Bayesian Kernel Tracking

Dorin Comaniciu

Real-Time Vision and Modeling Department
Siemens Corporate Research
755 College Road East, Princeton, NJ 08540, USA
comanici@scr.siemens.com

Abstract. We present a Bayesian approach to real-time object tracking using nonparametric density estimation. The target model and candidates are represented by probability densities in the joint spatial-intensity domain. The new location and appearance of the target are jointly derived by computing the maximum likelihood estimate of the parameter vector that characterizes the transformation from the candidate to the model. This probabilistic formulation accommodates variations in the target appearance, while being robust to outliers represented by partial occlusions. In this paper we analyze the simplest parameterization represented by translation in both domains and present a gradient-based iterative solution. Various tracking sequences demonstrate the superior behavior of the method.

Keywords: Real-Time Tracking; Maximum Likelihood Parameter Estimation; Joint Domain Density; Appearance Change.

1 Introduction

Visual object tracking is a task required by various applications such as perceptual user interfaces [4], surveillance [6], augmented reality [11], smart rooms [18], intelligent video compression [5], and driver assistance [13, 1]. In the general case, a visual tracker involves both *bottom-up* and *top-down* components. The former are represented by the target representation and localization, appearance change, and measurement model, while that latter regard the object dynamics, learning of scene priors, and hypothesis testing and verification.

Common techniques to model the target dynamics are the (extended) Kalman filter [2] and particle filters [15, 10]. The problem of target representation and localization is related to registration techniques [23, 20, 19]. The difference is that tracking assumes small changes in the location and appearance of the target in two consecutive frames. This property can be exploited to develop efficient, gradient based localization schemes, using the normalized correlation criterion [3]. Since the correlation is sensitive to illumination, Hager and Belhumeur [12] model explicitly geometry and illumination changes. The method is robustified by Sclaroff and Isidoro [21] by using M-estimators [14]. Learning of appearance models is discussed in [16] by employing a mixture of stable image structure, motion information and an outlier process. To efficiently accommodate non-rigid transformations, Comaniciu *et al.* [8] develop histogram-based tracking. Spatial gradient optimization becomes possible due to a spatially-smooth cost function produced by masking the target with an isotropic kernel.

In this paper we present a new approach for target representation and localization. Our motivation is to develop a framework that is optimal, efficient, robust to outliers, and that can be easily customized. We formulate the problem of *target localization* as a *classification* problem. Assuming that the probability density of the target model is known, we search for target candidates whose probability density under a parameterized transformation matches the density of the target. The matching criterion is derived by minimizing the probability of error in choosing the wrong candidate. Bayesian statistics are used to obtain the maximum likelihood estimates of the best target candidate and parameter vector. As another novelty, the probability densities characterizing the target and candidates are estimated in the *joint spatial-intensity* domain. This implies that location and target appearance are optimized simultaneously.

The organization of the paper is as follows. Section 2 describes the Bayesian alignment of probability densities under a general parameterized transformation. An explicit solution for the translation case is derived in Section 3. Section 4 formulates the density estimation in the spatial-intensity domain. Tracking experiments on different sequences are presented in Section 5.

2 Bayesian Alignment of Densities

Assume that the target model is specified by the d-dimensional sample $Q = \{\mathbf{x}_r, r = 1 \ldots N\}$ drawn i.i.d. from the probability density q. We hypothesize the existence of U target candidates generated by transforming a random variable $\mathbf{X}$ of density p under the parameterized transformation $T(\mathbf{X}; \boldsymbol{\theta}_u)$ where $u = 1 \ldots U$ and $\boldsymbol{\theta}_u$ is the parameter vector. In other words, starting from the sample $\{\mathbf{x}_i, i = 1 \ldots n\}$ drawn from p we obtain samples of the form $\{T(\mathbf{x}_i; \boldsymbol{\theta}_u), i = 1 \ldots n\}$, characterized by the density p_u, with $u = 1 \ldots \mathbf{U}$.

We want to determine a parameter vector $\boldsymbol{\theta}_v$ with $1 \leq v \leq U$ such that the probability that the model sample $Q = \{\mathbf{x}_r, r = 1 \ldots N\}$ and the candidate sample $\{T(\mathbf{x}_i; \boldsymbol{\theta}_v), i = 1 \ldots n\}$ belong to the same density source (i.e., $q = p_v$) is maximized (or, equivalently, the probability of error is minimized). This can be written as

$$v = arg\overset{u}{m}ax\ P(q = p_u|Q) = arg\overset{u}{m}ax\ P(Q|q = p_u)P(p_u) \tag{1}$$

where the last equality is obtain by applying the Bayes rule. The term $P(p_u)$ represents a priori information on the presence of candidate u. Depending on the tracking formulation $P(p_u)$ is obtained either by learning the motion dynamics or/and the appearance changes. This is a natural way to integrate priors on motion and appearance.

At this moment, we consider all the hypotheses equally probable, which is equivalent to maximizing the likelihood $P(Q|q = p_u)$. By taking into account that Q is drawn i.i.d. and applying the log function, it results that

$$v = \overset{u}{argmax}\ L_u = \overset{u}{argmax} \sum_{r=1}^{N} \log p_u(\mathbf{x}_r) \tag{2}$$

where L_u is the log-likelihood. [1] The kernel estimate of the density p_u computed at location $\mathbf{x}$ is given by

$$p_u(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - T(\mathbf{x}_i; \boldsymbol{\theta}_u)}{h}\right) \tag{3}$$

where h is the bandwidth of kernel K [22]. Hence, the log-likelihood is expressed by

$$L_u = \sum_{r=1}^{N} \log \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_r - T(\mathbf{x}_i; \boldsymbol{\theta}_u)}{h}\right) \tag{4}$$

The best target candidate is obtained by maximizing expression (4) as a function of $\boldsymbol{\theta}_u$. Note the optimality of the above formulation.

3 Translation Case

The transformation T is application dependent and related to the expected transformation of the target during tracking. We will show in this section how to maximize the log-likelihood for the translation case. The following computations are similar in strategy to those shown Section 4.2 of [8]. However, their significance is different: we deal here with the maximum likelihood alignment of densities, while in [8] the task was to maximize the Bhattacharyya coefficient between histograms.

The transformation $T(\mathbf{x}_i; \boldsymbol{\theta}_u)$ is replaced by $(\mathbf{x}_i - \mathbf{y})$ and the density p_u is now denoted by $p(\mathbf{x} + \mathbf{y})$, being expressed by

$$p(\mathbf{x} + \mathbf{y}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} + \mathbf{y} - \mathbf{x}_i}{h}\right) \tag{5}$$

For the convenience of notation, we introduce the profile of the kernel K as the function $k : [0, \infty) \rightarrow R$ such that $K(\mathbf{x}) = k(\|\mathbf{x}\|^2)$. Employing the profile notation we have

$$p(\mathbf{x} + \mathbf{y}) = \frac{1}{nh^d} \sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x} + \mathbf{y} - \mathbf{x}_i}{h}\right\|^2\right) \tag{6}$$

[1] By applying the law of large numbers [9, p.286] it can be easily shown that condition (2) is equivalent to minimizing the Kullback-Leibler distance $D(q||p_u)$. This is not required by our derivation. Note, however, that the other divergence, $D(p_u||q)$, is not appropriate for the task.

while log-likelihood (4) becomes

$$L_{\mathbf{y}} = \sum_{r=1}^{N} \log p(\mathbf{x}_r + \mathbf{y}) = \sum_{r=1}^{N} \log \frac{1}{nh^d} \sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x}_r + \mathbf{y} - \mathbf{x}_i}{h}\right\|^2\right) \tag{7}$$

Assume that the optimization is started with an initial value $\mathbf{y} = \mathbf{y}_0$. Using Taylor expansion around the values $p(\mathbf{x}_r + \mathbf{y}_0)$ the log-likelihood is approximated as

$$L_{\mathbf{y}} \approx \sum_{r=1}^{N} \log p(\mathbf{x}_r + \mathbf{y}_0) - N + \sum_{r=1}^{N} \frac{1}{p(\mathbf{x}_r + \mathbf{y}_0)} \frac{1}{nh^d} \sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x}_r + \mathbf{y} - \mathbf{x}_i}{h}\right\|^2\right) \tag{8}$$

The last term in (8) represents a weighted sum of density estimates computed at locations $\mathbf{x}_r + \mathbf{y}$. It is natural to employ the mean shift procedure [7] to maximize this term. By taking the gradient of this term with respect to $\mathbf{y}$, after some algebra the new value of $\mathbf{y}$ is obtained as

$$\mathbf{y}_1 = \frac{\sum_{r=1}^{N} \frac{1}{p(\mathbf{x}_r + \mathbf{y}_0)} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}_r) g\left(\left\|\frac{\mathbf{x}_r + \mathbf{y}_0 - \mathbf{x}_i}{h}\right\|^2\right)}{\sum_{r=1}^{N} \frac{1}{p(\mathbf{x}_r + \mathbf{y}_0)} \sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}_r + \mathbf{y}_0 - \mathbf{x}_i}{h}\right\|^2\right)} \tag{9}$$

where $g(x) = -k'(x)$ for the definition domain. At $\mathbf{y}_1$ the log-likelihood function is larger than that at $\mathbf{y}_0$. Expression (9) is computed iteratively until convergence. The new position is determined by weighted sums of local point differences. Since the measurements are local (due to the kernel weighting), the algorithm is robust to outliers in the data.

A more intuitive expression is obtained by replacing k and g by the normal profile and its derivative. The normal profile is

$$k(x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}x\right) . \tag{10}$$

hence, $g(x) = k(x)/2$. Since k is identical to g up to a constant, expression (9) simplifies to

$$\mathbf{y}_1 = \frac{1}{N} \sum_{r=1}^{N} \frac{\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}_r) k\left(\left\|\frac{\mathbf{x}_r + \mathbf{y}_0 - \mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x}_r + \mathbf{y}_0 - \mathbf{x}_i}{h}\right\|^2\right)} \tag{11}$$

This shows that the log-likelihood is maximized by computing weighted sums of local differences.

4 Density Estimation in the Joint Domain

The idea of density estimation in the joint domain is detailed in [7]. Each image pixel $\mathbf{z}$ is characterized by a location $\mathbf{x} = (x_1, x_2)^\top$ and a range vector $\mathbf{c}$.

The range vector is one dimensional in the case of gray level images or three dimensional in the case of color images. In other words, an input image of n pixels is represented as a collection of d-dimensional points $\mathbf{z}_i = (\mathbf{x}_i^\top, \mathbf{c}_i^\top)^\top$ with $i = 1 \dots n$. The space constructed as above is called the *joint spatial-intensity* domain or spatial-color domain. The concept can be extended by adding a temporal component. To estimate the probability density in the joint space we use a product kernel with bandwidth σ_s for the spatial components and σ_r for the range.[2]

Due to the use of product kernel, different transformations can be accommodated in the two spaces. For example one can define an affine transformation in the spatial domain and a translation in the range. The optimization, however, is performed jointly for both spaces, i.e., for both location and appearance.

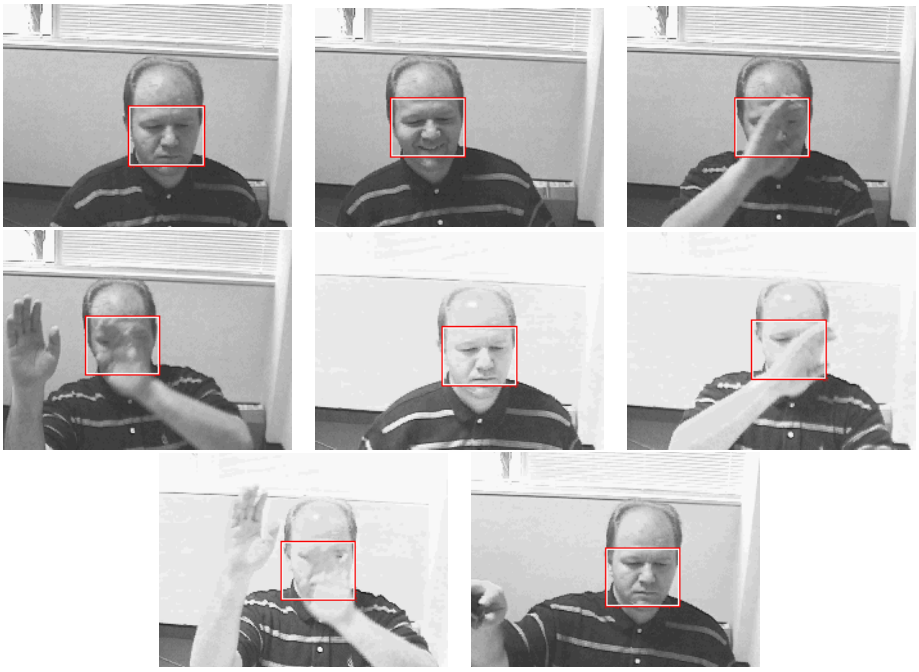

Fig. 1. *Face* sequence. 750 frames.

5 Experiments

We tested the new tracking framework for various sequences and the results are very promising. Although only translation in the joint domain (spatial/intensity

[2] The normal kernel is separable, so the idea of product kernel is implicit when a normal kernel is used.

or spatial/color) has been considered, the algorithm proved to be robust to illumination variations and high percentage of occlusions. For all the sequences we used $\sigma_s = 3$ and $\sigma_r = 20$. A three level pyramid was used for efficient implementation of the optimization. The tracker runs in real-time (30fps) on a 1GHz PC.

A gray level tracking sequence is shown in Figure 1. The optimization is performed in a three dimensional space (two spatial dimensions and one intensity dimension). The model is captured in the first frame. We tested the behavior of the algorithm with respect to outliers generated by hand occlusion. As one can see, a large amount of occlusion is tolerated. We also tested the adaptation of the algorithm to illumination changes. The changes were induced by applying the back-light correction of the camera. The model adapted gracefully to the new condition while the tracking continued unperturbed. We again tested the robustness to outliers within the new conditions. Finally, the back-light correction was stopped determining the model to adapt again.

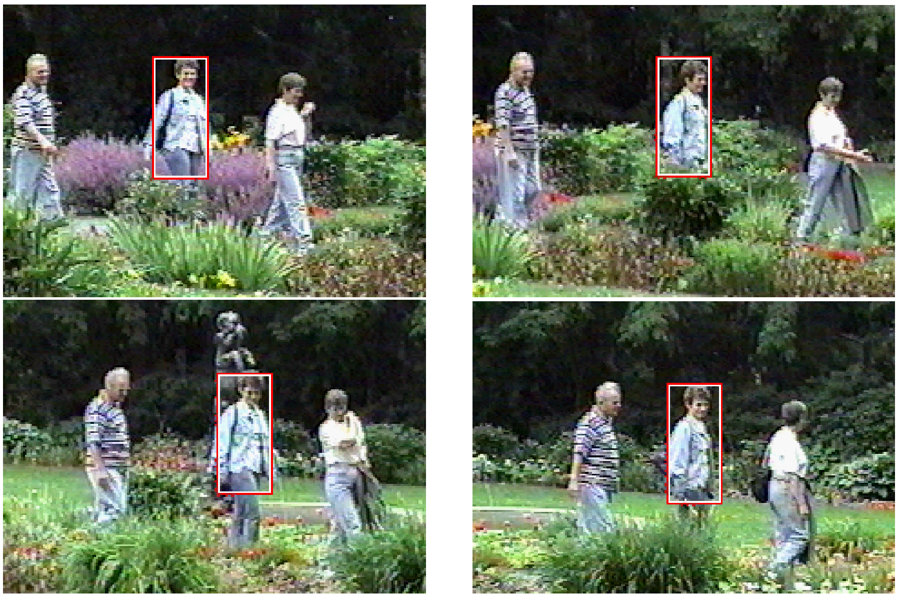

Fig. 2. *Walking* sequence. 540 frames.

We also tested two color sequences with natural illumination. The optimization is performed in a five dimensional space (two spatial dimensions and three color dimensions). In the first sequence we track a person walking in a garden (Figure 2). Partial occlusion is present from various flowers. In the second sequence we track a car at the exit from the tunnel (Figure 3). The camera gain adapts due to the change in illumination. In both sequences the tracker adapted correctly.

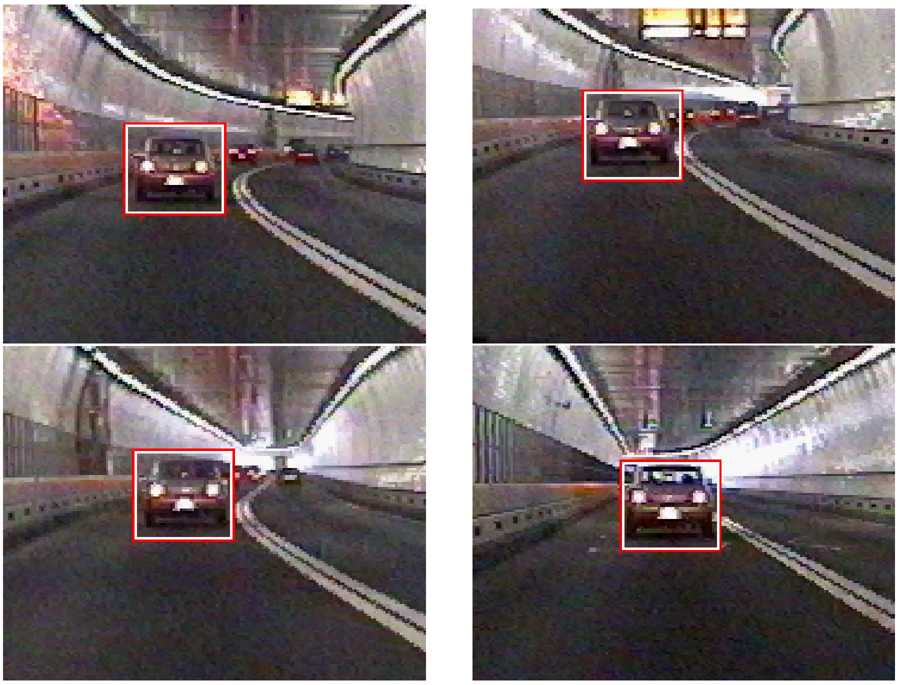

Fig. 3. *Pursuit* sequence. 300 frames.

6 Discussion

This paper presented a Bayesian approach to real-time tracking. Using a new formulation of the target representation and localization problem, we have developed a tracking framework that is both efficient and effective. It can naturally tolerate outliers and changes in the illumination. Results regarding optimization for other type of transformations such as similarity, affine and homography, will be reported in subsequent papers. Techniques that gradually introduce more complex transformation models can be employed [17].

Acknowledgments

I thank Visvanathan Ramesh and Huseyin Tek from Siemens Corporate Research for stimulating discussions on the subject.

References

1. S. Avidan. Support vector tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Kauai, Hawaii, volume I, pages 184–191, 2001.
2. Y. Bar-Shalom and T. Fortmann. *Tracking and Data Association.* Academic Press, 1988.

3. B. Bascle and R. Deriche. Region tracking through image sequences. In *Proc. 5th Intl. Conf. on Computer Vision,* Cambridge, MA, pages 302–307, 1995.
4. G. R. Bradski. Computer vision face tracking as a component of a perceptual user interface. In *Proc. IEEE Workshop on Applications of Computer Vision,* Princeton, NJ, pages 214–219, October 1998.
5. A. D. Bue, D. Comaniciu, V. Ramesh, and C. Regazzoni. Smart cameras with real-time video object generation. In *Proc. IEEE Intl. Conf. on Image Processing,* Rochester, NY, page to appear, 2002.
6. R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89(10):1456–1477, 2001.
7. D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(5):603–619, 2002.
8. D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Hilton Head, SC, volume II, pages 142–149, June 2000.
9. T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.
10. A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.
11. V. Ferrari, T. Tuytelaars, and L. V. Gool. Real-time affine region tracking and coplanar grouping. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Kauai, Hawaii, volume II, pages 226–233, 2001.
12. G. Hager and P. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* San Francisco, CA, pages 403–410, 1996.
13. U. Handmann, T. Kalinke, C. Tzomakas, M. Werner, and W. von Seelen. Computer vision for driver assistance systems. In *Proceedings SPIE*, volume 3364, pages 136–147, 1998.
14. P. J. Huber. *Robust Statistical Procedures.* SIAM, second edition, 1996.
15. M. Isard and A. Blake. Condensation - Conditional density propagation for visual tracking. *Intl. J. of Computer Vision*, 29(1), 1998.
16. A. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Hawaii, volume I, pages 415–422, 2001.
17. K. Kanatani. Image mosaicing by stratified matching. In *Proc. Statistical Methods in Video Processing Workshop,* Copenhagen, Denmark, 2002.
18. J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer. Multi-camera multi-person tracking for EasyLiving. In *Proc. IEEE Intl. Workshop on Visual Surveillance,* Dublin, Ireland, pages 3–10, 2000.
19. C. Olson. Image registration by aligning entropies. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Kauai, Hawaii, volume II, pages 331–336, 2001.
20. A. Roche, G. Malandain, and N. Ayache. Unifying maximum likelihood approaches in medical image registration. Technical Report 3741, INRIA, 1999.
21. S. Sclaroff and J. Isidoro. Active blobs. In *Proc. 6th Intl. Conf. on Computer Vision,* Bombay, India, pages 1146–1153, 1998.
22. D. W. Scott. *Multivariate Density Estimation.* Wiley, 1992.
23. P. Viola and W. Wells. Alignment by maximization of mutual information. *Intl. J. of Computer Vision*, 24(2):137–154, 1997.