Context-aware Anomaly Detection and Analysis using Spatial-Temporal Data

by

Qi Liu

B.E., Harbin Institute of Technology, 2010

M.S., University of Colorado Boulder, 2014

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Electrical, Computer and Energy Engineering

2018

This thesis entitled:
Context-aware Anomaly Detection and Analysis using Spatial-Temporal Data
written by Qi Liu
has been approved for the Department of Electrical, Computer and Energy Engineering

_____

Prof. Li Shang

_____

Prof. Qin Lv

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Liu, Qi (Ph.D., Electrical Engineering)

Context-aware Anomaly Detection and Analysis using Spatial-Temporal Data

Thesis directed by Prof. Li Shang

With the thriving of sensing and internet-of-things technologies, an increasing number of research communities and industries are stepping into the Era of Big Data. Following this technology trend, the amount and complexity of data generated by each domain are growing exponentially. The demand for automated monitoring, detecting and analyzing unusual events from those data are also increasing. These predictive analyses seek to identify and capture meaningful patterns in massive, highly heterogeneous data from various domains such as environmental sensing and cyber-physical systems. However, performing analysis such as anomaly detection faces a variety of challenges. For instance, the lack of prior knowledge regarding what is normal and what is abnormal, and the power consumption limitation for low-profile computing devices. These challenges constrain the flexibility of analysis methods. All these pose real problems to existing anomaly detection methods. Most existing techniques for anomaly detection only consider the content of the data source, i.e., the data itself directly gathered from sensing devices, not taking the context of the data into consideration. Therefore, anomalies under complicated settings are difficult to be identified. Hence, it is essential to design anomaly detection methods, especially the feature space design under a specific anomaly context. The context can be semantic, spatial, or temporal.

This thesis studies the context-aware data analysis approaches using spatial-temporal data. A general principle to design a context-aware data analysis framework for spatial-temporal data is proposed and investigated in three different problems: contextual anomaly detection in remotely sensed imagery, hierarchical context-aware fault diagnosis in photovoltaic systems and energy-efficient wearable computing empowered by context-aware predictive analysis. Results include: (1) an automated contextual anomaly detection approach is proposed and implemented. The method constructs and utilizes spatial-temporal neighborhood context. Average precision and recall of 98.1% and 95.7%

for contextual outlier detection are achieved. Also, meaningful and validated unusual events are detected from remotely sensed imagery. (2) A new hierarchical context-aware anomaly detection algorithm is proposed. With this algorithm, the fault detection accuracy of large-scale photovoltaic systems improves by 20% (from 63% to 83%) for top-100 detected anomalies, compared with existing solutions. (3) By identifying and predicting the intra-signal context, the proposed sparse adaptive sensing algorithm achieves 97.7% accuracy with 76.9% to 99% reduced energy consumption (83.6% average reduction under real-world testing).

These three studies demonstrate the utility of combining the spatial-temporal context in any future big data anomaly detection.

Dedication

To my beloved family.

Acknowledgements

Firstly, I would like to acknowledge and thank my Ph.D. advisor, Li Shang, and co-advisor Qin Lv. Their knowledge and vision guided me through many challenging moments during my research. And I would like to acknowledge and thank all the other professors on my Ph.D. committee. A lot of motivation and research points came from discussions with them. To the multitude that I have not mentioned, you have my lasting gratitude and thanks.

Finally, this work was supported in part by the National Science Foundation (NSF) under award No. 1251257 and No. 0910995, and the National Natural Science Foundation of China under award No. 61233016. I thank for their support.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Anomaly detection is the process of identifying individual items or events (groups of data points) which do not conform to an expected pattern or other items in the dataset [34]. Outliers, novelties can also be referred as anomalies. For instance, anomalies are usually referred to as frauds in bank and insurance industries, while they may be treated as novelties (e.g., unusual weather patterns) in environmental sensing communities. Anomaly detection is an important research area as the results of which are expected to provide actionable information for analysts on time or even ahead of time. For instance, an anomaly detection system can inform the emergence of unusual weather conditions for Earth scientists or report faulty operations in the manufacturing process to factory operators, etc. Typically, from the perspective of the demand on prior knowledge, anomaly detection techniques are categorized into three groups: unsupervised, supervised and semi-supervised [34]. The common deliverables from these models are the predicted labels for new data observations. For instance, the result can be a label for each data point, which indicates whether the data point is normal or abnormal. In this thesis, besides predicting the categories of data observations, further analysis of the characteristics of detected anomalies is conducted. For instance, whether an anomaly is caused by noise or induced by unusual events. Also, a different angle of detecting and utilizing contextual anomalies is investigated, providing insights and opportunities to reduce power consumption from sensing devices. For example, measurements acquired from sensors are usually with a predetermined sampling rate. However, regular patterns could be monitored under low sampling rate as those are easy to predict. Special or novel patterns could require higher sampling

rates as those are hard to predict. Therefore, this thesis also discusses how the signal's temporal context can be utilized to detect pattern changes and hence predict sampling rate for present and future data. This study can help optimize power consumption from sensing devices.

Before performing anomaly detection, the conceptual types of 'anomalies' needs to be determined. There are two key types of anomalies: point and contextual [34, 158]. Most previous research has focused on detecting **point anomalies** [34, 67, 23], which are individual data points that are considered globally anomalous (e.g., extreme low temperature). This thesis focuses on the less-studied **contextual anomalies** [158, 12], especially for spatial and temporal datasets, under dynamic conditions (e.g., environmental or human-driven). Contextual anomalies are relative anomalies under specific contexts. For example, a high air temperature trend in the summer may be usual, but if the same temperature trend occurs during a winter period, it could potentially be due to data defects or anomalous atmospheric processes [105, 96, 25]. The majority of the prior anomaly detection work focused on the point anomalies without considering the application-specific contextual information. However, the capability to incorporate contextual information from the studied domains can significantly improve the process of anomaly detection with low-complex model design and high accuracy. For instance, for spatial-temporal data collected by individual imaging sensor or a network of distributed sensors can have rich contextual information. Such information may include spatial, temporal locality across sensors or from a single sensor. Depending on the season, or weather condition, the definition of 'anomalies' can vary. For instance, a current sensor reading from a photovoltaic panel may have low values compared with its historical readings, standard point anomaly detection approach will recognize this as an anomaly, however, considering the weather condition, if the sensor readings come from a rainy or cloudy period, then the values would be normal.

1.1 Problem Motivation

As before mentioned, large amounts of spatial-temporal data are generated by various deployed sensors. These sensing systems enable researchers or analysts to 'sense' targets beyond the

natural capabilities of human eyes or ears. The sensing targets can vary from objects in the environment (e.g., weather, water), machines in cyber-physical systems (e.g., photovoltaic panels in solar farms), to activities of human (e.g., running, falling). For all these areas, it is important to determine whether and when anomalies, such as extreme weather, faulty photovoltaic panels, occur, providing necessary information to study, prevent or leverage those anomalies. However, for all those applications, there are several common challenges associated with the anomaly detection, not to mention the unique challenges from each application domain. Those shared challenges originate from the volume, velocity, and variety of the datasets.

- **Lack of prior knowledge**. Because of the cost of manual data labeling, the high volume, and high velocity of datasets, prior knowledge about normal or abnormal patterns hardly exist. This largely limits the options of anomaly detection techniques. Under such situations, unsupervised or semi-supervised techniques are preferable, and recently this has become a paradigm shift in machine learning research community. Supervised learning approaches are incapable of handling current or future Big Data problems. The unsupervised analysis is the ultimate learning solution.

- **Dirty data**. The resolution, accuracy and the effective lifetime of sensors are limited by the manufacturing processes and also the financial budget. Hence, sensor data are usually contaminated by noise and faulty readings. Additionally, sensors often operate in dynamics environment. The changes of environment can also introduce noise or cause the dysfunction of sensors. Therefore, noise, faulty readings from those interesting anomalies need to be taken into consideration while designing the anomaly detection methods.

- **Evolution of anomalies**. The concept of anomalies evolves with various factors, such as the weather conditions, machine operation status and so on. Hence, methods designed for points anomaly detection approaches are limited in accuracy for most spatial-temporal data applications.

- **Difficulty of adaptation**. Existing solutions are shown to be effective in their target domain, but adapting the methods to other domains is quite challenging. And the reason why this is challenging is that the natures of spatial-temporal datasets and anomalies are fundamentally divergent among different domains.

Therefore, in this thesis, we are particularly interested in developing anomaly detection solutions for complex spatial-temporal datasets, from environmental monitoring to human activity sensing. The objective is two-fold: (1) Conceptualize a general methodology to tackle contextual anomaly problems from feature engineering to model construction, and provide insights into adaptation to different domains using the similar feature engineering and modeling process. (2) Provide not only the categories of observations but also analyze the intrinsic properties of each anomaly, providing actionable information for further investigations and usage.

1.2 Research Contributions

Unsupervised Anomalous Event Detection Method for Large Satellite Datasets. Massive amounts of remotely sensed data are being generated at an unprecedented rate, offering new opportunities for data-driven scientific discovery in the Earth sciences and related domains. However, due to the sheer volume of remotely sensed data and the lack of practical data analysis tools, most data remain in the dark, with little to no quality assurance and limited access to high-level analytical tools. Anomaly detection aims to find scenarios that differ from the norm and is of particular importance when analyzing remotely sensed data. However, most previous work has focused on identifying individual anomalies, and required prior knowledge of the ground truth for supervised learning. To tackle the anomaly detection problem for spatial-temporal satellite data, an unsupervised anomaly detection algorithm is needed. More specifically, the algorithm shall not require prior knowledge and is capable of detecting anomalous events, which we define as groups of outlier objects differing contextually from their spatial and temporal neighbors. Such contextual anomalies can be useful in discovering both hidden quality issues in the data and real natural events

of significance. We demonstrate the effectiveness of our solution via Web-based tools developed to visualize and analyze such contextual anomalies, using two datasets. The techniques and tools developed in this project apply to a diverse set of satellite products.

Hierarchical Context-Aware Anomaly Diagnosis Solution for Large-Scale Photovoltaic Systems. Operation anomalies (e.g., faulty photovoltaic strings) are common phenomena in large-scale solar farms. Effective anomaly detection and classification is essential for improving the operation reliability and electricity generation of solar farms. However, this is a challenging task due to the high complexity and wide variety of often occurring anomalies. Furthermore, existing pre-installed supervisory control and data acquisition (SCADA) systems can only provide a limited amount of information regarding the healthy condition of solar farms. The limited information and data collection granularity make accurate anomaly detection and classification difficult.

We present a hierarchical context-aware anomaly detection and multi-modal classification solution, which can accurately detect and predict the type of a variety of photovoltaic system anomalies. The proposed solution does not require additional information beyond the pre-installed Supervisory control and data acquisition (SCADA) control system. More specifically, the proposed work consists of two methods: (1) a hierarchical context-aware anomaly detection method using unsupervised learning, and (2) a multi-modal anomaly type prediction method. As an experiment, the proposed solution has been deployed in two large-scale solar farms (39.36 MWp and 21.62 MWp). This thesis discusses the effectiveness and efficiency of the proposed solution under multi-month real-world operation.

Empowering Wearable Devices with Energy Efficient Context-Based Adaptive Sensing Algorithm As a third example of the concept, this research work demonstrates how the notion of contextual anomalies can be utilized to minimize sampling rate and hence reduce power consumption on sensing devices. Human-borne sensing systems such as wearable health monitors and activities trackers are typically built upon low-profile micro-controller and powered by small capacity batteries. To enable long-term sensing and onboard analysis for those systems,

efficient sampling strategy and low-complexity analysis algorithms are essential to reduce power consumption.

This research is different than the previous two. Contextual anomalous or novel patterns detected from this work is used to design an optimized sampling strategy further instead of reported as system defects. More specifically, Sparse Adaptive Sensing (SAS) is proposed, which selectively identifies the best sampling points to maintain high accuracy while significantly reducing sensing and analysis energy overheads. Evaluation of the methods is conducted with a wearable online analysis system - Gazelle [93, 171], which is designed for running and is compact, lightweight. Experimental results of the algorithm demonstrate 97.7% accuracy with 76.9% to 99% reduced energy consumption (83.6% average reduction under real-world testing) – a one-order-of-magnitude improvement over existing solutions. SAS enables long-term maintenance-free mobile analysis for running, with $>$ 200 days of continuous high-precision operation using only a coin-cell battery. Since 2014, Gazelle has been used by over 100 elite and recreational runners during daily training and at races like the Kona Ironman World Championships.

1.3 Thesis Organization

The remainder of this thesis is organized as follows.

- Chapter 2 outlines the background information associated with anomaly types and contextual anomaly detection for spatial-temporal data, and a literature review of the existing approaches to contextual anomalies. This chapter first introduces terminologies and preliminaries that are used throughout the rest of the thesis. Second, it provides a review of the state-of-the-art techniques for contextual anomaly detection, including an introduction to a variety of algorithms with their advantage and disadvantages. Finally, the chapter conceptualizes a general design methodology for contextual anomaly detection for various spatial-temporal data settings, from feature engineering and modeling.

- Chapter 3 presents an automated contextual anomaly detection approach for remotely

sensed imagery data. The details of spatial-temporal context are discussed, and corresponding contextual features are extracted for anomaly detection. More specifically, the development and evaluation of an extended Gaussian Mixture Model based anomaly detection algorithm are presented.

- Chapter 4 first discusses a hierarchical context-aware anomaly detection method for large-scale photovoltaic systems, leveraging the similar contextual anomaly detection modeling approach presented in Chapter 3. However, the design differences due to domain-specific problems are handled, such as the lack of apparent spatial context. Then, based on the anomalies detected, and existing anomaly labels, an anomaly type predictive model is built upon multi-modal features.

- Chapter 5 presents a different usage scenario of contextual anomaly detection. The contextual information from both intra-stride (temporal) and inter-stride (semantic) signals are investigated. Based on the context-aware critical pattern detection and prediction techniques, a real-time sparse adaptive sampling algorithm for reducing power consumption is proposed and presented.

- Finally, Chapter 6 concludes the thesis and discusses future research directions.

Chapter 2

Background and Literature Review

The goal of this chapter is two-fold: (1) Anomaly related concepts and modeling approaches are discussed. (2) An overview of the related research works for contextual anomaly detection are presented. The literature review includes prior studies specifically related to contextual detection in spatial-temporal data.

2.1 Contextual Anomalies

This section will introduce concepts and terminologies related to **anomaly detection**, **contextual attribute** and **behavior attribute**. This section serves as a primer for the concepts that are discussed further in the rest of the thesis.

2.1.1 Type of Contextual Anomaly

Contextual anomalies are context dependent. The concept of a context is induced by the structure in the dataset and needs to be specified as a part of the anomaly detection problem. Each data instance is defined using the following two sets of attributes: [34]

- **Contextual attributes**. The contextual attributes are used to define the context (or neighborhood) for a data point. For instance, in spatial datasets, the longitude and latitude of a location are the contextual attributes. In temporal data, time is a contextual attribute that determines the position of a data point on the entire time series.

- **Behavioral attributes**. The behavioral attributes define the non-contextual characteristics of a data point. For example, in a spatial dataset describing the surface temperature of the entire world, the temperature at any location is a behavioral attribute. The anomalous behavior is determined using the values for the behavioral attributes within a specific context. A data point might be a contextual anomaly in a given context, but an observation with the same behavioral attribute could be considered normal in a different context. This property is essential to identify contextual and behavioral attributes for a contextual anomaly detection technique.

Fig. 2.1 and Fig. 2.2 illustrates three types of contextual anomalies that meet the above definitions, which can be caused by unusual events, systematic or random errors.

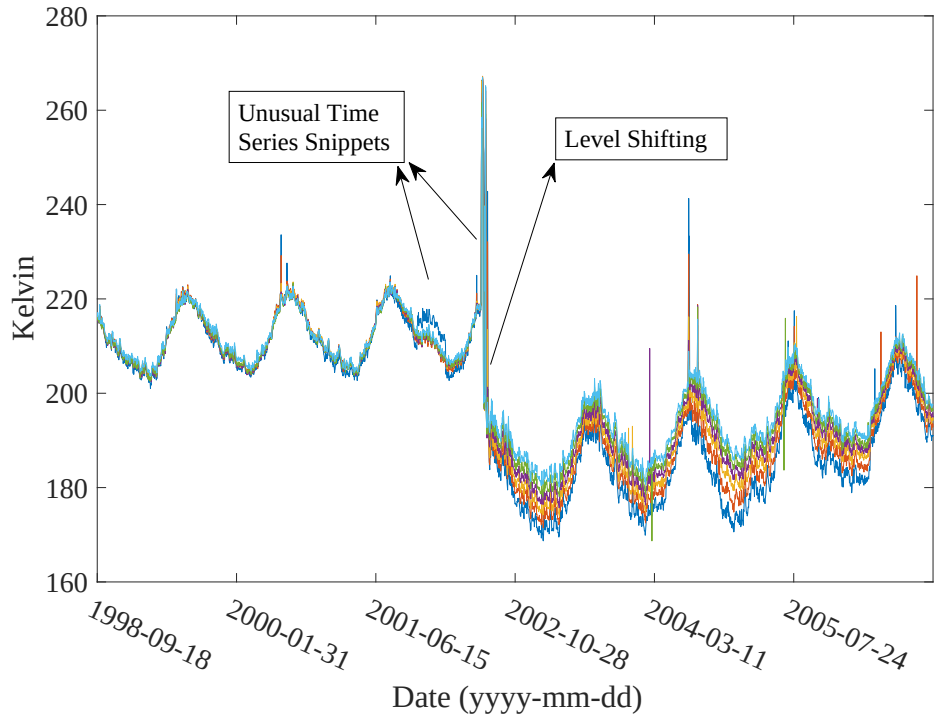

Fig. 2.1. Examples illustrating unusual time series and level shifting detected in the brightness temperature of several adjacent pixels.

- **Unusual time series snippets**. The Earth observation data usually has obvious cycles (e.g., diurnal or seasonal cycles). In addition, the period of a cycle and the average mag-

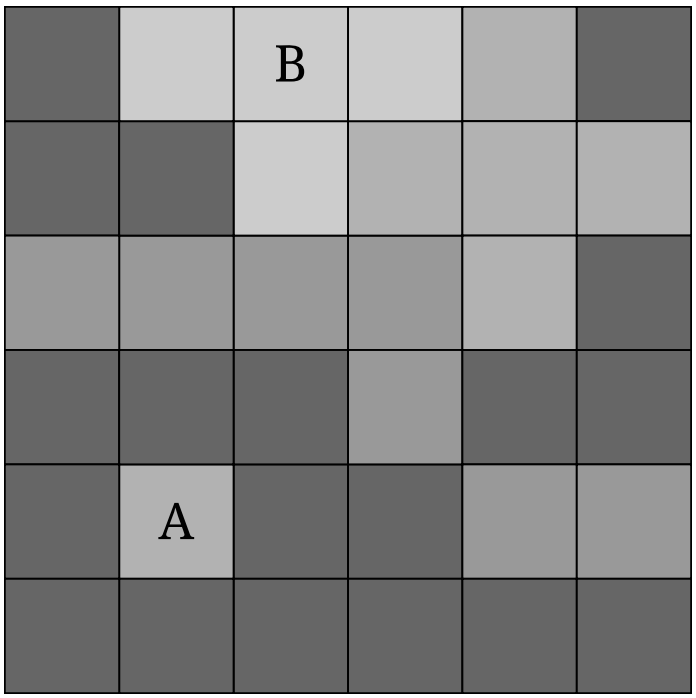

Fig. 2.2. An example of a local spatial outlier: Pixel A and B have the same value. However, pixel A is considered an outlier because of its behavior with respect to neighboring pixels whereas Pixel B is not because of its coherence with neighboring pixels.

nitude of the data in each cycle are relatively stable. However, there exist snippets in a time series that deviate from the stable pattern. For example, Fig. 2.1 shows a brightness temperature time series from several adjacent pixels. The duration of high brightness temperatures each summer is relatively stable. However, as noted in the figure, in one snippet the high brightness temperature persisted longer than usual, and in another snippet the data value was significantly higher than that of the previous summers. These unusual time series snippets can be caused by either unusual natural events or errors. Similar such anomalies also occur in the photovoltaic current time series and human running signals, which is illustrated in Chapter 4 and Chapter 5, respectively.

- **Level shifting**. Fig. 2.1 also shows a scenario when the values of a group of adjacent pixels significantly increase or decrease. This type of temporal discontinuity may appear normal when viewed spatially at a specific time, and can only be discovered when viewed as a time

series at a given location.

- **Local spatial outlier**. A pixel or an object (i.e., a block of $n \times n$ spatial pixels), which appears normal when viewed globally in an image, appears inconsistent when compared with its neighbors. As shown in Fig. 2.2, pixel A is an outlier concerning its neighbors, but normal when viewed globally. Pixel B has the same value as pixel A, but is not a local spatial outlier. Note that image data has explicit spatial context due to its nature, while though some other data collected from distributed sensor network does not have explicit spatial context, a conceptualized spatial neighborhood can be constructed. For instance, in a sensor network, sensors deployed adjacent are expected to behave more consistently than those has a longer distance. Hence, under specific application condition, sensors operated under similar environment can be clustered as a conceptualized spatial neighborhood.

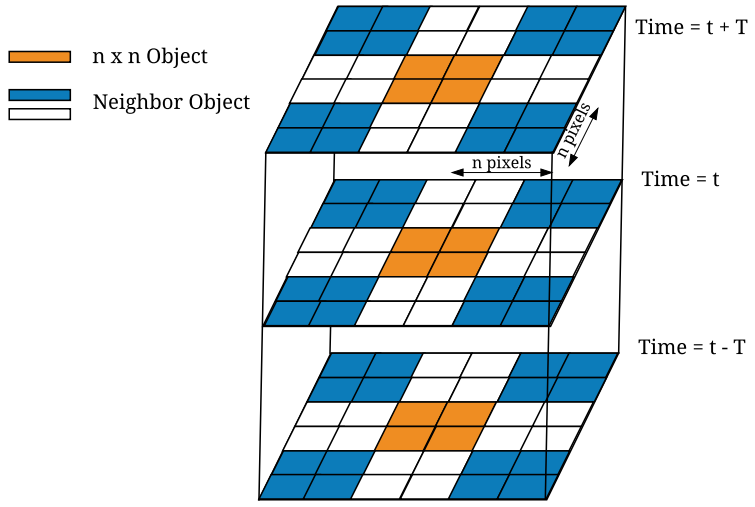

Fig. 2.3. An illustration of objects and local spatial-temporal neighborhoods. Each object is defined as a block of 2×2 pixels at a specific time. For the orange object at time t, its spatial neighbors include the 8 adjacent objects surrounding at time t (4 in blue and 4 in white), and its temporal neighbors are the other orange objects within the time range of $[t - T, t + T]$.

These three types of anomalies may all comply with the normal global data range and hence are invisible when using standard statistical analyses such as a two standard deviation criteria method against a normal data distribution. Moreover, a level shifting or an anomalous time series

snippet may be visible only from a temporal perspective. Therefore, to detect all these anomalies, both spatial and temporal contextual information is extracted from a pixel's local neighborhood. Specifically, as illustrated in Fig. 2.3, an object $o^t_{x,y}$ is comprised of a block of $n \times n$ pixels at a specific time t, where (x, y) is the location of the top left pixel of the object. Then the spatial neighborhood of the object refers to the eight spatially-adjacent objects at time t, and its temporal neighborhood refers to the set of objects $o^{t\prime}_{x,y}$ where $t\prime \in [t - T, t + T], t\prime \neq t$ and T is the window size parameter.

2.2 Problem Setting and Methodology

Since this thesis focuses on spatial-temporal data, in this section, we first formally define spatial-temporal outliers (ST-Outliers) and anomalous events. Second, feature space for spatial-temporal contextual data is proposed and described.

2.2.1 Spatial-Temporal Outliers and Events

Definition 1 *ST-Outliers: Given a set of objects $O = \{o^t_{x,y}\}$. An object $o^t_{x,y} \in O$ is considered an ST-Outlier if its non-spatial-temporal features differ significantly from its spatial or temporal neighbors in O.*

The non-spatial-temporal features represent an object's original physical value such as temperature, or derived values, e.g., temperature difference, temperature correlation and so on. Because ST-Outliers can emerge as a group due to the same natural event or systematic error, we also define anomalous events.

Definition 2 *Anomalous Events: Let D be a set of objects that are ST-Outliers, and r is an anomalous event that consists of a group of objects from D. The objects in r, which are spatial-temporally or solely temporally correlated, behave significantly different from the other objects in D regarding non-spatial-temporal features.*

2.2.2 Contextual Feature Engineering

With the definition of spatial-temporal neighborhood, we design three groups of features based on statistical parameters to describe the context of every object. Depending on the characteristics of the domain specific applications, a subset of those features or an extended set designed particularly for an application can be used as input to the anomaly detection model.

- Basic features. The mean $\mu(o_k^{(t)})$ and variance $s(o_k^{(t)})$ statistics are adopted for $o_k^{(t)}$ that containing n^2 pixels.

$$\mu_k^t = \frac{\sum o_k^{(t)}}{n^2} \tag{2.1}$$

$$s_k^t = \sqrt{\frac{1}{1/n^2 - 1} \sum (o_k^{(t)} - \mu)^2} \tag{2.2}$$

- Spatial context features. In order to capture the change of $o_k^{(t)}$ compared with its neighbors, a correlation vector $Corr_k^t = [corr_{k,1}^{(t)}, corr_{k,2}^{(t)}, \dots corr_{k,L}^{(t)}]$ and a difference vector $Diff_k^t = [diff_{k,1}^{(t)}, diff_{k,2}^{(t)}, \dots diff_{k,L}^{(t)}]$ are used to describe the spatial dynamics of $o_k^{(t)}$. $Corr_{k,l}^{(t)}, l \in [1, L]$ is the Pearson correlation between object $o_k^{(t)}$ and its neighbor $o_l^{(t)}$, while $Diff_{k,l}^{(t)}, l \in [1, L]$ is the direct difference between the object and one of its neighbors.

- Temporal context features. We use a gradient vector $Grad_{Corr,k}^{(t)} = [Corr_k^{(t,t-T)}, Corr_k^{(t,t+T)}]$ to capture the spatial correlation change of the object $o_k^{(t)}$. Where $Corr_k^{(t,t-T)} = Corr_k^t - Corr_k^{t-T}$ and $Corr_k^{(t,t+T)} = Corr_k^{t+T} - Corr_k^t$. We also use the gradient vectors of $Grad_{\mu,k}^{(t)} = [\mu_k^{(t,t-T)}, \mu_k^{(t,t+T)}]$ and $Grad_{s,k}^{(t)} = [s_k^{(t,t-T)}, s_k^{(t,t+T)}]$ to find the temporal change of object $o_k^{(t)}$.

2.2.3 Anomaly Detection Modeling

In this section, Gaussian Mixture Model (GMM) and Expectation Maximization (EM) algorithm [109] are introduced as part of the preliminaries.

2.2.3.1 Gaussian Mixture Model

The mixture model is used for density estimation. It is assumed that normal and abnormal data follow different generative models. In this thesis, mixture models are utilized to identify anomalous models. Given a dataset $D = \{x_1, x_2, ..., x_N\}$, where x_i is a d-dimension vector observation. Assume that the data points are independent and identically distributed, which are generated from an underlying density $p(x)$. As usually a dataset consists of different patterns, such as normal data, abnormal data, or sub-patterns of normality and abnormality. Hence, $p(x)$ can be defined as a finite mixture model with K linearly combined components:

$$p(x|\Theta) = \Sigma_{k=1}^{K} \alpha_k p_k(x|z_k, \theta), \Sigma_{k=1}^{K} \alpha_k = 1 \tag{2.3}$$

where:

- The $p_k(x|z_k, \theta_k)$ are mixture components, $1 \leq k \leq K$. Each is a distribution defined over $p(x)$, with parameters θ_k.
- $z = (z_1, ..., z_K)$ is a vector of K binary indicator variables that are mutually exclusive and exhaustive (i.e., one and only one of the z_k is equal to 1, and the others are 0). z is a random variable representing the identity of the mixture component that generates x. It is convenient for mixture models to represent z as a vector of K indicator variables.
- The $\alpha_k = p(z_k)$ are the mixture weights, representing the probability that a randomly selected x was generated by component k, where $\Sigma_{k=1}^{K} \alpha_k = 1$. The complete set of parameters for a mixture model with K components is $\Theta = \{\alpha_1, ..., \alpha_k, \theta_1, ..., \theta_1\}$.

Under the definition of Eq.2.3, the weight that an observation x_i belongs to cluster k is defined as,

$$w_{ik} = p(z_{ik} = 1|x_i, \Theta) = \frac{p_k(x_i|z_k, \theta_k)\alpha_k}{\Sigma_{m=1}^{K} \alpha_m p_k(x_i|z_m, \theta_m)} \tag{2.4}$$

In general, $p(x)$ can be any distributions, or densities. Gaussian distribution is by far the most popular. We can define a Gaussian Mixture model by making each of the K components a

multivariate Gaussian distribution,

$$p_k(x|\theta_k) = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\} \tag{2.5}$$

where μ is the mean vector and Σ is the covariance matrix for the Gaussian component. The parameters of Gaussian Mixture Models can be estimated using Expectation Maximization algorithm[109].

2.2.3.2 Maximum likelihood

To model the dataset $D = \{x_1, ..., x_N\}$ using Gaussian Mixture model. This dataset can be presented as an $N \times d$ matrix $\mathbf{X}$, where N is the number of data vectors and d is the dimension of the vector. Then the log-likelihood function is defined as

$$\log l(\Theta) = \Sigma_{i=1}^{N} \log p(x_i|\Theta) = \Sigma_{i=1}^{N} (\log \Sigma_{k=1}^{N} p_k(x_i|z_k, \theta_k)\alpha_k) \tag{2.6}$$

where $p_k(x_i|z_k, \theta_k)$ is the Gaussian density for the kth component.

EM algorithm for Gaussian mixtures is an iterative algorithm that starts from an initial estimation of Θ (e.g., randomly generated Gaussian distributions), and then iteratively update Θ untile convergence is detected. Each iteration consists of two steps: E-step and M-step.

E-step This step computes the membership weight w_{ik} defined in Eq.2.4 for all data observations $x_i, 1 \leq i \leq N$ in all mixture components $k, 1 \leq k \leq K$. This results an $N \times K$ matrix of membership weights.

M-step This step uses the membership weight matrix from E-step and the data D to compute and update mixture model parameters. The number of data points assigned to kth component is $N_k = \Sigma_{i=1}^{N} w_{ik}$. Hence, the updated mixture proportions are,

$$\alpha_k^{new} = \frac{N_k}{N}, 1 \leq k \leq K \tag{2.7}$$

Then the mean and covariance matrix are re-computed for each Gaussian component,

$$\mu_k^{new} = \frac{1}{N_k} \Sigma_{i=1}^{N} w_{ik} x_i \tag{2.8}$$

$$\Sigma_k^{new} = \frac{1}{N_k} w_{ik}(x_i - \mu_k^{new})(x_i - \mu_k^{new})^T \tag{2.9}$$

After all of the parameters are updated, the M-step is complete and iterates from E-step again until there is no significant change from one iteration to the next of the log-likelihood defined in Eq.2.6.

2.3 Literature Review

Anomaly detection has been a topic of active research [34, 67, 23], with the majority of research focusing on point anomaly detection. For example, one-class Kernel Fisher Discriminants [137] was proposed as a method of learning a discriminative boundary close to the normal instances, such that any test instance that does not fall within the learned boundary is considered anomalous. Knorr and Ng [83, 82] developed several distance-based outlier detection algorithms where the core methodology was to score a data instance by counting the number of nearest neighbors that are within a distance d; data instances with the lowest scores were considered outliers.

For more detailed survey for point anomaly detection algorithms, you can refer to the study of Chandola et al. [34], which summaries techniques for detecting point anomalies, including the following types: classification, nearest neighborhood, clustering, statistical modeling, information, and spectrum based. Point anomaly detection techniques are extensively used in scientific or industry data, such climate research to identify extreme events or financial institute to detect fraud behaviors. However, for contextual anomalies that fall within normal data ranges or hide in seasonal patterns, direct use of those classical outlier detection algorithms will fail. Hence, this thesis focuses on survey algorithms designed particularly for contextual anomalies using spatial-temporal data.

A series of anomaly detection methods leveraging spatial or temporal attributes have been proposed [34, 158, 12]. For instance, in the work of Vallis et al., long-term time series data was decomposed to remove seasonality before using statistical modeling to find anomalous points [162]. Spatial Local Outlier Measure (SLOM) proposed by Sun and Chawla [158] can capture the local

behavior of datum in its spatial neighborhood. Thus, local spatial outliers can be discovered, which are usually missed by global techniques like "three standard deviations away from the mean". This type of approach handles either temporal or spatial context.

In our work, we detect contextual anomalies in both spatial and temporal contexts, if both contexts exist for an application. The typical method for spatial-temporal outlier detection consists of three steps [67]: (1) Identify spatial objects from the input data. (2) Objects are analyzed to find spatial outliers. (3) Spatial outliers are then verified if they are also temporal outliers. This type of approach sequentially executes spatial and temporal outlier detection. Consequently, the output is the intersection set of spatial and temporal outliers. For example, Birant and Kut [24] proposed a density-based ST-Outlier detection method. They use DBSCAN [24] to identify spatial outliers first, then validate with their temporal neighbors. If no significant temporal difference was found, the candidate is abandoned. Similarly, Cheng and Li [38] proposed a four-step method to address the semantic and dynamic properties of geographic phenomena for ST-Outlier detection. However, to capture all possible data defects or interesting events, the union set of spatial and temporal outliers has to be detected. Therefore, we have proposed a single-step ST-Outlier detection algorithm using combined spatial-temporal features for the remotely sensed data, also proposed a derived version of hierarchical contextual anomaly detection algorithm for the distributed sensor systems (e.g., photovoltaic systems), with which we can get all the ST-Outliers and rank by the importance of those outliers, providing high-interest results for researchers or analysts and at the same time, a threshold associated with the ranking policy can also help reduce false alarms.

Furthermore, while most of the existing work only detects individual outliers, we aggregate ST-Outliers into anomalous events based on their spatial-temporal continuity, which on the one hand provide more insights into the data as they help reveal underlying processes that may have triggered groups of outliers, on the other hand, this aggregation can reduce false positives that are caused by transient environmental changes or system states transitions.

This study area of work is also related to collective anomalies [34], which have been investigated in some recent work, using statistical models. Das and Neil [49] used What's Strange About

Recent Events (WSARE) to detect anomalous clusters of counts in categorical data and performed testing to determine if a cluster is a significant anomalous pattern. And a Flexible Genre Model (FGM) was proposed by Xiong et al. [172] to discover anomalous behaviors of groups of points. In contrast to those supervised or semi-supervised methods that assume the availability of enough training data with ground truth, all those approaches proposed in this thesis are unsupervised and requires no prior knowledge of the datasets.

In summary, this literature survey serves as an overview of existing contextual anomaly detection techniques. For each domain problem studied in this thesis, domain related work and background are presented in each chapter.

Chapter 3

Unsupervised Contextual Anomaly Detection Approach for Big Remotely Sensed Data

This research partially fulfills the goal of the Condensate Database Project (NSF project number 1251257, *Condensate Database for Efficient Anomaly Detection and Quality Assurance of Massive Cryospheric Data*). Anomaly detection technique is a way to screen data quality issues and at the same time, can help construct a 'condensate database' – containing only 'interesting' data points or patterns. Hence, the overhead of storing and analyzing the 'Big Data' can be largely reduced. There are several contributors to this project, for instance, Glenn Grant's work [66] focused on the design and evaluation of a physical database and investigated various statistical point outlier detection approaches. In this thesis, we primarily focus on the design and evaluation of a general contextual anomaly detection approach, and its application to data quality assurance and rare natural events analysis.

3.1 Introduction

Recent advances in remote sensing technology have revolutionized the way remotely sensed (RS) data is acquired, managed, and analyzed [101, 130]. More than 200 on-orbit satellites are currently capturing continuous Earth observations [101], offering great opportunities for advancing the scientific understanding of the Earth's systems. However, as the proliferation of these products increases so does the complexity needed for processing them. The variety of data can vary greatly, even within individual data sets [88]. Therefore, human expert-driven data analysis, a laborious and

time-consuming process, remains the mainstream approach for data quality assessment [70, 65, 27] and scientific knowledge discovery [151, 59]. The sheer volume and complexity of RS data have hampered adequate quality assessment or higher-level analysis such as anomaly detection. While Earth scientists are very interested in studying anomalies such as climate extremes [45, 106, 55, 113], finding all such anomalies from massive data sets is challenging. Furthermore, RS data is often contaminated with noise or errors which need to be identified and then either corrected or eliminated. Thus, a high demand exists for effective and generic anomaly detection tools which require the minimal involvement of domain experts while having the ability to adapt to diverse data sets. Anomaly detection in RS data is challenging for several reasons. (1) Prior models may not exist for determining what constitutes anomalous data. Additionally, unknown types of anomalies may exist in the data. (2) Remotely sensed imagery is often contaminated with noisy pixels or missing data. (3) The dynamic nature of spatial and temporal variations in multiple frequency channels need to be considered. (4) Due to the high volume and variety of RS data, validated ground truth data sets are not often available for supervised learning. Additionally, there will always exist unusual anomalies in the data that exceed the expectations or prior knowledge of Earth scientists. Unsupervised approaches are thus preferred.

Furthermore, as contextual anomalies such as high temperature during cold seasons are usually of high importance in Earth sciences research, an effective anomaly detection algorithm that leverages both spatial and temporal contextual locality, referring the local coherence, is desired. The assumption is that in a natural environment, pixels nearby share similar morphology and evolve gradually over time, while anomalous pixels would have low coherence with their neighbors in space and time. Hence, in this chapter, a clustering-based framework for both point and contextual anomaly detection is presented, which requires no domain knowledge of the dataset and enables automated anomaly detection on diverse remotely sensed datasets.

Besides discovering individual objects ($n \times n$ pixels) that are contextual outliers relative to their spatial-temporal neighbors, it is also helpful to study these outliers collectively as **anomalous events**, which can potentially reveal unusual processes that lead to those outliers in the first place.

Such underlying processes can either be systematic errors (e.g., sensor calibration error), which require intervention for quality control, or natural events (e.g., extreme weather condition), which may lead to new knowledge [172, 146]. With the knowledge that anomalous behaviors caused by systematic errors or rare natural events can spread to a wide range of regions and last for a long period, we aggregate spatial-temporal outliers into anomalous events within a global spatial-temporal context and report those events with a ranking of their importance. Combining all the points above, we have developed a novel clustering-based solution for unsupervised detection of contextual anomalies in remotely sensed data.

The main contributions of this chapter are summarized as follows.

- The design of an unsupervised anomaly detection solution that (1) requires no prior knowledge of the data set, (2) identifies contextual outliers that differ from their spatial-temporal neighbors; and (3) groups contextual outliers into anomalous events to reveal possible underlying processes.
- Demonstration of the approach's effectiveness and efficiency using two different types of remote sensing data: brightness temperatures from Special Sensor Microwave Imager (SSM/I) and skin (surface) temperatures derived from Advanced Very High Resolution Radiometer (AVHRR).
- Identification and validation of data quality issues due to systematic or random errors as well as significant natural events.

This rest of this chapter is organized as follows. Section 3.2 presents an overview of the anomaly detection approach along with its design considerations, and also usage scenarios. Section 3.3 discusses the contextual anomaly detection in detail. Section 3.4 reports our evaluation of the proposed solution and presents case study results. Finally, Section 3.5 summarizes this work.

3.2 Anomaly Detection Framework

Climate extremes such as unusually warm and cold events are increasingly attracting the attention of Earth scientists [105, 96, 25]. Hence, both spatial-temporal outliers and anomalous events are the targets of our unsupervised contextual anomaly detection framework.

Our anomaly detection framework takes as input a time series of satellite images. Each image is usually processed at either the pixel level or the object level [69]. Noise or missing data usually appear as random, discontinuous pixels. However, interesting anomalies that represent natural events or systematic errors may often appear as a collection of adjacent pixels. As such, our proposed anomaly detection framework consists of both pixel-based and object-based analysis.

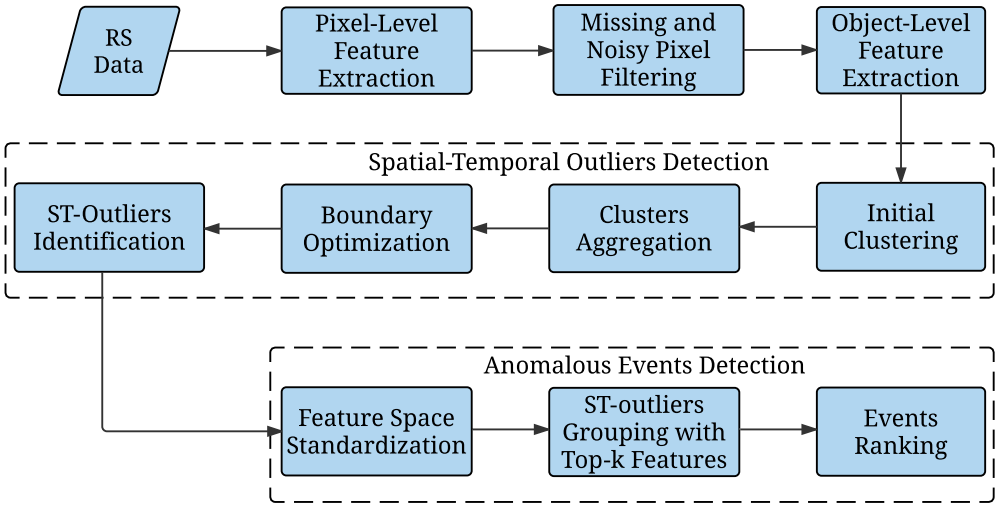

Fig. 3.1. An overview of the anomaly detection framework.

3.2.1 Overview

Fig. 3.1 gives an overview of our proposed anomaly detection solution, which consists of four main steps:

(1) **Missing and noisy pixel filtering**. RS imagery data may be contaminated by missing and noisy pixels, which would lead to skewed data distribution. For example, Fig. 3.2 shows two snapshots of the Advanced Very High-Resolution Radiometer (AVHRR) skin temperature

data for the South Pole [43]. The data values fluctuate from one location to another, as well as over time at the same location. Despite this spatial-temporal dynamic, the data record is also contaminated by random noise from clouds, instrumentation, and missing data. To reduce the bias or disturbance from noisy data when searching for interesting anomalies, we have developed a noisy pixel filtering algorithm and integrated it with the anomaly detection framework, to improve the quality of detected anomalies. This component can be used as either an independent tool for data cleansing or integrated into an anomaly detection process.

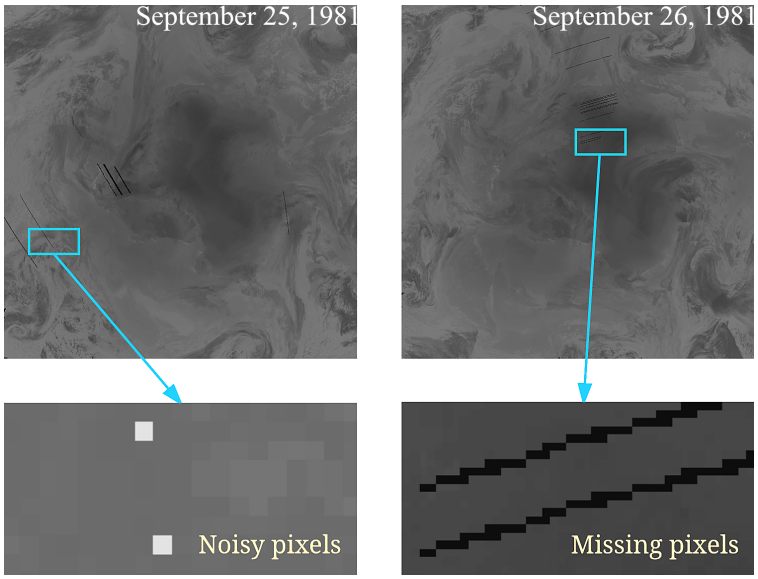

Fig. 3.2. AVHRR skin temperature data with noise and missing pixels. Examples shown for September 25 and 26, 1981.

(2) **Object-level feature extraction**. Each object consists of one or more pixels ($n \times n$). To capture the anomalous behaviors of an object, the radiometric values along with their derived features are extracted for each object, which are then compared with its neighbors in space and time to detect contextual anomalies.

(3) **ST-Outliers detection**. An object that has either low spatial or temporal coherence with its neighbors is identified as an outlier. This is accomplished through an unsupervised, clustering-based process.

(4) **Anomalous events detection**. In this step, we further group ST-Outliers that share similar anomalous behaviors (i.e., spatially and temporally correlated) as an anomalous event to help discover the underlying anomalous process, be it a natural event or systematic error.

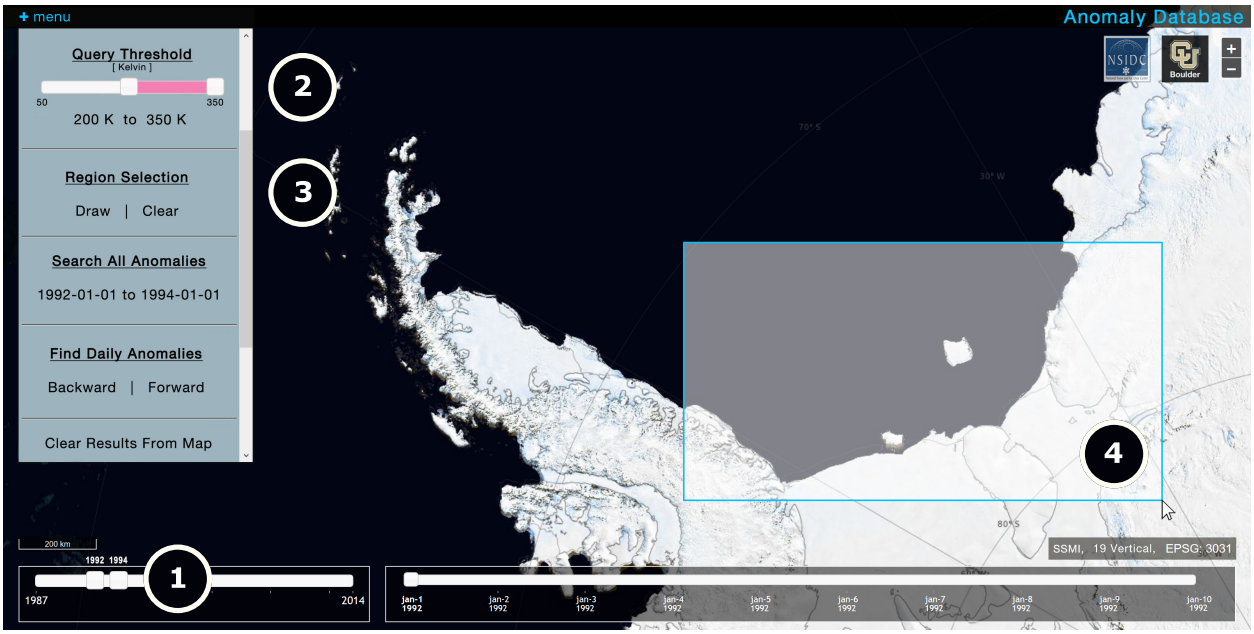

Fig. 3.3. Web-based user interface: Overall layout and key steps showing how anomalies are located.

3.2.2 Usage Scenarios

In the project of Condensate Database, a prototype of a web-based tool was also developed as a team effort [92]. Next, for the purpose of illustrating the usage scenarios of such anomaly detection tools, we briefly describe two usage examples with the web user interface.

User Interface Fig. 3.3 shows the web-based user interface (UI), which features a set of tools and enables data exploration using an intuitive mapping interface. The UI comprises of several key features that allow users to quickly select a set of parameters that include information such as the sensor type (e.g., SSM/I vs. AVHRR), frequency band (e.g., 19 GHz, 22 GHz), and polarization (e.g., vertical or horizontal), as well as select a sub-region within the map to search. Users can then

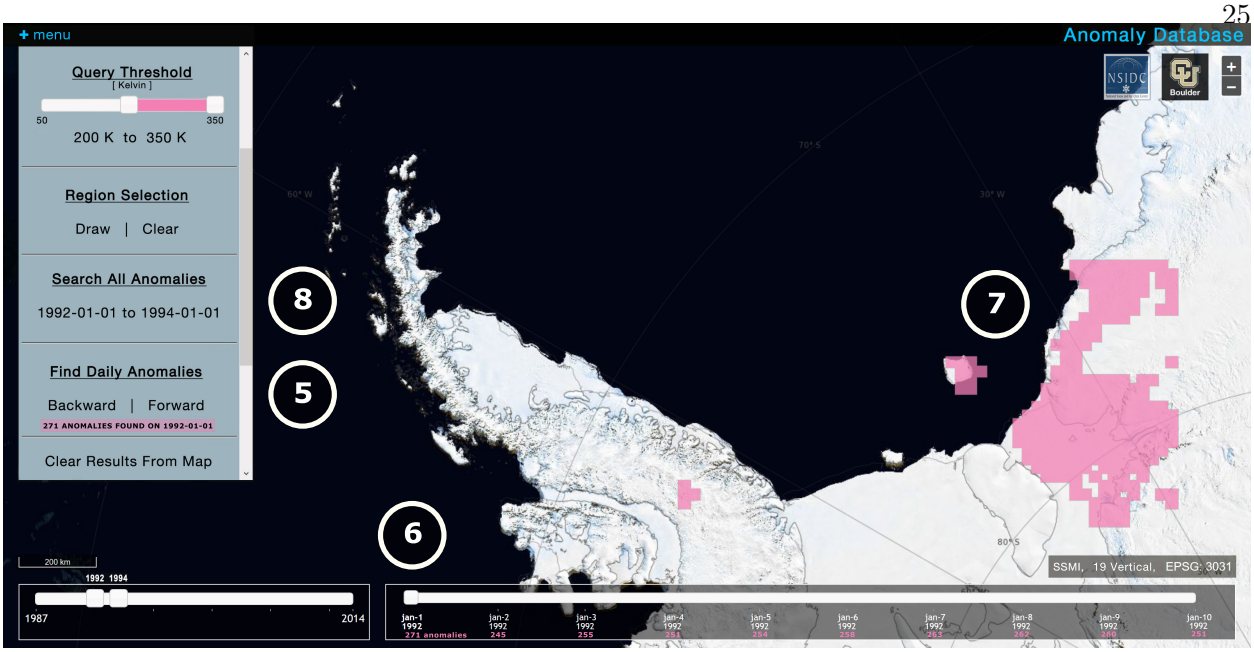

Fig. 3.4. Usage example: Querying the Weddell Sea and coast for anomalies.

explore their results in one of two ways. The first approach allows query results to be visualized day by day, where the daily anomalies and their associated metadata are used to help the user grasp dynamic changes within a particular region. The second approach is by searching for an aggregation of the data within the specified time frame. In this way, results are collected for each pixel and can be displayed to show information such as the average pixel value or the frequency with which anomalies occur at each location.

Usage Example We examine a coastal region to demonstrate how the tools can be used for exploring anomalies; the steps described are illustrated in Fig. 3.3 and Fig. 3.4. First, the user selects a date range of 1992 to 1994 of the South SSM/I 19 vertical polarization dataset, refining the query to search for anomalies above 200 Kelvin with the slider. Next, a rectangular region is selected from anywhere in the Antarctic region; the user opts to search an area within the Weddell Sea, drawing out a rectangle to partition the specific subregion they are interested in. Finally, the user then selects a query that will parse anomalies so that they can be reviewed along with the daily timeline. After the queried results are returned to the interface, they are overlaid within a layer of the map where the user has control over a temporal investigation of the data, allowing them

to traverse forwards and backward along the timeline to review results. Looking at the results, we can see that there are a significant number of events concentrated within the Filchner Ice Shelf extending up toward the Brunt Ice Shelf. This may prompt the user to refine his/her search to look at an aggregation of all anomalies during the specified two-year period, allowing the user to see where those anomalies are concentrated most, and the brightness temperatures associated with each pixel.

3.3 Anomaly Detection Algorithms

In this section, the backbone – anomaly detection algorithms in the framework are presented in detail.

3.3.1 Missing and Noisy Pixel Filtering

As mentioned earlier, most RS imagery contains missing and noisy pixels, which need to be properly identified and labeled. This information is then applied in the process of object-level feature extraction to reduce the bias introduced by those pixels. Additionally, we record noisy pixels in the anomaly database as random errors for users' reference. Missing pixels are easy to handle since a missing pixel's value is usually set to a special fill value such as 0. Hence, we focus primarily on filtering noisy pixels. Some of the noisy pixels are outside the normal data range (i.e., clear errors), and can be filtered easily using a threshold. However, some of the noisy pixels are within the normal range but obviously "wrong" when compared with their neighbors. To address these scenarios, we have designed a noisy pixel filtering algorithm, which can detect both types of noisy pixels (outside normal data range or not) in two steps: (1) identify objects which contain potential noisy pixels and (2) identify actual noisy pixels in each object.

Detect Potential Noisy Objects

An image is first divided into objects of $n \times n$ pixels each, which reduce the size of the feature space and improve computation efficiency. Note that in a real application of the framework, the size of each object depends on the actual resolution of the image and domain knowledge if exists.

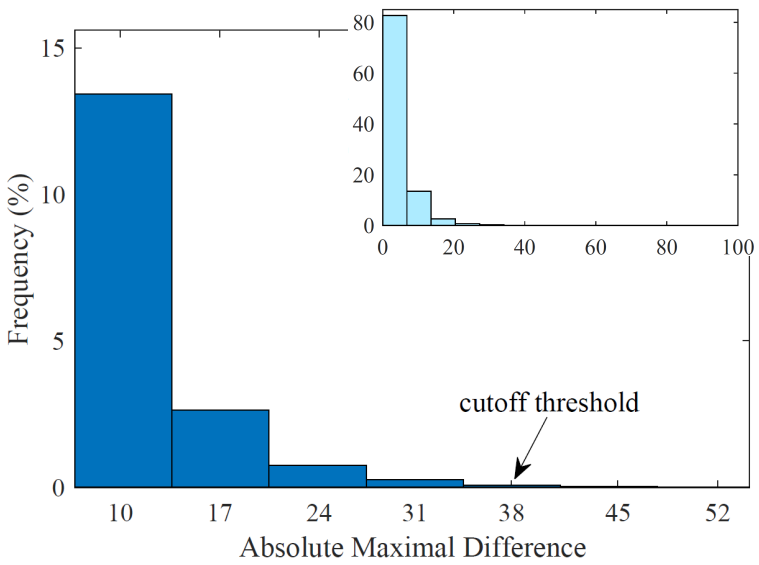

Fig. 3.5. Histograms showing the frequency distribution of absolute maximal differences between adjacent pixels in each object. The main plot (blue) highlights a subset of differences from the total distribution (cyan). The threshold for determining whether an object contains potential noisy pixels can be visually selected such that the probability of the absolute maximal difference converges toward zero.

For each object, we extract features such as the absolute maximal difference between every two adjacent pixels. Fig. 3.5 shows a distribution of the absolute maximal difference between every two adjacent pixels in an object from an AVHRR image. The cutoff value in the distribution is around 38. Empirically, this cutoff point can be utilized to find objects that contain potential noisy pixels. However, a fixed threshold is not general for all images in the AVHRR data or other data sets. Therefore, we have developed a clustering-based method to automatically determine the "cutoff" threshold. Since the the absolute maximal differences (AMD) follow a long tail distribution, as shown in Fig. 3.5, a common criterion to determine outliers from such distribution is where $value_{amd} > Q3 + m \times IQR$ or $value_{amd} < Q1 - m \times IQR$, where $Q1$ and $Q3$ are the first, third quantile, respectively. IQR is the interquartile for all AMDs. Typically, $m = 1.5$. However, it is difficult to decide a proper m for images generated from different days, as the sensing condition, and external environment changes over time. Therefore, a data-driven approach is proposed to automatically find such a threshold based on the property of data itself. The AMD feature data

is first clustered. In this study, we empirically set the number of clusters to 20. Then the second order difference of cluster centroids (sorted in the ascending order) is computed. As illustrated in Fig. 3.6, the first peak of the second order difference (shown in red) is detected, and the cluster centroid value (shown in blue) right after the peak is set as the threshold to detect objects that contain potential noisy pixels.

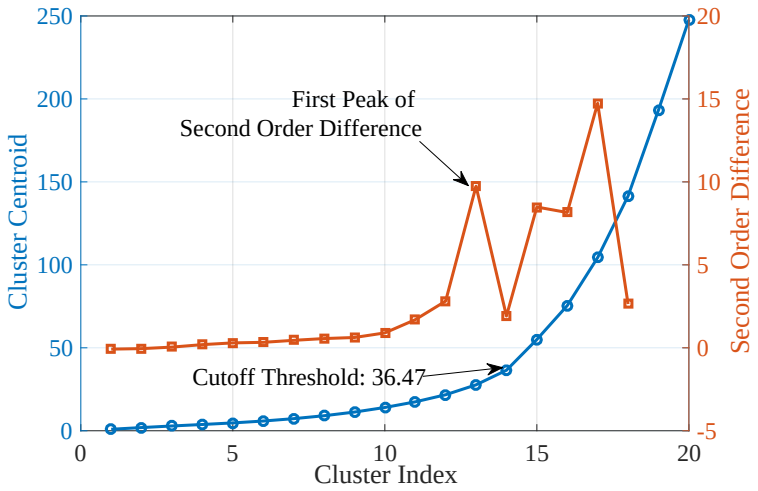

Fig. 3.6. A method for automatically determining the cutoff threshold. Step 1 finds the first peak of the second order difference (shown in red), here the cluster index is 13. Step 2 checks the cluster centroid series (shown in blue) to find the value 36.47 at index 14, that value is then used to detect objects containing potential noisy pixels.

Identify Noisy Pixels For each object detected above, we divide its pixels into two groups based on the similarity of their features, and pixels in the smaller group are identified as noisy pixels. However, there is one potential issue. As illustrated in Fig. 3.7, the pixels (P1 and P2 in the blue box) from the edge of the Antarctica Peninsula are identified as noisy pixels in an object B, but it looks normal in object A. To handle pixels that are located around edges or dynamic regions (e.g., ocean), we compute the absolute difference between each noisy pixel candidate and its neighbors. If a pixel is similar to the majority of its neighbors, the pixel is not noisy since we assume noisy pixels are random and do not occur together.

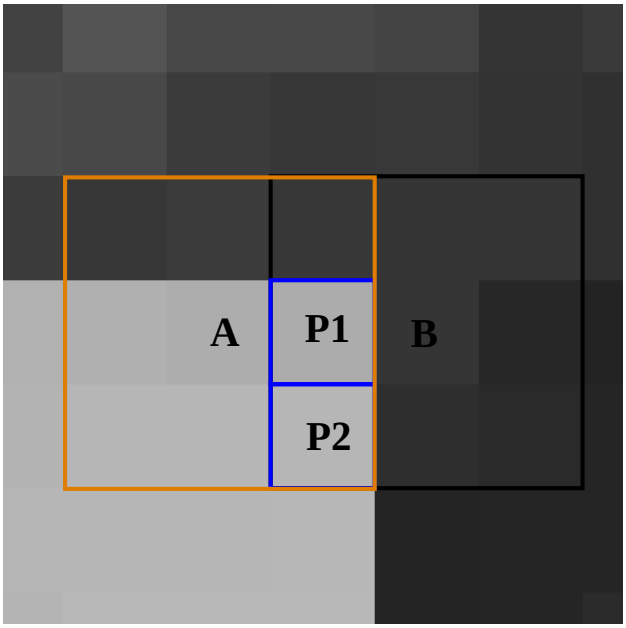

Fig. 3.7. One potential issue when identifying noisy pixels near edges and dynamic regions. Pixels P1 and P2 are observed as normal within object A (orange 3×3 pixels block), but are identified as outliers in object B (black 3×3 pixels block) due to the sharp transition at the edge of the Antarctic Peninsula.

3.3.2 Object-Level Feature Extraction

In this step, instead of extracting pixel-level features to emphasize the internal pixel-to-pixel variance of an object, we extract object-level features to describe an object with respect to its neighbors. Formally, we denote by $m^{(t)}$ the data at time point t, and by $m_{i,j}^{(t)}$ a single pixel at i^{th} row and j^{th} column in $m^{(t)}$, $o^{(t)}$ is an object constructed by $n \times n$ pixels. From the temporal perspective, t is an index of time and temporal neighborhood is within $[t - T, t + T]$, T is also a parameter depends on the temporal resolution of the data. As discussed in Chapter2, three types of contextual features: basic, spatial and temporal, can be extracted as input for anomaly detection model. For this problem, as the image data is spatial and temporal, all three features are extracted.

By transforming the satellite image time series into a feature space, we achieve three goals. (1) Reduce the impact from noisy and missing pixels: For instance, a missing pixel can be ignored by computing the mean and standard deviation of an object that contains multiple pixels; but if all pixels of an object are missing pixels, the object is abandoned. (2) Leverage the spatial features to

avoid the problem of spatial heteroscedasticity, i.e., the local distribution of the data is not uniform at different locations in an image. (3) Help remove the impact from cyclic patterns and highlight local outliers.

3.3.3 ST-Outliers Detection

Satellite imagery data often contain some cyclic patterns such as seasonal cycles. It is thus reasonable to assume that the data follows Gaussian mixture models (GMM). However, the number of clusters K requires prior knowledge and is very difficult to determine. Specifically, our case is even much complicated, because there exist various anomalies of which the exact number is unknown and random noise may not belong to any meaningful clusters. To address these issues, we propose an extended Expectation-Maximization (EM) algorithm. Fig. 3.9 illustrates the essential steps of our algorithm for detecting spatial-temporal outliers: cluster initialization, cluster aggregation, and boundary optimization. The rationale behind this approach is to focus on capturing the normal patterns, and treat small clusters and stand-alone data points as outliers because normal patterns occur more frequently than outliers and belong to denser clusters. Next, we describe the three steps in more detail.

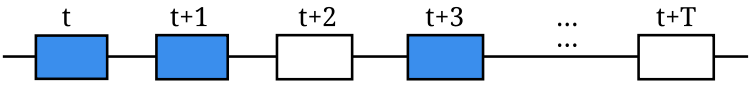

Fig. 3.8. A time series of objects at location (x, y) spanning time t to $t + T$. Normal objects are shown in blue and outlier objects are shown in white.

Cluster initialization In this stage, we detect outliers from all objects $\{o_{x,y}^{t_0}, \ldots, o_{x,y}^{t_T}\}$ at location (x, y) within the maximal temporal span, as shown in Fig. 3.8. Note that spatial-temporal outliers can be captured because both spatial and temporal features are also used here. For example, if one object is significantly different from its spatial neighbors, this object should be significantly far away from the others in the spatial feature space so that it will be assigned to a small cluster or be a stand-alone object. Moreover, since the exact number of clusters is unknown, we choose a relatively large cluster number ($K = 5$ as shown in Fig. 3.9 (a)) to perform the multivariate EM

algorithm [14, 86] in the object feature space. However, if K is too large, similar objects may be assigned to different clusters.

Cluster Aggregation To address this issue, we merge clusters with very similar statistical models. For example, in Fig. 3.9, the top-left two clusters shown in red and green in step (a) are merged into the larger red cluster in step (b). Let $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ be the initial clusters, and $\{\mathcal{N}_p(\mu_1, \Sigma_1), \ldots, \mathcal{N}_p(\mu_K, \Sigma_K)\}$ be the corresponding statistical models, where p is the dimension of the feature space, and μ, Σ are the mean vector and variance matrix respectively. We define the impact domain of model $\mathcal{N}_p(\mu_i, \Sigma_i)$ as $A^i_\alpha = \{x \mid \Pr[(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) \leq \epsilon] \leq 1-\alpha\}$. Intuitively, a greater α means a tighter impact domain. Based on the relation between $\mathcal{C}_i$ and A^j_α for any (i, j), we have the following three situations:

$$\mathcal{C}_i \subseteq A^j_\alpha \tag{3.1}$$

$$\mathcal{C}_i - \mathcal{C}_i \cap A^j_\alpha \neq \phi \tag{3.2}$$

$$\mathcal{C}_i \cap A^j_\alpha = \phi \tag{3.3}$$

The three situations are presented in Fig. 3.9 (a). Assuming the red triangles cluster represents $\mathcal{C}_j$, its impact domain A^j_α is within the dashed circle. In the case of 3.1, assuming $\mathcal{C}_i$ is

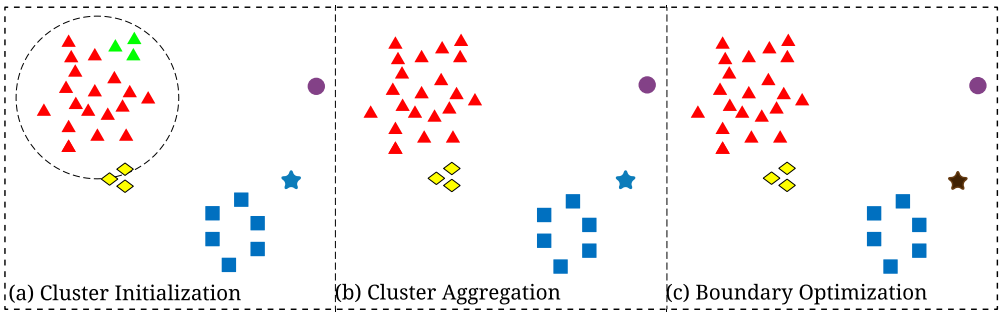

Fig. 3.9. Illustration of the three key steps of the proposed ST-Outliers detection algorithm using synthetic data. Object shapes represent the ground truth of different types of objects, and object colors indicate the clusters they belong to. In this example, step (a) identifies five different clusters; step (b) merges the green cluster into the red cluster; and step (c) separates the brown star object from the blue cluster.

the green triangles cluster, then $\mathcal{C}_i$ and $\mathcal{C}_j$ can be merged together, because with high probability, the observations in $\mathcal{C}_i$ may be from statistical model $\mathcal{N}_p(\mu_j, \Sigma_j)$. However, in the case of Eq. 3.3, assuming $\mathcal{C}_i$ is the yellow rhombuses cluster, then the probability of these two clusters are from the same statistical model is very low and cannot be merged. To deal with the remaining case of 3.2, we define a new statistic W:

$$W = \frac{|\mathcal{C}_i - \mathcal{C}_i \cap A_\alpha^j|}{|\mathcal{C}_i|} \tag{3.4}$$

If W is larger than a given threshold, then we merge the i^{th} and j^{th} clusters. In this way, The situation in (3.1) becomes a special case of the situation in (3.2).

Boundary optimization After cluster aggregation, we reestimate the statistical models based on the updated clustering results, and then optimize the boundary for each cluster. We remove an object from a cluster if the object does not follow the cluster's statistical model. For example, as shown in Fig. 3.9 (c), the blue hexagon is removed from its original cluster and becomes a stand-alone object.

To do so, we test every observation x from i^{th} cluster whether it is from population $\mathcal{N}_p(\mu_i, \Sigma_i)$.

$$H_0 : x \sim \mathcal{N}_p(\mu_i, \Sigma_i)$$

$$H_1 : x \nsim \mathcal{N}_p(\mu_i, \Sigma_i)$$

Since every observation can be treated as the estimation of the mean vector, we can use the $p-value$ of the alternate test.

$$H_0 : x = \mu_i$$

$$H_1 : x \neq \mu_i$$

Outlier Identification We repeat the procedure in cluster aggregation and boundary optimization until there is no further change. After that, the small clusters and isolated objects are considered as spatial-temporal outliers. Please note that the underlying assumption is that the

percentage of anomalous data in the whole data set is quite low. Thus, we treat as anomalies the cluster which contains less than 10% data in the empirical study section.

Algorithm 1 function STOutlierDetection $(S, K_{max}, Iter_{max})$

Input: Data S, maximum number of clusters: K_{max}, maximum iteration: $Iter_{max}$
Output: A set of outliers
$N \leftarrow$ Number of objects in S
clusterInit$(S, K_{max}, Iter_{max})$
clusterAgg$(roughClusters, \alpha)$
boundaryOptim$(aggreClusters, p)$
repeat step 3, 4 until no change

Algorithm 2 subroutine clusterInit $(S, K_{max}, Iter_{max})$

while Have empty clusters **do**
    for Each object count $i \in [1, \ldots, N]$ **do**
        randomly assign object i to a cluster k
    end for
end while
$models \leftarrow$ multivariate normal distribution set
$iter \leftarrow 0$
while $iter < Iter_{max}$ **do**
    Expectation step
    Maximization step
end whilereturn roughClusters

Algorithm 3 subroutine clusterAgg $(roughClusters, \alpha)$

sort roughClusters ascendingly by number of members
for Each cluster k in roughClusters **do**
    $centers[k] \leftarrow$ cluster center
    $radiuses[k] \leftarrow$ intra-cluster radius scaled by α
end for
$nCluster \leftarrow$ number of clusters in roughClusters
$nAggreCluster \leftarrow nCluster$ number of aggregated clusters
for k in $[1, nCluster]$ **do**
    for $k\prime$ in $[k+1, nCluster]$ **do**
        merge cluster k with cluster $k\prime$ if adjacent
        $nAggreCluster \leftarrow nAggreCluster - 1$
    end for
end forreturn aggreClusters

Algorithm 4 subroutine boundaryOptim ($aggreClusters$)

```
for i in [1, nAggreCluster] do
    for Each member x in cluster i do
        if x ~ N_p(μ_i, Σ_i) is false then
            Take x out of cluster i
            current section becomes this one
        end if
    end for
end for
```

3.3.4 Anomalous Events Detection

Usually, for domain experts, the ultimate goal of anomaly detection is not identifying the individual outliers, but to find out the underlying processes that cause those outliers. Thus, instead of raising an alarm for every single outlier, it is much more valuable to provide an ordered list of anomalous events along with their specific rankings of importance or level of interest. We accomplish this in the following three steps: feature space standardization, ST-Outliers grouping, and events ranking.

Feature Space Standardization In this step, we use z-score to compute the standardized score for each type of feature in the feature space:

$$F_{stad} = \frac{F - \mu_F}{\sigma_F}, \tag{3.5}$$

Then, we use the following criterion to categorize the standardized feature space.

$$C_F = \begin{cases} 1, \text{if } F_{stad} > F_{stad}^{thr} \\ -1, \text{if } F_{stad} < -F_{stad}^{thr} \\ 0, otherwise \end{cases} \tag{3.6}$$

where μ_F and σ_F are the mean and standard deviation of the feature, respectively. This step is needed because each feature type has a different value range. By using a z-score normalization, all features now fall within the same value range, thus allowing us to compare/group outlier objects using the top-k most significant features.

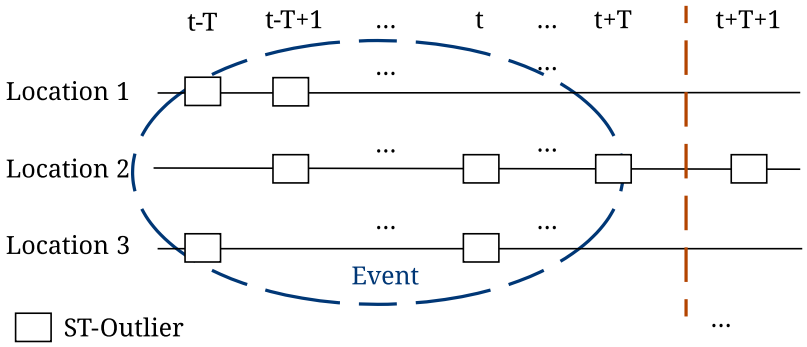

Fig. 3.10. An example of grouping ST-Outliers into anomalous events. ST-Outliers from three locations are detected at different time points and are similar in their top-k features. The outliers that occur within the same time window $[t - T, t + T]$ are then grouped together as a single event.

ST-Outliers Grouping and Events Ranking In this step, we sort the feature vector for each outlier by the absolute value of F_{stad}, then group the outliers as illustrated in Fig. 3.10, where the outliers in the same group have the same top-k categorical features. For each group, we merge every outlier and its spatial and temporal neighbors into an event until there is no further change. The intuition behind this grouping strategy is that the impact of an underlying anomalous process usually spans a continuous time period, and from the spatial perspective, more than one object is affected. Finally the events are ranked by the total number of outliers in each event and reported through the web-based UI.

3.4 Results and Discussion

In this section, we first evaluate the performance of the proposed anomaly detection framework with experiments carried out on two data sets: skin temperature derived from AVHRR data and DMSP SSM/I Daily Polar Gridded Brightness Temperatures. The details of parameter settings, and case studies of AVHRR and SSM/I data are discussed. Then, the computational efficiency of the proposed method is analyzed both theoretically and experimentally.

3.4.1 AVHRR Data

We used the South Pole AVHRR skin temperature data from July 24, 1981 to June 30, 2005, resolution is 5 km. During the process of creating the anomaly database we discovered that the AVHRR data is heavily contaminated with noise; this data set thus served as a good test case for assessing the performance of our noise pixel filtering algorithm.

Parameter Settings As mentioned in the algorithm design, it is inefficient to utilize clustering with individual pixels due to a large number of pixels in satellite images. Instead, we divide each image into objects of size $n \times n$ pixels to identify and filter out objects which could potentially contain noisy pixels. We experimented with different n values, and setting $n = 3$ achieves a good balance between accuracy and efficiency. The assumption is that within each 3×3 object the variance should be small. Once the absolute maximal difference is computed for each object, the data is transformed into a feature space where the algorithm described in Section 3.3.1 is applied.

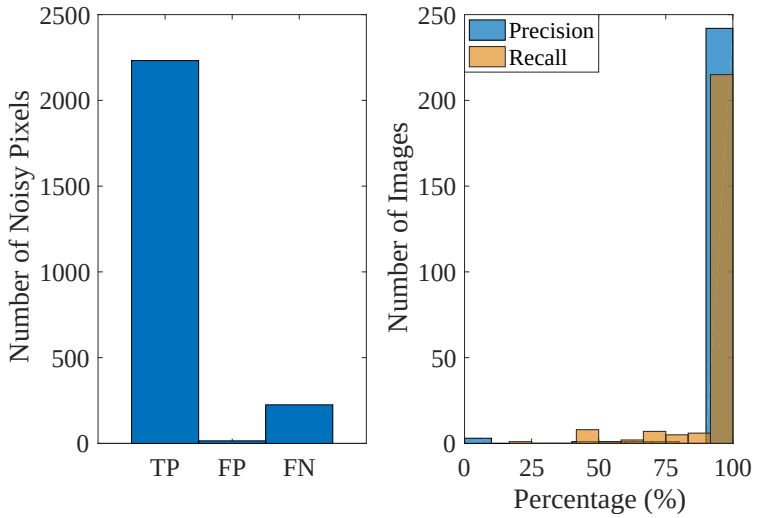

Fig. 3.11. Noisy pixel filtering in the AVHRR data set. Results show that our algorithm correctly identifies most of the noisy pixels (Left) and achieves high precision and recall for most of the images (Right).

Algorithm Performance To assess the accuracy of the noise pixel filtering algorithm, we can validate the detected pixels visually using a random sample of 10% of the AVHRR images. We then quantified the performance using two widely-used pattern recognition metrics: *precision* and

recall.

$$Precision = \frac{TP}{TP + FP} \tag{3.7}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.8}$$

True positive (TP) is the number of noisy pixels that are correctly detected as noise. False positive (FP) is the number of normal pixels that are incorrectly identified as noise by the algorithm. And false negative (FN) is the number of noisy pixels that are incorrectly classified as normal. Fig. 3.11 shows the distribution of each metric evaluated for the data set. The average *precision* and *recall* are 98.1% and 95.7%, respectively. This indicates that our noise filtering algorithm is effective, which can be used for data quality control and filtering, and can help reduce the bias introduced by such noisy pixels in the anomaly detection process.

3.4.2 SSM/I Data

SSM/I data is a primary resource for estimating sea ice concentrations and classifying sea ice types. While the data set has been continuously collected for over 30 years, from July 9, 1987 to present, it has been distributed without a thorough quality assessment. New data defects have been discovered by our framework and confirmed by specialists. Here, we use a case study from the North Pole data set to demonstrate the effectiveness of our approach.

Table 3.1: SSM/I Dataset Description

Region	Frequency (GHz)	Columns	Rows	Resolution (km)
North	85.5, 91.7	608	896	12.5
North	19.3, 22.2, 37.0	304	448	25

Parameter Settings For each image in the SSM/I data set, an object is defined as a 2×2 block of pixels, i.e., $n = 2$. Here we use a smaller block size because SSM/I data has a lower resolution than that of AVHRR data (25km vs 5km), and $n = 2$ is the minimum requirement for

removing the impact of missing or noisy pixels. With a vector size of four (four pixels in each object), we compute the spatial correlation between an object and its eight spatial neighbors. The temporal neighborhood spans two days before and after each object (i.e., $T = 2$) to help smooth out dynamic attributes such as clouds, which usually pass through the area in 1 to 2 days. The temporal neighborhood of 5 days can therefore reduce random noise without filtering out real, dynamic, or periodic fluctuations in the time series. For each object, we extract six features: the mean and standard deviation of the pixels, a spatial correlation vector and a difference vector between the object and its eight neighbors, and the mean and standard deviation between the object and its temporal neighbors (+/- 2 days). The object features are then used for the detection of ST-Outliers and anomaly events.

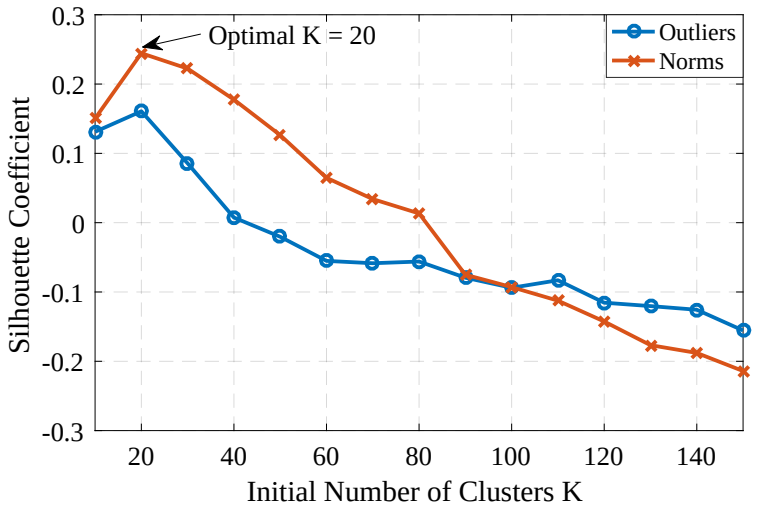

Fig. 3.12. Choosing the initial number of clusters K. Silhouette coefficients are computed for the normal and outlier clusters using varying numbers of initial clusters. The optimal number of clusters $K = 20$ is selected when Silhouette coefficients reach a maxima for both normal and outlier clusters.

As the inherent number of models of the data is uncertain, Silhouette coefficients [138] are used to evaluate the clustering's performance for differing initial conditions. Because the clustering quality is positively related to the Silhouette coefficient, the number of initial clusters is chosen where the Silhouette coefficient reaches a maximum. Fig. 3.12 shows how the Silhouette coefficient changes with the varying number of initial clusters with random samples of 10% of the SSM/I data.

Also, we used 10% as an upper bound for the total number of outliers. As with a relatively larger boundary, we maintain a high potential for including all anomalous events in the database. We aggregate outliers from the smallest size of cluster until the total number of outliers exceeds 10%. Thus, the resulted total outliers can be equal or less than 10%, which depends on the real partitions of norms and anomalies.

Table 3.2: List of Top Ranked Anomalous Events

Event Duration	Category
1990.01.01-1991.12.31	Sensor Failure
2012.07.09-2012.07.12	Natural Event
2012.07.27-2012.07.29	Natural Event
2002.06.27-2002.06.29	Natural Event
2003.10.24-2003.10.25	Natural Event
2011.02.01-2011.02.04	Unknown
2010.09.02-2010.09.04	Systematic Error

Case Studies Table 3.2 shows a partial list of top-ranked anomalous events discovered and reported by our framework. Because no ground truth exists for this data, we collaborated with other geoscientists and studied previous literature for the region to identify several of the most significant anomalous events; these events exhibited systematic error or evidence of natural events to help validate our technique. Here we discuss several of those events in detail.

Event 1: The first event was found within the 85.5 GHz channel. The 85.5 GHz vertical polarization channel exhibited a degradation in the signal from January 1, 1990, to December 31, 1991, while the horizontal channel degraded between January 1, 1991, to December 31, 1991. The origin of the event was a sensor failure. All the images collected during that period were corrupted with random noise, as shown in Fig. 3.13. This data defect could significantly affect prior analysis and computation of the region's climatology. While only part of the defect (degradation in 1991) was documented [103], our algorithm was able to uncover new errors within the 85.5 GHz channel. The issues with the 1990 data were reported to the NSIDC to help alert users.

Event 2: A sharp increase in the frequency of anomalies was discovered following 2010. The left image in Fig. 3.14 shows the spatial locations of ST-Outliers (orange squares), mainly detected by

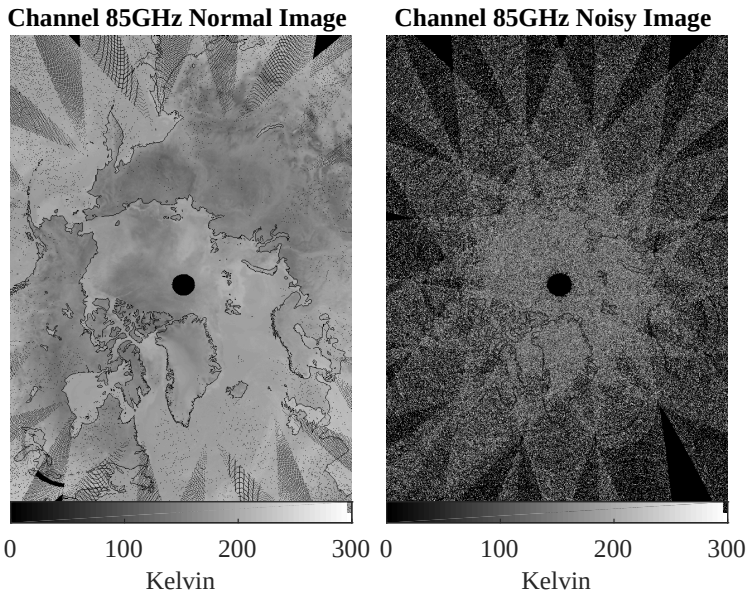

Fig. 3.13. SSM/I Data Defect: random noise within the image due to sensor failure.

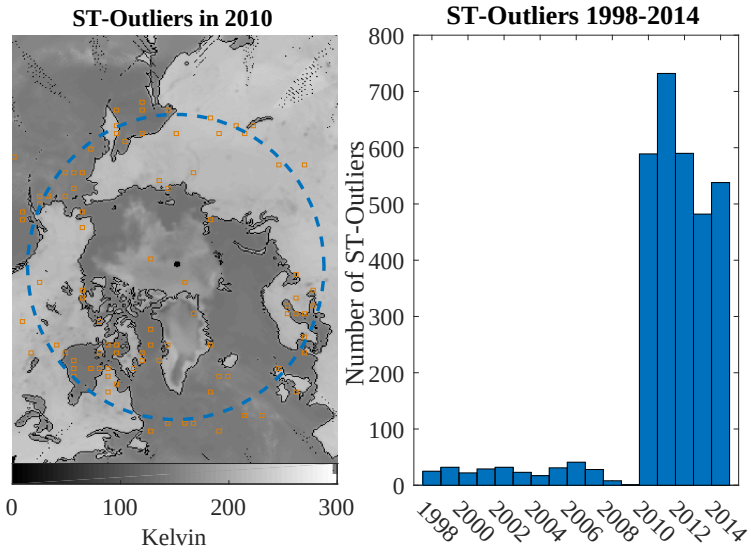

Fig. 3.14. A systematic error from 2010. (Left) The majority of ST-Outliers were detected around coastal regions. (Right) A significant shift in the number of ST-Outliers beginning in 2010.

temporal mean and standard deviation features in 2010. As seen in the figure, the majority of outliers are located around coastal areas. We determined that this surge of events was due to an inconsistency between measurements from the sensors on DMSP satellites F13 and F17 (where data from the F13 sensor was used until 2010, then transitioning to F17). The NSIDC conducted an

inter-comparison of the F13 and F17 data where products from the two sensors overlapped. Similar to our findings, larger differences, sometimes up to 10 K, were found in regions of sharp gradients of brightness temperature, usually around coastlines and sea ice extents [103]. In addition to the discovery of this systematic error, we were also able to generate a detailed report on the spatial-temporal locations of each outlier for the event. This last product could potentially accelerate the quality control process.

Event 3 and 4: Events from 2002 and 2012 are top-ranked, consistent with two extreme melt events that occurred during those years [114, 151]. As shown in Fig. 3.15 (a) and (c), the region outside of the red box regularly melts during the summer months, while in 2002 and 2012, that melting process abnormally expanded into part of the region within the red box. Our algorithm effectively detected the locations and dates of regions that normally would not melt when they

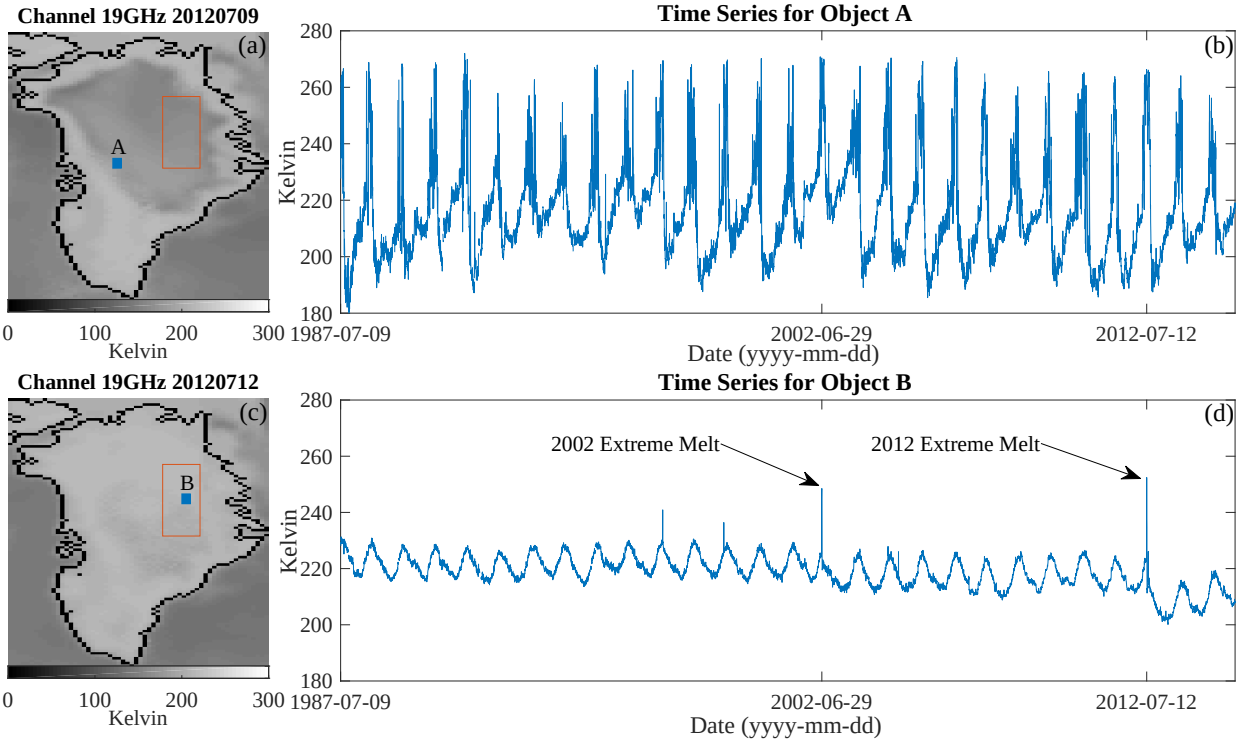

Fig. 3.15. SSM/I anomalous event: Summer extreme melt in 2002 and 2012. The red boxes in (a) and (c) represent regions which rarely melt. (b) The time series for object A, melting occurs regularly every summer. (d) The time series for object B, which was impacted by the extreme melt events in both 2002 and 2012.

exhibited abnormal behavior in 2002 and 2012. Fig. 3.15 (d) shows an example time series from one of those locations (object B) between 1987 and 2015. The brightness temperature reveals a sharp increase during the summers of 2002 and 2012. In Fig. 3.15 (c), most rare melt objects are located in the red box. Also, all rare melt locations detected were found to have averaged five melt days over the years. Thus, our framework can provide a way to explore potential interesting rare events without manually sifting through the data.

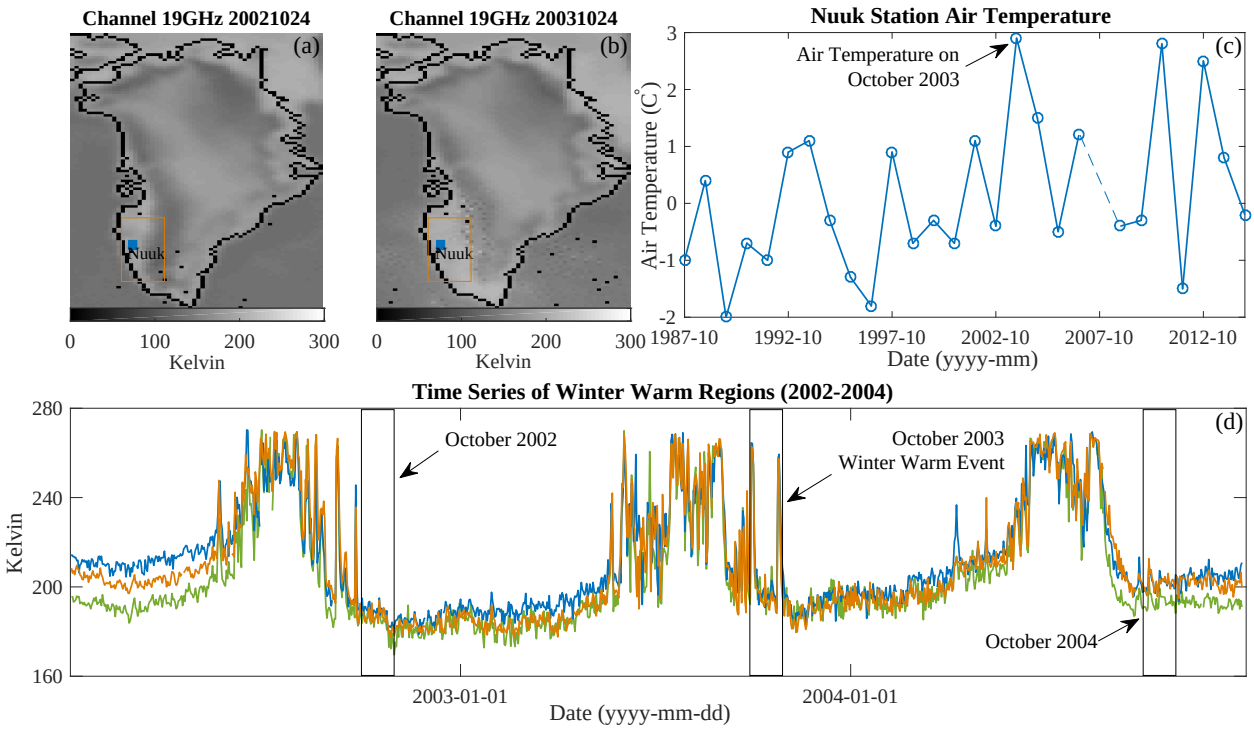

Fig. 3.16. SSM/I anomalous event: October 2003 winter warm event. (a) and (b) The region inside the box on October 24, 2003 has a higher average brightness temperature than that of the same region on October 24, 2002. (c) The warm event correlated with the air temperature at Nuuk station. (d) The brightness temperature of objects around Nuuk station show a sharp increase during October 2003.

Event 5: Besides the widely known extreme melt events in 2002 and 2012, the algorithm also detected an unusually warm event during October 2003. This event caused one location, which typically would begin to freeze during this time, to experience an extra month worth of melt days. Fig. 3.16 shows this unusually long melt event in October 2003. From the time series in this figure, spanning 2002 to 2004, the brightness temperatures of these three adjacent regions normally would

reveal a mean value which decreases during October. In 2003 though, there were sharp changes, reaching a maxima one would expect during the summer. Our algorithm accurately captured this event with the presence of seasonality and noise. The event was found to be related to the Atlantic Subpolar Gyre Warming (southwest to Nuuk at Greenland) in October 2003 [152]. Our finding is also consistent with a surface air temperature obtained from the Nuuk station where October 2003 was a record high between 1987 and 2015 (2007 data is missing), as seen in Fig. 3.16.

Other significant events detected by our framework have been shared with our collaborators for further investigation.

3.4.3 Computational Efficiency

The computational efficiency of the proposed method was analyzed with two phases: feature extraction and anomaly detection. For the feature extraction process, the computation complexity is $O(tmn)$, where t is the number of images and (m, n) are the number of columns and rows of each image. For anomaly detection, the algorithm complexity is $O(krL)$, where k is the number of clusters, r is the number of clustering iterations, and L is the length of a time series with extracted features. Table. 3.3 shows the average processing time for feature extraction on a SSM/I image, and the average processing time for anomaly detection for the longest time series (10,218 days). Since there are missing data in the SSM/I data, the time series have different length, which allows us to evaluate the real algorithm complexity for anomaly detection. Fig. 3.17 (a) shows the anomaly detection time as a function of time series length. The actual computation time is linear and consistent with theoretical complexity. Also, since there is no dependence among data files during feature extraction and anomaly detection, our method can be easily parallelized using multiple computers or CPUs. For example, Fig. 3.17 (b) shows the different anomaly detection time for a total of 24,356 time series (64 GB) with 1 to 4 CPUs.

Table 3.3: Computational Efficiency of Feature Extraction and Anomaly Detection

Feature Extraction	**Complexity**	**Image Size (m ,n)**	**Process Time (s)**	**I/O (s)**
	O(tmn)	(304, 448)	0.848	0.171
Anomaly Detection	**Complexity**	**Time Series Length**	**Process Time (s)**	**I/O (s)**
	O(krL)	10218	1.189	0.321

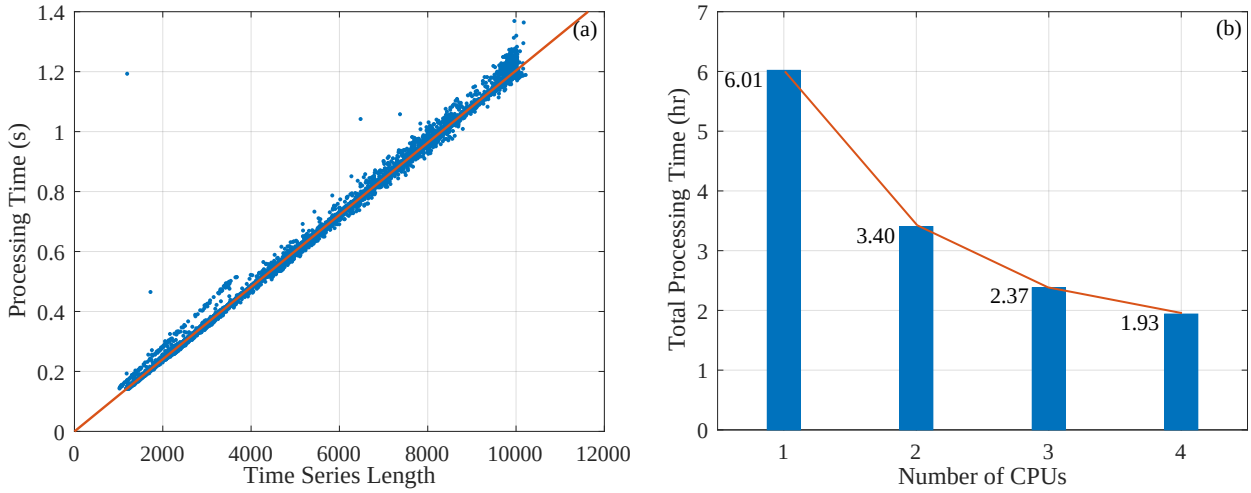

Fig. 3.17. (a) Illustration of the linear complexity of anomaly detection algorithm. (b) Demonstration of total anomaly detection time with multiple CPUs.

3.5 Chapter Summary

In this chapter, a novel unsupervised contextual anomaly detection framework is presented, which can effectively filter out noisy pixels, discover spatial-temporal outliers, and group those outliers into anomalous events. With this framework, significant data quality issues and natural events are successfully identified that were subsequently validated by geoscientists. We expect that our experience developing the framework will not only advance anomaly detection in remote sensing but also provide new approaches for speeding up scientific knowledge discovery, especially when combined with interactive data mining and visualization tools.

Chapter 4

Hierarchical Context-Aware Anomaly Detection and Multimodal Classification in Large-Scale Photovoltaic Systems

4.1 Introduction

The installation of photovoltaic (PV) systems has experienced rapid growth over the past decade [125]. Such aggressive deployment of solar farms raises serious challenges to system operation and maintenance (O&M), especially for large-scale PV systems [164]. As a large-scale PV system consists of a high volume of PV panels, sensors, high-complexity of internal architecture, anomalies can occur internally from any of the components. Furthermore, this type of system usually spans a large ground area. Due to the landscape variance, external environment can also introduce various anomalies. For instance, partial shading anomaly caused by multiple environmental factors, as well as aging due to inherent issues of the PV system itself. These anomalies, if not detected promptly, may degrade the PV system's electricity generation performance and further cause serious system hazards and failures [186].

Recent research has focused on developing anomaly detection and classification (ADC) methods to improve the reliability and safety of PV systems [186, 60]. An effective ADC solution can help capture PV system anomalies and make it possible to determine the right time for scheduling system O&M activities. Furthermore, it helps expedite PV system fault recovery and prevents further system deterioration. The recent study has demonstrated that PV system performance and reliability can be largely improved by adopting proper ADC solutions [60, 41].

The complexity of large-scale PV systems and the diversity of system anomalies are the

Table 4.1: Anomalies in PV Systems

Anomaly Type	**Anomalies**
Visual	partial shading [9](e.g., building shading, grass shading), surface soiling [15]
Thermal	hot spot [81]
Others	sensor bias, panel damaging, aging [60]

primary challenges when developing an effective ADC solution. Based on our literature survey, a wide range of anomalies occur in PV systems [15], as summarized in Table 4.1. Besides the variety of those anomalies, there are three other factors that make it challenging to design an ADC method. (1) The types of anomalies and the occurrence frequencies are affected by multiple variables, such as seasonality, PV panel location, PV system installation time, etc. For instance, one PV system in this study severely suffers from grass shading in July as weeds grow fast during summer. (2) Anomalies can be inherently related, such as a long-term partial shading can lead to hot spots. (3) Different types of anomalies require different treatments. For instance, a hot spot anomaly requires PV panel replacement, while a shading anomaly caused by cloud drift is transient and self-restoring. Existing ADC methods mostly focus on tackling specific anomaly types, hence with limited application scope [186]. Therefore, developing an effective ADC method that is capable of capturing anomalies of various types and categorizing anomalies into actionable types has remained an unconquered challenge. This is the primary focus of this chapter.

Data collection poses another challenge to ADC method design. Although supervisory control and data acquisition (SCADA) systems have been widely installed in solar farms and provide data to support PV systems' O&M, the information collected by the SCADA system is limited at certain granularity with a few variables. More specifically, as the architecture of a PV system illustrated in Fig. 4.1, a large-scale PV system can consist of thousands of PV panels, which are connected hierarchically – multiple PV panels are connected into a string, and multiple strings are connected together to a combiner box. However, only voltage and current information at PV string level, and temperature at the combiner box level are provided in existing SCADA systems. Since the voltage

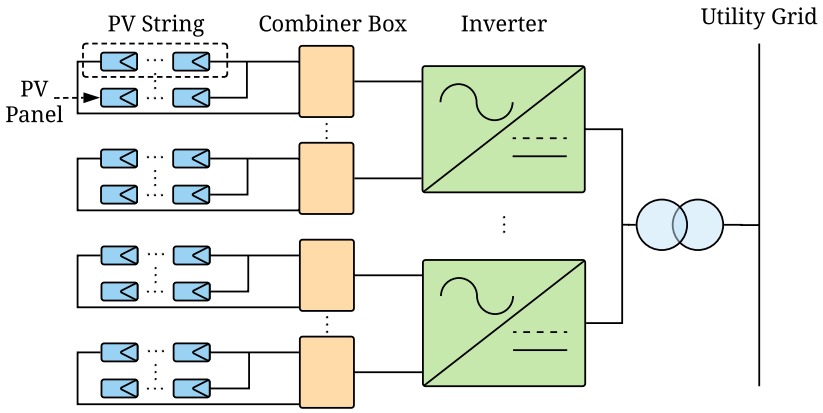

Fig. 4.1. Diagram of a grid-connected large-scale PV system.

is the same for all strings in parallel, current is the only unique parameter collected at string level. Such limited information type and granularity pose restrictions to existing ADC designs. For instance, some prior work only provides combiner box level or system level anomaly detection capability [60]. The anomaly information based on combiner box level or system level is insufficient for the utility operator to locate an anomaly. Recent work tried to perform anomaly detection at individual PV string level. Often, additional sensing and monitoring hardware installation are needed [125, 41, 174], introducing extra installation and maintenance overhead to utility operator and extra cost to the overall PV system. Compared with anomaly detection, accurate classification of the diverse types of anomalies (shown in Table 4.1) is more challenging due to the limited amount of information provided by the SCADA system.

This chapter presents a data-driven approach to perform high-accuracy PV string-level anomaly detection and classification, using information solely provided by the de facto installed SCADA system. The proposed anomaly detection method consists of two stages, namely local context-aware detection (LCAD) and global context-aware anomaly detection (GCAD). LCAD aims to identify all potential anomalous PV strings with current characteristics that are distinct from adjacent PV strings under similar environmental conditions. GCAD is designed to minimize false alarms across

the whole solar farm. Together, LCAD and GCAD can provide accurate string-level anomaly detection for solar farms. Furthermore, since it is difficult and expensive to obtain labeled anomaly data, the proposed anomaly detection method uses unsupervised machine learning techniques. The proposed anomaly classification method uses multimodal features. High-quality features are the first step towards efficient and accurate anomaly classification [87]. Therefore, the proposed method pays special attention to multimodal feature engineering. In our work, domain-specific features are identified. Then, to reduce computation complexity and improve classification performance, multimodal features are carefully designed and extracted. Next, a multimodal model training process is established, aiming to produce an accurate classification model tailored to specific classification scenarios. The contributions of this chapter are summarized as follows:

(1) This chapter presents an unsupervised learning based hierarchical context-aware anomaly detection method. Compared against existing anomaly detection methods, the proposed work improves anomaly detection accuracy by 20% (from 63% to 83%) for top-100 detected anomalies.

(2) As a byproduct of the anomaly detection, several anomaly classification methods are evaluated to provide actionable information for maintenance.

(3) The proposed anomaly detection solution has been adopted by two large-scale solar farms with DC nominal power of 39.36 MW and 21.62 MW, respectively. Multi-month operation demonstrates the effectiveness and efficiency of the proposed solution.

The rest of this chapter is organized as follows: Section 4.2 surveys the related works from the domain-specific angle. Section 4.4 presents the problem statement and method motivations. Section 4.5, Section 4.6 describes the proposed anomaly detection and classification methods, respectively. Section 4.7 presents experimental results. Finally, Section 4.9 summarizes this chapter.

4.2 Related Work in PV Fault Diagnosis

This section surveys the existing work in the area of PV fault diagnosis rather than the general anomaly detection research area, providing more background of the state-of-the-art PV fault diagnosis solutions and the motivation of the work presented in this chapter.

4.2.1 Anomaly Detection

Recent anomaly detection approaches in PV systems can be categorized into two types: model-based approaches and data-driven approaches.

4.2.1.1 Model-based Approaches

Model-based methods often require a-prior (physical) model based on domain knowledge to model specific types of anomalies [11].

Platon et al. proposed an online fault detection model to estimate the AC power production, in which solar irradiance and PV module temperature measurements are used to establish the model [125]. Garoudija et al. proposed a model-based fault detection method, in which temperatures and irradiance are used to detect faulty PV panels by predicting the healthy PV panel's maximum power [63]. Chouder et al. built a model to estimate the overall performance of PV systems by analyzing power loss, and detect faulty strings and partial shading anomaly [41]. Chen et al. used multiple online meters to monitor the voltage and power signals, which are then used for fault detection [36]. Dhimish et al. detected faulty PV module and faulty PV strings using two metrics: power ratio and voltage ratio [51]. Some of the recent works perform fault detection by analyzing the PV string electrical characteristics [29]. In these methods, extra monitoring equipment besides the de facto installed SCADA system is often required for model construction. The overall system maintenance cost thus increases.

4.2.1.2 Data-driven Approaches

Different from model-based approaches, data-driven methods mainly rely on the information provided by SCADA systems with a limited requirement of prior domain knowledge [141].

Mekki et al. used an artificial neural network (ANN) to estimate the output photovoltaic current and voltage to detect partially shaded conditions in a PV module [107]. Similar works, described in [39, 72], used ANN to detect faulty PV modules. In these methods, a large amount of labeled data are needed to train an accurate model. Yi et al. developed a method for line-to-line fault detection based on multi-resolution signal decomposition, and a two-stage support vector machine classifier is used to support decision making [176]. Other methods, such as Bayesian Neural Network [122] and decision tree [141], were also used in the past. Dhimish et al. presented an automatic fault detection and diagnosis solution using statistical methods. Their solution first uses voltage and power measurements of a PV system to evaluate the system performance. Then, they detect a fault by comparing the theoretical and measured performance [50]. Other statistical methods were also proposed in the past [185]. In summary, it is difficult and expensive to collect labeling data from real-world solar farms to build an accurate model using machine learning based methods. And statistical methods suffer from high false alarm issues since they ignore the spatially variant ambient environment in large-scale PV systems.

4.2.2 Anomaly Classification

Compared with anomaly detection in PV systems, anomaly classification is understudied [186, 41, 181, 119]. There are only a few research works that tackled the classification problem for PV systems.

Omran et al. presented an unsupervised learning based method to cluster similar segments of the output PV power [119]. Their method is built at a system level, which provides an overall performance evaluation of a PV system, but is incapable of providing the cause of an anomaly. Chouder et al. introduced an automatic supervised method to classify several types of faults in a

laboratory environment [41]. The method provided the cause of faults according to the energy loss. For instance, a string defect fault causes constant energy loss, and a shading fault causes short-term energy loss. Zhao et al. proposed a supervised learning based model to detect and classify fault types in PV arrays [181]. These fault types included line-line fault, open circuit fault, and shading fault. They later proposed a semi-supervised learning based method to classify the same types of fault while reducing the demand for labeled data [186]. The proposed anomaly classification method is different from the above methods in two aspects: (1) the proposed method has the capability of classifying anomalies into five types at PV string level based on SCADA systems; and (2) the design of multimodal features improves the classification performance, and reduces computational efficiency.

4.3 Problem Statement and Data

4.3.1 Problem Statement

Before data analysis and method motivations, an appropriate monitoring interval for large-scale PV systems needs to be determined. In general, there are two monitoring schemes: continuous monitoring and periodic monitoring [71]. Continuous monitoring is often computationally intensive and is prone to high alarm rate, which may put a huge burden on maintenance. Periodic monitoring is cost-effective but has the risk of failing to detect some anomalies which occur between successive inspections. This leads to increased safety risk in a PV system. In this work, we focus on daily anomaly detection and classification for two reasons. First, daily alarm report provides sufficient lead time to schedule maintenance activities according to the PV system requirements and hence reduce the safety risk. Second, intuitively, extra maintenance activities are not necessary if the duration of an occurred anomaly in a string is less than one day. Because these anomalies may be considered recoverable. For instance, shading anomaly caused by a drift of cloud can recover without maintenance. Therefore, we aim to address anomaly detection and classification with a daily alarm report on large-scale PV systems.

4.3.2 Data

In this part, data used for the study is described and based on the exploratory analysis, corresponding pre-processing procedures are presented, providing the necessary background of data used for algorithm development.

4.3.2.1 Data Collection

The dataset used in this work was collected from two PV systems (Site A and Site B[1]) with a nominal power of 39.36 MW and 21.62 MW respectively, located in a plain of China. The DC nominal capacity of Site A is generated by 131,184 300 W 72-cell panels connected to 553 combiner boxes. Each combiner box contains 4 to 16 strings. Site B is located in a mountain area, in which the DC nominal capacity of 21.62 MW is generated by 72,080 300 W PV panels connected to 4,240 PV strings, 294 combiner boxes. It is necessary to mention that some PV strings at site B are power limited due to various reasons[2] , and site A is not limited. The related parameters in the PV system are shown in Table 4.2. The data was collected every 1 minute by the SCADA system, from Jan 1st 2016 to Aug 15st 2016. Solar irradiance value and the string current value are used to develop the proposed anomaly diagnosis approach.

Table 4.2: Key Parameters in the PV System.

Parameters	**Symbol**	**Value**
Area of the PV module	A	1.941 m^2
Maximum Power	P_{mpp}	300 W
Maximum Power Voltage	V_{mpp}	36.50 V
Maximum Power Current	I_{mpp}	8.22 A
Open-circuit Voltage	V_{OC}	45.3 V
Short-circuit Current	I_{SC}	8.79 A

[1] Site A: Pingyuan, China. Site B: Naidong, China

[2] The specific reasons for abandoning solar power are different, the core reasons are the constraints of factors such as weak local power consumption capacity, poor power grid construction, and limited capacity of outbound power transmission channels.

4.3.2.2 Data Cleaning

String current data collected by the SCADA system is usually contaminated by errors due to malfunctions of the data collection system. These errors may include missing values, out-of-range values, misplacement values, etc. These errors should be removed before data analysis. Observations corresponding to zero current output are also removed.

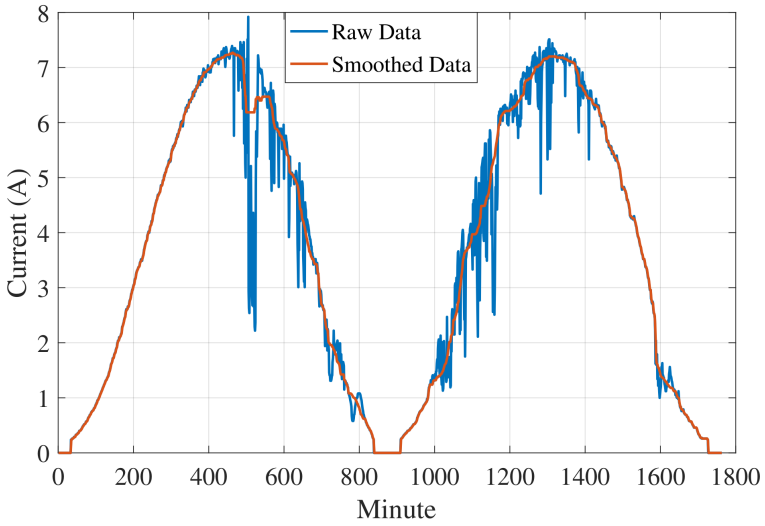

Fig. 4.2. Comparison of 1 minute sampled raw data and 60 minute smoothed data.

4.3.2.3 Data Filtering

The cleaned data still contained random noises, which should be filtered prior to data analysis. Figure 4.2 shows exemplary time series of string current signal over one day. As shown in Figure 4.2, the unfiltered signal at 1 minute interval exhibits high fluctuations. Therefore, an instantaneous current value should not be used to determine whether there is an anomaly or not. To reduce the random noises, the median filtering technique [16] is adopted here. Let $\{I(\cdot)\}$ be a discrete signal of a string current. At each instant n, an odd number of consecutive samples comprise the observation signals $\boldsymbol{I}(n)$:

$$\boldsymbol{I}(n) = [I(n - N_L), \cdots, I(n), \cdots, I(n + N_R)]^T \tag{4.1}$$

where N_L and N_R are nonnegative integers that stand for the left range and right range, respectively. The median filter f_m is applied on $\{I(\cdot)\}$ to produce smoothed current signal $\widetilde{I}$:

$$\widetilde{I}(n) = f_m[I(n - N_L), \cdots, I(n), \cdots, I(n + N_R)] \tag{4.2}$$

Since 1 hour averaging interval has been widely adopted and its accuracy has been verified in an existing fault detection study [125], we filter the data using the same interval. In most cases, the smooth window is symmetric. Thus we set $N_L = N_R = 30$ here. The red line in Figure 4.2 shows the smoothed current signals.

4.3.2.4 Data Downsampling

To reduce computation and memory cost in the proposed diagnosis approach, downsampling is adopted in this study. In prior works, 1 minute [182, 41], 5 minute [60], and 10 minute sampling interval [125] have been widely used in anomaly diagnosis in PV systems. Intuitively, a higher sampling rate would improve the model accuracy while ultimately leads to the increase of computation cost. We will show the diagnosis accuracy and computation time for each of the three widely used sampling intervals in Section 4.7.

4.4 Method Motivations

As illustrated in Fig. 4.1, a large-scale PV system consists of a massive number of PV panels, hierarchically connected into PV strings, combiner boxes, inverters, and finally the power grid through transformers [94]. Targeting such geographically distributed large-scale PV system, the proposed hierarchical context-aware anomaly detection, and multimodal classification method are motivated by the following observations.

4.4.1 Anomalous PV String Detection

First, given similar solar irradiance condition, healthy PV panels should produce similar amount of power. PV strings connected to the same combiner box are geographically close to each

other. Therefore, a malfunction PV panel/string can potentially be detected by comparing its power production against that of neighboring PV panels/strings connected to the same combiner box. On the other hand, PV strings located further away, e.g., connected to different combiner boxes, may exhibit distinct power production profiles due to spatially variant ambient environment.

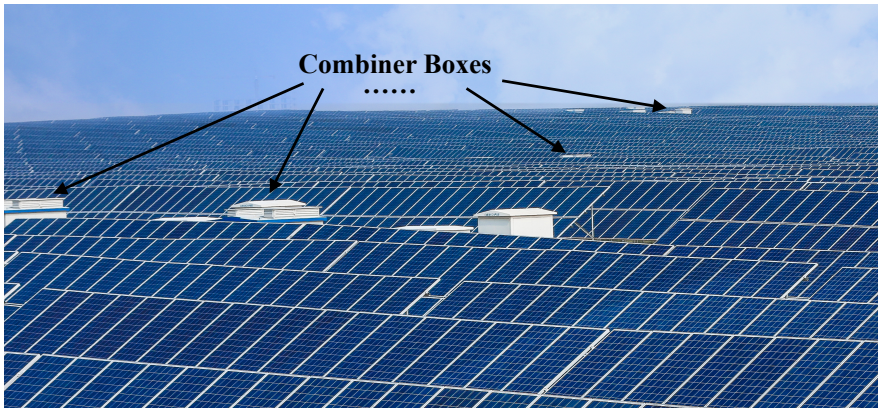

Fig. 4.3. Picture of a 39.36 MWp PV system located in China.

As shown in Fig. 4.3, in the 39.36 MWp PV system used in this study, the PV strings connected to different combiner boxes are far away from each other. As a result, a direct comparison between PV strings connected to different combiner boxes may fail to draw correct conclusions (e.g., high false negatives and false positives). As shown in Fig. 4.4, all PV strings connected to combiner box No. 1 operate properly, and one faulty PV string exists in combiner box No. 2. Using direct comparison of the power production of all the PV strings connected to the two combiner boxes, if the 3-Sigma rule is used for anomaly detection, normal strings in combiner box No. 1 will be detected as false positives, while the faulty one in combiner box No. 2 will be ignored as a false negative.

To understand this issue better, Figure 4.5 and Figure 4.6 show current time series for 16 strings in two combiner boxes. All strings operate properly in combiner box 1 (CB 1). A string that contains damaged modules and other strings that operate properly are shown in CB 2. Figure 4.13 shows two false alarm periods that appear once they are put together to conduct anomaly diagnosis. Hence, local-level analysis (i.e., combiner box level) is beneficial for detecting anomalies. (2) If only

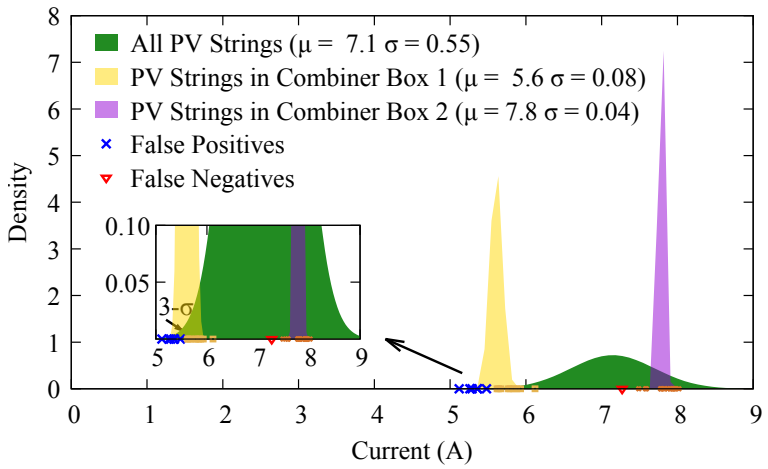

Fig. 4.4. Gaussian distributions of PV strings at the same time stamp for a 39.36 MWp PV system.

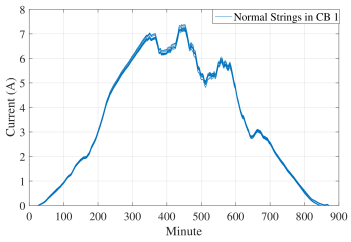

Fig. 4.5. Current variation within one day for 16 strings in combiner box 1 (CB 1).

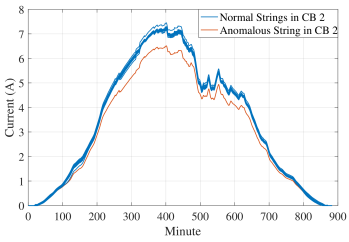

Fig. 4.6. Current variation within one day for 16 strings incombiner box 2 (CB 2).

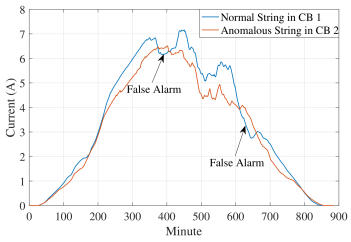

Fig. 4.7. Current variation within one day for 2 strings in different combiner boxes.

the local level is considered, the false positive rate can increase due to sensor noise or cloud drift. For instance, Figure 4.8 illustrates such a case, in which a statistical method (Hampel filter with 3σ rule [91]) is applied to diagnose anomalous strings in a combiner box. To reduce such false alarms, an approach at the global level (i.e., PV system level) is necessary.

Besides, those methods are built on a case-by-case basis that can not be utilized for general anomaly detection. For instance, the shading effect is determined by the relative angle of the sun in soft shading scenarios and the shading effect caused by surface soiling will not change with the angle of the sun in hard shading scenarios. The variation of some anomalous strings is even very similar to normal strings. Figure 4.9 shows three strings in different combiner boxes experiencing

shading, hot spot and damage anomalies. We can see that the current time series of strings that experience hot spot anomaly and damage anomaly are very similar to normal strings.

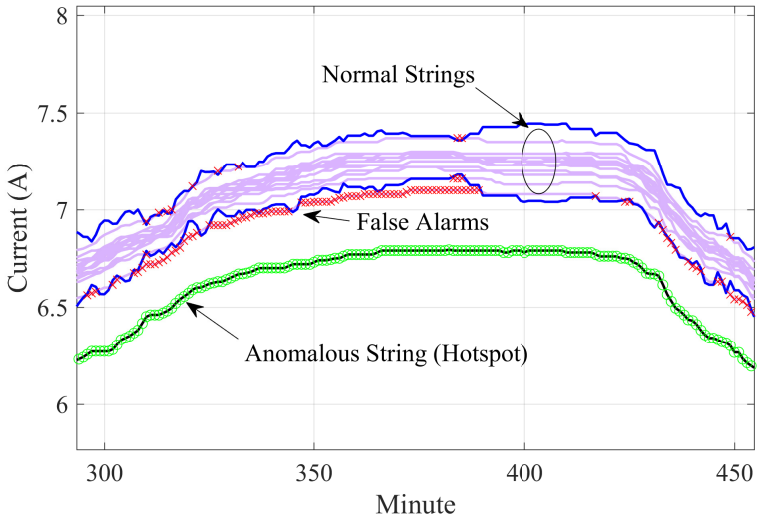

Fig. 4.8. Anomaly diagnosis using a statistical method for strings in the same combiner box.

On the other hand, the number of strings is large. Statistically, the majority of strings are expected to be fault-free most of the time, which motivate us to seek a global boundary to reduce false alarms. In summary, (1) when using statistical or other machine learning techniques to estimate the string currents model, the more strings considered, the higher the diagnosis accuracy, and (2) the heterogeneity among combiner boxes has to be taken into account, which can not only help reduce false alarms, but also make anomalous strings distinguishable from normal strings. Based on these findings, we have developed a hierarchical context-aware anomaly diagnosis approach, which is presented in the following Section 4.5.

4.4.2 Anomalous PV String Classification

The motivation of designing an anomaly classification method also comes from two perspectives: (1) To facilitate O&M, the different anomaly types' characteristics should be figured out to classify anomalies into as many types as possible. From the domain perspective, if a string is abnormal, its current value should deviate from the normal current value and for a significantly long

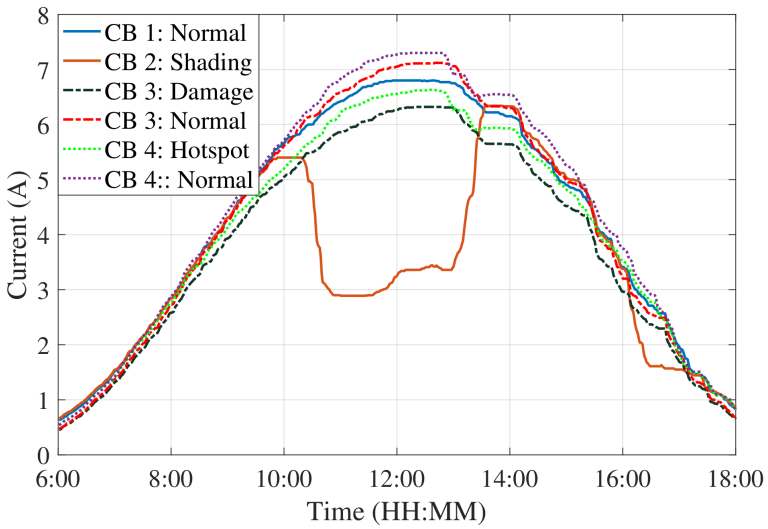

Fig. 4.9. Current variations caused by different anomalies for different strings in four combiner boxes.

period. For instance, the currents from an anomalous PV string caused by hot spots are lower than normal ones. Also, the deviations depend on anomaly types. For instance, the highest deviation can occur either in the morning or the afternoon for a building shading anomaly, which depends on both the building's position and the dynamic solar incidence angle. The long-period "deviations" can be viewed as the characteristics of different anomaly types. However, the key question is—how to use the characteristics to distinguish as many anomaly types as possible. (2) Different PV systems suffer from various anomalies thus requiring specific O&M activities, and the O&M activities may be seasonally changing, introducing complex classification scenario. For instance, a 39.36 MWp PV system in this study suffers from grass shading and hot spot anomalies during summer, while a 21.62 MWp PV system in this study suffers from sensor bias anomalies. How to design an optimal classifier for a specific classification scenario is still an unsolved problem.

4.5 Anomaly Detection Algorithm

This section details the proposed hierarchical context-aware anomaly detection method, which has two stages: LCAD and GCAD. The fundamental idea of the proposed method resides in its

ability to learn a normal operation status for all PV strings inside a combiner box in LCAD. The anomalies are then perceived as a long-period deviation from the normal operation status in GCAD. The illustration of the proposed method is shown in Fig. 4.10.

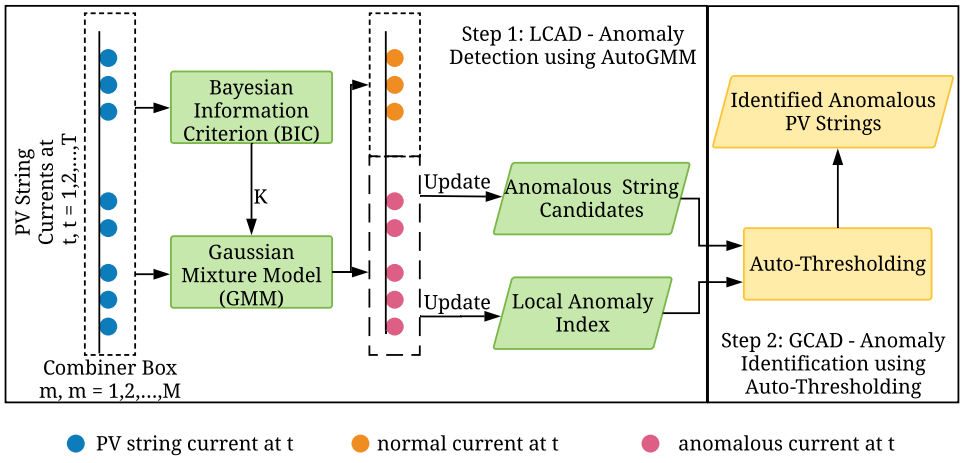

Fig. 4.10. Diagram of the anomaly detection process.

4.5.1 Local Context-Aware Anomaly Detection

As illustrated in Fig. 4.10, the goal of LCAD is to capture anomalous PV string candidates from each combiner box, leveraging the fact that PV strings in the same combiner box behave similarly except anomalous ones. To achieve this goal, an **AutoGMM** algorithm, which applies Gaussian Mixture Model (GMM) [166] to represent the behaviors of normal and anomalous PV strings at the combiner box level is proposed with the assumption that the currents measured from normal PV strings and anomalous PV strings follow different Gaussian distributions.

Let us consider a PV system composed by s sensors collecting PV strings currents and monitoring a time period n (n is the number of timestamps). A data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_s\}$ is represented a $n \times s$ matrix, in which each column vector $\mathbf{x}_i = [\mathbf{x}_{0,i}^{(j)}, \mathbf{x}_{1,i}^{(j)}, \ldots, \mathbf{x}_{n-1,i}^{(j)}]_{1\times n}^T$ denotes the current values generated from the ith PV string in the jth combiner box. At each timestamp, a mixture of $K_g^{(j)}$ Gaussian distributions $p^{(j)} = \sum_{i=1}^{K_g^{(j)}} \phi_i N(\mu_i, \sigma_i^2)$ is used to represent PV strings

distributions in the jth combiner box, where $N(\mu_i, \sigma_i^2)$ is the Gaussian component to describe the distribution of currents insider the combiner box, while μ_i and σ_i^2 are the mean and variance of the ith Gaussian component, respectively. The value of $K_g^{(j)}$ is limited by the number of PV strings inside the jth combiner box. As $K_g^{(j)}$ is an unknown parameter to be estimated, this study uses the Bayesian Information Criterion (BIC) [32] to determine the optimal value of $K_g^{(j)}$ automatically. The BIC value increases with the increasing of unexplained variations and the number of explanatory parameters in GMM, hence, the model with the lowest BIC is selected in this study. Eq. (4.3) shows the estimation of $K_g^{(j)}$.

$$\hat{K_g^{(j)}} = \arg\min_{K_g^{(j)}} m \cdot \ln(\sigma_e^2) + k \cdot \ln(m) \tag{4.3}$$

where m is the number of data samples, k is the number of free parameters to be estimated, and σ_e^2 is the model error variance. Besides, Expectation Maximization (EM) algorithm [109] is adopted to learn the parameters (e.g., means, variances) in GMM. In summary, Algorithm 5 describes the **AutoGMM** algorithm.

Algorithm 5 AutoGMM($\mathbf{x}$)

Require: $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ is a set of m PV string currents in a combiner box.

Initialize $ModelsNum \leftarrow m$
while $ModelsNum > 0$ **do**
    $clusters \leftarrow$ GMM($\mathbf{x}, ModelsNum$)
    $BIC_{ModelsNum} \leftarrow$BIC($clusters$)
    $ModelsNum \leftarrow ModelsNum - 1$
end while
$OC \leftarrow clusters$ with minimum $BIC_{ModelsNum}$
$NC \leftarrow$ the cluster with the maximal centroid in OC
$Cen \leftarrow$ the centroid of NC
return NC, Cen

The **AutoGMM** algorithm generates multiple clusters that include all PV strings currents in the same combiner box at a timestamp. Then, the cluster with the maximal centroid current is identified as the normal cluster (NC), and others are potential abnormal clusters. This is because the normal PV strings currents are higher than anomalous ones in the same combiner box at a timestamp. To quantify the anomalous level of the ith PV string, a local anomaly index (LAI) is

proposed and defined as:

$$LAI_i = \sum_{k=0}^{n-1} f(k)/n \tag{4.4}$$

where $f(k)$ is defined as:

$$f(k) = \begin{cases} 1 & \text{if } \mathbf{x}_{k,i}^{(j)} \notin NC \text{ at timestamp } k. \\ 0 & \text{otherwise.} \end{cases}$$

Here, LAI represents the percentage of time that a PV string current is considered abnormal. Theoretically, The higher the LAI is, the higher possibility the PV string is abnormal. Afterward, $LAI = \{LAI_i, \ldots, LAI_s\}$ are passed to the GCAD stage for further analysis.

4.5.2 Global Context-Aware Anomaly Detection

Due to temporal environmental conditions (e.g., cloud drift) and sensor noises, not all PV strings with positive LAIs are true anomalies. To reduce false alarms, a threshold is needed, and PV strings whose LAIs are less than this threshold will be filtered as normal PV strings. However, it is difficult to determine a proper threshold from day-to-day fault detection as the sensing conditions and external environment change over time. To address this issue, this subsection proposes a data-driven auto-thresholding algorithm.

Algorithm 6 AutoThresholding(LAI, K)

Require: A set of $LAI =< LAI_1, LAI_2, \ldots, LAI_s >$.
1: $Kclusters \leftarrow$ K-Means(K)
2: $< c'_1, c'_2, \ldots, c'_K > \leftarrow$ the ascendingly sorted centroids of the $Kclusters$
3: $thr \leftarrow 0$
4: Generating $c^* =< c^*_3, c^*_4, \ldots, c^*_K >$, with each $c^*_j \in c^*$ *and* $c^*_j = c'_j - 2c'_{j-1} + c'_{j-2}$
5: $thr \leftarrow$ the corresponding LAI value of the first peak in c^*
6: **return** thr

Algorithm 6 presents the auto-thresholding method. Firstly, K-Means clustering is used to partition all LAIs into K clusters. In Section 4.7, $K = 20$ is empirically set. Let c_i be the centroid LAI value of the ith cluster. The set of $c = \{c_1, c_2, \ldots, c_K\}$ is then sorted in ascending order into $c\prime = \{c'_1, c'_2, \ldots, c'_K\}$. Here, this study assumes that the centroid LAI values from abnormal clusters

are significantly larger than those from normal clusters. To capture this significant "divergence" from the sequence c', the second order difference (SOD) of this sequence is computed as SOD mathematically describes the rate of changes. Then, the centroid LAI corresponding to the first peak in the SOD sequence is used as the threshold thr.

4.6 Anomaly Classification

After anomaly detection, detected anomalous PV strings are further classified into multiple categories. To tackle the challenges introduced in Section 4.1, (1) multimodal features from both time and frequency domain are designed and extracted to represent the variant characteristics of different anomaly types and the invariant characteristics of the same anomaly types under the dynamic environmental conditions. (2) A classification model is produced tailored to specific classification scenarios discussed in Section 4.4.

4.6.1 Feature Extraction

As described in Section 4.4, currents of different anomalies exhibit distinct temporal, spatial, and spectral characteristics. The characteristics, originating from the long-term deviations from normal PV strings' currents, provide helpful information for classifying types of anomalies. However, as discussed in Section 4.4, the normal status of PV strings has a spatial variance, hence when deriving the deviations that characterize anomalies, spatial variance needs to be minimized. Specifically, in this study, the centroid of the normal cluster detected from the LCAD stage during the proposed anomaly detection process can be viewed as the expected current of normal PV strings. Thus, for the ith PV string in the jth combiner box, the deviation $D_i^{(j)}(k)$ as a function of discrete time k is defined in Eq. (4.5).

$$D_i^{(j)}(k) = Cen_k^{(j)} - \mathbf{x}_{k,i}^{(j)},\ k = 0, 1, \cdots, n-1 \tag{4.5}$$

where $Cen_k^{(j)}$ is the centroid of a normal cluster. It is necessary to mention that the n can be identified according to the real-time applicability. Since a daily alarm report is sufficient for the

O&M activities, a daily $D(k)$ sequence is used to describe the characteristics of an anomaly.

However, daily $D(k)$ sequence is a high-dimension feature vector, which is not effective and computational-efficient for classification. To reduce computation complexity and improve classification performance, lower-dimension feature space extracted from time and frequency domain of $D(k)$ is designed and presented in the following subsection.

4.6.1.1 Aggregation Features

Aggregation features are extracted from a temporal perspective and defined in Eq. (4.6).

$$\mathcal{F}_a = \{Mean(D(k)), Median(D(k)), Std(D(k)), Max(D(k))\}. \tag{4.6}$$

As shown in Eq. (4.6), $\mathcal{F}_a$ consists of the mean, median, standard deviation, and maximum of the $D(k)$ sequence. The aggregation features capture the unique temporal characteristics of different anomalies and invariant characteristics of the same anomalies under spatially variant ambient environment. For instance, two anomalies of the same type $D_i^{(j)}(k)$ and $D_p^{(q)}(k)$ $(j \neq q)$ may be different as the two anomalies are located under two combiner boxes. However, statistical values such as the mean, median, standard deviation, or maximum of daily $D_i^{(j)}(k)$ and $D_p^{(q)}(k)$ sequences are similar. Fig. 4.11 shows such a case. For two building shading anomalies, the highest scaled $D(k)$ can occur either in the morning (10AM for PV string No. 2) or the afternoon (2PM for PV string No. 1), which depends on both the PV strings location and the dynamic solar incidence angle.

4.6.1.2 Spectrum Features

Spectrum features represent the frequency properties of a $D(k)$ sequence. The intuition behind spectrum features is that the spectral energy of daily $D(k)$ sequences may be composed of different frequency components, depending on the anomaly types.

For example, the $D(k)$ sequence of a grass shading anomaly (PV string No. 4 in Fig. 4.12) may have fluctuations caused by environmental conditions, while daily $D(k)$ sequence of a hot spot anomaly (PV string No. 3 in Fig. 4.12) is more stable. In this study, Fast Fourier Transform (FFT)

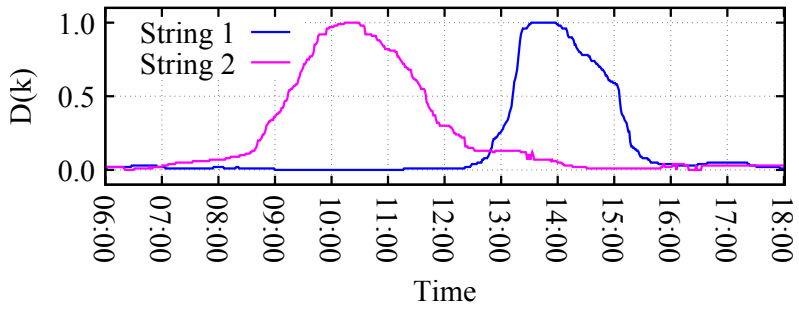

Fig. 4.11. Scaled $D(k)$ sequence examples for two building shading anomalies (string No. 1 and string No. 2).

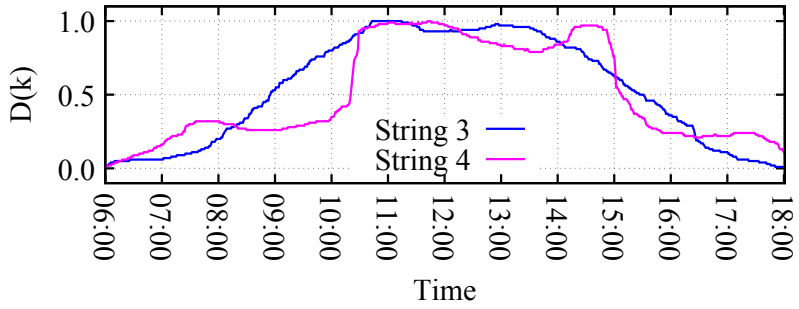

Fig. 4.12. Scaled $D(k)$ sequence examples for a hot spot anomaly (string No. 3), and a grassing shading anomaly (string No. 4).

is used to extract the spectrum features, which is defined as Eq. (4.7).

$$\mathcal{F}_s = \{g(u), u = 0, 1, \cdots, n-1\},\ \{g(u)\}_{u=0}^{n-1}. \tag{4.7}$$

$$g(u) = \sum_{k=0}^{n-1} D(k) e^{\frac{-j2\pi}{n} ku}. \tag{4.8}$$

Since the Fourier spectrum for $D(k)$ sequence is symmetric, this study only considers spectral values for $n/2$ frequencies.

4.6.2 Feature Selection

After the feature extraction, the feature dimension is reduced from n of $D(k)$ sequence to $n/2 + 4$ of extracted multimodal features. The dimension can be further reduced by selection, as the FFT spectrum of $D(k)$ sequence is dominated by a subset of frequency components. And

the remaining frequency components are of little importance for distinguishing anomaly types. To assess the importance of each feature and select the most important ones, this study first computes features importance scores using the ranking function of XGBoost [90]. Then, the features with positive importance scores are chosen.

4.6.3 Model Training

As commonly occurred anomaly types are affected by various factors, such as specific solar farms and seasonality, the best combination of features and classification model can vary. Thus, a suitable classifier given a set of pre-defined models and features needs to be identified. This study trains three classification models, including support vector machine (SVM) [44], Bagging [52], and XGBoost based on original $D(k)$ features and extracted multimodal features, respectively. The goal of the training procedure is to seek a model and the corresponding features with highest classification performance.

4.7 Experiments and Results

4.7.1 Evaluation Metrics and Experiment Setup

Anomaly Detection As there is no prior knowledge about the total number of anomalies, the top-k detection accuracy defined in Eq. 4.9 is used to quantify the effectiveness of the proposed anomaly detection method.

$$Detection\ Accuracy = \frac{k_{correct}}{k} \tag{4.9}$$

Where $k_{correct}$ represents the number of true anomalies in the top-k detected anomalies. The top-k detected anomalies are the k identified anomalous strings with the highest LAI from a daily report. For the three baseline methods used in the following experiments, the total number of alarms for each string is first counted within a day. Then the k strings with the most frequent alarms are chosen as the top-k detected anomalies.

The proposed method is compared against three state-of-the-art SCADA-based anomaly de-

tection methods for PV systems [185]: Hampel identifier, 3-Sigma rule, and Boxplot outlier rule. These methods aim to find and report anomalous strings using instantaneous currents of all strings at every timestamp (i.e., every minute). As one day is selected as the report interval for our proposed method. To make an equal comparison setup, first, preprocessed SCADA data is used for all methods. Second, daily anomaly reports from the three baseline methods are generated by counting the total number of anomaly alarms for each string within a day and sort their anomaly alarm numbers in descending order. Finally, the top-k detected anomalous strings are used to evaluate the performance of all methods.

Anomaly Classification In this study, the multilabel-based macro-averaging metric defined in Eq. (4.10) [180] is used to quantify the overall performance of the proposed classification method.

$$B_{macro}(h) = \frac{1}{L}\sum_{j=1}^{L} B(TP_j, FP_j, TN_j, FN_j) \tag{4.10}$$

where $B(TP_j, FP_j, TN_j, FN_j)$ represent binary classification metrics ($B \in \{Precision, Recall, F_1\}$), L is the number of anomaly types, in this study, $L = 5$. TP_j, FP_j, TN_j, and FN_j denote the number of *true positive*, $false$ *positive*, *true negative*, and $false$ *negative* test instances with respect to the j-th class label, respectively.

4.7.2 ADC Method Evaluation

4.7.2.1 Anomaly Detection Evaluation

Fig. 4.13 presents the detection accuracy for the top-k anomalies detected by different methods on a daily report. To show how the performance varies with different choices of k for each method, k is set to vary from 10 to 100. As shown in Fig. 4.13, our proposed method consistently outperforms the three other methods and the detection accuracies of the other methods decay more quickly than our proposed method with the increase of k. More specifically, the detection accuracy of the proposed method is higher than 83% when k is up to 100, while the accuracies of the other three methods are lower than 65% in the same condition. Also, as shown in Fig. 4.13, the data filtering

algorithm can help improve the performance of the anomaly detection methods, and the proposed method (Proposed*) outperforms the unfiltered case (Proposed).

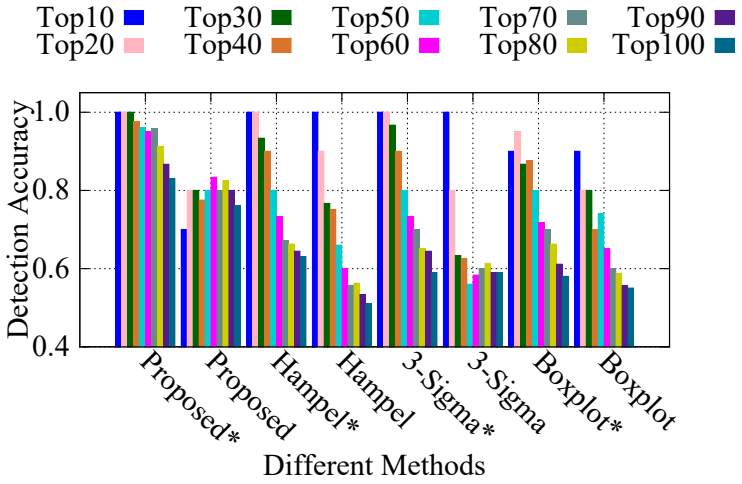

Fig. 4.13. Detection accuracies for site A data with top-k anomalous strings. (*: methods with filtered data)

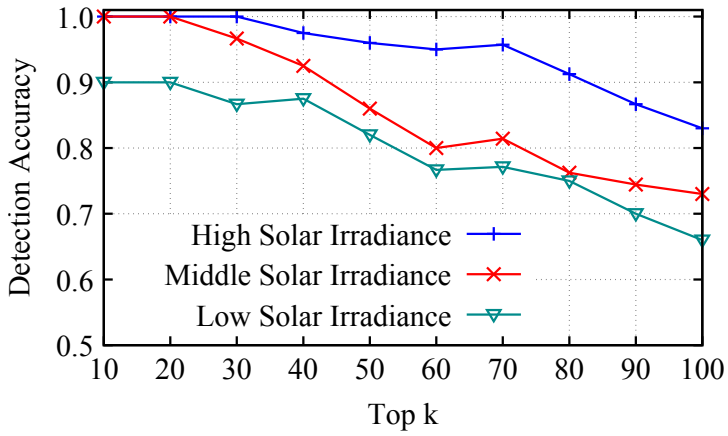

Fig. 4.14. Comparison of detection accuracies under the different solar irradiance.

Additionally, the detection performance under different weather conditions is studied. Typically, the detection accuracy is higher under higher solar irradiance conditions. This is because higher solar irradiance leads to higher current differences between normal and abnormal strings. Fig. 4.14 presents the detection accuracy of the top-k anomalies detected under different solar

irradiance conditions. Although some faults are nearly undetectable under low irradiance conditions [177], the proposed solution achieves over 90% detection accuracy for the top 20 anomalies and over 80% detection accuracy for the top 50 anomalies.

4.7.2.2 Case Study

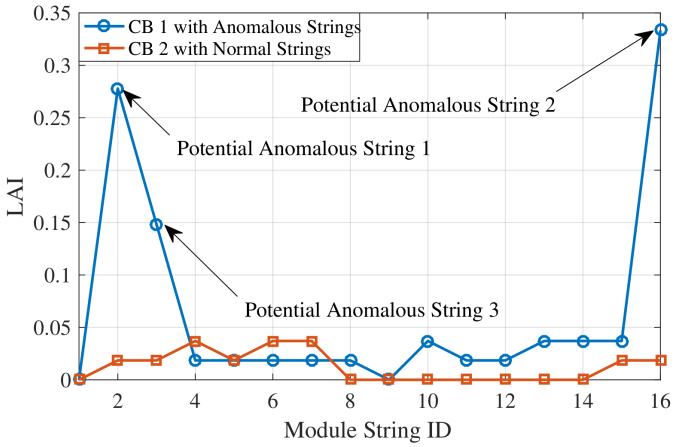

Fig. 4.15. A case study: LAIs for 32 strings in two different combiner box (CB 1 and CB 2).

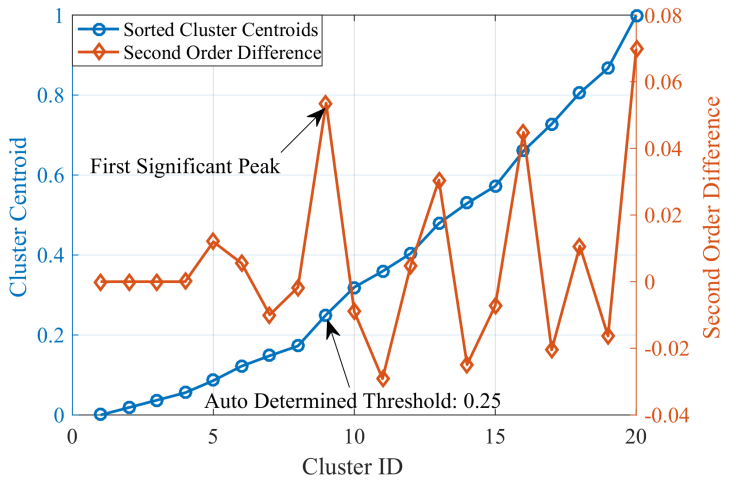

Fig. 4.16. An illustration of identifying LAI threshold automatically.

Potential Anomalies under Local Context The higher the LAI of a string, the more likely the string is anomalous. Fig. 4.15 shows how LAI values vary for 32 strings in two different

Table 4.3: Types of Anomalies Found in Two PV Systems

Property	Anomaly Source	Anomaly Type	Examples	Occurrence Frequency
unrecoverable	internal	type 1	sensor bias, aging	45.94%
	external	type 2	building shading	22.15%
	internal	type 3	hot spot, panel damaging	18.96%
recoverable	external	type 4	grass shading	12.57%
		type 5	surface soiling	0.39%

combiner boxed over a day (July 1st), in which three potential anomalous strings are identified: strings numbered 2, 3 and 16 in CB 1. Two strings are verified to be anomalies in later work (string numbered 2 and 16 in CB 1). One string numbered 3 in CB 1 is verified to be normal. As illustrated in this case, a local anomaly usually does not suggest the appearance of a real anomaly. Thus, the global context is needed to realize the anomaly diagnosis eventually.

Anomaly Identification under Global Context After LAIs of all strings are obtained, they will be clustered by the K-means method. Figure 4.16 shows the ascendingly sorted centroids from 20 clusters and the result of second order difference. The first significant peak appears at about 0.25. Thus, strings with LAIs greater than 0.25 are diagnosed as anomalies.

4.7.2.3 Anomaly Classification Evaluation

We collected 10-month operation data from the two solar farms, which is used to evaluate the proposed classification method. 1,034 anomalies were detected from the two solar farms during this period of time. These anomalies are further classified into five types, summarized in Table 4.3. Type 1 anomaly is due to sensor errors, e.g., bias, aging, and defects, affecting the accuracy of the collected data. Type 2 anomaly is due to external shading caused by ambient objects, e.g., light pole. Type 3 anomaly is due to faulty PV cells, which requires repair or replacement of PV panels. Type 4 and 5 anomaly are due to grass shading and surface soiling. Both of which can be recovered via routine maintenance, e.g., cleaning and mowing.

To evaluate the proposed classification method. The 10-month operation dataset collected from the two solar farms is randomly divided into training and test sets by fixing the ratio between

the training set and test set as 3:1. The results are averaged over 12 rounds of random training-test splits.

The proposed multimodal feature extraction process operates as follows. First, 541 features are extracted from each daily data sequence $D(k)$ (from 8 AM to 5 PM) with minute-level resolution. It then reduces the 541-dimension $D(k)$ data sequence into 274 features, among which 4 of them are aggregation features, and the rest 270 are spectrum features. Using XGBoost method, the importance of each feature is further assessed, resulting in 254 features with positive importance score. The 254 features are then fed into classifiers.

Fig. 4.17 evaluates the classification performance of the proposed feature extraction method. It first evaluates the performance of the proposed $D(k)$ feature sequence. As shown in this figure, SVM classifier achieves the best precision (92.0%) and recall (91.8%) using the proposed $D(k)$ feature sequence. Other classifiers, e.g., Bagging (BGG) and XGBoost (XGB), offer similar performance. In other words, the proposed $D(k)$ feature sequence consistently enables high-quality anomaly classification. Next, the proposed feature extraction method further reduces 541-dimension $D(k)$ feature sequence down to the final 254 multimodal features, offering 53.1% feature dimension reduction. As shown in this figure, using the reduced 254-dimension multimodal features, among the three classifiers, XBGoost offers the best classification precision (93.0%) and recall (92.8%). More importantly, it outperforms slightly against that of the 541-dimension $D(k)$ features. In other words, the proposed feature reduction method not only reduces classification computation complexity, but also maintains and slightly improves anomaly classification. Furthermore, using the 254-dimension multimodal features, other classifiers, i.e., SVM and Bagging (BGG) consistently offer similar classification performance.

The following study aims to gain further insights of the proposed feature extraction method. Fig. 4.18 and Fig. 4.19 illustrate the top-2 components of the proposed D(k) features and the final 254 multimodal features using t-distributed stochastic neighbor embedding (t-SNE) algorithm [22], respectively. It can be seen that both feature sets provide clean separation for anomalies belonged to different types. Fig. 4.18 shows the proposed D(k) features contribute more in classifying type

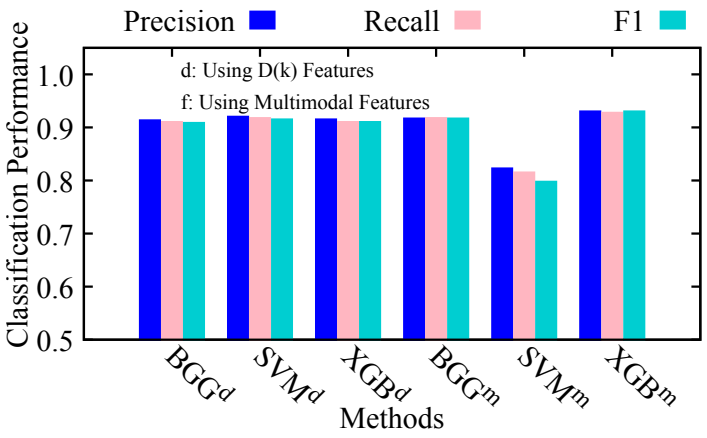

Fig. 4.17. Classification performance of different methods on features and the baseline.

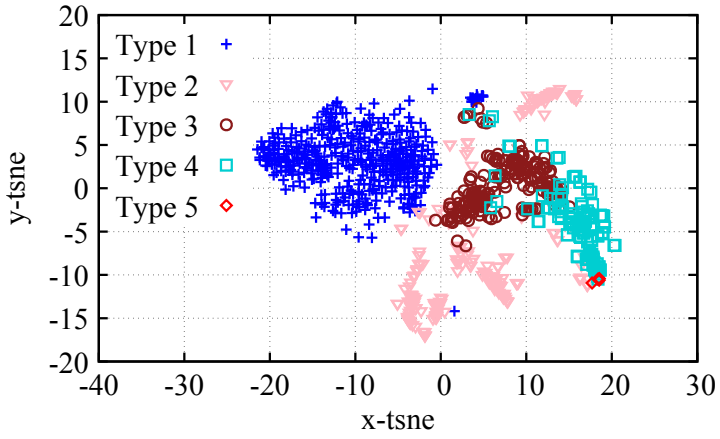

Fig. 4.18. Visualization of proposed D(k) features.

1 and type 3 anomaly. Compared against D(k) features, type 2, type 4, and type 5 anomaly are more accurately classified using the 254 multimodal features, which is shown in Fig. 4.19. Table 4.4 provides further study using confusion matrix. As can be seen, the classifier based on D(k) features misclassifies 9 testing samples of type 4 as type 3, while the classifier based on the 254 multimodal features misclassifies 4 testing samples of type 4 as type 3 and type 5.

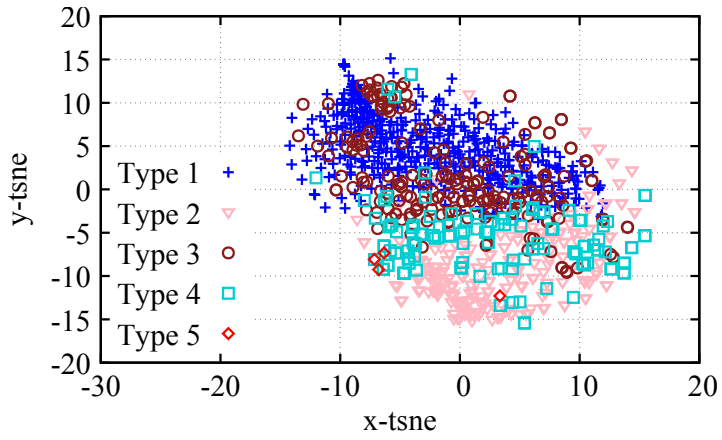

Fig. 4.19. Visualization of multimodal features.

Table 4.4: A Confusion Matrix Case for Individual Anomaly Type

		Predicted									
		D(k) Features					**Multimodal Features**				
		type 1	type 2	type 3	type 4	type 5	type 1	type 2	type 3	type 4	type 5
Actual	type 1	119	0	0	0	0	119	0	0	0	0
	type 2	7	51	0	0	0	0	52	2	0	0
	type 3	1	0	47	0	1	0	0	47	2	0
	type 4	0	0	9	24	0	0	0	3	29	1
	type 5	0	1	0	0	0	0	0	0	0	1

4.7.2.4 Efficiency Analysis of the ADC Method

Computation efficiency is critical to support daily maintenance activities. The computation time of processing the daily collected data is measured as follows. The computation time of the LCAD stage for each sampling interval (1-minute, 5-minute, and 10-minute) is 179 min, 36 min, and 15 min, respectively. When using 10-minute downsampling interval, the computation time for site B is approximate 9 minutes in the LCAD stage. The computation time of GCAD is the same for all sampling intervals, 2.2 seconds. Under different sampling intervals, the proposed anomaly detection method achieves over 83% accuracy of the top 100 anomalous PV strings. To reduce computation and memory cost, 10-minute downsampling interval is recommended in the anomaly detection method.

Fig. 4.20 shows the comparison of the computational efficiency of methods based on D(k)

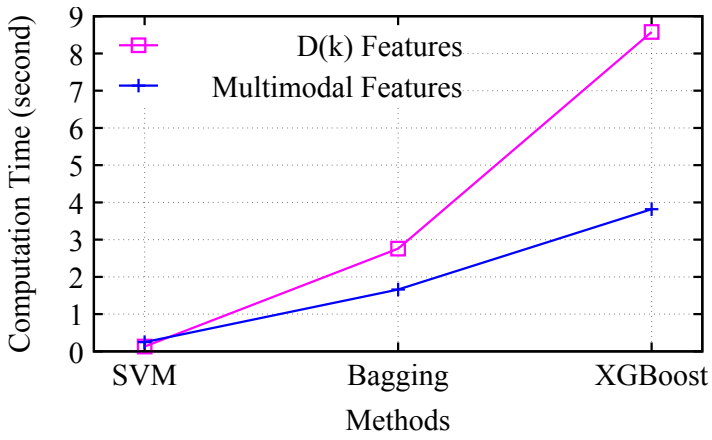

Fig. 4.20. Computational efficient of different methods on features and the baseline.

features and multimodal features in the test set. More specifically, the computation time in the test set for the proposed anomaly classification method with the best performance is less than 4.9 seconds (XGBoost method using multimodal features), which satisfies the real-time requirement of daily system O&M.

4.8 Acknowledgment

Qi Liu proposed, developed and evaluated the hierarchical context-aware anomaly detection algorithm. To further provide actionable information to maintenance activity, Qi Liu instructed Yingying Zhao to conduct classification on detected anomalies to further provide information for maintenance.

4.9 Chapter Summary

ADC plays an important role in large-scale PV systems in recent years. However, the diversity and complexity of commonly occurred anomalies in PV systems, and the limited measurements from SCADA systems pose significant challenges to prior ADC methods. To address these challenges, this study proposes a data-driven solution to accurately detect commonly occurred anomalies and further classify these anomalies into five types of large-scale PV systems via SCADA systems.

Two large-scale PV systems located in China have adopted the proposed solution. Comprehensive theoretical and experimental analysis demonstrates the efficiency and effectiveness of the proposed solution.

Chapter 5

Contex-Aware Sparse Adaptive Sensing for Running Wearables

5.1 Introduction

Running is the number one participatory sport. It is estimated that there are over 200 million regular runners in the world [1, 6]. Runners have a yearly injury rate of 50%–70% [48]. There is a consensus among physiologists that poor running form has a major impact on injury rates. Analyzing and improving running form can reduce injury rate and can also help runners to improve performance.

Sports physiologists and coaches have studied running form[1] for over a century [61]. Quantitative assessment of running form is mostly constrained to the laboratory environment. Sports physiology labs are commonly equipped with high-speed video cameras. To perform a test, markers are attached to various reference points on the runner's body. Calibration while standing is then performed. The test subject finally runs on a treadmill, while the 3D positional trajectory of each marker is determined over time [26]. This type of analysis has been limited to small-scale research studies and the support of elite athletes, due to the high equipment cost, the need for a special laboratory environment, and the lengthy setup and post-processing time. The data collected is of limited time duration and is collected in a static and controlled environment. Long-term running form effects, such as what occurs over the course of training plans lasting weeks and months, and effects due to a runner's negotiation of natural outdoor terrain and weather are not captured.

Economical microelectromechanical systems (MEMS) based inertial measurement units (IMUs),

[1] Running form refers to posture, cadence, etc.

such as accelerometers and gyroscopes, are widely used in mobile phones and can accurately sense motion, tracking the acceleration, velocity, and position of the human body. These technologies enable low-cost wearable kinematic-analysis [155, 154, 112, 147]. When paired with wireless data links, such as Bluetooth Low Energy, IMU sensor platforms enable real-time feedback to the user, allowing runners to learn from the result of form changes in-situ and on-the-fly. However, it is challenging to implement compact, accurate IMU-based kinematic analysis systems for running that both work in realtime and have long battery lifetimes.

Energy efficiency is, therefore, a foremost concern for wearables as 1) their compact form factors leave little space for large batteries, and 2) users do not prefer wearable devices needing frequent recharging. Compared with mobile phones, which are typically equipped with batteries storing thousands of mAh of energy, the batteries used in wearables only have tens of mAh to a few hundred mAh of energy capacity. Also, while people typically charge their smart phones every day, the expected battery lifetime for wearables ranges from weeks to months. For example, running foot pods now in the marketplace (primarily measuring a runner's speed and distance run) are simplistic in operation and work for one year without recharging. Users attach them to the shoe laces, and do not need to worry about them until it is time to replace the shoes themselves. The expectation of users has already been set. And device must adhere to this standard or be rejected by users. Overall, the energy budget for wearables is orders of magnitude smaller than that of mobile phones.

The energy consumption of mainstream economical MEMS IMUs sensors, although appropriate for mobile phones, is not suitable for ultra-compact wearables. Specifically, economical MEMS IMUs sensors have high active and/or idle currents. For instance, mainstream MEMS gyroscopes have active currents in the mA range, which would limit the battery lifetime of a wearable to a few days. More importantly, the power consumption of MEMS IMUs sensors is a function of sampling rate. As shown in Fig. 5.1, the active current of an accelerometer may increase by over an order of magnitude at high sampling rates. High-precision kinematic analysis potentially requires a high data sampling rate, imposing high computation and energy overheads; this is the primary barrier to wearable devices supporting high-precision running form analysis. There is a need for

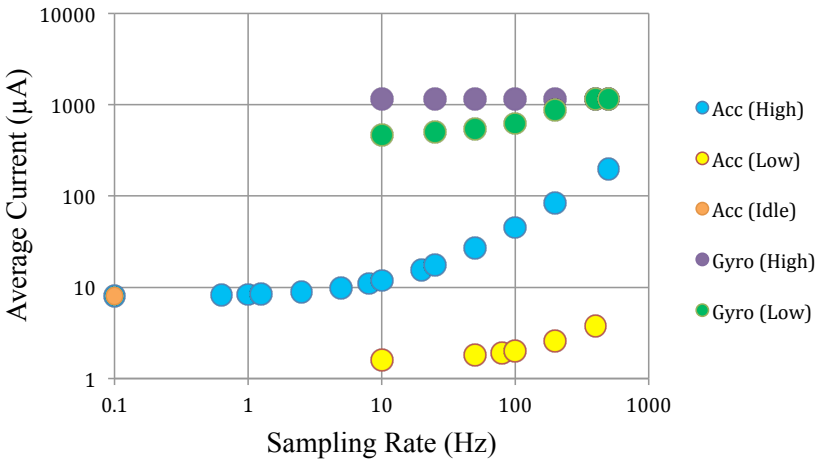

Fig. 5.1. Power consumption of MEMS IMU sensors: accelerometer, gyroscope, and low-power accelerometer currents are shown across frequency and operational mode.

energy-efficient sensing and analysis solutions to accommodate economical MEMS IMUs sensors technologies, yet providing high-precision running form analysis at runtime.

This chapter presents a sparse adaptive sensing (SAS) algorithm, which is inspired by the concept of contextual anomaly detection and related techniques. In the design of this algorithm, by analyzing the running signals, we extract a set of temporal features based on the context of running states and intra-stride signal variance. Along with the inter-running stride context, the SAS algorithm reduces the demand of samples, and manages the strengths of a low power and a high power accelerometers, greatly reduces data sensing and analysis overhead, yet maintains high running form analysis accuracy. Gyroscope is not used in the SAS algorithm due to its infeasible long startup time for intra-stride adaptive sensing. We can solve this problem by using inter-stride adaptive sensing for gyroscope, and we have achieved significant energy reduction with high running metric accuracy. However, a full discussion of this beyond the scope of this work and hence is not included in this chapter.

The proposed SAS algorithm is motivated by the fact that runners tend to maintain a consistent running form across many strides. Also, each running stride can be decomposed into several states, such as strike, toe-off, and airtime. Therefore, the sparse sensing process can be adaptive, i.e., we can vary the data sampling rate within a detected stride by detecting and predicting where

the critical points exist in time, further reducing the number of samples needed for accurate analysis. Our experimental study shows that SAS can reduce the data sensing and analysis overhead, hence the energy consumption, by 76.9% while maintaining 97.7% accuracy. This allows Gazelle to have a small form factor, with a total weight of less than 8 grams, yet offering over 200 days of use on a standard coin-cell battery.

This chapter makes the following contributions:

- The design of the **sparse adaptive sensing** (SAS) algorithm, which exploits the intra- and inter- stride context of the running signal to sample adaptively in time, thus reducing energy consumption yet still maintaining high accuracy for the Gazelle wearable system [93, 171].
- Real-world evaluation using in-lab experiments and pilot studies with runners during day-to-day training and racing, including our study of eight top professional and amateur athletes using the wearable system during the Kona Ironman World Championship race (October, 2014).

The rest of the chapter is organized as follows. Section 2 reviews prior work. Section 3 presents an overview of the running wearable system. Section 4 validates our running form analysis approach as compared with a laboratory kinematic analysis system. Section 5 describes our SAS algorithm. Section 6 presents the experimental results and pilot study results. Finally, Section 7 summarizes the work.

5.2 Related Work

Sports physiologists and coaches have long been studying running form and its impact on running performance and safety. High-speed video camera systems and floor-mounted force plates have been the de-facto equipment in sports physiology laboratories and have effectively supported running kinematic research [89, 19, 127, 135, 168, 26]. The limitations of such systems include high cost, time-consuming operation, and their use is confined to the indoor lab-testing scenario. Major sports brands have also developed pedometer-based wearable solutions to help people run

better [76, 108, 126, 116]. Gazelle offers longer battery lifetime with much more detailed and comprehensive running form analysis.

Recently, researchers have been using wearable sensing technologies to facilitate in-lab running kinematic analysis or out-of-lab studies [95, 140, 20, 173, 155, 154, 112]. Several wearable kinematic analysis prototypes have been developed using IMUs. These projects mainly used the wearable devices for data collection for offline analysis. There were few studies investigating the power consumption of an IMU-based kinematic analysis system, which showed limited battery lifetime of only a few days [112]. In the general motion or activity sensing area, there exists a lot of research on the problem of energy management [167]. There are mainly two categories of power saving methods: sensor duty-cycling and collaborative sensing with multiple sensors [21, 175, 77, 97]. For example, in the mobile sensing framework designed by Wang **et al** [167], only a minimum set of sensors were powered and appropriate sensor duty cycles were used to significantly improve device battery life. Ganti **et al** and Zhu **et al** also utilized sensor duty-cycle to minimize power consumption by detecting the active and idle state of user [62, 187]. In the E-Gesture work done by Park **et al**, the authors proposed a collaborative sensing technique that used accelerometer and gyroscope-based gesture detectors, and the gyroscope detector was only activated when a valid gesture was detected by the accelerometer detector to reduce energy consumption [121]. In our work, besides leveraging those power saving techniques, we also propose a sparse adaptive sensing algorithm with the collaboration of two accelerometers to reduce the sensor power consumption during active mode. Although our method is tuned for online running form analysis, it can also be applied to other sensing fields.

Regarding sparse or adaptive sampling algorithms at signal level, various model-based theoretical analysis has been conducted in signal processing and wireless communication [33, 78, 53, 56, 58]. These work utilized the sparsity of the signal, and the local signal time-frequency variance to minimize sampling overhead. For example, compressed sensing [33, 78, 53] does sparse, random sampling based on the sparsity of a signal in a sparse domain (e.g., frequency domain) though the signal may not be sparse in the time domain. As a result, though these work were used in wearable sensing

devices, only the sensing part can be executed on the wearable device, while the sampled data must be sent out to mobile phones or PCs with the high computing capability needed for reconstruction and analysis. The authors of [56, 58] proposed a time-domain adaptive sampling framework to predict the next sampling point based on historical sampled data and therefore reduce the power overhead for signal reconstruction. However, though running is a relatively consistent motion from stride to stride, the in-stride signal is non-deterministic, changes quickly, and varies across runners. It is therefore not practical to build a generic running signal model to predict future samples.

To the best of our knowledge, SAS algorithm is the first solution that enables low-power online running form analysis, with the consideration of the repetition and predictability of human running. The SAS algorithm works in realtime on-board out in the real world, and its performance and energy savings have been demonstrated through extensive in-lab experiments and outdoor use by real runners.

5.3 Wearable System Design

Although the primary focus of this chapter is the SAS algorithm, we first introduce the overall Gazelle system [93, 171] as the algorithm is initially motivated for it and evaluated with Gazelle. The Gazelle system was collaboratively designed and built with several other researchers in the project.

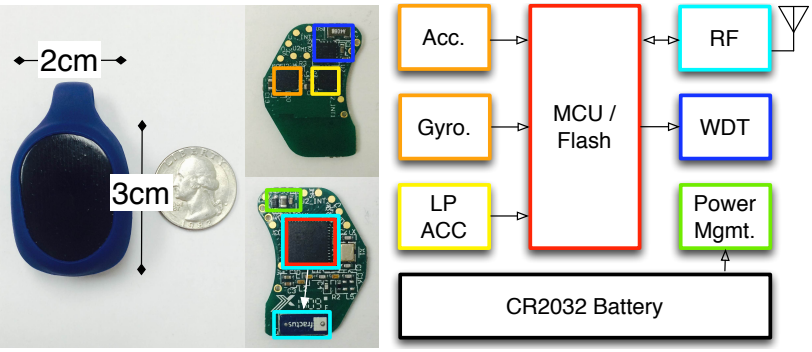

Fig. 5.2. The wearable sensor and system architecture.

The wearable system architecture is illustrated in Fig. 5.2. It consists of (1) a system-on-chip

Fig. 5.3. The example chest worn usage scenario of the mobile running analysis system.

with a 16 MHz low-power ARM Cortex-M0 and BLE/ANT+ wireless interface, (2) a 9-axis MEMS IMU suite with high-precision, high-power accelerometer (HHA), and gyroscope, (3) a standalone ultra-low-power, low-precision accelerometer (LLA), (4) an ultra-low-power watchdog timer, (5) a system power management unit, and (6) a standard CR2032 225 mAh coin-cell battery.

With a form factor of 2 cm×3 cm×1 cm and less than 8 grams of total weight, the system can be easily worn on different parts of a user's body, such as the chest, ankle, foot, or elsewhere. As shown in Table 5.1 below, depending on the specific worn body location, different running metrics can be obtained. The system's wireless interface, enables communication with a sports watch or mobile phone, which can provide voice or visual feedback as illustrated in Fig. 5.3.

5.3.1 Hardware

Processing and Communication: With form factor being a primary design driver, minimizing PCB size and power consumption is a first-order consideration in Gazelle's hardware design. The **nRF51422** is a System-on-Chip (SOC), equipped with a 32-bit ARM Cortex-M0 CPU and a 2.4 GHz ultra-low power RF front end. The RF front end supports concurrent Bluetooth Low Energy (BLE) and ANT+ protocol operation. The **nRF51422** allows onboard data processing and enables multi-platform (e.g., ANT+ Sport Watches & BLE Mobile Phones) data sharing. In

Table 5.1: Key Running Form Metrics

Metric	Definition	Chest	Hip	Foot	Ankle	Wrist
Stride Time (ST)	Duration of a stride	Y	Y	Y	Y	Y
Ground Contact Time (GCT)	Duration foot is in contact with ground	Y	Y	Y	Y	N
Vertical Oscillation (VO)	Amount of bounce up and down	Y	Y	N	N	N

addition, the **nRF51422** provides a flexible power management unit that can be used to further minimize power consumption. For example, depending on the user's usage pattern, Gazelle can switch between different states (e.g., idle or active).

Sensing: Measurement timing resolution (i.e., accuracy) and flexible sample rate control (i.e., power savings) are the two main driving factors in the design of the sensing hardware. Based on our studies of runners' walking and running signals, the maximum running acceleration can reach 16 g, which occurs when the foot strikes against the ground. We chose the **MPU9250** IMU as the main motion sensing unit because it is compact yet meets Gazelle's sensing precision requirements. The **MPU9250** includes an accelerometer and a gyroscope, supporting flexible individual sensor mode selection (e.g., standby, on/off), and quick adaption to changes in sensor sampling rate. However, one drawback of the **MPU9250** IMU is the high power consumption, e.g., 400 μA for the accelerometer in normal mode. Therefore, we added an ultra-low power, lower accuracy accelerometer whose power consumption is two orders of magnitude less than that of the **MPU9250** IMU. The **ADXL362** (3 μA at 400 Hz and 1.1 μA motion activated wake-up mode) is used to detect user status and running form changes. The information gathered from the **ADXL362** drives the configuration of the high power IMU. This control process is discussed in more detail in Section 3.2 and Section 5.4.2.

In addition to processing, sensing, and communication, 24/7 reliable operation is needed. Most of the time the system is idle in the OFF mode, and it continuously monitors the user's motion to trigger system wakeup. The **nRF51422** has an internal watchdog timer, but based on our testing, it was operational only in the higher current ON mode. Therefore, an external ultra-low power 100 nA watchdog timer, the **PCF2123**, is incorporated to ensure system health while keeping

accurate system time.

5.3.2 System Workflow

Gazelle's software is built on top of the **nRF51422**'s wireless protocol stack and SDK, taking less than 35 KB of flash memory. The software enables microsecond-resolution coordinated event-driven streaming operation, including system model checking, error handling, the operations of sensors, data processing, data storage, and wireless communication.

The Gazelle IMUs have built-in features to detect motion events, freeing the microprocessor from needing to actively read and process sensor data. For example, the ultra-low-power, lower-accuracy accelerometer **ADXL362** used in Gazelle can sample data and alert the microprocessor only when the acceleration has exceeded a predefined threshold for a predefined length of time. The microprocessor can keep track of time while in OFF mode between interrupts by reading the elapsed time of the watchdog timer. The microprocessor can dynamically change the threshold and time window in realtime. Taken together, an effective yet extremely low-power finite state machine classifier can be constructed. A simple rule-based approach can be used to classify user motion activity. To classify a walking/running pattern, the microprocessor can first configure the sensor to interrupt on a high-acceleration event, such as the impact due to a user's ground strike. Then, the microprocessor can reconfigure the sensor to look for a lower acceleration event, the toe-off, to occur after a minimum expected time duration, i.e., the time the foot spends on the ground. Appropriate time window durations and acceleration thresholds are tuned with walking/running datasets representing the majority set of walkers/runners.

When the user's running motion is detected by the system's low power classifier, the sensing hardware is reconfigured to capture running signals in high resolution. Captured running signal features are used to drive the sparse adaptive sensing (SAS) algorithm which 1) drives real-time IMU reconfiguration while running, and 2) constructs running metrics on board. Gazelle's wireless communication with either a sports watch or mobile phone is also triggered which allows the streaming of computed running form results to the user for on-the-fly feedback and post-run analysis.

The rest of the chapter will focus on the proposed SAS algorithm to enable energy-efficient, high-resolution running form sensing, and analysis.

5.4 Mobile Running Analysis

Kinematic analysis is used to quantitatively assess human locomotion. Running and walking motions are periodic. Stride by stride, force is produced by multiple muscle groups propelling the body forward and upward, while maintaining body kinematic stability. Gait can be broken down into a repetitive series of strides. A set of kinematic metrics can be measured, and then the musculoskeletal functions can be quantitatively evaluated. In this section, we demonstrate that the Gazelle system can capture such metrics for running with high accuracy when compared with traditional laboratory high-speed video camera systems and force plates. We then present the sparse adaptive sensing algorithm, by identifying those features intrinsic to running that uncover opportunities for significant reduction of energy consumption without a significant impact on accuracy.

5.4.1 Gazelle Sensor Accuracy Validation

To verify the Gazelle accelerometer accuracy is sufficient for running form analysis in the field, comparative experiments were conducted in a physiology laboratory equipped with a Vicon camera system and a treadmill instrumented with force plates. The Vicon system consists of an array of 8 high speed, high-resolution cameras placed in a ring to fully encircle the treadmill and runner under test. At multiple biometric landmarks, e.g. the ankle, knee, and chest, the runner was equipped with an infrared reflector and a Gazelle device.

In each experiment, Gazelle's high power accelerometer was sampled at 200 Hz while the Vicon cameras captured images at 200 fps and the force plate system ran at 1 kHz. Among the running metrics listed in Table 5.1, ST, GCT, and VO were each computed from raw Gazelle accelerometer data. To obtain ground truth for these metrics, data from the Vicon cameras and force plates system were processed as follows. Vertical oscillation was measured by subtracting the low to high points of the infrared reflector located on the runner's chest within each stride.

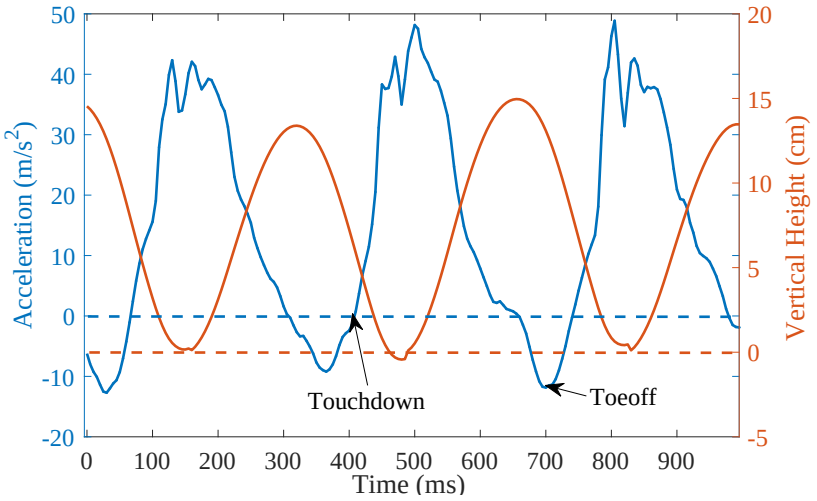

Fig. 5.4. Running stride acceleration from chest and vertical height.

Ground contact time was measured by computing the duration between foot touchdown and toe-off events. Touchdown and toe-off events were determined from force plate data by applying a threshold of 50 N for touchdown and 10 N for toe-off to the vertical force. Threshold in this range is recommended throughout the kinematic analysis literature to eliminate false detections due to force plate noise [169, 148, 110]. Stride time was obtained by subtracting step-by-step foot touchdown event. To extract those corresponding metrics from Gazelle, touchdown and toe-off events are also utilized. Fig. 5.4 shows a sample running acceleration collected from chest and the vertical height from acceleration integration. Touchdown event in the acceleration is identified by the zero-crossing right before the impact peak, and toe-off is identified as the negative minima after impact peak. Hence, ST and GCT can be computed in the same way as those obtained from force plates. VO is the difference between maximal height and minimal height, while vertical height is obtained by double-integrating the acceleration in which gravity is removed by a high pass filter.

The tests consisted of 9 different speed and cadence settings: the cross product of 5 mph, 6 mph, and 7 mph speeds with cadences of 160 spm, 175 spm, and 190 spm. Each setting was tested for 3 minutes in duration with the treadmill set for zero degrees of incline. In addition, a metronome was used during each test to assist runners to pace with the specified cadence. Gazelle was configured to stream raw data from HHA. In existing IMU-based kinematic analysis work [95, 140, 136], the

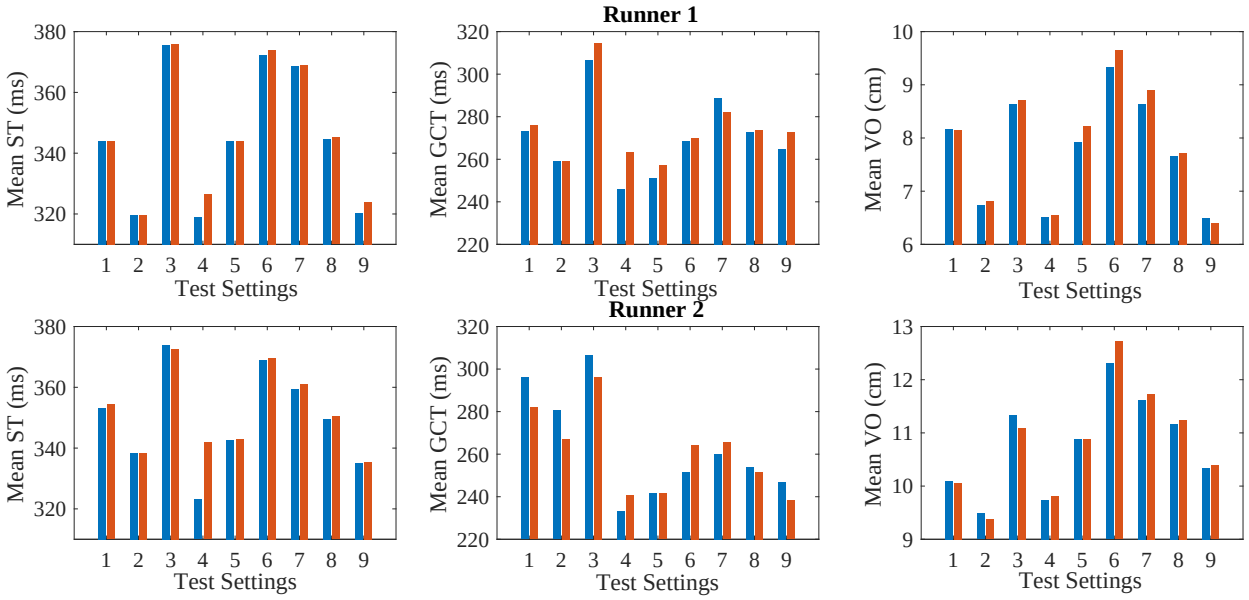

Fig. 5.5. Comparison of running form metrics captured by Gazelle and a physiology laboratory using Vicon camera and force plates system.

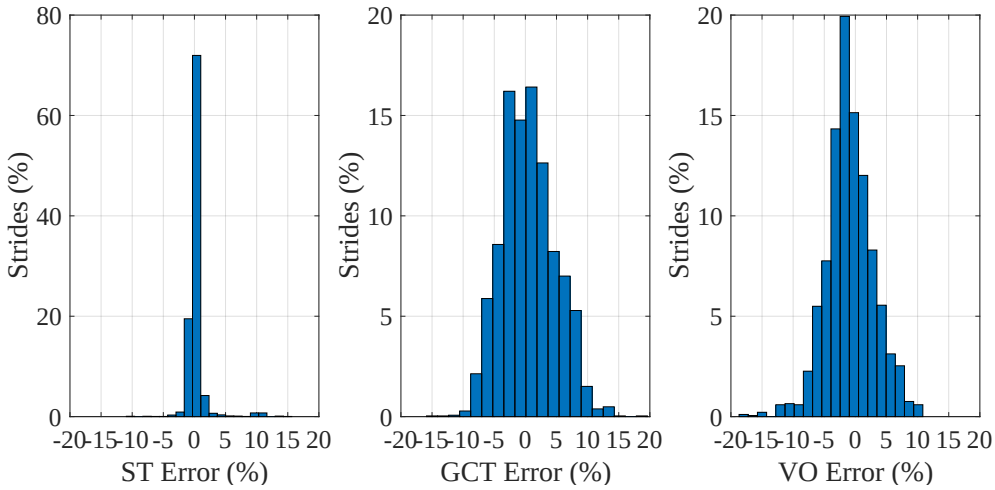

Fig. 5.6. Error distribution for ST, GCT, VO.

IMU sampling rate can vary from 100 Hz to 200 Hz, and at most 2000 Hz, depending on the degree of subtlety the running-form metric of interest. In our experiments, the HHA was configured to a 200 Hz sampling rate to sufficiently capture the running-form metrics. To compare the running metrics computed from Gazelle data to those computed from the sports physiology laboratory

camera system data, the definition of accuracy in Eqn. 5.1 was used.

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}(1 - \frac{|M_G^i - M_L^i|}{|M_L^i|})\times 100\% \tag{5.1}$$

where M_G^i and M_L^i are the running metric for each stride i computed from data measured by Gazelle and the laboratory camera system respectively. Fig. 5.5 shows representative results from two study participants and Fig. 5.6 shows the error distributions from all speed settings for each metric. This study demonstrates that when compared with the high-speed motion capture system, Gazelle offers over 99%, 98%, 97% accuracy on average for ST, VO, GCT respectively, at all nine test settings. The results from different settings illustrate that under changes of speed and cadence, Gazelle sensor has a similar stability of system accuracy as the laboratory-grade systems.

5.4.2 Opportunities for Energy Savings

Energy efficiency is of utmost importance when supporting online running analysis with wearable sensors. Having demonstrated that Gazelle can achieve high accuracy with regular sampling of acceleration at 200 Hz, we now consider techniques to further reduce the number of samples, and therefore relax the energy requirement, while maintaining high accuracy. The challenge ahead is to answer the following two-part question. **How many samples are minimally needed, and how to select the reduced sampling set?**

Stride-by-stride Variance is Low: Running form typically changes gradually over time. In real-world running, it is unnecessary to provide user feedback stride-by-stride. Instead, feedback on running metrics can be provided only when a form change is detected, or at a user-defined feedback interval. Therefore, it becomes possible to characterize the current running form by aggregating samples across many strides. Per stride, we can significantly reduce the required data sampling rate, thereby minimizing energy consumption, yet still, maintain high running form analysis accuracy. This motivates our design of **sparse sensing (SS)**, which consists of three key steps: (1) detect running form changes and group strides with similar running form together, (2) sparsely sample data within the same stride group, and (3) reconstruct a single stride from the sparse samples within

each stride group and compute the corresponding running metrics. Since the strides within each group have high similarity, the sparse samples we obtain from individual strides allow reconstruction of one representative stride for each stride group. Intuitively, there are two potential ways to get the representative stride: (1) Combine all samples to reconstruct a full stride signal and compute running metrics from it; (2) Since the results demanded by users are running metrics, metrics from selected strides in the same group can be computed and then the average for each metric can be calculated for user feedback.

Intra-stride Variance is Predictable: Given known contextual information, such as the foot touchdown, the significant event patterns within each stride are predictable in time. From Fig. 5.4 in Section 4.1, we can see that running acceleration is a periodic signal, and within one period, the signal changes sharply after the touchdown, while the change is more gradual around toe-off. Therefore, more samples are needed after touchdown, and less around toe-off, to capture sufficient information. The sampling rate can be adapted based on the variance pattern of running acceleration. Additionally, as is illustrated in Fig. 5.4, to compute ST, GCT, key points including consecutive zero-crossing points and minima are necessary to be captured. Therefore, instead of using a uniform high-frequency sampling rate, we can: (1) change the sampling rate adaptively by detecting and predicting the local variance within a single stride; and (2) based on this prediction adaptively sample only the points in time that are key to describe the selected running metrics of interest. The strategy for how to adaptively capture those key points varies based on a user's metric selection. For example, VO is computed through a double integration of the acceleration signal, presenting a more challenging scenario. Therefore, the tradeoff between lost accuracy and power savings from adaptive sampling when compared with the fully sampled acceleration signal must be identified and minimized per metric. This motivates our design of **adaptive sensing (AS)**, and when combined with SS, **sparse adaptive sensing (SAS)**, which consists of three key steps: (1) detect running form intra-variability, (2) adaptively adjust sampling rate based on the intra-variability, and (3) reconstruct a single running profile from the adaptive samples within a stride group and compute the corresponding running form metrics. Given the observations above,

we conducted theoretical analysis to understand the feasibility and potential performance of both **sparse sensing** and **adaptive sensing**, which we present in Section 5.5.

5.5 Context-Aware Sparse Adaptive Sensing (SAS)

This section describes Gazelle's sparse adaptive sensing (SAS), used to enable accurate and long-term running analysis under day-to-day real-world conditions. Firstly, we examine the theory behind SAS, then detailing the implementation of SAS. Lastly, we report our experimental results, showing that SAS maintains high accuracy and performance even when delivering an energy savings of from 76.9% to up to 99% over the continuous high-frequency sampling case.

5.5.1 Sparse Sensing (SS)

Human running acceleration signal can be represented in a sparse domain, e.g., using wavelets. Compressed sensing (CS) [33] can be used to estimate the number of samples required to reconstruct the signal. For example, we can derive the minimum number of samples required to ensure that the running metrics computed from the reconstructed running acceleration signal achieve $\geq 90\%$ accuracy compared with that computed from the 200 Hz uniformly sampled signal, as follows. Given a signal $S \in \mathbf{R}^n$, we can first decompose it using wavelets basis $\Psi = [\psi_1 \psi_2 ... \psi_n]$, as shown in Eqn. 5.2.

$$S = \sum_{i=1}^{n} c_i \psi_i \tag{5.2}$$

Assuming ΨS is k sparse, the number of samples required for reconstruction satisfies the following inequality,

$$m \geq C \cdot \mu^2(\Phi, \Psi) \cdot k \cdot \log n, \tag{5.3}$$

where C is a small positive constant and $\mu(\Phi, \Psi) = 1$. Then, $C \cdot k \cdot \log n$ samples are required for perfect signal recovery [33]. From our analysis, 5% (10 Hz on average) of the n samples need to be preserved to achieve 95% accuracy for ST, while around 25% (50 Hz on average) of the samples are needed to achieve 95% accuracy for GCT and VO. We therefore find theoretical opportunity

to reduce sampling and processing energy overheads from 75% to 95% whilst maintaining 95% accuracy.

5.5.2 Adaptive Sensing (AS)

Measuring the **intra-variability** of a running stride is an essential step in sparse adaptive sensing. Intra-variability is a measure of the local variance of a signal. In order to quantify intra-variability for use to adaptively control sensor sampling rate, we use wavelets to analyze the adaptive sampling rate required for different segments inside a stride signal. As described in Section 5.5.1, running acceleration can be decomposed into wavelets. To estimate the sampling rate, the first step is decomposing the signal S as below to get the approximate and detailed wavelets coefficients c_{low} and c_{high} [124, 102],

$$c_{low} = (S * h) \downarrow 2 \tag{5.4}$$

$$c_{high} = (S * g) \downarrow 2 \tag{5.5}$$

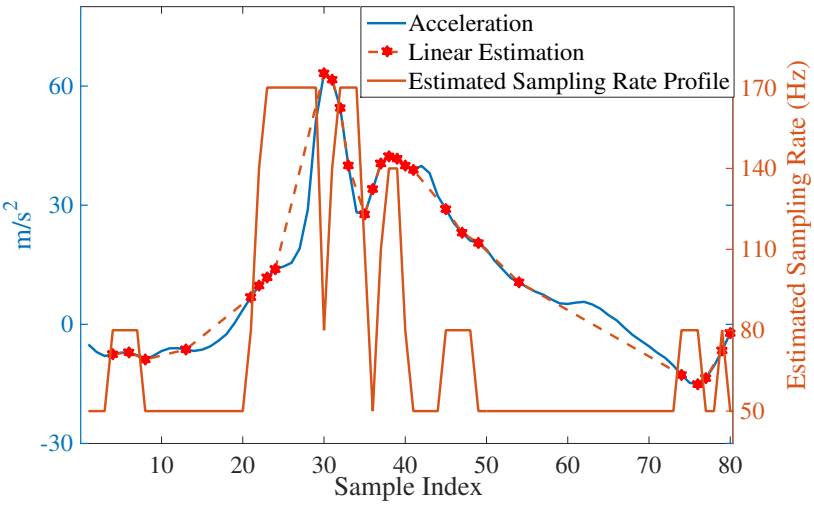

Fig. 5.7. Wavelet-based adaptive sampling rate estimation

c_{low} is then quantized in the range of 200 Hz to find adaptive sampling rates that correspond to the intra-variability of a running signal. Fig. 5.7 demonstrates a single stride acceleration, the

estimated adaptive sampling rates over time, and reconstructed signal based on linear interpolation. The sampled and reconstructed result can be seen to visually correspond to the dynamic changes across the original signal. When applied to our dataset, the wavelet-based sampling rate estimation shows that in order to achieve 90% accuracy for the running metrics computed from the reconstructed signal, on average, 80 Hz sampling rate is needed.

5.5.3 Limitations of CS and Wavelets

Our analysis from Sections 5.5.1 and 5.5.2 shows that both **sparse sensing** and **adaptive sensing** can be utilized to reduce the sampling rate yet still maintain high accuracy for running form analysis. However, CS and wavelets adaptive sensing are computationally intensive and not well adapted to the running signal.

High computational complexity: According to [53, 78], the complexity for CS reconstruction ranges from $O(M^2N^{1.5})$ to $O(\log(k)MN)$. Although the sparse sampling can be optimized to achieve only 5% CPU time for an 8 MHz wireless sensor node, the reconstruction required 30% CPU time on an **iPhone 3GS** with a 600 MHz processor [78], which is computationally intensive and not suitable for low-power CPUs. For runners who do not carry mobile phones, it is impractical to use CS on an ultra-low power 16 MHz CPU based wearable device. While the wavelets adaptive sensing reconstruction process can be as simple as performing a linear interpolation. To fit the restrictions of mobile kinematic analysis, we must further lower our reconstruction complexity.

Poor real-time adaptability: Another limitation of CS or wavelets adaptive sensing is when transforming the time domain information to a sparse domain, both cannot adaptively sample data based on running variability and the variability of a user's on-the-fly selection of running metrics of interest. For example, as demonstrated in Fig. 5.4, when only GCT is of interest to a runner, CS and wavelets adaptive sensing are not able to capture only the key points for computing GCT to achieve optimal sampling rate. Moreover, wavelets adaptive sensing requires offline processing with all signals known beforehand to build a sampling rates model, which works for efficient data storage and transmission, but is not feasible to reduce samples in real-time and hence to reduce

power consumption from sensing.

Additionally, based on the analysis in Sections 5.5.1 and 5.5.2, the required sampling rate is not low enough to achieve high energy reduction. Therefore, both methods are not well suited for realtime adaption to a real-world running signal, presenting key barriers to their use in a power-aware, low-profile wearable system.

5.5.4 SAS Algorithm Design

An alternative to overcome the limitations in Section 5.5.3 is to conduct all the analysis in the time domain and design an easily-configurable sensing algorithm which can adaptively optimize power and accuracy across the running metrics of interest. In this work, we have designed the SAS algorithm using direct time domain analysis to avoid the high computation complexity of time-frequency domain transformation and reconstruction processes, while preserving real-time adaptivity to different running metrics, thus enabling a novel and highly energy efficient long-term running form analysis on the Gazelle wearable device. Fig. 5.8 shows the overall SAS work flow. A zero-crossing (ZCR) detector and a sampling rate predictor (SRP) are used together to control HHA, and a linear interpolator is applied to reconstruct the samples from the HHA. The detailed design and implementation process are described in the following sections.

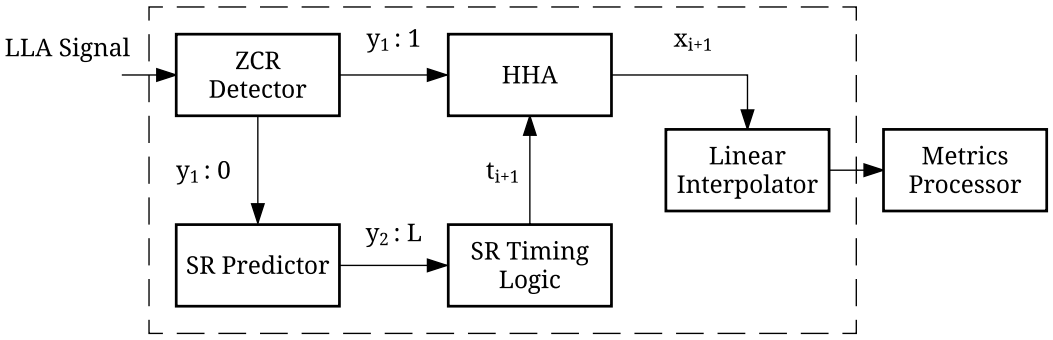

Fig. 5.8. SAS flow chart.

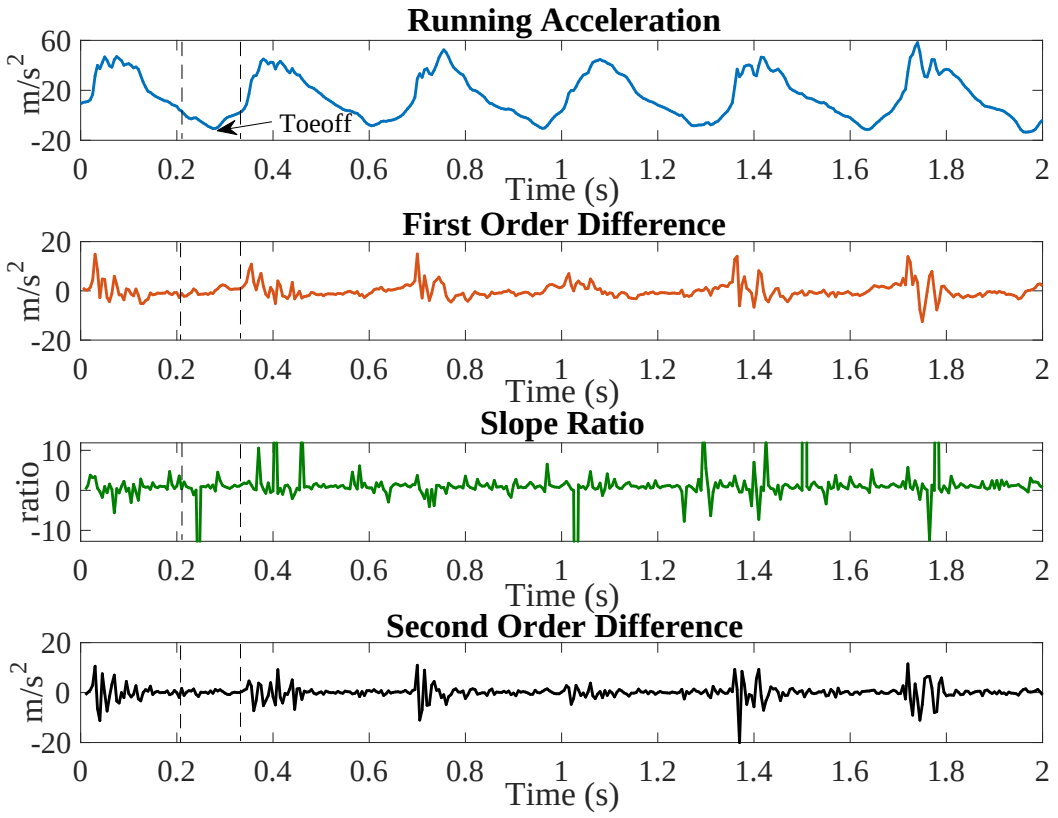

Fig. 5.9. SAS features.

5.5.4.1 SAS Design

The first question to tackle in the SAS flow is when to opportunistically acquire the next needed sample from the HHA. The largest time interval t_{i+1} between samples with the minimal loss of information is desirable. As mentioned earlier, the dependence lies in the variance pattern of the acceleration signal. The time interval can be chosen such that only the most critical points are captured for signal reconstruction. Thus we propose a method to determine an optimized t_{i+1}. First, we assume there is a finite set of intervals $\{T_1, T_2, ..., T_l\}$ to select from. Then, by constructing a projection from the predicted variance of the signal to the set of time intervals, the interval t_{i+1} can be determined. To predict which T_l should be used to acquire the next sample from the HHA, the local variance of the signal from the LLA, sampling at a higher frequency than the maximum HHA frequency, is utilized for prediction. To measure the LLA variance, three features are examined:

(1) **first-order difference (FOD)**, (2) **slope ratio (SRO)**, and (3) **second-order difference (SOD)**. FOD measures the sharpness of positive or negative slopes, SRO captures inflection points including local minima and maxima, and SOD estimates the slope rate of change. The FOD, SRO, and SOD features are computed as follows.

$$FOD = x_i - x_j \tag{5.6}$$

$$SRO = \frac{(x_i - x_j)/(i - j)}{(x_j - x_k)/(j - k)} \tag{5.7}$$

$$SOD = FOD(i) - FOD(i - 1) \tag{5.8}$$

Fig. 5.9 shows all three features along with running acceleration. FOD and SOD are sensitive to LLA acceleration when the foot is in contact with the ground, where most acceleration variance occurs. Additionally, we compared the standard deviation of FOD and SOD for the segment in each stride (between the two vertical dashed lines in Fig. 5.9) around toe-off events. Compared with SOD, FOD has higher standard deviation and hence more sensitive to toe-off events. Because FOD has less computation overhead and can cover those minima, maxima points that are primarily covered by SRO, FOD is preferred for driving the SR Predictor. However, signal variance around zero-crossings is not significant enough for FOD alone to predict critical samples; the zero-crossing points are often missed. Thus the ZCR Detector is added to augment the prediction. Combining the ZCR Detector and SR Predictor, high accuracy for all running metrics can be achieved.

Next, a set of proper sampling intervals, which can be considered as the pseudo sampling frequencies, is determined for the HHA. Here we refer to the multiplicative inverse of sampling intervals as pseudo sampling rates. This is because, in practice, an accelerometer sensor may not support the actual sampling rate needed. The one-shot operation is therefore utilized to attain the requisite pseudo sampling rate. A similar approach is used in the work of Feizi et al. [57], where the authors proposed the TANS with finite sample rate (TFR) method. In their work, an offline electrocardiograph (ECG) signal was divided into three repeating states, whereby each state was

strictly assigned a minimally needed sampling rate. TFR requires, for each state, a known signal starting point and an approximate number of samples for each state. Although running acceleration and ECG are both periodic, running acceleration has higher variance from stride to stride when compared with a beat to beat variance in ECG. For example, higher sampling rate may be required when a runner runs on a hard ground during ground contact time, while a lower sampling rate may be required when running on grass. Assigning a fixed sampling rate to a fixed segment within a stride of running acceleration, as done in TFR, limits the lowest sampling rate that can be achieved and not well adapts to the stride by stride running signal. Numerically, there are infinite combinations of possible HHA pseudo sampling rates. However, based on the target running signal, there are other further constraints: (1) The minimal sampling rate needs to ensure at least one sample can be obtained within a stride, and (2) the maximal pseudo sampling rate cannot exceed the sensor's maximal sampling rate with the consideration of the HHA sensor's measured startup delay. With those constraints in the design process, we further propose an empirical design criteria for the SR Predictor: We must minimize the number of sampling rates based on the patterns of the SR Predictor. For example, the FOD feature shown in Fig. 5.9 has the following clear patterns: (1) flat signal appearance and (2) dynamic signal changes with high amplitude. Therefore in our experiments in Section 5.6, two different boundary sampling rates are used. With this criteria and constraints identified, a set of pseudo sampling rates can be determined using the training data. The resulting average pseudo sampling rate, therefore, must satisfy the following equation:

$$\bar{sr} \leq \frac{(N_{T_m} \cup N_{zcr} \cup N_{T_t})}{\sum_{i=1}^{N} ST_i} \tag{5.9}$$

where N_{T_m} is the number of samples obtained with minimal interval T_m in the set $\{T_1, T_2, \ldots, T_l\}$, and N_{zcr} is the number of zero-crossing points. And, N_{T_t} is the number of transitions between any two different consecutive intervals. This augments to the SR Predictor design is based on the assumption that when an interval transition occurs, the samples close to this transition are important for describing the signal.

Fig. 5.10 demonstrates the reconstructed signals from the SAS algorithm as compared with

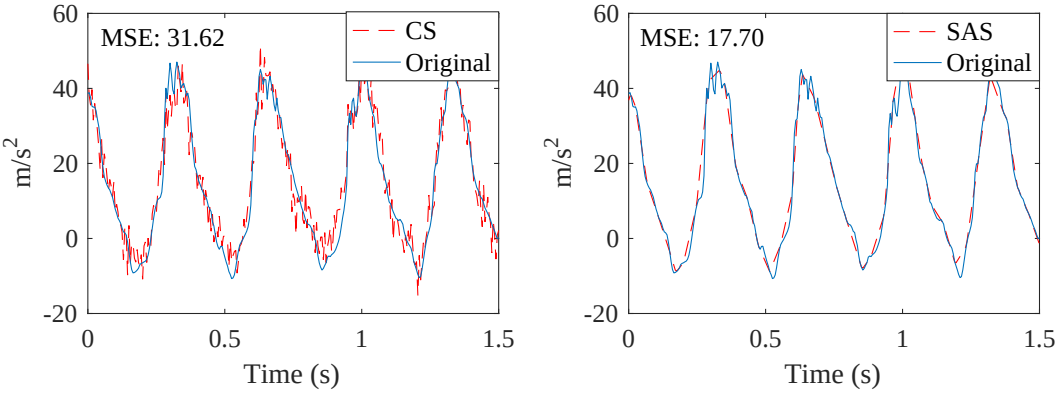

Fig. 5.10. Reconstructed signals from CS and SAS.

the compressed sensing method. The original 200 Hz signal was reduced to an average of 30 Hz for both algorithms. As can be observed in the figure, SAS outperforms CS with a lower mean squared error of 17.70. While CS can recover the overall shape and periodicity of the original signal, it does so with much lower signal to noise ratio. In Section 5.6.1, further comparison between SAS and CS are conducted.

5.5.4.2 SAS Implementation

As described in Section 5.3, Gazelle is equipped with a low-accuracy, ultra-low-power accelerometer (LLA) and a high-accuracy, high-power accelerometer (HHA). The LLA samples continually throughout a run. Even though the LLA suffers from high noise, it offers sufficient accuracy to continually detect the stride-by-stride timing structure and to estimate the similarity of running strides with low latency. Also, even though the LLA sensor cannot provide absolute accuracy for its acceleration measurement, velocity, or position-related metrics, it offers sufficient relative accuracy to detect changes of these metrics, and thus the change of running form.

The LLA consumes 3 μA and samples data at 400 Hz, to detect zero-crossings and estimate sampling rate beforehand, which are used to notify the host processor of such events. Although past work has shown lower sampling rates can be sufficient for accurate kinematic analysis, sampling the LLA at lower frequencies (1) negligibly improves battery life (e.g., 1.8 μA sampling at 100 Hz saves

Algorithm 7 SAS Algorithm

```
levels
maxSr
for newSample from LLA do
    preSr ← newSr
    if zero-crossing detected then
        Get a sample from HHA
    else
        get recent three |fods|
        fodMax ← max(|fods|)
        if fodMax > preMax then
            preMax ← (λ) × fodMax + (1 − λ) × preMax
        end if
    end if
    newSr ← (fodMax/preMax) × maxSr
    look up closest sampling rate in levels
    if preSr ≠ newSr then
        if |lastHHA − curLLA| > thr then
            Get a sample from HHA
        end if
    else
        Sample with newSr
    end if
end for
```

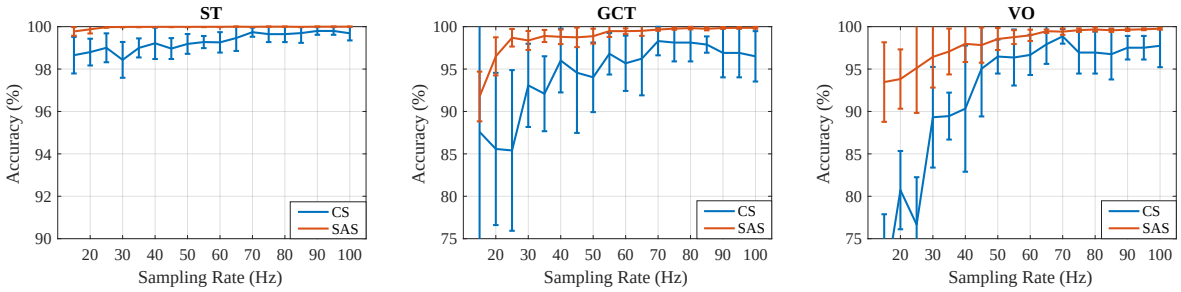

Fig. 5.11. Running metrics accuracy comparison between CS and SAS.

0.5% of **CR2032** capacity per 1000 hours of activity), and (2) reduces system accuracy by increasing the latency to trigger the sampling of the HHA. In order to ensure the HHA's sampled data is able to catch the acceleration feature detected by the LLA, the delay from both the LLA trigger and the startup time from the HHA must be lower than the sampled signal's bandwidth in Hz. From past work, a commonly used acceleration signal sampling frequency for low-power kinematic analysis is 100 Hz. Therefore, a delay from acceleration feature to LLA trigger to HHA sampling of 10 ms or

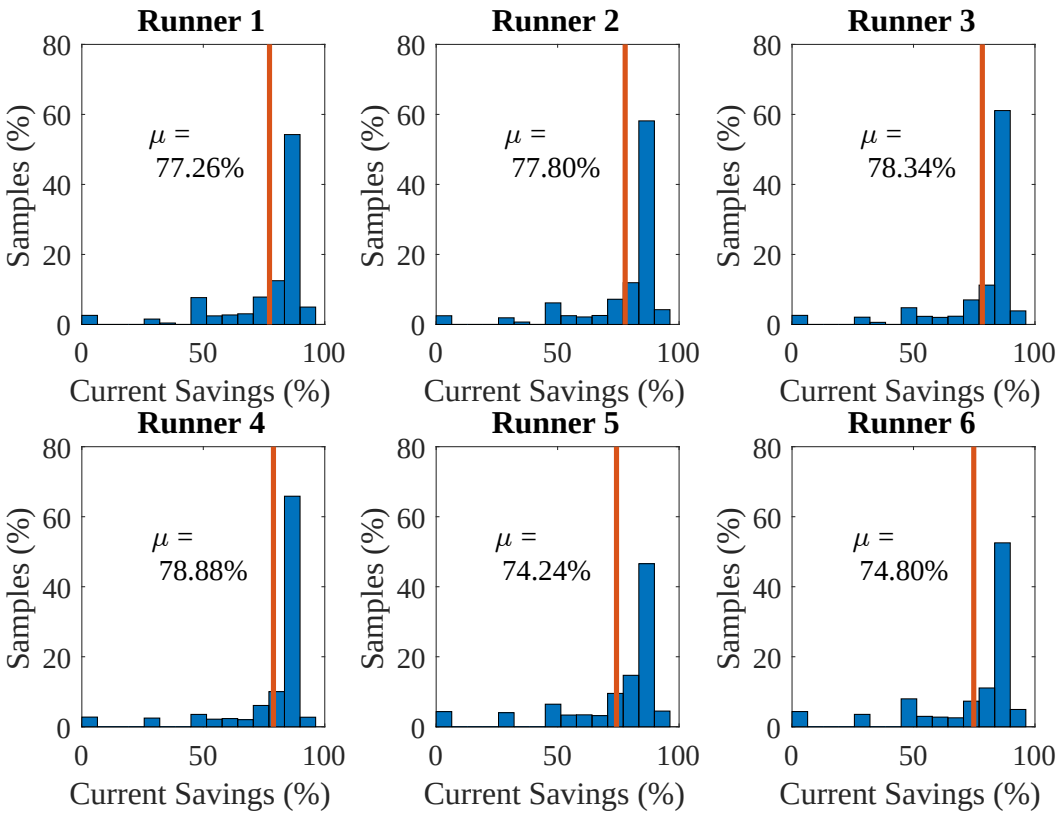

Fig. 5.12. Distributions of sample-by-sample current savings of adaptive SAS LLA + HHA sampling compared to constant 200 Hz HHA sampling, across 30 minute running sessions from six runners.

less will lose minimal fidelity. Due to this constraint, utilization of angular velocity data sensed by a gyroscope is not usable in the adaptive SAS algorithm as typically gyroscopes require 20 ms to 80 ms for start up. One solution to utilize a gyroscope with reduced power is by applying a constant duty-cycle. Due to space constraints, we do not consider such use of the gyroscope in this work. Operating the LLA at 400 Hz yields a 2.5 ms sampling delay, leaving up to 7.5 ms for HHA start up time to observe the 10 ms boundary. Therefore, we must find an accelerometer which can satisfy the start-up time constraint while maintaining high accuracy. While data from the **MPU9250** was used for the HHA during our algorithm design, the high precision accelerometer from the **MPU9250** has a maximal 25 ms startup time from sleep mode to active mode, which would then violate this constraint under a real-time implementation. Therefore the HHA used in the pilot study, which provides "first sample correct" and "zero-delay" capabilities, is the **LSM6DS3** [153]. The LSM6DS3

was measured to have a 2.38 ms delay from the start of SPI configuration commands while in power down mode to the first activated data ready interrupt signal, thereby meeting the overall real-time 10 ms constraint for signal feature to HHA sampling time delay interval.

The samples obtained by the LLA are then used to detect zero-crossings and predict pseudo sampling rates for this HHA used in our study. To achieve the adaptive selection of pseudo sampling rates, the most recent three consecutive absolute values of FOD are computed, and the maximal absolute FOD value is scaled by a global maximal absolute FOD. The scaled value is then used for looking up a proper pseudo sampling rate or time interval, as described in Algorithm 7. The HHA is then brought out of the power down mode and configured for operation at 400 Hz, and the first available sample is then acquired from HHA, achieving the selected pseudo sampling rate. The pseudo sampling rate is again updated when the absolute difference between the last HHA sample and current LLA sample exceeds a threshold. This threshold is optional and only used when lower average sampling rate is necessary. Setting the threshold to a low value can ensure key points are captured while reducing redundant points. For example, in Section 5.6.1, the threshold is set to $1.8\,m/s^2$. Additionally, a low pass filter can be applied to the global maximal absolute FOD to smoothly adapt to local changes in acceleration. Algorithm 7 summarizes the full SAS procedure.

Using the samples captured by our SAS algorithm, reconstruction methods can be applied to recover the running profile to compute all the running form metrics. Specifically, reconstruction is necessary because vertical oscillation double-integration of the single stride signal. In this paper, we choose linear interpolation as reconstruction method, which has low complexity, enabling on-board reconstruction. Note that the LLA is also used to estimate stride-by-stride running form changes based on stride time, and this information is used to group similar strides to further reduce sampling rate. For example, if every stride inside a group is close to the mean stride and the runner does not require stride-by-stride feedback, essentially, only one running stride needs to be processed to provide the running form metrics. However, as we will show in Section 5.6.2, the actual amount of energy saving depends on a runner's consistency, which varies by the experience and fitness of a runner.

5.6 Evaluation

To evaluate the energy efficiency and accuracy of the Gazelle wearable system for online running analysis, we conducted both in-lab experiments of the SAS algorithm and in-field pilot studies.

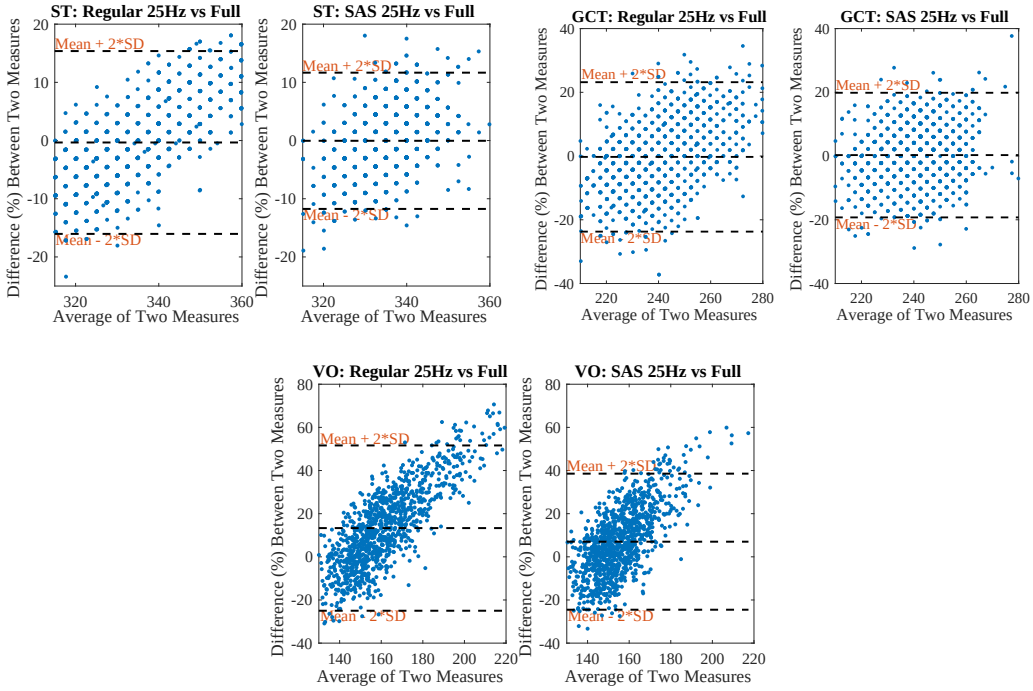

Fig. 5.13. Bland-Altman plots for regular 25 Hz sampling and SAS algorithm.

5.6.1 In-lab Experiments

For the in-lab experiments, we first compared the accuracy of our proposed SAS algorithm with that of the compressed sensing (CS). Although due to the intensive computation cost of CS, CS is not an optimal option for on-board sampling rate reduction without sufficient hardware support, CS is the leading approach to achieve high accuracy with a low sampling rate. Thus, in this experiment, we primarily compare SAS and CS from the perspective of reconstruction accuracy. In the experiment, seven 30 minute-long running datasets were recorded on an outdoor track. Each runner wore a chest band with the Gazelle device attached to the band in the center front location.

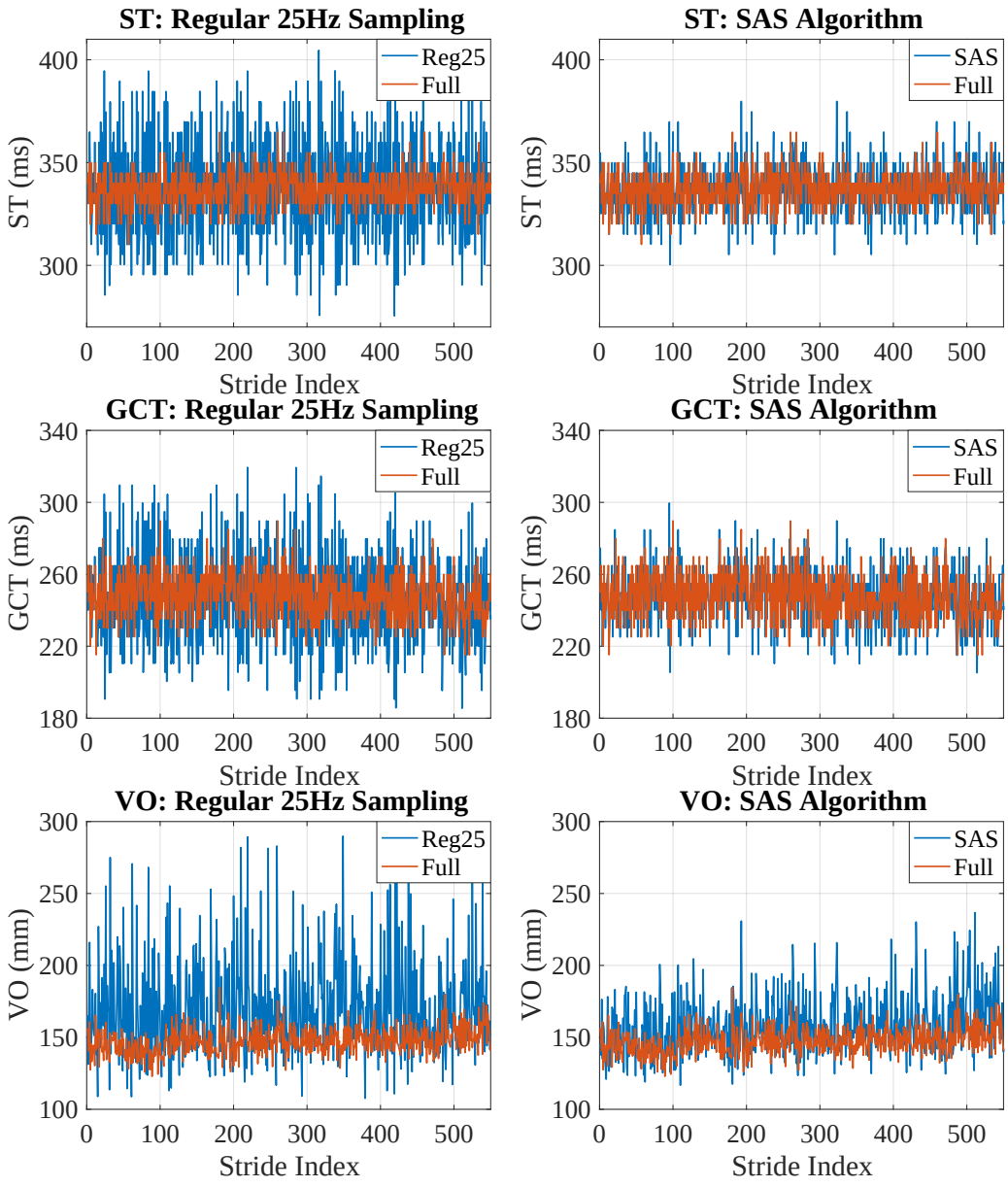

Fig. 5.14. Stride by stride performance.

In the test, both real-time running metrics and raw acceleration samples collected from HHA were streamed to a mobile phone for post validation. The key running metrics: ST, GCT, VO were computed as a comparative baseline from the raw data sampled from HHA over the entire running test. To determine the general trade-offs between sparse (adaptive) sensing rates and energy savings, we computed the average accuracy using stride-by-stride running form metrics, which did not include

the added benefits of grouping similar strides together. The accuracy was defined in Eqn. 5.10.

$$Accuracy_{avg} = \frac{1}{N}\sum_{n=1}^{N}(1 - \frac{|M^n_{\{a\}} - M^n|}{|M^n|}) \tag{5.10}$$

where $M_{\{a\}}$ is the metric computed from either SAS or CS resulted running signal, M^n is the metric computed from full 200 Hz sampled running signal. $n = 1, 2...N$ is the index of each stride for a specific running metric.

For CS, the sampling rate was fixed for each experiment; while for SAS, the sampling rate changed dynamically and the average sampling rate was used for comparison. With the results from Fig. 5.11, it can be referred that SAS outperforms CS in term of achieving lower sampling rate with sufficient accuracy, provide more potential to reduce energy consumption either for online signal processing or wireless transmission. Fig. 5.11 compares the accuracy between CS and SAS for different running metrics under different sampling rates of the HHA. We can see that SAS outperforms CS in almost all the scenarios. For ST, GCT, and VO, an average sampling rate of 25 Hz is sufficient to maintain higher than 99.0%, 98.6%, and 95.1% accuracy respectively, and this is sufficient for runners' feedback. Compared with our SAS method, CS achieves comparable accuracy for stride time, but has worse performance for GCT and VO, and cannot obtain an average of 90% accuracy when the sampling rate is lower than 30 Hz and 40 Hz, respectively. The major reason is demonstrated in Fig. 5.10: CS has much lower signal to noise ratio, and therefore error is accumulated when aggregating the ground contact time, and as vertical oscillation requires double integration, error is further accumulated.

We also conducted power modeling and analysis to determine the energy savings of the SAS approach as compared with the constant 200 Hz approach. Shown in Fig. 5.12, the current per sample was computed for SAS so we can compare the resulting dynamic sampling rate of the HHA and the static 3 μA of the LLA. The average current per sample of the HHA can be computed as a combination of the current cost for a single conversion of the HHA in high-resolution mode (240 μA) over the HHA start-up time, and the HHA power-down current cost (6 μA) for the remainder of the

sampled interval time for that sample. Overall, an average of 25 Hz sampling rate is required for SAS to achieve greater than 97.7% accuracy for all running metrics with over 76.9% energy savings. This represents one order of magnitude improvement over existing wearable running analysis devices while outperforming CS in accuracy and achieving significantly lower computational overhead by operating exclusively in the time domain. To further validate the effectiveness of SAS algorithm at 25 Hz, we compared its performance with the regular 25 Hz sampling approach. Fig. 5.13 and Fig. 5.14 demonstrates that: (1) Regular 25 Hz sampling results in comparable average accuracy compared with SAS algorithm at 25 Hz, however, it has larger error range and its performance varies significantly from stride to stride. (2) The regular 25 Hz sampling method has more than 7% error in average for VO than SAS algorithm. The reason is that regular 25 Hz sampling is not able to capture most of minima or maxima at sharp transitions. Thus, an adaptive, irregular sampling strategy like the SAS algorithm we proposed is necessary to reduce energy consumption while maintaining high measurement accuracy.

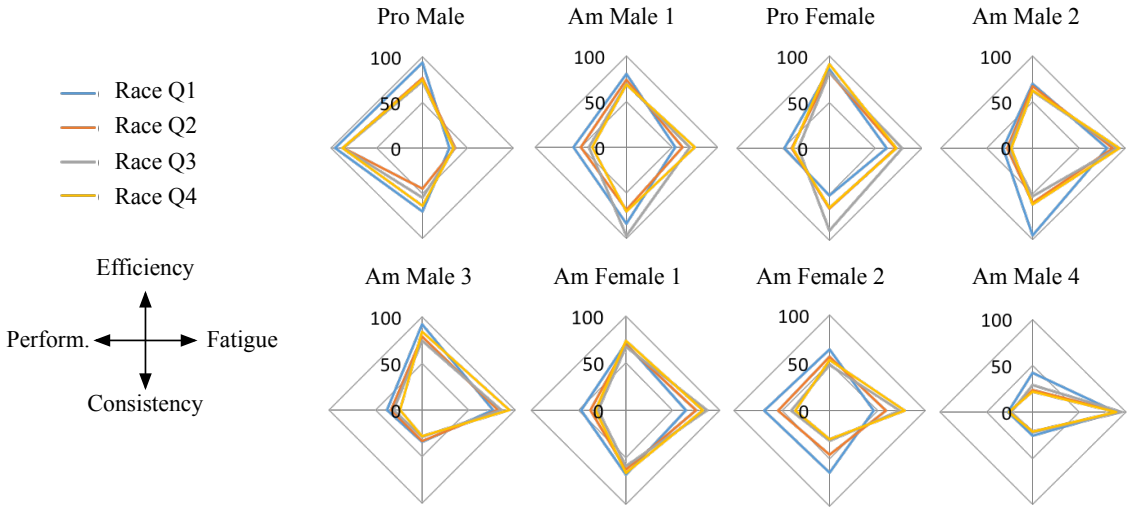

Fig. 5.15. Gazelle running analysis for top professional and elite triathletes at the Ironman World Championships in Kona, HI.

In addition, in an actual usage scenario, runners may have different demands of running metrics. Thus, the maximum energy savings can vary for different metric subsets. For example,

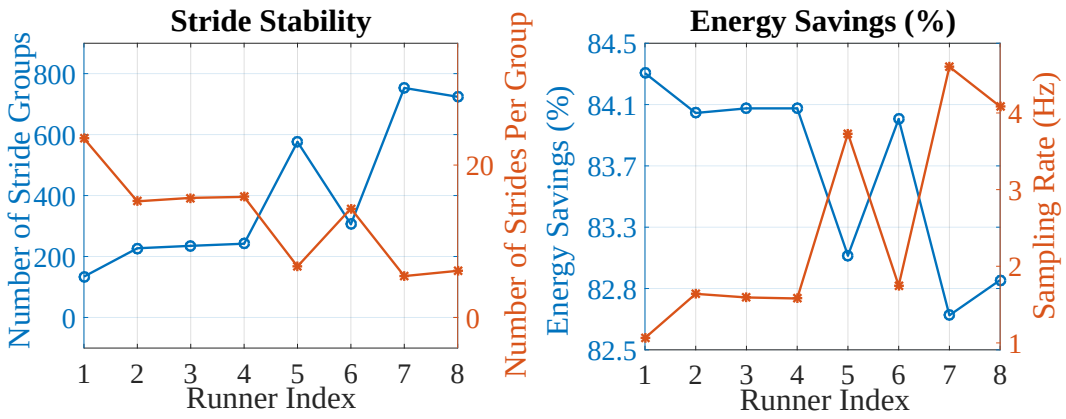

Fig. 5.16. Stride stability vs. energy savings for eight different runners in the Kona Ironman World Championships.

for stride time alone, the LLA active in interrupt-only mode is sufficient to capture these metrics at a 10 Hz sampling rate, and the energy savings can reach 99% compared with 200 Hz HHA. In future work, different usage scenarios can be studied. As shown, different running metrics require a different sampling rate to reach an accurate enough measurement. Therefore SAS can be designed to adapt to different sets of running metrics to further minimize the power consumption under various usage cases. In summary, our sparse adaptive sensing (SAS) algorithm is energy-efficient and accurate for running form analysis and feedback, and provide a solution for long-term running form study, and a potential guide for other similar applications.

Note that the accuracy and energy saving numbers above are for stride-by-stride running form analysis. Further sampling rate reduction can be achieved by grouping strides with a similar running profile, which depends on how consistently the runner is running. Next, we further evaluate the energy savings from runners with different experience levels based on pilot studies in real-world running races.

5.6.2 Pilot Study

In addition to laboratory testing and outdoor track testing, Gazelle was used in the Ironman World Championships in October 2014 Kona, Hawaii, the world's premier Ironman race event. In Kona, Gazelle monitored the marathon segments of two professional triathletes and six of the

world's best athletes in their age brackets. This section will focus on reporting and analyzing Gazelle's results for the eight athletes from this race. The focus of this pilot study was two-fold: 1) to test consistency of the metrics derived from the Gazelle wearable under the energy savings with SAS achieved in real-world running; and, 2) to understand Gazelle's metrics' overall usability in terms of running form information representation when compared across some of the world's best triathletes under race conditions.

Energy savings in real-world running: Stride-by-stride running-form consistency affects the performance and the energy savings of SAS. As described in the previous section, across ten runners data collected during in-lab experiments, an average of 25 Hz sampling rate was needed to achieve over 97% accuracy for all computed running form metrics. Running-form consistency varies among runners. Under the same stride time variance constraint, better running-form consistency leads to larger number of strides per group, hence lower data sampling rate and better energy savings. Fig. 5.16 shows the number of groups and the number of strides per group for each runner with 1% stride time variance. From this figure, Runner 1 shows the highest running-form consistency or minimal stride-by-stride variance, which leads to the largest number of strides per group, hence the lowest data sampling rate (1 Hz), and therefore largest energy savings (84.3%). On the other hand, Runner 7 shows the lowest running-form consistency, requiring the highest average data sampling rate (5 Hz), and resulting in the lowest energy savings (82.6%). Overall, an average energy savings of 83.6% was achieved across these eight runners.

Table 5.2: *RunQuality* scores vs race time

Runner	**RunQuality**	**Race Time**	**Level**
Pro Male	90.3	2h:58m:58s	4
Am Male 1	85.6	3h:14m:12s	3
Pro Female	86.2	3h:21m:34s	3
Am Male 2	80.6	3h:41m:51s	2
Am Male 3	74.7	3h:41m:51s	2
Am Female 1	80.5	3h:52m:38s	2
Am Female 2	75.0	4h:07m:16s	2
Am Male 4	62.5	5h:02m:54s	1

Metric report consistency: Based on the high-level metrics shown Fig. 5.15, the averaged RunQuality scores for all eight runners are summarized in Table 5.2 along with each of their race completion times. It can be seen that based on the race time, the runners can be classified into four run skill levels, and the *RunQuality* derived from the run form metrics measured by Gazelle is highly consistent with runners' actual race results, as well as the associated energy savings from Gazelle. This comparison serves to validate the feasibility and methodology of Gazelle wearable under real-world use. The following equations describe the high-level metrics, which are constructed post-race in terms of Gazelle's reported running metrics.

- $Efficiency = \frac{1}{t_{air} \times pace}$, *Efficiency* estimates how much energy is spent to propel the runner over the distance traveled.
- $Fatigue = \frac{t_{ground}}{t_{air}}$, *Fatigue* is an estimate of how tired the runner is.
- $Performance = Mean(\frac{t_{air}}{t_{ground}})$, *Performance* is an estimate for how much energy a runner is putting into the ground.
- $Consistency = StdDev(\frac{t_{air}}{t_{ground}})$.

Taken together, *RunQuality* is an aggregated measure of all the four high-level metrics described above. It is a simple unity weighted combination of the four, with the desirable set {*Efficiency, Consistency, Performance*} having positive unity weight and the undesirable set {*Fatigue*} having negative unity weight. The summation of the two sets together is a runner's RunQuality metric.

$$\begin{aligned} RunQuality &= Efficiency + Consistency \\ &\quad + Performance - Fatigue \end{aligned}$$

In the weeks following the Ironman World Championships at Kona, athletes and their coaches reviewed the running form metrics data that were generated by Gazelle. The feedbacks we received were consistent among most athletes and coaches that Gazelle was easy to use and the running form

metrics were useful for both understanding the precise places in the race where unexpected events occurred and for further improvement of the athletes' running form and racing strategy.

5.7 Chapter Summary

In this work, to tackle the challenges associated with the high energy consumption of high-precision motion sensing and analysis, we have developed an intelligent sparse adaptive sensing (SAS) algorithm. The design of SAS algorithm is based on two types of context information: the running state context (semantic) and intra-signal context (temporal). The algorithm provides a running form analysis solution, along with aggressive energy management techniques. Experiments using real-world running data demonstrate that, compared with uniform sensing at 200 Hz, SAS can achieve 97.7% accuracy and 76.9% energy saving with only an 25 Hz maximal sampling rate. As a result, together with the improvement in usable energy capacity due to lower average current draw, Gazelle can increase the battery life by one order of magnitude using a small coin-cell battery. Through our year-long pilot studies, Gazelle has been in use by over a hundred elite and recreational runners during day-to-day training and various racing events, with satisfactory results.

Chapter 6

Conclusions and Future Research

This chapter first concludes the research works presented in this thesis. The research in this thesis presents a general methodology to solve contextual anomaly detection problems using spatial-temporal data from various domains – remote sensing, photovoltaic systems, and low-power embedded systems. Additionally, this chapter discusses possible future studies for the context-aware anomaly detection techniques and its applicability to broader domains.

6.1 Thesis Summary

The main contributions of this thesis are the design, implementation, and evaluation of context-aware anomaly detection and analysis approach. The approach follows a modular, hierarchical design pattern from context conceptualization, feature engineering to modeling. This methodology enables flexible cross-domains and cross-scenarios adaptation. The detailed discussion about the contributions and their implications to broader applications are summarized as follows.

Approach to conceptualize anomaly 'context': The capability to identify a proper context is essential for achieving high-accuracy anomaly detection. This thesis proposes two methods to construct such contexts: (1) For a dataset that has explicit contextual attributes such as spatial, temporal attributes, the context can be constructed from a data object's spatial-temporal neighborhood. Depending on the focus of the study and the domain-specific settings, the granularity of the spatial-temporal neighborhood can change accordingly. For instance, in the work of contextual-anomaly detection for remotely sensed data, the data are images. Naturally, each data object

has nearby objects spatially and temporally. Hence a cube-like neighborhood can be constructed. (2) While for datasets such as SCADA data from the photovoltaic systems, only temporal information is explicitly defined. However, considering the importance of the spatial variance in such distributed sensor networks, a virtual spatial neighborhood for the modeling process is essential to improve detection accuracy and reduce false alarms. This thesis proposes and develops a hierarchical context-aware approach combining a conceptualized spatial-temporal neighborhood for such scenarios. The proposed approach effectively enhances the anomaly detection accuracy compared with existing methods that not apply context information.

Feature space design for contextual anomaly detection: A general approach to derive feature space from a spatial-temporal neighborhood is developed. The feature space contains both statistical and analytical information such as mean, standard deviation, differences (first, or higher orders), correlation and gradients of an object or within a neighborhood. This full feature space is applicable to various spatial-temporal datasets. Depends on a specific application, a subset of those features could be sufficient enough to help solve the anomaly problem.

Unsupervised anomaly modeling and ranking: This thesis tackles another major challenge – lack of prior knowledge. All the context-aware anomaly detection approaches in this thesis do not require priorly known models of normal or abnormal data. To further facilitate the usage of the anomaly detection results, ranking mechanisms for the remote sensing and the photovoltaic systems are designed based on the level-of-interest of each anomalous events and the spatial-temporal relationship among outliers. Experiment results show that top-ranked events in the remote sensing application are real data quality issues or significant natural events, while for the PV systems, 83% of top 100 anomalies are true positives, improved by 20% comparing existing anomaly detection methods in the PV domains.

On-demand efficient sampling using anomalies: Typical anomaly detection and analysis focus on the effect and prevention of anomalies. In the Gazelle human running analysis work discussed in this thesis, a different angle regarding 'anomalies' is presented. By utilizing the anomaly detection techniques, a sparse adaptive sensing algorithm is proposed and implemented to reduce

power consumption for wearable embedded systems. The whole methodology is to capture the 'interesting' patterns and hence to adjust the sampling rate. This sampling strategy improves the accuracy of measuring the patterns as well as reduces overall power consumed. This idea can be extended to general sensing techniques to not only reduce power but also reduce the overhead of data storage and data analysis, by only keeping 'interesting' samples.

6.2 Future Research

As mentioned above, overall, 'unsupervised' anomaly detection is the key direction in Big Data era. The ability to identify proper context for each domain-specific challenge can largely reduce the algorithm complexity while achieving high accuracy. The technology trend in data mining community will naturally improve the efficiency and effectiveness of knowledge discovery as time passes. In the meantime, there are several directions that we can investigate. For each of the domain problem discussed in this thesis, there are several directions to improve the algorithms' performance and expand the applicability.

Adaptive real-time learning to Gazelle's SAS algorithm: The current SAS algorithm is motivated by a thorough off-line data analysis, leveraging the intra-stride signal characteristics. A set of predetermined sampling rates is required for the algorithm to run real-time. Hence, a natural extension to the existing work would be to enable SAS to learn the set of sampling rates on-the-fly to fit sampling rate and duration to the individual and the real-time running gait characteristics. This proposed work could further reduce power consumption on average compared to the current SAS design and broaden the application areas beyond running activity tracking due to the automated adaptation on-the-fly.

Automated anomalous event clustering: The extended GMM clustering and event ranking mechanism can identify real anomalies caused by both data quality issues and natural events. An extension to the current solution is the automated categorization of anomalies to natural events and data quality problems, as users for such technologies can come from both data engineering and scientific researchers. With the proposed algorithm, both parties can benefit from it. Data engi-

neers and quality assurance team can quickly identify and hence fix data issues that can potentially introduce spurious scientific conclusions. While for researchers, instead of spending the majority of the time to clean data, they can focus more on scientific discoveries.

Another extension to this work is anomaly detection under distributed data management systems. The motivation for this proposal is that transferring and storing data in a centralized location is not cost-effective. Most Earth observations data nowadays are already stored in a distributed fashion, pulling all data to centralized cloud storage is not efficient and cost-effective. Hence, treating each data center as a node in a data network, and transferring concise, summarized local statistics or analytic patterns to a centralized data analysis center could drive down the cost and also potentially achieve country-level, event global anomaly detection promptly.

Adaptive and automated learning in distributed energy systems: Large-scale infrastructure systems, such as power plants, serve as the backbone of modern society. Accurate system modeling is essential for effective system management, optimized system performance, minimized system downtime, and extended system lifetime. Many of these large-scale systems are distributed in nature – the overall system is a hierarchical composition of numerous self-acting components, often with similar functionalities. For instance, a wind farm typically consists of tens of or over a hundred wind turbines; a solar farm may consist of over a hundred thousand solar panels, and an energy storage system may include thousands of rechargeable battery cells. For such large-scale distributed systems, the overall system characteristics, e.g., system utility, run-time system failures, and long-term system aging, are aggregated effects of individual sub-system components. Therefore, a more general learning and modeling system beyond anomaly detection at detailed component-level are required. For instance, accurate quantification of the dust accumulation level of individual solar panels can help optimize the cleaning schedule and maximize the overall solar farm profit. On the other hand, optimized control of individual components often requires a comprehensive view of the overall system condition and the ambient environment, collectively gathered by the rest of the sub-system components. For example, in a wind farm, the wind speed measurement by forefront wind turbines can enable feed-forward control for the rest of the wind turbines, thereby optimizing

their power production. Also, early-stage anomalies detected by individual wind turbines can help predict potential failures of the other turbines, and schedule system maintenance ahead of time, thereby minimizing the overall wind farm downtime.

Bibliography

[1] 80 million European runners reveal their reasons to run. http://www.prnewswire.com/news-releases.

[2] Adidas. http://micoach.adidas.com/speed_cell/.

[3] Garmin Foot Pod. https://buy.garmin.com/en-US/US/p/15516.

[4] Nike+ Foot Pod. https://www.apple.com/ipod/nike/run.html.

[5] Nordic Semiconductor, Bluetooth Smart and 2.4GHz proprietary SoC.

[6] Number of people who went jogging or running with the last 12 months in the United States. http://www.statista.com/statistics/227423/number-of-joggers-and-runners-usa.

[7] Suunto Foot Pod. http://www.suunto.com/Products/PODs/Suunto-Foot-POD-Mini/.

[8] International standard IEC 61724: Photovoltaic system performance monitoring–guidelines for measurements, data exchange and analysis. 1998.

[9] J. Ahmed et al. An accurate method for MPPT to detect the partial shading occurrence in PV system. IEEE Transactions on Industrial Informatics, PP(99):1–1, 2017.

[10] Cesare Alippi, Stavros Ntalampiras, and Manuel Roveri. A cognitive fault diagnosis system for distributed sensor networks. IEEE transactions on neural networks and learning systems, 24(8):1213–1226, 2013.

[11] Cesare Alippi, Stavros Ntalampiras, and Manuel Roveri. Model-free fault detection and isolation in large-scale cyber-physical systems. IEEE Transactions on Emerging Topics in Computational Intelligence, 1(1):61–71, 2017.

[12] Aïda Alvera-Azcárate et al. Outlier detection in satellite data using spatial coherence. Remote Sensing of Environment, 119:84–91, 2012.

[13] Kamiar Aminian et al. Gait analysis using shoe-worn inertial sensors: how is foot clearance related to walking speed? In Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, pages 481–485, 2014.

[14] T Andersson. Multivariate statistical analysis. John Wiley and Sons, Inc., New York, 1958.

[15] B. Andò et al. Sentinella: Smart monitoring of photovoltaic systems at panel level. IEEE Transactions on Instrumentation and Measurement, 64(8):2188–2199, 2015.

[16] Gonzalo R Arce. Nonlinear signal processing: a statistical approach. John Wiley & Sons, 2005.

[17] Çağlar Arı, Selim Aksoy, and Orhan Arıkan. Maximum likelihood estimation of gaussian mixture models using stochastic search. Pattern Recognition, 45(7):2804–2816, 2012.

[18] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[19] T.S. Ashok and A. Pardeshi Sanjay. Kinematic study of video gait analysis. In International Conference on Industrial Instrumentation and Control, pages 1208–1213, 2015.

[20] Joonbum Bae. Gait analysis based on a hidden markov model. In 12th International Conference on Control, Automation and Systems, pages 1025–1029, 2012.

[21] A.Y. Benbasat and J.A. Paradisio. Design of a real-time adaptive power optimal system. In Proceedings of IEEE Sensors, pages 48–51 vol.1, Oct 2004.

[22] Yoshua Bengio et al. Learning deep architectures for AI. Foundations and trends® in Machine Learning, 2(1):1–127, 2009.

[23] Kanishka Bhaduri, Bryan L Matthews, and Chris R Giannella. Algorithms for speeding up distance-based outlier detection. In Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining, pages 859–867. ACM, 2011.

[24] Derya Birant and Alp Kut. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data and Knowledge Engineering, 60(1):208–221, 2007.

[25] S Bokhorst, H Tømmervik, TV Callaghan, GK Phoenix, and JW Bjerke. Vegetation recovery following extreme winter warming events in the sub-Arctic estimated using NDVI from remote sensing and handheld passive proximal sensors. Environmental and Experimental Botany, 81:18–25, 2012.

[26] Jason Bonacci et al. Running in a minimalist and lightweight shoe is not the same as running barefoot: a biomechanical study. British Journal of Sports Medicine, 2013.

[27] Erik Borg, Bernd Fichtelmann, and Hartmut Asche. Assessment for remote sensing data: accuracy of interactive data quality interpretation. In International Conference on Computational Science and Its Applications, pages 366–375. Springer, 2011.

[28] Sid-Ahmed Boukabara, Fuzhong Weng, and Quanhua Liu. Passive microwave remote sensing of extreme weather events using NOAA-18 AMSUA and MHS. Geoscience and Remote Sensing, IEEE Transactions on, 45(7):2228–2246, 2007.

[29] M Bressan et al. A shadow fault detection method based on the standard error analysis of IV curves. Renewable Energy, 99:1181–1190, 2016.

[30] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In ACM Sigmod Record, volume 29, pages 93–104. ACM, 2000.

[31] B. Brooks. The bakersfield fire-a lesson in ground-fault protection. SolarPro Mag, pages 62–70, 2011.

[32] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding AIC and BIC in model selection. Sociological methods & research, 33(2):261–304, 2004.

[33] E.J. Candes and M.B. Wakin. An introduction to compressive sampling. IEEE Signal Processing Magazine, 25(2):21–30, 2008.

[34] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: a survey. ACM Computing Surveys, 41(3):15, 2009.

[35] Huanhuan Chen et al. Learning in the model space for cognitive fault diagnosis. IEEE transactions on neural networks and learning systems, 25(1):124–136, 2014.

[36] Leian Chen, Shang Li, and Xiaodong Wang. Quickest fault detection in photovoltaic systems. IEEE Transactions on Smart Grid, 2016.

[37] Leian Chen and Xiaodong Wang. Adaptive fault localization in photovoltaic systems. IEEE Transactions on Smart Grid, 2017.

[38] Tao Cheng and Zhilin Li. A multiscale approach for spatio-temporal outlier detection. Transactions in GIS, 10(2):253–263, 2006.

[39] W. Chine et al. A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. Renewable Energy, 90:501–512, 2016.

[40] DaeKi Cho et al. Autogait: A mobile platform that accurately estimates the distance walked. In IEEE International Conference on Pervasive Computing and Communications, pages 116–124, 2010.

[41] A. Chouder et al. Automatic supervision and fault detection of PV systems based on power losses analysis. Energy Conversion and Management, 51(10):1929–1937, 2010.

[42] C Christophe et al. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. Computational Statistics & Data Analysis, 41(3):561–575, 2003.

[43] Fowler Chuck, James Maslanik, Terry Haran, Ted Scambos, Jeffrey Key, and William Emery. AVHRR Polar Pathfinder Twice-daily 5 km EASE-Grid Composites V003, [July 1981 to October 1997]. Boulder, Colorado USA: National Snow and Ice Data Center. [January 2016 Accessed]., 2000, updated 2007.

[44] Corinna Cortes and Vladimir Vapnik. Support vector machine. Machine learning, 20(3):273–297, 1995.

[45] Dim Coumou and Stefan Rahmstorf. A decade of weather extremes. Nature Climate Change, 2(7):491–496, 2012.

[46] Thomas Cover et al. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1):21–27, 1967.

[47] P. Dagnely et al. A semantic model of events for integrating photovoltaic monitoring data. In IEEE 13th International Conference on Industrial Informatics, pages 24–30, 2015.

[48] A. Daoud et al. Foot strike and injury rates in endurance runners: a retrospective study. Medicine and Science in Sports and Exercise, 2012.

[49] Kaustav Das, Jeff Schneider, and Daniel B Neill. Anomaly pattern detection in categorical datasets. In Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining, pages 169–176. ACM, 2008.

[50] Mahmoud Dhimish and Violeta Holmes. Fault detection algorithm for grid-connected photovoltaic plants. Solar Energy, 137:236–245, 2016.

[51] Mahmoud Dhimish, Violeta Holmes, and Mark Dales. Parallel fault detection algorithm for grid-connected photovoltaic plants. Renewable Energy, 113:94–111, 2017.

[52] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, 40(2):139–157, 2000.

[53] A.M.R. Dixon et al. Compressed sensing system considerations for ECG and EMG wireless biosensors. Biomedical Circuits and Systems, IEEE Transactions on, 6(2):156–166, 2012.

[54] A. Drews et al. Monitoring and remote failure detection of grid-connected PV systems based on satellite observations. Solar Energy, 81(4):548–564, 2007.

[55] David R Easterling, Gerald A Meehl, Camille Parmesan, Stanley A Changnon, Thomas R Karl, and Linda O Mearns. Climate extremes: observations, modeling, and impacts. Science, 289(5487):2068–2074, 2000.

[56] S. Feizi, V.K. Goyal, and M. Médard. Locally adaptive sampling. In 48th Annual Allerton Conference on Communication, Control, and Computing, pages 152–159, 2010.

[57] Soheil Feizi et al. Backward adaptation for power efficient sampling. Signal Processing, IEEE Transactions on, 62(16):4327–4338, 2014.

[58] Soheil Feizi-Khankandi, Vivek K. Goyal, and Muriel Médard. Time-stampless adaptive nonuniform sampling for stochastic signals. IEEE Transactions on Signal Processing, 60(10):5440–5450, 2012.

[59] Craig R Ferguson and Gabriele Villarini. Detecting inhomogeneities in the Twentieth Century Reanalysis over the central United States. Journal of Geophysical Research: Atmospheres, 117(D5), 2012.

[60] S.K. Firth et al. A simple model of PV system performance and its use in fault detection. Solar Energy, 84(4):624–635, 2010.

[61] Otto Fischer and Wilhelm Braune. Der Gang des Menschen: Versuche am unbelasteten und belasteten Menschen. Hirzel Verlag, 1985.

[62] Raghu K Ganti et al. Satire: a software architecture for smart attire. In Proceedings of the 4th international conference on Mobile systems, applications and services, pages 110–123, 2006.

[63] Elyes Garoudja et al. Statistical fault detection in photovoltaic systems. Solar Energy, 150:485–499, 2017.

[64] Mark D Goldberg, Yanni Qu, Larry M McMillin, Walter Wolf, Lihang Zhou, and Murty Divakarla. Airs near-real-time products and algorithms in support of operational numerical weather prediction. IEEE Transactions on Geoscience and Remote Sensing, 41(2):379–389, 2003.

[65] Avid Roman Gonzalez and Mihai Datcu. Data cleaning: approach for Earth observation image information mining. In ESA-EUSC-JRC 2011 Image Information Mining: Geospatial Intelligence from Earth Observation Conference, pages 117–120, 2011.

[66] Glenn Edwin Grant. Exploring Antarctic Land Surface Temperature Extremes Using Condensed Anomaly Databases. PhD thesis, University of Colorado at Boulder, 2017.

[67] Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han. Outlier detection for temporal data. Synthesis Lectures on Data Mining and Knowledge Discovery, 5(1):1–129, 2014.

[68] R Hariharan et al. A method to detect photovoltaic array faults and partial shading in PV systems. IEEE Journal of Photovoltaics, 6(5):1278–1285, 2016.

[69] Masroor Hussain, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley. Change detection from remotely sensed images: From pixel-based to object-based approaches. ISPRS Journal of Photogrammetry and Remote Sensing, 80:91–106, 2013.

[70] David Isaac and Christopher Lynnes. Automated data quality assessment in the intelligent archive. White Paper prepared for the Intelligent Data Understanding Program, 17, 2003.

[71] A. Jardine et al. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing, 20(7):1483–1510, 2006.

[72] Kian Jazayeri, Moein Jazayeri, and Sener Uysal. Artificial neural network-based all-sky power estimation and fault detection in photovoltaic modules. Journal of Photonics for Energy, 7(2):025501–025501, 2017.

[73] Huaiguang Jiang et al. Fault detection, identification, and location in smart grid based on data-driven computational methods. IEEE Transactions on Smart Grid, 5(6):2947–2956, 2014.

[74] Huaiguang Jiang et al. Spatial-temporal synchrophasor data characterization and analytics in smart grid fault detection, identification, and impact causal analysis. IEEE Transactions on Smart Grid, 7(5):2525–2536, 2016.

[75] C Birk Jones et al. Automatic fault classification of photovoltaic strings based on an in situ IV characterization system and a gaussian process algorithm. In Photovoltaic Specialists Conference (PVSC), 2016 IEEE 43rd, pages 1708–1713. IEEE, 2016.

[76] NA Kane et al. Validity of the Nike+ device during walking and running. International Journal of Sports Medicine, 31(2):101–105, 2010.

[77] Seungwoo Kang et al. Seemon: Scalable and energy-efficient context monitoring framework for sensor-rich mobile environments. In Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services, pages 267–280, 2008.

[78] K. Kanoun et al. A real-time compressed sensing-based personal electrocardiogram monitoring system. In Design, Automation Test in Europe Conference Exhibition, pages 1–6, 2011.

[79] Jaya Kawale, Snigdhansu Chatterjee, Arjun Kumar, Stefan Liess, Michael Steinbach, and Vipin Kumar. Anomaly construction in climate data: Issues and challenges. In CIDU, pages 189–203, 2011.

[80] Eamonn Keogh. Exact indexing of dynamic time warping. In Proceedings of the 28th international conference on Very Large Data Bases, pages 406–417. VLDB Endowment, 2002.

[81] Katherine A Kim et al. Photovoltaic hot-spot detection for solar panel substrings using AC parameter characterization. IEEE Transactions on Power Electronics, 31(2):1121–1130, 2016.

[82] Edwin M Knorr, Raymond T Ng, and Vladimir Tucakov. Distance-based outliers: algorithms and applications. The International Journal on Very Large Data Bases, 8(3-4):237–253, 2000.

[83] Edwin M Knox and Raymond T Ng. Algorithms for mining distance based outliers in large datasets. In Proceedings of the International Conference on Very Large Data Bases, pages 392–403. Citeseer, 1998.

[84] Martin Kulldorff. A spatial scan statistic. Communications in Statistics-Theory and methods, 26(6):1481–1496, 1997.

[85] Alp Kut and Derya Birant. Spatio-temporal outlier detection in large databases. CIT. Journal of Computing and Information Technology, 14(4):291–297, 2006.

[86] Gyemin Lee and Clayton Scott. EM algorithms for multivariate gaussian mixture models with truncated and censored data. Computational Statistics and Data Analysis, 56(9):2816–2829, 2012.

[87] Yaguo Lei et al. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. IEEE Transactions on Industrial Electronics, 63(5):3137–3147, 2016.

[88] Songnian Li, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, et al. Geospatial big data handling theory and methods: A review and research challenges. ISPRS Journal of Photogrammetry and Remote Sensing, 115:119–133, 2016.

[89] Daniel E. Lieberman et al. Foot strike patterns and collision forces in habitually barefoot versus shod runners. Nature, 463:531–535, 2010.

[90] Guimei Liu et al. Repeat buyer prediction for e-commerce. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 155–164. ACM, 2016.

[91] Hancong Liu et al. On-line outlier detection and data cleaning. Computers & chemical engineering, 28(9):1635–1647, 2004.

[92] Q. Liu et al. Unsupervised detection of contextual anomaly in remotely sensed data. Remote Sensing of Environment, 2017.

[93] Q. Liu, J. Williamson, K. Li, W. Mohrman, Q. Lv, R. P. Dick, and L. Shang. Gazelle: Energy-efficient wearable analysis for running. IEEE Transactions on Mobile Computing, 16(9):2531–2544, Sept 2017.

[94] Qi Liu et al. Hierarchical context-aware anomaly diagnosis in large-scale PV systems using SCADA data. In Fifteenth International Conference on Industrial Informatics. IEEE, 2017.

[95] Tao Liu et al. Three-dimensional gait analysis system with mobile force plates and motion sensors. In 8th International Conference on Ubiquitous Robots and Ambient Intelligence, pages 107–110, 2011.

[96] Juan I López-Moreno, Ahmed El-Kenawy, Jesús Revuelto, César Azorín-Molina, Enrique Morán-Tejeda, Jorge Lorenzo-Lacruz, Javier Zabalza, and Sergio M Vicente-Serrano. Observed trends and future projections for winter warm events in the Ebro basin, northeast Iberian Peninsula. International Journal of Climatology, 34(1):49–60, 2014.

[97] Konrad Lorincz et al. Mercury: a wearable sensor network platform for high-fidelity motion analysis. In SenSys, volume 9, pages 183–196, 2009.

[98] Shane Lowe and Gearöid ÖLaighin. The age of the virtual trainer. Procedia Engineering, 34:242 – 247.

[99] C-T Lu, Dechang Chen, and Yufeng Kou. Algorithms for spatial outlier detection. In Third International Conference on Data Mining, pages 597–600. IEEE, 2003.

[100] George Luber and Michael McGeehin. Climate change and extreme heat events. American journal of preventive medicine, 35(5):429–435, 2008.

[101] Yan Ma, Haiping Wu, Lizhe Wang, Bormin Huang, Rajiv Ranjan, Albert Zomaya, and Wei Jie. Remote sensing big data computing: challenges and opportunities. Future Generation Computer Systems, 51:47–60, 2015.

[102] Stéphane Mallat. IX - an approximation tour. In A Wavelet Tour of Signal Processing (Second Edition), pages 376 – 433. Academic Press, 1999.

[103] J. Maslanik and J. Stroeve. DMSP SSM/I-SSMIS daily polar gridded brightness temperatures, version 4. [July 1987 to June 2015]. Boulder, Colorado USA: National Snow and Ice Data Center Distributed Active Archive Center. doi: http://dx.doi.org/10.5067/AN9AI8EO7PX0. [January 2016 Accessed]., 2004, updated 2016.

[104] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. Journal of the American statistical Association, 46(253):68–78, 1951.

[105] Heidrun Matthes, Annette Rinke, and Klaus Dethloff. Recent changes in Arctic temperature extremes: warm and cold spells during winter and summer. Environmental Research Letters, 10(11):114020, 2015.

[106] Aaron M McCright, Riley E Dunlap, and Chenyang Xiao. The impacts of temperature anomalies and political orientation on perceived winter warming. Nature Climate Change, 4(12):1077–1081, 2014.

[107] H Mekki, Adel Mellit, and H Salhi. Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules. Simulation Modelling Practice and Theory, 67:1–13, 2016.

[108] Mladen Milosevic, Aleksandar Milenkovic, and Emil Jovanov. mHealth@UAH: computing infrastructure for mobile health and wellness monitoring. XRDS, 20(2):43–49, 2013.

[109] Todd K Moon. The expectation-maximization algorithm. IEEE Signal processing magazine, 13(6):47–60, 1996.

[110] Carolyn F Munro, Doris I Miller, and Andrew J Fuglevand. Ground reaction forces in running: a reexamination. Journal of biomechanics, 20(2):147–155, 1987.

[111] Satoshi Murata, Masanori Suzuki, and Kaori Fujinami. A wearable projector-based gait assistance system and its application for elderly people. In Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, pages 143–152, 2013.

[112] Alvaro Muro-de-la Herran et al. Gait analysis methods: an overview of wearable and non-wearable systems, highlighting clinical applications. Sensors, 14(2):3362, 2014.

[113] Sina Muster, Moritz Langer, Anna Abnizova, Kathy L Young, and Julia Boike. Spatio-temporal sensitivity of MODIS land surface temperature anomalies indicates high potential for large-scale land cover change detection in Arctic permafrost landscapes. Remote Sensing of Environment, 168:1–12, 2015.

[114] SV Nghiem, DK Hall, TL Mote, Marco Tedesco, MR Albert, K Keegan, CA Shuman, NE DiGirolamo, and G Neumann. The extreme melt across the Greenland Ice Sheet in 2012. Geophysical Research Letters, 39(20), 2012.

[115] D. Nguyen et al. Performance evaluation of solar photovoltaic arrays including shadow effects using neural network. In Energy Conversion Congress and Exposition, pages 3357–3362. IEEE, 2009.

[116] J. Adam Noah et al. Comparison of steps and energy expenditure assessment in adults of Fitbit tracker and Ultra to the Actical and indirect calorimetry. Journal of Medical Engineering & Technology, 37(7):456–462, 2013.

[117] Stavros Ntalampiras. Fault diagnosis for smart grids in pragmatic conditions. IEEE Transactions on Smart Grid, 2016.

[118] Mahamed GH Omran, Ayed Salman, and Andries P Engelbrecht. Dynamic clustering using particle swarm optimization with application in image segmentation. Pattern Analysis and Applications, 8(4):332, 2006.

[119] W. A. Omran et al. A clustering-based method for quantifying the effects of large on-grid PV systems. IEEE Transactions on Power Delivery, 25(4):2617–2625, Oct 2010.

[120] Walid A Omran et al. A clustering-based method for quantifying the effects of large on-grid PV systems. IEEE Transactions on Power Delivery, 25(4):2617–2625, 2010.

[121] Taiwoo Park et al. E-gesture: a collaborative architecture for energy-efficient gesture recognition with hand-worn sensor and mobile devices. In Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems, pages 260–273, 2011.

[122] A. M. Pavan et al. A comparison between BNN and regression polynomial methods for the evaluation of the effect of soiling in large scale photovoltaic plants. Applied Energy, 108:392–401, 2013.

[123] A Massi Pavan and othersA. A comparison between RNN and regression polynomial methods for the evaluation of the effect of soiling in large scale photovoltaic plants. Applied energy, 108:392–401, 2013.

[124] Mile Petkovski, Sofija Bogdanova, and Momcilo Bogdanov. A simple adaptive sampling algorithm. In 14th Telecommunications Forum, pages 329–332, 2006.

[125] R. Platon et al. Online fault detection in PV systems. IEEE Transactions on Sustainable Energy, 6(4):1200–1207, Oct 2015.

[126] Justin P. Porta et al. Validating the Adidas miCoach for estimating pace, distance, and energy expenditure during outdoor over-ground exercise accelerometer. International Journal of Exercise Science: Conference Proceedings, 2(4), 2012.

[127] C. Prakash et al. Identification of spatio-temporal and kinematics parameters for 2-D optical gait analysis system using passive markers. In International Conference on Advances in Computer Engineering and Applications, pages 143–149, 2015.

[128] Quintic Corp, Quintic Bluetooth Smart Family.

[129] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4):3–13, 2000.

[130] Muhammad Mazhar Ullah Rathore, Anand Paul, Awais Ahmad, Bo-Wei Chen, Bormin Huang, and Wen Ji. Real-time big data analytical architecture for remote sensing application. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(10):4610–4621, 2015.

[131] Markus Reichstein, Michael Bahn, Philippe Ciais, Dorothea Frank, Miguel D Mahecha, Sonia I Seneviratne, Jakob Zscheischler, Christian Beer, Nina Buchmann, David C Frank, et al. Climate extremes and the carbon cycle. Nature, 500(7462):287–295, 2013.

[132] V. Reppa, M. M. Polycarpou, and C. G. Panayiotou. Decentralized isolation of multiple sensor faults in large-scale interconnected nonlinear systems. IEEE Transactions on Automatic Control, 60(6):1582–1596, June 2015.

[133] Douglas Reynolds. Gaussian mixture models. Encyclopedia of biometrics, pages 827–832, 2015.

[134] Eunice Ribeiro, Antonio J Marques Cardoso, and Chiara Boccaletti. Fault-tolerant strategy for a photovoltaic DC–DC converter. IEEE Transactions on Power Electronics, 28(6):3008–3018, 2013.

[135] Patrick O. Riley et al. A kinematic and kinetic comparison of overground and treadmill walking in healthy subjects. Gait & Posture, 26(1):17 – 24, 2007.

[136] Daniel Rodríguez-martín et al. A wearable inertial measurement unit for long-term monitoring in the dependency care area. Sensors, (10), 2013.

[137] Volker Roth. Outlier detection with one-class kernel fisher discriminants. In Advances in Neural Information Processing Systems, pages 1169–1176, 2004.

[138] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65, 1987.

[139] Benjamin D Santer, TML Wigley, GA Meehl, MF Wehner, C Mears, Matthias Schabel, FJ Wentz, C Ammann, J Arblaster, T Bettge, et al. Influence of satellite data uncertainties on the detection of externally forced climate change. Science, 300(5623):1280–1284, 2003.

[140] C. Senanayake and S.M.N.A. Senanayake. Human assisted tools for gait analysis and intelligent gait phase detection. In Innovative Technologies in Intelligent Systems and Industrial Applications, pages 230–235, 2009.

[141] Lucia Serrano-Luján et al. Case of study: Photovoltaic faults recognition method based on data mining techniques. Journal of Renewable and Sustainable Energy, 8(4):043506, 2016.

[142] S Mohammad Shahrokhy. Visual and statistical quality assessment and improvement of remotely sensed images. In Proceedings of the 20th Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS'04), pages 1–5. Citeseer, 2004.

[143] S. Silvestre et al. Automatic fault detection in grid connected PV systems. Solar Energy, 94:119–127, 2013.

[144] S. Silvestre et al. Analysis of current and voltage indicators in grid connected PV (photovoltaic) systems working in faulty and partial shading conditions. Energy, 86:42–50, 2015.

[145] Melinda D Smith. An ecological perspective on extreme climatic events: a synthetic definition and framework to guide future research. Journal of Ecology, 99(3):656–663, 2011.

[146] Xiuyao Song, Mingxi Wu, Christopher Jermaine, and Sanjay Ranka. Conditional anomaly detection. IEEE Transactions on Knowledge and Data Engineering, 19(5):631–645, 2007.

[147] M. Sousa et al. Human tracking and identification using a sensitive floor and wearable accelerometers. In IEEE International Conference on Pervasive Computing and Communications, pages 166–171, 2013.

[148] W Sparrow and O Tirosh. Identifying heel contact and toe-off using forceplate thresholds with a range of digital-filter cutoff frequencies. Journal of Applied Biomechanics, 19(2):178–184, 2003.

[149] S. Spataru et al. Diagnostic method for photovoltaic systems based on light I-V measurements. Solar Energy, 119:29–44, 2015.

[150] David J Spiegelhalter et al. The deviance information criterion: 12 years on. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(3):485–493, 2014.

[151] K Steffen, SV Nghiem, R Huff, and G Neumann. The melt anomaly of 2002 on the Greenland Ice Sheet from active and passive microwave satellite observations. Geophysical Research Letters, 31(20), 2004.

[152] M Stein. North Atlantic subpolar gyre warming–impacts on Greenland offshore waters. Journal of Northwest Atlantic Fishery Science, 36:43–54, 2005.

[153] STMicroelectronics. iNEMO inertial module: always-on 3d accelerometer and 3d gyroscope.

[154] C. Strohrmann et al. A data-driven approach to kinematic analysis in running using wearable technology. In 9th International Conference on Wearable and Implantable Body Sensor Networks, pages 118–123, 2012.

[155] C Strohrmann, H Harms, and G. Tröster. Out of the lab and into the woods: kinematic analysis in running using wearable sensors. In Proceedings of the 13th International Conference on Ubiquitous Computing, pages 119–122, 2011.

[156] C. Strohrmann, H. Harms, and G. Tröster. What do sensors know about your running performance? In 15th Annual International Symposium on Wearable Computers, pages 101–104, 2011.

[157] Sharmila Subramaniam, Themis Palpanas, Dimitris Papadopoulos, Vana Kalogeraki, and Dimitrios Gunopulos. Online outlier detection in sensor data using non-parametric models. In Proceedings of the 32nd International Conference on Very large Data Bases, pages 187–198. VLDB Endowment, 2006.

[158] Pei Sun and Sanjay Chawla. On local spatial outliers. In Fourth International Conference on Data Mining, pages 209–216. IEEE, 2004.

[159] Syafaruddin et al. Controlling of artificial neural network for fault diagnosis of photovoltaic array. In 16th International Conference on Intelligent System Applications to Power Systems, pages 1–6, 2011.

[160] Andrew P Tewkesbury, Alexis J Comber, Nicholas J Tate, Alistair Lamb, and Peter F Fisher. A critical synthesis of remotely sensed optical image change detection techniques. Remote Sensing of Environment, 160:1–14, 2015.

[161] Texas Instruments, SimpleLink Bluetooth Smart and Proprietary Wireless MCU.

[162] Owen Vallis, Jordan Hochenbaum, and Arun Kejariwal. A novel technique for long-term anomaly detection in the cloud. In 6th USENIX Workshop on Hot Topics in Cloud Computing, 2014.

[163] DW Van der Merwe and Andries Petrus Engelbrecht. Data clustering using particle swarm optimization. In Evolutionary Computation, 2003. CEC'03. The 2003 Congress on, volume 1, pages 215–220. IEEE, 2003.

[164] S. Vergura et al. Descriptive and inferential statistics for supervising and monitoring the operation of PV plants. IEEE Transactions on Industrial Electronics, 56(11):4456–4464, Nov 2009.

[165] Kiri Wagstaff et al. Constrained k-means clustering with background knowledge. In International Conference on Machine Learning, volume 1, pages 577–584, 2001.

[166] S. Wang et al. A randomized response model for privacy preserving smart metering. IEEE transactions on smart grid, 3(3):1317–1324, 2012.

[167] Yi Wang et al. A framework of energy efficient mobile sensing for automatic user state recognition. In Proceedings of the 7th international conference on Mobile systems, applications, and services, pages 179–192, 2009.

[168] Jaclyn R. Watt et al. A three-dimensional kinematic and kinetic comparison of overground and treadmill walking in healthy elderly subjects. Clinical Biomechanics, 25(5):444 – 449, 2010.

[169] Peter G Weyand et al. Faster top running speeds are achieved with greater ground forces not more rapid leg movements. Journal of applied physiology, 89(5):1991–1999, 2000.

[170] J. Williamson et al. Data sensing and analysis: challenges for wearables. In 20th Asia and South Pacific Design Automation Conference, pages 136–141, 2015.

[171] James Alexander Williamson. Low-Power System Design for Human-Borne Sensing. PhD thesis, University of Colorado at Boulder, 2016.

[172] Liang Xiong, Barnabás Póczos, and Jeff G Schneider. Group anomaly detection using flexible genre models. In Advances in Neural Information Processing Systems, pages 1071–1079, 2011.

[173] Wenyao Xu et al. Smart insole: A wearable system for gait analysis. In Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive Environments, pages 181–184, 2012.

[174] Imene Yahyaoui et al. A practical technique for on-line monitoring of a photovoltaic plant connected to a single-phase grid. Energy Conversion and Management, 132:198–206, 2017.

[175] Zhixian Yan et al. Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In 16th International Symposium on Wearable Computers, pages 17–24, 2012.

[176] Zhehan Yi and Amir Etemadi. Line-to-line fault detection for photovoltaic arrays based on multi-resolution signal decomposition and two-stage support vector machine. IEEE Transactions on Industrial Electronics, 2017.

[177] Zhehan Yi and Amir H Etemadi. Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems. IEEE Transactions on Smart Grid, 8(3):1274–1283, 2017.

[178] Fatma Ben Youssef and Lasaad Sbita. Sensors fault diagnosis and fault tolerant control for grid connected pv system. International Journal of Hydrogen Energy, 42(13):8962–8971, 2017.

[179] J. Yuventi. DC electric arc-flash hazard-risk evaluations for photovoltaic systems. IEEE Transactions on Power Delivery, 29(1):161–167, Feb 2014.

[180] M. Zhang et al. A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering, 26(8):1819–1837, 2014.

[181] Y. Zhao et al. Decision tree-based fault detection and classification in solar photovoltaic arrays. In Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition, pages 93–99, Feb 2012.

[182] Y. Zhao et al. Fault detection, classification and protection in solar photovoltaic arrays. PhD thesis, Northeastern University, 2015.

[183] Y. Zhao et al. Fault prognosis of wind turbine generator using SCADA data. In North American Power Symposium, pages 1–6, 2016.

[184] Ye Zhao et al. Decision tree-based fault detection and classification in solar photovoltaic arrays. In Twenty-Seventh Annual Applied Power Electronics Conference and Exposition, pages 93–99. IEEE, 2012.

[185] Ye Zhao et al. Outlier detection rules for fault detection in solar photovoltaic arrays. In Twenty-Eighth Annual Applied Power Electronics Conference and Exposition, pages 2913–2920. IEEE, 2013.

[186] Ye Zhao et al. Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays. IEEE Transactions on Power Electronics, 30(5):2848–2858, 2015.

[187] Shenggao Zhu, Hugh Anderson, and Ye Wang. Reducing the power consumption of an imu-based gait measurement system. In Advances in Multimedia Information Processing-PCM, volume 7674, pages 105–116. 2012.

[188] Gabriele Zini, Christophe Mangeant, and Jens Merten. Reliability of large-scale grid-connected photovoltaic systems. Renewable Energy, 36(9):2334–2340, 2011.

[189] X. Zou et al. Performance monitoring and test system for grid-connected photovoltaic systems. In Power and Energy Engineering Conference, 2012 Asia-Pacific, pages 1–4. IEEE, 2012.

[190] Jakob Zscheischler, Miguel D Mahecha, Stefan Harmeling, and Markus Reichstein. Detection and attribution of large spatiotemporal extreme events in Earth observation data. Ecological Informatics, 15:66–73, 2013.