Multi-Camera Trajectory Forecasting: Pedestrian Trajectory Prediction in a Network of Cameras

Olly Styles,[*] Tanaya Guha, Victor Sanchez
University of Warwick
o.c.styles | tanaya.guha | v.f.sanchez-silva
@warwick.ac.uk

Alex Kot
Nanyang Technological University
eackot
@ntu.edu.sg

Abstract

We introduce the task of multi-camera trajectory forecasting (MCTF), where the future trajectory of an object is predicted in a network of cameras. Prior works consider forecasting trajectories in a single camera view. Our work is the first to consider the challenging scenario of forecasting across multiple non-overlapping camera views. This has wide applicability in tasks such as person re-identification and multi-target multi-camera tracking. To facilitate research in this new area, we release the ***Warwick-NTU Multi-camera Forecasting Database (WNMF)****, a unique dataset of multi-camera pedestrian trajectories from a network of 15 synchronized cameras. To accurately label this large dataset (600 hours of video footage), we also develop a semi-automated annotation method. An effective MCTF model should proactively anticipate where and when a person will re-appear in the camera network. In this paper, we consider the task of predicting the next camera a pedestrian will re-appear after leaving the view of another camera, and present several baseline approaches for this. The labeled database is available online:* https://github.com/olly-styles/Multi-Camera-Trajectory-Forecasting.

1. Introduction

Predicting the future trajectory of objects in videos is a challenging problem with multiple application domains such as intelligent surveillance [13], person re-identification (RE-ID) [12], and traffic monitoring [4]. Existing works on this topic focus on only the single-camera scenario, that is, predicting the future trajectory of an object in the same camera in which the object is observed [1, 2, 10, 5, 11]. A critical drawback of such single-camera settings is that models cannot anticipate when new objects will enter the scene. A network of multiple cameras can be used to overcome this issue. To this end, we introduce the task of *multi-camera trajectory forecasting* (MCTF): Given the information about an object's location in a single camera, we want to predict its future location across the camera network, in other camera views. In particular, we want to identify the camera in which the object appears next. Fig. 1 presents an overview of the MCTF task.

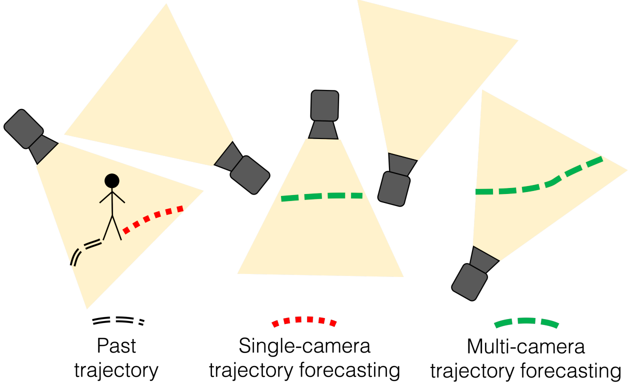
Past trajectory
Single-camera trajectory forecasting
Multi-camera trajectory forecasting

Figure 1. **Multi-camera trajectory forecasting (MCTF).** We introduce a novel formulation of the trajectory forecasting task which utilizes multiple camera views.

Tracking objects (pedestrians) across a large camera network require simultaneously running state-of-the-art algorithms for object detection, tracking, and person RE-ID. Running these algorithms simultaneously can be excessively computationally demanding. Processing videos at a lower image resolution or frame-rate may reduce the computational demands, but this often results in missed detections. Alternatively, we may choose to monitor only a subset of the cameras in the network, which also results in missed detections. A successful MCTF model can address this issue by preempting the location of an object-of-interest in a distributed camera network, thereby enabling the system to monitor only selected cameras intelligently. We envision an MCTF model to be an additional component of

[*] Portions of this work were completed while Styles was at NTU.

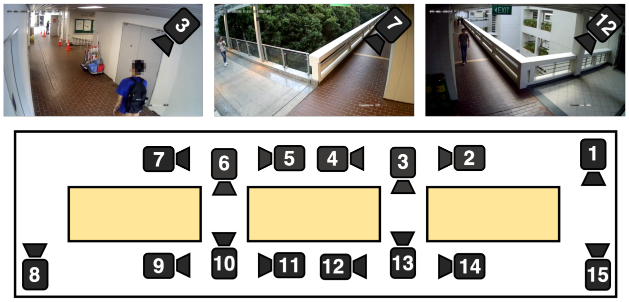

Figure 2. **Example frames and camera network topology.** Faces have been pixelated for privacy reasons.

a full multi-camera monitoring system, complementing the existing methods for detection, tracking, and person RE-ID.

Trajectory information has been used previously in multi-camera settings for tasks such as person RE-ID [12] and vehicle tracking [4]. These methods, however, are *reactive* to observations as they consider trajectory information to assist RE-ID or tracking only when an object has been observed in at least two cameras. In contrast, our proposed task of MCTF is *proactive* - involving predicting the future location of an object even before it enters the camera view. The predicted location serves as a prior for the object detection algorithm and may significantly reduce the search space for detection. Owing to the wide body of complementary literature on pedestrian detection [15] and person RE-ID [16], we focus on pedestrians for our MCTF task. Nevertheless, the task can be easily generalized to any moving object. To facilitate research in the newly formulated MCTF task, we collected a large dataset containing 600 hours of video footage using a network of 15 cameras. We present a semi-automated data annotation method that allows us to gather labels suitable for MCTF using minimal human supervision.

2. Data collection

Existing datasets commonly used for trajectory forecasting, such as ETH [8] and UCY [6], consist of just a single camera view, and are therefore unsuitable for MCTF. Other datasets, such as Duke-MTMC [9], are no longer publicly available. Due to the lack of datasets suitable for MCTF, we collect a new database of 600 hours of video footage from 15 overhead mounted cameras set up indoors on the *Nanyang Technological University* campus. Each camera is placed with a view of either a corridor or a junction. The footage is recorded for 3 weeks in 20-minute long segments collected evenly during the daytime. Example frames and the camera network topology is shown in Fig. 2. We describe our semi-automated data labeling method below.

Data annotation. Fully-manual annotation of data for MCTF would be prohibitively time-consuming as trajectories must first be labeled in single-camera views and then associated across cameras. To minimize the need for manual annotation, we propose a semi-automated method that uses a combination of off-the-shelf models for detection, tracking, and person RE-ID. These results are then manually verified to ensure that proposed tracks are accurate and correct cross-camera correspondences for pedestrians are found. An overview of this annotation method is shown in Fig. 3, which consists of the following three steps:

(i) We run pre-trained object detection [3] and tracking [14] models to locate and track pedestrians in each of the c cameras. The first 20 frames of a track form an entrance tracklet, $E^t = \{e^t, \cdots, e^{t+20}\}$, where e^t is the frame at timestep t. Similarly, the last 20 frames of a track form a departure tracklet $D^t = \{d^{t-20}, \cdots, d^t\}$. We define the first timestep of the entrance tracklet t_E, and the final timestep of a departure tracklet t_D.

(ii) We find cross-camera identity matches between all the departure and entrance tracklets. We use a person RE-ID model [7] to compute RE-ID features for each image and store the mean feature vector for the tracklet. We then compute the visual similarity between the entrance and departure tracklets by computing the squared difference in their RE-ID features, $(R(E^t) - R(D^t))^2$, for all entrance and departure tracklets found in step (i), where $R(x)$ denotes the RE-ID feature vector of tracklet x. We retain those with a squared difference below a manually specified threshold, $\delta = 0.002$. In addition, we constrain candidate tracklets within a manually specified time-window γ to cut down the search space of possible matches, i.e., we compare only those tracklets which satisfy $t_E - t_D < \gamma$. As we set $\gamma = 12$ seconds, the matches are generally from neighboring cameras in the network. We confirmed this by comparing the camera transitions with respect to the network topology in Fig. 4. Our annotation method results in a set of cross-camera transitions $T = \{(E^t, D^t)\}$.

(iii) Finally, we manually verify whether every match proposed by the algorithm is a true positive. The manual verification step assures annotation quality, as false matches

Table 1. **WNMF dataset statistics.**

Hours of footage	600
Number of cameras	15
Collection period	20 days
Time period	8:30am – 7:30pm
Image Resolution	1920 × 1080
Frames per second	5
Cross-camera matches	13.2K
Cross-camera matches after verification	3.2K

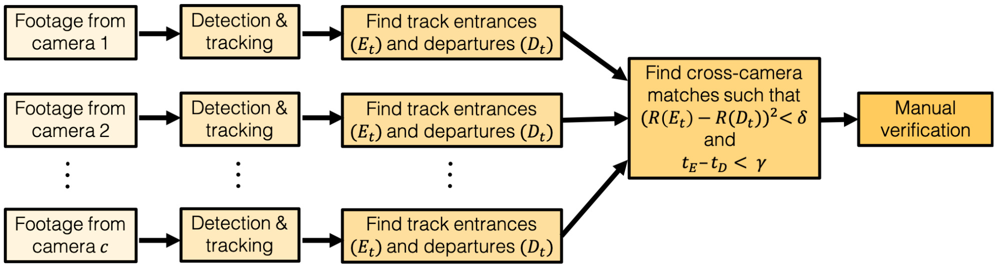
Footage from camera 1
Detection & tracking
Find track entrances (E_t) and departures (D_t)
Footage from camera 2
Detection & tracking
Find track entrances (E_t) and departures (D_t)
Footage from camera c
Detection & tracking
Find track entrances (E_t) and departures (D_t)
Find cross-camera matches such that $(R(E_t) - R(D_t))^2 < \delta$ and $t_E - t_D < \gamma$
Manual verification

Figure 3. **Annotation method.** The proposed method generates the labeled data required for MCTF with minimal human labor by using automated methods for detection, tracking, and person RE-ID before a final manual verification step.

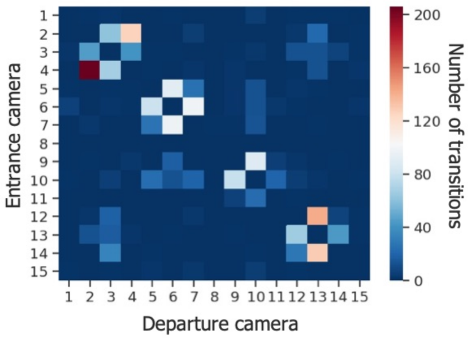

Figure 4. **Transition frequency between cameras** in the WNMF database. Intuitively, neighbouring cameras are transitioned between more frequently.

and bad detections are discarded (Table 1 shows 10K such bad matches were discarded). As the human annotator only has to verify the cross-camera matches rather than finding them from raw videos, the manual overhead is considerably lower than fully manual data annotation. Our annotation method produces a large set (3.2K) of verified departure-entrance pairs.

Data release. The WNMF database is annotated using the aforementioned method and is available online for the research community. We provide tracking data, pre-computed RE-ID features, full multi-camera trajectories, as well as the baseline methods described in the following section. In the interest of preserving privacy, we do not release the raw videos collected to create this dataset.

3. Next camera prediction

We evaluate the problem of predicting the next camera that a target person will re-appear, which we treat as a classification problem. The input is the past trajectory in a single camera view, and the output is a ranking that represents the next camera in the network that this person is most likely to re-appear. In this work, we focus on this next-camera prediction problem only. However, as full trajectory information is available in WNMF, it may also be used for more fine-grained MCTF in future works.

Existing trajectory forecasting methods such as Social-LSTM [1], Social-GAN [2], and SoPhie [10] are designed for single-camera forecasting. These methods do not forecast across multiple cameras; hence direct comparison between these methods for MCTF is not possible. For a fair comparison, we instead create the following baselines:

Shortest real-world distance. We use the physical distance between cameras in the real world, and predict the next camera as the one closest to the camera of the last observation.

Most frequent transition. Using the transition frequency matrix computed earlier (Fig. 4), we predict the next camera as the most frequent next camera of observation from its corresponding position.

Most similar trajectory. We find the most similar trajectory in the training set to the observed trajectory, and predict the next camera to be the same as that of the closest trajectory.

Hand-crafted features. Our hand-crafted feature vector contains velocity in x and y direction, acceleration in x and y direction, last observed bounding box height and width, and its four coordinates. The 10-dimensional features are classified using two fully-connected layers.

In addition, we implement 3 purely learned approaches using the normalized bounding box coordinates as inputs:

Fully connected network. A two-layer fully connected network with 128 hidden units in each layer.

Long short-term memory (LSTM). A standard LSTM with 64 hidden units.

Gated recurrent unit (GRU). A standard GRU with 64 hidden units.

4. Performance evaluation

We compute the top 1 and top 3 classification accuracy of each method introduced in Section 3.

Experimental setup. We evaluate each model using 3-fold

Table 2. **Camera classification.** Given observations from one camera, the next camera of re-appearance is predicted.

Model	Accuracy (%)	
	Top 1	Top 3
Shortest real-world distance	13.9	33.9
Most frequent transition	64.3	90.8
Most similar trajectory	68.3	92.8
Hand-crafted features	68.8	92.5
Fully-connected network	70.0	**93.0**
LSTM	71.8	92.6
GRU	**72.0**	92.7

cross-validation with our 15-camera dataset, using a challenging inter-day validation setup. Footage is recorded on different days in the validation and test sets than in the training set. In each fold, we select 10 days for training, and the remaining 5 days are split into equally sized validation and testing sets. Our neural network based methods are each trained for 10 epochs using a batch size of 16, a learning rate of 1×10^{-3}, and a dropout probability of 20% between fully-connected layers.

Discussion. Table 2 shows next camera prediction results. Predicting the correct camera in the top 3 is a relatively straightforward problem in our dataset, given the highly-structured camera setup and junctions with at most 3 exits. Predicting the most frequent transition using the transition matrix from the training data (Fig. 4) attains modest performance, although learned methods perform better, particularly in terms of top-1 accuracy. We suspect this is due to the past trajectory information in one camera view being informative of the person's future trajectory in a way that is not captured by other baselines. We observe moderate improvements in using recurrent models over a fully-connected network in terms of top-1 accuracy but no improvement in top-3 accuracy.

5. Conclusion

We have introduced a new task of human trajectory forecasting in a multi-camera scenario, which we call multi-camera trajectory forecasting (MCTF). To facilitate further research on MCTF, we present a large dataset, WNMF, which was labeled using a semi-automated data annotation method we developed. Additionally, we present several baseline results for predicting the next camera in which a target person re-reappears within a network. We believe our database and the preliminary results will facilitate and encourage research on this challenging problem.

Acknowledgements This work is funded by the UK EPSRC (grant no. EP/L016400/1) and the EU Horizon 2020 project IDENTITY (Project No. 690907). Our thanks to NVIDIA for their generous hardware donation.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016. 1, 3

[2] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, 2018. 1, 3

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[4] Hung-Min Hsu, Tsung-Wei Huang, Gaoang Wang, Jiarui Cai, Zhichao Lei, and Jenq-Neng Hwang. Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models. In *CVPR Workshop*, 2019. 1, 2

[5] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *CVPR*, pages 336–345, 2017. 1

[6] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 2

[7] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, 2019. 2

[8] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *International Conference on Computer Vision*, pages 261–268. IEEE, 2009. 2

[9] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 2

[10] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *CVPR*, pages 1349–1358, 2019. 1, 3

[11] Olly Styles, Tanaya Guha, and Victor Sanchez. Multiple object forecasting: Predicting future object locations in diverse environments. In *WACV*. IEEE, 2020. 1

[12] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. In *AAAI*, volume 33, pages 8933–8940, 2019. 1, 2

[13] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013. 1

[14] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 2

[15] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Towards reaching human performance in pedestrian detection. *IEEE TPAMI*, 40(4):973–986, 2017. 2

[16] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 2