Title

Quality of Information Driven Environment Crowdsourcing and its Impact on Personal Wellness Applications

Permalink

https://escholarship.org/uc/item/0pb6g34m

Author

Matthews, Jerrid E.

Publication Date

2014

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Quality of Information Driven Environment Crowdsourcing and its Impact on Personal Wellness Applications

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Jerrid E. Matthews

2014

Abstract of the Dissertation

Quality of Information Driven Environment Crowdsourcing and its Impact on Personal Wellness Applications

by

Jerrid E. Matthews

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2014

Professor Mario Gerla, Chair

Mobile devices with programmable embedded sensors and internet access have enabled a new paradigm of socially beneficial software applications. These devices may be stationary or mobile, and located sparsely across the globe operating under heterogeneous environments. These multi-lateral sensor data feeds produced by both autonomous and human sensing agents can be aggregated and transformed by a system to produce a human understandable spatiotemporal representation of a phenomenon (ie: event) in real-time. These data can then be disseminated using many different communication infrastructures (e.g. 4GLTE, WiFi). The study of how to efficiently organize these complex sensor data feeds is the primary contribution of this dissertation; in addition we present two health and wellness sensor data applications that leverage the sensor data feeds. Traditional sensor data platforms require the data publisher to associate a set of descriptive terms (also known as keywords or tags) with their data feed in order to organize the sensor data. Information operators must perform a keyword-based search in order to retrieve the data feeds of interest, which may require subject matter expertise to identify relevant keywords. The central theme of this thesis is the leveraging of personal sensor platforms, internet computing resources and crowdsourcing campaigns to achieve not only individual wellness but also community health maintenance. We contribute a new ontological data model for organizing and enriching sensor data with valuable QoI/VoI attributes. In addition, we combine theoretical models and systematic

measurements to show that it is possible to organize sensor data in such a way to retrieve relevant sensor data in order to measure a phenomenon of interest without tagging or human input.

The dissertation of Jerrid E. Matthews is approved.

Mani Srivastava

Rick Schoenberg

Alfonso Cardenas

Mario Gerla, Committee Chair

University of California, Los Angeles

2014

To my Lord and savior who imparted a talent in me and set me on a course to discover and grow. To my family and friends who have supported me through this endeavor. I would like to express my sincere gratitude to my advisor, Professor Mario Gerla, for his invaluable guidance and continuous support throughout my years at UCLA. My great appreciation goes to my other committee members as well. It is truly an honor to be under the academic lineage of the foremost internet pioneers, and a member of the Network Research Lab at UCLA.

Table of Contents

List of Figures

List of Tables

Acknowledgments

To my Lord and Savior who imparted a talent in me and set me on a course to discover and grow that talent. I would like to express my sincere gratitude to my advisor, Professor Mario Gerla, for his invaluable guidance and continuous support throughout my years at UCLA. He has been a strong motivator and advocate for my research throughout my matriculation through the Ph.D. program. His persistent pursuit of perfection and deep insight into various subjects has always been an inspiration to me. He has taught me show to conduct qualitative research and how to improve my writing and presentation skills.

My great appreciation goes to my other committee members as well: Professor Mani Srivastava, Professor Alfonso Cardenas, and Professor Rick Schoenberg. I thank them for kindly agreeing to be on my doctoral committee and for their helpful advice and suggestions.

It is truly an honor to be under the academic lineage of the foremost internet pioneers, and a member of the Network Research Lab at UCLA. I would also like to personally thank my colleagues, especially Lord Cole, Joshua Joy, Fabio Angus, and Jorge Mena for their personal support and for the enlightening discussions we had about various research topics. My deepest gratitude goes to my parents for their love, their support, and for their sacrifice over so many years. They have been the origin of my strength and will always be.

Vita

December, 2006	B.S. Computer Science, Michigan State University, East Lansing, MI
June, 2009	M.S. Computer Science, University of California, Los Angeles, Los Angeles, CA
December, 2014	Ph. D. Candidate, Computer Science, University of California Los Angeles, Los Angeles, CA

Publications

Matthews, J., Javadi, F., Rane, G., Zheng, J., Pau, G., and Gerla, M., Ultraviolet Guardian - Real Time Ultraviolet Monitoring, *MobileHealth Workshop (Mobihoc 2012).*

Matthews, J, Kulkarni, R, Gerla, M., and Massey, T., Rapid Dengue and Outbreak Detection with Mobile Systems and Social Networks, *MONET 2010.*

Matthews, J, Kulkarni, R, Gerla, M., and Massey, T.,, PowerSense: Power Awareness in Dengue Diagnosis Mobile Application Moving Towards a Power Conscious Computing Framework, *mHealthSys Workshop (ACM SenSys 2011).*

Matthews J., Bisdikian C., Kaplan L., Pham T., Ontology-based Quality of Information Library for Sensor Data, *MDM 2010.*

Amini, N. Matthews J, Dabiri F., Vahdatpour A., Noshadi H, Sarrafzadeh M, A Wireless Embedded Device For Personalized Ultraviolet Monitoring, *Biodevices 2009*.

Matthews, J, Kulkarni, R, Gerla, M., and Massey, T., A Light-weight Solution for Real-Time Dengue Detection using Mobile Phones, *MobiCase 2009*.

Marcondes, C. Matthews, J. Chen, R. Sanadidi, M.Y. Gerla, M., A Cross-Comparison of Advanced TCP Protocols in High Speed and Satellite Environments, *ASMS 2008*.

CHAPTER 1

Introduction

1.1 Motivation for Sensor Networks

As with many technologies, defense applications have been the driver for research and development in sensor networks. During the Cold War, the Sound Surveillance System (SOSUS) [2], a system of acoustic sensors (hydrophones) on the ocean bottom were deployed at strategic locations to detect and track Soviet submarines. Over the years, more sophisticated acoustic sensing agents have been developed and SOSUS was made available to the research community, such as the National Oceanographic and Atmospheric Administration (NOAA) for monitoring seismic and animal activity in the ocean [3]. Also during the Vietnamese War, Operation Igloo White [4] was a covert operation to deploy thousands of Air Delivered Seismic Intrusion Devices (ADSID) over the Ho Chi Minh Trail in an attempt to cut off strategic enemy supply routes in the dense forest. The device could sense vertical earth motion by the use of an internal geophone and could determine whether a man or a vehicle was in motion at a range of 30 meters and 100 meters respectively [5]. ADSID enabled US forces to track enemy movements in Laos, which facilitated the interdiction of North Vietnamese Army supply lines and staging areas used to resupply the military campaign conducted in South Vietnam.

Distributed sensor networks were proven in-valuable for intelligence, surveillance, target acquisition, and reconnaissance (ISTAR) missions. Later on, researchers at the Defense Advanced Research Projects Agency (DARPA) began investigating whether network communication principles from the ARPANET could be applied to low-cost networked autonomously operated sensor nodes. Later on researchers began developing prototype distributed wire-

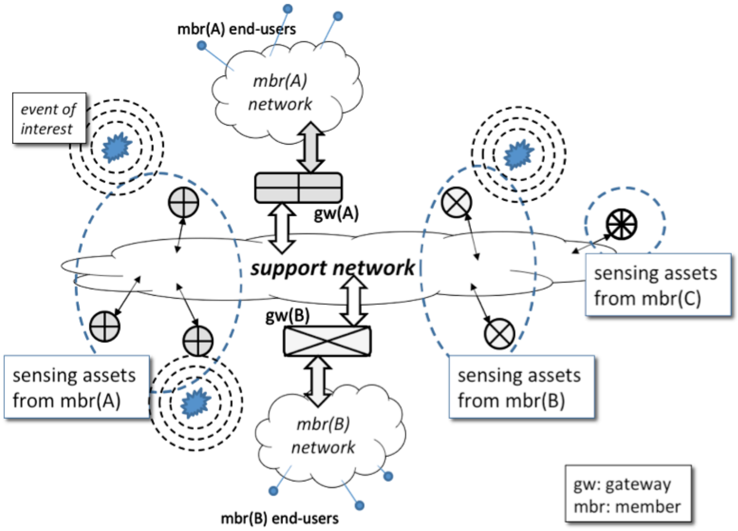


Figure 1.1: A sensor network enabled coalition use case.

less sensor networks (WSN's) and delay tolerant communication protocols [6–9]. During this period, the network of sensing agents that scientists deployed were analogous to modern wireless sensor networks, and the culmination of sensor data provided by these sensing agents improved the veracity of the reports that describe an event of interest provided by the fusion modules that process the sensor measurements. The more accurate information aided in better decision making and protected armed forces and allies stationed in high risk attack zones.

1.1.1 The Semantic Sensor Network

In modern society, a complex web of semantic sensor networks (SSN) and open sensor data publishing platforms have been erected to make sense of many worldly observable phenomena (ie: events) through information analytics and data fusion. Consider the use case outlined Figure 1.1. A collection of sensing resources belonging to a number of agencies (members) is deployed in a broad area of interest. These sensing assets monitor events of interest and feed their observations (directly or following various forms of information fusion) to end-users within these agencies. The sensor-originated information flows over a shared support (backbone) network, to which various agency networks couple. The use-case in the figure may represent any number of ad-hoc, or infrastructure-based cooperative, distributed sensor-enabled operations including disaster response situations, coalition military operations, and traffic monitoring services crossing multiple administrative domains. Sensor data applications require collection of data from both static and dynamic sensor networks. The devices that measure the required data must be capable of assessing these sensor data for relevancy to ensure effective operation of sensor-enabled operations and reliable business decisions.

In the next chapter, we discuss the legacy ontology based knowledge representation languages for characterizing sensor data, and propose a new ontology based data classification model that allows users to construct a sensor knowledge base that maps a phenomenon of interest (eg: a spatiotemporal event) to the appropriate set of sensors that are qualified to provide information about the phenomenon.

1.1.2 Contributions

This paper studies one of the many challenges in architecting software systems to deal with quality of information (QoI) evaluation operations for these systems in a reproducible fashion. The sensing assets under consideration are wireless sensor networks (WSNs), comprising a large number of low-cost untethered, yet interconnected, sensing nodes of (relatively) limited computing, communication, and energy resources. They form the basis for a broad set of smart applications, either smartening existing ones (such as building automation and environmental control) or enabling entirely new applications, such as low-overhead remote monitoring of habitat, farming fields, traffic conditions, etc. Assessing quality and the value of an information product when there is no a priori knowledge of the information sources that application will provide is indeed a challenge.

This is a crucial design challenge for the software system designer and developer that ideally would like to design and develop a system that can be easily copied and deployed in many occasions with as few customization adjustments to the core system code as possible. The latter system design comes at a cost and is notorious for being bug-prone. To achieve the above, a system development framework will be needed that remains valid from one deployment instance to another around which the core system code can be designed. In this paper, we study one aspect of such a design framework with particular interest in the design of a new ontology based data model for semantic sensor web (SSW) queries that classifies the many different types of sensing agents qualified to provide information about a phenomenon of interest (eg: a spatiotemporal event).

This thesis makes three principal contributions:

Contribution 1: *New Ontology Based Data Model for SSW Queries that classifies the many different types of sensing agents qualified to provide information about a phenomenon of interest.* We have extended the contributions of Bisdikian et. al. [10, 11] by developing an ontology based classification model for sensing agents and sensor data fusion platforms to enrich their sensor data with information about the data source and metrics describing the accuracy, precision, completeness, timeliness, reliability, and utility of sensor data.

Contribution 2: *Sensite: A knowledge-base Web Platform for SSW queries.* We have developed an open sensor data platform that enables users to upload an arbitrary sensor data annotated ontology based data model. Users can perform a spatiotemporal query for a phenomenon of interest (ie: rain) at a geographic coordinate (latitude, longitude) at a certain time and Sensite will return all relevant sensor information (which may consist of one or more sensor types) capable of providing information about the phenomenon of interest (whether directly or though sensor fusion of multiple sensor types). Conventional sensor data platforms such as OpenRTMS [12], Google Cloud Platform [13], SAFECAST [14] use a non-restricted keyword tagging based method that enables publishers to annotate keywords that describe information about the sensor data and/or the phenomenon being observed. The drawback to this approach is that the collection of tags may not provide a comprehensive set of contextually related classifiers that describe the heterogeneous environmental contexts that the sensor data may be applicable to providing information about. To our knowledge this is the first application of its kind.

Contribution 3: *Ultraviolet (UV) Guardian: A mobile application that leverages local crowdsourced UV irradiance information for pedestrian UV exposure estimation.* We have developed a mobile application that provides recommendations to protect Sportspersons from Sun over-exposure and gives recommendations for sunlight benefits, such as Vitamin D. This novel mobile application leverages environmental context sensor information and crowdsourced UV irradiance information (provided by institutions and other UV Guardian users) through Sensite in order to predict the amount of Sun exposure that the user will receive without requiring the user to wear a UV sensor. To our knowledge this is the also first application of its kind.

Contribution 4: *Dengue Detector Mobile Application (DDMA):* A mobile application that aims to provide a rapid and economical dengue diagnosis solution for territories unable to afford expensive dengue diagnosis testing kits. Moreover, DDMA is designed with Physicians and the Center for Disease Control (CDC) in mind to enable them to track dengue outbreaks in real-time through a website that queries our Sensite platform for crowdsourced dengue diagnosis test results uploaded by DDMA. DDMA aims to improve the quality of life in developing countries by providing disease diagnosis and surveillance on-site rather than waiting a few days with the conventional dengue diagnosis kits.

CHAPTER 2

Ontology Based Model for Assessing the QoI/VoI of Sensor Data

2.1 Introduction

Intelligence, Surveillance and Reconnaissance (ISR) networks, accurately assessing the quality of sensor information is key to making better decisions. In semantic web technology, ontologies continue to play a major role in the organization of sensor information to decide a course of action to take pertaining to data received from a source. In this chapter, we study how an ontology can be used to capture information about the many different types of sensing agents that are capable of relaying data about worldly phenomenons. We will discuss the design of our ontology based data model for assessing the quality and value of information (QoI and VoI) of sensor data and the data source publishing the information. Next, we will describe our implementation of an extensible software library that semantic sensor web (SSW) platform software developers can use to associate QoI error analysis algorithms that aid in assessing the QoI produced by data sources operating under heterogeneous environments in an ISR army simulation application. In chapter 3, we discuss how the proposed data model is applied to our Sensite Knowledgebase sensor data web platform for SSW data queries. Finally, the following chapters will describe novel applications for health and wellness monitoring that leverage crowdsourced sensor data provided by the Sensite Knowledgebase application.

2.2 Defining an Ontology based QoI/VoI Data Model

Traditionally, WSNs are designed, deployed and operate in rather “closed” set-ups, where WSNs are intimately tied to their applications. However, we envisage that a natural evolution to openness for their design, deployment, and operational architectures will come to prevalence (or at least attain significant penetration) as it has happened with many other distributed computing and communication technologies. Such openness will permit the economic reuse of sensing resources by multiple applications and facilitate the timely deployment of both smart sensing systems and even smarter sensor enabled applications that may search, select and bind (and unbind) dynamically to sensing systems that can best support their current information needs. Information needs relate to the what, where, and when properties of required information. Information providers (representing the who and how of the information sources) are selected to best match these. What “best” could be is characterized and assessed via the QoI and VoI of the desired information pieces. The distinction between QoI and VoI is necessitated from the fact that the same piece of information, e.g., an image which has given quality characteristics (e.g., resolution, area it covers, time it was taken, owner, etc.) may have many different uses and bring different value to each of these uses.

Consider the use-case outlined Figure 1.1, where a collection of sensing resources belonging to a number of agencies (members) is deployed in a broad area of interest. The sensing assets monitor events of interest and feed their observations (directly or following various forms of information fusion) to end-users within these agencies. The sensor originated information flows over a shared support (backbone) network, to which various agency networks couple. The use-case in the figure may represent any number of ad-hoc, or infrastructure based cooperative, distributed sensor enabled operations including disaster response situations, coalition military operations, and traffic monitoring services crossing multiple administrative domains.

2.3 Conventional Ontology Data Models for Sensor Data

Ontologies play an important role in the representation and organization of information for data retrieval. Semantic web languages such as Ontology Web Language (OWL) [15] and Suggested Upper Merged Ontology (SUMO) [16] attempt to define a high level taxonomic schema for defining terms for entities and their relationships. These ontological languages allow users to define terminologies that describe entities and their relationships. Research related to sensor networks and the semantic web has focused on using an ontology to define sensor instances, relationships between sensors in sensor networks, and organization of sensor data. OWL added another level of flexibility by allowing the user to define their own entities, taxonomies, and relationships. The SUMO schema is relatively specific to defining a schema for devices, and their relationships within the semantic web. However, these semantic representation languages fail to provide a solution that addresses the quality and value of the information attributes. Ontology based sensor metadata solutions also include SensorML, an XML-based language that describes sensing platforms, data, and processes [17].

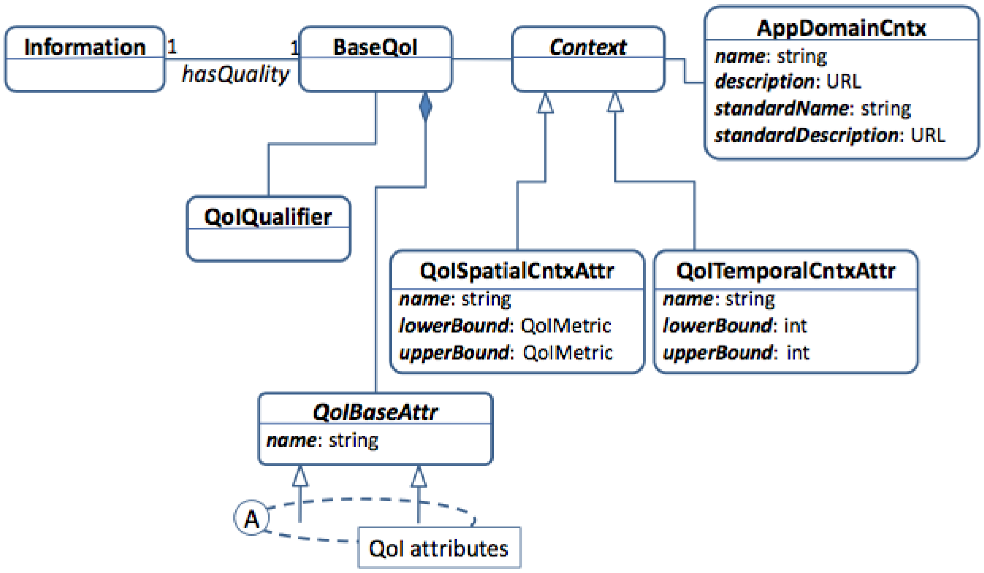


Figure 2.1: Bisdikian et. al. proposed the following ontology as a general purpose quality of information data model for sensor data.

In [11], Bisdikian et. al. proposed the ontology based data model shown in Figure 2.1

to describe the general body of attributes of Quality of Information (QoI) and Value of Information (VoI) for an arbitrary type of sensor data in Figure 2.2. A characterization of

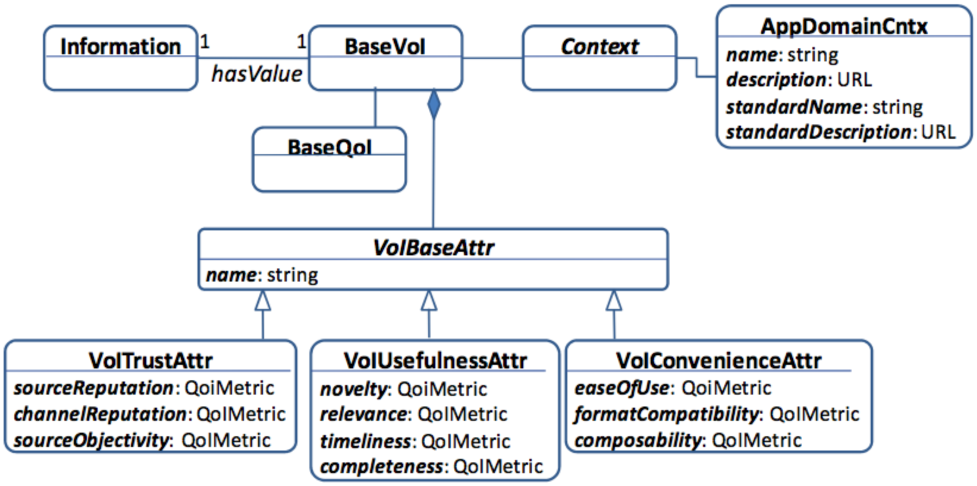


Figure 2.2: Bisdikian et. al. proposed the following ontology as a general purpose Value of Information (VoI) data model for sensor data.

QoI is useful in many contexts and can be invaluable in making decisions such as trusting, managing, and using the information in particular applications. However, the manner of representing QoI is highly application dependent, and incorporating algorithms on-the-fly to account for the operational and environmental context changes that these sensing agents may operate under can be cumbersome. Therefore, the authors proposed an application context agnostic ontology based data model for assessing the QoI and VoI of sensor information.

The authors define QoI and VoI as the following:

- **Quality of information (QoI)** represents the body of evidence (described by information quality attributes) used to make judgments about the fitness (or, utility) of the information contained in an information stream.
- **Value of information (VoI)** represents the utility of the information in an information stream when used in the specific application context of the receiver.

The flow of information in a typical SSW network goes from the sensing asset (human or

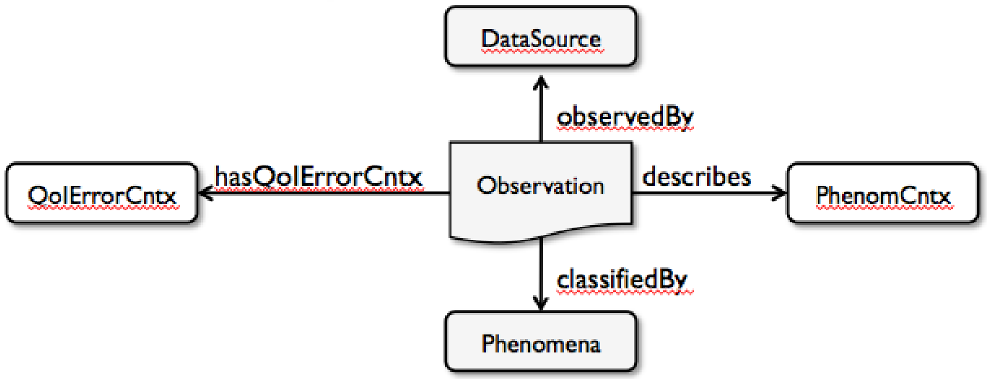


Figure 2.3: QoI Library: An ontology based model for associating QoI error analysis algorithms with the applicable sensing agents according to the operational environment context.

electronic device) to what the authors define as an information processing operator. Information operators may consume the sensor data directly or perform data fusion to produce a new piece of information and append metadata that describe the QoI and VoI attributes for sensor data when applied to a specific application context.

2.4 New Contribution in Defining QoI/VoI

In [18], we extended the data model described in section 2.3 and implemented an ontology based framework, referred to as the "QoI Library" for associating a library of QoI analysis algorithms that are specific to a data source.

Our proposed model (shown in Figure 2.3) is designed to be broad in scope for applications operating in any environment. Our data model follows the premise that an "*Observation*" of "*Phenomena*" (ie: worldly events) is observed by a "*DataSource*", operating within an environmental context denoted by *PhenomCntx* (eg: air, water, land), and may have some sort of (assumed detectable and correctable) error in measurement with respect to the specific operating environment. The following sections break down our data model in further detail.

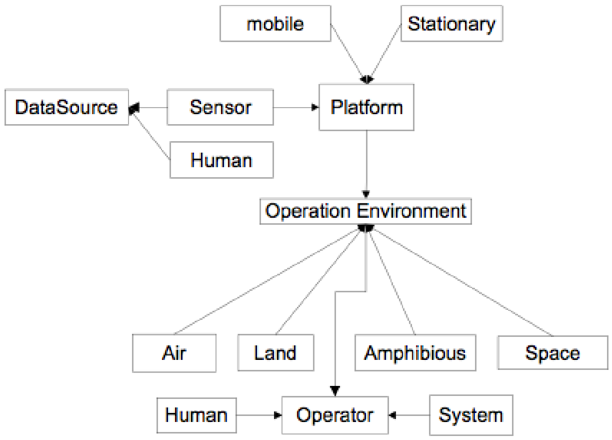


Figure 2.4: A *DataSource* (i.e. "data source") is an entity that disseminates and publishes information also known as an observation to its subscribers.

2.4.1 Sensing Agent

We studied the general characteristics of both human and electronic sensing agents (ie: *DataSources*) and the environmental contexts that these various sensing modalities operate under and describe our proposed ontology data model shown in Figure 2.4. In alignment with the use-case presented in Figure 1.1, sensing agents publish observations of a phenomenon typically as a sensor data stream that information operators can subscribe to in order to make decisions. Any process or agent that generates successive messages can be considered a source of information, where each observation published is treated as independent and somewhat stochastic, in a sense that all possible subsequent states of the sensor is determined by the predictable actions of the sensor plus the phenomena being observed.

We define the object to be classified as either a sensor or human. If a sensor, then we include such information as the type of platform used (stationary or mobile) and the sensors operation environment. Other information can also be extended from this framework as necessary to define identifying attributes for the software application.

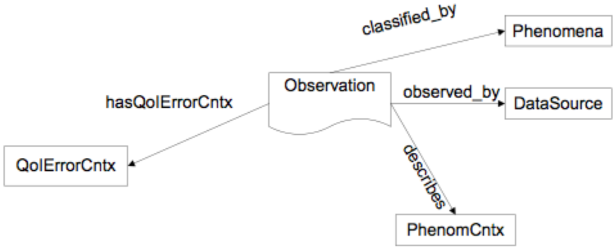


Figure 2.5: An *Observation* is a piece of information (e.g., sensor data) observed by and published from a data source describing an event of interest.

2.4.2 Observation

An *Observation* (see Figure 2.5) is a piece of information (e.g., sensor data) observed by and published from a data source describing an event of interest. An observation is usually measureable information that represents qualitative or quantitative attributes seen as the foundation of which knowledge is derived. The root framework of this library is tightly coupled in relationship to an observation.

The core of the ontology is comprised of predefined relationships which extend from an *Observation* object. An *Observation* describes information produced from a data source, denoted by the "*observed_by*" relationship. The event (eg: sniper shot) that triggered the Observation is classified or categorized by a unique term *Phenomena*, which we denote by the "$classified\_by$" relationship. A *Phenomena*, may produce many multifaceted observations that are mutually exclusive and potentially from a single or multiple data sources. We use the term mutually exclusive to state that each observation provides a qualitative or quantitative sensor measurement under a single context that is time bounded.

2.4.3 Phenomona

A *Phenomena* (see Figure 2.3) represents a uniquely identifiable name for categorizing any observable occurrence at the highest level. Essentially, the attribute is to be used for the naming and classification of an event that produced one or more *Observations* from a set of data sources. The *PhenomCntx* attribute provides information about the context of the

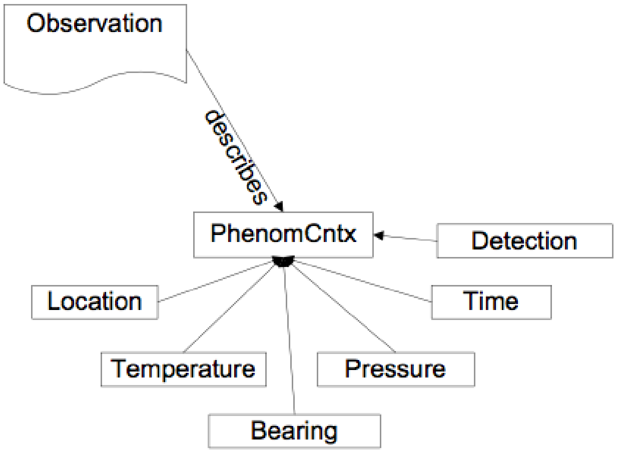


Figure 2.6: The *PhenomCntx* attribute provides information about the context of a specific observation. For example, an acoustic array can produce a bearing context event (*PhenomCntx*) as the result of a sniper shot (ie: transient phenomenon).

phenomenon being sensed. A *PhenomCntx* may span various spatiotemporal regions in relation to the phenomena that it describes. For example, consider the SniLoc scenario where a sniper shot (ie: transient phenomena) has occurred. A GPS sensor can produce a temporal observation under the context of location and an acoustic array can produce a direction of arrival (DOA) observation under the context of bearing. DOA events and GPS data are two separate entities each having their own unique context, although associated with the same parent phenomenon. The relationship between the *Phenomena* and *PhenomCntx* is unidirectional and hierarchical, with the *Phenomena* serving as the single parent instance that can generate multiple *PhenomCntx* events related to a unique context.

2.4.4 QoI Context

As mentioned in the earlier section, the QoI library, is an ontology based framework for mapping the relationship between analysis algorithms and a sensor data source (human or sensor) and its operational environment in order to aid in the improvement of QoI. In spite of the methods used to calibrate the sensors, random Gaussian noise may occur to bias the observations produced by the data sources. Therefore, we define a set of extensible abstract

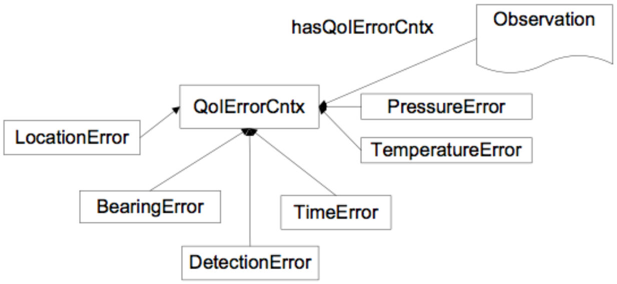


Figure 2.7: The *QoIErrorCntx* attribute is designed with extensibility in mind for SSW developers to categorize and extend QoI error analysis algorithms associated with a specific data source that provides information about a phenomenon of interest.

classes to represent all of the possible sensing modalities of a sensor. The software system designer and developer ideally would like to design and develop a system that can be easily copied and deployed as necessary with as few customization adjustments to the core system code as possible. Error analysis algorithms can be extended from the *QoIErrorCntx* classes (see Figure 2.7) by the software developer to perform analysis on observed and measureable events under specific contexts in the world. We stated that each sensor observes, measures, and reports occurring events under certain observable contexts in the world (*PhenomCntx*). Objects extending the *QoIErrorCntx* represent a taxonomic category of all feasible attributes that a sensor observation can represent, such as time, bearing, and temperature which when associated with the *QoIErrorCntx* will become a one to one mapping with their corresponding error object instance classes *TimeError*, *BearingError*, and *TemperatureError* respectively. The usage of the *QoIErrorCntx* is based upon the assumption that each extended object will define error analysis algorithms corresponding to assessing the capabilities of the data source that it corresponds to. The subclasses of the parent class can be augmented depending on the application domain.

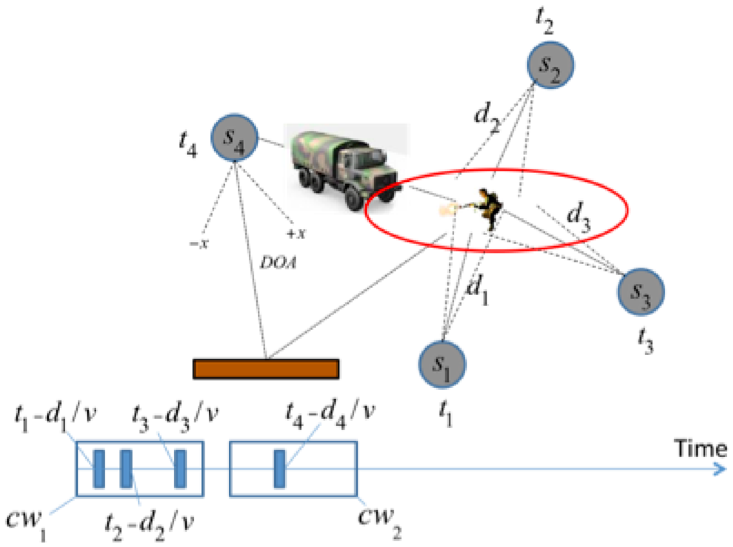


Figure 2.8: A collection of sensing resources belonging to a number of agencies (members) is deployed in a broad area of interest.

2.5 Use-case for the QoI Library

Consider the use-case outlined Figure 2.8, where we describe the sniper shooter localization (SniLoc) scenario. We consider a coalition operation environment with coalition partners pursuing common mission objectives; we can think of this as an instance of a multi-domain sensor driven use-case. The coalition partners collaborate at various mission tasks sharing along the way their sensing resources and senor originated information. Our interest here is in ISR (intelligence, surveillance, and reconnaissance) applications that depend on various sensor networks and sensing agents to monitor certain events. Based upon application needs, these sensing agents, which we shall call platforms, are designed with general purpose extensibility in mind. ISR sensor networks may include multiple sensor devices (GPS, barometer, acoustic, radar) residing on a single platform providing information feeds to applications.

With regard to our exemplar case, a shooter localization-sensing task is underway in support of a general ISR security mission. The task makes use of SniLoc, a sensor enabled application that analyzes sensor-originated information. SniLoc produces reports about sniper activity along with localization information produced by processing information derived from acoustic sensors. SniLoc is (assumed to be) deployed and running on a network such as the one in Figure 1.1, making use of the QoI metadata and the services provided by QoI annotators that create and process these metadata as discussed earlier. Suppose that SniLoc makes use of three acoustic sensors, noted as s_1, s_2, and s_3, for its localization operation. Suppose, the situation is such that sensors s_1 and s_3 are A-sensors while sensor s_2 is a non-A-sensor, i.e., it belongs and has been deployed by another coalition partner where the trust level of s_2 is questionable, so any localization report involving s_2 must account for this.

From a system design point of view, the above will involve some form of a sub-classing of QoI evaluation procedures that are applicable to the type of acoustic sensors that s_2 belongs to. For example, to calculate, say, the accuracy of the localization estimate provided by sensors s_1, s_2, and s_3, a software instance of a QoI annotator will need to access the QoI library for algorithms used to analyze and/or improve the quality of information produced from each sensor. Given a trace of historical observations of data produced from s_2, it is discovered that this sensor was configured to introduce an increased level of error in the measurements. However the error model follows a predictable pattern over a period of time. Error correction algorithms that specifically apply to one or more pieces of data produced from this sensor can be extended from the QoI library to enable real-time analysis or correction of produced information. With respect to SniLoc, the data fusion layer in Figure 2.9 aggregates/fuses acoustic sensor observations, such as DOA measurements, caused by shockwaves from the sniper fire. As a result of this aggregation, the layer produces localization reports that include not only the estimated location of the firing, but a statement regarding the goodness of the estimate that corresponds to the accuracy quality attribute.

The three properties mentioned above are relatively static over a long period of time, however dynamic properties, such as large military vehicles (Humvee, MRAP's, APC's and MBT's) temporarily residing in the path of the sensor also can affect DOA measurements from an acoustic sensor by delaying the time of arrival of sound waves. Figure 2.8 shows four sensors deployed in an area of interest that have detected a sniper shot. The red ellipse represents the region of the possible shooter locations. The variable d_i represents

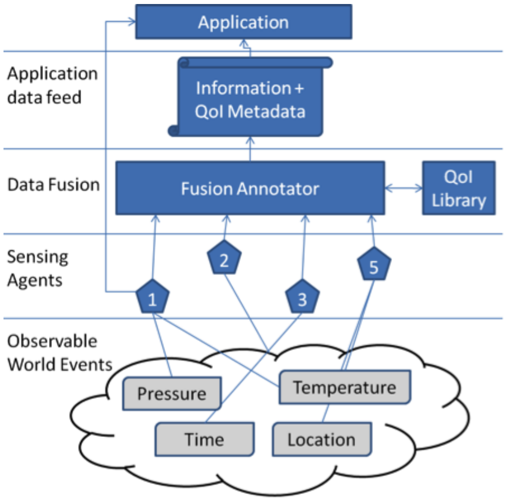


Figure 2.9: End to end architectural detail of the interaction between sensors observing an event and the application consuming sensor data.

the distance between the sensor and the shooters location; t_i represents the time that the sensor hears the event; cw_i (correlation window) represents the time window for processing all aggregated DOA measurements within the time period; and v represents the speed of sound. It is assumed that the position of each sensor is known. Given the location of each sensor, event timestamps, and DOA measurements (with error), the fusion annotator can estimate the location of the shooter. Upon receiving DOA measurements from each sensor, the fusion annotator can validate each sensors measurement by constructing a time-line of arrival at each sensor. Generally, sensing events will fall within a narrow window of each other denoted by cw_1. However, sensor s_4 is (presumably) located adjacent to a tall brick wall, with an elongated armored military vehicle impeding its line of sight to the sniper's location. In this situation, the sensor overhears the sound due to refraction against the wall, however the time of arrival of the event at the data fusion layer is within the next correlation time window cw_2, thus the measurement is excluded from the fusion process. A *Fusion Annotator* constructs a meaningful summary or transformation of its inputs in a format consumable by an application. The QoI library also resides within this layer to aid in the assessment of information to produce QoI metadata defining the accuracy of the information. The act of a sniper shot is considered a *Transient* event under the

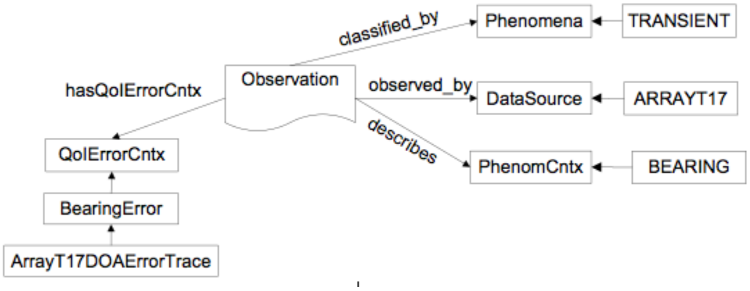


Figure 2.10: Example QoI Library scenario for shooter localization.

Phenomena attribute within the library (see Figure 2.10). The borrowed acoustic array, which has been configured by coalition members to reduce the accuracy of published data is named "ArrayT17". This array produces DOA measurements, also known as "Bearing" events, denoted by the *PhenomCntx*. The class "ArrayT17DOAErrorTrace", extends from the generic *BearingError* class, and performs a trace on DOA measurements reported for this sensor whenever a new DOA measurement is reported. The output of this algorithm is an assessment of trustworthiness given previous reports and their value to the application.

The *QoILibraryHandlerIFC* and *DataSourceHandlerIFC* is an interface implemented for the QoI library that manages the indexing of extended error calculation classes and data sources respectively.

```
QoILibraryHandlerIFC handler = QoILibraryHandler.getInstance();
DataSourceHandlerIFC datasource_handler = QoILibraryHandler.getInstance();
```

The new error algorithm class "ArrayT17DOAErrorTrace" to be added to the QoI Library will be extended using the following method *addNewErrorLibrary()*. This method serves as a triple key/value pair that provides an indexing scheme for the error analysis algorithm class in the library according to the following attributes denoted in the code below.

```
//Load library
QoIErrorCntx library = new ArrayT17DOAErrorTrace();
DataSource source = handler.addNewErrorLibrary(
DataSourceHandlerIFC.SENSOR,"ARRAYT17","TRANSIENT’’,
QoILibraryHandlerIFC.BEARING, library);
source.setName("ARRAYT17");

//Retrieve library
QoIErrorCntx lib = handler.retrieveErrorLibrary("ARRAYT17",
"TRANSIENT",QoILibraryHandlerIFC.BEARING);
```

Whenever an application needs to access the algorithm for assessing the reputation of measurements reported from “ArrayT17”, the DOA error analysis algorithm is retrieved by the aforementioned code.

The scenario described above is a subset of many other types of algorithms that provide usage for error analysis and assessment of sensor data.

2.6 Discussion

The proposed design (see Figure 2.3) is the first step towards providing an extensible library for system software developers to design and develop a QoI aware solution with the QoI Library. Future work will consider the addition of a model for data provenance. Data fusion layers may produce a new data stream summarizing a set of information reported from aggregated sensor data. It is our intent to maintain a historical record of ancestral information, appended as metadata to maintain a lineage of the transformation process of data produced. This research was sponsored by the US Army Research Laboratory and was accomplished under Agreement Number W911NF-06-3-0002-P00008.

CHAPTER 3

Sensite: Knowledge based Platform for Semantic Sensor Web Queries

The goodness and utility of information typically is assessed by how comprehensive the what, when, and where properties of the information are. In chapter 2, we studied how WSN's are designed, deployed and operated. We proposed our vision for WSN's to move from traditionally "closed" set-ups, where WSN's are intimately tied to their applications to an open platform where information operators can dynamically bind to web accessible sensing agents on-the-fly. We also discussed the design of our ontology based data model for assessing the quality and value of information (QoI and VoI) provided by a data source. The recent Internet of Things (IoT) and Web of Things (WoT) standard embodies the vision that we proposed in [18]. In this section, we describe *Sensite*, an open sensor data publishing and query web platform that applies the data model proposed in chapter 2 in order to add contextual relevance to the taxonomic relationship between sensors qualified to provide information about a worldly phenomena (ie: event) of interest.

3.1 Introduction

The *Internet of Things (IoT)* is a paradigm that defines a framework where multiple internet enabled intelligent embedded devices can be controlled and their data streams assessed. Recent applications in the IoT space have been implemented in domains such as home [19,20], logistics [21] and transportation [22, 23]. More recently, the *Web of Things (WoT)* builds on the IoT standard and mandates sensor devices to deploy representational state transfer (RESTful) API's to access the data streams that a sensing agent provides. The IoT and WoT are a proposed standard that brings SSW enabled devices closer to the realization of our vision described in chapter 2 and [18], where the data feeds from multi-purpose SSW sensor devices (when commercially available) are made available to a group of trusted members who are granted access. Information from these data streams are fused by information operators in order to make inferences and decisions about multiple phenomena of interest. We apply the *Observation* concept from our ontology based data model to a web based sensor data publishing platform named *Sensite*. We also discuss the research challenges faced when designing an unsupervised learning algorithm within Sensite that identifies what type of sensing agents are qualified to provide information about a phenomenon (directly or through data fusion).

3.2 Conventional SSW Sensor Data Platforms

COSM was a popular "open" SSW sensor data platform [24] that provided a repository for SSW developers to publish arbitrary types of sensor data streams (such as Arduino, Twitter feeds, or other feeds). In order to publish to COSM, the owner must assign a title (eg: Arduino Temperature) and descriptive keywords (eg: Arduino, temperature, outside) to use as search terms. Information operators could consume a feed by performing a query using the relevant keywords, then bind to the related sensor feed(s) of interest from the query results. COSM enabled users to store specific or arbitrary types of sensor data, however did not provide a comprehensive set of contextually related classifiers that describe the heterogeneous environmental contexts that the sensor data may be applicable to. In other words, these sensor data may be applicable to measuring and/or describing many other types of phenomena (eg: shooter localization, nuclear radiation, etc.), however there is no clear model that enables the user and/or application to obtain as much information as possible to make an informed decision about a certain phenomenon of interest without performing multiple queries for the different types of applicable sensor data and implementing algorithms to fuse these data. Moreover, having sufficient subject matter expertise of the applicable types of sensor data is also a requirement.

Recently, COSM has changed their name to Xively [25] and their business model to providing a service that enables seamless communication for arbitrary types of SSW devices. COSM also provides custom data analytics through dashboards in order to support the rapid decision making process by minimizing the time that it takes to process information. Other popular related commercial SSW platforms either enable seemless communication protocols for disparate sensors [25, 26] or provide analytics for a specific set of SSW enabled sensors [12, 27, 28].

We studied the principles of information theory and propose an unsupervised learning algorithm implemented within our Sensite Knowledgebase platform that learns the many phenomena that a sensor is qualified to provide information about by crawling unstructured text from the world wide web (www). Our algorithm enables a non-subject matter expert to make informed decisions about a phenomenon of interest without having to know beforehand all of the applicable types of sensor devices.

3.3 System Overview

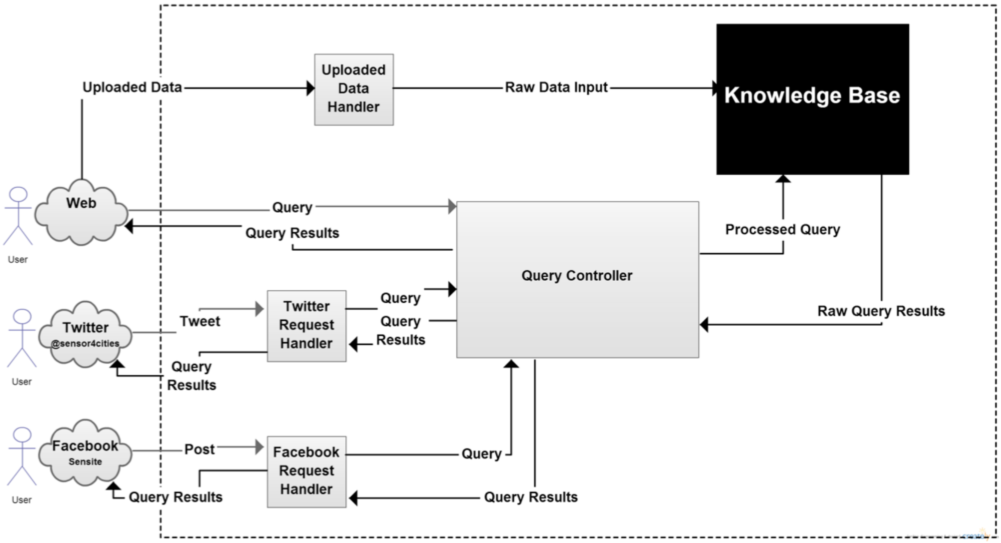


Figure 3.1: System architecture diagram for the Sensite sensor data knowledgebase web platform.

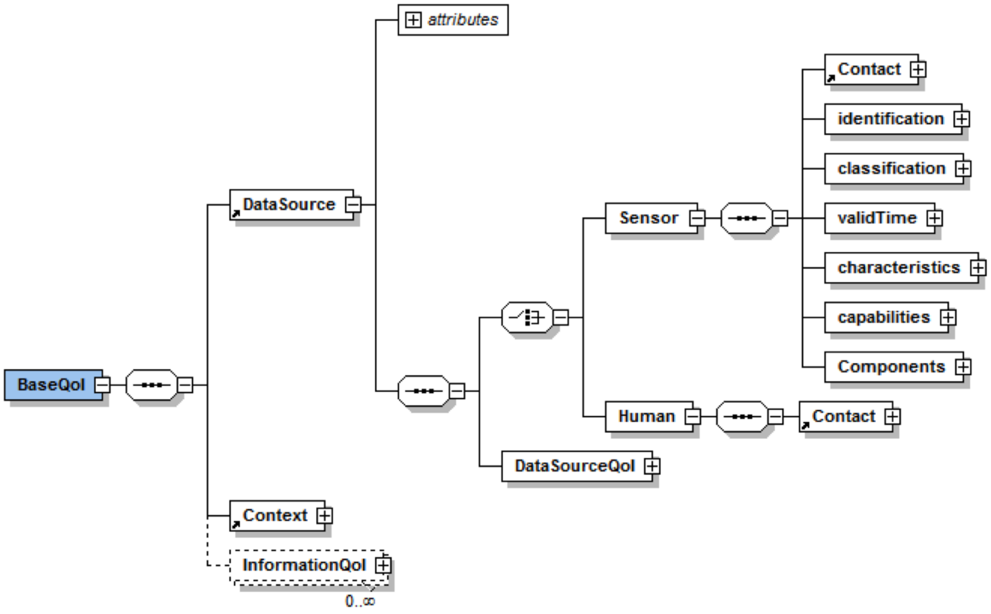


Figure 3.2: Ontology based data model that defines the *BaseQoI* schema for annotating a piece of sensor data with QoI information.

Sensite is an open sensor data web platform that enables information operators (human or software) to query for arbitrary types of sensor information that is “best” qualified to provide information about a phenomenon of interest at a given location, at a given time. Sensite can be used to query these sensor data feeds via the RESTful API, the webpage, Twitter or Facebook pages (see Figure 3.1). The metric that Sensite uses to gauge what sensors are “best” qualified to provide information either directly or indirectly (assumed though data fusion which is at the discretion of the user to perform) about a phenomenon of interest comes from an inference based learning algorithm that crawls webpages from the www and applies statistical algorithms to obtain a relevancy metric. Sensite’s system architecture is shown in Figure 3.1. Users can upload QoI/VoI annotated sensor data in batch or individually using the Sensite webpage or RESTful API, and Sensite’s Knowledgebase module will process the information about the publishing data source.

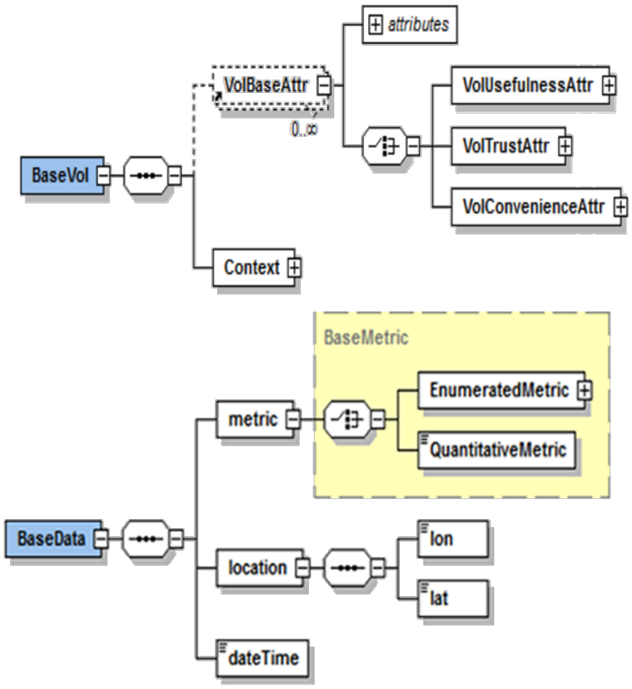
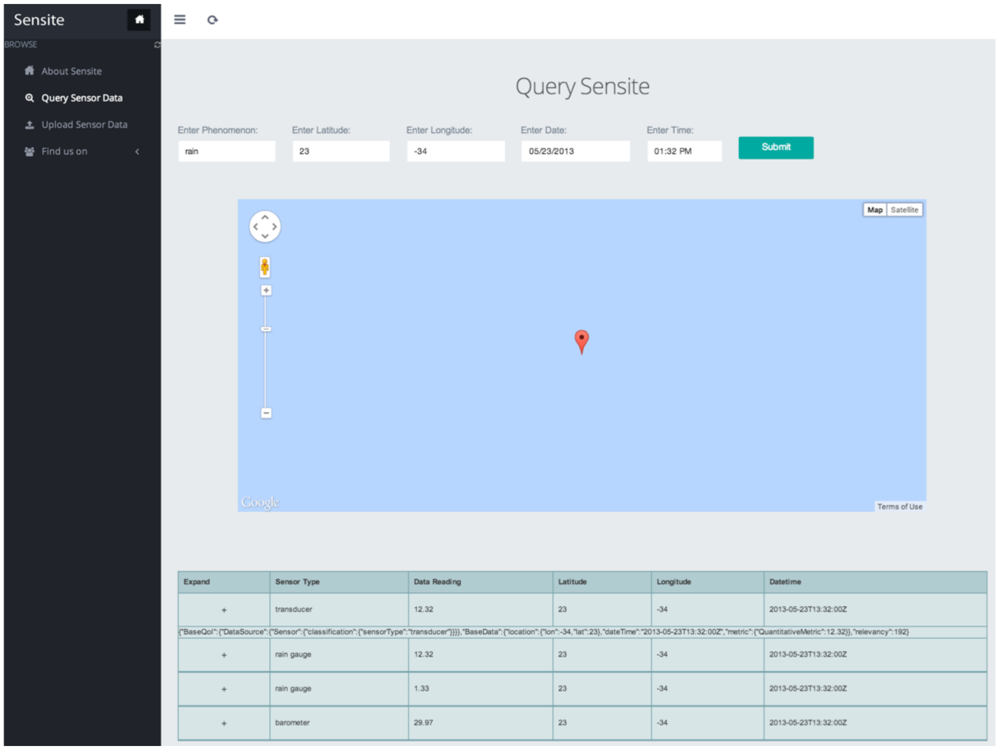


Figure 3.3: The *BaseVoI* class describes attributes related to the value of information. The *BaseData* class contains the sensor data and with spatiotemporal information about the observation.

3.4 Implementation of the Ontology Based QoI/VoI Data Model

We applied the ontolgy based model proposed in section 2.4 as an instantiable XML schema document (XSD) (see Figure 3.2), which the end user can convert to a Javascript Object Notation (JSON) format representation and append QoI/VoI metadata to the information produced from the sensor data feed(s). We extended the data model by providing the *BaseVoI* and *BaseData* attributes proposed in [11] (see Figure 3.3). *BaseVoI* describe attributes that are related to the value of information. Attributes that we extend are *VoITrustAttr*, which describe data trustworthiness and *VoIUsefulnessAttr/VoIConvenienceAttr* which describes usefulness and utility respectively of the data for the information operator to consume. *BaseData* contains the actual sensor datum point, along with spatiotemporal information about the observation. The sensor data returned by Sensite is in JSON format in order to make consuming the data easier for human and web service based information operators.



Enter Phenomenon: rain
Enter Latitude: 23
Enter Longitude: -34
Enter Date: 05/23/2013
Enter Time: 01:32 PM

Expand	Sensor Type	Data Reading	Latitude	Longitude	Datetime
+	transducer	12.32	23	-34	2013-05-23T13:32:00Z
{"BaseQoI":{"DataSource":{"Sensor":{"classification":{"sensorType":"transducer"}}}},"BaseData":{"location":{"lon":-34,"lat":23},"dateTime":"2013-05-23T13:32:00Z","metric":{"QuantitativeMetric":12.32}},"relevancy":192}					
+	rain gauge	12.32	23	-34	2013-05-23T13:32:00Z
+	rain gauge	1.33	23	-34	2013-05-23T13:32:00Z
+	barometer	29.97	23	-34	2013-05-23T13:32:00Z

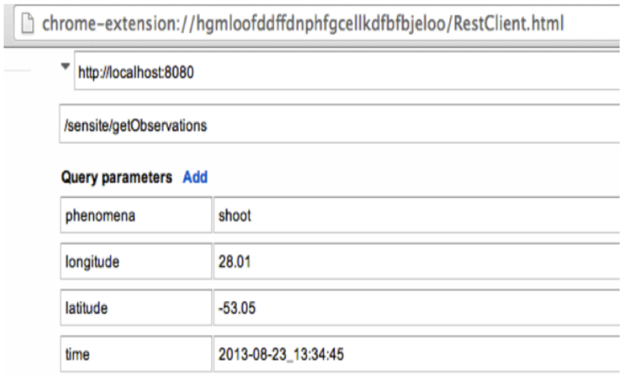
Figure 3.4: Webpage interface for querying sensor data.

3.5 How to Query/Upload Sensor Data

The Sensite application applies the corollary stated in section 2.4.2 that a *sensing agent (ie: Datasource) senses a phenomena occurring at a certain location, at a given time.* The query language that we use also adheres to the structure. In the next sections, we describe how the user can perform a query to obtain sensor data.

3.5.1 Webpage

The webpage interface, shown in Figure 3.4 enables the user to query the name of a phenomenon of interest, the latitude and longitude geographic location, and the date/time of the desired observation. When the user clicks "Submit", all relevant sensor data is returned in the table below. The user can click the "+" in the **Expand** column to see the JSON formatted metadata associated with that piece of sensor data.



chrome-extension://hgmloofddffdnphfgcellkdfbfbjeloo/RestClient.html

http://localhost:8080

/sensite/getObservations

Query parameters

phenomena	shoot
longitude	28.01
latitude	-53.05
time	2013-08-23_13:34:45

Figure 3.5: RESTful API for querying sensor data.

Sensor data can be uploaded by clicking the "Upload Sensor Data" link. The user will then be presented with a page that enables batch uploading of the sensor data to the Sensite Knowledgebase.

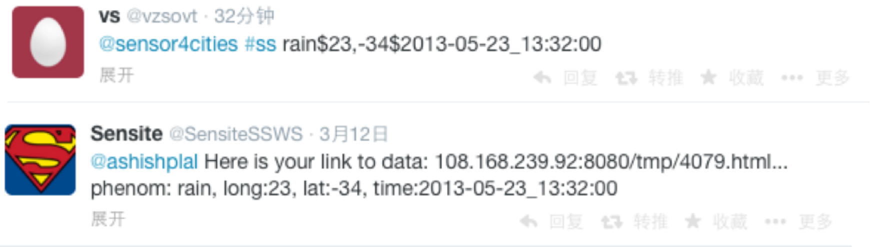
3.5.2 RESTful API Web Service

The RESTful API exposes two web service methods that enable users to upload or query sensor data. The query API, shown in Figure 3.5 requires the name of a phenomena of interest, the latitude and longitude geographic coordinates, and the date/time of the desired observation. The data returned will be JSON formatted metadata associated with that piece of sensor data. Sensor data can be uploaded from the REST API by accessing the URL to upload sensor data. This is also the same URL that the Sensite "Upload Sensor Data" webpage uses.

3.5.3 Twitter and Facebook

Many sensor data applications have leveraged social networking sites as a method of mining or publishing sensor data [29–31]. Twitter and Facebook are (in essence) fairly open sensor data publishing platforms. For instance, the Facebook Check-in feature uses sensor data to track the location of an individual.

We collaborated with Dr. Eduardo Cerqueira, from the Federal University of Para,



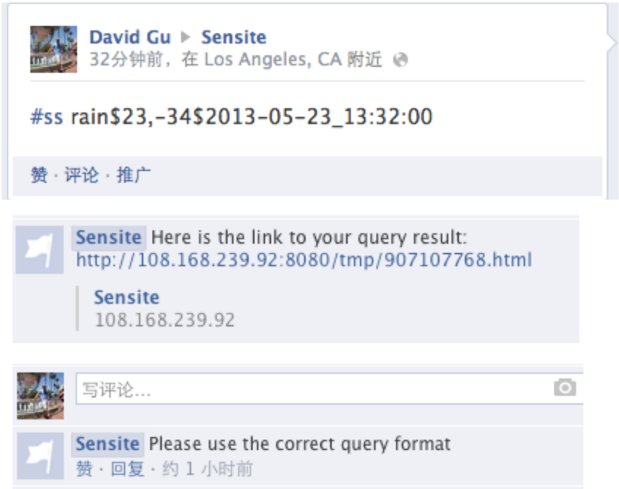
vs @vzsovt · 32分钟

@sensor4cities #ss rain$23,-34$2013-05-23_13:32:00

Sensite @SensiteSSWS · 3月12日

@ashishplal Here is your link to data: 108.168.239.92:8080/tmp/4079.html... phenom: rain, long:23, lat:-34, time:2013-05-23_13:32:00

Figure 3.6: Query sensor data via Twitter.



David Gu ▸ Sensite

32分钟前，在 Los Angeles, CA 附近

#ss rain$23,-34$2013-05-23_13:32:00

Sensite Here is the link to your query result: http://108.168.239.92:8080/tmp/907107768.html

Sensite
108.168.239.92

Sensite Please use the correct query format

赞 · 回复 · 约 1 小时前

Figure 3.7: Query sensor data via Facebook.

Brazil in order to integrate their existing Twitter social networking *Sensor4Cities* sensor data querying platform into our Sensite platform so that users can query for sensor data using the following query structures for Facebook and Twitter (examples given in Figures 3.6 and 3.7):

- **Twitter** - @sensor4cities #ss phenomenon$latitude,longitude$date_time (see Figure 3.6)
- **Facebook** - #ss phenomenon$latitude,longitude$date_time (see Figure 3.7)

Sensite will post a response containing a dynamic HTML link to the sensor data requested that the user can click on to see the raw sensor data. An incorrect query will result in a response message posted on the page.

3.6 Unsupervised Learning Algorithm

3.6.1 World Wide Web as a Data Source

The www hosts over one billion websites, each containing a wealth of multi-faceted information. We assume that each webpage represents a piece of information about our universe that we assume to be true (including the justification for those entanglements). Our algorithm associates mentioned sensors with the phenomena that they are qualified to provide information about by analyzing webpage text and extracting the contextual meaning of only complete (properly structured) sentences. Based upon the law of large numbers, we assume that our algorithm will eventually converge to the true relationships as new information is obtained from the number of related websites visited. We will go into further detail later in the chapter.

A big challenge that we faced was figuring out how to query webpages in the www. Due to economies of scale, we are not able to query the entire www. Therefore, we decided to use Google's search engine as a vehicle for webpage search and extracting information from webpages that contain sentences describing relevant contextual information. Our algorithm assumes the following information is known a priori:

a) the name (not including the synsets) of sensor devices capable of measuring a worldly phenomena

b) the name (not including the synsets) of worldly phenomena (ie: events) that are observable by a sensor device

c) knowledge of verbs (including the synsets) that are used in a complete sentence to ascertain the extent, dimensions, or quantity of a worldly phenomena

a) We store the names of a total of 139 sensors. Some of which are rain gauge, accelerometer, breathalyzer, acoustic sensor, etc. Sources such as Wikipedia [32] provide an exhaustive list of sensor devices that are both modern and obsolete. One could argue to simply assume that the sensor usage definition provided by Wikipedia is sufficient rather than learning the relationships by analyzing webpage text. However, the definition given by Wikipedia and the dictionary only describe the intended purpose of the sensor, and omit other contexts that the same sensor can be used to provide information about. For instance, an optical sensor is by definition a device that is used to measure ambient light. However, this definition alone omits the fact that an optical sensor can also be used to detect an explosion (a transient worldly phenomenon). Another exemplary usecase is using a speedometer to determine the speed of an object. However, speed can also be extracted using other sensors such as an accelerometer or a clock by applying a transform or sensor data fusion algorithm.
b) We assume knowledge a priori of the names of observable worldly phenomena (not including their synsets) because our inference algorithm is currently not capable of automatically deciphering the terms that describe a phenomenon. Therefore, we store the names of a total of 71 observable worldly phenomena in a database. Some which are rain, acceleration, fire, explosion, etc.

As mentioned previously, although we store the names of sensors and worldly phenomena, our knowledgebase initially is not aware of any of their relationships. This information is learned by analyzing the text from the www. As mentioned in section 3.6, no SSW

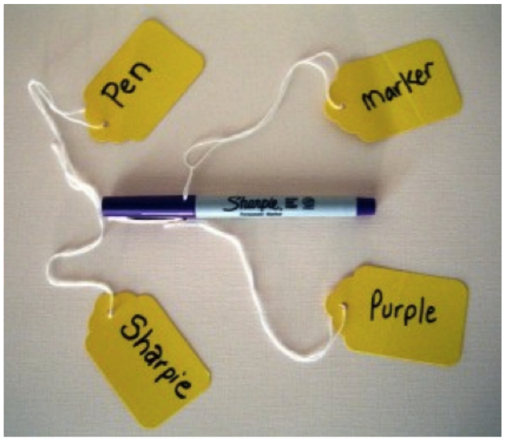
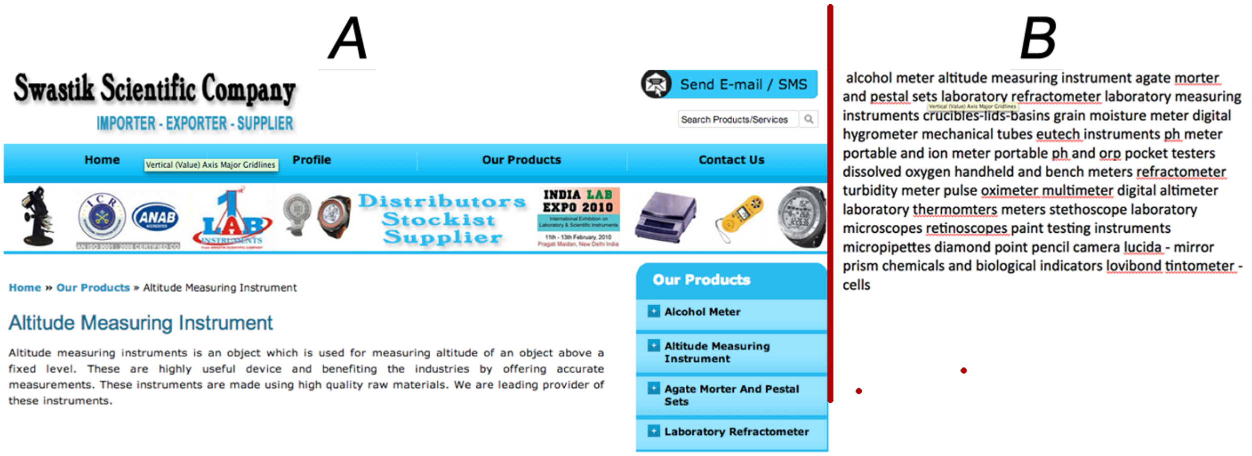


Figure 3.8: This object is tagged with keywords that describe what it is, but no keywords are listed that describe the contexts of how the object can be used (e.g. writing, coloring, hole punch, etc.). This is the problem with current SSW sensor data storage platforms.

sensor data platform provides a comprehensive set of contextually related classifiers that describe the heterogeneous environmental contexts that the sensor data may be applicable to: Requiring the information operator to have knowledge beforehand of all the necessary sensor data that their application needs. We use the example in Figure 3.8 to illustrate our main point in bullet point a), which is that most SSW data publishers tend to apply keywords that describe the sensor type and the current context to which the sensor is applied. However, when the data is uploaded to a SSW sensor data platform, the SSW data publisher usually neglects to add keywords to describe the other related contexts to which the sensor data can also be applicable to. Humans naturally categorize sensors according to contextual relevance, rather than by the sensor name or feed id. Given this, Sensite provides a query structure where users can perform spatiotemporal queries based on the phenomena of interest (see section 3.5), and recieve a manifest list of all the applicable sensor types (if data exists) that they can bind to. Our algorithm is implemented within the Knowledge Base module (see Figure 3.1) of Sensite as a service. To the best of our knowledge Sensite is the first SSW sensor data platform to apply this approach.

c) There are many ways that a sentence can be constructed to describe the same contextual relationship. There are also many different senses that a single word can be used, and synsets


A
Swastik Scientific Company
IMPORTER - EXPORTER - SUPPLIER
Send E-mail / SMS
Search Products/Services
Home
Vertical (Value) Axis Major Gridlines
Profile
Our Products
Contact Us
Distributors Stockist Supplier
INDIA LAB EXPO 2010
Our Products
Alcohol Meter
Altitude Measuring Instrument
Agate Morter And Pestal Sets
Laboratory Refractometer
B

Home » Our Products » Altitude Measuring Instrument

Altitude Measuring Instrument

Altitude measuring instruments is an object which is used for measuring altitude of an object above a fixed level. These are highly useful device and benefiting the industries by offering accurate measurements. These instruments are made using high quality raw materials. We are leading provider of these instruments.

alcohol meter altitude measuring instrument agate morter and pestal sets laboratory refractometer laboratory measuring instruments crucibles-lids-basins grain moisture meter digital hygrometer mechanical tubes eutech instruments ph meter portable and ion meter portable ph and orp pocket testers dissolved oxygen handheld and bench meters refractometer turbidity meter pulse oximeter multimeter digital altimeter laboratory thermomters meters stethoscope laboratory microscopes retinoscopes paint testing instruments micropipettes diamond point pencil camera lucida - mirror prism chemicals and biological indicators lovibond tintometer - cells

Figure 3.9: a) A human viewing this webpage in a web browser, can process the text in this webpage to understand what an altimeter is by reading the complete sentences. b) The Sensite Knowledgebase sees only unstructured text and must extract the sensor to phenomenon relationship. Our algorithm only processes text that is a complete sentence.

(ie synonyms) that describe the same hyponym. For instance:

- A pluviometer is used to quantify the amount of rainfall over a period of time.
- A udometer is a sensing device used to measure the amount of precipitation over a period of time.
- A rain gauge collects falling precipitation and funnels it to a rain measurement device. (Source: www.weathershack.com)

The aforementioned sentences use three different synsets to describe a rain gauge sensor and there are three different synsets used to describe rain. Therefore, we apply a semantic similarity algorithm to identify the relationships between the words of interest when analyzing the text. We will discuss our algorithm in the next section.

3.6.2 Understanding the Semantic and Grammatical Context of a Sentence

A human typically processes information from a webpage viewed in the browser by reading complete sentences and extracting the relevant context. However, when a computer program (ie: our Sensite knowledgebase) views the same webpage, it sees simply unstructured text. Therefore, the research challenge of interest is identifying the best method to process text within the webpage in order to identify the relevant context (see Figure 3.9B).

Considering this, we selected a few random sensor devices and searched for webpages that describe the phenomena that each sensor is qualified to provide information about. Our goal was to identify certain predictable characteristics of a properly structured sentence. We then viewed the same webpage from the perspective of what our Knowledgebase would see, and discovered that the best method to obtain valid information from the document corpus is to process only complete sentences. A complete sentence typically consists of a subject-verb pair, and present a complete thought. Sentences that describe a valid sensor/phenomena relationship contained on average 4 words at minimum and 24 words maximum. This average sentence length also concurs with LIWC [33, 34]; a comprehensive study to understand the physical and mental health of the author based upon their writing (web, article, etc.) style using sentiment analysis. Finally, we discovered that a sentence of interest must also contain either of the following hyponyms or their synsets; measure, detect, or quantify.

3.6.2.1 Overview of the N-gram Language Model

N-grams [35] are a very effective method to validate proper word and sentence structure. It is used in most text editors that offer features, such as spell check [36, 37] and sentence structure validation [38, 39]. A statistical N-gram language model describes the conditional probability of a word w given its history, described by the $n-1$ previous words, h. The general equation is given by

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)....P(w_n|w_1^{n-1}) \approx \prod_{k=1}^{n} P(w_k|w_{k-1}) \tag{3.1}$$

where the word sequence $w_1, w_2, ..., w_{n-1}$ is represented as w_1^{n-1}. The likelihood estimate of a word's existence is based on the samples obtained by a large text corpus consisting of (ideally) many different genera's of literature (eg: poems, legal documents, blogs, scientific articles, comedy, etc.). The more diverse the text corpus, the better the likelihood estimates. N-gram conditional probabilities can be estimated from raw text based on the relative frequency of word sequences. Popular N-gram combinations are unigram (N=1), bigram (N=2), and trigram (N=3).

To understand how N-grams work, lets take the following sentence as an example using a bigram: A sphygmomanometer is used to measure blood pressure. One would represent the conditional probability of this sentence by the following equation

```
P(<s> A sphygmomanometer measures blood pressure </s>) =
P(A|<s>)*P(sphygmomanometer|A)*P(measures|sphygmomanometer)*
*P(blood|measures)P(pressure|blood)*P(</s>|pressure)
```

Note that $< s >$ and $< /s >$ are standard symbols used to denote the start and end of a sentence respectively. Using the conditional probability equation for a bigram described in equation 3.1, the probability of a complete sentence is the product of the maximum likelihood estimates (MLE) of each bigram word combination. A single term can be thought of as the likelihood of seeing the current word (or symbol) w_n given the prior word (or symbol) w_{n-1}, as a factor of the emission probability that the prior word w_{n-1} has been seen in the training corpus.

3.6.2.2 Using N-grams to Validate Complete Sentences

In this study, we explored bigrams and trigrams to find the best method to recognize complete sentences. An N-gram language model must be trained using a large text corpus in order to estimate the parameter values. Therefore, we constructed a large training and testing text corpus of sample literature from the following sources [40,41]. These sources consist of news articles, editorials, reports, scientific papers, and legal trial transcripts. Next, we used the training data to establish a conditional probability that a word w_n will exist given the prior word order for the bigram and trigram using the following equations respectively

$$P(w_n|w_1^{n-1})_{bigram} = \frac{C(w_{n-1}, w_n)}{C(w_{n-1})} \tag{3.2}$$

$$P(w_n|w_{n-N+1}^{n-1})_{trigram} = \frac{C(w_{n-N+1}^{n-1}, w_n)}{C(w_{n-N+1}^{n-1})} \tag{3.3}$$

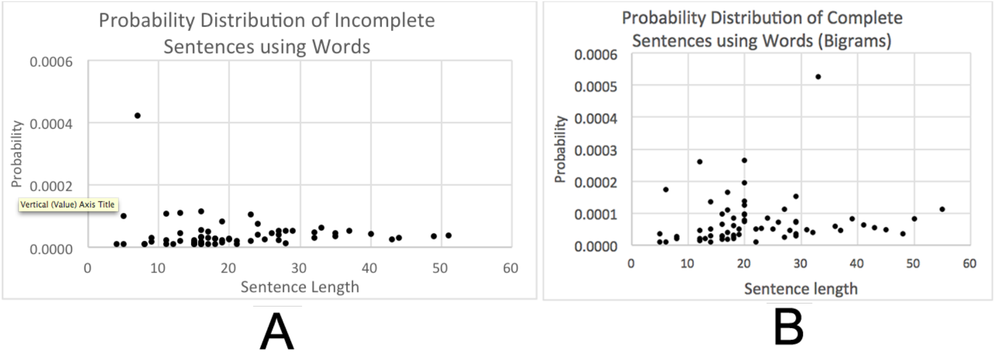


Figure 3.10: Chart's A and B show the probability distribution for sentences that we know to be complete or incomplete for bigrams respectively according to sentence length. The distribution is indistinguishable due to having a limited size training data set.

$C(x, y)$ is a function denoting the total number of times that the word w_n is preceded by the prior word w_{n-1} for a bigram or w_{n-N+1}^{n-1} for a trigram. $C(x)$ is a function denoting the total number of times that w_n exists in the training data.

After obtaining the MLE estimates for both bigrams and trigrams, we used the random sentences collected from webpage searches, which we knew to be either complete or incomplete in order to identify a threshold that distinguishes between a complete vs. incomplete sentence. Results show that the probability distribution describing the range of values for complete vs. incomplete sentences given sentences of various length are indistinguishable for bigrams (see Figure 3.10) due to having limited size training set. The probability distribution for trigrams are worse, with the majority of values converging close to zero for both sentence types, due to many non-existing trigram word combinations.

Generally, a "good" training set is comprehensive, such that it (ideally) includes all of the many ways that words can be structured to produce a complete sentence. Since the dataset used in this study is limited in size, an alternative approach was taken in order to predict complete sentences based on part of speech (PoS) rather than word combination. A complete sentence must contain a subject-verb pair and express a complete thought. In addition, a complete sentence usually follows a fairly structured ordering of words. For example, a complete sentence will not begin or end with a conjunction. Considering this, we re-trained the conditional probability tables for equations 3.2 and 3.3 respectively to produce MLE's according to sentence PoS and repeated our test to distinguish between complete vs. incomplete sentences. We found that setting $P(w_1^n)_{bigram} = .006$ results in 87% correct predictions of complete sentences. Given the amount of training data that we have access to (which is over 2GB in size), building a conditional probability table using PoS was the best solution to maximize the likelihood of correctly predicting a complete sentence. If given access to a larger data set, the accuracy of our algorithm can be improved greater.

The next step was implementing a webpage text parser that split the text in a webpage into a vector of elements according to punctuation marks. Then feed each element (i.e. sentence) into our Sensite Knowledgebase module, containing the bigram sentence detection algorithm. The text parser removes all extra space characters (including new lines) before feeding a line of text into the bigram complete sentence prediction algorithm. Reasons for an incomplete sentence are either the text from html links or ads were appended to a valid sentence, alternatively sentences listed as bullet points with no period at the end were appended together to produce a run-on sentence. If our algorithm predicts that a sentence is complete, then the next step is to evaluate the context of the sentence, which we discuss in the next section.

3.6.2.3 Understanding the Context of a Sentence

A complete sentence contains a subject-verb pair and expresses a complete thought. Prior to beginning our experiments, we studied the characteristics of complete sentences that describe a sensor-phenomenon relationship and discovered that each sentence contains at least two nouns and a verb. The common verbs found in the majority of sentences were "detect" or "measure" (or its synset). Since the sentence that we are looking for should contain a term that denotes quantifying or detection, our algorithm only performs contextual analysis if either a verb exists. If the term exists, then our algorithm searches for two nouns that denote the sensor-phenomenon relationship. As stated in section 3.6.1, the sensor and phenomena terms must exist in the database in order for our algorithm to track the relationship. In addition, the sensor and phenomenon terms in our database are all mutually exclusive (no synsets exist). Therefore, we applied the Wu & Palmer (wup) lexical word similarity algorithm [42] to gauge the relationships between words. We chose to use the wordnet similarity for Java (ws4j) due to their successful integration of Wordnet's lexical dictionary [43]. WUP uses a decimal value between 0 (no relationship) and 1 (strong) to denote the strength of the relationship between two words. The strength is based upon the number of overlapping synsets between two given words.

Other approaches considered were the Jaccard Index [44, 45] and the Cosine Similarity [46,47] algorithms. The weakness of both algorithms is that they measure sentence similarity by exactness rather than context. Therefore, in order for this algorithm to scale we must maintain a database of all possible ways to construct a complete sentence that describes the relationship of every valid sensor-phenomenon relationship that exists. This solution is not scalable.

Now that we are able to ascertain the sensor-phenomena relationship by evaluating sentences, we discuss our algorithm for a relevancy metric that determines the strength of the relationship.

3.6.2.4 Storing the Sensor to Phenomena Relationship

The Sensite Knowledgebase crawls random webpages from a Google search looking for complete sentences (predicted by our bigram algorithm described in section 3.6.2.1) that contain between 4 to 24 words. Given a valid sentence, our algorithm looks for any verb term that describes detection or quantification. If the term is found, then the sentence is processed further to extract at least two nouns that describe a sensor and a phenomenon listed in our database. If the two terms are found then we count the relationship in our database. Note that our algorithm is not yet capable of determining whether a sentence is communicating a "does not" relationship. This is the subject of future work. If the desired terms are found in a sentence, then our algorithm assumes a valid association.

We track the sensor-phenomenon associations in our database using the following JSON data structure:

```
{
    "phenomenon": "temperature",
    "observations": 11,
    "association": [
        {
            "sensor": "temperature gauge",
            "count": 1
        },
        {
            "sensor": "thermometer",
            "count": 6
        },
        {
            "sensor": "infrared thermometer",
            "count": 1
        },
        {
            "sensor": "multimeter",
            "count": 1
        },
        {
            "sensor": "thermocouple",
            "count": 2
        }
    ]
}
```

The “phenomenon” variable represents the worldly event that sets of sensors are qualified to provide information about (temperature in this case). The "association" variable is an array that contains the list of sensors that the Knowledgebase discovers are associated with temperature as a result of crawling random webpages, and the "observation" variable represents the total number of observed webpages that contain a complete sentence describing a sensor-phenomenon relationship. Finally, the "count" variable that is associated with each sensor represents the total number of times a sentence was encountered describing that specific relationship.

3.6.2.5 Association Algorithm to Determine Sensor to Phenomena Relationship

One can think of the way that we approach associations as treating the billions of webpages in the www as a universe. Each valid sentence found in the webpage is a piece of information in a universe that is assumed to be true, however the likelihood of actually being true is a factor of the new information also agreeing with the prior discovery as more webpages are visited. We tried various approaches to developing an algorithm that tabulates the observed sensor-phenomena relationships. We discuss each approach considered and their justification. *Bayes' Theorem:* Our first approach was to model the relationship using Bayes' Theorem. Let $\omega = s_1, s_2, ..., s_n$ be the set of unique sensors listed in our database, and $\phi = p_1, p_2, ...p_n$ be the set of unique phenomena in the database. H_i is our belief that s_i measures (or detects) p_i and D_i is the piece of information found in the webpage that describe the relationship. The likelihood of the sensor-phenomena relationship can be defined as

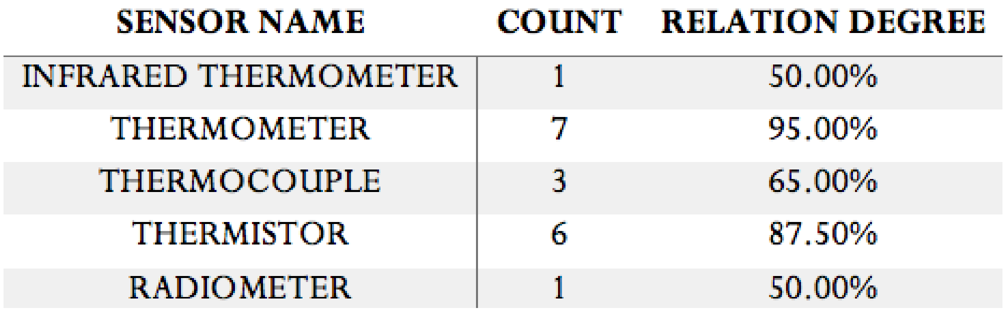
$$P(H_i|D_i) = \frac{P(H_i)P(D_i|H_i)}{P(D_i)} \tag{3.4}$$

This type of problem could be solved using Bayes' Theorem, however it is only solvable if a small number of sensors are listed in the database. As mentioned in section 3.6.1, our database maintains a total of 139 unique sensors and our Knowledgebase does not know about any sensor-phenomena relationships beforehand. Therefore, worse case we have to assume that either every sensor is equally likely ($P(H_i) = 1/139$) to be qualified to provide information about a sensor or that no sensor is qualified at all by setting $P(H_i)$ equal to a value near 0 (e.g. $P(H_i) = .000001$). One can choose either assumption to be true, however since the total number of sensors that Sensite tracks is so large that assuming equal likelihood for $P(H_i)$ is no different than the later assumption. $P(H_i)$ will be near 0 in both cases and the MLE $P(H_i|D_i)$ will not grow by a significant amount due to the low probability prior believe. Given this, Bayes' Theorem cannot be applied to this problem.

Frequency Based Statistics: We tried the simple approach of creating a histogram of the frequency count for all sensors s_i that are associated with a phenomenon p_i, and rank the associated sensors according to the strength of their relevance (ie: greatest count). In order to do this, the data needed to be normalized and then sorted. We then set a threshold value of .8 to filter out the low relevance sensors. This approach worked well when the number of observations was small. However, as new data D_i appears, the total number of observations grow causing the ratio to grow smaller and smaller to the point where eventually some relevant relationships filter below the threshold and are omitted. Consider the sensor-phenomenon relationship for temperature denoted in section 3.6.2.4. Lets assume that the total number of observations for temperature grows large (say 10,000), and the majority of the readings become skewed towards the “thermometer” (8,967 readings) such that a long tailed distribution occurs when comparing the remaining distribution among the other sensors. There is no transform that one can apply to cap the “thermometer” readings and re-balance the distribution so that the skew no longer exists. Therefore, we abandoned this approach.

Relationship Degree Normalization: We studied the pros and cons of the aforementioned approaches and devised an algorithm that effectively solves the problem. The goal of our algorithm is to ensure that we can effectively quantify the relationship between the sensors without any disproportionality as the number of observations grow. Ideally, we want an algorithm that gives equal likelihood to any sensor found to be associated with a phenomenon, and 0 likelihood to sensors that are not found to be associated at all. We do this by describing an algorithm that we describe as the “relationship degree”. We define the sensor-phenomena relationship degree as

$$R_{degree} = \frac{C(s_1^n)_{total} - C(s_1^n)_{min}}{C(s_1^n)_{max} - C(s_1^n)_{min}} \times \alpha + \beta \tag{3.5}$$



SENSOR NAME	COUNT	RELATION DEGREE
INFRARED THERMOMETER	1	50.00%
THERMOMETER	7	95.00%
THERMOCOUPLE	3	65.00%
THERMISTOR	6	87.50%
RADIOMETER	1	50.00%

Figure 3.11: Table describes a breakdown of the computed relationship degree for our proposed algorithm.

where s_1^n is shorthand for all sensors, $C(..)_{total}$ is the total number of observations, and $C(..)_{min|max}$ is the minimum and maximum sensor observation count respectively. α and β are paramaters that control the range of the weights for the relationship degree. We set $\alpha = .45$ and $\beta = .5$ because it constrains the range of R_{degree} degree from .50 to .95. β is set to this value because if D_i describes an assumed true sensor-phenomena relationship, then that sensor should have equal likelihood of association. As the counts increase, the weights will change accordingly but will always remain between .50 to .95. Figure 3.11, gives an example of the a visual description of the relationship degree. The following sensors are associated with the phenomenon temperature. Sensors with a single observation count have a 50% relevancy association, while the others with larger counts have a greater score. If no sensor is associated with a phenomenon, then that means the Knowledgebase did not find any webpage that describes the relationship and that sensor will not be in the result set when the user queries for a particular phenomenon. This algorithm was implemented in the Sensite Knowledgebase platform, with a default threshold for R_{degree} of .8. Any sensors that have 80% or more relevance are returned from the user's query.

3.7 Experiment

There is no SSW sensor database that we can benchmark our platform against, because current commercial systems require the user to manually query for sensor data feeds. This is

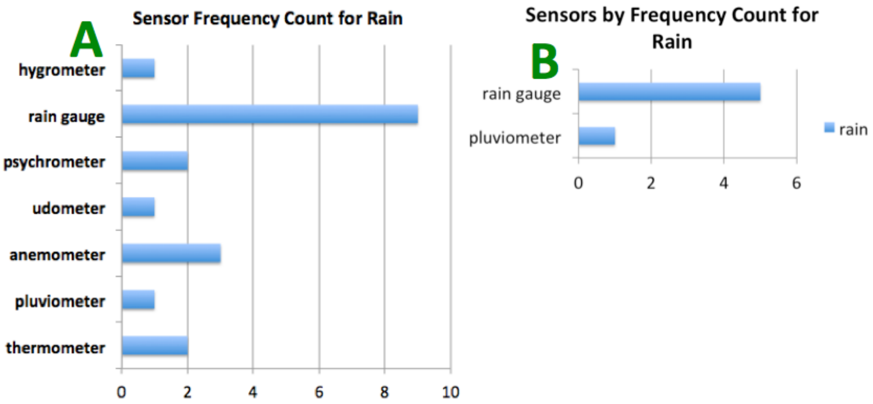


Figure 3.12: A) The sensor-phenomenon relationship accuracy is 42% (3 of 7 correct). However, a user's query will return rain gauge as the 1st rank and anemometer 2nd rank, which is incorrect. An anemometer measures wind speed. B) The accuracy of the sensor-phenomenon relationship for rain is 100%. A user query will return rain gauge as 1st rank and pluviometer 2nd rank. Incorrect sensors are omitted.

the exact thing that we are looking to mitigate the user from doing. Therefore, we compared the veracity of our proposed algorithm against an algorithm that took all mentioned sensors and phenomena in a webpage and stored the permutation of the relationships. Justification for storing the permutations is the assumption that the content of a typical webpage is usually focused on a main topic. Therefore assuming the law of large numbers, the true relationships will surface as the number of webpages visited increase. Figure 3.12A denotes the sensor to phenomena relationships found by permutation. Figure 3.12B denotes the relationships using our proposed algorithm. The results show that applying our algorithm significantly reduces number of incorrect sensor associations and greatly improves accuracy as a result. This is largely in part to the fact that complete sentences represent a complete thought. Therefore, by omitting run-on and incomplete sentences that can be produced by ads or hyperlinks appended to valid text we mitigate the error.

3.8 Discussion

To the best of our knowledge this is the first SSW sensor data platform that applies an unsupervised learning algorithm in order to understand the relationships between a sensor and the many phenomena that it is qualified to provide information about. Our Sensite platform enables the user to perform spatiotemporal queries for sensor data according to the phenomena, location and time of the readings of interest. A user interested in publishing sensor feeds to our platform does not need to bother tagging the sensor with arbitrary keywords. They simply need to annotate their sensor data with our QoI/VoI data model and include the sensor type, location and timestamp. Sensite will organize the sensor information and provide it to the community. We envision this type of sensor data platform to be the next generation SSW sensor data platform.

CHAPTER 4

Use Case: Ultraviolet Guardian Health & Wellness Application

4.1 Introduction

In 1987, former President Ronald Reagan had basal cell surgically removed and he survived. In 1977, Reggae legend Bob Marley, was diagnosed with malignant melanoma in the late stage and he did not survive [48,49]. It is projected that 1 in 5 Americans will develop some form of skin cancer in their lifetime [50]. Moreover, both Dermatologists and Epidemiologists that we have interviewed have confirmed that no concrete data exists to identify specifically what ultraviolet (UV) exposure patterns over time lead to a specific skin cancer type. Why? Because conventional dosimeters, which turn color shades to quantify personal UV exposure levels cannot capture the time in which the instantaneous exposure occurred. We have developed a mobile application that tracks your fine grain UV exposure without needing a UV sensor, provides recommendations to protect the user from Sun over-exposure, and allows the user to compare their relative exposure to others in their social circle. Our studies prove that our application can accurately estimate UV dosage for body parts, such as the vertex of the head, shoulders and feet that are always exposed to the sun (assuming the individual is standing upright). This fine grain spatiotemporal information is valuable for Epidemiologists to provide new insights about skin cancer, to influence Dermatologists, so that they can provide better care to their patients.

4.2 Background on Ultraviolet Solar Radiation (a basic review)

4.2.1 Ultraviolet Electromagnetic Spectrum

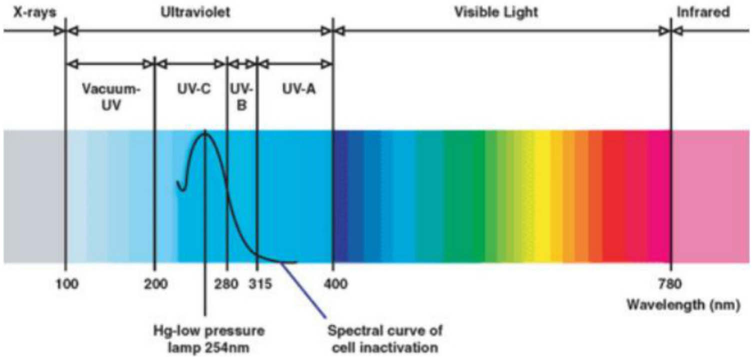


Figure 4.1: Spectral wavelengths emitted from the sun.

The UV portion of the solar spectrum (shown in Figure 4.1) plays an important role in many processes in the biosphere, such as initiating photosynthesis in plants. The UV spectrum is divided into three wavelengths; UVA 315-400nm, UVB 280-315nm, and UVC 200-280nm. Human skin cells are highly sensitive to UVC, however these wavelengths are blocked by the ozone. UVB wavelengths pass through the atmosphere, and are primarily responsible for skin reddening by damaging the epidermal layer. The skin is least sensitive to UVA wavelengths and is responsible for tanning by damaging the dermal layer. There are several benefits to sunlight, however may also be very harmful if an individual's personal UV exposure levels exceed "safe" thresholds. Some of the evidential effects include dark spots, tanning, redness of the skin, chaffing and carcinoma. To avoid effects such as these, people should take protective measures and limit their exposure to high solar radiation.

4.2.2 Atmospheric Properties

The diurnal and annual variability of solar UV radiation reaching the ground is governed by astronomical and geographical parameters and atmospheric conditions. Pollution such as carbon monoxide generated by humans also affect the atmosphere (see Figure 4.2 [51]) and the level of UV radiation reaching the ground. Roughly 30% of the radiation striking Earth's

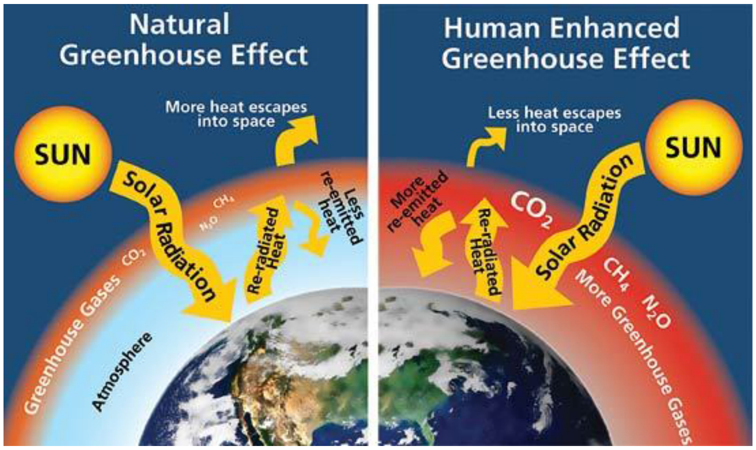


Figure 4.2: Left: Naturally occurring greenhouse gases; carbon dioxide (CO2), methane (CH4), and nitrous oxide (N2O) normally trap some heat from the Sun, keeping the planet from freezing. Right: Human activities, such as the burning of fossil fuels, increase greenhouse gas levels, leading to an enhanced greenhouse effect. The result is global warming and gradual climate change.

atmosphere is immediately reflected back into to space by clouds, ice, snow, sand and other reflective surfaces. The remaining 70% of incoming solar radiation is absorbed by the ocean, land and atmosphere [52]. As they heat up, the ocean, land and atmosphere release heat in the form of IR thermal radiation, which passes out of the atmosphere and into space. Cities and industrial areas typically release a higher amount of ozone affecting pollution, which in effect increases the ground level UV radiation.

Earth's atmosphere, shown in Figure 4.3, is roughly 600 kilometers (372 miles) from Earth's surface. The atmosphere absorbs the energy from the Sun, recycles water and other chemicals, and works with the electrical and magnetic forces to provide Earth's climate. The atmosphere also protects us from high-energy radiation and the frigid vacuum of space. The atmosphere is made up of the following:

***Troposphere*:** The troposphere extends from the Earth's surface to about 15 kilometers (9 miles) high. The trophosphere is the densest, and the temperature drops from about 17 to -52 degrees Celsius as you climb higher in this layer. Almost all weather is in this region.

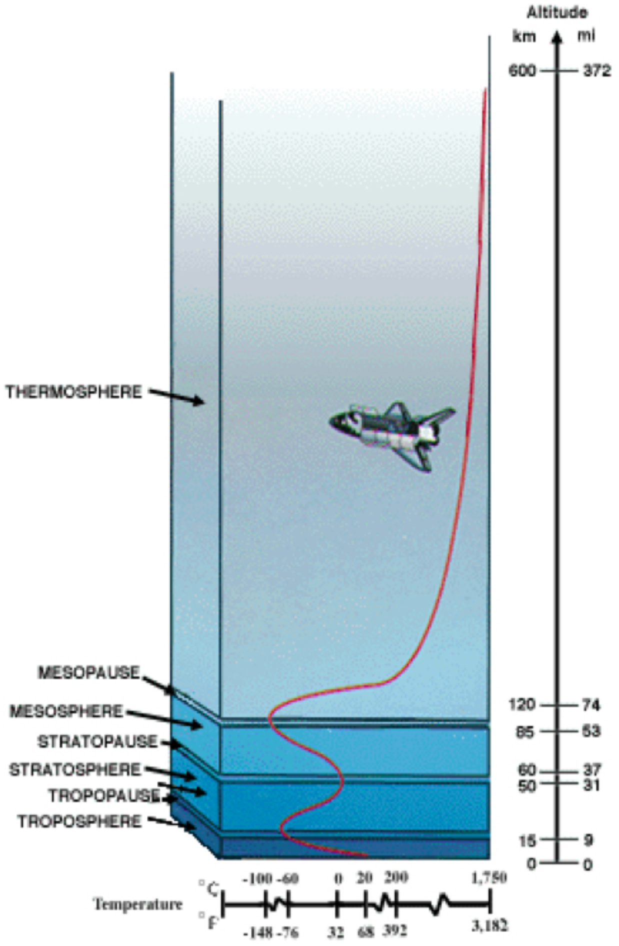


Figure 4.3: Earth's atmosphere is comprised of layers that work together to shield out harmful ultraviolet rays.

The tropopause separates the troposphere from the next layer. The tropopause and the troposphere are known as the lower atmosphere.

***Stratosphere*:** The stratosphere starts just above the troposphere and extends to 50 kilometers (31 miles) high. The stratosphere is dry and less dense. The temperature in this region increases gradually to -3 degrees Celsius due to the absorbtion of UV radiation. The ozone layer, which absorbs and scatters the solar UV radiation resides in the stratosphere. The stratopause separates the stratosphere from the next layer.

***Mesosphere*:** The mesosphere starts just above the stratosphere and extends to 85 kilometers (53 miles) high. In this region, the temperatures again fall as low as -93 degrees Celsius as you increase in altitude. The molecules residing in this layer absorb a lot of the Sun's shortwave (eg: xray energy), which cause the chemicals in this layer to be in an excited state. The mesopause separates the mesosphere from the thermosphere.

***Thermosphere*:** The thermosphere starts just above the mesosphere and extends to 600 kilometers (372 miles) high. The temperature goes up as you increase in altitude due to the Sun's energy. Temperatures in this region can go as high as 1,727 degrees Celsius. Chemical reactions occur much faster here than on the surface of the Earth. This layer is known as the upper atmosphere.

4.2.3 How is Solar Radiation Measured?

As stated in section 4.2.1, the UV spectrum is divided into three parts. The blue line in Figure 4.5 represents the typical intensity at each UV spectral wavelength at bandpass frequency. Since UVC is mostly blocked by the ozone, the intensity level for those spectral wavelengths are negligible and is reflected by the downward trend (although the UVC wavelength is not directly shown in the figure). UVA and UVB are the primary wavelengths that reach the pedestrian level.

The need to educate the public about UV exposure and its detrimental effects led scientists to define the ultraviolet index (UVI), a parameter used to indicate the severity of prolonged UV exposure for a typical Caucasian (see Figure 4.4). The erythmal weighting

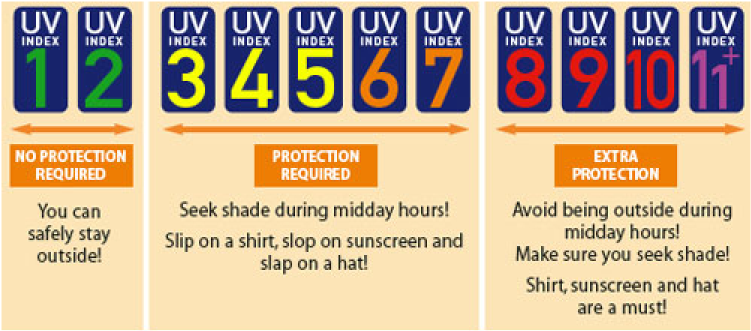

UV INDEX 1
UV INDEX 2
UV INDEX 3
UV INDEX 4
UV INDEX 5
UV INDEX 6
UV INDEX 7
UV INDEX 8
UV INDEX 9
UV INDEX 10
UV INDEX 11+
NO PROTECTION REQUIRED
You can safely stay outside!
PROTECTION REQUIRED
Seek shade during midday hours!
Slip on a shirt, slop on sunscreen and slap on a hat!
EXTRA PROTECTION
Avoid being outside during midday hours!
Make sure you seek shade!
Shirt, sunscreen and hat are a must!

Figure 4.4: The ultraviolet index (UVI) describes the level of solar UV radiation at the Earth's surface and ranges from 1 to 11+. The higher the UVI, the greater the potential for damage to the skin and eyes and the shorter the burn time.

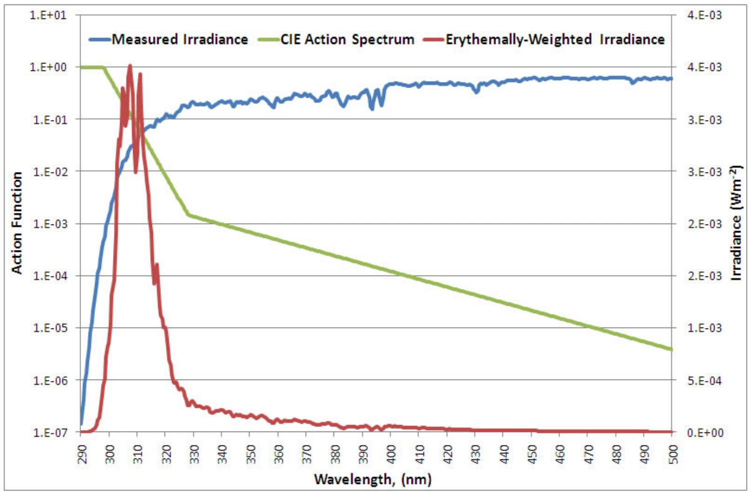


Figure 4.5: Sample data showing instantaneous UV irradiance levels at a single time instance. The blue line represents the UV irradiance measured by a bandpass spectroradiometer. The green line represents the erythemal weighting factor for each individual UV wavelength. The red curve represents the product of the green and blue lines to produce an erythemally effective total UV dosage the skin would receive per unit time.

factor, denoted by the green line in Figure 4.5, was originally proposed in 1987 by McKinlay and Diffey [53], and adopted as a standard by the International Commission on Illumination (CIE) in 1992. Total UV irradiance (ie. radiation) represents the cumulative energy that the skin receives per unit area and is defined by

$$UVir_{total} = \int_{200nm}^{400nm} I(\lambda)\omega(\lambda)d\lambda \tag{4.1}$$

where λ is the individual spectral wavelength measured in nanometers (nm), $I(\lambda)$ is the measured intensity at λ in (mW/m^2), and $\omega(\lambda)$ is a weighting function denoted by the green line in Figure 4.5 defined by

$$\omega(\lambda) = \begin{cases} 1 & 200 < \lambda \leq 298 \\ 10^{0.094(298-\lambda)} & 298 < \lambda \leq 328 \\ 10^{0.015(139-\lambda)} & 328 < \lambda \leq 400 \\ 0 & 400 < \lambda \end{cases} \tag{4.2}$$

that represents the sensitivity level of the typical Caucasian skin individual to reach erythema (sunburn) given I(λ). UV sensitive spectrometers perform integration typically over $\triangle\lambda = 0.5nm$ resolution in order to compute the total erythemally effective UV dosage that an individual would receive (denoted by the red line in Figure 4.5) at the ground level. This is also the total energy that the ground receives over time generally in the units of mW/sm^2. The UVI can finally be derived by the following equation

$$UVI = \lceil \frac{1}{25} * UVir_{total} \rceil \tag{4.3}$$

4.2.4 Sun Position

Figure 4.6 describes the geometric angles used to calculate the sun's hemispherical position. However, the actual computation is not as seemingly simple as the figure depicts. Many researchers have proposed various complex algorithms to calculate sun position [54–56]. Since the accuracy of each of the proposed algorithms are within 0.01° accuracy, we decided to use [54] for UVG since its accuracy is sufficient.

Solar elevation (ie: altitude) is the angle α between the horizon (as viewed by the person) and the Sun. The *solar zenith angle* (SZA) φ is often used in place of α, and is measured

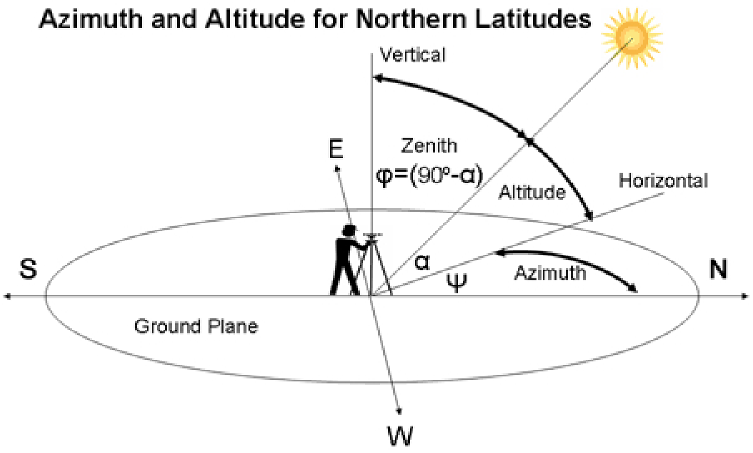


Figure 4.6: The position of the sun in the sky can be determined by the altitude θ and azimuth ϕ angles.

by $\varphi = 90° - \alpha$. The *azimuth angle* ψ is the angle between the horizontal direction of the Sun towards the horizon and the south direction of the Earth.

4.2.5 How is UV Irradiance Measured?

Initially, UV sensors located on top of base stations in sparse locations first measured UV irradiance across the globe, however coverage was limited to a small area. Later, the TOMS satellite became the conventional method for estimating UV irradiance due to its ability to cover roughly a $100km$ x $100km$ area. TOMS measures the amount of UV back-scattering from the Earth's atmosphere. In addition to considering cloud reflexivity, ozone, ground albedo and extra-terrestrial solar irradiance as parameters to estimate the ground level UV irradiance over a large area [57–59]. Measurements are taken once a day during solar noon and inaccurate estimations occur due to incorrect interpretation of environmental properties [58, 60–62]. Kalliskota et al [61] performed an extensive comparison of the calculated daily UV dosage between TOMS vs. ground based sensors. Results conclude that TOMS occasionally classifies snow as cloud cover causing an under estimation of ground level UV irradiance. In addition, the weekly average UV irradiances reported from TOMS are about 20% higher than ground based measurements during the summer. Their results show sup-

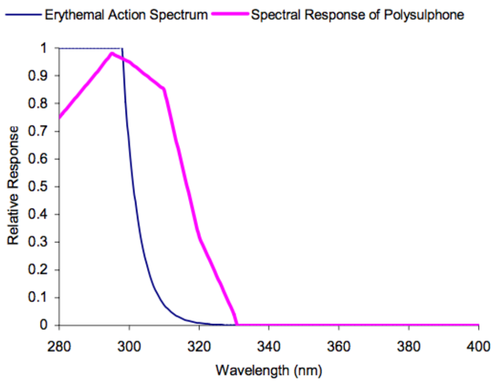


Figure 4.7: Comparison of the CIE erythemal action spectrum vs. the spectral response curve of polysulfone film. The curves do not align exactly, however industry experts agree that polysulfone provides a reasonable approximation for estimating UV dosage. Benchmarking the reading against a certified spectroradiometer improves accuracy further.

porting evidence that ground based UV measurements are more accurate. Moreover, ground based UV measurements are the only measurements reported in real-time and account for sudden environmental changes (eg: rain, thick clouds). As a result, we rely primarily on ground level UV sensor readings for UVG.

4.2.6 Polysulfone Dosemeters

Davis et. al. discovered that polysulfone and polyphenylene oxide darkened when exposed to UV radiation [63,64] while evaluating weathering characteristics of plastics. The material had the greatest responsivity primarily within the UVB electromagnetic spectrum, with peak sensitivity at the 300nm wavelength, and minimum sensitivity at 330nm. The spectral response of polysulfone (PS) film does not match exactly the erythemal action spectrum, as shown in Figure 4.7, however can be correlated by calibration against a spectroradiometer. This calibration should be performed seasonally in order to maintain a more precise estimate given UV irradiance levels vary with the season.

PS films have been the most widely adopted dosimeters to assess personal UV exposure [65–68] due to the similar spectral response characteristic of the erythemal action spectrum.

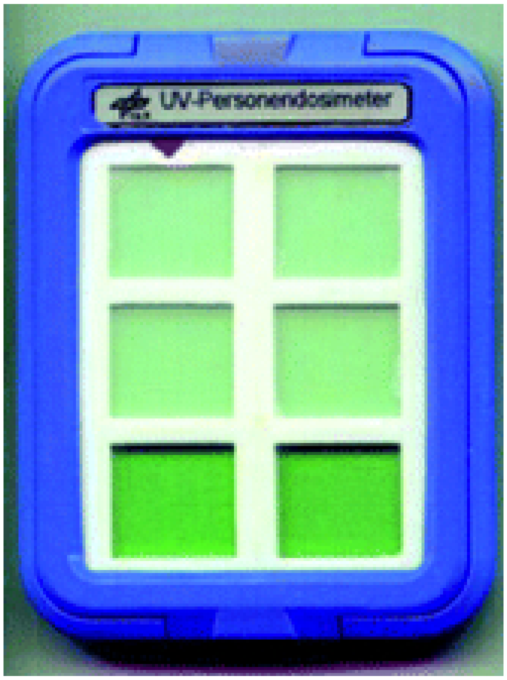


Figure 4.8: Polysulfone film dosimeter.

Users typically mount the film in a small cardboard holder (or plastic as in Figure 4.8), and commonly clipped on the lapel of a person's garment, arm, or shoulders. These devices are suitable for measuring anatomic body site specific cumulative exposure over a period of one day to one week.

The effectiveness of UV radiation leading to erythema is usually expressed in terms of the weighting function described in equation 4.2. When exposed to solar UV, the diphenyl sulfone group in polysulfone absorbs UV at wavelengths shorter than 330nm and undergoes a visual color change resulting in an increase in optical absorbency (Davis et al., 1976) due to energy of the UV wavelengths absorbed over time. PS dosimeters are capable of reasonably quantifying UV exposure, however are not able to store spatiotemporal information about the instantaneous UV exposure observation.

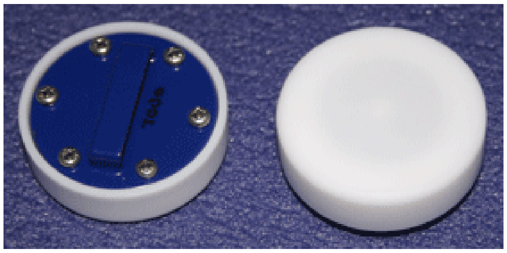


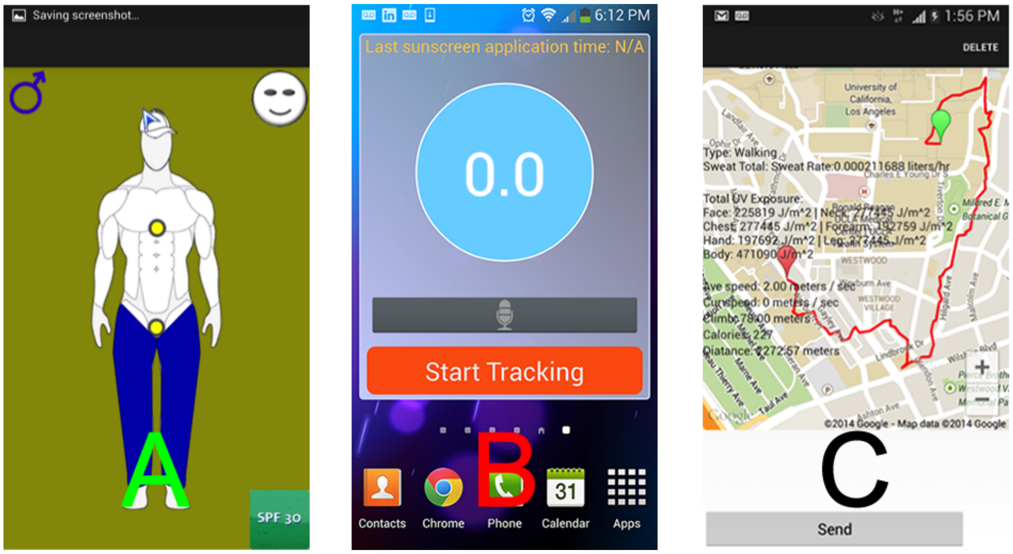
Figure 4.9: Digital UVB dosimeter manufactured by NIWA used for our extended experiments estimating body site UV exposure. One of the sensors used in our experiments.

4.2.7 Digital Dosimeters

In 1995, Diffey et. al. [69] proposed an embedded device that incorporates a miniature UVB sensitive sensor and a data logger that could be clipped to the lapel or on a waste belt in order to estimate cumulative UV exposure. Their device would be cumbersome to carry, however as technology evolved the digital dosimeter began to shrink in size, such as the one in Figure 4.9. The digital dosimeter in the figure was one of the sensors used for our experiments.

Digital dosimeters with integrated data loggers solve the problem of recording the time in which the instantaneous exposure occurred. However, similar to PS dosimeters, digital dosimeters are only capable of quantifying an exposure level for the single body site to which it is affixed, and these dosimeters cannot store the geographic location where the instantaneous reading was observed. We solve this problem by developing UV Guardian (UVG), a mobile application that tracks fine grain UV exposure comparable to a digital dosimeter, provides recommendations to protect the user from sun over-exposure, and allows the user to compare their relative exposure to others in their social circle. This fine grain spatiotemporal information is valuable for skin cancer researchers to provide new insights about skin cancer, to influence Dermatologists, so they can provide better care to their patients.

4.3 UV Guardian System Overview



Type: Walking
Sweat Total: Sweat Rate:0.000211688 liters/hr

Total UV Exposure:
Face: 225819 J/m^2 | Neck: 277445 J/m^2
Chest: 277445 J/m^2 | Forearm: 192759 J/m^2
Hand: 197692 J/m^2 | Leg: 277445 J/m^2
Body: 471090 J/m^2

Ave speed: 2.00 meters / sec
Curspeed: 0 meters / sec
Climb: 78.00 meters
Calories: 227
Diatance: 2272.57 meters

Figure 4.10: A) is the user profile screen where the user can inform UVG of their body exposure, sunscreen application and physical attributes. B) is a voice activated widget to control the activity monitor. C) is the path tracking feature showing an experiment where the participant performed a roughly 1 mile jog from Weyburne terrace to UCLA's campus during mid-day, with a digital UV dosimeter and the UVG mobile application affixed to their arm.

UV Guardian (UVG) shown in Figure 4.10 is a mobile application that helps establish fine grain UV exposure and skin cancer correlation. Moreover, UVG protects the user from sun over-exposure, while providing recommendations to enjoy sun light benefits such as Vitamin D. We propose a technique for tracking and estimating the UV exposure of a pedestrian traveling along a path in a mapped urban environment. The estimate can be computed either before or during the actual walk. If the user (e.g. a runner) affixes the smart phone facing frontward (e.g.: arm band) so that the sensor can measure the ambient light, fine grain UV exposure can be tracked by feeding the sensor light measurements to a model that maps light intensity to UV irradiance intensity. If the user (like many of us) keeps the phone in their pocket or purse, then exposure is computed using a more elaborate model that correlates travel path, environmental context (i.e.: buildings, trees) and sample UV irradiance readings. The latter method is also used to estimate exposure before the walk.

4.4 Prior Work

Our previous work [70] described a similar personal UV monitoring mobile application tethered to a Bluetooth UV sensor device. The mobile application recorded UV exposure information in the form of storing observed UVI readings taken per second as the pedestrian traveled outdoors. The application also implemented an algorithm to recommend the amount of time (in minutes) that the pedestrian should be outdoors before becoming over-exposed given their skin type and applied SPF suntan lotion. However the algorithm does not consider environmental properties that increase or reduce UV exposure time (eg: trees, buildings, pools). Since [70], our application has been refactored to become predictive, providing recommendations based upon real-time exposure and considering environmental context information such as indoors, outdoors and shade. To include, we leverage the QoI/VoI metadata data model (see section 2.4) to upload crowdsourced UV irradiance readings to our Sensite platform. UVG queries our Sensite sensor data platform (see chapter 3) in order to obtain the latest UV irradiance information according to the geography of interest.

4.5 QoI/VoI Metadata for Ultraviolet Guardian

4.5.1 Crowdsourcing Sensor Data with Personal Smart Weather Stations

Recently, mobile app enabled personal smart weather stations such as BloomSky [71] (outdoor station for the home and office), CliMate [72] and StormTag [73] (compact station carried on the person) are making headway into the market. These devices are designed to be affordable around $20.00 and stream weather information such as temperature, UV, humidity, rain and barometric pressure to arbitrary sources (including the mobile phone). These devices are the realization of the semantic sensor web (SSW) platforms we discussed in [18], and also in section 1.1.1. In the near future, these devices will become commonplace

Figure 4.11: Source: www.bloomsky.com
Envisioned picture of crowdsourced SSW of weather sensors reporting local weather information in their local environment, and streaming the sensor data to cloud based sensor data platforms, such as Sensite and the end user.

(as depicted in Figure 4.11), and UVG will transition from a research study to a commercially viable application that enables people to track their UV exposure and Vitamin D intake in real-time. These devices can easily upload their various sensor data to our Sensite platform by simply wrapping the data stream with our QoI/VoI data model format discussed in section 3.4.

4.5.2 Implementation of the Ontology Based QoI/VoI Data Model

In order for UVG to properly estimate spatiotemporal UV dosage, the following QoI attributes must accompany each UV irradiance datum point:

- Timeliness: Sensor data aggregated within 30 minute sliding window
- Completeness: Data must include GPS, UV irradiance and time
- Provenance: Transformation of UV Irradiance to UV Index (vice versa) (if applicable)

The format of the data is in Javascript Object Notation (JSON) and follows a similar structure as our dengue detector mobile application, described in Figure 5.15, with the exception of the *sensorType* attribute set to "uv_irradiance".

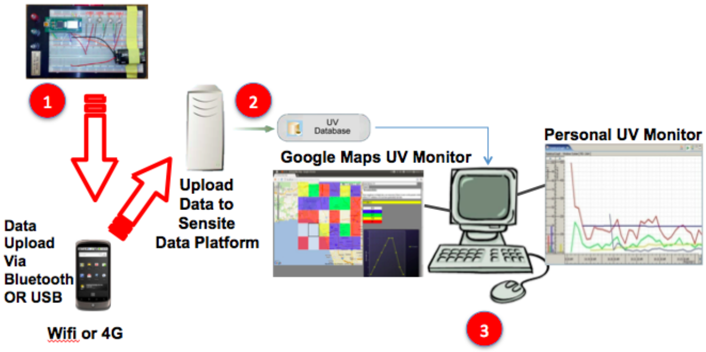


Figure 4.12: 1) UV irradiance measured by a UV sensor as the pedestrian travels. The sample is transmitted to the UVG Android Mobile Application. 2) Periodically, the most recent sample is uploaded to the Sensite sensor platform server from the mobile phone. 3) Spatiotemporal UV Irradiance information is queried from Sensite for analytics and viewable through website.

Volunteers wearing a UV sensor or users of the commercially available personal smart weather stations, such as the devices mentioned in section 4.5.1 can upload instantaneous sample UV irradiance information to the Sensite sensor platform. The UVG website performs a spatiotemporal query for sample UV irradiance data and UVG constructs a real-time UV irradiance map as a function of space and time relative to the Sun's position through participatory sensing, as shown in Figure 4.12.

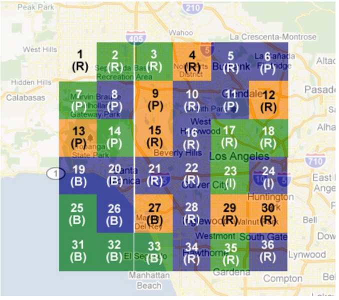


Figure 4.13: 36km x 36km region of Los Angeles centered at Latitude 34.06064 / Longitude -118.409271 is divided into 6km x 6km non-overlapping local habitats. Letters represent the habitat type.

The UV irradiance map is bounded and segmented according to the regions of interest denoted by the colors shown in Figure 4.13. Pedestrians periodically upload a randomly sampled GPS location, time-stamp, and UV irradiance reading as they travel outdoors using an Android mobile device running the UVG mobile application tethered to a Bluetooth enabled UV sensor. Later versions of the application leveraged a board with USB tethering capability. The UV sensor is not required in order for UVG to estimate their personal UV exposure, however users who carry the sensor contribute to constructing the global UV irradiance map. Before UVG uploads the sample UV reading, the datum point is annotated with QoI/VoI metadata (see section 2.4) to identify the data source, timestamp, location and confidence of the reading with respect to the prior. For instance, if our environment classifier algorithm identifies the environment as being "in_doors" and a sample UV irradiance reading is high, then the confidence will be low. Our experiments have shown that there is virtually no UV indoors. Following data annotation, the crowdsourced UV irradiance data is uploaded to a central server and aggregated within a time window until a certain tolerance is reached based upon the maximum population density within the region of interest. The data is then segmented according to the finite geographic region that the GPS location falls under to produce a model that maps the average UV irradiance for the region as a factor of the time of day. We also assume that the altitude angle of the sun is congruent with time. Factors that may affect the veracity of our model are sudden atmospheric property changes or thick clouds. However, the model will update itself as more real-time data is crowdsourced. In this paper, collective regions will be referred to as "habitats", and a particular region will be referred to as a "local habitat" from this point forward.

4.6 System Model

Whether the pedestrian takes a short walk or a long jog, they are subject to receiving a level of UV radiation proportional to their surrounding environmental properties. If two pedestrians travel the same distance and velocity, however one chooses a path with heavy vegetation and the other under direct sunlight. The pedestrian traveling under vegetation will receive less UV dosage.

In this thesis, we highlight a study performed in order to answer the following questions: *Can the pedestrians' total UV irradiance traveling outdoors be reasonably estimated?* To answer this question environmental context information is required. This section discusses the proposed algorithm for estimating pedestrian UV exposure along a path and its accuracy compared to the observations.

Estimating UV exposure requires a massive amount of environmental context information to be accessible to UVG. This includes information about the location and dimension of streets, buildings, and vegetation cover. The pedestrians path can be represented as a summation of line segments, where the total UV exposure E_{total} along each segment is defined by Equation 4.4 (in units mW/mm^2).

$$E_{total} = \int \omega(s,t)\ dt \tag{4.4}$$

The UV irradiance per unit time ω is a function of position s and time t (in seconds) relative to the Sun's position. The average UV irradiance per unit time $\bar{\omega}$ (in units mW/mm^2 per second) can be calculated by

$$\bar{\omega} = \frac{\int \omega(s,t)dt}{\int dt} \tag{4.5}$$

The pedestrians velocity is assumed constant and known. Environmental context information, such as tree dimension and leaf cover density along the path are also assumed known. Given these assumptions, Equation 4.4 can be represented by

$$E_{total} = \int \omega(s,t)\frac{dl}{V} \tag{4.6}$$

where $\int dl$ is the total walking distance. If the pedestrian travels a fraction of the total distance under a tree L_{tree} and the remainder in direct view of the Sun L_{open}, E_{total} can be expressed as:

$$E_{total} = \omega_{open}(s,t)\frac{L_{open}}{V} + \omega_{tree}(s,t)\frac{L_{tree}}{V} \tag{4.7}$$

where ω is the average observed UV irradiance per unit time within the local habitat under a tree $\omega_{tree}(s,t)$ or the Sun $\omega_{open}(s,t)$.

The error of the algorithms estimation can be calculated by

$$\tau = \frac{E_{total,obs} - E_{total,est}}{E_{total,obs}} \tag{4.8}$$

where $E_{total,obs}$ is the pedestrians observed total UV exposure, and $E_{total,est}$ is the pedestrians estimated total UV exposure by the algorithm. UVG assumes $\bar{\omega}$ to be a reasonable factor for approximating the instantaneous UV dosage that a pedestrian receives per unit time (per second), because the variance of the horizontal (ambient) instantaneous UV irradiance follows a Poisson distribution centered around $\bar{\omega}$. Given this, the error τ in the algorithms estimation is what we use as a way to express the deviation between our UV dosage estimate vs. the actual measurement, and the accuracy is measured by the following equation

$$Acc = 1 - \tau \tag{4.9}$$

We use the T-distribution to model the error distribution, because it is appropriate for smaller sample sizes. Since integration time for calculating cumulative UV dosage is per second, the unit for τ is also in seconds. In other words, τ is simply expressing the number of seconds that our UV exposure estimation algorithm $E_{total,est}$ deviates from the actual measurement.

4.7 Experiments

Our experiments were performed in two phases; In phase 1, described in section 4.7.1, random samples were collected to understand how UV radiation varies across a $36km$ x $36km$ region of Los Angeles centered at latitude 34.06064 / longitude -118.409271. We also collected UV irradiance samples under randomly selected trees with canopies that provided shade from the sun. Next, we used the collected data to perform experiments to gauge whether a linear model can be applied to reasonably estimate UV irradiance for the top of the head and shoulders given a travel path. In phase 2, described in section 4.7.3, we used the digital dosimeter from Figure 4.9 due to its convenient size in order to measure cumulative UV dosage along a travel path, then measured the correlation between UVG's cumulative dosage estimate and the digital dosimeter's for estimating the body site UV dosage on the

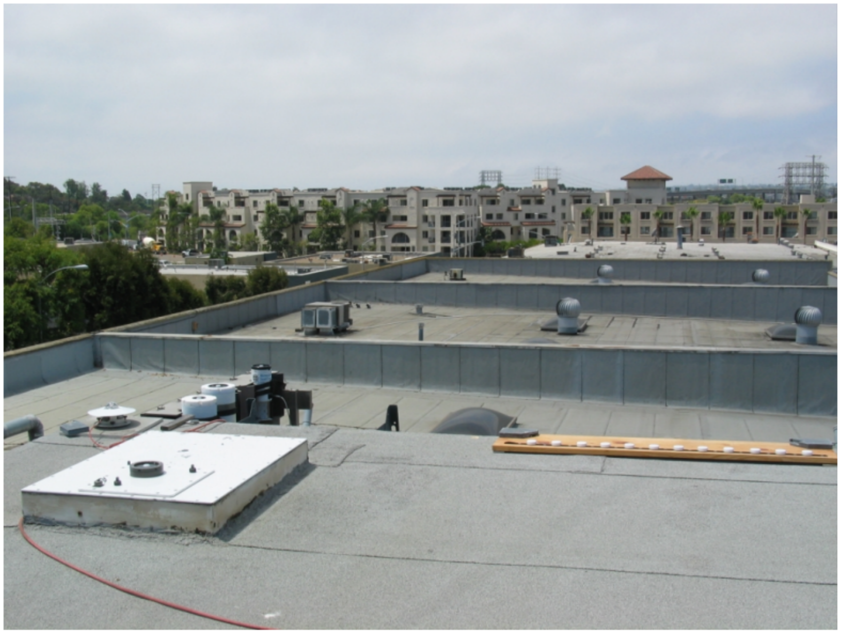


Figure 4.14: Roof platform at Biospherical Instruments, Inc with the top of the SUV-100 spectroradiometer on the left side and the board with all UV sensors on the right side. The collector of the SUV-100 spectroradiometer is the round object protruding the white box, which contains the SUV-100 instrument.

arm.

The typical user of UVG is a Sportsperson who is outdoors during the mid-day performing an outdoor recreational activity, such as jogging, biking or hiking where the user typically carries their phone. We began by shadowing these individuals and documenting where they placed their cellphones as they performed their activity, and discovered that the majority placed their phone on their arm and also had their arms exposed. The secondary benefit is we can leverage the sensors on the phone. Given this we chose to initially validate whether our algorithm can estimate UV dosage for the arm, and the output from the digital dosimeter is the control. The NSF ICORPS Research Grant IIP-1340385 supported phase 2 of this work.

The sensors used in our experiments were calibrated against a NIST certified SUV-100 spectroradiometer manufactured by Biospherical Instruments, Inc. Before calibration, their dynamic range was adjusted such that measurements would not be saturated when exposed to UV radiation levels encountered in California (as shown in Figure 4.14). When comparing the UVA and UVB sensors against each other respectively, each sensor reported consistent readings with minimal differences between devices.

4.7.1 Phase 1 - How does UV Radiation vary across a Large Geographic Area?

In [74] we studied the variation of UV irradiance levels across local habitats (ie: geographic regions). A two-stage cluster sampling method was applied to collect sensor data. The UV sensor used in this experiment was two pairs of Bereich Mikrotechnik JIC 119 UVA and JIC 129 UVB sensors. The sensors were connected to a uIceBlue2 Bluetooth programmable micro-controller and were programmed to report a single UV irradiance measurement per second once activated. The UV irradiance measurement is produced by averaging a total of 10 readings taken within the first 500ms, while holding the UV sensor parallel to the ground (ie. hemispherical view) in direct sunlight.

How does UV irradiance vary across local habitats? A two-stage cluster sampling method was applied in order to answer this question. At the first stage, a 36km x 36km region of Los Angeles centered at 34.06064 (latitude), -118.409271 (longitude) was divided into 6km x 6km non-overlapping clusters of local habitats as shown in Figure 4.13. In general, larger cluster sizes typically possess more heterogeneous elements and require larger samples to accurately estimate the population parameter. Conversely, smaller cluster sizes contain more homogeneous elements and require smaller sample sizes to estimate the population parameter. Significant changes in UV irradiance occur on an hourly basis, therefore the boundary size of the local habitats and the number of samples collected were chosen based upon the time to reach the sample locations within a reasonable timeframe.

Experiments were performed to measure the variance of UV irradiance under direct sun light at the pedestrian level both within and across the local habitats. The following nuisance factors in the environment were omitted; trees, pavement, and buildings. At the first stage, habitats 8, 15, 20, 21, and 23 were randomly chosen for observation. At the second stage, five

Table 4.1: Observed average UV irradiance from the Sun and deviation across randomly sampled local habitats

Habitat	Wave	$\bar{\omega}$	σ_{sample}
8	UVA	610.4222	5.055988
	UVB	705.8477	0.5335854
15	UVA	612.8682	3.485185
	UVB	706.059	0.1255518
20	UVA	608.8371	6.360273
	UVB	705.824	1.382525
21	UVA	608.8753	7.759923
	UVB	707.268	0.722
23	UVA	610.5026	1.883543
	UVB	707.2613	1.380255

Table 4.2: Overall average UV irradiance and deviation across the sampled habitats

Wave	μ_{pop}	σ_{pop}
UVA	610.3011	1.644800
UVB	706.452	0.747468

clusters were randomly chosen from each of the selected local habitats for sampling. Since the average UV irradiance measured within the sampled clusters were small, the cluster samples were treated as a collection of random samples in order to estimate the average UV irradiance per unit time $\bar{\omega}$ (see section 4.6) within the local habitat. The ideal minimum number of pedestrians equipped with the UVG application that is required to estimate $\bar{\omega}$ within $\pm 2mW/mm^2$ for both UVA and UVB, with 95% confidence are approximately 58 and 2 respectively. These values were obtained by using a Z score of 1.96 (95% confidence level), and taking the largest observed value of σ_{sample} for UVA and UVB in Table 4.1.

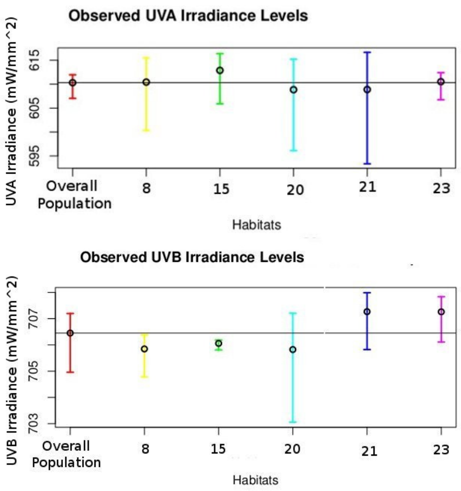


Figure 4.15: Quartile ranges of observed UV irradiance rates from the Sun across the randomly sampled habitats

Table 4.1 shows $\bar{\omega}$ and the sample deviation σ_{sample} of the observed UVA and UVB wavelengths for the local habitats. Table 4.2 shows the population mean μ_{pop} and the deviation σ_{pop} of UVA and UVB irradiance across the sampled habitats. Results show that the 95% confidence bounds for the estimate of μ_{pop} for UVA and UVB are $610.3011 \pm 3.2896 mW/mm^2$ and $706.452 \pm 1.494936 mW/mm^2$ respectively. Figure 4.15 shows the quartile ranges for the sampled local habitats, and the "Overall Population" quartile denotes the quartile ranges of $\bar{\omega}$ across all sampled habitats.

For UVA, results show a larger sample variance both within and across the local habitats, where the "Overall Population" quartile and the confidence bounds of the estimate μ_{pop} does not encompass the majority of the sampled observations. For UVB, results show a smaller variance within the local habitats. Moreover, the "Overall Population" and confidence bounds of the estimate μ_{pop} encompasses the majority of points, with the exception of habitat 20. Therefore, we conclude that UVB irradiance has a relatively "uniform" distribution across the sampled habitats, and that the UV intensity level does not vary significantly across the $36km^2$ area. For UVA, the variance is large both within and across the sampled local habitats, therefore can be considered "non-uniform" and readings must be readings must be taken in close proximity to be valid. UVA light is present throughout the day and there are many factors that could possibly attribute to causing this variance, such as changes in atmospheric properties or cloud thickness causing attenuation and/or light refraction.

For purposes of this application, “uniform” UV irradiance within a population is defined when $\sigma_{sample} < 1.5$ and $\sigma_{pop} < 1$. If this criterion is not met, the population is considered “non-uniform”. Justification for this metric is as follows:

- If deviation $\sigma_{sample} < 1.5$, then within the local habitat $\bar{\omega}$ lies around the mean, which in effect increases confidence in the UV exposure estimation.
- If deviation σ_{pop} is small across the population, then the assumption can be made that $\bar{\omega}$ will be normally distributed around the population mean no matter where the pedestrian travels.

4.7.1.1 Pedestrian UV Exposure Estimation

We restate that the aforementioned experiments was performed by holding the sensor horizontal to the ground (assuming a hemispherical view of the sky) as the pedestrian traveled outdoors. This sensor orientation is equivailant to measuring UV dosage at the top of the head and shoulders when the pedestrian is standing upright. The algorithm estimates the pedestrians’ UV exposure while traveling under direct sunlight and under trees assuming sunny days with very little cloud cover. The experiment begins by estimating the pedestrians’ UV exposure under a single tree, and then expands to a complex case that considers walking under multiple trees and direct sunlight.

4.7.1.2 UV Exposure Model for Trees

The experiments discussed in this section focus particularly on the samples collected in habitats 15 and 21. Trees with tall trunks (such as palm trees), small canopies (ie: crowns), or little leaf cover were omitted from this study because they do not provide shade to the pedestrian. The estimate $\bar{\omega}$ for the respective local habitats were used by our algorithm

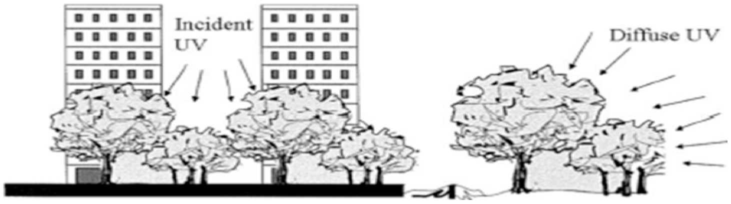


Figure 4.16: The total amount of UV radiation hitting the pedestrian level is comprised of incident and diffuse energy. Incident UV travels directly from the sky, and diffuse UV is scattered by reflection and refraction from atmospheric particles, clouds and objects in the environment such as buildings.

to estimate E_{total}. The UV dosage that the pedestrian receives under the tree canopy is proportional to the leaf cover density and diffuse component of UV [75]. We studied the UV radiation absorbance properties of leaves. We found that the UV radiation under the tree canopy is not only affected by the incident UV but also by the diffuse component. Dermatologists and skin cancer researchers alike have recommended that pedestrians seek shade under trees during the mid-day in order to avoid sunburn [76–78]. However, it is not recommended to seek shade under trees in the morning and afternoon due to the scattering of UV light as the wavelengths pass through more atmosphere [79]. On the contrary, the most damaging UVB wavelengths are not strongly present in the morning so sunburn is not a major concern (except those with photosensitive skin). Yang et. al. [75] studied the spectral reflectance and transmittance of UV radiation for different species of leaves in a laboratory environment and Yoshimura et. al. outdoors [80]. Yoshimura also compared the absorbance, reflectance, and transmittance of UV across leaves as they turn different color shades during their lifecycle (green, yellow, red and death). Results show that the UV transmittance through leaves is negligible for both the UVA and UVB wavebands, and UVA is the primary waveband that passes through leaves due to its longer wavelength. However, the UVB component is the strongest during the mid-day and seeking shade under a tree provides shelter from the incident component, but not from the diffuse (see Figure 4.16). Given this, we apply Grant's [81] equation to model the below-canopy relative irradiance on a horizontal surface I_p at the pedestrian level

$$I_p = \frac{(I_{b0} * P_0) + (I_{d0} * P_0^{`})}{I_{b0} + I_{d0}} \tag{4.10}$$

where I_{d0} is the diffuse radiation on a horizontal surface, I_{b0} is the direct beam radiation on a horizontal surface. P_0 is the probability that a direct beam of solar radiation will pass through the canopy unintercepted from the source (inside or outside the canopy) to any given point in the array of sub-canopies and is defined by

$$P_0 = e^{-G(\Omega(\phi,\theta))\rho S} \tag{4.11}$$

where Ω is the direction of radiation (with zenith angle θ and azimuth angle ϕ), G(Ω) is the fraction of foliage that is projected toward the radiation source (ie: percentage of leaves in the canopy with the surface facing the Sun), ρ the foliage density (foliage area per unit canopy volume), S is the distance through the canopy that the ray must pass, and $P_0^{`}$ is the probability that sky diffuse radiation will pass through the crown unintercepted, given by

$$P_0^{`} = \frac{\int_0^{2\pi} \int_0^{\pi/2} N(\phi,\theta) P_0 cos\theta sin\theta d\theta d\phi}{\int_0^{2\pi} \int_0^{\pi/2} N(\phi,\theta) cos\theta sin\theta d\theta \ d\phi} \tag{4.12}$$

where N is the isotropic sky radiance energy. In order to study the effective UV irradiance under a tree canopy, we conducted a test varying the parameters of ρ, and found .8 to be sufficient considering the trees near UCLA and adjacent neighborhoods have sufficient leaf cover density with leaves facing the Sun.

This model is reasonable to estimate UV irradiance under a tree canopy, however the next question is can the model's parameters be tuned to account for all trees planted within an urban neighborhood. In order to prove this we performed an experiment to estimate the pedestrian's UV exposure walking in the Sun and under trees. Two factors of interest were the following:

1. **Factor 1 - Leaf cover negligible:** UVB light is blocked, therefore we assume that the amount of UV irradiance the pedestrian receives under a tree canopy within any local habitat follows a Poisson distribution centered around the average UVA irradiance samples collected during our control experiments. In other words, we assume

that a pedestrian will receive the same UV irradiance energy standing under any tree outdoors.

2. **Factor 2 - Leaf cover not negligible:** We assume knowledge a priori of the exact amount UVA irradiance under the canopy of all trees within the local habitat along the pedestrian's travel path. To do this, measurements must be taken under the tree canopies along the walking path beforehand. In other words, we assume that trees offer varied levels of shade that must be quantified in order to reasonably estimate the pedestrians' exposure under the canopy.

With respect to our experiments: For Factor 1, we used measurements from 13 randomly sampled trees that provide various levels of shade and found $\bar{\omega}_{tree}$ to be 9.7 and 0 mW/mm^2 respectively for UVA and UVB, and fitted ρ to be .8. Measurements of these trees were taken as we performed the experiment described in section 4.7.1. Therefore, we assumed that the UV irradiance level under all tree canopies that the pedestrian can possibly walk under is $\bar{\omega}_{tree}$ for UVA and UVB respectively. For Factor 2, we took exact measurements of the UV irradiance under each tree canopy for all trees along the path. We chose trees with dense tree canopies that let in minimal light, therefore ρ was set to 1 (ie: no sunlight passes through) and actual UV irradiance measurements show 0 for UVA and UVB irradiance under the canopy.

4.7.1.3 Single Tree Path Walk Results

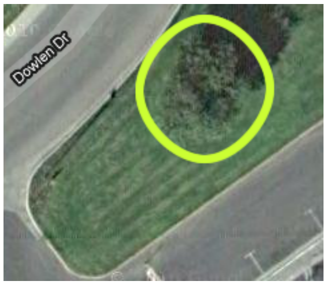


Figure 4.17: Single tree selected for the single tree path walk experiments in habitat 21.

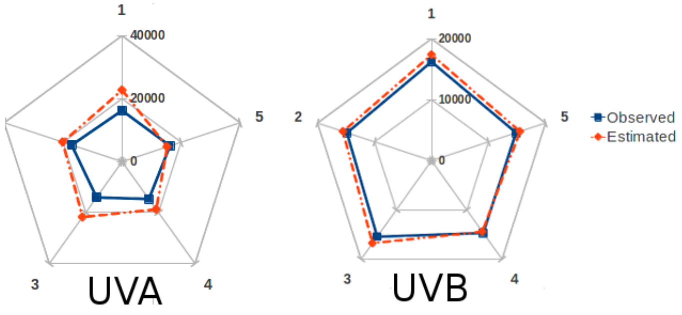


Figure 4.18: Single Tree Path Walk - Accuracy of the UV exposure estimates in Table 4.3.

Five experiments were conducted during a sunny day between the hours of 12pm and 2pm with the tree shown in Figure 4.17. The total walking distance was $86ft$, with the pedestrians average walking velocity $3.15ft/sec$ and no stopping along the path. The tree is located in habitat 21 at latitude 34.052098 and longitude -118.454799. The canopy dimensions are $34ft$ x $33ft$, and the tree's foliage completely blocks UV light penetration (ie $\rho = 1$).

Figure 4.21, shows the quartile ranges of τ for the observations collected in the experiment. Remember that the error τ (described in section 4.6) is measured in seconds because our algorithm estimates UV dosage on a per second basis, therefore multiplying by $\bar{\omega}$ gives you E_{total}, the estimated cumulative UV dosage. The dosage received under the tree canopy is a secondary expression in linear model with its own separate value for $\bar{\omega}_{tree}$. Please refer to section 4.7.1.2 for more information. Experiments show that for the majority of points, the values of τ lies between 0.705 to 1.36 seconds for UVB. For UVA, results of the factor 1 experiment show that for the majority of points, τ lies between 2.05 to 13.044 seconds, and between -6.01 to 0.52 seconds for factor 2. Factor 2 clearly minimizes error the most for UVA exposure experiments.

The accuracy of the total UV exposure estimates provided by the algorithm against the values observed for each test is shown numerically in Table 4.3 and graphically in Figure 4.18. Factor 2 was used to measure the total UVA exposure. For the majority of the tests, the total UVB exposure estimates provided by the algorithm are 94% accurate. For UVA, the estimates provided by the algorithm are 68% accurate. As shown in section 4.7.1, these results are explained by the efficiency of the sampling method that the algorithm uses to

Table 4.3: Percent accuracy of UV exposure estimates for the single tree path walk experiment

	$E_{total,obs}$	$E_{total,est}$	$\bar{\omega}$	Acc. %
		UVA		
1	14926.79	22700.36	617.13	0.48
2	17113.83	20142.39	613.87	0.82
3	14026.36	21743.80	612.24	0.45
4	14671.54	18729.32	614.62	0.72
5	16226.99	15276.34	617.13	0.94
		UVB		
1	16226.99	17395.27	704.98	0.93
2	14827.36	15518.34	706.12	0.95
3	15435.71	16757.48	706.63	0.91
4	14671.54	14399.46	706.00	0.98
5	14827.36	15518.34	706.12	0.95

estimate the UV irradiance within the local habitat.

4.7.1.4 Multiple Tree Path Walk Results

The next day experiments were performed in habitat 15 on a sunny day between 12-2pm.

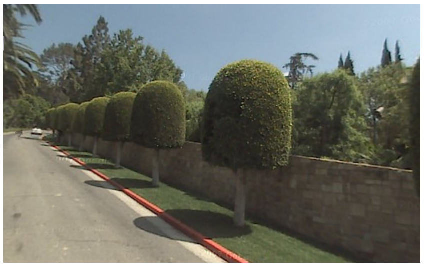


Figure 4.19: Location in Bel-Air selected for multiple tree path walk experiments.

The path shown in Figure 4.19 is located at latitude 34.086999 and longitude -118.44313. Five experiments were performed walking a distance of 100.5ft under five equally sized spherical tree canopies roughly 7.5ft x 7.5ft in diameter. UV light penetration is completely blocked by the foliage for each tree along the path. The pedestrians average walking velocity was 3.35ft/sec. Figure 4.22, shows the quartile ranges of τ for the observations collected in the experiment. Remember that the error τ (described in section 4.6) is measured in seconds because our algorithm estimates UV dosage on a per second basis, therefore multiplying by $\bar{\omega}$ gives you E_{total}, the estimated cumulative UV dosage. The dosage received under the tree canopy is a secondary expression in linear model with its own separate value for $\bar{\omega}_{tree}$. Please refer to section 4.7.1.2 for more information. For the majority of points, the values of τ lies between -1.47 to -0.70 seconds for UVB. For UVA, results of the factor 1 experiment show that for the majority of points, τ lies between 4.07 to 7.29 seconds, and between -4.75 to -2.57 seconds for factor 2. Factor 2 clearly minimizes error the most for UVA exposure experiments.

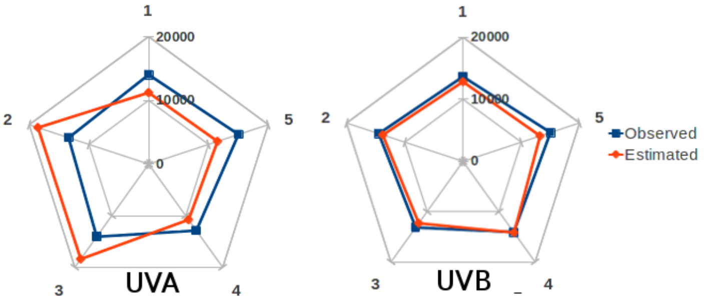


Figure 4.20: Multiple Tree Pathwalk - Accuracy of the UV exposure estimates in Table 4.4.

The accuracy of the total UV exposure estimates provided by the algorithm against the values observed for each test is shown numerically in Table 4.4 and graphically in Figure 4.20. Factor 2 was used to measure the total UVA exposure. For the majority of the tests, the total UVB exposure estimates provided by the algorithm are 94% accurate. For UVA, the estimates provided by the algorithm are 74% accurate. The significant findings of this experiment are in agreement with findings stated in the previous section.

Table 4.4: Percent accuracy of UV exposure estimates for the multiple tree path walk experiment

	$E_{total,obs}$	$E_{total,est}$	$\bar{\omega}$	Acc. %
		UVA		
1	14026.36	11205.08	616.96	0.80
2	13400.36	18523.04	618.12	0.62
3	14087.89	18415.05	615.40	0.69
4	12881.35	10792.45	617.44	0.84
5	15020.16	11524.57	613.80	0.77
		UVB		
1	13649.27	12824.62	708.42	0.94
2	14471.68	13769.12	709.23	0.95
3	13147.02	12310.77	708.99	0.94
4	14087.89	14187.10	707.42	0.99
5	15020.16	13244.86	707.74	0.88

4.7.1.5 Discussion

For all pedestrian path walk experiments conducted, results show that the proposed algorithm estimates the pedestrians UVB exposure with 94% accuracy. The algorithm estimates UVA exposure with 71% accuracy. Results also show that the error in the proposed UV exposure estimation algorithm for the majority of points is no larger than ±1.5 seconds for UVB and ±7.29 seconds for UVA. These results are achieved due to the efficiency of the two stage cluster random sampling method that the algorithm uses to estimate the UV irradiance within the local habitat under direct sunlight. Results also show that tree canopies provide different levels of shade and knowing the leaf cover density for each tree canopy reduces error when estimating exposure under the tree.

In Los Angeles (LA), the Department of City Planning maintains a GIS database (DB)

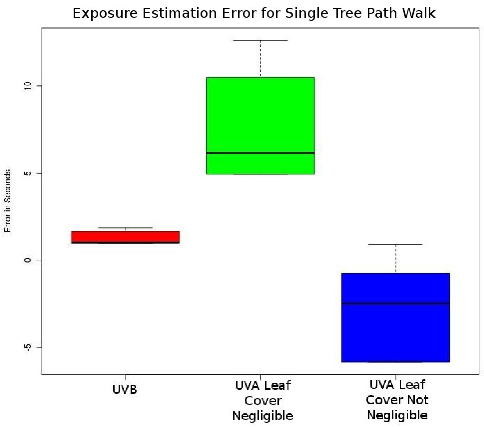


Figure 4.21: Results for single tree path walk in habitat 21.

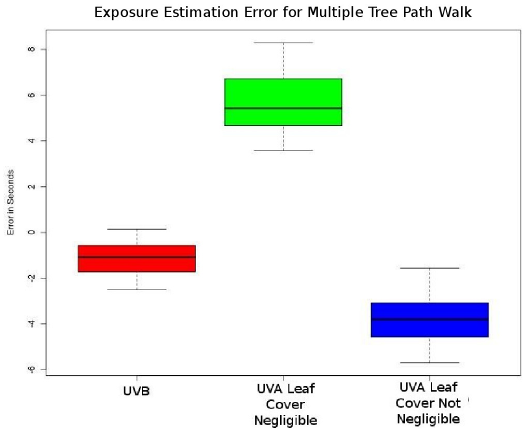


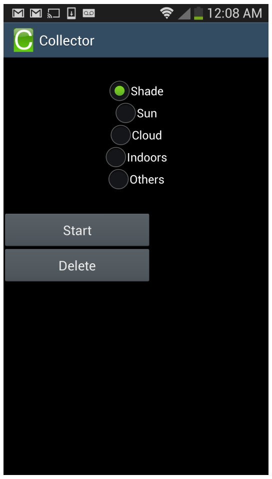
Figure 4.22: Results for multiple tree path walk in habitat 15.

of the location of all trees planted in LA since 1990. If this DB were made publicly accessible we could identify where a tree is located along the path, but information such as the tree's height and canopy thickness is omitted, making it not useful for UVG. Moreover, leafs fall off the tree seasonally and residents prune trees, making it hard to scale our proposed tree model so we dropped it. In the next section, we discuss an alternate approach to understand environmental context.

4.7.2 Identifying Outdoor Environmental Context (Sun, Shade, or Indoors)

It is imperative to know whether the user is indoors, in shade, or in direct sunlight. Typically, when indoors the user receives no UV radiation unless standing in front of a window or under a UV emitting light source such as a fluorescent light bulb (which emits a negligible amount of UVA). In outdoor environments, the amount of UV radiation that the pedestrian receives is a factor of the atmospheric properties and the objects in the environment that can either decrease or increase the pedestrian's UV exposure level. Trees and buildings provide shade during mid-day, however a single tree with a large dense canopy or multiple trees with overlapping canopies (typically found in neighborhoods) provide virtually no UV exposure (similar to an indoor environment), and shade from buildings simply scale down the energy.

We seek to design and implement a classifier algorithm that identifies the environment that the pedestrian is in by studying the characteristics of ambient light intensity levels within the aforementioned environments in order to discriminate against the instantaneous UV irradiance energy received within the direct sun, shade and indoor environments.



● Shade
○ Sun
○ Cloud
○ Indoors
○ Others

Figure 4.23: A data collector Android application that we used to collect light intensity readings under various outdoor and indoor lighting conditions.

The light sensor on the cellphone has a dynamic range (between 0 - 60,000 lux) that can differentiate between the radiant flux levels within each environmental context (eg: indoors vs outdoors). For example, the cellphone screen brightens or dims as a factor of radiant flux (ie: light intensity). Over a one month period, we performed experiments between 8:30am-6pm to collect light intensity data under the following environmental conditions:

- Indoors
- Shade
- Cloudy
- Sun

using our data collector mobile application shown in Figure 4.23. Initially, we also affixed a light sensor on each shoulder to learn whether placing sensors on the shoulder provided better discrimination in-case the cellphone's light sensor was not sufficient. Light sensor measurements were sampled between 60hz and 100hz for the Arduino and Android phone respectively, and readings were grouped into non-overlapping windows of 42 and 60 readings respectively. We eventually abandoned the shoulder study since the phone's light sensor had sufficient dynamic range and we did not want to burden the user with auxiliary devices that they would not naturally carry.

Next, we investigated various methods that could provide the best discrimination between environments, such as the mean, mean absolute deviation, standard deviation, the maximum and minimum values. Figure 4.24 shows a subset of the data collected. We found that taking the maximum value within a window works best because the intensity flux for each environment is vastly different and do not overlap. We also found that distinguishing between shade and cloudy conditions was not possible due to there being many different levels of cloud thickness that overlap with other shade values. For instance, thick clouds (similar to rainfall clouds) can produce a similar intensity flux as shade conditions. As a result, we limited our environmental condition options to indoor, shade, and sun. Therefore, we use the following environment classifier algorithm for environmental discrimination on the Samsung Galaxy S4

$$Environment = \begin{cases} \text{Indoors if } x \leq 200Lux \\ \text{Shade if } x > 200Lux \text{ \&\& } x \leq 4800Lux \\ \text{Sun if } x > 4800Lux \end{cases} \tag{4.13}$$

enabling real-time UV exposure estimation.

Next, rather than considering objects such as buildings and trees as factors for grouping UV irradiance measurements, as previously done in section 4.7.1.2, we grouped UV irradiance measurements according to the measured ambient light intensity for that environment, and assign $\bar{\omega}_{environment}$ accordingly. This means that if the pedestrian travels under a tree with a large dense canopy that blocks virtually all UV light, then the environment will be considered indoors and $\bar{\omega}$ will be set to 0 for that reading. As the user exits the tree canopy back into the direct sun $\bar{\omega}$ will assume the crowdsourced average UV irradiance value.

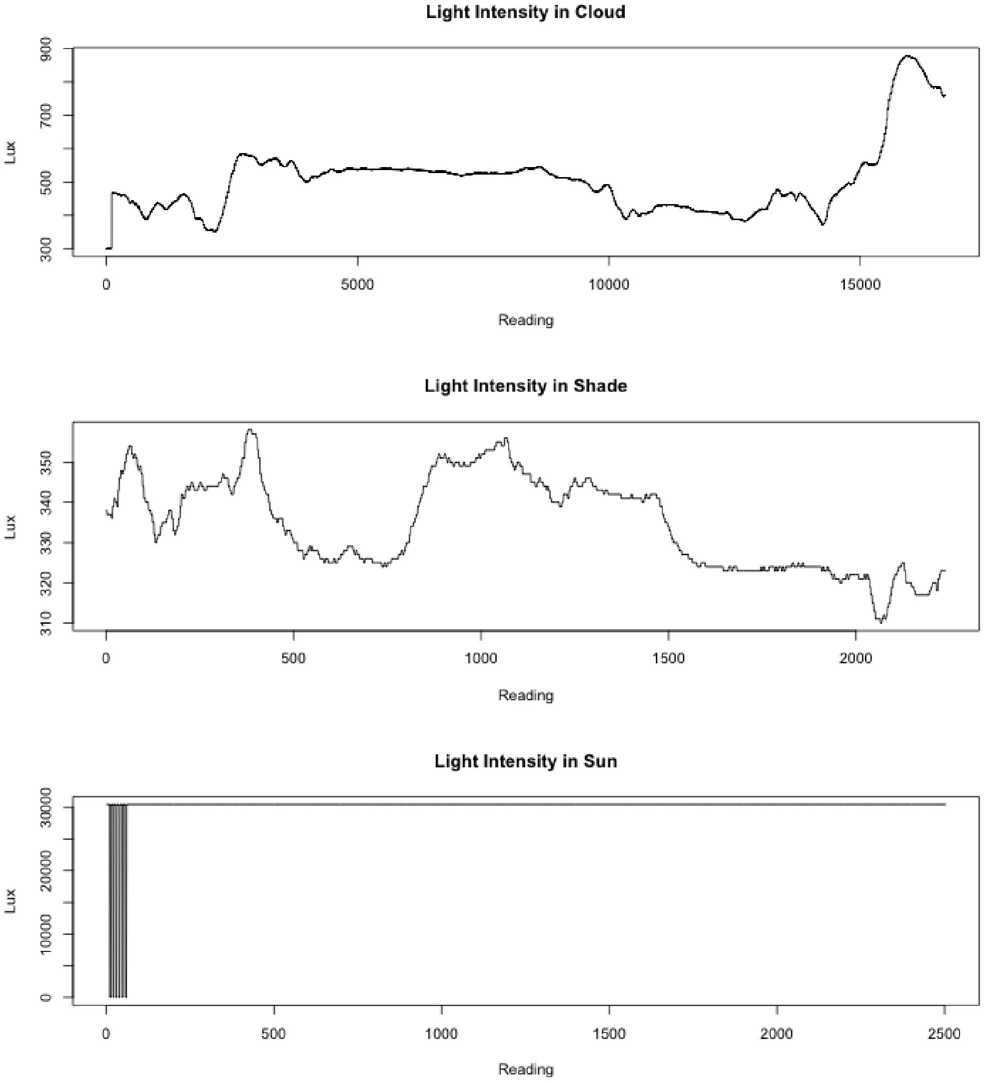


Figure 4.24: A subset of sample light intensity data used to train our environment classifier algorithm.

4.7.3 Phase 2 - Can Anatomic Body Site Ultraviolet Exposure be Estimated Comparable to a Dosimeter?

Since our primary objective is to provide proactive recommendations to protect pedestrians from erythema, and UVA contributes the least towards it, we chose to focus our algorithm on the UVB wavelengths in the phase 2 experiments. Moreover, most digital dosimeters are only sensitive to UVB since its the spectrum primarily responsible for causing skin damage. The UVB polysulfone dosimeters, described in section 4.2.6 were the golden standard adopted by the research community to quantify body site specific cumulative UV exposure. However, recently digital dosimeters retrofitted with flash memory to store exposure history overtime are becoming the new standard. Leveraging crowd sourced sample UV irradiance readings from digital dosimeters streamed to our Sensite platform, we performed experiments to validate whether our algorithm can estimate UVB exposure for the arm comparable to the dosimeter assuming the user is standing upright.

4.7.3.1 Setup

The smartphone device used in this experiment was the Samsung Galaxy S4 Active mobile phone running the UVG mobile application for Android. Once the UV exposure tracker is activated (see Figure 4.10B) UVG periodically performs a spatiotemporal query for sample UV irradiance data and uses it as input into a model that estimates the pedestrian's UV dosage as they travel outdoors. The mobile device was affixed to the subject's arm using an arm band and the NIWA UVB dosimeter used for comparison was wrapped around the arm band as shown in Figure 4.25. The dosimeter's small size made it feasible to measure UV irradiance for vertical body parts.

4.7.3.2 Method

Many researchers [82–84] performed experiments to estimate how much radiation that each anatomic body site receives when standing outdoors. The total amount of UV radiation received per unit area of the skin is a factor of the total body surface area directly exposed

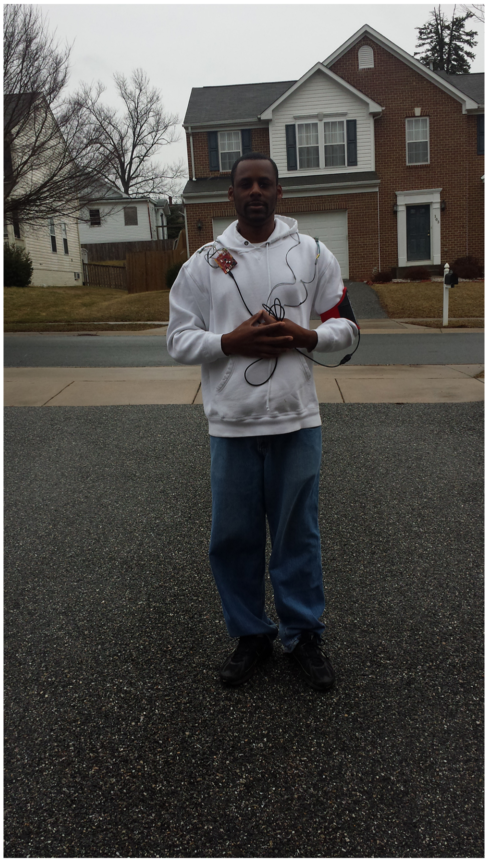


Figure 4.25: UVG mobile application running on Samsung Galaxy S4 phone affixed to the arm. The light sensors affixed to the participant's shoulders were used to gauge whether the shoulder is an appropriate place to gauge whether the user is in the sun, shade or indoors.

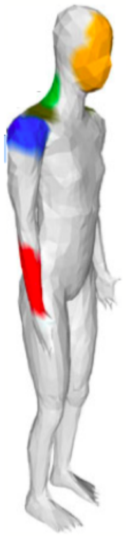


Figure 4.26: An image denoting the typical body sites for exposure to ultraviolet radiation.

to the sky. Figure 4.26 demonstrates a picture of the human body. Typically, our forearm, hands, neck and face are always exposed to the sun and that roughly represents about 1.5, 1, and 2.5% of our total body skin surface area [85]. It is also apparent that the horizontal areas of the body receive a higher level of exposure than the vertical (non-sky facing) body parts such as arms and legs. We use the ratios provided by Downs et. al [86] to estimate UV exposure for each body site relative to the sampled horizontal UV irradiance since it agrees with other prior works.

In other words we assume that anytime our classifier algorithm discovers that the user is in the shade, we apply a ratio algorithm to estimate the amount of exposure each anatomic body part will receive with respect to the horizontal $\bar{\omega}$. One could argue that there is different levels of shade that affects UV intensity, however it is not possible to obtain knowledge of the instantaneous UV irradiance in all locations. Therefore, we have placed sensors both in the shade and sun and found that on average the amount of shade provided by any vertical object tall enough to eclipse the sun from the pedestrian (such as a building) follows a general sin ratio with respect to altitude angle θ (see section 4.2.3).

In our experiments, we gauged whether our algorithm could reasonably estimate UV exposure comparable to the dosimeter, when the dosimeter is placed directly over the phone's armband. Doing so places both the phone and the dosimeter's sensor in plain sight of (we assume) roughly the same light angle measurements. It should be noted that rays from the Sun travel in a straight line towards the ground rather than spreading outward as a typical light source close to us does due to the Sun's distance.

4.7.3.3 Experiments

Our participant performed a roughly 1 mile jog around the outskirts of UCLA's campus (as shown in Figure 4.10)C during mid-day with a digital dosimeter and the UVG mobile application affixed to their arm. Affixing the device to their arm provides an estimate of the UV dosage that the pedestrian would receive on their arm. The participant was asked to run as they naturally would observing all traffic rules, walk and/or take breaks as necessary. The neighboring area of UCLA has sidewalks in front of buildings that provide partial shade in some areas, and trees on some streets. Therefore, there are a variety of lighting conditions affecting the pedestrian's UV dosage.

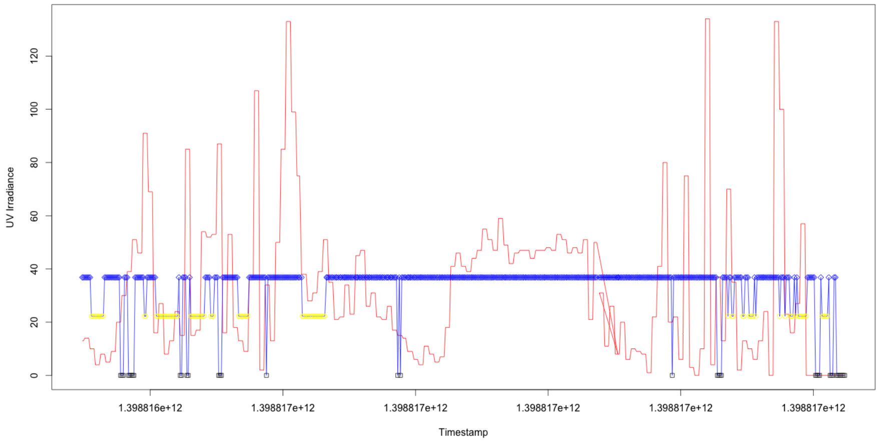


Figure 4.27: Participant performed a roughly 1 mile jog from Weyburne terrace to UCLA's campus during mid-day, with a digital UV dosimeter and the UVG mobile application affixed to their arm.

Figure 4.27 plots the estimated instantaneous UV exposure that UVG assumes the pedestrian received on their arm (denoted by the blue line) vs. the recorded UV dosage of the dosimeter (denoted by the red line). The black square on the blue line represents the assumed $\bar{\omega}$ indoors, the yellow circle represents the assumed $\bar{\omega}$ in the shade, and the blue diamond represents $\bar{\omega}$ in the Sun.

In this experiment, we measure UV exposure for the arm in order to gauge how well our algorithm can estimate UV exposure for a vertical facing body part. The amount of UV exposure that the arm receives is a factor of arm position. Readings were taken every 1 second for the digital dosimeter. Since readings for the arm are cyclical in nature due to arm swing, we used the Pearson correlation to measure the linear correlation between the dosimeter's reading and our estimate. A correlation closer to 1 denote the functions are close to each other, values closer to 0 denote the functions have a weak correlation. Results show for our experiments a correlation coefficient of .09 with a significance (P-Value) of .0247, proving that our algorithm cannot accurately estimate UV dosage for the arm due to the cyclical nature of the exposure levels as the arm moves. This would also be an issue if we tried to measure leg exposure.

4.8 Discussion

The reason for such a weak correlation is that the intensity level of UV irradiance on the arm varies as the arm swings given the user's natural gait. Our algorithm assumes that the user will receive an average UV irradiance integrated over time for a time period. Therefore, the variance in the readings due to natural arm swing is not accounted for introducing large error. Possible solutions that would improve the correlation would be to account for angular changes in the arm's position using the phone's accelerometer or gyro sensor. This information can be provided in real time, and an algorithm could be used to interpolate the instantaneous UV intensity level that the arm would receive in natural gait. We were unable to perform further experiments due to time limitations.

To explain the contrast in the experiments performed in phases 1 vs. 2. The phase 1 experiments were performed to understand the following:

- How does UV irradiance vary across a large geographic area
- How does UV irradiance vary under tree canopies
- Can our proposed algorithm be used to estimate UV dosage for body parts exposed to hemispherical UV irradiance, such as shoulders and the vertex of the head

The results from phase 1 experiments show that it is possible for UVG to accurately estimate UV dosage for sky facing body parts (assuming the pedestrian stands upright). Therefore, UVG is comparable to most commercial UV exposure estimation applications on the market such as Netatmo June [87]. In phase 2, we performed experiments to gauge whether our algorithm can accurately estimate UV dosage for each anatomic body part. To do this, we studied prior research and amended our UV exposure estimation algorithm. Next, we performed experiments and discovered that our algorithm is not capable of estimating UV dosage for vertical moving body parts. Prior explanation was given in the first paragraph.

Our studies prove that our application can estimate the pedestrians' UV dosage for body parts that are exposed to the sun (assuming the individual is standing upright) with sufficient accuracy (for most applications), such as the vertex of the head, shoulders and feet. This information can also be used to estimate vitamin D intake. The amount of vitamin D naturally produced by the skin is a factor of how much skin is exposed to the sun. Figure 4.10A shows the widget that users can use to select their clothing that they are wearing. If the user chooses not to select cloths we will assume an outfit for them based on the temperature, activity and time of day.

UVG is a mobile application that tracks your fine grain UV exposure without needing a UV sensor, provides recommendations to protect the user from Sun over-exposure, and allows the user to compare their relative exposure to others in their social circle. This fine grain spatiotemporal information is valuable for Epidemiologists to provide new insights about skin cancer, to influence Dermatologists, so that they can provide better care to their patients. We envision this app to be ported to devices such as Google Glass, and worn as the user travels outdoors during a sunny day.

CHAPTER 5

Use Case: Dengue Detector Mobile Application for Health & Wellness

5.1 Introduction

Dengue is a virus transmitted through the bite of an infected mosquito (also referred to as *aedes aegypti*). Innovative solutions have been developed to combat outbreaks, but these solutions are not affordable or easily accessible in developing countries. Additionally, traditional approaches are slow to diagnose and treat the virus. We present Dengue Detector Mobile Application (DDMA) [88–90], a mobile application that uses the vision sensors in cellular phones, a lightweight object identification algorithm to diagnose the dengue virus, and the QoI/VoI data model proposed in section 2.4 in order to crowdsource accurate dengue outbreak information for healthcare providers and the Center for Disease Control (CDC) to take action. DDMA leverages a novel microfluidic paper-based analytical device (mPAD) technology developed by researchers at the Harvard University Department of Chemistry [91], but was never commercially released.

This work shows what is possible when SSW enabled dengue diagnosis solutions are made available. Our approach improves the quality of life in developing countries by rapidly and economically detecting dengue and providing data to the CDC for monitoring of epidemics.

5.2 Facts about the Dengue Virus

The pathway into the body is from the bite of a mosquito with dengue infected blood. The virus infects nearby skin cells called keratinocytes, the most common cell type in the skin.

The dengue virus also infects and replicates inside of the dendritic cells. The infected cells display dengue viral antigens on their surface, which activate the innate immune response by alerting two types of white blood cells, called monocytes and macrophages to fight the virus. Normally, monocytes and macrophages ingest and destroy pathogens, but instead of destroying the dengue virus, both types of white blood cells are targeted and infected by the virus. As the infected monocytes and macrophages travel through the lymphatic system, the dengue virus spreads throughout the body. In another adaptive immune response, cytotoxic T cells, or killer T cells, recognize and kill the cells that are infected with the dengue virus. Together, the innate and adaptive immune responses attempt to neutralize the dengue infection, however if the ratios of these antibodies become low the patient risks getting dengue fever.

As the adaptive immune response starts fighting the dengue infection, B cells produce antibodies called IgM and IgG that are released in the blood and lymph fluid, where they specifically recognize and neutralize the dengue viral particles. Generally, it takes a few days for the virus to become detectable using a PCR based NS_1 antigen. Over the next few days, the affected individual will experience “dengue fever”. Early diagnosis is important for rapid dengue control measures. NS1 and/or PCR testing is vital in early cases of suspected dengue, because the antibody may not be present. The following week, the risk of shock or hemorrhage occurs placing the individual in danger of “dengue hemorrhagic fever”, which may lead to death if not diagnosed early. Dengue IgM testing is of limited use very early in the illness, as it only becomes detectable between about day 3 and day 7. IgG rises steeply a few days after onset, often with minimal or transient IgM [92].

Each year, there are approximately 100 million cases of dengue fever or dengue hemorrhagic fever worldwide [93]. Dengue is also the most common arthropod-borne infection worldwide with 50 to 100 million cases annually [94]. This mosquito born viral disease spread in developing countries due to sub-standard housing, inadequate waste and water management, immigration, airborne travel, and deteriorating disease prevention programs [95]. Disease prevention and control measures have been established for the early detection and monitoring of outbreaks. However, the lack of organized resources and capital in some coun-

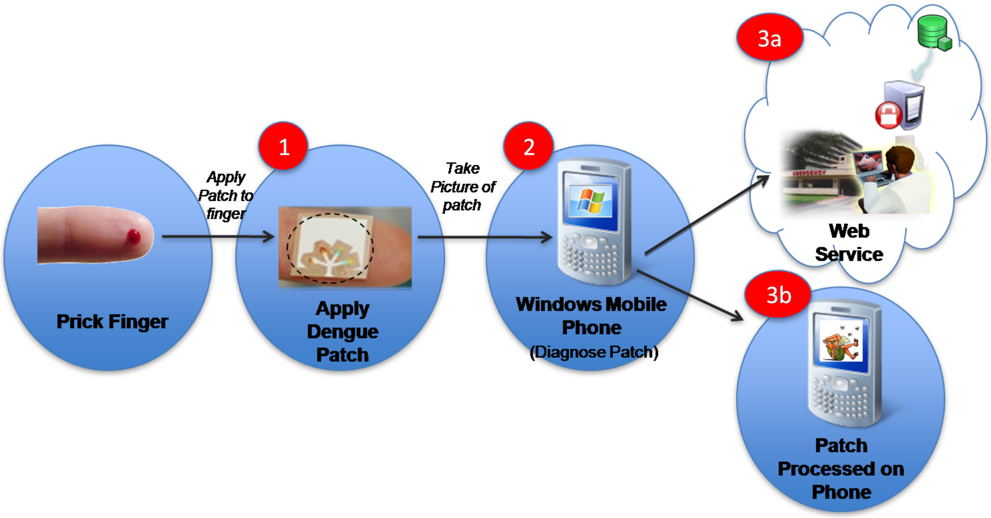


Figure 5.1: Interaction between mPAD, DDMA, and DDMA-WS resource for patch analysis. Architectural overview of dengue detection mobile application (DDMA) dengue detection for 1) dengue template 2) embedded system platform 3) central server at Center for Disease Control

tries has resulted in a number of increasing dengue viral outbreak cases [96]. Cost effective measures to accurately identify dengue can be combined with rigorous efforts to adequately treat patients and reduce the number of mosquito breeding sites. Accurate diagnosis of infection and effective preventive measures can reduce the number of outbreaks by as much as 30% [97].

We address the challenge of rapid and affordable detection of dengue disease in countries with limited resources through a combination of low cost hardware and an innovative medical bioassay patch developed by researchers at Harvard University Department of Chemistry [91]. Our approach leverages optical sensors on a cellular phone to analyze the patch results with a color identification algorithm that uses reference color shades to classify the level of dengue infection. The results of the test are displayed to the healthcare provider. An architectural diagram that shows the relationship between the medical patch, camera phone, and CDC is shown in Figure 5.1.

Our proposed system can have a significant impact on how dengue is treated in countries with limited resources. The main contributions of our system are:

- a light weight image processing algorithm that uses a $0.20 USD medical patch and a cellular telephone camera to rapidly diagnose the level of dengue infection
- real time crowd-sourced dengue outbreak statistics intended for the CDC and other prevention and control agencies for surveillance and additional testing purposes leveraging the QoI/VoI data model proposed in section 2.4 in order to ensure high quality outbreak statistics

The overall goal of mobile dengue detection is to improve the quality of life in developing countries through providing disease diagnosis and surveillance on-site rather than waiting a few days with the conventional dengue diagnosis kits. In this study, we describe our approach for dengue detection using the mobile phone and a lightweight image processing algorithm. Lastly, we summarize with an analysis of our process and conclusion.

5.3 Diagnostic Support for the Dengue Virus

Dengue and dengue hemorrhagic fever are viral diseases transmitted by mosquitoes that have the potential to cause significant illness, particularly if undetected. The mosquito has a predilection for urban areas, particularly in developing nations where breeding regulations may be lacking. The incidence of dengue infections is increasing. It is estimated that there are 100 million infections annually. Five million of these infections are serious enough to require hospitalization [93]. No vaccine is currently available for the disease. Treatment consists of early identification of the disease combined with intensive surveillance and fluid support as necessary. Significant morbidity occurs when the disease is not detected in a timely fashion to allow for resuscitation efforts to proceed.

5.3.1 Conventional Diagnostic Support

The workflow in Figure 5.2 describes the conventional methodology for treatment of a patient suspected of having the dengue virus. The process involves an initial clinical assessment

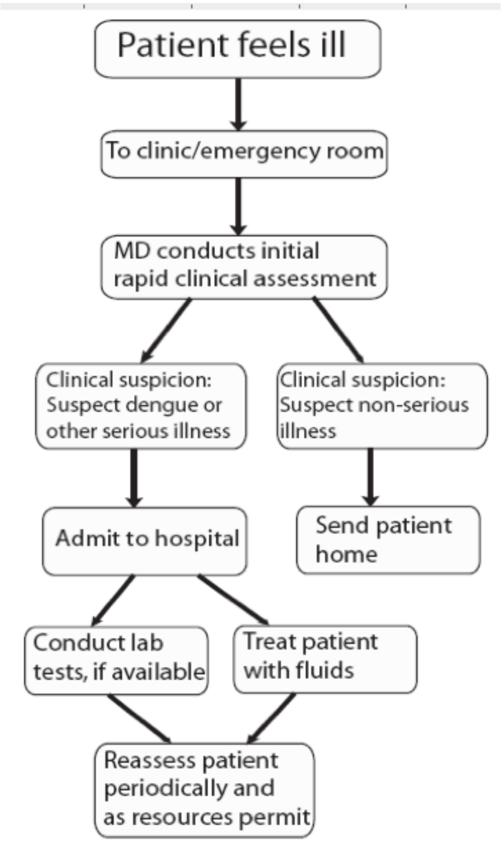

Patient feels ill
To clinic/emergency room
MD conducts initial rapid clinical assessment
Clinical suspicion: Suspect dengue or other serious illness
Clinical suspicion: Suspect non-serious illness
Admit to hospital
Send patient home
Conduct lab tests, if available
Treat patient with fluids
Reassess patient periodically and as resources permit

Figure 5.2: Workflow diagram of how dengue is traditionally treated.

where the physician admits the patient in for initial testing. A sample blood specimen is then taken for testing and a PCR based assay test would be administered to extract the viral components of dengue from the patient's serum sample. The test would be administered multiple times each test isolating a different viral strand to gauge the level of infection. This process can take as long as five days to identify the level of infection. Following this process, if the physician has suspicion of illness, the patient is immediately admitted to the hospital where further tests and viral antibodies are administered to fight the virus.

The apparent problem with this methodology is there is no concise way to diagnosis the level of infection quickly to prescribe the necessary treatment. With each day that passes, the level of viral infection can grow worse, so a more accurate and timely system is needed to save lives. Our proposed DDMA system rapidly and economically diagnoses dengue and can provide valuable information to healthcare providers within a few minutes by transmitting valuable information on location and quantity of detections rapidly to the Center for Disease Control (CDC).

5.3.2 Mobile Diagnostic Support for Dengue Detection with DDMA

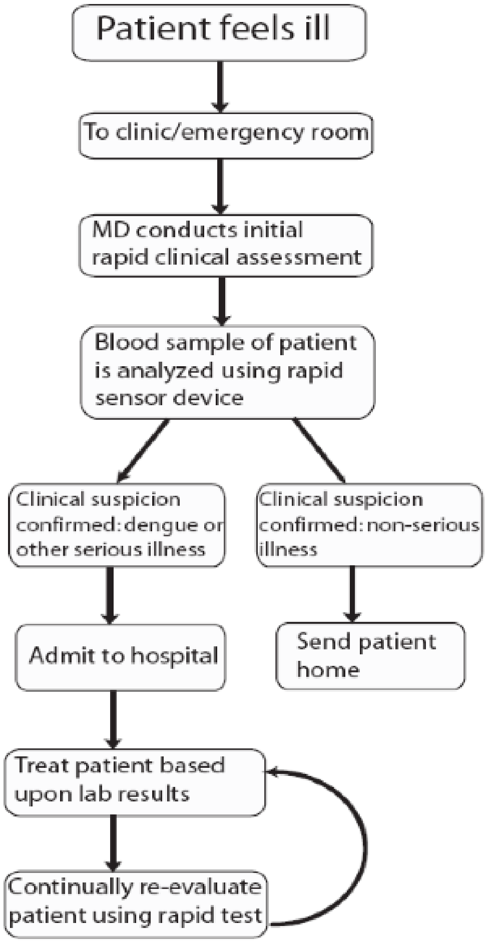

Patient feels ill
To clinic/emergency room
MD conducts initial rapid clinical assessment
Blood sample of patient is analyzed using rapid sensor device
Clinical suspicion confirmed: dengue or other serious illness
Clinical suspicion confirmed: non-serious illness
Admit to hospital
Send patient home
Treat patient based upon lab results
Continually re-evaluate patient using rapid test

Figure 5.3: Workflow diagram of our proposed system with DDMA.

We propose a new workflow described in Figure 5.3. An ill patient suspected to have dengue infection is identified and a small blood sample is taken from the patient. The sample would be applied to the microfluidic paper-based analytical patch (mPAD) manufactured by Diagnostics For All (shown in Figure 5.4A). The analytes in the patients blood are allowed to diffuse into the different chambers of the patch. Reaction occurs with the reagents in each well and the device is inspected after half an hour. Gross abnormalities in the patients blood

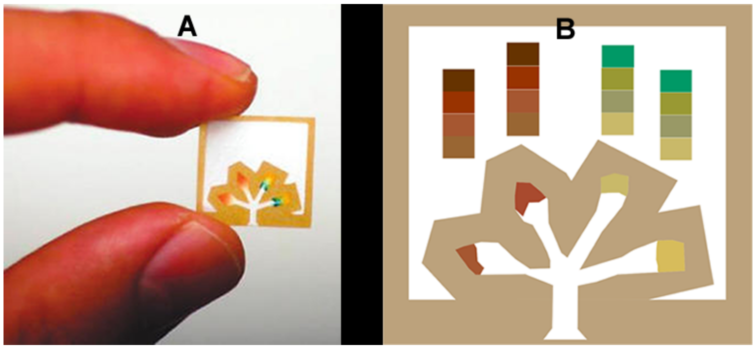


Figure 5.4: (A) Microfluidic paper-based analytical patch manufactured by Diagnostics For All. (B) Mock-up of envisioned patch design for DDMA using reference colors for image analysis.

sample can be detected by inspecting the color shades of the patch with the unaided eye. However, individuals interpret color shades differently which may lead to bias and incorrect diagnosis. Therefore, we propose the mockup patch (shown in Figure 5.4B) adding reference color shades to the patch for an image processing algorithm to provide a non-biased dengue diagnosis by analyzing the color shades of the patch. This image is then processed and the color levels in each of the wells are compared to the reference ranges, yielding a quantitative result of the analyte levels. Based upon the results, the patient can be triaged and given appropriate treatment. Proper treatment may include fluids, additional interventions, and admission to the hospital for more intensive care and follow-up.

The image processing algorithm that we are using was developed using Matlab and later implemented on a cellular phone. The resolution of the image used for testing was 240 x 320. As the image scaled the picture clarity was reduced, so until we receive the actual patch, we are unable to perform in depth analysis of the patch detection errors. However, the photo suffices as a basis for algorithm development.

We propose the use of reference colors because ambient light conditions may degrade the richness of the original colors in the patch (see Figure 5.5A). In order to solve this, we propose the usage of reference color shades above each well. The degradation of the reference color shades will be uniform enabling fast color matching.

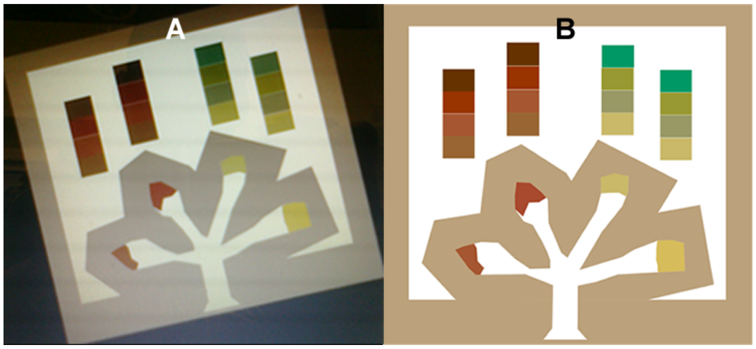


Figure 5.5: (A) Image of Mock-up mPAD taken with Windows Mobile camera. (B) Mock-up patch design. Distortion of color in A vs original image B validates need for reference colors.

5.4 Mobile System Platform

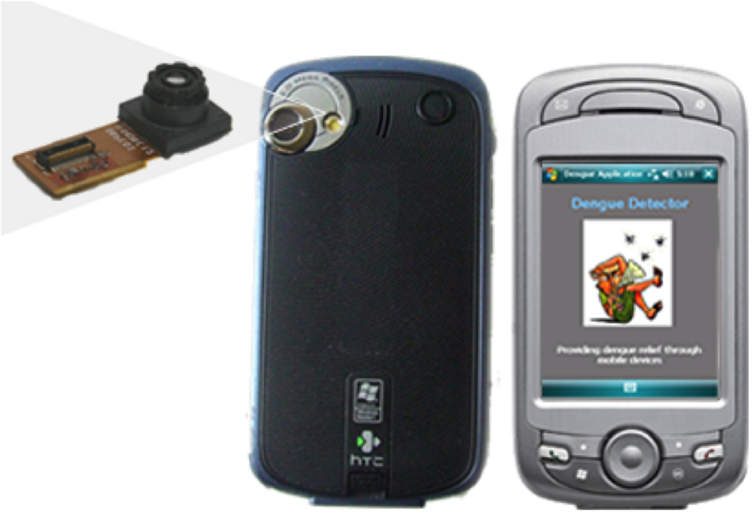


Figure 5.6: HTC Mogul 6800 cellular platform with embedded optical sensor.

The HTC Mogul 6800 Windows Mobile Smartphone device is one of the HTC Company's flagship smart phones (Figure 5.6). The mobile device is small, lightweight and uses the Qualcomm 400 MHz MSM750 ARM Processor. The device has 64 MB of RAM and 512 MB of flash memory. The device operates under the Windows Mobile 6.1 Operating System, and contains a 2.0 Megapixel CMOS camera embedded in the device. The image of the device is shown below. The software used to program the Dengue Detector Mobile application

(DDMA) is written in C#.

5.5 Light-weight Image Processing Algorithm

In the past decade, cell phones have transformed from simple mobile communication devices to mid-range scalable computing devices. The processing power of cellular phones has increased dramatically enabling many technological innovations. The cellular phone has increased processing capability, memory, and external sensors (camera, accelerometer, gyroscope, etc.).

In this work, we discuss the algorithms implemented on the cell phone. In addition, we analyze several object identification algorithms in the challenges section that are constrained by limited memory and low processing power.

5.5.1 Greedy Scanning

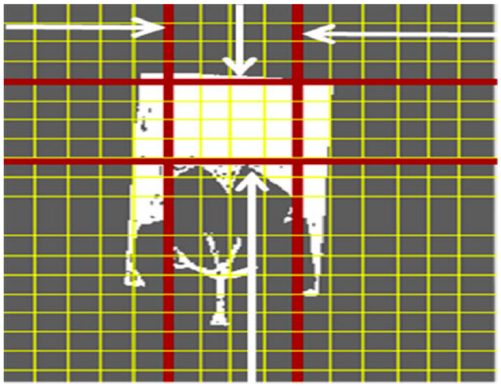


Figure 5.7: Direction of scan lines to define localized area of patch where highest average pixel value is encountered.

The first step of our algorithm is to isolate the patch from background noise. We assume that the patch is clearly identifiable in the foreground so that a binary copy of the image can be reproduced (see Figure 5.7). To localize the patch, a greedy scan is applied beginning with a vertical scan of every ten pixels across the X axis, then a horizontal scan across the Y axis in the image. The scan starts from the outside moving inward until the highest average pixel value is obtained in each respective scan direction. Once the two columns (from the X axis) and rows (from the Y axis) are obtained for that respective scan area, the first two high gradient edge points encountered are stored. Figure 5.7 shows the horizontal and vertical scan directions. The following algorithm identifies the gradient points:

- *X axis* scan gradient points obtained by a scan beginning outward starting at the top position. Then the scan moves inward and downward on the patch. The scan stops at the first largest gradient encountered in each scan direction.

- *Y axis* scan gradient points obtained by a scan beginning outward starting at the left position. Then, the scan moves inward to the right. The scan stops at the first largest gradient encountered in each scan direction.

These points are obtained as the average pixel value and are calculated for each scan line. The two points are then used to form a line segment that cuts through the patch. The figure below marks the direction of the scan lines as they cut through the patch area. The red lines from Figure 5.7 correspond to the index where the greatest average pixel value was seen from respective scan directions. Given this information, we now have a localized area where the patch resides. Our implementation has the two following assumptions.

- There exists stronger edge approximation accuracy when the patch orientation is a few degrees within the ranges of 0, 90, 180, or 270 degrees.

- The mobile phone camera is parallel to the patch to avoid skewed areas.

For example a patch oriented around a 45° angle will reduce the algorithms ability to accurately identify edge regions. In most cases, the patchs edges will be unidentifiable because the edge points will form a V shape instead of a straight line. The distorted shape would invalidate the algorithms ability to predict an approximate line segment.

5.5.2 Approximating Patch Edges

Now that the scan lines have identified a local area of the patch, the edges must be identified in order to approximate the location of the corner points. To approximate the edges we trace the line segment formed between the gradient points as shown below.

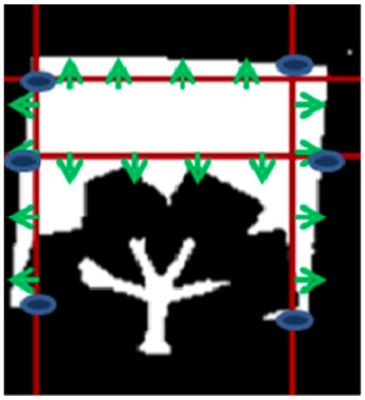


Figure 5.8: Shows gradient points marked from each respective scan and stencil scan to define the edge. The blue circle represents a strong gradient point, G$_i$(x,y), that forms a line segment cutting through patch. The green arrow represents the stencil scan direction to the approximate edge of the patch.

Figure 5.8 gives a visual picture of the algorithm used to estimate the edge regions of the patch. The blue oval represents the gradient points selected for each scan area containing the highest average pixel value. The edge detection stencil scan extends outward, depicted by the green arrow towards the outer edges of the patch. The next section describes the edge region approximation algorithm in more detail.

5.5.3 Edge Region Approximation

The below algorithm is the weighted neighboring stencil equation that we developed. A stencil is used to find the sharp gradient point that represents the location where the image section stops being white and sharply changes to black.

$$S_x = [(L(x_i, y_{j-2}) + 2L(x_i, y_{j-1})) - (2L(x_i, y_{j+1}) + L(x_i, y_{j+2}))]/4 \qquad (5.1)$$

$$S_y = [(L(x_{i-2}, y_j) + 2L(x_{i-1}, y_j)) - (2L(x_{i+1}, y_j) + L(x_{i+2}, y_j))]/4 \tag{5.2}$$

$$L(x_i, y_j) = (1-t)G_1 + tG_2, t = [0..1] \tag{5.3}$$

S_x is the stencil equation used for scans across the X axis (Eq. 5.1). S_y is used for scans across the Y axis (Eq. 5.2). In Eq. 5.3, the scan begins at a starting point, L(x_i, y_j). The starting point lies along the line segment between gradient points, G_1G_2, defined by the blue oval for each respective line scan. The gradient points extend perpendicular to the segment in the direction of the associated green arrow until a sharp gradient is found. Each initial starting point L along G_1G_2 comes from the line equation:

If an edge is not found, the next intermediary point along the line segment is evaluated. We chose to use a weighted four point stencil to assign higher weights to adjacent pixel points.

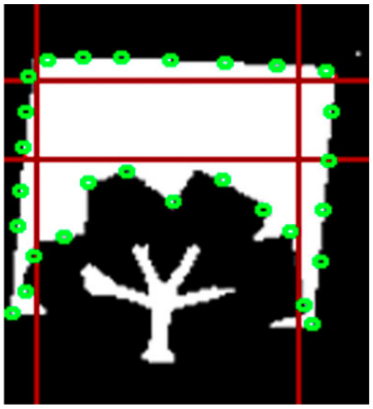
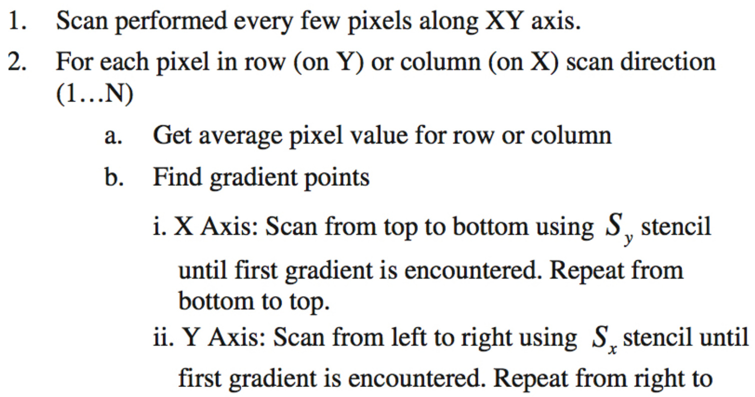


Figure 5.9: Shows edge point traces of outward stencil scan.

Figure 5.8 shows the direction of the stencil scan, and Figure 5.9 above shows the resultant edge points marked to outline the approximated edges of the patch. Figure 5.10 shows pseudo code of the stencil scan algorithm. Following the line segment scan, a trend line algorithm such as the Least Squares Fitting can be used to construct a trend line for each side. Only three sides are necessary to identify the patch. Those sides are the ones that form perfect straight lines. The side tracing the well area will contain points that do not construct a



```
1. Scan performed every few pixels along XY axis.
2. For each pixel in row (on Y) or column (on X) scan direction
   (1…N)
      a. Get average pixel value for row or column
      b. Find gradient points
            i. X Axis: Scan from top to bottom using $S_y$ stencil
               until first gradient is encountered. Repeat from
               bottom to top.
            ii. Y Axis: Scan from left to right using $S_x$ stencil until
               first gradient is encountered. Repeat from right to
```

Figure 5.10: Pseudo-code for algorithm to obtain highest average pixel rows/columns and gradient points.

perfect line and reveal the orientation of the patch. Once a line equation is constructed for each of the three sides, the two corner points connecting each side to form the square can be identified by finding the intersection points from the segments.

5.5.4 Patch Orientation

Our approach assumes the side with the highest margin of error given the trend line to be where the wells reside. Another enhancement that provides further accuracy is to split the square into four quadrants. Then, the top two highest average pixel values are the top half of the patch. The two quadrants with the lowest values are mostly black and contain the wells.

5.5.5 Well Identification and Detection

Given that the patch is square, we know the length of the sides by measuring the distance between the two approximated coordinate points. This distance measurement is then used as a scaling factor for our reference point measurements. To identify the location of each well, we manually measured the distance between each well and our chosen reference point. A scaling factor of how many pixels corresponded to one unit of measure. We chose our reference point to be the top left of corner of the patch opposite of the wells was created.

Given the reference point measurements, our length scale measurements, and our scaling factor we have all the information we need to process the patch on the mobile phone very quickly.

5.5.6 Patch Angle Transformation and Well Identification

Given an acceptable patch orientation within range of the accepted degree values mentioned above, we applied the Jacobian Transform to identify the wells given any slight offset of the patch from the normal and a measurement. We found the offset from the normal angle by taking the arctan of the two corner points.

$$\theta = arctan(y_2 - y_1/x_2 - x_1) \tag{5.4}$$

Once the offset is found the Jacobian Theorem can be applied. For our equation, we used the following theorem to identify the well location given our reference point.

$$B = cos^2\theta - sin^2\theta \tag{5.5}$$

$$1/B = \begin{bmatrix} cos\theta & -sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x^{'} \\ y^{'} \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} \tag{5.6}$$

($\text{x}^{'}$, $\text{y}^{'}$) represents the original pixel distance from the reference point to a desired well. (x, y) represents the new x, y coordinates given the angle offset θ. Given this equation, we can plug in our reference measurements for each well and find the exact location in the image.

5.5.7 Well Color Detection

Our implementation assumes that a certain amount of infrastructure is available. Below is a list of our assumptions:

- The CMOS optical sensor is fairly sensitive enough to produce a good quality image

- There are no variations in illumination

- The image is a taken at an angle parallel to the patch

After locating a well, its color can be contrasted with corresponding reference template colors by comparing their luminosity. Luminosity analysis analyzes the intensity of the pixels while maintaining the original colors (Figure 5.11a). Figure 5.11b shows the image after the luminosity analysis. After luminosity, the colors are further segmented into red green and blue yellow planes. Then, the colors are clustered so that colors that are the closest to each other are grouped together. A range metric is also used to reduce noise by ensuring that there is a minimum distance between colors in a cluster. Figure 5.12 shows the pseudo code

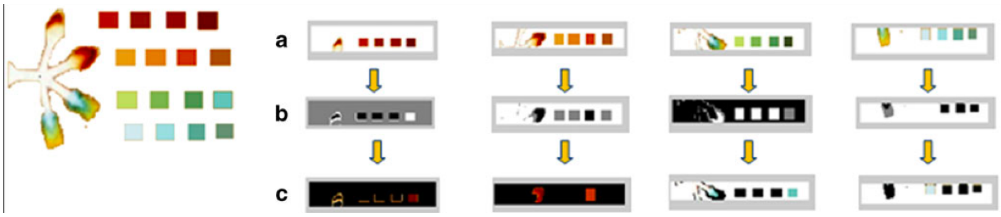
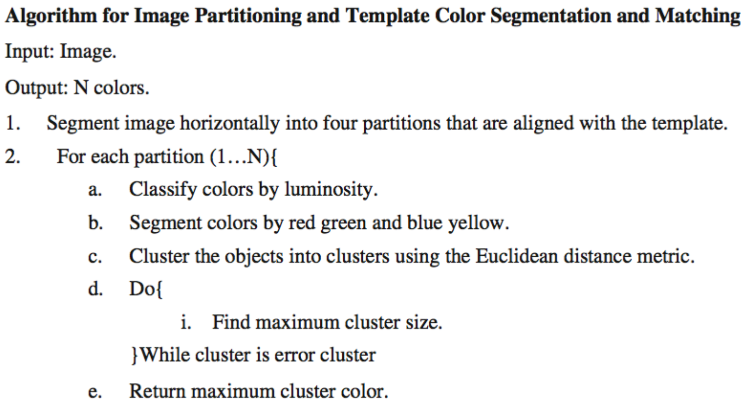


Figure 5.11: Image partitioning and template color segmentation and matching. a) Image partitioning by square template matching. b) Luminosity analysis by pixel intensity. c) Maximum color size by cluster.



```
Algorithm for Image Partitioning and Template Color Segmentation and Matching
Input: Image.
Output: N colors.
1.  Segment image horizontally into four partitions that are aligned with the template.
2.  For each partition (1…N){
        a.  Classify colors by luminosity.
        b.  Segment colors by red green and blue yellow.
        c.  Cluster the objects into clusters using the Euclidean distance metric.
        d.  Do{
                i.  Find maximum cluster size.
            }While cluster is error cluster
        e.  Return maximum cluster color.
```

Figure 5.12: Pseudo-code for image partitioning and template color segmentation and matching.

for well color detection.

Some reference colors are error colors that appear in the well but are not the valid test result. If the maximum cluster color is an error color, then the algorithm searches for the next maximum cluster size. The result of the maximum cluster color is shown in Figure 5.11c.

5.6 Image Processing Challenges

One of the biggest challenges encountered during this project involved patch localization algorithms. Due to having a very restricted execution environment in terms of memory and processing power, this section describes the challenges faced with analyzing images on the camera.

5.6.1 Grayscale versus Binary

Processing an image given an intensity scan, also known as gray scaling is a common technique in image processing. Intensity scans work well to transform the dimensions of a color image into a single dimension of gray shades. The intensity then can be used by edge detection algorithms such as Sobel to find the gradient of images.

What we found during experimentation was that although gray scaling works well to normalize images, this technique alone does not get rid of background noise. To create a simplified image with virtually no background noise, a binary image must be constructed. The binary image that we constructed classifies the inner area of the patch area as pure white, and removes virtually all background noise. A binary image simplifies the images color scheme, and makes it easier to identify the patch. Please see Figures 5.9 and 5.13 for reference.

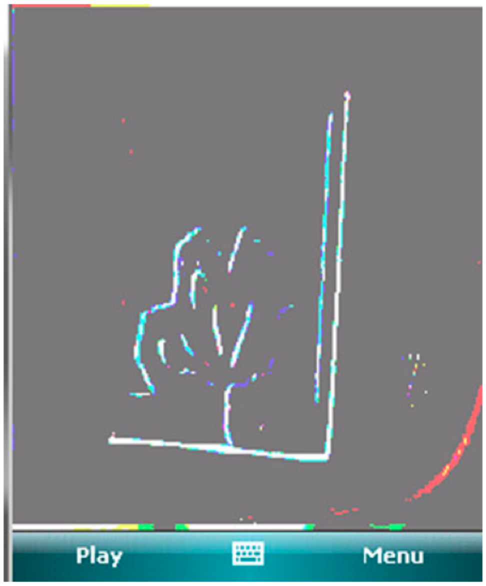


Figure 5.13: Gradient image of patch with weak edge definition from an intensity scan (gray scale).

5.6.2 Sobel Edge Detection

The first approach was to use grayscale image techniques to identify the patch using intensity scans, followed by edge detection algorithms to isolate the edge regions of the patch. The algorithm used was the Sobel Algorithm defined below.

$$G_x = A * \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} G_y = A * \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 1 & -2 & -1 \end{bmatrix} \tag{5.7}$$

$$G(x,y) = \sqrt{G_x + G_y} \tag{5.8}$$

The goal was to utilize the edge map produced from the gradient edges to identify and isolate the outer corner region of the patch. However, we found that in certain instances edges were not clearly defined to have an influence on our bounding box localization algorithm.

For example, the image shown in Figure 5.13 is a result of the Sobel algorithm using a 3 x 3 convolution kernel in both the X and Y direction. G$_x$ and G$_y$ produced the gradient image G. The variable A corresponds to a 3 x 3 subsection of the image to be masked. When the subjects skin color resembles the color of the outer edges of the patch, the edges were less defined. In addition, if a picture was taken of the patch at certain angles, the edges were also less defined.

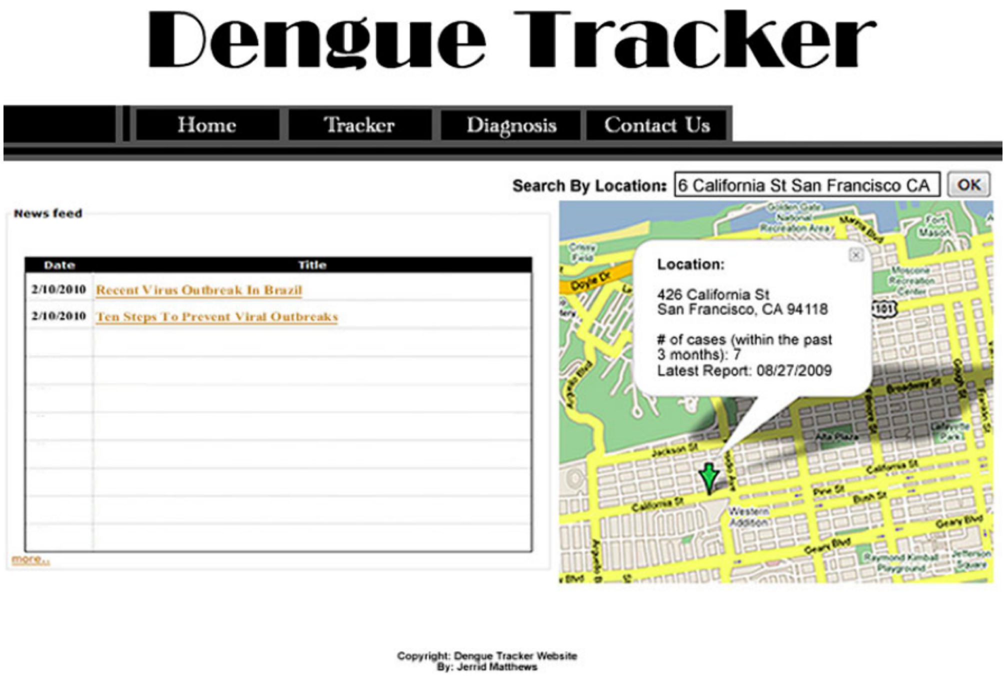
5.7 Dengue Outbreak Webpage

An important aspect of our application is to perform a complete diagnosis on the cellular phone. The real-time tracking of outbreak information enables individuals and organizations, such as the Center for Disease Control (CDC), to act accordingly to prevent further outbreak occurrences. We provide real-time tracking of outbreaks following the users decision to upload information to the web service. The DDMA Web Service (DDMA-WS) was developed with two major goals in mind.

The first goal is to enable real-time tracking of dengue outbreaks. The second goal is to facilitate monitoring and record keeping of test results. The purpose of the Dengue Outbreak Tracker webpage is to display the latest dengue outbreak information in real-time to interested parties. The dengue website consists of the following features:

Search for outbreaks: The Google Maps API allows data retrieval based upon unique location facts. The search bar widget enables querying of the latest dengue outbreak information according to a particular location of interest. The resultant information returned from the query will be the total number of individuals diagnosed as infected with the dengue virus within a 3 month timeframe. Additionally, a search widget enables querying of the latest dengue outbreaks within a radius of a given location.

Dengue outbreak tracker: The Dengue Tracker Google Maps Application, shown in Figure 5.14, displays graphically the impact factor of the outbreak in a region of interest. The impact factor is represented by a marker containing information about the total number of reported infections in the area. The information only relays the sum and does not provide information on individuals on a case by case basis. Additionally, a headline news feed column gives periodic updates of the latest reported news regarding dengue outbreaks.



Search By Location: 6 California St San Francisco CA

News feed

Date	Title
2/10/2010	Recent Virus Outbreak In Brazil
2/10/2010	Ten Steps To Prevent Viral Outbreaks

Location:

426 California St
San Francisco, CA 94118

# of cases (within the past 3 months): 7
Latest Report: 08/27/2009

Copyright: Dengue Tracker Website
By: Jerrid Matthews

Figure 5.14: Screenshot of the dengue tracker webpage. Users are able to query the number of dengue outbreak cases within a three month window and view a news feed of latest dengue news.

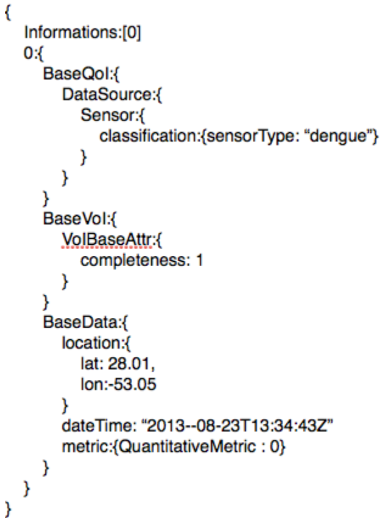
Real-time data dissemination An important property of our DDMA system is that it can perform a complete diagnosis on the cellular phone. The real-time tracking of diagnoses enables individuals and organizations, such as the Center for Disease Control (CDC), to act accordingly to prevent further outbreak occurrences. We provide real-time tracking of outbreaks following the users decision to upload information to the web service. During the diagnosis, DDMA retrieves the GPS location where the test was administered. We assume that the reported GPS location corresponds to or is in close proximity to the location of the actual infection. These data is annotated with QoI/VoI metadata (see section 2.4) that describe the following:

- Relevance: Conventionally there are 3 phases of dengue infection with a specific diagnosis test for each phase
- Completeness: Sensor data must include location and time

5.7.1 Implementation of the Ontology Based QoI/VoI Data Model

In [89], we described the envisioned usecase for DDMA, where a team of Physicians are deployed to regions of interest where high incidences of dengue are reported. The Physicians are equipped with: (a) numerous Microfluidic paper-based analytical devices (mPADs) capable of diagnosing dengue infection, (b) a Windows Mobile phone with the Dengue Detector Mobile Application (DDMA) to analyze the mPAD, and (c) a vehicle equipped with network access. Since [89], the DDMA Web Service (DDMA-WS) has been replaced with our Sensite data platform due to the platform performing the same function.

Sensite (see chapter 3) is designed to be a generic sensor data platform. Therefore, no logic exists to discriminate or perform custom actions based on the content of the metadata. It simply stores any uploaded sensor data. It is the role of the DDMA-WS to perform a RESTFul query against Sensite to obtain the relevant sensor data, then filter these sensor data according to a radius based region of interest. The data is returned in JSON format like the following in Figure 5.15.



```
{
   Informations:[0]
   0:{
      BaseQoI:{
         DataSource:{
            Sensor:{
               classification:{sensorType: “dengue”}
            }
         }
      }
      BaseVoI:{
         VoIBaseAttr:{
            completeness: 1
         }
      }
      BaseData:{
         location:{
            lat: 28.01,
            lon:-53.05
         }
         dateTime: “2013--08-23T13:34:43Z”
         metric:{QuantitativeMetric : 0}
      }
   }
}
```

Figure 5.15: QoI/VoI metadata that is created from DDMA and uploaded to the Sensite platform for sensor data storage.

The *QuantitativeMetric* attribute stores the dengue diagnosis information in either string or decimal format. The website receives a JSON array containing the sensor data results, then filters out the complete tests and displays the latest outbreak information to the user on the dengue outbreak tracker website (see Figure 5.16). The goal is to use web services for a medical mobile health infrastructure at the point of testing.

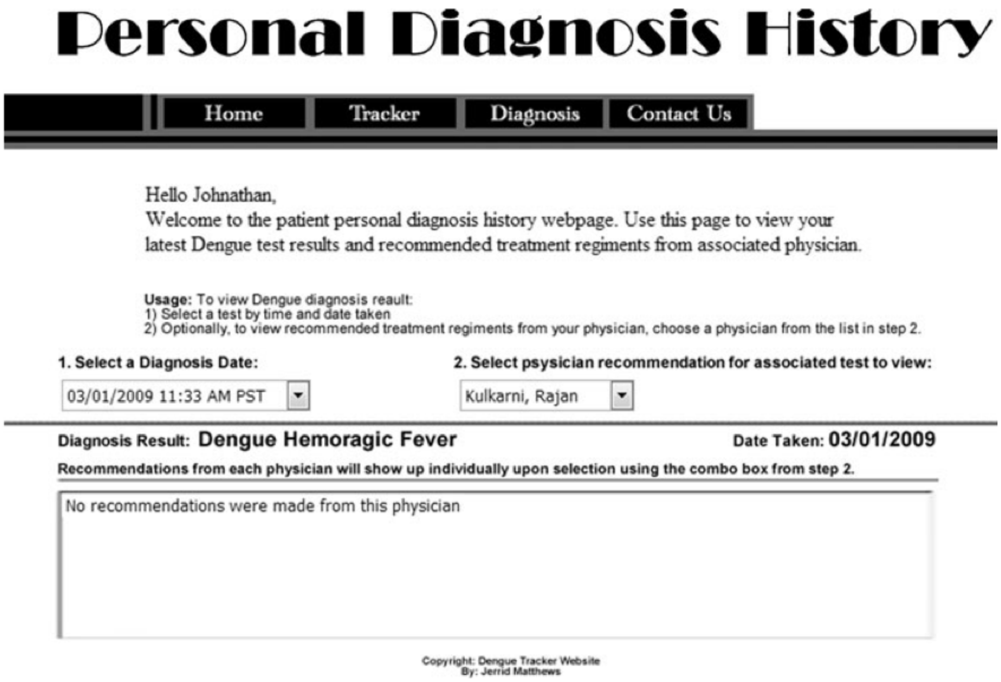
5.7.2 Social Networks

Social networking has become a widely popular medium to create and build relationships. Research has shown that social networks can describe relationships on many levels, play a critical role in determining various options for solving problems, and determine the value that one may gain from the relationship [98–100]. We leverage social networking concepts to build a website user profile that models the connections made between the following entities. These entities represent the key elements necessary to facilitate patient diagnosis, treatment, and patient monitoring.

- Patient
- Physician
- Epidemiologist

5.7.2.1 Patient

The Patients profile is comprised of basic information for representing a person using the Dengue patch for purposes of diagnosis. General information under this profile includes the patients name, sex, weight, age, home address, and their diagnosis history. On the Dengue mobile application, test results are anonymous and only reveal the presence of an infection in the region (Figure 5.16). In a developing country, the patients likely will not have access to a cellular phone. While kiosks have been proposed in developing countries, the patient functionality is a feature that more likely to be used in more industrialized countries.



Hello Johnathan,
Welcome to the patient personal diagnosis history webpage. Use this page to view your latest Dengue test results and recommended treatment regiments from associated physician.

Usage: To view Dengue diagnosis reault:
1) Select a test by time and date taken
2) Optionally, to view recommended treatment regiments from your physician, choose a physician from the list in step 2.

1. Select a Diagnosis Date: 03/01/2009 11:33 AM PST

2. Select psysician recommendation for associated test to view: Kulkarni, Rajan

Diagnosis Result: **Dengue Hemoragic Fever** **Date Taken:** **03/01/2009**

Recommendations from each physician will show up individually upon selection using the combo box from step 2.

No recommendations were made from this physician

Copyright: Dengue Tracker Website
By: Jerrid Matthews

Figure 5.16: Screenshot of the Personal Diagnosis History webpage. This webpage enables the user to view their latest dengue diagnosis history and optionally recommended treatments from a physician.

5.7.2.2 Physician

The Physicians profile is used to display information about the physician, such as his/her name, medical practice, specialty, and location of practice. The Physicians profile has the *careGiverOf* relationship that allows the physician to seamlessly interact with the Patient, identifies the physician as the caregiver, and allows the physician to upload information on the Patient. For each test result, a physician has the ability to create a dialogue with the patient for recommended treatment under the patients profile.

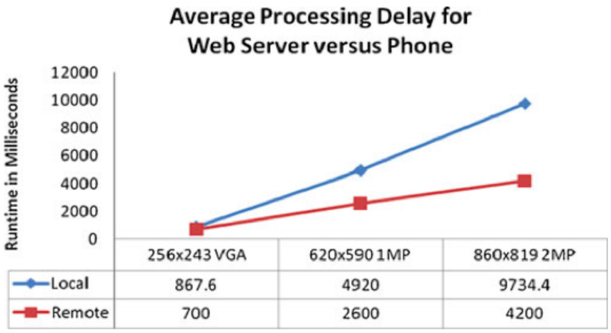
5.7.2.3 Epidemiologist

The Epidemiologists profile was created to allow the analysis of the frequency and distribution of dengue in a population or region. Epidemiologists identify disease outbreaks and recommend public health policy. These recommendations may include approaches to maintain an outbreak. The information stored under this profile is reports and alerts from analysis of data on the dengue outbreak webpage. Also, a forum for epidemiologists will be present for analyst to discuss patterns in detections.

5.8 Results and Analysis

Our image partitioning and color segmentation and matching algorithm allows a camera phone to diagnose dengue in a patient. Diagnosis is done by matching the color results to reference colors for a respected well on the medical patch. We discuss the methodology used to create the metrics used for the processing delays and power consumption of images of the following resolutions (VGA, 1MP, 2MP).

5.8.1 Image Processing Delay

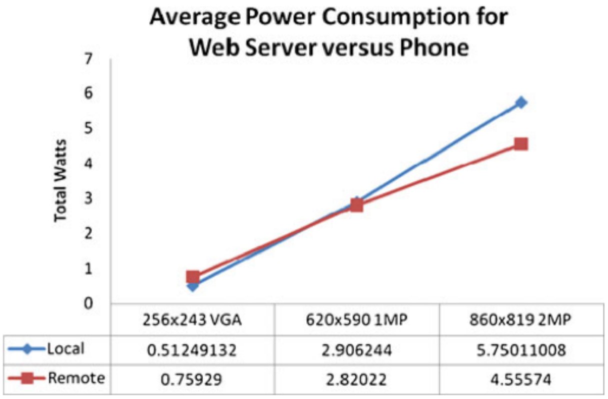


	256x243 VGA	620x590 1MP	860x819 2MP
Local	867.6	4920	9734.4
Remote	700	2600	4200

Figure 5.17: Diagram of the average time variation between end-to-end process completion of web server versus phone local processing time.

To gain an estimate of the average processing time for images of various resolutions, we ran the algorithm ten times for images of various resolutions, and then recorded the average processing time for both experiments. In Figure 5.17, the average time of both resources (smart-phone and web service) spend performing actual image processing is shown. From the chart it is clear that the laptop performs orders of magnitude faster than the smart-phone in terms of processing speed, which makes the laptop processor a golden resource to the smart-phone. However, although processing capability varies greatly between devices, there are more important factors to consider before deciding to offload processes to an external resource. Smart-phones are often equipped with proximity accelerometer, CMOS light, and GPS sensors. However what still remains conventional is battery lifetime. If these sensor devices are used simultaneously, the battery life of the phone is decreased more rapidly. In addition, overhead incurred to invoke the web service is also a detail to consider, which we studied further in [89] and developed *PowerSense*, an extensible module for DDMA that proactively facilitates the management of processes for web service assisted mobile applications. We omit discussion of PowerSense for sake of brevity. The intent is to obtain minimal power consumption by leveraging an adjacent web service infrastructure for processing if the resource results in a smaller resource footprint on the smart-phone device.

5.8.2 Power Consumption



	256x243 VGA	620x590 1MP	860x819 2MP
Local	0.51249132	2.906244	5.75011008
Remote	0.75929	2.82022	4.55574

Figure 5.18: Diagram of the average power consumption on phone given images of various resolutions.

Physicians have preference of several resolutions and the ability to use a web service as an extra resource for processing thus minimizing the power consumption on the smart-phone device. We took the average runtime to process an image under a certain resolution. Images of VGA and 1MP quality show a minimal power savings difference if processed via the web service versus locally. Thus, if the user prefers high quality the tradeoff is time (which the user has the ability to manipulate). For example, if the user has preference for faster image processing time, then they can offload processing to the web service. However, if the user is using an image of lower quality resolution, the benefits of a web service are negligible and the transmission overhead consumes more energy. Our experiments show that under WIFI wireless operating conditions, images above 1 MP gain greater time and energy savings by using the web service. Images of 1 MP resolution have generally equal power consumption assuming ideal wireless network conditions (Figure 5.18).

5.8.3 Implications

There are several important implications on how this would affect dengue diagnosis in developing countries. Due to the portable nature of the cell phone diagnosis can be done at the patients home. Mobile diagnosis aids in the treatment of patients in rural areas. Subsequently, patients will be provided with the recommended treatment for dengue to prevent death. The mobile diagnosis will have the largest impact on children who are the most susceptible to the serious side effects of dengue.

The transmission of positive test results to the CDC can also aid in the control and prevention of dengue. The cellular communication also contains information on the corresponding tower of the transmission. This location information can be used to determine areas that are susceptible to epidemics. If various agencies have accurate information on the number, location, and type of dengue outbreaks, they can work on improving the waste and water management in those particular regions. These prevention and control techniques can mitigate the further spread of the disease and improve the quality of peoples lives in the developing country. Additionally, affected travelers will be aware of their condition and be advised not to travel to aid in recovery and prevent propagation of the disease.

When a new technology is introduced into the healthcare arena, it is important to note its limitations as well as its capabilities. The portable nature of the cell phone and medical patch can also raise difficulties in maintaining the equipment. In developing countries with high crime, cellular phones may be susceptible to thieves. The cellular phone is a valuable resource in developing countries and could be used for other purposes. However, a key factor that should be considered that contributes to the effectiveness of our solution is the wavelength sensitivity of the CMOS optical sensor. The sensors level of sensitivity does play a role in the quality of an image.

Additionally, the cost of the cellular service was not put into the cost analysis. The assumption that the cellular network is functioning in the remote region also may not be a valid assumption for some developing countries. The cost of using the cellular service and the availability of the service can vary greatly depending on the countrys resources.

5.9 Future Directions

To further enhance the capabilities of our application, we are continuing research on better image processing algorithms to effectively identify the patch given background noise. The active contouring algorithm is very effective at identifying objects within images. However, the algorithm involves many complex sets of operations to bind an object of interest. Further research will involve taking into account general concepts of how the algorithm works to create a hybrid version given that the only object of interest is our patch in an image.

Lastly, an area of special interest resides in the security and authenticity of data from the cellular phone. Ensuring the integrity of a patients medical record is a very important consideration in addition to ensuring that a patients medical data remains safe if the patients cell phone is stolen or lost. Researchers at the Wireless Health Institute at UCLA have developed a medical device called a Gateway, which serves as a secure data warehouse for medical data, in addition to a communication controller device for medical information uploads to a third party. This device is small enough to clip on the belt or reside in their pocket of a patient. This device can ensure data integrity by facilitating the upload of dengue outbreak information to the CDC, and encrypting a patients medical information before uploads to ensure integrity.

5.10 Discussion

In conclusion, we have presented a novel approach for detecting dengue using the processing power of mobile phones. This approach provides a method for detecting dengue that is highly cost effective. Using light weight image processing algorithms, the cell phone is used to analyze the patch and accurately determine the disease state with limited processing capabilities and memory. Our system requires no off board processing and provides the user with simple and timely feedback. Additionally, a web service displays spatial information of outbreaks and facilitates the monitoring and diagnosis of outbreaks. However, in order for maintain a high level of accuracy for the information presented, QoI/VoI metrics must be followed so that the information processor (DDMA-WS) can provide accurate information to the user. We envision that our patch and similar real-time mobile diagnosis systems will be affordable and available in large quantities at local retailers. The availability and affordability of these systems will allow testing to be performed quickly and efficiently when a patient experiences symptoms of an infectious disease.

CHAPTER 6

Conclusion

6.1 What do Dengue and UV have in Common?

Both Dengue and Skin Cancer are prevalent issues facing our society, and we have presented two health and wellness monitoring applications that crowdsource sensor data feeds in order to measure a phenomenon of interest (namely Dengue and UV irradiance). These applications are designed with a narrow focus on specific types of sensor data feeds, and QoI is important. In the simple case, DDMA would leverage three separate sensors each diagnosing a specific phase of the dengue virus and returning a boolean value (assuming the commercially available sensors), and in the complex case UVG can leverage different types of sensors: One that reports the UVI and the other UV irradiance. The UVI can be converted into a truncated UV irradiance value, but with lesser precision.

Rather the traditional practice of developing a "closed" SSW platform where applications such as the aforementioned operate using closed data storage with coupled analytics for a single domain, these applications can publish their sensor data streams to our Sensite platform, and be presented with a manifest list of relevant sensor data streams to bind to. These applications can then use our QoILibrary to associate sensor data analysis algorithms on the fly, and they can simply re-use their views (eg. using portlets) to display analytics without bringing down their system to refactor code.

6.2 General Purpose Extensible Framework for SSW Developers

Traditionally, WSNs were designed, deployed and operate in rather “closed” set-ups, where WSNs are intimately tied to their applications. The recent Internet of Things (IoT) and Web of Things (WoT) standard took the first step enabling a more “open” setup, provided the applications have knowledge of the communication protocol that the SSW device uses to stream data. These solutions do not address the challenge of providing the right sensor data that the information processor needs in order to make better decisions. What if SSW developers could be presented with a manifest list of relevant sensor feeds to bind to ordered by *Relevance* or *Resolution* and they simply extend processing algorithms to a library (such as our QoI Library) to process the feeds that they choose to bind to? This type of solution solves the problem that we described in [18], enabling general purpose extensibility for SSW developers using IoT enabled sensors in an open deployment to build applications that may search, select and bind (and unbind) dynamically to sensor data streams that can best support their current information needs.

6.3 Other Applications

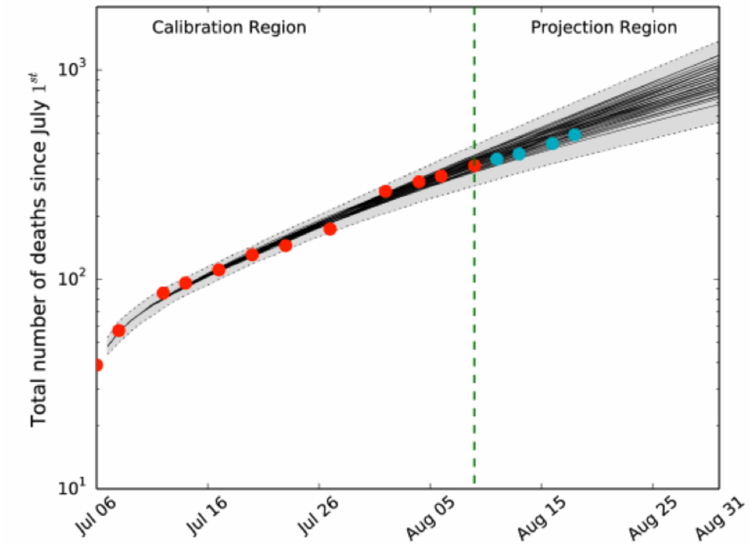


Figure 6.1: Cumulative number of EVD deaths in West Africa as of 1 July 2014 [1].

In 2013, the recent outbreak of Ebola virus disease (EVD) that started in West Africa has spread to other territories such as Guinea, Liberia and Sierra Leone. In August 2014, the death toll in those countries has reached roughly 844 EVD confirmed deaths [1, 101], with a projected rate of infection slated to continue rising (see Figure 6.1).

More recently, outbreaks have been confirmed in the USA [102]. The same dengue outbreak application that we presented in chapter 5 can be modified with minimal overhead in order to track EVD outbreaks. Our Sensite sensor data platform can be leveraged to store relevant QoI/VoI attributes to assess the confidence of the diagnosis, and Physicians can be deployed strategically to key outbreak sites in order treat patients and save lives by treating patients early.

CHAPTER 7

Related Work

7.1 Ontology Based Models for Sensor Data

The structure of our QoI library framework leverages technologies developed for SSW [103]. These are recast and extended to serve the specific purpose of our framework. In SSW, sensing systems and tasks are described through domain specific ontologies and sensor data are enriched with metadata annotations. The purpose of the SSW is to create manageable and extensible frameworks that use semantic technologies (such as the OWL semantic language [15]) to facilitate human computer and machine-to-machine interactions and reasoning pertaining the abundance of sensory data and processes. This is an active research area with forums exclusively dedicated to it. Regarding sensor metadata, the SensorML is an XML-based language that describes sensing platforms, data, and processes [17].

Researchers in [104,105] consider leveraging concepts from OWL and SensorML to create an ontology for representing the capabilities of sensors, their operational platforms, and their operating environment. In [104], the authors develop an ontology based language for dynamically assigning predefined tasks, such as detecting, distinguishing, and identifying predefined objects (e.g., buildings or vehicles) of interest to available sensors. In [105], the authors use similar concepts as in [104], however they create an interaction layer between humans and sensors by defining a semantic model to fuse and consume information from various sensors, in addition to defining a language for managing resources connected to sensor platforms.

Finally, regarding quality computations for localization [106] studies the quality of trilateration based localization and proposes an iterative technique the localization accuracy.

7.2 SSW Data Publishing Platforms

Madden et. al. [107] developed Tiny Aggregation (TAG), a web based sensor data aggregation platform for ad hoc networks of TinyOS motes. TAG provides a web interface for publishing, and an SQL based language for querying sensor data. Sensor data queries are performed by a root node and distributed across a closed sensor network of TinyOS motes. Their contribution shows that an SQL-like query syntax can offers greater flexibility for retrieving sensor and can enrich the capabilities of Senite if integrated.

Central Nervous System for the Earth (CeNSE) [28], developed by Hewlett Packard is a closed setup enterprise sensor data platform that consists of a network of nano-scale sensors designed to feel, taste, smell, see, and hear what is going on in the world. When deployed, these sensors will collect and transmit these sensor data to information operators, which will analyze and act upon the information in real time. Example scenarios for CeNSE include smelling a gas leak, monitoring traffic conditions (eg: speed and volume of traffic flow), infrastructure wear and tear, or tracking the spread of viruses. CeNSE's web interface enables users to perform queries for phenomena of interest and receive annotated sensor data. This is the exact thing that Sensite does. However, our platform does not operate in a closed environment. Sensite embraces openness where sensing agents can publish any type of sensor data, and our knowledgebase will associate sensor with all of the phenomena that the sensor is qualified to provide information about.

GeoSense [108] from the MIT Media Lab is an open sensor data publishing platform for the visualization, social sharing and analysis of sensor data. Cosm [24] is a company that is dedicated to providing a platform for users with internet enabled Geiger radiation detectors to upload sensor data in real-time. BISHAMON [109] aims to provide an web based platform for individuals and communities to visualize and assess pre-processed Geiger detector sensor data logs that are provided by a remote controlled vehicle-mounted radiation monitoring system.

These applications provide a platform for storing specific or arbitrary types of sensor data, however does not provide a comprehensive set of contextually related classifiers that describe the heterogeneous environmental contexts that the sensor data may be applicable to. In other words, these sensor data may be applicable to measuring and/or describing many other types of phenomena (eg: shooter localization, nuclear radiation, etc.), however there is no clear model that enables the user and/or application to obtain as much information as possible to make an informed decision about a certain phenomenon of interest without having to perform multiple queries for the different types of applicable sensor data, and implement algorithms to fuse the data. Moreover, having sufficient subject matter knowledge of the applicable types of sensor data is also a requirement. We have developed an ontology based model that enables a community of users to publish sensor data, extend algorithms to assess the QoI and VoI of the sensing agent and/or its sensor data. In addition, enabling subject matter experts to classify the qualified sensing agents that are capable of providing applicable context related information to describe a phenomenon.

7.3 Applications Related to Dengue Detector Mobile Application

There is a wide and diverse body of related research that investigates computational approaches to detect, model, and prevent the spread of contagious diseases and explores lightweight approaches to image processing on resource constrained embedded systems. We restrict our attention only on the most directly related research and developmental results in the detection of dengue in countries with limited resources and image processing on mobile systems.

Lensless Ultra-wide-field Cell monitoring Array platform based on Shadow imaging (LUCAS): Recently, Ozcan et al. developed a novel blood analysis tool to detect HIV and Malaria with the Lensless Ultra-wide-field Cell monitoring Array platform based on Shadow imaging (LUCAS) and a cellular phone [110]. Researchers introduced a small blood sample to three individual stacked trays placed on top of the CMOS sensor of a cell phone. The light from a Light Emitting Diode (LED) exposes distinctive signatures of blood cells by filtering the wavelength of light that passes through the trays. After light filtering, images of the shadows cast from blood cells are then uploaded to the LUCAS processing platform server and processed using template matching algorithms to identify certain signatures of diseased blood cells. Previous approaches involve a tedious process of using expensive microscopes to analyze blood cells and examining each tray one by one to detect anomalies.

LUCAS provides a novel approach to viral detection through microscopy. However further research has found that polymerase chain reaction (PRC) assays are far more effective for detecting specific types of dengue viral strands. This is evident through the ability to construct specialized primer sequences, based upon the genomic sequence of specific types of dengue viral strands. These primers anneal to their respective viral genome sequence to isolate a specific dengue type [111]. LUCAS can reveal dengue viral strands on the cell surface in addition to sometimes inside the cell, however is not able to classify the viral strand types as efficiently as our dengue patch. This enables our patch to distinctly diagnose the level of dengue infection.

CMUcam: Another area of research that has been gaining attention is image processing on mobile devices. Researchers at Carnegie Mellon have spearheaded a project called CMUcam. The objective of this project is to provide a framework for image processing with a small CMOS color camera module, coupled with a high speed embedded ARM processor. The embedded device is a fully programmable open source library tailored for general image processing tasks [112].

This device provides high speed processing, but lacks networking capability. The cost of this device is roughly $100.00, which is comparable to the cost of a cell phone. CMUcam provides a good foundation for robotics and artificial intelligence frameworks [112]. However, the solution may not be viable in a developing country due to the need for an additional computer to program the interface. In contrast, our simple algorithm is capable of performing analysis of the patch without an additional processor.

Dengue Testing Kits: A number of groups have attempted to develop accurate dengue tests to detect the presence of dengue antibodies or electrolyte abnormalities indicating the presence of a serious disease [113]. However, many of these tests are unavailable to developing nations due to their high cost or the requirement for complicated electronics and machinery. To overcome this limitation, researchers are developing a novel, low-cost patch to detect the presence of infection and associated abnormalities [?, 91, 93, 111, 113]. This paper-based sensor utilizes colorimetric detection schemes to note the presence of abnormalities in the patients blood.

Martinez developed a chromatography paper with a hydrophobic, UV-sensitive polymer pattern to detect biomarkers in small samples as low as 5μl. The hydrophobic pattern directs the flow of the sample along the hydrophilic paper to regions within the overlay that has enzymatic detecting reagents. The paper detects glucose and protein biomarkers and has small pores that prevent the contamination of foreign substances such as sediment or dirt [91, 114–117]. The techniques described in this paper use the mPAD prototype as input to detect clinically relevant concentrations of glucose and protein in artificial urine. The authors state that they envision their current process of using Physicians to manually analyze the color shades of the patch to eventually be replaced by mobile phones scanning the patch and diagnosing the patient [114]. Our work builds upon their study by providing a solution that removes the manual process of analyzing the color shades of the patch using a lightweight image processing algorithm on the cellphone.

Dengue fever rapid test devices, also known as one-step dengue tests, are typically PCR or immuno-chromatographic based assays for the rapid, qualitative and differential detection of dengue IgG and IgM antibodies to dengue fever virus in human blood [118]. A medical kit developed in India adopts a version of PCR bioassay methodology to detect common strands of Dengue. This novel invention is called Erba Den-GO [118]. Transasia Bio-Medical developed this kit to detect strands of dengue in India. Erba Den-GO is a simple testing kit that uses PCR based assays, which is a process where the viral strands in the DNA are separated by synthesis of oligonucleotide primers (genomic sequences designed to anneal to their respective viral genomes), amplified and then analyzed to detect dengue viral strands. The box includes a small hand held testing device with a small reservoir in the center. In this reservoir, the user mixes approximately 10μl of blood with a few drops of a chemical compound used to isolate the viral strands. After about 15 min, the patch reveals lines to reflect the stage of dengue.

Rapid detection kits are a good solution for detecting dengue, however kits such as these are priced between $200 (Dengue Fever Rapid DipStick Test) to $700 (Dengue Fever IgG/IgM Card Test Kit). The above methodologies offer a good approach to dengue detection. However, our solution provides a methodology to detect dengue that is more affordable to third world countries. In addition to performing diagnosis on the cell phone, the paper based patch that can cheaply be manufactured in bulk. The current cost estimate for the patch is 20 cents.

Mobile Image Processing: Our assumption is that medical personnel in the developing country have at least a cell phone device with an integrated CMOS camera. The algorithm that we use to analyze the patch does not require heavy processing power. Our algorithm is also simple enough to run on cell phone devices with limited processing capability.

Our research presents an easy method to detect dengue using the processing power of a cell phone and a small medical patch that turns different color shades. Our solution involves a bio-assay based test, and a lightweight image processing algorithm to diagnose the level of disease. Alternative solutions typically involve traditional tests using bio-assays, in addition to using microscopy to detect the viruses. Further detail regarding traditional dengue detection methods are defined in later sections.

7.4 Applications Related to Ultraviolet Guardian

In 1995, Diffey et. al. [69] proposed an embedded device that incorporates a miniature UVB sensitive sensor and a data logger that could be clipped to the lapel or on a waste belt in order to estimate cumulative UV exposure. To our knowledge this is the first proposed personal UV exposure monitoring device capable of recording exposure on a per second basis. Their goal was to better understand the outdoor activities of humans in Sunlight. The controller itself could be placed in the pocket or clipped to a belt buckle. The device has a cosine weighted angular response, therefore making it comparable to Polysulfone dosimeters. Erythemal irradiance samples were taken every two seconds for up to two hours max.

Herlihy et. al. [84] conducted a series of experiments to quantify the percentage of UV irradiance that each anatomical body site receives with respect to horizontal irradiance measurements. The 94 participants wore polysulfone badges on their cheek, hand, shoulder, back, chest, thigh, and calf. In addition, the same badge, denoted as a surface badge, was used as a control in order to correlate the site specific badges against the horizontal (ie: ground) measurements. This work contributes useful information in order to understand the exposure percentage at relative body sites after the activity is performed, however they lack information about the SZA at the time the measurements were taken. For instance, if the SZA is at 0° (ie: solar noon) then 0% of the torso is exposed to the visible sky, compared to a SZA of 70% (ie: sun nearing horizon morning/late afternoon). UV radiation at the later SZA is more direct as the rays hit a smaller vertical body surface area. Therefore, its not possible to accurately use their body site relative exposure percentage estimates for body site exposure estimation.

Vernez et. al. [119] UVSimEx is a simulation tool developed by et. al. that applies computer graphics algorithm to generate a 3D model of UV exposure sites given information about the environment, the Sun's position, and the user. Assuming certain general properties, their algorithm estimates cumulative body exposure over the whole body by integrating slices of the body in order to calculate the total energy per unit area that the skin has received. Our work differentiates from theirs by using sensors and a light weight UV exposure estimation model capable of running on a cellphone. This enables our algorithm to provide real-time estimates.

CHAPTER 8

Future Work

This work lays the foundation for an unsupervised learning algorithm that discovers the relationships between sensors and phenomena without users having to annotate sensor data with a bag of terms. There is still many areas that the research community can build on top of to further improve our algorithm.

SQL-like Queries: The contributions of [107, 120, 121] has proven that an SQL-like query syntax is an effective method for querying sensor data using complex queries. Currently, the Sensite platform implements a simple query language that enables the user to recall a single phenomenon that instantaneously occurred at a given spatiotemporal moment. The SQL query language has evolved to a multifaceted set of commands that can be combined in such a way to retrieve a range of events. Therefore, integrating a structured query engine into the Sensite platform will enable information operators to perform complex sensor data queries.

Complete Sentence Prediction: Due to the size limitation of our dataset, we were unable to train our N-gram based algorithm to predict the likelihood of a complete sentence using words. However, our algorithm still achieved 87% correct predictions. The research community can contribute by applying a larger and more comprehensive training dataset. Companies such as Microsoft and Google [122–124] have implemented N-Gram based algorithms for sentence prediction and have access to a large corpus of unstructured text that can greatly improve the accuracy of our algorithm. The accuracy of our algorithm can be vastly improved if a more comprehensive training data set is applied.

Context Recognition by Semantic Similarity: We have laid the groundwork for a knowledge-base that learns the relationships between sensors and the many phenomena that they are qualified to provide information about by using semantic similarity algorithms to interpret the meaning of unstructured text. Our list contains only a small percentage of the many possible sensor types and worldly phenomena. The research community can extend this work by applying an algorithm capable of automatically learning the names of terms that describe a sensor or phenomenon. This will enable Sensite to cover a broader spectrum of relationships.

References

[1] Marcelo FC Gomes, AP Piontti, Luca Rossi, Dennis Chao, Ira Longini, M Elizabeth Halloran, and Alessandro Vespignani. Assessing the international spreading risk associated with the 2014 west african ebola outbreak. *PLOS Currents Outbreaks*, 2014.

[2] Helge Kragh. Jack s. goldstein, a different sort of time: The life of jerrold r. zacharias, scientist, engineer, educator. cambridge, mass, and london: MIT press, 1992. pp. xviii + 373. ISBN 0-262-07138. $31.50. *The British Journal for the History of Science*, 26(02):255–255, 1993.

[3] Kathleen M. Stafford, Christopher G. Fox, and David S. Clark. Long-range acoustic detection and localization of blue whale calls in the northeast pacific ocean. *The Journal of the Acoustical Society of America*, 104(6):3616–3625, 1998.

[4] ChristopherP. Twomey. The McNamara line and the turning point for civilian scientist-advisers in american defence policy, 1966-1968. *Minerva*, 37(3):235–258, 1999.

[5] Jacob Van Staaveren. *Interdiction in Southern Laos, 1960-1968: The United States Air Force in Southeast Asia.* Center for Air Force History, 1993.

[6] Richard F. Rashid and George G. Robertson. Accent: A communication oriented network operating system kernel. *SIGOPS Oper. Syst. Rev.*, 15(5):64–75, December 1981.

[7] R. Rashid, D. Julin, D. Orr, R. Sanzi, R. Baron, A. Forin, D. Golub, and M. Jones. Mach: a system software kernel. In *COMPCON Spring '89. Thirty-Fourth IEEE Computer Society International Conference: Intellectual Leverage, Digest of Papers.*, pages 176–178, February 1989.

[8] C. Myers, A. Oppenheim, Randall Davis, and W. Dove. Knowledge based speech analysis and enhancement. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, volume 9, pages 162–165, March 1984.

[9] Chee-Yee Chong, Kuo-Chu Chang, and S. Mori. Distributed tracking in distributed sensor networks. In *American Control Conference, 1986*, pages 1863–1868, June 1986.

[10] Chatschik Bisdikian, Joel Branch, Kin K Leung, and Robert I Young. A letter soup for the quality of information in sensor networks. In *Pervasive Computing and Communications, 2009. PerCom 2009. IEEE International Conference on*, pages 1–6. IEEE, 2009.

[11] Chatschik Bisdikian, Lance M Kaplan, Mani B Srivastava, David J Thornley, Dinesh Verma, and Robert I Young. Building principles for a quality of information specification for sensor information. In *Information Fusion, 2009. FUSION'09. 12th International Conference on*, pages 1370–1377. IEEE, 2009.

[12] OpenRTMS. *An Open Source Real Time Mobile Sensor Platform.* http://openrtms.org/.

[13] Google. *Google Cloud Platform.* http://data-sensing-lab.appspot.com/.

[14] SafeCast. *Open Sensor Data Platform.* http://blog.safecast.org/maps/.

[15] Deborah L McGuinness, Frank Van Harmelen, et al. OWL web ontology language overview. *W3C recommendation*, 10(2004-03):10, 2004.

[16] Adam Pease, Ian Niles, and John Li. The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, volume 28, 2002.

[17] M Botts. OGC implementation specification 07-000: OpenGIS sensor model language (SensorML)-open geospatial consortium. Technical report, Tech. Rep, 2007.

[18] J. Matthews, C. Bisdikian, L.M. Kaplan, and Tien Pham. An ontology-based framework for designing a sensor quality of information software library. In *2010 Eleventh International Conference on Mobile Data Management (MDM)*, pages 264–269, 2010.

[19] Mohsen Darianian and Martin Peter Michael. Smart home mobile RFID-based internet-of-things systems and services. In *Advanced Computer Theory and Engineering, 2008. ICACTE'08. International Conference on*, pages 116–120. IEEE, 2008.

[20] Ali Ziya Alkar and Umit Buhur. An internet based wireless home automation system for multifunctional devices. *Consumer Electronics, IEEE Transactions on*, 51(4):1169–1174, 2005.

[21] Huan-Sheng Ning and Qun-Yu Xu. Research on global internet of things' developments and it's construction in china. *Dianzi Xuebao(Acta Electronica Sinica)*, 38(11):2590–2599, 2010.

[22] John Markoff. Google cars drive themselves, in traffic. *The New York Times*, 10:A1, 2010.

[23] Mark Campbell, Magnus Egerstedt, Jonathan P How, and Richard M Murray. Autonomous driving in urban environments: approaches, lessons and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928):4649–4672, 2010.

[24] Cosm. *How it Works.* https://cosm.com/how_it_works.

[25] LogMein. *Xively.* https://xively.com/?from_cosm=true.

[26] Antonio Pintus, Davide Carboni, and Andrea Piras. Paraimpu: a platform for a social web of things. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 401–404. ACM, 2012.

[27] ThingSpeak. *ThingSpeak.* https://thingspeak.com/.

[28] HP labs' central nervous system for the earth project aims to build a planetwide sensing network: HP labs feature article (november 2009).

[29] Murat Demirbas, Murat Ali Bayir, Cuneyt Gurcan Akcora, Yavuz Selim Yilmaz, and Hakan Ferhatosmanoglu. Crowd-sourced sensing and collaboration using twitter. In *World of Wireless Mobile and Multimedia Networks (WoWMoM), 2010 IEEE International Symposium on a*, pages 1–9. IEEE, 2010.

[30] Charu C Aggarwal and Tarek Abdelzaher. Integrating sensors and social networks. In *Social Network Data Analytics*, pages 379–412. Springer, 2011.

[31] Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. # earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1):124–147, 2013.

[32] Wikipedia. *List of Sensor Devices.* http://en.wikipedia.org/wiki/List_of_sensors.

[33] Martha E Francis and James W Pennebaker. LIWC: Linguistic inquiry and word count. *Dallas, T.: Southern Methodist University*, 1993.

[34] LIWC. *LIWC Output Variable Information.* http://www.liwc.net/descriptiontable3.php.

[35] William B Cavnar, John M Trenkle, and others. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

[36] Mark Kantrowitz and Shumeet Baluja. *Method for rule-based correction of spelling and grammar errors.* Google Patents, September 2003. US Patent 6,618,697.

[37] Karen Kukich. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4):377–439, 1992.

[38] Jacob Eisenstein. 1 recap of supervised machine learning. 2012.

[39] Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37, 2009.

[40] D Lewis, Fan Li, Tony Rose, and Yiming Yang. The reuters corpus volume i as a text categorization test collection. *Journal of Machine Learning Research*, 2003.

[41] W Nelson Francis and Henry Kucera. Brown corpus manual: Manual of information to accompany a standard corpus of present-day edited american english for use with digital computers. *Brown University, Providence, Rhode Island, USA*, 1979.

[42] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[43] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics, 2004.

[44] Lieve Hamers, Yves Hemeryck, Guido Herweyers, Marc Janssen, Hans Keters, Ronald Rousseau, and Andr Vanhoutte. Similarity measures in scientometric research: the jaccard index versus salton's cosine formula. *Information Processing & Management*, 25(3):315–318, 1989.

[45] Paul Jaccard. *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques rgions voisines.* Rouge, 1901.

[46] Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. The evaluation of sentence similarity measures. In *Data Warehousing and Knowledge Discovery*, pages 305–316. Springer, 2008.

[47] Taher H Haveliwala, Aristides Gionis, Dan Klein, and Piotr Indyk. Evaluating strategies for similarity search on the web. In *Proceedings of the 11th international conference on World Wide Web*, pages 432–442. ACM, 2002.

[48] *President Ronald Reagan : Health & Medical History.* http://www.doctorzebra.com/prez/g40.htm.

[49] SkinCancer.org. *Skin Cancer and Skin of Color - SkinCancer.org.* http://www.skincancer.org/prevention/skin-cancer-and-skin-of-color.

[50] June K Robinson. Sun exposure, sun protection, and vitamin d. *Jama*, 294(12):1541–1543, 2005.

[51] National Park Service. *What is Climate Change?* http://www.nps.gov/goga/naturescience/climate-change-causes.htm.

[52] Kevin E. Trenberth, John T. Fasullo, and Jeffrey Kiehl. Earth's global energy budget. *Bulletin of the American Meteorological Society*, 90(3):311–323, March 2009.

[53] AF McKinlay and BL Diffey. A reference action spectrum for ultraviolet induced erythema in human skin. *CIE j*, 6(1):17–22, 1987.

[54] Robert Walraven. Calculating the position of the sun. *Solar Energy*, 20(5):393 – 397, 1978.

[55] Joseph J. Michalsky. The astronomical almanac's algorithm for approximate solar position (19502050). *Solar Energy*, 40(3):227 – 235, 1988.

[56] Roberto Grena. An algorithm for the computation of the solar position. *Solar Energy*, 82(5):462 – 470, 2008.

[57] TF Eck, PK Bhartia, PH Hwang, and LL Stowe. Reflectivity of earth's surface and clouds in ultraviolet from satellite observations. *Journal of Geophysical Research: Atmospheres (19842012)*, 92(D4):4287–4296, 1987.

[58] JR Herman, R Hudson, R McPeters, R Stolarski, Z Ahmad, X-Y Gu, S Taylor, and C Wellemeyer. A new self-calibration method applied to TOMS and SBUV backscattered ultraviolet data to determine long-term global ozone change. *Journal of Geophysical Research: Atmospheres (19842012)*, 96(D4):7531–7545, 1991.

[59] TF Eck, PK Bhartia, and JB Kerr. Satellite estimation of spectral UVB irradiance using TOMS derived total ozone and UV reflectivity. *Geophysical Research Letters*, 22(5):611–614, 1995.

[60] Antti Arola, Stelios Kazadzis, Nickolay Krotkov, Alkis Bais, Julian Grbner, and Jay R Herman. Assessment of TOMS UV bias due to absorbing aerosols. *Journal of Geophysical Research: Atmospheres (19842012)*, 110(D23), 2005.

[61] Sari Kalliskota, Jussi Kaurola, Petteri Taalas, Jay R Herman, Edward A Celarier, and Nikolay A Krotkov. Comparison of daily UV doses estimated from nimbus 7/TOMS measurements and ground-based spectroradiometric data. *Journal of Geophysical Research: Atmospheres (19842012)*, 105(D4):5059–5067, 2000.

[62] JR Herman, N Krotkov, E Celarier, D Larko, and G Labow. Distribution of UV radiation at the earth's surface from TOMS-measured UV-backscattered radiances. *Journal of Geophysical Research: Atmospheres (19842012)*, 104(D10):12059–12076, 1999.

[63] A. DAVIS, G. H. W. DEANE, and B. L. DIFFEY. Possible dosimeter for ultraviolet radiation. *Nature*, 261(5556):169–170, May 1976.

[64] A. Davis, B. L. Diffey, and T. K. Tate. A PERSONAL DOSIMETER FOR BIOLOGICALLY EFFECTIVE SOLAR UV-b RADIATION. *Photochemistry and Photobiology*, 34(2):283–286, 1981.

[65] B Diffey. Personal ultraviolet radiation dosimetry with polysulphone film badges. *Photo-dermatology*, 1(3):151–157, June 1984.

[66] Michael G. Kimlin. Techniques for assessing human UV exposures. volume 5156, pages 197–206, 2003.

[67] AV Parisi, LR Meldrum, MG Kimlin, JC Wong, J Aitken, and JS Mainstone. Evaluation of differences in ultraviolet exposure during weekend and weekday activities. *Physics in medicine and biology*, 45(8):2253–2262, August 2000.

[68] Marcelo de Paula Correa, Sophie Godin-Beekmann, Martial Haeffelin, Colette Brogniez, Franck Verschaeve, Philippe Saiag, Andrea Pazmino, and Emmanuel Mahe. Comparison between UV index measurements performed by research-grade and consumer-products instruments. *Photochemical & Photobiological Sciences*, 9(4):459–463, 2010.

[69] B. L. Diffey and P. J. Saunders. BEHAVIOR OUTDOORS AND ITS EFFECTS ON PERSONAL ULTRAVIOLET EXPOSURE RATE MEASURED USING AN AMBULATORY DATALOGGING DOSIMETER. *Photochemistry and Photobiology*, 61(6):615–618, 1995.

[70] Navid Amini, Jerrid E Matthews, Foad Dabiri, Alireza Vahdatpour, Hyduke Noshadi, and Majid Sarrafzadeh. A wireless embedded device for personalized ultraviolet monitoring. *BIODEVICES*, 9:200–205, 2009.

[71] BloomSky. *World's 1st Smart Weather Camera.* http://www.bloomsky.com/.

[72] CliMate. *Create Your Own Friendly Environment.* http://rootilabs.com/.

[73] Stormtag. *StormTag - Bluetooth Weather Station.* http://hex3.co/products/stormtag-bluetooth-personal-weather-station.

[74] Jerrid Matthews, Farnoosh Javadi, Gauresh Rane, Jason Zheng, Giovanni Pau, and Mario Gerla. Ultraviolet guardian - real time ultraviolet monitoring: Estimating the pedestrians ultraviolet exposure before stepping outdoors. In *Proceedings of the 2Nd ACM International Workshop on Pervasive Wireless Healthcare*, MobileHealth '12, pages 45–50, New York, NY, USA, 2012. ACM.

[75] Xiusheng Yang, Gordon M Heisler, Michael E Montgomery, Joseph H Sullivan, Edward B Whereat, and David R Miller. Radiative properties of hardwood leaves to ultraviolet irradiation. *International journal of biometeorology*, 38(2):60–66, 1995.

[76] Martin W Allen, Kimberley Miller, and Myles Cockburn. SunSmart UV dosimetry programmes: From southern californian to new zealand schools.

[77] Alfio V Parisi, Michael G Kimlin, Joe CF Wong, and Meegan Wilson. Solar ultraviolet exposures at ground level in tree shade during summer in south east queensland. *International journal of environmental health research*, 11(2):117–127, 2001.

[78] Peter Gies, Robin Elix, David Lawry, Jennifer Gardner, Trevor Hancock, Sarah Cockerell, Colin Roy, John Javorniczky, and Stuart Henderson. Assessment of the UVR protection provided by different tree species. *Photochemistry and photobiology*, 83(6):1465–1470, 2007.

[79] Peter G Parsons, Rachel Neale, Penny Wolski, and Adele Green. The shady side of solar protection. *The Medical Journal of Australia*, 168(7):327–330, 1998.

[80] Haruka Yoshimura, Hui Zhu, Yunying Wu, and Ruijun Ma. Spectral properties of plant leaves pertaining to urban landscape design of broad-spectrum solar ultraviolet radiation reduction. *International journal of biometeorology*, 54(2):179–191, 2010.

[81] Richard H. Grant, Gordon M. Heisler, and Wei Gao. Estimation of pedestrian level UV exposure under trees. page 75, 2002.

[82] Martin Allen and Richard McKenzie. Enhanced UV exposure on a ski-field compared with exposures at sea level. *Photochemical & Photobiological Sciences*, 4(5):429–437, 2005.

[83] JM Elwood, JAH Lee, SD Walter, T Mo, and AES Green. Relationship of melanoma and other skin cancer mortality to latitude and ultraviolet radiation in the united states and canada. *International journal of epidemiology*, 3(4):325–332, 1974.

[84] Ellen Herlihy, Peter H. Gies, Colin R. Roy, and Michael Jones. PERSONAL DOSIMETRY OF SOLAR UV RADIATION FOR DIFFERENT OUTDOOR ACTIVITIES. *Photochemistry and Photobiology*, 60(3):288–294, 1994.

[85] A. Milon, J.-L. Bulliard, L. Vuilleumier, B. Danuser, and D. Vernez. Estimating the contribution of occupational solar ultraviolet exposure to skin cancer. *British Journal of Dermatology*, 170(1):157–164, 2014.

[86] Nathan Downs and Alfio Parisi. Measurements of the anatomical distribution of erythemal ultraviolet: a study comparing exposure distribution to the site incidence of solar keratoses, basal cell carcinoma and squamous cell carcinoma. *Photochemical & Photobiological Sciences*, 8(8):1195–1201, 2009.

[87] Netatmo. *Netatmo June.* https://www.netatmo.com/en-US/product/june.

[88] Jerrid Matthews, Rajan Kulkarni, George Whitesides, Majid Sarrafzadeh, Mario Gerla, and Tammara Massey. A light-weight solution for real-time dengue detection using mobile phones. *MOBICASE 2009*, 2009.

[89] Jerrid Matthews, Max Chang, ZhiNan Feng, Ravi Srinivas, and Mario Gerla. PowerSense: Power aware dengue diagnosis on mobile phones. In *Proceedings of the First ACM Workshop on Mobile Systems, Applications, and Services for Healthcare*, mHealthSys '11, pages 6:1–6:6, New York, NY, USA, 2011. ACM.

[90] Jerrid Matthews, Rajan Kulkarni, Mario Gerla, and Tammara Massey. Rapid dengue and outbreak detection with mobile systems and social networks. *Mobile Networks and Applications*, 17(2):178–191, April 2012.

[91] Andres W Martinez, Scott T Phillips, Manish J Butte, and George M Whitesides. Patterned paper as a platform for inexpensive, low-volume, portable bioassays. *Angewandte Chemie International Edition*, 46(8):1318–1320, 2007.

[92] Sharon de T Martins, Guilherme F Silveira, Lysangela R Alves, Claudia Nunes Duarte dos Santos, and Juliano Bordignon. Dendritic cell apoptosis and the pathogenesis of dengue. *Viruses*, 4(11):2736–2753, 2012.

[93] Duane J Gubler. Dengue/dengue haemorrhagic fever: history and current status. In *Novartis foundation symposium*, volume 277, page 3. Chichester; New York; John Wiley; 1999, 2006.

[94] Hilde M van der Schaar, Michael J Rust, Chen Chen, Heidi van der Ende-Metselaar, Jan Wilschut, Xiaowei Zhuang, and Jolanda M Smit. Dissecting the cell entry pathway of dengue virus by single-particle tracking in living cells. *PLoS pathogens*, 4(12):e1000244, 2008.

[95] Mara G Guzman and Gustavo Kouri. Dengue and dengue hemorrhagic fever in the americas: lessons and challenges. *Journal of Clinical Virology*, 27(1):1–13, 2003.

[96] Mary E Wilson and Lin H Chen. Dengue in the americas. *Dengue Bulletin*, 26:44–61, 2002.

[97] Organizacin Panamericana de la Salud. *Nueva generacin de programas de prevencin y control del dengue en las Amricas*. OPS Washingtonˆ eD. CDC, 2002.

[98] Phillip R Shaver and Kelly A Brennan. Attachment styles and the" big five" personality traits: Their connections with each other and with romantic relationship outcomes. *Personality and Social Psychology Bulletin*, 18(5):536–545, 1992.

[99] Michal Kosinski, Yoram Bachrach, Pushmeet Kohli, David Stillwell, and Thore Graepel. Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, 95(3):357–380, 2014.

[100] Tracii Ryan and Sophia Xenos. Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage. *Computers in Human Behavior*, 27(5):1658–1664, 2011.

[101] World Health Organization Regional Office for Africa. Ebola virus disease.

[102] Michael Klompas, Daniel J Diekema, Neil O Fishman, and Deborah S Yokoe. Ebola fever: Reconciling ebola planning with ebola risk in US hospitals. *Ann Intern Med*, 2014.

[103] Amit Sheth, Cory Henson, and Satya S Sahoo. Semantic sensor web. *Internet Computing, IEEE*, 12(4):78–83, 2008.

[104] Geeth De Mel, Murat Sensoy, Wamberto Vasconcelos, and Alun David Preece. Flexible resource assignment in sensor networks: A hybrid reasoning approach. 2009.

[105] Jean-Sbastien Brunner, Jean-Franois Goudou, Patrick Gatellier, Jrme Beck, and Charles-Eric Laporte. SEMbySEM: a framework for sensors management. In *1st International Workshop on the Semantic Sensor Web (SemSensWeb 2009)*, page 19, 2009.

[106] Zheng Yang and Yunhao Liu. Quality of trilateration: Confidence-based iterative localization. *Parallel and Distributed Systems, IEEE Transactions on*, 21(5):631–640, 2010.

[107] Samuel Madden, Michael J. Franklin, Joseph M. Hellerstein, and Wei Hong. TAG: A tiny AGgregation service for ad-hoc sensor networks. *SIGOPS Oper. Syst. Rev.*, 36(SI):131–146, December 2002.

[108] Anthony Anthony Lawrence DeVincenzi. *GeoSense: An open publishing platform for visualization, social sharing, and analysis of geospatial data*. PhD thesis, Massachusetts Institute of Technology, 2012.

[109] Yoh Kawano, David Shepard, Yugo Shobugawa, Jun Goto, Tsubasa Suzuki, Yoshihiro Amaya, Masayasu Oie, Takuji Izumikawa, Hidenori Yoshida, Yoshinori Katsuragi, and others. A map for the future: Measuring radiation levels in fukushima, japan. In *Global Humanitarian Technology Conference (GHTC), 2012 IEEE*, pages 53–58. IEEE, 2012.

[110] Sungkyu Seo, Ting-Wei Su, Anthony Erlinger, and Aydogan Ozcan. Multi-color LUCAS: Lensfree on-chip cytometry using tunable monochromatic illumination and digital noise reduction. *Cellular and Molecular Bioengineering*, 1(2-3):146–156, 2008.

[111] Robert S Lanciotti, Charles H Calisher, Duane J Gubler, Gwong-Jen Chang, and A Vance Vorndam. Rapid detection and typing of dengue viruses from clinical samples by using reverse transcriptase-polymerase chain reaction. *Journal of clinical microbiology*, 30(3):545–551, 1992.

[112] Anthony Rowe, Charles Rosenberg, and Illah Nourbakhsh. A low cost embedded color vision system. In *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, volume 1, pages 208–213. IEEE, 2002.

[113] Philippe Dussart, Laure Petit, Bhety Labeau, Laetitia Bremand, Alexandre Leduc, David Moua, Sverine Matheus, and Laurence Baril. Evaluation of two new commercial tests for the diagnosis of acute dengue virus infection using NS1 antigen detection in human serum. *PLoS neglected tropical diseases*, 2(8):e280, 2008.

[114] Andres W Martinez, Scott T Phillips, Emanuel Carrilho, Samuel W Thomas III, Hayat Sindi, and George M Whitesides. Simple telemedicine for developing regions: camera phones and paper-based microfluidic devices for real-time, off-site diagnosis. *Analytical Chemistry*, 80(10):3699–3707, 2008.

[115] Andres W Martinez, Scott T Phillips, and George M Whitesides. Three-dimensional microfluidic devices fabricated in layered paper and tape. *Proceedings of the National Academy of Sciences*, 105(50):19606–19611, 2008.

[116] Andres W Martinez, Scott T Phillips, Benjamin J Wiley, Malancha Gupta, and George M Whitesides. FLASH: a rapid method for prototyping paper-based microfluidic devices. *Lab on a Chip*, 8(12):2146–2150, 2008.

[117] Andres W Martinez, Scott T Phillips, George M Whitesides, and Emanuel Carrilho. Diagnostics for the developing world: microfluidic paper-based analytical devices. *Analytical chemistry*, 82(1):3–10, 2009.

[118] Erba Den-Go. *Erba Den-Go Rapid Action Force for Dengue.* http://transasia.co.in.204-11-59-220.mdus-pp-wb10.webhostbox.net/Common/Uploads/DMS/120600.pdf.

[119] David Vernez, Antoine Milon, Laurent Francioli, Jean-Luc Bulliard, Laurent Vuilleumier, and Laurent Moccozet. A numeric model to simulate solar individual ultraviolet exposure. *Photochemistry and Photobiology*, 87(3):721–728, 2011.

[120] Wesley W. Chu, Alfonso F. Crdenas, and Ricky K. Taira. KMeD: A knowledge-based multimedia medical distributed database system. *Information Systems*, 20(2):75 – 96, 1995. Scientific Databases.

[121] John DN Dionisio and Alfonso F Crdenas. MQuery: a visual query language for multimedia, timeline and simulation data. *Journal of Visual Languages & Computing*, 7(4):377–401, 1996.

[122] Liang-Chih Yu, Chung-Hsien Wu, Andrew Philpot, and EH Hovy. OntoNotes: sense pool verification using google n-gram and statistical tests. In *Proc. of the OntoLex Workshop at the 6th International Semantic Web Conference (ISWC 2007)*, 2007.

[123] Kuansan Wang, Christopher Thrasher, Evelyne Viegas, Xiaolong Li, and Bo-june Paul Hsu. An overview of microsoft web n-gram corpus and applications. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 45–48. Association for Computational Linguistics, 2010.

[124] Adam Pauls and Dan Klein. Faster and smaller n-gram language models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 258–267. Association for Computational Linguistics, 2011.