Group LSTM: Group Trajectory Prediction in Crowded Scenarios

Niccoló Bisagno[1], Bo Zhang[2], and Nicola Conci[1]

[1] University of Trento (Trento, Italy)
{niccolo.bisagno, nicola.conci}@unitn.it
[2] Dalian Maritime University (Dalian City, China)
bzhang@dlmu.edu.cn

Abstract. The analysis of crowded scenes is one of the most challenging scenarios in visual surveillance, and a variety of factors need to be taken into account, such as the structure of the environments, and the presence of mutual occlusions and obstacles. Traditional prediction methods (such as RNN, LSTM, VAE, etc.) focus on anticipating individual's future path based on the precise motion history of a pedestrian. However, since tracking algorithms are generally not reliable in highly dense scenes, these methods are not easily applicable in real environments. Nevertheless, it is very common that people (friends, couples, family members, etc.) tend to exhibit coherent motion patterns. Motivated by this phenomenon, we propose a novel approach to predict future trajectories in crowded scenes, at the group level. First, by exploiting the motion coherency, we cluster trajectories that have similar motion trends. In this way, pedestrians within the same group can be well segmented. Then, an improved social-LSTM is adopted for future path prediction. We evaluate our approach on standard crowd benchmarks (the UCY dataset and the ETH dataset), demonstrating its efficacy and applicability.

Keywords: group prediction; crowd analysis; trajectory clustering; social-LSTM

1 Introduction

Crowd analysis is a hot topic in computer vision, covering a wide range of applications in visual surveillance. The main challenges in crowd analysis include: crowd dynamics modeling [43, 5]; crowd segmentation [4]; crowd activity classification [33]; abnormal behavior detection [16, 25]; density estimation [30]; and crowd behavior anticipation [2].

Among them, crowd behavior anticipation is an emerging task, which has drawn a fair amount of attentions, due to the rapid development in machine learning, and particularly the deep learning techniques applied to time series analysis (such as RNN [34], GRU [9], LSTM [18], and VAE [22]).

Different from crowd behavior recognition, the prediction task has its distinguished characteristics, which is generally addressed by observing the motion histories of the subjects moving in the scene. In some specific applications (i.e., early warning, abnormal event detection, collision avoidance), prediction plays a more relevant role comparing to activity recognition, as dangerous behaviors should be warned in advance.

Traditional methods can merely make one-step forecasting (e.g., Kalman filter, particle filter, Markov chains); thanks to deep learning, long term prediction is becoming applicable gradually.

At the beginning, researchers merely focused on anticipating individual's future path. The corresponding models highly rely on the precise motion history of a pedestrian, thus being generally intractable in very dense environments, due to the instability of object tracking algorithms in presence of frequent mutual occlusions.

However, continuous and precise frame-based tracking might not be essential. In fact, in most cases, people pay more attention on the whole dynamics of the scene. People gathering and behaving together will generate and exhibit macroscopic salient features, which are instead worth being observed. Such coarse-level information usually maps densely and sparsely populated areas, including direction and flow characteristics, as well as the final destinations. Therefore, in such scenarios, it makes more sense to focus on group activities instead of individuals. It is well known that people moving in the crowds usually tend to follow a series of implicit social rules [28]. For instance, individuals tend to speed up or slow down their paces in order to avoid collisions when a vehicle or another group of people is approaching; people prefer to preserve personal space, thus keeping a certain distance from their neighbors; pedestrians tend to follow people in their front especially in presence of crowded situations, to prevent collisions.

Focusing on grouping, it is very common that friends/couples/families tend to move in accordance with a coherent motion pattern. Based on this assumption, we propose a novel approach to predict future trajectories at the group level, in order to further analyze crowded scenes from a holistic point of view. Firstly, by exploiting the motion coherency, we cluster trajectories that have similar motion trends. In this way, pedestrians within the same group can be highlighted and segmented. Finally, an improved social-LSTM is proposed to estimate the future path prediction.

The main contributions of this work are summarized as follows:

- we propose a novel framework for group behavior prediction;
- we exploit an improved coherent filtering to enhance the trajectory clustering performance;
- we propose a strategy for long term prediction of pedestrians, which leverages on group dynamics.

The rest of the paper is organized as follows: Section 2 briefly reviews the related work in the field of crowd analysis. The proposed framework, called Group LSTM for conciseness, is described in Section 3, including the steps of trajectory clustering and group path prediction. The experimental results are provided in Section 4. Conclusions and future work are summarized in Section 5.

2 Related work

A detailed literature on the recent works in crowd analysis, especially regarding the topics of crowd dynamic modeling, social activity forecasting, and group segmentation, can be found in some recent surveys [24][13][20]. In the next paragraphs, we will concentrate on two specific sub-topics, namely, group analysis and forecasting.

2.1 Group analysis in crowds

In the early approaches, trajectories were adopted to represent low level motion features in the crowd. By clustering trajectories with similar motion trends, pedestrians can be gathered into different groups. In [42], the traditional k-means algorithm was exploited to learn different motion modalities in the scene. In [21], support vector clustering was exploited to group pedestrians. In [44], coherent filtering was presented to detect coherent motion patterns in a crowded environment[40].

As far as the representation of collective activities is concerned, Ge et al. [12] worked on the automatic detection of small individual groups who are traveling together. Ryoo et al. [31] introduced a probabilistic representation of group activities, for the purpose of recognizing different types of high-level group behaviors.Yi et al. [41] investigated the interactions between stationary crowd groups and pedestrians to analyze pedestrian's behaviors, including walking path prediction, destination prediction, personality classification, and abnormal event detection. Shao et al. [32] proposed a series of scene-independent descriptors to quantitatively describe group properties, such as collectiveness, stability, uniformity, and conflict. Bagautdinov et al. [7] presented a unified end-to-end framework for multi-person action localization and collective activity recognition using deep recurrent networks.

2.2 Social activity forecasting

Forecasting social activities has lately gained a relevant amount of attentions, especially as far as crowd analysis is concerned. This research domain is rather diversified and it involves trajectory prediction, interaction modeling, and contextual modeling. Among the pioneering research in social activity analysis, Helbing et al. [17] introduced the well known Social Force Model (SFM), which is able to describe social interactions between humans [23, 27]. Other models, such as the continuum crowds model [36] and the Reciprocal Collision Avoidance [37], are capable to reproduce human interactions using priors. In [3], the Social Affinity Maps (SAM) features and the Origin and Destination (OD) priors were proposed to forecast pedestrians' destinations using multi-view surveillance cameras. Robicquet et al. [29] introduced a large scale dataset that contains various types of targets (pedestrians, bikers, skateboarders, cars, buses, and golf carts) using aerial cameras, in order to evaluate trajectory forecasting performance in real outdoor environments. In [1] [26], contextual information is taken into account as well, to model the static configuration and the dynamic evolution of the scene.

More recently, neural networks have been employed to predict events in crowded videos. In particular, with the emerging of deep generative models (such as RNN, LSTM, VAE), the sequence-to-sequence generation problem can be solved properly, making it possible to handle the long-term prediction task directly. Alahi et al. [2] proposed the so-called social-LSTM to model the interactions among people in a neighborhood by adding a new social pooling layer; In [22], Lee et al. presented a deep stochastic IOC RNN encoder-decoder framework to predict the future paths of multiple interacting agents in dynamic scenes. Ballan et al. [8] considered both the dynamics of moving agents and the scene semantics to predict scene-specific motion patterns.

Social activities are often ruled not only by the motion dynamics, but are also driven by human factors. Jain et al. [19] adopted a structural RNN that combines spatio-temporal graphs and recurrent neural networks to model motion and interactions in the scene. Fernando et al. [38] applied both the soft attention and the hard-wired attention on the social LSTM, and significantly promote the trajectory prediction performance. Varshneya et al. [6] presented a soft attention mechanism to forecast individual's path, which exploits the spatially aware deep attention model. Vemula et al. [39] proposed a novel social attention model that can capture the relative importance of each person when navigating in the scene.

3 Group LSTM

The motion of pedestrians in crowded scenes is highly influenced by the behavior of other people in the surroundings and their mutual relationships. Stationary groups, groups of pedestrians walking together, people coming from opposite directions, will exert different effects on the action that one pedestrian takes. Thus, it becomes necessary to take people in the neighborhood into account when forecasting the behavior of an individual in the crowd.

To achieve this goal, we propose a framework, which is able to consider whether the subject of interest is walking coherently with the pedestrians in his surroundings or not. By exploiting the coherent filtering approach [44], we first detect people moving coherently in a crowd, and then adopt the Social LSTM to predict future trajectories. In this way, we are able to improve the prediction performance, accounting for the interactions between socially related and unrelated pedestrians in the scene.

3.1 Pedestrian trajectory clustering

Coherent motion describes the collective movements of particles in a crowd. The coherent filtering studies a prior meant to describe the coherent neighbor invariance, which is the local spatio-temporal relation between particles moving coherently. The algorithm is based on two steps. First, it detects the coherent motion of pedestrians in the scene. Then, points moving coherently are associated to the same cluster. Point clusters will continue to evolve, and new clusters will emerge over time. Finally, each pedestrian i is assigned to a cluster s_i. The outputs of the coherent filtering are consist of the sets s_i ($i = 1, 2, \cdots, n$) of people moving in a coherent manner. If a pedestrian is not moving or it does not belong to any coherent group, it is considered as belonging to its own set.

The coherent filtering originally relies on the KLT tracker [35], aiming at detecting candidate points for tracking and generating trajectories, which will then be used as the input of the algorithm. The KLT tracker may detect many key points for each pedestrian, thus there is no clear correspondence between the number of key points and the number of pedestrians. Our objective is to cluster pedestrians into groups, where each individual in a group is represented using a single point, as shown in Fig. 1. For this purpose, and without loss of generality, we apply the coherent filtering algorithm directly on the ground truth of pedestrian trajectories.

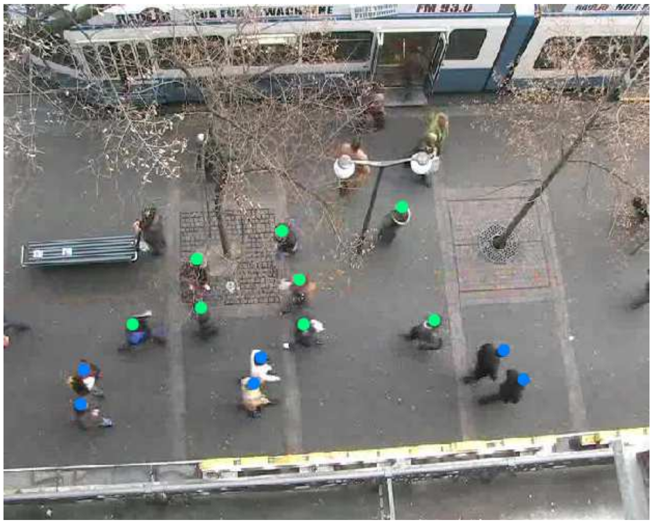


Fig. 1. Each pedestrian is represented by a single keypoint. Pedestrians walking in the same direction are clustered into one group s_i. In this example, two sets of pedestrians going in opposite directions are identified.

3.2 Group trajectory prediction

We extend the work of Alahi et al. [2], which models the relationships of pedestrians in the neighborhood by introducing a so-called social pooling layer. In the Social LSTM model, the pedestrian is modeled using an LSTM network as displayed in Fig. 2. Furthermore, each pedestrian is associated with other people in his neighborhood via a social pooling layer. The social pooling layer allows pedestrians to share their hidden states, thus enabling each network to predict the future positions of an individual based on his own hidden state and the hidden states in the neighborhood.

The i^{th} pedestrian at time instance t in the scene is represented by the hidden state h_t^i in an LSTM network. We set the hidden-state dimension to D and the neighborhood size to N_0, respectively. The neighborhood of the i^{th} agent ped^i is described using a tensor H_t^i as in Eq. 1, with dimensions of $N_0 \times N_0 \times D$:

$$H_t^i(m,n,:) = \sum_{j \in N} 1_{mn}[x_t^j - x_t^i, y_t^j - y_t^i] 1_{ij}[s_i \neq s_j] h_{t-1}^j \tag{1}$$

where $1_{mn}[x,y]$ is an indicator function to select pedestrians in the neighborhood. It is defined as in Eq. 2:

$$1_{mn}[x,y] = \begin{cases} 0 & \text{if } [x,y] \notin \text{cell mn} \\ 1 & \text{if } [x,y] \in \text{cell mn} \end{cases} \tag{2}$$

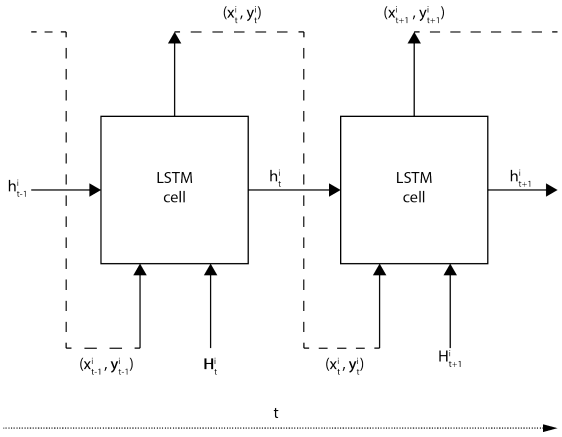


Fig. 2. The figure represents the chain structure of the LSTM network between two consecutive time steps, t and $t + 1$. At each time step, the inputs of the LSTM cell are the previous position (x^i_{t-1}, y^i_{t-1}) and the Social pooling tensor H^i_t. The output of the LSTM cell is the current position (x^i_t, y^i_t).

If two pedestrians i and j belong to the same coherent set s_i, they will not be taken into account when computing the social pooling layer for each of them. The function $1_{ij}[i \in s_i, j \in s_i]$ is an indicator function defined as in Eq. 3:

$$1_{ij}[s_i \neq s_j] = \begin{cases} 0 & \text{if } i \in s_i, j \in s_i \\ 1 & \text{if } i \in s_i, j \notin s_i \end{cases} \tag{3}$$

Doing so, the social pooling layer of each pedestrian contains information only about pedestrians, which are not moving coherently with him.

Once computed, the social hidden-state tensor is embedded into a vector a^i_t. The output coordinates are embedded in the vector e^i_t. Following the recurrence defined in [2], we can predict our trajectories gradually.

4 Results

4.1 Implementation details

In the first place, we need to configure the coherent filtering to cluster pedestrians. To this aim, we use $K = 10$, $d = 1$ and $\lambda = 0.2$ according to the original implementation.

For our LSTM network, we adopt the following configuration. The embedding dimension for the spatial coordinates is set to 64. The spatial pooling size, which corresponds to an area of $4{\times}4$ m^2, is set to 32. The pooling operation is performed using a sum pooling window of size 8×8 with no overlaps. The hidden state dimension is

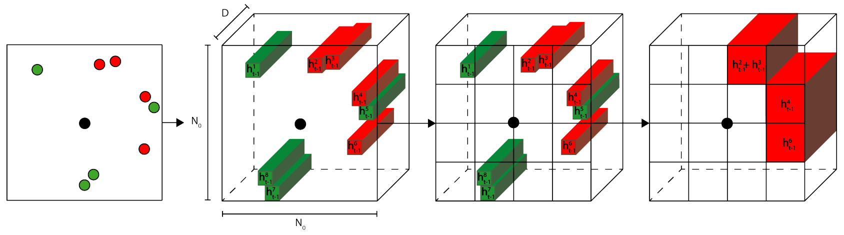


Fig. 3. Representation of the Social hidden-state tensor H^i_t. The black dot represents the pedestrian of interest ped_i. Other pedestrians ped_j $(\forall j \neq i)$ are shown in different color codes, namely green for pedestrians belonging to the same set, and red for pedestrians belonging to a different set. The neighborhood of ped_i is described by $N_0 \times N_0$ cells, which preserves the spatial information by pooling spatially adjacent neighbors. Pedestrians belonging to the same set are not used for the final computation of the pooling layer H^i_t.

128. The learning rate is set to 0.003, and RMS-prop [11] is used as the optimizer. The model is trained on a single GPU using a PyTorch[3] implementation.

4.2 Quantitative results

Our experiments are carried out on two publicly available datasets, commonly used as the standard benchmarks for crowded scenarios, namely, the UCY dataset [23] and the ETH dataset [27].

The two datasets present a rather large set of real-world trajectories covering a variety of complex crowd behaviors that are particularly interesting for our research.

In the same way as other works [27, 2], we evaluate our results with the following two metrics:

- *Average Displacement Error (ADE)*, namely the average displacement error (in meters) between each point of the predicted path with respect to the ground truth path.
- *Final Displacement Error (FDE)*, namely the distance (in meters) between the final point of the predicted trajectory and the final point of the ground truth trajectory.

In our experiments, we follow the same evaluation procedure as adopted in [2]. The model is trained and validated using the leave-one-out strategy. We train on 4 videos and test on the remaining one to obtain the prediction results. For both training and validation, we observe and predict trajectories using a time interval of 0.4 seconds. We observe trajectories for 8 time steps and predict for the next 12 time steps, meaning that we observe trajectories for $t_{obs} = 3.2$ seconds and predict for the next $t_{pred} = 4.8$ seconds. In the training phase, only trajectories that remain in the scene for at least 8 seconds are considered.

We compare our method with the Social LSTM model [2] and its most recent variant [14]. We also compare our model with a linear model, which uses the Kalman filter to

[3] http://pytorch.org

predict future trajectories under the assumption of linear acceleration, as also reported in [2]. The numerical results are shown in Table 1.

Our method performs on average better or equal than other methods, especially on the UCY dataset. This is due to the characteristics of crowd flows in the scene, which usually consist of easily identifiable groups walking in opposite directions. However, for the ETH dataset, the motion patterns are more varied and chaotic.

Our results show that the prediction performance can be improved when considering pedestrians that are not moving coherently. We argue that the change of motion and the evolution of trajectories are mainly influenced by pedestrians which move in different directions with respect to the pedestrian of interest. People walking together, instead, loosely influence each other, as they behave as in a group.

Table 1. Quantitative results using our Group-LSTM and the mentioned baseline approaches on the UCY and ETH datasets, respectively. Two error metrics, namely, the Average Displacement Error (ADE) and the Final Displacement Error (FDE) are reported (in meters) for an observation interval $t_{obs} = 3.2$ seconds and a prediction of subsequent $t_{pred} = 4.8$ seconds. Our model outperforms other approaches, especially in terms of average error.

Metric	***Dataset***	***Lin.*[2]**	***Social-LSTM*[14]**	***Social-GAN*[14]**	**Group-LSTM**
ADE	ETH [27]	1.33	1.09	0.81	0.28
	HOTEL [27]	0.39	0.86	0.72	0.28
	ZARA1 [23]	0.62	0.41	0.34	0.23
	ZARA2 [23]	0.77	0.52	0.42	0.34
	UCY [23]	0.82	0.61	0.60	0.56
	AVERAGE	0.79	0.70	0.58	0.34
FDE	ETH [27]	2.94	2.41	1.52	1.12
	HOTEL [27]	0.72	1.91	1.61	0.89
	ZARA1 [23]	1.21	1.11	0.84	0.91
	ZARA2 [23]	1.48	1.31	1.26	1.49
	UCY [23]	1.59	0.88	0.69	1.48
	AVERAGE	1.59	1.52	1.18	1.18

4.3 Qualitative results

In Section 4.2 we have shown that considering only pedestrians not moving coherently can improve the prediction precision. In this section we will further evaluate the consistency of the predicted trajectories.

As a general rule, the LSTM-based approaches for trajectory prediction follow a data-driven approach. Furthermore, the future planning of pedestrians in a crowd are highly influenced by their goals, their surroundings, and their past motion histories. Pooling the correct data in the social layer can promote the prediction performance in a significant way.

In order to guarantee a reliable prediction, we not only need to account for spatio-temporal relationships, but also need to preserve the social nature of behaviors. According to the studies in interpersonal distances [15, 10], socially correlated people tend to stay closer in their personal space and walk together in crowded environments as compared to pacing with unknown pedestrians. Pooling only unrelated pedestrians will focus more on macroscopic inter-group interactions rather than intra-group dynamics, thus allowing the LSTM network to improve the trajectory prediction performance. Collision avoidance influences the future motion of pedestrians in a similar manner if two pedestrians are walking together as in a group.

In Tables 2, 3 and Fig. 4, we display some demos of predicted trajectories which highlight how our Group-LSTM is able to predict pedestrian trajectories with better precision, showing how the prediction is improved when we pool in the social tensor of each pedestrian only pedestrians not belonging to his group.

In Table 2, we show how the prediction of two pedestrians walking together in the crowd improves when they are not pooled in each other's pooling layer. When the two pedestrians are pooled together, the network applies on them the typical repulsion force to avoid colliding with each other. Since they are in the same group, they allow the other pedestrian to stay closer in they personal space.

In Fig. 4 we display the sequences of two groups walking toward each other. In Table 3, we show how the prediction for the two groups is improved with respect to the Social LSTM. While both prediction are not very accurate, our Group LSTM perform better because it is able to forecast how pedestrian belonging to the same group will stay together when navigating the environment.

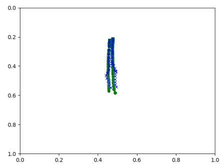
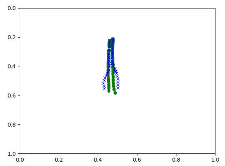
Name	Scene	Our Group-LSTM	Social-LSTM
ETH Univ Frame 2425			

Table 2. ETH dataset: the prediction is improved when pooling in the social tensor of each pedestrian only pedestrians not belonging to his group. The green dots represent the ground truth trajectories; the blue crosses represent the predicted paths.

5 Conclusion

In this work, we tackle the problem of pedestrian trajectory prediction in crowded scenes. We propose a novel approach, which combines the coherent filtering algorithm with the LSTM networks. The coherent filtering is used to identify pedestrians walking together in a crowd, while the LSTM network is used to predict the future trajectories by exploiting inter and intra group dynamics. Experimental results show that the proposed Group LSTM outperforms the Social LSTM in the prediction task on two public

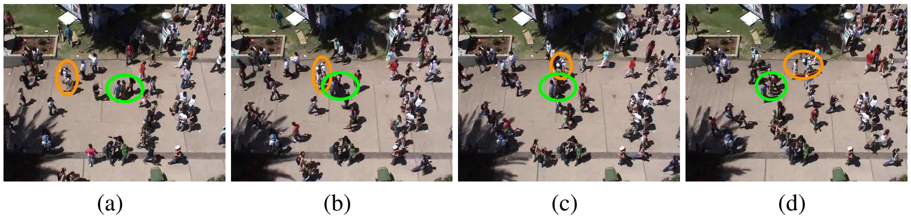


Fig. 4. Sequences taken from the UCY dataset. It displays an interaction example between two groups, which will be further analyzed in Table 3.

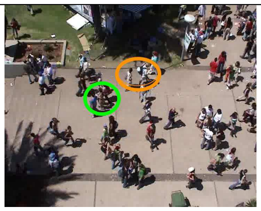
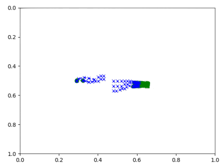
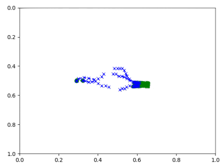
Name	Scene	Our Group-LSTM	Social-LSTM
UCY Univ Frame 1025			

Table 3. We display how the prediction is improved for two groups walking in opposite directions. The green dots represent the ground truth trajectories, while the blue crosses represent the predicted paths.

benchmarks (the UCY and ETH datasets). For the future work, we plan to further investigate social relationships and how fixed obstacles will influence the behaviors of other pedestrians.

6 Acknowledgement

This work is partly supported by the National Natural Science Foundation of China (Grant No.61702073), and the Fundamental Research Funds for the Central Universities (Grant No.3132018190).

References

1. Context-aware trajectory prediction in crowded spaces. In: Proceedings of the
2. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. pp. 961–971. IEEE (2016)
3. Alahi, A., Ramanathan, V., Li, F.F.: Socially-aware large-scale crowd forecasting. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. pp. 2203–2210. IEEE (2014)
4. Ali, S., Shah, M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. pp. 1–6. IEEE (2007)

5. Allain, P., Corpetti, T., Corpetti, T.: Crowd flow characterization with optimal control theory. In: Proceedings of the Asian Conference on Computer Vision. pp. 279–290. Springer (2009)
6. A.Vemula, K.Muelling, J.Oh: Social attention: Modeling attention in human crowds. https://arxiv.org/abs/1710.04689 (2017)
7. Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. pp. 3425–3434. IEEE (2016)
8. Ballan, L., Castaldo, F., Alahi, A., Palmieri, F., Savarese, S.: Knowledge transfer for scene-specic motion prediction. In: Proceedings of the European Conference on Computer Vision. pp. 697–713. Springer (2016)
9. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
10. Conci, N., Bisagno, N., Cavallaro, A.: On modeling and analyzing crowds from videos. In: Computer Vision for Assistive Healthcare, pp. 319–336. Elsevier (2018)
11. Dauphin, Y., de Vries, H., Bengio, Y.: Equilibrated adaptive learning rates for non-convex optimization. Advances in Neural Information Processing Systems (NIPS) pp. 1504–1512 (2015)
12. Ge, W., Collins, R.T., Ruback, R.B.: Vision-based analysis of small groups in pedestrian crowds. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(5), 1003–1016 (2012)
13. Grant, J., Flynn, P.: Crowd scene understanding from video: A survey. ACM Transactions on Multimedia Computing, Communications, and Applications **13**(2) (2017)
14. Gupta, A., Savarese, S., Alahi, A., et al.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). No. CONF (2018)
15. Hall, E.T., Birdwhistell, R.L., Bock, B., Bohannan, P., Diebold Jr, A.R., Durbin, M., Edmonson, M.S., Fischer, J., Hymes, D., Kimball, S.T., et al.: Proxemics [and comments and replies]. Current anthropology **9**(2/3), 83–108 (1968)
16. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: Real-time detection of violent crowd behavior. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–6. IEEE (2012)
17. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Physical review E **51**(5), 4282 (1995)
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
19. Jain, A., Zamir, A., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. pp. 5308–5317. IEEE (2015)
20. Kok, V., Mei, K., Chan, C.: Crowd behavior analysis: A review where physics meets biology. Neurocomputing **177**, 342–362 (2016)
21. Lawal, I., Poiesi, F., Aguita, D., Cavallaro, A.: Support vector motion clustering. IEEE Transactions on Circuits and Systems for Video Technology **27**(11), 2395–2408 (2017)
22. Lee, N., Choi, W., Vernaza, P., Choy, C., Torr, P., Chandraker, M.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. pp. 2165–2174. IEEE (2017)
23. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. vol. 26, pp. 655–664. Wiley Online Library (2007)
24. Li, T., Chang, H., Wang, M., Ni, B., Hong, R., Yan, S.: Crowded scene analysis: A survey. IEEE Transactions on Circuits and Systems for Video Technology **25**(3), 367–386 (2015)

25. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(1), 18–32 (2013)
26. Ma, W., Huang, D., Lee, N., Kitani, K.M.: A game-theoretic approach to multi-pedestrian activity forecasting (2016)
27. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: Proceedings of the International Conference on Computer Vision. pp. 261–268. IEEE (2009)
28. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory understanding in crowded scenes. In: Proceedings of the European Conference on Computer Vision. pp. 549–565. Springer (2016)
29. Robicquet, A., Alahi, A., Sadeghian, A., Anenberg, B., Doherty, J., Wu, E., Savarese, S.: Forecasting social navigation in crowded complex scenes (2016)
30. R.Stewart, M.Andriluka, Ng, A.: End-to-end people detection in crowded scenes. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. pp. 2325–2333. IEEE (2016)
31. Ryoo, M., Aggarwal, J.: Stochastic representation and recognition of high-level group activities. International Journal of Computer Vision
32. Shao, J., Loy, C.C., Wang, X.: Learning scene-independent group descriptors for crowd understanding. IEEE Transactions on Circuits and Systems for Video Technology **27**(6), 1290–1303 (2017)
33. Solmaz, B., Moore, B.E., Shah, M.: Identifying behaviors in crowd scenes using stability analysis for dynamical systems. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(10), 2064–2070 (2012)
34. Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems. pp. 3104–3112 (2014)
35. Tomasi, C., Kanade, T.: Detection and tracking of point features (1991)
36. Treuille, A., Cooper, S., Popović, Z.: Continuum crowds. vol. 25, pp. 1160–1168. ACM (2006)
37. Van Den Berg, J., Guy, S.J., Lin, M., Manocha, D.: Reciprocal n-body collision avoidance. In: Robotics research, pp. 3–19. Springer (2011)
38. Varshneya, D., Srinivasaraghavan, G.: Human trajectory prediction using spatially aware deep attention models. https://arxiv.org/abs/1705.09436 (2017)
39. Vemula, A., Muelling, K., Oh, J.: Social attention: Modeling attention in human crowds. arXiv:1710.04689 (2017)
40. Yamaguchi, K., Berg, A.C., Ortiz, L.E., Berg, T.L.: Who are you with and where are you going? In: Proceedings of the International Conference on Computer Vision and Computer Vision. pp. 1345–1352. IEEE (2011)
41. Yi, S., Li, H., Wang, X.: Understanding pedestrian behaviors from stationary crowd groups. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. pp. 3488–3496. IEEE (2015)
42. Zhong, J., Cai, W., Luo, L., Yin, H.: Learning behavior patterns from video: a data-driven framework for agent-based crowd modeling. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. pp. 801–809 (2015)
43. Zhou, B., Tang, X., Wang, X.: Learning collective crowd behaviors with dynamic agent. International Journal of Computer Vision **111**(1), 50–68 (2015)
44. Zhou, B., Tang, X., Wang, X.: Coherent filtering: Detecting coherent motions from crowd clutters. In: Proceedings of the European Conference on Computer Vision. pp. 857–871. Springer (2012)