Convolutional Neural Network Based Forecasting Pedestrian Trajectory with Sequence Modelling

A. Roselin M.C.A.,
Assistant Professor
Department of Computer Science with Data Analytics
Dr. N.G.P. Arts and Science College
Coimbatore - 641 048,
Mail-id: *roselinarokyasamy22@gmail.com*

Dr. P. Prabhusundhar M.C.A., Ph.D.,
Assistant Professor
PG & Research Department of Computer Science
Gobi Arts & Science College
Gobichettipalayam – 638 453
Mail-id: drprabhusundhar@gmail.com

Dr. N. V. Poornima
Assistant professor
Symbiosis centre for management studies, Bengaluru
Symbiosis international University, pune.
Mail-id: poornima@scmsbengaluru.siu.edu.in.

Abstract

Predicting the future path of pedestrians is a vital task for enhance the safety and reliability of autonomous technologies, include self-driving vehicle, intelligent surveillance platforms and robotics. The complex and dynamic nature of human movement, which is influenced by social interactions and environmental context, is often considered a significant challenge to accurate forecasting. Nowadays deep learning - based prediction methods have demonstrated better performance to conventional approaches for various trajectory data. These approaches face challenges in improving accuracy, efficiency, and dependability. In this work, an optimized Convolutional Neural Network (CNN) is proposed for forecasting human future trajectories. A compact deep learning-based CNN architecture is developed for the extraction of spatial-temporal dependencies, while scene semantics and social dynamics are incorporated to enhance contextual understanding. Computational efficiency has been maintained throughout the model design to ensure suitability for real-time applications. The model has been demonstrated to perform well compared to state-of-the-art techniques even though its light weight in nature. Moreover, faster inference has been achieved, facilitating its adoption in real-time environments. The proposed optimized CNN model has exhibited superior performance, highlighting its effectiveness even with reduced computational demands. ADE and FDE scores of 1.0 and 1.28, respectively, have been recorded. In addition, reduced inference time was observed, further confirming the model's suitability for real-time pedestrian trajectory forecasting. These findings confirm that accurate and reliable predictions can be delivered through the proposed model, while maintaining a computationally efficient and lightweight architecture.

Keywords: Convolutional Neural Network, Trajectory Prediction, Spatial-Temporal, Scene Semantics, Social Dynamics

1 Introduction

Predicting pedestrian trajectories is a critical task in has been recognized as essential in various autonomous systems, particularly in applications such as self-driving vehicles, robotics navigation, and crowd analysis. In densely populated urban environments, anticipating human movement is essential for ensuring safety and seamless human-machine interaction. But human movement is hard to predict because it can change quickly and is often influenced by other people nearby. In recent years, the field of trajectory prediction has been significantly advanced by deep learning approaches through the capture of complex spatiotemporal patterns in pedestrian movements. Among these, the utilization of CNNs has been explored due to their ability to extract spatial features effectively from image-based data, have shown promising capabilities in modelling sequential and spatial data. Their ability to extract hierarchical features and local dependencies makes them well-suited for learning motion dynamics from trajectory data.

This research investigates the effectiveness of optimized CNN-based architectures for predicting future pedestrian trajectory. We propose a model that leverages convolutional layers to learn rich representations of observed motion sequences, capturing both local temporal dependencies and spatial movement patterns. By applying CNNs to trajectory prediction, we focus on achieving both accurate predictions and low computational overhead. paving the way for real-time applications in dynamic environments. Research in this area has been aimed at developing models that can learn pedestrian behavioural patterns under varying environmental conditions. The results obtained from experiments have been analysed to validate the effectiveness of our approaches. Challenges related to occlusions, dynamic environments, and variability in human behaviour have been addressed through recent advancements.

2 Review of related work

Pedestrian future trajectory prediction has been highly studied due to its significance in autonomous systems, including autonomous vehicles, robotics, and surveillance system. Early approaches primarily relied on physics-based models, including the Social Force Model introduced by (Helbing., 1995) which simulated pedestrian dynamics based on attractive and collision avoidance interactions. These methods struggled to capture the human behavior in dynamic environments. According to (Antonini, G., 2006) Varaity of studies have modeled pedestrian walking as a sequence of discrete actions within a dynamically discretized space. Discrete choice models have been employed to capture behavioral elements such as direction changes, acceleration, and interaction with nearby pedestrians. Approaches like the cross-nested logit model (CNL) have effectively represented spatial correlation and have been verified onextensive large-scale real-world datasets.

Yamaguchi K et.al., (2011) proposed an agent-based behavioural model where pedestrians are treated as decision-makers influenced by personal, social, and environmental cues. Their method employs behaviour prediction through energy minimization while accounting for latent variables such as destination intent and social group dynamics. By integrating these hidden influences into the model, they achieved improved tracking and prediction accuracy, especially in scenarios with limited observations.

LSTM based pedestrian trajectory prediction method was proposed by Alahi et al. (2016) treating it as a sequence generation task informed by past movements. Unlike traditional approaches such as the Social Forces model, their method learns motion behavior directly from data without hand-crafted rules. The model achieved superior results across several public datasets, showcasing its effectiveness in capturing intricate patterns of human movement.

Zamboni, S., et al. (2020) proposed a innovative approch using a two-dimensional convolutional neural network was employed for the prediction of pedestrian trajectories, moving beyond conventional RNN-based architecture. Their approach capitalized the performance on benchmark datasets like ETH and TrajNet by leveraging spatial feature extraction, accurate position representation, and robust data augmentation techniques. In addition, their investigation intooccupancy-based social modeling reveled its limited effectiveness in capturing complex social interactions amoung pedestrians.

Yang J., et al. (2022) tackled the issue ofvisual blind spots in pedestrian interaction modeling by proposing a spatio-temporal graph-based convolutional network was used for trajectory prediction. Their approach constructs aspatiotemporal graph to extract pedestrian properties, eliminates superfluous connections based on visual boundaries, and employs a TXP-CNN to forecast future trajectories.

Bai et al. (2018) conducted a comprehensive comparison between convolutional and recurrent architectures for sequence modelling tasks. According to their study demonstrated that basic convolutional networks can outperform traditional recurrent models like LSTMs across various benchmarks, longer effective memory and improved performance. These results challenge the default reliance on recurrent networks and suggest that convolutional architectures could serve as a powerful alternative for sequence modelling.

Despite the success of recurrent and graph-based architectures, Trajectory prediction has also been investigated using convolutional neural networks (CNNs). CNNs are recognized for efficiently extracting local spatial and temporal features. Recent studies have started to leverage CNNs for modelling pedestrian trajectories by treating motion sequences as spatiotemporal patterns, enabling fast and parallelizable inference. Some works have integrated CNNs with attention modules or graph structures to improve prediction performance. Nevertheless, there remains a need for further investigation into optimized CNNs architectures specifically tailored for trajectory prediction tasks.

3 Trajectory prediction method

This section begins by outline the dataset used, evaluation metrics, and implementation specifics. Following this, we report the experimental outcomes of training the proposed model and baseline approaches using the different pre-processing methods.

3.1 Definition of the Problem

Pedestrian trajectory prediction seeks to forecast a pedestrian's future positions based on where they have been moving next. Specifically in a given scenario with multiple pedestrian's positions are observed over a certain period of time, their observed coordinates are referred as $\mathbf{X}^i_{\text{obs}} = \{(x^i_1, y^i_1), (x^i_2, y^i_2), \ldots\ldots, (x^i_{T_{\text{obs}}}, y^i_{T_{\text{obs}}})\}$ and the goal is to predict their future positions from time $T_{\text{obs}} + 1$ upto T_{pred} assuming time starts at 0.Time is treated as discrete, meaning the time gap between each frame is constant. Each pedestrian's Positions are captured using (x, y) coordinates measured in meters, relative to a scene-specific fixed reference point. The future path of pedestrian i is denoted as:

$$\widehat{\mathbf{P}}^{\boldsymbol{i}}_{Pred} = \left\{\left(\widehat{\boldsymbol{x}}^{\boldsymbol{i}}_{\boldsymbol{T_{obs}+1}}, \widehat{\boldsymbol{y}}^{\boldsymbol{i}}_{\boldsymbol{T_{obs}+1}}\right),\ \ldots,\ \left(\widehat{\boldsymbol{x}}^{\boldsymbol{i}}_{\boldsymbol{T_{pred}}}, \widehat{\boldsymbol{y}}^{\boldsymbol{i}}_{\boldsymbol{T_{pred}}}\right)\right\},$$

Actual (ground truth) future trajectory is

$$\mathbf{P}^i_{\text{pred}} = \left\{\left(x^i_{T_{\text{obs}}+1}, y^i_{T_{\text{obs}}+1}\right),\ \ldots,\ \left(x^i_{T_{\text{pred}}}, y^i_{T_{\text{pred}}}\right)\right\}$$

We use $\boldsymbol{X}^i$ to represent the past trajectory, $\widehat{\mathbf{P}}^i$ represents the predicted positions of the future trajectory, ground truth future trajectory positions as $\mathbf{P}^i$.

3.2 Data Pre – Processing

To achieve low error rate, data must be properly pre-processed before training the model. In this method we applied the Kalman filter approach as the part of the pre – processing pipeline to reduce noisy trajectory data and handle the missing values. As a recursive state estimation algorithm,the Kalman Filter is effective for modelling systems with uncertain and noisy observations, making it particularly well suitable pedestrian trajectory data due to its ability to model linear motion with Gaussian noise.

3.2.1 Kalman Filter

The Kalman Filter works two main core phrases: Prediction and update. It relies on the assumption that the system dynamics can be modelled using linear equations and that the noise follows a Gaussian distribution.

3.2.1.1 State representation

At every time step t, the state vector x_t represents the hidden state of the system. In the Context of pedestrian motion, the state usually includes position and velocity in two-dimensional space.

$$x_t = \begin{bmatrix} x_t \\ y_t \\ v_{x_t} \\ v_{y_t} \end{bmatrix} \quad (1)$$

Where x_t, y_t: Position coordinates, v_{x_t}, v_{y_t}:Velocity components

3.2.1.2 Prediction phase

A constant velocity motion modelis applied to forecast the pedestrian's next state is predicted from their current state.

$$x_{t\ |t-1} = F \cdot x_{t-1} \quad (2)$$

$$p_{t\ |t-1} = F \cdot p_{t-1}.F^T + Q \quad (3)$$

Where $x_{t\ |t-1}$: Estimated state prediction, $p_{t\ |t-1}$: Estimated covariance of prediction error, F: State transition matrix, Q: Process noise covariance.

For constant velocity, F might look like:

$$F = \begin{bmatrix} 1 & 0 & \Delta_t & 0 \\ 0 & 1 & 0 & \Delta_t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4)$$

3.2.1.3 Update Step (Correction)

When an observation Z_t is available (e.g., position), the predicted state is corrected:

Kalman Gain: $K_t = P_{t\ |t-1}H^T(HP_{t\ |t-1}H^T + R)^{-1}$ (5)

State Update: $x_t = x_{t\ |t-1} + K_t(Z_t - Hx_{t\ |t-1})$ (6)

Covariance Update: $P_t = (I - K_tH)P_{t\ |t-1}$ (7)

Where K_t: Gain matrix used for state correction (Kalman gain), H: Matrix representing observations R: Covariance of the observation noise,

Z_t: Observed position. (Welch, G., & Bishop, G. (1995))

3.2.2 Data Normalization

In pedestrian trajectory prediction, spatial coordinates (x, y) are scene dependent to address the variability in coordinate range across different scenes, min-max normalization is applied to all spatial data.This inconsistency can hinder model generalization and learning efficiency. To ensure model consistency and model performance this, we apply min-max normalization on each spatial coordinate axis across the dataset.Given a set of trajectory points, $X = x_1, x_2 \cdot \cdots x_n$ and $Y = y_1, y_2 \cdot \cdots y_n$ normalization is performed independently as follows.

$$x_i^{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}, \quad y_i^{norm} = \frac{y_i - y_{min}}{y_{max} - y_{min}} \quad (8)$$

This transformation scales all coordinate values to the [0,1] range, which improves the model stability and convergence of the training phase. The temporal information

(timesteps) unaltered to maintain the correct sequence order. The scaling process carried out using the MinMaxScalar.

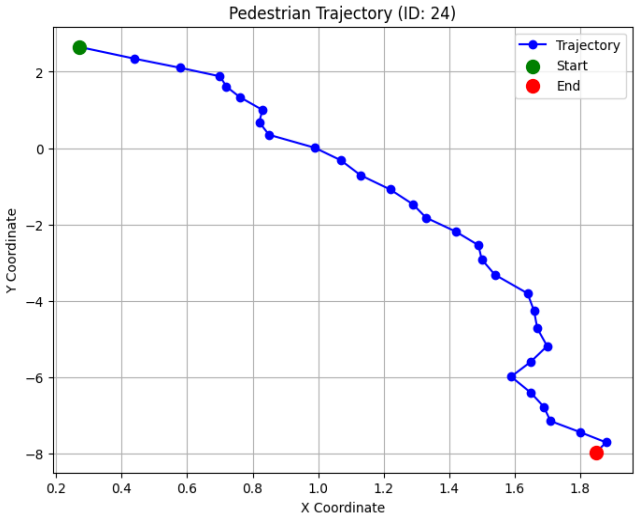

Fig 1: Visualization of actual trajectories of pedestrians.

3.3 Model Based on Convolutional Neural Networks

The sequential structure of pedestrian trajectories has been modelled through a convolutional neural network (CNN) architecture, rather than traditional LSTM-based methods where temporal dependencies are captured sequentially. This approach enables fully parallel and feed-forward processing. The model has been designed as a sequence-to-sequence framework, mapping a fixed-length sequence of observed positions into a fixed-length output sequence of predicted future coordinates. The input to the network consists of sequences of eight observed (x, y) coordinate pairs, reshaped into a (8, 2, 1) tensor. To identify meaningful patterns in the spatio-temporal features, three convolutional layers have been stacked. Each layer uses 2D kernels, with a ReLU activation function applied after each convolution. Padding has been used to preserve the temporal length of the sequences throughout the convolutional blocks. After the extraction of features, the resulting feature maps are flattened and then fed through fully connected layers with ReLU activation function. To enhance generalization and prevent over fitting, dropout regularization is applied following each fully connected layer. The final output layer consists of a dense layer with a linear activation function, projecting the features into a set of predicted future positions. The extracted features obtained from the final convolutional layer are combined and then subsequently fed through a fully connected layer, enabling the simultaneous prediction of all position coordinates $(x^{t+1}, y^{t+1}), (x^{t+2}, y^{t+2}), \dots, \left(x^{t+t^{pred}}, y^{t+t^{pred}}\right)$.

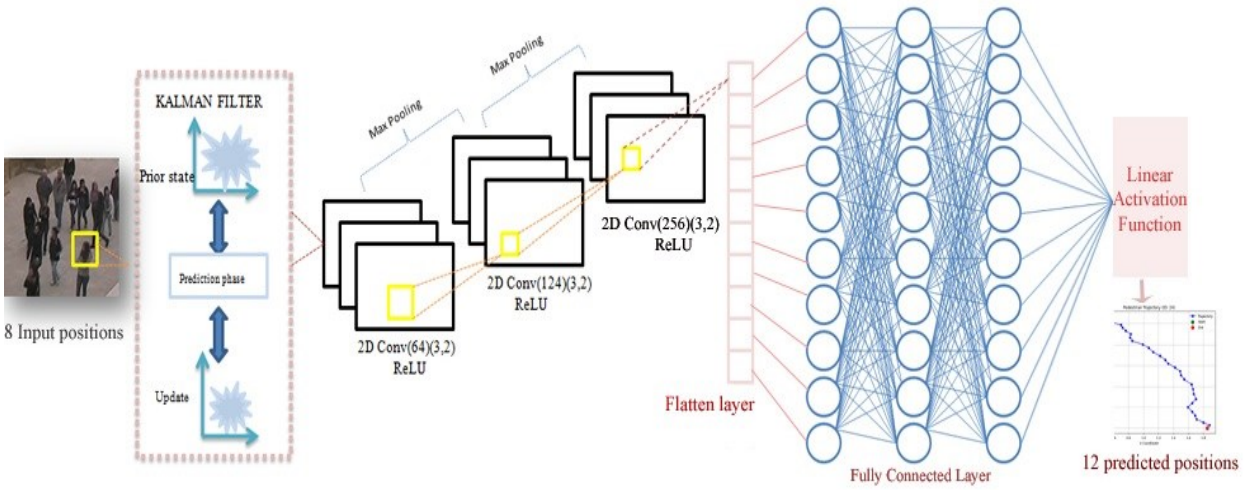

Fig 2: Kalman Filter-Augmented 2D CNN Architecture for Pedestrian Trajectory Prediction

4 Experiments

Experimental evaluations have been carried out on publicly available pedestrian trajectory datasets.In these experiments, the model leverages eight historical observations, corresponding to a temporal window of 3.2 seconds,to predict the subsequent twelve future positions, covering a time horizon of 4.8 seconds.

4.1 Sources of Data

The publicly available ETH (Pellegrini, S.et.al, 2009) and UCY (Lerner, A., 2007) datasets are utilized as widely recognized benchmarks for pedestrian trajectory prediction. Both datasets includes annotated trajectories of real-world pedestrians engaging in various social scenarios. These datasets contains nontrivial movements such as pedestrian collisions, collision avoidance behaviour, and group movement. The datasets consist of five unique scenes in total: Zara1, Zara2, and Univ (name derived from the University of Cyprus), and ETH and Hotel (name taken by the ETH Zurich University). Top-view images and 2D positions of each individual relative to world coordinates are provided for each scene. One image is used per scene as the cameras remain fixed. Each scene takes place in a relatively open outdoor environment, minimizing the impact of physical barriers. **Figure 3** shows a scene from the ETH and UCY datasets.

Fig 3: A Sample top-view frames from the ETH - UCY datasets showing pedestrian interactions and social behaviours in outdoor environments.

4.2 Performance Metrics

In the existing literature, it is a standard practice to set the observation length $\boldsymbol{T_{obs}}$ = 8 and the prediction length $\boldsymbol{T_{pred}}$ = 12. This configuration has been consistently employed in several influential works (Alahi, A., 2016, Gupta, A., 2018) among many others. To ensure fair comparison with prior methods, the same settings are used across all experiments presented in this study.

Established performance metrics are used to conduct the evaluation of predicted pedestrian trajectories. The primary and most widely utilized metric is the Average Displacement Error (ADE). ADE measures the mean Euclidean distance between the predicted trajectory points and the corresponding ground truth points, calculated from $\boldsymbol{T_{obs}}$ to $\boldsymbol{T_{pred}}$, and averaged over all pedestrians in the dataset.

The mathematical formulation of ADE is given as:

$$\mathbf{ADE} = \frac{1}{n(T_{\text{Pred}} - T_{\text{obs}})} \sum_{i=1}^{n} \sum_{t=T_{\text{obs}}+1}^{T_{\text{pred}}} |\widehat{\mathbf{P}}_t^i - \mathbf{P}_t^i| \quad (9)$$

$\boldsymbol{n}$**:** Number of trajectories, $\boldsymbol{T_{\text{obs}}}$**:** The last observed time step, $\boldsymbol{T_{\text{pred}}}$: The final predicted time step, $\widehat{\mathbf{P}}_t^i$**:** Predicted position of trajectory i at time t, $\mathbf{P}_t^i$: Ground truth position of trajectory i at time t,$\left|\widehat{\mathbf{P}}_t^i - \mathbf{P}_t^i\right|$ Euclidean distance between predicted and true positions.

The second performance metric considered is the **Final Displacement Error (FDE)**. FDE calculates the Euclidean distance between the predicted and actual final positions of a pedestrian at the time step t = $\boldsymbol{T_{pred}}$. For each pedestrian, this distance is computed and subsequently averaged over the entire dataset. The FDE is used to assess how accurately the final destinations of pedestrians are predicted by the model. The mathematical expression for FDE is given by:

$$\mathbf{FDE} = \frac{1}{n} \sum_{i=1}^{n} |\widehat{\mathbf{P}}_{T_{Pred}}^{i} - \mathbf{P}_{T_{pred}}^{i}| \quad (10)$$

5 Result and Discussions

The performance of the proposed Convolutional Neural Network (CNN)-based model was evaluated across the benchmark ETH and UCY datasets using the standard metrics of Average Displacement Error (ADE) and Final Displacement Error (FDE). These metrics were computed over a 12 frame prediction horizon based on 8-frame historical observations.

Table 1 presents a comparative summary of the ADE and FDE values obtained across different dataset scenes. The proposed model demonstrates superior performance compared to baseline approaches, confirming the efficacy of the spatial-temporal feature extraction and efficient model architecture.

METRICS	(Nikhil, N.,2018)		(Zamboni, S.,2022)		Kalman + 2DCNN	
	ADE	FDE	ADE	FDE	ADE	FDE
ETH	1.04	2.07	0.559	1.114	0.24	0.28
HOTEL	0.59	1.17	0.240	0.464	0.42	0.52
UNIV	0.57	1.21	0.581	1.225	0.12	0.17
ZARA1	0.43	0.90	0.456	0.993	0.11	0.16
ZARA2	0.34	0.75	0.347	0.751	0.11	0.15

Table 1: ADE and FDE Performance of Proposed CNN Model Across Datasets

High predictive accuracy was demonstrated in structured and semi-structured environments such as ZARA1, ZARA2, and UNIV. In these scenes, significantly lower error values were achievedADE. These results indicate that consistent spatial and motion patterns were effectively captured by the CNN architecture. The model's ability to learn and generalize trajectory behavior in predictable pedestrian interactions was thus validated.

In contrast, scenes characterized by dynamic and unpredictable pedestrian behavior, such as the HOTEL dataset, posed greater challenges. Higher error values were recorded in this setting. These deviations were attributed to the presence of abrupt trajectory changes and non-linear pedestrian motion, which reduced prediction reliability. Despite these challenges, the model-maintained performance within a competitive margin when compared to state-of-the-art approaches.

One of the primary strengths of the model lies in its lightweight design, which ensures real-time inference without compromising accuracy. The integration of Kalman filtering and min-max normalization during data preprocessing contributed to improved noise reduction and stability during model training, thereby enhancing generalization performance. The results of the proposed CNN-based model have been evaluated using the ETH Hotel dataset. The model's performance has been assessed using the widely accepted metrics ADE (Average Displacement Error) and FDE (Final Displacement Error). It has been observed that the model has achieved an ADE of 0.10 and an FDE of 1.28 on the dataset, indicating high prediction accuracy. CNN-based trajectory prediction model has demonstrated strong performance across multiple benchmark scenes, excelling particularly in structured environments while remaining resilient in complex ones. Its design has proven to be both effective and efficient, validating its suitability for real-time pedestrian forecasting tasks in dynamic environments. Overall, the CNN-based trajectory prediction model has been demonstrated to perform robustly across diverse pedestrian scenarios, offering a reliable and efficient solution for real-time trajectory forecasting tasks in intelligent transportation and surveillance systems.

Figure 4 provides a comprehensive visualization of the pedestrian trajectory prediction. The observed positions (blue), real future positions (green), and predicted future positions (red) are plotted together, showcasing the high spatial alignment between the predicted and actual future trajectories. The close proximity between the green and red lines indicates the model's strong forecasting accuracy across the 12 prediction steps.

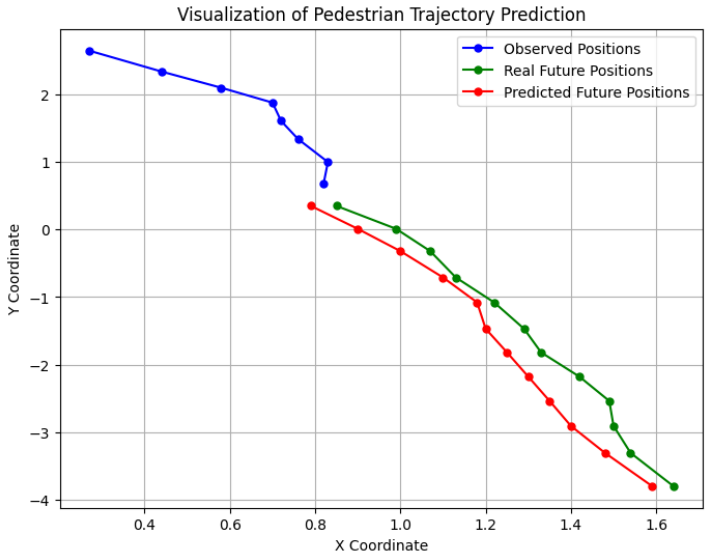

Fig 4: Visualization of Observed, Ground Truth, and Predicted Pedestrian Trajectories

Conclusion

In this research, we proposed an optimized Convolutional Neural Network (CNN) framework for pedestrian trajectory forecasting, spatio-temporal sequence modeling was incorporated to accurately predict future positions in dynamic environments. Kalman filter-based preprocessing was applied to improve the model, enabling noisy trajectory data to be smoothed and stability in both short-term and long-term predictions to be increased. The performance of the model was evaluated on benchmark datasets (ETH and UCY), achieving competitive results in terms of Average Displacement Error (ADE) and Final Displacement Error (FDE) compared to existing baseline methods. Our results demonstrate the CNN model's capability to capture temporal dependencies and spatial context, even without explicit graph or recurrent structures. By balancing model complexity with predictive accuracy, the proposed approach was found to be appropriate for real-time applications in autonomous navigation and crowd analytics.

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 961-971).

Antonini, G., Bierlaire, M., & Weber, M. (2006). Discrete choice models of pedestrian walking behavior. Transportation Research Part B: Methodological, 40(8), 667-687.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Becker, S., Hug, R., Hübner, W., & Arens, M. (2018). An evaluation of trajectory prediction approaches and notes on the trajnet benchmark. arXiv preprint arXiv:1805.07663.

Berg, R. V. D., Kipf, T. N., & Welling, M. (2017). Graph convolutional matrix completion. arXiv preprint arXiv:1706.02263.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social gan: Socially acceptable trajectories with generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2255-2264).

Helbing, D., & Molnar, P. (1995). Social force model for pedestrian dynamics. Physical review E, 51(5), 4282.

Kothari, P., Kreiss, S., & Alahi, A. (2021). Human trajectory forecasting in crowds: A deep learning perspective. IEEE Transactions on Intelligent Transportation Systems, 23(7), 7386-7400.

Lerner, A., Chrysanthou, Y., &Lischinski, D. (2007, September). Crowds by example. In Computer graphics forum (Vol. 26, No. 3, pp. 655-664). Oxford, UK: Blackwell Publishing Ltd.

Nikhil, N., & Tran Morris, B. (2018). Convolutional neural network for trajectory prediction. In Proceedings of the European conference on computer vision (ECCV) workshops (pp. 0-0).

Pellegrini, S., Ess, A., Schindler, K., & Van Gool, L. (2009, September). You'll never walk alone: Modeling social behavior for multi-target tracking. In 2009 IEEE 12th international conference on computer vision (pp. 261-268). IEEE.

Wang, G. (2009). Signal extraction from long-term ecological data using Bayesian and non-Bayesian state-space models. Ecological Informatics, 4(2), 69-75.

Welch, G., & Bishop, G. (1995). An introduction to the Kalman filter.

Yang, J., & Han, C. (2022, December). Pedestrian Trajectory Prediction Based on Improved Social Spatio-Temporal Graph Convolution Neural Network. In Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing (pp. 63-67).

Yamaguchi, K., Berg, A. C., Ortiz, L. E., & Berg, T. L. (2011, June). Who are you with and where are you going?. In CVPR 2011 (pp. 1345-1352). IEEE.

Zamboni, S., Kefato, Z. T., Girdzijauskas, S., Norén, C., & Dal Col, L. (2022). Pedestrian trajectory prediction with convolutional neural networks. Pattern Recognition, 121, 108252, DOI: https://doi.org/10.1016/j.patcog.2021.108252.