Mobile Sensing Enabled Robust Detection of Security Threats in Urban Environments

Jie Yang[1], Jerry Cheng[2], and Yingying Chen[1]

[1] Department of Electrical and Computer Engineering
Stevens Institute of Technology
Castle Point On Hudson, Hoboken, NJ, 07030, USA
{jyang,yingying.chen}@stevens.edu
[2] Department of Statistics
Columbia University
1255 Amsterdam Ave., New York, NY, 10027, USA
jcheng@stat.columbia.edu

Abstract. Mobile sensing enables data collection from large numbers of participants in ways that previously were not possible. In particular, by affixing a sensory device to a mobile device, such as smartphone or vehicle, mobile sensing provides the opportunity to not only collect dynamic information from environments but also detect the environmental hazards. In this paper, we propose a mobile sensing wireless network for surveillance of security threats in urban environments, e.g., environmental pollution sources or nuclear radiation materials. We formulate the security threats detection as a significant cluster detection problem. To make our approach robust to unreliable sensing data, we propose an algorithm based on the Mean Shift method to identify the significant clusters and determine the locations of threats. Extensive simulation studies are conducted to evaluate the effectiveness of the proposed detection algorithm.

Key words: Mobile sensing, security threats, Mean Shift Clustering

1 Introduction

Mobile sensing has become increasingly popular in recent years as it enables data collection from large number of participants in ways that previously were not possible [1]. In particular, by affixing sensory devices to mobile devices, such as smartphones or vehicles, mobile sensing provides an opportunity to not only collect and share dynamic information at an urban-scale but also perform data analysis to detect security threats presenting in urban environments. In this work, we consider security threats as environmental sources hazardous to civilian's daily lives, for example, air pollution sources and nuclear radiation sources.

We show the importance of detecting the hazardous environmental sources by using the following two examples. The environmental air pollution directly affects the public health in a metropolitan area. According to reports of the

QSHINE 2010, November 17-19, Houston, United States

DOI 10.4108/eai.17-11-2010.2245428

World Health Organization, over 4.6 million people are estimated to die annually from the direct impact of air pollution - more than those from car accidents every year [2]. Meanwhile radiological sources have become commonplace at research, industrial and medical facilities. As a result, we are facing a growing danger that nuclear materials might be acquired by terrorist organizations. For example, there are numerous cases of lost or stolen radioactive materials have been reported [3, 4]. Moveover, major metropolitan areas are attractive targets for placement of radiation materials by attackers due to their dense population and economic importance. Previous detection of such security threats are carried out mostly by specialized teams or fixed sensors, and may not perform frequently in a large-scale due to the limited resources and high cost involved. However, because of their growing danger and severe harm, detecting these security threats through daily activities is not only desirable but also feasible by the increasing popularity of mobile sensors. In this paper, we develop a mechanism that exploits sensing data obtained from mobile sensing to detect such security threats.

In our framework, real time sensor readings together with participants' locations will be reported back through existing wireless networks to a central surveillance center, where the data is analyzed to detect and localize the security threats sources. The ubiquitous nature of smartphones, vehicles and other portable electronic devices makes it possible to constantly sense threats sources. However, there are a number of underlying challenges to be overcome before the accurate detection can occur. For example, sensitivity of a sensor is diminishing with distance from the source; a mobile sensor may fail to correctly sense an existing threat (causing false negative), or report a reading when there is none threat source (causing false positive). In addition, the background noise may vary significantly from one region to another. For instance, a person who just came out from a radioactive therapy or a bag of cat litter may even set off false alert of a mobile sensor.

To enable robust detection of security threat sources under the presence of unreliable sensor readings, we propose to detect these sources using sensor readings collaboratively. In our work, the security threats source detection is formulated as significant cluster detection problem, in which we aim to identify whether one or more spatial clusters exist in the area significantly from background noise. When a source emits energy at a certain location, the mobile sensors within the detection range of the source would have higher probability to be activated, hence create a cluster of points in the physical space, while the background false positive readings would be likely randomly distributed throughout the area. We develop a detection mechanism grounded on Mean Shift Clustering procedure to both detect the significant clusters and localize the threat sources.

To validate the effectiveness of the proposed security threat detection mechanism, we simulated a mobile sensing wireless network in an area of similar size as metropolitan Manhattan in New York City. Our simulation results show that our proposed framework can achieve over 90% detection rate for both single and multiple sources with small localization errors. Thereby this shows that our approach is effective in detecting treat sources.

The rest of the paper is organized as follows. We place our work in the context of related research in Section 2. The system overview of the mobile sensing enabled security threat detection is presented in Section 3. We detail our detection mechanism by focusing on the Mean Shift procedure in Section 4 and describing robust threat source localization in Section 5. We then present our simulation results in Section 6 and conclude our work in Section 7.

2 Related Work

Although mobile sensing has been gaining popularity in various applications [5–7], there is relatively less work focusing on detecting security threat sources. The PEIR project [7] proposed by UCLA is prototyped to measure how often people are exposed to high levels of air pollution by using sensing data collected by mobile phones. As a part of the project of Use of Mobile Technology for Social Change [1], taxis equipped with a carbon monoxide sensor and a GPS device are used to collect environmental data such as readings of carbon monoxide levels, which is then displayed on a city map for visulization. However, none of these work performed systematic data analysis to detect the presence of security threat sources.

In the area of radiation detection, using individual and fixed sensors to detect radiation sources have been well studied [8, 9]. These approaches have limited application to identify portable nuclear radiation sources placed by attackers. The recent progress has shown some success of using a network of sensors for detecting and tracking radiation sources [10–13]. In [11, 12], a linear arrangement of fixed detectors has been considered to detect the radioactive sources. In [13], a latent model is proposed to detect multiple nuclear materials by using a mobile sensor network approach. In addition, the Radiation Laboratory at Purdue University used a network of cell phones with GPS capabilities to detect and track radiation [14]. The difference between our work to previous studies is that we propose to use mobile sensing that enables an extensive coverage in a metropolitan area to detect and localize the security threats sources with the presence of unreliable sensor readings.

Our approach is based on Mean Shift Clustering method. Traditional statistical methods for detecting a cluster of events in spatial data is using a class of Scan Statistics [15–17]. The most commonly used scan statistics is the maximum number of cases within a fixed size window that scans through the study area. Based on this scan statistics, a generalized likelihood ratio test has been developed to test the null hypothesis whether all the information are uniformly distributed in the area (no cluster). Scan statistics procedures are mainly used in detecting a single significant cluster, and they also have had some success in detecting multiple clusters of fixed sizes.

However, problems arise for detecting multiple clusters of varying sizes. In recent years, procedures have been developed to overcome the difficulty. One of the well known approaches is a stepwise regression model combined with model selection procedures to locate and determine the number of clusters [18]. These

approaches rely on a weighted least square formulation, although the response variable (gaps between incidents) is typically non-Gaussian. In our approach, we utilized Mean Shift Clustering algorithm, a nonparametric clustering technique, does not assume any prior knowledge of the number of clusters, and can handle arbitrarily shaped clusters. Thus, it is suitable for handling clusters of arbitrary shape and number for detecting security threats using mobile sensing data.

3 System Overview

In this section, we first present the threat model that describes the transmission behavior of the threat source. We then provide our network model and sensing model enabled by mobile sensors. Finally, we overview the key components of our threat source detection mechanism.

3.1 Threat Model

We consider security threats in urban environments that can cause biological hazards to civilians, for example, environmental pollution sources or nuclear radiation materials. A threat source can be either static or mobile. For simplicity, we start with assuming the impact of the threat source travels in spherical waves. Thus, the impact intensity T decreases by the inverse square of the distance r: $T(r) = c/r^2$, where the constant c is a factor related to the energy of the source. This simple model is also used to describe nuclear radiation emitting from a threat source [19].

In addition, there might be multiple threat sources, whose impact region may overlap. In this work, we assume we know the number of threat sources and consider the sources of the same type, e.g., environmental pollution sources that can trigger the reading of mobile sensors measuring pollution level; or nuclear radiation materials that can trigger the reading of mobile sensors measuring nuclear radiation. When these sources assume the same energy spectrum, the overall impact intensity at a particular location is an aggregation of individual ones: $T_{total} = \sum_{i=1}^{N} c_i/r_i^2$ where N is the total number of threat sources of the same type. The sources of different types will be considered in our future work.

3.2 Network Model

We build our solution of detecting threat sources in urban environments upon the wireless mobile sensor networks with the following characteristics.

Mobility. Mobile sensors can be either mounted or built-in on smart phones or vehicles. This enables the mobile sensors to move randomly or in some routine patterns in urban cities. For example, the participating vehicles can be taxicabs, police patrol vehicles, or city buses.

Location-Aware. Each mobile sensor knows its physical location at all times during moving. This is a reasonable assumption as most of wireless devices (e.g., mobile phones or vehicles) are equipped with GPS or some other approximate but less burdensome localization algorithms [20].

Sensing Model. When mobile sensors move within a certain range of a threat source, the reading of the sensors will be triggered. We define the reading of a sensor S using a threshold model: $S = \mathbf{1}_{\{T(r)\geq\tau\}} = \mathbf{1}_{\{c/r^2\geq\tau\}}$ where τ is a threshold for detection and $\mathbf{1}_{\{.\}}$ is the indicator function. That is, if the intensity $T(r)$ at the sensor location is greater than the threshold τ, the sensor reports a positive reading (i.e., detection of a threat source), otherwise the sensor reports a negative reading. In the case of multiple threat sources of the same type, the threshold model can be represented as: $S = \mathbf{1}_{\{T_{total}\geq\tau\}} = \mathbf{1}_{\{\sum_{i=1}^{N} c_i/r_i^2\geq\tau\}}$. Moreover, as with any sensing device, the reading of mobile sensors may not be 100% accurate. Additionally, transient scenarios, e.g., a person who is walking on the street just went through a radioactive therapy, may also trigger false alarms. In this work, we do not explore how to improve the reliability of sensor readings. Instead, to live with unreliable sensor readings, we focus our work on design robust detection mechanisms with readings from multiple sensors to collaboratively detect the presence of the threat sources.

3.3 Mobile Sensing Enabled Threat Detection

Our framework employs a mobile sensing approach to detect security threats in urban environments. The readings of mobile sensors (e.g., equipped with smartphones or mounted on vehicles) will be reported to a central monitoring server along with the positions of the sensors when readings occur. The report of sensor readings will utilize the existing wireless infrastructure by sending the data over a cellular uplink or making use of the WiFi connections depending on cost/delay tradeoffs.

The detection mechanism running at the back-end server comprises three main components: *Mean Shift Procedure*, *Building Clustering Hierarchy*, and *Threat Source Identification*. The Mean Shift procedure is the core element during the detection process. It finds the clustering of the densest area in the sensing data through mode finding of a density function. The accuracy of modes detection relies on appropriate bandwidth selection in Mean Shift clustering. Since the impact range of a threat source is unknown, it is a challenging task to choose an appropriate bandwidth. To find the optimal bandwidth, we develop a technique that produces a hierarchy of clusters by repeatedly applying the Mean Shift procedure under different bandwidths. Finally, threat sources are identified by choosing the clusters with top number of sensor readings under the optimal bandwidth. This approach has the main advantage of filtering out the clusters that are formed by unreliable sensor readings and makes the detection robust. We describe these components in detail in the following sections.

4 Threat Source Detection Using Mean Shift

The Mean Shift procedure is a key component in our threat source detection framework. Detecting a threat source is equivalent to the mode detection in the Mean Shift procedure. There have been many applications using Mean Shift such as image analysis [21], texture segmentation [22], objective tracking [23,24] and data fusion [25]. The Mean Shift Clustering is a nonparametric clustering technique using the mode finding Mean Shift procedure, which is ideal to identify the presence of threat sources based on the readings of mobile sensors together with spatial information. Because it does not assume any prior knowledge of the number of clusters, and can handle arbitrarily shaped clusters. In particular, Mean Shift is a procedure for locating the maxima of a density function given discrete data sampled from that function [26]. We next detail the kernel density estimation to derive the density estimator and the density gradient estimation to find the modes for detecting threat sources based on the description and notation in [27].

4.1 Kernel Density Estimation

Kernel density estimation is a non-parametric way of estimating the probability density function of a random variable [28, 29]. Let $\mathbf{x}_i \in R^d, i = 1, 2, .., n$, be n independent and identically distributed d-dimensional data points. The multivariate kernel density estimator with kernel $K(\mathbf{x})$ and a symmetric positive definite $d \times d$ bandwidth matrix $\mathbf{H}$ is given by:

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i), \tag{1}$$

where

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-\frac{1}{2}} K(\mathbf{H}^{-\frac{1}{2}}\mathbf{x}). \tag{2}$$

The d-variate kernel $K(\mathbf{x})$ is a bounded function with compact support satisfying the following properties:

$$\int_{R^d} K(\mathbf{x}) d\mathbf{x} = 1; \int_{R^d} \mathbf{x} K(\mathbf{x}) d\mathbf{x} = 0;$$

$$\lim_{\|\mathbf{x}\| \to \infty} \|\mathbf{x}\|^d K(\mathbf{x}) = 0; \int_{R^d} \mathbf{x}\mathbf{x}^T K(\mathbf{x}) d\mathbf{x} = c_K I.$$

where c_K is a constant.

In practice, the bandwidth matrix $\mathbf{H}$ is chosen proportional to the identity matrix $\mathbf{H} = h^2\mathbf{I}$, with $h > 0$. Employing the bandwidth parameter h, the kernel density estimator (1) becomes

$$\hat{f}_h(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right), \tag{3}$$

where $K(\mathbf{x})$ is a class of radially symmetric kernels satisfying:

$$K(\mathbf{x}) = c_{k,d}k(\|\mathbf{x}\|^2). \tag{4}$$

The function $k(x)$ is called the profile of the kernel and $c_{k,d}$ is the normalization constant, which makes $K(\mathbf{x})$ integrate to one.

The commonly used kernels in literature are Epanechnikov Kernel, Uniform Kernel and Gaussian Kernel. In this work, we used Uniform Kernel whose profile is given by:

$$\text{Uniform Kernel:} \qquad k_U(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & x > 1 \end{cases} \tag{5}$$

Using the profile notation, the density estimator (3) can be written as

$$\hat{f}_{h,K}(\mathbf{x}) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} k\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right). \tag{6}$$

The modes of the density function $f(\mathbf{x})$ are located at the zeros of the gradient function $\nabla f(\mathbf{x})$. The Mean Shift procedure can locate these modes without estimating the underlying density.

4.2 Density Gradient Estimation

The gradient of the density estimator (6) is

$$\nabla \hat{f}_{h,K}(\mathbf{x}) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_i) k'\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right), \tag{7}$$

where $k'(x)$ is the derivative of the profile.

If we define $g(x) = -k'(x)$, and introducing $g(x)$ into (7), we have

$$\begin{aligned} \nabla \hat{f}_{h,K}(\mathbf{x}) &= \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}) g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \\ &= \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \right] \left[\frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right)} - \mathbf{x} \right]. \end{aligned} \tag{8}$$

The first term of the above equation is proportional to the density estimate at $\mathbf{x}$ computed with kernel $G(\mathbf{x}) = c_{g,d}g(\|\mathbf{x}\|^2)$

$$\hat{f}_{h,G}(\mathbf{x}) = \frac{c_{g,d}}{nh^d} \sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right). \tag{9}$$

The second term

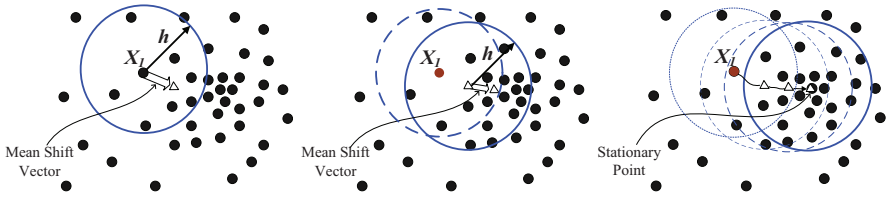

Fig. 1. Illustration of Mean Shift procedure using a uniform kernel. (a) Start from a data point $\mathbf{x}_1$, calculate the Mean Shift vector $\mathbf{m}_{h,G}(\mathbf{x}_1)$, which is the difference between $\mathbf{x}_1$ and weighted mean of data points within the window with a radius of bandwidth h. (b) Shift the window by $\mathbf{m}_{h,G}(\mathbf{x})$, recompute the Mean Shift vector $\mathbf{m}_{h,G}(\mathbf{x}_1)$. (c) Reach the stationary point in the densest area, which is the mode.

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x} \tag{10}$$

is the *Mean Shift* which is the difference between the weighted mean and $\mathbf{x}$, the center of the kernel window. The Mean Shift vector always points toward the direction of the maximum increase in the density. Therefore, the Mean Shift procedure is guaranteed to converge to a point where the gradient of density function is zero. This point is the mode (i.e., the security threat source) obtained from running Mean Shift by going through the following steps iteratively:

1. Compute the Mean Shift vector $\mathbf{m}_{h,G}(\mathbf{x})$,
2. Translate the kernel window of $G(\mathbf{x})$ by $\mathbf{m}_{h,G}(\mathbf{x})$, and re-compute the weighted mean,
3. Stop iteration if gradient is close to zero.

4.3 Cluster Formulation

For each reported reading of a mobile sensor, the Mean Shift procedure under a fixed bandwidth will find the mode (i.e. the point where the gradient of density function is zero) in the density function that the sensor reading is associated with. The mode of the density function is equivalent to the location of the densest area of the data set of all the reported sensor readings, i.e., the location of the security threat source. To form a cluster, all the sensor readings associated with the same mode belong to the same cluster. And the modes could be used to represent the cluster center. Alternately, we can define the *basin of attraction* of the mode as the set of all mobile sensor readings that converge to the same mode [27]. The sensor readings which are in the same basin of attraction is associated with the same cluster. The number of clusters is obtained by the number of modes. Figure 1 illustrates the cluster formulation for detecting security threat sources. The detailed algorithm flow is described in Algorithm 1. ClusterCenters[] is a one dimensional matrix that holds the clusters formulated under a specific fixed bandwidth.

Algorithm 1 Threat Source Detection Using Mean Shift Clustering under a Fixed Bandwidth

Let $S = (s_1, .., s_n)$ be the sensor reading dataset, h be the bandwidth;
Let ClusterCenters be the modes;
Let ClusterMembership be the cluster membership of each point;
ClusterCenters=[]; ClusterMembership=[];
$i \leftarrow 0$;
$j \leftarrow 0$;
repeat
  For the data point s_i, run Mean Shift procedure to get the mode M ;
  if (M is not in the ClusterCenters) **then**
    $j \leftarrow j + 1$
    ClusterCenters[j] = M;
    ClusterMembership[i] = j;
  else
    if (M is at the lth position of the ClusterCenters) **then**
      ClusterMembership[i] = l;
    end if
  end if
  $i \leftarrow i + 1$;
until i==n
return ClusterCenters;
return ClusterMembership;

5 Robust Threat Source Localization

5.1 Challenges on Bandwidth Selection

The bandwidth selection in Mean Shift directly affects the performance of the Mean Shift Clustering, and consequently affects the accuracy of detecting security threat sources. Since the impact range of the threat sources is unknown, it is hard, if not possible, to identify the significant clusters with the optimal bandwidth during Mean Shift Clustering. If the bandwidth h is estimated too large, it will produce an oversmoothed density estimate, resulting in only one mode in the estimated density even if there are multiple treat sources present. On the other hand, a too small bandwidth h will cause the density estimate to produce too many clusters, which is called undersmoothed. To address these challenges, we develop a technique that perform threat source localization by using adaptive bandwidth. Our technique is described as follows.

5.2 Building the Cluster Hierarchy

Researchers summarized four different techniques for bandwidth selection [27] including plug-in rule [30], Least Squares Cross Validation [31] and stability of the decomposition [32]. In order to achieve robust detection of security threat sources, we designed a technique to find the optimal bandwidth with significant clusters by building a hierarchy of clusters. The repeated clustering of the

Algorithm 2 Building the Cluster Hierarchy

1: Let $S = (s_1, .., s_n)$ be the dataset, h be the bandwidth and p be the bandwidth step;
2: Set h= minimum non-zero distance between any two points in S;
3: $i \leftarrow 0$;
4: ClusterCenters{i}=S; ClusterMembership{i}=$1 : n$;
5: **repeat**
6: (newClusterCenters, newClusterMembership) = MeanShiftClustering(S, h);
7: ClusterCenters{$i + 1$} $\leftarrow$ newClusterCenters;
8: ClusterMembership{$i + 1$} $\leftarrow$ newClusterMembership;
9: $i \leftarrow i + 1$;
10: $h \leftarrow h + p$;
11: **until** size(ClusterCenters{i})==1
12: return ClusterCenters;
13: return ClusterMembership;

data is accomplished by iteratively running Mean Shift with increasingly larger bandwidths on top of the mobile sensing data. This iterative procedure is closely related to the stability of decomposition [32] for a density shape estimate. The optimal bandwidth is taken as the center of the largest operating range of the bandwidth over which the same cluster set is obtained for the given sensing data. This means that the shapes of the estimated densities are unchanged over this operating range of the bandwidth. This technique can yield a suitable bandwidth estimate by finding all cluster centers, i.e., threat sources, over the chosen operating range.

In particular, for the initial bandwidth value h_{min} (i.e. the smallest bandwidth value), we use the minimum non-zero distance between any two points in the reported mobile sensing data set. This is the lowest level of the hierarchy consisting of individual sensing data points. In this case, each individual data point either remains as its own cluster of size one or it is merged into a larger cluster when mobile sensors have same readings (e.g., two people walked together side by side or two vehicles passed the same location within a very short time period). In the subsequent iteration of Mean Shift, the bandwidth is increased by the step p, which is the bandwidth increment. With the increased bandwidth, Mean Shift is applied to the original data points to produce a new set of clusters. We continue this process until there is only one cluster left. This is the algorithm stop criteria that we employ in our technique of building cluster hierarchy. This criteria means that the bandwidth is increased large enough so that no new cluster can be generated beyond this cluster hierarchy. Intuitively, we are repeatedly blurring the data by using a larger bandwidth. Algorithm 2 shows the procedure to build the cluster hierarchy. ClusterCenters{} is a two dimensional matrix. ClusterCenters{i} holds the clusters formulated in the ith iteration. The constructed cluster hierarchy will be used in the next section to perform multiple threat source localization.

Algorithm 3 Threat Source Localization Using Adaptive Bandwidth

1: Build cluster hierarchy using Algorithm 2;
2: Set L be the threats source candidate list;
3: Let k be the number of the threats source;
4: $i \leftarrow 0$;
5: **repeat**
6: For each cluster C in ClusterCenters$\{i\}$, calculate the number of data points in C from ClusterMembership$\{i\}$;
7: Add the top clusters with top k number of points into the candidate list L;
8: $i \leftarrow i + 1$;
9: **until** size(ClusterCenters$\{i\}$)==k
10: For each cluster C in the candidate list L, calculate the lifetime of C;
11: Return the cluster centers with top k lifetime,

5.3 Localizing Multiple Threat Sources Using Adaptive Bandwidth

The challenging task is to identify the significant clusters, i.e., multiple threat sources, embedded in the constructed cluster hierarchy. We adopt the concept of the *lifetime* of a cluster to identify multiple threat sources from the built cluster hierarchy and localize their positions. The concept of the *lifetime* of a cluster is proposed in [33], which measures the range of bandwidth over which a cluster survives, i.e., the difference between the bandwidth when the cluster is formed and the bandwidth when the cluster is merged with other clusters. Specifically, given the cluster C_i formed at the ith iteration and the cluster C_{i+1} formed at the $(i + 1)$th iteration, we define the cluster C_i is survived if the distance between C_i and C_{i+1} is less than the half of the bandwidth at ith iteration. And the distance between two clusters is measured as the distance between these two centers of the clusters.

By using the definition of *survive*, we can calculate the lifetime of each cluster in the cluster hierarchy. Recall that in the mobile sensing data, there can be many isolated data points due to the unreliable sensing devices and the environmental noise. Each of those isolated data points may form one cluster and the lifetime of these clusters could be large. It is thus necessary to filter out these clusters whose number of data points is small. To accurately identify the existence of multiple threat sources, when given the number of the threat source is k, our technique first chooses clusters with top k longest lifetimes at each Mean Shift Clustering iteration (i.e., with different bandwidths) and puts them into a candidate list. We then return cluster centers with the top k longest lifetimes from the constructed candidate list, the top clusters chosen from all the iterations, i.e., under all the bandwidths, as the identified security threat sources. The positions of these cluster centers are the estimated locations of these sources. The threats source localization algorithm is described in Algorithm 3.

6 Simulation Evaluation

6.1 Methodology

To validate our mechanism, we simulated the scenarios of mobile sensing with sensors mounted on city taxicabs. The simulation area is about 25 by 25 street blocks, similar to the size of downtown Manhattan in New York City. The length of each square block is 20 units, which scale to 200 feet in real distance. Since total number of taxicabs in the New York City is about 13,000, it is reasonable to assume that there are 1,500 to 2,000 taxicabs operating in this area. For illustration, we deploy 1,500 taxicabs installed with detection sensors of 5% error rates (false positive rate and false negative rate).

We first randomly generate positions of these taxicabs. We then randomly put one or two security threat sources in the study region. Based on the models described in Section 3, we can assign a probability of positive detection for each mobile sensor. We only focus on the data points with positive sensor readings. We declare a correct detection if the detected cluster center is within the impact range of the security threat source. We repeat the same simulation steps 500 times and compute the percentage of times that the proposed algorithm correctly detects true sources as the detection rate of our mechanism.

6.2 Results

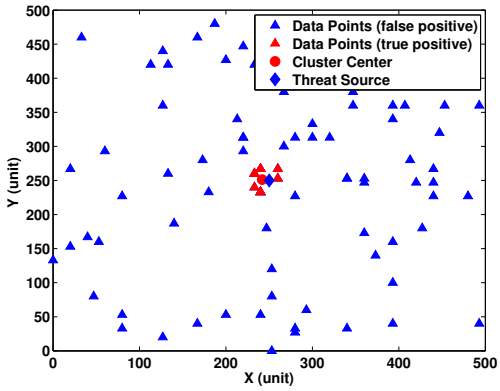

Fig. 2. An illustration of threat source detection using Mean Shift Clustering.

Figure 2 illustrates a set of data points with positive mobile sensor readings depicted as triangles. The detected cluster center obtained from our mechanism is shown as red circles. In this simulation scenario, there is only one threat source located at (250, 250) shown as a blue diamond and the impact range of the source is 20 units (i.e. 200 feet). The red triangles represent sensors with true positive readings while the blue triangles represent false positive readings. We simulated the sensor readings with low quality (e.g. 5% error rates) together with

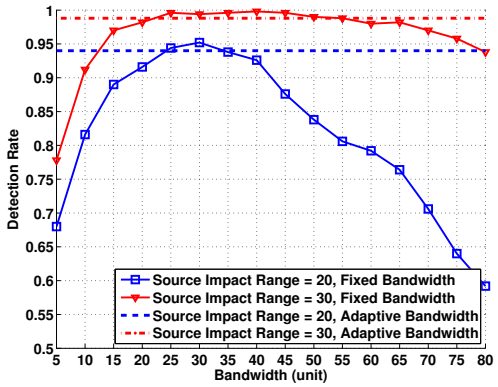

Fig. 3. One security threat source scenario: detection rate vs. bandwidth.

environmental noises. As a result, a large amount of false positive readings occur. This shows the challenges in accurate threat source detection. Moveover, the cluster center detected by our mechanism using the Mean Shift Clustering with bandwidth of 30 units is located extremely close to the true threat source. This indicates the feasibility of achieving accurate detection by using our mechanism.

Single Threat Source Detection. We first study the detection rate when there is only one security threat source present with an impact range of 20 units and 30 units respectively. The results of applying Algorithm 1 with various fixed bandwidths and Algorithm 3 using adaptive bandwidth (with the bandwidth increment step $p = 5$) are displayed in Figure 3.

The two curves of detection rate obtained from Algorithm 1 under different impact ranges initially increase with bandwidths, reach maximum detection rate, and then decrease. This is because a smaller bandwidth causes an undersmooth density estimate, whereas a bigger one causes an oversmooth density estimate. Both of the estimates result in inaccurate clustering of the sensing data. Based on the curves shown in the figure, the optimal bandwidth, where we achieve the best detection rate, is 30 units and 40 units for the impact source range of 20 units and 30 units respectively. This suggests that we would choose a bandwidth larger than the source impact range. At the optimal bandwidth, the detection rate is 95.2% and 99.8% for impact source range of 20 units and 30 units respectively. In general, the detection rate of a bigger impact source range is higher than that of a smaller impact range. This is because most likely we will obtain a larger area with more true positive sensor readings to make cluster detection more accurate under a larger source impact range. Furthermore, we observed that the detection results are less sensitive to the bandwidth under larger impact source range, i.e., the detection rate under the impact range of 30 units does not drop as much after its optimal point as the rate under the impact range of 20 units.

Moreover, our Algorithm 3 returns the detection rate of 94% for the impact range of 20 units and 98.8% for the impact range of 30 units. These two detection rates are plotted horizontally in the Figure 3 to compare with the results from Algorithm 1 under various fixed bandwidths. The key observation is that the

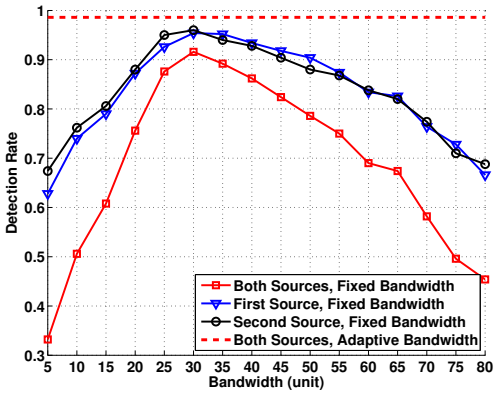

Fig. 4. Two security threat sources scenario: detection Rate under different bandwidths.

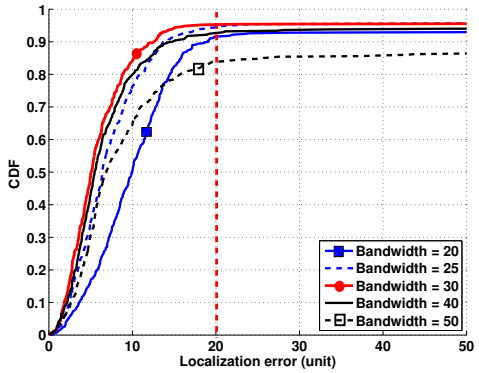

Fig. 5. Localization error under the source impact range of 20.

results of our technique is very close to that obtained under the optimal bandwidth with a fixed bandwidth. Therefore, without knowing the source impact range, our proposed method achieves similar high detection rate and is more feasible in practice than using the Mean Shift with a fixed bandwidth.

Multiple Threat Source Detection. We then move on to the scenario of two threat sources with impact range of 20 units. Figure 4 displays detection rates of correctly identifying each individual source and both under various fixed bandwidths using Algorithm 1 and Algorithm 3 respectively. Examining the curves of detection rate constructed using various fixed bandwidths, the detection rate for either one of individual source is much higher than that of both threat sources. Consistent with the single threat source case in Figure 3, the best detection rate is reached at bandwidth of 30 units under source impact range of 20 units. In addition, we attain an optimal rate of 91.6% for detecting both sources. Furthermore our proposed adaptive bandwidth method achieves a 98.6% detection rate which outperforms all the results using fixed bandwidths. This shows that our mechanism is highly effective in capturing multiple number of security threats.

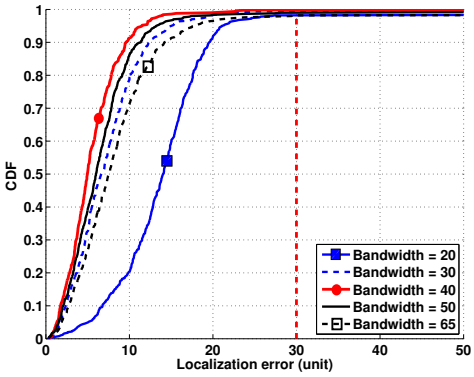

Fig. 6. Localization error under the source impact range of 30.

We next turn to examine the localization error resulted from our threat source localization. Figure 5 shows the cumulative distribution function (CDF) of the localization error under various fixed bandwidths for the single threat source scenario. A localization error is defined as the distance between the detected cluster center to the true source location. The dashed vertical line shows cut off the CDF curves at 20 units which is the source impact range. We observed that we have better localization accuracy when the bandwidth is close to the optimal bandwidth (i.e. 30 units). This is consistent with what we observed in Figure 3 in which a bandwidth value closer to the optimal bandwidth results in a better detection rate. Additionally the median error of the optimal bandwidth is 5 units (scaled to 50 feet) whereas the 90th percentile error is 12.6 units (scaled to 126 feet). Moreover, we observed that the localization error increases sharply when we failed to detect the threat source.

A similar plot for source impact range of 30 units is depicted in Figure 6. Again, we have the similar observation as Figure 5: a bandwidth value closer to the optimal bandwidth results in a better localization result. Moreover, the localization error of impact range of 30 units is smaller than that of the impact range of 20 units under the optimal bandwidth (i.e., 40 units). Specifically, the median errors of these two cases are about the same, however, the 90th percentile error of the impact range of 30 units is only 9.5 units, which is 30 feet shorter that of the impact range of 20 units. This is due to a longer impact range.

7 Conclusion

In this paper, we proposed to use mobile sensing, which utilizes the ubiquitous nature of mobile devices, for surveillance of security threats in urban environments, e.g., environmental pollution sources or nuclear radiation materials. We proposed to detect the security threats sources using all the sensors collaboratively under the presence of unreliable sensor readings. We formulated security threat source detection as a significant cluster detection problem, in which we identify whether one or more spatial clusters exist in the area significantly from

the background noise. We developed a detection mechanism grounded on Mean Shift Clustering procedure to both detect and localize the threat sources. In the proposed detection mechanism, a clustering hierarchy is built and the clusters who have the longest lifetime are returned as the threat sources. We evaluated the effectiveness of our mechanism by simulating mobile sensing using taxicabs in a metropolitan area. Our results obtained from simulations show that our detection mechanism can achieve over 90% detection rate for both single and multiple threat sources with low median localization errors, thereby strongly indicating the feasibility of detecting threat sources in urban environments using our approach.

Acknowledgments

This research was supported in part by NSF grants CNS-0954020 and CCF-1018270.

References

1. Kinkade, S., Verclas, K.: Wireless technology for social change: Trends in mobile use by NGOs. United Kingdom: UN Foundation–Vodafone Group Foundation Partnership. Retrieved November **15** (2008) 2008
2. World Health Organization (WHO): WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide: Global Update 2005 (2006)
3. Panofsky, W.: Nuclear proliferation risks, new and old. Issues in Science and Technology **19**(4) (2003) 73–74
4. IAEA: Trafficking in Nuclear and Radioactive Material in 2005 (2006) Available at http://www.iaea.org/NewsCenter/News/2006/traffickingstats2005.html.
5. Campbell, A., Eisenman, S., Lane, N., Miluzzo, E., Peterson, R.: People-centric urban sensing. In: Proceedings of the 2nd annual international workshop on Wireless internet, ACM (2006) 18
6. Johnson, P., Kapadia, A., Kotz, D., Triandopoulos, N., Hanover, N.: People-centric urban sensing: Security challenges for the new paradigm. Technical report, Citeseer (2007)
7. Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., Srivastava, M.: Participatory sensing. In: World Sensor Web Workshop. (2006) 1–5
8. Archer, D., Beauchamp, B., Mauger, G., Nelson, K., Mercer, M., Pletcher, D., Riot, V., Schek, J., Knapp, D.: Adaptable radiation monitoring system and method (June 21 2004) US Patent App. 10/874,127.
9. Glenn, F.: Radiation detection and measurement. John Wiley&Sons New York-Chichester-Brisbane-Toronto-Singapore (1989)
10. Morelande, M., Ristic, B., Gunatilaka, A.: Detection and parameter estimation of multiple radioactive sources. In: International Conference on Information Fusion. (2007)
11. Nemzek, R., Dreicer, J., Torney, D.: Distributed sensor networks for detection of mobile radioactive sources. In: Nuclear Science Symposium Conference Record, 2003 IEEE. Volume 3., IEEE (2004) 1463–1467

12. Brennan, S., Mielke, A., Torney, D., Maccabe, A.: Radiation detection with distributed sensor networks. IEEE Computer **37**(8) (2004) 57–59
13. Cheng, J., Xie, M., Chen, R., Roberts, F.: A mobile sensor network for the surveillance of nuclear materials in metropolitan areas. Technical report, DIMACS Technical Report, Rutgers University (2009)
14. Purdue University: Cell phone sensors detect radiation to thwart nuclear terrorism. ScienceDaily 24 January 2008
15. Glaz, J., Naus, J., Wallenstein, S.: Scan statistics. Springer Verlag (2001)
16. Balakrishnan, N., Koutras, M.: Runs and scans with applications. Wiley, New York (2002)
17. Fu, J., Lou, W.: Distribution theory of runs and patterns and its applications: a finite Markov chain imbedding approach. World Scientific Pub Co Inc (2003)
18. DematteIš, C., Molinari, N., Daurès, J.: Arbitrarily shaped multiple spatial cluster detection for case event data. Computational Statistics & Data Analysis **51**(8) (2007) 3931–3945
19. Wein, L., Wilkins, A., Baveja, M., Flynn, S.: Preventing the importation of illicit nuclear materials in shipping containers. Risk Analysis **26**
20. Langendoen, K., Reijers, N.: Distributed localization in wireless sensor networks: a quantitative comparison. Comput. Networks **43**(4) (2003) 499–518
21. Comaniciu, D.: An algorithm for data-driven bandwidth selection. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(2) (2003) 281–288
22. Yang, X., Liu, J.: Unsupervised texture segmentation with one-step mean shift and boundary Markov random fields* 1. Pattern Recognition Letters **22**(10) (2001) 1073–1081
23. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2003) 564–575
24. Peng, N., Yang, J., Liu, Z.: Mean shift blob tracking with kernel histogram filtering and hypothesis testing. Pattern Recognition Letters **26**(5) (2005) 605–614
25. Chen, H., Meer, P.: Robust fusion of uncertain information. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **35**(3) (2005) 578–586
26. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(8) (1995) 790–799
27. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on pattern analysis and machine intelligence **24**(5) (2002) 603
28. Parzen, E.: On estimation of a probability density function and mode. The annals of mathematical statistics (1962) 1065–1076
29. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics **27**(3) (1956) 832–837
30. Sheather, S., Jones, M.: A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological) **53**(3) (1991) 683–690
31. Park, B., Marron, J.: Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association **85**(409) (1990) 66–72
32. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Pr (1990)
33. Leung, Y., Zhang, J., Xu, Z.: Clustering by scale-space filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence (2000) 1396–1410