Adaptive and Cost-Effective Collection of High-Quality Data for Critical Infrastructure and Emergency Management in Smart Cities—Framework and Challenges

ELISA BERTINO and MOHAMMAD R. JAHANSHAHI, Purdue University, Indiana

CCS Concepts: • **Information systems → Mobile information processing systems**;

Additional Key Words and Phrases: Civil engineering, edge computing, device swarms

ACM Reference format:
Elisa Bertino and Mohammad R. Jahanshahi. 2018. Adaptive and Cost-Effective Collection of High-Quality Data for Critical Infrastructure and Emergency Management in Smart Cities—Framework and Challenges. *J. Data and Information Quality* 10, 1, Article 1 (May 2018), 6 pages.
https://doi.org/10.1145/3190579

1 INTRODUCTION

In future smart cities, many decision processes in critical infrastructure and emergency management will be based on machine learning techniques. One particular application will be the processing of large datasets of visual images for defect assessment where the data is collected by a swarm of mobile sensing agents (e.g., unmanned aerial vehicles). In this context, examples of defective regions are corrosion and cracks in buildings and facilities [1], and potholes on roads. A critical requirement for the success of such assessment processes is the reliable detection, quantification, and localization of defective regions. Furthermore, in such applications, the real-time assessment is often critical so that the swarm can decide regarding the optimum strategy and corresponding actions for effective data collection in unknown environments (e.g., robots that will be used for earthquake reconnaissance and rescue where they enter buildings whose plan is unknown to the robot). On the other hand, the reliability of the assessments requires data of good quality, since poor data may negatively affect the accuracy of classification and predictions, and consequently, may introduce additional costs and time overhead.

In general, acquiring such data and making sure that the data is of high quality, especially for real-time decisions, is expensive due to difficulty of reaching the regions where the objects of interest are located and the need for human-intensive assessment. However, today we have many technologies that can be leveraged to devise effective and inexpensive solutions, including: deep neural networks for image analysis; image processing techniques; mobile image data acquisition

This work was partially supported by the NSF Award IIS-1636891 "BD Spokes: Planning: MIDWEST: Cyberinfrastructure to Enhance Data Quality and Support Reproducible Results in Sensor Originated Big Data."
Authors' addresses: E. Bertino, CS Dept., Purdue University, West Lafayette, IN, 47907; email: bertino@purdue.edu; M. R. Jahannshahi, Civil Engineering School, Purdue University, West Lafayette, IN, 47907; email: jahansha@purdue.edu.

https://doi.org/10.1145/3190579

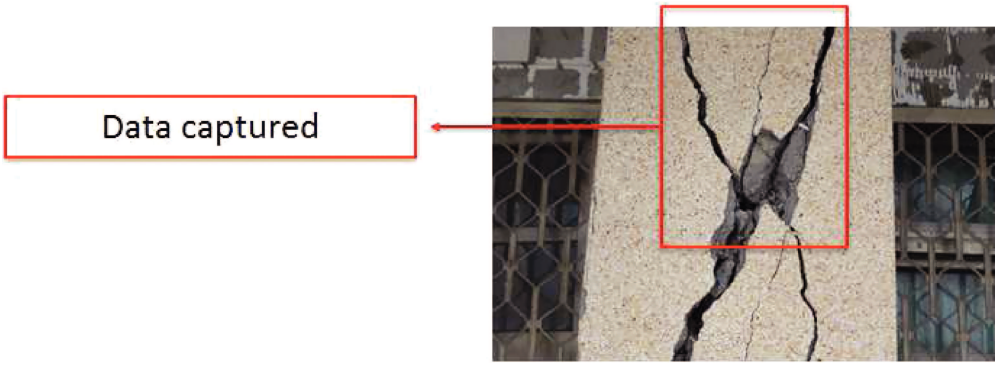

Fig. 1. A spatially incomplete object.

agents (mobile phones, small drones, robots, sensors); 5G networks and edge computing processing [2]; crowdsourcing.

In this article, we first briefly discuss relevant data quality requirements related to applications in the area of critical infrastructure and emergency management, although this framework can be extended to other applications. We then present a comprehensive framework for a real-time, adaptive, and cost-effective collection of high-quality data for such applications that leverage many of the above technologies, and elaborate on a few research challenges.

2 DATA QUALITY REQUIREMENTS

Data quality is usually characterized by many different dimensions [3]. In our context, e.g., objects extracted from image data, key requirements include:

- Spatial Completeness: The objects of interest should be "fully covered" by the image data. For example, an image reporting only half of a building crack would not have satisfactory spatial completeness (see Figure 1 for an example of a spatially incomplete object).
- Temporal Completeness: The temporal evolution of the objects of interest should be covered as it is critical for accurate prediction.
- Precision: The object images should be sharp and have high resolution.
- Traceability: Information about the entire process, according to which data of interest was collected, processed, and transmitted, should be recorded; this is critical for identifying errors that lead to poor quality data about the objects of interest.
- Minimality: The presence of non-relevant objects should be minimized.

It is, however, important to remark that other quality requirements, such as currentness and consistency, are also relevant in our context.

3 DATA COLLECTION FRAMEWORK

Our framework (see Figure 2) is based on two conceptual parties: *data collection coordinator* (referred to as base station (BS)); and *data collectors* (e.g., agents in charge of data gathering). The data collection coordinator is the interface system that coordinates the data acquisition tasks and data quality assessment. It interfaces on one side with the data users (e.g., end-users and applications) and on the other with data collectors. Given a data acquisition task and geographical area of interest, it allocates a number of data collectors, based on the capabilities of collectors, for the execution of the task, by also trying to optimize the cost of data acquisition and minimize

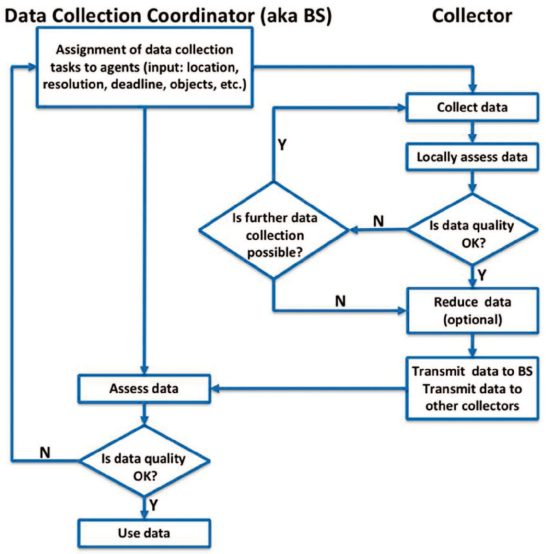

Fig. 2. Data collection framework.

the response time. Such allocation decisions can be basically supported by optimization techniques developed in the area of operation research. The main challenge is to determine the most suitable optimization techniques for dynamic contexts. The data collection coordinator must also assess the quality of data with respect to specific quality requirements provided as input by the data users. Since a data collection task may often be split among data collectors, the coordinator may have to integrate the various collected data to see whether, overall, the data meets the specified quality requirements or not. The coordinator may also support data enrichment, for example, by using GIS data [5] and data linkage with other sources. The data collectors carry out the basic tasks of collecting data, assessing the quality of the collected data, and, based on this assessment, collecting more data. Notice that data collectors may have different capabilities. For example, some collectors may have equipment for very high-resolution imagery with powerful computing capabilities and can run machine learning tools that require large storage size and GPU. These collectors may thus be able to perform a high accurate data quality analysis. On the other hand, other collectors are very small and thus can easily move very close to the objects and take images from very short distances; however, their capability for data quality assessment may be very limited. Finally, other collectors may be equipped with mechanical devices to take samples from the environment, such as a sample of soil or water, or perform active testing through injecting dynamic disturbances by the collector actuators at selected locations (e.g., exciting the structure by a hammer and collecting the propagated wave characteristics for damage detection). The decision about the right combinations of data collectors for a data acquisition task is taken by the data collection coordinator based on the knowledge of the capabilities of each data collection device. However, as research in the area of distributed decision making for autonomous systems progresses, such decisions could be even taken autonomously by swarms of data collectors.

Our framework is based on the notion of *data collection cycle*, which is organized according to a continuous loop consisting of two main phases: (a) data collection; (b) data quality assessment. Once data is collected, it is assessed for quality. If quality is insufficient, further data is collected.

Further data collection is typically tailored to improve the quality. For example, a data collector may be required to collect data of higher resolution for a specific object.

Data quality assessment is executed at three levels: (1) locally at the data collector; (2) collaboratively within the data collector swarms; (3) globally at the data collection coordinator. Assessments (1) and (2) may not always be possible. Assessment (1) may not be possible as the data collector may not have the capabilities to assess data. Assessment (2) may not be possible if the swarm does not have capabilities to assess the data or if a data collector is isolated from the rest of the swarm. However, Assessment (2) may be highly desirable when connections with the BS are fragmented/unreliable. Adaptation capabilities are thus crucial to deal with those situations. *It is important to notice that a critical challenge is to develop approaches for automatically assessing the quality of collected data and automatically determining which additional data needs to be further collected to refine/complete/enhance the quality of the initial data.* In particular, when data collection is performed by a swarm of data collectors, the swarm should automatically assess the data and decide further data collection.

The development of such framework requires addressing several challenges:

— *Optimized data-quality driven allocation of data collection tasks to agents*: Data collectors are typically heterogeneous with respect to hardware and software capabilities and with respect to special equipment for data acquisition—for example, a drone may have equipment for acquiring images at very high resolution. Also, data collectors may be located in different geographical regions. Data collection also depends on the quality requirements; for example, when performing an initial assessment, data of low quality may be fine. Therefore, it is important to design approaches that are able to support the optimal allocation of data acquisition tasks based on different constraints, requirements, and data collectors' capabilities and status. Furthermore, it is important that each data collector has the capability of autonomously deciding which data to collect based on its own local assessment of data that have already been collected. Thus, the allocation of data collection tasks is a combination of centralized decisions with decisions local to the data collectors and/or data collector swarms.

— *Automatic (collaborative) data quality assessment*: Techniques are needed to automatically assess the quality of the collected data with respect to the specific quality requirements. Techniques based on machine learning are relevant here. The main issue is that such assessment may be carried out at three different levels (see previous section) and thus tradeoff may be needed between accuracy and resource usage. For example, at the level of the data collectors, resource usage should be minimized. However, resource use minimization may lead to less accurate decisions. It is also critical to devise approaches by which such assessments can be carried out by data collector swarms. Finally, for assessments to be carried out at the BS level, it is important to assess the "optimal data transmission strategy," namely whether the bulk data should be sent from the data collectors to the BS, or whether the data collectors should perform some local data reduction and then send the reduced data based on the desired tradeoff between accuracy, communication costs, and data collectors' resource usage. We use here the term data reduction with a broad meaning to indicate techniques to reduce the amount of data to be transmitted. Examples of such techniques include extracting features from images and sending only these features, discarding images that do not include objects of interest, discarding images of poor quality, and selecting relevant frames from videos. Data reduction is important when the computation, memory, power, and transmission bandwidth constraints of the data collectors are considered, particularly

for large infrastructure systems, such as dams, where archiving of the whole raw data is not a viable and efficient solution.

— *Automatic (collaborative) specification of data to be further acquired*: Techniques are needed to support an automatic generation of the specification of data to be further acquired; examples include acquisition of data at finer resolution and at different angles and acquisition of missing portions of objects of interest. A language must be devised according to which such specifications are encoded and also algorithms must be designed to automatically generate such specifications based on the analysis of previously acquired data as well as the specific data quality requirements.

4 RELATED WORK

There is a large body of work focusing on the use of big data technologies and machine learning algorithms for critical infrastructure and emergency management [1, 4–7]. However, most previous approaches do not address the problem of data quality in its many data dimensions. They typically focus on other issues such as modeling very complex domains and assessing the performance of different machine learning algorithms for specific application domains. Perhaps, the work that is more closely related to our goal of enhancing data quality is by Cervone et al. [8]. They recognize that data provided by remote sensing techniques, such as Landsat data, may have geo-temporal gaps and, thus, suggest the use of unmanned aerial vehicles (UAVs) and volunteered data collected by users to fill in these gaps. Such suggestions align with our framework in that we also suggest the use of mobile devices, such as UAVs, and crowdsourcing approaches for data collection. However, such previous work does not include frameworks or techniques that would make it possible for devices to autonomously decide which additional data to collect to enhance the data quality. They mention, however, that there is a need to triage and optimize inspections by humans or tasks for additional data collection. Our goal is exactly to automatically decide about inspections and additional data collection tasks, and possibly have these inspections and tasks being autonomously executed by devices.

5 CONCLUDING REMARKS

We have outlined a framework for dynamic and adaptive data acquisition aimed at applications in the area of critical infrastructure and emergency management. Our proposed framework is particularly suited for such applications because many decisions in these applications will be increasingly based on the use of big data and machine learning algorithms. However, the availability of detailed data of good quality is critical. At the same time, the acquisition of good quality data should not be expensive and, in particular, it should not require high human operator costs. Our framework addresses exactly such requirements and presents a vision and its related challenges concerning how to push the state-of-the-art techniques so that devices can autonomously decide further data acquisition in order to enhance data quality. In emergency management situations, the availability of a system based on our proposed framework would be very critical as in these situations human resources may be scarce and emergency management decisions often need to be taken in a very short time—for example, in the case of a rapidly spreading forest fire with high variability due to changing winds. In addition, from a technical point of view, our framework would be inexpensive to deploy and could easily use an aerial mobile ad-hoc network to rapidly transmit the collected data by devices autonomously. It could use devices specialized for exploring hazardous areas for humans (e.g., Fukushima Daiichi nuclear disaster in 2011). In many emergency situations, our framework could also be easily extended to recognize humans in danger and rapidly assess the gravity of the danger, especially if devices are equipped with sophisticated techniques, such as transfer learning approaches, for rapid training of neural networks to recognize new objects.

We are currently investigating the tradeoff concerning analytics on the edge compared with centralized approaches. This is an important first step toward addressing the challenges outlined in the article.

REFERENCES

[1] F. C. Chen and M. R. Jahanshahi. 2017. NB-CNN: Deep learning-based crack detection using convolutional neural network and naïve Bayes data fusion. *IEEE Transactions on Industrial Electronics* 6, 5, 4392–4400.

[2] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu. 2016. Edge computing: Vision and challenges. *IEEE Internet of Things Journal* 3, 5, 637–646.

[3] C. Batini and M. Scannapieco. 2016. *Data and Information Quality: Dimensions, Principles and Techniques*. Springer.

[4] A. Giretti, A. Carbonari, and B. Naticchia. 2012. A spatio-temporal Bayesian network for adaptive risk management in territorial emergency response operations. In *Bayesian Networks*, Wichian Premchaiswadi (Ed). Retrieved from https://www.intechopen.com/books/bayesian-networks/a-spatio-temporal-bayesian-network-for-adaptive-risk-management-in-territorial-emergency-response-op.

[5] M. Khouj, C. López, S. Sarkaria, and J. Marti. 2011. Disaster management in real time simulation using machine learning. In *Proceedings of the 24th Canadian Conference on Electrical and Computer Engineering (CCECE'11)*. 001507–001510. DOI: 10.1109/CCECE.2011.6030716

[6] C. Yang, G. Su, and J. Chen. 2017. Using big data to enhance crisis response and disaster resilience for a smart city. In *Proceedings of the IEEE 2nd International Conference on Big Data Analysis (ICBDA'17)*. 504–507. DOI: 10.1109/ICBDA.2017.8078684

[7] S. Lee, L. Chen, S. Duan, S. Chinthavali, M. Shankar, and B. A. Prakash. 2016. URBAN-NET: A network-based infrastructure monitoring and analysis system for emergency management and public safety. In *Proceedings of the IEEE International Conference on Big Data (Big Data'16)*. 2600–2609. DOI: 10.1109/BigData.2016.7840902

[8] G. Cervone, E. Schnebele, N. Waters, M. Moccaldi, and R. Sicignano. 2017. Using social media and satellite data for damage assessment in urban areas during emergencies. In *Seeing Cities through Big Data*, P. Thakuriah et al. (Ed.). Springer Geography.

Received December 2017; revised February 2018; accepted February 2018