EFFECTIVE AND ROBUST OBJECT TRACKING IN CONSTRAINED ENVIRONMENTS

Junda Zhu, Yuanwei Lao, and Yuan F. Zheng

Department of Electrical and Computer Engineering
The Ohio State University
2015 Neil Avenue, Columbus, OH 43210
E-mail: {zhuj, laoy, zheng}@ece.osu.edu

ABSTRACT

We present a new scheme for efficient and robust video object tracking in constrained environments. It finds its application in security surveillance, traffic monitoring, etc. In these applications movements of objects are restricted by the environments; therefore, environment constraints can be exploited as heuristic information for improving the performance of tracking. In this paper we use the distance field to represent environment constraints and integrate it into the framework of particle filtering. Experiments on some video surveillance sequences demonstrate the effectiveness of our approach.

Index Terms— Object tracking, video surveillance, distance transform, particle filtering.

1. INTRODUCTION

The research into video object tracking has received much attention in recent years. Various methods have been proposed to track moving objects in video sequences, such as Kalman filter and its variants [1], mean shift [2], and particle filter [3]. Among these methods, particle filter provides robust solutions to the problems where linearizations and Gaussian approximations are not applicable. But due to its algorithm nature that multiple hypotheses need to be maintained and evaluated, its computational complexity is comparatively higher than the other methods.

Study of effective tracking of objects in constrained environments is of significant importance since it has a wide application scope ranging from access control of communities and buildings, traffic monitoring to military purposes. In these applications, movements of objects are often restricted by surveillance environments such as streets, roads, walls, etc. Although existing generic tracking methods may work in these applications, overlooking the impact of environment constraints results in low tracking accuracy and low efficiency. Some works [4] [5] try to address this problem by offline learning from many training video sequences, and use the learned motion pattern as heuristic information for online tracking. Hu et al. [5] proposed a system for learning statistical motion patterns and applied it into traffic scenes. There is a major drawback for these learning based approaches: a large number of training sequences are required for good performance, which are not always available in practice.

In this paper, we address this problem by mining environment constraints from the surveillance video and represent the environment constraint information by the distance field using the *distance transform*. The generated distance field is integrated into the framework of particle filtering for robust and effective tracking. Although the distance transform was invented in the 1960s [6], and has found its application in many areas such as image analysis [7], robot path planning [8], etc., to the best of our knowledge, our approach is the first to integrate the distance transform into video object tracking in constrained environments.

The paper is organized as follows. Section 2 gives an overview of our proposed scheme. Section 3 studies the details about extracting environment constraints, which includes partitioning the video scene and the representation of the environmental information by distance field. Section 4 discusses the integration of distance fields into particle filtering for robust and effective tracking. Section 5 provides the experimental results, and Section 6 concludes the paper.

2. OVERVIEW OF THE PROPOSED SCHEME

Our approach for effective and robust tracking in constrained environments is based on two fundamental observations: (a) the environment under surveillance can be partitioned into several disjointed regions according to their characteristics and semantic meanings, and (b) objects of interest either move within a certain region or pass across the regions. For example, an outdoor surveillance scene often consists of streets, buildings, lawns, etc., while a traffic monitoring scene is comprised of roads, sidewalks, etc. The topological configuration of these regions forms the environmental constraints upon the object motions. Objects of interest such as pedestrians and vehicles exhibit certain motion patterns under the influences of such constraints. When one moves within a region, such as the cases of people walking along the street and vehicles traveling along the lane, it tends to keep a certain distance from the boundary of the regions. When one passes across regions, its distance from the boundary of the regions undergoes gradual increment/decrement. Based on these observations, the

distance from the object of interest to the boundary of the regions serves as an important heuristic information for describing the movement of objects under environment constraints. Although due to different configurations and view angles of surveillance cameras, the measured distance in the acquired videos/images may not be the same as the real distance, it can still effectively reflect the relative position of the object in the environment.

Our approach consists of two major steps:

1. Extracting environment constraints.

   Our scheme first partitions the scene into several regions, which are semantically meaningful such as roads, lawns, buildings, etc. Then distance transforms are performed upon the partitioned scene, which yield the distance field map of the scene.

2. Particle filtering with a hybrid motion model.

   Distance field is integrated into the motion model of the particle filter to generate more effective particles so that the computational cost can be reduced and the tracking will be more accurate and robust.

3. EXTRACTING ENVIRONMENT CONSTRAINTS

3.1. Scene Partitioning

In order to extract the environment constraints from the video, we first decompose the whole scene into semantically meaningful regions. Let us consider a scene $\mathcal{S} = \{s(x, y)\}$ composed of N pixels. The objective of scene partitioning is to decompose this scene into a tessellation of R regions, namely,

$$\mathcal{S} = \cup_i \Omega_i, i \in \{1, 2, 3, \cdots, R\}, s.t.\ \Omega_i \cap \Omega_j = \phi, \forall\, i, j. \quad (1)$$

Each region Ω_i should correspond to one of the meaningful environmental entities, such as roads, lawns, walls, etc. The boundary between these regions can be easily obtained once the regions have been segmented, which can be represented by a set of boundary pixels $\Gamma = \cup_i \partial\Omega_i, i \in \{1, 2, 3, \cdots, R\}$.

In this paper, we accomplish this task in two steps. Firstly, we apply the image segmentation algorithm on the video scene, which breaks the whole scene into several parts according to image features. By comparing several popular segmentation algorithms, we adopt the mean-shift based segmentation method [9] due to its low over-segmentation rate. By choosing appropriate color/spatial bandwidth and large value for the minimum region area, accurate segmentation results are obtained for the testing sequences. For algorithm details, we refer the readers to Comaniciu & Meer [9].

Secondly, a supervised region merging and boundary smoothing step is applied to the segmentation output from the first step. Since the segmentation algorithm itself does not take the semantic meanings of the segmented regions and the smoothness of the boundary into consideration, the scene is possibly over-segmented, and the generated boundary is often in a ragged manner. This is not consistent with the real environment constraints and also not convenient for further processing. At this step, user inputs are incorporated as high level information for merging over-segmented regions. The extracted boundary is filtered with a neighborhood averaging filter for better smoothness.

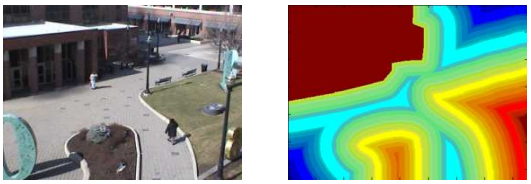


Fig. 1. A typical surveillance scene and its distance field representation.

3.2. Distance Transforms

Distance Transform (DT) maps a 2-D image or a 3-D shape into a distance field map, in which the value at each point corresponds to its distance to the nearest boundary point. DT was first developed as a tool for image analysis [6] [7]. Jarvis [8] successfully extended the DT to robot path planning, in which a collision-free path in a structured environment could be determined by following the steepest descent direction in the distance field map. In this paper, we utilize the DT for tracking in constrained environments, which is the first attempt of applying the DT to video object tracking in constrained environments to the best of our knowledge. The application of DT in our new tracking approach is described as follows.

After the scene partitioning, we have obtained a partition of the scene $\mathcal{S} = \cup_i \Omega_i, i \in \{1, 2, 3, \cdots, R\}$, and also a boundary pixel set Γ. The generic Euclidean distance field [7] has been defined as the value that equals to the shortest Euclidean distance from a pixel $\mathbf{p}$ to the closest point in Γ:

$$\mathsf{d}_\Gamma(\mathbf{p}) = \min_{\mathbf{x} \in \Gamma} \|\mathbf{x} - \mathbf{p}\|. \quad (2)$$

In this paper, we employ a linear time algorithm based on Voronoi diagram construction, proposed by Breu et al. [10]. For the purpose of our tracking application, not only are we interested in the value of the absolute distance, but also want the distance field to help distinguish different regions. Therefore, a weighted distance field representation is employed in this paper. The value of the weighted distance field at a certain point is determined by both the shortest distance to the boundary and the assigned weight of the region to which this point belongs. It can be represented by the following equation:

$$\mathsf{d}_{\Gamma,\Omega_i}(\mathbf{p}) = \mathsf{w}(\Omega_i) \min_{\mathbf{x} \in \Gamma, \mathbf{p} \in \Omega_i} \|\mathbf{x} - \mathbf{p}\|. \quad (3)$$

In this paper, we simply assign the weights according to the properties of the regions, namely,

$$\mathsf{w}(\Omega_i) = \begin{cases} -1 & \text{if } \Omega_i \text{ is a pathway} \\ 1 & \text{if } \Omega_i \text{ is an accessible region} \\ \infty & \text{if } \Omega_i \text{ is an inaccessible region.} \end{cases}$$

The above assignment creates a distance field map which associates each point of the scene with a distance field value and is continuous everywhere except for at the boundaries of the inaccessible regions. A typical video surveillance scene of a constrained environment and its distance field map is shown in Fig. 1. Without ambiguity, the subscripts of $\mathsf{d}_{\Gamma,\Omega_i}$ are dropped, and the weighted distance field is called distance field in the remainder of this paper.

4. OBJECT TRACKING

4.1. Generic Particle Filter

Video object tracking is essentially an estimation problem in which states of objects are to be estimated based on video observations. The particle filter (PF) provides an approximate Bayesian solution to the video object tracking problem by recursively updating an approximate discrete description of the posterior probability.

In the setting of video object tracking, we use $\mathbf{x}_t$ to denote the state of the object of interest at time t, which usually includes the position and/or velocity information, and use $\mathbf{Z}_t = \{\mathbf{z}_i\}_{i=1}^t$ to represent the set of the measurements from the beginning until time t. Depending on the selection of measurement model, $\mathbf{z}_t$ may be computed using color histogram, active contour, PCA coefficients, classification score, or the combination of some of these models. The particle filter approximates the posterior density $p(\mathbf{x}_t|\mathbf{Z}_t)$ using a set of N weighted particles $\{\mathbf{x}_t^i\}_{i=1}^N$ with corresponding weights $\{w_t^i\}_{i=1}^N$, by $p(\mathbf{x}_t|\mathbf{Z}_t) \approx \sum_{i=1}^N w_t^i \cdot \delta(\mathbf{x}_t - \mathbf{x}_t^i)$ .Therefore, the minimum mean square error (MMSE) estimator of the object state is straightforwardly the posterior mean, which is given by $\hat{\mathbf{x}}_t = \sum_{i=1}^N w_t^i \cdot \mathbf{x}_t^i$.

How to generate particles effectively so that the collection of the particles can well describe the characteristic of the posterior distribution is the central issue for effective particle filtering trackers. In the PF framework, particles are propagated throughout time by sampling from the proposal distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$. As for the widely used Sampling Importance Resampling (SIR) PF, the motion model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is directly employed as the proposal distribution. Therefore, a proposal distribution or motion model which can precisely depict the motion of objects is desirable for effective tracking.

4.2. The Hybrid Motion Model

In view of the shortcomings of existing tracking methods, we integrate the environmental information into the particle filtering framework, which improves the efficiency as well as the robustness of existing PF for tracking in constrained environments.

Since the state of the moving object is not only determined by the previous state, but also influenced by the environment constraints, it is natural to bring the environment constraints, which is represented by the distance field, into the motion analysis. Following this idea, we first augment the state vector $\mathbf{x}$ by incorporating the distance field value d at that position,

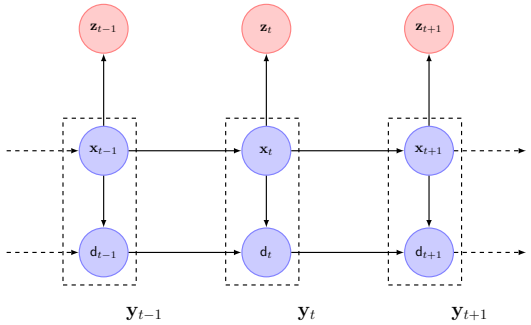


Fig. 2. The graphical probabilistic model for the particle filter with the hybrid motion model.

i.e. create a hybrid state vector $\mathbf{y} = \{\mathbf{x}, \mathsf{d}\}$. Note that d is a variable dependent on $\mathbf{x}$. The probabilistic dependencies among the variables are represented by a graphical model, shown in Fig. 2. A Markovian assumption is made, which implies that the current distance field value d_t is independent with the previous distance field values except for d_{t-1}.

The generic motion model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is then replaced by a hybrid motion model $p(\mathbf{y}_t|\mathbf{y}_{t-1})$. From the Bayes rule, it follows that

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{y}_{t-1}) &\equiv p(\mathbf{x}_t, \mathsf{d}_t|\mathbf{x}_{t-1}, \mathsf{d}_{t-1}) \\ &= p(\mathbf{x}_t|\mathsf{d}_t, \mathbf{x}_{t-1}, \mathsf{d}_{t-1}) \cdot p(\mathsf{d}_t|\mathbf{x}_{t-1}, \mathsf{d}_{t-1}) \\ &= p(\mathbf{x}_t|\mathbf{x}_{t-1}) \cdot p(\mathsf{d}_t|\mathsf{d}_{t-1}, \mathbf{x}_{t-1}), \end{aligned} \tag{4}$$

where we also use the conditional independence property $p(\mathbf{x}_t|\mathsf{d}_t, \mathbf{x}_{t-1}, \mathsf{d}_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$, which can easily be determined from the graphic model shown in Fig. 2.

From Eq. 4, it can be seen that the hybrid motion model is proportional to the product of the motion model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and an "environmental prior" term $p(\mathsf{d}_t|\mathsf{d}_{t-1}, \mathbf{x}_{t-1})$. By manipulating this environmental prior, the environmental constraints are imposed upon the generation of particles. In the constrained environment, an object tends to move either along a path with approximately equal distance field, or following a path with gradually increasing/decreasing distance field values. Therefore, for any $\mathbf{x}_{t-1}$, $p(\mathsf{d}_t|\mathsf{d}_{t-1}, \mathbf{x}_{t-1})$ should have larger value in the neighborhood of d_{t-1} and decrease when $|\mathsf{d}_t - \mathsf{d}_{t-1}|$ gets larger. In this paper, given the state vector at previous time step $\mathbf{x}_{t-1}$, we simply choose a uniform distribution which centers at d_{t-1} with a small interval $2h$, namely, $p(\mathsf{d}_t|\mathsf{d}_{t-1}, \mathbf{x}_{t-1}) \sim \mathcal{U}(\mathsf{d}_{t-1} - h, \mathsf{d}_{t-1} + h)$, which discards the particles with large deviations of distance field value from that of the previous time step. Other simple forms of probability distribution can also be used, e.g. Gaussian distribution.

5. EXPERIMENTAL RESULTS

Experiments are conducted on some real video surveillance sequences, and performance comparisons are made between our scheme and the generic particle filter. The first order kinematic model and color histogram based measurement model are used for both schemes, and all the designing parameters

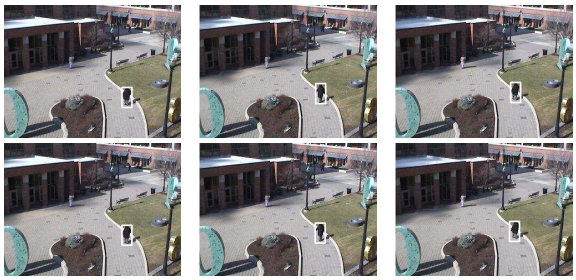


Fig. 3. Test results on the OTCBVS sequence. Top row: the generic particle filter with 80 particles. Bottom row: our approach which uses 30 particles.

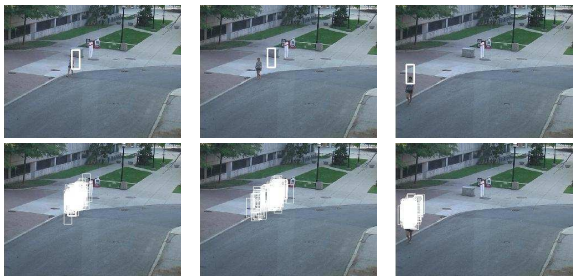
are kept the same. The parameter h of the environmental prior $p(\mathsf{d}_t|\mathsf{d}_{t-1})$ is chosen to be 5.

The first test sequence is a surveillance video taken from IEEE OTCBVS WS Series Bench [11], which can be found at http://www.cse.ohio-state.edu/OTCBVS-BENCH/bench.html. A person walking along a pathway is tracked throughout a 400-frame-long sequence recorded at a resolution of 320× 240 at 30 Hz. The tracking results of both schemes are shown in Fig. 3. As can be seen, both approaches can track the object well for this simple scenario. However, our proposed approach only requires 30 particles on average for stable tracking, far more efficient than the generic particle filter which uses 80 particles in our test.

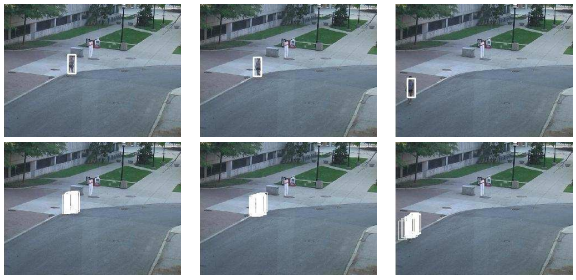
The second test sequence consists of 208 frames sampled at 5 Hz with a resolution of 320×240 pixels, recording a person walking along a road on OSU campus. This sequence is more challenging than the first one, since the color of the tracked person happens to be quite similar with some background objects. The tracking results and particle distributions are shown in Fig. 4. For the generic particle filter, a total of 150 particles are used. In frame 99 and 112, a large portion of the particles are attracted by a background object, which results in inaccurate tracking and temporary losing of track. For our distance field based scheme, an average of 30 effective particles suffices for good tracking results throughout the whole sequence even when the generic particle filter fails. As can be seen from Fig. 4(b), the distributions of the particles of our scheme effectively cover the possible path of the tracked object and exclude the areas which the object is unlikely to visit, in contrast with the uninformed particle distributions in the generic scheme.

6. CONCLUSIONS

In this paper we have proposed a new approach for efficient and robust video object tracking in constrained environments. Inspired by the observation that the object motion is restricted by the environmental constraints, we first extract such constraints from video by scene partitioning followed by the distance transform, and then integrate this information, in terms of distance field, into the tracking framework of the particle filter. Experimental results have shown the computational efficiency of the proposed scheme in reducing the number of particles and its robustness against background clutters.



(a) Generic particle filter



(b) Our approach

Fig. 4. Tracking results (top row) and particle distributions (bottom row) of the 2nd test case (frame # 99, 112, 188).

7. REFERENCES

[1] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*, New York: John Wiley & Sons, 2001.

[2] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, May 2003.

[3] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*, Artech House, 2004.

[4] D. Makris and T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 35, no. 3, pp. 397–408, June 2005.

[5] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1450–1464, Sept. 2006.

[6] A. Rosenfeld and J. Pfaltz, "Sequential operations in digital picture processing," *Journal of the ACM*, vol. 13, no. 4, pp. 471–494, 1966.

[7] G. Borgefors, "Distance transformations in digital images," *Comput. Vision Graph. Image Process.*, vol. 34, no. 3, pp. 344–371, 1986.

[8] R. Jarvis, "Distance transform based path planning for robot navigation," in *Recent Trends in Mobile Robots*, Y. Zheng, Ed. World Scientific, NJ, 1993.

[9] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.

[10] H. Breu, J. Gil, D. Kirkpatrick, and M. Werman, "Linear time Euclidean distance transform algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 529–533, May 1995.

[11] J.W. Davis and V. Sharma, "Fusion-based background-subtraction using contour saliency," in *Proc. 2005 IEEE Conference on Computer Vision and Pattern Recognition*, June 2005, vol. 3, pp. 11–11.