Regular paper | Journal of the KIECS. pp. 1319-1326, vol. 19, no. 6, Dec. 31. 2024, t. 128, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.6.1319

다중 헤드 어텐션과 결합한 convLSTM을 활용한 군중 밀도 예측

파텔 더너스리 수레쉬* · 타히라** · 실라** · 박장식***

Crowd Density Prediction Using convLSTM Combined with Multi-Head Attention

Patel Dhanshree Suresh* · Nusrat Jahan Tahira** · Sheilla Wesonga** · Jang-Sik Park***

요약

정확한 군중 밀도 예측은 군중 관리, 공공 안전, 도시 계획과 같은 응용 분야에 매우 중요한다. 기존 방법은 비디오 시퀀스의 시간적 종속성과 동적 변화를 포착하는 데 어려움을 겪는 경우가 많아 예측의 신뢰성이 떨어진다. 밀집되고 역동적인 군중은 폐색, 모션 블러 및 다양한 관점을 포함하여 정확한 밀도 추정을 복잡하게 만드는 추가적인 문제를 제시한다. 이러한 문제를 해결하기 위해 우리는 convLSTM(convolutional Long Short-Term Memory) 네트워크를 multi-head attention 메커니즘과 통합하는 새로운 접근 방식을 제안한다. 이 결합된 모델은 복잡한 시나리오에서도 시간적 패턴과 공간적 관계를 효과적으로 캡처하여 예측 정확도를 향상시킨다. 광범위한 실험을 통해 우리의 제안방법이 기존 기술보다 훨씬 뛰어나며 실제 응용 분야에서 정확한 군중 밀도 예측을 위한 강력한 솔루션을 제공한다는 것을 보여준다.

ABSTRACT

Accurate crowd density prediction is crucial for applications such as crowd management, public safety, and urban planning. Traditional methods often struggle with capturing temporal dependencies and dynamic changes in video sequences, leading to less reliable forecasts. Dense and dynamic crowds present additional challenges, including occlusions, motion blur, and varied perspectives, which complicate accurate density estimation. To address these issues, we propose a novel approach that integrates convolutional Long Short-Term Memory (convLSTM) networks with Multi-Head Attention mechanisms. This combined model improves prediction accuracy by effectively capturing temporal patterns and spatial relationships, even in complex scenarios. Extensive experiments show that our method significantly outperforms traditional techniques, offering a robust solution for precise crowd density prediction in real-world applications.

Keywords
Crowd Dynamics, Attention mechanisms, Density Prediction, convLSTM
군중 역학, 주의 메커니즘, 밀도 예측, convLSTM

* 주저자 : 경성대학교 석사(dhanashri.s.patel@gmail.com)
** 공저자 : 경성대학교 박사과정(tahira2@kyungsung.ac.kr)
** 공저자 : 경성대학교 박사과정(sheilla2@kyungsung.ac.kr)
*** 교신저자 : 경성대학교 전자공학과
· 접 수 일 : 2024. 09. 11
· 수정완료일 : 2024. 10. 27
· 게재확정일 : 2024. 12. 12
· Received : Sep. 11, 2024, Revised : Oct. 27, 2024, Accepted : Dec. 12, 2024
· Corresponding Author : Jang-Sik Park
Dept. of Electronic Engineering, Kyungsung University
Email : jsipark@ks.ac.kr

Ⅰ. INTRODUCTION

Crowd density prediction plays a crucial role in managing large events, ensuring public safety, and guiding urban planning efforts. The rising frequency of large-scale events, such as political rallies and concerts, underscores the critical need for robust crowd control measures [1]. Recent tragedies, such as the 2022 Korea halloween crowd crush, which resulted in numerous fatalities and injuries, highlight the severe consequences of inadequate crowd management [2]. Overcrowding in venues often leads to dangerous situations marked by fear and panic. Traditional methods for predicting crowd density struggle with dynamic environments due to issues like occlusions, motion blur, and varied perspectives. further complicate the estimation process, leading to less reliable forecasts [3]. Addressing these issues requires innovative solutions capable of modeling both temporal patterns and spatial relationships of crowds effectively.

Therefore, we propose a novel method for crowd density prediction from combining Convolutional Long Short-Term Memory (convLSTM) networks with Multi-Head Attention mechanisms [4][5]. Our approach leverages convLSTM's strength in capturing temporal patterns and multi-head attention's capability to focus on critical spatial features, enhancing prediction accuracy. By integrating these advanced techniques, our model aims to improve the precision of crowd density forecasts. Enhanced prediction capabilities can aid in preventing crowd-related incidents. Our proposed method represents a significant step forward in crowd density prediction, offering a more reliable and accurate solution for managing complex crowd dynamics.

Ⅱ. RELATED WORKS

In recent years, crowd analysis has become increasingly vital within the field of computer vision, particularly in video surveillance, public safety, and urban planning [6]. Accurate crowd behavior prediction helps mitigate risks such as mob violence, traffic congestion, riots, and stampedes [7]. To tackle the complexities of crowd dynamics various approaches have been developed.

2.1 Crowd Density Estimation Approaches

Traditional methods include crowd-oriented techniques, regression models, and density map-based approaches. Crowd-oriented techniques focus on detecting and counting individuals within a scene but struggle with accuracy in high-density situations due to occlusions and perspective distortion. Regression-based models [8]. predict crowd density but struggle with complex, non-linear crowd data.

2.2 Density Map Oriented Approaches

Density map-based methods generate visual representations of crowd distribution and are essential for analyzing traffic and crowd movement [9]. A common method is the Gaussian-based approach but has limitations in highly crowded areas, where it can be difficult to distinguish between closely packed individuals [10]. To overcome this, the Focal Inverse Distance Transform (FIDT) map enhances head localization by refining visibility around densely packed individuals, making it effective [11] in crowd monitoring and management.

2.3 Deep Learning Oriented Approaches

Deep learning models like Convolutional Neural Networks (CNNs) [12] and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks [13][14] have significantly improved crowd density estimation by capturing spatial and temporal dependencies in crowd data. However, traditional CNNs and RNNs often struggle with capturing complex

spatial-temporal interactions in crowded scenes. The CNN-LSTM combination [15], improves spatial and temporal modeling. However, challenges like increased computational complexity, sensitivity to motion blur, and the need for extensive annotated data persist.

III. PROPOSED CROWD DENSITY PREDICTION FRAMEWORK

The proposed crowd density prediction framework consists of two approaches, the series method and the parallel method.

3.1 General Methodology Framework

The methodology involves developing a deep-learning framework that combines convLSTM with multi-head attention to predict future crowd density. The process begins with generating crowd density maps using the FIDT method, which enhances the precision and reliability of traditional density map representations. These density maps serve as input to the deep learning models.

The convLSTM model integrates the strength of CNN and LSTM efficiently capturing both spatial and temporal dependencies. CNNs are effective at modeling short-term spatial features, while LSTMs exploits sequential information. So adding both in a single model captures the intricate dependencies in an efficient way compared to individual models.

To further improve data representation, the multi-head attention mechanism is applied. this technique enhances the model's ability to focus on different parts of the input sequence, capturing long-term dependencies by analyzing weight vectors across multiple time instants. By tapping into latent features, model achieves more accurate predictions. The hybrid model of convLSTM and multi-head attention addresses the challenges of crowd movement uncertainties by effectively capturing short-term and long-term dependencies, improves the accuracy of crowd density prediction.

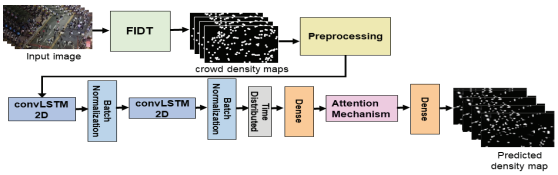

Fig. 1 Model architecture of proposed method (series).

3.2 The Series Method

In the series method as shown in Fig. 1, the serial addition of multi-head attention is applied after convLSTM processing to tap the models ability to focus on non-linearities and long-term dependencies. The meat data generated by convLSTM is refined through the attention mechanism, which identifies critical time steps and spatial regions relevant to crowd density prediction. this process improves the prediction accuracy by focusing on the weight vectors of long term sequences. The refined output features are then passed through dense layers, followed by a final prediction layer.

3.3 The Parallel Method

The Parallel Method extends the capabilities of the Series Method by processing the dataset into the convLSTM and multi-head attention layers simultaneously, as depicted in Fig. 2 capturing both short-term and long-term dependencies concurrently. The output from the convLSTM is combined with the output from the multi-head attention layer, enhancing the model's understanding of the input crowd data. This combined representation is then processed through dense layers, leading to an enriched representation and improved crowd density prediction.

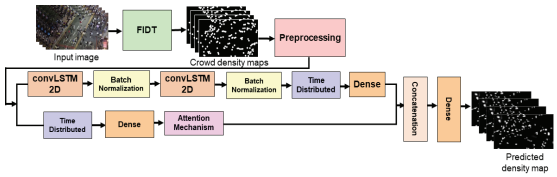

Fig. 2 Model architecture of proposed method (parallel).

Ⅳ. Experimental Results

For our experimental analysis, we utilized the Fudan-ShanghaiTech (FDST) dataset [16], which offers a comprehensive set of images and videos for crowd counting. The dataset includes 150,000 images from 100 videos across 13 diverse scenes. from this we used a total of 9000 images. 70% for training, 20% for validation and 10% for testing.

The experiments were conducted on a system equipped with NVIDIA Quadro p5000 GPU. The models were implemented using tensorflow 2.6 with CUDA 11.4 running on python 3.7. The Adam optimizer was used with learning rate of 0.001. To assess the performance of our deep learning model, we use two key metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE). MAE measures average prediction error and is less affected by outliers, making it suitable for non-normal residual errors. MSE, on the other hand, penalizes larger errors more heavily, emphasizing robustness in sequential predictions.

The performance comparison in Fig. 3 shows that the proposed models, both in parallel and series configurations, outperform baseline convLSTM and LSTM models. The parallel model achieved the lowest MAE (0.024) and MSE (0.020), demonstrating superior accuracy in crowd density prediction by effectively addressing non-linearity and sparsity. The series model also showed strong performance with an MAE of 0.025 and MSE of 0.023, though slightly higher than the parallel model. The series model is simpler and less prone to over fitting, making it potentially better for less sparse data.

The visual crowd density predictions in Fig. 4, compare proposed parallel and series models, CNN-LSTM, and convLSTM models against the ground truth. The proposed parallel model closely matches the gt, effectively capturing spatio-temporal dependencies.

The series model also performs well, showing similar density trends. In contrast, CNN-LSTM and convLSTM models show more deviation, with

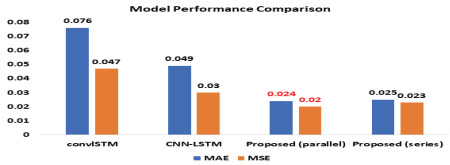

Fig. 3 Model performance of crowd density prediction.

convLSTM struggling to capture spatiotemporal dependencies, leading to underestimations in certain regions. The CNN-LSTM, while more generalized, still falls short compared to the proposed models.

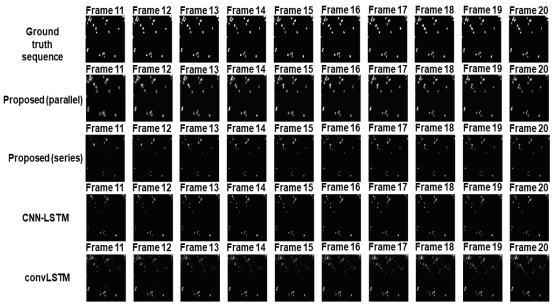

Fig. 4 Comparison results of crowd density prediction.

The detail comparison of the proposed models with other DL architectures is shown in the table 1. based on the criteria, series model captures detailed dependencies but is complex and inflexible, while parallel convLSTM provides better flexibility and broader feature representation but increases

complexity. convLSTM efficiently integrates spatiotemporal data with moderate complexity, while CNN-LSTM, being simpler, separates spatial and temporal tasks, making it more efficient but less powerful in capturing temporal features. series model is computationally expensive, whereas CNN-LSTM is more resource-efficient.

Table 1. Comparison of proposed models with existing architectures.

Criteria	Series model	Parallel model	convLSTM	CNN-LSTM
Design	Sequential connection	Parallel connection	Stacked convLSTM layers	Separate CNN and LSTM
Feature Integration	Hierarchical spatio temporal features	Simultaneous spatial and temporal features	Combines temporal and spatial data	Combines CNN spatial, LSTM temporal
Complexity of architecture	High, due to deep layering	Higher due to two parallel paths	Moderately complex	Simple
Flexibility and design	Less flexible	Flexible	Rigid	Flexible
Complexity	High	Moderate	Moderate	Least
Performance	Detailed hierarchical feature capture	Broader feature representation	Lower compared to attention	Less powerful than convLSTM
Efficiency	Computationally expensive	More efficient with reduced time	Slower	More Efficient

V. Conclusion

This work compares CNN-LSTM, convLSTM, and the proposed spatio-temporal multi-head attention-based series and parallel models for crowd density prediction. The proposed models, leveraging advanced attention mechanisms, excel in capturing both long-term and short-term dependencies, resulting in significantly improved accuracy. The parallel model achieved 13.04% better MSE than the series model, 57.44% better than convLSTM, and 59.18% better than CNN-LSTM. Future work will focus on optimizing these models for real-time applications by reducing computational overhead and enhancing generalizability across diverse datasets.

Acknowledgment

This research was supported by Development of Social Complex Disaster Response Technology through the Korea Planning & Evaluation Institute of Industrial Technology funded by Ministry of the Interior and Safety in 2023. (Project Name: Development of risk analysis and evaluation technology for high reliability stampede accidents using CCTV and Drone imaging, Project Number: 20024403).

References

[1] S. Yoon and J. Song, "A study on the interaction factors in implementing virtual reality to solve safety problems in public toilets," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 17, no. 06, Dec. 2022, pp. 1167-1174. doi: 10.13067/JKIECS.2022.17.6.1167.

[2] Seoul halloween crowd crush, Wikipedia, jul. 2023, URL: https://en.wikipedia.org/wiki/Seoul_Halloween_crowd_crush.

[3] A. Patwal, M. Diwakar, V. Tripathi, and P. Singh, "Crowd counting analysis using deep learning: a critical review," *Proc. Computer Science,* vol. 218, 2023, pp. 2448-2458. doi: 10.1016/j.procs.2023.01.220.

[4] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. K. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems 28, NIPS,* Dec. 2015. doi: 10.5555/2969239.2969329.

[5] A. Vaswani, N. Shazeer, Niki Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L.

Kaiser, I. Polosukhin “Attention Is All You Need,” Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems,* Dec. 2017.
doi: 10.48550/arXiv.1706.03762.

[6] J. C. Silveira Jacques Junior, S. R. Musse, and C. R. Jung, “Crowd analysis using computer vision techniques,” *IEEE Signal Processing Magazine,* vol. 27, no. 05, Sept. 2010, pp. 66-77.
doi: 10.1109/MSP.2010.937394.

[7] S. Park, I. Young, S. Won, “Conv-LSTM-based Range Modeling and Traffic Congestion Prediction Algorithm for the Efficirnt Transportation System,” *J. of The Korea Institute of Electronic Communication Sciences,* Apr. 2023, pp. 321-328.

[8] A. B. Chan and N. Vasconcelos, “Counting people with low-level features and bayesian regression,” *IEEE Transactions on Image Processing,* vol. 21, no. 4, Apr. 2012. pp. 2160-2177.
doi: 10.1109/TIP.2011.2172800.

[9] J. Kim, S. Lee, J. Ko, and J. Park, “Traffic control algorithm for periodic traffics in WSN,” *J. of The Korea Institute of Electronic Communication Sciences,* vol. 05, no. 01, 2010, pp. 44-50.

[10] J. Wan and A. Chan. “Adaptive Density Map Generation for Crowd Counting,” *Proceeding IEEE/CVF International Conference,* 27, Oct. 2019, pp. 1130-1139.
doi: 10.1109/ICCV.2019.00122.

[11] D. Liang, W. Xu, Y. Zhu, and Y. Zhou, “Focal inverse distance transform maps for crowd localization,” *IEEE transactions on multimedia,* vol. 25, Sept. 2023, pp. 6040-6052. doi: 10.1109/TMM.2022.3203870.

[12] J. Choi and H. Choi, “Prediction of wind power generation using deep learning,” *J. of The Korea Institute of Electronic Communication Sciences,* vol. 16, no. 02 May 2021, pp. 329-338.
doi: 10.13067/JKIECS.2021.16.2.329.

[13] H. Jang, Y. Moon, S. Im, “Very Short- and Long-Term Prediction Method for Solar Power,” *J. of The Korea Institute of Electronic Communication Sciences,* vol. 18, no. 06 Dec 2023, pp. 1143-1150.

[14] S. Hochreiter and J. Schmidhuber, “Long short-term Memory,” *Neural Computation,* vol. 9, no. 8, Nov. 1997, pp. 1735 – 1780.
doi: 10.1162/neco.1997.9.8.1735.

[15] K. Wu, J. Wu, L. Feng, B. Liang, S. Yang, R. Zaho, “An attention based CNN LSTM BiLSTM model for short term electric load forecasting in integrated energy system,” *International Transactions on Electrical Energy Systems,* vol. 31, no. 1, Sept. 2020, pp. e12637.
doi: 10.1002/2050-7038.12637.

[16] Y. Fang, B. Zhan, W. Cai, S. Gao, and B. Hu, “Locality-constrained spatial transformer network for video crowd counting,” *IEEE international conference on multimedia and expo (ICME),* July 2019*, pp. 814-819*.
doi: 10.48550/arXiv.1907.07911.

AUTHORS

박장식(Jang-Sik Park)

1992년 부산대학교 전자공학과 졸업(공학사)
1994년 부산대학교 대학원 전자공학과 졸업(공학석사)
1999년 부산대학교 대학원 전자공학과 졸업(공학박사)
1997년 ~2011년 동의과학대학 전자과 교수
2011년 ~현재 경성대학교 전자공학과 교수
※ 관심분야 : 신호처리, 기계학습, 컴퓨터비전, 심층학습, 임베디드시스템

파텔 더너스리 수레쉬(Patel Dhanshree Suresh)

2015년 북마하슈트라대학교 전자공학과 졸업(공학사)
2024년 경성대학교 대학원 전자공학과 졸업(공학석사)
※ 관심분야 : AI, 데이터 분석, 영상 처리

타히라(Nusrat Jahan Tahira)

2021년 R.P.사하대학교 컴퓨터공학과 졸업(공학사)
2023년 경성대학교 대학원 전자공학과 졸업(공학석사)
2023년 ~현재 경성대학교 대학원 전자공학과 (공학박사과정)
※ 관심분야 : AI, 데이터 분석, 영상 처리

실라(Sheilla Wesonga)

2017년 음바라라대학교 컴퓨터공학과 졸업(공학사)
2021년 경성대학교 대학원 전자공학과 졸업(공학석사)
2021년 ~현재 경성대학교 대학원 전자공학과 (공학박사과정)
※ 관심분야 : AI, 데이터 분석, 영상 처리