Intention-Aware Multi-Human Tracking for Human-Robot Interaction via Particle Filtering over Sets

Aijun Bai
Univ. of Sci. & Tech. of China
baj@mail.ustc.edu.cn

Reid Simmons
Carnegie Mellon Univ.
reids@cs.cmu.edu

Manuela Veloso
Carnegie Mellon Univ.
mmv@cs.cmu.edu

Xiaoping Chen
Univ. of Sci. & Tech. of China
xpchen@ustc.edu.cn

The Approach

The ability for an autonomous robot to track and identify multiple humans and understand their intentions is crucial for socialized human-robot interactions in dynamic environments (Michalowski and Simmons 2006). Take CoBot (Rosenthal, Biswas, and Veloso 2010) trying to enter an elevator as an example. When the elevator door opens, suppose there are multiple humans occupied, CoBot needs to track each human's state and intention in terms of whether he/she is going to exit the elevator or not. For the purposes of safely and friendly interacting with humans, CoBot can only make the decision to enter the elevator when any human who intends to exit is believed to have exited.

Most multi-object tracking (MOT) methods follow a *tracking-by-detection* paradigm (Yilmaz, Javed, and Shah 2006; Andriluka, Roth, and Schiele 2008). In this setting, an object detector runs on each frame to obtain a set of detections as inputs for a tracker. Tracking-by-detection algorithms can be roughly classified into two groups: *online* and *offline*. Online tracking intends to recursively estimate the current situation given past observations in a Bayesian way; offline tracking is typically formulated as a global optimization problem to find optimal paths given the whole sequence of observations. In this paper, we focus on online tracking which is more suitable for applications on robots.

Provided with an inevitable imperfect human detector (with false and missing detections occasionally), we model the intention-aware online multi-human tracking problem as a hidden Markov model (HMM). Formally, a joint state S is defined as a set of all humans, $S = \{h_i\}_{i=1:|S|}$, where a human is represented as a high-dimensional vector $h = (s, i)$. Here $s = (x, y, \dot{x}, \dot{y})$ is the physical state, and $i \in \mathcal{I}$ is an intention. In our experiments, we introduce a *moving* intention to move to a potential goal, and a *staying* intention to stay almost in the same area.

In MOT domain, most existing approaches assume one or more hypotheses of data associations between observations and targets, and apply Bayesian filtering on each target separately (Fortmann, Bar-Shalom, and Scheffe 1983; Reid 1979; Yilmaz, Javed, and Shah 2006). It is difficult for these methods to recover from wrong assumptions. Our approach avoids directly performing observation-to-target association, by using a joint state to encode the entire multi-target state including the number of targets, the state of each target, and implicitly all possible hypotheses.

Motion Model with Intention. We associate human intentions with different motion models. An intention $i \in \mathcal{I}$ gives the probability $\Pr(a \mid s, i)$ of choosing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. We define $\Pr(i' \mid s, i)$ as the probability of changing from intention i to intention i' in state s. The joint intention-aware motion model for a single human is then given as

$$\Pr(s', i' \mid s, i) = \sum_{a \in \mathcal{A}} \Pr(s' \mid s, a) \Pr(a \mid s, i') \Pr(i' \mid s, i). \quad (1)$$

Observation Model. An observation is defined as a set of detections (i.e. reported human positions), $O = \{(x_i, y_i)\}_{i=1:|O|}$. We assume that false and missing detections are mutually independent, and the numbers of false and missing detections are following Poisson processes with parameters ν and $|S|\xi$ per second respectively. Let $F = \{(x_i, y_i)\}_{i=1:|F|} \subseteq O$ and $M = \{h_i\}_{i=1:|M|} \subseteq S$ be the set of false and missing detections, and let $O \circ S = \{\langle F_i, M_i \rangle\}_{i=1:|O \circ S|}$ be the set of all possible F-M pairs, satisfying $|O - F| = |S - M|$. Suppose the update time interval is τ, the observation function that gives the probability of observing set O given state S is

$$\Pr(O \mid S) = \sum_{\langle F, M \rangle \in O \circ S} \Pr(O - F \mid S - M) \cdot (\nu\tau)^{|F|} e^{-\nu\tau} \prod_{o \in F} P_f(o) \frac{(|S|\xi\tau)^{|M|} e^{-|S|\xi\tau}}{|M|!} \frac{1}{\binom{|S|}{|M|}}, \quad (2)$$

where P_f is a background distribution stating that a false detection occurring in position $o = (x, y)$ has a probability density of $P_f(o)$. Notice that, the probability density $\Pr(O \mid S)$ is defined over sets.

Let Ψ_{S-M}^{O-F} be the set of all possible assignments from $S - M$ to $O - F$, and assume conditional independence between observations, we have

$$\Pr(O-F \mid S-M) = \sum_{\psi \in \Psi_{S-M}^{O-F}} \prod_{h \in S-M} P_o(\psi(h) \mid h), \quad (3)$$

where $P_o(o \mid h)$ gives the probability density of observing o given human state h. The resulting full expression of Equation 2 has $\Omega((\frac{\max\{|O|,|S|\}}{e})^{\min\{|O|,|S|\}})$ terms. Approximations are made in practice.

Particle Filtering. To infer in the resulting HMM, we propose a probabilistic approach named particle filtering over sets (PFS), which avoids directly performing observation-to-target association by using a set as a joint state to encode the entire multi-human state.

Particle Refinement. For particle X, let $\tilde{X}$ be the directly proposed particle based on motion model. For observation $o \in O$, if it does not match well with existing humans in $\tilde{X}$, then a refined proposal $\hat{X} \leftarrow \tilde{X} \cup h$, where $h \sim \Pr(\cdot \mid o)$, will probably result a better proposal in terms of o. The proposal and motion weights are then approximated by using a Bayesian density estimation method developed for sets.

Human Identification. Although a set of updated particles encodes completely the joint posterior distribution, a human identification process is developed to report each individual human from updated particles, which is more useful for high-level tasks. We perform data associations within each particle, link human states to observations, reduce the problem to a linear number of best assignment sub-problems, and repeat this process until converged or a maximal number of steps is reached in an expectation-maximization (EM) way.

Experiments

To compare with existing MOT algorithms, we evaluate PFS with only a random-walk motion model in the S2L1 sequence of the challenging PETS2009 dataset (Ferryman and Shahrokni 2009). The experimental results are shown in Table 1, evaluated in terms of CLEAR MOT metrics with distance threshold of 1.0m (Bernardin and Stiefelhagen 2008). We also report IDS which is the number of identity switches. Only identified humans with confidence higher than 0.4 are considered for evaluation. Segal and Reid (2013) model MOT as a Switch Linear Dynamical System and take advantage of a trained pedestrian and outlier detector in the target PETS dataset. Zamir, Dehghan, and Shah (2012) utilize Generalized Minimum Clique Graphs to solve the data association problem by incorporating both motion and appearance information. Milan (2014) formulate MOT as an optimization problem over splines given an energy function. Breitenstein et al. (2011) track each human separately with greedy data association in a particle filtering framework.[1]

Algorithm	MOTA	MOTP	IDS
PFS[1] (**proposed**)	0.93	0.76	3.6
PFS[1 2] (**proposed**)	0.91	0.75	4.8
Milan (2014)	0.90	0.80	11
Segal and Reid (2013)	0.92	0.75	4
Segal and Reid[2] (2013)	0.90	0.75	6
Zamir, Dehghan, and Shah[2] (2012)	0.90	0.69	8
Breitenstein et al.[2] (2011)	0.56	0.80	-

[1]averaged over 16 runs.
[2]evaluated within tracking region not cropped.

Table 1: Results in PETS2009 S2L1 dataset.

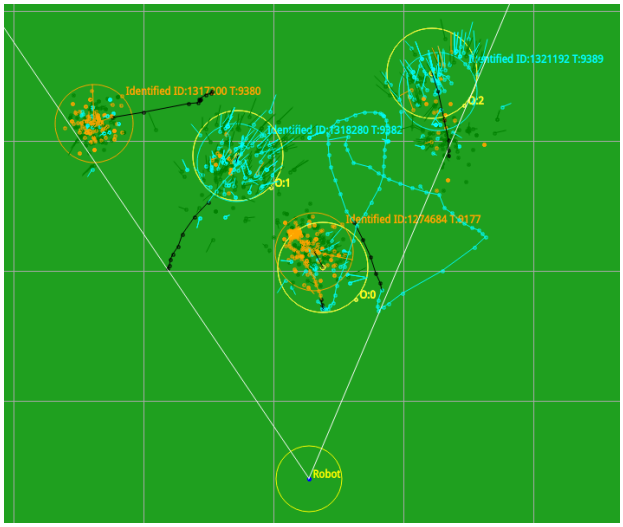

Figure 1: A snapshot of running PFS.

We conduct the real robot experiments on CoBot, by deploying the robot to be tasked by users in a university building (Biswas and Veloso 2013). A short video showing the running PFS taken from one deployment can be found via http://goo.gl/K70inP (13.7MB). A snapshot is also shown in Figure 1. The robot is depicted by a yellow circle labeled with "Robot". Observations from the human detector are represented as yellow circles labeled with sequential numbers, for example "O:0". Each identified human is illustrated as a circle for the current state and linked points for respective trajectory. Dominant intentions associated with states and trajectories are shown with different colors: cyan for *moving* intention and orange for *staying* intention. Black means no dominant intention classified yet.[2]

[1]A short video showing the whole results compared with ground truth is available at http://goo.gl/35UOy1.

[2]A video showing CoBot following a human for approximately 10 minutes based on the results provided by PFS can also be found at http://goo.gl/T3UtlS.

Conclusion

We present a novel particle filtering over sets (PFS) approach to the intention-aware multi-human tracking problem in the domain of human-robot interaction. From a multi-object tracking point of view, our approach avoids directly performing observation-to-target association by using a set formulation. The experiments in PETS2009 dataset show that the overall tracking performance is robust. The real robot experiments indicate that a robot integrated with this approach is able to track humans, and understand their intentions in terms of *moving* and *staying*. In future work, we plan to apply PFS in complex social tasks with multiple intentions.

Acknowledgments

This work is supported in part by the National Hi-Tech Project of China under grant 2008AA01Z150 and the Natural Science Foundation of China under grant 60745002 and 61175057. The authors want to thank Joydeep Biswas and Brian Coltin for their help on setting up CoBot.

References

Andriluka, M.; Roth, S.; and Schiele, B. 2008. People-tracking-by-detection and people-detection-by-tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

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

Biswas, J., and Veloso, M. M. 2013. Localization and navigation of the cobots over long-term deployments. *The International Journal of Robotics Research* 32(14):1679–1694.

Breitenstein, M. D.; Reichlin, F.; Leibe, B.; Koller-Meier, E.; and Van Gool, L. 2011. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(9):1820–1833.

Ferryman, J., and Shahrokni, A. 2009. PETS2009: Dataset and challenge. In *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, 1–6. IEEE.

Fortmann, T. E.; Bar-Shalom, Y.; and Scheffe, M. 1983. Sonar tracking of multiple targets using joint probabilistic data association. *Oceanic Engineering, IEEE Journal of* 8(3):173–184.

Michalowski, M. P., and Simmons, R. 2006. Multimodal person tracking and attention classification. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 347–358. ACM.

Milan, A. 2014. *Energy Minimization for Multiple Object Tracking*. PhD, TU Darmstadt, Darmstadt.

Reid, D. B. 1979. An algorithm for tracking multiple targets. *Automatic Control, IEEE Transactions on* 24(6):843–854.

Rosenthal, S.; Biswas, J.; and Veloso, M. 2010. An effective personal mobile robot agent through symbiotic human-robot interaction. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, 915–922. International Foundation for Autonomous Agents and Multiagent Systems.

Segal, A. V., and Reid, I. 2013. Latent data association: Bayesian model selection for multi-target tracking. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2904–2911. IEEE.

Yilmaz, A.; Javed, O.; and Shah, M. 2006. Object tracking: A survey. *Acm computing surveys (CSUR)* 38(4):13.

Zamir, A. R.; Dehghan, A.; and Shah, M. 2012. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *Computer Vision–ECCV 2012*. Springer. 343–356.