Information Fusion 36 (2017) 208–224

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Sheaves are the canonical data structure for sensor integration

Michael Robinson

Mathematics and Statistics, American University, Washington, DC, USA

ARTICLE INFO

Article history:
Received 4 March 2016
Revised 1 December 2016
Accepted 4 December 2016
Available online 5 December 2016

Keywords:
Sheaf
Sensor integration
Mosaic
Semantic fusion
Heterogeneous data source
Cohomology

ABSTRACT

A sensor integration framework should be sufficiently general to accurately represent many sensor modalities, and also be able to summarize information in a faithful way that emphasizes important, actionable information. Few approaches adequately address these two discordant requirements. The purpose of this expository paper is to explain why sheaves are the canonical data structure for sensor integration and how the mathematics of sheaves satisfies our two requirements. We outline some of the powerful inferential tools that are not available to other representational frameworks.

1. Introduction

There is increasing concern within the data processing community about "swimming in sensors and drowning in data," [1] because unifying data across many disparate sources is difficult. This refrain is repeated throughout many scientific disciplines, because there are few treatments of unified models for complex phenomena and it is difficult to infer these models from heterogeneous data.

A sensor integration framework should be (1) sufficiently general to accurately represent all sensors of interest, and also (2) be able to summarize information in a faithful way that emphasizes important, actionable information. Few approaches adequately address these two discordant requirements. Models of specific phenomena fail when multiple sensor types are combined into a complex network, because they cannot assemble a global picture consistently. Bayesian or network theory tolerate more sensor types, but suffer a severe lack of sophisticated analytic tools.

The mathematics of *sheaves* partially addresses our two requirements and provides several powerful inferential tools that are not available to other representational frameworks. This article presents (1) a sensor-agnostic measure of data consistency, (2) a sensor-agnostic optimization method to fuse data, and (3) sensor-agnostic detection of possible "systemic blind spots." In this article, we show that sheaves provide both theoretical and practical tools to allow representations of locally valid datasets in which the datatypes and contexts vary. Sheaves therefore provide a common, canonical language for heterogeneous datasets. We show that sheaf-based fusion methods can combine disparate sensor modalities to dramatically improve target localization accuracy over traditional methods in a series of examples, culminating in Example 24. Sheaves provide a sensor-agnostic basis for identifying when information may be inadvertently lost through processing, which we demonstrate computationally in Examples 30–32.

Other methods typically aggregate data either exclusively globally (on the level of whole semantic ontologies) or exclusively locally (through various maximum likelihood stages). This limits the kind of inferences that can be made by these approaches. For instance, the data association problem in tracking frustrates local approaches (such as those based on optimal state estimation) and remains essentially unsolved in the general case. The analysis of sheaves avoids both of these extremes by specifying information where it occurs – locally at each data source – and then uses global relationships to constrain how these data are interpreted.

The foundational and canonical nature of sheaves means that existing approaches that address aspects of the sensor integration problem space already illuminate some portion of sheaf theory without exploiting its full potential. In contrast to the generality that is naturally present in sheaves, existing approaches to combining heterogeneous quantitative and qualitative data tend to rely on specific domain knowledge on small collections of data sources. Even in the most limited setting of pairs quantitative data sources, considerable effort has been expended in developing models of their joint behavior. Our approach leverages these pairwise models into models of multiple-source interaction. Additionally, where joint models are not yet available, sheaf theory provides a context for understanding the properties that such a model must have.

E-mail address: michaelr@american.edu

http://dx.doi.org/10.1016/j.inffus.2016.12.002

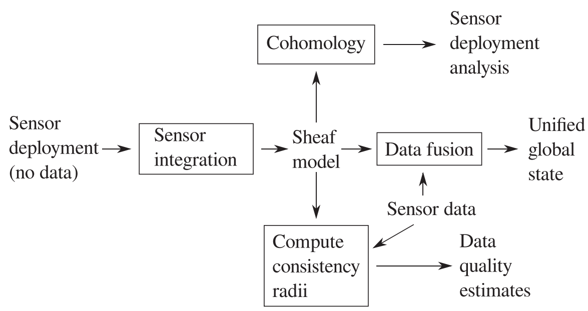

Fig. 1. The sheaf-based multi-sensor workflow.

1.1. Contribution

Sometimes data fusion is described methodologically, as it is in the Joint Directors of Laboratories (JDL) model [2,3], which defines operationally-relevant "levels" of data fusion. A more permissive – and more theoretically useful – definition was given by Wald [4], "data fusion is a formal framework in which are expressed the means and tools for the alliance of data originating from different sources." This article shows that this is *precisely* what is provided by sheaf theory, by carefully outlining the requisite background in sheaves alongside detailed examples of collections of sensors and their data. In Chapter 3 of Wald's book [5], the description of the implications of this definition is strikingly similar to what is presented in this article. But we go further, showing that we obtain not just a full-fidelity representation of the data and its internal relationships, but also a jumping-off point for analysis of both the integrated sensor system and the data it produces.

We make the distinction between "sensor integration" in which a collection of *models* of sensors are combined into a single unified model (the *integrated sensor system*), and "sensor data fusion" in which *observations* from those sensors are combined. This article presents the sheaf-based workflow outlined by Fig. 1, which divides the process of working with sensors and their data into several distinct stages:

1. *Sensor integration* unifies models of the individual sensors and their inter-relations into a single system model, which we will show has the mathematical structure of a *sheaf*. We emphasize that no sensor *data* are included in the sensor system model represented by the sheaf.
2. *Consistency radius computation* uses the sheaf to quantify the level of self-consistency of a set of data supplied by the sensors.
3. *Data fusion* takes the sheaf and data from the sensors to obtain a new dataset that is consistent across the sensor system through a constrained optimization. The consistency radius of the original data places a lower bound on the amount of distortion incurred by the fusion process.
4. *Cohomology* of the sheaf detects possible problems that could arise within the data fusion process. Although cohomology does not make use of any sensor data, it quantifies the possible impact that certain aspects of the data may have on the fusion process.

Through a mixture of theory and detailed examples, we will show how this workflow presents solutions to four distinct problem domains:

1. Formalizing the description of an integrated sensor system as a well-defined mathematical entity (defined by Axioms 1–6 in Section 3),
2. Quantifying the internal consistency of the data from individual sensors (the *consistency radius* in Definition 20 in Section 4),
3. Deriving globally consistent data from the data provided by the entire collection of sensors (solving Problem 19, *sheaf-based data fusion* in Section 4), and
4. Measuring the impact of the relationships between sensors on what data will be deemed consistent (using *cohomology* in Section 5).

1.2. Historical context

There are essentially two threads of research in the literature on data fusion: (1) physical sensors that return structured, numerical data ("hard" fusion), and (2) unstructured, semantic data ("soft" fusion) [6]. Hard fusion is generally performed on spatially-referenced data (for example, images), while soft fusion is generally referenced to a common ontology. Especially in hard fusion, the spatial references are taken to be absolute. Most *sensor* fusion is performed on a pixel-by-pixel basis amongst sensors of the same physical modality (for instance [7–9]). It generally requires image registration [10] as a precondition, especially if the sensors have different modalities (for instance [11,12]). When image extents overlap but are not coincident, *mosaics* are the resulting fused product. These are typically based on pixel- or patch-matching methods (for instance [13]). Because these methods look for areas of close agreement, they are inherently sensitive to differences in physical modality. It can be difficult to extend these ideas to heterogeneous collections of sensors.

Like hard fusion, soft fusion requires registration amongst different data sources. However, since there is no physically-apparent "coordinate system," soft fusion must proceed without one. There are a number of approaches to align disparate ontologies into a common ontology [14–17], against which analysis can proceed. There, most of the analysis derives from a combination of tools from natural language processing (for instance [18–20]) and artificial intelligence (like the methods discussed in [21–23]). That these approaches derive from theoretical logic, type theory, and relational algebras is indicative of deeper mathematical foundations. These three topics have roots in the same category theoretic machinery that drives the sheaf theory discussed in this article.

Weaving the two threads of hard and soft fusion is difficult at best, and is usually approached statistically, as discussed in [24–26]. Unfortunately, this "clouds" the issue. If a stochastic data source is combined with another data source (deterministic or not), stochastic analysis asserts that the result will be stochastic. This viewpoint is quite effective in multi-target tracking [27,28] or in event detection from social media feeds [29], where there are sufficient data to estimate probability distributions. But if two *deterministic* data sources are combined, one numeric and one textual, why should the result be *stochastic*?

Regardless of this conundrum, information theoretic or possibilistic approaches seem to be natural and popular candidates for performing sensor integration, for instance [30–32]. They are actually subsumed by category theory [33,34] and arise naturally when needed. These models tend to rely on the homogeneity of sensors in order to obtain strong theoretical results. Sheaf theory extends the reach of these methods by explaining that the most robust aspects of networks tend to be topological in nature. For example, one of the strengths of Bayesian methods is that strong convergence guarantees are available. However, when applied to data sources arranged in a feedback loop, Bayesian updates can converge to the wrong distribution or not at all! [35] The fact that this is possible is witnessed by the presence of a topological feature – a loop – in the relationships between sources.

Sheaf theory provides canonical computational tools that sit on the general framework and has been occasionally [36–38] used in applications. Unfortunately, it is often difficult to get the general machine into a systematic computational tool. Combinatorial ad-