions, L the number of homogeneous regions, and $M = I + J + K + L$ is the total number of features in the model.

Interest Point Observation Density First, centered at the hypothesized interest point $\mathbf{x}_t^i$, a square-shaped search region is defined. Then, the response of the combined detector of section 2 is computed at each pixel within the search region. Writing d_{min} for the distance from the center of the search region to the closest above-threshold interest point, the observation likelihood $p_I(\mathbf{z}_t|\mathbf{x}_t^i)$ for the interest point $\mathbf{x}_t^i$ is given by $p_I(\mathbf{z}_t|\mathbf{x}_t^i) \propto exp\left(-\frac{(d_{min}^i)^2}{2\sigma_I^2}\right)$. In order to prevent an unreliable measurement (e.g. due to occlusion or noise) from having to much influence on the joint likelihood we introduce a residual probability for it. The same approach is also used for the other features types dscribed below. Note that the high number of observation density evaluations per frame requires a fast interest point detector. This backs up our choice from section 2.

Straight / Curved Edge Observation Density First, at R regularly space points along the hypothesized edge $\mathbf{x}_t^j$, line segments normal to the edge are cast into the image. Next, a one-dimensional edge detector is applied to the image intensity along each of these R measurement lines. Writing $\nu_{min}(r)$ for the distance on the rth measurement line from the normal base point to the nearest above-threshold edge, the observation likelihood for the edge $\mathbf{x}_t^j$ becomes $p_E(\mathbf{z}_t|\mathbf{x}_t^j) \propto \prod_{r=1}^{R} exp\left(-\frac{(\nu_{min}^j(r))^2}{2\sigma_E^2}\right)$.

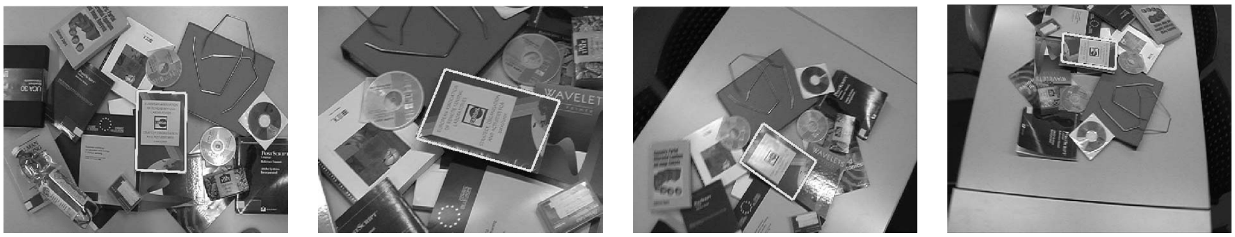
Textured Region Observation Density First, the HV color histogram for the hypothesized region $p(\mathbf{x}_t^k)$ is computed, as described in section 2. In order to compare the histogram of the hypothesized region (i.e. the region according to the affinity of the sample) $p(\mathbf{x}_t^k)$ with the reference histogram q from the initial model, a similarity measure based on the Bhattacharrya coefficient $\rho[p(\mathbf{x}_t^k), q]$ [3] is used to define the observation likelihood for the textured region $\mathbf{x}_t^k$ as follows: $p_T(\mathbf{z}_t|\mathbf{x}_t^k) \propto exp\left(-\frac{(d_B^k)^2}{2\sigma_T^2}\right)$.

Homogeneous Region Observation Density The comparison of a hypothesized homogeneous region $\mathbf{x}_t^l$ with a reference homogeneous region is done by computing the difference of their average colors. Attention has to be payed to the case of "hue" regions because the hue channel is an angular representation. In this case, the distance between the two average colors H_1 and H_2 is given by $d_{hue} = sin\left(\frac{(H_1 - H_2)}{2}\right)$. Then, the observation likelihood for the homogeneous region $\mathbf{x}_t^l$ can be written as: $p_H(\mathbf{z}_t|\mathbf{x}_t^k) \propto exp\left(-\frac{(d_{hue/value}^k)^2}{2\sigma_H^2}\right)$.

4 Results

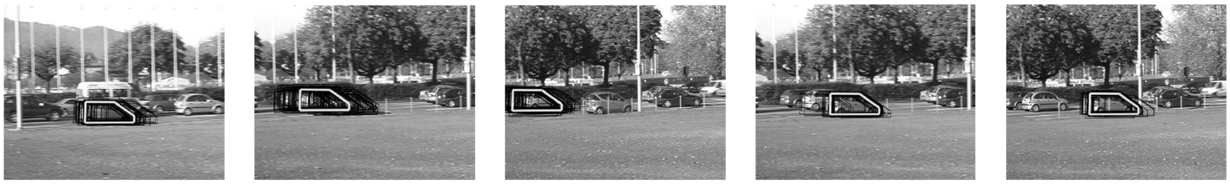
This section shows two sequences which demand the use of multiple features in order to robustly and precisely track the objects in question. No specific assumptions about the object and camera motion are made. All parameters, except the ones for the dynamical model and the number of samples, have been fixed to the same value for all experiments.

Figure 2 shows the tracking of a book that undergoes large scale changes and out-of-plane rotations as well as in-plane rotation and translation. The contour of the book alone, as used in [7], did not suffice for robust tracking under the rather abrupt affine transformations and in the presence of similarly shaped distractors. Also, if the tracking is solely based on color cues, as in [9], it fails due to the strong lighting changes like the specular reflections in frame 270. However, the use of multiple, complementary features (see fig. 1) allows to robustly and precisely track the book throughout the sequence. The next example sequence, shown in figure 3, shows the capability of the proposed tracker to handle very similar objects. The silver-gray car crosses a similar vehicle. The first three frames show the failure of a tracker based on a colour histogram alone [9]. After the crossing in frame 34 the tracker switches from the car in front to the similar car on the second lane, because during several frames the second car matches the color histogram better. In contrast, as can be seen from the fourth and fifth frames, the multi-cue tracker is able to distinguish between the two cars due to its more discriminative object model that makes also use of shape information. Although the system, in general, is able to robustly track an object using the complete object model, there are situations where



frame 1 frame 200 frame 270 frame 410

Figure 2. The *book* sequence shows the robustness of the proposed tracker against large affine shape changes and specular reflections. The thick contour indicates the object's mean state.



frame 1 frame 34 frame 54 frame 34 frame 54

Figure 3. 1,2,3: Simple color histogram tracker; 4,5: New multi-feature tracker. The thick contour is the object's mean state and the thin ones are the samples approximating the posterior distribution.

some of the automatically extracted features act as distractors rather than to support the system. This happens, for instance, with texture regions over the windows of cars. As the background changes all the time, they do not offer reliable information. Therefore, by making use of some prior knowledge the user can select the most promising features from the automatically extracted ones in order to obtain an even better object model. Furthermore, by throwing away the distracting features the speed performance of the tracker is improved. Work is underway to have the tracker decide for itself which features to give preference.

5 Conclusions and Future Work

The proposed tracking method adds the robustness of opportunistic low-level features to that of particle filtering. The system extracts the different cues automatically and combines them in a flexible manner to build a characteristic representation of an object. The multi-cue tracker is applicable in many areas. Moreover, it can handle affine transformations. Our research interests now focus on adaptive models which also exploit spatial configurations of features. The photometrical description of the regions could be adapted to also deal with heavily changing lighting conditions. Robustness against partial occlusions can be improved by devising a system that is able to detect and temporarily switch off occluded features.

Support by SNF NCCR IM2 is gratefully acknowledged.

References

[1] A. Baumberg, *Reliable feature matching across widely separated views*, CVPR, pp. 774–781, 2000.

[2] J. Canny, *A computational approach to edge detection*, PAMI, 8(6):679–698, 1986.

[3] D. Comaniciu, V. Ramesh, and P. Meer, *Real-time tracking of non-rigid objects using mean shift*, CVPR, pp. 142–149, 2000.

[4] V. Gouet, P. Montesinos, R. Deriche, and D. Pele, *Evaluation de detecteurs de points d'interet pour la couleur*, Reconnaissance des formes et Intell. Artificielle, pp. 257–266, 2000.

[5] E. Hayman, and J. Eklundh, *Figure-ground segmentation of image sequences from multiple cues*, ECCV, pp. 661–675, 2002.

[6] C. G. Harris, and M. J. Stephens, *A combined corner and edge detector*, Alvey Vision Conference, pp. 147–151, 1988.

[7] M. Isard, and A. Blake., *Condensation – conditional density propagation for visual tracking*, IJCV, 29(1):5–28, 1998.

[8] M. Isard, and A. Blake, *ICondensation: Unifying low-level and high-level tracking in a stochastic framework*, ECCV, pp. 893–908, 1998.

[9] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, *Color-based probabilistic tracking*, ECCV, pp. 661–675, 2002.

[10] C. Rasmussen, and G. D. Hager, *Probabilistic data association methods for tracking complex visual objects*, PAMI, 23(6):560–576, 2001.

[11] C. Schmid, R. Mohr and Christian Bauckhage, *Evaluation of interest point detectors*, IJCV, 37(2):151–172, 2000.

[12] J. Shi, and C. Tomasi, *Good features to track*, CVPR, pp. 593–600, 1994.

[13] S. M. Smith and J. M. Brady, *SUSAN - a new approach to low level image processing*, IJCV, 23(1):45–78, 1997.

[14] J. Triesch, and C. von der Malsburg, *Self-organized integration of adaptive visual cues for face tracking*, IEEE Conf. on Automatic Face and Gesture Recognition, pp. 102–107, 2000.

[15] Y. Wu, and T. S. Huang, *A co-inference approach to robust visual tracking*, ICCV, pp. 26–33, 2001.