External Cameras and a Mobile Robot for Enhanced Multi-person Tracking

A. A. Mekonnen[1,2], F. Lerasle[1,2] and A. Herbulot[1,2]
[1]*CNRS, LAAS, 7 avenue du Colonel Roche, F-31400 Toulouse, France*
[2]*Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France*

Keywords: Multi-target Tracking, Multi-sensor Fusion, Automated Human Detection, Cooperative Perception Systems.

Abstract: In this paper, we present a cooperative multi-person tracking system between external fixed-view wall mounted cameras and a mobile robot. The proposed system fuses visual detections from the external cameras and laser based detections from a mobile robot, in a centralized manner, employing a "tracking-by-detection" approach within a Particle Filtering scheme. The enhanced multi-person tracker's ability to track targets in the surveilled area distinctively is demonstrated through quantitative experiments.

1 INTRODUCTION

Automated multi-person detection and tracking are indispensable in video-surveillance, robotic and similar systems. Unfortunately, automated multi-person perception is very challenging due to variations in human appearance. These challenges are further amplified in robotic platforms due to mobility, limited Field-Of-View (FOV) of on-board sensors, and limited on-board computational resources. Relatively successful multi-person perception systems have been reported in classical video-surveillance frameworks that rely on visual sensors fixed in the environment (Hu et al., 2004). Even though these systems benefit from global perception from wall-mounted cameras, they are still susceptible to occlusions and dead-spots. To circumvent these shortcomings, we propose a cooperative multi-person perception system consisting of a mobile robot and two wall-mounted fixed-view cameras. This system benefits from the global perception of the wall-mounted cameras and additionally, from the mobile platform which provides local perception, a means for action, and as it can move around, the ability to cover dead spots and possibly alleviate occlusions resulting in enhanced perception capabilities. Similar systems have been proposed in (Chia et al., 2009) and (Chakravarty and Jarvis, 2009). Contrary to both works, our proposal fuses cooperative information in a centralized manner. The proposed system has the ability to complement local perception with global perception and vice-versa, enhancing each individual approach through cooperation. To the best of our knowledge this cooperative framework has not been addressed in the literature.

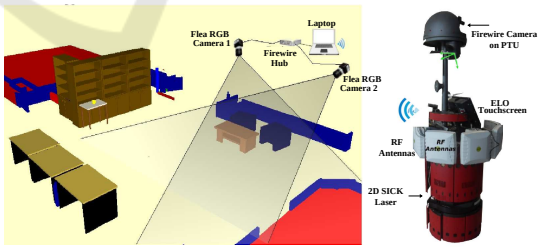

Figure 1: Perceptual platform; static cameras (with rough positions and fields of view) and Rackham.

This paper is structured as follows: architecture of the cooperative system is presented in section 2. Section 3 describes the different detection modalities that drive the multi-person tracker (presented in section 4). Evaluations and results are presented in section 5 followed by concluding remarks in section 6.

2 ARCHITECTURE

Our cooperative framework is made up of a mobile robot and two fixed view wall-mounted RGB flea2 cameras (figure 1). The cameras have a maximum resolution of 640x480 pixels and are connected to a dual-core Intel Centrino Laptop *via* a fire-wire cable. The robot, called Rackham, is an iRobot B21r mobile platform. It has various sensors, of which its SICK Laser

Mekonnen A., Lerasle F. and Herbulot A..
External Cameras and a Mobile Robot for Enhanced Multi-person Tracking.
DOI: 10.5220/0004294904110415
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 411-415
ISBN: 978-989-8565-48-8

Range Finder (LRF) is utilized in this work. Communication between the mobile robot and the computer hosting the cameras is accomplished through a wi-fi connection.

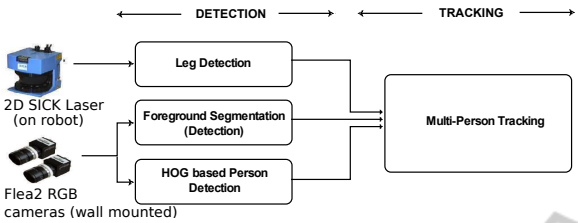

Figure 2: Multi-person detection and tracking system block diagram.

Figure 2 shows block diagram of the envisaged multi-person perceptual system. It has two main parts. The first part deals with automated multi-person detection. The second part is dedicated for multi-person tracking. It takes all detections as input and fuses them in a Particle Filtering framework. Each of these parts are discussed in detail in subsequent sections. It is worth mentioning here that the entire system is calibrated with respect to a global reference frame. Both the intrinsic and extrinsic parameters of the fixed cameras are known and in addition the mobile robot has localization module that localizes its pose with respect to the reference frame using laser scan segments.

3 MULTI-PERSON DETECTION

The perceptual functionalities of the entire system are based on various detections. The detection modules are responsible for automatically detecting persons in the area. Different person detection modalities are utilized depending on the data provided by each sensor.

Leg Detection with LRF: the LRF provides horizontal depth scans with a 180^o FOV and 0.5^o resolution at a height of $38cm$ above the ground. Person detection, hence, follows by segmenting leg patterns within the scan. In our implementation a set of geometric properties characteristic to human legs and outlined in (Xavier et al., 2005) are used.

Person Detection from Wall Mounted Cameras: to detect persons using the wall mounted cameras, two different modes are used. First, a foreground segmentation using a simple Σ-Δ background subtraction technique (Manzanera, 2007) is used. The mobile robot is masked out of the foreground images using its position from its localization module. Second, Histogram of Oriented Gradients (HOG) based person detection (Dalal and Triggs, 2005) is used. This method makes no assumption of any sort about the scene or the state of the camera (mobile or static). It detects persons in each frame using HOG features. Both detections are projected to yield ground positions, $(x,y)_G$ with associated color appearance information in the form of HSV histograms (Pérez et al., 2002), of individuals in the area.

4 MULTI-PERSON TRACKING

Multi-person tracking in our context, is concerned with the problem of tracking a variable number of persons, possibly interacting, in the ground plane. The literature in multi-target tracking contains different approaches but when it comes to tracking multiple interacting targets of varying number (Khan et al., 2005) has clearly shown that Reversible Jump Markov Chain Monte Carlo - Particle Filters (RJMCMC-PFs) are more appealing taking performance and computational requirements into consideration. Inspired by this, we have used RJMCMC-PF, adapted to our cooperative perceptual strategy, for multi-person tracking driven by the various heterogeneous detectors. The actual detectors are: the LRF based person detector, the foreground segmentation (detection) and HOG based detections from each wall mounted camera. Implementations choices crucial to any RJMCMC-PF are briefly discussed below.

State Space: the state vector of a person i in hypothesis n at time t is a vector encapsulating the id and (x,y) position of an individual on the ground plane with respect to a defined coordinate base, $x^n_{t,i} = \{Id_i, x^n_{t,i}, y^n_{t,i}\}$.

Proposal Moves: RJMCMC-PF accounts for the variability of the tracked targets by defining a variable dimension state space. Proposal moves propose a specific move on each iteration to guide this variable state space exploration. In our implementation, four sets of proposal moves, $m = \{\texttt{Add}, \texttt{Update}, \texttt{Remove}, \texttt{Swap}\}$, are used. The choice of the proposals privileged in each iteration is determined by q_m, the jump move distribution. These values are determined empirically and are set to $\{0.15, 0.8, 0.02, 0.03\}$ respectively. Equation 1 shows computation of the acceptance ratio, β, of a proposal X^* at the n^{th} iteration. It makes use of the jump move distribution, q_m; proposal move distribution, $Q_m()$, associated with each move; the observation likelihood, $\pi(X^n_t)$; and the interaction model, $\Psi(X^n_t)$.

$$\beta = \min\left(1, \frac{\pi(X^*)\, Q_{m*}\left(X_t^{n-1}|X^*\right) q_{m*}\Psi(X^*)}{\pi\left(X_t^{n-1}\right) Q_m\left(X^*|X_t^{n-1}\right) q_m \Psi\left(X_t^{n-1}\right)}\right) \quad (1)$$

where $m \in \{\texttt{Add}, \texttt{Update}, \texttt{Remove}, \texttt{Swap}\}$ and m^* denotes the reverse operation. `Update` and `Swap` moves are self reversible.

`Add`: the add move, randomly selects a detected person, x_p, from the pool of provided detections and appends its state vector on X_t^{n-1} resulting in a proposal state X^*. The proposal density driving the `Add` proposal, $Q_{Add}\left(X^*|X_t^{n-1}\right)$, is then computed according to equation 2.

$$Q_{Add}\left(X^*|X_t^{n-1}\right) = \sum_d k_d . \sum_{j=1}^{N_d} \mathcal{N}\left(x_p; z_{t,j}^d, \Sigma\right) . \left(1 - k_m \sum_{j=1}^{N_t} \mathcal{N}\left(x_p; \hat{X}_{t-1,j}, \Sigma\right)\right) \quad (2)$$

where d represents the set of detectors, namely: from laser (l), fixed camera 1 (c_1), and fixed camera 2 (c_2); $d \in \{l, c_1, c_2\}$, N_d the total number of detections in each detector, k_d is a weighting term for each detector such that $\sum_d k_d = 1$, N_t is the number of targets in the MAP, and k_m is a normalization constant. When a new person is added, its appearance is cross-checked with the appearance of persons that have been tracked for re-identification.

`Remove`: this move randomly selects a tracked person x_p from the particle being considered, X_t^{n-1}, and removes it, proposing a news state X^*. Contrary to the add move, the proposal density used when computing the acceptance ratio, $Q_{Remove}(X^*|X_t^{n-1})$ (equation 3), is given by the distribution map from the tracked persons masked by a map derived from the detected passers-by.

$$Q_{Remove}\left(X^*|X_t^{n-1}\right) = \left(1 - \sum_d k_d . \sum_{j=1}^{N_d} \mathcal{N}\left(x_p; z_{t,j}^d, \Sigma\right)\right) . \left(k_m \sum_{j=1}^{N_t} \mathcal{N}\left(x_p; \hat{X}_{t-1,j}, \Sigma\right)\right) \quad (3)$$

`Update`: here, the state vector of a randomly chosen passer-by is perturbed by a zero mean normal distribution. The update proposal density, $Q_{update}(X^*|X_t^{n-1})$, is a normal distribution with the position of the newly updated target as mean.

`Swap`: the swap move handles the possibility of id switches amongst near or interacting targets. When this move is selected, the ids of the two nearest tracked persons are swapped and a new hypothesis X^* is proposed. The acceptance ratio is computed similar to the `Update` move.

Interaction model ($\Psi(.)$): is used to maintain tracked person identity and penalize fitting of two trackers on the same object during interaction. A Markov Random Field (MRF), similar to (Khan et al., 2005), is adopted to address this.

Observation Likelihood ($\pi(.)$): the observation likelihood, in equation 1, is derived from all detector outputs except the laser for which blobs formed from the raw laser range data, denoted as l_b, are considered. If the specific proposal move is an `Update` or `Swap` move, a Bhattacharyya likelihood measure is also incorporated. Each detection is represented as a Gaussian, $\mathcal{N}(.)$, centered on the detection. Representing the measurement information at time t as z_t, the observation likelihood of the n^{th} particle X_t^n at time t is computed as shown in equation 4.

$$\begin{aligned}
\pi(X_t^n) &= \pi_B(X_t^n).\pi_D(X_t^n) \quad (4)\\
\pi_B(X_t^n) &= \begin{cases} \prod_{i=1}^{M}\prod_{c=1}^{2} e^{-\lambda B_{i,c}^2} & \text{, if } m = \texttt{Update or Swap} \\ 1 & \text{, otherwise} \end{cases}\\
\pi_D(X_t^n) &= \frac{1}{M}\sum_{i=1}^{M}\left(\sum_d k_d.\pi\left(x_i|z_t^d\right)\right), \sum_d k_d = 1\\
\pi\left(x_i|z_t^d\right) &= \frac{1}{N_d}\sum_{j=1}^{N_d}\mathcal{N}\left(x_i; z_{t,j}^d, \Sigma\right)
\end{aligned}$$

Above, B_i represents the Bhattacharyya distance computed between the appearance histogram of a proposed target i in particle X_t^n and the target model in each camera c. M represents the number of targets in the particle, and N_d the total number of detections in each detection modality d, $d = \{l_b, c_1, c_2\}$, in this case including the measures from the laser blobs. k_d is a weight assigned to each detection modality taking their respective accuracy into consideration and x_i represents the position of target i in the ground plane.

5 EVALUATIONS AND RESULTS

To evaluate the performance of our RJMCMC-PF multi-person tracker, three sequences acquired using Rackham and the wall mounted cameras are used. Each sequence contains a laser scan and video stream from both cameras. Sequence I and II contain 200 frames each and consist of two and three targets consecutively. Sequence III is 186 frames long containing four targets moving in the vicinity of the robot. The evaluation is carried out using the CLEAR MOT metrics (Bernardin and Stiefelhagen, 2008), Multiple Object Tracking Accuracy (MOTA) and Precision (MOTP). To clearly observe the advantages of each

Table 1: Multi-person tracking evaluation results.

Sequence	Laser-only				Fixed Cameras only				Cooperative			
	MOTP		MOTA		MOTP		MOTA		MOTP		MOTA	
	μ	σ	μ	σ	μ	σ	μ	σ				
I	15.62	2.34	0.41	0.05	19.80	0.14	0.79	0.03	17.01	1.87	0.84	0.03
II	19.90	1.66	0.27	0.07	22.79	1.35	0.70	0.05	17.73	0.50	0.79	0.03
III	21.94	1.75	0.20	0.07	28.44	1.60	0.46	0.07	21.30	1.34	0.54	0.04

sensor modality, the evaluation is carried out by doing the tracking using (1) laser-only information, (2) vision-only data from the two wall mounted cameras, and finally (3) laser and the two cameras cooperatively. A hand labeled ground truth with (x, y) ground positions and unique id for each person is used in the evaluation. Each sequence is run eight times to account for the stochastic nature of the filter. Results are reported as mean value and associated standard deviation in table 1.

The results presented in table 1 clearly attest the improvements in perception brought by the cooperative fusion of laser and wall mounted camera percept. The cooperative system consisting of laser and two wall mounted cameras exhibit an MOTA of 0.841 when tracking two targets, 0.793 for three targets. These results clearly indicate the enhanced performance of this system. Sample tracking sequences from sequence II are shown in figure 3[1]. Evidently, the LRF-only has low accuracy owing to the mistakes made with leg like structures in the environment, sensitivity to occlusion, and lack of discriminating information amongst tracked passers-by. The results obtained using the wall mounted cameras show major improvements though their position tracking precision is relatively lower compared to those which include laser measurement. The final tracker runs at $1 fps$. Most of the computation time, $\approx 700ms$, is spent on HOG based person detection.

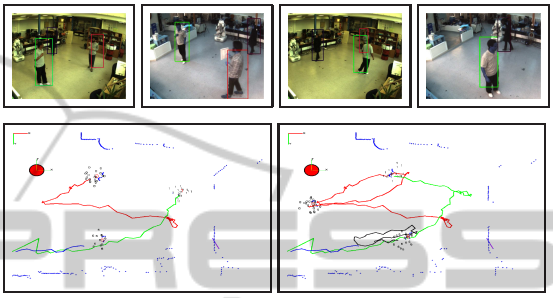

Figure 3: Multi-person tracking illustrations taken from sequence II at a) frame 60, and b) frame 94.The top row images show camera streams and the bottom shows the ground floor with tracked persons' trajectories superimposed [1].

6 CONCLUSIONS

The work presented herewith makes its main contribution in the vein of multi-person tracking by proposing a cooperative scheme between overhead cameras and sensors embedded on a mobile robot in order to track people in crowds. Our Bayesian data fusion framework with the given sensor configuration enhances typical surveillance systems with only fixed cameras and complete embedded systems without wide FOV and straightforward (re)-initialization ability. The presented results are a clear indication of the framework's notable tracking performance.

[1]For complete run, visit the URL homepages.laas.fr/aamekonn/videos/

REFERENCES

Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008:1:1–1:10.

Chakravarty, P. and Jarvis, R. (2009). External cameras and a mobile robot: A collaborative surveillance system. In *Australasian Conf. on Robotics and Automation (ACRA'09)*, Sydney, Australia.

Chia, C., Chan, W., and Chien, S. (2009). Cooperative surveillance system with fixed camera object localization and mobile robot target tracking. In *Advances in Image and Video Technology*, volume 5414, pages 886–897. Springer Berlin / Heidelberg.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA.

Hu, W., Tan, T., Wang, L., and Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst., Man, Cybern., Syst., Part C: Applications and Reviews*, 34(3):334 –352.

Khan, Z., Balch, T., and Dellaert, T. (2005). Mcmc-based particle filtering for tracking a variable number of in-

teracting targets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(11):1805–1918.

Manzanera, A. (2007). Sigma - delta background subtraction and the zipf law. In *Iberoamericann Congress on Pattern Recognition (CIARP'07)*, Valparaiso, Chile.

Pérez, P., Hue, C., Vermaak, J., and Gangnet, M. (2002). Color-based probabilistic tracking. In *Europ. Conf. on Computer Vision (ECCV'02)*, Copenhagen, Denmark.

Xavier, J., Pacheco, M., Castro, D., and Ruano, A. (April, 2005). Fast line, arc/circle and leg detection from laser scan data in a player driver. In *Int. Conf. on Robotics and Automation (ICRA'05)*, Barcelona, Spain.