Online semi-supervised multi-person tracking with gaussian process regression

Baobing Zhang[*], *Zhengwen* Huang , *Babak H.* Rahi, *Qicong* Wang, and *Maozhen* Li

Department of Electronic and Computer Engineering, Brunel University London, Uxbridge, UB8 3PH, UK

Abstract. Most existing multi-person tracking approaches are affected by lighting condition, pedestrian pose change abruptly, scale changes, real-time processing to name a few, resulting in detection error, drift and other issues. To cope with this challenge, we propose an enhanced multi-person framework by introducing a new observation model, which adaptively updates fully online to avoid the loss of sample diversity and learning in a semi-supervised manner. We fuse prior information for tracking decision, meanwhile extracted knowledge from current frame is used to assist to make tracking decision, which can be viewed as a transfer learning strategy, and both aspects can ameliorate the tendency to drift. The new approach does not need any calibration or batch processing. Experimental results show that the approach yields comparable or better performance in comparison with the state-of-the-arts, which do calibration or batch processing.

1 Introduction

Tracking-by-detection approach is very popular during recent years [1] [2]. In practice most tracking-by-detection approaches are still limited to special scenarios and affected by occlusion, scale change, real-time processing etc. Moreover harmed by false and missing detection, some methods employ occlusion reasoning to smoothing the trajectories [3]. However, these methods are sensitive to detection error, because they build the trajectory based on two consecutive frames. Thus, during long-term occlusion or abrupt changes in pose, a danger is the tracked target trends to drift.

To deal with this problem, dynamic and observation model are combined used for the tracking problem. Dynamic model take pedestrian behaviour into account, moreover often used for estimating the new location of pedestrian. However, most existing dynamic model use only the previous one state information for predicting and lack of utilizing prior information, when the pedestrian motion change abruptly tend to incorrect estimation. Observation model represents the pedestrian's appearance change, particularly when adapted online take account for the gradual appearance change. Most of the existing observation models gather past appearance information over time, but in fact, these methods lack utilize current state information of the pedestrian's appearance, inevitably lead to drift problem.

[*] Corresponding author: Baobing.Zhang@brunel.ac.uk

In this paper, we introduce a new observation model to cope with these problems mentioned above in several aspects. First, we fuse prior information for tracking decisions. Second, the observation model is learnt in a semi-supervised manner through using both labelled and unlabelled sample. Third, background information is taking into consideration in the process of observation model updating. Fourth, re-weighting knowledge is used for tracking decision can be viewed as a transfer learning strategy. All the aspects mentioned above tend to alleviate drift. The main contributions of our work are:

- The new observation model, updated adaptively avoids the loss of sample diversity and learnt in a semi-supervised manner.
- We extract re-weighting knowledge from the current pedestrian status information and used for tracking inference, can be viewed as a transfer learning strategy.

We test our method on two multi-person tracking benchmark sequences. Our method achieved promising results better than previously tested state-of-the-arts. The rest of the paper is organized as follows. Section 2 presents the new observation model used for tracking. Section 3 evaluates the performance of the observation model in comparison with a number of typical methods. Section 4 concludes the paper and points out some future work.

2 Tracking with Gaussian process regression

2.1 New Observation model

In this section, we present the process of tracking. At each frame f_t, the current location information of a tracker is stored in a bounding box $X_t = [x_t^i, y_t^i, w_t^i, h_t^i,]$, which is estimated by the observatio T_i n model based on the previous information. Once a tracker is initialized, $\mathcal{I}_t^{T_i} = \{\mathbf{I}_1, \cdots, \mathbf{I}_t\}$ is a set of observation images, we aim to inference the hidden state variable $X_t^{T_i}$ in a Markov model which is given by equation (1).

$$p\left(X_t^{T_i} \middle| \mathcal{I}_t^{T_i}\right) \propto p\left(\mathbf{I}_t \middle| X_t^{T_i}\right) \int p\left(X_t^{T_i} \middle| X_{t-1}^{T_i}\right) p\left(X_{t-1}^{T_i} \middle| \mathcal{I}_{t-1}^{T_i}\right) dX_{t-1}^{T_i} \tag{1}$$

We use Kalman filter's prediction as input to stochastically generate a set of pedestrian candidate location in the current frame, which is $\mathcal{X}_U^{T_i} = \{X_t^{T_i,j}, j = 1, 2, ..., n_U\}$. Tracking results of tracker T_i can be estimated by MAP as shown in equation (2).

$$X_t^{\hat{T}_i} = \underset{X_t^{T_i,j}}{\arg\max}\, \mathrm{P}\left(X_t^{T_i,j} \middle| \mathcal{I}_t^{T_i}\right) \tag{2}$$

For each sample, we introduce an indicator variable $y_j \in \{-1, +1\}$ to indicate positive sample $(y_j = +1)$ or negative $(y_j = -1)$ of $X_t^{T_i,j}$. $\mathcal{X}_U^{T_i}$ is the unlabeled sample set of tracker T_i. From the tracking results up to the (t-1)-th frame for each tracker, we extract n_L labelled training samples with indicator variables, and then we divide n_L labelled training samples into two groups. First is $\mathcal{D}_T$ which consist n_T samples, refers target sample set gathered from the most recent frame, the other is $\mathcal{D}_A$ which consist n_A samples, we called auxiliary sample set which collected every few intervals, and $n_L = n_T + n_A$. Then, the regression function for the indicators of unlabelled samples y_U can be written as equation (3).

$$\mathcal{R} = P\left(y_U = 1 \middle| \mathcal{X}_U, \mathcal{D}_A, \mathcal{D}_T\right) \tag{3}$$

We introduce two real valued latent vectors l_A and l_U , corresponding to the label y_A and y_U respectively. We connect regression and classification by using a sigmoid output model. The Gaussian process model restricted to the auxiliary data and unlabelled data is as shown in equation (4).

$$P\left(l_A, l_U \middle| \mathcal{X}_A, \mathcal{X}_U, \mathcal{D}_T\right) \sim \mathcal{N}\left(\mu, \widetilde{\Delta}^{-1}\right) \tag{4}$$

2.2 Graph Laplaacians

We construct the prior covariance matrix based on the weighted graph $\mathcal{G} = \left(V, E\right)$, which has the node set *V*and edge *E*, corresponding to all samples in the way similar with [4]. We explore the manifold structure of all samples. Furthermore we define weight matrix *W* of graph $\mathcal{G}$ using the method proposed by [5]. Finally prior covariance matrix is defined by the inverse graph Laplacian $\widetilde{\Delta}^{-1}$.

Because of the sigmoid noise label output model, the $P\left(l_A, l_U \middle| \mathcal{X}_U, \mathcal{D}_A, \mathcal{D}_T\right)$ is no longer Gaussian and has no closed form solution. Assuming $P\left(l_A, l_U \middle| \mathcal{X}_U, \mathcal{D}_A, \mathcal{D}_T\right)$ is a uni-modal function, we use its Laplace approximation to get the optimal estimation of l_A and l_U .

Because we construct prior covariance matrix depending on all samples, the correlated structure of the labelled samples and unlabelled samples has a significant effect on the latent real-valued output. The latent variable l_A is the re-weighting knowledge extracted from the Regression can be a soft replacement of indicator label y_A , and is better for ameliorating sample misalignment problem, less sensitive to noisy compare with the indicator variable.

2.3 Tracker's birth and death

For the purpose of maintaining the tracker, we divided trackers into two groups based on the template it owes. Once a tracker is born we call it Novice, it will accumulate templates throughout the tracking process, after **K** template accumulated over a period of robust tacking. Novice would be promoted to Expert, conversely an Expert demoted to a Novice when it loses template less than **K**, we set K to 5. Each tracker keep at most N$_{\max}$ reliable templates by discarding the lower score template, we set N$_{\max}$ to 10.

A tracker candidate is activated when its detection rate is above the ξ_{init} . On the contrary, a tracker would be killed when its detection rate is less than ξ_{term} . Both of them is given by in equation (5) and (6).

$$\xi_{\mathrm{term}} = \Omega - \gamma_2\sqrt{\Omega} \quad \xi_{init} = \Omega - \gamma_1\sqrt{\Omega} \tag{5}\tag{6}$$

where γ_1 and γ_2 is the scale factor, we set to 1 and 2 respectively, for each tracker's detection rate is defined as equation (7).

$$\Omega_i = \frac{\Delta N_i^{matched}}{\Delta t} \tag{7}$$

where $\Delta N_i^{matched}$ is the number of detections matched with T_i in a sliding window of length Δt

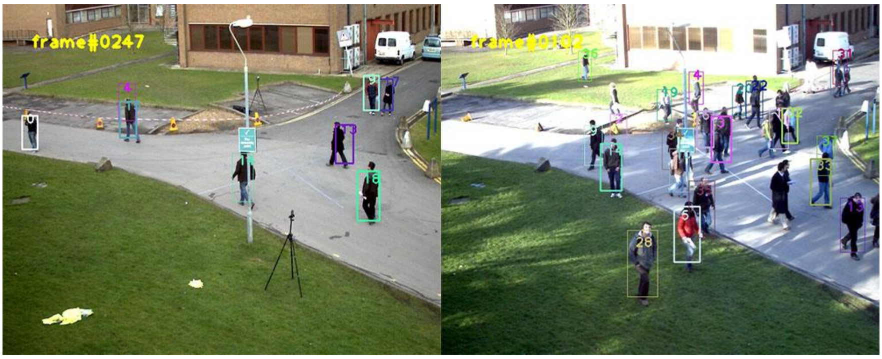

Fig. 1. Representative tracking results of PET09 sequence.

3 Experiments

Through several experiments, compared with the state-of-the-arts, our approach shows his unique advantages.

3.1 Datasets and Ground Truth

We currently test our algorithm on two sequences, one is the sequence S2L1, which is taken from the VS-PETS benchmark 2009 [6], representative tracking results can be seen in Fig 1. This sequence is filmed by 7 cameras and show up to 8 people the resolution is (768×576), we only use the first viewpoint. Most people wear similar dark clothes, which make colour-based observation model for tracking difficult. We also test the new algorithm under a crowed environment (S2L2), which has a lot of pedestrians within a confined space make tracking difficulty even for the individual detect. A brief description for these two sequences is as shown in Table 1, the ground-truth used for evaluation is public available[1].

Table 1. Datasets description.

	Frame Rate	Number of Frames	Number of Id	HOG Detector	
				Precision	Recall
PETS2009 S2L1	7	795	19	0.87	0.81
PETS2009 S2L2	7	436	43	0.90	0.59

3.2 Experimental Environments

All experiments were tested on a computer with 2.8GHz Octa-core CPU, 16GB memory. We use C++ implementation and rely on the OpenCV and Eigen library. Runtime performance about 2 fps per second with the new observation model employed for tracking. We believe that with GPU implementation or more optimized code could achieve real-time performance.

3.3 Evaluation Metrics

There is no standard established protocol to measure multi-object tracking performance, we use the current best practice which calculates the CLEAR-MOT metrics proposed in [7], FP means false positive, MS considers number of missed detections and ID.S takes account for

[1] http://www.gris.informatik.tu-darmstadt.de/%CB%9Caandriye/data.html

the switches of identities. The multiple objects tracking accuracy (MOTA) was defined as equation (8).

$$MOTA = 1 - \frac{\sum_{t=1}^{N_{frames}} \left(c_m(m_t) + c_f(fp_t) + c_s(ID - SWITCHES_t) \right)}{\sum_{t=1}^{N_{framees}} N_G^{(t)}} \tag{8}$$

The multiple objects tracking precision (MOTP) was given by equation (9).

$$MOTP = \frac{\sum_{i=1}^{N_{mapped}} \sum_{t=1}^{N_{frames}^{(t)}} \left[\left| \frac{G_i^{(t)} \cap D_i^{(t)}}{G_i^{(t)} \cup D_i^{(t)}} \right| \right]}{\sum_{t=1}^{N_{framees}} N_{mapped}^{(t)}} \tag{9}$$

Note that both MOTA and MOTP, higher values of the output indicate better performance (see [7] for detail).

Table 2. Results Comparison.

Sequence	Method	MOTP(%)	MOTA(%)	FP	MS	ID.S
PETS2009-S2L1	Ours	62.53	67.69	537	839	126
	[8]	68.22	60.27	28	1804	15
	[9]	-	67	-	-	-
PETS2009-S2L2	Ours	52.97	49.45	664	4118	420
	[8]	61.34	24.46	31	7696	47

As shown in table 2, we compare our method with [8] on the PETS2009 sequence the results of [8] is tested by ourselves, for fair comparison we use the same detector as our method and same protocol to evaluate the outputs. Note that the results of [8] are slightly different from the original paper. It may be influenced by the parameter tuning, pre-treatment optimization and other factors. We also show the results in [9] when available. As shown method [9] made the calibration action and perform batch processing of the data, our method achieve higher MOTA score success surpass both [8] and [9] on the S2L1 and S2L2 sequence. Compare with [8] our method have less missing detection and potentially increases the number of false positive. We believe the slightly lower MOTP score was caused by the update of the sample set not perfectly adapt the scale change over time. We also noticed that with the increase of density of people in the sense there are few veterans the ratio of veterans is much higher in sequence S2L1 than S2L2, it can be explained by there is more occlusion issues in S2L2 than S2L1.

Table 3. Methods Comparison.

Sequence	Tracker	Rcll(%)	Prcn(%)
PETS2009-S2L1	[10]	83.9783	80.9783
	Ours	81.957	87.6495
	[11]	81.6559	83.4139
	[12]	81.2903	84.2434
	[13]	75.1828	91.5663
PETS2009-S2L2	[14]	54.8387	90.4224
	Ours	59.9883	90.2896
	[15]	49.611	82.4797
	[16]	38.2014	82.045
	[17]	50.8039	81.3081
	[18]	35.5461	78.3673

Table 4. Results Comparison with the most recent state-of-the-arts.

Sequence	Tracker	FP	IDSW	MOTA(%)	MOTP(%)
PETS2009 -S2L1	[10]	910	348	56.2581	71.119
	Ours	537	126	67.6989	62.5369
	[11]	755	31	64.7527	70.1903
	[12]	707	239	60.9462	71.1903
	[13]	322	99	66.129	71.5962
PETS2009 -S2L2	[14]	560	238	46.5616	67.6273
	Ours	664	420	49.4559	52.9748
	[15]	1016	139	37.631	65.9038
	[16]	806	240	27.3519	67.361
	[17]	1126	190	37.1538	67.6956
	[18]	946	162	24.0535	67.5983

We further compare our method with most recently state-of-the-arts [10] [11] [12] [13] [14] [15] [16] [17] [18], the comparison results are as shown in Table 3 and Table 4. Compare with other methods, the Recall rate and Precision rate of our method success surpass most methods in both S2L1 and S2L2 sequence. We also got the best MOTA score compare with other methods. Because the tracking decision mainly depend on the observation model, the observation model update process has a significant impact on MOTP score, we will do further optimization to improve the MOTP performance. The speed of this algorithm is proportional to the sample sampling size. While reducing the number of samples, we also consider the balance of accuracy of tracking and algorithm speed. In order to test the sensitivity of our algorithm to parameters, we conduct experiments which parameter setting from the baseline each parameter floating up and down 40% while keeping other parameters fixed. As shown in figure 2, the performance of our algorithm changes within a reasonable range for all the sequence. This indicates that our algorithm is relatively robust to the setting of parameters.

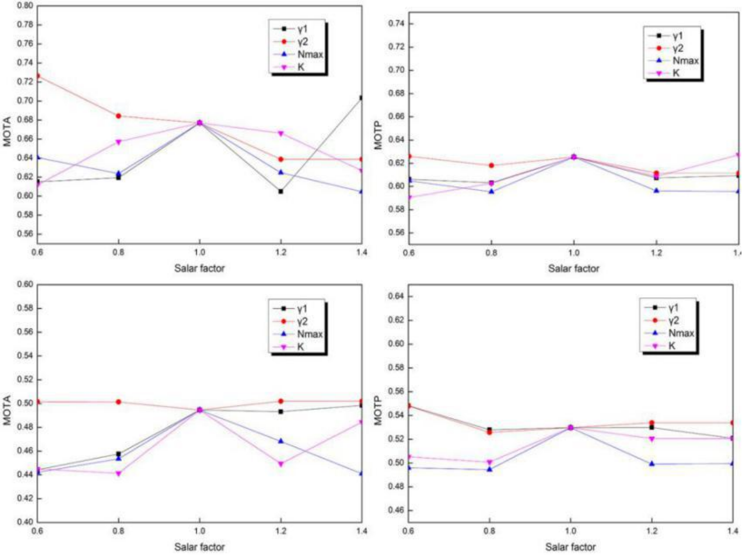

Fig. 2. Algorithm Sensitivity test on several major parameters. First row is the effect of parameters on MOTA and MOTP of S2L1 sequence, Second row is the effect of parameters on MOTA and MOTP of S2L2 sequence.

4 Conclusion

In this paper, we have presented a novel multi-person tracking algorithm. The new observation model adopts graph Laplacian, meanwhile prior gram matrix is constructed based on all samples. In this way unlabelled samples have strong influence on the prior can be viewed as a transfer learning strategy. We divided trackers into two categories base on the number of templates it holds, experimental results show that our algorithm hold obvious advantage compared with other methods.

A future work will research on re-identification scheme in our algorithm to help account for people re-identification. We will extend this framework to Multi-Target, Multi-Camera Tracking.

Acknowledgment

This work was supported by the Principal Foundation of Xiamen University under grant 20720180075.

References

1. Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E., & 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.
2. Avidan, S. (2007). Ensemble tracking. *IEEE transactions on pattern analysis and machine intelligence*, **29**(2).
3. Andriyenko, A., Roth, S., & Schindler, K. (2011, November). An analytical formulation of global occlusion reasoning for multi-target tracking. *In Computer Vision Workshops (ICCV Workshops)*, 2011 IEEE International Conference on (pp. 1839-1846). IEEE.
4. Gao, J., Ling, H., Hu, W., & Xing, J. (2014, September). Transfer learning based visual tracking with gaussian processes regression. *In European Conference on Computer Vision* (pp. 188-203). Springer, Cham.
5. Hu, W., Li, X., Luo, W., Zhang, X., Maybank, S., & Zhang, Z. (2012). Single and Multiple object tracking using log-Euclidean Riemannian subspace and block-division appearance model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(12), 2420-2440.
6. Ferryman, J., & Shahrokni, A. (2009, December). Pets2009: Dataset and challenge. In Performance Evaluation of Tracking and Surveillance (PETS-Winter), *2009 Twelfth IEEE International Workshop* on (pp. 1-6). IEEE.
7. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., ... & Zhang, J. (2009). Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(2), 319-336.
8. Zhang, J., Presti, L. L., & Sclaroff, S. (2012, September). Online multi-person tracking by tracker hierarchy. *In Advanced Video and Signal-Based Surveillance (AVSS)*, 2012 IEEE Ninth International Conference on (pp. 379-385).
9. Leal-Taixé, L., Pons-Moll, G., & Rosenhahn, B. (2011, November). Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. *In Computer Vision Workshops (ICCV Workshops)*, 2011 IEEE International Conference on (pp. 120-127). IEEE.

10. Pirsiavash, H., Ramanan, D., & Fowlkes, C. C. (2011, June). Globally-optimal greedy algorithms for tracking a variable number of objects. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (pp. 1201-1208). IEEE.
11. Bae, S. H., & Yoon, K. J. (2014). Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. *In Proceedings of the IEEE conference on computer vision and pattern recognition*(pp. 1218-1225).
12. Geiger, A., Lauer, M., Wojek, C., Stiller, C., & Urtasun, R. (2014). 3d traffic scene understanding from movable platforms. *IEEE transactions on pattern analysis and machine intelligence*, **36**(5), 1012-1025.
13. Dicle, C., Camps, O. I., & Sznaier, M. (2013, December). The way they move: Tracking multiple targets with similar appearance. In Computer Vision (ICCV), 2013 IEEE International Conference on (pp. 2304-2311). IEEE.
14. Leal-Taixé, L., Fenzi, M., Kuznetsova, A., Rosenhahn, B., & Savarese, S. (2014). Learning an image-based motion context for multiple people tracking. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3542-3549).
15. Rezatofighi, S. H., Milan, A., Zhang, Z., Shi, Q., Dick, A. R., & Reid, I. D. (2015, December). Joint Probabilistic Data Association Revisited. *In ICCV* (pp. 3047-3055).
16. Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016, September). Simple online and realtime tracking. In Image Processing (ICIP), 2016 IEEE International Conference on(pp. 3464-3468). IEEE.
17. Yoon, J. H., Yang, M. H., Lim, J., & Yoon, K. J. (2015, January). Bayesian multi-object tracking using motion context from multiple objects. *In Applications of Computer Vision (WACV)*, 2015 IEEE Winter Conference on (pp. 33-40). IEEE.
18. Fagot-Bouquet, L., Audigier, R., Dhome, Y., & Lerasle, F. (2015, September). Online multi person tracking based on global sparse collaborative representations. In Image Processing (ICIP), 2015 IEEE International Conference on(pp. 2414-2418). IEEE.