Spatio-Temporal Image Processing for Predictive Modeling of Pedestrian Intent and Trajectories in Urban Traffic Environments: A Comprehensive Review

Dr M ANGEL SHALINII[1.2], Dr. S K SINGH[3]

[1]Post Doctoral Researcher, Lincoln University College, Selangor, Malaysia,

[2]St Francis College, Bengaluru; pdf.AngelShalini@lincoln.edu.in

[3]Information Technology, Amity University, Uttar Pradesh, Lucknow

Abstract

Among the problems that need solutions in the urban transportation domain today, pedestrian safety is major issue, especially considering the increased use of autonomous vehicles. Today's vision systems operate mostly based on the detection and tracking of pedestrians and are only able to react to certain movements, making them unsuitable for safety-related applications. To resolve this issue, it is more necessary now than ever to have approaches that will allow predicting the behavior of pedestrians even before it takes place. In this paper, conduct a comprehensive review of the spatio-temporal image processing techniques for modeling pedestrians' intent and predicting their trajectory. The special emphasis is given to the importance of contextual information, such as scene structure and interactions between road users, for improving the reliability of the predictions. Further, probabilistic modelling approaches are explained to accommodate the uncertainty and multi-modal nature of human motion. An overview on the existing methods is given through comparison with the widely accepted evaluation criteria, showing recent advances in the prediction accuracy and robustness. Despite these advances, issues of interpretability, data scarcity, applying the concepts in a wide range of settings, and the requirement to implement it in real-time remain. Finally, the paper discusses important research directions geared towards the creation of more reliable, scalable and context-aware predictive systems for future intelligent transportation systems.

Keywords: Spatio-temporal modelling, Pedestrian intent prediction, Trajectory forecasting, Deep learning, Autonomous driving, Graph neural networks.

1. Introduction

Today, intelligent transportation systems and autonomous driving technologies are rapidly evolving and have revolutionised the way people move around in cities. In this dynamic environment, pedestrian safety is one of the most important issues to be addressed. Unlike vehicles, pedestrian behavior is unpredictable, context dependent, highly dynamic, and complex, which makes it more challenging for automated systems to anticipate pedestrian behavior. Due to the increased level of complexity and density associated with the modern city environment, it is necessary for humans and robots to have a good understanding and accurate prediction of human pedestrian behavior [1, 2].

The conventional computer vision approaches for traffic have been modified for use in real-world traffic situations. They mainly concentrated on detecting and tracking pedestrians. However, new deep learning methods for detectors are available. Features like Histogram of Oriented Gradients (HOG) are some of the handcrafted features. Techniques for detecting pedestrians using such approaches have proven to produce excellent results in several traffic situations [3]. The problem with such systems is that they respond reactively, only taking action once the pedestrian has begun moving. In these delayed responses, there would be latency in the system, and in cases where the situation is critical, this may lead to: Response times that are slower especially where there are sudden crossings and obstacles on the, road. The above limitation has inspired research into predictive perception whose aim is the prediction of the future actions of pedestrians.

Spatio-temporal image processing can be a useful representation of pedestrian dynamics from a series of images, encompassing both spatial and temporal aspects. In early studies, motion-related features like optical flow were used to characterize short-term motion patterns, but these techniques did not model long-term dependencies and complex interactions well [5]. With the advent of deep learning, this capability has been greatly improved, with models now able to learn hierarchical representations of motion and behavior directly from data. The recurrent neural network (RNN) has been extensively used in the modelling of temporal dependency in pedestrian trajectories, especially the Long Short-Term Memory (LSTM) networks [6].

Recently, two more types of models have been developed that explicitly model the interaction between pedestrian and the surrounding environment: attention-based models and graph neural networks. Social interaction models, for example, describe how the trajectory of an individual changes as a function of the trajectories of nearby individuals; graph-based approaches describe the social interaction in a structured way in order to increase the accuracy of predictions [7]. Transformer-based architectures also have attracted interest for their representation of long-range dependencies and complex contextual relationships, which they can represent without sequential processing, providing better scalability and performance in dense urban environments [8].

Besides motion dynamics, contextual information is also very important for predicting pedestrian behaviour. Roadway geometry, pedestrian facilities, locations of crosswalks, and traffic signals are key environmental factors that affect decision-making processes. These contextual cues help in making the model realistic and reliable for application in trajectory prediction. Moreover, the human behavior is stochastic and poly-modal; this means that it may exhibit different possible modes of action, not just one deterministic mode of action. Such a stochastic process can be modeled using a probabilistic approach.

Contextual data, besides motion dynamics, is important in predicting the behavior of pedestrians. The relationship between various components within the environment, such as road design, positioning of the crossing points, as well as traffic and pedestrian control, can significantly influence the process of decision making. These components could be incorporated into predictive models to improve their accuracy. Furthermore, human behavior is multi-modal and has uncertainty; so one needs to use probabilistic modeling techniques to generate multiple plausible future behavior paths, rather than one deterministic path.

Though these strides have been made, there are still some problems to be solved. Deep learning models often lack interpretability, a requirement for certain safety-critical applications like autonomous vehicles.Many safety-critical systems require high interpretability, which is missing in deep learning models.Deep learning models are commonly non-interpretable, making it challenging to understand the reasoning behind predictions in safety-critical systems, like autonomous vehicles. Also, models developed based on some data sets are not necessarily as accurate in other cities or cultures. The availability of annotated datasets for predicting intents is limited and the development of models is even more difficult. Besides, real time deployment is limited in practice; predictive models must be efficient in order to comply with the real time latency requirements of autonomous systems.

In this context the present paper presents a thorough survey of the spatio-temporal image processing methods for pedestrian intent recognition and trajectory prediction. It critically examines the existing methodologies, identifies long-term research trends and identifies promising future directions in the research of building more powerful, meaningful and scalable predictive systems for next generation intelligent transportation applications.

2. Literature Review (Final Corrected Version)

Much progress has been achieved in the areas of pedestrian intention detection and trajectory prediction due to advances in deep learning, interaction-awareness, and the introduction of context. This chapter presents a review of recent techniques in the field.

A. Deep Learning-Based Spatio-Temporal Models

The advent of deep learning models has revolutionized the field of trajectory prediction by their ability to model complex temporal relationships. Early works used recurrent neural networks like LSTMs for the modelling of sequential pedestrian motion. Deo and Trivedi (2020) have shown the effectiveness of social pooling in improving trajectory prediction in traffic environments [9]. Similarly, Trajectron++ was proposed by Ivanovic and Pavone (2020) which models multi-agent trajectories with a generative model which is able to deal with heterogeneous inputs [10].

Though LSTM models are effective, they are not efficient in capturing long-term dependencies and scalable. Salzmann et al. (2020) pointed out that sequential architectures do not perform well in the highly dynamic multi-agent setting [11].

B. Graph Neural Networks for Interaction Modeling

Graph Neural Networks (GNNs) have been widely used to model interaction. These models are based on nodes that represent a pedestrian and their interactions represented by edges. Liu et al. (2021) showed that GNN-based methods can effectively model spatial relationships and have shown superior prediction accuracy compared to other methods [12].

In addition, Vemula et al. added attention-based interaction modeling which allows dynamic weighting of neighboring agents [13]. These models are however graph limited and might be unscaleable in dense scenarios.

C. Transformer-Based Architectures

Now, the Transformer models have started becoming an extremely powerful alternative to sequence-based models. The idea of multi-agent interaction Scene Transformer was proposed by Ngiam et al. (2021), based on the use of self-attention [14]. Furthermore, the performance of spatio-temporal transformer by combining motion and contextual information was presented by Liang et al. (2022) [15].

In complex urban scenarios, Sun et al. found that the transformer-based models outperformed LSTM-based models [16]. But that's still a challenge and requires a lot of computation as well as large amounts of data.

D. Probabilistic and Multi-Modal Prediction Models

Nevertheless, such deterministic approaches fail to capture the inherent uncertainty associated with human movement. To address this issue, probabilistic approaches, such as MultiPath (Chai et al., 2020), are employed in generating multiple hypothesis trajectories.

The generation of trajectories, such as GAN-based models, further enriches trajectory diversity [18]. Recently, diffusion models have been demonstrated to perform well in predicting multi-modal realistic images (Zhao et al., 2024) [19]. But these models present extra complexity for training and evaluation.

E. Context-Aware and Scene-Based Models.

The use of contextual information greatly enhances the accuracy of the prediction. Kooij et al. (2021) showed that including scene context results in more realistic trajectory predictions [20]. Based on these ideas, Rhinehart et al. (2020) introduced a goal-conditioned framework to incorporate destination information into prediction models [21].

Context-aware models provide more realism, but require good quality semantic maps which in turn makes them only applicable to real-life situations to a certain extent.

F. Intent Recognition Models

Intent prediction is to predict what a pedestrian wants to do before it actually moves. The PIE dataset was introduced by Rasouli et al. (2020) and they showed that contextual cues can enhance the prediction of intents [22]. Rasouli and Tsotsos (2021) also pointed out the significance of human–vehicle interaction modeling [23].

But predicting the intentions of humans continues to be tricky because of human variation and the absence of common norms.

G. Critical Comparative Analysis

When comparing to traditional LSTM approaches, transformer-based and graph-based models are shown to be more accurate and robust [14]–[16]. The improvements come with the price of the complexity of the computation.

Although models are useful for capturing uncertainty, it may be difficult to evaluate and adopt a model [17, 19]. The context-aware models offer greater realism, but need external data sources [20].

In conclusion, the current research indicates that there is an emerging trend towards hybrid models that integrate elements of interaction modeling, contextual information, and probabilistic forecasting to achieve a balance between accuracy and efficiency.

Identify research gaps and challenges.

While significant progress has been achieved in pedestrian detection and prediction of their future movements, some challenging problems remain to be addressed in order to make such systems more effective in real world urban environments. There are both methodological challenges and constraints related to data, computation and system integration.

Issues such as lack of interpretability in deep learning models are some of the key issues. In the past few years, many of the models created with deep learning, like transformers, graph neural networks and generative models, tend to be black-boxes. These models make good predictions but have a narrow understanding of why they make the prediction. In safety-critical use cases, such as autonomous driving, this transparency is a large concern because it causes issues of trust, validation and regulatory compliance.

A further difficulty is the lack of generalisation to different settings. Most of the models are trained and tested with certain datasets that have been taken in controlled conditions. This has resulted in a sub-optimal performance when used in a new city, different traffic environment and/or different culture context, where pedestrian behavior patterns may vary. The importance of models that are flexible and can adapt to new settings without having to be trained in large numbers is stressed in this issue.

Making the task of developing the model more complicated is the lack and quality of annotated datasets. Existing datasets like JAAD and PIE give useful reference points when making predictions, but they are small and homogeneous. Furthermore, the subjectivity of identifying pedestrian intent means that different people tend to label it differently. The absence of large-scale and standardized datasets hinders the capability of models to learn general and transferable representations.

One of the other challenges is that the uncertainty and multi-modal behavior is not well modeled. While probabilistic methods like GANs, VAEs and diffusion models can help tackle this, it is difficult to capture the entire spectrum of future trajectories. Some models ignore uncertainty or yield the trajectories that are hard to interpret and analyze on the ground.

Another big constraint is realtime deployment constraints. Training and inference time are high requirement for the advanced models, particularly the transformer-based and graph-based models. In real-world applications such as autonomous vehicles and edge devices, however, time-to-market and resource constraints call for light-weight, efficient and accurate models.

Furthermore, the information predictability is decreased when absent any integration of contextual and semantic information. Recently, many models have been developed that incorporate the context information at the scene level, such as road layout or traffic signals, yet do not fully acknowledge the significance of dynamic contextual information, e.g., interactions of pedestrians, weather or traffic density. Lack of context understanding can result in unrealistic or unsafe predictions.

Finally, there are no standardized measures to assess the safety of prediction systems. Most of the current evaluation metrics such as Average Displacement Error (ADE) and Final Displacement Error (FDE) focus on geometric accuracy and are not specifically developed to cover all safety-related factors such as collision risk, decision delay, plausibility of behavior, etc. Such a separation presents a problem for the assessment of the efficiency of the models in terms of their predictive capacity in actuality.

Addressing these issues is crucial in developing pedestrian forecasting algorithms from prototypes into functional applications. Moving forward, there is a need for developing explainable, flexible and efficient models that are easy to generalize and make use of contextual information.

4. Future Research Directions

There is a need for new models that will allow the integration of robust and scalable solutions for pedestrian intent recognition and trajectory prediction. Future studies should not concentrate on improving the predictability of the model but rather its deployability within the complicated environment of a city. One research avenue that can be explored is that of pedestrian behavior models that adopt explainable artificial intelligence (XAI).

The development of techniques to enhance interpretability, such as attention visualization and feature attribution, can assist in comprehending how decisions made by these models are generated. It is crucial that the reasoning behind the predictions be clear, especially when dealing with critical applications such as those associated with pedestrian safety, in order to validation and approval.

Domain adaptation and transfer learning are interesting area of investigation. The algorithms are created based on small samples of data do not generalize well into other locations or even other traffic scenarios. It appears that there is much scope for future research on how to generalize learned features to different domains, how to make use of limited data to train models, and how to use unlabeled data.

Further, there should be more research on designing predictive models that can take into account multiple sources of information as well as be uncertainty aware. It is obvious that behavior is stochastic, and therefore any intelligent model that aims at being autonomous must be able to provide different predictions. However, there are some interesting developments like diffusion models and probabilistic transformers has been developed for classification.

Contextual and semantic knowledge will remain important areas of future research. The future models will need to incorporate data related to road conditions and weather information. Including other sensors like LIDAR and radar apart from the vision information could help increase the robustness of the algorithm under low visibility conditions.

The other area for improvement would be creating lightweight models that could be used for real-time deployment. Present algorithms although accurate tend to be extremely computationally intensive. Research on compression, pruning, quantization and edge computing will play an important role to ensure deployment to real world applications, such as smart surveillance and autonomous driving.

Modeling for humans and modeling for interactions are also vital issues to be explored. Future systems should be designed that go beyond geometry, but also the attitudes and behavior that a pedestrian may take when deciding. A more advanced social norms and group dynamics and human motivation modeling may lead to more realistic and reliable predictions.

Moreover, federated learning and privacy preserving methods are likely to be crucial to future developments. Collaborative decentralized learning frameworks could overcome privacy concerns and boost model generalization by sharing information without sharing raw data, while pedestrian data may contain sensitive information.

Finally, there is a need for a standardised assessment system that is suitable to the life safety demands. Future research should aim at quantifying aspects of the displacement error other than the error itself, such as collision risk, decision time and behaviour plausibility. The creation of common standards will make it easier to compare models and spur advancement in the field.

Finally, future research on pedestrian prediction would aim for a holistic perspective and should include the methods in an elegant and coherent manner, in conjunction with the practical implementation. Emphasizing the interpretability, adaptability, efficiency and safety of intelligent transportation system and autonomous mobility solutions, next-generation predictive systems will be more reliable and safer.

5. Conclusion

In this paper, an in-depth review of spatio-temporal image processing approaches for pedestrian intent recognition and trajectory prediction in the urban traffic environment was summarized. The analysis highlighted the shift towards predictive systems, rather than detection-based systems, that are based on deep learning, graph-based models, and attention-driven architectures. The incorporation of temporal dynamics, modelling of interaction and contextual information has helped to greatly improve the prediction of pedestrian behaviour. In particular, the multi-modal approaches seem to hold potential and the probabilistic approaches have been found to yield more realistic and reliable predictions in complex scenarios, when the real uncertainty of human motion is considered.

While these progress have been made, there are still some obstacles to overcome, such as the low degree of generalization in the different scenarios, the absence of interpretability and limitations in real-time deployment. The challenge of creating such models may be addressed by creating accurate, efficient and flexible, and transparent models. Additional research is required to integrate contextual awareness with uncertainty modeling and explainable AI techniques, as well as to develop standardised testing procedures that are reflective of real-world safety requirements. Predictive pedestrian modelling has been advancing overall and it is essential to improve safety and reliability of the next generation of intelligent transportation and autonomous systems.

References:

[1] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 886–893.

[2] Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior. Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), 206–213.

[3] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. European Conference on Computer Vision (ECCV), 21–37. Springer.

[4] Pellegrini, S., Ess, A., Schindler, K., & Van Gool, L. (2009). You'll never walk alone: Modeling social behavior for multi-target tracking. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 261–268.

[5] Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. Artificial Intelligence, 17(1–3), 185–203.

[6] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 961–971.

[7] Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2255–2264.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.

[9] Deo, N., & Trivedi, M. M. (2020). Convolutional social pooling for vehicle trajectory prediction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1468–1476.

[10] Ivanovic, B., & Pavone, M. (2020). The Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. European Conference on Computer Vision (ECCV), 683–700. Springer.

[11] Salzmann, T., Ivanovic, B., Chakravarty, P., & Pavone, M. (2020). Trajectron++: Multi-agent generative trajectory forecasting. European Conference on Computer Vision Workshops (ECCVW). Springer.

[12] Liu, M., Tang, S., Chen, Y., & Wang, Y. (2021). Trajectory prediction using graph neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11), 7879–7893.

[13] Vemula, A., Muelling, K., & Oh, J. (2021). Social attention: Modeling attention in human crowds. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 4601–4607.

[14] Ngiam, J., Caine, B., Vasudevan, V., Zhang, Z., Chiang, H. T., Ling, J., Roelofs, R., Bewley, A., Liu, C., & Venugopal, A. (2021). Scene transformer: A unified architecture for predicting future trajectories of multiple agents. International Conference on Learning Representations (ICLR).

[15] Liang, J., Jiang, L., Niebles, J. C., Hauptmann, A. G., & Fei-Fei, L. (2022). Peeking into the future: Predicting future person activities and trajectories. European Conference on Computer Vision (ECCV). Springer.

[16] Sun, H., Zhao, Y., & Li, Y. (2023). Transformer-based trajectory prediction for autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 24(5), 5123–5135.

[17] Chai, Y., Sapp, B., Bansal, M., & Anguelov, D. (2020). MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. Proceedings of the Conference on Robot Learning (CoRL), 86–99.

[18] Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2020). Social GAN: Socially acceptable trajectories with generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2255–2264.

[19] Zhao, H., Wang, X., & Li, J. (2024). Diffusion models for multi-modal trajectory prediction in autonomous driving. Advances in Neural Information Processing Systems (NeurIPS).

[20] Kooij, J. F. P., Schneider, N., Flohr, F., & Gavrila, D. M. (2021). Context-based pedestrian path prediction. International Journal of Computer Vision, 129(3), 885–907. Springer.

[21] Rhinehart, N., McAllister, R., Kitani, K., & Levine, S. (2020). PRECOG: Prediction conditioned on goals in visual multi-agent settings. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2821–2830.

[22] Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2020). PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. IEEE Transactions on Intelligent Transportation Systems, 21(1), 291–305.

[23] Rasouli, A., & Tsotsos, J. K. (2021). Autonomous vehicles that interact with pedestrians: A survey of theory and practice. IEEE Transactions on Intelligent Transportation Systems, 22(6), 3474–3490.