Region Tracking via HMMF in Joint Feature-Spatial Space

Yuan XiaoTong[1] Yang ShuTang[2] Zhu HongWen[2]
1 School of Information Security Engineering, Shanghai Jiao Tong University
2 Department of Electronic Engineering, Shanghai Jiao Tong University
1954 Huashan Road, Shanghai, 200030 China
E-mail: { yxt, guolihua, styang }@sjtu.edu.cn

Abstract

Region-based tracking in a temporal image sequence is described as a segmentation of current frame into a set of non-overlapping regions: the tracking regions and the non-tracking region. The segmentation is viewed to be a Markov labeling process. Based on the key idea of using a doubly stochastic prior model, the optimal estimation for the label field is found by the minimization of a differentiable function. We exploit the feature-spatial probabilistic representation of a region as the conditional distribution in the Bayesian framework, which makes our tracker robust to local deformation and partial occlusion. The continuity of the objective function leads to a much faster numerical implementation. Very promising experimental results on some real-world sequences are presented to illustrate the performance of the presented algorithm.

1. Introduction

In recent years, interest in tracking has increased with advances in computer vision methodology and video processing capabilities. Tracking of regions is often connected with difficult problems such as video surveillance and object-based coding. Generally, the approaches to region tracking can be divided into three dominant classes: approaches based on feature points [1] and edges [2], approaches based on region information [3], and the combination of both contour-based and region-based approaches [4]. Although numerous methods have been developed to date, most of them suffer from severe constrains imposed on the nature of the image sequence. Some assume the additional priori motion model, while others constrain the region tracked have uniform or sharp edge. To make the approach more applicable, Mansouri [5] proposes a tracking algorithm based on region-competition model and solves it via Level Set PDEs. The assumption used in [5] is the continuity of luminance/Chrominance, which is a very basic assumption for tracking problems, hence the avoidance of motion field or motion parameters computation. However, this approach is very sensitive to the initial contour that encloses the region to be tracked. In the motion tracking issues, background pixels may sometimes be mistracked.

In this paper, we further investigate tracking approaches without motion computation and we introduce the hidden Markov random field models to handle both spatial and appearance properties of the tracking and non-tracking regions. The tracked objects are located in each new frame by searching a label field of the pixel lattice that maximizes a post-probability. We exploit the Feature-Spatial distribution of a region representing a non-rigid object as a conditional distribution in the Bayesian framework. Given a sample from a region representing the objects, we estimate the Feature-Spatial joint distribution using kernel density estimation. The hidden measure vector field introduced by the doubly stochastic model permits the characterization of the solution for complex field labeling problem in terms of a differentiable energy function, which can be solved efficiently by Newtonian descent scheme. The introduction of Feature-Spatial distribution and the exploitation of doubly stochastic model make our algorithm distinguish itself from other region-based tracking approaches.

The structure of this paper is organized as follows: In section 2, we formulate the tracking problem in a probabilistic framework. In section 3, we present the joint Feature-spatial representation and the tracking objective function. In Section 4, we discuss the scheme for the minimization of objective function. In section 5, some experimental results and comparisons with other tracking algorithms are provided. Finally, some conclusions are drawn in section 6.

2. Hmmf Tracking Framework

Let (I^k) represents a sequence of images observed from the pixel lattice Ω and indexed by k. Assume that there are $M-1$ tracking regions $\{R_1^n,\ldots,R_{M-1}^n\}$ and one non-tracking region (or background) R_M^n in image I^n,

Proceedings of the IEEE Workshop on Motion and Video Computing (WACV/MOTION'05)
0-7695-2271-8/05 $ 20.00 IEEE

such that $\Omega = \bigcup_{l=1}^{M} R_l^n$; $R_i^n \cap R_j^n = \phi$, $i \neq j$. The tracking of $\{R_1^n, \ldots, R_{M-1}^n\}$ from time interval n to $n+1$ can be formulated as the problem of segmenting image I^{n+1} into $\{R_1^{n+1}, \ldots, R_M^{n+1}\}$, given I^n , I^{n+1} and $\{R_1^n, \ldots, R_M^n\}$. The way to achieve this goal can be naturally regarded as a labeling process. As is, to some extent, similar to the region-based static image segmentation. Some statistical approaches, such as Markov Random Field (MRF) method, have been proposed in the last decades and proved to be powerful for such segmentation problem. Let l be the label field associated with I^{n+1}. In classical MRF model $l(r) \in \{1, \ldots, M\}$, denoting that pixel $r \in \Omega$ belongs to the region $R_{l(r)}^{n+1}$. The estimation of l is always described as a discrete optimization problem, which is normally solved, with very expensive computational cost by SA-like or EM-like algorithms. Recently Marroquin *et al.*[6] presented the HMMF model for image segmentation. HMMF constructs a doubly stochastic model with an additional hidden Markov random measure field. It has achieved great improvement over classical MRF model in both accuracy and computational complexity. In this paper, we adopt HMMF model in the tracking labeling process. The formal description of the model is presented as follows:

Let $\Theta = \{(e_1, \ldots, e_M) \mid \sum_{i=1}^{M} e_i = 1, e_i > 0)\}$, f is the hidden measure vector field associated with the discrete label field l , $f(r) = (f_1^r, \ldots, f_M^r) \in \Theta$ and f_i^r is the probability that pixel r belongs to region R_i^{n+1}. Denote $R^n = (R_1^n, \ldots, R_M^n)$. The goal of labeling is to compute a posterior probability distribution $P(f, \theta \mid I^{n+1}, I^n, R^n)$. Through Bayesian rule, we get

$$P(f,\theta \mid I^{n+1}, I^n, R^n) = \frac{1}{Z} P(I^{n+1} \mid I^n, R^n, f, \theta) P_f(f) P_\theta(\theta)$$

$$= \frac{1}{Z} \prod_{r \in \Omega} P(I^{n+1}(r) \mid I^n, R^n, f, \theta) P_f(f) P_\theta(\theta)$$

$$= \frac{1}{Z} \prod_{r \in \Omega} (\sum_{i=1}^{M} P(I^{n+1}(r) \mid l(r) = i, I^n, R^n, \theta) P(l(r) = i \mid f)) P_f(f) P_\theta(\theta)$$

$$= \frac{1}{Z} \prod_{r \in \Omega} (\sum_{i=1}^{M} P_i(I^{n+1}(r) \mid \theta_i) f(r)) P_f(f) P_\theta(\theta)$$

, $P_i(\bullet \mid \theta_i)$ is the probability distribution function that generate region R_i (region R_i^n may be viewed to be a group of feature samples generated by this distribution) , θ_i is the associated parameters. $P_f(f) = \frac{1}{Z_f} e^{-\sum_C W_C(f)}$. $W_C(f)$ is given potential function and C are the cliques of a given neighborhood system. If we consider cliques C of size 2, a simple quadratic potential is expressed as:

$$W_C(f) = W_{rs}(f(r), f(s)) = \beta \sum_{i=1}^{M} (f_k^r - f_i^s)^2$$

, where $< r, s >$ are neighboring sites in Ω and β is some positive constant. $P_\theta(\theta)$ is the prior distribution of θ, a non-informative (constant) prior may be used if there are no prior constrains on θ. Z and Z_f are normalized constants.

Let's denote $P(f, \theta \mid I^{n+1}, I^n, R^n) = \frac{1}{Z} e^{-U(f,\theta)}$, then optimal estimation of f must minimize the following energy function:

$$U(f,\theta) = -\sum_{r \in \Omega} \log(v(r) \cdot f(r)) + \sum_{C} W_C(f) - \log P_\theta(\theta) \quad (1)$$

, where $v(r) = (v_1^r, \ldots, v_M^r)$, $v_i^r = P_i(I^{n+1}(r) \mid \theta_i)$,

To obtain the optimal estimator l^* for the label field, we use the following procedure:

1. Minimize the $U(f, \theta)$ given by (1), subject to the constrains: $f(r) \in \Theta$, for all $r \in \Omega$;
2. Find the mode for each discrete measure $f^*(r)$ in a decoupled way:

$$l^*(r) = \arg\max_k f_k^*(r)$$

3. APPEARANCE REPRESENTATION

Now we will define the distribution function $P_i(\bullet \mid \theta_i)$. To a given region R_i, we view both the feature and the feature location to be probabilistic random variables . Given the tracked region of R_i^n in the frame I^n, we can model the feature-spatial joint probability distribution of R_i as follows [7]:

$$P_i(x, u \mid (\theta_{i1}, \theta_{i2})) = \frac{1}{N_i} \sum_{k=1}^{N_i} K_{\theta_{k1}}(x - x_{ik}) G_{\theta_{k2}}(u - u_{ik}) \quad (2)$$

, where $\{(x_{ik}, u_{ik})\}_{k=1,\ldots,N_i}$ are the samples that form R_i^n, x is the 2-dimensional location variable, $u(x)$ is the d -dimensional feature vector at location x . $K_{\theta_{k1}}$ is a 2-dimensional kernel with a bandwidth θ_{i1} and $G_{\theta_{k2}}$ is a d -dimensional kernel with a bandwidth θ_{i2}. The bandwidth in the spatial dimensions represents the variability in feature location due to

the local deformation or measurement uncertainty while the bandwidth in the feature dimensions represents the variability in the value of the feature. We absorb the normalization constants into the kernels for convenience.

The outline of our tracking process is shown in fig. 1:

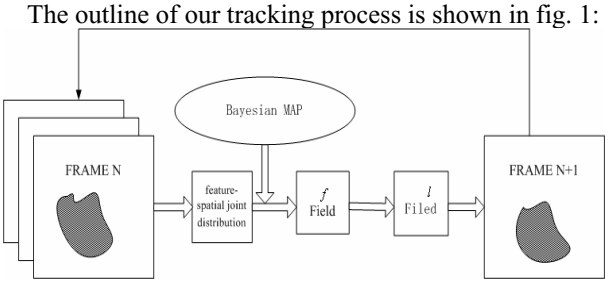


Fig.1 Tracking based on HMMF via Joint Feature-Spatial distribution

4. Energy Minimization Algorithm

4.1 Estimation of The Distribution Functions

From energy function (1), we can see that vector $v(r)$ has to be calculated for every pixel in current frame. Let N be the total number of pixels on lattice Ω, and we have $\sum_{i=1}^{M} N_i = N$. From expression (2) it is clear to see that the total calculation complexity for estimate v is $O(N^2)$. This is extremely high for real-time tracking purpose. We now make a major simplifying assumption which will allow us to approximate the probability distribution $P_i(x,u\,|\,(\theta_{i1},\theta_{i2}))$. We assume that the distribution is highly peaked as a function on x. In other words, we assume that $P_i(x,u\,|\,(\theta_{i1},\theta_{i2}))$ is concentrate on a small neighborhood of x (see Fig. 2), that is:

$$K_{\theta_{k1}}(x-x_{ik}) = \begin{cases} \sup K_{\theta_{k1}} & \|x-x_{ik}\| \le \delta \\ 0 & \|x-x_{ik}\| > \delta \end{cases}$$

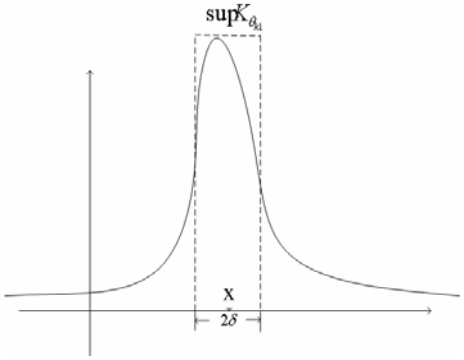


Fig. 2 Approaching kernel function by bound function

Then the joint feature-spatial distribution can be expressed as:

$$P_i(x,u\,|\,(\theta_{i1},\theta_{i2})) = \frac{1}{N_i}\sum_{\|x-x_{ik}\|\le\delta} \sup K_{\theta_{i1}} G_{\theta_{i2}}(u-u_{ik})$$

With such simplification, the computational complexity for v is reduced to $O(N\delta^2)$, and typically we have $\delta << N$. $G_{\theta_{k2}}(\bullet)$ is chosen to be the Gaussian distribution function with zero-mean and variance θ_{i2} in this paper. In natural tracking problems, the regions of the objects are always singly connected and the translation and deformation are small between successive frames. The HMMF model can be just applied on a rectangle area that surrounds the tracked region in current frame, as is illustrated in Fig. 3. If the tracking region is relatively smaller comparing to the whole image, the computational complexity can be further reduced. These rectangles need to be updated for every tacking object in each frame.

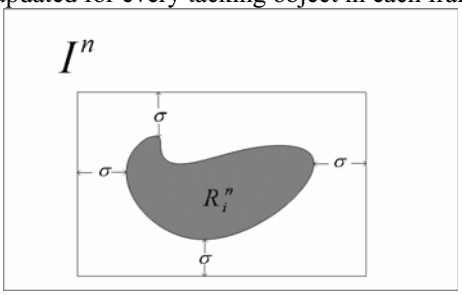


Fig. 3 Tracking region and the local rectangle surrounding it

4.2 Energy Minimization Algorithm

As soon as the vector field v has been determined in the current frame, the minimization of (1) may be effected using any general purpose constrained optimization technique. We have found, however, that due to the simplicity of the constraint to f, by Lagrange multiplier, the optimization goal may be expressed as maximizing following energy function without additional constraints:

$$U(p,\lambda) = -\sum_{r\in\Omega}\log(v(r)\cdot p(r)) + \sum_{C} W_C(p) + \sum_{r\in\Omega}\lambda_r\Big(\sum_{i=1}^{M}(p_i^r)^2 - 1\Big) \quad (3)$$

, where $p_i^r = (f_i^r)^{\frac{1}{2}}$, $i = 1,\ldots,M$, which is introduced to guarantee that the resulting $(f_i^r)^*$ satisfies $(f_i^r)^* > 0$. $\Lambda = (\lambda_r \,|\, r \in \Omega)$ is the set of Lagrange parameters. Here we consider a non-informative prior for θ, thus $\log P_\theta(\theta)$ can be ignored. The maximization of U can be solved effectively by multi-scale gradient projection Newtonian descent (GPND) algorithm [4]:

$$\ddot{p} = -\nabla_p U - 2\gamma\,\dot{p}$$

$$\ddot{\lambda} = -\nabla_{\lambda} U - 2\gamma \dot{\lambda}$$

The discretization of these equations gives an iterative gradient descent algorithm with inertia:

$$p^{t+m} = \frac{2}{\gamma m+1} p^{t} + \frac{\gamma m-1}{\gamma m+1} p^{t-m} - \frac{m^2}{\gamma m+1} \nabla_{p} U(p^{t}, \lambda^{t})$$

$$\lambda^{t+m} = \frac{2}{\gamma m+1} \lambda^{t} + \frac{\gamma m-1}{\gamma m+1} \lambda^{t-m} - \frac{m^2}{\gamma m+1} \nabla_{\lambda} U(p^{t}, \lambda^{t})$$

, where m is the time interval of iteration， γ is a constant weight.

5. Experimental Results

In this section, we illustrate our tracking algorithm on two real image sequences with natural motion. In our experiments, initial tracking regions are manually selected in frame 0(I^0) of the sequence. This region is then tracked from I^0 to I^1, from I^1 to I^2， and so on until the last frame in the sequence. The results are shown in form of the minimal bounding box (MBB) of the tracked region. The first tracking experiment is performed on the *Fish-Tank* sequence (320×240 RGB images, 97 frames). In this case, we use five dimensional feature-spatial space (2D location and 3D RGB color). Firstly, we concentrate on tracking only one distinct fish in the scene and compare the results with that of the condensation algorithm [8]. Fig.4 shows some tracking results of our HMMF-based algorithm. The bandwidth parameters are $(\delta, \theta_{12}) = (8,10)$ and $\sigma = 5$. During the first 60 frames, the background is pure on the motion trajectory of the target white fish. From frame 70, the scenario is much more challenging for the tracker. Our algorithm performs well in both the simple and highly cluttered scenes and is adapted to the shape deformation of the target. As a comparison, we give the tracking results of the condensation tracker [8] on same sequence (Fig. 5). We choose gradient as the image feature. The tracker clearly fails after frame 70. The features throughout the scenario draw the contour tracker away from the true target.


2 **40**


60 **70**


80 **95**

Fig.4 Tracking a fish in the tank


2 **70**


80 **95**

Fig.5 Mistracking of the condensation tracker

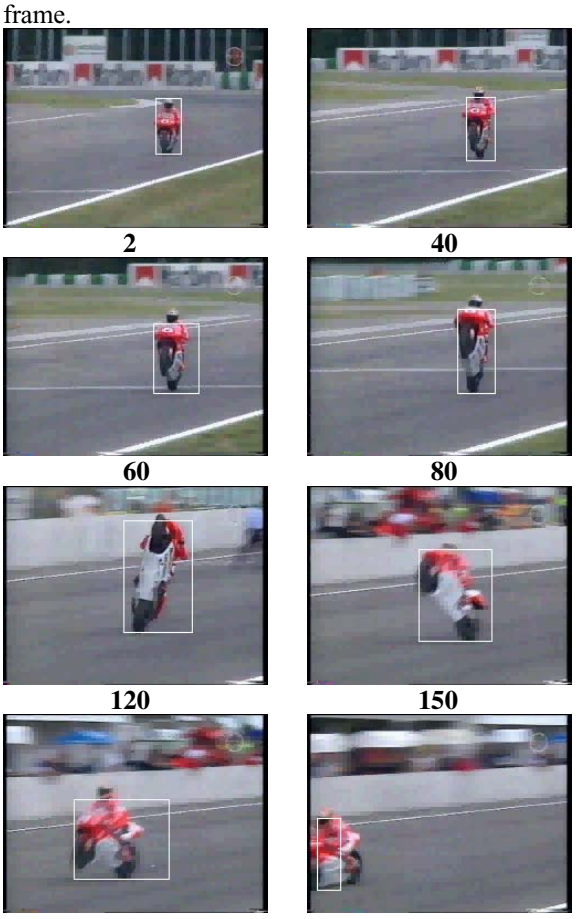
Secondly, we will concentrate on the tracking of two fishes on the same sequence. The two red fishes begin to overlap from frame 60 and then separate again after frame 80. Fig.6 shows the tracking results of the two targets. The tracker also performs well during the whole sequence in spite of the overlapping and deformation in the appearance due to the flowing. The bandwidth parameters are $(\delta, \theta_{12}, \theta_{22}) = (6,20,20)$ and $\sigma = 5$


46 **60**


83 96

Fig.6 Tracking of two fish under mutual occlusion

The second experiment is performed on a sequence called *Bike-God* (240×180 RGB images, 199 frames). In this case, we still use five dimensional feature-spatial space (2D location and 3D RGB color). The bandwidth parameters are $(\delta,\theta_{12})=(8,0.5)$ and $\sigma=8$. This sequence shows tracking where the target is moving fast and the camera is unfixed. The tracked motor undergoes different variant kinds of poses and the orientation of the camera lens is changing in the sequence. From fig.7 we can see that the tracker successfully located the target at each new frame.



2 40 60 80 120 150 160 175

Fig.7 Tracking a motor from moving camera

6. Discussion

In our paper, the tracking is driven by the region structure as well as the region appearance distribution. If we consider $\theta_{i1}\to\infty$ and $\theta_{i2}=0$, which means the ignorance of the spatial constraint and the feature kernel $G_{\theta_{k2}}(\bullet)$ is reduced to a kronecker delta function, then distribution $P_i(x,u\,|\,(\theta_{i1},\theta_{i2}))$ becomes the exact histogram of region R_i^n. We have tried just a histogram representation of regions in our experiments, the realization is rather simple, however, the results are not very satisfying. This is partly because histogram is a very simple global measurement and replies heavily on the samples. Further more, it may be quite similar for the pixel in the boarding area thus background pixels may sometimes be mistracked . Hence the asymptotic behavior of $\theta_{i1}\to\infty$ is not very suitable for tracking purpose. The rigorist spatial constraint is very essential for the success of our HMMF based tracker.

7. Conclusion

We have presented a novel region-based image tracking approach. The novelty of the method lies in the fact that region is formulated as a labeling problem and solved through Bayesian estimation. No motion model is assumed nor any dense motion field is computed. The shape of the region being tracked is not constrained to belong to a particular family of shapes, nor should the region exhibit strong contrast with respect to the background. The only basic assumption of this algorithm is that the object's appearance is generated from some certain probability distribution. This distribution is formulated in the joint Feature-Spatial space via kernel function. The exploitation of doubly stochastic model makes the estimation of optimal labeling field much quicker and more accurate comparing to the traditional MRF method. The experimental results validate our proposed approach. Our current research is aimed at providing better simplification to the kernel-based distribution function.

Acknowledgement

This research is partly supported by the the National High Technology Development 863 Program of China under Grant No. 2002AA145090.

Reference

[1] I. K. Sethi and R. Jain, "Finding Trajectories of Feature Points in a Monocular Image Sequence", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 9, no. 1, pp. 56-73, 1987.

[2] R. Deriche and O. D. Faugeras, "Trackingline segments", In First European Conference on Computer Vision, pp. 259-268, Antibes, Fance, Apr., 1990.

[3] R. Deriche and T. Blaszka, "Recoveringand Characterizing Image Features Using an efficient Model Based Approach", In Computer Vision and Pattern Recognition, New-York, June 14-17, 1993.

[4] N. Paragios and R. Deriche, “Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 3, pp. 266-280, Mar., 2000.

[5] A. R. Mansouri, “Region Tracking via Level Set PDEs without Motion Computation”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 947-961, July 2002.

[6] J. L. Marroquin, E. A. Santana and S. Botello, “Hidden Markov Measure Field Models for Image Segmentation”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 11, pp. 1380-1387, Nov., 2003.

[7] A. Elgammal, R. Duraiswami, L. Davis. Probabilistic tracking in joint feature-spatial spaces, Proceedings IEEE Conference of Computer Vision and Pattern Recognition, Wisconsin, Madison, June 2003. Vol. 1, pp. 781 -788.

[8] M. Isard, and A. Blake, Condensation – conditional density propagation for visual tracking. Internal Journal of Computer Vision, 1998, 28(1):5-28.