MULTI-OBJECT FILTERING FROM IMAGE SEQUENCE WITHOUT DETECTION

Reza Hoseinnezhad[1], *Ba-Ngu Vo*[1], *David Suter*[2] *and Ba-Tuong Vo*[3]

[1]Department of Electrical & Electronic Engineering, The University of Melbourne
VIC 3010, Australia, {rezah,bnvo}@unimelb.edu.au
[2]School of Computer Science, The University of Adelaide
SA 5005, Australia, dsuter@cs.adelaide.edu.au
[3]School of Electrical, Electronic & Computer Engineering, The University of Western Australia
WA 6009, Australia, btv@ee.uwa.edu.au

ABSTRACT

Almost every single-view visual multi-target tracking method presented in the literature includes a detection routine that maps the image data to point measurements relevant to the target states. These measurements are commonly further processed by a filter to estimate the number of targets and their states. This paper presents a novel visual tracking technique based on a multi-object filtering algorithm that operates directly on the image observations without the need for any detection. Experimental results on tracking sport players show that our proposed method can automatically track numerous interacting targets and quickly finds players entering or leaving the scene.

Index Terms— visual tracking, object filtering, Bayesian estimation, multi-target tracking, random finite sets

1. INTRODUCTION

To the best of our knowledge, all single-view visual tracking techniques, appearing in the literature so far, include a *detection* module that generates point measurements from the images in the video sequence. The point measurements are then usually used as inputs by a *filtering* module that estimates the number of targets and their states (properties such as location and size) from the results of detection.

Detection has been and still is an integral component of visual tracking systems, with a large body of literature on models and techniques for detecting targets based on various background and foreground models. One of the most popular approaches is the detection of targets based on matching color histograms of rectangular blobs [1, 2]. Other recent methods include a game-theoretic approach [3], using human shape models [4, 5], multi-modal representations [6], sample-based detection [7], range segmentation [8] and a multi-step detection scheme including median filtering, thresholding, binary morphology and connected components analysis [9].

Although some tracking methods do not include any filtering routine (for the actual tracking of the targets) and they suffice with the detection results [3, 6, 7, 9], such methods include a search routine to find the best blobs (*e.g.* to optimize a cost function to find a MAP or ML estimate of targets states). However, in applications where numerous targets are to be tracked (*e.g.* sport players tracking), there is a high likelihood that some targets are missed. Therefore, in such applications, stochastic filtering is required to keep track of all targets including those missed in the detection process.

This paper presents a novel multi-target visual tracking method, formulated in a random finite set framework, that tracks multiple moving targets directly from the image information embedded in a video sequence, without the need for extracting any point measurements. A sequential Monte Carlo implementation of the multi-Bernoulli approximation to the Bayesian filter is explained and examined in two case studies involving tracking of multiple sport players.

2. BACKGROUND

In order to jointly track the number of targets and their state values, we represent the collection of states, referred to as the multi-target state, as a *finite set*. As in the Bayesian estimation paradigm, the state and measurement are treated as realizations of random variables, the finite-set-valued (multi-target) state X is modeled as a *random* finite set (RFS). The Finite Set Statistics (FISST) framework, developed by Mahler [10], provides the practical mathematical tools for dealing with RFSs. Using the FISST notion of integration and density, we can compute the posterior probability density $\pi(\cdot|y)$ of the multi-target state from the prior density based on Bayes rule:

$$\pi(X|y) = \frac{g(y|X)\pi(X)}{\int g(y|X)\pi(X)\delta X} \tag{1}$$

This work was supported by Australian Research Council through the ARC Discovery Project grant DP0880553.

where $g(y|X)$ is probability density (likelihood) of observation y given the multi-target state X and the integral over the space of finite sets is defined as follows:

$$\int f(X)\delta X \triangleq \sum_{i=0}^{\infty} \frac{1}{i!} \int f(\{x_1, \ldots, x_i\}) dx_1 \ldots dx_i. \quad (2)$$

In this paper, the finite set of targets, X, is modeled by a *multi-Bernoulli* RFS which is defined as the union of M independent RFSs $X^{(i)}$ where M is the maximum number of targets. In this representation, each $X^{(i)}$ is either empty or a singleton with probabilities $1-r^{(i)}$ and $r^{(i)}$, respectively. In case $X^{(i)}$ is a singleton, its only element is distributed according to a probability density $p^{(i)}(\cdot)$. Thus, a complete representation of the multi-target state is given by $\{(r^{(i)}, p^{(i)}(\cdot))\}_{i=1}^{M}$.

It was shown in [11] that if the likelihood function has the following separable form:

$$g(y|X = \{x_1, \ldots, x_n\}) = f(y) \prod_{i=1}^{n} g(x_i, y) \qquad (3)$$

and the multi-target RFS has a multi-Bernoulli prior distribution $\{(r^{(i)}, p^{(i)})\}_{i=1}^{M}$, then the posterior distribution of X, given by Bayes rule (1), is also multi-Bernoulli with the parameters $\{(r_{\text{updated}}^{(i)}, p_{\text{updated}}^{(i)})\}_{i=1}^{M}$ where:

$$r_{\text{updated}}^{(i)} = \frac{r^{(i)} \langle p^{(i)}(\cdot), g(\cdot, y)\rangle}{1 - r^{(i)} + r^{(i)} \langle p^{(i)}(\cdot), g(\cdot, y)\rangle} \qquad (4)$$

$$p_{\text{updated}}^{(i)}(\cdot) = \frac{p^{(i)}(\cdot) g(\cdot, y)}{\langle p^{(i)}(\cdot), g(\cdot, y)\rangle} \qquad (5)$$

and $\langle f_1, f_2 \rangle$ denotes the standard inner product $\int f_1(x) f_2(x) dx$.

In the next section, we present a likelihood function in the above separable form, formulated based on HSV color histograms. Using the update results (4) and (5), we devise a complete filtering scheme take takes the raw image sequence as input to directly track multiple targets.

3. VISUAL TRACKING

Without loss of generality, in our method, each target is represented as a rectangular blob and a state vector comprising the location and size of the blob in pixels. Consider the image y in one frame from a video sequence. Given a multi-target state $X = \{x_1, \ldots, x_n\}$, we derive a separable measurement likelihood function $g(y|X)$. From the image y, we can compute the HSV histogram of the image blob corresponding to each target location x_i. We denote the histogram values for the i-th target by the vector v_i. To ensure that histogram values represent probability distributions, each vector is normalized to sum to 1. For the set of pixels that do not belong to any target (hypothetically background pixels), the HSV histogram is also computed, and the values recorded are denoted by the vector v_b. It is important to note that the values recorded in the vectors $v_1, \ldots, v_n$ and v_b depend on both the target states and the image y.

We assume that the histograms of individual targets and the histogram of the background are mutually independent from each other, noting that as long as the targets do not largely occlude each other, they do not substantially affect each other's color histogram. Thus, the likelihood function can be formulated as follows:

$$g(y|X) = g_b(v_b) \prod_{i=1}^{n} g_i(v_i) \qquad (6)$$

where $g_b(v_b)$ is the likelihood of background histogram to be given by v_b, and $g_i(v_i)$ is the likelihood that a target is present in the image y with state x_i.

The color histogram of the background can reasonably be assumed to have time-invariant statistics. Note that this is not same as assuming a static background. Indeed, the background may change, but the portions of HSV colors contributing to it are assumed to change very slightly. In applications where numerous targets are to be tracked and each target covers a relatively small portion of the image, movement of the targets will not change the color histogram of the unoccupied parts of the image (background) drastically, and we can reasonably assume that each component of color histogram values vary in a narrow band (they are almost constant) with a uniform distribution. Hence, $g_b(\cdot)$ is constant and plays the role of $f(y)$ in the separable likelihood form (3).

The likelihood terms $g_i(\cdot)$ can be computed using kernel density estimation over a database of training data. The database contains n_{train} vectors $\{v_j^*\}_{j=1}^{n_{\text{train}}}$ and each vector corresponds to the HSV color histogram of a training blob. For a given histogram v_i, the likelihood function is then given by the following kernel density estimate:

$$g_i(v_i) = \frac{\xi}{n_{\text{train}}\, h^N} \sum_{j=1}^{n_{\text{train}}} \kappa\left(\frac{d(v_i, v_j^*)}{h}\right) \qquad (7)$$

where $\kappa(\cdot)$ is the kernel function (Gaussians used in our experiments), h is the kernel bandwidth, N is the total number of bins in each histogram and $d(v_i, v_j^*)$ is the Bhattacharyya distance [1]:

$$d(v_i, v_j^*) = \left(1 - \sum_{r=1}^{N} \sqrt{v_{jr}^* v_{ir}}\right)^{\frac{1}{2}} \qquad (8)$$

and ξ is the normalizing factor to ensure that the kernel density integrates to 1.

It is important to note that the likelihood function $g_i(v_i)$ defined in (7) depends both on the image observation y (because the blobs are extracted from the image) and on the target state x_i (because the location and size of the blob is determined from the target state). Thus, $g_i(v_i)$ represents a formulation for the $g(x_i, y)$ term in (3).

3.1. Sequential Monte Carlo Implementation

In the following, a sequential Monte Carlo implementation of a multi-Bernoulli filter is reviewed. The algorithm is based on the method presented in [11], adapted to the likelihood function defined in (6) for multi-target visual tracking.

Suppose that at time $k-1$, the posterior density $\{r^{(i)}_{k-1}, p^{(i)}_{k-1}\}_{i=1}^{M_{k-1}}$ is given and each $p^{(i)}_{k-1}$ is represented by a set of weighted samples (particles) $\{w^{(i,j)}_{k-1}, x^{(i,j)}_{k-1}\}_{j=1}^{L^{(i)}_{k-1}}$. More precisely,

$$p^{(i)}_{k-1}(x) = \sum_{j=1}^{L^{(i)}_{k-1}} w^{(i,j)}_{(k-1)} \delta_{x^{(i,j)}_{k-1}}(x). \qquad (9)$$

We assume a constant survival probability P_S, and consider a predefined model for birth particles denoted by known parameters $\{r^{(i)}_{\Gamma}, p^{(i)}_{\Gamma,k}\}_{i=1}^{M_\Gamma}$ where the density $p^{(i)}_{\Gamma,k}$ is represented by the particles $\{w^{(i,j)}_{\Gamma,k}, x^{(i,j)}_{\Gamma,k}\}_{j=1}^{L_\Gamma}$. In our experiments, we assume that with a constant probability of 0.02, one target appears in each of the four quarters of the image planes, with the location of the target being uniformly distributed within the quarter. Thus, $M_\Gamma = 4$, $r^{(1)}_\Gamma = \cdots = r^{(4)}_\Gamma = 0.02$ and the birth particles are sampled with uniform distribution and weights.

Similar to many other particle filtering schemes, in each iteration, the particles are predicted then updated. In the prediction step, the birth particles are generated according to the birth model parameters. The multi-Bernoulli parameters from the previous iteration, $\{r^{(i)}_{k-1}, w^{(i,j)}_{k-1}, x^{(i,j)}_{k-1}\}$, are propagated forward:

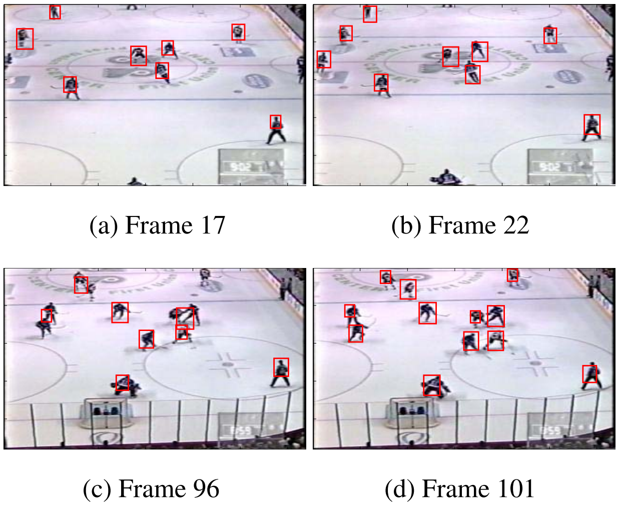
$$\begin{aligned} x^{(i,j)}_{k|k-1} &\sim f_{k|k-1}(\cdot|x^{(i,j)}_{k-1}) & (10)\\ r^{(i)}_{k|k-1} &= P_S\, r^{(i)}_{k-1}\ ;\ w^{(i,j)}_{k|k-1} = w^{(i,j)}_{k-1}. & (11) \end{aligned}$$

The proposal density equals the state transition density $f_{k|k-1}(\cdot|x_{k-1})$. In our experiments, the targets are modeled by rectangular blobs and the target state is a 4-tuple vector comprising the x and y location and width and height. The target dynamic is modeled by $x(k+1) = x(k) + e(k)$ where $e(k)$ is a 4-dimensional Gaussian variable with zero mean and variance $\Sigma = \mathrm{diag}(\sigma_x^2, \sigma_y^2, \sigma_h^2, \sigma_w^2)$. Thus, $f_{k|k-1}(x|x_{k-1}) = \mathcal{N}(x, \Sigma)$.

In the update step, the predicted multi-Bernoulli parameters are updated using the likelihood function (7) and update formulas (4) and (5) which translate to:

$$\begin{aligned} r^{(i)}_k &= r^{(i)}_{k|k-1}\varrho^{(i)}_k \Big/ \left(1 - r^{(i)}_{k|k-1} + r^{(i)}_{k|k-1}\varrho^{(i)}_k\right) & (12)\\ w^{(i,j)}_k &= w^{(i,j)}_{k|k-1}\ g_{y_k}(x^{(i,j)}_{k|k-1}) / \varrho^{(i)}_k & (13) \end{aligned}$$

where $\varrho^{(i)}_k = \sum_{j=1}^{L^{(i)}_{k|k-1}} w^{(i,j)}_{k|k-1}\ g_{y_k}(x^{(i,j)}_{k|k-1})$ [11].


(a) Frame 17

(b) Frame 22

(c) Frame 96

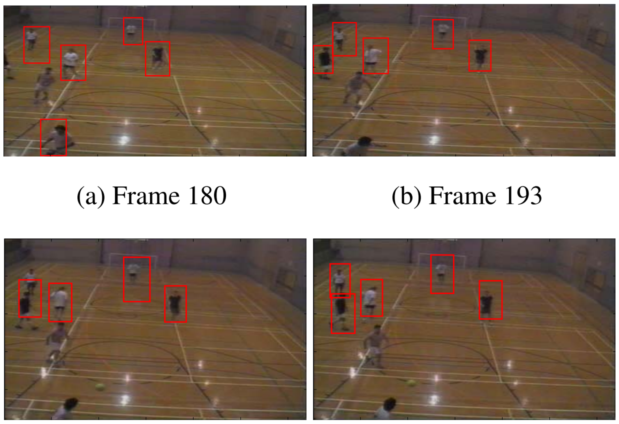
(d) Frame 101

Fig. 1. Four snapshots of the results of hockey player tracking: (a) A new player is entering the scene from left side (b) The new player is detected and tracked (c) Three couples of very close players are merged into single targets (d) The merged targets are separately tracked as soon as the players get separated.

Similar to the MeMBer filter [12], the updated particles are resampled with the number of particles reallocated in proportion to the probability of existence as well as restricted between a minimum $L_{\min}$ and maximum $L_{\max}$. To reduce the growing number of multi-Bernoulli parameters, those with probabilities of existence less then a small threshold (set at 0.01) are removed. In addition, the targets with substantial overlap are merged. Finally, the number of targets and their states are estimated via finding the multi-Bernoulli parameters with existence probabilities larger than a threshold (set at 0.5 in our experiments). Each target state estimate is then given by the weighted average of the particles of the corresponding density.

4. EXPERIMENTAL RESULTS

We examined our method to track multiple players in two video sequences. The fist video sequence includes 101 frames of a hockey game (benchmarked in [1]) and the second includes over 700 frames of an indoor football game. Snapshots of the tracking results are presented in Figures 1 and 2. With the hockey game, we recorded the HSV histograms of 1500 training rectangular blobs, each manually selected to contain a player. With the football game, this number was 2200. In addition, in the case of football game, we deliberately did not select any blob containing one specific player (the one with the light pink stripy shirt). The main purpose of this exclusion was to examine the selectiveness of tracking (*e.g.* tracking the players and not the referee). As it appears in the tracking results, that player is not picked by the tracker at any time.



(a) Frame 180 (b) Frame 193

(c) Frame 200 (d) Frame 205

Fig. 2. Four snapshots of football player tracking: (a) A new player enters the scene (b) The player is detected and tracked – see the left side of the image (c) Two very close players are merged into a single target (d) They are separately tracked when the players get apart.

Figures 1and 2 show that our method is capable of tracking numerous targets at the same time. In addition, Figures 1(a)-(b) and 2(a)-(b) demonstrate the ability of the method to inherently detect and track the arrival of new targets.

When there is a substantial occlusion, the merging step in our tracking method results in a single blob for the occluding targets. However, as Figures 1(c)-(d) and 2(c)-(d) show, as soon as the moving targets separate from each other, the tracker corrects its results.

Full video of the tracking results for the hockey game can be downloaded and viewed from the following link: `www.ee.unimelb.edu.au/people/rezah/hockey.avi` and the video showing the tracking results for the football game is available to download from: `www.ee.unimelb.edu.au/people/rezah/football.avi`.

5. CONCLUSIONS

A novel visual multi-target tracking method, capable of direct tracking from image observations without the need for any detection, was presented. A separable likelihood function in terms of color histograms of hypothesis targets was formulated and used in the context of a MeMBer filter. The method was evaluated in two sport player tracking experiments, showing that numerous players can be tracked accurately.

6. REFERENCES

[1] K. Okuma, A. Taleghani, N. De Freitas, J.J. Little, and D.G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *ECCV'04*, 2004, vol. 3021, pp. 28–39.

[2] M. Kristan, J. Per, M. Pere, and S. Kovacic, "Closed-world tracking of multiple interacting targets for indoor-sports applications," *Computer Vision and Image Understanding*, vol. 113, no. 5, pp. 598 – 611, 2009.

[3] M. Yang, T. Yu, and Y. Wu, "Game-theoretic multiple target tracking," in *ICCV'07*, Rio de Janeiro, Brazil, 2007, http://dx.doi.org/10.1109/ICCV.2007.4408942.

[4] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *IJCV*, vol. 75, no. 2, pp. 247 – 266, 2007.

[5] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *PAMI*, vol. 30, no. 7, pp. 1198 – 211, 2008.

[6] S. Apewokin, B. Valentine, R. Bales, L. Wills, and S. Wills, "Tracking multiple pedestrians in real-time using kinematics," in *CVPR'08 Workshops*, Anchorage, AK, United states, 2008, http://dx.doi.org/10.1109/CVPRW.2008.4563149.

[7] L. Zhu, J. Zhou, and J. Song, "Tracking multiple objects through occlusion with online sampling and position estimation," *Pattern Recognition*, vol. 41, no. 8, pp. 2447 – 2460, 2008.

[8] E. Parvizi and Q.M. Jonathan Wu, "Multiple object tracking based on adaptive depth segmentation," in *Canadian Conference on Computer and Robot Vision – CRV 2008*, Windsor, ON, Canada, 2008, pp. 273 – 277.

[9] R.G. Abbott and L.R. Williams, "Multiple target tracking with lazy background subtraction and connected components analysis," *Machine Vision and Applications*, vol. 20, no. 2, pp. 93 – 101, 2009.

[10] R.P.S. Mahler, *Statistical multisource-multitarget information fusion*, Artech House, 2007.

[11] B.-N. Vo, B.-T. Vo, N.-T. Pham, and D. Suter, "Bayesian multi-object estimation from image observations," in *Proc. 12th Annual Conf. Information Fusion*, R. Lynch and C. Chong, Eds., Seatle, Washington, 2009, ISIF.

[12] B.-T. Vo, B.-N. Vo, and A. Cantoni, "The cardinality balanced multi-target multi-Bernoulli filter and its implementations," *IEEE Trans. Signal Processing*, vol. 57, no. 2, pp. 409–423, 2009.