Social and Scene-Aware Trajectory Prediction in Crowded Spaces

Matteo Lisotto Pasquale Coscia Lamberto Ballan
Department of Mathematics "Tullio Levi-Civita", University of Padova, Italy
{pasquale.coscia,lamberto.ballan}@unipd.it

Abstract

Mimicking human ability to forecast future positions or interpret complex interactions in urban scenarios, such as streets, shopping malls or squares, is essential to develop socially compliant robots or self-driving cars. Autonomous systems may gain advantage on anticipating human motion to avoid collisions or to naturally behave alongside people. To foresee plausible trajectories, we construct an LSTM (long short-term memory)-based model considering three fundamental factors: people interactions, past observations in terms of previously crossed areas and semantics of surrounding space. Our model encompasses several pooling mechanisms to join the above elements defining multiple tensors, namely social, navigation and semantic tensors. The network is tested in unstructured environments where complex paths emerge according to both internal (intentions) and external (other people, not accessible areas) motivations. As demonstrated, modeling paths unaware of social interactions or context information, is insufficient to correctly predict future positions. Experimental results corroborate the effectiveness of the proposed framework in comparison to LSTM-based models for human path prediction.

1. Introduction

Human trajectory forecasting is a relevant topic in computer vision due to numerous applications which could benefit from it. Socially-aware robots need to anticipate human motion in order to optimize their paths and to better comply with human motion. Delivery robots could reduce energy consumption to get to their destinations avoiding people and obstacles as well. Moreover, anomalous behaviors could be detected using fixed cameras in urban open spaces (e.g., parks, streets, shopping malls, etc.) but also in crowded areas (e.g., airports, railway stations). Despite meaningful results attained using recurrent neural networks for time-series prediction, many problems still remain. In this context, data-driven approaches are usually unaware of surrounding elements which represent one of the main reasons of direction changes in a urban scenario.

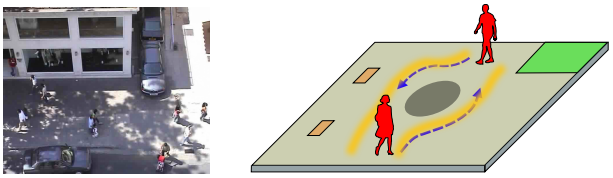

Figure 1. Our goal is to predict future positions of pedestrians in an urban scenario. Since human motion is guided by intentions, experience and the surrounding environment, such elements are encapsulated in our framework along with learned social rules to forecast a social and semantic compliant motion in the crowd.

When approaching their destination, people tend to conform to observed patterns coming from experience and visual stimuli to avoid threats or select the shortest route. Moreover, when walking in public spaces, they typically take into account which kind of *objects* encounter in their neighborhood. Several factors may also lead to velocity direction changes in many situations. For example, when approaching roundabouts (see Fig. 1), people adjust their path to avoid collisions. In some cases, they use different paces according to weather conditions or crowded areas. In this context, LSTM networks have been extensively used over the last years due to their ability to learn, remember and forget dependencies through gates [11, 7]. Such characteristics have made them one of the most suitable solution for solving sequence to sequence problems. To address the limitations of previous works which mainly focuses on modelling *human-human* interactions [1, 9, 8], we propose a comprehensive framework for predicting future positions which are also locally-aware of surrounding space combining social and semantic elements. Scene context information is crucial to improve prediction of future positions adding physical constraints and providing more realistic paths, as demonstrated by early works focused on exploiting *human-space* interactions [12, 13, 2].

In this paper, we propose a data-driven approach allowing an LSTM-based architecture to extract social conventions from observed trajectories and augment such data with semantic information of the neighborhood. More specifi-

cally, our work is built upon the *S*ocial-LSTM model proposed by Alahi *et al*. [1], in which the network is only aware of nearby pedestrians, by embedding new factors encoding also *human-space* interactions in order to attain more accurate predictions. More specifically, we encompass prior motion about the scene as a navigation map which embodies most frequently crossed areas and scene context using semantic segmentation to restrain motion to more plausible paths.

The remainder of the paper is organized as follows. Section 2 reviews main work related to human path prediction. Section 3 describes the proposed model. Section 4 provides our findings while conclusions and suggestions for future work are summarized in Section 5.

2. Related Work

We briefly review main work on human path prediction considering two main kinds of interactions, namely *human-human* and *human-human-space*. The former only models interactions among pedestrians; the latter, takes also into account interactions with surrounding elements, i.e. fixed obstacles, which kind of area is crossed (*e.g*., sidewalk, road) and nearby space.

Human-human interactions. Helbing and Molnár [10] introduced the Social Force Model to describe social interactions among people in crowded scenarios using handcrafted functions to form coupled Langevin equations. More recent works based on LSTM networks mainly rely on the model proposed in [1] where a "social" pooling mechanism allows pedestrians to share their hidden representations. The key idea is to merge hidden states of nearby pedestrians to make each trajectory aware of its neighbourhood. Nevertheless, pedestrians are unaware of nearby elements, such as benches or trees, which could be primary reasons for direction changing when they do not interact with each others. [5] detects groups of people moving coherently in a given direction which are excluded from the pooling mechanism. [8] uses a Generative Adversarial Network (GAN) to discriminate between multiple plausible paths due to the inherently multi-modal nature of trajectory forecasting task. The pooling mechanism relies on relative positions between two pedestrians. The model captures different styles of navigation but does not make any differences between structured and unstructured environments. [21] handles prediction using a spatio-temporal graph which models both position evolution and interaction between pedestrians. [9] embodies *vislet* information within the social-pooling mechanism also relying on mutual faces orientation to augment space perception.

Human-human-space interactions. Sadeghian *et al*. [18] adopt a similar approach to ours, by taking into account both past crossed areas and semantic context to predict social and context-aware positions using a GAN. [3] introduces attractions towards static objects, such as artworks, which deflect straight paths in several scenarios (*e.g*., museums, galleries) but the approach is limited to a reduced number of static objects. [2] proposes a Bayesian framework based on previously observed motions to infer unobserved paths and for transferring learned motions to unobserved scenes. Similarly, in [6] circular distributions model dynamics and semantics for long-term trajectory predictions. [19] uses past observations along with bird's eye view images based on a two-levels attention mechanism. The work mainly focuses on scene cues partially addressing agents' interactions.

Some relevant approaches do not fall into the above two categories. For example, [20] focuses on transfer learning for pedestrian motion at intersections using Inverse Reinforcement Learning (IRL) where paths are inferred exploiting goal locations. [4] attains best performance on the challenging Stanford Drone Dataset (SDD) [17] using a recurrent-encoder and a dense layer. [22] predicts future positions in order to satisfy specific needs and to reach latent sources.

3. Our model

Pedestrian dynamics in urban scenarios are highly influenced by static and dynamic factors which guide people towards their destinations. For example, grass is typically less likely to be crossed than sidewalks or streets for pedestrians. Benches are turned around by people walking with accelerated paces. Moreover, only doors are used to enter buildings. Hence, to forecast realistic paths, it is important to allow human dynamics to be influenced by surrounding space, not only in terms of other people in their neighborhood, but also considering semantics of crossed areas as well as past observations which can represent our experience. To this aim, we extend the Social-LSTM model proposed in [1], as schematized in Fig. 2. More specifically, our framework models each pedestrian as an LSTM network interacting with the surrounding space using three pooling mechanisms, namely *Social, Navigation* and *Semantic* pooling. *Social* pooling mechanism takes into account the neighborhood in terms of other people, merging their hidden states. *Navigation* pooling mechanism exploits past observations to discriminate between equally likely predicted positions using previous information about the scene. Finally, *Semantic* pooling uses semantic scene segmentation to recognize not crossable areas.

Given the i^{th} pedestrian, his/her complete trajectory is represented by the 2-D sequence $\mathcal{T}^i$ =

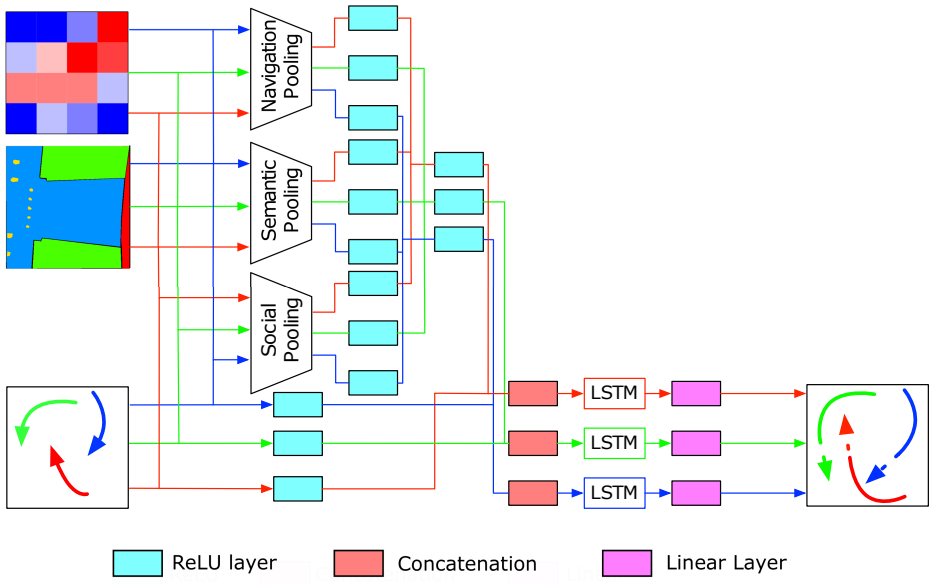

Figure 2. Overview of the proposed model. Trajectories, navigation map and semantic image are fed to the LSTM network and combined using three pooling mechanisms, namely social, navigation and semantic pooling. Future positions are obtained using linear layers to extract key parameters of a Gaussian distribution.

$\{(x_1^i, y_1^i), (x_2^i, y_2^i), ..., (x_{T_{obs}}^i, y_{T_{obs}}^i), (x_{T_{obs}+1}^i, y_{T_{obs}+1}^i), ..., (x_{T_{pred}}^i, y_{T_{pred}}^i)\}$ where T_{obs} and T_{pred} represent the last observation and prediction timestamps, respectively. Each trajectory is then associated to an LSTM network which is described by the following equations:

$$\begin{aligned}
\mathbf{f}_t^i &= \sigma(W_f \mathbf{x}_t^i + U_f \mathbf{h}_{t-1}^i + b_f) \\
\mathbf{i}_t^i &= \sigma(W_i \mathbf{x}_t^i + U_i \mathbf{h}_{t-1}^i + b_i) \\
\mathbf{o}_t^i &= \sigma(W_o \mathbf{x}_t^i + U_o \mathbf{h}_{t-1}^i + b_o) \\
\mathbf{c}_t^i &= \mathbf{f}_t^i \odot \mathbf{c}_{t-1}^i + \mathbf{i}_t^i \odot tanh(W_c \mathbf{x}_t^i + U_c \mathbf{h}_{t-1}^i + b_c) \\
\mathbf{h}_t^i &= \mathbf{o}_t^i \odot tanh(\mathbf{c}_t^i) \qquad (1)
\end{aligned}$$

where $\mathbf{f}_t^i$, $\mathbf{i}_t^i$, $\mathbf{o}_t^i$ are the forget, input and output gates, respectively; $\mathbf{c}_t^i$ is the cell state and $\mathbf{h}_t^i$ is the hidden state. $\odot$ indicates the element-wise product.

The above model, named *Vanilla* LSTM, is unaware of what happens nearby the monitored agent, such as the presence of other people, encountered obstacles and most frequently crossed areas. In fact, such a network could only learn motion patterns and dependencies potentially present in the training set's trajectories. To consider a more rich input representation, *Vanilla* LSTM is enhanced concatenating the following tensors:

Social Tensor. As proposed in [1], we firstly exploit a social pooling mechanism in order to make people aware of their neighbors. More specifically, hidden states of people in the neighborhood are taken into account using an $N_o \times N_o \times D$ social tensor:

$$H_t^i(m, n, :) = \sum_{j \in \mathcal{N}_i} \mathbf{1}_{mn}[x_t^j - x_t^i, y_t^j - y_t^i] h^j{}_{t-1} \qquad (2)$$

where N_o is neighborhood size and D is the dimension of the hidden state. The indicator function $\mathbf{1}_{mn}(x, y)$ checks if (x,y) is inside the (m, n) cell.

Navigation Tensor. People tend to reach building entrances or opposite sidewalks using a limited number of paths. Some areas would, indeed, be more likely to be crossed than others. On the contrary, areas corresponding to obstacles or buildings would not (or less likely to) be crossed making them not *eligible* for generating new position candidates. To measure such crossing probability, we define the *Navigation Map* $\mathcal{N}$ which counts the crossing frequency of squared patches. A smoothing linear filter (i.e., average pooling) is then used to reduce "sharp" frequency transitions. An example of such map in shown in Fig. 3. Given the *Navigation Map* $\mathcal{N}$, we define the rank-2 *Navigation* tensor $N_t^i \in N_n \times N_n$ as follows:

$$N_t^i(m, n) = \mathcal{N}_{mn}, \qquad (3)$$

which extracts the neighborhood's frequency of the i^{th} pedestrian for the (m, n) cell considering all the past observations for such cell.

Semantic Tensor. People may also manifest direction changes due to a number of reasons; for example, they

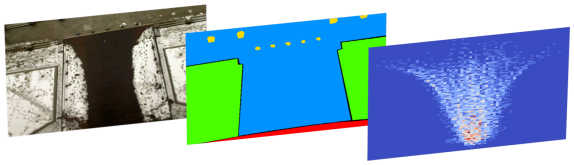

Figure 3. Semantic map is generated from the reference image while the Navigation map is obtained from observed data. The image shows an example of such maps for ETH dataset.

could be approaching a fixed obstacle which must be circled or could avoid streets preferring sidewalks. The aim of the *Semantic* tensor is to capture *why* specific dynamics emerge related to the semantics of surrounding space. Since our datasets do not provide any semantic annotations to model human-space interactions, we define the following semantic classes $\mathcal{C}$ = {grass, building, obstacle, bench, car, road, sidewalk}. A one-hot encoding is used to represent semantic of pixels image. For example, assuming that a pixel represents grass, a location j is represented by a vector $\mathbf{v}_j \in \mathbb{R}^7 = [1\ 0\ ...\ 0]$ according to $\mathcal{C}$.
Given a neighborhood size of N_s, we define a $N_s \times N_s \times L$ tensor S_t^i for the i^{th} pedestrian as follows:

$$S_t^i(m,n,:) = \frac{1}{|\mathcal{S}_{mn}|} \sum_{j \in \mathcal{S}_{mn}} \mathbf{v}_j \tag{4}$$

where $\mathbf{v}_j$ represents the semantic vector of location j and $\mathcal{S}_{mn}$ represents the locations within the (m,n) cell of i^{th} pedestrian. $|\mathcal{S}_{mn}|$ is the number of locations within the (m,n) cell. In other words, for each cell, we extract the occurring frequency of each semantic class.

The above tensors are embedded into three vectors, namely a_t^i, n_t^i, s_t^i while the spatial coordinates into e_t^i. The embedded vectors are concatenated and used as input to the LSTM cell as follows:

$$e_t^i = \Phi(x_t^i, y_t^i; W_e)$$

$$a_t^i = \Phi(H_t^i; W_a)$$

$$n_t^i = \Phi(N_t^i; W_n)$$

$$s_t^i = \Phi(S_t^i; W_s)$$

$$g_t^i = \Phi(concat(a_t^i, n_t^i, s_t^i); W_g)$$

$$h_t^i = LSTM(h_{t-1}^i, concat(e_t^i, g_t^i); W_h) \tag{5}$$

where Φ represents the ReLU activation function and W_h are the LSTM weights. Fig. 4 depicts our pooling mechanisms.

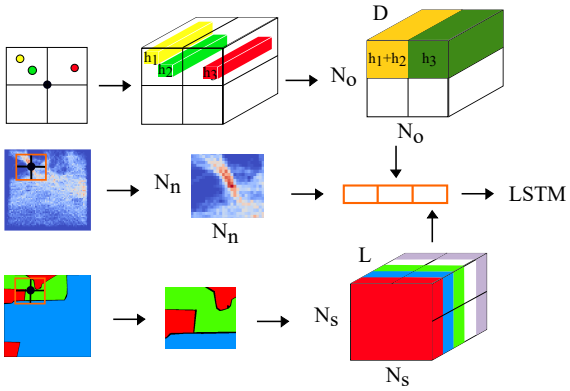

Figure 4. Overview of the pooling mechanisms. Three tensors take into account social neighborhood, past observations and semantics of surrounding space, respectively. Tensors are finally concatenated, processed by ReLU layers and fed to LSTM networks along with embedded positions. Figure also highlights dimensions of each introduced tensor.

Loss Function. Positions are predicted using a bi-variate Gaussian distribution whose parameters are obtained using a $D \times 5$ linear layer as follows:

$$[\mu_t^i, \sigma_t^i, \rho_t^i] = W_l h_{t-1}^i, \tag{6}$$

$$(\hat{x}_t^i, \hat{y}_t^i) \sim \mathcal{N}((x,y); \mu_t^i, \sigma_t^i, \rho_t^i). \tag{7}$$

Finally, the parameters of the network are obtained minimizing the negative log-Likelihood loss L^i for the i^{th} pedestrian as follows:

$$L^i(W_e, W_a, W_n, W_s, W_g, W_h, W_l) = - \sum_{t=T_{obs}+1}^{T_{pred}} log(\mathbb{P}(x_t^i, y_t^i | \mu_t^i, \sigma_t^i, \rho_t^i)). \tag{8}$$

The above loss is minimized for all the trajectories in our training sets.

4. Experiments

In this section, we describe the used datasets along with the evaluation protocol. Next, we present a quantitative analysis to show the effectiveness of our model. Finally, we show some qualitative results of predicted trajectories for challenging situations.

Datasets. For our experiments, we use two datasets: ETH [15] and UCY [14]. ETH contains two scenes (*ETH* and *HOTEL*) while UCY contains three scenes (*UNIV/UCY*, *ZARA-01*, *ZARA-02*). They are captured from a bird's-eye view and involve numerous challenging situations, such as interacting pedestrians, standing people and highly non-linear trajectories. We use a leave-one-out-cross-validation

Metric	Scene	Vanilla LSTM	Social-LSTM [1]	SN-LSTM	SS-LSTM	SNS-LSTM
ADE	ETH [15]	0.52	0.51	**0.47**	0.48	0.58
	HOTEL [15]	0.33	0.31	0.44	**0.24**	0.30
	UNIV [14]	0.52	0.55	0.39	0.43	**0.37**
	ZARA-01 [14]	0.41	0.36	0.29	0.33	**0.28**
	ZARA-02 [14]	0.27	**0.25**	0.28	0.31	0.26
	Average	0.41 ± 0.11	0.40 ± 0.13	0.37 ± 0.09	$\mathbf{0.36 \pm 0.10}$	0.36 ± 0.13
FDE	ETH [15]	2.84	2.82	2.55	2.57	**2.43**
	HOTEL [15]	1.90	1.67	2.25	**1.38**	1.58
	UNIV [14]	2.92	3.04	2.10	2.54	**2.08**
	ZARA-01 [14]	2.35	2.05	1.56	1.81	**1.53**
	ZARA-02 [14]	1.48	**1.42**	1.59	1.63	1.44
	Average	2.30 ± 0.61	2.20 ± 0.71	2.01 ± 0.43	1.99 ± 0.54	$\mathbf{1.81 \pm 0.43}$

Table 1. Quantitative results of our architecture compared to baseline models for ETH and UCY datasets. The errors are reported in meters. Our models attain on average best performance due to the combination of learned social rules, past observations and acquired information of the surrounding space. For the ADE metric our two models, namely SS-LSTM and SNS-LSTM, show comparable results. By contrast, the best FDE value is attained considering the combination of all factors.

approach training the models on N-1 scenes and testing on the remaining one. We average the results over the five datasets.

Evaluation Protocol. To perform an ablation study, we separately test the effect of navigation and semantic factors both along with social interactions, defining two different models, namely SN-LSTM and SS-LSTM, respectively. Finally, we combine all the above effects defining the SNS-LSTM model. As proposed in [15], a pedestrian trajectory is observed for $3.2s$ in order to predict the next $4.8s$. At frame level, we train the network on 8 frames and predict the next 12 frames. As error metrics, we report the *Average Displacement Error* (ADE) and the *Final Displacement Error* (FDE). ADE represents the average Euclidean distance between the predicted and ground-truth positions, while FDE consists in the average Euclidean distance between the final predicted and ground-truth position. The above metrics are defined as follows:

$$ADE = \frac{\sum_{i \in \mathcal{P}} \sum_{t=T_{obs}+1}^{T_{pred}} \sqrt{((\hat{x}_t^i, \hat{y}_t^i) - (x_t^i, y_t^i))^2}}{|\mathcal{P}| \cdot T_{pred}}, \quad (9)$$

$$FDE = \frac{\sum_{i \in \mathcal{P}} \sqrt{((\hat{x}_{T_{pred}}^i, \hat{y}_{T_{pred}}^i) - ({x_{T_{pred}}}^i, {y_{T_{pred}}}^i))^2}}{|\mathcal{P}|}, \quad (10)$$

where $\mathcal{P}$ is the set of pedestrians and $|\mathcal{P}|$ its cardinality, $(\hat{x}_t^i, \hat{y}_t^i)$ are the predicted coordinates at time t and (x_t^i, y_t^i) are the groud-truth coordinates at time t.
We compare our models against two LSTM-based methods, i.e., Vanilla LSTM and Social-LSTM [1].

Implementation Details. The number of hidden units for each LSTM cell and the embedding dimension of spatial coordinates are set to 128 and 64, respectively. The model is trained on a single GPU using TensorFlow library[1] . The learning rate is set to 0.003 and we use RMS-prop as optimizer with a decay of 0.95. Models are trained for 50 epochs. N_o, N_n, and N_s are set to 8, 32 and 20, respectively.

4.1. Quantitative Results

Table 1 shows quantitative results for ETH and UCY datasets. As expected, the worst model is the one which does not take into account any internal/external factor, namely Vanilla-LSTM. The lack of any kind of interactions does not allow the model to reproduce realistic paths. We also notice that SN-LSTM model improves the performance compared to S-LSTM model due to the introduction of discriminative regions especially when two or more path are plausible. A significant improvement is also obtained when semantics is introduced with SS-LSTM, especially for HOTEL scene. Interestingly, S-LSTM performs better on ZARA-02 scene, where results are slightly better compared to SNS-LSTM. The main reason could be ascribed to a number of non-moving pedestrians of such dataset where navigation and semantic factors may not much influence the prediction. Navigation map allows better predicted trajectories on ETH dataset compared to HOTEL dataset. For the latter, the effect of only semantic pooling mechanism reduces the error metrics of $\sim 50\%$. However, the combination of our proposed factor seems to confirm the importance of introducing navigation and semantic factors to achieve more robust predictions.

[1]The code is released at https://github.com/Oghma/sns-lstm/.

4.2. Qualitative Results

To perform a qualitative study, we firstly show static comparisons between our models and baselines for some trajectories and then show several predicted trajectories over time evaluating predictions at successive timestamps.

Static visualization. Fig. 5 shows some examples of predicted trajectories drawn from HOTEL dataset. More specifically, the first column shows cases when our models are able to correctly predict the ground-truth both when people move through the crowd or when they approach the tram. The predicted paths of our models appear able to better capture complex dynamics showing more *natural* motions without continuously adjusting moving directions. We also note that our models are closer to the ground-truth while the baselines end first. Since SN-LSTM and SNS-LSTM are typically similar, our models rely on the navigation map when multiple trajectories can be considered. On the contrary, second column shows cases where SNS-LSTM model appears unable to capture the correct path reaching different destinations. A challenging situation where a standing pedestrian is captured is also shown. In such case, most of the models do not move from the initial position.

Temporal visualization. To temporally evaluate our predictions, we firstly visualize the observed path (*e.g.*, 8 frames) and then select four successive frames, namely $9^{th}, 13^{th}, 17^{th}$ and 19^{th} frame, of each trajectory. For the sake of simplicity, we only report results for SNS-LSTM and S-LSTM models, and ground-truths. More specifically, Fig. 6 shows observed paths and multiple predicted positions at successive timestamps for trajectories drawn from both HOTEL and ETH datasets, respectively. First column demonstrates the effect of the navigation pooling mechanism which avoids deviations from common paths and, at the same time, social pooling mechanism avoids collisions with nearby pedestrians. Second column shows an example of anomaly trajectories which do not frequently occur in the dataset, such as people suddenly stopping after accelerating. Such cases are not properly modelled by both models since predictions tend to move away from the ground-truths. Third column reports the effect of the semantic pooling mechanism which avoids the collision with an object (obstacle surrounded by snow) and predicts more realistic positions. In the last column, our SNS-LSTM model predicts more accurate positions than the S-LSTM model due to the combined effect of navigation and semantic pooling mechanisms.

The above temporal visualization confirms the effectiveness of the introduced elements to better capture complex human behaviors in crowded scenarios where also static objects contribute to the path generation process.

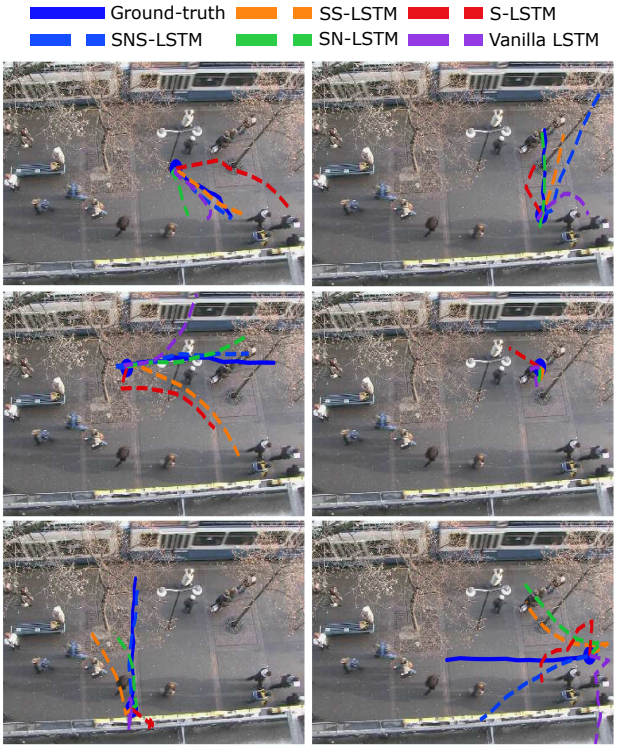

Figure 5. Some examples of predicted trajectories for HOTEL dataset. Ground-truths are shown as solid lines, while predicted trajectories as dashed lines. First column shows cases where predicted positions are very close to the real paths. Second column shows cases where the SNS-LSTM appears not able to correctly predict future positions.

5. Conclusion

We proposed a comprehensive framework to model human-human and human-space interactions for trajectory forecasting in challenging scenarios. The SNS-LSTM merges past observations about the scene and semantics of crossed areas using Navigation and Semantic pooling mechanisms. Such an approach favours more reliable predicted paths when multiple choices are simultaneously possible to reach desired destinations. Previously observed paths and semantic labels of nearby cells allow a pedestrian to be aware of surrounding space and change his/her direction accordingly.

Experimental results show better performance than state-of-the-art methods which do not use context information. Our method, indeed, significantly reduce the error for common metrics and, from qualitative point of view, shows direction changes even when a pedestrian is not influenced by other humans in its neighborhood. Our future work will investigate different datasets (*e.g.*, Stanford Drone

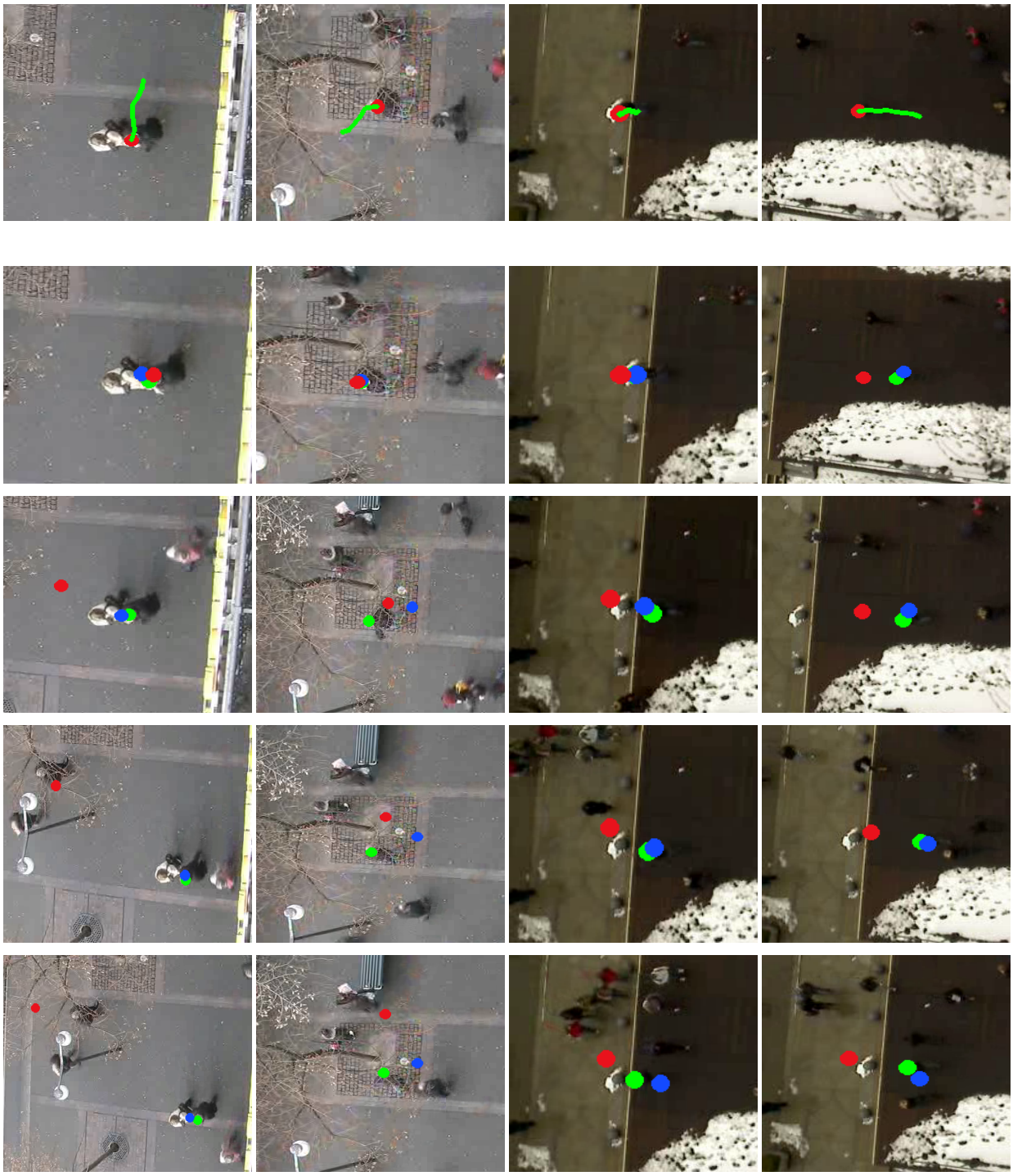

Figure 6. Temporal sequences visualization for different tracks drawn from both HOTEL and ETH dataset. The circles represent ground-truth (green), SNS-LSTM model (blue) and S-LSTM model (red), respectively. For each column, the first image shows the observed path (in green) which corresponds to 8 frames (the three circles are superimposed), while the remaining ones show the 9^{th}, 13^{th}, 17^{th} and 20^{th} predicted frames, respectively. In case of significant accelerations (1^{st} column) our SNS-LSTM model remains close to the ground-truth compared to S-LSTM baseline avoiding obstacles and other pedestrians. By contrast, anomaly behaviors, i.e.. stopping after accelerating, (2^{nd} column) are not correctly predicted by both models. Semantic pooling mechanism avoids our model to collide with obstacles (3^{th} column). Relying mainly on both social and navigation pooling mechanisms, our model predicts better positions than the S-LSTM model (4^{th} column). Images are cropped to focus on monitored pedestrians' neighborhood.

Dataset [16]) where more complex dynamics are captured as well as multiple agents settings since several elements, such as cars, cyclists or skateboarders, typically share the same environment.

Acknowledgements. This work is partially supported by MIUR (PRIN-2017 PREVUE grant) and UNIPD (BIRD-SID-2018 grant). We gratefully acknowledge the support of NVIDIA for their donation of GPUs used in this research.

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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 3, 5

[2] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese. Knowledge transfer for scene-specific motion prediction. In *European Conference on Computer Vision*, pages 697–713. Springer, 2016. 1, 2

[3] F. Bartoli, G. Lisanti, L. Ballan, and A. D. Bimbo. Context-aware trajectory prediction. *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1941–1946, 2018. 2

[4] S. Becker, R. Hug, W. Hübner, and M. Arens. Red: A simple but effective baseline predictor for the trajnet benchmark. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 138–153, Cham, 2019. Springer International Publishing. 2

[5] N. Bisagno, B. Zhang, and N. Conci. Group lstm: Group trajectory prediction in crowded scenarios. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 213–225, Cham, 2019. Springer International Publishing. 2

[6] P. Coscia, F. Castaldo, F. A. Palmieri, A. Alahi, S. Savarese, and L. Ballan. Long-term path prediction in urban scenarios using circular distributions. *Image and Vision Computing*, 69:81 – 91, 2018. 2

[7] T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Soft + hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural Networks*, 108:466 – 478, 2018. 1

[8] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 1, 2

[9] I. Hasan, F. Setti, T. Tsesmelis, A. D. Bue, F. Galasso, and M. Cristani. Mx-lstm: Mixing tracklets and vislets to jointly forecast trajectories and head poses. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6067–6076, 2018. 1, 2

[10] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Phys. Rev. E*, 51:4282–4286, May 1995. 2

[11] A. Karpathy, J. Johnson, and F.-F. Li. Visualizing and understanding recurrent networks. *CoRR*, abs/1506.02078, 2015. 1

[12] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012. 1

[13] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila. Context-based pedestrian path prediction. In *European Conference on Computer Vision*, pages 618–633. Springer, 2014. 1

[14] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. *Comput. Graph. Forum*, 26:655–664, 2007. 4, 5

[15] S. Pellegrini, A. Ess, K. Schindler, and L. V. Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268, 2009. 4, 5

[16] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 549–565, Cham, 2016. Springer International Publishing. 8

[17] A. Sadeghian, V. Kosaraju, A. Gupta, S. Savarese, and A. Alahi. Trajnet: Towards a benchmark for human trajectory prediction. *arXiv preprint*, 2018. 2

[18] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. *CoRR*, abs/1806.01482, 2018. 2

[19] A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese. Car-net: Clairvoyant attentive recurrent network. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[20] M. Shen, G. Habibi, and J. P. How. Transferable pedestrian motion prediction models at intersections. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4547–4553, 2018. 2

[21] A. Vemula, K. Muelling, and J. Oh. Social attention: Modeling attention in human crowds. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7, 2018. 2

[22] D. Xie, T. Shu, S. Todorovic, and S. Zhu. Learning and inferring dark matter and predicting human intents and trajectories in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7):1639–1652, July 2018. 2