Accepted Manuscript

A Survey on Data Fusion in Internet of Things: Towards Secure and Privacy-Preserving Fusion

Wenxiu Ding, Xuyang Jing, Zheng Yan, Laurence T. Yang

PII: S1566-2535(18)30473-1
DOI: https://doi.org/10.1016/j.inffus.2018.12.001
Reference: INFFUS 1059

To appear in: Information Fusion

Received date: 5 July 2018
Revised date: 29 November 2018
Accepted date: 3 December 2018

Please cite this article as: Wenxiu Ding, Xuyang Jing, Zheng Yan, Laurence T. Yang, A Survey on Data Fusion in Internet of Things: Towards Secure and Privacy-Preserving Fusion, Information Fusion (2018), doi: https://doi.org/10.1016/j.inffus.2018.12.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Highlights

- Propose a number of IoT data fusion requirements to comment existing work.
- Review the data fusion methods in three typical application domains.
- Discuss security and privacy issues of data fusion in IoT.
- Discover open issues and propose a number of future research directions.
A Survey on Data Fusion in Internet of Things: Towards Secure and Privacy-Preserving Fusion

Authors: Wenxiu Ding\textsuperscript{1}, Xuyang Jing\textsuperscript{1}, Zheng Yan\textsuperscript{1,2*}, Laurence T. Yang\textsuperscript{3}

1. The State Key Laboratory on Integrated Services Networks, School of Cyber Engineering, Xidian University, Xi’an, China  
2. Department of Communications and Networking, Aalto University, Espoo, Finland  
3. Department of Computer Science, St. Francis Xavier University, Antigonish, Canada  

*Corresponding author: Zheng Yan (zyan@xidian.edu.cn; zhengyan.pz@gmail.com)

Abstract
Internet of Things (IoT) aims to create a world that enables the interconnection and integration of things in physical world and cyber space. With the involvement of a great number of wireless sensor devices, IoT generates a diversity of datasets that are massive, multi-sourcing, heterogeneous, and sparse. By taking advantage of these data to further improve IoT services and offer intelligent services, data fusion is always employed first to reduce the size and dimension of data, optimize the amount of data traffic and extract useful information from raw data. Although there exist some surveys on IoT data fusion, the literature still lacks comprehensive insight and discussion on it with regard to different IoT application domains by paying special attention to security and privacy. In this paper, we investigate the properties of IoT data, propose a number of IoT data fusion requirements including the ones about security and privacy, classify the IoT applications into several domains and then provide a thorough review on the state-of-the-art of data fusion in main IoT application domains. In particular, we employ the requirements of IoT data fusion as a measure to evaluate and compare the performance of existing data fusion methods. Based on the thorough survey, we summarize open research issues, highlight promising future research directions and specify research challenges.

Keywords: Data Fusion, Internet of Things, Data Privacy, Security, Smart Home, Smart Grid, Smart Transportation

1 INTRODUCTION

IoT aims to create a world that enables the interconnection and integration of things in physical world and cyber space [1]. As an emerging technology, it is expected to connect all items (such as, physical devices [2, 3], vehicles [4], home appliances [5-7], etc.) and enable them to exchange data. IoT plays an important role in various practical systems for remote perception and control. Specifically, it has permeated into many real scenarios, including but not limited to healthcare, environment, transportation, agriculture and education, for the purpose of providing intelligent services and applications.
To benefit from IoT, perceiving and gathering data about environments and physical objects is the first step for intelligent services. Multiple sensors can help improve the breadth of collected information, which also is the only way to enhance the depth of data fusion result because a single sensing modality is usually inadequate [8, 9]. Thus, many sensors/devices are normally deployed to sense and perceive various data (such as, environmental data, medical data, infrared data, etc.). But due to multi-source heterogeneity and large amount of sensory data [10], it is not wise to distribute all data, which obviously wastes network bandwidth and device power energy. Hence, data fusion becomes a significant technique to dig out vital data from the widely sensed or collected data to improve data quality and facilitate decision making.

There exist many definitions of data fusion [11-13]. Bostrom et al. [14] gave the definition of information fusion as “Information fusion is the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making”. Concretely, data fusion can reduce the size and dimension of data, optimize the amount of data and extract useful information from them. It helps eliminate the imperfect of data and overcomes the complicit of sensed data from different sensors. Data fusion in IoT has been widely researched in recent years. However, it still incurs some issues and challenges, such as privacy leakage and power consumption.

Some existing papers have tried to give an overview of the efforts in data fusion. Lee et al. [15] provided a review on data fusion techniques, but they did not give a discussion on the data fusion specifically designed for IoT big data. Alam et al. [16] focused on the data fusion in smart ubiquitous environments but ignored the security and privacy issues in IoT. The literatures [17, 18] investigated data fusion methods. Pires et al. [19] mainly discussed identification of activities, while Wang et al. [20] reviewed and analyzed existing work according to an evaluation framework that consists of ten features but ignores security and privacy issues. Faouzi et al. [21] mainly discussed the application of data fusion in intelligent transportation. Though there exist many survey papers about multisensory data fusion as listed above, they failed to provide a comprehensive overview on IoT data fusion [1, 22] and neglected the security and privacy issues in the process of data fusion.

In this paper, we perform a thorough survey on data fusion in IoT. We first specify data fusion requirements based on IoT data properties by paying attention to the security and privacy issues. We then review and discuss existing work in three main IoT application domains (i.e., smart home, smart grid, and smart transportation) by employing the requirements as a measure to comment pros and cons of each work. We also compare key issues and data fusion techniques in the three domains and emphasize security and privacy issues in data fusion of IoT by providing some possible solutions. Different from prior arts, we aim to provide a generic and comprehensive review on data fusion methods in IoT, offer a deep discussion on the existing work in specific IoT application domains for the purpose of specifying open issues, especially those related security and privacy and proposing emerging research
trends in this field.

Table 1 Comparison of surveys about IoT data fusion

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Objectives and Topics</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>[15]</td>
<td>Focus on data fusion algorithms and theories</td>
<td>Not highly related to IoT data fusion</td>
</tr>
<tr>
<td>[16]</td>
<td>Focus on data fusion methods in specific IoT environments</td>
<td>Ignore privacy and security in IoT data fusion</td>
</tr>
<tr>
<td>[17]</td>
<td>Classification of data fusion methods: stage-based methods, feature-level-based methods, semantic-based methods</td>
<td>Do not touch concrete IoT application scenarios</td>
</tr>
<tr>
<td>[18]</td>
<td>Classification of data fusion methods according to challenging problems of input data</td>
<td>Do not discuss concrete IoT applications</td>
</tr>
<tr>
<td>[19]</td>
<td>Focus on sensing and fusion methods for identification of activities in daily life</td>
<td>Do not touch the privacy issues and the difference and relationships among different applications in IoT</td>
</tr>
<tr>
<td>[20]</td>
<td>Analyze existing work according to an evaluation framework with ten features</td>
<td>Do not cover any security and privacy issues</td>
</tr>
<tr>
<td>[21]</td>
<td>Focus on the challenges and issues in intelligent transportation</td>
<td>Do not touch the privacy issues and the difference and relationships among different applications in IoT</td>
</tr>
<tr>
<td>Our work</td>
<td>Specify data fusion requirements, analyze existing work in different IoT application domains; point out the differences and characteristics of three popular IoT application domains; emphasize security and privacy issues in data fusion of IoT and provide some possible solutions</td>
<td>Do not touch too much on detailed design of fusion methods</td>
</tr>
</tbody>
</table>

Table 1 compares existing related surveys with our work. We can see the difference of our work from existing ones and its novelty. Specifically, the contributions of this paper can be summarized as follows:

- We propose a number of IoT data fusion requirements by paying attention to security and privacy and employ them as a measure to comment existing work for the purpose of figuring out open issues and proposing future research trends.
- We review the data fusion methods in three typical application domains (i.e., smart home, smart grid, and smart transportation), which are chosen according to data features in terms of privacy sensitivity, sensor number and sensor distribution. In particular, we point out the differences and focuses of data fusion research in each domain during survey.
- We discuss security and privacy issues of data fusion in the above three application domains and present possible solutions of security and privacy issues.
We discover the open issues in IoT data fusion and propose a number of future research directions in order to guide future research, especially the issues highly related to security and privacy.

The rest of this paper is organized as follows. In Section 2, we introduce popular applications of IoT, specify IoT data properties, and propose a number of requirements of IoT data fusion. We provide a detailed survey on existing data fusion methods applied in three main IoT application domains in Section 3. Based on our literature survey, we further discuss a number of open research issues and propose future research directions in Section 4. Finally, a conclusion is presented in the last section.

2 IoT DATA PROPERTIES AND DATA FUSION REQUIREMENTS

In this section, we give a classification of IoT application domains, specify the characteristics of IoT data, introduce some basic knowledges and applications of data fusion, and further summarize a number of data fusion requirements in order to employ them as a measure to comment and discuss existing work.

2.1 IoT Applications

IoT enables objects to be sensed or even controlled remotely and offers the opportunities for the integration of physical world into network systems. It results in high efficiency, accuracy and economic benefit with reduced human interactions [23-25]. Currently, there is a rapidly growing interest in applying IoT technologies in various applications, such as industry, home, government, etc. Due to its wide usage, it is impossible to envisage all possible applications of IoT. Herein, we first classify the IoT applications into different domains and then choose some typical application scenarios in each domain for further discussions. In this paper, we mainly classify the IoT applications according to the characteristics of sensors and collected data into such application domains as smart home, smart industry and smart transportation. In this subsection, we introduce these three IoT application domains by specifying their specific data characteristics and sensor characteristics.

Smart home aims to improve our daily life, such as inhabitant monitoring or home appliance management. Home-related data are highly privacy-sensitive and crucial in making some decision or performing context inference [26]. For example, the data about lights can indicate whether a householder is at home or not. For another example, the data used for medical analysis carry individual health information, which are extremely private. Low-quality data fusion may lead to an incorrect health alarm, which may cause a serious risk to a patient. Thus, the applications in the home domain highly request data confidentiality and fusion accuracy. On the other hand, they normally contain a fixed number of sensors that seldomly changes.

As a typical example in industry, smart grid is a modern electric power-grid infrastructure [27]. It aims to provide improved efficiency, reliability, and safety with smooth integration of renewable and alternative energy sources through automated control and modern communication technologies [28]. Sensors are employed for each consumer and usually cover a larger area than a smart home. A smart grid system connects millions of smart meters and produces high-dimensional and complex data.
The data are transmitted over open communication networks, which would be vulnerable to various data attacks, such as false data injection. Although the data collected in the smart grid are less privacy-sensitive than home-related data, they also carry some private information and data integrity should be ensured.

Smart government aims to provide convenient services to citizens, such as smart education and smart transaportation, etc. Data confidentiality in the applications of this domain does not come first, but reliability and availability of services are becoming important. Thus, the completeness of collected data is of great importance. Usually, widely collected data are taken into consideration for offering services with high qualities. Taking an important application, smart transportation, as an example, various kinds of sensors (such as road sensors and devices on board of vehicles) are deployed in this kind of applications. In order to offer real-time transportation services (e.g., real-time traffic monitoring and incident detection), real-time, accurate and reliable data collection and processing are highly expected. Different from smart grid, though smart transportation also deploys lots of sensors, it usually takes into consideration of sensors in a wide/medium-sized area, such as one city or one street.

Table 2 summarizes the characteristics of data and sensors in the above three typical IoT applications. We can see that the data in IoT applications in different domains hold such a typical characteristic as high privacy-sensitive. Besides data fusion reliability, each domain has some different key security issue.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Smart home</th>
<th>Smart grid</th>
<th>Smart transportation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy sensitivity</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Sensor distribution</td>
<td>Small area</td>
<td>Wide area</td>
<td>Wide/Medium Area</td>
</tr>
<tr>
<td>Security goals</td>
<td>Confidentiality</td>
<td>Integrity</td>
<td>Timeliness</td>
</tr>
</tbody>
</table>

### 2.2 IoT Data Properties

IoT enables a huge number of sensors and devices to communicate seamlessly to facilitate our daily life. The number of sensors is expected to reach 50 billions by 2020. Various sensors are deployed to perceive/collect useful data and further provide a better understanding of surroundings through data fusion and analytics. IoT data hold different characteristics from those collected in other scenarios (e.g., trust recommendation in pervasive social network). Before providing a comprehensive discussion on data fusion in IoT, we first summarize the characteristics of IoT data highly related to data fusion as follows:

1) **Big data**: A large number of sensors keep monitoring/observing physical objects and generate large amounts of data [29, 30], which results in “big data”. Thus, big data fusion becomes crucially essential in IoT. But it also faces a number of challenges caused by limited storage and energy of sensor devices and other reasons: a) it is difficult to distinguish spurious data; b) fusion accuracy suffers from over/under confidence on data dependency and correlation; c) wide distribution of data incurs high communication burden and serious delay in...
centralized data fusion models.
2) Multi-modality: The involvement of multi-modality data from various sensors (e.g., wearables, cameras and microchips) enhances the breadth of collected data. But the diversity of data with different types, forms, representations, scales and densities makes it hard to fuse the data directly.
3) Privacy-sensitive: Observed data may always carry some personal information of an observed target. For example, some sensed data may disclose the location of a user, or even his/her consuming habits or preferences. Fusing multisensory data increases the risk of privacy invasion. How to preserve user privacy during data fusion and at the same time ensure fusion accuracy is an important research issue.
4) Imprecise: Data perceived by various sensors may be imprecise, inaccurate and uncertain due to data loss or data source unreliability, which brings additional challenges for data fusion caused by data imperfection, data conflict, data ambiguity and inconsistency.
5) Dynamic: Data sensing is time-variant because context changes over time. Hence, the freshness of collected data becomes a vital issue, which is also a key factor that impacts the quality of data fusion [31].

2.3 Traditional Data Fusion Methods
Many mathematical theories are available for fusing data. The literatures [15-18] have provided deep discussion and introduction to these methods. Herein, we just give a brief introduction to some popular data fusion methods. According to the discussion in [17, 19], data fusion methods can be categorized as probability-based methods, evidence reasoning methods, and knowledge-based methods.

**Probability-based methods (PBM):** Probability distribution or density function were introduced to cope with data imperfection, which can express the dependency between random variables and set up the relationships among different datasets. Probability-based data fusion methods include Bayesian inference [32, 33], state-space models, Markov models, belief propagation, maximum likelihood methods, possibility theory, evidential reasoning and, more specifically, evidence theory, and least square-based estimation methods, e.g., Kalman filtering, optimal theory, regularization and uncertainty ellipsoids. But probabilistic fusion algorithms mainly have the following drawbacks [34]: (1) difficult to obtain a density function and define priori probabilities; (2) limited performance when dealing with complex and multivariate data; (3) cannot handle uncertainty.

**Evidence reasoning methods (EBM):** This type of methods includes Dempster-Shafer (D-S) theory and recursive operators. Compared with Bayesian inference, D-S theory has such advantages as representing ignorance due to the lack of information and aggregating beliefs when collecting new evidences [35, 36]. It introduces the concepts of belief and plausibility to represent the uncertainty in the real world and enable the inference in dynamic situation, where belief represents the degree of belief to which a certain evidence supports a certain event and plausibility indicates the degree of belief to which a certain evidence fails to refute a certain
event. It also introduces a mass function to represent the distribution of belief. However, it incurs a difficult issue of estimating mass function, which constrains its applications.

**Knowledge-based methods (KBM):** This kind of methods enables the fusion center to gain data from imprecise big data, which has no need to get density/distribution function. Knowledge-based methods include intelligent aggregation methods, machine learning, fuzzy logic and so on. Machine learning is mainly divided into three categories: (i) supervised learning, e.g., K-Nearest Neighbor (KNN), Support Vector Machines (SVM) and Naive Bayes (NB) that are classification algorithms, and Artificial Neural Network (ANN) that simulates a biological learning system; (ii) unsupervised learning, e.g., Density-Based Spatial Clustering of Applications with Noise (DBSCAN) that is a density-based clustering method, Chameleon that is a hierarchical clustering method, K-Means that is a partitioning clustering method and (iii) semi-supervised learning. Fuzzy logic [34] is based on the theory of fuzzy set, which can cope with imprecision and uncertainty owing to its intrinsic ability to integrate numerical control and symbolic planning, but is not flexible enough to incorporate prior knowledge.

Table 3 summarizes the advantages and disadvantages of the above popular data fusion methods and their use cases.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Use cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBM</td>
<td>Bayesian Inference</td>
<td>Simple, efficient and good performance in analyzing prior knowledge.</td>
<td>Activity recognition, traffic anomaly [32, 33]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Need a prior probability; high complexity in case of multivariate data and large amount of involved data; unable to deal with uncertainty</td>
<td></td>
</tr>
<tr>
<td>Kalman Filter</td>
<td>Simple, easy implementation, optimal in a mean-squared error sense</td>
<td>Be in a linear form and with zero-mean Gaussian noise; Sensitive to corruption of outliers.</td>
<td>Real-time traffic system [37]</td>
</tr>
<tr>
<td>Markov Models</td>
<td>Accurate prediction</td>
<td>Not suitable in long-term prediction</td>
<td>Appliance control [38]</td>
</tr>
<tr>
<td>Belief Propagation</td>
<td>Can deal with large amount of data</td>
<td>Easily lead to loopy graphs and difficult to converge when dealing with non-linear data</td>
<td>State estimation [39]</td>
</tr>
<tr>
<td>EBM</td>
<td>D-S Theory</td>
<td>Deal with missing information and complementary hypotheses; solve uncertainty problem</td>
<td>Occupancy sensing [40], fire detection [41], activity recognition[35]</td>
</tr>
<tr>
<td>KBM</td>
<td>Fuzzy Logic</td>
<td>Have the ability of deal with inaccurate and uncertain data</td>
<td>Fall detection [34]</td>
</tr>
<tr>
<td>Model</td>
<td>Description</td>
<td>Challenges</td>
<td>Applications</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>------------</td>
<td>--------------</td>
</tr>
<tr>
<td>ANN</td>
<td>High accuracy and precise prediction, suitable for nonlinear data fusion</td>
<td>Need various features in training process; have a high computation cost; unable to explain decisions.</td>
<td>Crowd flows forecasting [42]</td>
</tr>
<tr>
<td>SVM</td>
<td>Efficient to deal with huge datasets, have the ability of handling high dimensional data</td>
<td>Poor processing capacity for oversized data sets, sensitive to incomplete data.</td>
<td>Vehicle localization [4]</td>
</tr>
<tr>
<td>KNN</td>
<td>Model is easy to understand, and building the model is fast.</td>
<td>Slow prediction, cannot process datasets with many features and biased datasets.</td>
<td>Parking spot detection [43]</td>
</tr>
<tr>
<td>NB</td>
<td>Fast training and predicting speed, easy to understand, can be used for high-dimensional sparse datasets.</td>
<td>Cannot apply to attribute-related datasets.</td>
<td>Activity recognition [44]</td>
</tr>
<tr>
<td>K-Means</td>
<td>Simple and easy to understand and implement, low time complexity</td>
<td>Need to set an initial threshold, and very sensitive to noise and outliers.</td>
<td>Data injection detection [45]</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>Can identify the noise point and find the spatial clustering of any shape, high clustering speed</td>
<td>When density is uneven, clustering quality is poor.</td>
<td></td>
</tr>
<tr>
<td>Chameleon</td>
<td>No need to preset the number of clusters</td>
<td>Large calculation and storage requirements</td>
<td>Data injection detection [45]</td>
</tr>
</tbody>
</table>

### 2.4 IoT Data Fusion Requirements

Data fusion can be efficiently applied to reconstruct and fill missing values or data [2, 46, 47], to conduct prediction [48, 49], to identify the causality of specific events and give a conclusion about causal connections [50, 51], and to detect anomalies [52], etc. Though data fusion provides many advantages for data processing and analytics by enhancing dataset quality and reducing the volume of data transmission, the characteristics of IoT data as described above come up with new challenges for data fusion in IoT. First, collected data will be transmitted over the Internet, which may undermine privacy and be vulnerable to false data injection. Second, the large amount of data collected in IoT makes data fusion difficult, which increases the complexity and even introduces inconsistency and conflicted data. Third, the multi-modality and heterogeneity of data collected from various sensors further complicate data fusion. Herein, we figure out a number of requirements for secure and reliable data fusion in IoT as described below:

- **Context-Aware (CA):** Data fusion is desired to be context-aware in order to support adaptive and flexible services with high intelligence. Context specifies
any information that can be used to characterize the background or situation of involved entities [53], while the ability to identify and adapt to context is treated as context-awareness. The context information may not intrinsically relate to the entities, it may change over time. For example, the most widely used context information, location, determines which sensors should be taken into consideration in smart transportation. Context-awareness is a significant requirement of IoT applications, which should be also considered in IoT data fusion.

- **Privacy Preserving (PP):** Data fusion for further analysis would possibly reveal private information, such as user location, user preference, etc. Especially, the home-related data are extremely private. Data fusion without any consideration on privacy preservation may greatly affect user acceptance of an IoT system. Thus, privacy preservation is also regarded as an essential requirement for data fusion in IoT.

- **Reliable (Re):** Data fusion results often directly lead to a specified decision, such as diagnosis or emergency response. Unreliable fusion result may cause an unacceptable danger to a related user. Thus, reliability is one of the most basic but vital requirements in practical use of data fusion. The reliability and validity evaluation metrics in classification include several elements: 1) True Positive (TP) that indicates the number of positive samples predicted to be positive; 2) False Positive (FP) that indicates the number of negative samples predicted to be positive; 3) False Negative (FN) indicating the number of positive samples predicted to be negative; 4) True Negative (TN) representing the number of negative samples predicted to be negative. Hence, several metrics based on the four elements are computed and employed to evaluation the performance of fusion algorithm:
  1) Accuracy (ACC=(TP+TN)/(TP+FP+TN+FN));
  2) Precision Rate (PR=TP/(TP+FP));
  3) Recall Rate or Sensitivity (RR=TP/(TP+FN));
  4) F-Measure=2*PR*RR/(PR+RR);
  5) False Positive Rate (FPR=FP/(TN+FP));
  6) False Negative Rate (FNR=FN/(FN+TP));
  7) Specificity or true negative rate (Spec=TN/(TN+FP)).

Different from classification, clustering has several different metrics:
  1) Cluster accuracy;
  2) Rand Index;
  3) Compactness; etc.

- **Efficient (Ef):** Efficient data fusion can optimize the computation overhead of fusion center or sensors and further support fast response on data analysis. As the sensors continuously observe and collect data for data fusion, inefficient fusion method would make the fusion become a bottleneck in IoT services. To evaluate the performance of a data fusion method or scheme, several detailed aspects should be evaluated, such as training time, testing time, and communication cost for data collection and objective monitoring. As most existing work ignores the
performance evaluation, we only list the training time and testing time in the following analysis.

- **Real Time (RT):** Context changes with time, which calls for real-time data fusion and analytics within a short time period with quick response, especially in emergency (e.g., fall detection for heart patients [54]).

- **Robust (Ro):** Data fusion should be robust to resist various attacks (e.g., false data injection [52]) in IoT due to the adoption of open networks. Otherwise, it would aggravate the imperfection of collected data and produce an incorrect result.

- **Verifiable (Ve):** Verifiability means that the fusion result can be verified by a fusion center or the public. On one hand, this requirement enables service consumers to evaluate the correctness of data fusion result. On the other hand, it helps the fusion center or the public to check the quality of collected data. Especially, data fusion with verifiability can help judge whether a piece of collected data actually contributes to a final decision.

3 **DATA FUSION IN IoT**

In this section, we review and discuss the recent advances of data fusion in three main IoT application domains. In particular, we compare and discuss existing works by employing the proposed data fusion requirements as a measure for the purpose of figuring out open issues and assisting us in directing future research in this field.

3.1 **Smart Home**

Data fusion is employed in smart home to provide inhabitants with a remote central control for their appliances [55] or a remote monitoring of elderly people or patients in home. Here, we take healthcare monitoring as an example to discuss fusion methods in smart home, which has fewer sensors and smaller amount of data than smart industry and smart transportation.

3.1.1 **Data fusion in smart home**

Monitoring the status of the elderly or chronically ill patients in their own homes is an essential requirement for delivering more effective pervasive healthcare, which can help doctors/family members to follow their status and guarantee they can live independently and safely. A diversity of sensors (wearable sensors based on accelerometers [56, 57], pulse oximeters [57], temperature [56], and bend sensors [58, 59]) are deployed in smart home to collect data, which will be forwarded to the gateway for fusion as shown in Figure 1. Besides the medical-related data, environmental variables collected by ambient sensors are also significant factors in healthcare monitoring. Moreover, the visual data can provide real-time and rich data for detecting the human shape from a scene and analyzing human activity, which can achieve a more correct inference. Many researches have contributed their efforts to healthcare and activity recognition [60, 61] in smart home based on different types of sensors and methods.
Different from other two application scenarios, lots of personal home-related data are collected but extremely privacy-sensitive [62]. Hence, we mainly focused on the existing efforts in three aspects: privacy-preserving data fusion, reliable and robust data fusion, and context-aware data fusion.

1) Privacy-preserving data fusion

The collected data in home can be transmitted based on a network protocol, such as ZigBee, WiFi, etc. But the fusion relies on the honest of gateway and the fusion result needs to be shared through open networks, which threatens user privacy. With regard to data fusion in smart home, color cameras and depth sensors provide effective detection of human shape by extracting information from sensed data. However, visual data affect the privacy of inhabitants, which calls for privacy preserving approaches in videos and images [54, 63-65]. Julien et al. [32] proposed to use wearable e-AR sensor [57] and blob sensors to collect data, and then perform data fusion based on Gaussian Bayes Expectation Maximization (EM) classifier, which can classify nine classes of activities. The blob sensor containing on-board processor extracts signal metrics (e.g., silhouette) and local motion in optical flow form in real time, which can reduce sensitivity and preserve inhabitant privacy but incur computation overhead to ambient sensors. Moreover, the extraction time could affect the efficiency and timeliness for constrained computation capability of local sensors. Chaaraoui et al.[63] improved the accuracy of accurate activity recognition through multi-view of cameras and proposed the concept of privacy by context, which points out several levels of privacy preservation and related variables including identity, location, appearance etc. It applies weighted feature fusion scheme over the fusion result of each view, while privacy preservation is achieved through image processing during remote connection but ignored the security and privacy of fusion results. Hence, the privacy depends on a trusted fusion center. With regard to efficiency, it achieves accuracy rate of 91.4% in 207 Frames Per Second (FPS) on the IXMAS dataset. Moreover, its extraction stage took about 84% of the total processing time,
which results in an insignificant recognition time and supports almost real-time recognition.

Hong et al. [36] introduced anonymous binary sensors and employed D-S theory to recognize the inhabitant activity, which collects data from several contact sensors installed in cupboards, the fridge, the microwave etc., and passive infrared sensors in the ceilings of the home. Moreover, it considered the impact of the numbers of sensors and the reliability of each sensor on the overall results in decision making process, which is a guideline for measuring the confidence in the fusion result or identifying the thresholds of the minimum number of sensors in real environment. But it failed to provide the implementation result.

The work above can achieve data privacy, but the fusion center is still able to infer about some private hypotheses if its statistical data are available. In order to further enhance the privacy, Sun et al. proposed to [66] to introduce the concept of an empirical normalized risk and design a nonparametric privacy-aware optimization framework, which can limit the detection rate of private hypotheses without impairing that of public hypotheses. This scheme can be widely used in various IoT applications (e.g., recognition activity, Census Data analysis) but ignore the data fusion design.

The previous work [32] provides a high accuracy when moving but it has only 47% correct detection of sitting. In order to overcome the limitation of motion sensors, Nesa et al. [40] involved some different sensed data including temperature, humidity, light and CO2 to detect the occupancy in a room. They took into consideration of the probability density functions of each kind of readings and calculated the Probability Mass Assignment (PMA) for each parameter $y$ to be in class $x$ as follows:

$$m_y(x) = \frac{\rho_y(x)}{\sum_x \rho_y(x) + n_x (1 - \alpha_y \beta_y)}$$

where $\rho_y(x)$ is probability density function, $n_x$ is the number of training samples in class $x$, $\alpha_y$ and $\beta_y$ are maximum probability density function value and distributing factor of parameter $y$, respectively. Then the Dempster’s combination rule was introduced to combine the PMA values and make a final decision. Through experiments and comparison, this method shows superiority to several classification models including random forest, gradient boosting machines, and linear discriminant analysis, especially recall rate and specificity. Though the authors claimed it can support real-time detection, they did not provide results on efficiency and ignored the concern on robustness and verifiability.

Furthermore, Zimmermann et al. [67] proposed to use environmental sensors (such as Total Volatile Organic Compounds (TVOC) and CO2 sensors) to detect occupancy and estimate the number of occupants, which are non-intrusive to user privacy. The authors further tested several fusion methods to test the effectiveness of environmental sensors and obtained the accuracies of 81.1% in occupancy detection and 64.7% in occupant estimation. But it also ignored context awareness and robustness.

2) Reliable and robust data fusion

Fleury et al. [61] proposed to first filter the noise before data fusion, which can
improve the accuracy and also reduce the communication cost. Based on the activities of volunteering actors, it results in 72% of well-classified movements, which is not convincing enough. In order to improve the reliability further, Hamid and Jerome [34] proposed to apply fuzzy logic to detect the fall of elderly people. The intrinsic nature of fuzzy logic to deal with data imprecision and uncertainty helps increase the system reliability and robustness when facing temporary sensor malfunctions. This scheme collects multi-modality data from more sensors in home including microphones, wearable device, infrared sensors and demotic sensors for environment condition monitoring, which obtains a higher accuracy than another fuzzy logic based scheme [68] that only considers posture, location and duration information. However, they ignore privacy issues.

To further enhance recognition accuracy, Mahjoub et al. [44] used Kinetc camera to capture data and design a hybrid data fusion scheme. First, it obtains the feature fusion result and then apply naive-Bayes to combine the scores from three classifiers: Collaborative Representation Classifier (CRC), Sparse Representation Classifier (SRC) and Kernel based Extreme Learning Machine (KELM). It provides a high recognition accuracy of 90.5% over datasets enclosing 27 actions. However, none of them can preserve the privacy and the work [44] does not provide robustness.

3) Context-aware data fusion

In order to prompt inference in a dynamic situation, Han et al. [69] focused on the change of belief of each focal element to provide urgent decision. They designed the Basic Probability Assignment (BPA) for sensed values at an interval of 10 seconds as follows:

\[
m(T'_i) = \frac{\sum_{T_{i-1} \cap T_{i-0} \neq \emptyset} m(T_{i-1}) * m(T_i)}{1 - \sum_{T_{i-1} \cap T_{i-0} \neq \emptyset} m(T_{i-1}) * m(T_i)}, i = 1, 2, ..., n
\]

where \( m(T_i) \) means the value of BPA function at the time slot \( T_i \). Moreover, this work tried to reduce the calculations for recognizing the moving object and obtaining the infer result quickly by only considering several time slots, which improved the real-time response. But this work did not provide implementation to show its accuracy in detailed use cases.

Chahuara et al. [38] proposed to apply Markov Logic Network (MLK) to make the context-aware decision in smart home, which also introduced multiple sensors to construct the system. It can aim to classify the sitting, tidying, washing up dishes and so on, which gains a higher accuracy of 54% in case of sitting than [32]. But it is still not reliable enough. Furthermore, Paola et al. [70] designed a three-tier architecture to achieve a context-aware and adaptive system for data fusion, which is composed of collection tier, fusion tier and self-optimization tier. Dynamic Bayesian Network (DBN) is applied in fusion tier to deal with the dynamicity of phenomena and cover the past states, which has no restrictions on the conditional probability distributions. The belief about a specific state \( x \) in the time slot \( t \) is defined as \( Bel(x) = P(x|E_{1:t}, C_{1:t}) \), where \( E \) is the sensory readings and \( C \) is the context information. To improve the quality of involved sensors, Kullback-Leibler (KL)
divergence [71] was introduced to evaluate the importance of sensors to help select the most important sensors for data collection, which can help enhance the reliability and robustness. Moreover, it can reduce the execution time and take about 81% of that when all sensors are involved. However, these schemes do not consider user privacy.

In general, the schemes above can only satisfy partial fusion requirements pre-defined in Section 2.4. Specifically, most work ignore robustness and verifiability, which is significant to realize secure and reliable IoT data fusion. The privacy and accuracy are carefully taken into consideration in most work. In detail, Table 4 shows a summary and comparison of data fusion schemes in the smart home domain. We evaluate and compare the existing work in terms of the following criteria, which is applied in all tables throughout this paper:

- **Context-Aware (CA):**
  - Yes (Y): It can adapt to various scenarios.
  - No (N): It does not consider context awareness.

- **Privacy-Preserving (PP):**
  - Yes (Y): It considers the privacy issues and take some countermeasures to solve it.
  - No (N): It does not mention it.

- **Reliable (Re):** As most work did not provide all metric values, we will not indicate those metrics in our tables if no work shows them. These metric values evaluate the reliability and validity of data fusion, which will be denoted by concrete values from 0% to 100%. As some clustering schemes do not evaluate their performances according to the metrics, hence we directly use accuracy (ACC) to measure them.

- **Efficient (Ef):** Efficiency is measured by training time (Tr) and testing time (Te) that are obtained from original paper.

- **Real-Time (RT):**
  - Yes (Y): It considers the real-time objective in scheme design.
  - No (N): Real-time requirement is not considered in the scheme.

- **Robust (Ro):**
  - Yes (Y): It can deal with the issue of temporary sensor malfunction or dishonest sensor nodes.
  - No (N): It neglects robustness and is vulnerable to attacks.

- **Verifiable (Ve):**
  - Yes (Y): Verifiability is considered in the design.
  - No (N): Do not discuss about verifiability.

- **:-** not applicable, given or mentioned
<table>
<thead>
<tr>
<th>Ref.</th>
<th>Scenarios</th>
<th>Collected data type</th>
<th>Method</th>
<th>Datasets</th>
<th>Requirements</th>
<th>CA</th>
<th>P</th>
<th>ACC</th>
<th>PR</th>
<th>RR</th>
<th>F-measure</th>
<th>FP</th>
<th>FN</th>
<th>Sp</th>
<th>R</th>
<th>T</th>
<th>R</th>
<th>V</th>
<th>T</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>[32]</td>
<td>Activity classification</td>
<td>e-AR data; blob sensor data</td>
<td>Bayes</td>
<td>Two actors in a built simulated home</td>
<td>N</td>
<td>Y</td>
<td>75</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>[63]</td>
<td>Activity recognition</td>
<td>Camera data</td>
<td>Weighted feature fusion</td>
<td>Weizmann Dataset, MuHAVi Dataset, IXMAS Dataset</td>
<td>N</td>
<td>Y</td>
<td>95.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[36]</td>
<td>Activity recognition</td>
<td>Contact, tilt, passive infrared sensor data</td>
<td>D-S Theory</td>
<td>-</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>[66]</td>
<td>IoT (e.g., activity recognition)</td>
<td>-</td>
<td>Bayesian error probability</td>
<td>UCI repository</td>
<td>-</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[40]</td>
<td>Occupancy sensing</td>
<td>Temperature, humidity, light, CO2 levels</td>
<td>D-S Theory</td>
<td>Real data from a coffee room</td>
<td>N</td>
<td>Y</td>
<td>92.24</td>
<td>89.1</td>
<td>89.5</td>
<td>89.33</td>
<td>6.24</td>
<td>10.4</td>
<td>93.7</td>
<td>-</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Topic</td>
<td>Data Sources</td>
<td>Method</td>
<td>Results</td>
<td>Other Notes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>--------------</td>
<td>--------</td>
<td>---------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[67]</td>
<td>Occupancy detection</td>
<td>TVOC, CO2, air humidity</td>
<td>Naive Bayes classifier</td>
<td>Real data of a student dormitory</td>
<td>N Y 81.1 - - - - - Y N N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[61]</td>
<td>Sound classification, activity recognition</td>
<td>Microphone, magnetic kinematic sensor data</td>
<td>Gaussian Mixture Model classifier</td>
<td>Testing data from 15 actors</td>
<td>N N 72 - - - - - N Y N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[34]</td>
<td>Fall detection</td>
<td>Acoustical, medical, infrared sensor data</td>
<td>Fuzzy Logic</td>
<td>HOMECAD [77]</td>
<td>Y N 95 97 97 - - 96 - - Y Y N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[44]</td>
<td>Activity Recognition</td>
<td>Kinetic camera</td>
<td>PCA, CRC, SRC, KELM, Naive Bayes</td>
<td>UTD-MHA D[51]</td>
<td>Y N 90.5 - - - - - N N N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[69]</td>
<td>Context Inference</td>
<td>Noise, ultrasonic, ultrasonic sensor data</td>
<td>D-S theory</td>
<td>-</td>
<td>Y - - - - - - - Y - -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[38]</td>
<td>Intelligent controller</td>
<td>Infrared presence detector, contact sensors, cameras, microphones</td>
<td>MLK</td>
<td>15 actors to produce data for 11 hours</td>
<td>Y N 73 - - - - - N N N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[70]</td>
<td>Activity recognition</td>
<td>Motion, door, temperature sensor data</td>
<td>Dynamic Bayesian Network</td>
<td>Aruba dataset [78]</td>
<td>Y N 79.5 40.8 - - - - - 0.81 x N Y N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.1.2 Discussions on data fusion in smart home

As discussed above, there exist many efforts on secure and reliable data fusion schemes in smart home. But some issues still need to be solved. From Figure 1, data fusion involves several steps: 1) data collection in smart home; 2) data fusion at gateway; 3) access control of fusion results. Here we list some problems according to the comparison and summary in Table 4.

1) Data Collection

How to collect enough useful data for fusion becomes a primary issue because home-related data are privacy-sensitive and essential for safety of property and inhabitant. First, from Table 4, we can observe that there is a diversity of sensors available in smart home and that visual sensory data can help improve the accuracy [44, 63]. But visual sensory data fusion increases the computation overhead, which needs to extract features and then reduce dimensions through Principal Component Analysis (PCA). Moreover, though visual sensors offer rich information [79] in activity recognition, it carries more private information, such as identity, location, appearance, etc. Environmental sensors perform well in occupancy detection without impairing privacy of inhabitant, but it failed to realize the activity recognition. In fact, most existing work mentioned above ignore the privacy preservation in smart home, which needs further research.

Second, it has been proved that appropriate selection of sensors can help improve efficiency in previous work [70], thus the reliability and trust of sensors still needs further researches to enhance the performance of data fusion.

Generally, it is a primary and vital step to deploy suitable sensors which can balance the fusion efficiency and privacy preservation in the smart home.

2) Data Fusion

First, from the discussion above, we can find that feature extraction can be used to reduce the communication cost caused by various sensors, especially visual sensors which contributes much to the reliability. But how to balance the communication cost and the computation overhead at the local sensor becomes a new issue.

Second, there are several datasets available for training (such as, UCI repository [75], HOMECAD [77], UTD-MHAD [51] and Aruba dataset [78], etc.), while some work collected data from concrete scenarios with some limited volunteering actors which makes the fusion accuracy unconvincing. The quality of datasets influences the accuracy of data fusion, but sometimes it is difficult to obtain enough data. Hence, it is still an issue to generate behavior datasets for various objectives in smart home [80], which can help overcome the weakness of experiments in real world: imperfection of sensed data, short learning periods and insufficient volunteers for training [81].

3) Access Control

Current existing work spared no efforts to improve the accuracy of data fusion but did not take into consideration of access control over the fusion results. The verifiability of fusion result, even off-line fusion result which has lower requirement on real-time response, have not been considered in existing work. Actually, fusion result carries more meaningful data than original data, which makes it more significant to control the access to fusion results. Generally, it has become an essential
requirement in the highly distributed IoT system.

3.2 Smart Grid

As a typical application example of smart industry, smart grid provides such capabilities as power generation, power delivery, and power utilization for autonomous power distribution and management. It can effectively deliver power and quickly respond to wide ranging events. But with the continuous development of smart grid, the scale and structural complexity of grid are continually growing. Multi-source data monitored during grid operation become a very important part of measuring grid service and the maintenance of the whole grid system. How to effectively analyze these data has become an urgent problem.

Moreover, as massive, heterogeneous and multi-source data (such as, measured states and control actions) are collected and exchanged over the interconnected and open communication network, they are vulnerable to several types of attacks: vulnerability attack, data injection attack, and intentional attack [82, 83]. Hence, besides the privacy issue, data fusion comes up with robustness requirement in smart grid. In addition, real-time data fusion also becomes serious due to the massive multi-modality of data.

3.2.1 Data fusion in smart grid

Different from home-related data fusion, the amount of data in smart grid is much larger, which introduces many new challenges: false data injection, high communication cost, different sampling rates, etc. Hence, many schemes are designed for reliable and efficient data fusion in smart grid.

1) Reliable data collection

Due to the highly distributed and open network in smart grid, the data generated by millions of sensors/devices bring heavy burden to data transmission and increase the communication cost. Moreover, it becomes one attractive target of many attackers, which can distort real energy demand and then obtain an erroneous analysis result. Reliable data collection mainly includes two parts: one is to reduce the transmitted data, another is to solve bad data processing that covers the detection of measurement errors and bad data injection.

In order to reduce the amount of data to be transmitted in the wireless channel and saving energy, Farias el al. [84] revised the basic algorithm and designed an Enhanced Bayesian Inference (EBI), an Enhanced Dempster-Shafer inference (EDS) and an Enhanced Fault Tolerant Interval (EFTI) for information fusion in Shared Sensor and Actuator Networks (SSANs) scenarios. They further tested their enhanced multisensory data fusion methods in the overhead power line monitoring and the battery monitoring, which showed that the efficiency and accuracy are increased compared with the traditional methods. EBI accuracy is reduced in non-static scenarios, while EDS can adjust its belief function to support context-aware scenarios. But it only considers the fusion of single-modality data using different fusion algorithms and does not provide the hybrid multi-sensor data fusion function.

Besides the reduce of data amount, to improve the reliability of collected data is also an effective way, which has been researched in many works. Liu et al. [45]
proposed a method to detect the bad data injection and locate the attacks in a small area. L-bounded Graph Partition Method (LGPM) [85] based on K-means is applied to partition the weighted undirected graph of power system into subgraphs and then Chi-squares test [86] is used to detect bad data in the subgraph. In subgraph, the threshold corresponding to a specified detection confidence is lower than that of the whole system and the redundancy of measurements can be reduced. Experiments showed its success in detecting bad data injection.

To further deal with the dynamics and heterogeneities in smart grid, Chen et al. [87] proposed a lightweight approach based on spatial-pattern recognition and temporal-pattern-consistency evaluation to balance efficiency and scalability. A spatial-correction consistency region can be obtained, which represents possible correct estimation pairs from different smart components and can be directly used to check consistency. Furthermore, two rounds of trust-based voting are applied to identify the anomalies. Kordestani et al. [88] presented a new fusion method for fault diagnosis. Discrete Wavelet Transform (DWT) and Radial Basis Functions Artificial Neural Network (RBF-ANN) were utilized to detect fault, respectively. And then the ordered weighted averaging (OWA) operator combines the diagnosis to make a final decision and locate faults. However, both of the above methods focus on reliability but ignore user privacy and timeliness that could be affected by a complex procedure.

Reliable data collection resisting various attacks [52, 89, 90] is of paramount significance to provide efficient and reliable power transmission and distribution, which can help improve the security and trustworthiness. It lays the basis for further data analysis and fusion.

2) Data aggregation

As one of the most basic computation and fusion algorithm, data aggregation has been researched in various scenarios. But some is not suitable in smart grid for the largescale aggregation. Ruj et al. [91] designed a decentralized framework for smart grid. It simultaneously solves the issues of privacy preservation and access control. All data are encrypted through homomorphic encryption and access control with revocation is realized through Attribute-Based Encryption (ABE). Besides the privacy issue, data integrity is also an important factor which can help resist accidental errors and internal/external attacks. Hence, homomorphic signature was introduced in [92] to protect the data integrity and verification is designed for detecting and tracing the anomaly in this scheme. However, both the two schemes focused on only one-dimensional data. Lu et al. [93] proposed to support multi-dimensional user report via homomorphic encryption and realized privacy-preserving data aggregation.

There still exist many reliable and privacy-preserving schemes [94-96] for aggregating consumer data. However, all these works focused on single-modality data, which is not enough for complicated application scenarios, especially the risk assessment and state estimation in smart grid.

3) Risk assessment

Compared with traditional grid systems, smart grid becomes more vulnerable to cyber-attacks owing to the increased access points for smart components to connect the local networks or the Internet. Cheng et al. [97] used Radial Basis Functions
Neural Network (RBFNN) to fuse multi-source data to provide a quantitative risk assessment of power information system. They first collected data from network management tools such as Intrusion Detection System (IDS), and then formatted the data into a uniform style and extracted the features values. The optimal choice of the feature values was considered as input neuron of network. Continuously training the RBFNN based on an improved genetic algorithm until it meets the given error precision. The method can predict the value of threats and provide credible information for dynamic risk assessment with a low error rate of 0.113. But it does not touch privacy preservation.

Wang et al. [98] proposed a deep data fusion model to fuse cross-domain Energy Management System (EMS) data for risk perception and coordinated control of smart grid. The model includes a kernel principal component analysis algorithm used to find the efficient state representation, a coordinated control algorithm used to obtain optimal solution, and a mutual backup strategy used to automatically switch the critical system. The efficiency and accuracy of the proposed model are verified on real-world large-scale smart grid data. It also ignores the privacy issue and verifiability.

4) State estimation

The aim of state estimation of power systems is to provide a global perspective of the grid for specific automation and control purposes [99, 100]. Traditional state estimation system of grid faces a serious challenge in integrating Phasor Measurement Units (PMUs) and Supervisory Control and Data Acquisition (SCADA) data due to the different sampling rates of them. Ghosal and Rao [101] presented an improved multi-rate multi-sensor data fusion algorithm to fuse PMUs and SCADA data for dynamic state estimation in power system. First, they adopted Kalman filtering to deal with each type of data at different sampling rates. Then, a simpler multi-time step prediction approach is adopted to scale down the data at different time-scale. Finally, two types of data can be fused based on Bar-Shalom-Campo fusion formula [102]. The experiment results showed that the improved fusion has a low estimation error covariance of about 10%. Though this scheme can overcome the weakness of static state estimation, it depends on the linearized model of power system dynamics and measures. In order to solve this problem, the authors further proposed a new scheme based on Unscented Kalman filtering in [103], which relaxes the limitation that the ratio of the two sampling rates should be an integer.

Belief propagation (BP) algorithm for non-linear power flow models easily leads to loopy graphs and is difficult to converge. In order to overcome this issue, Fusco et al. [39] proposed a novel data fusion computational framework for power systems state estimation, which can reduce the number of possible loops and become scalable for very large and complex grid topologies by using complete state estimation problems (or parts of) to represent individual factor nodes. Moreover, it can reduce the effect of noise and overcome the occurrences of missing or erroneous data. The experiments showed that it obtains a better solar generation forecast. This scheme solves the problem of loops and decomposes the matrix inversion of dimension $n_1 + n_2$ to four inversions of dimension $n_1$ and $n_2$. It saves computation cost if the dimensionality
is high. However, none of the work considers the issue of privacy. Having the same format as in Table 4, we summarize and compare data fusion methods in smart grid in Table 5. From the table, we can see that most of work still cannot satisfy all requirements, which encourage us to work out more general data fusion methods in smart grid. Especially, verifiability is ignored almost in all work while privacy preservation is only considered in data aggregation schemes. Generally, privacy preservation is still an open issue in multi-modality data fusion.

3.2.2 Discussions on data fusion in smart grid
As discussed above, data fusion in smart grid also faces some challenges. Here, we give a discussion according to the analysis in Table 5.

1) Data collection
Different from data collection in smart home, the data in smart transportation are highly distributed and transmitted in a more open network, which is more vulnerable to various attacks. Thus, it is more important to guarantee the reliability of data collection, which can be solved in two ways according to previous discussions. One is to apply data fusion methods to reduce the amount of transmitted data and improve the data quality of data, another is to enhance the reliability of data by detecting false data injection. From Table 5, we can observe that those schemes of detecting false data injection as the first step of data fusion ignore the demand of privacy preservation.

2) Data fusion
There are mainly two types of data in smart grid in state estimation: PMU data sets and SCADA data sets. In the case of simple data aggregation of PMU data, many schemes have been proposed to achieve privacy preservation by introducing decentralized system. Moreover, the homomorphic signature can help verify the correctness of aggregation result. But when it comes into two kinds of data sets, data sampling at different rates becomes one of the most serious issues in smart grid. Kalman filtering and belief propagation are efficient to fuse the large amount of data, while the scaling down of data with different sampling rates is applied to cope with different datasets. However, they ignore the privacy preservation in complex data fusion methods, which needs further research.

3) Access control
We can observe that access control of fusion result is only considered in data aggregation and that verifiability of data aggregation is achieved via homomorphic signature. The access control of fusion result from more complex methods is still vacant.
Table 5 Summary and comparison of data fusion methods in smart grid

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Scenarios</th>
<th>Collected data type</th>
<th>Method</th>
<th>Datasets</th>
<th>Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>[84]</td>
<td>Power line monitoring and battery monitoring</td>
<td>Device communication data</td>
<td>EBI</td>
<td>Simulation</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>EDS</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>EFTI</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>[45]</td>
<td>Data injection detection</td>
<td>Line power flow</td>
<td>K-means or EM</td>
<td>IEEE 39 bus</td>
<td>-</td>
</tr>
<tr>
<td>[87]</td>
<td>False data injection detection</td>
<td>State estimation</td>
<td>Spatiotemporal cyber-state correlation</td>
<td>US smart grid</td>
<td>Y</td>
</tr>
<tr>
<td>[88]</td>
<td>Fault Diagnosis</td>
<td>Circuit Breakers data and voltage data</td>
<td>OWA, ANN,</td>
<td>IEEE 14bus test case [104]</td>
<td>-</td>
</tr>
<tr>
<td>[91]</td>
<td>Consumer data aggregation</td>
<td>Consumer data</td>
<td>Aggregation</td>
<td>-</td>
<td>Y</td>
</tr>
<tr>
<td>[92]</td>
<td>Consumer data aggregation</td>
<td>Consumer data</td>
<td>Aggregation</td>
<td>-</td>
<td>Y</td>
</tr>
<tr>
<td>[93]</td>
<td>Consumer data aggregation</td>
<td>Multi-dimensional consumer data</td>
<td>Aggregation</td>
<td>-</td>
<td>Y</td>
</tr>
<tr>
<td>[97]</td>
<td>Risk assessment</td>
<td>IDS data</td>
<td>RBFNN</td>
<td>800 samples</td>
<td>N</td>
</tr>
<tr>
<td>[98]</td>
<td>Risk perception</td>
<td>EMS data</td>
<td>Deep data fusion model</td>
<td>East China project</td>
<td>N</td>
</tr>
<tr>
<td>[101]</td>
<td>Dynamic state estimation</td>
<td>PMU and SCADA Data</td>
<td>Kalman filtering+Bar-Shalom-Campo fusion</td>
<td>IEEE 14bus test case</td>
<td>Y</td>
</tr>
<tr>
<td>Reference</td>
<td>State Estimation</td>
<td>Data Source</td>
<td>Estimation Method</td>
<td>Systems</td>
<td>Y</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------</td>
<td>-------------</td>
<td>-------------------</td>
<td>---------</td>
<td>---</td>
</tr>
<tr>
<td>[103]</td>
<td>Nonlinear Dynamic State Estimation</td>
<td>PMU and SCADA Data</td>
<td>Unscented Kalman filtering + weighted sum</td>
<td>IEEE 9, 39, 57 and 118-bus systems</td>
<td>Y</td>
</tr>
<tr>
<td>[39]</td>
<td>State estimation</td>
<td>SCADA Data</td>
<td>Belief propagation</td>
<td>Open Power System Data[105]</td>
<td>Y</td>
</tr>
</tbody>
</table>
3.3 Smart Transportation

Different from the scenarios above, smart transportation emphasizes real-time requirement more owing to the fast speed of vehicles. Intelligent transportation systems are the most significant application of the smart government, which provide accurate traffic information through analyzing multi-source traffic data. Traffic data can be divided into temporal data and spatial data. Normally, temporal traffic data are collected from road sensors (mostly use inductive loop detectors) embedded around the road. It carries basic traffic information such as traffic volume, occupancy and speed at a specific point. Spatial traffic data are usually collected from road cameras, remote traffic microwave sensor, satellites, etc. To get an accurate and comprehensive traffic situation, multi-source data fusion is expected.

3.3.1 Data fusion in smart transportation

Due to the involvement of moving vehicles, the scenarios of smart transportation become more dynamic and complex, compared to other IoT applications. The research on data fusion is applied to support smart transportation, in such applications as parking spot detection, position estimation, and traffic monitoring. Different applications of data fusion raise different requirements on timeliness. In this subsection, we review existing work on data fusion in smart transportation according to the purposes of fusion.

1) Static parking spot detection

One application of data fusion in smart transportation is to detect parking spot, which does not involve vehicles with high speed and has lower requirement of timeliness. Moreover, the parking information is less sensitive than those in smart home and smart grid. But it needs high accuracy.

Sevillano et al. [43] proposed a smart traffic management system that supports parking spot detection by fusing the data collected by widely-deployed video surveillance camera networks. In order to capture different scene parameters, they extracted global features that describe the whole annotated region and local features that describe salient points within the annotated region. Global features consist of Histograms of Oriented Gradients (HOG) and Gabor histogram that have the ability of capturing gradient orientations and texture. They carried out a feature-level fusion by the concatenation of the 9-dimensional HOG and 8-dimensional Gabor descriptors into a single feature vector. A binary classification whether a parking spot is occupied or vacant is conducted by applying KNN and SVM respectively with global and local features as input. Through experiments, HOG coupled with KNN can result in a good compromise between accuracy and context-awareness. But its performance evaluation was also not given. Carnelli et al. [106] presented a parking activity detection system called ParkUs, which collected data from low-power sensors such as the magnetometer and accelerometer sensors to find on-street parking in congested urban areas and used a median filter to remove any outlier data. A novel sensor fusion feature called the Orthogonality Error Estimate (OEE) was generated from various
extracted features. ParkUs was able to detect parking activities with high accuracy by analyzing the values of OEE. Both the works aim to detect parking spot, which is not so privacy sensitive.

2) Context-aware traveler information estimation

Different from parking spot detection, more context information are taken into consideration for obtaining a good journal recommendation. Usually it is not so strict on timeliness.

Yu et al. [107] designed a journey planning system called JPlanner to carry out journey recommendation. In order to provide effective and accurate recommendations, JPlanner utilized more comprehensive urban data including road map information, public transportation information, traffic information and weather information. JPlanner also solved the accuracy problem caused by rapid changes of traffic condition through conducting a speed fusion module. This module adopted a weighted sum fusion method to fuse two datasets collected by stationary speed cameras and probing vehicles to estimate the road link level speed. Moreover, the confidence value is introduced to evaluate the confidence of observations with the exponential function \( f(x) = e^{-x/\alpha} \). The authors compared the journey recommendation results of JPlanner with two popular journey planners (Google Maps and Gothere) to demonstrate the advantages of JPlanner.

Traveler information estimation offers users integrated traveler information. In traveler information estimation, various types of traffic data that contain detailed traffic conditions are used to derive indicators to assist in user guidance [108]. Zhang et al. [109] proposed a novel data fusion system to estimate dynamic road travel time. The fusion system consists three parts: (1) data normalization, the function of this part is to unify the description of traffic information collected from multi-sensors with different scale. (2) the fusion framework, this part fuses multiple data sources in a uniform Spatio-Temporal context and gets the optimal source. (3) the Uncertainty Feedback (UNIF) method, UNIF is applied to detect and clean outliers according to change interval determined by optimal source. Furthermore, they compared the accuracy of this method with other three fusion methods including simple average, Kalman filtering and ANN to show its higher accuracy and robustness.

The scenario above changes fast and has a high requirement of context awareness. Compared with parking spot detection, more sensors are involved, which emphasize the importance of multi-sensor data fusion for higher accuracy in dynamic scenarios.

3) Real-time estimation

One prominent feature of smart transportation is its rapid changes, which leads to a urgent demand of real-time estimation.

Position estimation plays an important role in smart transportation, which enables the vehicle trace and location-based recommendations. Belhajem et al. [37] combined Extended Kalman filtering (EKF) and neural networks (NN) to achieve real-time and reliable vehicle position estimation. EKF is used to fuse raw data collected from odometer and gyrometer sensors and then compute the vehicle predicted position.
When new Global Positioning System (GPS) measurements arrive, the EKF updates the predicted position. When no GPS information arrives, position errors are eliminated by NN which have been trained. An improved work with SVM was proposed in [4], which achieves improvement of up to 94% over the simple EKF predictions in case of GPS failures. But the position of vehicle is detected through data fusion, and the data are collected and analyzed in raw format, which brings threats to user privacy.

Besides vehicle position estimation, traffic monitoring is also one of the most important part in intelligent transportation systems and many techniques have been developed for this goal. Sufficient traffic data collection and accurate traffic prediction schemes are the premise of traffic monitoring. Huang et al. [110] used simulation based Dynamic Traffic Assignment (DTA) models to fuse multi-sensor multi-source surveillance traffic data collected from cameras, loop and toll counters. They made use of the correlation relationship among traffic data to yield consistent approximation of the network condition. Moreover, their framework has the ability to estimate the parameters of simulator from the costly optimization of an equation incorporating model parameters. Kong et al. [111] introduced a fusion based system with the help of Sydney Coordinated Adaptive Traffic System (SCATS) to realize real-time traffic state surveillance. The system consists of three modules: 1) SCATS data processing that converts traffic information collected by SCATS into a kind of link-based spatio-temporal mean-speeds using traffic wave theory; 2) GPS data processing that uses GPS-equipped taxis as probe vehicles to obtain the real-time mean-speed of travel through each road section in that road network; and 3) Fusion module that combines the Federated Kalman filtering and the D-S evidence theory to fuse the two results generated from other two modules. Hence, a more comprehensive and accurate estimation of traffic state is obtained. Robert [112] presented a new framework to monitor traffic state of an intersection. As a pre-processing, the sensors (video cameras and inductive loops) are projected onto a satellite map and calibrated. Then a vehicle detection algorithm is applied to extract vehicle position and kinematics from camera frames. Moreover, the author used a deterministic fusion method to fuse data features extracted from video cameras and inductive loops to get an additional feature: the vehicle length. The state of intersection is aligned and displayed on satellite maps by analyzing these three features. The framework also employed a Kalman filtering to linearly track the vehicles along the lanes.

The work above takes in more collected data to improve the accuracy of traffic monitoring. But how to balance the accuracy and the efficiency is still an open issue as both measurements are highly related to the number of different types of collected data. Privacy is not an issue for traffic monitoring is public for all drivers and collected data is not easily connected to specific vehicle.

Incident detection aims to recognize road events to have an emergency response, which also need to obtain real-time and accurate fusion result. Jayarajah et al. [113] proposed an architecture that supports to detect and localize anomalous events from urban informatics. The degree of traffic camera-based congestion and public bus occupancy levels were estimated by analyzing the data collected by physical sensors.
At the same time, Twitter feedbacks were also estimated in order to enhance the accuracy of event localization at social sensors. Finally, a fusion component gave a fused detection decision through taking all analysis results into account. After verifying the architecture in real scenario, they showed that it has a high accuracy and fine granularity to detect and localize urban anomalies by fusing such physical and sensor data.
<table>
<thead>
<tr>
<th>Ref.</th>
<th>Scenarios</th>
<th>Collected data type</th>
<th>Method</th>
<th>Datasets</th>
<th>Requirements</th>
<th>CA</th>
<th>PP</th>
<th>RF (%)</th>
<th>EF</th>
<th>RT</th>
<th>Ro</th>
<th>Ve</th>
</tr>
</thead>
<tbody>
<tr>
<td>[43]</td>
<td>Parking spot detection</td>
<td>HOG+Gabor</td>
<td>KNN</td>
<td>Real-world video data</td>
<td>Y N 90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SVM</td>
<td>Real-world video data</td>
<td>Y N 85,76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[106]</td>
<td>Parking activity detection</td>
<td>Car and GPS data</td>
<td>Wavelet entropy</td>
<td>Real-world data</td>
<td>Y N 98 98</td>
<td></td>
<td></td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td>0.9 min</td>
</tr>
<tr>
<td>[107]</td>
<td>Traveler information estimation</td>
<td>Traffic, weather</td>
<td>Linear weighted sum</td>
<td>Arterial real data in Singapore</td>
<td>Y N - - - - -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[109]</td>
<td>Traveler information estimation</td>
<td>Traffic data</td>
<td>Weighted average method</td>
<td>Beijing traffic data</td>
<td>Y N - - - - -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[37]</td>
<td>Position estimation</td>
<td>Car and GPS data</td>
<td>Extended Kalman filtering</td>
<td>Institute Pascal Data Sets</td>
<td>Y N - - - - -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[110]</td>
<td>Traffic monitoring</td>
<td>Cameras, loop and toll counters data</td>
<td>Dynamic Traffic Assignment model</td>
<td>Real-world motorway data</td>
<td>N N - - - - -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[111]</td>
<td>Traffic monitoring</td>
<td>SCATS data, GPS data</td>
<td>Extended Kalman filtering + D-S Evidence Theory</td>
<td>Real-world data provided by traffic police headquarters</td>
<td>N N 97.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[112]</td>
<td>Traffic monitoring</td>
<td>Camera and loop data</td>
<td>Deterministic fusion decision</td>
<td>Real-world monitor data</td>
<td>N N 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[113]</td>
<td>Incident detection</td>
<td>Social feeds and Decision fusion</td>
<td>real-world</td>
<td>Y N 90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Having the same format as in Table 4, we summarize and compare data fusion methods in smart transportation in Table 6. From the table, we can see that most existing work only considers the requirement of timeliness in smart transportation. This is inspired by the demand that we need to monitoring traffic situation in real-time. Moreover, some work ignores context awareness and robustness of data fusion in smart transportation. When some emergencies or some attacks occur, the performance of data fusion will be dramatically reduced, and hence easily leads to a serious inaccurate data fusion result.

### 3.3.2 Discussions on data fusion in smart transportation

As discussed above, data fusion in smart transportation still faces some other challenges. Here, we give a discussion according to the analysis in Table 6.

1) **Data Collection**

   Most work ignored the privacy issue. With regard to parking spot detection, data collection has no need to emphasize too much on privacy preservation. But in position estimation, it would still be a serious issue for motivating vehicles to contribute their own data. In addition, how to balance the efficiency and the accuracy still needs consideration. The selection of data types directly affects the timeliness of response.

2) **Data Fusion**

   Timeliness is regarded as a significant issue in existing work. But the robustness of data becomes an issue for the dynamic changes of sensors or data contributors/vehicles on the road. Owing to the severe consequences of traffic accidents, it has a more urgent expectation of high robustness of data fusion in smart transportation, which can overcome the issues of missing data or false data injection for traffic monitoring and accident indication. From the above work, we can find confidence analysis and correlation analysis are potential countermeasures to improve the observations.

3) **Access Control**

   The location of vehicle is privacy sensitive, while the widely distributed sensors over the road for monitoring traffic status makes it a serious issue, especially when transmitted in the highly open network. Thus, it also needs further research to control the access of fusion result in smart transportation.

### 3.4 Comparison and analysis of data fusion in three application domains

As discussed above, data fusion in each scenario has its own emphasis and involves different sets of data. In order to provide a more comprehensive evaluation, here we present a brief summary and comparison in Table 7.

Some work in smart grid concentrates on the consumer data fusion. The single-modality data with privacy preservation can be easily achieved. However, it is not effective for data fusion in other scenarios with multi-modality data. Home-related data are privacy sensitive, which need to be protected. The data in smart transportation may be less privacy sensitive.

Besides the reliability of data fusion methods in all scenarios, the data quality is also an important factor on improve the reliability. Due to the highly distribution of sensors in smart grid and smart transportation, the protocols connecting networks in
smart home (such as ZigBee, Bluetooth, etc.) become unavailable. But trust evaluation or KL divergence may be a useful way to improve the quality of data sources or sensors.

Table 7 Comparison of three application domains of data fusion

<table>
<thead>
<tr>
<th></th>
<th>Smart home</th>
<th>Smart grid</th>
<th>Smart transportation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collected data</td>
<td>Multi-modality</td>
<td>Single-modality, sometimes multi-modality</td>
<td>Multi-modality</td>
</tr>
<tr>
<td>Sensitive data</td>
<td>All kinds of data</td>
<td>Consumer data</td>
<td>Vehicle position</td>
</tr>
<tr>
<td>Primary concern</td>
<td>Privacy concern</td>
<td>Data integrity</td>
<td>Real-time</td>
</tr>
<tr>
<td>Robustness and reliability</td>
<td>Secure transmission channel, such as ZigBee, Bluetooth, etc.</td>
<td>False data injection detection; sensor selection based on trust</td>
<td>Data fusion methods</td>
</tr>
<tr>
<td>Access Control</td>
<td>Rely on gateway</td>
<td>Decentralized fusion nodes</td>
<td>--</td>
</tr>
<tr>
<td>Methods</td>
<td>Bayes; D-S theory; Fuzzy logic, etc.</td>
<td>Kalman filtering; Belief propagation; Neural network, etc.</td>
<td>D-S theory; Extended Kalman filtering, etc.</td>
</tr>
<tr>
<td>Challenges</td>
<td>Privacy preservation; high accuracy</td>
<td>Different sampling rates, dynamic injection, robustness</td>
<td>Real-time issue; location privacy</td>
</tr>
</tbody>
</table>

In smart home, data security and privacy preservation can be partially guaranteed by the secure channels, but it becomes highly related to the honesty of gateways. In smart grid, cluster-based scheme can be employed to balance the fusion efficiency and privacy. However, it becomes a serious issue for highly distributed and rapidly changed scenarios in smart transportation, which has not been researched.
Finally, an overview of fusion framework and theoretic methods is given in Figure 2. During data collection or sensing, trust evaluation or KL divergence can be applied to improve the reliability, robustness and privacy preservation. In addition, context awareness can be realized in this step. Various fusion techniques can be applied to realize feature extraction, classification, and decision fusion, etc. By improving the fusion algorithms, the five fusion requirements (i.e., efficiency, reliability, robustness, timeliness, and context awareness) can be satisfied or improved. After decision fusion, verifiability and access control can be executed in the last step. Homomorphic encryption/signature is a possible technique for verifiability, while ABE may be useful for fine-grained access control.

<table>
<thead>
<tr>
<th>Fusion Framework</th>
<th>Requirements</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensor selection and control</td>
<td>Reliability, Robustness, Context awareness</td>
<td>1) Trust evaluation or KL divergence; 2) Adaptive sensor selection; 3) False data injection detection, etc.</td>
</tr>
<tr>
<td>Feature extraction/reduction</td>
<td>Efficiency, Reliability, Robustness, Timeliness, Context awareness</td>
<td>1) HOG; LBP; 2) PCA; LDA; GBM; 3) CRC, SRC, KELM; 4) K-Means; EM; DBSCAN; 5) DE combination rule; Naïve bayes; Maximization; Weighted feature fusion, etc.</td>
</tr>
<tr>
<td>Preprocessing</td>
<td>Privacy, Preservation</td>
<td>1) Homomorphic encryption/signature; 2) ABE, etc.</td>
</tr>
<tr>
<td>Decision level fusion</td>
<td>Classifier/Clustering</td>
<td>1) Decision level fusion</td>
</tr>
<tr>
<td>Decision</td>
<td>Verifiability, Access Control</td>
<td>1) Homomorphic encryption/signature; 2) ABE, etc.</td>
</tr>
</tbody>
</table>

Figure 2 Overview of fusion framework and theoretic methods

4 OPEN ISSUES AND FUTURE RESEARCH DIRECTIONS

4.1 Open Issues

According to the discussion above, we figure out a number of open issues as below.

First, sensor selection is still a serious issue in IoT data fusion. The accuracy of data fusion is influenced by the number of sensors and the reliability of selected sensors. A great number of sensors are employed to perceive data. Some are of less significance than others while some are essentially worthless. Moreover, different sensors generate data in various formats and with different sampling rates, which cause the difference of data resolution, accuracy and reliability. Analyzing and processing multi-modality...
data incur high computation and communication cost, which makes it difficult to support real-time services. Thus, it is not a wise choice to involve all types of sensors for different purposes, which could be a good way to reduce computation cost and communication overhead at gateways where the collected data are fused. Xu et al. [114] first designed a source selection principle based on internal-confidence and external-confidence degrees to judge the reliability of information sources. Data utility was introduced by Bijarbooneh et al. [2] to measure the quality of data from selected sensors and reduce the communication consumption. How to setup suitable sensors and adaptively select enough sensors is a still a challenge.

Second, privacy preservation is always neglected in existing work. As discussed above, motion sensors and video sensors often carry too much personal information, which increases the risk of privacy invasion. Though environmental sensors without motion enables a highly accurate detection of occupancy, it cannot deal with activity recognition. In smart grid, privacy preservation of consumer data in aggregation has been researched, but existing work ignores the privacy issue in multi-sensor data fusion, which could bring threats to users. Moreover, how to achieve privacy preservation in a rapidly changed context, such as smart transportation, is still an open issue. Though some efforts tried to apply encryption algorithms [115-118] to achieve privacy preservation in extracting features from images, the timeliness requirement in smart transportation makes it inappropriate to apply encryption and seriously obfuscates this issue.

Third, how to balance communication cost and computation overhead for data fusion is still an open issue, which is highly related to concrete application scenarios. A central node can be applied to fuse all collected data in a distributed system while different fusion algorithms can be also implemented in sensors to perform local data fusion. However, due to the limited capability of sensor nodes, fusion impacts the lifetime of sensor battery and may result in unsatisfactory performance [119]. Data fusion at gateways can overcome the above issues and have such advantages as improved reliability and robustness, and increased quality of data. Only the gateway that gets all collected data for fusing can reduce the risk of privacy leakage if it is trusted. However, the gateway needs to collect all data and undertakes all computation overhead, which would become a bottleneck. Hence, how to reduce the overhead of gateways is a serious issue. The effort in [120] proposed one countermeasure to solve it. First, it sets up a threshold. Only the data satisfying the threshold requirement are transmitted to a fusion node, which carries out some initial data fusion. Then, it adopts a structural pattern recognition algorithm to get the value of the threshold at a computing center and continuously update the threshold at the sensor node. But this method is not suitable for a highly distributed system.

Fourth, hybrid data fusion still needs further research. Multi-sensor data are always involved in data fusion for improving the reliability of fusion result. But it comes up with two new challenges. One is that different modality complicates data fusion. The other is that different sampling rates makes it difficult to fuse data together. Hybrid fusion methods can help choose the best method to fuse data in each modality. Moreover, it can solve the issue of fusing data at different sampling rates. Though
some hybrid data fusion [101] in smart grid has been proposed, it only supports the data fusion over two sets of data and its performance is highly related to data types. Hybrid data fusion generically suitable for more application scenarios is worth our efforts for further investigation.

4.2 Future Research Trends

Based on the open issues, we further propose some interesting research directions in IoT data fusion for the purpose of guiding future research.

- IoT data fusion based on edge computing
  
  IoT data fusion at edge devices should be studied for the purpose of achieving real-time, reliable, privacy-preserving and robust data fusion. With abundant computing and storage resources, cloud computing can help users reduce heavy computation and storage burdens [121]. However, it incurs new challenges in data fusion scenarios [122]. First, it is not flexible and efficient enough to support distributed IoT systems due to high communication cost. Second, the cloud-assisted schemes incur high delay due to possible congestion in the Internet, while local fusion increases computation overhead of local sensors. Third, cloud computing as a semi-trusted system poses a threat to data confidentiality and data privacy. Different from cloud computing, edge computing [123-126] as a decentralized paradigm is proposed to support real-time services, location awareness and mobility, which is appropriate for various real-time application scenarios, such as smart grid and smart traffic [127], etc. Data fusion in IoT applications always needs to be real-time, which can offer quick response to some specific cases (e.g., emergent healthcare cases). Thus, edge computing may be a good choice to balance the efficiency and timeliness of IoT data fusion [128-130]. In addition, the deployment of non-colluding and distributed edge nodes as gateways in smart home can also be a possible method to help preserve user privacy.

- Privacy-preserving data fusion
  
  Privacy-preserving data fusion is still a hot topic in the near future. Privacy-friendly sensors cannot provide enough information for satisfying application design goals. How to guarantee user privacy is still an open issue and needs deep research. In smart grid, state measurement is more sensitive than pricing information or control policy. Homomorphic encryption can be used for privacy-preserving data analytics and machine learning, but its high computation cost makes it difficult to be employed in real-time application scenarios. In a highly distributed system, secret-sharing based schemes may be a possible solution to solve this issue by fusing data in clusters with reduced communication cost. But the literature still requests more advanced solutions that can be widely deployed in practice. Without any doubt, privacy-preserving data fusion is an interesting and challenging research topic.

- Adaptive sensor selection and flexible data fusion
  
  Adaptive sensor selection and flexible data fusion would be a real demand and interesting research topic in the research field of IoT data fusion. As discussed above, the adoption of motions sensors has a low accuracy of sitting detection but suitable for activity recognition, while environmental sensors can achieve highly accurate occupancy detection. If occupancy for lighting management is detected, the activity recognition is needed for fall detection or other purposes. In smart transportation [43],
the spot detection accuracy even decreases when the Gabor features is involved. Generally, the data chosen for fusion is really an issue. Moreover, how to adaptively select suitable sensors and flexible perform data fusion based on the selection is a practical research topic.

- Access control over fused results

Data fusion algorithms should offer the properties to support access control over fused results when performing data fusion in a privacy-preserving way. Apart from real-time data fusion, there also exists off-line fusion, which does not need quick response [131]. Data fusion extracts meaningful information from multi-modality and massive data, which provides an intuitive view over a target object. Due to privacy and security concern, secure and flexible access control over fusion result for multi-user access would be an interesting research topic [132, 133]. For example, a fine-grained access control scheme based on specific attributes is highly needed since it can preserve the privacy of inhabitants and reduce the risk of emergence by sharing the fusion result to authorized users. This kind of schemes are highly requested in the smart home and other IoT applications. However, existing works mainly focus on the research of data fusion or data access, but they ignore the issue of access control over data fusion results which needs further research.

- Verifiable data fusion

Verifiable data fusion is a novel and challenging research issue, which can help tracing back the source of fusion problem and locating attacks on data fusion. As a fusion center or node may not be fully trusted by service consumers, besides the reliability of fusion algorithm, whether the fusion center acts honestly is also an issue. Hence, we need to ensure the correctness of data fusion, especially when security and privacy should be considered in the data fusion. The reliability of data fusion would be a new issue. How to motivate the fusion center to act honestly becomes interesting and significant. Verifiable computation has been deeply researched in cloud computing [134, 135], but big data in IoT makes it a difficult problem. Some research has been done in data aggregation that is one of the most basic algorithms in data fusion [136] by using aggregate signature. Homomorphic signature [92] is also an efficient technique to verify the aggregation. But fully homomorphic computation has a higher computation overhead than partially homomorphic computation, which is inefficient for IoT. So far, it is still unable to check the correctness of complex fusion algorithm. Verifiable data fusion is a significant research topic in terms of data fusion performed at a third party. This topic could attract special attention in future research, especially in the context of distributed IoT systems.

5 Conclusions

In this paper, we give a thorough review on the data fusion technologies in IoT, especially focusing on three main IoT application domains: smart home, smart grid and smart transportation. We proposed a number of data fusion requirements and reviewed the existing data fusion methods in the above three main IoT domains by employing the requirements to study their pros and cons. Based on our survey and discussion, we figure out a number of open research issues and further highlight
promising future research directions to guide future research towards secure and privacy-preserving data fusion in IoT.

Acknowledgment

This work is sponsored by National Postdoctoral Program for Innovative Talents (grant BX20180238), the NSFC (grants 61672410, 61802293 and U1536202), the Academy of Finland (grant 308087), the National Key Research and Development Program of China (grant 2016YFB0800704), the Project Supported by Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2016ZDJC-06), the Key Lab of Information Network Security, Ministry of Public Security under grant No. C18614, the Project funded by China Postdoctoral Science Foundation (grant 2018M633461), and the China 111 project (grants B16037 and B08038).

References


Mechatronics, 8 (2007) 53-60.


2007, pp. 208-212.


[78] D.J. Cook, Learning Setting-Generalized Activity Models for Smart Spaces, IEEE


[87] P.Y. Chen, S. Yang, J.A. McCann, J. Lin, X. Yang, Detection of false data injection attacks


[95] C.I. Fan, S.Y. Huang, Y.L. Lai, Privacy-Enhanced Data Aggregation Scheme Against


[111] Q.-J. Kong, Y. Chen, Y. Liu, A fusion-based system for road-network traffic state surveillance: a case study of Shanghai, IEEE Intelligent Transportation Systems


[117] M. Schneider, T. Schneider, Notes on non-interactive secure comparison in "image feature extraction in the encrypted domain with privacy-preserving SIFT", in, 2014, pp. 135-140.


