Peeking into the Future: Predicting Future Person Activities and Locations in Videos

Junwei Liang[1]* Lu Jiang[2] Juan Carlos Niebles[3,2] Alexander Hauptmann[1] Li Fei-Fei[3,2]
[1]Carnegie Mellon University [2]Google AI [3]Stanford University
{junweil,alex}@cs.cmu.edu, lujiang@google.com, {feifeili,jniebles}@cs.stanford.edu

1. Introduction

Deciphering human behaviors to predict their future paths/trajectories and what they would do from videos is important in many applications. With the advancement in deep learning, systems now are able to analyze an unprecedented amount of rich visual information from videos. An important analysis is forecasting the future path of pedestrians, called future person trajectory prediction. This problem has received increasing attention in the computer vision community [5, 1, 3]. It is regarded as an essential building block in video understanding for many applications like self-driving cars, socially-aware robots [9, 6, 4], *etc*.

Humans navigate through public spaces often with specific purposes in mind, ranging from simple ones like entering a room to more complicated ones like putting things into a car. Such intention, however, is mostly neglected in existing work. Consider the example in Fig. 1, the person (at the top-right corner) might take different paths depending on their intention, *e.g*., they might take the green path to *transfer object* or the yellow path to *load object into the car*. Intuitively, humans are able to read from others' body language to anticipate whether they are going to cross the street or continue walking along the sidewalk. In this example, the man at the bottom left corner is waving at the person. Based on common sense, we may agree that the person will take the green path instead of the yellow path. Inspired by this, this paper is interested in modeling the future path jointly with such intention in videos. We model the intention in terms of a predefined set of routine activities such as "loading", "object transfer", *etc*.

To this end, we propose a multi-task learning model called *Next* [1] which has prediction modules for learning future paths and future activities simultaneously. As predicting future activity is challenging, we introduce two new techniques to address the issue. First, unlike most of the existing work [5, 1, 3, 12] which oversimplifies a person as a point in space, we encode a person through rich semantic features about visual appearance, body movement and interaction with the surroundings, motivated by the fact that humans derive such predictions by relying on similar visual cues. Second, to facilitate the training, we introduce an auxiliary tasks for future activity prediction, *i.e*. activity location prediction. In the auxiliary task, we design a discretized grid which we call the Manhattan Grid as location prediction target for the system.

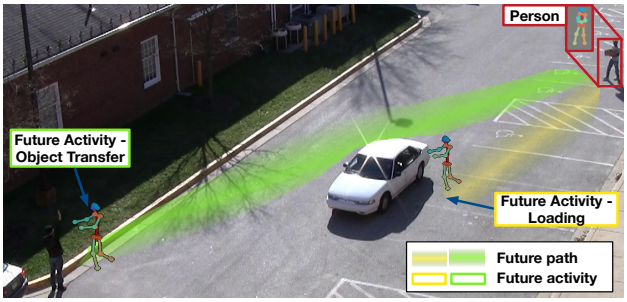

Figure 1. Joint future person path and activity prediction.

To the best of our knowledge, our work is the first on joint future path and activity prediction in streaming videos, and more importantly the first to demonstrate such joint modeling can considerably improve the future path prediction. We empirically validate our model on two benchmarks: ETH & UCY [11, 7], and ActEV/VIRAT [10, 2]. Experimental results show that our method outperforms state-of-the-art baselines, achieving the best-published result on two common benchmarks and producing additional prediction about the future activity.

2. Approach

Problem Formulation: Following [1, 3, 12], we assume each scene is first processed to obtain the spatial coordinates of all people at different time instants. Based on the coordinates, we can automatically extract their bounding boxes. Our system observes the bounding box of all the people from time 1 to T_{obs}, and objects if there are any, and predicts their positions (in terms of xy-coordinates) for time T_{obs+1} to T_{pred}, meanwhile estimating the possibilities of

*Work partially done during a part-time research program at Google.
[1]This is an extended abstract. Full paper, code and models can be found at https://next.cs.cmu.edu

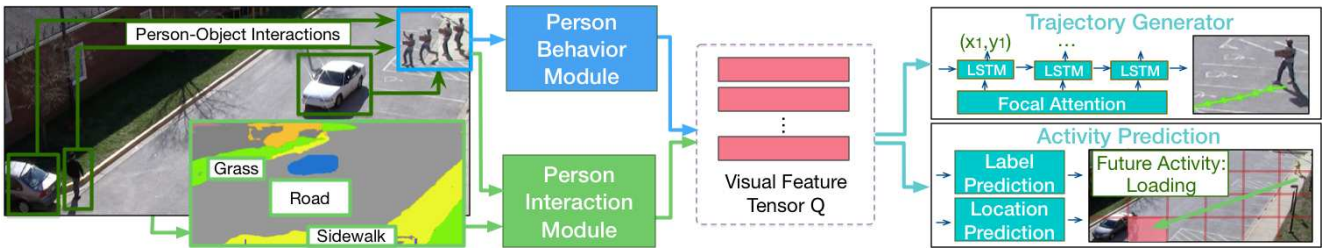

Figure 2. Overview of our model. Given a sequence of frames containing the person for prediction, our model utilizes person behavior module and person interaction module to encode rich visual semantics into a feature tensor.

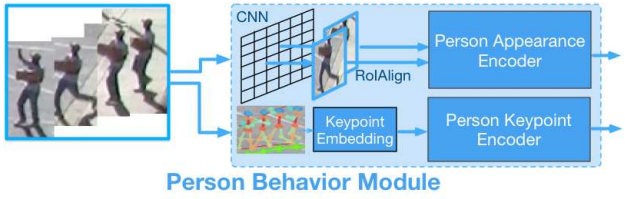

Figure 3. Person behavior module.

future activity labels at time T_{pred}.

2.1. Network Architecture

Fig. 2 shows the overall network architecture of our *Next* model. Unlike most of the existing work [5, 1, 3, 12] which oversimplifies a person as a point in space, our model employs two modules to encode rich visual information about each person's behavior and interaction with the surroundings. *Next* has the following key components:

Person behavior module extracts visual information from the behavioral sequence of the person. As opposed to oversimplifying a person as a point in space, we model the person's the appearance and body movement. To model appearance changes of a person, we extract CNN features for each person bounding box. To capture the body movement, we extract person keypoint information. See Fig. 3.

Person interaction module looks at the interaction between a person and their surroundings, *i.e*. person-scene and person-objects interactions. See Section 2.2 for details.

Trajectory generator summarizes the encoded visual features and predicts the future trajectory by the LSTM decoder with focal attention [8].

Activity prediction utilizes rich visual semantics to predict the future activity label for the person. See Section 2.3.

2.2. Person Interaction Module

Person-scene. To encode the nearby scene of a person, we first extract pixel-level scene semantic classes such as roads and sidewalks for each frame. We apply two convolutional layers on the semantic features with a stride of 2 to get the *scene CNN features* in two scales. Given a person's xy-coordinate, we pool the scene features at the person's current location from the convolution feature map. As the example shown at the bottom of Fig. 4, the red part of the convolution feature is the discretized location of the person

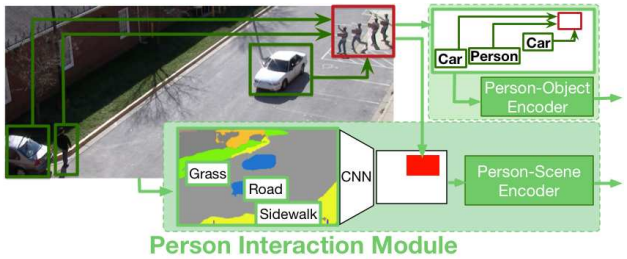

Figure 4. The person interaction module includes person-scene and person-objects modeling. See Section 2.2.

at the current time instant. The receptive field of the feature at each time instant, *i.e*. the size of the spatial window around the person which the model looks at, depends on which scale is being pooled from and the convolution kernel size. In our experiments, we set the scale to 1 and the kernel size to 3, which means our model looks at the 3-by-3 surrounding area of the person at each time instant.

Person-objects. Unlike previous work [1, 3] which relies on LSTM hidden states to model nearby people, our module explicitly models the *geometric relation* and the *object type* of all the objects/persons in the scene. At any time instant, given the observed box of a person (x_b, y_b, w_b, h_b) and K other objects/persons in the scene $\{(x_k, y_k, w_k, h_k)|k \in [1, K]\}$, we encode the geometric relation into $\mathcal{G} \in \mathbb{R}^{K\times 4}$, the k-th row of which equals to:

$$\mathcal{G}_k = [\log(\frac{|x_b - x_k|}{w_b}), \log(\frac{|y_b - y_k|}{h_b}), \log(\frac{w_k}{w_b}), \log(\frac{h_k}{h_b})] \quad (1)$$

This encoding computes the geometric relation in terms of the geometric distance and the fraction box size. We use a logarithmic function to reflect our observation that human trajectories are more likely to be affected by close-by objects or people. For the object type, we simply use one-hot encoding to get the feature.

2.3. Activity Prediction

Since the trajectory generation module outputs one location at a time, errors may accumulate across time and the final destination would deviate from the actual location. Using the wrong location for activity prediction may lead to bad accuracy. To counter this disadvantage, we introduce an auxiliary task, *i.e*. activity location prediction, in addition to predicting the future activity label of the person. We describe the two prediction modules in the following.

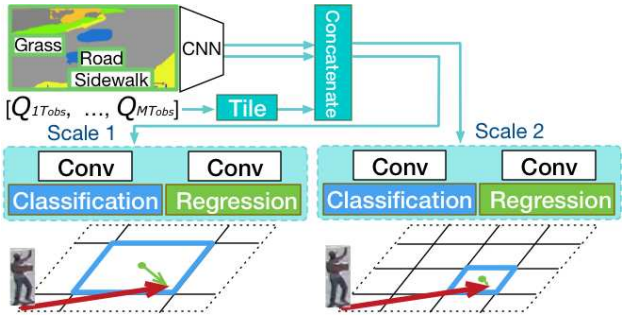

Figure 5. Activity location prediction with classification and regression on the multi-scale Manhattan Grid. See Section 2.3.

Activity location prediction with the Manhattan Grid. To bridge the gap between trajectory generation and activity label prediction, we propose an activity location prediction module to predict the final location of where the person will engage in the future activity. The activity location prediction includes two tasks, *location classification* and *location regression*. As illustrated in Fig. 5, we first divide a video frame into a discretized $h \times w$ grid, namely *Manhattan Grid*, and learn to classify the correct grid block and at the same time to regress from the center of that grid block to the actual location. Specifically, the aim for the classification task is to predict the correct grid block in which the final location coordinates reside. After classifying the grid block, the aim for the regression task is to predict the deviation of the grid block center (green dots in the figure) to the final location coordinate (the end of green arrows). The reason for adding the regression task are: (i) it will provide more precise locations than just a grid block area; (ii) it is complementary to the trajectory prediction which requires xy-coordinates localization. We repeat this process on the Manhattan Grid of different scales and use separate prediction heads to model them. These prediction heads are trained end-to-end with the rest of the model.

As shown in Fig. 5, we first concatenate the scene CNN features (see Section 2.2) with the last hidden state of the encoders . For compatibility, we tile the hidden state $Q_{:T_{obs}:}$ along the height and width dimension resulting in a tensor of the size $M \times d \times w \cdot h$, where $w \cdot h$ is the total number of the grid blocks. The hidden state contains rich information from all encoders and allow gradients flow smoothly through from prediction to feature encoders. The concatenated features are fed into two separate convolution layers for classification and regression. The convolution output for grid classification $\mathsf{cls}_{grid} \in \mathbb{R}^{w \cdot h \times 1}$ indicates the probability of each grid block being the correct destination. In comparison, the convolution output for grid regression $\mathsf{rg}_{grid} \in \mathbb{R}^{w \cdot h \times 2}$ denotes the deviation, in the xy-coordinates, between the final destination and every grid block center. A row of rg_{grid} represents the difference to a grid block, calculated from $[x_t - x_{ci}, y_t - y_{ci}]$ where (x_t, y_t) denotes the predicted location and (x_{ci}, y_{ci}) is the center of the i-th grid block. The ground truth for the grid regression can be computed in a similar way. During training, only the correct grid block receives gradients for regression.

	Method	ADE	FDE	move_ADE	move_FDE
Single Model	Linear	32.19	60.92	42.82	80.18
	LSTM	23.98	44.97	30.55	56.25
	Social LSTM	23.10	44.27	28.59	53.75
	SGAN-PV	30.51	60.90	37.65	73.01
	SGAN-V	30.48	62.17	35.41	68.77
	Ours	**17.99**	**37.24**	**20.34**	**42.54**
20 Outputs	SGAN-PV-20	23.11	41.81	29.80	53.04
	SGAN-V-20	21.16	38.05	26.97	47.57
	Ours-20	**16.00**	**32.99**	**17.97**	**37.28**

Table 1. Comparison to baseline methods on the ActEV/VIRAT validation set. Top uses the single model output. Bottom uses 20 outputs. Numbers denote errors thus lower are better.

Activity label prediction. Given the encoded visual observation sequence, the activity label prediction module predicts the future activity at time instant T_{pred}. We compute the future activity probabilities using the concatenated last hidden states of the encoders. The future activity of a person could be multi-class, *e.g.* a person could be "walking" and "carrying" at the same time.

3. Experiments

We evaluate the proposed *Next* model on two common benchmarks for future path prediction: ETH [11] and UCY [7], and ActEV/VIRAT [2, 10]. Here we only show results on ActEV/VIRAT. Please refer to the full paper for full results.

Baseline methods. We compare our method with the two simple baselines and two recent methods: ***Linear*** is a single layer model that predicts the next coordinates using a linear regressor based on the previous input point. ***LSTM*** is a simple LSTM encoder-decoder model with coordinates input only. ***Social LSTM*** [1]: We train the social LSTM model to directly predict trajectory coordinates instead of Gaussian parameters. ***SGAN*** [3]: We train two model variants (PV & V) detailed in the paper using the released code from Social-GAN [3] (https://github.com/agrimgupta92/sgan/).

Aside from using a single model at test time, Gupta *et al*. [3] also used 20 model outputs per frame and selected the best prediction to count towards the final performance. Following the practice, we train 20 identical models using random initializations and report the same evaluation results, which are marked "20 outputs" in Table 1.

Main Results. Table 1 lists the testing error, where the top part is the error of a single model output and the bottom shows the best result of 20 model outputs. The "ADE" and "FDE" columns summarize the error over all trajectories, and the last two columns further detail the subset trajectories of moving activities ("walk", "run", and "ride_bike"). We report the mean performance of 20 runs of our single model at Row 7. The standard deviation on "ADE" met-

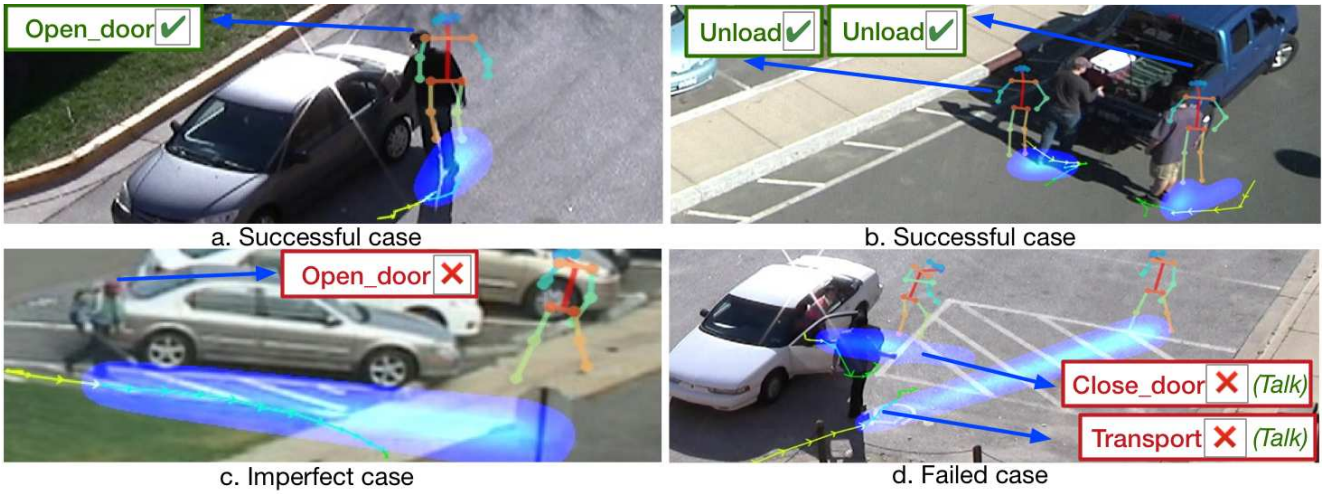

Figure 6. (Better viewed in color.) Qualitative analysis. Yellow path is the observable trajectory and green path is the ground truth trajectory during the prediction period. Predictions are shown as blue heatmaps. Our model also predicts the future activity, which is shown in the text and with the person pose template.

ric is 0.043. As we see, our method performs favorably against other methods, especially in predicting the trajectories of moving activities. For example, our model outperforms Social-LSTM and Social-GAN by a large margin of 10 points in terms of the "move_FDE" metric. The results demonstrate the efficacy of the proposed model and its state-of-the-art performance on future trajectory prediction.

Qualitative analysis. We provide a qualitative analysis of our model predictions. (i) Successful cases: In Fig 6(a) and 6(b), both the trajectory prediction and future activity prediction are correct. (ii) Imperfect case: In Fig 6(c), although the trajectory prediction is mostly correct, our model predicts that the person is going to open the door of the car, given the observation that he is walking towards the side of the car. (iii) Failed case: In Fig 6(d), our model fails to capture the subtle interactions between the two persons and predicts that they will go separate ways, while in fact they are going to stop and talk to each other.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016. 1, 2, 3

[2] G. Awad, A. Butt, K. Curtis, J. Fiscus, A. Godil, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, J. Magalhaes, D. Semedo, and S. Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *TRECVID*, 2018. 1, 3

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

[4] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *CVPR*, 2015. 1

[5] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012. 1, 2

[6] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2016. 1

[7] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Computer Graphics Forum*, pages 655–664. Wiley Online Library, 2007. 1, 3

[8] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. Hauptmann. Focal visual-text attention for visual question answering. In *CVPR*, 2018. 2

[9] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras. People tracking with human motion predictions from social forces. In *ICRA*, 2010. 1

[10] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011. 1, 3

[11] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*, 2012. 1, 3

[12] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. *arXiv preprint arXiv:1806.01482*, 2018. 1, 2