UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Graph-based Geospatial Prediction and Clustering for Situation Recognition

Permalink

https://escholarship.org/uc/item/3nh0z3jt

Author

Tang, Mengfan

Publication Date

2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Graph-based Geospatial Prediction and Clustering for Situation Recognition

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Mengfan Tang

Dissertation Committee:
Professor Ramesh Jain, Chair
Professor Gopi Meenakshisundaram
Professor Charless Fowlkes

2017

© 2017 Mengfan Tang

DEDICATION

to my family for their constant support and love.

TABLE OF CONTENTS

Page

LIST OF FIGURES vi

LIST OF TABLES viii

LIST OF ALGORITHMS ix

ACKNOWLEDGMENTS x

CURRICULUM VITAE xi

ABSTRACT OF THE DISSERTATION xiii

1 Introduction 1
- 1.1 Situation Recognition in Geospatial Multimedia Data . . . 2
- 1.2 Contributions . . . 6
- 1.3 Dissertation Overview . . . 7

2 Geospatial Heterogeneous Data Prediction 12
- 2.1 Introduction . . . 12
- 2.2 Related Work . . . 16
  - 2.2.1 Spatial System Tools . . . 17
  - 2.2.2 EventShop . . . 17
  - 2.2.3 Spatial Prediction Models . . . 18
  - 2.2.4 PM2.5 prediction . . . 20
- 2.3 Framework . . . 20
  - 2.3.1 Problem Definition . . . 20
  - 2.3.2 Spatial Gaussian Process Regression . . . 21
  - 2.3.3 Spectral Features . . . 23
- 2.4 Experiment . . . 26
  - 2.4.1 Data and Performance Metrics . . . 26
  - 2.4.2 Details and Setup . . . 28
  - 2.4.3 Performance Comparison with A Physical Chemical Model . . . 29
  - 2.4.4 Comparison with State-of-the-art Approaches . . . 29
  - 2.4.5 Point-based Performance Analysis . . . 32
- 2.5 Situation Recognition on EventShop . . . 33

2.5.1 Asthma Risk Zone Study . . . 33

3 Asthma Risk Factors From Personal Data and Environmental Data 35
3.1 Framework . . . 36
3.2 Environmental Data and Personal Data . . . 38
3.3 Feature and Data Analysis . . . 40
3.3.1 Personal Life Events Features: F_P . . . 40
3.3.2 Emission Factors Features: F_E . . . 42
3.3.3 Air Pollution Features: F_A . . . 43
3.4 Models . . . 44
3.5 Results . . . 45
3.6 Discussion . . . 46

4 A Graph-based Spatial Prediction Model 48
4.1 Motivation . . . 48
4.2 Background of Graph-based Methods . . . 49
4.2.1 Graph . . . 50
4.2.2 Graph Random Walk . . . 50
4.2.3 Graph Laplacian . . . 51
4.3 Parameters Optimized Graph Regression . . . 52
4.4 Experiments . . . 55
4.4.1 Effectiveness of Features . . . 56
4.4.2 Parameter Analysis . . . 58
4.5 Discussion . . . 58

5 A Noise Resistant Graph Embedding for Photo Clustering 60
5.1 From Photo Clustering to Situation Recognition . . . 60
5.2 Photo Clustering with Outliers . . . 61
5.3 Capped Norm Graph Embedding . . . 63
5.3.1 Problem Formulation and Motivation . . . 64
5.3.2 The Graph-based Capped Norm Method . . . 65
5.3.3 Optimization Algorithm . . . 66
5.4 Experiments . . . 68
5.4.1 Data Description . . . 69
5.4.2 Experimental Settings . . . 69
5.4.3 Results . . . 70
5.5 Discussion . . . 73

6 A Graph-regularized PCA for Out-of-sample Photo Clustering 74
6.1 Introduction . . . 74
6.2 Related Models . . . 76
6.3 The Proposed Model . . . 78
6.3.1 Algorithm Convergence Analysis . . . 80
6.3.2 Connection to LDA . . . 81
6.4 Experiments . . . 82

6.4.1 Dataset Description . . . 82
6.4.2 Experimental Settings . . . 83
6.4.3 Convergence Study of the Proposed Method . . . 85
6.4.4 Results . . . 87
6.4.5 Computational Time Analysis . . . 88
6.5 Discussion . . . 89

7 From Photo Streams to Evolving Situations **90**
7.1 Examples of Evolving Situations . . . 92
7.2 The Proposed Framework . . . 94
7.2.1 Problem Formulation and Motivation . . . 94
7.2.2 Semi-supervised Learning Method for Unknown Labels . . . 96
7.2.3 Noise Robust Semi-Supervised Learning Methods for Unknown Labels 97
7.2.4 Optimization Algorithm . . . 99
7.3 Experiments . . . 102
7.3.1 Data Preprocessing . . . 102
7.3.2 Semi-supervised Clustering with Unknown Labels . . . 103
7.3.3 Evolving Situation Detection . . . 104
7.4 Discussion . . . 105

8 Conclusion and Future Research **106**
8.1 Conclusion . . . 106
8.2 Future Research . . . 109

Bibliography **111**

LIST OF FIGURES

Page

1.1 The high level architecture of Social Life Networks (SLN). 3

2.1 The locations of air pollution sensory stations 13
2.2 Four heterogeneous data sources, (a) US EPA Air Quality System, (b) Community Modeling & Analysis System (CMAQ), (c) MODIS Aerosol Optical Depth and (d) Traffic Flow, for PM2.5 concentration analysis over California on Feb 12, 2008, which demonstrates different spatial granularity ranging from points to grids. 14
2.3 The framework of multiresolution prediction 21
2.4 Prediction results using Spatial Gaussian Process on all data sources. 28
2.5 Observed vs prediction values on testing samples 31
2.6 Asthma risk situations across California detected by EventShop 34

3.1 Expected county level emissions from agriculture, road dust, and mining from CHAD 38
3.2 Average maximum, mean, and minimum PM2.5 concentration values for the month of June for counties from CHAD 39
3.3 25 randomly chosen diary entries from CHAD. Each coloured row on the y-axis represents a day's activities for one person. The color marks the activities, and the x-axis shows time in 15 minute increments 40
3.4 Relative importance of features as determined by a gradient boosting tree . . 45

4.1 Values of various data sources across United States, plotted on $12km \times 12km$ grid 56
4.2 Effects of k on graph-based methods with different features. 58

5.1 Comparison of capped ℓ_p-norm and ℓ_1-norm 62
5.2 Examples of outliers in the YFCC data 71

6.1 Flowchart of the proposed method 79
6.2 Representative images of MNIST, ORL, and Yale data 83
6.3 Convergence analysis on five data sets 85

7.1 Evolving concepts in Beijing in the year 2008 93
7.2 Semi-supervised learning framework for evolving situation recognition under the condition of unknown-labeled data 95

7.3 The evolving situation of “Hanami” 104
7.4 The situation of “Holi” in space and time 104

LIST OF TABLES

Page

2.1 Data sources for PM2.5 concentration analysis with diverse spatial and temporal granularities . . . 15
2.2 Geospatial prediction models evaluation . . . 30

3.1 Extracted personal profile features . . . 41
3.2 Emission factors . . . 42
3.3 Classifier performance against feature set . . . 45

4.1 Performance and standard deviation of CMAQ, Ordinary Kriging, Linear Regression(LR), Spatial Gaussian Process (SGP), and the proposed model (POGR) on PM2.5 prediction for 20 runs. . . . 57

5.1 Data Descriptions . . . 70
5.2 Clustering accuracy using K-means, Rcut [16], Ncut [101], L1G [80] and the proposed model CGE on five data sets. . . . 71
5.3 Clustering NMI K-means, Rcut [16], Ncut [101], L1G [80] and the proposed model CGE on the five data sets. . . . 72

6.1 Datasets Descriptions . . . 83
6.2 Clustering average accuracy and standard deviation using K-means on PCA, LPP, FME, and grlrPCA on seven datasets for 20 runs . . . 86
6.3 Clustering average normalized mutual information and standard deviation using K-means on PCA, LPP, FME, and grlrPCA on seven datasets for 20 runs 86
6.4 Clustering average accuracy and normalized mutual information using K-means on glPCA, MMF, and grlrPCA on training datasets for 20 runs . . . 87
6.5 Computational time analysis for grlrPCA, FME, and glPCA on seven datasets in seconds. . . . 88

7.1 Accuracy (in percentage) for label-known data and label-unknown data: GSS is Semi-supervised Learning Method for Unknown Labels Model (Eq. (7.1)), SSL is ℓ_1-norm Semi-supervised Learning Method (Eq. (7.2)), SSC is ℓ_p-norm Semi-supervised Learning Method (Eq. (7.3)) . . . 103

LIST OF ALGORITHMS

		Page
1	Spectral Spatial Gaussian Process Prediction	25
2	Multiresolution Graph-based Prediction	54
3	A General Re-weighted Optimization Algorithm	66
4	Graph-based Capped Norm Optimization	68
5	Graph-regularized PCA Iterative Optimization Method	80
6	Semi-supervised ℓ_1-norm Method	100
7	Semi-supervised ℓ_p-norm Method	102

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Ramesh Jain, for supporting me during my Ph.D. life. Ramesh is someone you will instantly love and want to meet soon. He is one of the most energetic and smartest people I know. I hope I could be lively and enthusiastic as Ramesh. He has been supportive and has given me the freedom to pursue my research interests without objection. He has also provided insightful discussions about the research. I am greatly indebted to Ramesh for all that I learned from both academics and beyond. I am very grateful to have him as my academic father.

I would like to thank my committee members, Professor Gopi Meenakshisundaram and Professor Charless Fowlkes, for their constructive comments and helpful career advice.

I would like to thank Professor Sandy Irani. I worked as her teaching assistant for nine quarters. She helped me go through difficult times in finding financial support.

I would like to thank Prof. Feiping Nie. This study also benefited from conversations with him.

I would like to acknowledge support of Donald Bren Foundation through their funding of Prof. Jain's in my last one and half years at UCI.

I would like to thank my lab-mates, Siripen Pongpaichet, Pranav Agrawal, Laleh Jalali, Jordan Oh, and Nitish Nag for their collaborations. I would like to also thank my fellow Ph.D. students, Wei Ping, Yi Li, Yuxiao Wang, and Wenshan Wang for their helpful discussions.

I would like to thank Zhaoqiang Bai, Qiguang and Qin Wang, Zhenqiu Huang, Meng Chen for the sharing of their Ph.D. life, which encourages me to progress. My special thanks go to Timothy and Jean Chen for their consistent shepherding. I would like to also thank Joseph and Doris Shie for being supportive throughout my time here and for preparing countless tasty love feasts.

Last, I would like to thank my family for their unchanging love and constant support, without which I would not have completed the Ph.D. study.

CURRICULUM VITAE

Mengfan Tang

EDUCATION

Doctor of Philosophy in Computer Science **2017**
University of California, Irvine *Irvine, California*

Master of Science in Computer Science **2017**
University of California, Irvine *Irvine, California*

Bachelor of Science in Mathematics **2010**
University of Science and Technology of China *Hefei, China*

RESEARCH EXPERIENCE

Graduate Research Assistant **2016–2017**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2012–2016**
University of California,Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Integration of Diverse Data Sources for Spatial PM2.5 Data Interpolation **2017**
IEEE Transacations on Multimedia

A Graph Regularized Dimension Reduction Method for Out-of-sample Data **2017**
Neurocomputing

Low-Rank Regularization for Learning Gene Expression Programs **2013**
PLOS ONE

REFEREED CONFERENCE PUBLICATIONS

From Photo Streams to Evolving Situations **July 2017**
4th IEEE International Workshop on Mobile Multimedia Computing

Geospatial Multimedia Data for Situation Recognition **Oct. 2016**
The 2016 ACM Multimedia Conference

Research Challenges in Developing Multimedia Systems for Managing Emergency Situations **Oct. 2016**
The 2016 ACM Multimedia Conference

Capped ℓ_p-Norm Graph Embedding for Photo Clustering **Oct. 2016**
The 2016 ACM Multimedia Conference

Using Photos as Micro-Reports of Events **Jun. 2016**
The 2016 ACM International Conference on Multimedia Retrieval

A Graph Based Multimodal Geospatial Interpolation Framework **Jul. 2016**
The 2016 IEEE International Conference on Multimedia & Expo

Geospatial Interpolation Analytics for Data Streams in EventShop **Jul. 2015**
The 2015 IEEE International Conference on Multimedia & Expo

Habits vs Environment: What Really Causes Asthma? **Jun. 2015**
The 7th International ACM Web Science Conference 2015

Personicle: Personal Chronicle of Life Events **Sep. 2014**
VLDB 2014 Workshop on Personal Data Analytics in the Internet of Things

Observing Real-World Phenomena through EventWeb over Space, Time and Theme **Jun. 2014**
The 2st International workshop on Building Web Observatories

ABSTRACT OF THE DISSERTATION

Graph-based Geospatial Prediction and Clustering for Situation Recognition

By

Mengfan Tang

Doctor of Philosophy in Computer Science

University of California, Irvine, 2017

Professor Ramesh Jain, Chair

Big data continues to grow and diversify at an increasing pace. To understand constantly evolving situations, data is collected from various location-based sensors as well as people using effective participatory sensing. Static sensors are placed at particular locations, monitoring and measuring important variables from the environment. Additionally, people contribute data in the form of mobile streams through participatory sensing. To process such disparate data for situation recognition, we need to address two major research challenges, namely data quality and data analysis.

This dissertation presents graph-based models as a general framework to combine diverse data sources for situation recognition. To improve spatial data quality, we propose a multimodal geospatial prediction model that integrates heterogeneous data into a unified space-time-value format, extracts multi-resolution spectral features, and predicts values at unobserved locations. We demonstrate the effectiveness of the proposed model using air pollution data. We then combine predicted air pollution data with pollen data to analyze the situation of asthma risk across California. The proposed model based on our predicted values provides greater accuracy in detecting high asthma risk areas compared to previous studies.

Given the wide use of mobile technology, humans also act as sensors and increasingly contribute photos as mobile data streams. We investigate the use of visual concepts for situation

recognition and develop new models to tackle the technical challenges of noisy data and processing real-time data, which are problems in existing approaches. First, we propose a graph-regularized linear regression PCA to address photo clustering in a real-time setting. Second, to overcome the noisy data that degrades graph quality, leading to poor recognition results, we incorporate a capped norm into a graph embedding method to remove the adverse effects of outliers. Our models not only outperform existing approaches in overcoming the problems of noisy data and real-time data processing, but also allow us to utilize a new cluster of data through a soft label method. Finally, we apply the proposed approaches to Yahoo Flicker 100 Million photos for detecting evolving situations.

Chapter 1

Introduction

We are in an era of big data, which is fundamentally different from the data we had a decade ago. Big data is mostly heterogeneous data streams and collected by sensors with space and time information. In addition, people are becoming sensors as well as actuators, namely in the wide spread use of mobile technology. However, the quality of data needs to be determined and considered in its use. Much of the new data collected to capture the dynamic nature of the world utilizes disparate sensors that usually have different spatial resolutions. Integrating this data requires multimedia techniques. Using cybernetic principles combined with multimedia technology, we can develop effective frameworks for using diverse multimedia data for situation recognition, and determining and communicating appropriate actionable guidance to people.

In the study of computer science and social sciences, the term "situation" is defined as the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future [26]. Situations in the physical world are the result of interactions among several related events. Events can be observed by measuring physical attributes using sensors. To understand constantly

evolving situations, data may be collected using disparate sensors, social media, cyberspace interactions, Internet-of-Things (IoT), as well as from people using effective participatory sensing. Using these precise observations, we can model and understand the world around us in order to better control adverse situations or cultivate a more optimal situation. Many emerging applications increasingly rely on assimilating complex heterogeneous data streams or big data.

1.1 Situation Recognition in Geospatial Multimedia Data

In order to connect people's needs to resources effectively, efficiently, and promptly, Social Life Networks (SLN) [50] has been developed. SLN extends the power of current social networks by dynamically monitoring and detecting existing situations, surrounding contexts, and their combinations. With these situations, SLN computes and matches an individual's needs to resources. The basic structure of SLN is illustrated in Fig. 1.1. The global situation is first detected from heterogeneous environmental data sources. The individual situation is then recognized from various types of personal data sources. Finally, an individual's needs are identified and the individual is connected to resources.

The quality of human life can be enhanced by connecting people's needs to resources. This is a fundamental but very challenging task. Therefore, SLN faces various challenges ranging from situation recognition, sensor fusion, multi-modality data integration, predictive analysis, to recommendation [132, 88, 38]. For example, by observing the symptoms, physiological states, and activities of asthma patients, a health persona can be created through wearable/portable sensors and self-reported information on social media. With this persona and knowledge about the patients' locations and contextual information, such as air quality, weather, and pollen counts, high asthma risk areas are detected and sent to asthma patients. Asthma patients can take action to avoid the high risk areas and thus asthma attacks can

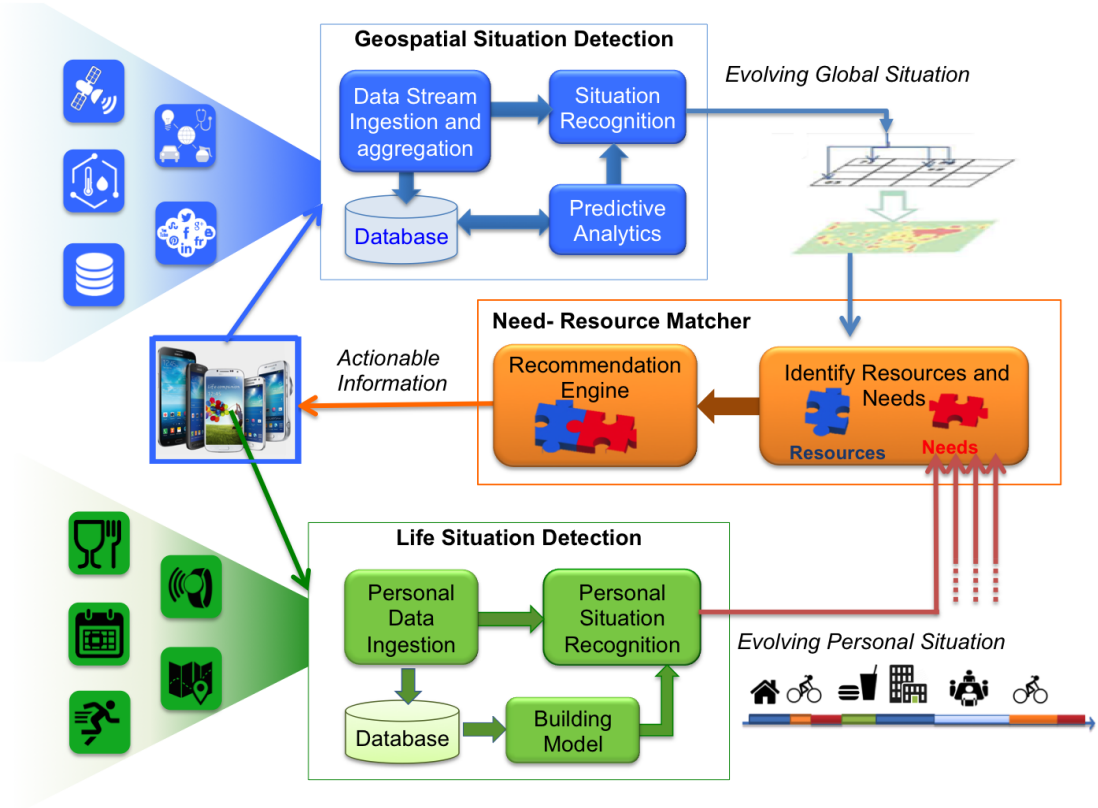


Figure 1.1: The high level architecture of Social Life Networks (SLN).

be effectively prevented [89, 91].

There have been several studies related to SLN's goal of using data for connecting people to resources. Plenario [1] is designed to handle smart open data. It allows users to assemble and download data from multiple, independent data sources and unite all datasets along a single spatial and temporal index for complex aggregations with one query. London Traffic System [75] is designed as a complex event processing system for smart city applications. It includes weather, public events, and traffic condition data to discover the influence of social events on traffic. When the users perform experiments using this system, they realize there is a missing data problem, that is, 27% of the data is missing. However, these systems do not have any component to improve the quality of the sensory data.

Besides sensor data, Twitter data has become one of most widely used data sources for

[1] http://plenar.io/

social and journalistic applications in the last decade. A great number of research papers have shown how tweets can be used for important social analysis. Among them, the project JeffCo IMT is built to analyze flood-related data and provides insight into how to meet the information needs of a community [104]. In [121], visual reports from physical surveillance cameras and social tweets are combined to detect evolving real-world events. In the study of [103], air pollution data and pollen data are combined with Twitter data in order to study the situation of asthma risk. The accuracy of the situation detection depends on the quality of data. Despite considerable success, there are three most critical problems in using this type of social media to understand situations: missing/misleading context, noise, and subjectivity [5, 23, 62, 99].

Therefore, in SLN, to understand constantly evolving situations, data is collected from various sensors and people using effective participatory sensing. These diverse data sources are divided into two categories: **static sensor data** and **citizen-contributed data**. First, static sensors are relatively fixed at some locations, monitoring and measuring important variables from the environment. For example, air quality sensors are placed to collect meteorological data and air pollution data commonly used for weather prediction and air pollution situation detection. Environment dynamics can be analyzed using this data. Second, citizen-contributed data comes as mobile streams. People are mobile and use their phones and other devices for contributing to this data. The synergy of data in both categories gives powerful insight into the understanding of evolving situations, in which participatory sensing is integrated with the surrounding environment.

Static Sensor Data

In the real world, environmental sensors are often spatially sparse, preferentially located and often without complete measurements in the targeted domain [25, 134, 58]. Reliance on these

measurements restricts the geographical region of situation recognition [86]. Consequently, there is demand for spatial prediction with additional information to increase the availability of this data across space [69, 96, 70, 12, 47]. For example, one data stream is provided as grid cells, while other data streams are only available at point locations. It is exceptionally challenging to combine data sources that arise from different spatial resolutions for accurate predictions [15].

Mobile Data

Mobile data has greatly increased, given the rapid development of wireless communication and mobile computing technologies and navigational systems. This data calls for new computing technologies for storage, preprocessing, retrieving, mining, and exploring its broad applications. Travelers make records of their travel routes with GPS for memorizing and sharing experiences. Bicyclers, joggers, and hikers record their trails for activity analysis. Foodies post their favorite food using check-ins in Four-square and Instagram. People are mobile and actively recording space and time information, as well as the activities they are engaged in. People might not realize that they are contributing to mobile data recording because almost everyone uses a mobile phone and generates many spatial trajectories. Such massive mobile data creates new possibilities to analyze and understand various aspects of human behavior and experiences, ranging from personal entertainment preferences to high risk environments. Understanding such behavior and experiences will help realize the goal of connecting people to resources, effectively, efficiently, and promptly.

Citizen-contributed Photos

Among this diverse mobile data, the citizen-contributed photos is one of the biggest data sources. Every minute, millions of photos are uploaded by people to social media. There has

been a great number of studies on processing and using photos for various applications. One of the major efforts in recent years is automatic semantic classification of media contents into a number of predefined concepts [17]. These concepts and their relations are defined in an ontology from the results of user studies or data mining techniques applied to detect events. For example, photos of music festivals may contain a large number of objects like people and instruments, actions like singing as well as scenes like stages and posters. Event detection has achieved considerable advancement in recent years. However, detection and understanding of a complex multimedia situation has not been explored. Various deep learning frameworks have been developed in video and image analysis, such as TensorFlow [1], Caffe [54], Theano[6], Mxnet [19], and Cuda-convnet [63] have been available as open source. These concept detector frameworks provide visual concepts for informative and objective micro-reports along with their location and time. Given the accuracy of concept detections, can we convert photos to powerful data streams for situation recognition?

We use air pollution, asthma risk factors, and multimedia micro-reports as examples for situation recognition because of their common occurrence and important effects on human lives. Most issues discussed for air pollution are easily generalizable and adaptable to other environmental situations. The framework to deal with asthma risk factors can be extended to other health care applications.

1.2 Contributions

The above challenges motivate this dissertation. The following specific issues are addressed to deal directly with data quality and data analysis of both static sensor data and citizen-contributed photos.

- Observations from static sensors are typically sparse in space. Reliance on these sparse

sensory measurements restricts the geographical region of situation recognition. We propose statistical spatial prediction approaches to increase the availability of static sensor data across space. We introduce spectral features to further improve prediction accuracy.

- To effectively use multiple data sources for situation recognition, we propose graph-based methods as a principled framework and derive several graph-based models for static data prediction and citizen-contributed photo clustering. First, a parameter optimized graph regression model is presented to reduce the complexity of computation in spatial prediction. Second, noise resistant graph embedding methods are introduced for both unsupervised and semi-supervised learning tasks.
- Photographs are considered as objective and informative micro-reports from citizen sensors. Using advances in deep learning, we derive accurate visual concepts from photos. Combining space and time information of the photos, we formulate event and trend detection problems into photo clustering tasks by introducing "bag of concepts." We then apply the developed graph-based models to explore photos as a powerful information stream for detecting evolving situations.
- Traditional semi-supervised learning methods are not able to detect unknown situations in out-of-sample data, which poses difficulties in directly applying these methods to out-of-sample data for situation recognition. To address the problem, we introduce a soft label method that enables the traditional semi-supervised learning framework to accurately predict predefined labels as well as effectively form new clusters.

1.3 Dissertation Overview

This dissertation addresses various technique research challenges in static sensor data and user-contributed photos for situation recognition. From Chapter 2 to Chapter 7, we discuss

our approaches in detail. We then conclude this dissertation in Chapter 8.

In Chapter 2, we present a geospatial prediction framework. Previous frameworks have confronted significant data problem, that is, the geographical regions of situation recognition are limited by environmental data which is collected by environmental sensors. These sensors are usually sparsely and preferentially located. Spatial prediction is an effective way to solve the problem of data sparsity, and demands the availability of related data sources. However, these data sources are usually in different resolutions, distributions, scales, and densities, which poses a major challenge for data integration. To address this problem, we present a novel spatial interpolation framework to incorporate diverse data sources and explicitly model the spatial processes at multiple resolutions. Spectral analysis is deployed to generate features at multiple spatial resolutions and to improve the interpolation accuracy at unobserved locations. A statistical operator based on the spatial Gaussian Process is implemented and integrated into a geospatial situation recognition system, which can analyze heterogeneous spatio-temporal data streams derived from sensors. First, to verify the effectiveness and efficiency of the proposed framework, this framework is applied to the PM2.5 data. Our experiments conducted in California demonstrate that the proposed method outperforms state-of-the-art approaches. Second, we then use the PM2.5 data that predicted by the proposed method, combine with pollen count, in order to detect situations of asthma in California. This chapter is a revision of [110], [115], and [107].

Following the asthma situation recognition study in Chapter 2, in Chapter 3 we explore personal asthma risk factors using both environmental data and profile data. Asthma is a common and potentially life threatening respiratory disorder disease. Since there is no clear understanding of asthma causes and no definite cure, it has become increasingly the topic in health care studies. Existing studies are often limited to either profile data or environmental data. We present a data-based approach to estimate personal asthma risk using profile data, personal life events data, and environment data, and thus to extract personal and

environmental features that determine a person's susceptibility to asthma. Features derived from our analysis agree with the current general consensus in the medical community about factors affecting asthma. This chapter was originally published in [108].

In Chapter 4, in order to reduce the computational complexity compared with the spatial Gaussian Process models used in Chapter 2, we propose a novel spatial prediction model, capable of incorporating multimodal data sources and comprehensively modeling the spatial processes at multiple resolutions. The framework transforms the spatial interpolation problem into a graph structure learning problem, based on the latent structure of the data. The graph-based model integrates hyper-parameter learning into graph structure learning. This enables more efficient and accurate predictions at unobserved locations. We demonstrate the effectiveness of our approach by testing it on air pollution interpolation. This chapter was originally published in [109].

We have discussed using data from static physical sensors for situation recognition. Additionally, given the wide use of mobile technology, humans also act as sensors in that people themselves, in carrying and using smartphones, can contribute their measurements or observations for situation recognition. For example, every minute, millions of photos are uploaded by people to social media. These photos and their context are a natural source for situation recognition. Starting with Chapter 5, we will explore the possibility of using geospatial streams of photos for situation recognition. Situations are defined as a bag of photo visual concepts in space and time. By computing the similarity between the situation model with an evolving bag of visual concepts, similar situations can be detected. The situation detection problem is then transformed to a photo clustering problem. To effectively transform citizen contributed photos into evolving situations, we must deal with noisy and real-time photos.

In Chapter 5, We present several approaches to deal with outliers in photos. We design a capped ℓ_p-norm graph-based model to reduce the impact of outliers. This is accomplished by

allowing the data graph to self adjust as part of the graph embedding. Furthermore, we derive an iterative algorithm to solve the objective function optimization problem. Experiments on four real-world benchmark datasets and the Yahoo Flickr Creative Commons dataset show the effectiveness of this new graph-based capped ℓ_p-norm clustering method. This chapter was originally published in [111].

In Chapter 6 we introduce a graph-based method to handle real-time data streams. The proposed graph-regularized Principal Component Analysis (PCA) that extends traditional graph-regularized dimension reduction methods for out-of-sample data. Among various dimension reduction techniques, PCA is used for treating vector data, whereas Laplacian embedding is often employed for embedding graph data. Moreover, graph regularized PCA, a combination of both techniques, has also been developed to assist the learning of a low dimensional representation of vector data by incorporating graph data. However, these approaches are confronted by the out-of-sample data problem: each time when new data is added, it has to be combined with the old data before being fed into the algorithm to re-compute the eigenvectors, leading to enormous computational cost. In order to address this problem, we extend the graph regularized PCA to the graph-regularized linear regression PCA (grlrPCA). grlrPCA eliminates the redundant calculation on the old data by first learning a linear function and then directly applying it to the new data for its dimension reduction. Furthermore, we derive an efficient iterative algorithm to solve the grlrPCA optimization problem and show the close relatedness of grlrPCA and unsupervised Linear Discriminant Analysis at an infinite regularization parameter limit. The experiments using multiple metrics on seven realistic datasets demonstrate that grlrPCA outperforms established unsupervised dimension reduction algorithms and grlrPCA is an effective method for dealing with real-time data streams. This chapter was originally published in [112].

In Chapter 7, we extend graph-based methods introduced in Chapter 5 and Chapter 6 to explore evolving situation recognition. We use visual concepts in photos together with space

and time information for situation detection. A semi-supervised learning framework and graph-based models are proposed to solve the problem. To extend the method for unknown situations, we introduce a soft label method that enables the traditional semi-supervised learning framework to accurately predict predefined labels as well as effectively form new clusters. To overcome the noisy data that degrades graph quality, leading to poor recognition results, we take advantage of two of the noise-robust norms introduced in Chapter 5, which can eliminate the adverse effects of outliers in visual concepts and improve the accuracy of situation recognition. This chapter is a revision of [113].

In Chapter 8, we summarize motivations and contributions of this dissertation study. We also discuss several challenges for future research.

Chapter 2

Geospatial Heterogeneous Data Prediction

2.1 Introduction

Big data arises from continuous monitoring and observations in the world around us. Understanding the data allows us to predict and control dynamic situations with greater dexterity. There are tremendous opportunities in big data to help society and science. Many emerging problems increasingly rely on assimilating complex heterogeneous data streams. In this chapter, we address an important problem rising in integrating such heterogeneous data, namely data sparsity problem.

In order to detect situations in our daily environment, static sensors are deployed to continuously monitor and measure certain environmental variables at relatively fixed locations. For air pollution monitoring, air quality sensors are placed to collect meteorological and air pollution data, which is commonly used for weather predictions and air pollution measurements. Environmental dynamics can be analyzed and used to predict emergent situations, and then

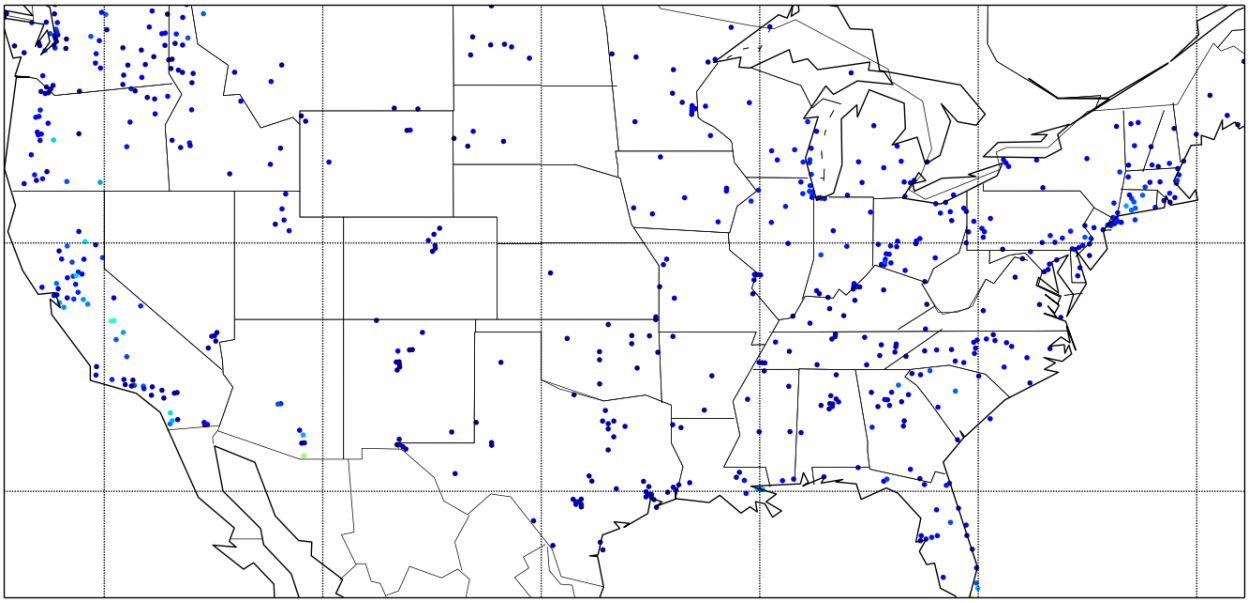
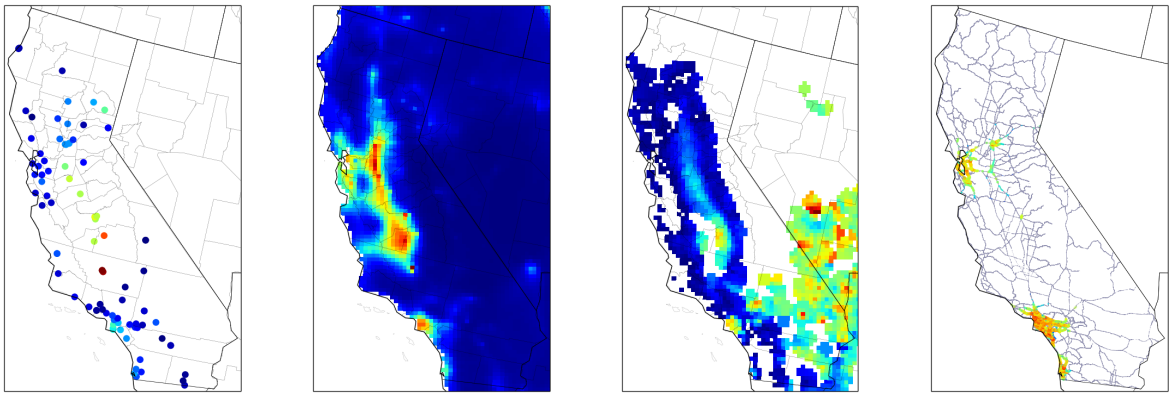


Figure 2.1: The locations of air pollution sensory stations

appropriate actions can be taken. Unfortunately, data generated from static sensors usually have significant problems, such as spatial sparsity or missing measurements [13, 110]. Figure 2.1 depicts the problem of spatial sparsity given there are vast swathes of territory without static sensors nearby. There are 738 monitoring stations operating in the year 2015 [27]. Around 78% of the counties in the United States are not covered by the monitoring stations. Reliance on spatially sparse data results in geographical restrictions for applications where data cannot be obtained. In general, spatial data prediction is used to predict or estimate values at unobserved locations from a limited amount of observed location data. Data from a single source is possibly insufficient for predicting values at unobserved locations due to its limited capability in characterizing various spatial properties. Consequently, spatial prediction requires auxiliary information as a proxy to increase the availability of data that cannot be directly measured.

In this chapter, we use air pollution spatial prediction as an example to illustrate how the data and our proposed framework contribute to connect people to their needs in the given situations of air pollutions. Environment and air pollution have become serious issues for



(a) (b) (c) (d)

Figure 2.2: Four heterogeneous data sources, (a) US EPA Air Quality System, (b) Community Modeling & Analysis System (CMAQ), (c) MODIS Aerosol Optical Depth and (d) Traffic Flow, for PM2.5 concentration analysis over California on Feb 12, 2008, which demonstrates different spatial granularity ranging from points to grids.

human health and public welfare in recent years. In particular, PM2.5, which is suspended particulate matter smaller than 2.5 μm in aerodynamic diameter, is the most well-known particle pollutant and has become a world-wide problem, especially in England, China, India, and the United States. Some cities in China and India have PM2.5 concentrations above 800 $\mu g/m^3$. In the United States, the Environmental Protection Agency (EPA) classifies levels above 150 $\mu g/m^3$ as very unhealthy. Such high levels can cause harmful cardiovascular problems and even premature death. Epidemiological studies have shown that prolonged exposure to PM2.5 results in adverse respiratory health responses. PM2.5 also has a strong association with most types of respiratory illnesses, causing a wide range of diseases including lung cancer, asthma, and strokes. In China alone, more than 1.6 million people die each year due to air pollution [97]. In the United States, around one out of fifteen Americans suffer from asthma. Obtaining accurate PM2.5 estimates is an important problem to solve. However, measurement sensors are typically sparsely located. Figure 2.1 shows the locations of air pollution monitoring stations in the United States.

Table 2.1: Data sources for PM2.5 concentration analysis with diverse spatial and temporal granularities

Theme	PM2.5 value	PM2.5 Estimation	Aerosol Concentration	Traffic Flow
Source	US EPA AQS[1]	CMAQ[2]	MODIS AOD[3]	Caltrans PeMS[4]
Temporal	hourly	daily	aperiodic	5 mins
Spatial	point	10 km × 10 km grid	12 km × 12 km grid	isolated point

For PM2.5 studies, various data sources have been commonly used for simulating and monitoring air pollution dynamics, including CMAQ[2], MODIS AOD[3], and traffic flow[4], which are listed in Table 2.1. CMAQ is an air quality physical model which is used to simulate the distribution of PM2.5. MODIS AOD is a product that captures satellite images of ambient aerosol optical thickness, reflecting the levels of particle matter in the air. Traffic flow measures speed and congestion in a spatial area, which can help infer the emission levels of PM2.5. These three data sources are correlated with PM2.5 and thus can be treated as auxiliary information for PM2.5 prediction at locations that are not covered by monitoring stations. However, this auxiliary information usually comes from different data sources that are diverse in terms of granularity, coverage, availability, and reliability. The spatial and temporal granularities of four observations related to the concentration of PM2.5 are also listed in Table 2.1. Temporal resolution ranges from minutes to days. Spatial resolution and coverage also range from points to grids. Observations from the U.S. EPA are collected from point sensors. The outputs from CMAQ and MODIS AOD are in grid formats. Traffic flow is a point measurement along a line. As shown in figures in Table 2.1, the coverage (presented in color) of each data source is different, and from this we can see the overlapped regions are inconsistent. Effective utilization of these diverse data can help better infer PM2.5 values at unobserved locations. The challenges lie in the effective integration and modeling of heterogeneous data of various forms, such as points, grids and lines. Existing frameworks (e.g., [134]) often combine spatial data into a fixed grid resolution. This grid resolution is

[1]US EPA Air Quality System: http://www.epa.gov/airquality/airdata/
[2]CMAQ (The Community Multi-scale Air Quality modeling system): https://www.cmascenter.org/
[3]MODIS Aerosol Optical Depth: http://modis-atmos.gsfc.nasa.gov/
[4]CalTrans PeMS: http://pems.dot.ca.gov/

selected without any application-specific justification. However, using such a setting usually causes unpredictable errors. This motivates us to propose a scheme of multiple resolution data fusion.

To handle diverse data from static sensors, we propose a novel multi-resolution learning framework for geospatial data prediction. Given a set of heterogeneous data across different domains, this framework seamlessly integrates these correlated data into spatially unified grids, and efficiently extracts spectral features at multiple resolutions, effectively associates observations with auxiliary features, and explicitly models spatial process at feature levels in a joint manner. The prediction task is formulated into a regression problem where observations are predicted by auxiliary data. A spatial Gaussian Process is proposed to solve the regression problem by incorporating heterogeneous features at multiple resolutions. The framework has been successfully applied to the estimation of PM2.5 concentrations. Experiments conducted on four air pollution related data sources from California have demonstrated that the proposed approach outperforms state-of-the-art methods and significantly improves the accuracy of geospatial prediction.

The remainder of this chapter is organized as follows: we discuss related work in Section 2 and explain predictive methods as a component in a geospatial situation recognition system in Section 3.

2.2 Related Work

This study is closely related to multiple research fields, ranging from spatial statistics, multimedia, and data mining. We group related studies into three categories and survey the literature in detail.

2.2.1 Spatial System Tools

Geographic Information Systems (GIS) have been successfully leading the research related to all types of spatial or geographical data. Abundant tools, both open-source and commercial, are available to manage, store, manipulate, analyze, and visualize geospatial data. Most GIS applications store their data in vector form, in which each element corresponds to an observation [77]. This makes elementary operations such as union, intersection and pixel-pixel computations between two vectors with different support systems fast and scalable. However, their abilities are limited when used for advanced problems like prediction and prediction via sophisticated models. Existing solutions to these problems are often ad-hoc and very scenario-specific, and therefore, not optimally generalizable. Additionally, in GIS, many spatial prediction methods have been developed to predict unknown values at unsampled locations from the known values of nearby sampled points. Their formulations are typically based on geo-statistical concepts, smoothness (splines) assumptions, and ad-hoc functions. Different methods can produce quite different predicted results. As such, a comprehensive understanding of the phenomenon is required when searching for suitable prediction methods for GIS applications [73]. These approaches are also restricted in the types of data they can process.

2.2.2 EventShop

To handle diverse data from static sensors, an open source project EventShop[1] has been developed [89]. It provides a generic infrastructure for domain experts or application developers to experiment with heterogeneous spatio-temporal data streams, aiming to recognize situations and define actions such as in sending out personalized alerts. Situations on EventShop can vary from daily/routine situations, such as a traffic jam to unusual situations, such as

[1]Source code can be found at: github.com/spongpai/eventshop_beta

an Ebola outbreak. Data can be streamed from diverse sets of sources, such as social media, sensor networks, distributed cameras, and knowledge bases like censuses and maps.

EventShop indexes and integrates all spatio-temporal heterogeneous data streams into a common representation called an E-mage. An E-mage is a collection of grids, where each cell captures a value associated with a particular theme at a particular spatio-temporal coordinate [102]. These E-mages are inputs to a tree of mathematical operators that create new E-mages. Data streams are converted to raw data E-mages by the Ingestor and transferred to the Stream Processing Engine which processes the data E-mages. Motivated by image processing methods, EventShop currently can apply a spatial convolution operator for prediction. Based on situation models defined by the users, event processing operators are applied on the E-mage stream to detect situations. Finally, the Action Control Unit allows an application developer or domain expert to visualize the detected situations and perform specified actions in response [89]. EventShop has been used by a number of groups for various applications such as identifying weather patterns and demand hot-spots of business products, detecting flu outbreaks, wildfires, and hurricane detection, providing allergy risk prevention recommendation and flood loss mitigation [89, 114].

2.2.3 Spatial Prediction Models

Under certain smoothness assumptions, spatial prediction models track back to various techniques to estimate the unobserved values, such as inverse distance weighting, splines and Kriging. These models assume no auxiliary data and are easy to implement, but they are not very accurate. Other approaches focus on finding correlated data, and making use of this additional information. All available data is jointly modeled in a spatial process [30], but auxiliary data may dominate the prediction results. Average weighted auxiliary data methods [11, 87] are used to smoothly model the output, resulting in a better prediction,

especially when predicting observations geographically distant from others. The Multitask Gaussian process [13] assumes a Gaussian prior over latent functions to directly induce correlations between tasks. The multiscale spatial process proposed in [78] models the bias of each data source, for multiple auxiliary data.

Because of the computation of the inverse matrix in the Gaussian Process, the complexity of the exact inference hinders large-scale applications. To adjust to large-scale data sets, many studies have focused on solving the so called "Big n" problem. Covariance tapering and replacing small covariance matrix values with 0 are used to simplify the computation [61]. Various statistical methods or numerical methods are applied to approximate likelihood so the matrix is easier to compute [3]. Various low rank approximations [92, 123] were proposed to perform exact computations on a simplified model of reduced rank matrix to covariance matrix inference. Proper simplification based on auxiliary information from the data also leads to parameter space dimension reduction. Data fusion [39, 130] via space alignment integrates multi-sensor data into one common space. The data fusion problem has been studied in the field of dimensionality reduction of high dimensional data [98, 127]. Graph data and vector data can be combined by manifold learning [64]. Models are being developed to learn a low-dimensional data representation that incorporates the clustering information encoded in graph data. Several models [133] are proposed to incorporate the intrinsic nonlinear dimension reduction into a linear low-rank factorization. Given a set of reference correspondence data that correlated across different domains, space alignment attempts to place the semantic related data in a common space [20]. [126] shows the equivalence of graph-based methods with Gaussian Process models.

2.2.4 PM2.5 prediction

Remote sensing data such as aerosol optical depth has been used to estimate ground-level PM2.5 [118]. Various atmospheric transportation models are developed to predict PM2.5 [14, 94]. However, these chemical models are impossible to incorporate other data sources and very hard to run by non-experts. Machine learning and statistical techniques such as neural networks and inverse distance weighting on probable factors are also used for PM2.5 prediction [134]. In [71], the authors apply computer vision techniques on user-generated photos to compute haze level for air pollution monitoring in cities. This study efficiently transforms online user photos to air pollution participatory sensing, which is the new direction of multimedia.

2.3 Framework

We propose a multimodal geospatial prediction framework. It consists of three components: data pre-processing, spectral feature extraction, and prediction. Figure 2.3 gives an illustration. The framework integrates diverse data into a unified space-time-value format and represents the data as a spatial and time dependent thematic grid. It provides a feature extractor and prediction operator to efficiently generate multi-resolution features and predict values at unobserved locations. The result is improved by explicitly modeling the associations of these spectral features. We will introduce each component in the following subsections.

2.3.1 Problem Definition

In the EventShop, E-mage and Theme are used to present data in an unified format. Here, we use E-mage and Theme to present the data as well. E-mage is denoted as a collection

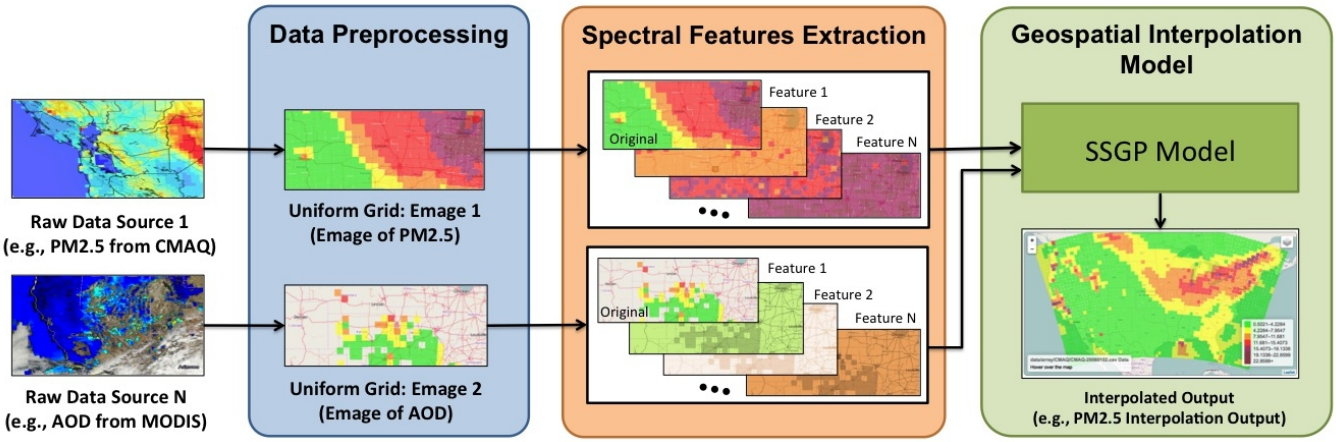


Figure 2.3: The framework of multiresolution prediction

of grids, where each grid cell captures a value associated with a particular data source at a particular spatio-temporal coordinate. Theme refers to a topic for a given data source, such as the PM2.5, aerosol concentration, or traffic flow used in this study. We define the prediction problem using both E-mage and Theme. $E = \{E_1^t, E_2^t, \cdots, E_{m+1}^t\}$ is a set of E-mages representing data collected on $m+1$ themes at time t, where $0 \leq t \leq T$. S is the set of all locations in E-mage. Each location $s \in S$, denotes the collection of grids on m themes, as $X(s) = \{x_1^t(s), x_2^t(s), \cdots, x_m^t(s)\} \in \mathcal{X}$ and the $(m+1)^{th}$ theme as $z^t(s) \in \mathcal{Z}$, where $\mathcal{Z}$ is the theme set of the prediction target, and $\mathcal{X}$ is the set with auxiliary data. The problem is converted to a supervised learning problem.

2.3.2 Spatial Gaussian Process Regression

Because the observation of PM2.5 is a continuous value, the prediction problem is naturally a regression problem. Thus, we approach the prediction of $z \in \mathcal{Z}$ by learning an efficient mapping from inputs $x \in \mathcal{X}$ to output z based on samples. All the spatio-temporal heterogeneous data streams are indexed and integrated into a grid structure (e.g., $5km \times 5km$) in their spatial domain.

The mapping function learning problem is formulated as a Spatial Gaussian Process regression learning problem. More specifically, the goal is to predict a variable Z at location $s \in D$ using data $X(s)$. The spatial process at location $s \in D$ can be defined as

$$z(s) = \beta^T X(s) + w(s) + \varepsilon(s), \tag{2.1}$$

where $X(s)$ is a vector of m features associated with each site s, and β is an *m-dimensional* vector of associated coefficients. The residual is partitioned into two parts, non-spatial $\varepsilon(s)$, and spatial $w(s)$, where $\varepsilon(s) \sim N(0, \tau^2)$ for any location s. Spatial dependence can be imposed by modeling $\{w(s) : s \in D\}$ as a zero mean stationary Gaussian process with covariance function, $C(s, s', \phi)$, specifically, the $(i, j)^{th}$ element $C_{ij}(\phi) = \sigma^2 \rho(s_i, s_j; \phi)$, where $\rho(; \phi)$ is a correlation function between any two locations in $s_1, \cdots, s_n \in D$. Let $w = [w(s_1), \cdots, w(s_n)]^T$, and we have

$$w|\sigma^2, \phi \sim N(0, \sigma^2 C(\phi)). \tag{2.2}$$

Here, we give an example of a covariance function. Consider the $(i, j)^{th}$ element of exponential covariance function $C(\phi)$, $C_{ij}(\phi) = exp(-||s_i - s_j||/\phi)$, and $||s_i - s_j||$ is the Euclidean distance between locations s_i,$s_j \in D$. This model is interpreted as follows: τ^2 is the variance of the non-spatial error, say from measurement error or from a micro-scale stochastic source associated with each location, and σ^2 and ϕ indicate the scale and range of the spatial dependence, respectively. It is clear that the dependence between two locations decreases as the distance between them increases. The likelihood is given by $Z \sim N(\beta^T X, \Sigma)$, with $\Sigma = \sigma^2 C(\phi) + \tau^2 I_n$ where $Z = (z(s_1), \cdots, z(s_n)$, X is a matrix of regressors, and I_n is an identity matrix.

2.3.3 Spectral Features

This approach is motivated by spectral downscaling [95]. In the spectral downscaling, outputs from the first numerical models can be utilized to generate features at higher spatial resolution. The generated features then can be used for the second models to improve prediction accuracy. For example, in climate research, simulations from circulation models are used as inputs for regional climate models. Spectral downscaling can obtain large-scale spatial trends from the outputs of the first model and improve fine-resolution predictions [72].

Instead of modeling the covariance of Z and X spatially, we model the covariance of their discrete *Fourier* transforms in the spectral domain, thus allowing the model to incorporate greater complexities in the $X \sim Z$ relationship. We "split" covariance of the transformations into k parts in the spectral domain. Modeling each of these sections in the spectral domain is equivalent to modeling the relationship at different scales in the spatial domain. Spectral features significantly reduce computation time compared to approaches where the latent field is modeled using all available data. The complexity of calculating spectral features is $\mathcal{O}(n)$, where n is the number of grid cells. In this subsection, we introduce methods to generate spectral features from data.

Let $X(s)$ be the measurements of auxiliary data, and $Z(s)$ be the observed measurements at location s. $Z(s)$ is sparsely observed. Suppose the spatial correlation between $X(s)$ and $Z(s)$ is determined by the spectral density $f(\omega)$. G is Gaussian with $E(G(\omega)) = 0$, and $Var[G(\omega)] = \delta^2 f(\omega)$. $X(s) = \int_{R^2} exp(-i\omega^T s) G_1(\omega) d\omega$, and $Y(s) = \int_{R^2} exp(-i\omega^T s) G_2(\omega) d\omega$, where G_1 and G_2 is Gaussian with $E(G_i(\omega)) = 0$, $Var[G_i(\omega)] = \delta_i^2 f_i(\omega)$, $i = 1, 2$. Suppose $X(s)$ is completely observed. $G_1(s)$ can be obtained by the inverse Fourier transform, $G_1(\omega) = \int_{R^2} exp(i\omega^T s) X(s) ds$. In the spectral domain, because of the normal distribution, $G_2(\omega) = \beta(\omega) G_1(\omega) + G^*(w)$, where $\beta(\omega) = Cor(G_1(\omega), G_2(\omega)) \frac{\delta_2 \sqrt{f_2(\omega)}}{\delta_1 \sqrt{f_1(\omega)}}$. G^* is independent

over ω. Conditional on $X(s)$, the mean of $Z(s)$ is

$$Z(s) = \int_{R^2} exp(-i\omega_c^T s)\beta(\omega)G(\omega)d\omega.$$

The relationship of $X(s)$ and $Z(s)$ is modeled at multi-resolution by $\beta(\omega)$. In order to easily compute the integral, the method in [95] is adopted to model $\beta(\omega)$, where $\beta(\omega)$ can be represented by a combination of M basis functions $B_k(\omega)$:

$$Z(s) = \sum_{m=1}^{M} \lambda_m \int_{R^2} exp(-i\omega^T s)B_m(\omega)G(\omega)d\omega.$$

Fixing the basis functions, the conditional mean of $Y(s)$ is a linear combination of the spectral features $\bar{X}_k(s)$,

$$\bar{X}_k(s) = \int_{R^2} exp(-i\omega^T s)B_k(\omega)G(\omega)d\omega. \tag{2.3}$$

Spectral features can be efficiently computed by fast *Fourier* transformation on grid cells data. Derived from the raw data, the low frequencies can capture large-scale trends and high frequencies can capture high resolution variations. We use these covariates as additional k features for our model. Since spectral features are incorporated in this work, this model is called the Spectral Spatial Gaussian Process model.

With models and features in mind, we present Algorithm 1 to perform the prediction task.

First, data is integrated into unified grid cells. For each cell, all data sources that intersect with it are averaged using inverse squared weighting on the distance from the cell center. More concretely, let the value of the data source X at location l and time t be $X(l,t)$. Let the set of all locations with data in X at time t that lie in cell i with center p_i be L, and r_{ab} be the spatial distance between locations a, and b. Then, the value of X at cell i at time t

Algorithm 1 Spectral Spatial Gaussian Process Prediction

Data: X: a collection of d data sources, $X = (x_1, \cdots, x_d)$, z: prediction target data
Input: Basis function $B_m(\omega)$, number of spectral features k
Output: predicted z
Step 1: Spectral Feature Extraction
Integrate data into uniformed space-time grid cells.
Compute spectral features for each data source which is completely observed in grid cells using Equation (2.3).
Step 2: Learning SSGP
Specify priors for model's parameters, $\Omega = (\beta, \tau^2, \sigma^2, \phi)$.
Inference proceeds by performing Gibbs sampling from $p(\Omega|Z)$.
Step 3: prediction
Prediction at an arbitrary location s_0 is performed by sampling $p(z(s_0)|Z)$ with posterior draws of Ω

is given by

$$X(p_i, t) = \sum_{l \in L} \frac{1}{r^2_{lp_i}} X(l, t). \tag{2.4}$$

In order to complete the Spectral Gaussian Process model, we first specify the form of basis function B_k in Step 1 of Algorithm 1. Given the fact that Bernstein polynomials are commonly used to approximate any continuous function over a prescribed interval [28], we choose Bernstein basis polynomials as our basis functions for Equation (2.3). The Bernstein basis polynomials are expressed as

$$B_m(\omega) = B_m(\delta) = \binom{M-1}{m-1} t^{m-1}(1-t)^{M-m},$$

where $t = \frac{\delta}{2\pi} \in [0, 1]$. Here, we use these Bernstein basis functions to generate additional features as described above. For instance, in the PM2.5 prediction task at location s, we denote PM2.5 values by $z(s)$, CMAQ by $x_1(s)$, AOD by $x_2(s)$, traffic data by $x_3(s)$, and the generated features by $x_4(s), x_5(s), \cdots, x_{k+3}(s)$.

Let $Z = (z(s_1), \cdots, z(s_n))$. The likelihood of the model is given by $Z \sim N(\beta^T X, \Sigma)$, with

$\Sigma = \sigma^2 C(\phi) + \tau^2 I_n$. X is the matrix of regressors, with the $(i,j)^{th}$ element, $X_{ij} = x_i(s_j)$. Eventually, the prediction can be realized using Algorithm 1.

2.4 Experiment

2.4.1 Data and Performance Metrics

We provide a synopsis of the data used in Table 2.1 and visualizations for one day. Our experiments are performed over California. The data sources are described below.

- **US EPA PM2.5 Observation Stations:** The US EPA provides hourly PM2.5 concentrations ($\mu g/m^3$) as measured by 82 National Air Monitoring sites across California. Each site has multiple sensors. Hourly values are averaged to daily PM2.5 concentrations.

- **Traffic Data:** California Department of Transportation provides near real-time traffic data. The data is collected from a network of loop detectors and traffic monitoring stations operated by the Caltrans Performance Measurement and Enhancement System. It is collected every 30 seconds and aggregated over 5 minutes on traffic occupancy, speeds, delays, etc. We use aggregated daily traffic flow data measured as vehicles per day.

- **CMAQ Modeling System:** CMAQ is an open source project from the Atmospheric Science Modeling Division of the Environment Protection Agency. Its aim is to develop chemical models to simulate the transportation of various pollutants and estimate their spatial distributions. Daily estimates of PM2.5 concentration values by this model are in $12km \times 12km$ grids.

- **MODIS AOD Data:** Moderate Resolution Imaging Spectro-radiometer is a remote sensor in NASA's Terra satellite. It captures the image of the Earth in 36 different wavelengths. Optical depth of aerosols measured at 550 nm wavelength can be used to get accurate estimates of PM2.5 concentrations in the air [22]. This data is available in the size of $10km \times 10km$.

We evaluate the performance of our models by comparing the estimated values with the actual observed values at some of the monitored locations. To make sure that the measurement is an accurate reflection of the model's performance, these monitored locations are randomly chosen, and not used for training. The errors are evaluated in terms of Mean Absolute Percentage Error (MAPE), which is defined as

$$\text{MAPE} \;=\; \frac{1}{N}\sum_{s\in M} \mid \frac{Z(s)-Y(s)}{Y(s)} \mid \times 100, \tag{2.5}$$

where M is the set of monitored locations chosen for evaluation, Y, and Z are the measured and estimated values respectively. Other metrics include Prediction Mean Squared Error (PMSE) and Prediction Mean Absolute Error (PMAE), defined as

$$\text{PMSE} \;=\; \frac{1}{N}\sum_{s\in M} (Z(s)-Y(s))^2, \tag{2.6}$$

and

$$\text{PMAE} \;=\; \frac{1}{N}\sum_{s\in M} \mid Z(s)-Y(s) \mid . \tag{2.7}$$

For all these three metrics, smaller numbers present greater accuracy.

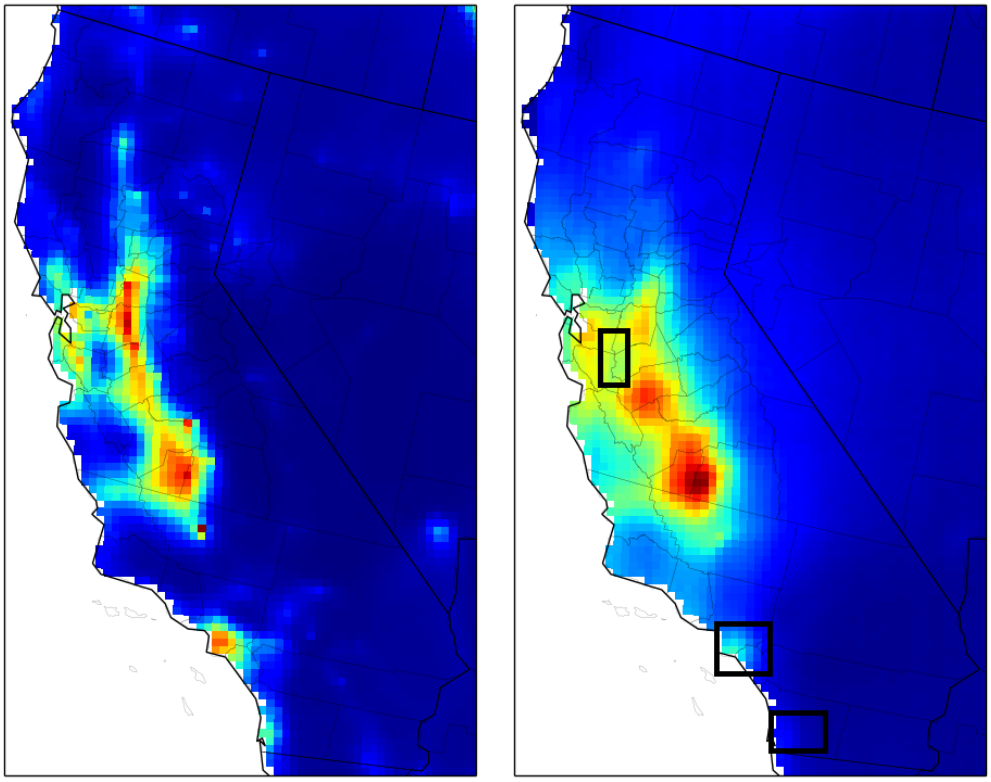


Figure 2.4: Prediction results using Spatial Gaussian Process on all data sources.

2.4.2 Details and Setup

We model daily PM2.5 concentration for the year 2008. Spatially, we use the $12km \times 12km$ CMAQ grid as our underlying grid for two reasons. First, this grid stays the same across all days, and second, CMAQ has complete measurements over the spatial domain. In the following sections, all cells refer to CMAQ grid cell. For each cell, all data sources that intersect with a CMAQ grid cell are averaged using inverse squared weighting on the distance from the cell center using Equation 2.4.

Priors for parameters are first assigned as $\beta_j \sim N(0, 10^2)$. Parameters σ^2 and τ^2 follow $IG(2, 1)$ and $IG(2, 0.2)$ priors, respectively. We use an exponential covariance function and assign $\phi \sim N(0, 10^2)$. The prior values for the parameters above are obtained from [4]. We denote parameters by Ω, where $\Omega = (\beta, \tau^2, \sigma^2, \phi)$. Inference proceeds by performing Gibbs sampling from $p(\Omega|Z)$. Prediction at an arbitrary location s_0 is performed by sampling

$p(z(s_0)|Z)$ with posterior draws of Ω.

2.4.3 Performance Comparison with A Physical Chemical Model

Physical chemical model is commonly used in the field of air pollution. It involves functions of emission factors, meteorology, and landscape based on a few empirical assumptions and parameters. We combine the diverse data sources described earlier and predict PM2.5 values across California. The results between the Physical Chemical Model and the proposed model are visually compared. Figure 2.4 shows that both the physical chemical model and the proposed model successfully capture the pollution distributions in the map. However, by incorporating MODIS AOD and traffic data, the proposed approach predicts results more accurately than the CMAQ physical chemical model. The results are particularly significantly improved around Los Angeles, Orange County, San Diego, Ventura County, Sacramento, and the San Joaquin Area, where the predicted values are closer to the observed station values than CMAQ. Notably, these areas are some of the most populated areas of the state and are home to around half of California's population. We also highlight two areas using solid boxes. In these two areas, the physical chemical model predicts values smaller than ground truth in Northern California but predicts values greater than ground truth in Southern California. The proposed data-driven model outperforms the physical chemical model in local areas.

2.4.4 Comparison with State-of-the-art Approaches

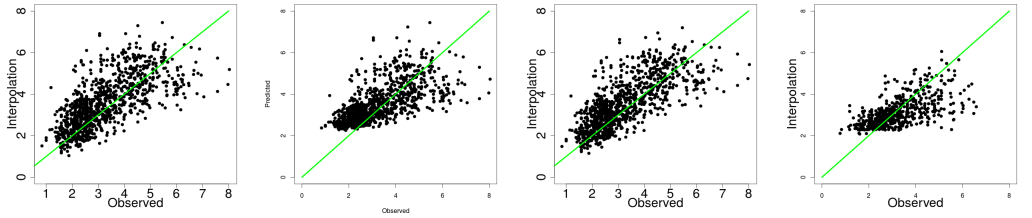
Prediction methods are grouped into two categories: deterministic and statistical. On one hand, we compare the performance of the proposed method with the state-of-the-art statistical approaches in spatial predictions: Ordinary Kriging or a spatial Gaussian Process. On the other hand, we compare with a deterministic method: Inverse Distance Weighted

Table 2.2: Geospatial prediction models evaluation

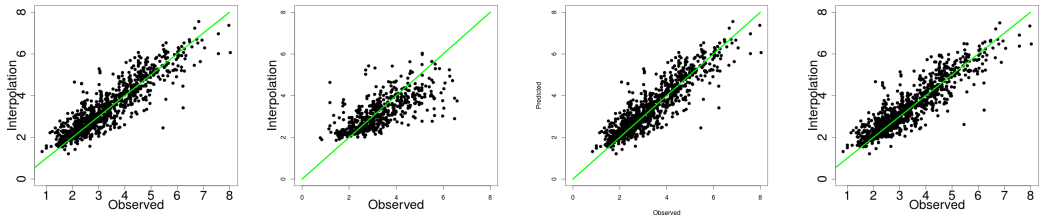
Data	Model	PMSE	PMAE	MAPE
CMAQ	—	1.0619	0.7918	27.2873
CMAQ	LR	0.9586	0.7590	27.1077
Station	IDW	1.2304	0.8933	27.5009
Station	Kriging	0.9077	0.6997	22.9672
CMAQ	IDW	0.7737	0.6531	21.4347
CMAQ	Spectral IDW	0.6210	0.5699	18.5722
CMAQ	SSGP	0.3468	0.4177	**14.2727**
ALL	IDW	0.6159	0.5675	18.0187
ALL	SGP	0.3006	0.3937	**13.5109**
ALL	SSGP	0.2858	0.3823	**13.1087**

LR: Linear Regression
SGP: Spatial Gaussian Process
IDW: Inverse Distance Weighted Method
SSGP: Spectral Spatial Gaussian Process
Spectral IDW: Spectral Inverse Distance Weighted Method
CMAQ, MODIS AOD, and traffic data
The unit of PMSE and PMAE is $\mu g/m^3$
The unit of MAPE is percentage.

Method (IDW). IDW predicts value at unknown locations using weighted known neighborhood values, and the weights are inversely proportional to the distance to the target location. Similar to IDW, Kriging uses neighboring values to predict unknown values. However, the calculation of the weights differs from IDW because it uses kernel functions and covariation between observed data. Kriging is a generic term for spatial prediction models. Various types of Kriging have been state-of-the-art in many geo science disciplines such as natural resources [37, 43], remote sensing [60], and environmental sciences [68]. The most commonly used is Ordinary Kriging [85]. Ordinary Kriging gives estimates by using weighted linear combinations of the available data. It aims at minimizing the variance of the errors [48]. Thus, we compare the proposed method with Ordinary Kriging as well as IDW. Besides the spatial regression, we also include linear regression in comparisons because linear regression is treated as a baseline without considering spatial correlation.



(a) CMAQ Data (b) Linear Regression (c) "Station" data - Kriging Model (d) CMAQ data - Spectral IDW Model



(e) "CMAQ" data - SSGP Model (f) "ALL" data - IDW Model (g) "ALL" data - SGP Model (h) "ALL" data - SSGP Model

Figure 2.5: Observed vs prediction values on testing samples

We benchmark the performance of our model using two-fold cross validation. In particular, we randomly remove half of the stations across California and predict their PM2.5 concentration values as measured at the stations, using CMAQ, MODIS AOD, and traffic data from the other half. Each model is trained using the training data and is then tested on the remaining half of the data. We use mean squared error and mean absolute error to measure the differences between values predicted by each model and station measurements. The performance of CMAQ, linear regression, Ordinary Kriging, the IDW and the Spatial Gaussian Process and the Spectral Spatial Gaussian Process on PM2.5 values is shown in Table 2.2. The Spectral Spatial Gaussian Process model reduces mean absolute percentage error from 27.29% of CMAQ, to 13.11%. Our models significantly outperform Ordinary Kriging and the IDW as well.

2.4.5 Point-based Performance Analysis

To show the performance of the prediction at point level, we compare predicted values and ground truth values using each test sample. The comparison of different models on testing samples is shown in Figure 2.5. Each dot indicates one test sample. The X-axis represents observations of ground truth, whereas the Y-axis represents predicted values. The perfect prediction results should lie in the "green" line, which means that prediction values are equal to ground truth values. In (a), the results of the CMAQ data are not very accurate with many dots far way from the "green" line. In (b), without including spatial correlation, linear regression does not perform well, either. In (c), Ordinary Kriging takes in station measurements and spatial correlation, and performs better than CMAQ because there are fewer "outliers". We show the effectiveness of spectral features in (d). In this plot, it's clear that more dots go towards the "green" line compared to (a), (b), and (c). In (e) and (f), by incorporating MODIS AOD and traffic data, the Spatial Gaussian Process works much better and by using spectral features our proposed method further improves prediction accuracy with more dots closer to the "green" line. Our Spectral Spatial Gaussian Process model predicts results more accurate than CMAQ, Ordinary Kriging, and Inverse Distance Weighted Method at 78%, 72% and 70% of testing samples, respectively.

This framework can be applied to geospatial prediction problems such as traffic estimation, noise pollution estimation, etc. For the traffic estimation problem, this framework can be used to predict the traffic conditions where there are no traffic sensors, using data streams like GPS based projections from cars, public events and points of interest.

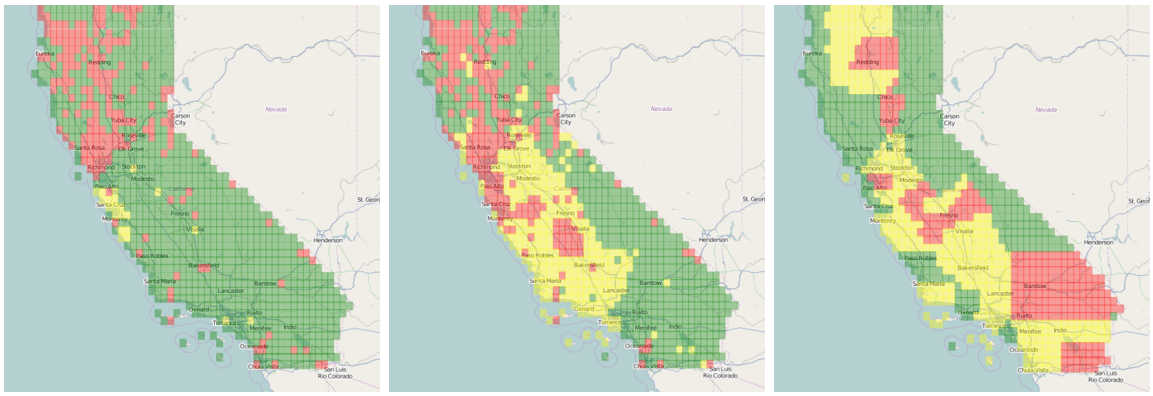
2.5 Situation Recognition on EventShop

We use EventShop to balance between efficiency and accuracy by integrating not only numerical model outputs but also various spatio-temporal heterogeneous data streams. EventShop provides operators for data stream ingestion, integration, visualization, and situation characterization. We use the proposed predictive framework and communicate with EventShop for detecting the situation of asthma risk.

2.5.1 Asthma Risk Zone Study

Epidemiological studies have shown that overexposure to PM2.5 results in adverse respiratory health responses such as an asthma attack. In this sense a method for accurate estimation of the PM2.5 concentration is of critical necessity in asthma risk situation recognition but has been absent up until this point. The spatial coverage of environmental data used in the previous studies was spatially sparse because the data is collected merely from monitoring stations. The ambient assessments are typically obtained by averaging PM2.5 concentrations. By applying our predictive model to predict PM2.5 in space in the year 2008 using monitoring data, traffic data, CMAQ, and MODIS AOD data, we demonstrate our approach to be more comprehensive and accurate. We then use our new PM2.5 values from the operator with the asthma alert application on EventShop to analyze the asthma risk situation across California. The asthma risk model aggregates data from three sources: pollen level [2], air quality index [3], and PM2.5, and then classifies the result into three risk categories low (green), medium (yellow) and high(red). The screenshots of the E-mages from EventShop are shown in Figure 2.6. We compared the risk indications of the two models to the number of asthma related hospitalizations per 10,000 people, by county, in the year 2008 [4]. The model based on our

[2]Pollen Concentrations: http://pollen.com
[3]AirNow: http://airnow.gov
[4]Data from http://californiabreathing.org



(a) Earlier EventShop (b) predictd PM2.5 (c) Hospitalizations

Figure 2.6: Asthma risk situations across California detected by EventShop

predicted values clearly provides a better coverage and higher accuracy in detecting asthma risk level. To measure the improvement, we calculated the correlation of our risk values with the number of hospitalizations per 10,000 people by county. The model using our operator has a correlation of 0.45 with the hospitalizations data, while the old model has a correlation of 0.17, thus showing a significant improvement over the previous model.

After the situation is detected, an appropriate action can be taken. In this case, EventShop can send out an alert tweet[5] mentioning that the asthma index is high to people who subscribed to the handle, and have mentioned asthma in their tweets and are currently staying in that location.

[5]EventShop Twitter Handle: https://twitter.com/EventShopSLN

Chapter 3

Asthma Risk Factors From Personal Data and Environmental Data

We have experienced the power of data in heath care. Physicians make decisions using their clinical expertise and data from various medical tests. However, big data is transforming health care. Emerging mobile technologies and diverse data provide the possibility of future health in which individual's lifestyle and environment are continuously quantified for preventive care and patients can actively participate in their own care and focus on their needs, resources, and preferences. Following the previous chapter that we discuss an example of using environmental data to help asthma patients, in this chapter, we introduce a more comprehensive study using both activity, profile data and environmental data for asthma risk analysis.

Although asthma is a potentially life-threatening lung disease and has been studied for a long time, asthma causes are still unclear. With an increasing amount of data, such as Electronic Health Records, social media, environmental sensory data, smartphone and wearable sensor data, a data driven solution holds promise for helping to solve this problem

[93]. Thus, many studies have been performed using various forms of data. [67] predicted asthma attacks by considering bio-signals of patients and environmental data. [59] studied the association between the spatial distribution of allergy prevalence and air pollutants such as PM2.5, as well as living distance from point of interests such as parks and roads. [31] detected asthma risk from personal profiles. [33] designed tools for discovering dynamic changes in body sensor network data streams of asthma patients. [2] studied occupation as a factor in asthma. [44] researched associated risk factors of air pollution and weather, [46] revealed the impact of lifestyle and behavior on asthma vulnerability.

In this chapter, we combine environmental data and patient survey data for a more comprehensive study. Environmental data is collected from open data on the Internet. The survey data is obtained from EPA. We propose a data-driven method which combines both data sources for discovering comprehensive risk factors, as well as aiding in understanding asthma.

To identify these risk factors within integrated diverse data sources, the framework is built on top of Personal EventShop [52] and EventShop [32] to address this problem because of their capability for integrating and analyzing heterogeneous data sources. EventShop is a generic infrastructure for application developers to analyze various spatio-temporal data streams. Personal EventShop [51] is a unified framework for aggregating personal data streams to analyze life events and personal situations. Details of this framework are described in the next section.

3.1 Framework

The proposed framework is based on Personal EventShop and EventShop. The three main parts of the framework are briefly discussed.

Data ingestion

Personal EventShop receives data from wearable sensors, mobile apps such as calendar and smartphone sensors like accelerometers and GPS. It correlates these data streams and determines users' activity levels and life events such as exercising, sleeping, and working. This chronicle of life events, called "Personicle" is stored in a database with attributes of user-id, location, time, and life events. This framework also collects data from global sources like EventShop, which can provide aggregated environmental data from different sensors. In the context of this dissertation, EventShop can provide interpolated PM2.5 values inferred from sources like air pollution monitoring sites, satellite imagery and traffic [110]. All of this data is collected and sent for aggregation and matching.

Data aggregation

Ingested data has a wide range of granularities in both space and time. For example, in the context of this dissertation, emission inventory data was reported on county level once every three years, activity data was reported over a period of several years, and varying levels of spatial granularity and pollution data were collected at station locations, and were reported as daily values. Personal EventShop takes in all this data and matches it to a common spatio-temporal scale. This data is then sent to the Analysis Engine.

Analysis Engine

This part uses aggregated data, and performs desired analyses on it. For example, these analyses can be co-occurrence pattern detection, personal situation recognition and life event detection. In this case, the aggregated data from the sources described in section 3.2 are transformed to features as described in section 3.3. The features with the highest discrim-

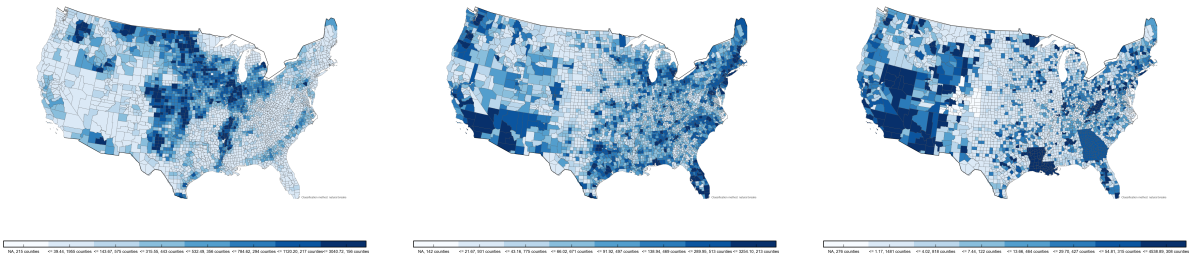


Figure 3.1: Expected county level emissions from agriculture, road dust, and mining from CHAD

inatory power are detected using the model described in section 3.4 and then used to rank the asthma risk factors.

3.2 Environmental Data and Personal Data

Our asthma risk analysis approach requires various personal and environmental features at the same granularity. However, the data is at different resolutions, such as point data and region data. Data sources are described in this section and the methodology to obtain the features from the data is discussed in the section below.

Consolidated Human Activity Data

Consolidated Human Activity Database Master (CHAD) [105] is a collection of profile and activity information from 22 studies, and contains data for over 54,000 days combined from over 700 counties across the United States. Each personal profile contains information like zipcode, county, age, gender, race, history of asthma, cardiovascular illness, whether the person is a smoker or lives with a smoker, employment, education level, and income. Every person has a set of diary entries. Each diary entry is an activity record containing activity (like walking, exercising, sleeping, and leisure), location (home, office, and traveling), and

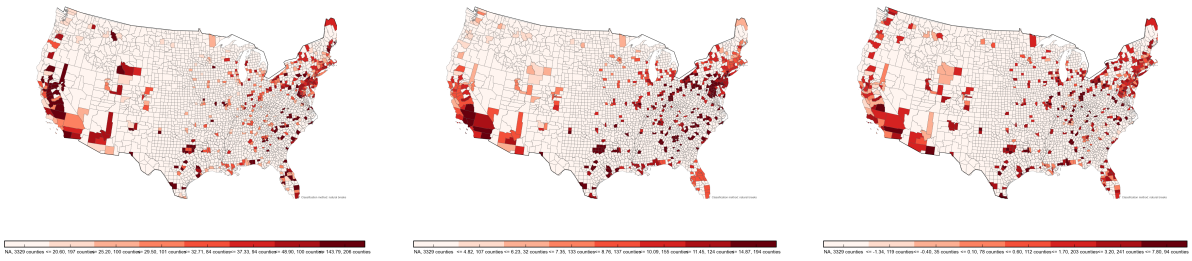


Figure 3.2: Average maximum, mean, and minimum PM2.5 concentration values for the month of June for counties from CHAD

start and end times. Every diary entry also contains flags showing whether the subject was breathing heavily and whether the subject was smoking during each activity.

Emission Factors

The National Emission Inventory data was collected in 2011 by the EPA. This is an estimate of PM2.5 contributions of all the sources of air pollution within the United States in *metric tons/year*. Sources are categorized into classes like Agricultural, Industrial, Dust, Fuel, and Mobile. For example, Fig. 3.1 shows the distributions of agriculture, road dust, and mining.

Historical Air Pollution Data

Historical air pollution data is the actual concentrations of pollutants in the environment. This data is available as daily average values recorded at each of the individual monitoring stations across the United States. The stations also record the daily average weather conditions: temperature, pressure, wind speed, and wind direction. For example, Fig. 3.2 shows the distributions of PM2.5 in the month of June.

In the next section, we describe the features extracted from these data sources and how they

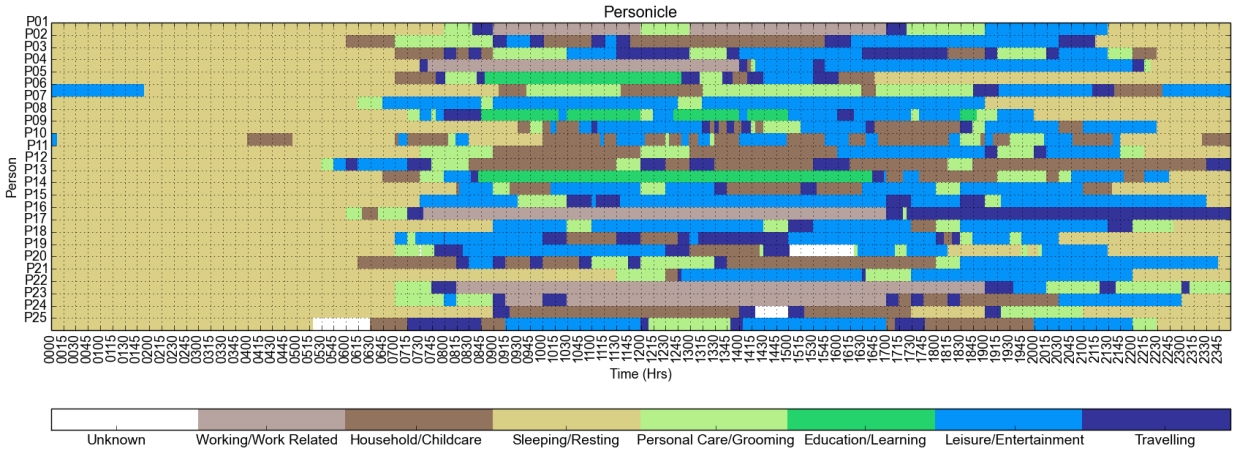


Figure 3.3: 25 randomly chosen diary entries from CHAD. Each coloured row on the y-axis represents a day's activities for one person. The color marks the activities, and the x-axis shows time in 15 minute increments

are used.

3.3 Feature and Data Analysis

3.3.1 Personal Life Events Features: F_P

We use the profile information provided in CHAD for people who have location information at the county level and have answered whether or not they had asthma. The number of such people was 11,000, with 24,000 days of activity data. Of these, around 950 people reported having asthma, which is close to the asthma prevalence rate of 8.7% in the United States [53]. For analysis, we use the asthmatic subjects and an equal number of non-asthmatic subjects. The non-asthmatic subjects are randomly chosen to balance the dataset.

Features extracted and their justifications, as mentioned in [53] are described in Table 3.1.

Activity data is in the form of a set D of daily diary entry tuples. Each tuple x for the day j contains the activity's start time, duration $d_{x,j}$, activity code $a_{x,j}$, location code $l_{x,j}$, and flags

Table 3.1: Extracted personal profile features

Feature	Comment
Age	Children are more susceptible than adults
Gender	Females are more susceptible than males
Active & Passive Smoking	Smokers are more susceptible
Occupation, Income, & Education	Lower income and lower education correlate with asthma
Hours of work & Employment Status	Measure of stress
Gas stove ownership, heating & fuel type	Exposure to smoke

for whether the person was smoking ($s_{x,j}$) and whether the person was breathing heavily ($hb_{x,j}$) during the activity. First, locations are classified into a set L of five non-exclusive categories:

$$L = \{work,\ travel,\ home,\ indoor,\ outdoor\}$$

Activities are classified into a set A of six non-exclusive categories:

$$A = \{sleep,\ work,\ exercise,\ walking,\ cycling,\ leisure\}$$

.

For every person p_i for day j, in the set of people P, let the set of corresponding diary entries be $D^j_{p_i}$. For every person p_i, and every location category L_k, the average time spent daily at locations in that category $t^i_{L_k}$ is computed for diary entries x in $D^j_{p_i}$ over all j:

$$t^i_{L_k} = \underset{j}{\text{mean}}(\sum_x d_{x,j}), \qquad \{\forall j, x : x \in D^j_{p_i}\ and\ l_{x,j} \in L_k\}$$

. Similarly, for activities, the average time spent daily by p_i performing activities in each activity category A_k, denoted by $t^i_{A_k}$, is computed. Also, the average amount of time in the

Table 3.2: Emission factors

Category	Emission Factor
Mobile	Aircraft, Marine Vessels, Locomotives, Equipment, Heavy Duty Vehicles, Light Duty Vehicles
Industrial	Agricultural, Mining, Oil & Gas Production, Storage & Transportation, Other
Dust	Construction, Paved Road Dust, Unpaved Road Dust
Fires	Agricultural Field Burning, Prescribed Fires, Wildfires
Fuel	Biomass, Coal, Natural Gas, Oil, Residential Wood, Other
Miscellaneous	Waste Disposal, Agriculture, Commercial Cooking

entries where p_i reports heavy breathing is computed as t^i_{hb}:

$$t^i_{hb} = \underset{j}{\text{mean}}(\sum_x d_{x,j}), \qquad \{\forall j, x : x \in D^j_{p_i} \; and \; hb_{x,j} = 1\}$$

. The average amount of time in the entries where p_i reports smoking, is also calculated as t^i_s, in the same way.

The average count of the number of times p_i reports heavy breathing, is recorded as (n^i_{hb}):

$$n^i_{hb} = \underset{j}{\text{mean}}(\sum_x hb_{x,j}), \qquad \{\forall j, x : x \in D^j_{p_i} \; and \; hb_{x,j} = 1\}$$

. Likewise, the average count of activities per day where the smoking flag was 1 is calculated as n^i_s. Figure 3.3 shows twenty five randomly chosen daily diaries from the data.

3.3.2 Emission Factors Features: F_E

PM2.5 emission sources from the national emission inventory were used at the county level. Different kinds of PM2.5 emissions have different health effects on people. The emission amount by each type of pollution source such as mobile, industrial, dust, fire, and fuel, was used as a feature. The various emission factors are listed in Table 3.2.

3.3.3 Air Pollution Features: F_A

Historical air pollution data records are in the form of daily average pollutant concentration values at the measuring station. Since all of our analysis is at the county level, we interpolate this point data to region data. This can be done using a variety of techniques, ranging from pure chemical transportation models to enhanced data based techniques [110]. For this experiment, linear interpolation was used.

The data used is between the years 2001 and 2014. The pollutants considered are PM2.5, Ozone, Carbon Monoxide, Sulphur Dioxide, and Nitrogen Dioxide. The weather factors considered are temperature, pressure, and wind speed.

For each pollutant and weather factor f, the following features are extracted:

$$f^m_{max} = \underset{y}{\text{mean}}(\max_d(v^{m,y}_{d,f}))$$

,

$$f^m_{mean} = \underset{y}{\text{mean}}(\underset{d}{\text{mean}}(v^{m,y}_{d,f}))$$

,

$$f^m_{min} = \underset{y}{\text{mean}}(\min_d(v^{m,y}_{d,f}))$$

, where f belongs to $\{PM2.5,\ SO_2,\ NO_2,\ O_3,\ CO,\ temperature,$
$pressure,\ windspeed\}$, y is the year between 2001 and 2014, m is the month, and d is the day of the month.

3.4 Models

Consider a feature vector $x \in R^d$ in the d-dimensional feature space, y^+ is the label for subjects with asthma, y^- is the label for subjects without asthma. Our feature ranking model is based on a gradient boosting regression tree method [29]. Risk factors are ranked by analyzing importance of features from this model. The gradient boosting tree framework generates an ensemble of weak regression tree models and combines all the weak learners to produce a strong classifier. The training algorithm performs gradient descent in the function space to minimize a differentiable loss function. Suppose $F(x)$ is the model, we have

$$F(x) = \sum_{i=1}^{M} \lambda_i f_i(x), \tag{3.1}$$

where $f_i(x)$ is the i^{th} weak learner, and λ_i is the weight associated with that weak learner.

One of the advantages of using this model for ranking features in asthma study is that no normalization of data is needed, and thus it is better at handling categorical data with discrete and continuous data.

The model hyper-parameters for the gradient boosting classifier (depth and number of trees) are selected using five-fold cross validation. The depth was varied between 1, 2, and 3, and the number of trees was chosen from 50, 100, and 150. After choosing the best hyper-parameters, the learned model was then applied to the remaining subset.

A K nearest neighbor classifier was used to predict whether a person is asthmatic or not using their features. The parameter K was also cross-validated on the training data using five folds to select the best K.

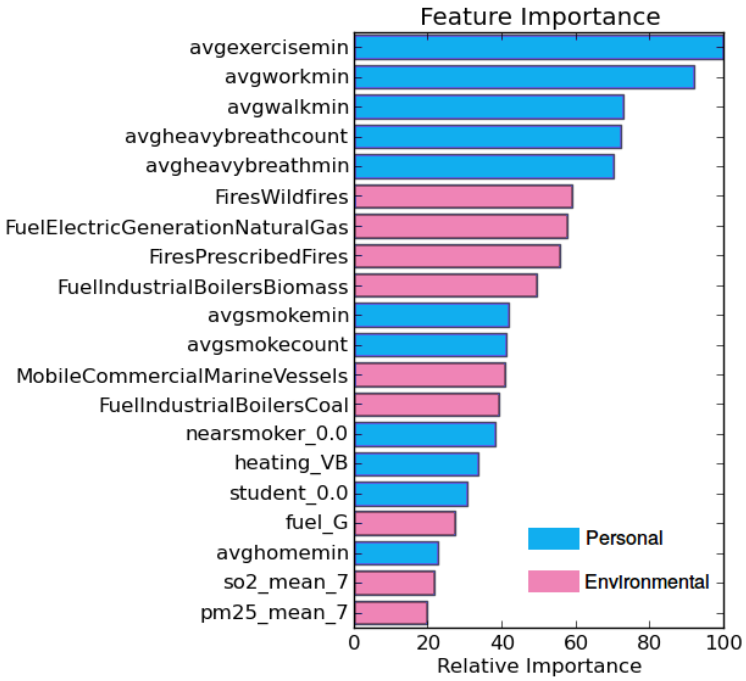


Figure 3.4: Relative importance of features as determined by a gradient boosting tree

3.5 Results

In this section we present our results based on the features extracted from the data sources, showing the effectiveness of the identified data sources.

Table 3.3: Classifier performance against feature set

Features	**Precision**	**Recall**	**AUC**
F_P	0.818	0.789	0.807
F_P+F_A	0.846	0.864	0.853
F_P+F_E	**0.902**	0.807	0.860
F_P+F_E+F_A	0.898	**0.927**	**0.911**

The most significant contribution of this chapter is the determination of the relative importance of factors in causing asthma. The twenty most effective features out of around 400 extracted features and their relative importance is shown in Figure 3.4. The top five features correspond to personal factors such as physical exertion and work stress. Following that, environmental features like forest fires and industrial fuel burning show up. Other features

that appear in the analysis are smoking habits, passive smoking, and personal exposure to smoke from domestic fuels and heating. Lastly, PM2.5 and SO_2 concentration during summer (July) also appear on the list. These are all independently known to be important factors affecting asthma susceptibility. Our ordering suggests that personal factors like physical exertion and stress are strong indicators of asthma risk, followed by exposure to PM2.5 and smoke from the environment. The personal features obtained here are particularly useful because they do not require extensive diary keeping by the subjects. Physical exercise time, working hours, time at home, and breath rate can be easily obtained from wearable sensors and smartphones. Other personal features such as smoking habits, fuel type, and occupation can be obtained from one-time questions.

The effectiveness of our features is shown by the good performance of an asthma classifier trained on them. The four K-nearest neighbor classifiers are trained on the various subsets of the data: one on just F_P, one on F_P and F_A, one on F_P and F_E and one F_P, F_E, and F_A. The Area Under the Curve metric, which is a commonly used metric in clinical studies [137] is used to evaluate the performance of the classifier. The performance is shown in Table 3.3. The table clearly shows that asthma prediction based on both environmental features and personal features significantly outperforms using any subset of these features. It may be noted that the recall of the system improves significantly, on moving from $F_P + F_E$ to $F_P + F_E + F_A$, at the cost of a small drop in precision.

3.6 Discussion

There is no clear understanding of the causes of asthma and no definite cure, making the study of asthma draw increasing attention in health care studies. Existing studies are often performed on either habit data or environmental data. We proposed a framework on top of EventShop and Personal EventShop, integrating environmental data and profile and activity

data, and extract both features that determine a person's susceptibility to asthma. A ranking of these features is given based on their potential in causing asthma. The features derived from our analysis agree with the current general consensus in the medical community about factors affecting asthma.

Chapter 4

A Graph-based Spatial Prediction Model

4.1 Motivation

Recent years have witnessed a dramatic increase in collecting heterogeneous data from disparate sensors, social media, cyberspace interactions, and IoTs. Many emerging applications have been developed, ranging from smart city to health care. In these applications, we are confronting significant challenges in leveraging the vast amount of data, including data processing, data storage, system design, model design, and data mining. In this chapter, we address a particular problem, namely model complexity problem.

In Chapter 2, we apply a Gaussian Process model for predicting missing values in spatial data. However, the complexity of the Gaussian Process model is high. Motivated by equivalence between the Gaussian Process model and graph-based semi-supervised learning framework, we propose an efficient graph based learning framework for geospatial data prediction. Given a set of proxy data, correlated across different domains, this framework attempts to place the

correlated data in a manifold space, and allows effective associations between proxy data and observations at multiple resolutions. The model integrates adaptation for weights in graph based manifold learning using similarity functions. Instead of assuming a fixed graph, we use a data-driven estimator of the graph structure that exploits correlations among the observed feature values. By incorporating adaptive multiscale features of observed covariance into the prediction, this framework is able to improve results of geospatial prediction tasks as well as reduce the computational time. The proposed graph-based model has two advantages. First, the proposed method alleviates the problem of over-fitting, compared to graph based Kriging, due to the computation of weights in a similarity matrix, and integrates hyper-parameter learning into graph structure learning. Second, the proposed method reduces the computational complexity compared to the Spatial Gaussian Process models where the latent field is modeled.

Similar to Chapter 2, to handle the missing data problem, we go through three steps: data preprocessing, spectral feature extraction, and prediction.

This framework receives prediction target data and proxy data from a satellite and environmental sensors like air pollution and temperature sensors. It correlates these data streams and integrates them into grid cells, similar to [110]. On a rectangular grid, we apply an efficient method to compute spectral covariates, and provide features for modeling the target values and proxy data at multiple resolutions.

4.2 Background of Graph-based Methods

Before we discuss graph-based methods, we first introduce some definitions in graph-based methods such as, graph, random walk, and graph Laplacian.

4.2.1 Graph

Given a graph $G = (V, E)$ with n vertices, V is the set of vertices and E is the set of edges. Consider G is a weighted and directed graph. This graph G can be represented by an adjacency matrix $W = (w_{ij})_{n \times n}$. In this adjacency matrix, if there is an edge from v_i to v_j, $w_{ij} > 0$, otherwise, $w_{ij} = 0$. Vertex v_i has in-degree $w_{+i} = \sum_{j=1}^{n} w_{ji}$ and out-degree $w_{i+} = \sum_{j=1}^{n} w_{ij}$. The volume of the graph is defined as $vol(G) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$. For an undirected graph, the in-degree and out-degree are the same. In such a case, we use d_i to represent w_{+i} and w_{i+}.

In social networks, when the data is in the form of a graph, the weight of the graph can be interpreted as a binary value based on friendship. For example, if Person A is a friend of Person B, then the weight of Person A and Person B is 1, otherwise, the weight is 0.

When the relationship between data points cannot be naturally used, we need to use functions to construct the weight matrix. There are some requirements for the functions. The functions must be symmetric and non-negative, so that they can quantify the affinity between two nodes. For example, the Gaussian kernel function is a popular choice, where the weight between node i and node j can be computed as $\exp(-||x_i - x_j||^2/\sigma^2)$. More examples of functions can be seen in [136].

4.2.2 Graph Random Walk

Given $G = (V, E)$, we can have a random walk associated with the graph. The probability of a transition from node i to node j is

$$P_{ij} = \frac{w_{ij}}{w_{i+}}. \tag{4.1}$$

In other words, the random walk at node i randomly follows its outlink by the weights. In practice, we want this random walk to be irreducible and aperiodic. In some cases, adding a small valve to all the nodes can help make the random walk irreducible and aperiodic. The advantage of an irreducible and aperiodic walk is that the walk has a unique stationary distribution on all the nodes. On a undirected graph, a closed-form solution of the stationary distribution is

$$\pi_i = \frac{d_i}{vol(G)}. \tag{4.2}$$

4.2.3 Graph Laplacian

With all the notations, we introduce the graph Laplacian, another representation of the graph G. The graph Laplacian is also the basis for graph-based learning frameworks. The mathematical theory of the graph Laplacian has been well studied in the field of spectral graph theory. For the proposed methods based on graph, we give the definition of the graph Laplacian.

In an undirected graph, we consider unnormalized and normalized graph Laplacians. The unnormalized graph Laplacian matrix is

$$L = D - W. \tag{4.3}$$

The normalized graph Laplacian matrix is represented as

$$\hat{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}. \tag{4.4}$$

Both unnormalized and normalized graph Laplacians are symmetric and positive semidefinite matrices. Using the graph Laplacian matrix, we can formulate graph-based regression models

for addressing spatial prediction problems.

4.3 Parameters Optimized Graph Regression

Suppose there are m observed data $(x_1, y_1), \cdots, (x_m, y_m)$, and u unobserved data $x_{m+1}, x_{m+2}, x_{m+u}$. Denote $n = u + m$ as the number of locations. Consider a connected graph, $G = (V, E)$. Nodes in V correspond to the n data points. Nodes in $M = (1, 2, \cdots, m)$ are observed data. Nodes in $U = (m+1, \cdots, m+u)$ are unobserved data. The edge E in the graph is described in a similarity matrix $S \in R^{n \times n}$, where S_{ij} is the similarity measurement of a pair of vertices, x_i and x_j. The similarity matrix S_{ij} can be

$$S_{ij} = \begin{cases} \exp(-\sum_{d=1}^{p} \frac{(x_{id}-x_{jd})^2}{\sigma_d^2}) & j \in \mathcal{N}_i \text{ or } i \in \mathcal{N}_j \\ 0, & \text{otherwise} \end{cases} \tag{4.5}$$

where $x_i = (x_{i1}, \cdots, x_{id})$, x has p dimensions of features. $\mathcal{N}_i$ is a set of indices of x_i's neighbors, σ_d $(d = 1, \cdots, p)$ are parameters associated with features. Based on the similarity matrix S, the graph Laplacian matrix $L \in R^{n \times n}$ is defined by $L = D - S$, where D is a diagonal matrix with diagonal entries as $D_{ii} = \sum_j S_{ij}, \forall i$. The key idea of the graph-based method is that the nodes connected by a large weight in the graph have similar values. In other words, observations $y_1, \cdots, y_n$ change smoothly on the graph. Based on the smoothness assumption, the task is to assign values to nodes U. [135] proposed Gaussian fields and harmonic functions to study a label propagation problem in a classification task. Motivated by this study, we formulate the geospatial prediction into the following objective function:

$$\min_{\substack{f_l = y_l \\ \forall l \in M}} \sum_{i,j}^{m} S_{ij}(f_i - f_j)^2, \tag{4.6}$$

where y_i and f_i are the known and computed values at location i, respectively. Let $y = (y_1, y_2, \cdots, y_m)$, be the vector of observed values, and $\hat{f} = (f_{m+1}, f_{m+2}, \cdots, f_n)$ be the vector of computed values at the unobserved locations. For simplicity of notation, let $f = (f_1, f_2, \cdots, f_m, f_m + 1, \cdots, f_n)$ be the vector of computed values. In the matrix multiplication form, the objective function can be rewritten as

$$\min_{\substack{f_l = y_l \\ \forall l \in M}} f^T L f, \tag{4.7}$$

where $L = D - S$ is the Laplacian matrix. If we split the Laplacian matrix into matrix blocks, we obtain $(y, \hat{f}) \begin{bmatrix} L_1 & L_2 \\ L_2 & L_3 \end{bmatrix} (y, \hat{f})^T$, where $L_2 \in R^{u \times m}$, $L_3 \in R^{m \times m}$. The solution of Eq. (4.6) is

$$\hat{f} = -L_2^{-1} L_3 y. \tag{4.8}$$

Compared to Kriging, this solution shows that when conditioned on y, $\hat{f}$ is sampled from a multivariate normal distribution. In fact, this graph-based semi-supervised learning is equivalent to Kriging based on a fixed covariance matrix [126]. It can also be represented as a local linear approximation.

$$\hat{f}_i = \sum_j S_{ij} y_j / D_{ii}. \tag{4.9}$$

The accuracy of graph-based methods heavily depends on the graph [135] because the structure of the manifold of data is presented by the graph. We propose an adaptive model for capturing the manifold structure. We use the same similarity matrix S_{ij}. Parameters $\{\sigma_d\}_{d=1}^p$ are optimized in the model by

$$\min_{\{\sigma_d\}_{d=1}^p} \sum_{i=1}^n ||x_i - \frac{1}{D_{ii}} \sum_{j \in \mathcal{N}_\rangle} S_{ij} x_j||. \tag{4.10}$$

We refer to this model as **Parameters Optimized Graph Regression** (POGR).

The optimal solution of this model can be obtained by gradient descent. Because the objective function is not convex, it may not converge to the global minimum. The minimum value depends on the initialization of σ_d. One of the ways to overcome the problem is to initialize σ_d for a number of times and select the best solution. We use the same methods introduced in Chapter 2 to generate spectral features from data. We use Algorithm 2 to solve the spatial prediction problem.

Algorithm 2 Multiresolution Graph-based Prediction

Data:X: a collection of d data sources, $X = (x_1, \cdots, x_d)$, Y: prediction target data
Input: Basis function $B_m(\omega)$, K-nearest neighbor in computing similarity function
Output: Interpolated Y
Step 1: Spectral Feature Extraction
1. Integrate data into uniformed space-time grid cells
2. Compute spectral features for each data source which is completely observed in grid cells using Equation (2.3)
Step 2: Learning POGR
1. Initialize $\{\sigma_d\}_{d=1}^{p}$
2. Find the best structure of the graph using Equation (4.10)
Step 3: prediction
Interpolate using Equation (4.8)

The proposed framework reduces computation considerably when compared to the studies in the related works. Because the number of edges is k, the similarity matrix can be computed by $\mathcal{O}(npk)$. The graph is optimized by $\mathcal{O}(np^2k)$. The derivative of spectral features is $\mathcal{O}(n \log n)$. Thus, the entire process is $\mathcal{O}(np^2k)$. In practice, k is usually a small number such as $k = 10$, thus, this framework is much more computationally convenient when compared to $\mathcal{O}(n^3p^3)$ in [13].

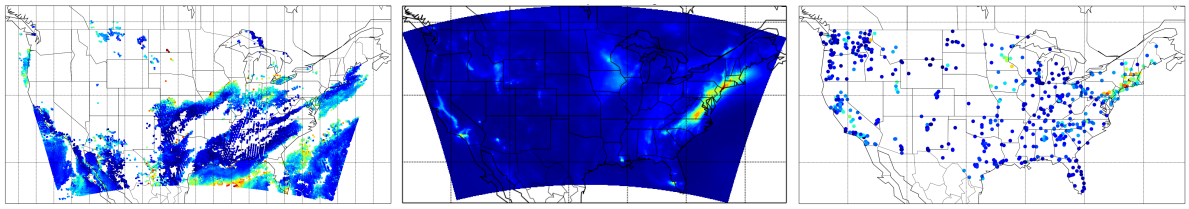
4.4 Experiments

To show the effectiveness of our framework, we apply it to daily PM2.5 concentration predictions.

We use the same data in Chapter 2. Data is from three main sources: EPA observation stations, Community Multiscale Air Quality Modeling System (CMAQ) simulations, and 550 nm satellite imagery of the Aerosol Optical Depth (AOD) from the MODIS Terra satellite. The EPA provides hourly PM2.5 concentrations (in $\mu g/m^3$) as measured by over 20,000 air monitoring stations across the United States. We use average daily values from all stations that report at least once a day. CMAQ is an open source chemical transportation model developed by EPA. We use daily estimates of this system.

MODIS satellite images are originally made available in $10km \times 10km$ grids, for days that the satellite passed over a point. The grid has a different orientation for every image, and its availability, is neither consistent over space, nor over time. The visualizations of three data sources are shown in Figure 4.1.

We model daily PM2.5 concentration for the year 2008 across the contiguous 48 states of the United States. Spatially, we use the $12km \times 12km$ CMAQ grid as our underlying grid, as this grid stays the same across all days, and has complete measurements over the spatial domain. In the following sections, all cells refer to cells in that CMAQ grid. For each cell, all data sources that intersect with it, are averaged using inverse squared weighting on the distance from the cell center. More concretely, let the value of the data source X at location l and time t be $X(l,t)$. Let the set of all locations with data in X at time t that lie in cell i with center p_i be L, and r_{ab} be the spatial distance between locations a, b. Then, the value



(a) MODIS AOD image (b) CMAQ Model Estimates (c) PM2.5 Stations

Figure 4.1: Values of various data sources across United States, plotted on $12km \times 12km$ grid

of X at cell i at time t is given by

$$X(p_i, t) = \sum_{l \in L} \frac{1}{r_{lp_i}^2} X(l, t).$$

Temporally, we average all values from each data source to daily values, which is a common temporal scale used in atmospheric sciences.

4.4.1 Effectiveness of Features

The objective of the experiments is to evaluate the proposed framework, particularly, in two aspects: (1) whether using spectral features can improve predictions and (2) whether hyper-parameter learning is efficient at improving the structure of the learned graph. For this purpose, we extract three sets of features for each cell, from our data sources:

1. Aerosol Optical Depth (F_a)
2. CMAQ daily PM2.5 estimates (F_c)
3. Spectral features from CMAQ estimates (F_s)

Spectral features are computed in Step 1 of Algorithm 2, and Bernstein basis polynomials

Table 4.1: Performance and standard deviation of CMAQ, Ordinary Kriging, Linear Regression(LR), Spatial Gaussian Process (SGP), and the proposed model (POGR) on PM2.5 prediction for 20 runs.

Model (Features)	**MSE**	**STD**
CMAQ	55.09	5.44
Ordinary Kriging	33.42	4.02
LR ($F_c + F_a + F_s$)	29.94	4.22
SGP (F_c)	30.38	1.97
SGP ($F_c + F_s$)	23.74	1.93
SGP ($F_c + F_a + F_s$)	21.62	2.41
POGR (F_c)	26.01	2.17
POGR ($F_c + F_s$)	21.12	2.72
POGR ($F_c + F_a + F_s$)	**19.97**	**2.56**

are chosen as the basis functions for Equation (2.3):

$$B_m(\omega) = B_m(\delta) = \binom{M-1}{m-1} t^{m-1}(1-t)^{M-m},$$

where $t = \frac{\delta}{2\pi} \in [0, 1]$. We use $M = 8$ spectral features. In Step 2, the best result from 20 random initializations is selected. We use 2-fold cross validation by using half of the cells with PM2.5 stations for training, and then test on the remaining half. We use mean square error (MSE) to measure the differences between interpolated values and station measurements. The best results of 20 runs of each model are reported in Table 4.1. There is only one parameter, k, in the framework when constructing the similarity graph. The effect of k on the performance of the graph based models and SGP [110] is shown in Figure 4.2. We see that the proposed model consistently outperforms SGP for the same feature set plotted in the same color in the figure. We also notice that the performance of the model is robust when k is greater than 7.

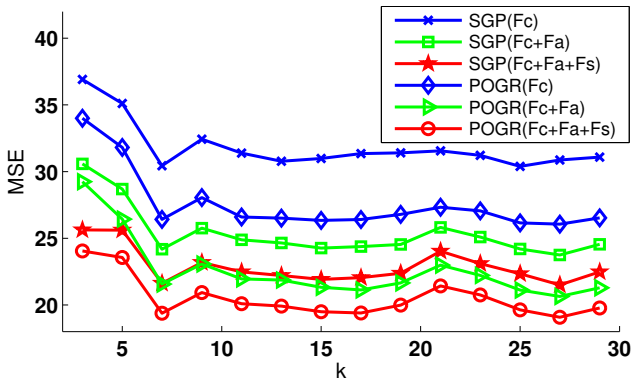


Figure 4.2: Effects of k on graph-based methods with different features.

4.4.2 Parameter Analysis

With respect to the MSE, for a given k, the proposed framework outperforms all other compared methods. For example, it is 63% better than just the chemical transportation model CMAQ, and 7.6% better than the Spectral Spatial Gaussian Process from [110]. More importantly, the model performs around 18.8% better when trained with the spectral features ($F_c + F_s$) than when trained without them (F_c). In agreement with [110], the performance also improves when the model is trained with all the data sources. Therefore, these results clearly demonstrate the effectiveness of POGR in learning form the latent data structure and solving the prediction problem.

4.5 Discussion

We propose a multimodal geospatial interpolation framework. This framework integrates heterogeneous data into a unified space-time-value format and represents the data as a spatial grid with time-dependent thematic data. It provides a feature extractor and predictive model to efficiently generate multiscale features and predict values at unobserved locations. We demonstrate the effectiveness of the proposed framework using air pollution data. This

graph-based predictive model can also be applied to other domains in order to estimate the latent spatial process. Examples of such areas are traffic condition estimation, city noise pollution prediction, and meteorological interpolation.

Chapter 5

A Noise Resistant Graph Embedding for Photo Clustering

5.1 From Photo Clustering to Situation Recognition

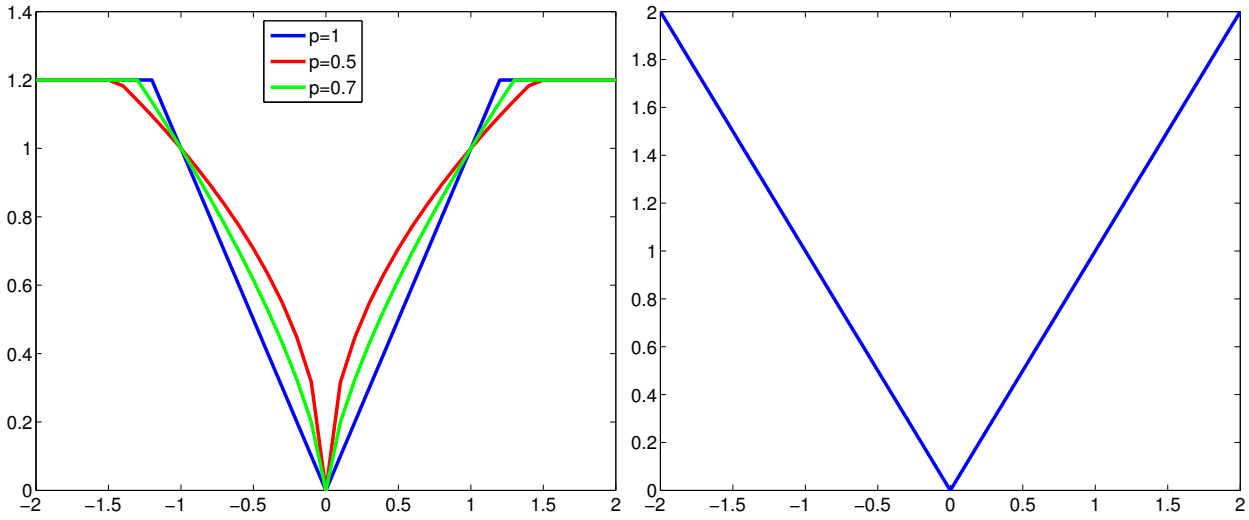
To understand constantly evolving situations, data is collected from disparate sources, including static sensors, social media, cyberspace interactions, Internet-of-Things, and people using effective participatory sensing. We have discussed using data from static physical sensors for situation recognition. In the following chapters, we will discuss how, given the wide use of mobile technology, humans can also act as sensors and contribute their measurements or observations for situation recognition by carrying and using smartphones. In the last decade, Twitter emerged as the most popular participatory sensing mechanism for finding information about events as well as understanding situations and trends. However, obtaining objective information from tweets has been a problem. As the web continually becomes more visual [49], photos serve dual roles. They are important for capturing, saving, sharing, and remembering memories of events and people. Moreover, photos have become a crucial

objective resource for capturing events [90]. They have been used for emerging applications, such as event recognition [122], trend analysis [56], and cultural dynamics [45]. We will explore the possibility of using geospatial streams of photos and their context for situation recognition. A situation model can be defined as a bag of photo visual concepts in a volume of space and time. By computing the similarity between the situation model with evolving bags of visual concepts, similar situations can be detected. Thus, the situation detection problem is transformed into a photo clustering problem. To effectively transform citizen-contributed photos to evolving situations, we must mitigate noise in the data and real-time photos. Thus, we present approaches to remove the adverse effects of outliers for situation recognition in Chapter 5. In Chapter 6, we formulate the real-time data problem into an out-of-sample data problem and propose a graph-based regularized method to address the problem. In Chapter 7, based on the approaches we develop, we explore evolving situation recognition using photos.

5.2 Photo Clustering with Outliers

Social events and trends are usually detected by concurrences of photo visual concepts. For example, "Music Festival" is always associated with concepts of people, instruments, stages, and light, etc., evolving in a certain pattern. The event model of Music Festival can be defined as bag of these photo visual concepts. By computing the similarity between the Music Festival model with evolving bag of visual concepts, similar events can be detected. Detection of events is then transformed into a photo clustering problem, in which each cluster represents an event [90].

In the context of event and trend detection, photo clustering assigns photos to groups which share the same semantic concepts with the contents of the photos. As an important topic in multimedia and vision, clustering analysis has been well studied. Many approaches from



(a) capped ℓ_p-norm loss with $\theta = 1$ (b) ℓ_1-norm loss

Figure 5.1: Comparison of capped ℓ_p-norm and ℓ_1-norm

various angles have been proposed for efficient clustering analysis, such as K-means, support vector clustering, and spectral clustering. Among these approaches, spectral clustering which has shown to have the state-of-the-art performance, models data into an efficient manifold. For instance, ratio cut clustering can be approximated by Laplacian embedding [16], and normalized cut clustering can be approximated by generalized eigenvectors of a Laplacian matrix [101]. Thus, graph-based learning provides an efficient way for modeling data in clustering problems. An important advantage of working with a graph is its capacity to incorporate diverse types of information, such as the relationship between data points [109]. Some of the existing graph-based models use a quadratic form of graph embedding. However, a major drawback is that the results are sensitive to outliers. A robust graph-based unsupervised learning method can alleviate this drawback through manipulation of metrics in noisy data. Among these methods, an important improvement is ℓ_0-norm. The ℓ_0-norm of vector F is defined as $\|F\|_0 = |\{i : f_i \neq 0\}| = n$, where n represents the number of important variables in the measurement. These important variables actually play roles in clustering. Models using this norm have demonstrated effective performance [128]. Unfortunately, computational expense increases when subset selection problems find the number of subset of cardinality

n. To reduce the computational burden, the ℓ_0-norm has been replaced by some relaxations which approximate it. For example, ℓ_1-norm is a standard convex relaxation for ℓ_0-norm, in which ℓ_1-norm based objective functions lead to sparse and direct clustering results [129]. Other studies also use a $\ell_{2,1}$-norm based objective function to reduce the impact of outliers [79]. Still these methods suffer from extreme data outliers. Thus, the performance of event and trend detection using these methods is also easily influenced by noisy outliers.

We propose a novel noise resistant model. This model learns an efficient graph embedding by utilizing capped ℓ_p-norm to remove data outliers. As shown in Figure 5.1, the capped ℓ_p-norm loss function is $\min(|x|^p, \theta)$, where $0 < p \leq 2$. This function gives the value of $|x|^p$, when $|x|^p < \theta$, and θ, otherwise. The capped ℓ_p-norm effectively removes outliers and noise beyond the parameter θ. We use this property in projecting data into a manifold where the similarity of data points is adjusted without input from outliers. If the distance of a label vector is large, the corresponding similarity value is not updated. Additionally, an efficient iterative algorithm is adopted to solve the optimization problem. To verify the effectiveness and efficiency of the proposed model, we apply it to four benchmark datasets and Yahoo Flickr Creative Commons data. We show that this proposed method outperforms state-of-the-art approaches.

5.3 Capped Norm Graph Embedding

In this section, we present the noise resistant capped ℓ_p-norm unsupervised learning model. First, we introduce graph Laplacian unsupervised learning. Second, we give the formulation of the proposed model, followed by the optimized solutions.

5.3.1 Problem Formulation and Motivation

Suppose we have n data points $\{x_1, x_2, \cdots, x_n\} \in \mathcal{R}^{d\times 1}$. The main task of clustering is to group these n data points into c different clusters. The cluster assignment matrix is denoted by $F = [f_1, \cdots, f_n]^T \in \{0,1\}^{n\times c}$, where f_i is the cluster assignment vector for the data point x_i. The j-th element of f_i is 1 if x_i belongs to cluster j, and is 0, otherwise. Given the data, a similarity matrix $W \in \mathcal{R}^{n\times n}$ can be defined. The element w_{ij} reflects the affinity of x_i and x_j, which is the ij-th entry of matrix W. The Laplacian matrix L is defined as $L = D - W$, where D is a diagonal matrix with diagonal elements, $D_{ii} = \sum_j w_{ij}, \forall i$. $Tr(L)$ id denoted as the trace operator of matrix L. The multi-way graph ratio-cut clustering is formulated as

$$\min_F Tr(F^T L F). \tag{5.1}$$

The minimization is not a trival problem because the cluster assignment matrix is constrained to be an integer solution. With the relaxed continuous clustering assignment matrix $F \in \mathcal{R}^{n\times c}$, the relaxation form of the multi-way graph ratio-cut clustering is written in the following form:

$$\min_{F^T F=I} \sum_{i,j=1}^{n} w_{ij} \left\| f^i - f^j \right\|_2^2. \tag{5.2}$$

Another similar spectral clustering is Normalized Cut, and it can be written as

$$\min_{F^T DF=I} \sum_{i,j=1}^{n} w_{ij} \left\| f^i - f^j \right\|_2^2. \tag{5.3}$$

The ideal solution of F is $f^i = f^j$ if x_i and x_j are in the same cluster. In this case, rows of F are equal, and thus F represents the clustering structure. These two graph-based models use a quadratic form of graph embedding, thus the results are sensitive to outliers. In order to overcome this problem, more robust graph-based unsupervised learning methods

are proposed to handle such noisy data. Because ℓ_1-norm is more robust to outliers, some studies build models on top this norm. In [80], the authors propose a ℓ_1-norm of spectral embedding in the following form:

$$\min_{F^T DF=I} \sum_{i,j=1}^{n} w_{ij} \left\| f^i - f^j \right\|_2 . \tag{5.4}$$

However, this model still suffers from the effect of extreme outliers. Besides that, a ℓ_1-based model is not stable compared to ℓ_2-based model [125].

5.3.2 The Graph-based Capped Norm Method

Capped norm based loss functions have been used for various purposes. For example, The capped ℓ_1-norm is used for sparsity regularization [131]. The capped $\ell_{1,1}$-norm is proposed for multi-task learning [36]. Here, we use the ℓ_p-norm as a robust and stable loss function to resist outliers and propose a new graph embedding model using this norm.

The capped norm based loss function of a c-dimensional vector $u \in \mathcal{R}^{1\times c}$ is $\min(\|u\|_2^p, \theta)$, where θ is a parameter. The value of this loss function is $\|u\|_2^p$, if $\|u\|_2^p$ is smaller than θ, and is θ, otherwise. This loss function is more robust when there are outliers than ℓ_1-norm, because it has threshold θ for outliers.

The proposed capped ℓ_p-norm method is expressed as

$$\min_{F^T F=I} \sum_{i,j} w_{ij} \min(\|f_i - f_j\|_2^p, \theta), \tag{5.5}$$

where w_{ij} is the similarity measurement in the graph between x_i and x_j, $0 < p \leq 2$ and θ is a parameter.

The above formulation benefits from the control of the input data. The clustering results are dependent on the quality of input data graph. Most of the time, they are sensitive to the particular graph construction methods. We overcome this problem in two stages, graph initialization and graph similarity adaptation. Because the objective function is not convex when $p < 1$, it may not converge to the global minimum. The minimum value depends on the initialization. One of the ways to overcome the problem is to select "good" graph initializations. We will introduce a method to give a "good" initialization. For the graph similarity adaptation, if the distance of a label vector is large, then the corresponding similarity value is not updated because of the capped norm.

5.3.3 Optimization Algorithm

In this subsection, we introduce the optimization algorithm to solve Problem (5.5). Because the proposed model is a weighted sum of a concave function, it can be solved in a general re-weighted optimization framework [84].

The general re-weighted optimization problem is

$$\min_{x \in \mathcal{C}} f(x) + \sum_i h_i(g_i(x)), \tag{5.6}$$

where $h_i(x)$ is an arbitrary concave function with the domain of $g_i(x)$. This general problem can be solved by Algorithm 3.

Algorithm 3 A General Re-weighted Optimization Algorithm

Initialize $X \in \mathcal{C}$;
Repeat:
Calculate $D_i = h_i'(g_i(x))$
Solve the following problem: $\min_{x \in \mathcal{C}} f(x) + \sum_i Tr(D_i^T g_i(x))$
Until convergence

For Problem (5.5), denote $h(x) = \min(x^{\frac{p}{2}}, \theta)$ and $x = \|f_i - f_j\|_2^2$. Because $h(x)$ is a concave

function with respect to x, the supergradient of $h(x)$ can be obtained as follows:

$$h'(x) = \begin{cases} \frac{p}{2}x^{\frac{p-2}{2}} & x^{\frac{p}{2}} \leq \theta \\ 0 & \text{otherwise} \end{cases}. \tag{5.7}$$

Problem (5.5) can be solved using Algorithm 3, because $h(x)$ is concave. The proposed model can be written in the following form:

$$\begin{aligned} \min_{F^TF=I} \sum_{i,j} w_{ij}h'(x)x &= \min_{F^TF=I} \sum_{i,j} w_{ij}s_{ij} \left\|f_i - f_j\right\|_2^2 \\ &= \min_{F^TF=I} \sum_{i,j} \tilde{s}_{ij} \left\|f_i - f_j\right\|_2^2 \\ &= \min_{F^TF=I} Tr(F^T L_{\tilde{s}} F), \end{aligned} \tag{5.8}$$

where $\tilde{s_{ij}} = w_{ij}s_{ij}$.

In the objective function, because an initial graph similarity matrix $W \in \mathcal{R}^{n\times n}$ is required and the construction of the graph plays key a role in the performance of graph-based clustering. In order to obtain an accurate graph, we use the following Constrained Laplacian Rank method [81] to generate initial graph for the proposed model. The model is

$$\min_{w_i^T\mathbf{1}=1, w_i\geq 0, w_{ii}=0} \sum_{j=1}^{n} \left\|x_i - x_j\right\|_2^2 w_{ij} + \lambda \sum_{j=1}^{n} w_{ij}^2,$$

where w_i is the i-th row of similarity matrix W. If we denote $e_{ij} = \left\|x_i - x_j\right\|_2^2$ and use the Lagrangian method, the optimal similarities can be obtained:

$$\hat{w_{ij}} = \begin{cases} \frac{e_{i,m+1} - e_{ij}}{k e_{i,k+1} - \sum_{m=1}^{k} e_{ik}} & j \leq k \\ 0 & j > k \end{cases}, \tag{5.9}$$

where k is the number of neighbors. Computing the similarity matrix always involves parameters, such as Gaussian functions. The above method has only one integer parameter: the number of neighbors, which is easier to tune. Besides that, because of the simplicity of computing $\hat{w_{ij}}$ compared to Gaussian functions, it can also be applied to construct large-scale graphs.

According to Algorithm 3, we have the following Algorithm 4 to solve Problem (5.5).

Algorithm 4 Graph-based Capped Norm Optimization

Initialize s_{ij} and set $s_{ij} = 1$;
Construct graph for $w_{ij} \in W$ using Equation (5.9).
Repeat:
Calculate $\tilde{s}$ by $\tilde{s} = w_{ij}s_{ij}$
Update F by $\min_{F^TF=I} Tr(F^T L_{\tilde{s}} F)$;
Calculate $s_{ij} = \begin{cases} \frac{p}{2} \|f_i - f_j\|_2^{p-2} & \|f_i - f_j\|_2^p \leq \theta \\ 0 & \text{otherwise} \end{cases}$
Until convergence

We apply the iterative algorithm to solve the optimization Problem (5.5). One of the most important aspects of the iterative algorithm is convergence. Because the objective function is fit into the general optimization problem. The general optimization problem has been proven to converge [84]. Thus, Algorithm 4 is guaranteed to converge.

5.4 Experiments

In this section, we perform experiments on photo data sets to demonstrate the effectiveness of the proposed approach. For simplicity, we denote the proposed capped ℓ_p-norm graph embedding method as CGE. The proposed model is compared with K-means, Ratio Cut [16], Normalized Cut [101] and ℓ_1-norm graph [80].

5.4.1 Data Description

Four benchmark data sets, COIL20 [76], BinAlpha [18], USPS [66] and MNIST [74], and one recent released multimedia data set YFCC100M [117] are used.

BinAlpha data contains 26 binary hand-written alphabets and 39 photos for each alphabet.

USPS is a handwritten digit database. We randomly select 100 photos for each digit, and use 1000 photos in total for our experiments.

COIL20 contains different objects imaged at different angles in a 360 rotation.

MNIST is a handwritten digit database. We randomly sample 200 photos from each group, and 2000 photos in total for the experiments.

YFCC contains around 99.2 million photos and nearly 800,000 videos from Flickr. For the purpose of our experiments, we sample 3000 photos from ten concepts, "architecture", "plant","sport", "car", "food", "animal", "bike", "field game", "landscape" and "cloud." In addition, we use CaffeNet Hybrid-CNN to generate 4096 features for the experiments [10].

A summary of the data is shown in Table 5.1.

5.4.2 Experimental Settings

In this section, we present an experimental study on various unsupervised methods such as K-means, ratio cut, normalized ratio cut, graph-based ℓ_1-norm model, and the proposed model for clustering.

There are two parameters, θ and p in CGE. We cross validate to select the best parameter combination from the range $\theta = (0.1, 1, 10, 100, 1000)$ and $p = (0.5, 0.7, 1, 1.5, 1.7, 2)$.

Table 5.1: Data Descriptions

Data set	# sample	# feature	# class
USPS	1000	784	10
BinAlpha	1014	320	36
COIL20	1440	1024	20
MNIST	2000	784	10
YFCC	3000	4096	10

The clustering algorithm K-means is used to obtain the final clustering results. Because the results of K-means depend on the initializations, we repeat K-means with 100 random initializations and report the results with the best objective values. We set K as the number of classes in the data sets. Clustering accuracy and normalized mutual information (NMI) are used to evaluate the performances. Accuracy is defined as

$$accuracy = \frac{1}{n}\sum_{j=1}^{n}\delta(l_j, p_j),$$

where for data point j, l_j is the ground true label and p_j is the label assigned by the clustering results, and δ is 1 if l_j and p_j are the same, otherwise 0.

The NMI of two clusters C_1 and C_2 is defined as

$$NMI(C_1, C_2) = \frac{MI(C_1, C_2)}{max(H(C_1), H(C_2))},$$

where $H(C) = -\sum_{c_i \in C} p(c_i)log(p(c_i))$, and $MI(C_1, C_2) = \sum_{c_i \in C_1, c_j \in C_2} p(c_i, c_j)log\frac{p(c_i, c_j)}{p(c_i)p(c_j)}$, where $p(c_i)$ is a probability of the data point belongs to cluster c_i.

5.4.3 Results

Next, we evaluate the performance of the proposed model on four benchmark data sets by comparing it with other established models. Accuracy and normalized mutual information

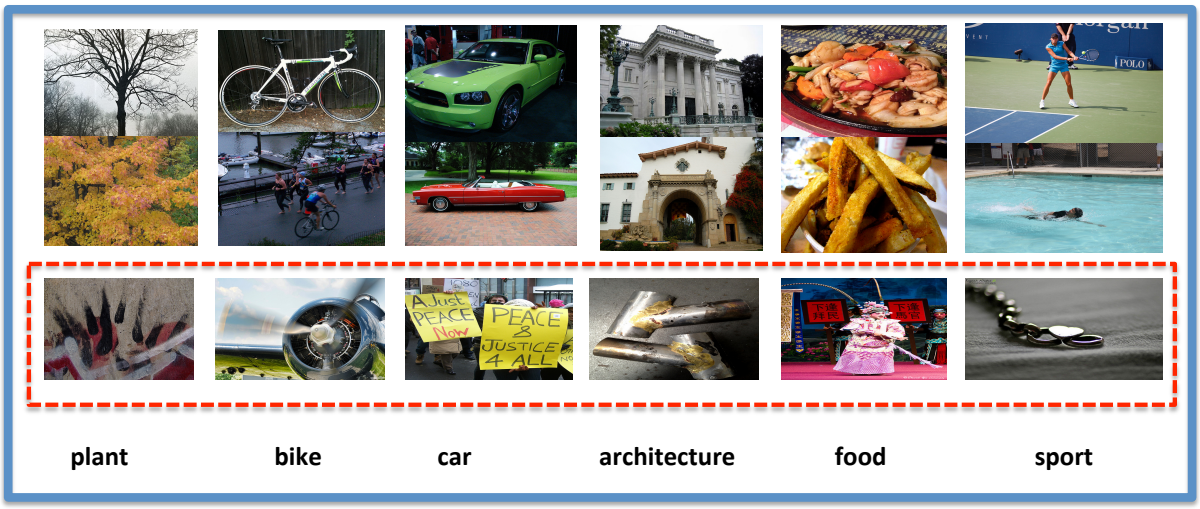


Figure 5.2: Examples of outliers in the YFCC data

Table 5.2: Clustering accuracy using K-means, Rcut [16], Ncut [101], L1G [80] and the proposed model CGE on five data sets.

Data	Accuracy				
	K-means	Rcut	Ncut	L1G	CGE
COIL20	0.6299	0.7896	0.8174	0.8229	**0.8451**
USPS	0.5910	0.6330	0.6220	0.6320	**0.6420**
BinAlpha	0.4131	0.5164	0.4986	0.4466	**0.5321**
MNIST	0.6015	0.6660	0.6605	0.6380	**0.6825**
YFCC	0.6534	0.7259	0.6099	0.5789	**0.7827**

Table 5.3: Clustering NMI K-means, Rcut [16], Ncut [101], L1G [80] and the proposed model CGE on the five data sets.

Data	Normalized Mutual Information				
	K-means	Rcut	Ncut	L1G	CGE
COIL20	0.7704	0.8950	0.8960	0.9052	**0.9136**
USPS	0.5827	0.7351	0.7136	0.7407	**0.7470**
BinAlpha	0.5758	0.6457	0.6107	0.6054	**0.6537**
MNIST	0.5042	0.6530	0.6508	0.6160	**0.6608**
YFCC	0.5940	0.6821	0.5486	0.5478	**0.6910**

as listed in Tables 6.2 and 6.3 are utilized to measure the differences between the labels predicted by each model and the benchmark labels actually observed. With respect to the accuracy, the proposed model CGE outperforms all other compared methods. For example, compared with the baseline Normalized Ratio Cut, CGE reduces testing error by 3.39%, 3.22%, 6.72% and 3.33%, respectively, on the four data sets. These results demonstrate the effectiveness of CGE in learning the intrinsic manifold projection. Similar to the performance for the accuracy, CGE holds promise for enhancing NMI accuracy as well. Compared with its baseline Normalized Ratio Cut, the proposed model reduces error by 1.96%, 4.68%, 7.04% and 1.54%, respectively, as presented in Table 6.3, which further proves the effectiveness of our method in learning the intrinsic manifold projection.

We also experiment with data with extreme outliers, in particular, YFCC data. Each photo in YFCC is given a concept label; however, some photos are mislabeled. We call these photos outliers. Figure 5.2 shows an example of outlier photos from each concept in a box with red dash border. It is very clearly shown that outlier photos are irrelevant to the concept cluster they belong to. Regarding the established methods, the L1G method obtains lower accuracy than K-means, and is not robust on YFCC. The main reason is that L1G is defined on the ℓ_1-norm. Theories and experiments in [125] show that the ℓ_1-based model cannot be stable. The proposed method leverages both stability and robustness to outliers to improve clustering of photos by using a capped ℓ_p-norm. For example, compared with the baseline Ratio Cut, CGE reduces clustering error by 7.82%.

5.5 Discussion

We present a new graph-based capped ℓ_p-norm model for photo clustering. It utilizes the properties of capped norm to remove the effect of outliers, and explore the underlying data structure in a manifold to improve the performance of clustering. We develop an efficient iterative algorithm to solve the optimization problem for the proposed objective function. This algorithm updates the similarity matrix in an adaptive way by ignoring the outliers and noise and performs effective data mapping in the nonlinear manifold. The experiments on photo data demonstrate the effectiveness of the proposed method. A stream of photos from different people and different places creates a mosaic of world wide events as they happen. Accurate photo clustering will facilitate situation recognition in emerging applications on the Internet.

Chapter 6

A Graph-regularized PCA for Out-of-sample Photo Clustering

6.1 Introduction

A significant trend in mobile technology has been the dominance of real-time user generated data. Users upload data to the Internet for sharing their thoughts and exchanging information. To effectively use the information, the real-time data must be processed in time which requires that data mining algorithms can accommodate an ever increasing data load instantaneously. In the real-time data model, algorithms must process high speed data under constraints such as computation time. Consequently, data streams pose several challenges for machine learning algorithms. First, the algorithms consider limited time for quick response. Second, they must deal with the fact that distributions of the data can change over time. In this chapter, we propose a novel model to address the out-of-sample problem in the real-time data setting.

In last chapter, we have discussed proposed graph-based methods for addressing the noisy

data problem. To effectively transform citizen contributed photos into evolving situations, we must deal with both noisy and real-time photos. The computation or processing large volumes of real-time data poses a challenge for typical machine learning models. However, conventional data mining algorithms operate in a batch mode that requires to have all the data at once. The models can work well on off-line data, but they cannot be directly applied to real-time data. To perform real-time photo clustering, we develop new models in this chapter. Many techniques have been developed for handling offline photos for clustering. In clustering, dimension reduction is a necessary step because data is usually in high dimension. A great number of models for dimension reduction have been proposed. Principal Component Analysis (PCA) [57], which is the linear projection of data along the direction of maximum variance, is popular for its simplicity and effectiveness. However, data is not always in a linear manifold. To discover the intrinsic manifold structure of data, many non-linear dimension algorithms were developed. These algorithms, such as Isomap [116], Locally Linear Embedding(LLE) [98], and Laplacian Eigenmap [8], assume that sampled data lies in a low dimensional nonlinear manifold. A framework of sparse unsupervised dimension reduction for multiple view data is also proposed in [40].

However, these algorithms are not suitable for updating the nonlinear low-dimensional projections when the data is in real-time. New data is treated as out-of-sample data and must be processed. Locality Preserving Projection (LPP) [42] and Local Learning Projection (LLP) [124] learn a linear projection function such that the function can be applied to the out-of-sample data. A framework for extending LLE, Isomap, and other unsupervised learning methods is also introduced in [9]. Flexible Manifold Embedding (FME) [83] is a unified manifold learning framework for unsupervised dimension reduction by using a linear regression function.

Recently, models have been developed for learning a low-dimensional data representation that incorporates the clustering information encoded in graph data. Among them, several

models [133, 64, 129] have been proposed to incorporate the intrinsic nonlinear dimension reduction into a linear low-rank factorization. The idea is to add a regularization term to the non-negative matrix factorization model in order to find a part-based representation of the data in which the neighborhood connections are preserved. When vector data and graph data are available, [55] proposed a graph Laplacian PCA to learn a graph embedded low dimensional representation of data. The new representation includes the clustering information from graph data. The goal of these models is to provide a data representation that takes into account the intrinsic nonlinear structure of data.

Nevertheless, neither low-rank matrix approximation with manifold regularization nor graph Laplacian PCA are capable of handling the out-of-sample data. Instead, both of them have to recompute data representations upon the input of new data, which leads to high computational cost. In order to solve this problem, we propose a new graph-regularized PCA. Similar to the idea in [41] that a model can learn a mapping function for out-of-sample data, the new model learns an mapping function from original data space to the nonlinear manifold. The out-of-sample problem can be overcome by applying the learned function to these data without recomputing eigenvectors. The learned function is obtained by an efficient iterative algorithm for optimizing the model. Moreover, we use the theoretical analysis to reveal that the model closely related to the unsupervised Linear Discriminant Analysis (LDA).

6.2 Related Models

Principle Component Analysis projects a data matrix, $X = (x_1, x_2, \cdots, x_n) \in R^{p \times n}$, onto an optimal low-dimension (k-dim) subspace defined by the principal directions, $U = (u_1, u_2, \cdots, u_k) \in R^{p \times k}$. The projected data set in the new k-dim subspace is $V^T = (v_1, v_2, \cdots, v_n)$. PCA

searches for U and V via the following minimization process:

$$\min_{U,V^TV=I} \left\|X - UV^T\right\|_F^2 s.t. V^TV = I. \tag{6.1}$$

PCA is limited to embedding data in the linear regime. However, in many applications, data lies in a non-linear manifold, for which the graph Laplacian-based embedding is valid. Given the pairwise similarity data matrix $S \in R^{n\times n}$, which contains the edge weights on a graph with n nodes, the embedding coordinates of the n data points, $Q^T = (q_1, q_2, \cdots, q_n) \in R^{k\times n}$, can be obtained by minimizing

$$\min_{Q^TQ=I} \sum_{i,j=1}^{n} \|q_i - q_j\|^2 S_{ij} = Tr(Q^T(D-S)Q), \tag{6.2}$$

where $D = diag(d_1, d_2, \cdots, d_n)$ and $d_i = \sum_j S_{ij}$. Laplacian embedding preserves the local geometrical relationship and maximizes the smoothing of a low embedding space.

Graph regularized PCA (glPCA) [55] incorporates graph structures encoded in S into a low dimensional representation of data and gives a closed-form solution to the non-convex problem:

$$\min_{V^TV=I,U} \left\|X - UV^T\right\|_F^2 + \gamma Tr(V^TLV), \tag{6.3}$$

where $L = D - S$.

Matrix approximation with Manifold learning (MMF) [133], another example of Laplacian embedding, is a low-rank factorization model used to approximate the data matrix and also preserves the nonlinearities of the data.

$$\min_{U^TU=I_r,V} \left\|X - UV^T\right\|_F^2 + \gamma Tr(V^TLV). \tag{6.4}$$

Despite their merits, neither algorithm is suitable for updating the nonlinear low-dimensional projections when the new data is involved.

6.3 The Proposed Model

To solve this problem, we develop a model extending graph Laplacian and present a straightforward solution for out-of-sample data. This model not only solves the out-of-sample problem but also performs dimension reduction and cluster optimization together.

Suppose $X \in R^{p\times n}$ is the centered training data in the original space, where the mean values across n samples are $0 \in R^{p\times 1}$. X is labeled by c classes. A linear transformation is applied to the original data, and the data is mapped to a new space of r dimensions. By minimizing the objective function, we obtain $W \in R^{p\times r}$, which can be used to map the out-of-sample data $\bar{X}$ into $W^T\bar{X}$, which is the dimension reduced data matrix. The new model combines both the linear transformation and the nonlinear dimension reduction. We call the model graph regularized linear regression PCA (grlrPCA):

$$\min_{\substack{W^TS_tW=I\\U,V^TV=I}} \left\|W^TX - UV^T\right\|_F^2 + \gamma Tr(V^TLV), \tag{6.5}$$

where $S_t = XX^T$, $U \in R^{r\times c}$, $V \in R^{n\times c}$.

Unlike [55] and [133] where the data matrix is factorized into V, i.e., the projected data, our model treats V as a continuous solution of the cluster membership indicators in the K-means clustering method, as it is treated in [24]. In fact, the discretization of $V \in R^{n\times c}$ is the label indicator matrix of c classes. Next, we introduce how to derive an iterative algorithm to solve this optimization problem.

Taking the derivative of (6.5) with respect to U, and setting the derivative to zero, we have

$$W^TXV - UV^TV = W^TXV - U = 0.$$

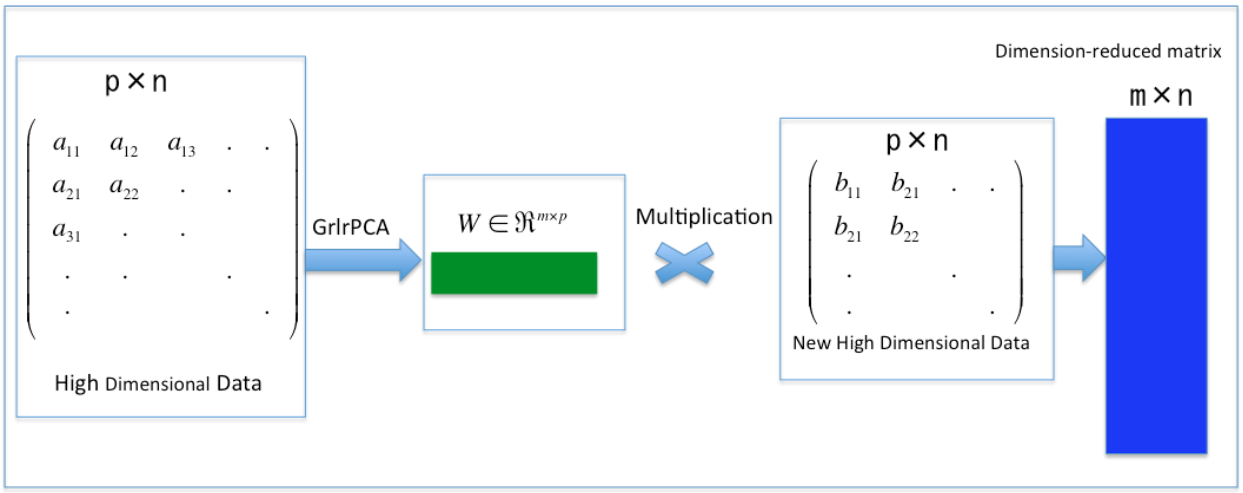


Figure 6.1: Flowchart of the proposed method

U can be expressed by W, X, and V in the following form,

$$U = W^T X V.$$

Substituting U with W and V in (6.5), the optimization problem is

$$\min_{\substack{W^T S_t W = I \\ V^T V = I}} \left\| W^T X - W^T X V V^T \right\|_F^2 + \gamma Tr(V^T L V). \tag{6.6}$$

When V is fixed, the original optimization problem is

$$\min_{W^T S_t W = I} Tr(W^T X (I - V V^T) X^T W).$$

When W is fixed, we will solve the following problem:

$$\max_{V^T V = I} Tr(V^T (X^T W W^T X - \gamma L) V).$$

The optimization procedure is described in Algorithm 5. The two-step algorithm alternately optimizes the model parameters W and V.

As Figure 6.1 shows, the out-of-sample problem is handled by computing a learned coefficient matrix W, using training data. When the new data points come in, multiplying the new data matrix by W will result in a dimension-reduced data matrix.

Algorithm 5 Graph-regularized PCA Iterative Optimization Method

Initialize V;
Repeat:
Update W by $\min_{W^T S_t W=I} Tr(W^T X(I - VV^T)X^T W)$
Update V by $\max_{V^T V=I} Tr(V^T(X^T W W^T X - \gamma L)V)$
Until convergence

6.3.1 Algorithm Convergence Analysis

We adopt a coordinate descent method to find the minimum value of function (6.6), denoted by $F(V, W)$. The process is performed by minimizing the objective function along one direction each time and alternately on V and W. Suppose at the k-th step,

$$W_{k+1} = \arg\min_{W^T S_t W=I} Tr(W^T X(I - V_k V_k^T)X^T W),$$

$$V_{k+1} = \arg\max_{V^T V=I} Tr(V^T(X^T W_{k+1} W_{k+1}^T X - \gamma L)V),$$

where $k \geq 0$. Because $F(V, W) \geq 0$ for any V and W, and $F(V_k, W_k)$ is a decreasing sequence.

$$F(V_k, W_{k+1}) \geq F(V_{k+1}, W_{k+1}) \geq F(V_{k+1}, W_{k+2}) \cdots \geq 0$$

Thus,

$$\lim_{k\to\infty} F(V_k, W_k) = inf\{F(V_k, W_k)\}$$

6.3.2 Connection to LDA

In this subsection, we show that when $\gamma \to \infty$, the resulting optimization Problem (6.5) is closely related to the LDA problem. We summarize it in the following proposition.

Proposition 6.1. *When $\gamma \to \infty$, problem (6.5) is closely related to the problem of LDA, where labels of the data are also optimized as variables.*

Proof. Suppose Y is the label indicator matrix of centered data X, when $\gamma \to \infty$,

$$\min_{\substack{W^T S_t W = I \\ V^T V = I}} \left\| W^T X - W^T X V V^T \right\|_F^2 + \gamma Tr(V^T L V) \tag{6.7}$$

The above optimization problem becomes

$$\min_{W^T S_t W = I} \left\| W^T X - W^T X V_s V_s^T \right\|_F^2, \tag{6.8}$$

where $V_s = [v_1, v_2, \cdots v_c]$, c is the number of clusters and $[v_1, v_2, \cdots v_c]$ are the first eigenvectors corresponding to the first c smallest eigenvalues from matrix L.
In order to derive our target equation, we transform $\left\| W^T X - W^T X V_s V_s^T \right\|_F^2$ into $Tr(W^T X(I - V_s V_s^T) X^T W)$. Because V_s is a continuous solution of Y, $X V_s V_s^T X$ is closely related to $S_b = XYY^T X^T$, the between-class scatter matrix defined in the LDA. Furthermore, the total-class scatter matrix is $S_t = XX^T$, and the within-class scatter matrix is $S_w = S_t - S_b$. With these notations, Eq. (6.8) can be further reduced,

$$\begin{aligned}
&\min_{W^T S_t W = I} Tr(W^T (XX^T - X V_s V_s^T X^T) W) \\
&\approx \min_{W^T S_t W = I} Tr(W^T (S_t - S_b) W) \\
&= \min_{W^T S_t W = I} Tr(W^T S_w W)
\end{aligned}$$

$\min_{W^T S_t W = I} Tr(W^T S_w W)$ is exactly the optimization problem of LDA. So grlrPCA model

is closely related to the unsupervised LDA. □

6.4 Experiments

6.4.1 Dataset Description

To show the effectiveness of the proposed model, we use seven data sets, USPS, BinAlpha [7], ORL [100], YALE [34], MNIST [65], NIST Topic Detection and Tracking corpus (TDT2) [21] and CMU World Wide Knowledge Base (WebKB) [35].

MNIST is a handwritten digit database. We randomly select 500 images, 50 for each digit.

BinAlpha data contains 26 binary handwritten alphabets and 39 images for each alphabet.

USPS is a handwritten digit database. We randomly select 50 images for each digit, and in total 500 images from the database for our experiments.

ORL database consists of 10 images for each of 40 distinct subjects. For some subjects, the images were taken with varied lighting, facial expressions, and facial details.

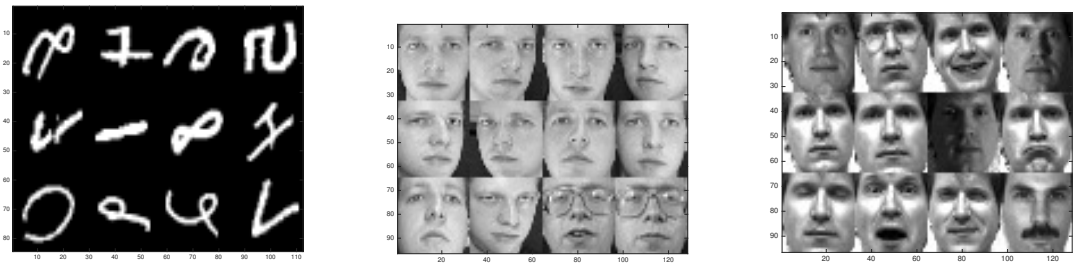
Yale database contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration.

TDT2 consists of data collected from 6 sources, including 2 newswires (APW, NYT), 2 radio programs (VOA, PRI), and 2 television programs (CNN, ABC). In this subset, 21 to 30 topics are kept in order.

WebKB data is Web pages, which was collected from the websites of computer science departments at Texas universities by the World Wide Knowledge Base project.

Table 6.1: Datasets Descriptions

Dataset	# sample	# feature	# class
MNIST	500	784	10
BinAlpha	1014	320	36
USPS	500	256	10
ORL	400	1024	40
Yale	165	1024	15
TDT2	653	36771	10
WebKB	814	4029	7



(a) MNIST (b) ORL (c) Yale

Figure 6.2: Representative images of MNIST, ORL, and Yale data

A summary of the data is shown in Table 6.1, and representative images of the data are shown in Figure 6.2.

6.4.2 Experimental Settings

In this section, we present an experimental study on various unsupervised methods such as PCA, LPP [42], FME [83], MMF [133], and glPCA [55] for dimension reduction and clustering using real-world data sets. We use the Gaussian function to calculate S and L, where the graph similarity matrix is obtained as $S_{ij} = exp(-||x_i - x_j||^2/t)$, if x_i is among the k nearest neighbors of x_j; $S_{ij} = 0$, otherwise. For fair comparison, we set t according to the setting in [119].

In order to avoid a large conditional number when performing eigenvalue decomposition, it

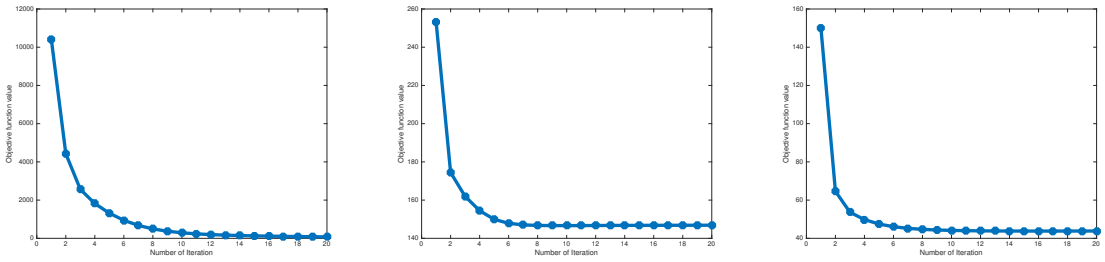
is necessary to preproject the data into a lower dimensional space before applying various models. In this regard, data is first preprocessed using PCA to preserve 95% of energy. FME aims to learn a linear regression to solve the regularized manifold learning problem, which has the following objective function:

$$(F^*, W^*, b^*) = arg \min_{\substack{F,W,b \\ F^T V F = I}} Tr(F^T M F) + \mu \|W\|^2 + \gamma \left\| X^T W + \mathbf{1} b^T - F \right\|^2 .$$

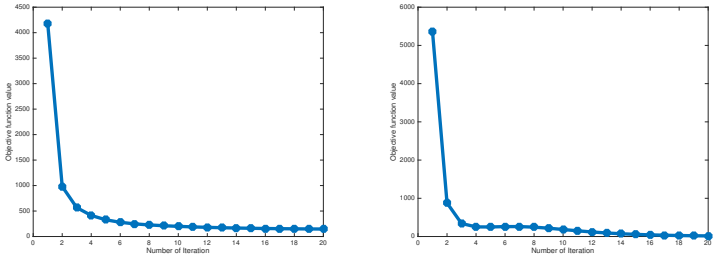
There are two parameters, μ and γ, in FME. We perform cross validation to select the best parameter combination from the range $(10^{-5}, 10^{-4}, 10^{-3}, \cdots, 10^3, 10^4, 10^5)$, which is the same as [119]. The number of combinations of μ and γ is 121. In addition, glPCA, MMF, and the proposed method have one parameter γ. The parameter γ is set within the range $(10^{-5}, 10^{-4}, 10^{-3}, \cdots, 10^3, 10^4, 10^5)$. The rank r in MMF is set to the number of classes in the data set, which is the same as [133].

We repeat experiments 20 times for each method on each dataset and report the average results and the standard deviation. For each dataset, we randomly select 50% of the data as the training and the other 50% as testing. Each model is trained on one half of the data, and is tested on the remaining half. We use K-means for evaluation, and set K as the number of classes in the data sets. K-means with random initializations is run 50 times, and the average clustering results are reported. We report the best accuracy from each model.

We use clustering accuracy and normalized mutual information (NMI) to evaluate the performance of our model.



(a) MNIST (b) USPS (c) BinAlpha



(d) ORL (e) Yale

Figure 6.3: Convergence analysis on five data sets

6.4.3 Convergence Study of the Proposed Method

We use an iterative algorithm to solve the optimization problem (6.5). Convergence is one of the most important aspects of iterative algorithms. We have shown in the previous section that this algorithm is theoretically proven to converge. In this subsection, we empirically study the property of convergence of the proposed algorithm. Figure 6.3 gives the objective function value of the proposed method at each iteration on five datasets. As the figure shows, the objective values consistently decrease in the iterative process which is consistent with theoretical analysis. In addition, the algorithm converges to a stable value within 100 iterations. In the experiments, we set 10^{-5} as the stop criterion, which is sufficient for the convergence of the algorithm.

Table 6.2: Clustering average accuracy and standard deviation using K-means on PCA, LPP, FME, and grlrPCA on seven datasets for 20 runs

Data	Accuracy				
	Original	PCA	LPP	FME	grlrPCA
MNIST	0.4283	0.4559	0.4648	0.4713	**0.4885**
USPS	0.6492	0.6852	0.7149	0.7168	**0.7387**
BinAlpha	0.3845	0.4323	0.3934	0.4450	**0.4689**
ORL	0.5310	0.5625	0.5655	0.5835	**0.6036**
TDT2	0.3758	0.3909	0.4263	0.4052	**0.4368**
WebKB	0.4475	0.4458	0.5794	0.4729	**0.5812**

Table 6.3: Clustering average normalized mutual information and standard deviation using K-means on PCA, LPP, FME, and grlrPCA on seven datasets for 20 runs

Data	Normalized mutual information				
	Original	PCA	LPP	FME	grlrPCA
MNIST	0.4246	0.4518	0.4452	0.4655	**0.4705**
USPS	0.6160	0.6260	0.6365	0.6555	**0.6582**
BinAlpha	0.5824	0.6059	0.5744	0.6133	**0.6273**
ORL	0.7424	0.7705	0.7503	0.7743	**0.7828**
YALE	0.5486	0.5824	0.5511	0.5977	**0.6404**
TDT2	0.3305	0.3909	0.4274	0.4256	**0.4384**
WebKB	0.1182	0.1439	0.1989	0.1634	**0.2030**

Table 6.4: Clustering average accuracy and normalized mutual information using K-means on glPCA, MMF, and grlrPCA on training datasets for 20 runs

Data	Accuracy			Normalized mutual information		
	glPCA	MMF	grlrPCA	glPCA	MMF	grlrPCA
MNIST	0.4743	0.4474	**0.4914**	0.4510	0.4373	**0.4650**
BinAlpha	0.4492	0.4206	**0.4574**	0.6179	0.6017	**0.6237**
USPS	0.7384	0.6816	**0.7422**	**0.7004**	0.6209	0.6700
ORL	**0.6415**	0.5685	0.6117	**0.8071**	0.7647	0.7906
YALE	0.4644	0.4600	**0.4967**	0.5613	0.5679	**0.6027**
TDT2	0.4341	0.4212	**0.4384**	0.4278	0.3965	**0.4412**
WebKB	0.4689	0.4606	**0.5145**	0.2305	0.2264	**0.2345**

6.4.4 Results

Next, we evaluate the performance of grlrPCA using seven real-world data sets by comparing with other established models on these datasets. The models are trained with training data before the evaluation is performed on the testing data. Accuracy and normalized mutual information (NMI), as listed in Tables 6.2 and 6.3, are utilized to measure the differences between the labels predicted by each model and the benchmark labels actually observed. With respect to the Accuracy, our proposed method outperforms all other compared methods. For example, compared with the baseline PCA, grlrPCA reduces testing errors by 7.15%, 7.81%, 8.47%, 7.31%, 12.38%, 11.74% and 30.3% , respectively, on the seven data sets. These results demonstrate the effectiveness of grlrPCA in learning the intrinsic manifold projection. Regarding the established methods, LPP obtains lower accuracy than PCA, and is not robust on BinAlpha and Yale. The main reason for this lack of robustness is that LPP is defined on the graph weighted squared l_2-norm distances in the projected space and l_2-norm is prone to outliers [120]. Similar to the performance on the AC accuracy, grlrPCA enhances NMI accuracy as well. Compared with its baseline PCA counterpart, our method reduces testing errors by 4.14%, 5.14%, 3.53%, 1.6%, 9.96%, 12.15% and 41.07%, respectively, as presented in Table 6.3, which further proves the effectiveness of our method in learning the intrinsic manifold projection. The performances of the proposed method and FME are similar on the

Table 6.5: Computational time analysis for grlrPCA, FME, and glPCA on seven datasets in seconds.

data	grlrPCA	FME	glPCA
MNIST	22.6	32.8	6.7
USPS	15.5	19.6	4.3
BinAlpha	104.5	140.9	32.6
ORL	13.1	15.5	3.7
YALE	5.2	5.5	1.9
TDT2	21.8	35.2	6.8
WebKB	3.8	4.9	2.0

USPS data set, but for the rest of the six data sets, the proposed method obtains better performances.

The fact that glPCA and MMF lack the capability to handle out-of-sample problem motivates us to further compare our proposed model with glPCA and MMF on the training data set. As can be seen in Table 6.4, grlrPCA outperforms MMF on all seven data sets. Compared with glPCA, grlrPCA also obtains overall comparable or even better results on all data sets except the ORL. The reason for the exception is that, in our algorithm, V is a continuous solution of the cluster membership indicators, and closely relates to the labels of the data, and V is updated with W. The increase of the number of clusters may cause difficulties to obtain the effective V, which is manifested in the ORL with the most classes.

6.4.5 Computational Time Analysis

The experiments are performed on a single machine, Intel(R) Core(TM) i5-4570 CPU@3.20G. Table 6.5 shows the CPU-time used for training to accumulate 10 runs. It is clear that the proposed method grlrPCA is more efficient than FME which is the main method we compare with. In Algorithm 5, the initialization of V is given by the result of glPCA, and then W is optimized by the gradient decent method, which takes more time to perform the optimization. However, the goal of the grlrPCA model is to handle the out-of-sample data,

which can not be solved by the glPCA model.

6.5 Discussion

We presented a graph regularized linear regression PCA, a novel graph Laplacian model that incorporates the manifold structure of the data. This model is able to address the out-of-sample problem which is missing in previous studies including graph-regularization models for non-negative factorization and PCA. The proposed model provides a mapping for the new data by a linear regression function and effectively maps the sampled data to a manifold. This model is proven to be closely related to the unsupervised LDA model by theoretical analysis. The extensive experiments on seven datasets demonstrate that the proposed method achieves outstanding performance on both training and testing data. This model can also be used for the real-time situation recognition using photos where the proposed model can first perform dimension reduction and then detect situations by grouping similar photos together.

Chapter 7

From Photo Streams to Evolving Situations

In the study of social multimedia and situational awareness, the term "situation" is defined as the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future [26]. Situations in the physical world are the result of interactions among several related events. Events can be observed by measuring physical attributes using sensors. To understand constantly evolving situations, data may be collected using disparate sensors, social media, cyberspace interactions, and Internet-of-Things (IoT), as well as by people using effective participatory sensing. Using these precise observations, we can model and understand the world around us in order to better control adverse situations or cultivate a more optimal situation.

Deep learning frameworks have been successfully used in video and image analysis. Visual concepts can be accurately recognized. One of the major efforts in recent years is automatic semantic classification of media contents into a number of predefined concepts [17]. These

concepts and their relations are defined in an ontology from the results of user studies or data mining techniques applied to detect events. For example, photos of music festivals may contain a large number of objects like people and instruments, actions like singing as well as scenes like stages and posters. Event detection has achieved considerable advancement in recent years. Unlike event detection which is based on data at one particular space and time, situation is defined as the perception of elements in the environment within a volume of time and space and the projection of their status in the near future. Detection and understanding of a complex multimedia situation has not been explored. Given the accurate concept detector frameworks and considerable success in object recognition and event detection in computer vision and multimedia, it is a good time to explore the use of geo-tagged photos for situation recognition.

We use photos as the information source for radical improvement in the quality of participatory sensing data. Photo data has associated geo-spatial and other useful information. With information about space and time, the objective and factual nature of photos makes them the best resources for situation recognition. Given known situation photos, situation labels, and unlabeled photos, many traditional methods such as K-means, support vector machine, and spectral clustering can be used to infer a photo's label using labeled photos and unlabeled photos. In particular, by using labeled and unlabeled data together, semi-supervised learning can assign labels to the remaining photos by assuming that neighboring data points are likely to have the same label. However, traditional semi-supervised learning frameworks require at least one data sample for each label and lack abilities to discover new situations.

Graph-based semi-supervised learning methods have been the state-of-the-art in photo annotation and photo understanding [106]. [135] propose Harmonic Energy Minimization and use Gaussian fields to propagate label information to unlabeled data. As the proposed method is one kind of random walks, the output can be interpreted as the probabilities of classifying the data points to given labeled clusters. The classification heavily depends on

the labeled data which makes it sensitive to noise in labeled data. Some of the existing graph-based models use a quadratic form of graph embedding. However, a major drawback is the sensitivity of results to outliers. A robust graph-based learning method can overcome this drawback through the use of noise-robust norms. Among these methods, models using the ℓ_0-norm have demonstrated effective performance [128]. Unfortunately, computational expense increases with subset selection problems. To reduce the computational burden, the ℓ_0-norm has been replaced by some relaxations, such as the ℓ_1-norm.

In this chapter, we combine the advantages of noise-resistant norms and soft label methods and propose new graph-based models. This model learns efficient graph embedding by utilizing a ℓ_1-norm and a capped ℓ_p-norm to remove data outliers. The soft label method empowers the semi-supervised learning framework to accurately predict labels as well as discover new clusters. Furthermore, efficient iterative algorithms are adopted to solve the proposed optimization problem. The proposed framework provides a powerful tool to investigate situation recognition under the condition of noisy and unknown-labeled data. To verify the effectiveness and efficiency of the proposed models, we apply them to Yahoo Flickr Creative Commons data.

7.1 Examples of Evolving Situations

Situations have varying temporal and spatial dimensions, and varying social and economic attributes. They range from flexible space and time to fixed space and time, from intense disasters to common entertainment. In this section, we give two examples of evolving situations based on their different spatio-temporal natures.

The first example is "Olympic Games." The Olympics Games are always associated with concepts of "people", "sport", "bar", and "stadium", etc, evolving in a certain pattern. Thus,

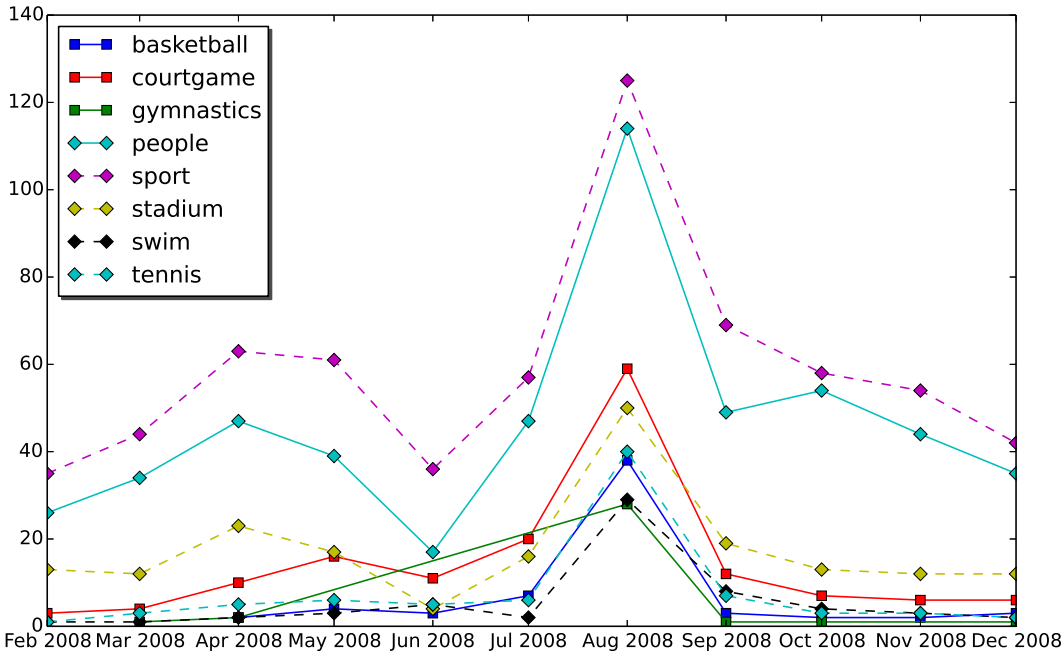


Figure 7.1: Evolving concepts in Beijing in the year 2008

the situation model of the Olympics Games can be defined as a bag of these photo visual concepts. The evolving situation is clearly shown in Figure 7.1 where the number of these ten concepts "basketball","courtgame","gymnastics","people", "sport","stadium","swim", and "tennis" reaches its peak at the same time of Olympic Games. Unlike event detection of either swimming competitions or gymnastics competitions which are based on data at one particular space and time, Olympic Games are defined as the perception of a larger number of visual concepts. By computing the similarity between the Olympics Games situation model with an evolving bag of visual concepts, similar situations can be detected. As we know, the Beijing Olympic Games were held in Beijing in the summer of 2008. The space and time information for situations such as the Olympic Games is fixed before they happen.

The second example is "Hanami." Its space and time are flexible. The Hanami is the Japanese traditional custom where people enjoy the transient cherry blossom from the end of March to early May all over Japan. The Hanami celebrations usually involve eating, drinking, and

playing music. The first cherry blossoms happen in southern Japan and start blooming in the northern part later. People as well as television and newschapter medias closely follow the "cherry blossom front" because it slowly moves from south to north. Unlike the Olympic Games, this kind of situation evolves in the environment within a volume of flexible time and space.

When a public situation occurs, people take photos related to this situation. People are acting as sensors enabling prompt detection of situation occurrence. Suppose we have millions of photos uploaded to the Web every day. Can we use the photos to observe real-world situations like the Olympic Games and Hanami?

7.2 The Proposed Framework

7.2.1 Problem Formulation and Motivation

We develop a framework that treats photos as micro-reports to detect evolving situations. The framework makes use of visual concepts, space and time information and includes three components: data collection, clustering, and filtering by space and time. An overview of the proposed framework is shown in Figure 7.2. The framework first collects historical photos and their labeled situations. For each photo, a deep learning concept detector is applied to acquire visual concepts. We do not aim at developing new concept detection methods or try to improve the existing concept detector. Instead, we use these concepts as features to detect situations. At the data collection stage, photos are associated with visual concepts and situation labels. It is commonly known that the amount of labeled data is far less than unlabeled data on the web. Semi-supervised learning methods can be naturally adopted to target the problem of recognizing new photo labels. However, traditional semi-supervised learning methods can not be directly applied because they require that photos

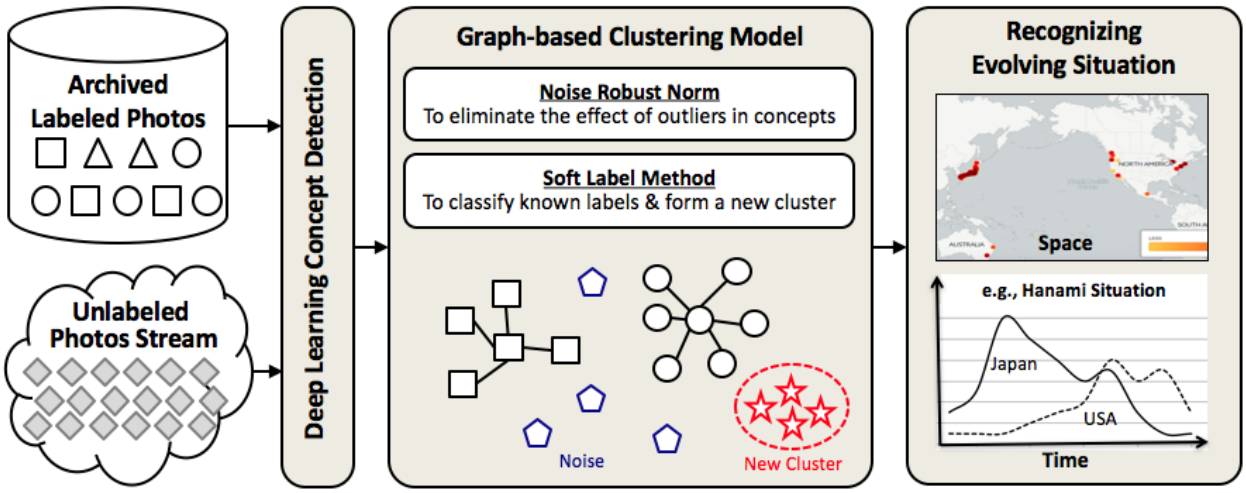


Figure 7.2: Semi-supervised learning framework for evolving situation recognition under the condition of unknown-labeled data

must belong to at least one predefined cluster. In real world applications, unknown or new situations may exist in unseen photos. One the other hand, noise may also exist in the new data. Thus, in the clustering stage, we need to handle both the unknown situation problem and the noise problem. We introduce a soft label method which can effectively form new clusters as well as accurately predict known labels. We incorporate noise robust norms to eliminate the adverse effects of outliers in visual concepts and thus improve the accuracy of situation recognition. At the end of the second stage, photos are labeled by situations. At the last stage, we use space and time information to further understand when and where these situations happen. For example, given photos of the situation "Olympic Games", time information can be used to understand if it is the "Winter Olympic Games" or the "Summer Olympic Games". Another example from space perspective is "Holi." Holi is an original Indian festival. Gradually it has been celebrated in the United States and Europe. The proposed framework can show how the Holi trend evolves at different locations.

7.2.2 Semi-supervised Learning Method for Unknown Labels

Given n photos $\{x_1, \cdots, x_m, x_{m+1}, \cdots, x_n\}$ and the labeled situation set $U = \{1, \cdots, u\}$, $\{x_1, \cdots, x_m\}$ are labeled with known situations $y \in U$, and the remaining photos $\{x_{m+1}, \cdots, x_n\}$ are not labeled. The goal is to use the visual concepts as features to predict the labels of unlabeled photos and discover an unknown situation if one exits in the photos. Predicting the labels of unlabeled photos using both labeled and unlabeled data is a semi-supervised learning problem. To enable the traditional semi-supervised learning framework to discover new situations, we introduce one more label variable denoted as $u+1$, where $\hat{U} = \{1, \cdots, u+1\}$. This simple setting will solve the new label discovery problem which traditional semi-supervised learning frameworks cannot solve.

Consider a connected graph of photos, $G = (V, E)$. Nodes in V correspond to n data points. Nodes in $M = (1, 2, \cdots, m)$ are situation-labeled photos. Nodes in $U = (m+1, \cdots, m+u)$ are unknown-situation label photos. The edge E in the graph is described in the similarity matrix $W \in R^{n \times n}$, where w_{ij} is the similarity measurement of a pair of vertices, x_i and x_j. For the similarity matrix w_{ij}, one example using the Gaussian function is

$$w_{ij} = \begin{cases} \exp(-\sum_{z=1}^{p} \frac{(x_{iz} - x_{jz})^2}{\sigma_z^2}) & j \in \mathcal{N}_i \text{ or } i \in \mathcal{N}_j \\ 0, & \text{otherwise} \end{cases},$$

where $x_i = (x_{i1}, \cdots, x_{iz})$, x has p dimensions of features, $\mathcal{N}_i$ is a set of indices of x_i's neighbors, and $\sigma_z (z = 1, \cdots, p)$ are parameters associated with features.

Given a similarity matrix W, the key idea of the graph-based method is that the nodes connected by a large weight in W in the graph have similar values. Based on the smoothness assumption, the task is to assign labels to nodes U. [82] propose a general graph-based

semi-supervised learning method for new class discovery:

$$\min_{F} \sum_{i,j} \hat{w}_{ij} \left\| f_i - f_j \right\|_2^2 + \sum_{i=1}^{n} u_i \hat{d}_i \left\| f_i - y_i \right\|_2^2 , \tag{7.1}$$

where the normalized weights $\hat{w_{ij}}$ can be computed by $\hat{w}_{ij} = w_{ij}/\sqrt{d_i d_j}$, $d_i = \sum_j w_{ij}$. By optimizing Eq. (7.1), we can obtain the soft label matrix $F \in \mathcal{R}^{n \times (c+1)}$, where c is the number of known situations. Then the label of x_i can be calculated as

$$y_i = \arg \max_{j \leq c+1} F_{ij}.$$

This model has two terms. The first term is a regularization, which controls the smoothness of the predicted labels on the graph. The second term determines the degree of label matching between the predicted labels and initial labels. Two parameters u and d are used to balance the trade-off between these two terms.

7.2.3 Noise Robust Semi-Supervised Learning Methods for Unknown Labels

Traditional methods such as (7.1) are built on the ℓ_2-norm of graph embedding. The quadratic form makes these methods sensitive to noise or outliers. To overcome the noisy data which degrades graph quality, leading to poor recognition results, we use noise robust norms which can eliminate the adverse effects of outliers in visual concepts and improve the accuracy of situation recognition. In particular, we propose one ℓ_1-norm method and one capped norm method. Capped norm based loss function has been used for various purposes. For example, the capped norm has been used for unsupervised photo clustering [111]. Here, we use the ℓ_p-norm as a robust and stable loss function to resist outliers in a semi-supervised

learning setting. We use this property in projecting data into a manifold where the similarity of data points is adjusted without input from outliers. If the distance of a label vector is large, the corresponding similarity value is not updated. We have the following objective function:

$$\min_{F} \sum_{i,j} \hat{w}_{ij} \left\| f_i - f_j \right\|_2 + \sum_{i=1}^{n} u_i \hat{d}_i \left\| f_i - y_i \right\|_2^2 . \tag{7.2}$$

The capped norm based loss function of a c-dimensional vector $u \in \mathcal{R}^{1\times c}$ is $\min(\|u\|_2^p, \theta)$, where θ is a parameter. The value of this loss function is $\|u\|_2^p$, if $\|u\|_2^p$ is smaller than θ, and is θ, otherwise. This loss function is more robust to outliers than the ℓ_1-norm because it has a threshold θ for outliers.

The proposed capped ℓ_p-norm method is expressed as

$$\min_{F} \sum_{i,j} \hat{w}_{ij} \min(\|f_i - f_j\|_2^p, \theta) + \sum_{i=1}^{n} u_i \hat{d}_i \left\| f_i - y_i \right\|_2^2 , \tag{7.3}$$

where w_{ij} is the similarity measurement between x_i and x_j in the graph, $0 < p \leq 2$ and θ are parameters.

The above formulation benefits from the control of the input data. The clustering results are dependent on the quality of input data graph. Most of the time, they are sensitive to the particular graph construction methods. We overcome this problem by graph initialization and graph similarity adaptation. We will introduce a method to give a "good" initialization. For the graph similarity adaptation, if the distance of labeled vector is large, then the corresponding similarity value is not updated because of the capped norm. By optimizing Eq.(7.2) and Eq. (7.3), we can obtain the soft label matrix $F \in \mathcal{R}^{n\times(c+1)}$, where c is the

number of labeled situations. Similar to Eq.(7.1), the label of x_i can be calculated as

$$y_i = \arg\max_{j \leq c+1} F_{ij}.$$

7.2.4 Optimization Algorithm

In this subsection, we introduce the optimization algorithm to solve Problem (7.2) and Problem (7.3). To optimize the objective function easily, we rewrite Eq.(7.2) into a matrix form:

$$\min_{F} \sum_{i,j} \hat{w}_{ij} \left\| f_i - f_j \right\|_2 + Tr(F - Y)^T U \hat{D} (F - Y), \tag{7.4}$$

where $\hat{D}$ is a diagonal matrix with diagonal entries as $\hat{D_{ii}} = \sum_j \hat{W}_{ij}, \forall i$.

Taking the derivative of Eq.(7.4) with respect to F, and setting the derivative to 0, we have

$$\bar{L}F + U\hat{D}(F - Y) = 0,$$

where $\bar{L}$ is the Laplacian matrix of $\bar{W}$, and the ij-th element of $\bar{W}$ is defined by

$$\bar{W}_{ij} = \frac{W_{ij}}{2 \left\| f_i - f_j \right\|_2}.$$

Thus,

$$F = (\bar{L} + U\hat{D})^{-1} U\hat{D}Y.$$

It is noted that $\bar{L}$ is dependent on F; we propose an iterative algorithm to obtain the solution.

We initialize the similarity matrix $W \in \mathcal{R}^{n \times n}$ in the objective function before performing the optimization; therefore the structure of the graph plays important roles in the performance of the graph-based clustering. We use the same method in Chapter 5 to generate the initial graph for the proposed model.

Algorithm 6 Semi-supervised ℓ_1-norm Method

Initialize W_{ij};
Construct graph for $w_{ij} \in W$ using Equation (5.9).
Repeat: Calculate $\bar{L}_t = \bar{D}_t - \bar{W}_t$,
where $(\bar{W}_t)_{ij} = \frac{W_{ij}}{2\|(f_t)_i - (f_t)_j\|_2}$
Calculate $F_{t+1} = (\bar{L}_t + U\hat{D})^{-1} U\hat{D}Y$
t = t+1;
Until convergence
Assign labels to x_i, $y_i = \arg\max_{j \leq c+1} F_{ij}$

To easily optimize the objective function in Problem (7.3), we rewrite it into a matrix form:

$$\sum_{i,j} \hat{w}_{ij} \min(\|f_i - f_j\|_2^p, \theta) + + Tr(F - Y)^T U \hat{D} (F - Y).$$

Because the proposed model is a weighted sum of concave functions, it can be solved in the general re-weighted optimization framework from Chapter 5.

The general re-weighted optimization problem is

$$\min_{x \in \mathcal{C}} f(x) + \sum_i h_i(g_i(x)), \tag{7.5}$$

where $h_i(x)$ is an arbitrary concave function with the domain of $g_i(x)$. This general problem can be solved by Algorithm 3.

For Problem (7.3), if we denote $h(x) = \min(x^{\frac{p}{2}}, \theta)$ and $x = \|f_i - f_j\|_2^2$, $h(x)$ is a concave function with respect to x, the supergradient of $h(x)$ can be obtained as

$$h'(x) = \begin{cases} \frac{p}{2}x^{\frac{p-2}{2}} & x^{\frac{p}{2}} \leq \theta \\ 0 & \text{otherwise} \end{cases}.$$

The first part of the proposed model can be written in the following form:

$$\begin{aligned} \min_{F^TF=I} \sum_{i,j} w_{ij}h'(x)x &= \min_{F^TF=I} \sum_{i,j} w_{ij}s_{ij} \left\| f_i - f_j \right\|_2^2 \\ &= \min_{F^TF=I} \sum_{i,j} \tilde{s}_{ij} \left\| f_i - f_j \right\|_2^2 \\ &= \min_{F^TF=I} Tr(F^T L_{\tilde{s}} F), \end{aligned}$$

where $\tilde{s_{ij}} = w_{ij}s_{ij}$.

After fixing $\tilde{s_{ij}}$, the objective function becomes

$$\min_F Tr(F^T L_{\tilde{s}} F) + Tr(F - Y)^T U \hat{D} (F - Y). \tag{7.6}$$

After taking the derivative of Eq.(7.6) with respect to F and setting the derivative to 0, we have

$$L_{\tilde{s}} F + U\hat{D}(F - Y) = 0.$$

Thus,

$$F = (L_{\tilde{s}} + U\hat{D})^{-1} U\hat{D}Y.$$

We propose an iterative algorithm to obtain the solution. The algorithm is guaranteed to converge because it is an application of a general optimization framework.

According to Algorithm 3, we have the following Algorithm 7 to solve Problem (7.3).

Algorithm 7 Semi-supervised ℓ_p-norm Method

Initialize s_{ij} and set $s_{ij} = 1$;
Construct graph for $w_{ij} \in W$ using Equation (5.9).
Repeat: Calculate $\tilde{w}$ by $\tilde{w} = w_{ij}s_{ij}$
Update F by $F = (L_{\tilde{s}} + U\hat{D})^{-1}U\hat{D}Y$;
Calculate $s_{ij} = \begin{cases} \frac{p}{2}\|f_i - f_j\|_2^{p-2} & \|f_i - f_j\|_2^p \leq \theta \\ 0 & \text{otherwise} \end{cases}$
Until convergence
Assign labels to x_i, $y_i = \arg\max_{j \leq c+1} F_{ij}$

7.3 Experiments

7.3.1 Data Preprocessing

Yahoo Flickr Creative Commons 100M (YFCC100M) dataset [117] is used to demonstrate our idea. As we described in previous sections, deep learning approaches from Clarifai are used to detect 1,570 concepts in all the sampled photos. A list of top-30 concepts is shown as "people", "nature", "indoor", "sport", "landscape", "plant", "architecture", "music", "performer", "tree", "demonstrator", "vehicle", "building", "concert", "cherry blossom", "water", "musician", "blossom", "crowd", "sakura", "outfit", "ensemble", "text", "stage", "slope", "riverbank", "animal", "lake", "road", and "sign."

We perform two experiments to verify the idea of situation recognition and the proposed models. The first one is to prove that the proposed noise-robust norms are effective in reducing the adverse function of outliers. The second experiment is to apply the verified models to YFCC100M data and show the evolving situation recognition.

Table 7.1: Accuracy (in percentage) for label-known data and label-unknown data: GSS is Semi-supervised Learning Method for Unknown Labels Model (Eq. (7.1)), SSL is ℓ_1-norm Semi-supervised Learning Method (Eq. (7.2)), SSC is ℓ_p-norm Semi-supervised Learning Method (Eq. (7.3))

Method	**Label-known Data**	**Label-unknown Data**
GSS	92.80	91.40
SSL	93.64	97.40
SSC	96.06	95.10

7.3.2 Semi-supervised Clustering with Unknown Labels

We use six situations to demonstrate the idea and the proposed framework. The six situations include "Hanami", "Olympic Games","protest", "flood", "blues fest", and "Holi." We use the users' tags as situation labels. We randomly select 500 photos for each situation before the year 2010 as the labeled data, then we randomly select 3000 photos after the year 2010 as testing data. Among the unlabeled data, 500 photos are not in the categories of predefined situations as noise. We use 3000 labeled photos and 3000 unlabeled photos. We present an experimental study on GSS [82], Semi-supervised ℓ_1-norm method (SSL), and Semi-supervised ℓ_p-norm method (SSC) for situation detection. There are parameters u_i in the proposed methods. For fair comparison, we set the compared methods and the proposed methods with the same parameter setting. We use one u for the labeled data and another u for unlabeled data to control the smoothness term and the fitting term for the model, $u = (1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100)$. We also perform cross validation to select the best parameter combination for SSC from the range $\theta = (0.01, 0.1, 1, 10)$ and $p = (0.5, 0.7, 1, 1.5, 1.7, 2)$.

As Table 7.1 shows, two proposed models, SSL and SSC outperform the ℓ_2-norm based GSS model. For the known labeled data, compared with GSS, SSC improves the clustering accuracy by 4.3%, and SSC improves the accuracy from 92.80% to 93.64%. For the unknown labeled data, both SSL and SSC consistently improved over the benchmark. Methods such as

traditional semi-supervised learning methods are not able to assign unknown labels to data, thus the accuracy of those methods on unknown labeled data is 0. Compared with GSS, the proposed methods SSC and SSL reduce testing error by 6.5% and 4.0%, respectively, which further proves the effectiveness of our methods in discovering unknown labels.

7.3.3 Evolving Situation Detection

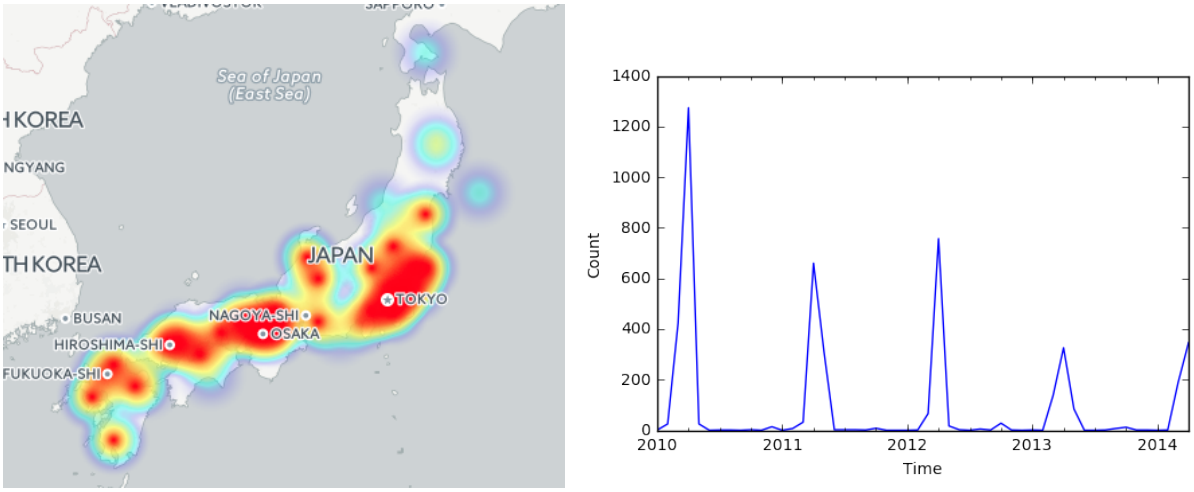


Figure 7.3: The evolving situation of "Hanami"

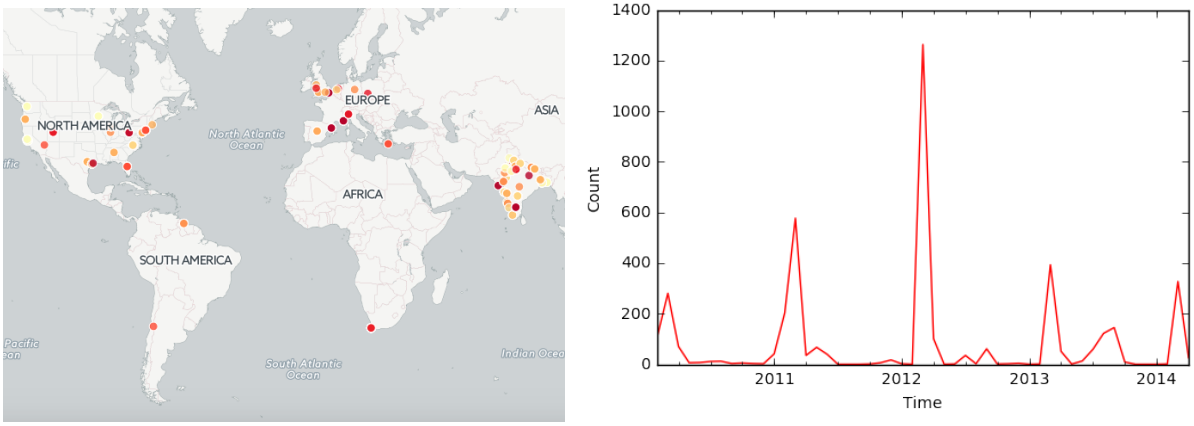


Figure 7.4: The situation of "Holi" in space and time

As we mentioned, unlike event detection which is based on data at one particular space and time, situation is defined as the perception of elements in the environment within a volume of time and space and the projection of their status in the near future. In order

to detect evolving situations, we further analyze the space and time context of the labeled photo outputs from our clustering method and apply the proposed framework to the data after the year 2011 to detect situations. Figure 7.3 shows that we can successfully detect the situation of "Hanami" in Japan. Tokyo, Osaka, and Hiroshima are the three most popular places for Hanami, which happens in late March and early April every year. Similarly, we demonstrate it on detection of "Holi". Figure 7.4 shows that Holi is celebrated not only in India but is widely celebrated in the United States and Europe as well in March every year.

7.4 Discussion

Situation recognition is important in many real world applications ranging from disaster response to social events. With the explosive growth of user-contributed photos and videos, there is a great opportunity to detect evolving situations and respond to them. In this chapter, we combine geo-tagged photo streams and concept detection to explore situation recognition. We develop new methodologies to tackle some technical challenges, i.e., overcoming noisy data and processing real-time data. Situation recognition will be enriched by developing methods to combine visual concepts with physical sensory data.

Chapter 8

Conclusion and Future Research

In this chapter, we summarize the major research results presented in this dissertation, and discuss some future directions for situation recognition.

8.1 Conclusion

Social Life Networks, which aim to connect people to resources effectively, efficiently, and promptly, requires assimilating multimedia heterogeneous data for situation recognition. In recent years, the multimedia research community has been gradually shifting its emphasis from analyzing traditional media data such as, image, video, and text, to exploring the opportunities to select and combine diverse information sources from geo sensors, given the rapid development of sensor technology and participatory sensing. To develop strategies for processing geospatial multimedia data for situation recognition, we need to address two major research challenges, i.e., data quality and data analysis.

This dissertation presents a graph-based model as a general framework to combine diverse data sources for prediction and clustering. We propose a multimodal geospatial prediction

framework that integrates heterogeneous data into a unified space-time-value format and represents the data as a spatial grid with time-dependent thematic data. This framework provides a feature extractor and prediction model in order to efficiently generate multiscale features and predict values at unobserved locations. We demonstrate the effectiveness of the proposed framework using air pollution data. We then use the predicted air pollution data and combine pollen data to analyze the situation of asthma risk across California. The proposed model based on our predicted values provides greater accuracy in detecting high asthma risk areas compared to previous studies. To further understand asthma, we consider both environmental data and personal profile and activity data and extract personal and environmental features that determine a person's susceptibility to asthma. A ranking of these features is given based on their potential for causing asthma. The features derived from our analysis agree with the current consensus in the medical community about factors affecting asthma.

Similar to health-related issues, such as asthma, many personal situation recognition problems can benefit from collecting more personal data. Given the wide use of mobile technology, humans also act as sensors in that people themselves, in carrying and using smartphones, can contribute their measurements or observations. With the explosive growth of user-contributed photos and videos, there is an excellent opportunity to detect evolving situations. We investigate the use of visual concepts for situation recognition and develop new methodologies to tackle the technical challenges of noisy data and processing real-time data, which are problems in existing approaches. We propose several graph-based models to address the above challenges. First, we propose a graph-regularized linear regression PCA to address photo clustering in a real-time setting. This graph Laplacian model incorporates the manifold structure of the data and performs dimension reduction. This model provides a mapping for unseen data through a linear regression function and effectively copes with the data sampled from nonlinear manifold. Additionally, we conduct theoretical analysis and show that this model is intrinsically related to unsupervised LDA model.

Furthermore, in order to overcome the technique challenge of noisy data, we present a capped norm model for photo clustering. It utilizes the properties of capped norm to remove the effect of outliers, and explores the underlying data structure in manifold to improve the performance of clustering. We develop an efficient iterative algorithm to solve the optimization problem for the proposed objective function. This algorithm updates the similarity matrix in an adaptive way by ignoring the outliers and noise, and performs effective data mapping in the nonlinear manifold. A stream of photos from different people and different places creates a mosaic of world wide events as they happen. Accurate photo clustering will facilitate situation recognition in emerging applications on the Internet. To extend the method for unknown situations, we introduce a soft label method that enables the traditional semi-supervised learning framework to accurately predict predefined labels as well as effectively form new clusters. To overcome the noisy data that degrades graph quality, leading to poor recognition results, we take advantage of two of the noise-robust norms, which can eliminate the adverse effects of outliers in visual concepts and improve the accuracy of situation recognition.

To evaluate the performance of the proposed approaches, we provide a thorough study on the YFCC100M data. We consider our approaches to be relevant beyond the multimedia research community, contributing to the data mining community for developing better algorithms. Therefore, we want to emphasize that although the approaches developed in this dissertation are applied to air pollution prediction, asthma risk analysis, and multimedia micro-reports, the contributions and potential applications of the proposed approaches are not only limited to these domains.

8.2 Future Research

Social Life Networks offer many new opportunities to develop efficient approaches for combining heterogeneous data sources. As demonstrated in previous chapters, we address the air pollution prediction problem by integrating multiple data sources and we explore the use of photos for multimedia micro-reports by combining space and time information. But the approaches we describe in this dissertation only reveal "the tip of iceberg" when it comes to the research problems. That is, many interesting future research directions can be explored, some of which are detailed below.

Multimedia Participatory Sensing: Participatory sensing utilizes the eyes and ears of citizens to collect information about a situation. We use visual concepts from photos and combine space and time information to detect evolving situations. More data sources from environment and IoT can be integrated for exploring how we can detect current situations and more importantly predict emerging situations.

Evolving Situation Recognition: Cameras and sensors are deployed at different locations to continuously monitor target conditions, for example, traffic flow, air pollution, and temperature. When such data streams are combined to provide actionable information, discovering the spatially correlated and frequently evolving patterns is challenging: First, informative patterns are often flooded by trivial fluctuations. Second, the pattern search space is extremely large. A pattern can contain an arbitrary number of sensors, and the matching time intervals are at different scales. This results in an exponential pattern search space, calling for novel and efficient pattern search methods. Situation recognition may very well be the next major research challenge for machine learning after recent progress in concept recognition.

Real-time Data Processing: In some situations, response time is critical. The computation or processing time of large volumes of data can take too much time to be useful.

A real-time visualization can serve as a first step to build a real-time situation recognition engine. The engine must have the capability to do real time visual analytics for evolving situations. The research challenges here are to develop new architectures for real time processing of massive volumes of heterogeneous streaming data.

Personalized Recommendation: In order to connect needs and resources, and optimize the need matching, we develop various techniques for situation recognition. The locations and behaviors of citizens might be continuously changing. Tracking the locations and behaviors of citizens are important for allocating resources to aid them. Using multimedia data like photos, audio, and signals from wearable devices to detect activity and emotion can be critical for personalized responses. The research challenge here is to develop multimodal recommendation engines that are not only multi-dimensional but also time and location aware.

Health-related Problems: We discuss the use of both personal data and environmental data for asthma risk factor analysis. The personal data is from surveys and survey data usually is not objective. To accurately monitor patients, future health requires the integration of emerging data technology to obtain data for quantifying lifestyle and environment. Decision making can be powered by event correlation analysis. Preventive health system then can be built to constantly predict health status to offer targeted intervention.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] A. A. Arif, G. L. Delclos, and C. Serra. Occupational exposures and asthma among nursing professionals. *Occup Environ Med*, 66(4):274–278, Apr. 2009.

[3] E. Aune, D. P. Simpson, and J. Eidsvik. Parameter estimation in high dimensional gaussian distributions. *Statistics and Computing*, 24(2):247–263, 2014.

[4] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

[5] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010.

[6] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*, 2012.

[7] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, July 1997.

[8] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.*, 15(6):1373–1396, June 2003.

[9] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Advances in neural information processing systems*, 16:177–184, 2004.

[10] J. Bernd, D. Borth, C. Carrano, J. Choi, B. Elizalde, G. Friedland, L. Gottlieb, K. Ni, R. Pearce, D. Poland, et al. Kickstarting the commons: The yfcc100m and the yli corpora. In *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*, pages 1–6. ACM, 2015.

[11] V. J. Berrocal, A. E. Gelfand, and D. M. Holland. A spatio-temporal downscaler for output from numerical models. *J. AGR. BIOL. ENVIR. ST.*, 15(2), 2010.

[12] A. Beutel, T. Mølhave, P. K. Agarwal, A. P. Boedihardjo, and J. A. Shine. Terranni: natural neighbor interpolation on a 3d grid using a gpu. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 64–74. ACM, 2011.

[13] E. V. Bonilla, K. M. Chai, and C. Williams. Multi-task gaussian process prediction. In *Advances in neural information processing systems*, pages 153–160, 2007.

[14] D. W. Byun, J. Ching, and et.al. *Science algorithms of the EPA Models-3 community multiscale air quality (CMAQ) modeling system*. US EPA, Office of Research and Development.

[15] F. Castanedo. A Review of Data Fusion Techniques. *The Scientific World Journal*, 2013:1–19, 2013.

[16] P. K. Chan, M. D. Schlag, and J. Y. Zien. Spectral k-way ratio-cut partitioning and clustering. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 13(9):1088–1096, 1994.

[17] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo. Large-scale multimodal semantic concept detection for consumer video. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 255–264. ACM, 2007.

[18] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[19] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

[20] J. Choo, S. Bohn, G. Nakamura, A. M. White, and H. Park. Heterogeneous data fusion via space alignment using nonmetric multidimensional scaling. In *SDM*, pages 177–188. SIAM, 2012.

[21] C. Cieri, D. Graff, M. Liberman, N. Martey, S. Strassel, et al. The tdt-2 text and speech corpus. In *Proceedings of the DARPA Broadcast News workshop*, pages 57–60, 1999.

[22] S. Crumeyrolle, G. Chen, L. Ziemba, A. Beyersdorf, L. Thornhill, E. Winstead, R. H. Moore, M. A. Shook, C. Hudgins, and B. E. Anderson. Factors that influence surface PM2.5 values inferred from satellite observations: perspective gained for the US baltimorewashington metropolitan area during DISCOVER-AQ. *Atmospheric Chemistry and Physics*, 14(4):2139–2153, Feb. 2014.

[23] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*, pages 198–206, 2013.

[24] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM, 2004.

[25] R. Durichen, M. Pimentel, L. Clifton, A. Schweikard, and D. Clifton. Multitask Gaussian Processes for Multivariate Physiological Time-Series Analysis. *IEEE Transactions on Biomedical Engineering*, 62(1):314–322, Jan. 2015.

[26] M. R. Endsley. Situation awareness global assessment technique (sagat). In *IEEE Aerospace and Electronics Conference*, 1988.

[27] EPA. *http://www.epa.gov/airquality/airdata*.

[28] R. T. Farouki. The bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29(6):379–419, 2012.

[29] J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, Oct. 2001.

[30] M. Fuentes and A. E. Raftery. Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1).

[31] S. P. Galant, L. J. R. Crawford, T. Morphew, C. A. Jones, and S. Bassin. Predictive value of a cross-cultural asthma case-detection tool in an elementary school population. *Pediatrics*, 114(3):e307–316, Sept. 2004.

[32] M. Gao, V. K. Singh, and R. Jain. Eventshop: From Heterogeneous Web Streams to Personalized Situation Detection and Control. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 105–108, 2012.

[33] M. K. Garg, D.-J. Kim, D. S. Turaga, and B. Prabhakaran. Multimodal Analysis of Body Sensor Network Data Streams for Real-time Healthcare. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, pages 469–478, New York, NY, USA, 2010. ACM.

[34] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):643–660, 2001.

[35] R. Ghani, R. Jones, D. Mladenic, K. Nigam, and S. Slattery. Data mining on symbolic knowledge extracted from the web. In *Workshop on Text Mining at the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 56, 2000.

[36] P. Gong, J. Ye, and C.-s. Zhang. Multi-stage multi-task feature learning. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2012.

[37] P. Goovaerts. Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma*, 89(1):1–45, 1999.

[38] A. Gupta and R. Jain. Social life networks: A multimedia problem? In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 203–212, New York, NY, USA, 2013. ACM.

[39] D. L. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.

[40] Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, and J. Jiang. Sparse unsupervised dimensionality reduction for multiple view data. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(10):1485–1496, 2012.

[41] Y. Han, Z. Xu, Z. Ma, and Z. Huang. Image classification with manifold learning for out-of-sample data. *Signal Processing*, 93(8):2169–2177, 2013.

[42] X. He, X. He, and P. Niyogi. Locality Preserving Projections. 2002.

[43] T. Hengl, G. B. Heuvelink, and A. Stein. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120(1):75–93, 2004.

[44] W. Ho, W. Hartley, L. Myers, M. Lin, Y. Lin, C. Lien, and R. Lin. Air pollution, weather, and associated risk factors related to asthma prevalence and attack rate. *Environmental Research*, 104(3):402–409, July 2007.

[45] N. Hochman and L. Manovich. Zooming into an instagram city: Reading the local through social media. *First Monday*, 18(7), 2013.

[46] C. Y. Hong, T. P. Ng, M. L. Wong, K. T. C. Koh, L. G. Goh, and S. L. Ling. Lifestyle and behavioural risk factors associated with asthma morbidity in adults. *Qjm*, 87(10):639–645, 1994.

[47] F. Hui, B. Xu, H. Huang, Q. Yu, and P. Gong. Modelling spatial-temporal change of poyang lake using multitemporal landsat imagery. *International Journal of Remote Sensing*, 29(20):5767–5784, 2008.

[48] E. H. S. Isaaks, M. R. E. H. Isaaks, and M. R. Srivastava. *Applied geostatistics*. Number 551.72 ISA. 1989.

[49] R. Jain. Let's weave the visual web. *MultiMedia, IEEE*, 22(3):66–72, 2015.

[50] R. Jain and D. Sonnen. Social life networks. *IT Professional*, 13(5):8–11, 2011.

[51] L. Jalali, D. Huo, H. Oh, M. Tang, S. Pongpaichet, and R. Jain. Personicle: Personal chronicle of life events. In *Workshop on Personal Data Analytics in the Internet of Things (PDA@ IOT) at the 40th International Conference on Very Large Databases (VLDB), Hangzhou, China*, 2014.

[52] L. Jalali and R. Jain. Building Health Persona from Personal Data Streams. In *Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia*, PDM '13, pages 19–26, New York, NY, USA, 2013. ACM.

[53] M. JE, A. LJ, and B. CM. National Surveillance of Asthma: United States, 2001-2010. *Vital Health Statistics*, Nov. 2012.

[54] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, volume 2, page 4, 2014.

[55] B. Jiang, C. Ding, B. Luo, and J. Tang. Graph-Laplacian PCA: Closed-Form Solution and Robustness. pages 3492–3498. IEEE, June 2013.

[56] X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han. The wisdom of social multimedia: using flickr for prediction and forecast. In *Proceedings of the international conference on Multimedia*, pages 1235–1244. ACM, 2010.

[57] I. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[58] Juan Ye, G. Stevenson, and S. Dobson. Using temporal correlation and time series to detect missing activity-driven sensor events. pages 44–49. IEEE, Mar. 2015.

[59] Y. Kanani Sadat, F. Karimipour, and A. Kanani Sadat. Investigating the Relation Between Prevalence of Asthmatic Allergy with the Characteristics of the Environment Using Association Rule Mining. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-2/W3:169–174, Oct. 2014.

[60] J. Kang, R. Jin, and X. Li. Regression kriging-based upscaling of soil moisture measurements from a wireless sensor network and multiresource remote sensing information over heterogeneous cropland. *IEEE Geoscience and Remote Sensing Letters*, 12(1):92–96, 2015.

[61] C. G. Kaufman, M. J. Schervish, and D. W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, 2008.

[62] E. Kouloumpis, T. Wilson, and J. D. Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11:538–541, 2011.

[63] A. Krizhevsky. cuda-convnet. https://code.google.com/p/cuda-convnet/, 2010.

[64] D. Kuang, C. Ding, and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In *In Proceedings of the 2012 SIAM International Conference on Data Mining*, 2012.

[65] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[66] Y. LeCun, J. S. Denker, S. A. Solla, R. E. Howard, and L. D. Jackel. Optimal brain damage. In *NIPs*, volume 89, 1989.

[67] C.-H. Lee, J. C.-Y. Chen, and V. S. Tseng. A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring. *Computer Methods and Programs in Biomedicine*, 101(1):44–61, Jan. 2011.

[68] J. Li and A. D. Heap. Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, 53:173–189, 2014.

[69] L. Li and P. Revesz. Interpolation methods for spatio-temporal geographic data. *Computers, Environment and Urban Systems*, 28(3):201–227, 2004.

[70] X. Li, G. Cheng, and L. Lu. Spatial analysis of air temperature in the qinghai-tibet plateau. *Arctic, Antarctic, and Alpine Research*, 37(2):246–252, 2005.

[71] Y. Li, J. Huang, and J. Luo. Using user generated online photos to estimate and monitor air pollution in major cities. In *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, page 79. ACM, 2015.

[72] P. Liu, A. Tsimpidi, Y. Hu, B. Stone, A. Russell, and A. Nenes. Differences between downscaling with spectral and grid nudging using wrf. *Atmospheric Chemistry and Physics*, 12(8):3601–3610, 2012.

[73] L. Mitas and H. Mitasova. *Spatial interpolation*. Wiley, 1999.

[74] T. M. Mitchell. Machine learning. *Computer Science Series (McGraw-Hill, Burr Ridge, 1997) MATH*, 1997.

[75] A. Moraru and D. Mladenic. Complex event processing and data mining for smart cities. In *Conference on Data Mining and Data Warehouses (SkiDD 2013), Held at the 15th International Multiconference on Information Society (IS-2012), 8th October*, 2012.

[76] S. A. Nene, S. K. Nayar, H. Murase, et al. Columbia object image library (coil-20). Technical report, technical report CUCS-005-96, 1996.

[77] M. Neteler. *Open source GIS: a grass gis approach, third edition*. Springer, New York, 2007.

[78] H. Nguyen, N. Cressie, and A. Braverman. Spatial statistical data fusion for remote sensing applications. *J. AM. STAT. ASSOC.*, 107(499).

[79] F. Nie, H. Huang, X. Cai, and C. H. Ding. Efficient and robust feature selection via joint l2,1-norms minimization. In *Advances in neural information processing systems*, pages 1813–1821, 2010.

[80] F. Nie, H. Wang, H. Huang, and C. Ding. Unsupervised and semi-supervised learning via l1-norm graph. In *ICCV*, pages 2268–2273. IEEE, 2011.

[81] F. Nie, X. Wang, M. I. Jordan, and H. Huang. The constrained laplacian rank algorithm for graph-based clustering. In *AAAI*, 2016.

[82] F. Nie, S. Xiang, Y. Liu, and C. Zhang. A general graph-based semi-supervised learning with novel class discovery. *Neural Computing and Applications*, 19(4):549–555, 2010.

[83] F. Nie, D. Xu, I. Tsang, and C. Zhang. Flexible Manifold Embedding: A Framework for Semi-Supervised and Unsupervised Dimension Reduction. *IEEE Transactions on Image Processing*, 19(7):1921–1932, July 2010.

[84] F. Nie, J. Yuan, and H. Huang. Optimal mean robust principal component analysis. In *ICML*, pages 1062–1070, 2014.

[85] M. A. Oliver. *Geostatistical applications for precision agriculture*. Springer Science & Business Media, 2010.

[86] M. A. Osborne, S. J. Roberts, A. Rogers, and N. R. Jennings. Real-time information processing of environmental sensor network data using bayesian gaussian processes. *ACM Trans. SE. NETW.*, 9(1).

[87] C. J. Paciorek. Combining spatial information sources while accounting for systematic errors in proxies. *J. R. STAT. SOC.*, 61(3).

[88] A. Paladugu, P. Chandakkar, P. Zhang, and B. Li. Supporting navigation of outdoor shopping complexes for visuallyimpaired users through multi-modal data fusion. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–7, July 2013.

[89] S. Pongpaichet, V. K. Singh, M. Gao, and R. Jain. EventShop: Recognizing situations in web data streams. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, pages 1359–1368, 2013.

[90] S. Pongpaichet, M. Tang, L. Jalali, and R. Jain. Using photos as micro-reports of events. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 87–94. ACM, 2016.

[91] S. Ponpaichet, M. Tang, L. Jalali, and R. Jain. Observing real-world phenomena through eventweb over space, time and theme. In *2nd International Workshop on Building Web Observatories, ACM Web Science*, volume 2014, 2014.

[92] J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[93] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze. Predicting Asthma-Related Emergency Department Visits Using Big Data. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–1, 2015.

[94] G. Razeghi, M. Carreras-Sospedra, T. Brown, J. Brouwer, D. Dabdub, and S. Samuelsen. Episodic air quality impacts of plug-in electric vehicles. *Atmospheric Environment*, 137:90–100, 2016.

[95] B. J. Reich, H. H. Chang, and K. M. Foley. A spectral method for spatial downscaling. *Biometrics*, June 2014.

[96] P. Revesz and S. Wu. Spatiotemporal reasoning about epidemiological data. *Artificial Intelligence in Medicine*, 38(2):157–170, 2006.

[97] R. A. Rohde and R. A. Muller. Air pollution in china: Mapping of concentrations and sources. *PloS one*, 10(8):e0135749, 2015.

[98] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[99] H. Saif, Y. He, and H. Alani. Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org), 2012.

[100] F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pages 138–142. IEEE, 1994.

[101] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[102] V. K. Singh, M. Gao, and R. Jain. Social pixels: Genesis and evaluation. In *Proceedings of the International Conference on Multimedia*, pages 481–490, 2010.

[103] V. K. Singh, M. Gao, and R. Jain. Situation Recognition: An Evolving Problem for Heterogeneous Dynamic Big Multimedia Data. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1209–1218, New York, NY, USA, 2012. ACM.

[104] L. A. St Denis, L. Palen, and K. M. Anderson. Mastering social media: an analysis of jefferson countys communications during the 2013 colorado floods. In *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM 20014)*, 2014.

[105] M. T, G. G, S. L, and L. Y. The National Exposure Research Laboratorys Consolidated Human Activity Database. *Journal of Exposure Analysis and Environmental Epidemiology*, 10(6):566–578, 2000.

[106] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain. Image annotation by k nn-sparse graph-based label propagation over noisily tagged web images. *ACM TIST*, 2(2):14, 2011.

[107] M. Tang. Geospatial multimedia data for situation recognition. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1430–1434. ACM, 2016.

[108] M. Tang, P. Agrawal, and R. Jain. Habits vs environment: What really causes asthma? In *Proceedings of the ACM Web Science Conference*, page 30. ACM, 2015.

[109] M. Tang, P. Agrawal, F. Nie, S. Pongpaichet, and R. Jain. A graph based multimodal geospatial interpolation framework. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016.

[110] M. Tang, P. Agrawal, S. Pongpaichet, and R. Jain. Geospatial interpolation analytics for data streams in eventshop. In *Multimedia Expo. ICME, 2015 IEEE Int. Conf.* IEEE.

[111] M. Tang, F. Nie, and R. Jain. Capped lp-norm graph embedding for photo clustering. In *ACM Multimedia*, 2016.

[112] M. Tang, F. Nie, and R. Jain. A graph regularized dimension reduction method for out-of-sample data. *Neurocomputing*, 225:58–63, 2017.

[113] M. Tang, F. Nie, S. Pongpaichet, and R. Jain. From photo streams to evolving situations. In *4th IEEE International Workshop on Mobile Multimedia Computing*, 2017.

[114] M. Tang, S. Pongpaichet, and R. Jain. Research challenges in developing multimedia systems for managing emergency situations. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 938–947. ACM, 2016.

[115] M. Tang, X. Wu, P. Agrawal, S. Pongpaichet, and R. Jain. Integration of diverse data sources for spatial pm2. 5 data interpolation. *IEEE Transactions on Multimedia*, 2016.

[116] J. B. Tenenbaum. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, Dec. 2000.

[117] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[118] A. Van Donkelaar, R. V. Martin, and R. J. Park. Estimating ground-level PM2. 5 using aerosol optical depth determined from satellite remote sensing. *J. GEOPHYS. RES-ATMOS.*, 111.

[119] D. Wang, F. Nie, and H. Huang. Large-scale adaptive semi-supervised learning via unified inductive and transductive model. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 482–491. ACM, 2014.

[120] H. Wang, F. Nie, and H. Huang. Learning robust locality preserving projection via p-order minimization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[121] Y. Wang. Socializing multimodal sensors for information fusion. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 653–656. ACM, 2015.

[122] Y. Wang and M. S. Kankanhalli. Tweeting cameras for event detection. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1231–1241. International World Wide Web Conferences Steering Committee, 2015.

[123] C. K. Wikle. Low-rank representations for spatial processes. *Handbook of Spatial Statistics*, pages 107–118, 2010.

[124] M. Wu, K. Yu, S. Yu, and B. Schölkopf. Local learning projections. In *Proceedings of the 24th international conference on Machine learning*, pages 1039–1046. ACM, 2007.

[125] H. Xu, C. Caramanis, and S. Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):187–193, 2012.

[126] Y. Xu, J. S. Dyer, and A. B. Owen. Empirical stationary correlations for semi-supervised learning on graphs. *The Annals of Applied Statistics*, pages 589–614, 2010.

[127] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, Jan. 2007.

[128] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 625–632. IEEE, 2011.

[129] G. Ye, M. Tang, J.-F. Cai, Q. Nie, and X. Xie. Low-rank regularization for learning gene expression programs. *PloS one*, 8(12):e82146, 2013.

[130] J. Zhang. Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion*, 1(1):5–24, 2010.

[131] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:1081–1107, 2010.

[132] Y. Zhang, G. Li, L. Chu, S. Wang, W. Zhang, and Q. Huang. Cross-media topic detection: A multi-modality fusion framework. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, July 2013.

[133] Z. Zhang and K. Zhao. Low-Rank Matrix Approximation with Manifold Regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1717–1729, 2013.

[134] Y. Zheng, F. Liu, and H. Hsieh. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1436–1444, 2013.

[135] X. Zhu, Z. Ghahramani, J. Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.

[136] X. Zhu, J. Lafferty, and R. Rosenfeld. *Semi-supervised learning with graphs*. Carnegie Mellon University, language technologies institute, school of computer science, 2005.

[137] M. H. Zweig and G. Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.