Tweeting Cameras for Event Detection

Yuhui Wang
NUS Graduate School for Integrative Sciences and Engineering
National University of Singapore, Singapore
wangyuhui@u.nus.edu

Mohan S. Kankanhalli
School of Computing
National University of Singapore, Singapore
mohan@comp.nus.edu.sg

ABSTRACT

We are living in a world of big *sensor* data. Due to the widespread prevalence of visual sensors (e.g. surveillance cameras) and social sensors (e.g. Twitter feeds), many events are implicitly captured in real-time by such heterogeneous "sensors". Combining these two complementary sensor streams can significantly improve the task of event detection and aid in comprehending evolving situations. However, the different characteristics of these *social* and *sensor* data make such information fusion for event detection a challenging problem. To tackle this problem, we propose an innovative multi-layer tweeting cameras framework integrating both physical sensors and social sensors to detect various concepts of real-world events. In this framework, visual concept detectors are applied on camera video frames and these concepts can be construed as "camera tweets" posted regularly. These tweets are represented by a unified probabilistic spatio-temporal (PST) data structure which is then aggregated to a concept-based image (Cmage) as the common representation for visualization. To facilitate event analysis, we define a set of operators and analytic functions that can be applied on the PST data by the user to discover occurrences of events and to analyse evolving situations. We further leverage on geo-located social media data by mining current topics discussed on Twitter to obtain the high-level semantic meaning of detected events in images. We quantitatively evaluate our framework with a large-scale dataset containing images from 150 New York real-time traffic CCTV cameras, university foodcourt camera feeds and Twitter data, which demonstrates the feasibility and effectiveness of our proposed framework. Results of combining camera tweets and social tweets are shown to be promising for detecting real-world events.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741634.

General Terms

Design; Experimentation; Performance

Keywords

tweeting cameras; social sensors; event detection; content analysis; social media

1. INTRODUCTION

We are witnessing a world of big *social* and *sensor* data. From visual sensors, wearable sensors, to humans as sensors [31], multiple sensor media streams constantly provide observations on the real world that are utilized in many aspects of our life: from industrial process control, robotics, surveillance, smart houses to situation awareness [34]. Examples of sensors include (a) ***physical sensors*** like cameras, accelerometers, gyroscopes, mobile phones, RFID tags, temperature sensors, humidity sensors and (b) ***social sensors*** like social networking sites containing user-generated content reporting events in all kinds of formats (text, image, video, etc.). On one hand, visual sensors, either static video cameras or mobile cameras embedded in smartphones, are rapidly increasing in numbers around the world. On the other hand, many types of social media platforms like Twitter, Weibo, Facebook, Wechat, Youtube, Flickr etc. allow humans as social sensors to report daily events or individual opinions, disseminate breaking news, discuss trending topics and discover social events. Therefore, our physical world is being monitored by these increasing numbers of physical and social sensors. These multi-modal streams of data can therefore facilitate event discovery since they implicitly capture the evolving situations around the world.

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. Since camera and social streams provide different facets of events or a situation, we argue that combining the two different but complementary streams would greatly improve event detection and further aid in comprehending evolving situations. Taking inspiration from the "things that tweet" [24, 13], we incorporate cameras into a social network to build a network of ***tweeting cameras***. This camera network "tweets" information to facilitate event detection – our key idea is to apply visual concept detectors on the camera data. The detected visual concepts at a camera can be considered as "tweets" posted by it. Multiple camera "tweet-

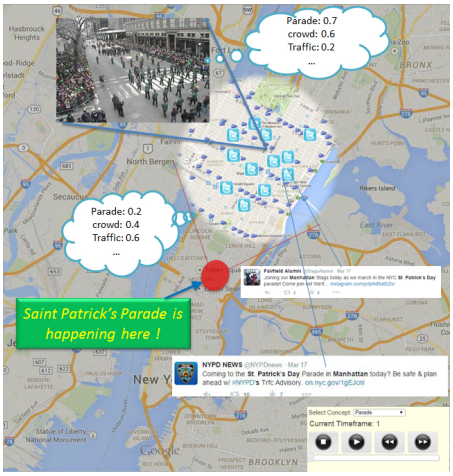


Figure 1: Tweeting Cameras and Twitter Tweets for Event Detection.

s" are then combined with twitter textual tweets from the vicinity to detect events occurring in that geographical area. The conceptual illustration of our idea is shown in Figure 1.

This introduces several challenging research problems:

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 camera data and the sheer volume of social media data, how can we handle the uncertainty and noise of these data?

In this work, we present an innovative approach to perform fusion of physical and social sensor data. 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. Concept detectors indicate the confidence of a detected concept in an image using a probability value. We consider these detected concepts as "Camera Tweets" with associated confidence values. Camera tweets at a particular geo-location can then be visualised as "concept pixels". Spatially aggregating such "concept pixels" creates a powerful and intuitive situation visualization interface, *concept-based image (Cmage)*, where both social information and sensor information are further fused to represent situations. For this purpose, we propose a multi-layer tweeting camera framework where tweets from cameras are analysed at different levels such that multi-level information is extracted and subsequently combined with social information to derive situational knowledge. In the first layer of the framework, we consider individual 'tweets' at the lowest level by applying a variety of concept detectors on the raw video data so as to extract low-level semantic information from visual features. After applying concept detectors on every camera in a spatial region, low-level semantic information is then represented by the proposed PST data structure and mapped to the *Cmage*. To facilitate pattern mining and mid-level information extraction, in the second layer, we aggregate all the processed low-level tweets and define a set of filtering operators and analytic functions which could be applied on the PST data to obtain mid-level processed camera tweets (e.g. concepts with high probabilities for a region). Mid-level tweets are aggregated information from several cameras which give insights into the overall situation pattern or general event trend by applying predefined analytic functions. At the highest level, information is extracted from social media (e.g. twitter data) to facilitate cross media analysis that provides the highest level semantic information about an event. The high-level processed 'tweets' analysed in the third layer allow us to infer the social meaning of a particular event. We tackle the social event detection problem with signal detection theory to separate the event 'signals' from social 'noise', and improve the detection accuracy by fusing the two complementary information sources (camera feeds and social feeds).

Contributions: To summarize, our main contributions in this work are:

1. Defining a unified probabilistic spatio-temporal data structure to handle uncertainty of physical sensors.

2. Proposing a multi-layer tweeting cameras framework containing innovative *concept image (Cmage)* and a set of filtering & analytic operators for user to query different levels of information from both physical and social sensors.

3. The aggregation of physical sensors and social sensors to overcome the unreliability of individual sensors and which improves the overall performance of event detection.

We demonstrate 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), and conduct evaluation experiments with four instances of real-world events.

The remainder of the paper is organized as follows. Section 2 presents works related to our research. Section 3 describes the proposed multi-layer tweeting cameras framework. Section 4 elaborates on the details of defined operators and functions, signal detection theory for event detection, as well as the social information extraction algorithm. Section 5 gives qualitative demonstrations and comprehensive statistical evaluation on the data. Section 6 discusses issues related to our work and Section 7 concludes the paper.

2. RELATED WORK

Event Detection using Visual Sensors

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 utilized for phenology study, scene and environment understanding [19, 6]. A multi-tier network SensEye of heterogeneous cameras has been proposed to overcome the disadvantage of single-tier networks in surveillance application, performing object detection, recognition and tracking tasks [25]. Distributed smart cameras [11], 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 [10] or abnormal events [1] occur. Event is an elementary concept not only for humans but also in multimedia applications. Several works propose event-related models [39], as well as the concept of atomic and compound events [5] in video application. Tasks of identifying and localizing specified spatio-temporal patterns in video, such as waving hands or picking up objects are tackled in crowd video event detection, where actions are matched with a predefined event template, which limits the

general use of event detection in real-world scenarios [22]. In addition, unusual event detection has also drawn much attention. To bridge the gap between machine-oriented low level features and human-friendly high level semantics, a number of works concern image captioning [15, 21, 36] and video concept detection [20, 9], where the task is to assign concept labels to an input image/video along with their associated probabilities. However, this introduces uncertainty and confidence problems [8] in sensor readings. Moreover, multi-modal sensor fusion has been well studied for combining multiple physical sensor modalities for various multimedia tasks [4]. Since the fusion of video and text is not well explored, we contribute to this part by providing a unified spatio-temporal data representation that can easily incorporate social media text analysis.

Event Detection using Social Sensors

Many online social network services are prevalent nowadays by which users share personal opinions, breaking news and interesting stories. Twitter, as one of the most important social sensors, has attracted a large number of works for event detection [12], topic discovery [18], as well as content analysis [26]. A news processing system, TwitterStand [32], is 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. Also, Twitter users are regarded as social sensors [31] in detecting and tracking earthquakes, typhoons or traffic jams. Twitter 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 [7]. [38] proposes a situation awareness algorithm to detect geospatial events in a given monitored geographic area, which offers good 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 situation in a holistic view. Twitris [33] 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, [35] 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. Besides, data abstraction and tools are designed to analyze spatio-temporal pattern of situation. Based on the "Emage", a media-processing approach and a declarative query mechanism are defined for the end user to query and understand large scale social situation. Though many of these previous works on social sensors contribute to event detection, they do not consider the fusing of information from physical sensors (e.g. surveillance cameras) which we argue would offer complementary and powerful understanding of evolving situations. Besides, the fusion of mobile computing and social networking services for traffic anomaly detection has been proposed in [28], where GPS data are analysed for abnormal traffic patterns, and social information collected from

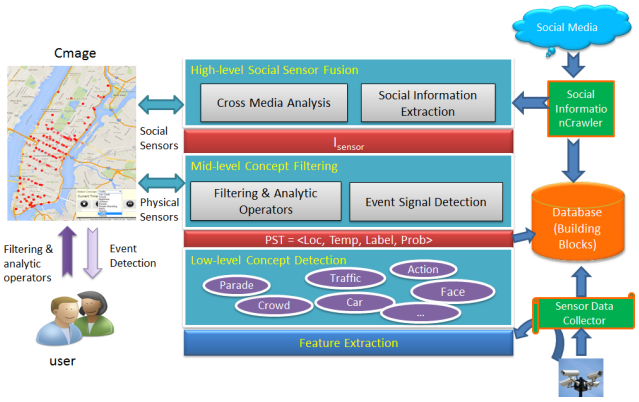


Figure 2: Overview of Proposed Multi-layer Tweeting Cameras Framework. In the first layer, low-level concepts are detected from raw sensor images through applying various concept detectors. A unified probabilistic spatio-temporal (PST) data structure represents the low-level concepts. Those camera "tweets" 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.

microblogs are analysed to offer high-level semantic explanations for the anomalies. In addition, crowd-sourced sensing and collaboration systems over Twitter are designed and implemented which opens publish-subscribe infrastructure for sensors to be combined with social platform [14].

Summing up, many works have been done towards event detection using either visual sensors or social sensors. However, they explicitly consider either physical sensors or social sensors and do not deal with fusing these two types of sources. To the best of our knowledge, there is no previous work that builds a multi-layer tweeting camera framework that fuses both physical and social information for event detection, which is the main contribution of this paper.

3. TWEETING CAMERAS FRAMEWORK

In this section, we describe the proposed multi-layer tweeting cameras framework as shown in Figure 2. The framework consists of three layers, namely low-level concept detection layer, mid-level concept filtering layer, and high-level social sensor fusion layer. 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 facilitate event detection. By aggregating multiple individual sensor feeds via a common representation, the framework provides a visualization interface as well as information filtering tools for users to gain global understanding of the events occurring by manipulating a set of predefined analytic functions. In the following, we detail how (where, when, what) camera tweets are analysed and combined with social media data in this framework.

3.1 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 con-

cept information represented by the unified data structure (defined in Section 3.2) is also stored in the database for querying and further analysis.

3.2 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 [20, 17], object detectors [16], face detectors [37] are examples of these advances. In the first layer, we incorporate a set of concept detectors to detect a variety of concepts 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 [20] which can detect 374 general concepts defined in LSCOM [23] 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 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 [39]. 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, 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 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 import for higher information filtering and are indexed in the repository (stored in MongoDB) for querying and pattern mining. Therefore, cameras whose feeds are analysed in the framework are constantly tweeting low-level concepts in the first layer and simultaneously pushing PST data in to database for indexing.

3.3 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.

3.3.1 Filtering Operators and Analytic Functions

A set of pre-defined filtering operators and analytic operators has been designed to analyse 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 on 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.

3.3.2 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 [27] 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".

3.4 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 on to 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, hash-tag, etc). We utilize the location and time information obtained from physical sensors to filter out non-concept 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 4). 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 summarise, 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.

4. PROCESSING FRAMEWORK

In this section, we elaborate on the probabilistic spatio-temporal data structure, the aggregation format Cmage, as well as the 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 utilised to complement each other for event detection.

4.1 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 [20] 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".

$$pst = [loc, temp, label, prob, pointer] \quad (1)$$

where

- $loc = [lat, lon]$ 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.
- *temp* 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.
- *prob* is the confidence value in [0,1] representing the output of a concept detector as a probability value.
- *pointer* points to 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.

4.2 Cmage

Extending the idea of "Emage" which represents aggregated social interest of users [35], 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 pixel intensity is the probability value of the concepts. 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.

Figure 3: *Crowdedness* Cmage of NUS Foodcourts at 12:00 on 24^{th} March 2014. 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 location.

4.3 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.

4.3.1 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.

$$\{T_1, T_2, \ldots, T_j\} = \tau_{task}\left(T, CAM_i, l_i\right), where : T_j \in T \quad (2)$$

4.3.2 Query Operators

Query operator 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 $P\_filter(exp)$:

$$\Theta_{P\_filter(exp)}(S) = \{(l_j^{CAM_i}, p_j^{CAM_i})^t | P\_filter(exp) = True\} \quad (3)$$

where P_filter are predicates on an element of PST data.

These filtering functions can retrieve from the PST stream by taking user-defined pst related expression as 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 "parade" at $5th$ Avenue with a confidence higher than 0.8:
Query: $\Theta_{P\_PROP \wedge P\_LAB \wedge P\_LOC \wedge P\_TEMP}(S)$
where: $P\_PROP = P\_prop(0.8 \leq p)$
$P\_LAB = P\_lab(label = traffic \vee parade)$
$P\_LOC = P\_loc(CAM_i) = 5^{th} Avenue)$
$P\_TEMP = P\_temp(t = March\ 17^{th})$

4.4 Data Analysis Functions

As the PST element is a numeric data presentation with spatial, temporal and symbolic information about going-on

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.

4.4.1 Statistical functions

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 $g_t^{\{l_1,l_2,\ldots,l_k\}}$ 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 spatial region or with close data in 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 [2]. 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 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`

Abnormal data pattern is regarded as important information that deserves an alert for the tweeting camera system. A $\mathbb{S}_{outlier}(D)$ 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 abnormal pattern in terms of σ.

Example: show the time when the crowd concept has an abnormal intensity during a particular period.

$\mathbb{S}_{outlier}(D,\sigma)(\Theta_{P\_lab(crowd)\wedge P\_tem(t_1\leq t\leq t_2)}(S))$

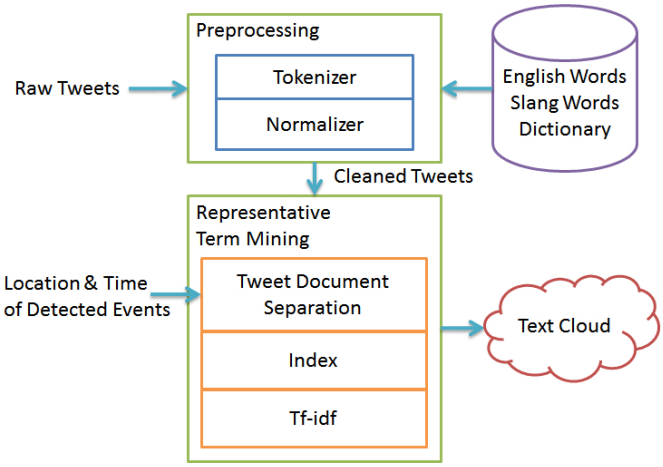


Figure 4: Architecture of Twitter Data Processing.

4.4.2 Data Mining with PST

Computing spatial clusters/segments helps in better characterizing the situation across regions [3]. 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.

4.4.3 Density Function

The 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 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 a intuitive information about the situation.

Example: calculate the number of cameras that detected "person" concept between time t_1 and t_2.

$\mathbb{C}_{CAM}(\Theta_{P\_lab(person)\wedge P\_pro(p=1)\wedge P\_tem(t_1\leq t\leq t_2)}(S))$

4.5 Social Information Fusion

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 4.

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 4 adopted from [28] is used to calculate the weight

of extracted terms that could best describe the event.

$$w_{term} = tf(term, T_C) \times idf(term, T_H)$$
$$s.t. \begin{cases} tf(term, T_C) = \dfrac{f(term, T_C)}{\max\{f(w, T_C), \forall w \in T_C\}} \\ idf(term, T_H) = \log \dfrac{|T_H|}{|\{th \in T_H : term \in th\}|} \end{cases} \quad (4)$$

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 weight 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 5 to fuse them to derive final event signal intensity.

$$Es(e) = w_{se} * I_{se}(e) + w_{so} * I_{so}(e) \quad (5)$$

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

4.6 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.

5. EXPERIMENTS

5.1 Datasets

5.1.1 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.

5.1.2 NUS Foodcourt CCTV Camera

The NUS (National University of Singapore) foodcourt video dataset consists of feeds from 73 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.

5.1.3 Twitter Data

We have crawled tweets using Twitter Streaming API from October 08, 2014 to November 2, 2014, with the geographic bounding box of [40.698770, -74.021248, 40.872932, -73.905459] which includes Manhattan, and collected a total of 1,510,025 records. Each record is stored in the database with original set containing all data fields such as time of created, geo-location, text etc.

5.2 Evaluation Approach

In this study, we analyse 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.

5.2.1 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 1.

Table 1: 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

5.2.2 Measurement

To demonstrate the effectiveness of the framework, we consider the detection rate of each event listed in Table 1. As per the signal detection theory, the threshold for the corresponding concepts is evaluated in terms of detection rate.

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.

5.3.1 Event Detection based on SDT

In order to model the noise and real signal of each camera, we consider concept confidence values as the signals for detection. We use noise and signal to model concept output values of non-event and event respectively and automatically select an optimal detection threshold. Note that the distribution is only valid between 0 and 1 since the confidence value is a probability value. Figure 5 shows the distribution of "parade" signal from camera in 5th Avenue 57 Street.

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

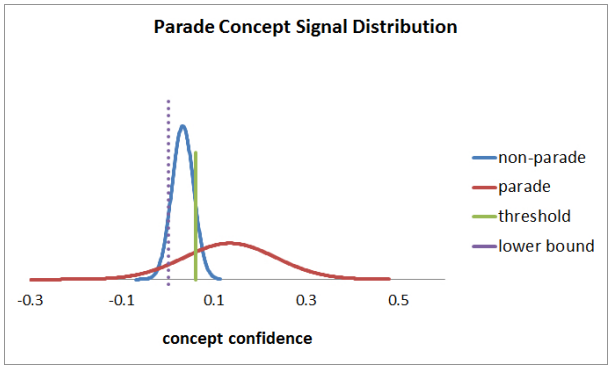


Figure 5: "Parade" Signal Distribution in 5 Ave @ 57 Street.

The distribution is determined by calculating mean and standard deviation of concept results from images captured on October 13, 2014, from 2pm 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 analysed to determine an optimal threshold for a particular concept detector for a camera. Note that since different cameras usually capture different scenes, optimal threshold of a particular detector are 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 6.

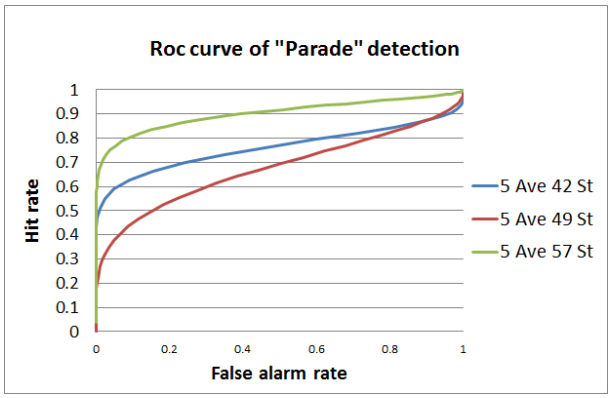


Figure 6: ROC Curve of "Parade" Signal in Three Locations.

This figure shows ROC curve of the parade detector for three different cameras. The threshold is set to the value that maximizes the hit rate as well as minimizes the false alarm rate. Therefore, the cusp point in the ROC curve i.e. the point that minimises 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 of concept detectors is set, it is fixed and applied on images to automatically trigger event alerts. However, the user can also manually reconfigure this threshold if required. Using this threshold, we examine the detection performance of two parade events in terms of f1-score; the analysis is shown in Figure 7 and 8 for two cameras.

We use the threshold to analyse the results of both "Columbus Day Parade" and "Hispanic Parade" event, and compare our thresholding results with baseline which is given by the detectors in the label field based on a fixed value 0.5 as

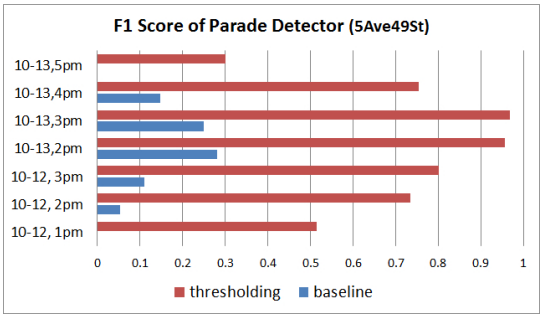


Figure 7: F1 Score for Camera in 5 Avenue at 49 Street

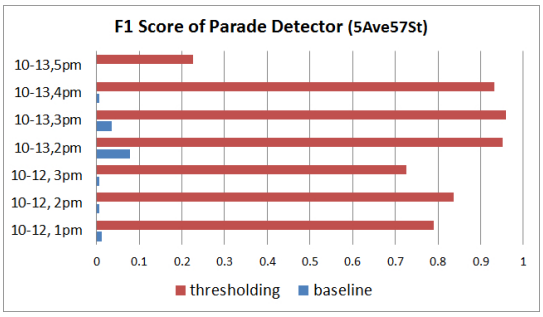


Figure 8: F1 Score for Camera in 5 Avenue at 57 Street

threshold. As can be seen, having an adaptive threshold significantly improves the performance.

5.3.2 *Applying Analytic Functions*

Once concepts' confidence is obtained for 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.

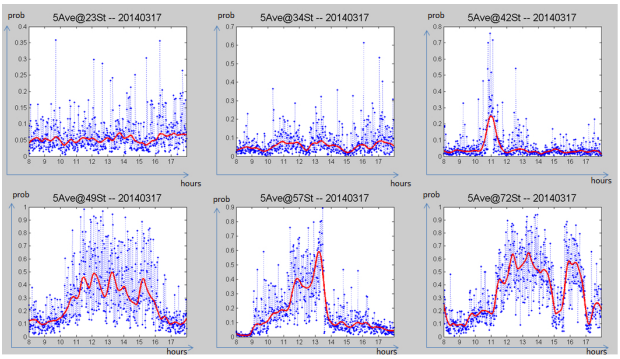


Figure 9: *People Marching* Concept Results from 8:00 to 18:00 in March 17th, during Saint Patrick's Day Parade Event

Figure 9 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], higher value representing higher intensity of crowdedness. Therefore, we are able to convert the image to the unified probabilistic spatio-temporal data point (loc_{cam_A}, t, "*crowd*", *probability*), where the location is the canteen having the camera. Therefore, the cameras would tweet crowd information twice every minute.

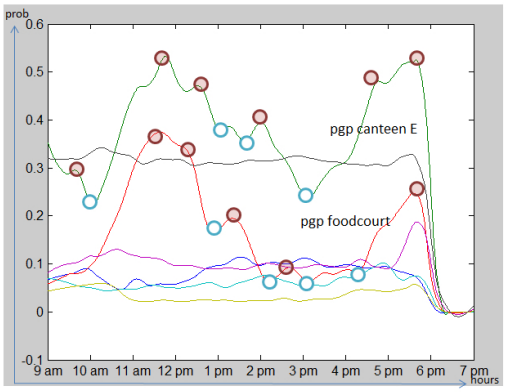


Figure 10: The Crowd Extremes at 7 Foodcourts on Sunday.

Figure 10 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 idea of where and when to go for lunch and dinner.

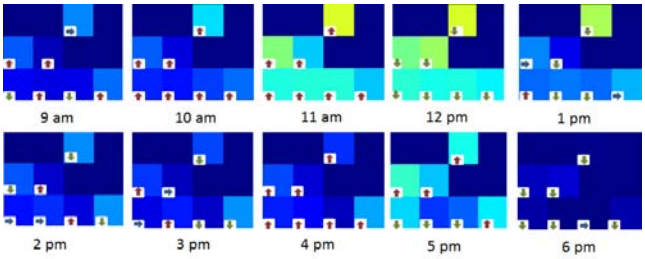
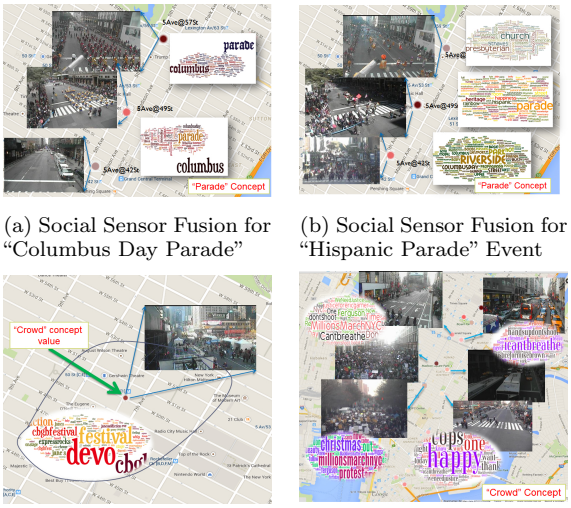


Figure 11: A Campus Foodcourt Cmage *Crowdness* with Arrows indicating Ascending, Descending and Plateau Trend.

In addition, we show the crowd density Cmage trend (with labels) of campus canteens from 9 am to 6 pm on 21th March 2014 in Figure 11. 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.

5.3.3 *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 1pm to 5pm 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 12.



(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 12: 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 high-level semantic meaning of the event, e.g. Figure 12 (a) and (b) confirm the events are indeed the "Columbus Day Parade" and "Hispanic Parade" events respectively. (c) offers details of 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.

Table 2 shows the comparison between our approach of social information mining with baseline in terms of number of tweets. As represented, our framework utilize 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 number of geo-tagged tweets crawled before and during the event respectively. Our approach TH and TC are the number of geo-tagged tweets crawled *around the event location* before and during the event respectively.

Table 2: 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 13, according to the equation defined in section 4.5 the final parade event signal intensity is im-

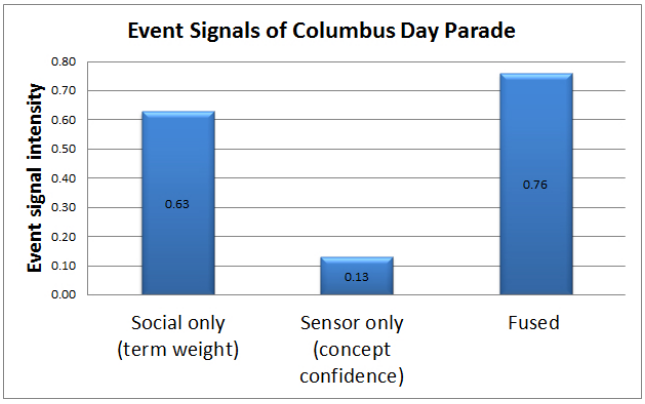


Figure 13: Comparison among Fused, Physical and Social Sensor Results of Detecting the "Columbus Day Parade" Event

proved 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 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 user the ability to analyse 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.

6. DISCUSSION

Our experiments have verified the significant correspondence between physical sensors data and virtual social information. While the vocabulary of camera tweets is limited in our experiments, it can be scaled up. A more comprehensive set of concept detectors including face detectors and image caption labelers can be used in the first layer. 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. Without loss of generality, the framework is also able to detect unplanned events as long as the events emit visual and social signals. The occurrence of an unplanned event will then lead to changes in the observed signal levels, when compared to the times when there are no events occurring. The unified data representation for visual sensors and analytic operators allows users to flexibly specify events characteristic for detection. Thus outlier detection can potentially help detect unplanned events. For an unplanned (or planned) event that doesn't include visual information (e.g. video) the framework will have to rely on text processing only. While this can still be useful, it will not demonstrate the full power of the proposed approach. One of 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. Since we are exploring sensor data concerning real world situations, it is worth considering the degree of privacy loss for the people who happen to appear in the camera images. To prevent any breach of privacy, we currently use images of low resolution so that people's faces are not easily recognizable. For a more comprehensive approach towards privacy, the identity leakage model [29] could be used to calculate the privacy loss due to the face identity as well as the various side-channels. Furthermore, adaptive robust privacy protection methods [30] could be incorporated into our proposed framework.

7. CONCLUSIONS

In this work, 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[2] including code, data, and relsults.

For future work, we plan to integrate information from social media sensors, which include trend analysis and topic mining, to improve the quality of the camera tweets. In addition, we also would explore the possibility of creating an interactive framework that allows for the integration of other types of sensors into the framework. Finally, machine learning based approaches can be employed to learn the correlation between physical sensor tweets and social media tweets.

8. ACKNOWLEDGEMENT

This research was conducted at the NUS-ZJU SeSaMe Centre. It is supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO. We thank the Ambient Intelligence Laboratory, NUS for sharing the NUS foodcourts camera data and associated analysis. We also would like to thank our SeSaMe colleague Dr. Christian Von Der Weth for his insightful comments and careful proof-reading.

9. REFERENCES

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *In PAMI*, 30(3):555–560, 2008.

[2] T. Althoff, D. Borth, J. Hees, and A. Dengel. Analysis and forecasting of trending topics in online media streams. In *ACM Multimedia*, pages 907–916, 2013.

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *In PAMI*, 33(5):898–916, 2011.

[4] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli. Multimodal fusion for multimedia

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

analysis: a survey. *Multimedia Systems*, 16(6):345–379, 2010.

[5] P. K. Atrey, M. S. Kankanhalli, and J. B. Oommen. Goal-oriented optimal subset selection of correlated multimedia streams. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1), 2007.

[6] R. Babari, N. Hautiere, E. Dumont, N. Paparoditis, and J. Misener. Visibility monitoring using conventional roadside cameras - emerging applications. *Transportation Research Part C: Emerging Technologies*, 22(0):17 – 28, 2012.

[7] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain*, 2011.

[8] C. Bhatt and M. Kankanhalli. Probabilistic temporal multimedia data mining. *ACM Trans. Intell. Syst. Technol.*, 2(2):17:1–17:19, 2011.

[9] C. A. Bhatt, P. K. Atrey, and M. S. Kankanhalli. A reward-and-punishment-based approach for concept detection using adaptive ontology rules. *ACM Trans. Multimedia Comput. Commun. Appl.*, 9(2):10:1–10:21, 2013.

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

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

[12] M. Cataldi, L. Di Caro, and C. 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.

[13] H. Cramer and S. Buttner. Things that tweet, check-in and are befriended. two explorations on robotics amp; social media. In *HRI*, pages 125–126, 2011.

[14] M. Demirbas, M. Bayir, C. Akcora, Y. Yilmaz, and H. Ferhatosmanoglu. Crowd-sourced sensing and collaboration using twitter. In *WoWMoM*, pages 1–9, 2010.

[15] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *PAMI*, 35(8):1915–1929, Aug 2013.

[16] R. B. Girshick, P. F. Felzenszwalb, and D. A. McAllester. Object detection with grammar models. In *Advances in Neural Information Processing Systems*, pages 442–450, 2011.

[17] A. Habibian, K. E. van de Sande, and C. G. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, pages 89–96, 2013.

[18] L. Hong, A. Ahmed, S. Gurumurthy, and et al. Discovering geographical topics in the twitter stream. In *WWW*, pages 769–778, 2012.

[19] N. Jacobs and et al. The global network of outdoor webcams: Properties and applications. In *ACM SIGSPATIAL*, pages 111–120, 2009.

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

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

[22] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, pages 1–8, 2007.

[23] L. Kennedy and A. Hauptmann. Lscom lexicon definitions and annotations (version 1.0). 2006.

[24] M. Kranz, L. Roalter, and F. 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.

[25] P. Kulkarni, D. Ganesan, P. Shenoy, and Q. Lu. Senseye: A multi-tier camera sensor network. In *ACM Multimedia*, pages 229–238, 2005.

[26] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *WWW*, pages 251–260, 2012.

[27] D. Los Angeles Thomas et al. *Elementary signal detection theory*. Oxford University Press, 2001.

[28] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *ACM SIGSPATIAL*, pages 344–353, 2013.

[29] M. Saini, P. Atrey, S. Mehrotra, and M. Kankanhalli. W3-privacy: understanding what, when, and where inference channels in multi-camera surveillance video. *Multimedia Tools and Applications*, 68(1):135–158, 2014.

[30] M. Saini, P. K. Atrey, S. Mehrotra, and M. Kankanhalli. Adaptive transformation for robust privacy protection in video surveillance. *Advances in Multimedia*, page 14, 2014.

[31] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[32] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: News in tweets. In *ACM SIGSPATIAL*, pages 42–51, 2009.

[33] A. Sheth, A. Jadhav, P. Kapanipathi, and et al. Twitris: A system for collective social intelligence. In *Encyclopedia of Social Network Analysis and Mining*, pages 2240–2253. 2014.

[34] J. Shin and et al. Asap: A camera sensor network for situation awareness. *Principles of Distributed Systems*, 4878:31–47, 2007.

[35] V. K. Singh, M. Gao, and R. Jain. Social pixels: Genesis and evaluation. In *ACM Multimedia*, pages 481–490, 2010.

[36] O. Vinyals and et al. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[37] P. Viola and M. Jones. Robust real-time face detection. *In IJCV*, 57(2):137–154, 2004.

[38] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *European Conference on Advances in Information Retrieval*, pages 356–367, 2013.

[39] U. Westermann and R. Jain. Toward a common event model for multimedia applications. *IEEE MultiMedia*, 14(1):19–29, Jan. 2007.