International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n1.15
Journal homepage: https://wepub.org/index.php/IJCSIT/index

A Review of Multi-source Heterogeneous Twin Data Processing Methods in Traffic Scenes

Yu Yan

School of Information Engineering, Henan University of Science and Technology, Luoyang 471000, China

ABSTRACT

As the complexity of intelligent public transportation systems continues to increase and the degree of interconnection of equipment at all levels gradually deepens, building a digital twin system will generate a large amount of multi-source heterogeneous data. Effective processing and in-depth mining of data can provide real-time and intelligent decision-making optimization for the construction of digital twin systems, thereby efficiently handling complex emergencies. This paper conducts a systematic review on the processing methods and technologies of multi-source heterogeneous twin data in traffic dynamic scenarios. First, the content and classification of multi-source heterogeneous twin data in the operation process of intelligent bus systems are clarified; secondly, the multi-source The data processing methods and technologies applied in each stage of data collection, data storage, data fusion and data analysis in heterogeneous twin data processing are analyzed, and the advantages, disadvantages and applications of various methods and technologies are analyzed; finally, corresponding to the multiple This paper summarizes the methods and technologies for processing multi-source heterogeneous twin data, and points out the challenges and development trends faced by the current multi-source heterogeneous twin data processing methods and technologies.

KEYWORDS

Twin Data; Data Collection; Urban Transportation; Data Fusion; Data Analysis

1. INTRODUCTION

With the development and maturity of information communication technology (ICT), digital intelligent society construction solutions such as the Internet and the Internet of Intelligence are gradually being launched around the world, such as the "Industrial Internet" in the United States, "Industry 4.0" in Germany, "Made in 2025". However, as the complexity of intelligent public transportation systems continues to increase and the degree of interconnection of equipment at all levels gradually deepens, the construction of intelligent networked transportation systems also faces many challenges [1]: 1) The problem of fusion and mining of massive data from Internet of Vehicles equipment; 2) There is the problem of mutual separation between digital systems and physical systems; 3) The problem of coordination of multi-source heterogeneous resources. In response to the above problems, there is an urgent need to develop a new generation of ICT and intelligent technologies to support future construction and development.

2. CURRENT STATUS OF DIGITAL TWIN RESEARCH AND DEVELOPMENT

2.1. The Origin of Digital Twin Development

The concept of digital twins can be traced back to the "mirror space model" [1] proposed by Professor Grieves in the Product Lifecycle Management (PLM) course at the University of Michigan in 2003. It is defined as including physical products, A three-dimensional model of a virtual product and the connection between the two. As shown in Figure 1. Due to the limitations of technology and cognitive level at that time, this concept was not taken seriously [2], and no relevant results were published in the following ten years. It was not until 2010 that NASA introduced the concept of digital twins for the first time in its space technology roadmap [3], in order to use digital twins to achieve comprehensive diagnosis and maintenance of flight systems. In 2011, the U.S. Air Force Laboratory clearly proposed a digital twin paradigm for future aircraft, pointing out that a complete virtual mapping of the aircraft should be constructed based on the aircraft's high-fidelity simulation model, historical data, and real-time sensor data to achieve an understanding of the aircraft's health status and remaining life. and prediction of task accessibility [4]. Since then, the concept of digital twins has begun to attract widespread attention, and relevant research institutions have begun research on related key technologies [5].

There are few research results and there is also a lack of corresponding review and summary, which is not conducive to the subsequent application and development of digital twins in the field of smart transportation.

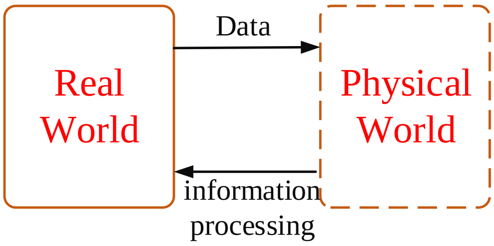

Figure 1. Mirror space model

2.2. Current Status of Digital Twin Research and Development

2.2.1. Smart Highway

Problems existing on current highways include entrances and exits, non-stop toll collection, non-stop overflow monitoring, how to improve the driving channels and road conditions, and how to solve rain and fog environments, road diseases, etc. Therefore, the all-weather access system came into being, which is one of the key points of construction based on digital twin technology. Some companies use digital twin technology to build all-weather access systems. In addition, the road information traveled by vehicles will be simultaneously uploaded to the digital twin visualization platform to help traffic managers make early warning judgments on the road environment [6].

2.2.2. Vehicle-road Rollaboration

The platform includes functions such as data fusion and docking, infrastructure cloud platform, big data center, vehicle-road collaborative business supervision and management, etc., to create a standardized, systematic and intelligent intelligent network business application display center and supervision and management operation center; proactive and automated prediction and identify risks to minimize operational safety hazards. It ultimately solves traffic problems such as waste of traffic resources, rigid signaling system functions, and unpredictable and rapid response to traffic events [7].

2.2.3. Autopilot

Today, autonomous driving digital twins can improve the accuracy of intelligent driving tests. By building a 1:1 digital twin scene of the real world, the operating rules of the physical world are restored, meeting the training needs of artificial intelligence algorithms in intelligent driving scenarios, and greatly improving training efficiency and safety. For example, by collecting laser point cloud data, building high-precision maps, constructing autonomous driving digital twin models, completing centimeter-level road restoration, and at the same time structurally processing road data and turning it into machine-understandable information, by generating a large number of actual traffic accident cases , train the ability of autonomous driving algorithms to handle unexpected scenarios, and ultimately achieve algorithm testing and detection verification of high-precision autonomous driving [8].

3. MULTI-SOURCE HETEROGENEOUS TWIN DATA IN URBAN TRAFFIC SCENES

Multi-source heterogeneous data refers to data from different data sources with different structures, formats, semantics and attributes, and there are differences and incompatibilities in data types, representation and processing methods. These data sources can be multiple different systems, databases, files, sensors, etc. Mining and analysis of multi-source heterogeneous data is a challenging task, which requires overcoming data heterogeneity and integration to achieve global consistency and model accuracy to meet the needs of various application scenarios [9].

Multi-source heterogeneous data in traffic scenes refers to data from different devices, different sources, and different spatio-temporal environments in the traffic system, including but not limited to videos, images, text, sensors, GPS trajectories, traffic flow, road conditions, etc. type of data. These data have the characteristics of heterogeneous data formats, data accuracy, data dimensions, etc., and there are rich connections and correlations between the data. The complex structures and patterns in traffic scenes can be revealed through data fusion and analysis [10]. But at the same time, the processing and analysis of multi-source heterogeneous data also faces various challenges and difficulties such as inconsistent data quality, data source integration, feature extraction and modeling.

4. MULTI-SOURCE HETEROGENEOUS TWIN DATA PROCESSING IN URBAN TRAFFIC SCENES

Data preprocessing mainly implements database storage of data, preprocessing such as data cleaning, conversion, and dimensionality reduction, and the construction of massive related databases to provide cleaned data for the data fusion stage. The data fusion stage aims to fuse twin data from different sources. Provide preprocessed data sources for the data analysis phase. Data analysis mainly uses technologies such as correlation analysis, classification and clustering, and deep learning to realize the value of data. The general flow of multi-source heterogeneous twin data processing is shown in Fig. 2 [11].

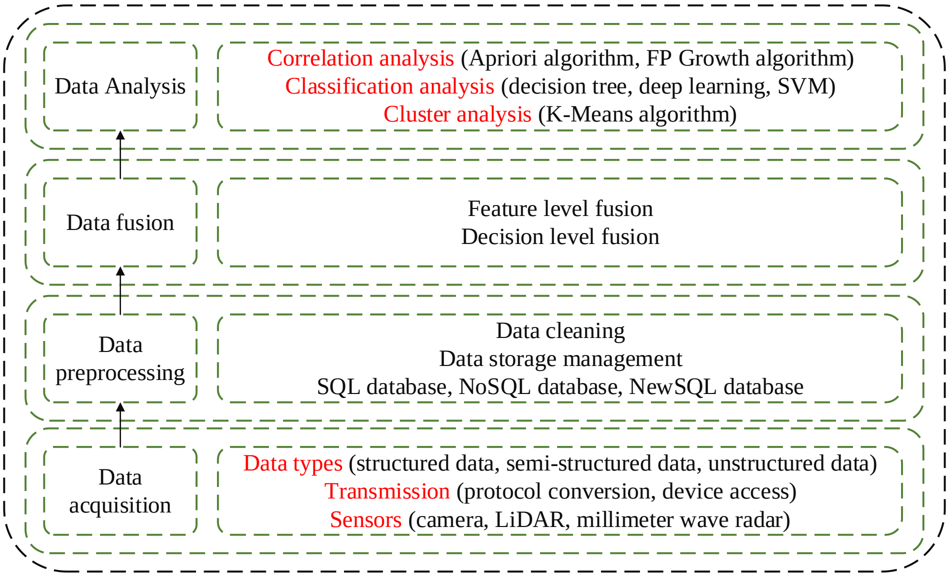

Figure 2. Multi-source heterogeneous twin data processing flow chart

4.1. Data Collection

Data collection is the basis for multi-source heterogeneous twin data processing. Only by accurately and real-time collecting a large amount of raw data in complex traffic scenarios and transmitting it to the data storage management platform can we make predictions about congestion, accident prediction, and public transportation. planning for real-time feedback and intelligent decision-making [12].

According to the different types of complex data in traffic scenes, the data in traffic scenes can be divided into structured data and unstructured data. High-precision maps can be used to obtain structured data [13], which can perform environmental sensing, positioning, and Path planning and vehicle control, high-precision maps have higher requirements for real-time data. The ranging accuracy of lidar is very high, which can basically reach plus or minus one or two centimeters, or even to the millimeter level, and the resolution is also very high. Mechanical lidar can rotate 360 degrees and has higher angular resolution than other radars. However, the current mechanical rotating lidar has a relatively high cost and is easily affected by sunlight, rain, fog and mutual interference. Millimeter wave radar uses electromagnetic waves with a wavelength of 1~10mm and a frequency of 30GHz~300GHz to calculate the distance by measuring the time difference of the echoes. It has the advantages of all-weather and long detection range. Currently, the mainstream vehicle-mounted millimeter wave radar frequency bands on the market are 24GHz (for short- and medium-range radar, 15~30m) and 77GHz (for long-range radar, 100~200m) [14]. The detailed comparison of the advantages and disadvantages of lidar, millimeter wave radar and cameras is shown in Table 1 below.

4.2. Data Preprocessing

In traffic scenarios, numerous multi-source heterogeneous twin data are involved, such as vehicle flow, traffic signals, sensor data, etc. These data come from various sources and types. There may be interference and noise between some data, which affects the data quality and signal-to-noise ratio. Therefore, multi-source heterogeneous twin data preprocessing is required. Multi-source heterogeneous twin data preprocessing is to clean, transform, and standardize multi-source heterogeneous twin data to obtain more accurate and reliable data for subsequent analysis, modeling, and decision support.

4.2.1. Data Cleaning

Different data cleaning methods can be given for different problem data collected in traffic scenes. Since multi-source heterogeneous twin data in traffic scenes often come from multiple data sources, each data source usually has a different database system, interface service, etc., so the data has the characteristics of diverse structure types and inconsistent expression forms, which leads to the collection. For real data, in most cases it is necessary to fill in manually, and in some cases missing data can be processed through statistical learning methods. Cao Lin [15] proposed a missing value cleaning framework for regression interpolation for data sets with clustering characteristics. For erroneous data, statistical analysis methods are first used to identify possible erroneous values, and then the erroneous data can be cleared to achieve the purpose of data cleaning. For inconsistent data, potential data errors can be detected based on the consistency between related data and repaired to complete the cleaning of multi-data source data [16].

Shawn R. jeffery et al. established a scalable sensor stream processing and cleaning system. This system is a data cleaning system that has appeared in the early stages of the development of the Internet of Vehicles and is currently widely used. The system contains multiple programmable streamlined cleaning modules, each of which is connected to one another and is also called a system pipeline. The ESP system mainly contains 5 analysis modules. After the sensor collects data, it is analyzed by 5 modules and finally collected upward. During the data transmission process, each module will independently clean the data. The type and type of data cleaned by each module are the purposes are different and independent of each other, so during the actual cleaning process, the system cleaning process can be adjusted according to the specific situation [17].

Saul Gill et al. collected 39 kinds of data in categories based on the relevant data of location and environmental sensors installed on the top of city buses, established a complete distributed data cleaning system, and also simulated linear models, polynomial models and generalized processing. Multiple regression models such as sexual models were also used to evaluate the data cleaning results of the system [18]. Abdulhakim et al. established a FAHES system that can detect data and fill in missing values. It mainly detects missing data through the attribute value distribution density, and then fills in the missing data through the data model. Using data integration technology to bring together various types of data to form a large database is a trend in data processing, and repeating data in the database Detecting and cleaning data inconsistencies and removing dirty data helps to improve the quality of data in the database analysis process [19].

4.2.2. Data Storage Management

Oracle developed by Oracle Corporation of the United States is an efficient and high-throughput relational database system. It is used in the steel [20], coal [21], and automobile manufacturing industries with large amounts of data and high requirements for stable system performance. widely. DB2 developed by IBM in the United States has the characteristics of good scalability, good query performance and good backward compatibility. It is suitable for the storage and management of massive data. It is widely used in governments, banks, etc., and is also used in Baosteel[22], Benxi Iron and Steel and other steel companies. Enterprises also have applications. The massive multi-source heterogeneous twin data collected in traffic scenes includes structured, semi-structured and unstructured data. Due to the inherent limitations of traditional relational databases for structured data in terms of scalability, fault tolerance, and scalability [23], it is difficult to meet the requirements for storage and management of massive multi-source heterogeneous data when used alone. Therefore, NoSQL Database has become a hot spot in current research and application.

4.3. Data Fusion

In traffic scenarios, since there are many sources and types of data involved, these data need to be integrated into a unified data model according to certain rules to form a multi-source heterogeneous twin data fusion method. Data fusion refers to using appropriate algorithms and technologies to fuse data from different sources, different forms, and different accuracies together to form a data model that is consistent with the actual situation to support better application scenarios and data analysis.

4.3.1. Feature Level Fusion

In data fusion, fusion based on feature level first preprocesses the original data, then extracts features from the original data, clusters the extracted feature information, synthesizes multi-source data, generates feature vectors, and finally performs feature vector fusion. Liu K, Cui M Y and others used coil detection data and floating car data as data sources and used the Bayesian method for data fusion [24]. Alfredo Nantes, Dong Ngoduy and others used data from multiple sources, especially data from loop detectors and partial observation data from Bluetooth and GPS devices, to fuse heterogeneous data sources through Bayesian estimation methods to achieve flexible and Robust traffic state estimation [25].

The former achieves almost the same improvement in detection accuracy compared to raw data-level fusion solutions, which provide the ability to dynamically adjust the size of feature maps to be transferred [26]. Wei Sun et al. [27] proposed a two-level fusion model for fatigue driving recognition based on multi-source information. They first determined the three most effective contextual features, namely continuous driving time, sleep duration and current time, to facilitate real-time (online) to identify fatigue status. In the decision-level fusion based on D-SET, by modifying the BPA and combining the fatigue status identified in the previous time step, the evidence conflicts between multiple pieces of evidence during evidence combination are resolved and the reliability of the identification model is improved.

4.3.2. Decision-level Integration

This method of decision-level fusion first preprocesses the original data, and then extracts features from the original data; clusters multi-source data according to the feature information to generate feature vectors; uses pattern recognition methods to process the feature vectors to obtain the data cut set about Description of the goal; classify the data according to the unified goal, and then perform data fusion. El Faouzi and Lefevre proposed a classifier fusion method based on D-S evidence theory [28]. Zeng Yiliang, Lan Jinhui and others proposed a spatial matching fusion model to estimate the average travel speed of the road based on the spatial complementarity of data based on rover detectors and fixed detectors. The fusion decision-making model of rough sets and clouds was used to fuse traffic data such as flow, speed, and occupancy detected between Beijing Deshengmen Bridge and Gulou Bridge to evaluate road traffic conditions [29].

It can be seen from the above research that the research on multi-source heterogeneous twin data processing methods has never stopped. Whether it is basic theory, application fields, academic research, etc., they are the main research directions of current data processing methods [30]. And compared with the research on multi-source heterogeneous twin data processing methods in complex traffic scenarios, it is still a hot topic. Through the fusion analysis of data from multiple sensors, the valuable information in each sensor can be effectively used, and thus the smart traffic digitalization can be improved. The global decision-making and judgment capabilities of twins [31], therefore, it is necessary to conduct research on multi-source heterogeneous twin data processing methods in complex traffic scenarios.

4.4. Data Analysis

4.4.1. Correlation Analysis

The Apriori algorithm first determines frequent item sets by traversing the database, then prunes them based on the support threshold, and finally calculates credibility based on the support to determine the association rules. It is a widely used association rule mining algorithm. In response to the high probability of emergencies and frequent accidents in the scene, Fan Hong [32] proposed a dynamic traffic alarm processing method based on the data mining Apriori algorithm, which reduced the number of repeated alarms and improved the processing of alarm events. efficiency. However, this algorithm still has the problem of frequently traversing the database to generate a large number of candidate sets. In response to this problem, Zhou Kai et al. [33] proposed an improved Apriori algorithm that only needs to scan the database once, which can effectively improve the efficiency of generating effective frequent itemsets. In addition, Liu Fang and Wu Guangchao [34] proposed a method to convert the database into matrix form to improve algorithm efficiency by reducing the size of the candidate item set and reducing the generation of useless candidate item sets. The FP-Growth algorithm is the most classic improvement on the Apriori algorithm. It uses frequent pattern trees (FP-tree) to store frequent item sets and reduce the number of database scans.

In addition to the above two correlation algorithms, literature proposed a new method to deeply investigate driver speed selection behavior in severe weather, using trajectory-level data obtained from the SHRP2 natural driving study, using non-parametric association rule mining and parametric ordered logistic regression. In order to study the impact of severe weather on driver speed selection behavior in the natural environment, two analysis methods, non-parametric association rule mining and parametric ordered logistic regression, were used. Due to the large SHRP2NDS and road information datasets, speed selection, especially in bad weather, is a complex driving behavior, considering that unlike parametric models, it has no predefined assumptions and can reveal the relationship between variables in big data. There are unclear and complex relationships between them, so association rule mining methods are used [35]. To select interesting rules and determine the strength of the association, various measures are used, including support, confidence, and lift. Support represents the frequency of an item set appearing in the data set, which can be defined as the percentage of transactions (i.e., speed selection behavior in this study) in the entire data set covered by the rule [36].

Although the association analysis method can solve most problems, there are many shortcomings in the current association analysis method. How to use the association rule algorithm to effectively process unstructured data, and how to combine the association rule algorithm with other decision-making methods to achieve more accurate data. Analysis, etc., are subject to further research and development.

4.4.2. Classification Analysis

For data analysis in urban traffic dynamic scenes, data classification technology is one of the most important methods to achieve data information mining and result prediction. Classification refers to dividing data into defined categories through algorithms. Commonly used classification algorithms include decision tree algorithms, rule-based classification methods, artificial neural network algorithms, deep learning algorithms, support vector machine (SVM) algorithms, Bayesian algorithms, etc.

A. Ceccarelli uses multiple data sources (such as sensor data, social media data, etc.) for classification to further improve the accuracy of data classification [37]. Du Yanping reviewed a variety of methods for classification and analysis of multi-source heterogeneous twin data, and compared and summarized each method [38]. These methods include genetic programming and ensemble learning, fusing multi-sensor data, leveraging multiple data sources, and more. These methods improve the accuracy and efficiency of data classification and provide better support and assistance for realizing intelligent transportation systems and traffic management.

4.4.3. Cluster Analysis

Clustering is to group similar data into one category. The principle is to maximize the similarity of each category of data. Commonly used clustering algorithms include four categories: partition-based clustering methods, hierarchical-based clustering methods, density-based clustering methods and model-based clustering methods.

Q. Ye proposed a multi-objective optimization method for urban transportation network division based on multi-source heterogeneous data, combining information such as road segment flow, speed and traffic accident data [39]. J. Aslam proposed a traffic anomaly detection method based on multiview clustering, which clusters and detects anomalies in traffic flow by fusing sensor data and GPS trajectory data. D. Kim proposed a traffic detection and clustering method based on multi-sensor data, which improves the accuracy of traffic detection and clustering by fusing multiple sensor data such as video, microwave radar, and sonar [40]. Although parallel clustering algorithms have the advantages of high efficiency and good scalability for big data, the algorithm implementation is more complicated. Simple, efficient, and highly scalable classification and clustering algorithms for big data that do not consume more software and hardware resources are the main research and optimization directions in the future.

5. SUMMARY

With the rise and development of the concept of digital intelligence, the sources of multi-source heterogeneous twin data in traffic scenarios are more extensive and the data structures are more diverse. First, there are problems in data collection: Due to the wide range of multi-source heterogeneous twin data in traffic scenes, the challenges faced by data collection work are very complex. Secondly, there are problems in heterogeneous data fusion: different data sources in traffic scenarios have different data types and formats, which makes data fusion very complicated. During data fusion, it is necessary to consider the differences in data quality between different data sources and how to unify the format of data between different data sources. Finally, there are problems in data analysis and mining: when analyzing and mining multi-source heterogeneous twin data, the correlation between data needs to be considered. However, due to the multiple connections between data, the relationship between data is very small. Difficult to accurately characterize and excavate.

In order to solve the above difficulties, the future can be based on the following perspectives to strengthen the collection and processing technology of multi-source heterogeneous twin data and optimize data quality and accuracy. Develop more reliable heterogeneous data fusion algorithms and techniques to coordinate data formats and processing methods among different data sources. Delve into the relationship between multi-source twin data and explore the most effective data analysis and mining methods. Combining deep learning, reasoning, image processing and other technologies to focus on exploring the establishment of digital twin models to better support data integration and decision-making in traffic scenarios.

REFERENCES

[1] Lee E A. Cyber physical systems: Design challenges [C]//2008 11th IEEE international symposium on object and component-oriented real-time distributed computing (ISORC). IEEE, 2008: 363-369.

[2] Grieves M W. Product lifecycle management: the new paradigm for enterprises [J]. International Journal of Product Development, 2005, 2(1-2): 71-84.

[3] Glaessgen E, Stargel D. The digital twin paradigm for future NASA and US Air Force vehicles [C]//53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference 20th AIAA/ASME/AHS adaptive structures conference 14th AIAA. 2012: 1818.

[4] Kim Y, Kim C M, Han Y H, et al. An efficient strategy of nonuniform sensor deployment in cyber physical systems [J]. The Journal of Supercomputing, 2013, 66(1): 70-80.

[5] Mandolla C, Petruzzelli A M, Percoco G, et al. Building a digital twin for additive manufacturing through the exploitation of blockchain: A case analysis of the aircraft industry [J]. Computers in industry, 2019, 109: 134-152.

[6] Tuegel E J, Ingraffea A R, Eason T G, et al. Reengineering aircraft structural life prediction using a digital twin [J]. International Journal of Aerospace Engineering, 2011, 2011.

[7] Korth B, Schwede C, Zajac M. Simulation-ready digital twin for realtime management of logistics systems [C]//2018 IEEE international conference on big data (big data). IEEE, 2018: 4194-4201.

[8] Schluse M, Rossmann J. From simulation to experimentable digital twins: Simulation-based development and operation of complex technical systems [C]//2016 IEEE International Symposium on Systems Engineering (ISSE). IEEE, 2016: 1-6.

[9] Kharlamov E, Martin-Recuerda F, Perry B, et al. Towards semantically enhanced digital twins [C]//2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018: 4189-4193.

[10] Merkle L, Segura A S, Grummel J T, et al. Architecture of a digital twin for enabling digital services for battery systems [C]//2019 IEEE international conference on industrial cyber physical systems (ICPS). IEEE, 2019: 155-160.

[11] Soni R, Bhatia M. Digital twin: intersection of mind and machine [J]. International Journal of Computational Intelligence & IoT, 2019, 2(3).

[12] Haag S, Anderl R. Digital twin–Proof of concept [J]. Manufacturing letters, 2018, 15: 64-66.

[13] Schroeder G N, Steinmetz C, Pereira C E, et al. Digital twin data modeling with automationml and a communication methodology for data exchange [J]. IFAC-PapersOnLine, 2016, 49(30): 12-17.

[14] Schleich B, Anwer N, Mathieu L, et al. Shaping the digital twin for design and production engineering [J]. CIRP annals, 2017, 66(1): 141-144.

[15] Desert island. Intelligent road transportation in the 5G era [M]. BEIJING BOOK CO. INC., 2020.

[16] Duan H, Shen Y, Ding P. Research on construction technology of real digital twin traffic scene based on edge-cloud collaboration [C]//Third International Conference on Computer Science and Communication Technology (ICCSCT 2022). SPIE, 2022, 12506: 1119-1124.

[17] Wang S, Zhang F, Qin T. Research on the construction of highway traffic digital twin system based on 3D GIS technology [C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1802(4): 042045.

[18] Li Y, Zhang W. Traffic flow digital twin generation for highway scenario based on radar-camera paired fusion [J]. Scientific reports, 2023, 13(1): 1-15.

[19] Ivanov S, Nikolskaya K, Radchenko G, et al. Digital twin of city: Concept overview [C]//2020 Global Smart Industry Conference (GloSIC). IEEE, 2020: 178-186.

[20] Wang Z, Yu G, Zhou B, et al. A train positioning method based-on vision and millimeter-wave radar data fusion [J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(5): 4603-4613.

[21] Cai H, Zhu J, Zhang W. Quality deviation control for aircraft using digital twin [J]. Journal of Computing and Information Science in Engineering, 2021, 21(3).

[22] Yumnam A S, Sreeram Y C, Naeem S A. Overview: Weblog mining, privacy issues and application of Web Log mining [C]//2014 International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2014: 638-641.

[23] El Saddik A. Digital twins: The convergence of multimedia technologies [J]. IEEE multimedia, 2018, 25(2): 87-92.

[24] Karadeniz A M, Arif İ, Kanak A, et al. Digital twin of eGastronomic things: A case study for ice cream machines [C]//2019 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2019: 1-4.

[25] Pargmann H, Euhausen D, Faber R. Intelligent big data processing for wind farm monitoring and analysis based on cloud-technologies and digital twins: A quantitative approach [C]//2018 IEEE 3rd international conference on cloud computing and big data analysis (ICCCBDA). IEEE, 2018: 233-237.

[26] Steger S, Mair V, Kofler C, et al. Correlation does not imply geomorphic causation in data-driven landslide susceptibility modelling–Benefits of exploring landslide data collection effects [J]. Science of the total environment, 2021, 776: 145935. [32]

[27] Kai Liu, Meng-Ying Cui, Peng Cao, Jiang-Bo Wang. Iterative Bayesian Estimation of Travel Times on Urban Arterials: Fusing Loop Detector and Probe Vehicle Data [J]. PLOS ONE, 2016, 11(6).

[28] Nantes A, Ngoduy D, Bhaskar A, et al. Real-time traffic state estimation in urban corridors from heterogeneous data [J]. Transportation Research Part C: Emerging Technologies, 2016, 66: 99-118.

[29] Ibrahim M, Rjabtšikov V, Gilbert R. Overview of digital twin platforms for EV applications [J]. Sensors, 2023, 23(3): 1414.

[30] Semeraro C, Lezoche M, Panetto H, et al. Data-driven invariant modelling patterns for digital twin design [J]. Journal of Industrial Information Integration, 2023, 31: 100424.

[31] Yi yang Yao, Yingjie Xia, Zhenyu Shan, Zhengguang Liu. Learning for Traffic State Estimation on Large Scale of Incomplete Data [P]. International Conference on Multimedia Retrieval, 2016.

[32] Li Q, Chen L, Li M, et al. A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios [J]. IEEE Transactions on Vehicular Technology, 2013, 63(2): 540-555.

[33] Chen Q, Ma X, Tang S, et al. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds [C]//Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. 2019: 88-100.

[34] Sun W, Zhang X, Peeta S, et al. A real-time fatigue driving recognition method incorporating contextual features and two fusion levels [J]. IEEE transactions on intelligent transportation systems, 2017, 18(12): 3408-3420.

[35] Keechoo Choi, YounShik Chung. A Data Fusion Algorithm for Estimating Link Travel Time [J]. Journal of Intelligent Transportation Systems, 2002, 7(3-4).

[36] Faouzi N. Data-driven aggregative schemes for multisource estimation fusion: a road travel timeapplication [J]. Proceedings of Spie the International Society for Optical Engineering, 2004, 5(1) :351-359.

[37] Zeng Y, Lan J, Ran B, et al. A Novel Multisensor Traffic State Assessment System Based on Incomplete Data [J]. The Scientific World Journal, 2014, 2014: 1-13.

[38] Javaid M, Haleem A, Suman R. Digital twin applications toward industry 4.0: a review [J]. Cognitive Robotics, 2023, 3: 71-92.

[39] Liu X, Jiang D, Tao B, et al. A systematic review of digital twin about physical entities, virtual models, twin data, and applications [J]. Advanced Engineering Informatics, 2023, 55: 101876.

[40] Li X, Feng M, Ran Y, et al. Big Data in Earth system science and progress towards a digital twin [J]. Nature Reviews Earth & Environment, 2023, 4(5): 319-332.