A Bayesian Approach: Fusion of Laser and Vision for Multiple Pedestrians Tracking

Xuan Song, Jinshi Cui, Huijing Zhao and Hongbin Zha

Key Laboratory of Machine Perception (Ministry of Education), Peking University
Beijing, 100871, China
{songxuan,cjs,zhaohj,zha}@cis.pku.edu.cn

Abstract

We present a promising system to simultaneously detect and track multiple pedestrians in the outside scene using laser and vision. The useful information of laser and vision is automatically extracted and combined in a Bayesian formulation. In order to compute MAP estimation, an effective Probabilistic Detection-based Particle Filter (PD-PF) has been proposed. Experiments and evaluations demonstrate that not only can our system perform robustly in real environments, but also obtain better approximation of MAP than previous methods in most complex situations.

Key words: Multi-target tracking, Sensor fusion, Bayesian estimation.

1. Introduction

Detection and tracking of multiple interacting people is a problem that arises in a variety of different contexts. Examples include intelligent surveillance for security purposes, scene analysis for service robot, crowds’ behavior analysis for human behavior study, traffic flow analysis and many others.

In this paper, we present a promising system using laser and vision that aims at reliable and real-time monitoring and tracking of multiple people in crowds. Laser scanner provides us robust human detection map and video camera provides rich color information. Therefore, our system incorporates the advantages of laser and vision to obtain a better tracking performance. The fusion in our system contains two aspects: (1) Feature level fusion: As illustrated in Figure 1, the useful information of laser and vision is extracted automatically and combined in a Bayesian framework, and then we perform the maximum a posteriori (MAP) estimation to obtain the tracking results. (2) Strategy level fusion: As illustrated in Figure 2, since the laser scanner can provide robust detections, therefore we can perform the detection-based visual tracking. With the help of the robust detections, tracking of pedestrians becomes more reliable.

Moreover, in order to compute MAP estimation and deal with the interactions and occlusions among pedestrians, we propose a Probabilistic Detection-based Particle Filter (PD-PF). In our method, we design a mixed proposal distribution for sampling particles. This proposal incorporates the interacting detections and information of dynamic model, which can quite effectively model the interactions and occlusions among targets by adjusting the weights of the components in it. Moreover, we run independent PD-PF for each target, such that particles are sampled in a small state space, thus our method not only obtains a better approximation of posterior than joint particle filter or independent particle filter when interactions occur, but also has an acceptable computational complexity. Experiments and results prove the efficiency of our proposed method.

The remainder of this paper is organized as follows: Section 2 gives a brief review of previous work. Section 3 introduces our system architecture. Section 4 introduces the formulation of laser and vision. An Effective MAP Estimation method is present in section 5. Experiments
and results are given in section 6, and the paper is summarized in section 7.

2. Related Work

In the field of computer vision, automatically detecting and tracking people has become a very important topic. Over the last several years, a vast body of literature exists [1-10]. However, most of these algorithms face three challenges: the change of lighting condition, the interaction or occlusion between targets and the vast time-consuming. Thus these methods are difficult to utilize in the real environments, where lighting conditions are not controlled and temporary occlusions can occur at any time.

On the other hand, laser-based detection and tracking systems can provide reliable automatic detection of human in variable scenes. They are insensitive to lighting conditions and laser data processing does not consume much time. Lots of laser-based people tracking systems [11-14] have been developed in the past years. However, the limitations of these systems are inherent and obvious. They cannot provide information such as color of objects, so it is difficult to obtain a set of features that uniquely distinguish one object from another.

Recently, several systems [15-18] have been proposed in robotic area to detect and track people based on a combination of distance information obtained from laser range data and visual information obtained from a camera. [15,18] obtain laser range data to detect moving objects, then the position information is used to perform face detection. In [16], face-tracking module is used to verify that non-confirmed feet belong to people. [17] makes a fusion of panoramic vision and laser range data for human tracking in a stationary robot. However, all the above systems can only detect and track one or few people, which cannot be applied in a real surveillance and monitoring environment. [19] makes a new attempt to fuse the laser and vision for multiple pedestrians detection and tracking in real environment, which is the original version of this paper.

3. System Architecture

A typical tracking system using laser and camera is illustrated in Figure 3. The goal of our system is to sufficiently incorporate the laser observation and visual observation into one framework for obtaining more accurate tracking results. Moreover, we can exploit the powerful detection capability of laser scanners to easily extract the visual information from each target. Human detection is mainly done by laser scanner. After detection, the information of the two sensors is fused into a Bayesian framework, and an efficient maximum a posteriori (MAP) estimation method is utilized. The overall tracking system is depicted in Figure 4.

3.1 Laser-based Subsystem

In this subsystem, laser range scanners, LMS291, produced by SICK are exploited. LMS291 has a maximum range of 80-metre and an average distance error of 3cm. It is set on the tripod (about 120cm above the ground surface) performing horizontal scanning. The data of moving (e.g. human body) as well as still objects (e.g. buildings, trees and so on) are obtained in a rectangular coordinate system. Moving objects are subsequently obtained by background subtraction.

3.2 Vision-based Subsystem

Once the human body is detected using laser data, corresponding body region in video image is localized from average man height, body position and calibrated camera model. The video camera is calibrated to the global coordinate system using Tsai’s model, where both internal and external parameters are calculated using at least 11 control points. More details can be found in [20]. Rectangle is utilized to model body region in the image plane, which can be described by \( p^x \) and \( s \). The \( p^x \) is the center position of a body in the image plane, \( s \) the scale factor of the rectangle width and height. Both of the two parameters can be computed by camera model easily, which are also the state to be estimated. The width and height of the rectangle are depending on the average man height and width with a constant ratio.
4. Bayesian Formulation

We formulate the sequential tracking problem as computing the maximum a posteriori (MAP) $X^*$ such that $X^* = \arg \max_{X \in X} P(X | Y)$, where $X = (X_0, ..., X_t)$ is the state sequence and $Y = (Y_0, ..., Y_t)$ is the observation sequence.

A basic Bayesian sequential estimation can be described as two step recursion [21]

**Prediction step:**
$$p(x_t | y_{t-1}) = \int p(x_t | x_{t-1})p(x_{t-1} | y_{t-1})dx_{t-1}, \quad (1)$$

**Filtering step:**
$$p(x_t | y_t) = \frac{p(y_t | x_t)p(x_t | y_{t-1})}{p(y_t | y_{t-1})}, \quad (2)$$

The recursion requires the specification of a dynamic model describing the state evolution $p(x_t | x_{t-1})$, and a model for the likelihood in light of the current measurements $p(y_t | x_t)$. The recursion is initialized with some distribution for the initial state $p(x_0)$. Therefore, in this section, we give details about the state model, dynamic model and observation model in our tracking system.

4.1 State Model

An individual person’s state at time $t$ is modeled using a 5-dimensional state vector $x_{t,k} = [p^k_{t,s}, p^k_{t,b}, s]$, where $k$ is the person’s identifier, $p^k_{t,s}$ denote the barycenter position of the laser point cluster, while $p^k_{t,b}$ denote body position.

Each of them is one 2-dimensional vector. $s$ is the scale factor for each person.

4.2 Dynamic Model

In our tracking system, we use a constant velocity model which can be best described by a second order autoregressive equation
$$x_{t+k} = Ax_{t+k-1} + Bx_{t+k-2} + CN(0, \Sigma), \quad (3)$$

Matrices $A$, $B$, $C$ and $\Sigma$ could be learned from a set of representative sequences where correct tracks have been obtained in our experiment. $N(0, \Sigma)$ is a Gaussian noise with zero mean and standard deviation of 1.

4.3 Laser Based Observation Model

The observation of laser is the “human ellipse” which is composed by some point sets generated by clustering as shown in Figure 5.

In order to extract “human ellipse”, we firstly get foreground laser scan image (Figure 5) via background subtraction firstly. Then we use Parzen window for clustering to extract body of persons from discrete foreground points. Parzen window density estimation is a well-known non-parametric method to estimate distribution from sample data; we utilize it to convert discrete sample points to a continuous density function.

The general form of the density is
$$p(I) = \frac{1}{m} \sum_{i=1}^{m} G(I - I_i, h), \quad (4)$$

in which, $\{I_1, ..., I_m\}$ is a set of d-dimensional samples, $G(\cdot)$ is the Gaussian window function and $h$ is the window width parameter which is depending on the size of body region. The results of detected human body are shown in Figure 5.
The interaction among persons who are nearby causes missed detections for some persons. For instance, the “human ellipse” for two persons becomes one large ellipse because of the close distance between them. To address this issue, we use a potential function to constrain one person ellipse. For an isolated person whose state is \( x_t \), we evaluate the likelihood of the observation \( y_t \) given the cluster \( C \),

\[
P_l(x_t | x_{t-k}) = I(p_t) \varphi(p_{t-l}, p_{t-v}),
\]

(5)

where \( I(p) \) is the intensity at \( p \) of laser image after Parzen window process and \( \varphi(p_{t-l}, p_{t-v}) \) is the potential function:

\[
\varphi(p_{t-l}, p_{t-v}) = \begin{cases} 
\exp(-\frac{\max_{p_{t-l}, p_{t-v}} \| p_{t-l} - p_{t-v} \|}{w}), & \text{max}_{p_{t-l}, p_{t-v}} \| p_{t-l} - p_{t-v} \| > w, \\
1, & \text{otherwise} 
\end{cases}
\]

(6)

where \( p_{t-l} \) is the sample point in the cluster, \( \| p_{t-l} - p_{t-v} \| \) the distance between two sample points in the cluster, and \( w \) the distance limit depending on average human width.

4.4 Vision Based Observation Model

For the observation of vision, we adopt a multi-color model [22] based on Hue-Saturation-Value color histogram because the HSV color space can decouple chromatic information from shading effects. The observation of the person is represented by an \( N \)-bin color histogram extracted from the region \( R(p_{t-s}) \) of camera image (as shown in Figure 6). It is denoted as \( Q(p_{t-s}) = \{q(n; p_{t-s})\}_{n=1}^{N} \), where

\[
q(n; p_{t-s}) = K \sum_{k \in R(p_{t-s})} \delta(b(k) - n),
\]

(7)

where \( \delta \) is the Kronecker delta function to decide whether location \( k \) is in the region \( R(p_{t-s}) \). \( K \) is a normalization constant ensuring \( \sum_{n=1}^{N} q(n; p_{t-s}) = 1 \).

The data likelihood must favor candidate color histograms \( Q(p_{t-s}) \) close to the reference histogram \( \hat{Q}(p_{t-s}) \); we should choose a distance \( D \) on the HSV color distributions. In [22], \( D \) is derived from the Bhattacharyya similarity coefficient, and defined as

\[
D(q^*, q(\hat{p}_{t-s})) = \left[ 1 - \sum_{n=1}^{N} \sqrt{q^*(n; \hat{p}_{t-s}) q(n; p_{t-s})} \right]^2.
\]

(8)

The likelihood is then evaluated as

\[
P_v(y_t | x_{t-s}) \propto e^{-D(q^*, q_{t-s})},
\]

(9)

4.5 Joint Observation

We make the assumption that the laser-based observation is independent to the vision-based one. For an isolated person, the joint observation is

\[
P_{j}(y_t | x_{t-s}) \propto P_l(y_t | x_{t-k}) P_v(y_t | x_{t-s}),
\]

(10)
5. Probabilistic Detection-based Particle Filter

Particle filter has been a successful numerical approximation technique for Bayesian sequential estimation with non-linear, non-Gaussian models. This is due to its efficiency, flexibility and easiness of implementation. In the special context of multi-target tracking, most algorithms have been proposed using joint particle filters [1,3,5]. However, with the increasing number of tracking targets, these methods cannot obtain good approximation because of the limitation of sampling in high-dimensional state spaces. Although independent particle filters [6] can obtain a good approximation of posterior, it cannot maintain good tracking when interactions and occlusions occur. Hence, in our system, we propose a Probabilistic Detection-based Particle Filter (PD-PF), which incorporates probabilistic laser detections and information of dynamic models. Our method not only can obtain good approximation for each person, but also deal with the interactions and occlusions between persons.

5.1 Filtering

The basic idea behind the particle filter is very simple. Starting with a weighted set of samples \( \{ w_n^t, \mathbf{x}_n^t \} \), approximately distributed according to \( p(\mathbf{x}_{t-1} | y_{t-1}) \), new samples are generated from a suitably designed proposal distribution, which may depend on the old state and new measurements. To maintain a consistent sample the new importance weights are set to

\[
    w_n^{t+1} \propto w_n^t \frac{p(y_t | x_n^t) p(x_n^t | y_{t-1})}{q(x_n^t | x_{t-1}, y_t)}, \quad \sum_{n=1}^N w_n^{t+1} = 1, \quad (11)
\]

From time to time it is necessary to resample the particles to avoid degeneracy of the importance weights. The resample procedure essentially multiplies particles with high importance weights, and discards those with low importance weights, more details can be found in [23].

One of the crucial design issues in particle filter is the choice of the proposal distribution \( q(x_n^t | x_{t-1}, y_t) \). In our system, we run an independent particle filter for each person and the proposal distribution for each one is

\[
    q_k(x_n^t | x_{t-1}, y_t) = \sum_{j} \beta_k q_j(x_n^t | x_{t-1}, y_t) + (1 - \sum_j \beta_j) p_k(x_n^t | x_{t-1}) \tag{12}
\]

where \( q_j(x_n^t | x_{t-1}, y_t) \) is the information of detections which have influence on the target \( k \); \( \beta_k \) their weights and \( p_k(x_n^t | x_{t-1}) \) the dynamic model.

When the target does not interact with others, this mixed proposal has one detection or a high weighted detection, which is like to the simple detection-based method [6] (See Figure 7-a). When the interactions occur, the mixed proposal has more than one detections, these detections competed in this mixed proposal. Therefore, this proposal quite effectively models interactions and occlusions among targets (See Figure 7-b). For instance, when two similar appearance targets interacted on each other, the proposal for each target became a multiple peak distribution. Sampling particles in these proposals is like a blending step. On the other hand, although the two targets may share the same detections, the proposals of them are different to each other. There are two reasons: 1) the weights \( \beta_k \) are computing in a joint space (see next subsection), for each target, the weights of detections are different to each other. 2) The two targets have different information of dynamics. Therefore, for each target, particles will have high weights in its real state, and the posterior is approximate to its optimization. When two targets merge, we do not lose any target because a specific detection has high weight in respective proposal of each target.

Moreover, because each particle filter samples in a small space, we can obtain better approximation and significantly reduce computational cost than joint particle
filter. In addition, we computed $\beta_k$ in each recursion, thus our method can automatically adapt to the different complex tracking situations. In the next subsection, we will provide details about how to compute $\beta_k$.

5.2 Probabilistic Detections

The weight $\beta_k$ of detection should be a probability which reflects the similarity to the target. Thus, we call these detections “Probabilistic Detections”. In order to consider the interactions between targets, we use joint probabilistic model which is motivated by Joint Probabilistic Data Association (JPDA) [24] to compute $\beta_k$. We define an association event $\theta$ expressed as a vector with dimension $d_n$. Here $d_n$ is the number of detections. Each $\theta$ uniquely determines how each detection is assigned to a specific person. The vector can be drawn from a set of numbers as $\{0,1,2,...\}$ and $\theta_j = k$ means detection $j$ is from target $k$. Therefore, for each event $\theta$, the joint likelihood is

$$L(\theta) = \prod_{j=1}^{n} P_{\theta|\mathbf{x}}(y'_j | x^{i,j}_k) ,$$

Hence the $\beta_k$ is

$$\beta_k = \alpha \sum_{\theta \in \Theta_k} L(\theta) \times P_{\theta|\mathbf{x}} ,$$

where $\Theta_k$ is the set of joint association events that include all the case detection $j$ being from target $k$. $\alpha$ is a normalization factor ensuring that $\beta_k$ sums up to one over all $\theta$, and $P_{\theta|\mathbf{x}}$ prior knowledge which depends on the correctness of detections.

5.3 Algorithm Summary

Therefore, the Probabilistic Detection-based Particle Filter can be depicted as:

Initialization: $t=0$

1. Laser detecting to get all the persons $\{x_{k,0}\}_{k=1,2,...,m}$.

2. For all persons, create particle set $\{w^{i,j}_0, x^{i,j}_0\}_{i=1}$ by sampling from $p(x_{k,0})$.

For $t=1,...,T$

1. Proposal distribution construction
   a. For all persons, predict state by Eq. (3).
   b. Laser scanning to get detections.
   c. For all persons, compute joint likelihood by Eq.(10)
   d. For each person, find interacting detections by the 2-D statistical distance (circle or ellipse).
   e. Targets addition or removal.
   f. For each person, compute $\beta_k$ by Eq. (14)
   g. Targets removal.
   h. For all persons, construct proposal distribution by Eq. (12).

2. Filter
   a. Importance sampling by $x^{i,j}_t \sim q(x_t | x^{i,j}_t, y_t)$.
   b. Weight update by Eq. (11).
   c. Deterministic resample.

3. Output

Fig. 8: Tracking results of our system. The first row is the laser scanner data, and the second row is the tracking results with laser and vision. With the help of laser scanner, we can deal with the new coming pedestrians very easily (see frame 90 and frame 135). Our system can maintain the correct tracking where the shadow and occlusion frequently take place.
For each target, $E(x_{t,i}) = \sum_{j=1}^{\infty} w_{j}^{s} x_{j,i}$

6. Experimental Results

We evaluated our tracking system in a real scene at the park whose lighting conditions changed frequently. One laser scanner and one camera were utilized. Camera was set on the second floor of a building (about 5m high) and laser scanner was set on the tripod (about 120cm above the ground surface). Both of them captured the data at 15fps. We test our system with 8-minute long data. In this section, we present our tracking results and make some comparisons with other methods.

6.1 Tracking Performance

Figure 8 shows the tracking results of our system in strong lighting condition, where shadow and occlusion frequently take place. We can see that our system can deal with the new coming pedestrians very easily with the help of laser scanner. From frame 40 to 215, our system performs robust in real environment.

Figure 9 shows another example. The first row is the tracking results using laser and vision, and the second row is the results with camera only [25]. The system with no laser often fails because of interactions, occlusion or the influence of shadow and lighting. In the frame 341, interactions occurred between two similar appearance people, system with no laser failed to maintain the correct tracking (frame 345-360).

Figure 10 shows our tracking results under more complex situations. In this scene, the air is cloudy and there are a lot of crossovers, collisions and occlusions among multiple humans. However, our system still maintains good performance.

6.2 Quantitative Evaluation

We made a statistical survey of 200 continued frames and made a comparison among three methods: PD-PF, joint
particle filter [1] and independent particle filter [6]. The person number of these selected frames is variable, and the maximum number in these frames is nine. We define that if two rectangles of persons have the overlapping region, one interaction would be counted. Figure 11 shows the quantitative evaluation of three methods in the condition of interactions. Tick marks (vertical line) of Figure 11 (a) show the number of correct tracking. Figure 11 (b) shows the number of interactions in these frames. We can see that with the increasing number of interactions, independent particle filter cannot maintain correcting tracking, and our method have a better performance than joint particle filter.

6.3 Time-consuming Comparison

The three methods were all performed at PentiumIV 2.4G and 1000 joint particles and 100 particles per single filter were used. The time-consuming was shown in Figure 12. The tick marks (vertical line) was the average time-consuming per frame, and tick marks (horizontal line) was average target number per frame. We can see that with the increasing number of targets, the time-consuming of joint particle filter grew exponentially and our methods only slightly more time-consuming than independent particle filter because of the interactions between targets.

7. Conclusion

In this paper, we present a promising system to simultaneously detect and track humans in a real scene using laser and vision. Our contribution in this work is: 1) extract useful information from laser and vision, and combine them into a Bayesian framework; 2) propose an effective MAP estimation method—Probabilistic Detection-based Particle Filter (PD-PF). Experiments and evaluations show that our tracking system can perform robustly in a real scene; our proposed method can obtain better approximation than joint particle filter or independent particle filter under interacting situations and have an acceptable computing complexity.

In the future, this work could be extended as the following. 1) A prior probabilistic model on the correctness of laser detections should be embedded. 2) With the increasing number of tracking targets, a Markov chain Monte Carlo (MCMC) strategy for computing joint association probability should be considered.

Acknowledgments

This work was supported in part by the NKBPRC (No. 2006CB303100), NSFC Grant (No. 60333010), NSFC Grant (No. 60605001), the NHTRD 863 Grant (No. 2006AA01Z302), Key Grant Project of Chinese Ministry of Education (No. 103001) and NHTRD 863 Grant (2007AA11Z225).

References

Xuan Song Xuan Song received the BS degree in information engineering from Jilin University, China, in 2005. He now is a Ph.D. candidate in Key Laboratory of Machine Perception (Ministry of Education), Peking University. His research is mainly in computer vision and robotics, especially on multi-target tracking, sensor fusion and with application to intelligent surveillance.

Jinshi Cui Jinshi Cui received the BSc (1999) and Ph.D. (2004) degrees in computer science from Tsinghua University in Beijing, China. Since 2004, she has been a faculty in the department of Machine Intelligence in Peking University. She is now an associate professor in Peking University. Her current research interests are human body detection, tracking and pose estimation, with application to intelligent surveillance and driver assistance.

Huijing Zhao Huijing Zhao received B.S. degree in computer science in 1991 from Peking University, China. From 1991 to 1994, she was recruited by Peking University in a project of developing a GIS platform. She obtained M.E. degree in 1996 and Ph.D. degree in 1999 in civil engineering from the University of Tokyo, Japan. After several years’ post-doctoral research as the same university, in 2003, she was promoted to be a visiting associate professor in Center for Spatial Information Science, the University of Tokyo, Japan. In 2007, she joined Peking Univ. as a professor at the School of Electronics Engineering and Computer Science. Her research interest covers image sensing, computer vision and spatial data handling.

Hongbin Zha Hongbin Zha received the BE degree in electrical engineering from the Hefei University of Technology, China, in 1983 and the MS and PhD degrees in electrical engineering from Kyushu University, Japan, in 1987 and 1990, respectively. After working as a research associate at Kyushu Institute of Technology, he joined Kyushu University in 1991 as an associate professor. He was also a visiting professor in the Centre for Vision, Speech, and Signal Processing, Surrey University, Unite Kingdom, in 1999. Since 2000, he has been a professor at the State Key Laboratory of Machine Perception, Peking University, China. His research interests include computer vision, digital geometry processing, and robotics. He has published more than 200 technical publications in journals, books, and international conference proceedings. He received the Franklin V. Taylor Award from the IEEE Systems, Man, and Cybernetics Society in 1999.