FUSING PHYSICAL AND SOCIAL SENSORS FOR SITUATION AWARENESS

YUHUI WANG

(M.Eng.)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE
SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2017

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Wang Yuhui

Yuhui Wang

25 February 2017

Acknowledgements

First and foremost, I would like to express my sincere and deepest gratitude to my advisor Professor Mohan S Kankanhalli. Thank you so much for your professional and inspiring guidance during my whole Ph.D. study. I will never forget your continuous encouragement, especially the motivating stories, you have given me that makes me overcome all the obstacles in the past four years. It has been a great privilege to work with you and learn from you about how to become a mature, self-motivated, hard-working yet humble man.

Second, I would like to show my gratitude to the members of my thesis committee, Professor Roger Zimmermann and Professor Qi Zhao. Thank you for all your insightful comments and valuable inputs in my TAC meetings, which greatly help to shape my thesis and most importantly, make me confident and excited about my research. Also, thanks to my thesis examiner Professor Terence Sim for the precious comments and suggestions for the final version. I also appreciate my cousin Justin Wang's proof reading of thesis for correcting the mistakes and offering better modification suggestions.

My research journey would not have been as exciting without the countless discussions with so many prestigious professors and excellent researchers including Professor Ramesh Jain, Vivek Singh, Dr. Christian von der Weth and Dr. Thomas Winkler. Thank you very much for your valuable time, helpful comments and excellent ideas shared with me.

I am grateful for NGS and SoC offering me such a stimulating academic environment in NUS. The considerable student support, diverse inspiring seminars, first-class facilities and financial support, all leave me with an enjoyable study experience here.

My lab mates and friends have also supported me at different great times. I would like to especially thank Dr. Gan Tian, Dr. Prabhu, Dr. Padmanabha, Dr. Yongkang and my friends Huang Zhi, Francesco, Luo Yan, Zhongtao, Yinan, Lin Hang, Lian Yi, Mulong and all my other friends for their encouragement and unconditional companionship in my school life.

Lastly, I would like to thank my parents, for all the love and unyielding support they have given me during my study abroad. My debt to them is beyond measure; without their affection, support and sacrifices, I would have never obtained such a great opportunity to receive a higher education which shapes me to who I am today.

Contents

Summary

With the prevalence of physical sensors and social sensors, we are now living in a world of big sensor data. There are mainly two types of sensors that are constantly monitoring our surroundings: (a) **physical sensors** such as CCTV cameras, accelerometers, gyroscopes, mobile phones, RFID tags, temperature sensors, humidity sensors, etc. and (b) **social sensors** like social networking sites (e.g., YouTube, Youku, Facebook, Twitter, Weibo, Wechat) containing user-generated content reporting events in all kinds of formats (text, image or video).

Though theses sensors generate heterogeneous data, they often provide complementary information about the surroundings; and their ambient sensing capabilities provide the opportunity for humans and machines to work together to make sense of ongoing situations. Independently analyzing either one of these two types sensor stream data will result in an incomplete understanding of evolving situations. Therefore, such massive amount of various information from diverse sensor sources calls for a fusion mechanism that combines their information to provide a more holistic view of what is happening.

However, the fusion of physical sensors and social sensors for situation awareness is still in its infancy. It lacks a unified framework to aggregate and composite real-time media streams from diverse sensors and social network platforms. Heterogeneous data from different modalities, different spatio-temporal resolution, and sensor noise in the huge volume information pose a big challenge.

In this thesis, we aim to fuse physical sensors (i.e. CCTV camera, weather stations) information with social sensors (Twitter) information through different frameworks and methods, so as to detect the occurrence of large-scale events, enable spatial prediction of situations as well as enhance situation understanding.

First, we proposed an innovative multi-layer tweeting cameras framework integrating CCTV camera feeds with surrounding geo-tagged Twitter content to detect various concepts of real-world events. Specifically, we applied visual concept detectors on cameras and construed detected concepts as regularly posted “camera tweets”. We represented these camera tweets using a unified data structure named probabilistic spatio-temporal (PST), which was then aggregated to a concept-based image (Cmage) as a common representation for visualization. In addition, a set of operators and analytic functions were defined so that a user can apply them on the PST data to discover occurrences of events or analyze evolving situations. Mining emerging topics discussed on Twitter, we obtain the high-level semantic meaning of detected events in images. The conceptual framework is implemented and showcased by a Raspberry Pi based “tweeting camera” that is able to sense, analyze, learn and “tweet” on Twitter, with humans in the sensing loop.

Second, we proposed a novel hybrid fusion strategy which, based on the Cmage representation from our first work, models sparse sensor information using Gaussian Process, fuses event signals with a Bayesian approach, and incorporates spatial relations between sensor and social observations. This work shows that the proposed approach can reduce the sensor-related noise, locate event place, improve event detection reliability, and add semantic context for further interpretation of events.

Third, we proposed a novel unified matrix factorization based model to fuse physical and social sensor signals for spatio-temporal analysis. Readings of physical sensors signals are represented by a spatio-temporal situation matrix, which then incorporates social content that can provide explanations for the physical signal strengths. This work improves the detection of an event and enables situation prediction by leveraging correlation of the two types of sensors.

To sum up, the three works have explored fusion method for physical and social sensors from different aspects. The conducted experiments suggest that fusing complementary sensor information can help humans have a better understanding of evolving situations. In the end, the thesis concludes with findings and gives possible future directions for multimodal sensor fusion research.

List of Tables

List of Figures

Chapter 1

Introduction

We are currently witnessing an explosion of digital data in the age of big data [72], in particular, big sensor data. With the prevalent use of sensing devices, the rise of internet-of-things, and the rapid growth of social media [75], humans and devices are more connected than ever before, and society is becoming increasingly more instrumented, generating vast amounts of information describing ongoing situations and occurring events. On one hand, sensors are widely distributed due to the decrease in cost and the development of embedded systems and integrated chips. These sensors endowed with sensing, processing and communicating capabilities, are constantly monitoring our environment and detecting real-world events in a non-invasive and timely manner. From visual sensors to wearable or mobile sensors, we consider these sensing devices as ***physical sensors*** including cameras, accelerometers, gyroscopes, mobile phones, RFID tags, temperature sensors, humidity sensors, etc. Applications of physical sensors in situation awareness range from surveillance [85, 97], smart homes [108], to event detection [51]. On the other hand, fast-growing online social networks services and platforms such as Twitter, Weibo, Facebook, Youtube, Wechat and etc., are resulting in an increasing amount of user-generated content. Humans using such social media platforms to post timely reports can be regarded as ***social sensors*** [98], which also enable a wide range of situation awareness applications such as large-scale event detection [23], noise pollution detection [44], citizen sensing [102] and various spatio-temporal pattern analysis [105].

Although these two types of sensors observe our surrounding utilizing different mechanisms with respect to their sensing rate, spatial distribution, signal presentation, they both monitor the same environment and view situations from different but complementary perspectives. This demands the combination and fusion of these sources' information for a holistic view of ongoing situations. However, due to the diversity of these sources and different modalities of their sensing data, fusing their information together is a big challenge. This dissertation identifies and tackles the challenges in the field of physical and social sensor fusion. It proposes different methods of combining and fusing them in order to provide a better understanding of ongoing situations.

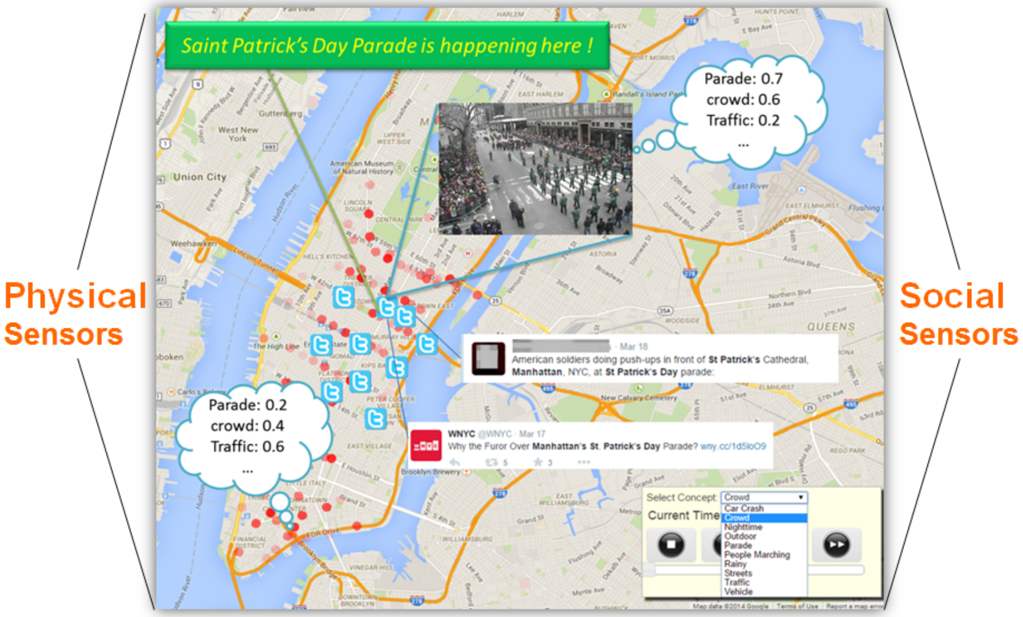
In the rest of this chapter, we first provide the background of situation understanding using physical sensors and social sensors in Section 1.1, and a brief introduction of multimodal sensor fusion. Section 1.2 describes the motivations and the scientific problems in our works, followed by our research contributions in Section 1.3. In the end, we provide an outline of this thesis.

1.1 Background

1.1.1 Situation Awareness and Event Detection

"Situation" has considerably varying definitions across different domains such as robotics [78], context awareness [81], control system [74], surveillance [122], and military [33]. A general definition describes situational awareness as "the perception of the elements in the environment within a volume of time and space, comprehension of their meaning and the projection of their status in the near future" [32]. It is described by a model containing three hierarchical phases [34], namely, perception of the elements in the environment (level 1), comprehension of the current situation (level 2), and projection of future status (level 3). From perception to decision making, a situation has been recently computationally defined as "an actionable abstraction of observed spatio-temporal descriptors" [104], by which humans are able to generate actionable insights from diverse data streams. Situation awareness (SA) applications deal with recognizing when sensed data could lead to actionable knowledge [90].

Summarizing the above definitions, SA in human/machine systems involves two aspects: external and internal. The external aspect describes information within the region of time and space that is observed by sensors or environment measuring devices, while the internal aspects relate to human perception and inferences drawn from external information and predictions. Specifically, we refer external information to the streaming data captured by physical sensors monitoring the physical world and the internal information resides in the social sensors detecting events derived from human interpretation. The distinction is respect to the whole system with human interaction involved, and the decision maker is a user who uses the system which fuses multi-source information. To a human, external information means the physical sensors which capture objective event signals, whereas internal information means the relative subjective social sensors information, which is easier for the user to interpret. Central to the problem of situation awareness is that of detecting, analyzing and predicting occurring events that characterize particular spatial, temporal as well as semantic patterns. We adopt the definition of SA from [104] as "An actionable abstraction of observed spatio-temporal descriptors." Specifically, the "abstraction" is represented by semantic concepts extracted from physical sensors and trending topics from social sensors. An event is composed of a set of correlated concepts from a particular spatio-temporal point and the situation is the understanding and evolvement of an event across a larger area in a longer time series. The conceptual illustration of our idea of fusing physical and social sensors for situation awareness is shown in Figure 1.1. It represents fused sensor streams for understanding ongoing situations, which include different elements in the pipeline of situation understanding from initial data collecting to final event visualization:



[redacted] · Mar 18
American soldiers doing push-ups in front of **St Patrick's** Cathedral, **Manhattan**, NYC, at **St Patrick's Day** parade:

WNYC @WNYC · Mar 17
Why the Furor Over **Manhattan's St. Patrick's Day** Parade? wny.cc/1d5loO9
5 3

Select Concept: Crowd

Figure 1.1: Tweeting Cameras and Twitter Tweets for Event Detection.

1. **Data Capture**

   The data used can be from either physical sensors or social sensors. Different multimedia streams entail diverse data collection mechanisms to capture heterogeneous sensor measurements. The data streams can originate from a personal mobile device, social network updates, videos captured by cameras, new information from websites or archived data sources, describing dynamics of a region (e.g., state of the civil, occurring events, or transportation and information infrastructures). Therefore, a fusion framework should be durable to support a large number of different types of raw data relevant to a particular situation.

2. **Information Aggregation**

   The data captured from different sensor sources should be combined and aggregated in a unified format so that they are suitable for further analysis. Modules processing data of various sensor types should be coupled and integrated seamlessly for information analysis in a multi-sensor situation awareness framework.

3. **Spatio-temporal-semantic Dynamics**

   Based on the properties of situation awareness (SA), time, space and semantic information are the most significant elements that constitute the description of the situation. On one hand, temporal information indicates when events happen as well as the status evolution in a time frame. On the other hand, spatial information provides knowledge of hot spots of the events. In addition, the comprehension of dynamic situations should be expressed by the information with semantic meaning.

4. **Multimodal Fusion**

   To detect or recognize dynamic situations, a set of appropriate data operators and analysis tools aiming at pattern mining and event learning should be built in the situation awareness system. Fusing data from different sensors involves converting data from raw format to low-level features and then high-level decisions; it is necessary to create operators or data analyser targeting at extracting multi-level information from both physical and social sensor data streams.

5. **Event Visualization** To facilitate human understanding about a situation, it is necessary to utilize a suitable visualization tool for situation representation. Possible visualization tools include maps, timeline, or storyboard. The visualization should ideally present the evolving situ-

ation or social dynamic in an efficient and direct way, in which basic situation elements such as time, location, event name, and information quantity are presented in a holistic view.

A situation is usually implicitly captured by signals from streams of different types of sensors. Understanding fused situation requires detecting and inferring events in each individual modality. While multimodal event signals could come from all kinds of media sources, this thesis focuses on the physical and social sensors with respect to the works of event detection.

1.1.2 Event Detection Using Physical Sensors

Physical sensors are now embedded with increasingly powerful processing and are able to communicate with each other through the Internet. They are widely distributed and used for the task of event detection and situational information collection and understanding in many areas such as phenology study, surveillance scene and environment understanding [15,47]. For example, a multi-tier network SensEye of heterogeneous cameras has been proposed to overcome the disadvantage of single-tier networks in a surveillance application, performing object detection, recognition and tracking tasks [63]. Distributed smart cameras [22], which combine video sensing, processing, and communication on a single embedded platform, are also being widely used in camera sensor networks to produce alerts if certain types of unusual behaviour [20] or abnormal events [2] occur. In multimedia applications, *event* is an elementary concept and event-related models [117], as well as the concept of atomic and compound events [12] are proposed in several works focusing on video applications. The tasks of event detection include identifying and locating specified spatio-temporal patterns, such as waving hands or picking up objects in crowd [53], detecting unusual events [127], or analysing human action behaviours [2,80]. Also, a number of works about image captioning [36,52,111] and video concept detection [18, 50], are trying to bridge the gap between machine-oriented low-level features and human-friendly high-level semantics. Moreover, with the sensor types becoming diverse and their sensing capability increasing, fusing different modalities of sensor information has drawn large attention in solving various multimedia problems [10].

1.1.3 Situation Understanding Using Social Sensors

Many online social network services are prevalent nowadays by which users share personal opinions, disseminate breaking news and discuss trending topics. The concept of "social sensor" was first introduced by the work of event detection using Twitter for detecting and tracking earthquakes, typhoons or traffic jams [98]. Twitter, as one of the most important social sensors, has attracted a large number of works for event detection [24], topic discovery [42], as well as content analysis [69]. A news processing system, TwitterStand [100], has been built to capture and investigate latest breaking news. By analyzing news related tweets, it automatically obtains breaking news and current hot topics, filtering out noise that does not belong to the news domain. Social streams with a specific set of keywords are monitored and classified into events and non-events. Events in the work, however, are only recognized for specific predefined keywords, which limits its usage for general automated event detection. Similarly, a framework constituted by event clustering, feature extraction and classification steps has been proposed to distinguish real-world event and non-event twitter messages [16]. [113] proposes a situation awareness algorithm to detect geo-spatial events in a given monitored geographic area, which offers a detailed summary of events. However, the events detected are limited to a small local area. An overall situation cannot be inferred due to the geographic limitation. Aggregating large-scale social information streams from various locations into a unified platform allows users to understand evolving situations in a holistic view. Twitris [102] captures spatio-temporal-thematic properties in processing large scale social data, and integrates semantic context from multiple web resources, which facilitates social sensing in a broad variety of application domains. To understand various events, [105] takes social media data which express social interest of users as “social pixels” and spatially aggregates them into “Emage”, an event data based analogy of image. All these works look into situation awareness from different perspectives. A more comprehensive literature survey on social sensors in situation understanding is given in Chapter 2.

1.1.4 Multimodal Sensor Fusion

Multimodal fusion refers to the integration of multiple media, their associated features, or the intermediate decisions in order to perform an analysis task [10], such as semantic concept detection, audio-visual speaker detection or human tracking. It has been well studied in combining different modalities of same sources or from synchronized multiple physical sensors. Examples include a framework of heterogeneous cameras for object detection, recognition and tracking [63], multi-level audio-video integration of camera networks and microphone arrays for semantic event processing [110], mixture of text and video analysis for video retrieval [123], or fusing cameras with depth sensors for person re-identification [83]. While traditional multimodal fusion mainly focuses on physical sensors information, with the emergence of social platform, social information is combined with physical sensors to provide conceptual details or semantic meanings of detected events and situations. For example, EventNet [125] utilizes tag keywords, meta-data and surrounding text of a YouTube video to automatically learn a large set of event-specific concepts for procedural and social events. Lanagan et al. [67] combine the tweet information with sports video shot boundaries to detect events such as goals and penalties. In the crowd sensing area, geo-social media is used to combine with mobile data [120] and GPS data [87] to analyze human mobility. We also proposed a Tweeting Camera Framework [115] integrating both physical sensors and social sensors to detect various concepts of real-world events, which will be elaborated in Chapter 3.

1.2 Motivation and Research Problems

Streams of data from multiple modalities provide complementary information to each other in facilitating event discovery and situation awareness. However, due to the diversity of these sources, physical sensors and social sensors capture information separately in their individual silos. The information captured by sensors of different modalities is not combined or fused which impedes event detection and understanding in a comprehensive manner. Different properties of these two types of sensors make the fusion of heterogeneous information a challenging problem:

- **Multiple modalities:**

  Different modalities may produce signals with different formats and properties: physical sensors (e.g., CCTV cameras or a temperature sensor) may produce sensed numeric data (e.g., geo-coordinates, pixels, temporal sound waves), while social sensors are often in rich of text. Although several works try to bridge them [87] by shared proximities (e.g., similar locations and time durations), fusing these two different modalities in a unified model is a hard problem due to their heterogeneous representation and different levels of information.

- **Multiple sources:**

  Even if they are of different modalities, signals from the same source are explicitly correlated (for example the frame and the audio signal at a specific time point). Fusing signals from different sources is more challenging because the correlation between physical and social sensors could be weak since humans do not necessarily post about what physical sensors capture.

- **Different spatio-temporal density:**
  From the spatial aspect, due to their cost, physical sensors are usually distributed sparsely in a region, while humans have no limitations in choosing where they want to post a message, and the user generated content is usually spatially denser than the physical sensors especially when some large-scale events are going on. From temporal aspect, physical sensors have advantages because they are designed to sense the environment continuously. In contrast, this is not applicable to social sensors by the nature of humans beings; most people only "tweet" spontaneously.

- **Approximate sensing:**
  Physical sensors often produce noisy observations for a number of reasons such as the failure of sensors, environmental changes or the maintenance of devices, making it more difficult to rely on the readings and exploit the correlation between the two different modalities.

These problems raise several **research questions** we need to solve:

1. How to integrate heterogeneous data from both physical sensors and social sensors to detect real-world events?

2. What kind of processing framework should be adopted in order to extract meaningful situational information from multi-modal media streams?

3. Given the intrinsic unreliability of individual sensor data and the sheer volume of social media data, how can we handle the uncertainty and noise of these data?

4. How to utilize multimodal complementary information from multiple sensors distributed in a large place to suppress the sensing noise for situation understanding?

5. What kinds of fusion strategies should be adopted to make appropriate use of properties and characteristics of such multimodal spatio-temporal-semantic data?

This thesis aims at solving these questions by proposing a framework and two methods that integrate and fuse physical and social sensors information in a unified data structure and representation.

1.3 Scope and Contributions

1.3.1 Aims

This thesis' objective is to provide novel multimodal sensor fusion framework and methods which can:

1. obtain appropriate information from physical sensor networks;
2. utilize social information to enhance event understanding;
3. fuse information from both physical sensors and social sensors for event detection and situation prediction.

1.3.2 Contributions

This thesis contributes towards the problem of physical and social sensor fusion for situation awareness and event detection. The main contributions are as follows:

- **A multilayer tweeting cameras framework** We design a multilayer tweeting cameras framework that integrates physical sensors (CCTV cameras) with social information (Twitter Tweets). It automatically detects and broadcasts high-level semantics, which are more intelligible for humans. The framework reduces network traffic because high-bandwidth videos are not broadcast, but only short semantic compressions. The proposed unified probabilistic spatio-temporal data structure can handle the uncertainty of physical sensors and the aggregation of physical and social sensors addresses the unreliability issue of individual sensors and thus improves the overall performance of event detection. In addition, the proposed novel tweeting camera paradigm enables event learning and notifying, which complements the classic streaming-based approach. This paradigm not only protects the privacy of the detected objects by generating only semantic information but also facilitates architecture that allows the sensing process to be customized for different applications or particular purposes with humans in the loop.

- **A Concept image based hybrid multimodal fusion method** We propose a concept image (namely Cmage) based hybrid fusion method featuring sensor decision and spatial information; Spatial sparsity issue due to physical sensors distribution is solved with embedded Gaussian Process and heterogeneity problem is addressed with a Bayesian fusion method to combine event decisions from both physical and social sensor Cmages. Our proposed fusion method provides not only a better visualization of event summary but also the convenience of manipulation of the event-related sensor and social signals. The fusion strategy can effectively remove noise from the data streams, accurately locate the event place and offer more detailed situational semantics.

- **A matrix factorization based fusion method** We introduce an innovative way of fusing social and physical sensors by taking account of numerical readings and semantic text simultaneously. Our proposed matrix factorization based fusion model opens the possibility of fusing spatio-temporal numeric and semantic data for various tasks, by utilizing the correlation between physical and social sensors. This fusion method applied on two different large real-world datasets suggests how the social sensor can contribute to event analysis tasks in the fusion process. It results in higher performance in event classification and prediction and is generalizable to spatio-temporal data fusion tasks.

1.3.3 Significance

This work contributes to the fusion of physical and social sensor by applying statistical models and mathematical based fusion techniques (Bayesian fusion, Gaussian Process and Matrix Factorization) and utilizing the correlation between these two types of sensor modalities. The study will have significant impact on multimodal sensor fusion since 1) it can demonstrate the feasibility of framework and methods that fuse both physical and social sensors; 2) it will provide the mechanism of smart socialized sensors network; 3) it could facilitate people's response to the emergency by providing situation oriented information, which humans can refer to for proper action takings.

1.3.4 Scope

The work is limited to physical sensors and social sensors that can both observe same physical event or situation. This implies that analyzing the trending topics captured by the social sensor but not seen from physical sensors, or detecting physical events that contain no social feeds are beyond the scope of this thesis. In addition, distribution of sensors for better coverage of situation is another research question beyond our scope.

1.4 Thesis Organization

The thesis is organized as follows. Chapter 2 surveys a breadth of situation awareness using physical sensors and social sensors as well as multimodal fusion techniques. Chapter 3 presents a tweeting camera framework which combines CCTV camera feeds with social information and a new paradigm of tweeting camera network. Chapter 4 discusses a hybrid fusion method on top of a unified Cmage-based representation. Chapter 5 presents a matrix factorization based method of fusing physical sensors and social sensors for situation prediction and event detection enhancement. Chapter 6 concludes the thesis with suggested future work.

Chapter 2

Related Works

This chapter presents a survey of research works related to multimodal sensor fusion for situation awareness. The survey mainly focuses on event detection using physical sensors and social sensors. We first briefly review physical sensor fusion area, listing works related to event detection using physical sensors, operators required in manipulating sensor streams, and the internet of things. Secondly, we give a comprehensive survey on social sensors for situation awareness, which mainly investigates how the fastest-growing microblogging service Twitter is utilized to conduct event detection, breaking news as well as trending topic detection. In the last section, we review the works on the fusion of heterogeneous information, which give different fusion techniques that handle data from multiple sources, of different modalities and data representations.

2.1 Physical Sensors Fusion for Situation Understanding

With an increasing number of sensors having capabilities of sensing, processing, communicating, usage of sensors in event detection and situation awareness is spreading. Massively distributed visual sensors (webcams) are being utilized for phenology study, scene and environment understanding [15,47]. Data from these ambient sensors are being fused and analyzed to detect real-world occurring events and evolving situations

2.1.1 Event Detection using Physical Sensors

Kulkarni et al. [63] proposed a multi-tier network SensEye of heterogeneous cameras to overcome the disadvantage of single-tier networks in a surveillance application, performing object detection, recognition and tracking tasks. The network consists of 3 tiers, each of which contains homogeneous sensors while they are heterogeneous between tiers. Fusion of different sensors is likely to achieve less energy consumption. One of the multi-sensor fusion problems is to select a proper model from correlated sensor data streams. Different sources give different confidences in detecting events. Atrey et al. [12] p-resented a novel framework that finds the optimal subset of media streams in order to achieve the system goal under specified constraints. A dynamic programming approach is used to find the optimal subset of media streams based on three different criteria regarding 1) probability of achieving the goal, 2) confidence in the achieved goal and 3) cost to achieve the goal. Event detection in surveillance scenario (including events such as running/walking, knocking/talking/shouting) with two cameras and microphones demonstrates the feasibility of model selection strategies.

In the machine perception area, human activities and interactions will be effectively and efficiently supported because of embedded multimodal sensory systems and databases of semantic events. Trivedi et al. [110] developed a multimodal system which integrates different modalities. The sensory information of this system comes from camera networks and microphone arrays. The camera networks include 4 omnidirectional cameras, four rectilinear cameras, and the arrays include 12 microphones. The system contains a semantic event database which can retrieve activities at high levels of semantic granularities. However, the system is not generally designed and limited to the special structure of an intelligent room.

Events are an elementary concept not only for the human brain but also in multimedia applications. Different multimedia applications such as eChronicles, life logs, and event-centric media management, content analysis, surveillance all have different event definitions and processing mechanisms, sharing a various established notion and model of events. Therefore, a common multimedia event model such as [117] is required to offer unified base representation, media indexing, common event management infrastructure, exploration and visualization. In addition, unusual event detection has also drawn much attention. Examples include a framework that consists of unsupervised clustering and the Coupled Hidden Markov Model [127], a method that extracts and processes low-level observations from local "monitors" [2] and a system that involves detection, tracking and behaviour analysis in airborne moving platform [80]. To bridge the gap between machine-oriented low-level features and human-friendly high-level semantics, a number of works concern image captioning [36, 52, 111] and video concept detection [18, 50], where the task is to assign concept labels to an input image/video along with their associated probabilities. Moreover, multi-modal sensor fusion has been well studied for combining multiple physical sensor modalities for various multimedia tasks [10].

2.1.2 Streaming Data Operators

To manipulate sensor data for event detection, a proper set of operators should be well defined to query for different situations. A lot of work has been done towards efficiently processing query streams of relational data. A various number of systems have been proposed that support an SQL-like query language for querying relational data streams: TelegraphCQ [25], Aurora/StreamSQL [1,48], CQL/Logical Stream Algebra [60], and similar systems. Other works focus on the fundamentals of query processing by extending or adopting the notions of the Relational Algebra to new required concepts. For example, [79] describes temporal algebras to address the time aspect of the stored data. However, these works do not consider streaming data, i.e., queries are processed against the currently available data. Existing stream processing systems as mentioned above map the SQL-like query languages or stream-specific extensions of SQL to an algebraic level. This typically involves, adopting the definitions of existing operators to the streaming paradigm as well as the definition of new operators to accommodate processing tasks not being handled by existing operators. [41] proposes an algebra that takes both the temporal and streaming characteristics of the data into consideration.

2.1.3 Smart Cameras and Internet of Things

Today's smart camera systems are usually designed with larger memory, considerable computing power, and wired or wireless communication interfaces. A number of smart cameras with onboard image processing and analysis capabilities have been built to accomplish visual analysis oriented tasks such as object detection and classification, event detection and situation awareness [95,103]. Some of the pervasive smart camera prototypes are implemented based on standard, off-the-shelf components [101]. However, most of the smart camera-related works focus on the research issues such as the architecture of smart devices and networks [30], privacy protection in visual sensor networks [97,119], or visual processing [3]. Rather than letting cameras work passively, we consider taking a camera node as an active sensor participating in a social-cyber-physical world, which establishes a connection with humans and offers useful real world situation information to people.

With the Internet of things (IoT), everyday objects now have the ability to interconnect not only among themselves but also humans. Social networking concepts have been integrated into IoT [13], which establishes Social Internet of Things (SIoT). An architecture for SIoT has been presented in [14]. Things not only sense but start to update their status on social networks. Kranz et al. [61] make both humans and technical systems together to form a socio-technical network by describing cognitive office, where the states of the plant, windows and doors are posted via Twitter accounts. Many accounts of similar function have been created. For example, @VedamsIoTEdison tells if room lights are off or on; @VedamsIoTRPi tweets when there was a power cut in office hours; @MoneyPlantTrack not only posts a plant's temperature, humidity, status but also uploads images captured by the camera that is monitoring the plant.

2.1.4 Video Concept Detection

Another stream of research on event detection is the semantic analysis of events, including detecting concepts from videos streams. The aim of video concept detection is to rank video shots according to the presence of semantic concepts (e.g., "sports", "crowd","people marching", etc.), which can act as semantic filters for different multimedia applications.

There are many works on concept detection, including Columbia374 [124], VIREO-374 [50], CU-VIREO374 [49], NUS-WIDE [27], EventNet [125] and Mediamill-101 [106]. Table 2.1 summarizes the works on concept detection based on the number of concepts, classifiers, and the concept examples.

These works release different set of concept detectors trained using various features (colour moment, texture, bag-of-works feature points, etc.) or various fusing techniques (late decision fusion or early feature fusion) and can be directly used for various multimedia applications.

Concept detectors learn from training samples, the mapping between a set

Table 2.1: Summary of Concept Detection Works

Work	**Number of Concepts**	**Classifier & Features**	**Source**	**Concept Examples**
Columbia374 [124]	374	SVM edged direction histogram (EDH), Gabor (GBR), grid color moment (GCM).	broadcast news videos	Events, Objects Locations, People
VIREO-374 [50]	374	DoG detector, SIFT descriptor (bag-of-words)	broadcast news videos	Events, Objects Locations, People
CU-VIREO374 [49]	374	local keypoint features global features	broadcast news videos	Events, Objects Locations, People
Mediamill-101 [106]	101	SVM Textual Feature color-texture Feature	broadcast news videos	Events, Objects Locations, People
EventNet [125]	4490	CNN deep learning	YouTube	Knife, Food, Hand, Pets, Sandwich
NUS-WIDE [27]	81	k-NN color histogram, color correlogram, edge direction histogram, wavelet texture, blockwise color moments, bag of words (SIFT descriptions)	Flickr	Airport,Animal Beach, Birds, Person, Ocean, Protest

of low-level visual features (local descriptors, color, texture, etc.) to a concept with particular semantic (e.g., parade, crowded, face, etc.), but usually achieve low detection accuracy due to the so-called *semantic gap* between the image features and the conveyed meanings perceived by a human being. Many works [35, 86] have been devoted to improving the detection accuracy and adding more classifiers. To explore dynamic spatiotemporal correlations among primitive concepts, Bhatt et al. [18] proposed utilizing dynamic spatiotemporal correlations of the given ontology rules; based on the accuracy of detection of concepts, these ontology rules could be updated. The method considers a new paradigm of mutual learning between multimedia data mining and ontology to define and discover *Adaptive Ontology Rules* (AOR), which can adapt to dynamic correlations, primitive concept detection inaccuracy and uncertainty in spatiotemporal relations.

2.2 Social Sensors for Event Detection

Online social media platforms (e.g., Facebook, Twitter, Weibo, Youtube, Youku and etc.) have revolutionized our way of communicating with each other. Social media embedding higher-level comprehension patterns and forecast operators have demonstrated the ability to enhance situation awareness [126]. Previously these social network services focus mainly on establishing connections amongst people and helping them share information in their networks [31]. Nowadays, due to the blossoming of all kinds of microblogging and media sharing services, people use them to report daily news/events [6, 64, 70, 71, 94, 98, 113, 116], disseminate breaking news [46, 66, 100], discuss trending topics [5, 17, 68, 76, 128] or express feelings/opinions [4, 59, 89, 96, 114]. This results in a huge volume of user-generated content (UGC) that has drawn significant attention from researchers.

2.2.1 Event Detection Using Twitter

Twitter as one of the fastest-growing microblogging services has become a mature tool for real-time event detection and situation monitoring. A well-known example is to use it to detect and track earthquakes, typhoons or traffic jams in Japan using Twitter streams [98]. In this work, Twitter users are considered as "social sensors" that sense the situation of surroundings and give the measurement in terms of tweets. To detect specific events (e.g., earthquake, typhoons), a classifier is devised to classify if the tweets are event or non-event related, based on tweets features including target events keywords, the number of words and their contexts.

In addition, Kalman filtering and particle filtering are applied to estimate events location and to trace how situations evolve across space and time. An earthquake reporting system is then designed to promptly detect an earthquake by monitoring the tweets. Due to the rapidity and ambiance properties of the tweets related to the earthquake, the system is able to detect an earthquake with high accuracy and deliver notifications even faster than the announcements broadcasted by the meteorological agency.

Focusing on crime and disaster related events (CDE) (e.g., shootings, car accidents, tornado), Li et al. proposed [71] a Twitter-based Event Detection and Analysis System (TEDAS) which not only detects and analyses new events' spatial and temporal patterns, but also identifies the importance of the detected events. The system contains components of crawling, Twitter-specific and CDE-specific features based classification, importance ranking, location extraction as well as a map visualization, and allows a user to interact with it by specifying spatial, temporal or topic queries. However, the system relies heavily on manually predefined CDE terms, and not able to detect events in real-time.

Predefining rules or selecting keywords/hashtags as done in the works above limits the generality of event detection. Many works, therefore, try to detect events using unsupervised or semi-supervised methods. For example, Becker et al. [16] used online clustering techniques to group together tweets that are similar in topics. Features are extracted in cluster level including temporal (volume of frequent cluster terms), social (retweets, mentions, replies), topical, and Twitter-centric (hashtags) features, and a classifier is trained with the cluster based feature using support vector machines. Walther et al. [113] proposed first clustering spatially and temporally similar tweets together and then classifying the results into event or non-event clusters by decision tree based classifier trained on manually labelled clusters. A bunch of cluster level features (e.g., common theme, tweets counts, present tense, etc.) are defined to differentiate positive and negative samples. The system supports real-time processing of tweets and provides GUI showing where the events are happening. However, the detection focuses only on a relatively small geographic region which should be pre-specified.

To detect emerging events in a fully automatic framework, Weng et al. [116] tackled event detection with clustering of wavelet-based signals (EDCoW). In order to filter out meaningless "babbles" in twitter streams, this method builds signals for individual words which can be quickly computed by applying wavelet analysis on the frequency of the words. After trivial words with low auto-correlation in terms of similarity being filtered out, the remaining words are then clustered to form events with a modularity-based graph partitioning technique. However, since the semantics of words are not explored, words associated with different real-life events are potential to be grouped together. Similarly, Li et al. proposed [70] a segment-based event detection system for tweets, *Twevent*, which detects bursty tweet segments that are then clustered as event candidates. Different from previous work that relies solely on Twitter data, the system integrates Wikipedia to identify the realistic events and provides the description of the identified events. Also based on clustering method, Kumar et al. [64] proposed an emergency event detection framework to handle the informal nature, the high volume, and high-velocity characteristics of Twitter text streams. The inclusion of a temporal model improves the efficiency of event detection and supports identifying sub-events within a larger event.

Apart from public events, wellness information posted in many social platforms is also utilized to detect and analyze personal wellness events [40, 92]. For example, Akbari et al. [6] presented a novel supervised model to categorize tweets into different wellness event taxonomies including three main events diet, exercise, health and other sub-events. In addition, relations between event categories are also analyzed to learn task-specific and task-shared features. Reis et al. [92] used user tweets to estimate the effect of exercise on mental health. The volume of a users' tweets expressing anxiety, depression, and anger is estimated and compared to two group of people to mine the relationship between mental health and physical exercise.

To sum up, Twitter provides greatly valuable user-generated content that can be transformed into actionable situational knowledge. A large number of works have been conducted using tweets to detect events. Different works tackle a variety of tasks and detect different types of events including general events, crisis or emergency events, personal heath events, etc. Also, a variety of features are extracted from the text and different classifiers are used to detect events.

2.2.2 Breaking News and Trending Topic Detection

If the spatial information is ignored, event detection is technically similar to another two areas in social media analysis: news detection and trending topic detection. The difference is that breaking news are not necessarily events occurring in our physical world, but simply featuring bursty occurrence of some particular topics.

Teitler et al. built a news processing system, TwitterStand [100], to capture and investigate latest breaking news from Twitter feeds. By analyzing news related tweets, the system can automatically obtain breaking news and current hot topics, filtering out noise that does not belong to the news domain via online clustering. In contrast to traditional news wire services, sources of the system come from many identities of the contributors/reporters who are not known in advance.

Kwak et al. [66] conducted the first quantitative study on the entire Twittersphere and explored how trending information is diffused in Twitter. Trending topics were classified into exogenous and endogenous categories, indicating breaking news or self-reporting. User participation and active period of trending topics are also investigated. Compared with trends in other media, the work shows that the majority (over 85%) of topics are headline news or news persistent in nature. Further supporting the statement, Twitter has been proven as the first source to the breaking news and has convinced a large number of its audience before mainstream media reported the news [46]. Hu et al. [46] provided an in-depth analysis of how the news broke and spread on Twitter and discovered that individuals affiliated with media played, mass media and celebrities are the main groups of people that play the key role in news diffusion. In this work, bag-of-words are extracted from tweets to be classified via SVM into "certain", "uncertain" or "irrelevant" categories describing the reliability of a given tweet.

In the textual data mining area, emerging trends is a topic area that is growing in interest and utility over time [56]. With the prevalence of social media platform, an enormous number of research works focus on detecting topics that were previously unseen or rapidly growing in importance online. For example, TwitterMonitor [76] is the first system that performs trend detection over the Twitter stream in real time. A trend is identified as a set of bursty keywords that suddenly appear in tweets at an unusually high rate and occur frequently together. Besides trend identification, the system extracts frequently mentioned entities to discover multiple aspects of the trend. The trending order is potentially adjustable through user interaction with the system via GUI.

To understand what detected trending topics are about, Lee et al. [68] classified these topics into 18 general categories such as sports, politics, technology, etc. Naive Bayes multinomial classifier with bag-of-words and tf-idf weights were used to classify the tweets. A decision tree was used to categorize similar topics by the number of common influential users between the given topic and its similar topics. Similarly, Zubiaga et al. [128] introduced a topology to classify trending topics into four general categories: news, current events, memes, and commemoratives. A set of predefined language-independent features such as bag-of-words, retweet, hashtags and etc, were used to discriminate trending topics via an SVM-based classifier.

Benhardus et al. [17] defined trending topic as "a word or phrase that is experiencing an increase in usage, both in relation to its long-term usage and in relation to the usage of other words." In this work, standard features such as tf-idf, unigrams, bigrams, and etc. were used in detecting and identifying trending topics in Twitter streams. The trending topic identification results demonstrated the ability and feasibility to extract and identify relevant information from a continuously changing corpus with an unconventional structure. To summarize information originating from social sources, Aiello et al. [5] conducted a comprehensive comparison of six topic detection methods on three datasets related to large scale events or topics including the FA Cup and US elections. The work found that the volume of activity over time, the sampling procedure and the pre-processing of the data, as well as the type of used detection method (e.g., LDA, FPM, BNgram and SFPM) all greatly, affect the quality of detected topics.

In summary, Twitter, as one of the most popular microblogging services, provides rich content for situation awareness including detecting and diffusing events, breaking news and trending topics. Those events detection methods are sometimes similar to the approaches to breaking news or trending topic detections. Events are normally detected by classification or clustering techniques with all kinds of features of different levels. A more comprehensive review of techniques in finding real-world occurrences of events that unfold over space and time could be found in [9]. Table 2.2 summarizes the abovementioned works in terms of different properties including main features extracted from the data, the classifier in the tasks, learning methods, event type, spatio-temporal as well as categories.

2.3 Fusing Heterogeneous Information

Real world phenomena are now being observed by multiple complementary sensors streams. Deriving actionable insights of evolving situation, therefore, requires fusing heterogeneous information in terms of data characteristics. Such heterogeneity is reflected by data from different sources, of different modalities, distributed in different locations and having different reliabilities. For example, sensors reading streams usually give data in the numeric form (such as image pixels, humidity values, pm2.5 readings, sound waves, etc.), while social feeds streams contain mainly symbolic text (such as tags in Flickr, posts in Twitter or Facebook). In recent years, more and more research works in the multimedia area are trying to fuse these heterogeneous data with various information fusion techniques or frameworks, generating more holistic or global view of ongoing situations.

Table 2.2: Summary of Works on Situation Understanding using Social Sensors

Works	Main Features	Classifier	Learning Methods	Event Type	Spatial& Temporal	Category
[98]	keywords, #words, context	SVM	supervised	Specified	both	Emergency Event Detection
[16]	#words, retweet, reply, mention, topic, hashtag	Online Clustering, SVM	semi-supervised	Unspecified	temporal	General Event Detection
TEDAS [71]	URL, hashtag mention, time, location	Nearest Neighbour	unsupervised	Unspecified	both	General Event Detection
[113]	text, #tweet of clusters	Decision Tree	semi-supervised	Unspecified	both	General Event Detection
[116]	tf-idf,wavelet energy, entropy, H-Measure	Wavelet Analysis	unsupervised	Unspecified	temporal	General Event Detection
Twevent [70]	term/user frequency, content	kNN graph	unsupervised	Unspecified	N.A.	General Event Detection
[6]	nGrams, hashtags	Bootstrap	supervised	Specified	N.A.	Wellness Event Detection
[92]	unigrams, bigrams, lexicon, emoticons	Logistic Regression	supervised	Specified	N.A.	Wellness Event Detection
[64]	hashtag, #tweets, #users	Nearest Neighbour	semi-supervised	Specified	temporal	Emergency Event Detection
TwitterStand [100]	tf-idf	Online Clustering	unsupervised	Unspecified	both	Breaking News Processing
[66]	reply, mention, retweet	Page Rank	N.A.	Unspecified	N.A.	Breaking News Processing
[46]	bag-of-words	SVM	supervised	Unspecified	N.A.	Breaking News Processing
[76]	#words, keywords	PCA, SVD	unsupervised	Unspecified	temporal	Trending Topic Detection
[68]	bag-of-words tf-idf	Naive Bayes Multinomial Decision Tree	supervised	Unspecified	temporal	Trending Topic Detection
[128]	bag-of-words, retweet, tf-idf hashtag, tweet length, reply	SVM	supervised	Unspecified	temporal	Trending Topic Detection
[17]	tf-dif, entropy unigrams, bigrams,	Nearest Neighbour	unsupervised	Unspecified	temporal	Trending Topic Detection
[5]	tf-dif, unigrams, bigrams,	LDA, FPM, BNgram, SFPM	unsupervised	Unspecified	temporal	Trending Topic Detection

2.3.1 Multi-source Fusion

To understand various events, patterns and emerging situation, Singh et al. [105] designed abstraction and tools to analyze spatio-temporal patterns of a situation using social media data. Taking humans as sensors, the work aggregates social media data which express social interest of users about a particular theme from any particular location into “social pixels”, a unified spatio-temporal-thematic data structure. The social pixels were combined spatially to form E-mage, an event data based analogy of image, for situation visualization. Besides, a declarative set of operators and media processing operations upon such data structure were defined to analyze aspects of the situation including important parameters, the patterns of the situation, or macro events that were relevant to the application domain. Based on the E-mage representation, a system named EventShop [37] was built to combine streams from heterogeneous data sources, to process them to detect situations. Such spatio-temporal-thematic data structure was also proposed by Sheth et al. and used in the system Twitris [84,102] which captured spatio-temporal-thematic properties in processing large scale twitter data, and integrated semantic context from multiple web resources, facilitating social sensing in a broad variety of application domains. The system comes with a web mashup application that is able to explore social signals by tracking particular events and providing event related popular topics with semantics.

Mining multimodal geo-social media data from different social media platform, Hsieh et al. [44] developed a joint inference and visualization system that integrated multimodal features that were able to reason and present urban noise pollution. The data come from New York City related to noise from social reports or posts. A GUI was provided to allowed users to understand the noise composition in a particular region. New Your City has the huge amount of all kinds of social or sensor data and is full of diverse activities. Kuo et al. [65] explored various perspectives of the dynamic of the city from rich and diverse social media content including Instagram, Flickr, TripAdvisor, Twitter and open data). Specifically, a broad spectrum of life aspects including trends, events, food, wearing and transportation were analyzed through multi-source and multi-modality data. Activities in New York City across social media in both visual and semantic perceptions were discovered and a number of interesting applications revealing patterns related to urban dynamics (e.g., traffic pattern, sentiment, human activities and fashion styles) of NYC were also demonstrated.

2.3.2 Multi-modality Fusion

Thanks to the fast-growing online social networking servings and the proliferation of user-generated content, social information has been demonstrated contributive to the tasks used to be accomplished by physical sensors. Incorporating social feeds with semantics facilitates a better understanding of events or ongoing situations. *EventNet* [125] utilizes tag keywords, meta-data and surrounding text of a YouTube video to automatically learn a large set of event-specific concepts for a procedural event and social event. The work provides a large-scale structural event ontology from social tags for the video captured by physical sensor cameras, which offers great potential for unseen event retrieval and browsing. Similarly, tags of Flickr images are used to train event-driven concepts [26] for various event detection tasks. Lanagan et al. [67] combined the tweet information with sports video shot boundaries to detect events such as goals and penalties. The volume of tweets during the game were shown to be an effective and accurate mean of event detection, and meanwhile, the discussed content offered the semantic sense of what people were talking about at event moments. However, the method is only applicable to specific events due to predefined features, selectively filtered keywords and hashtags, which limits the generalization of event detection.

The works mentioned above only look at semantic concepts or temporal patterns evolution, yet spatial information as a critical aspect revealing geo-locations of events or situations is not considered. Due to the proliferation of geo-enabled smartphones and GPS sensors, geographical locations recorded in the form of latitude and longitude are becoming basic meta-data for newly user generated content. In crowd sensing area, geo-social media is used to combine with mobile data [120] and GPS data [87] to analyze human mobility. Specifically, [120] combines mobility data and geographically surrounding tweets together to understand semantically why and where a person travels to in particular time. In [87], GPS trajectories of vehicles and geo-tagged tweets from microblogging service Weibo are also combined to identify traffic anomalies. When there are significantly different traffic patterns being discovered, social media is leveraged to annotate the unusual pattern. Besides human mobility analysis, geo-tagged social information is also used to combine with traffic cameras for event detection. However, the methods in these works can not handle the noise in the data and do not provide a unified model to fuse physical and social sensors. They could only be considered as shallow integration, in the sense that they focus on obtaining the semantic explanation of physical sensors and there is no unified method that simultaneously takes these two sources of information into account, which is what we propose in this thesis. The data heterogeneity of different sources in our problem is reflected by the numerical data representation from physical sensor readings and the semantic text representation from social sensor feeds. In the multimedia event detection area, this problem has never been considered. However, the fusion of such heterogeneous data is extensively studied in recommendation systems area [57], where multiple aspects of user and items (e.g., movies, music, products) are combined to find the patterns of users interest in products for a personalized recommendation. In the recommendation system works, a set of users give ratings to different items and this is formalized in a rating matrix where rows and columns represent different users and items respectively and the value of each cell represents the rating given by a user to a corresponding item. Matrix factorization [99] as a collaborative filtering algorithm is a common way of discovering the latent associations between the user and matching items. The advantages of matrix factorization is that besides the raw ratings, it allows incorporating additional information such as user social relations [109], reviews [77] independently [73] or simultaneously [45]. Such additional information (e.g., item reviews, implicit feedbacks), in terms of data format, is in line with social sensor content in our problem. This gives a hint of using matrix factorization as a alternative and complementary method to the fusion of multi-modal data, especially physical sensor and social sensor data.

2.3.3 Summary

Various methods of fusing heterogeneous data from physical sensors to social sensors are presented in this section. A summarization and comparison of these works in terms of different situation aspects and source properties is provided in Table 2.3. Although situation awareness applications have been investigated extensively with either physical sensors or with social sensors, to the best our knowledge there is no previous work that analyses these two modalities simultaneously in a unified framework that takes into account all situational aspects including spatial, temporal and semantic information, and meanwhile considers multiple independent sources with noise, which are the main goal of this thesis.

Table 2.3: Summary of Fusion Methods and Sources

Works	Multiple Sources	Multiple Modality	Spatio -temporal Fusion	Semantic Fusion	Fusion Method	Sources	Applications
Eventshop [105] [37]	x	x	x	x	unified data structure	Twitter, Flickr	personalized alert system disaster management business analysis
Twitris [84] [102]	x		x	x	unified data structure	Twitter	event detection situation awareness semantic analysis
[44]	x	x	x		integration system	Foursquare Twitter Flickr Gowalla	noise pollution detection
[65]	x	x	x	x	integration system	Instagram Flickr TripAdvisor Twitter open data	traffic pattern analysis sentiment analysis human activities fashion style
EventNet [125] [26]		x		x	deep learning	YouTube	event concept construction
[67]	x	x		x	temporal correlation	Twitter Video	sports activity analysis
[120]		x	x	x	spatio-temporal correlation	Twitter mobility data	human moving pattern analysis
[87]	x	x	x	x	spatio-temporal correlation	Weibo taxi GPS	human moving pattern analysis
Tweeting Cameras [115]	x	x	x	x	spatio-temporal correlation	Twitter CCTV camera	event detection situation understanding
[57]		x		x	matrix factorization	product ratings, review	recommendation
[109]		x		x	matrix factorization	product ratings, review social relationship	recommendation

Chapter 3

Tweeting Cameras for Event Detection

3.1 Overview

Thanks to the widely distributed visual sensors (e.g., surveillance cameras) and the prevalence of social sensors (e.g., Twitter feeds), many events are implicitly captured in real-time by these sensors of heterogeneity. However, due to the diversity of these sources, physical sensors and social sensors capture information separately in their individual silos. Since the camera and social streams provide different facets of events or a situation, the accuracy of event detection and evolving situations comprehension can be significantly improved if these two complementary sensor streams are fused together for evaluation or analysis. In addition, sensors are not passively observing environment but starting to report surrounding situation if it goes abnormal [29, 62]. Many works propose future IoT architectures that integrate physical sensors and make them perform a social behavior when they are connected [14].

In this chapter, we present an innovative multi-layer tweeting cameras framework to combine physical and social sensor data; we also implement a new tweeting camera paradigm that connects cameras with a human to facilitate event detection and enhance situation understanding. In order to process social data and sensor data, we define a unified probabilistic spatio-temporal (PST) data structure to represent semantic concepts information, which aids handling the uncertainty and noise in sensor and social streams respectively. We apply concept detectors on the images captured by cameras to indicate the confidence of a detected concept from the image. We consider these detected concepts as “Camera Tweets” with associated confidence values indicating the probability of happening event containing such concepts. Camera tweets (CamTweets) at a particular geo-location can then be visualized as “concept pixels” since they represent concept signals emerging from a particular geo-location if we consider a broad region (having distributed signals of different strength) as a situation image [105]. Spatially aggregating such “concept pixels” creates a powerful and intuitive situation visualization interface, *concept-based image (Cmage)*, which enable fusing both social information and sensor information in an image-based representation (discussed in Chapter 4). To achieve this goal, we propose a multi-layer tweeting camera framework where tweets from cameras are analyzed at different levels such that multi-level information is extracted and subsequently combined with social information to derive situational knowledge.

Images or videos from visual sensors (e.g., CCTV cameras) are important information sources for all kinds of monitoring, surveillance, and observation tasks and serve as the biggest contributor to the phenomenon of Big Data. The handling and processing of such data typically require a good number of hardware resources. To complement the idea of tweeting camera, we design and implement a real smart camera that can sense, analyze, “tweet” and learn. Like smart cameras, we pre-process the data to avoid sending the raw camera feed and to significantly increase the level of privacy. We introduce an architecture that can easily be customized for different needs and applications. The proposed smart camera uses individually trained classifiers for the detection of a set of events such as “lights on/off” and “meeting in progress yes/no”. As a result, we can transform low-level visual data into a high level of information on the camera itself which can be viewed as a form of “semantic compression”. The level of information allows for sharing camera data in a new way. By following a camera over Twitter, not only humans but also systems can subscribe to camera output for further analysis. Moreover, by replying the “Camtweet” of a tweeting camera, a human can rectify or confirm the information posted by the camera and the camera will automatically update its knowledge by re-adjusting its learning model.

The rest of this chapter is organized as follows. Section 3.2 introduces our proposed multilayer tweeting camera framework that combines sensor readings with social information. Section 3.3 describes the detail of the data processing procedure in the framework. Section 3.4 illustrates our proposed new paradigm of socialized smart camera as an implementation of proposed framework. Section 3.5 demonstrates the feasibility of our work using three different datasets (New York real-time traffic camera feeds, NUS foodcourts' camera feeds and Twitter data from New York City), conducts evaluation experiments with four instances of real-world events as well as a face recognition application and discusses issues related to our work. Section 3.6 concludes the chapter.

3.2 Multi-layer Tweeting Camera Framework

We propose a multi-layer tweeting cameras framework as shown in Figure 3.1. The framework consists of three layers, namely low-level concept detection layer, mid-level concept filtering layer, and high-level social sensor fusion layer. In the first layer, low-level concepts are detected from raw sensor images through applying various concept detectors. A unified probabilistic spatio-temporal (PST) data structure represents the low-level concepts. Those camera “tweet-

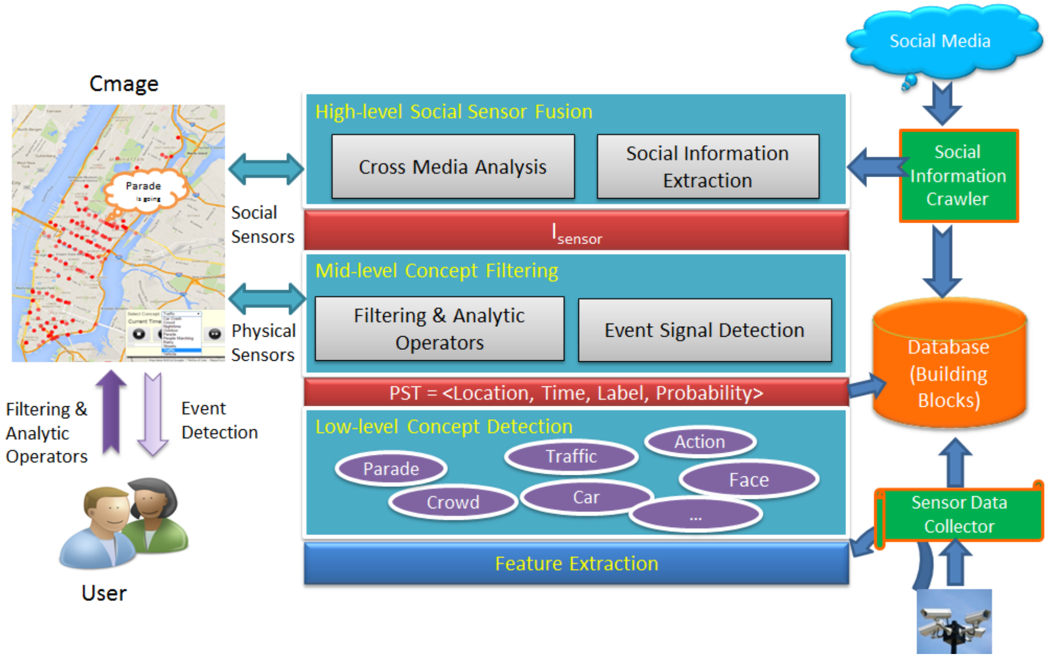

Social Media
Cmage
Parade
High-level Social Sensor Fusion
Cross Media Analysis
Social Information Extraction
Social Information Crawler
Social Sensors
I_{sensor}
Mid-level Concept Filtering
Filtering & Analytic Operators
Event Signal Detection
Physical Sensors
Database (Building Blocks)
PST = <Location, Time, Label, Probability>
Filtering & Analytic Operators
Event Detection
Low-level Concept Detection
Action
Traffic
Parade
Face
Car
Crowd
...
Sensor Data Collector
Feature Extraction
User

Figure 3.1: Overview of Proposed Multi-layer Tweeting Cameras Framework.

s" of low-level concepts are then aggregated and processed at the second layer using a set of predefined operators and functions. Signal detection theory is applied to detect abnormal patterns indicating occurring events. In the third layer, social information and the processed camera tweets are fused to derive high-level semantics. It provides an effective data processing pipeline to convert the raw media streams (either live camera feeds or real-time tweets) to different abstraction levels and finally facilitates event detection. By aggregating multiple individual sensor feeds via a common representation, the framework provides a visualization interface as well as information filtering tools based on a set of predefined analytic functions customized for PST data processing. This enables a user to be able to gain a global understanding of occurring events by manipulating the sensor feeds. In the following, we detail how (where, when, what) camera tweets are analyzed and combined with social media data in this framework.

3.2.1 Data Collector and Storage

The data collector component is required for pulling in raw data, by which we crawl raw sensor data (e.g., image sequences from surveillance cameras), and obtain social media (e.g., tweets from Twitter) using their respective APIs. After obtaining the data, we use MongoDB to store the crawled images and tweets from Twitter. In addition, low-level concept information represented by the unified data structure (defined in Section 3.3.1) is also stored in the database for querying and further analysis.

3.2.2 Low-level Concept Detection

To bridge the gap between low-level features and human interpretable meaning, a wide variety of detectors have been created in many works to extract semantic information from images or videos. Concept detectors [39,50], object detectors [38], face detectors [112] are examples of these advances. In the first layer, we incorporate a set of concept detectors to detect a variety of concept-s from the camera feeds. Here the concepts could be faces, objects, actions or general entities with semantic meanings. These concept detectors, which are essentially statistical models or classifiers, assign text labels (tags) to the sensed data. Specifically, we adopt the VIREO-374 detectors [50] which can detect 374 general concepts defined in LSCOM [54] including “parade”, “crowd”, “traffic”, “people marching” etc. These detectors, with a mean average precision of 16%, are not yet capable of providing accurate performance and are associated with uncertainty. The uncertainty is represented as a probabilistic confidence score indicating the probability that an observation is correctly classified into the concept category. If concept detectors are periodically (e.g once every 10 seconds) applied to the camera data, we can consider the camera to be *tweeting* these labels and a set of cameras in a geographic region can be considered to be a *network of tweeting cameras*.

In addition, spatial and temporal aspects have been found to be critical for describing a situation or event [117]. Therefore, this layer models the outputs of concept detectors (camera tweets) in a unified data representation called probabilistic spatio-temporal data (PST data), which contains four elements, including camera location information, temporal information, the label of the detected concept and the associated probability as the confidence value of the label. We consider the confidence of detecting a concept at any location to be like the intensity of pixels in an image, and term it as a "Concept Pixel". Such "Concept Pixel" represents a basic concept of an event and is considered as a small signal that provides a clue about the holistic situation. Therefore, spatially aggregated data from a set of cameras can be construed to be an image.

Moreover, tweets represented by the PST data serve as the input for higher information filtering and are indexed in the repository (stored in MongoDB) for querying and textual pattern mining. Therefore, cameras whose feeds are analyzed in the framework are constantly tweeting low-level concepts in the first layer and simultaneously pushing PST data into database for indexing.

3.2.3 Mid-level Concept Filtering

In this layer, PST data from each camera is aggregated for the holistic representation. Specifically, "Concept Pixels" from a geographic region of interest are visualized in a map-based form called the "Concept Image" (Cmage). Concept filtering operators can then be applied on the Cmage to facilitate event detection.

Filtering Operators and Analytic Functions: A set of pre-defined filtering operators and analytic operators has been designed to analyze such integrated information. For example, a user can use filtering operators to query situational information about a specific concept from a particular location given specified time and probability threshold. In addition, we formally define basic analytic functions for statistics, such as min, max, sum, count, smooth, extremes, trend, abnormal, clustering and density function that can be applied to the PST data as well as the Cmage. For example, a user can check the weak signal trend of a particular concept, can obtain knowledge of when an event occurs and which region has a higher confidence of detecting specific concepts as well as how such concept confidence rises and falls with time.

Event Detection: Cameras in open environments do sensing under different and often noisy ambient conditions. Therefore, PST data is usually noisy. To overcome this problem, we use Signal Detection Theory [118] which models the detection task by checking objects/concepts being present or absent with the threshold set by "observers". It consists of two distributions, namely "noise" distribution and "signal + noise" distribution. We model detector results using Gaussian distributions where non-event results are considered as "noise", and event results are considered as "signals". Given this, we define event detection goal as separating the "signal" from "noise".

3.2.4 High-level Social Sensor Fusion

At the third level of this framework, we integrate both sensor information and social information so as to obtain a high-level semantic information of events which can be used for decision making and action. In such cross-media analysis, we try to leverage information from social media onto physical sensor data and vice versa. For example, when the tweeting cameras sense an unusual number of concepts from a specific location, we try to mine representative terms from the social media like the geo-located messages posted on Twitter. Tweets in the camera regions are collected and grouped into different clusters based on message content (topic, keywords, hashtag, etc). We utilize the location and time information obtained from physical sensors to filter out nonconcept related posts to enhance the efficiency. We then calculate the most dominant cluster (that contains most similar tweets) as the emerging topic in that particular location and compare current frequent terms in tweets with historical tweets to discover most discussed topics for the rising intensity of particular concept detector results (details provided in Section 3.3.5). Therefore, high-level knowledge is obtained in the form of social context (hot topics and messages in tweets) combined with mid-level information from physical sensors.

To summarize, in the proposed framework, camera tweets in the first layer are represented as probabilistic spatio-temporal data coming from multiple cameras, which describe low-level concepts and emit weak signals of events. Mid-level information is obtained at the second layer by aggregating and processing individual low-level tweets using various filtering operators and analytic functions. High-level knowledge is derived by fusing both sensor information and social sensors information for situation understanding.

3.3 Processing Framework

In this section, we elaborate on the probabilistic spatio-temporal data structure, the aggregation format Cmage, as well as the filtering operators and analytic functions that can be applied on the data structure. We show how signal detection theory is leveraged upon to detect abnormal events. In addition, we illustrate how physical sensors and social sensors are utilized to complement each other for event detection. The notation we use throughout this chapter is defined in Table 3.1

Table 3.1: Tweeting Camera Framework Notation

Symbol	Description
D_i	concept detector i
l_i	semantic label extracted by concept detector i
N	number of cameras
CAM_i	camera i
$p_j^{CAM_i t}$	the confidence value of label j detected from camera i at time t
pst	probabilistic spatio-temporal element
Θ	query operators
$\mathbb{S}_{func\_name}$	statistical functions

3.3.1 Probabilistic Spatio-Temporal Data

Let D be a detection system that consists of a set of r concept detectors $D = \{D_1, D_2, \ldots, D_r\}$. Let l_i be the corresponding semantic label extracted by various concept detectors D_i, $L = \{l_1, l_2, \ldots, l_r\}$. For example, D can be VIREO-374 [50] where $r = 374$. The N cameras in the system are defined by $\mathbb{CAM} = \{CAM_1, CAM_2, \ldots, CAM_N\}$. As we assume that each concept detector D_i will assign a symbolic label l_i as well as the probability value p_i, we define $0 \leq p_j^{CAM_i t} \leq 1, (1 \leq i \leq N, 1 \leq j \leq r)$, be the confidence value of label l_j from by detector D_j, being applied to the raw media data captured by camera CAM_i at time t. Let $S = \{S_1, S_2, \ldots, S_n\}^t$ be the processed *probabilistic spatio-temporal stream* from whole camera network, where $S_i = \{(l_1^{CAM_i}, p_1^{CAM_i})^t, (l_2^{CAM_i}, p_2^{CAM_i})^t, \ldots, (l_r^{CAM_i}, p_r^{CAM_i})^t\}$ represents concept information detected from each individual camera.

Definition (PST: Probabilistic Spatio-Temporal Data)

The fundamental building block for low-level concept representation is the probabilistic spatio-temporal element pst.

$$pst = [location, time, label, probability, pointer] \tag{3.1}$$

where:

- $location = [lat, lon]$ represents the geo-location – latitude and longitude – of the camera location. We assume that the camera is static here but it naturally can be extended to mobile cameras as well.
- *time* stores the time information of captured data.
- *label* represents semantic concept such as *car*, *human*, *crowd*, *parade*, etc., detected in the stream. Generally, these concepts express low-level abstraction of information which could be semi-reliably detected by existing detectors or classifiers.
- *probability* is the confidence value in [0,1] representing the output of a concept detector as a probability value.
- *pointer* points to the actual raw data stream. While our intention is to abstract the raw data into a concept level data structure, it is also necessary to store the reference to the real data for further validation.

An example of pst data describing the "crowdedness" situation at 5th Avenue of 11 Street Manhattan on November 24th, 2014, 15:10:16 is as following:

$[5Ave@11Street, 20141124 : 151016, crowdedness, 0.9, image\_path]$;

3.3.2 Cmage

Extending the idea of "Emage" which represents aggregated social interest of users [105], we project the probabilistic spatial temporal data with attached concepts onto the spatial map to form a *Cmage (concept image)*, to provide an intuitive visualization. The Cmage is a pseudo image presentation of an ongoing situation in a region, constituted by concept pixels indicating a confidence value of a particular semantic concept. Each pixel in a Cmage is a PST point that contains the probability of that concept being detected by concept detectors from an image captured at a particular spatio-temporal unit. Note that there is one Cmage for every concept. Let X be a 2D point set (lat, lon). A Cmage on X at a given time t is any probabilistic spatial temporal element $(L \times V)^X$, where L is the concept labels set and V is a probability value set of real values between 0 and 1. A Cmage is denoted as: $g^l = \{(x, p) | x \in X = \Re^2, 0 \leq p \leq 1, l \in L)\}$. A Cmage example is shown in Figure 3.2. The campus is divided into 3 × 4 grids. The 7 foodcourts are at cells numbered from 1 to 7. Higher pixels intensity means higher confidence of "Crowdedness" in that spatio-temporal point.

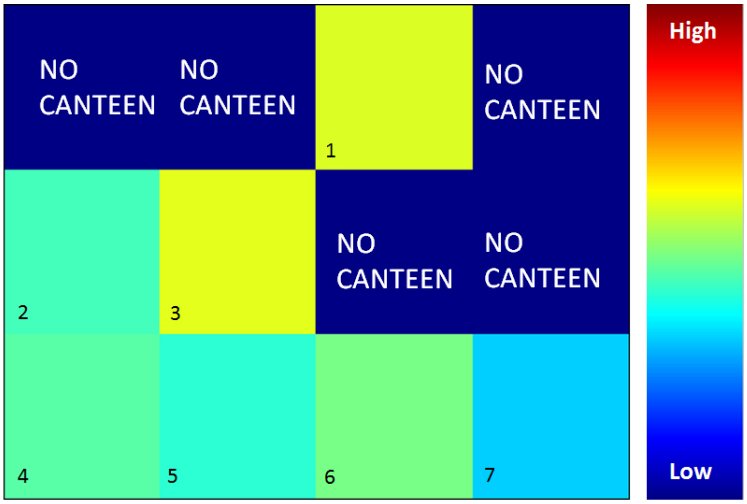


Figure 3.2: *Crowdedness* Cmage of NUS Foodcourts at 12:00 on 24th March 2014.

3.3.3 PST Data Filtering Operators

In order to efficiently retrieve relevant PST data by describing the data properties, we provide a set of basic operators for a user to query based on specific PST elements. The user can also apply analytic functions after obtaining the relevant subset of PST data. Hence the framework is envisaged to be used interactively by the user to gather insights.

Context-based Selection of Detector: Based on the prior knowledge of cameras (location, camera properties, history pattern or new information from other sources, e.g., social media), we define a concept selector τ_{task} to allow selecting a subset of cameras for achieving specific concept detection tasks.

$$\{T_1, T_2, \ldots, T_j\} = \tau_{task}\left(T, CAM_i, l_i\right), where : T_j \in T \tag{3.2}$$

Query Operators: Θ select a subset of data from the stream as a filter based on user specification. We use Predicate P as boolean function applied on a "pixel" (PST data point) of Cmage. Based on the four elements of PST data, we provide filtering by the functions indicated by $P\_filter(exp)$:

$$\Theta_{P\_filter(exp)}(S) = \{(l_j^{CAM_i}, p_j^{CAM_i})^t | P\_filter(exp) = True\} \tag{3.3}$$

where P_filter are predicates on an element of PST data.

These filtering functions can retrieve from the PST stream by taking a user-defined pst related expression as the parameter. Once the expression satisfies the predicate, a subset of $S' \subseteq S$ will be returned. Note that since the filter operations are based on predicates, we can combine multiple atomic predicates on pst data to form compositional queries.

Examples: Show the March 17th data for the concept of "traffic" or "parade" at the $5th\ Avenue$ with confidence value higher than 0.8:

Query: $\Theta_{P\_PROP \wedge P\_LAB \wedge P\_LOC \wedge P\_TEMP}(S)$

where:

$P\_PROP = P\_prop(0.8 \leq p)$

$P\_LAB = P\_lab(label = traffic \vee parade)$

$P\_LOC = P\_loc(CAM_i) = 5^{th} Avenue)$

$P\_TEMP = P\_temp(t = March\ 17^{th})$

3.3.4 Data Analysis Functions

As the PST element is a numeric data presentation with spatial, temporal and symbolic information about ongoing events or situations, we define a set of functions and arithmetic operations that could be applied on the PST values to extract characteristics or features of happening events.

Statistical functions: $\mathbb{S}_{func\_name}$ are used to analyse the PST dataset so as to explore the patterns or the nature of the stream by calculating extreme values, mean, trends, change points or other statistic-related indicator or parameters. The set of functions defined in our work is as follows:

a) `mean, max, min, sum`

$\mathbb{S}_{mean}(D)$ calculates the average value of a given set of data. Here we consider the data of the same label. The input data D could be a Cmage set $g_t^{\{l_1,l_2,\dots,l_k\}}$ or a subset of PST stream $S' \subseteq S$ of label l. The function gives an averaged Cmage g_t^m with pixels $cp = (lat, log, t, l, prob_{mean})$, where $prob_{mean}$ is calculated by taking the average probability value along the temporal axis. e.g., showing the average intensity of concept c in Cmage between t_1 and t_2 : $g^m_{t_1->t_2} = \mathbb{S}_{mean}(\Theta_{P\_lab(label=c) \wedge P\_tem(t_1 \leq t \leq t_2)}(S))$. $\mathbb{S}_{max}(D)$, $\mathbb{S}_{min}(D)$, and $\mathbb{S}_{sum}(D)$ are computed in a similar way.

b) `extremes`

Extended to the max and min functions, $\mathbb{S}_{extremes}(D)$ calculates the PST data's local minima and maxima along the temporal axis as well as among spatial regions, corresponding to the probability values. The results are computed by comparing current data points with nearby data in a spatial region or with close data in the temporal axis. The output are the PST data with a $tag_{extreme} \in \{crest, trough, plateau\}$. **Example:** show the peak hours in foodcourt A: $\mathbb{S}_{extremes}(D)(\Theta_{P\_lab(crowd) \wedge P\_loc(can_A)}(S))$

c) `trend`

A tweeting camera keeps sensing the environment at all times, so it would be helpful to design a function $\mathbb{S}_{trend}(D)$ to discover the social trend or changes from certain concepts pattern along time [7]. The function calculates the gradient of every data point along time series and returns every PST data with a $tag_{trend} \in \{ascending, descending, plateau\}$ as well as the trending rate $r \in \Re$. **Example:** show the trend of crowdedness in foodcourt B: $\mathbb{S}_{trend}(D)(\Theta_{P\_lab(crowd) \wedge P\_loc(can_B)}(S))$

d) `smooth`

Along with the temporal dimension, a concept detector (e.g., car detector) may be unreliable due to the environmental changes (e.g., illumination change or occlusion); therefore, the PST data generated by the sensor and hence the low-level concept detectors contain noise that may affect the further analysis. The $\mathbb{S}_{smooth}(D)$ function smooths the PST data with Gaussian filter and convolution operator, so as to remove the noise in the data.

e) `outlier`

The abnormal data pattern is regarded as important information that deserves an alert for the tweeting camera system. A $\mathbb{S}_{outlier}(D)$ the function is defined for extracting statistically abnormal data points from PST dataset. The function uses normal distribution model to fit the dataset and calculates the mean and variance of the observation. After that, it allows the user to specify a threshold as an abnormal pattern in terms of σ. **Example:** show the time when the crowd concept has an abnormal intensity during a particular period. $\mathbb{S}_{outlier}(D, \sigma)(\Theta_{P\_lab(crowd) \wedge P\_tem(t_1 \leq t \leq t_2)}(S))$

Data Mining with PST: Computing spatial clusters/segments help in better characterizing the situation across regions [8]. We define the clustering function $\mathbb{CL}$ to group a set of PST data or 'pixels' of a Cmage in various dimensions (spatial, temporal, concept) based on the probability values of each data points. For example, $\mathbb{CL}_{loc}(\Theta_{P\_lab(c)}(S)) = \{loc_1^{gl}, loc_2^{gl}, \ldots loc_n^{gl}\}$ gives the locations of each group $gl \in \{gr_1, gr_2, gr_3\}$, where in gr_i the probability values of the subset data points $(l_r^{CAM_i}, p_r^{CAM_i})^t$ of concept l are close to each other.

Density Function: $\mathbb{C}$ takes the PST dataset and calculates the number of elements that satisfy a predefined requirement. The set could be a Cmage, the whole PST dataset or a sub-stream of PST dataset selected by the filtering operations described in the previous section. It can be used on various dimensions of data for deriving the characteristics of that particular dimension. For instance, when a specific event happens, the number of the cameras capturing the concept of this event gives us an intuitive information about the situation.

Example: calculate the number of cameras that detected “person” concept between time t_1 and t_2: $\mathbb{C}_{CAM}(\Theta_{P\_lab(person)\wedge P\_pro(p=1)\wedge P\_tem(t_1\leq t\leq t_2)}(S))$

3.3.5 Incorporating Social Information

Once interesting PST data characteristics has been detected, camera location information is utilised to query social media tweets posted around the camera location, and the time interval during event (e.g., when an anomaly was detected (i.e., $[t_1, t_2]$)) is used to compare current highly frequent tweets with historical tweets, so as to obtain textual information that best describes an event using social media. The text analysis architecture including tweets preprocessing and representative term mining is shown in Figure 3.3.

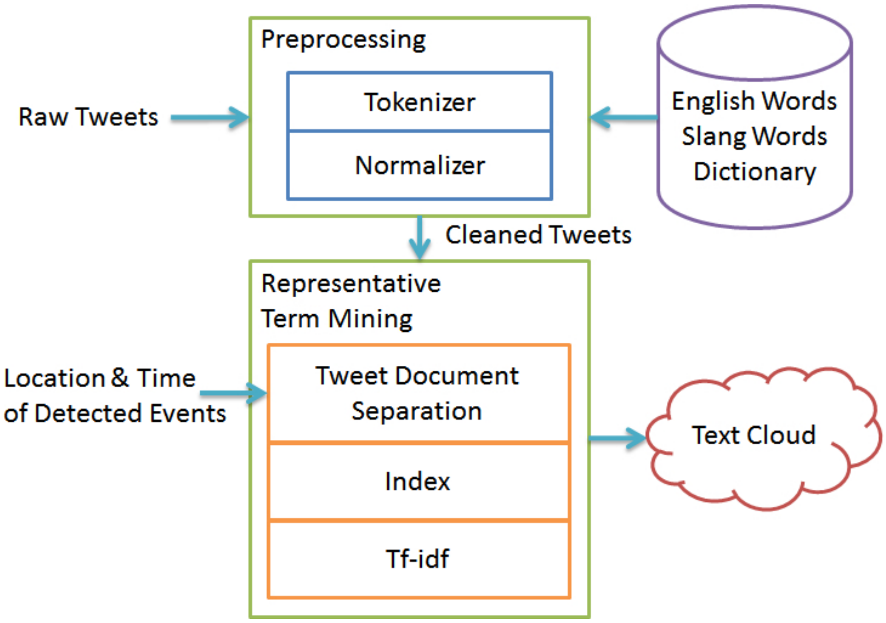


Figure 3.3: Architecture of Twitter Data Processing.

All the tweets posted during $[t_1, t_2]$ are considered as a document denoted as T_C, and the historical tweets denoted by T_H refers to all the documents of other than the event time in the past days; tf-idf is used to analyse the relevance of each term among them once both T_H and T_C are obtained. Equation 3.4

adopted from [87] is used to calculate the weight of extracted terms that could best describe the event.

$$
\begin{aligned}
&w_{term} = tf(term, T_C) \times idf(term, T_H) \\
&s.t. \quad \begin{cases} tf(term, T_C) = \dfrac{f(term, T_C)}{\max\{f(w, T_C), \forall w \in T_C\}} \\ idf(term, T_H) = \log \dfrac{|T_H|}{|\{th \in T_H : term \in th\}|} \end{cases}
\end{aligned}
\tag{3.4}
$$

where *tf* is the function to calculate the frequency of the term in the current tweet document (T_C), and *idf* refers to the calculation of inverse document frequencies in all the historical tweets documents (T_H). Therefore, terms with high weights mean that they are highly discussed in current document (event related topics) and less discussed in the whole collection of historical tweets. To fuse information from both physical and social sensors, we define event signal $Es_e =< I_{se}(e), I_{so}(e) >$ where I_{se} stands for event sensor signal and I_{so} stands for event social signal. Here we take confidence value of a particular concept **c** as I_{se} and term weights of **tw** as I_{so}, in which **c** and **tw** are closely related or the same content. Then we adopt equation 3.5 to fuse them to derive final event signal intensity.

$$
Es(e) = w_{se} * I_{se}(e) + w_{so} * I_{so}(e) \tag{3.5}
$$

where w_{se} and w_{so} are considered as weights of sensors that can be specified by users.

3.3.6 Cameras Tweeting Rate

Since we use the polling method for fetching camera data, this determines how frequently a camera tweets. Currently, we have set the cameras to regularly tweet once every 10 seconds. However, we provide flexibility to the user in setting this camera tweeting rate by $ctr = T(x)$ posts/second.

3.4 A New Paradigm of Socialized Sensors: Realization of Tweeting Camera

Cameras represent one of the most utilized physical sensors to monitor our world. They are also the main contributor to the phenomenon of Big Data. However, the level of detail provided by cameras often raises privacy concerns. Both challenges currently impede the most wide-spread sharing of data stemming from cameras which would enable new types of services and applications. This section introduces a novel smart tweeting cameras paradigm that connects human with pervasive visual sensors through social networks. In the nutshell, we have developed a smart camera that maps the visual data into higher-level concepts using customized classifiers. This avoids the bandwidth-hungry sending of raw camera feeds and intrinsically enables a much higher level of privacy preservation. Using additional light-weight processing on a camera, it only outputs information in case user-defined events occur. We make that low-volume, high-level information available by letting cameras directly tweet their outputs. These outputs can be regarded as the examples of "CamTweets" discussed in section 3.2.2. With that, users or applications can "follow" cameras alongside other tweeting objects, joining a novel type of social cyber-physical ecosystem. In addition, we designed a self-learning module in which enables natural human-camera interaction through a social network (Twitter). In such natural interaction, humans can tell a camera which specific concept (e.g., person name of a new face) it has captured when it thinks having detected abnormal/new events. The replies of CamTweets from humans are considered "labels" for the "training data" that are used for the learning process of the cameras. In the long run, these socialized cameras in our proposed paradigm can automatically update their models and become smarter by learning more "labelled" knowledge with the help of human efforts in the loop. Figure 3.4 and 3.5 show the idea of tweeting camera paradigm and working flow that embeds learning process involving human interactions. In Figure 3.5, each tweet from the "SeSaMeCamera" account represents an example of "CamTweet" introduced in section 3.2.2. Here we disclose the raw image only for illustration purpose. In a real scenario, "CamTweet" could be easily configured not to release raw images.

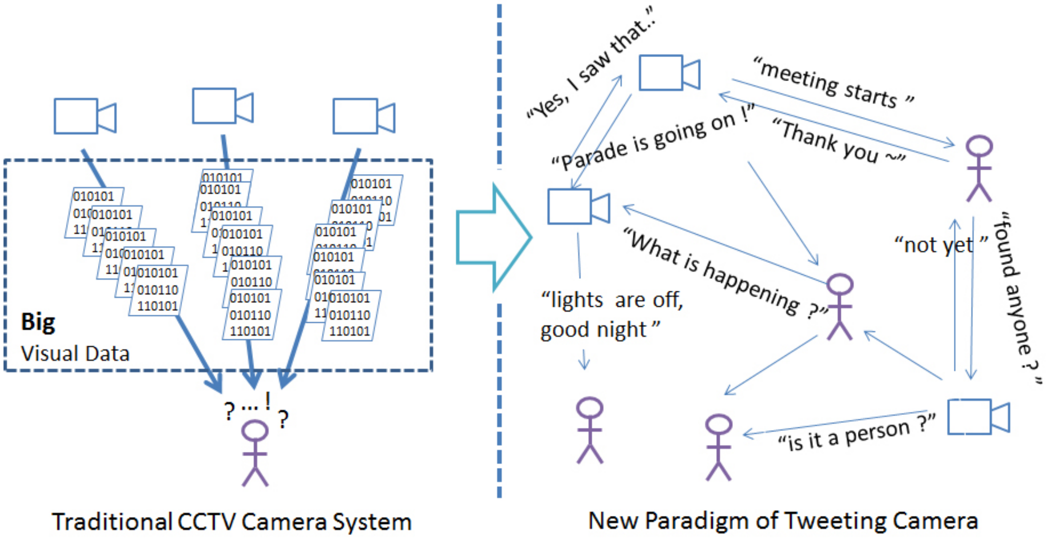
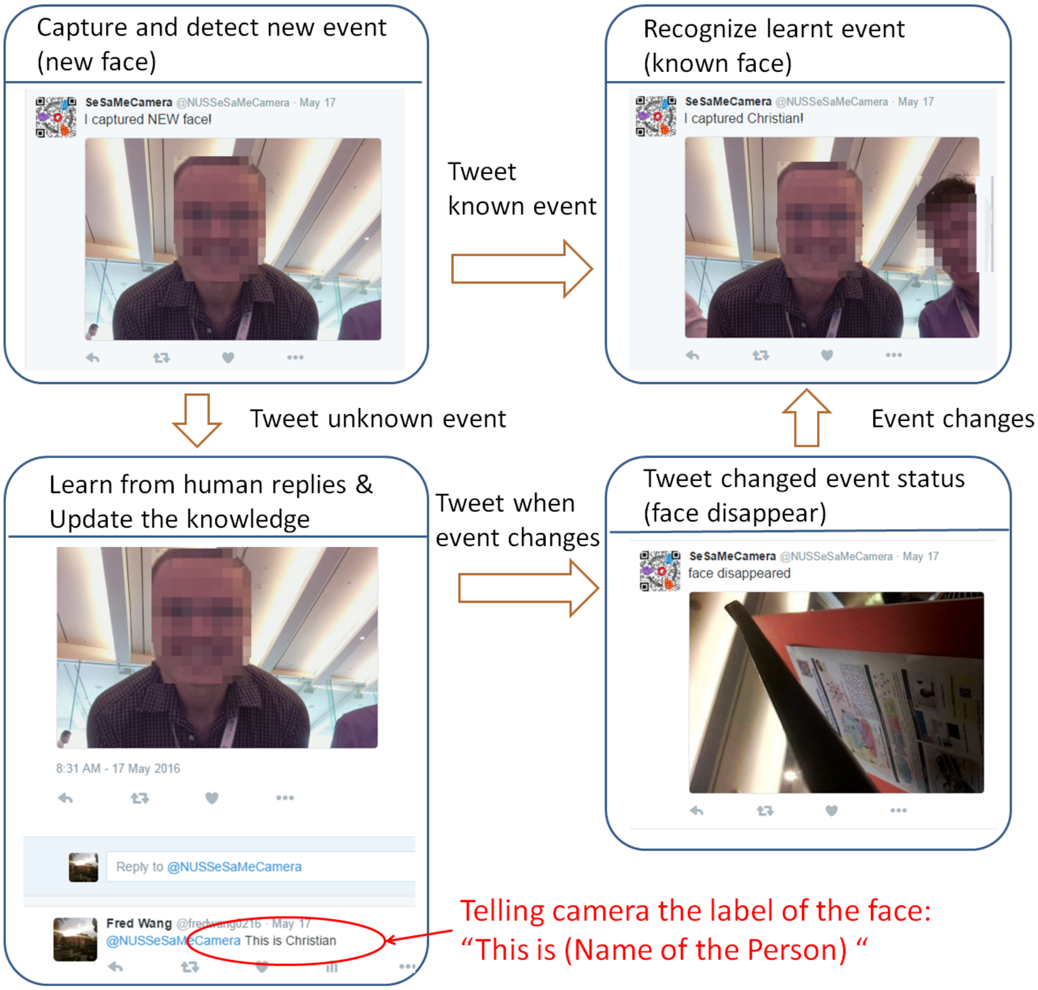


Figure 3.4: From Traditional Camera Network to New Sensing Paradigm made from Tweeting Cameras

3.4.1 Hardware Components

Our camera prototype (Fig. 3.6) is based on a Raspberry Pi 3 single board computer which is equipped with a 1.2GHz 64-bit quad-core ARMv8 CPU clocked at 900 MHz per core, 1 GB of SDRAM and wireless module. A Pi camera or a Sony IMX219 8-megapixel sensor can be connected to the board via the Camera interface (CSI) or USB respectively. Both cameras are capable of maximum resolution of 2592x1944 pixels static images. The system runs an embedded Raspbian Jessie OS booted from a 64G microSD card.


Capture and detect new event (new face)
Tweet known event
Recognize learnt event (known face)
Tweet unknown event
Event changes
Learn from human replies & Update the knowledge
Tweet when event changes
Tweet changed event status (face disappear)
Telling camera the label of the face: "This is (Name of the Person) "

SeSaMeCamera @NUSSeSaMeCamera · May 17
I captured NEW face!

SeSaMeCamera @NUSSeSaMeCamera · May 17
I captured Christian!

8:31 AM - 17 May 2016

Reply to @NUSSeSaMeCamera

Fred Wang @fredwang0216 · May 17
@NUSSeSaMeCamera This is Christian

SeSaMeCamera @NUSSeSaMeCamera · May 17
face disappeared

Figure 3.5: Tweeting Cameras Working Flow

3.4.2 Software Architecture

The tweeting camera contains three software components, namely Event Handler, Logic Processing and Data Communication (Fig 3.7):

The **Event Handler** detects abnormal events or new faces and recognizes known events or faces based on the images captured by the camera. To achieve a high accuracy, we use Visual Recognition of the IBM Bluemix cloud platform [107] for event classification and the Open Biometric Verification library [55] for the face recognition task. The training step is triggered when the camera receives human replies and the camera tweets are generated based on

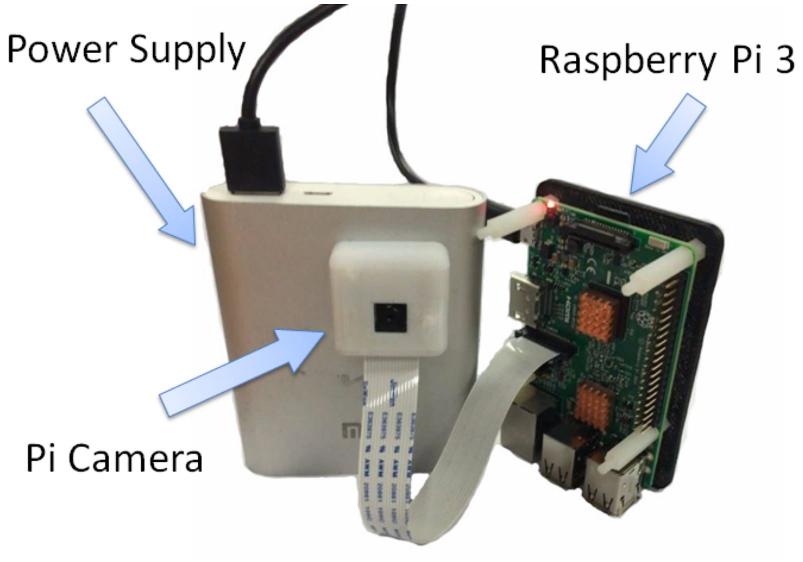


Figure 3.6: Prototype of A Tweeting Camera

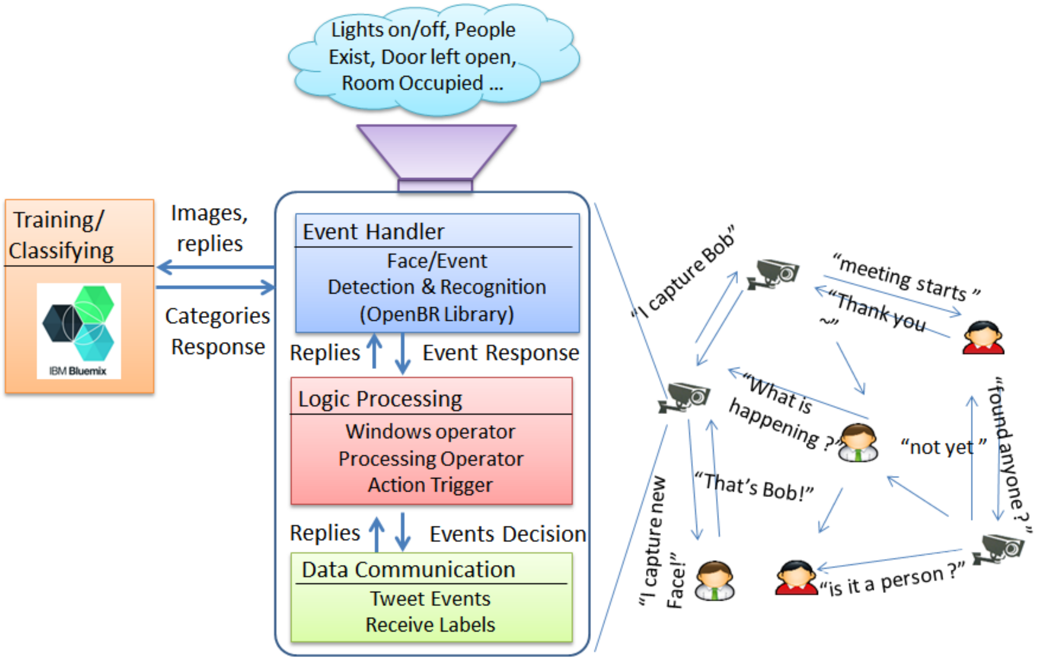


Figure 3.7: Tweeting Camera Software Architecture

the response of cloud services. The **Logic Processing** component processes the responses to notify only about "interesting" events, where a pipeline is implemented using a query algebra to process and manipulate the data, as well as triggers to invoke user-defined actions. The **Data Communication** component implements a Python-based Internet connection with an associated Twitter account. The camera calls the Twitter API to post information from the Logic Processing component as well as to receive replies from humans to trigger a new learning process.

3.5 Experiments and Discussion

This section conducts analysis on our proposed framework that integrates physical camera information with social information for large scale situation awareness and demonstrates how our proposed tweeting camera paradigm could facilitate event detection for daily use in the real physical world.

3.5.1 Datasets

NYC Traffic CCTV Camera

We have crawled live feeds from 150 public CCTV traffic cameras distributed on the roads all over the Manhattan district of New York City, which are under the management of the Department of Transportation. The live cameras provide frequently updated still images from several locations in the five boroughs. The update frequency varies from 1 second to 5 seconds. We have collected our data (resolution of 352×240) during March 13, 2014 to March 19, 2014, June 24, 2014 to August 20, 2014, and October 3, 2014 to October 22, 2014, with a total size of data being 1.23 TB, to ensure sufficient variety. This dataset is denoted as NYC traffic in the remaining section.

NUS Foodcourt CCTV Camera

The NUS (National University of Singapore) foodcourt video dataset consists of feeds from 73 standard CCTV surveillance cameras located at 9 different foodcourts on the NUS campus. Each foodcourt has several cameras facing either seating area or the food stalls areas. The data has been recorded over six months.

Twitter Data

We have crawled tweets using Twitter Streaming API from October 08, 2014 to August 14, 2016, with the geographic bounding box of [40.698770, -74.021248, 40.872932, -73.905459] which includes Manhattan, and collected a total of 42,778,483 records. Each record is stored in the database with the original set containing all data fields such as time of created, geo-location, text etc.

3.5.2 Evaluation Approach

In this study, we analyze the effectiveness and capacity of our framework to detect different events. We evaluate our framework by comparing detected events with ground truth shown in the next section, and illustrate the semantic meaning of the change of sensor data pattern by mining social information.

Events Ground Truth We use the notices posted on the "Weekend Traffic Advisory" website of the New York City Department of Transportation for obtaining the ground truth.[1] This website details traffic alerts in terms of locations of road construction and other events that will affect the flow of traffic for the coming weekend. The ground truth for the events that we try to detect is shown in Table 3.2.

Table 3.2: Real-world Events Ground Truth

Event	***Date***	***Time***	***Location***
CBGB Music Festival	12 Oct	10am-7pm	Broadway 51 Street
Hispanic Parade	12 Oct	12pm-5pm	5th Avenue
Columbus Day Parade	13 Oct	11am-5pm	5th Avenue
Saint Patrick's Day Parade	17 Mar	12pm-5pm	5th Avenue
Million March NYC Protest	13 Dec	2pm-5pm	Washington Square Park, 5th Avenue,Foley Square

[1] http://www.nyc.gov/html/dot/html/motorist/wkndtraf.shtml

Measurement

To demonstrate the effectiveness of the framework, we consider the detection rate of each event listed in Table 3.2. As per the signal detection theory, the threshold for the corresponding concepts is evaluated in terms of detection rate.

3.5.3 Results

We evaluate our framework by examining the detection results and social information fusion results on the four events shown above. For event detection using sensors, we look at the concept results and demonstrate the usage of proposed analytic functions as well as visualization of Cmage. In the use of social information, we compare event relevant tweets during event happening time with ordinary non-event time in terms of the tf-idf values as word's importance weight.

Event Detection based on SDT We use concept confidence, values of specific visual concept extracted from the sensor data, to represent the event signal. Signal Detection Theory is then applied to model event noise and event signal with respect to the confidence values for event detection. Note that the distribution is only valid between 0 and 1 since the confidence value is a probability value. Figure 3.8 shows the distribution of "parade" signal from camera in 5th Avenue 57 Street.

The distribution is determined by calculating the mean and standard deviation of concept results from images captured on October 13, 2014, from 2 pm to 3 pm to when a "Columbus Day Parade" event was happening. The non-parade curve depicts the concept results from same time but on different days when there are no parade events. These data are analyzed to determine an optimal threshold for a particular concept detector for a camera. Note that since different cameras usually capture different scenes, the optimal threshold

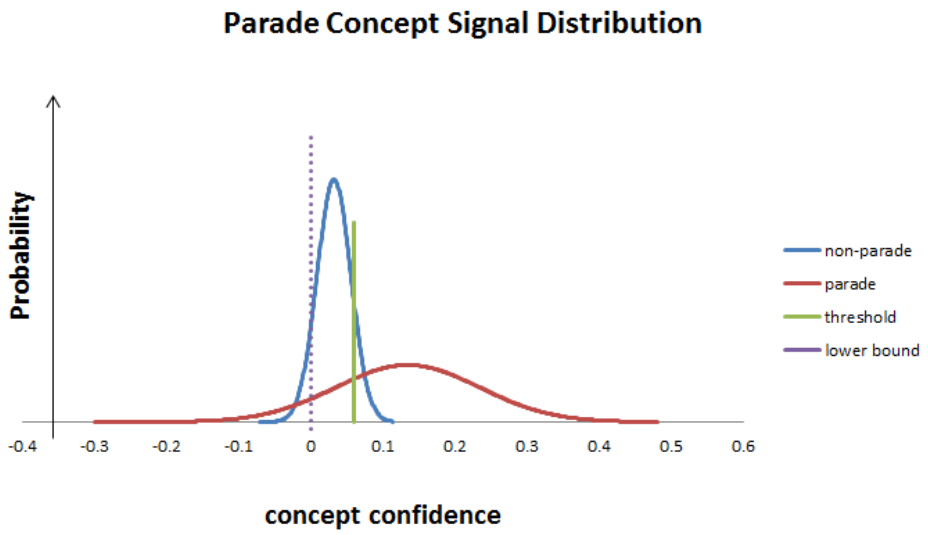


Figure 3.8: "Parade" Signal Distribution in 5th Avenue at 57 Street.

of a particular detector is not always the same. Also, for a particular camera, varying the threshold would cause different hit rate as well as false alarm, as depicted in Figure 3.9.

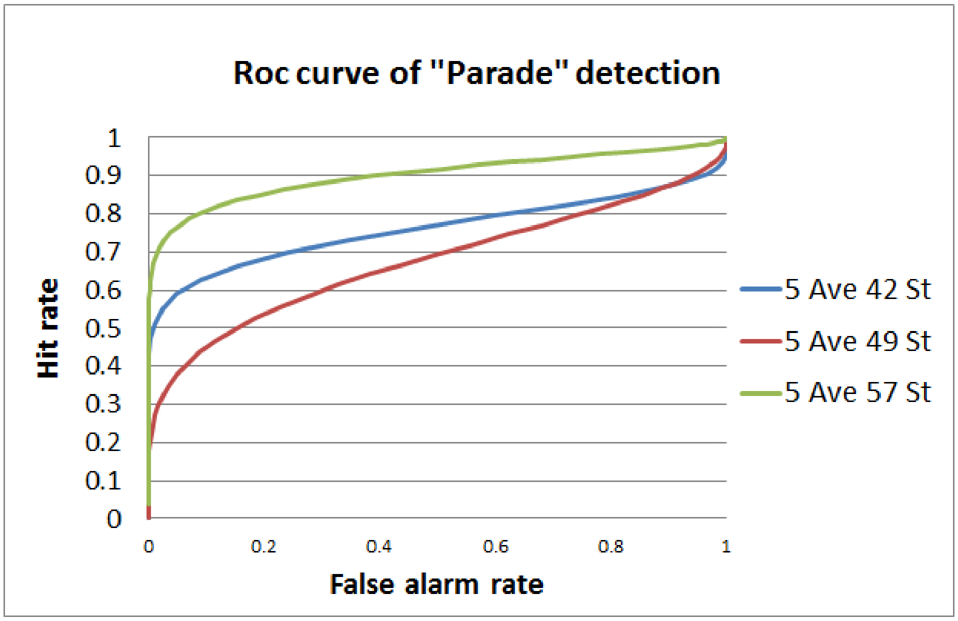


Figure 3.9: ROC Curve of "Parade" Signal in Three Locations.

This figure shows ROC curve of the parade detector for three different cameras. The threshold is chosen from the point that is closest to the left upper corner of the ROC curve, which trades off the hit rate and the false alarm rate. Therefore, the cusp point in the ROC curve i.e. the point that minimizes false alarm while maximizing hit rate is chosen as the threshold. For example, the threshold for the parade concept in Camera of 5th Avenue 57 St is computed as 0.07. Once the threshold is set, it is fixed for the specific camera to automatically trigger event alerts. The "fixed" value of the threshold is according to each concept detectors applied on the camera. However, for different cameras, since they capture different scenarios, the threshold could also be manually reconfigured accordingly. Using this threshold, we examine the detection performance of two parade events in terms of f1-score; the analysis is shown in Figure 3.10 and 3.11 for two cameras.

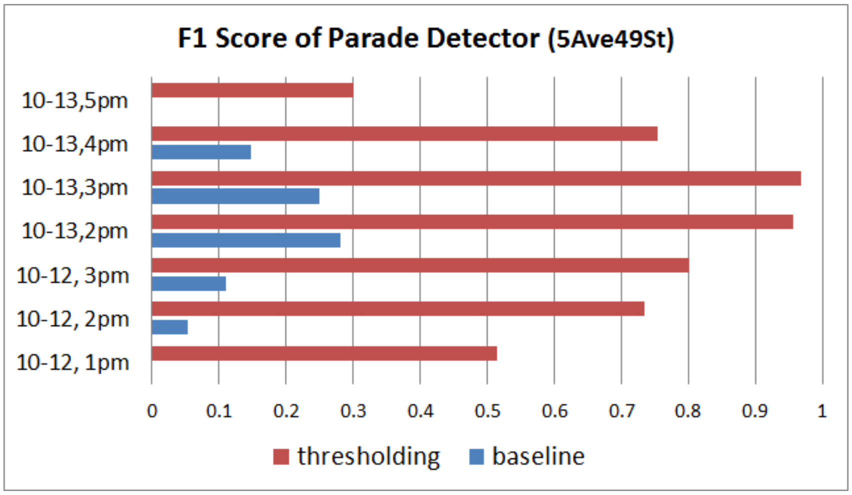


Figure 3.10: F1 Score for Camera in 5th Avenue at 49 Street

We use the threshold to analyze the results of both "Columbus Day Parade" and "Hispanic Parade" event, and compare our thresholding results with the baseline which is given by the detectors in the label field based on a fixed value 0.5 as the threshold. As can be seen, having an adaptive threshold significantly improves the performance.

Applying Analytic Functions Once concepts' confidence is obtained for

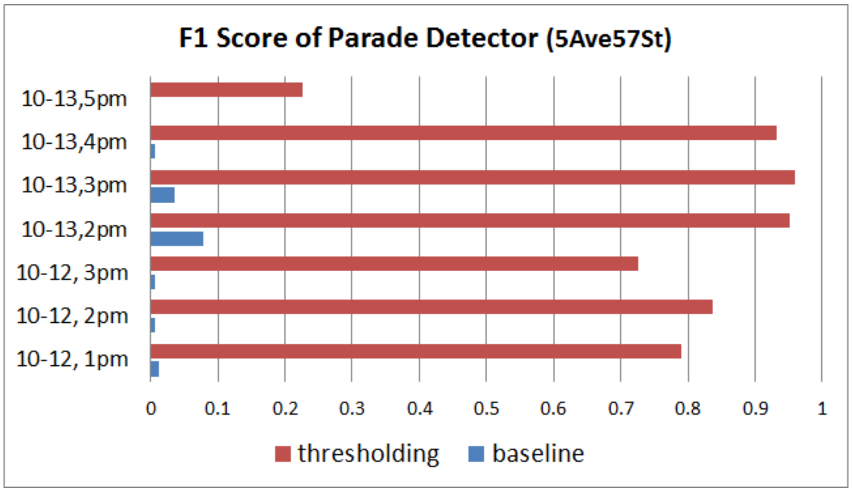


Figure 3.11: F1 Score for Camera in 5th Avenue at 57 Street

image snapshots of cameras, predefined analytic functions such as "smooth", "extreme", "trend" can be applied to obtain meaningful information such as event pattern, concept trending by interacting with Cmage in the second layer of our framework.

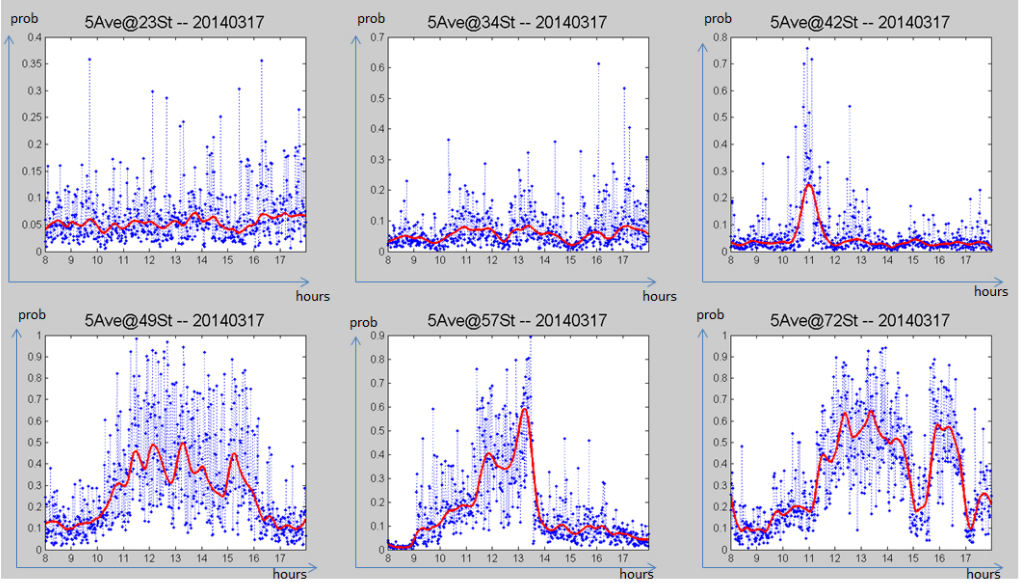


Figure 3.12: *People Marching* Concept Results from 8:00 to 18:00 in March 17th, during Saint Patrick's Day Parade Event

Figure 3.12 shows the *people marching* concept detailed results with smoothing function from 8:00 to 18:00 in March 17^{th} Saint Patrick's Day. It is shown that the peaks occur from 11:00 in cameras at 5th Avenue 42, 49, 57 and 72 streets. This demonstrates that the function performs reasonably well in providing a smooth curve for the concept and effectively reducing the impact of sensor uncertainty.

The Foodcourt videos are separated into frames and the crowd volume index for the frames is calculated every 30 seconds through background subtraction. Given a snapshot taken at time t from Camera A, the crowd analyser will return a crowd index in the range of [0, 1], a higher value representing a higher intensity of crowdedness. Therefore, we are able to convert the image to the unified probabilistic spatio-temporal data point

$(loc_{cam\_A}, t, \text{“}crowd\text{”}, probability)$,

where the location is the canteen having the camera. Therefore, the cameras would tweet crowd information twice every minute.

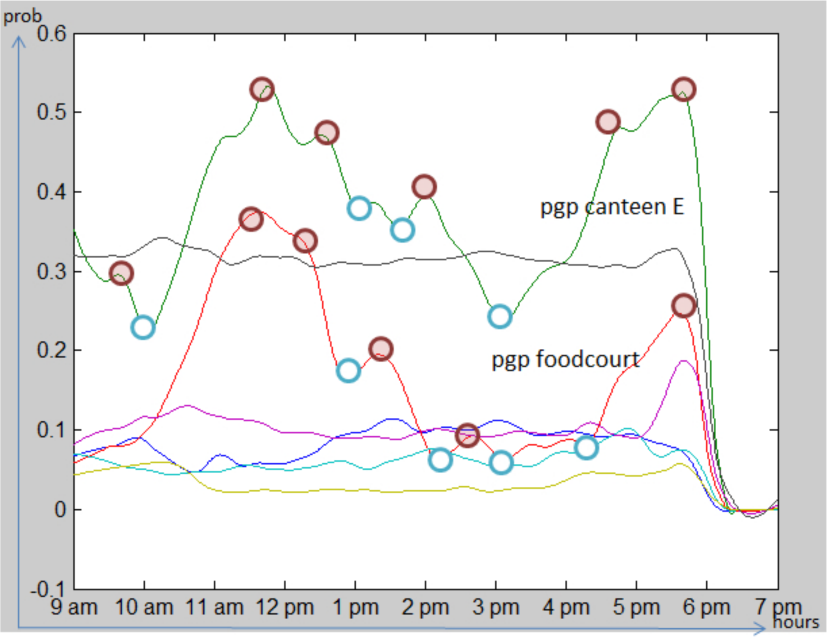


Figure 3.13: The Crowd Extremes at 7 Foodcourts on Sunday.

Figure 3.13 shows the crowd intensity on Sunday. Being applied with the *extreme* function, the two curves labelled with circles show extremes of the crowd intensity. These are foodcourts near student dorms. As can be seen, the curves match with real situation, sketching two major peaks at lunch time and dinner time and provide the information of foodcourts that remain open on Sundays so that user could have the idea of where and when to go for lunch and dinner.

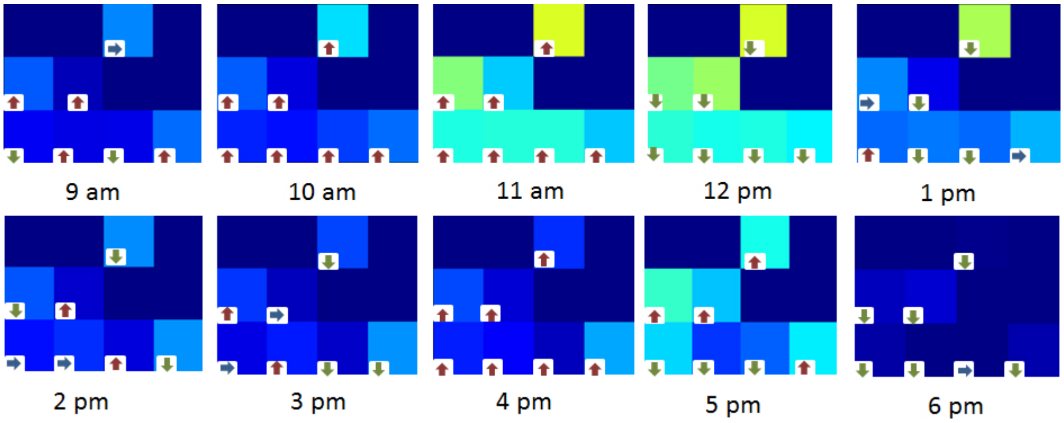
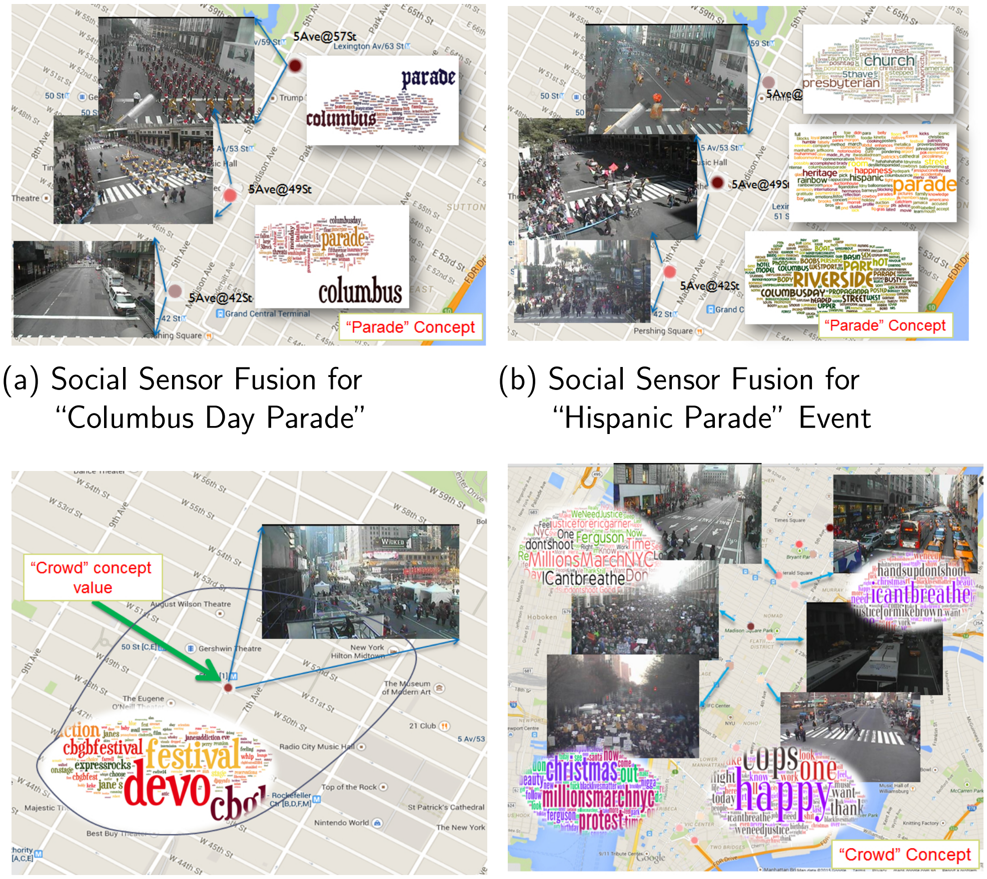


Figure 3.14: A Campus Foodcourt Cmage *Crowdedness* with Arrows indicating the Trend of Ascending, Descending and Plateau.

In addition, we show the crowd density Cmage trend (with labels) of campus canteens from 9 am to 6 pm on 21st March 2014 in Figure 3.14. The *trend* function is applied after data are smoothed with the *smooth* function and the trend is calculated by taking the gradient value of a time point. If the gradient is below a given threshold t, the Cmage pixel will be labelled "plateau" at that time. If defined by user preference, a tweeting alert could be triggered with Cmage sending a notification to an end user.

Cross Media Analysis & Social Sensor Fusion To extract the relevant semantic information from tweets in order to fuse with the camera information, we conduct term frequency analysis from social media by using tweets posted nearby the places where an event occurs. All the tweets are separated into different documents in terms of each hour and distance to a camera location.

The time span of a document could be several hours depending on the start time and end time of detected events. For example, tweets posted from 1 pm to 5 pm within a geo-circle centered in 5th Avenue 42 Street are stored as one document. Here we set the radius as 0.01 in terms of coordinates. High-frequency terms of a given location and tweets during the events are shown to the user through the framework interface. Examples are shown in Figure 3.15.



(a) Social Sensor Fusion for "Columbus Day Parade"

(b) Social Sensor Fusion for "Hispanic Parade" Event

(c) Crowd Concept and Salient Topic Words during "CBGB Musical Festival"

(d) Sensor and Social Information of "Million March NYC" Protest Event

Figure 3.15: Social Sensor Fusion for Real-world Events

We calculate the term weight for each word posted from a specific location. Words of bigger size represent higher weight. As can be seen, most tweets posted near an event location are able to give a high-level semantic meaning of the event, e.g., Figure 3.15 (a) and (b) confirm the events are indeed the “Columbus Day Parade” and “Hispanic Parade” events respectively. (c) offers details of a musical band (DEVO) that participate in CBGB musical festival, and in (d) “Million March NYC” protest event is captured by both tweets and camera feeds in terms of “crowd” concept. Video demos of “Million March NYC” protest event could be found in Youtube links [2] [3] [4]

Table 3.3 shows the comparison between our approach to social information mining with baseline in terms of a number of tweets. As represented, our framework utilizes sensor information (where and when an event is detected) to significantly (from 10^5 to 10^3) reduce the noise in the tweets. The baseline TH and TC are the numbers of geo-tagged tweets crawled before and during the event respectively. Our approach TH and TC are the numbers of geo-tagged tweets crawled *around the event location* before and during the event respectively.

Table 3.3: Comparison based on Number of Tweets Analyzed

Event	Base line #tweets		our approach	
	—TH—	***—TC—***	***—TH—***	***—TC—***
Hispanic Parade	86,714	24,357	1,496	225
Columbus Day Parade	111,071	21,891	1,573	335
CBGB Musical Festival	86,714	24,357	1,321	369

As shown in Figure 3.16, according to the equation defined in section 4.5 the final parade event signal intensity is improved when compared to using each individual sensor. Here the $I_{se}(\text{“}parade\text{”})$ here is the average confidence value in time span tweeted by the cameras. Though here only the “parade” is detected from both, this equation could also be generally applied when concepts discussed in two type of sensors are similar or correlated (e.g., a “crowd” concept detected could be related to a musical festival CBGB), which

[2] https://www.youtube.com/watch?v=UdCXzDuJtec
[3] https://www.youtube.com/watch?v=USjC263XSvE
[4] https://www.youtube.com/watch?v=AcBvfh8wKCU

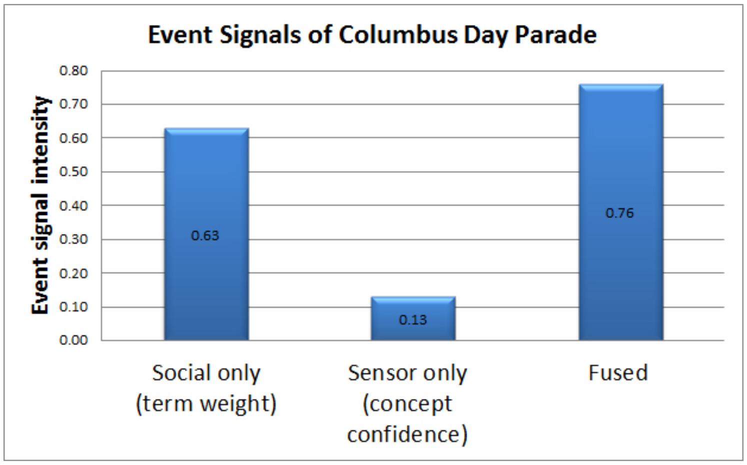


Figure 3.16: Comparison among Fused, Physical and Social Sensor Results of Detecting the "Columbus Day Parade" Event

would be our future exploration.

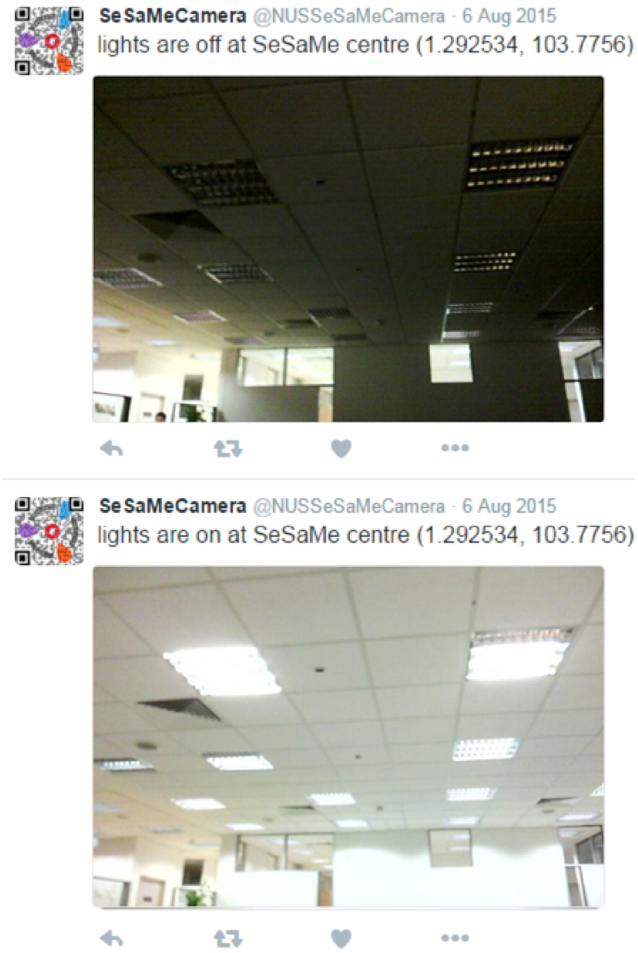
To conclude, by automatically computing the concept detection threshold, we are able to improve the event detection rate and offer a user the ability to analyze event patterns. Based on the spatial and temporal information provided by the sensors, we can leverage on the social information to give more detailed information of events.

3.5.4 Tweeting Camera Paradigm

Real Scenarios

In the following, we describe four application scenarios we have implemented so far. Figure 3.17 shows a screen shot of our camera's Twitter account "@NUSSeSaMeCamera".

Lighting status: We implement a function for checking if the lights in our office are on or off. By facing the camera in different angles, we collected positive and negative training samples with the lights turned on and off. We extract HSV color features from the images and calculate the probability of "lights on" by measuring the distances between a new sample to the center of two categories ("lights on/off") using the L2 norm. The probability calculation



SeSaMeCamera @NUSSeSaMeCamera · 6 Aug 2015
lights are off at SeSaMe centre (1.292534, 103.7756)

SeSaMeCamera @NUSSeSaMeCamera · 6 Aug 2015
lights are on at SeSaMe centre (1.292534, 103.7756)

Figure 3.17: Camera Tweeting about Lighting Condition of An Office

formula is:

$$P\_on = \frac{|s_n - s_{coff}|}{|s_n - s_{coff}| + |s_n - s_{con}|} \tag{3.6}$$

To avoid sensor noise and make detection reliable, we average over the last ten measurements. When the state of the lights changes, we send a tweet with the last recorded image.

Meeting event: In our lab, if the meeting room is taken up for a discussion, the door is usually closed and the lights are on. When the door is left open and lights are out then there is no meeting. Again, we collected the training data accordingly. We extract HoG features from the training set and use an SVM to train the concept detector. Since a meeting event is typically longer running, we can compensate occasional misclassification by only sending a new tweet if the last n (e.g., $n = 5$) results are identical. Due to the complexity of events and the involved processing steps, images are taken and analyzed every 10 seconds. Such a sampling rate is valid for events such as meetings, presentations or home monitoring where the scenarios do not change significantly in a short time.

Face & people detection: We adopt existing face detectors and people detectors provided by OpenCV [21]. Similar to meeting detection, when the camera detects a new face or a person (or the disappearance of a formerly detected face or person) over a period of n measurements, it sends a corresponding tweet.

Event Learning - New Face and Group Meeting: In face recognition scenario, a captured image is first sent to the Event Handler for face recognition as well as event classification. We use OpenBR to return a response indicating a new face has been detected; the Logic Processing component will then determine if this information should be tweeted. A tweet is only generated once the status of physical world changed (e.g., from non-face to face appearing, or from event category c_1 to category c_2). The owner of the camera can reply to a tweet telling the camera who the person on the embedded image is. The Data Communication component is constantly receiving new replies and passes them to the Event Handler to update the face database. When the same face is captured by the camera again, it will be recognized and the camera will tweet about a known face (with the name of this person) being detected. For a group meeting event learning, the camera uses IBM Bluemix cloud platform to learn the concept from newly captured images and is able to tweet if there is a meeting going on or not. A video illustrating face recognition as well as meeting event recognition (i.e., meeting starts, meeting finishes, room being occupied) can be found in our project website [5].

Performance Evaluation

We measure the tweeting camera's performance by examining the memory usage and the speed of processing in terms of seconds per frame. The concept detectors are implemented in C++; the data processing pipeline is written in Python. The resolution of each captured image is 640x480. The performance for different tasks is shown in Figure 3.18. The memory usage is the same for either individual detection task or a combination of the four tasks, which costs 14 MB. The memory usage for data processing is less than for concept detection. In terms of processing speed, we can see that the light and meeting detection requires minimal runtime while the face and people detection naturally demands more computation due to the complexity of learning model.

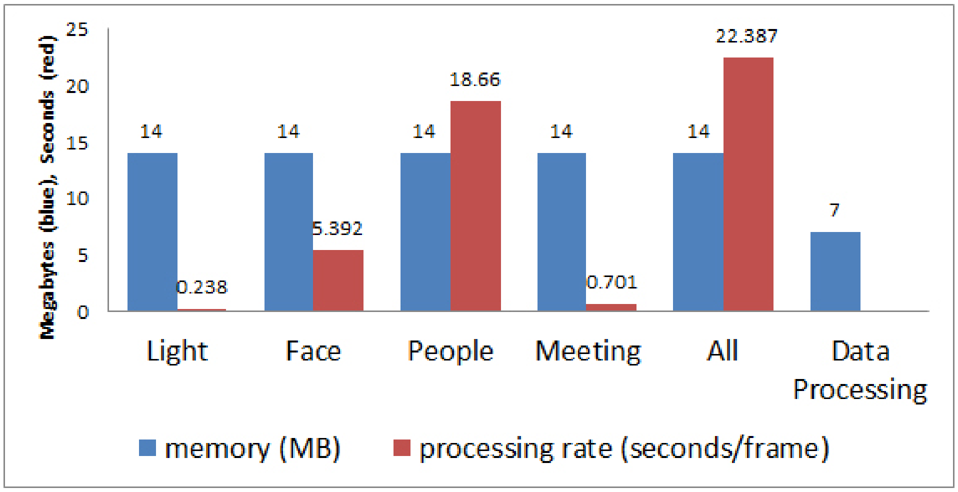


Figure 3.18: Memory Usage and Processing Frame Rate

In addition, the size of each captured image is 120KB and Figure 3.19 shows how processing time decreases with different resolutions for each individual concept and the combination of all. Apparently, the data size reduces drastically from kilobytes to key-value pairs for each concept, which is on-

[5] https://sites.google.com/site/tweetingcamera/

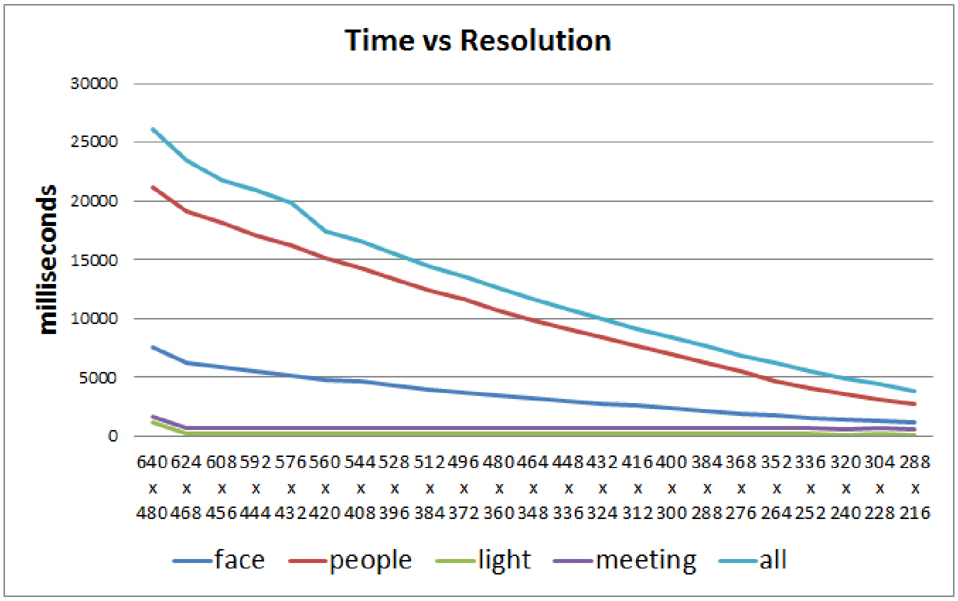


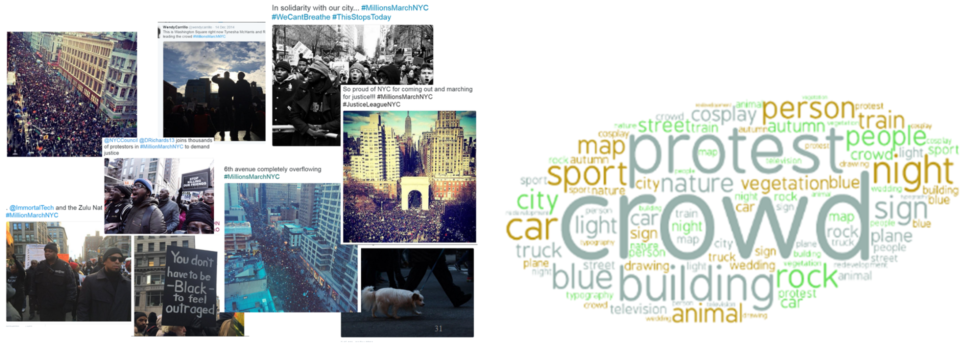
Figure 3.19: Running Time versus Captured Image Resolution

ly a few bytes, yet provides high-level information via semantic compression. This meets our goal of greatly reducing the huge amount of data to a human readable level of information and protecting the privacy of captured identities.

3.5.5 Discussion

Our experiments have verified the significant correspondence between physical sensors data and virtual social information. Any improvement in the quality and scope of concept detectors will immediately benefit the proposed method. In the second layer, more filtering options, and analytic operators to query various types of event can be provided. The unified data representation for visual sensors and analytic operators allows users to flexibly specify events characteristic for detection. One of the areas of improvement is the efficiency in performance. Concept detection requires considerable computation power and time to generate the camera tweets, which may not be able to keep up with the frame rate of camera feeds. Subsampling of the feeds could be one of the solutions. However, dropping of frames might lead to loss of critical information for events of short duration. Thus, improving the quality of camera tweets and generating them real-time is an open problem and such trade-off is also discussed in [58].

Additionally, it would be interesting to see if anything useful can be extracted from Twitter images. Given this, figure 3.20(a) shows the sample images posted in Twitter during "MillionMarchNYC" protest event and 3.20(b) shows the concepts extracts from these images, bigger words meaning they are posted more frequently during the event. As can be seen, the Twitter images could also provide events-related information for the situation.


In solidarity with our city... #MillionsMarchNYC #WeCantBreathe #ThisStopsToday

WendyCarrillo @wendycarrillo · 14 Dec 2014
This is Washington Square right now Tynesha McHarris and R leading the crowd #MillionsMarchNYC

So proud of NYC for coming out and marching for justice!!! #MillionsMarchNYC #JusticeLeagueNYC

@NYCCouncil @DRichards13 joins thousands of protestors in #MillionMarchNYC to demand justice

6th avenue completely overflowing #MillionsMarchNYC

. @ImmortalTech and the Zulu Nat #MillionMarchNYC

(a)

(b)

Figure 3.20: Images Posted in Twitter during MillionsMarchNYC Event

Letting a camera output high-level information as a result of processing raw image data on the camera itself reduces the volume of transmitted data drastically. The accuracy of the customized classifiers trained for a specific setting is generally very high. This flexibility and accuracy, however, comes with a price. Commonly available cameras are typically very easy to use since they simply output the raw camera feed. The performance of classifiers currently heavily depends on third party online cloud services such as Bluemix and could be rather poorly depending on the actual environment. Graphical tools for formulating queries of relational data are already common, and services that allow training classifiers online also exist.[6]

[6] https://www.metamind.io/vision/train

3.6 Summary

In this chapter, we propose a novel *multi-layer tweeting cameras framework*, which uses visual sensors to tweet semantic concepts for event detection. We define a unified Probabilistic Spatio-Temporal (PST) data structure to integrate the low-level concepts from the network of visual sensors. A number of filtering operators and analytic operators are also defined for the user to apply on such PST data so as to derive mid-level concepts that are suitable for higher level data visualization. We also discussed how information from the physical sensors and social media sensors can be fused to infer high-level semantics. Experiments on three real-world datasets have confirmed the effectiveness of our proposed framework. More information is available on our project website[7] including code, data, and results.

Continuously streaming sensor data to a central server or the Cloud is the state-of-the-art approach for storing, processing and sharing sensor data. Although this is suitable for simple scalar values such as temperature, humidity, voltage, etc., it poses significant challenges for streaming visual data. Many monitoring use cases, however, do not require continuous updates but rely on the detection of events of interest. With proposed tweeting camera platform, we perform a semantic compression to address both challenges. The core idea is to equip each camera with additional logic that maps the raw data into low-volume, high-level information. We extend the concept of tweeting objects and let cameras tweet detected events. Such new tweeting paradigm utilizes human knowledge to make cameras constantly update its learning models and become smarter in detecting and recognizing events. We believe that fusing tweets from both physical sensors and social sensors (i.e., humans) facilitate an exciting ecosystem for monitoring and observing our surroundings.

[7] https://sites.google.com/site/fredyuhuiwang/home

Chapter 4

Cmage Based Hybrid Fusion of Physical and Social Sensors

4.1 Overview

Event signals from physical or social sensors usually reflect various spatio-temporal and semantic patterns [87]. Fusing and exploring these multimodal sensor streams, which capture different perspectives of an event, could, therefore, result in locating, semantic interpretation of various events with higher accuracy. However, due to the heterogeneous data representations, different spatio-temporal densities, and the inherent noise in each modality, the fusion of such multimodal data remains a challenge. Thus there is a need for not only a unified data representation format but also a sophisticated framework that can combine and analyze such multimodal data for better event detection.

Identifying space and time as the unifying axes for multimodal event data, we extend the Cmage concept introduced in Chapter 3, and propose to use this image-like representation for designing a generic hybrid fusion framework for heterogeneous event signals. Such representation provides a generic way to model heterogeneous spatio-temporal data and also allows for the use of a rich repository of image processing algorithms (e.g., convolution) to easily derive semantically useful event information from such data. From a human user perspective, image, as an artifice that depicts or records visual perception, is also an intuitive way to visualize and understand different phenomena. In particular, we generate sensor Cmage and social Cmage as building blocks for hybrid fusion, from physical sensors and social sensors respectively.

A sensor Cmage is generated by aggregating information coming from multiple physical sensors based on their spatial distribution. A sensor decision is considered the confidence of specific visual concept extracted from the sensor. For example, a crowded being detected from an image captured by a camera with some confidence x (valued between 0 and 1); or a semantic word "protest" detected from the social stream as an occurring event with high frequency posted in an area. They represent a sensor confirming some event's occurring. Higher sensor decision values mean higher confidence of detecting such event. The geo-locations of the sensors define the corresponding pixel's positions in the image and the intensity of each pixel is computed by extracting higher information from the sensor readings (e.g., using concept detectors [50]). For example, Figure 4.1 shows how a "crowd" image is generated from Manhattan CCTV cameras readings. The left shows distributed physical sensors (CCTV Cameras) detecting particular concept ("Crowd") with different probabilities. The middle shows corresponding conversion into *sensor Cmage* with sparse pixels. The right part simulates dense sensor readings by applying Gaussian Process model to predict missing pixels in a given region. Similarly, a social Cmage is generated by aggregating geo-tagged social information (e.g., tweets) in an area and the pixel intensity denotes the popularity of particular social terms, such as trending hashtags in that area.

As can be seen, though such a (pseudo) image representation allows for intuitive visualization of the situation, due to the intrinsic properties of the sensor and social information, event Cmages usually contain noise to differ-

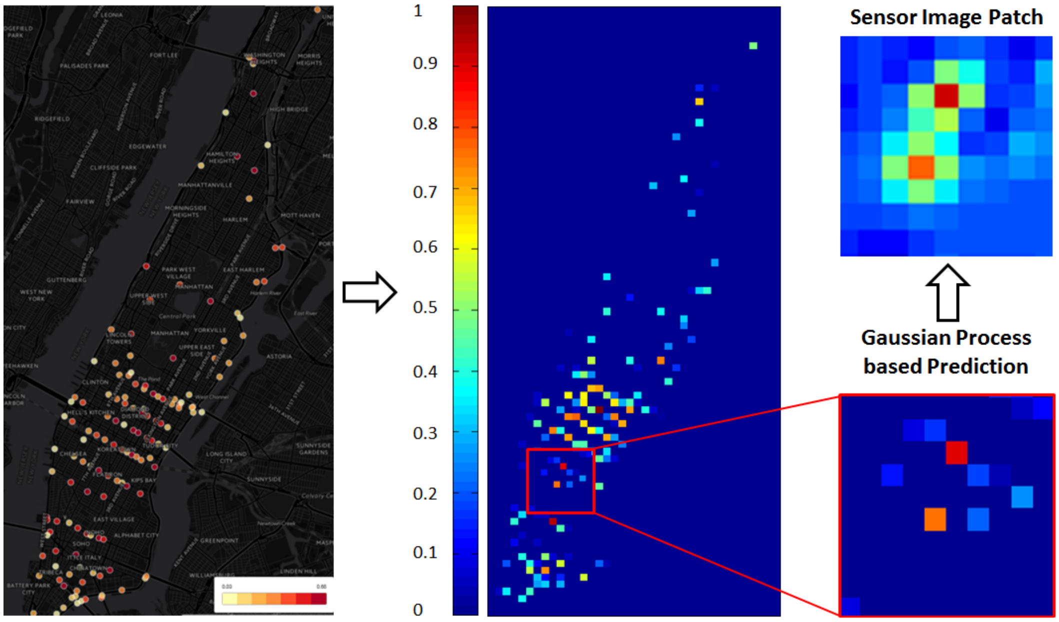


Figure 4.1: Generating Sensor Cmage from Sensors Map.

ent extents. For example, the location in a social Cmage may be incorrect if people discuss an event at locations other than the event's origin. In addition, the distributed physical sensors have different sparsity compared with the social feeds, which makes event information unavailable at some locations (pixels). Such noise and sparsity properties make event locating and situation understanding a hard problem even when dealing with multiple channels of information. To tackle this problem, we design a hybrid fusion framework including decision fusion and spatial fusion to fuse sensor Cmage and social Cmage, so as to eliminate noise and meanwhile uncover semantic details of a particular situation (such as a description of the event and its most related concepts or topics). In order to achieve these, we first predict the signals for places with no sensors using a Gaussian Process model. Second, we use a Bayesian method to fuse images at pixel-level to generate event decisions by combining corresponding pixels from both sensor and social Cmage. Third, in the spatial fusion step, we also take the nearby sensors' locations to a reference sensor into account. A location's event decision is updated by its nearby sensors and social event signals according to the distance between them. The final fusion result is evaluated by the ability of the fused image to cluster potential events candidates and the accuracy of locating events. Concepts and terms of corresponding sensor Cmage and social Cmages are then used to uncover semantic details for the event clusters.

The rest of this chapter is organized as follows. Section 4.2 describes our hybrid fusion strategies including Bayesian fusion of event signals in terms of event decisions, and spatial fusion that considers nearby event signals. Section 4.3 demonstrates the effectiveness of our hybrid fusion strategy using CCTV camera data and Twitter tweets for three large-scale marching and protest events. Section 4.4 concludes the chapter.

4.2 Methodology

We begin by defining the transformation from events signals to event C-mages, followed by the hybrid fusion method description. The notation we use throughout this chapter is defined in Table 4.1

4.2.1 From Event Signals to Event Cmages

Sensor Event Cmage: C^{sen} can be generated from a set of M physical sensors $S^{sen} = \{SEN_1, ...SEN_M\}$ in a region bounded by upper-left corner $P_{ul} = (lat_{ul}, lon_{ul})$ and down-right corner $P_{dr} = (lat_{dr}, lon_{dr})$ in terms of geo-coordinates in the physical world. Each sensor $SEN_m = G_m \times R_m$ is composed of its geo-location $G_m = (lat_m, lon_m)$ and its environment reading R_m (e.g., image captured by a camera, humidity value measured by a weather sensor, and so on). A region is then separated into grids, using a user-defined grid size r_{sen}, to form the sensor event Cmage $C^{sen} = [e^{sen}_{ij}]_{H\times W}$, where $H = (lat_{ul} - lat_{dr})/r_{sen}$ and $W = (lon_{ul} - lon_{dr})/r_{sen}$ and sensor pixel $e_{ij} = \mathbb{F}(SEN_m)$; $\mathbb{F}$ is

Table 4.1: Cmage Fusion Notation

Symbol	***Description***
M	number of locations (physical sensors)
G_m	geo-location of sensor m
C^{sen}	sensor event Cmage
C^{soc}	social event Cmage
r_{sen}	user defined grid size
R_m	reading of physical sensor m
$\mathbb{F}$	function transforming sensor information into numeric values
$\mathbb{LAT}$	mapping from geo-location to image x coordinate
$\mathbb{LON}$	mapping from geo-location to image y coordinate
$term_c$	posted word
$POST_m$	the m^{th} tweet
SOC_i	social observations in location i
$\boldsymbol{p}_i$	the position of pixel i
e_i	the intensity of pixel i
$\mathbf{e}$	observed pixel values
$\hat{\mathbf{e}}$	predicted pixel values
$\mathcal{P}$	observed feature matrix
$\mathcal{P}_\star$	predicted feature matrix
σ_n^2	noise variance
$K(\cdot,\cdot)$	covariance matrix of pixels positions
F_n	overall confidence values of n sensor streams
f_{ij}	fused event confidence

the function that transforms the sensor readings into numeric values, such as a concept detector [50] for an image, a direct copy of air quality index [43], etc., representing the strength of event signal with particular semantic meanings. The mapping from sensor SEN_m location to corresponding image coordinate is defined by $i = \mathbb{LAT}(G_m) = |lat_{ul} - lat_m|/r_{sen}$, $j = \mathbb{LON}(G_m) = |lon_{ul} - lon_m|/r_{sen}$.

Social Event Cmage: C^{soc} is generated from a set of social observations $S^{soc} = \{SOC_1, ...SOC_M\}$ in the same region as physical sensors, where each observation $SOC_m = G_m \times POST_m$ contains its corresponding geo-location and the content $POST_m$ (e.g., the tweet text). We define such posted content $POST = \{term_1, ..., term_c\}$ as a set of terms (or words). Different from pixels of sensor Cmage where the value of each pixel is derived directly from corresponding one physical sensor, the pixel values in social Cmage related to nearby social observations. We propose using two methods to represent the "social pixel" [105]: (1) density based signals; and (2) term frequency based signals. The density based signal method considers the density of nearby posts that contains the particular term. Specifically, given $C^{soc} = [e^{soc}_{ij}]_{H \times W}$ of $term_x$ and a radius r, for social observation SOC_m, $e_{ij} = \mathbb{F}(term_x, SOC_m, r)$; where $\mathbb{F} = \sum_{k=1, dist(G_k, G_m)<r}^{|S^{soc}|} \mathcal{H}(POST_k, term_x)$ is the number of surrounding social observations whose post $POST_k$ contains the term $term_x$, indicated by $\mathcal{H}(POST_k, term_x)$ and $dist(G_k, G_m)$ is the Euclidean distance of two social observation locations. The mapping from social SOC_m location to corresponding pixel coordinate is defined similar to that in sensors Cmage, but with grid size r_{soc}. For term frequency based method, the pixel value e_{ij} is calculated as TF-IDF values defined in Chapter 3, which generate each term's weight by considering history posts in the same location. Terms of higher weight mean that they are frequently discussed currently but seldom discussed in the past days, which we consider as a good indication of occurring events.

4.2.2 Hybrid Event Cmage Fusion

Sensor pixel value estimation using noisy and sparse observations

Due to the intrinsic characteristic and sparse spatial distribution of sensors, a sensor Cmage will be generated with many empty and noisy pixels, which causes the problem for the later fusion with social Cmages. In particular, fusing social pixel with a false empty pixel (e.g., the one between the two red pixels in the bottom-right magnified patch of Fig. 4.1) will result in an empty pixel reflecting there is no event, which is not the truth. To solve this problem, we assume the sensor readings over an urban area to be realized from a Bayesian non-parametric model, Gaussian Process (GP) [91], which incorporates noise model and allows the spatial correlation of sensor readings (sensor pixels) to be formally characterized in terms of their locations in the image. This property enables predicting empty pixels using observed sensor readings. Specifically, assuming a sensor Cmage $C^{sen}_{H\times W}$ is defined as $\{(\boldsymbol{p}_i, e_i)|i = 1, ..., H \times W\}$, where $\boldsymbol{p}_i$ and e_i are the pixel's position ($\boldsymbol{p}_i = [lat_i, lon_i]$) and intensity respectively, we model the joint distribution of Q observed pixel values $\mathbf{e} = [e_1, ..., e_Q]^\top$ and predicted pixel values $\hat{\mathbf{e}} = [e_{Q+1}, ..., e_{W\times H}]^\top$ at the test locations under the prior as: $e_i \in [0, 1]$

$$\begin{bmatrix} \mathbf{e} \\ \hat{\mathbf{e}} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathcal{P},\mathcal{P}) + \sigma^2 I & K(\mathcal{P},\mathcal{P}_\star) \\ K(\mathcal{P}_\star,\mathcal{P}) & K(\mathcal{P}_\star,\mathcal{P}_\star) + \sigma^2 I \end{bmatrix}\right) \tag{4.1}$$

where $\mathcal{P} = [\boldsymbol{p}_1, ..., \boldsymbol{p}_Q]$, and $\mathcal{P}_\star = [\boldsymbol{p}_{Q+1}, ..., \boldsymbol{p}_{H\times W}]$ are the observed and predicted feature matrix respectively; σ_n^2 is the noise variance; and the elements in covariance matrix $K(\cdot, \cdot)$ reflecting correlation between two pixels positions $\boldsymbol{p}_m$ and $\boldsymbol{p}_n$ are defined by the covariance function:

$$k(\boldsymbol{p}_m, \boldsymbol{p}_n) = \sigma_s^2 exp(-\frac{1}{2}\sum_{d=1}^{2}(\frac{\boldsymbol{p}_{m,d} - \boldsymbol{p}_{n,d}}{l_d})^2) \tag{4.2}$$

where $\boldsymbol{p}_{m,d}(\boldsymbol{p}_{n,d})$ is the d-th component of 2D (lat and lon) vector $\boldsymbol{p}_m$ and $\boldsymbol{p}_n$, and the hyperparameters σ_s^2, l_1, l_2 are signal variance, and length-scales respectively that can be learned using maximum likelihood estimation. Note that term $\boldsymbol{p}_{m,d} - \boldsymbol{p}_{n,d}$ measures the geographic distance of two locations in terms of latitude or longitude.

Having this covariance matrix, values of predictive pixels can be defined by the Gaussian Process regression equations:

$$\hat{\mathbf{e}} = K_\star^\top (K + \sigma^2 \boldsymbol{I})^{-1} \mathbf{e} \tag{4.3}$$

$$e_i = f(\boldsymbol{p}_i) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.4}$$

where $K_\star$ is the $Q \times (W \times H - Q)$ covariances matrix between predicted pixels and observed pixels, and $K = K(\mathcal{P}, \mathcal{P})$ is the $Q \times Q$ covariance matrix of observed pixels; $\mathbf{e}$ is $Q \times 1$ observation vector.

Event decision fusion We undertake a pixel-by-pixel fusion between the sensor and social Cmages. Specifically, we adopt a Bayesian approach based confidence fusion based on [11], where a general fusion of multiple streams in terms of confidence values is computed according to the following formula:

$$F_n = \frac{F_{n-1} \cdot f_n}{F_{n-1} \cdot f_n + (1 - F_{n-1})(1 - f_n)} \tag{4.5}$$

where F_n is the overall confidence values considering n sensor streams and f_n is the confidence value of n^{th} sensor stream. In a two stream case, each pixel e_{ij}^{soc} (e_{ij}^{sen}) represents event decision's confidence from corresponding locations and the fused confidence f_{ij} is computed as:

$$f_{ij} = f(e_{ij}^{soc}, e_{ij}^{sen}) = \frac{e_{ij}^{soc} \cdot e_{ij}^{sen}}{e_{ij}^{soc} \cdot e_{ij}^{sen} + (1 - e_{ij}^{soc})(1 - e_{ij}^{sen})} \tag{4.6}$$

Using this fusion method, fusing two pixels of high confidence will result in a higher confidence. Fusing high confidence with low confidence pixels (which means conflicting observation) will result in a value close to the lower one and two low confidence pixels will result in a much lower value than either one of them.

Spatial fusion

For each social observation or sensor reading, spatial fusion considers its surrounding signals which could contribute to the considered event signal. The range to consider is defined by a fixed reference window $W^{w\_size \times w\_size}$. Windows size is set to be flexible so that users can specify the size of area based on the specific type of the events. Given a reference signal, each signal within the window will be assigned a weight based on their distance to the referenced signal. This is valid since a geographically closer signal has a higher influence. Such a spatial fusion model is then defined by the fusion function F as follows:

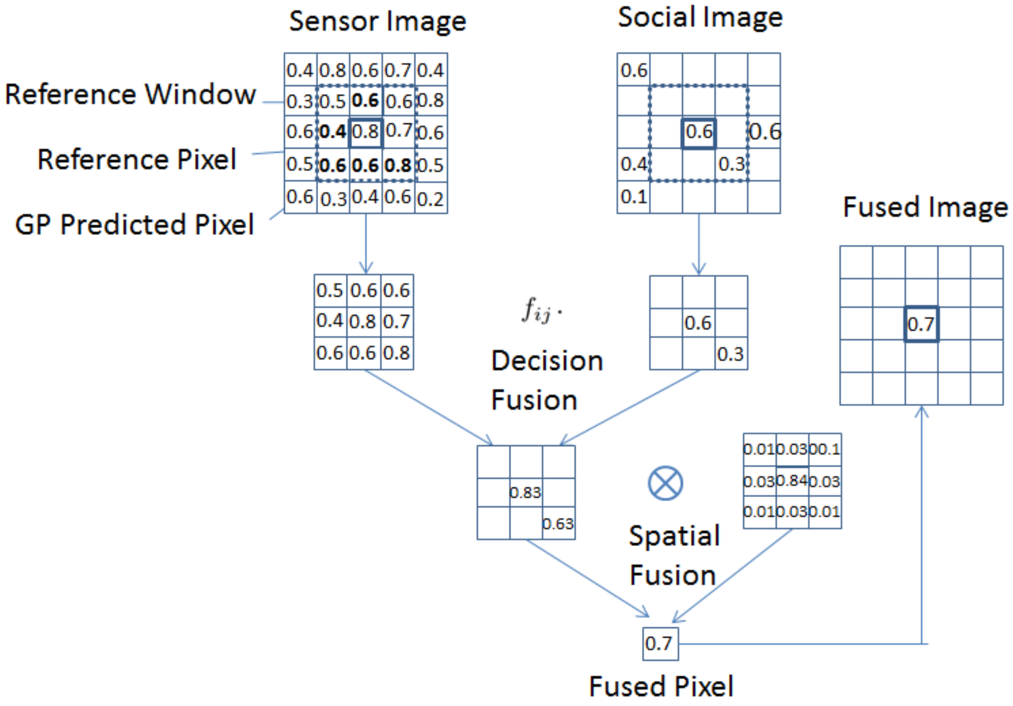


Figure 4.2: Illustration of Decision and Spatial Fusion

$$F(C^{soc}, C^{sen}, w\_size) = \{e_{ij}^{fus}\} \tag{4.7}$$

$$e_{ij}^{fus} = \sum_{xy \in W^{ij}} w_{xy} \cdot f_{ij} \tag{4.8}$$

where $|x-i| \leq \frac{w\_size-1}{2}, |y-j| \leq \frac{w\_size-1}{2}$

$$w_{xy} = \alpha \cdot e^{-\sqrt{(x-i)^2+(y-i)^2}} \tag{4.9}$$

w_{xy} is the weight given to the neighbouring pixels of referenced pixel e_{ij}^{fus} based on their distances to it. An illustration of integrating decision and spatial fusion is shown in Figure 4.2 with reference window size set to 3×3; the sensor Cmage is preprocessed with Gaussian process resulting in a "Sensor Cmage Patch" similar to the one shown in Figure 4.1. Note that Gaussian Process is not applied to social Cmage, because unlike physical sensors readings that vary smoothly across a region, human posts can appear anywhere much more spontaneously, making social signals too noisy to be modelled with a smoothing kernel (e.g., Eq. 4.2) as in GP.

4.3 Experiments and Discussion

We tested our fusion approach on the same dataset used in our first work. It contains continuous image snapshots from 149 CCTV traffic cameras across Manhattan, New York City, and geo-tagged tweets. In this short paper, we demonstrate the efficacy of the proposed approach based on three popular events ("ColumbusDayParade", "MillionMarchNYC" and "StPatricksDayParade") with large spatio-temporal coverage that is examined in work [115].

4.3.1 Evaluation Metrics

Saliency Metric: Events shown in the image should appear "natural" and "sharp" to a human interpreter [88]. To this point, the fused images are supposed to preserve the salient information and enhance the contrast for visualization. In order to objectively evaluate our hybrid fusion algorithm, we would need a "saliency metric" measure describing how events signals are concentrated in a small dense region. This "saliency metric" is obtained by averaging the spatial distance of the points belonging to the same cluster with respect to the centroid of the cluster for each cluster. Such clusters can be obtained by mean-shift clustering [28] aiming to discover "blobs" in a smooth density of samples. Given an image I, *saliency metric* $\mathbf{S}$ is defined by:

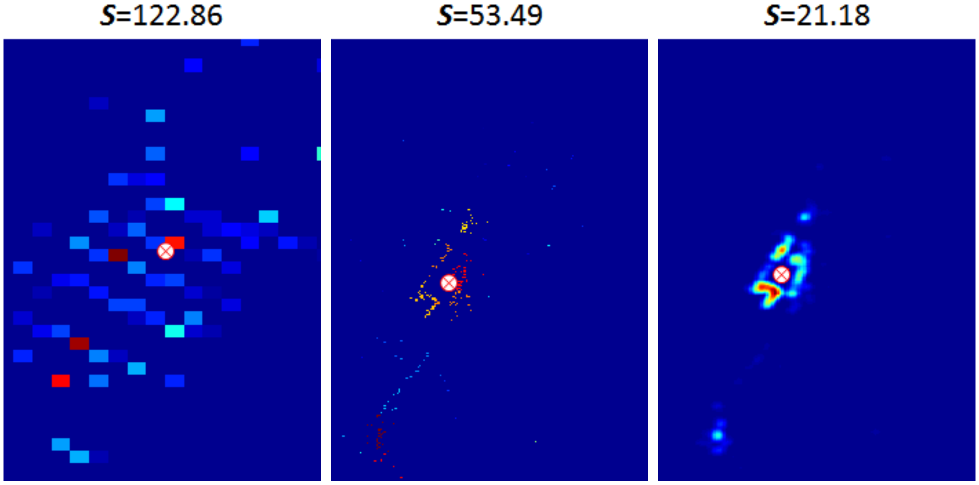
$$\mathbf{S}(I) = \sum_{i=1}^{C} \sum_{\mathbf{p}_m \in CL(c_i)} w_{im} * Dist(\mathbf{p}_m, c_i) \tag{4.10}$$

where c_i is the cluster centroid of cluster $CL(c_i)$ given by mean-shift clustering and $w_{im} = \frac{e_m}{\sum_{p_n \in CL(c_i)} e_n}$ is the normalized weight for each pixel; for each cluster $\sum w_{im} = 1$. A lower value of $\mathbf{S}$ means a more salient and concentrated region, therefore a better image for visual analytics purposes.

MSE of Ground Truth: To demonstrate the efficiency of noise removing in fusion process, we evaluate how much the fused event Cmage is matched with the manually labelled ground truth discussed in Chapter 3.

4.3.2 Noise Removal & Saliency Enhancement

Figure 4.3 shows different Cmages for the "MillionMarchNYC" protest event. Figure (a) is obtained by applying the "Marching" concept detector on the CCTV camera recordings, generating a low-resolution sensor Cmage. Figure (b) is obtained by calculating the TF-IDF weight of term "MillionMarchNYC" according to [115], resulting in a high-resolution social Cmage. The intensity of the pixels represent the signal strengths at the corresponding locations. Red crosses are the centroids of clusters given by the mean-shift algorithm. Saliency metric values are shown on top of the figures. As can be seen, Figure (c) effectively enhances the contrast and saliency of event candidates than that of Figure (a) and Figure (b), which look noisy. The fused image tells exactly where this marching event is happening. This demonstrates that the proposed Bayesian-based fusion in Section 4.2.2 can help enhance the signal if both sources contribute to the confirmation of events and meanwhile eliminates the noise based on their disagreement.



(a) Sensor Cmage (b) Social Cmage (c) Fused Results

Figure 4.3: Event Cmages of "Million March NYC" Event: (a) Low Resolution Sensor Cmage of "*Marching*" Concept ; (b) High Resolution Social Cmage of "MillionsMarchNYC"; (c) Fused Image

4.3.3 Semantic Details Mining

The effectiveness of fusion is also demonstrated by extensive experiments with different combinations of sensor concepts and social terms (sConcept-term), shown in Figure 4.4. Blue, red and green bars are the saliency metric S of the sensor, social and fused images respectively. They are ordered by value S of fused images. Rather than presenting only a loosely defined concept, such orderings help users to find the best matching semantic details of ongoing events. For example, details about the "Marching" concept is best described by the social term "blacklivesmatter", which is a popular hashtag posted during the protest. This shows that the fusion will have a good performance if two concepts have similar spatial distributions in terms of their event signal.

Conducting experiments for two more events, we generated Table 4.2 showing the average improvement in S values based on the proposed fusion method for different combinations in terms of best matches (e.g., "parade" with "blacklivesmatter", "stpatricksday", "green", "columbus" etc.). The Enhancement

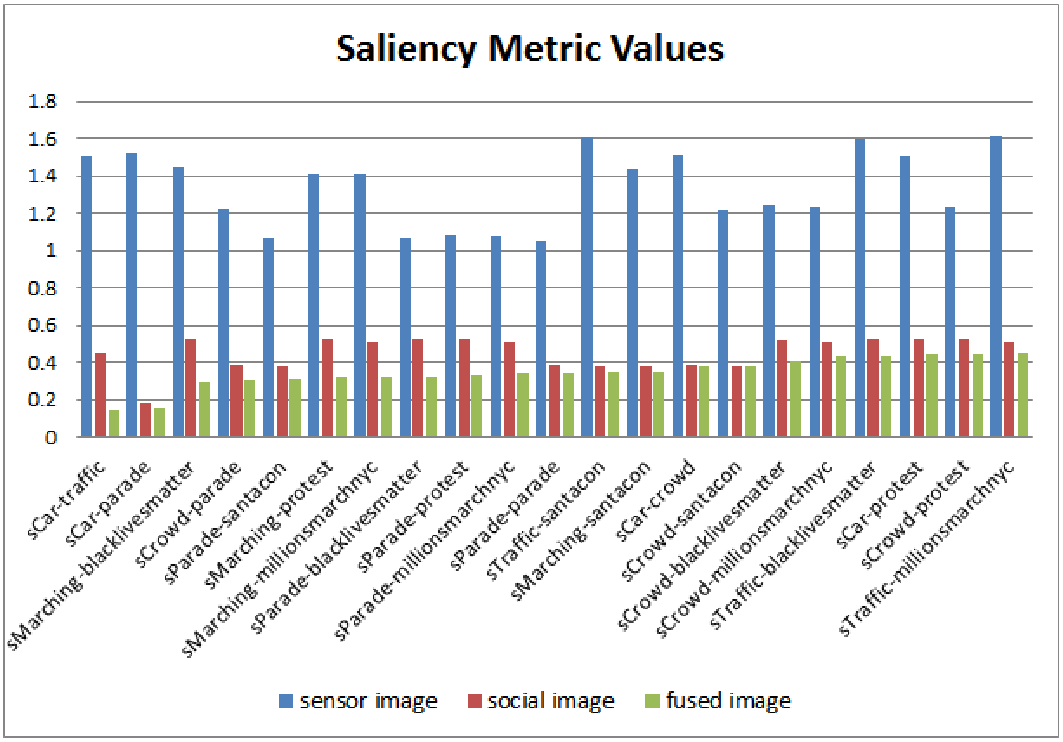


Figure 4.4: Saliency Metric Values **S** of Different Sensor or Social Event Cmages and Fused Results

Table 4.2: **S** Values for Different Events

Events	Sensor Image	Social Image	Fused	Enhancement Rate on Average
ClumbusDayParade	1.24	0.43	0.34	0.47
MillionMarchNYC	1.24	0.47	0.40	0.41
StPatricksDayParade	1.49	0.61	0.53	0.39

Rate measures how much the fused image enhances the saliency for sensor and social on average.

We compare the sensor, social and fused Cmage with ground truth, which is binary picture illustrating the location of this protest event. There are 6 locations where from the camera feeds, we are sure about the event happening and generate a ground truth Cmage accordingly. All Cmages are compared with the ground truth Cmage in terms of MSE. The result is shown in Figure 4.5.

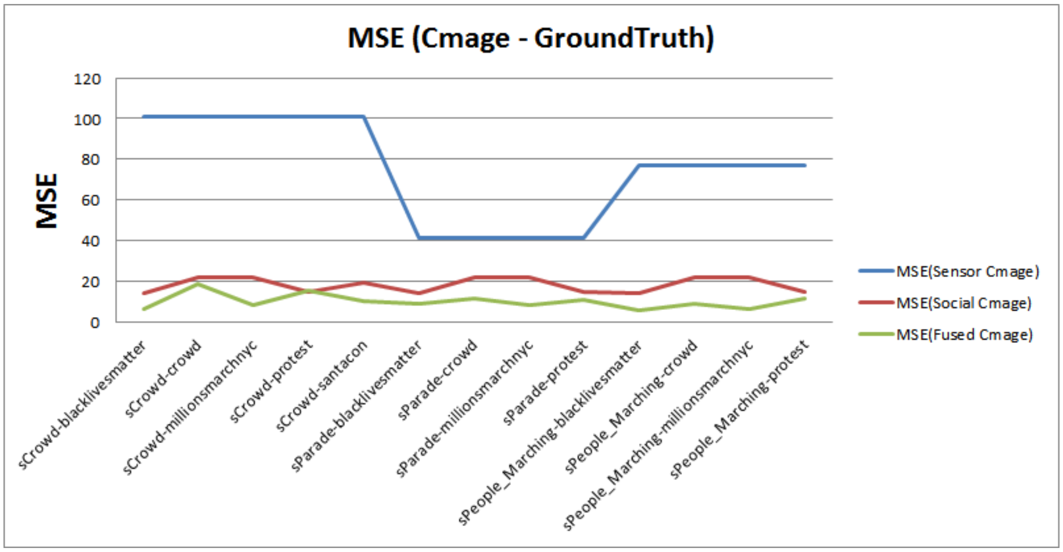


Figure 4.5: MSE of Sensor, Social and Fused Cmage Compared with Ground Truth Cmage

Since we have detected the events and mined the related semantic words of this situation, we specifically examine the Cmages of concepts that are closely related to this events, including: “crowd”, “parade”, “people marching”, “blacklivesmatters”, “millionmarchnyc” and “protest”. As can be seen from Figure 4.5, the fused Cmages have less MSE compared to non-fused Cmage, either sensor Cmage of social Cmage. This is because the Bayesian fusion utilizes the agreement the event signals from both sources and the Gaussian Process enhances the signal of event locations given their nearby signals contribute to the confirmation of occurring events.

4.3.4 Effectiveness of Gaussian Process

Sensors sparsity problem is handled by Gaussian Process with σ_s^2 set to 0.90 and l_d set to 0.89. The effectiveness of Gaussian Process for the fusion process is shown in Figure 4.6, where red line shows the S of fusion without GP and the blue line is the fusion with GP. For the best matches, the fusion will results in better performance (lower S) if Gaussian Process is incorporated in the fusion process. However, the fusion of some particular combinations performs better if no GP is applied. A plausible explanation is that the social term (e.g., “santacon”) is not semantically related to the sensor concept (describing two different events), so the prediction could not contribute to the fusion.

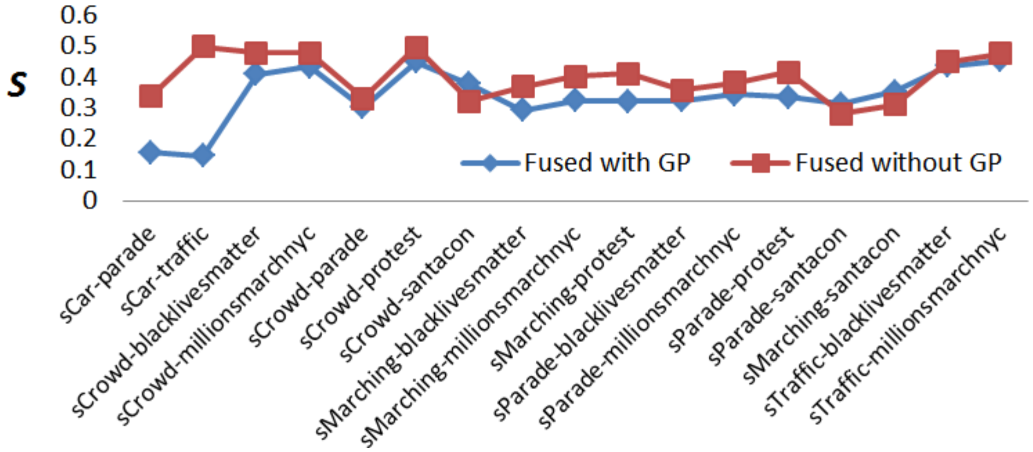


Figure 4.6: Comparison of Fusion with GP and without GP

4.4 Summary

In this chapter, we present an image-based hybrid fusion framework to fuse different modalities of the physical sensor and social data, considering both the event signal strength and their spatial relations. Image-based representations of different data streams provide not only a better visualization of situations but also the convenience of manipulation of the event-related sensor and social signals. The results demonstrate that the fusion strategy can effectively remove noise from the data streams, locate the event place and offer situational semantics details.

Chapter 5

A Matrix Factorization Based Framework for Fusion of Physical and Social Sensors

5.1 Overview

Beyond spatial information that is mainly considered in the fusion strategies discussed in Chapter 4, semantic information, which expresses high-level human knowledge, also plays an important role in interpreting holistic situations. When physical sensors offer some signals, for instance, numeric values of weather conditions or confidence values of particular concepts being detected by a camera, social sensors' rich content could usually reflect and explain why such signals are given. Therefore it is necessary to utilize such implicit semantics in the fusion of these two sources. This leads to the goal of this chapter: fusing physical sensor readings with social sensor feeds that considers spatio-temporal information and semantics simultaneously. However, different modalities, multiple sources, various spatio-temporal density and approximate sensing properties of these two types of sensors make the fusion of heterogeneous information a challenging problem.

Fortunately, a fusion of these different data representations could be inspired by the works of another emerging area: collaborative filtering [57], where the numeric ratings given by users to particular items are fused with additional contextual information such as reviews or comments, and the predicting of missing ratings can be obtained by matrix factorization techniques [99].

We propose an innovative way of combining physical sensors and social sensors' spatio-temporal data using matrix factorization. Specifically, given a time window, we first extract numeric sensor readings from physical sensors (e.g., PSI (Pollutant Standards Index) stations, CCTV Traffic cameras) and build a matrix where for each time point and location the corresponding reading is inserted. An illustration image of such situational matrix is shown in Figure 5.1. Pixels represent concepts confidence values given by physical sensors in different locations from different times. High intensity denotes high confidence. Word clouds on top are social information collected from the cor-

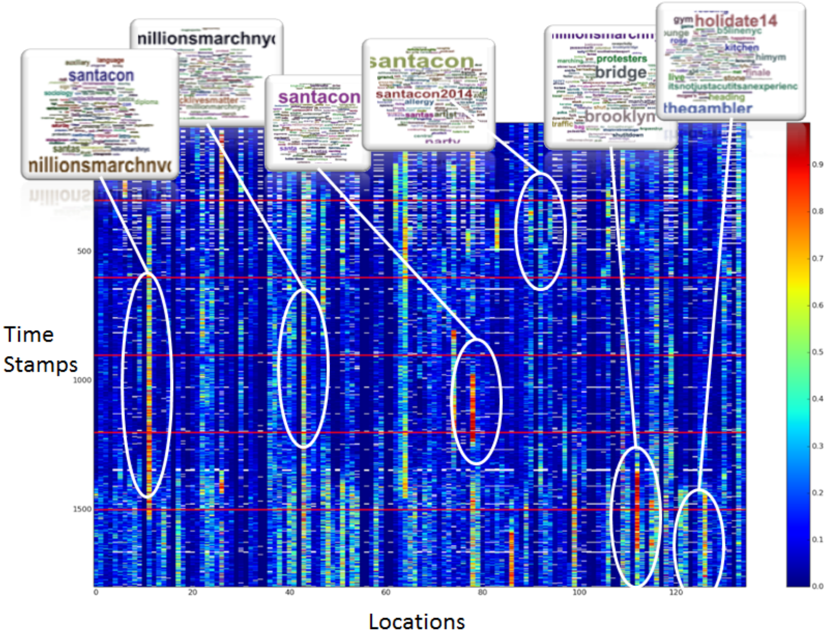


Figure 5.1: Spatial Temporal Fusion of Physical and Social sensors.

responding locations and time stamps. We conduct matrix factorization on the situation matrix for filtering out the noise of the original physical signals, while simultaneously we combine the information extracted by social networks (Twitter for the current experiment). Such topics are detected from social sensors with LDA [19]. We applied the designed framework for two situation awareness tasks:

- *Missing Readings Prediction*: given a geographic area, physical sensor distribution can fail to cover some locations. We use our framework to infer from social sensors the readings in those points. For this task, we use a dataset of PSI reading of Singapore, as some parts of the country are far from any meteorological station and people will tweet when PSI index is extremely high (signifying haze).
- *Event Noise Filtering*: from a traffic camera, it is possible to monitor large-scale outdoor events like a parade, marathon or a protest. However, due to the fixed camera field of view and the limitations of visual concept detectors, physical signals are extremely noisy. The proposed matrix factorization can be used for factoring out the noise and hence improve the results of detection of events.

Comparing with a baseline where only physical sensors are used, experiments demonstrate that incorporating social sensor information will improve the performance of both the tasks.

The rest of this chapter is organized as follows. Section 5.2 elaborates problem formalization and fusion algorithm. Section 5.3 gives a statistical evaluation of our method on real-world data as well as qualitative and quantitative analysis for situation prediction and event filtering. Section 5.4 concludes the chapter.

5.2 Methodology

In this section, we elaborate the matrix factorization (MF) based fusion framework. We first state the problem and notations, then describe matrix factorization model which could embed social information. Subsequently, we describe how MF could be used for situation prediction and for handling noise and conflicting values in the data. The notation we use throughout this chapter is defined in Table 5.1.

Table 5.1: Matrix Factorization Framework Notation

Symbol	**Description**
N	number of locations
M	number of timestamp
S^{δ}	situation matrix
S^{δ}_{ij}	sensor reading extracted at location j for time stamp i
LD^{r}_{j}	social feeds collected from location j within distance of r
S^{δ}	physical sensor stream
$\mathbb{S}^{\delta}$	social sensor stream
t_i	temporal latent vector at time i
l_j	spatial latent vector at location j
β_i	temporal bias
β_j	spatial bias
μ	situation matrix average reading
$\theta_{LD_j,k}$	probability of seeing topic k in location document LD_j
ϕ_k	word distribution for topic k
$z_{LD_j,q}$	topic assignment of each word q for a location document LD_j

5.2.1 Matrix Factorization Based Fusion Framework

Problem Statement and Notations

Physical Information-First Modality: we consider the situation as the set of observations from a collection of physical sensors. Formally, each observation includes physical sensor reading signals as well as spatio-temporal information. Suppose we have a temporal window δ, N physical sensors from N locations $j = \{1, ..., N\}$ respectively, given their observations in M time stamps $i = \{1, ..., M\}$, the situation can be defined by a *situation-matrix* $S^{\delta} \in \mathbb{R}^{M \times N}$ where S^{δ}_{ij} represents the reading extracted from the sensor situated at location j for time stamp i. We focus on situations where some readings may be missing (e.g., in some locations there are no physical sensors, or for some time one or more sensors can be faulty) and where data is highly affected from noise or conflict readings (due to the nature and collocation of sensors).

Social Information-Second Modality: a location is often associated with geo-tagged social information (e.g., Tweets, Facebook posts, or Flickr images from the location). We denote the social information as $\mathbb{S}^{\delta} = \{LD^{\delta,r}_1, ..., LD^{\delta,r}_N\}$. During δ, LD^r_j is a set containing the social feeds collected from a geographical circle with radius r centered at location j. Given a radius r, we name LD "location document" and $LD_j = \{p_1, ...p_k\}$ are the set of geo-tagged records (e.g., a Twitter tweet or a Facebook post) collected from social media, where one record $p_k = \{w_1, ..., w_r\}$ contains a set of words $w \in D$.

Problem Statement: the fusion problem can be formalized in line with the intuition that the goal is to find a combination of the two modalities that is able to produce better performance on a certain task compared to using the two sources individually.

Given physical sensor stream S^{δ}, social sensor stream $\mathbb{S}^{\delta}$ and a performance evaluation function *performance* for a specific task, the goal of the fusion problem is to find a fusion function f such that

$$performance(f(S^{\delta}, \mathbb{S}^{\delta})) > performance(S^{\delta})$$

For the rest of this chapter , δ is omitted.

5.2.2 MF on Physical Signals: Basic Model

We propose to model the fusion function f as a factorization of the matrix S. Matrix factorization is the state-of-the-art method that is widely used in recommendation systems (RS) for filling missing ratings in a user-item matrix. Taking inspiration from this approach, we utilize it to find latent relations between temporal and spatial responses on situations or occurring events by factorizing the situation matrix. Specifically, we use matrix factorization to map time and location to a K-dimension joint space, where each time point i and location j are associated with latent vectors $t_i, l_j \in \mathbb{R}^K$ respectively. Matrix factorization approach seeks to approximate the situation matrix S by a multiplication of K-rank factors $S \approx T^\top L$, where $T = [t_1, ..., t_m] \in \mathbb{R}^{K\times M}$ and $L = [l_1, ..., l_n] \in \mathbb{R}^{K\times N}$. Formally, the predicted situation signals at location l at time t is equal to the inner product of the corresponding spatial and temporal latent vectors:

$$\hat{s_{ij}} = t_i^\top l_j + \mu + \beta_i + \beta_j \tag{5.1}$$

where β_i and β_j are bias terms regarding to time i and location j respectively, and μ is the overall average reading. Here the observation of physical sensor is broken down into four components: latent features of time stamps and locations, global average, time bias and location bias. The objective is to compute the parameters $\hat{\Gamma} = \{\mu, \beta_i, \beta_j, t_i, l_j\}$ by minimizing the following regularized squared error on the set of situation signals, denoted by S_{phy}.

$$S_{phy} = \sum_{(i,j)\in\mathcal{K}} (S_{ij} - \hat{s_{ij}})^2 + \Omega(\Gamma) \tag{5.2}$$

where $\Omega(\Gamma) = \lambda_{reg}(\|t_i\|^2 + \|l_j\|^2 + {\beta_i}^2 + {\beta_j}^2)$ is regularization function to control over-fitting, and $\mathcal{K}$ is the set of time and location pairs for which S_{ij} is used for training set. This formulation is based on the assumption that similar locations will have similar readings for similar time-stamps, where similarity is based on the situation matrix.

5.2.3 MF incorporating Social Signals: Latent Topics

When a physical event is happening, there are usually social discussions or on-the-spot posts generated by people in a variety of social media platform. Therefore a location can be easily associated with social information which could implicitly indicate events and offer an explanation of the situations. Looking at the social feeds, we are not only able to obtain the semantic explanation of physical sensor signals, but also to establish the relation between physical readings and social feeds. Location situations (i.e., emerging topics) are hidden in local social feeds and can be represented by semantically associated information such as topic modelling or sentiment analysis.

We use Latent Dirichlet Allocation (LDA) to uncovers hidden dimensions in the social information of a particular location, which could then be in line with the spatial latent vector of that location. Specifically, we model LDA to associate location document LD with a K-dimension topic distribution θ_{LD}, each dimension of which denotes the fraction of words in LD that discuss each of the K topics. We denote $\theta_{LD_j,k}$ as the probability of seeing topic k in location document LD_j. LDA models each topic k with word distribution ϕ_k, which encodes the probability a particular word is used in that topic, which is drawn from the Dirichlet distribution. The topic assignment of each word q for a location document LD_j is denoted as $z_{LD_j,q}$. Therefore, the likelihood of seeing the whole social information $\mathbb{S}$ is the multiplication of word distribution for each topic ϕ_k and the topic distribution for each location document LD_j across all the locations:

$$p(\mathbb{S}|\Theta, z) = \prod_{j} \prod_{q=1}^{N_{LD_j}} \theta_{LD_j, z_{LD_j,q}} \phi_{z_{LD_j,q}, w_{LD_j,q}} \tag{5.3}$$

where $\Theta = \{\theta, \phi\}$ and $w_{LD_j,q}$ is the q^{th} word in location document LD_j. Parameters relations are shown in Figure 5.2. Our idea is that during the matrix factorization the temporal response (time latent feature t_i) not only relates to the spatial latent feature (l_j), but also relate to the topic distribution θ_{LD_j} for that location. Therefore, we fuse the physical sensor information and social sensor information with a goal of fitting the situation matrix S from physical sensor readings and simultaneously maximizing the likelihood of seeing $\mathbb{S}$ being generated by the estimated LDA model. To achieve this we minimize the following objective function:

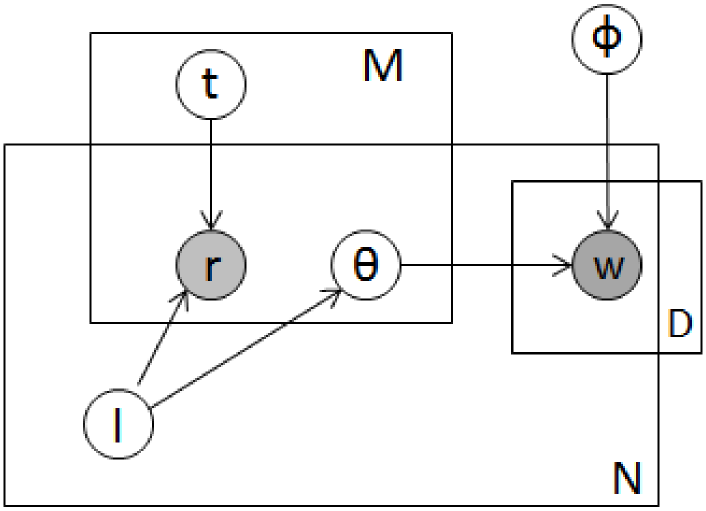


Figure 5.2: Parameters Relationship between Physical and Social Signals

$$f(S, \mathbb{S}|\Gamma, \Theta, z) = S_{phy} - \lambda_{social} p(\mathbb{S}|\theta, \phi, z) \tag{5.4}$$

For each location the latent features and topic distributions are aligned with the following equation:

$$\theta_{j,f} = \frac{exp(kl_{j,f})}{\sum_f exp(kl_{j,f})} \tag{5.5}$$

This loss function fuses both physical and social information in a unified model, and λ_{social} is the parameter to tune the weight of social information. Note that we implicitly assume that emerging situations will attract people's attention, resulting in a strong correlation between the topics and the physical sensor readings. This assumption will be proved in the experiment section.

5.2.4 Situation Prediction for Missing Readings

As discussed, the spatio-temporal density is different between these two sources. Physical sensors have geographically sparser readings than social sensors, which will result in a situation where there is social information in some locations but no physical sensors at all. We can exploit social information for predicting the readings for such *uncovered* locations. This is similar to a "cold-start" problem in the recommendation system, which cannot be effectively solved by traditional matrix factorization. To this end, we modify the model by taking account of the social information even if there are no readings. In details, among the set of $< timestamp, location, readings >$ triples used for training the model, a fraction will have no reading. For this reason, whenever the loss needs to be computed for the actual parameters we only consider the complete triples, because the difference between the reading and the prediction should be computed in order to perform gradient descent and update the parameters. However, if a location has no readings, the corresponding document which is built from social sensors will affect the latent feature based on equation 5.5.

When applying the framework for such scenarios the loss function is also modified in:

$$S_{phy} = \sum_{(i,j)\in\mathcal{K}} (S_{ij} - \hat{s_{ij}} - \beta_j)^2 + \lambda_{reg}(\|t_i\|^2 + \|l_j\|^2) \tag{5.6}$$

This means that we don't model biases for locations. The reason is that unlike locations with readings, uncovered locations will always have zero biases. Hence even for two locations having similar latent features, the bias difference will diminish their similarities, generating different reconstructed readings.

5.2.5 Handling Data Noise and Conflicts

Even if MF can result in large performance for very sparse matrices, a limitation is that it is highly sensitive to noise. Since we are dealing with physical sensor data, the data could be very noisy for a variety of reasons such sensor failure, environment change and so on. In addition, sensors may have conflicting readings due to their false positive or false negative sensing. In order to filter out the noise, we incorporate the temporal pattern of social information and content similarity into our framework before doing MF.

Sub-sampling using Social Content Similarity: As values are significantly affected by noise, traditional factorization approach is not directly applicable to such input. Such a noisy input will result in the latent representations of locations and times to equally fit the correct and incorrect entries. Hence, before factorizing the situation matrix a sub-sampling is performed in order to decide which observations should be used for training. Sub-sampling consists of defining the set $\mathcal{K}$ in equation 5.2, in other words selecting which observations S_{ij} to use during MF.

Intuitively, the sub-sampling should remove entries from the situation matrix minimizing conflicting signals. Locations that are geographically close to each other are highly likely to share similar topic distribution since people could post similar words in a small area, especially when events are happening. If one or more of such sensors is faulty or for other reason is not able to capture the reason that causes people to talk about a particular topic, there will be conflicting data for sensor readings, resulting in a weaker correlation between the two modalities. Sub-sampling is designed in order to detect similar situations and solve the conflicts.

How the situation matrix is sub-sampled is described in details in the

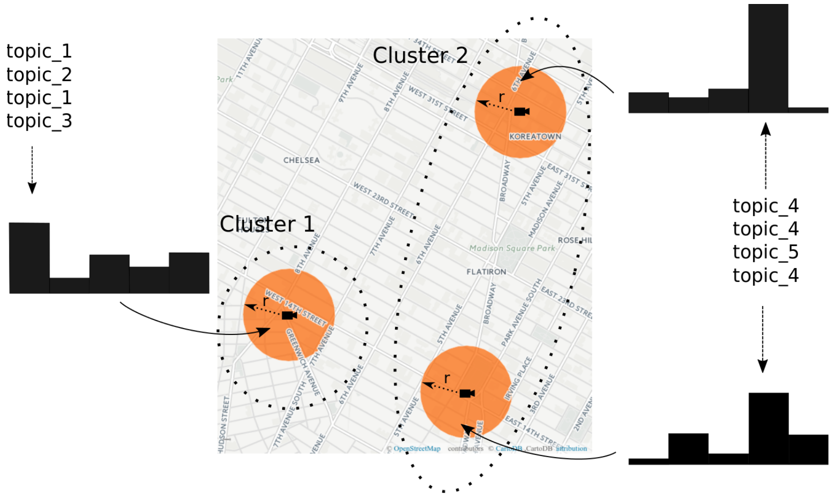


Figure 5.3: Clustering for Sub-sampling: cluster 2 contains locations where social sensors share a particular topic, while the leftmost location is assigned to cluster 1, as topics are different

following. Locations are clustered based on their surrounding social content similarity, that is estimated for each location by the social signals. Specifically, LDA topic modelling is applied on LD_j for each location j, extracting the topics distribution for each location document. The number of topics can be empirically set. For each topic, a cluster of locations is created, where the topic index of the largest response is considered for assigning a location to a cluster. Figure 5.3 shows an example of how two locations are associated to two different clusters. In order to distinguish when two locations in the same clusters have conflicting readings, a smoothing filter with a Hanning window is applied, resulting in a smooth signal. After that, within each cluster, if there is at least one location of which readings are larger then $\mathcal{T}$ for a sufficient time interval, the cluster will be marked as containing conflicting readings. In such cases, any reading lower than $\mathcal{T}$ will be omitted from the location for the whole temporal window since our framework is able to reconstruct the missing readings based on social sensors.

Physical Signal Correction using Social Impulse Pattern: When a large-scale event is happening, it can be detected not only by physical sensors but can also be reflected by social signals. Many features could be extracted from social sensors to detect events [113, 121]. In order to avoid taking false alarm values from physical sensors for the training set, we correct physical signals by incorporating social information temporal pattern. Specifically, we give each S_{ij} a event confidence weight $c_{ij} \in [0,1]$ of it original values. The event confidence values are derived from the more or less abrupt change in the number of social information at the same location j, and is calculated by:

$$c_{ij} = \left| \Omega(LDj^{\delta}) - \frac{1}{ND} \sum_{d=1}^{ND} \Omega(LDj^{\delta_d}) \right| \tag{5.7}$$

where LDj^{δ_d} is the location document in other days during the same time window δ, and function Ω counts the total number of records in the location document. In such case, the term $S_{ij} - \hat{s_{ij}}$ in equation 6 will change to $c_{ij}S_{ij} - \hat{s_{ij}}$ High values of c_{ij} will have no big impact on S_{ij}, while low values will regularize false alarms from physical sensor readings. The underlying assumption is more posts will be generated in social media when abnormal situations are occurring.

5.3 Experiments

In this section, we evaluate our proposed fusion method for different tasks: spatio-temporal situation prediction and event detection. We first introduce two real-world datasets and report experimental set-ups and data cleaning procedure. After that, we analyze the correlation between physical sensors and social sensors. Finally, we demonstrate how fusing physical sensor data with social sensor data can result in better performance in these tasks.

5.3.1 Datasets

SG Haze Data: we crawled Historical PSI Readings from Singapore National Environment Agency[1] for analyzing haze-related data. The data are collected from 5 stations[2] in corresponding five districts of Singapore, and span from 3 weeks: 1st-7th, 12th-19th August when there was no haze in Singapore, and 22nd-29th September when there were severe PSI values. Meanwhile, we collect geo-tagged tweets in Singapore during the same temporal window. We will make this dataset available upon request. For each week we focus on the meteorological situation for seven locations. A more clear visualization is shown in Figure 5.4. We choose three (l_1, l_3 and l_5) of these locations and leave the other four for prediction task. The PSI readings rate is one per hour and we average consecutive three hours (starting at 1 am) for each time stamp. 7 days of readings will result in 56 ($24/3 \times 7$) rows. Thus, we form a 56×21 PSI *Event Matrix* where each line represents PSI readings of a 3 hours temporal window and the columns denote 7 locations in 3 weeks. Note that we treat locations of different weeks as different locations; this is valid since we PSI readings are different from week to week for the same locations. With this representation, we simulate a scenario with a larger number of sensors with geographically diverse readings. The PSI situation matrix is shown in Figure 5.5.

NYC Traffic Data: we used the dataset shared by the authors of work [115], which contains continuous image snapshots from 149 CCTV traffic cameras across Manhattan, New York City and geo-tagged tweets. Each tweet contains the text content, the time when it was posted and the geographical coordinates (in the form of latitude and longitude). This data set contains dozens of events in various types including protests, festivals, parades, marathons and

[1] http://www.nea.gov.sg/anti-pollution-radiation-protection/air-pollution-control/psi/historical-psi-readings

[2] http://aqicn.org/map/singapore/

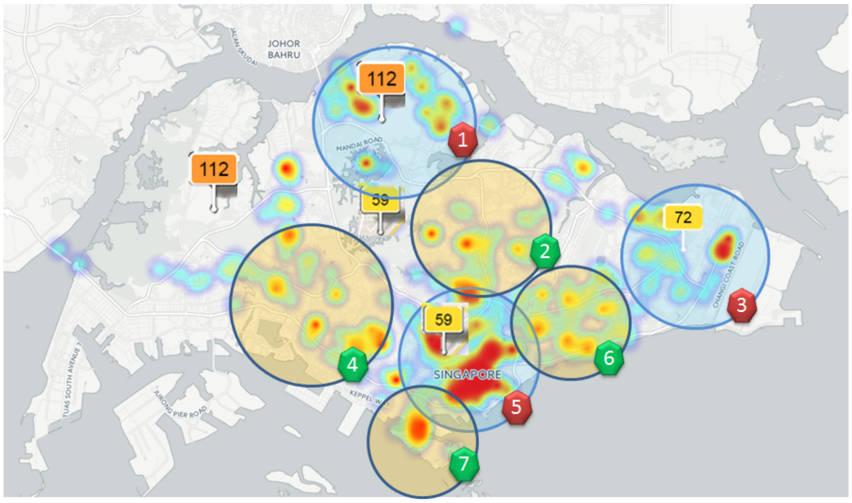


Figure 5.4: Haze Data Selection

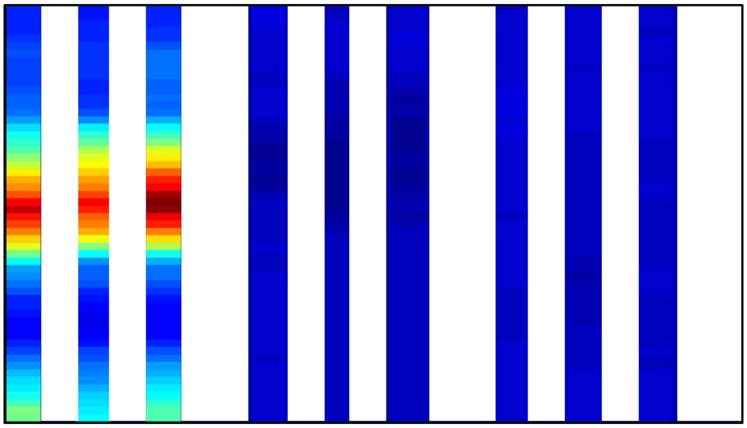


Figure 5.5: Haze Situation Matrix

etc. Different concepts detectors (e.g., parade, people marching, crowd, car etc) are applied to the images, resulting in concept confidence value (ranging from 0 to 1) in each spatial-temporal unit. For our experiments, we chose a subset of 135 cameras and selected a temporal window from 3 pm and 4 pm, 13th December 2014, in order to detect the large-scale "Millions March NYC" protest event [3].

Tweets Cleaning: in order to effectively apply topic modelling over texts, we need a clean tweet without non-standard tokens such as URLs, emojis, e-

[3] http://www.millionsmarchnyc.org/

moticons or slang words. This requires pre-processing tasks such as tokenizing and normalizing to be performed reliably. However, due to the informal language, spacing errors, punctuation errors, typos, etc. (e.g., *"Matchball...cant believe it!!! :-))) #sg50 http://< url >"*), off-the-shelf solutions perform poorly on tweets for these tasks. We therefore implemented a text preprocessing pipeline optimized to handle informal writing style of social media messages. First, we split tweets into words and other parts with a distinct meaning, meanwhile labelling tokens according to their type (e.g., words, numbers, user mentions, hashtags, email addresses, URLs, emoticons). Second, we apply our own normalizer that can normalize the common concept of expressive lengthening. After this pre-processing, the number of tweets and cleaned words of each spatio-temporal cell are shown in Table 5.2.

Table 5.2: Tweets Density in Situation Matrix

	#tweets	#words	#words per location
SGHaze	19073	178825	8515
NYCTraffic	3335	30127	223

5.3.2 Situation Awareness: Singapore Haze

We performed two experiments in order to evaluate the performance of the task described in the previous sections and to validate the assumptions of correlations between physical and social. To prove the effectiveness of the fusion, all the experiments' performance is analyzed by comparing with a baseline that consists in using signals from only physical sensors.

Correlations between Physical and Social Sensors:

Our work is based on the assumption that given an occurring spatio-temporal physical event, the social sensor information gives explanations to physical sensors readings, which means there is a correlation between the two sources. Our first experiment is to prove the existence of this correlation, and show how it can be measured quantitatively.

We form a dense haze matrix $E^{hz}_{56\times 9}$ similar to Figure 5.5 without locations with empty values. $E^{hz}_{56\times 9}$ is then factorized using the code from [77] into temporal latent matrix $T = [t_1, ..., t_{56}]^T \in \mathbb{R}^{20\times 56}$ and spatial latent matrix $L = [l_1, ..., l_9] \in \mathbb{R}^{20\times 9}$. A PSI reading for the element of haze matrix is the multiplication of corresponding time feature and location feature in the latent spaces. After a proper tuning of parameters, the dimension of latent factors is set to 20 and the regularizers λ_{reg} and λ_{social} are set to 0.1 and 0 (meaning tweets are not considered) respectively. This is because only a high number can represent the topic diversity in the tweets.

During the factorization, the proportion of training, validation and test set is 80%, 10% and 10% respectively. After factorization, we measure the similarity of two latent factors l_i, l_j using Spearman's rank correlation coefficient ρ_{ij}, which assesses statistical independence between two variables. A high ρ_{ij} means there is a high correlation between location i and j in terms of latent features. We consider all the tweets posted around locations as a bag of words for the location and incorporate tweets LDA in the matrix factorization. After this, all the spatial latent factors are embedded with topics information from social sensors. We examine how such social embedded latent factors can reflect better representation for original PSI readings. To do this, we compute ρ^{psi}_{ij} for all locations combinations (9 locations generating 36 pairs) in terms of PSI readings and take them as the ground truth correlations between locations. Similarly, we compute ρ_{ij} for all locations combinations in terms of latent factors with and without considering tweets LDA, denoted as ρ^{l}_{ij} and ρ^{l+tw}_{ij} respectively. For each locations pairs (e.g., l_1, l_3), we calculate *Dist(*ρ^{l}_{ij} *,* ρ^{psi}_{ij} *)* and *Dist(*ρ^{l+tw}_{ij} *,* ρ^{psi}_{ij} *)*. If there is correlation between tweets and PSI readings, the *Dist(latentLDA, PSI)* is supposed to be smaller than *Dist(latent, PSI)*. Figure 5.6 shows these two ρ_{ij} distances for all the locations pairs.

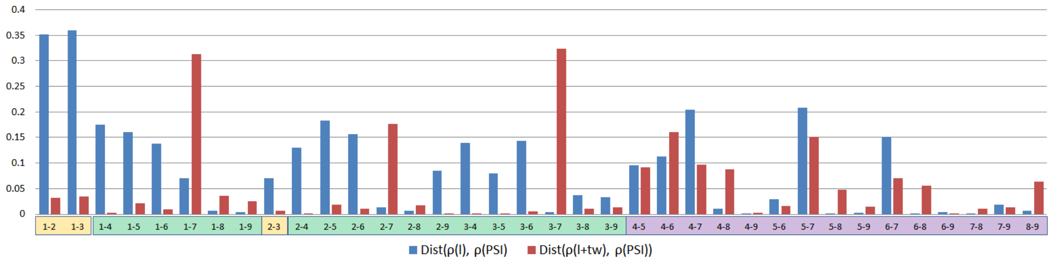


Figure 5.6: Location Pair Correlations Comparison of Standard Latent Vector and Social Embedded Latent Vector

As can be seen, in most of the locations pairs, incorporating tweets LDA in MF results in smaller distance to PSI readings in terms of the correlations between locations. This suggests that social sensors' contribution in describing physical sensor signals. Moreover, according to the ground truth, we categorize the location pairs into 3 clusters with regarding if the locations in each pair both observe events (yellow segments), neither observe events (purple segments), or only one observes events (green segments). We calculate for each category, the number of pairs where tweets incorporated ρ_{ij}^{l+tw} is closer to the ground truth than ρ_{ij}^{l}. Table 5.3 shows the percentage of pairs for the three categories.

Table 5.3: Number of Location Pairs Where Tweets Having Better Explanation in Different Categories

	Both Observe Event	Neither Observe Event	Only One Observes Event
#Location Pairs	3	15	13
#Better Correlation	3	12	6
Percentage	100%	80%	46.15%

It is shown that, in the first category, ρ_{ij}^{l+tw} is better than ρ_{ij}^{l} in all location pairs and, closer to the ground truth in 80% of pairs; while in the third category, ρ_{ij}^{l+tw} is better for 64.16 % of the location pairs. This implies that there is a strong correlation between social sensor and physical sensor in the place where events are occurring, and the tweets are good at describing reasons difference in event and non-event physical signals of different locations. However, for the places where no events are happening, social information has less ability to explain why there are no event signals. This is due to the diversity of content in tweets, especially in the places where any kinds of topics could be discussed.

Spatio-Temporal Situation Prediction:

One of the main strengths of the matrix factorization framework is that it is able to infer from social sensors the readings for locations which are not covered by physical sensors, provided the existence of a correlation between the two modalities, which is shown in the previous experiment.

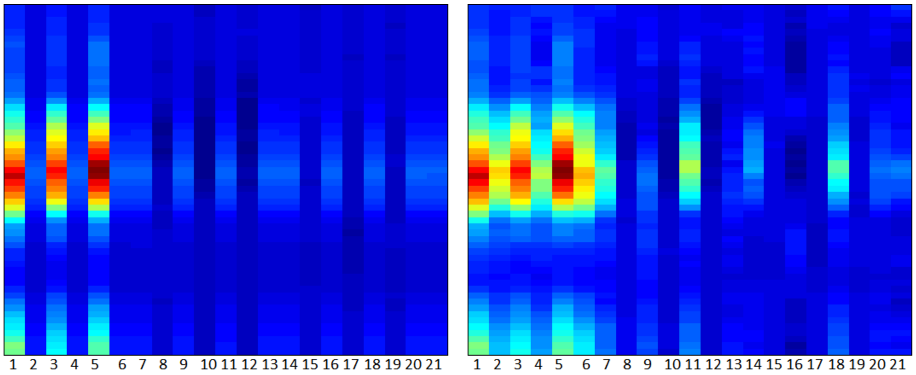
The following experiments consist of applying the MF framework on the entire dataset, comprehending for locations without readings. We choose to use 19 as the number of dimension for latent vectors and used the experimental setting as shown in Table 5.4.

Table 5.4: Experimental Setting for Situation Prediction

	latentReg	lambda	numTopics
Physical + Social	0.001	10	13
Physical Only (baseline)	0.1	0	19

What we expect is that for such locations high PSI readings will be predicted hours when people tweeted about haze. Providing a quantitative evaluation is not trivial in this case because a ground truth is not available for such missing locations. For this reason, we both provide a qualitative analysis of results and a root mean square error on a hand-built reasonable ground truth.

Figures 5.7(a) and 5.7(b) show the predictions for PSI readings obtained using only physical sensors (baseline) and both the modalities respectively. This qualitative analysis confirms that using only physical sensors the framework is not able to predict the situation in uncovered locations: such values can be inferred only from the time biases learned during the factorization. On the contrary, adding social information in the framework allows to learn semantically meaningful latent features for the locations, and the similarities of topics between different locations help predicting expected responses for the locations without physical readings. As can be seen from the figure, for the first week (columns from 1 to 7) the temporal trend of PSI readings for uncovered locations is similar to the others. However, in the other two weeks (remaining columns) it is easy to notice that for two day segments the readings are mistakenly predicted as large PSI observations (columns 11 and 18). A possible explanation for those false positives is that the topic (or topics) that were associated by LDA to "haze" contained also terms that were accidentally contained in the tweets observed in those locations and weeks.



(a) Only Physical Sensor (b) Physical + Social Sensor

Figure 5.7: Comparison of Prediction using Different Sources

For the quantitative evaluation, a hand-built ground truth is created assuming that PSI values are approximately constant among nearby locations in a three hours temporal window. In this evaluation, we want to measure how much for an uncovered location the predicted value is similar to the average of the available nearby physical readings. This ground truth matrix is then compared with the reconstructed matrix after factorization and RMSE (shown in Table 5.5) is computed over the whole matrix; It is shown that MF of fused with social performs much better than that of using only physical readings. This is because social information help to constrain the MF in tuning social

Table 5.5: RMSE of Reconstructed Matrix

	RMSE
Physical	40.84
Physical + Social	**25.87**

embedded latent factors, for the locations with similar physical readings.

3-Fold Cross Validation:

The validation of prediction above is demonstrated in a 3-fold cross-validation where in each fold, we use two of the three locations with data to predict the situation of the third one and compare the predicted physical sensor readings with ground truth. From each location, we filter out the words that appear frequent in the past (e.g., locations names, stop words), so as to make the location documents more representative for the location. The comparison of using only physical readings and incorporating tweets are shown in Table 5.6 in terms of MSE with ground truth

Table 5.6: 3-Fold Cross Validation Measured by MSE

	Loc 1	**Loc 2**	**Loc 3**
noTweets	51.38	49.45	57.73
withTweets	**42.48**	**26.80**	**22.86**

As can be seen, fusing social information gives us results representing the truth. This is because the topics of social information from a particular location can generally indicate the reason of the physical sensor readings in that place.

Topic Discovering:

Even if matrix factorization ignores any semantics that words may have, visualizing topics help to have an intuition on the results. For this reason, some top words of four topics after LDA after matrix factorization are here listed:

1. hazy, hari, haze, gardens, uffc

2. internationalcosplaydaysingapore, icds, sghaze
3. the, and, with, for
4. iphone, airport, changi, terminal

Topic number 4 contains mostly terms related to Singapore airport and topic number 3 clusters stop-words, while the first two topics contain words related to haze. We observe that the latent features of locations in the first week are likely to have much a stronger response for topics 1 and 2 than that of the other two weeks. This is because during haze days many people tweeted about haze in specific locations, hence both in a covered and uncovered location, the corresponding latent features will then be shaped to be similar, resulting in similar predictions.

5.3.3 Noise Filtering: NYC Large Scale Events

Unlike the clear PSI reading in SG Haze, NYC Traffic Data contain a huge number of noise in readings from cameras due to the unstable status of cameras (such as under maintenance, covered by raindrops). For this reason, we need to build a testing set from external knowledge. We manually set a binary label for every spatio-temporal point in the matrix. A binary ground truth matrix was built where values were set as positive if an event was observed in the video frames or if tweets around a locations contained a large enough number of occurrences of specific keywords, which indicated that the event was effectively there (e.g., “I am at the millionsmarchnyc protest”). Two evaluations are performed: the first is based on RMSE and the second uses metrics from classification tasks.

Evaluation with Root Mean Squared Error:

RMSE based evaluation will measure how much combining physical and social sensors reduces the noise, compared with cameras readings only. Instead of performing RMSE on the entire matrix we compute two different errors.

A first one considers only the locations where positive labels were put during ground truth creation. This measurement ($RMSE_{pos}$) will evaluate the contribution of social sensors in confirming positive readings. Another error is then computed for locations that were labelled as negative in the ground truth, but only for those who have significantly different tweets from positive ones. The motivation is that if a location shares similar tweets with another for which we are sure the event took place, then it becomes difficult to label it either as positive or negative. We then exclude such locations from our evaluation, considering only those for which we can tell with high confidence that there was not an event. This second error ($RMSE_{neg}$) will evaluate if the fusion is able to remove false positives in the camera readings. To demonstrate the effectiveness of our fusion model, we conduct our experiment on a large-scale dataset containing a variety of situations and events happening in New York City, including "MillionsMarchNYC" (M), "St Patrick's Day Parade" (S) and "ColumbusDay Parade" (C). We apply different concept detectors including "people marching" and "Crowd" etc to obtain the physical sensor readings. Results of these events of different times are given in Table 5.7.

Table 5.7: RMSE of Different Experiments

	$RMSE_{pos}$		$RMSE_{neg}$	
Events &Times	physical only	physical +social	physical only	physical +social
M14-15	0.17	**0.16**	0.16	**0.06**
M15-16	0.75	**0.51**	0.14	**0.04**
M16-17	0.72	**0.59**	0.20	**0.15**
S11-12	0.61	**0.47**	1.35	**0.09**
S12-13	0.61	**0.57**	0.11	**0.03**
S13-14	0.63	**054**	0.15	**0.06**
S14-15	**0.67**	0.76	0.16	**0.13**
S15-16	0.71	**0.7**	0.15	**0.1**
C12-13	0.76	**0.67**	**0.11**	0.15

The table shows in most cases, fusion of physical and social sensor outperforms the baseline of using only physical sensors for filtering noise by the measurements of RMSE. One exception happens in "St Patricks Day Parade" 14-15. A plausible explanation for our observations is that in one of the locations (5 Avenue at 72 Street) there are also significant values from cameras telling about "people marching". However, that location (a relatively wild place outside the central park) lacks tweets that are able to correct the noise, hence results in a large prediction error comparing with ground truth. It is also worth noticing that the different is less significant in the positive ground truth than that of negative. This is because the sub-sampling step takes advantages of the temporal change of tweets numbers to correct false alarms in the camera readings.

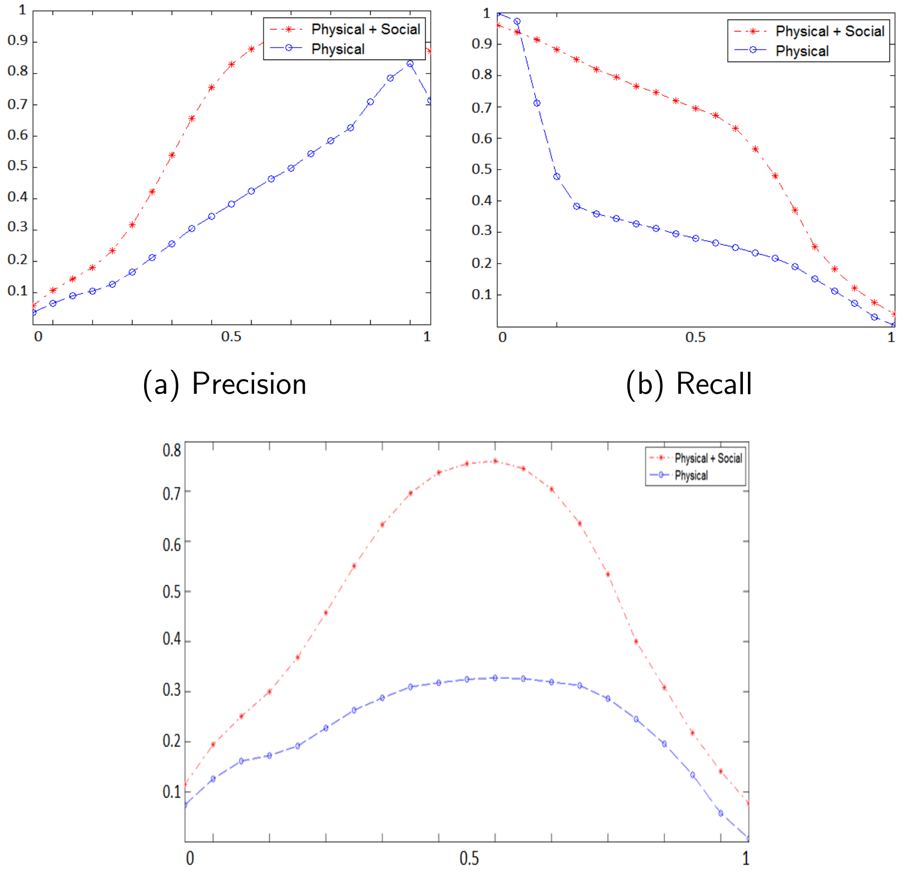
Event Classification Evaluation:

In the previous evaluation, the output is not binary like the ground truth, hence it is useful for comparing the performance of the fusion with the baseline, but it is not able to tell how much a result is good independently. For this reason, we translated the noise filtering problem in a classification one, where the goal is to classify spatio-temporal points into positive or negative instances (according to if the protest is happening). In order to get a binary output we define a threshold in $[0, 1]$ and compute precision, recall and F_1 score on the entire ground truth matrix. Figure 5.8 shows the results for different thresholds. As can be seen, social sensors significantly improve the performance in all the metrics: F_1, Precision, and Recall. A threshold of 0.55 will produce a F_1 of 0.76.

Table 5.8: F_1 Comparison with Previous Work and Baseline

	Baseline	Work [115]	Our Result
F_1 Score	0.11	0.74	**0.76**

We compare our result with that shown in Chapter 3. The baseline is using the threshold of 0.5, and the work in Chapter 3 tunes this threshold to 0.07 for



(a) Precision

(b) Recall

(c) F_1

Figure 5.8: Evaluation Comparison between Incorporating Tweets and Without Tweets

2 only cameras. Table 5.8 shows that without special tuning of the threshold for each individual camera, a global threshold for the cameras outperforms their results in terms of F_1 score.

5.4 Discussion and Summary

In this chapter, we proposed an innovative way of fusing spatial and temporal physical and social information in a unified matrix factorization based framework to handle multimodal and multi-source analysis. Our framework focuses on handling missing readings, noise and conflicting data. We applied it for situation awareness and event filtering on two real-world datasets and proved the correlation between these two modalities. Experiments showed that by incorporating social information the performance of both tasks can be largely improved. Compared with the Cmage based fusion method proposed in Chapter 4, this method doesn't require a dense network of physical sensors where the Gaussian Process can result in a better prediction. Besides, Cmage based fusion encloses a social term selection process, where only event-related social terms are used to fuse with physical sensors. This requires a prior knowledge of the events and therefore the method is suitable for planned events. Instead, the MF based fusion method in this chapter doesn't require preselected terms in the fusion process. They are implicitly fused with physical sensors readings. The method tries to automatically predict the information using the correlation between physical sensor signals and social terms by incorporating topic distribution in the social network. So this method could be applied when handling unplanned events such as the "haze situation" example shown in the experiment.

Chapter 6

Conclusion

6.1 Summary

In this thesis, we have investigated a new and emerging research topic: fusing physical sensors with social sensors for situation awareness. We conducted a thorough study on how information from these two sources of different modalities can be combined in order to provide humans a better situation understanding or prediction. Research questions raised in Chapter 1 are answered accordingly: i) we designed a multilayer process framework (Chapter 3) to integrate heterogeneous data from both physical and social sensor data; ii) the multi-layer tweeting camera framework contains concept detection and data analysis layers to extract meaningful information from both sources; iii) noise and uncertainty in each individual sensor data are handled via Cmage data representation and the hybrid fusion techniques discussed in Chapter 4; iv) multimodal complementary information is utilized through Matrix Factorization techniques (Chapter 5) to suppress the sensing noise for situation understanding; v) the Matrix Factorization technique is shown to be suitable for fusing spatio-temporal-semantic data.

First, we designed a multi-layer tweeting camera framework that integrates physical sensor feeds from CCTV Traffic cameras with social sensor information from Twitter tweets, to detect various concepts related to occurring events. Specifically, we considered visual concepts detected from images captured by cameras as "camera tweets" that are sent by cameras. They are represented by a unified data structure probabilistic spatio-temporal (PST) and aggregated into a concept-based image (Cmage) for better visualization of the overall situation in a geographical region in the real world. We defined a set of operators and analytic functions in a query-based mechanism to monitor the pattern or change of these PST data, so as to discover abnormal situation from evolving physical sensor signals. These physical signals are then explained by geo-tagged surrounding tweets that give higher level semantic meanings for the physical sensor readings. We demonstrated this conceptual framework by a smart sensing device "tweeting camera". Its sensing, analyzing and learning ability shows that it is interesting and promising to make physical sensors join human social networks and meanwhile bring humans in the loop for situation understanding.

Second, we extended the conceptual Cmage representation proposed in the first work and designed a hybrid fusion method that handles the sparsity and noise of sensor information, the heterogeneity of physical and social signals, and the influence of signals based on different geographical locations. Specifically, we applied the Gaussian Process model to interpolate event situation where there are no physical sensors, using nearby sensor readings. Also, we used Bayesian approach to eliminate noise from each source and enhance the detection of events leveraging the agreement of these two sources. The fusion results are evaluated according to the semantics of fused Cmage (i.e. the concept of the Cmage) which could be to understand the relationship of different concepts for a specific event.

Third, we proposed a unified fusion technique based on matrix factorization. We argue that matrix factorization used in the recommendation systems area is well suited to our problem that considers physical sensors reading in numerical format and social sensor feeds that are in symbolic representation. When we formalize the physical sensor readings in a temporal-spatial event matrix, the surrounding tweets implicitly gives the reason of physical readings, which could be aligned to specific topics for particular locations during factorization. Our conducted experiments found the correlation between these two sources, which suggests that fusing them can lead to a better performance in noise reduction, event detection, as well as situation prediction.

To sum up, the works in this thesis investigate the fusion of these two sources from different perspectives, and the experiments demonstrate the feasibility and efficiency of complementary sensors information.

6.2 Future Work

There are several interesting directions that are worth exploring for the future work:

- To extend the multi-layer tweeting camera framework, we plan to integrate information from social media sensors that include trend analysis and topic mining, to improve the quality of the camera tweets. By mining the main topics and monitoring the evolving of specific topics, we can be more precise in predicting or correcting the noise given by physical sensors. In addition, we also would explore the possibility of creating an interactive framework that allows for the use of general user as well as the integration of other types of sensors into the framework. More comprehensive set of concept detectors including face detectors and image caption labellers can also be used in the first layer of the framework.

- In the hybrid fusion method, we could investigate more sophisticated ways to reflect the relatedness between concepts and terms. The relatedness of different visual concepts or social terms could be utilized in the fusion process in order to result in a cleaner situation Cmage. One way to exploit the semantic relatedness is to consider the ontology from various lexical databases, e.g., WordNet [82]. Besides, fusion results might also give a guide on building a dynamic ontology for new situations or events. We envision our system to provide detailed information about so far unknown events, i.e., to automatically find patterns and correlations across semantics. The main challenge here is to calculate the correct relatedness between unknown terms (e.g., new popular hashtags in tweets) existing terms and concepts.

- For the fusion based on matrix factorization, as currently we only consider situation prediction in the spatial dimension, it would be interesting to consider the temporal evolving factors in order to make short-term predictions of ongoing situations. Comparing with other matrix factorization methods such as [93] is also worth exploring. Moreover, substituting topic distribution with other text features involving, for example, semantics or sentiments, is also a direction worth exploring. It is also worth considering more diverse datasets (e.g., YouTube, Flickr, Instagram) for more events to further prove the correlation of these two sources and also apply our methods to other areas such as semantic analysis, sentiment analysis or stock analysis that has similar properties as our problem (i.e. spatio-temporal dynamics, numeric and symbolic readings) At last, how to make our detection and prediction process in real-time is also an important direction that needs further investigation.

Bibliography

[1] Daniel J. Abadi, Don Carney, Ugur Çetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, and Stan Zdonik. Aurora: A new model and architecture for data stream management. *The VLDB Journal*, 12(2):120–139, 2003.

[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):555–560, March 2008.

[3] Amir Afrah, Gregor Miller, Donovan Parks, Matthias Finke, and Sidney Fels. Hive: A distributed system for vision processing. In *ACM/IEEE International Conference on Distributed Smart Cameras*. IEEE, 2008.

[4] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38, 2011.

[5] Luca Maria Aiello, Georgios Petkos, Christian Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Goker, Ioannis Kompatsiaris, and Aldo Jaimes. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282, 2013.

[6] Mohammad Akbari, Xia Huc, Nie Liqianga, and Tat-Seng Chua. From tweets to wellness: Wellness event detection from twitter streams. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[7] Tim Althoff, Damian Borth, Jörn Hees, and Andreas Dengel. Analysis and forecasting of trending topics in online media streams. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 907–916, 2013.

[8] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011.

[9] Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

[10] Pradeep K Atrey, M.Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, 2010.

[11] Pradeep K Atrey, Mohan S. Kankanhalli, and Ramesh Jain. Information assimilation framework for event detection in multimedia surveillance systems. *Multimedia Systems*, 12(3):239–253, 2006.

[12] Pradeep K Atrey, Mohan S Kankanhalli, and John B Oommen. Goal-oriented optimal subset selection of correlated multimedia streams. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1):2, 2007.

[13] L. Atzori, A. Iera, and G. Morabito. From ”smart objects” to ”social objects”: The next evolutionary step of the internet of things. *IEEE Communications Magazine*, 52(1):97–105, 2014.

[14] Luigi Atzori, Antonio Iera, Giacomo Morabito, and Michele Nitti. The social internet of things - when social networks meet the internet of things: Concept, architecture and network characterization. *Computer Networks*, 56(16), 2012.

[15] Raouf Babari, Nicolas Hautière, Éric Dumont, Nicolas Paparoditis, and James Misener. Visibility monitoring using conventional roadside cameras–emerging applications. *Transportation research part C: emerging technologies*, 22:17–28, 2012.

[16] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. In *International AAAI Conference on Web and Social Media*, 2011.

[17] James Benhardus and Jugal Kalita. Streaming trend detection in twitter. *International Journal of Web Based Communities*, 9(1):122–139, 2013.

[18] Chidansh A. Bhatt, Pradeep K. Atrey, and Mohan S. Kankanhalli. A reward-and-punishment-based approach for concept detection using adaptive ontology rules. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 9(2):10:1–10:21, 2013.

[19] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[20] Oren Boiman and Michal Irani. Detecting irregularities in images and in video. *International Journal of Computer Vision*, 74(1):17–31, 2007.

[21] G. Bradski. *Dr. Dobb's Journal of Software Tools.*

[22] M. Bramberger, A. Doblander, A. Maier, B. Rinner, and H. Schwabach. Distributed embedded smart cameras for surveillance applications. *Computer*, 39(2):68–75, 2006.

[23] Hongyun Cai, Yang Yang, Xuefei Li, and Zi Huang. What are popular: Exploring twitter features for event detection, tracking and visualization. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 89–98, 2015.

[24] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *International Workshop on Multimedia Data Mining*, pages 4:1–4:10, 2010.

[25] Sirish Chandrasekaran, Owen Cooper, Amol Deshpande, Michael J. Franklin, Joseph M. Hellerstein, Wei Hong, Sailesh Krishnamurthy, Samuel R. Madden, Fred Reiss, and Mehul A. Shah. Telegraphcq: Continuous dataflow processing. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pages 668–668, 2003.

[26] Jiawei Chen, Yin Cui, Guangnan Ye, Dong Liu, and Shih-Fu Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *Proceedings of International Conference on Multimedia Retrieval*, pages 1:1–1:8, 2014.

[27] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of ACM Conference on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009.

[28] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[29] Henriette Cramer and Sebastian Büttner. Things that tweet, check-in and are befriended.: two explorations on robotics & social media. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 125–126, 2011.

[30] Andreas Doblander, Andreas Zoufal, and Bernhard Rinner. A novel software framework for embedded multiprocessor smart cameras. *ACM Transactions on Embedded Computing Systems*, 8(3):24:1–24:30, 2009.

[31] Nicole B Ellison et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

[32] Mica R Endsley. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 32, pages 97–101, 1988.

[33] Mica R Endsley. A survey of situation awareness requirements in air-to-air combat fighters. *The International Journal of Aviation Psychology*, 3(2):157–168, 1993.

[34] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.

[35] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010.

[36] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[37] Mingyan Gao, Vivek K Singh, and Ramesh 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.

[38] Ross B. Girshick, Pedro F. Felzenszwalb, and David A. McAllester. Object detection with grammar models. In *Advances in Neural Information Processing Systems*, pages 442–450, 2011.

[39] Amirhossein Habibian, Koen E.A. van de Sande, and Cees G.M. Snoek. Recommendations for video event recognition using concept vocabularies. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, pages 89–96, 2013.

[40] Carleen Hawn. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*, 28(2):361–368, 2009.

[41] Mohamed A Helala, Ken Q Pu, and Faisal Z Qureshi. A stream algebra for computer vision pipelines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 786–793, 2014.

[42] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J Smola, and Kostas Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *WWW*, pages 769–778, 2012.

[43] Hsun-Ping Hsieh, Shou-De Lin, and Yu Zheng. Inferring air quality for station location recommendation based on urban big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 437–446, 2015.

[44] Hsun-Ping Hsieh, Tzu-Chi Yen, and Cheng-Te Li. What makes new york so noisy?: Reasoning noise pollution by mining multimodal geo-social big data. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 181–184, 2015.

[45] Guang-Neng Hu, Xin-Yu Dai, Yunya Song, Shu-Jian Huang, and Jia-Jun Chen. A synthetic approach for recommendation: Combining ratings, social relations, and reviews. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1756–1762, 2015.

[46] Mengdie Hu, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu Ma. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2751–2754, 2012.

[47] Nathan Jacobs, Walker Burgin, Nick Fridrich, Austin Abrams, Kylia Miskell, Bobby H. Braswell, Andrew D. Richardson, and Robert Pless. The global network of outdoor webcams: Properties and applications. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 111–120, 2009.

[48] Namit Jain, Shailendra Mishra, Anand Srinivasan, Johannes Gehrke, Jennifer Widom, Hari Balakrishnan, Uğur Çetintemel, Mitch Cherniack, Richard Tibbetts, and Stan Zdonik. Towards a Streaming SQL Standard. *Proceedings of the VLDB Endowment*, 1(2):1379–1390, 2008.

[49] Yu-Gang Jiang, Akira Yanagawa, Shih-Fu Chang, and Chong-Wah Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection. Technical report, 2008.

[50] Yu-Gang Jiang, J. Yang, Chong-Wah Ngo, and A.G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42–53, 2010.

[51] Krasimira Kapitanova, Sang H Son, and Kyoung-Don Kang. Using fuzzy logic for robust event detection in wireless sensor networks. *Ad Hoc Networks*, 10(4):709–722, 2012.

[52] Andrej Karpathy, Armand Joulin, and Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems 27*, pages 1889–1897. 2014.

[53] Yan Ke, Rahul Sukthankar, and Martial Hebert. Event detection in crowded videos. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[54] Lyndon Kennedy and Alexander Hauptmann. Lscom lexicon definitions and annotations (version 1.0). 2006.

[55] J. C. Klontz, B. F. Klare, S. Klum, A. K. Jain, and M. J. Burge. Open source biometric recognition. In *International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, Sept 2013.

[56] April Kontostathis, Leon M. Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. *Survey of Text Mining: Clustering, Classification, and Retrieval*, chapter A Survey of Emerging Trend Detection in Textual Data Mining, pages 185–224. 2004.

[57] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[58] Pavel Korshunov and Wei Tsang Ooi. Critical video quality for distributed automated video surveillance. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 151–160, 2005.

[59] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11:538–541, 2011.

[60] Jürgen Krämer and Bernhard Seeger. Pipes: A public infrastructure for processing and exploring streams. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, pages 925–926, 2004.

[61] Matthias Kranz, Luis Roalter, and Florian Michahelles. Things that twitter: Social networks and the internet of things. In *in What can the Internet of Things do for the Citizen (CIoT) Workshop at The Eighth International Conference on Pervasive Computing (Pervasive*, 2010.

[62] Matthias Kranz, Luis Roalter, and Florian Michahelles. Things that twitter: Social networks and the internet of things. In *What can the Internet of Things do for the Citizen (CIoT) Workshop at International Conference on Pervasive Computing*, 2010.

[63] Purushottam Kulkarni, Deepak Ganesan, Prashant Shenoy, and Qifeng Lu. Senseye: A multi-tier camera sensor network. In *Proceedings of the 13th ACM International Conference on Multimedia*, pages 229–238, 2005.

[64] Shamanth Kumar, Huan Liu, Sameep Mehta, and L. Venkata Subramaniam. From tweets to events: Exploring a scalable solution for twitter streams. *CoRR*, abs/1405.1392, 2014.

[65] Yin-Hsi Kuo, Yan-Ying Chen, Bor-Chun Chen, Wen-Yu Lee, Chun-Che Wu, Chia-Hung Lin, Yu-Lin Hou, Wen-Feng Cheng, Yi-Chih Tsai, Chung-Yen Hung, Liang-Chi Hsieh, and Winston Hsu. Discovering the city by mining diverse and multimodal data streams. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 201–204, 2014.

[66] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web*, pages 591–600, 2010.

[67] James Lanagan and Alan F Smeaton. Using twitter to detect and tag important events in live sports. *Artificial Intelligence*, pages 542–545, 2011.

[68] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, pages 251–258, 2011.

[69] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *WWW*, pages 251–260, 2012.

[70] Chenliang Li, Aixin Sun, and Anwitaman Datta. Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 155–164, 2012.

[71] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. Tedas: A twitter-based event detection and analysis system. In *IEEE 28th International Conference on Data engineering*, pages 1273–1276, 2012.

[72] Steve Lohr. The age of big data. *New York Times*, 11, 2012.

[73] Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King. Sorec: Social recommendation using probabilistic matrix factorization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 931–940, 2008.

[74] Ladislav Madarász, Rudolf Andoga, Ladislav Fozo, and Tobiáš Lazar. *Towards Intelligent Engineering and Information Technology*, chapter Situational Control, Modeling and Diagnostics of Large Scale Systems, pages 153–164. 2009.

[75] Lev Manovich. Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, 2:460–475, 2011.

[76] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158, 2010.

[77] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *ACM Recommender Systems*, pages 165–172, 2013.

[78] John McCarthy. *Philosophical Logic and Artificial Intelligence*, chapter Artificial Intelligence, Logic and Formalizing Common Sense, pages 161–190. 1989.

[79] L. Edwin McKenzie, Jr. and Richard Thomas Snodgrass. Evaluation of Relational Algebras Incorporating the Time Dimension in Databases. *ACM Computing Surveys*, 23(4):501–543, December 1991.

[80] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):873–889, 2001.

[81] Ulrich Meissen, Stefan Pfennigschmidt, Agnès Voisard, and Tjark Wahnfried. *Current Trends in Database Technology - EDBT 2004 Workshops*, chapter Context- and Situation-Awareness in Information Logistics, pages 335–344. 2005.

[82] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[83] A. Møgelmose, T. B. Moeslund, and K. Nasrollahi. Multimodal person re-identification using rgb-d sensors and a transient identification database. In *2013 International Workshop on Biometrics and Forensics (IWBF),*, pages 1–4, April 2013.

[84] Meenakshi Nagarajan, Karthik Gomadam, Amit P. Sheth, Ajith Ranabahu, Raghava Mutharaju, and Ashutosh Jadhav. *Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences*, pages 539–553. 2009.

[85] Prabhu Natarajan, Pradeep K. Atrey, and Mohan Kankanhalli. Multi-camera coordination and control in surveillance systems: A survey. *ACM Transactions on Multimedia Computing, Communications*, 11(4):57:1–57:30, 2015.

[86] Paul Over, Jon Fiscus, Greg Sanders, David Joy, Martial Michel, George Awad, Alan Smeaton, Wessel Kraaij, and Georges Quénot. Trecvid 2011-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID 2011-TREC Video Retrieval Evaluation Online*, 2011.

[87] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 344–353, 2013.

[88] Gemma Piella. Image fusion for enhanced visualization: a variational approach. *International Journal of Computer Vision*, 83(1):1–11, 2009.

[89] Hemant Purohit and Amit P Sheth. Twitris v3: From citizen sensing to analysis, coordination and action. In *International AAAI Conference on Web and Social Media*, 2013.

[90] Umakishore Ramachandran, Kirak Hong, Liviu Iftode, Ramesh Jain, Rajnish Kumar, Kurt Rothermel, Junsuk Shin, and Raghupathy Sivakumar. Large-scale situation awareness with camera networks and multimodal sensing. *Proceedings of the IEEE*, 100(4):878–892, 2012.

[91] Carl Edward Rasmussen. Gaussian processes for machine learning. MIT Press, 2006.

[92] Virgile Landeiro Dos Reis and Aron Culotta. Using matched samples to estimate the effects of exercise on mental health from twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 182–188, 2015.

[93] Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000, 2010.

[94] Timo Reuter, Symeon Papadopoulos, Giorgos Petkos, Vasileios Mezaris, Yiannis Kompatsiaris, Philipp Cimiano, Christopher de Vries, and Shlomo Geva. Social event detection at mediaeval 2013: Challenges, datasets,

and evaluation. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, 2013.

[95] B. Rinner, T. Winkler, W. Schriebl, M. Quaritsch, and W. Wolf. The evolution from single to pervasive smart cameras. In *ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–10, Sept 2008.

[96] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *Proceedings of the 11th International Conference on The Semantic Web*, pages 508–524, 2012.

[97] Mukesh Saini, Pradeep K. Atrey, Sharad Mehrotra, and Mohan Kankanhalli. W3-privacy: Understanding what, when, and where inference channels in multi-camera surveillance video. *Multimedia Tools and Applications*, 68(1):135–158, 2014.

[98] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860, 2010.

[99] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1257–1264, 2008.

[100] Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, and Jon Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51, 2009.

[101] W. Schriebl, T. Winkler, A. Starzacher, and B. Rinner. A pervasive smart camera network architecture applied for multi-camera object classification. In *Third ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–8, 2009.

[102] Amit Sheth, Ashutosh Jadhav, Pavan Kapanipathi, Chen Lu, Hemant Purohit, Gary Alan Smith, and Wenbo Wang. Twitris: A system for collective social intelligence. In *Encyclopedia of Social Network Analysis and Mining*, pages 2240–2253. 2014.

[103] Junsuk Shin, Rajnish Kumar, Dushmanta Mohapatra, Umakishore Ramachandran, and Mostafa Ammar. Asap: A camera sensor network for situation awareness. In *Principles of Distributed Systems*, volume 4878. 2007.

[104] Vivek K Singh. *Personalized Situation Recognition*. PhD thesis, University of California, Irvine, CA, USA, 2012.

[105] Vivek K Singh, Mingyan Gao, and Ramesh Jain. Social pixels: genesis and evaluation. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 481–490, 2010.

[106] Cees Snoek, Marcel Worring, Jan van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th ACM International Conference on Multimedia*, pages 421–430, 2006.

[107] Alisa Sotsenko, Janosch Zbick, Marc Jansen, and Marcelo Milrad. Flexible and contextualized cloud applications for mobile learning scenarios. In *Mobile, Ubiquitous, and Pervasive Learning*, pages 167–192. Springer, 2016.

[108] Qingquan Sun, Weihong Yu, Nikolai Kochurov, Qi Hao, and Fei Hu. A multi-agent-based intelligent sensor and actuator network design for smart house and home automation. *Journal of Sensor and Actuator Networks*, 2(3):557–588, 2013.

[109] Jiliang Tang, Xia Hu, Huiji Gao, and Huan Liu. Exploiting local and global social context for recommendation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 2712–2718, 2013.

[110] Mohan Trivedi, Kohsia Huang, and Lvana Mikic. Intelligent environments and active camera networks. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 2, pages 804–809, 2000.

[111] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[112] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[113] Maximilian Walther and Michael Kaisser. Geo-spatial event detection in the twitter stream. In *Advances in Information Retrieval*, pages 356–367. 2013.

[114] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120, 2012.

[115] Yuhui Wang and Mohan S. Kankanhalli. Tweeting cameras for event detection. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1231–1241, 2015.

[116] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. *International AAAI Conference on Web and Social Media*, 11:401–408, 2011.

[117] Utz Westermann and Ramesh Jain. Toward a common event model for multimedia applications. *IEEE MultiMedia*, 14(1):19–29, January 2007.

[118] Thomas D Wickens. *Elementary signal detection theory.* Oxford university press, 2001.

[119] Thomas Winkler and Bernhard Rinner. Security and privacy protection in visual sensor networks: A survey. *ACM Computing Surveys*, 47(1):2:1–2:42, 2014.

[120] Fei Wu, Zhenhui Li, Wang-Chien Lee, Hongjian Wang, and Zhuojie Huang. Semantic annotation of mobility data using social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1253–1263, 2015.

[121] Ke Xie, Chaolun Xia, Nir Grinberg, Raz Schwartz, and Mor Naaman. Robust detection of hyper-local events from geotagged social media data. In *Proceedings of the Thirteenth International Workshop on Multimedia Data Mining*, pages 2:1–2:9, 2013.

[122] Akito Yamasaki, Hidenori Takauji, Shun'ichi Kaneko, Takeo Kanade, and Hidehiro Ohki. Denighting: Enhancement of nighttime images for a surveillance camera. In *19th International Conference on Pattern Recognition*, pages 1–4, 2008.

[123] Rong Yan, Jun Yang, and Alexander G Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *Proceedings of the 12th ACM international conference on Multimedia*, pages 548–555, 2004.

[124] Akira Yanagawa, Shih-Fu Chang, Lyndon Kennedy, and Winstonr Hsu. Columbia university's baseline detectors for 374 lscom semantic visual concepts. Technical report, 2007.

[125] Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 471–480, 2015.

[126] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27(6):52–59, Nov 2012.

[127] Hanning Zhou and D. Kimber. Unusual event detection via multi-camera video mining. In *International Conference on Pattern Recognition*, volume 3, pages 1161–1166, 2006.

[128] Arkaitz Zubiaga, Damiano Spina, Víctor Fresno, and Raquel Martínez. Classifying trending topics: a typology of conversation triggers on twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2461–2464, 2011.