Multimodal layer surveillance map based on anomaly detection using multi-agents for smart city security

Hochul Shin¹ | Ki-In Na¹ | Jiho Chang¹ | Taeyoung Uhm²

1Department of Intelligent Robotics, Electronics and Telecommunications of Research Institute, Daejeon, Republic of Korea
2Intelligent Robotics R&D Division, Korea Institute of Robotics and Technology Convergence, Pohang, Republic of Korea

Correspondence
Hochul Shin, Department of Intelligent Robotics, Electronics and Telecommunications of Research Institute, Daejeon, Republic of Korea. Email: creatrix@etri.re.kr

Funding information
MSIP/IITP, Grant/Award Number: 2017-0-00306

Abstract
Smart cities are expected to provide residents with convenience via various agents such as CCTV, delivery robots, security robots, and unmanned shuttles. Environmental data collected by various agents can be used for various purposes, including advertising and security monitoring. This study suggests a surveillance map data framework for efficient and integrated multimodal data representation from multi-agents. The suggested surveillance map is a multi-layered global information grid, which is integrated from the multimodal data of each agent. To confirm this, we collected surveillance map data for 4 months, and the behavior patterns of humans and vehicles, distribution changes of elevation, and temperature were analyzed. Moreover, we represent an anomaly detection algorithm based on a surveillance map for security service. A two-stage anomaly detection algorithm for unusual situations was developed. With this, abnormal situations such as unusual crowds and pedestrians, vehicle movement, unusual objects, and temperature change were detected. Because the surveillance map enables efficient and integrated processing of large multimodal data from a multi-agent, the suggested data framework can be used for various applications in the smart city.

KEYWORDS
multimodal analysis, security robot, smart-city security, surveillance system

1 | INTRODUCTION

Smart cities are expected to provide residents with convenience through multiple agents such as CCTV, delivery robots, security robots, and unmanned shuttles. During their tasks, these agents can collect environmental data and be used for various purposes, including advertising and security monitoring.

For security applications, an effective service can be provided if various agents are combined, for example, if a security service is provided using fixed CCTV and delivery robots, costs can be reduced, and blind spots can be eliminated. Although there have been attempts to provide security services using multi-agents [1], the processing of large multi-modal data has a limit. Because of the limitations of wireless communication, most security robots performed most of their data processing within the robot. In many cases, the primary function of a robot is to perform a specific function such as intruder detection. However, these approaches have limitations when dealing with multi-agent systems in complex and dynamic environments.
However, although various data sets are emerging for data collection in smart cities, research on security monitoring services using multi-modal data from mobile platforms in a smart city environment is still lacking. When a multi-agent system is used in a smart city, considerable amount of data are generated, necessitating the use of an efficient method for dealing with the large amounts of data generated by multiple agents.

This study presents a data framework capable of providing efficient and integrated data processing by converting large multi-modal data from multiple agents into a multi-layered surveillance map. Moreover, using the framework, we propose an abnormal situation detection algorithm for security services.

2 RELATED WORK

2.1 Security robot applications

Although the sensor configuration is somewhat different, the security robot usually detects people and objects using RGB cameras and laser scanners. Meghana [2] used RFID tags and metal detector sensors to detect metal bombs in outdoor environments such as large facilities, critical infrastructure, and borders. A wireless camera mounted on the robot rotates in a 360-degree circle and continuously streams a defined outdoor area and employs infrared lighting to ensure surveillance even in complete darkness. Chakravarty [3] proposed a mobile robot that repeats a manually trained path and detects visual anomalies that did not exist during training. Mobile robots acquire panoramic images in a complex laboratory, corridor, and office and distinguish between images captured during autonomous driving and reference images captured during training to detect anomalies such as intruders. Zhang [4] proposed a robot system to patrol around a dam using a multi-sensor-based system. For precise movement in complex situations around the dam, data from wheel encoders, IMUs, LiDARs, and GNSS receivers were used. Deep learning was used to detect cracks, people, and anomalies using images from RGB cameras.

In the abovementioned applications, only an individual sensor was used to detect objects and abnormal situations, but more improved results can be obtained with multiple sensors. Di Paola [5] proposed an autonomous mobile robot for surveillance in an indoor environment and implemented a multi-sensor platform, including a monocular camera, a laser scanner, and an RFID sensor. The sensor platform was used to provide detection of lost objects for surveillance, human detection, and tracking and autonomous navigation.

2.2 Datasets and applications for smart city

Smart cities will collect considerable amount of city-based data from various sensors in multiple domains. For example, the US government public data site has a dataset of over 183 500 US cities, with an average of 2791 new datasets growing every month [6]. Consequently, in smart cities, modeling for application aims and efficient data management are important research topics. There is considerable amount of data in different domains such as traffic, environment, emergency and public safety, and energy and social sensing. Crowd analysis, behavior recognition, fire and smoke detection, traffic prediction, and vehicle tracking are certain datasets related to computer vision applications.

Deep learning technology is extensively used for solving problems for images or videos from CCTV. Behavioral recognition is an important process for detecting abnormal behavior, and many studies are ongoing. The following datasets can be used in the security of smart cities. There are datasets for human recognition, such as SCFace [7], UTKFace [8], and IMDB-WIKI [9], over various places and cameras. To research group classification and person tracking using crowd scenes of CCTV, datasets such as UCSD [10], UCFCC50 [11], and TRANCOS [12] are used. There is a dataset that focuses on human interactions, such as the TV Human Interaction Dataset [13], as well as a dataset that includes interactions with objects and sports such as UCF101 [14], Mivia [15], Bilkent, and Cetin are datasets on fire and smoke generation, which aim to detect from general RGB images, and not thermal imaging cameras. There are BIT Vehicle [16] and GTI Vehicle [17] datasets related to vehicles, and research on vehicle tracking, including rough classification (such as bus, SUV, sedan, and truck) and vehicle detection, is in progress. The abovementioned datasets are typically images or videos obtained from stationary cameras.

A method for transmitting and processing large amounts of data from multiple agents is necessary for real-time applications. Many studies aim to develop an approach to efficiently transmit such a large amount of data [18]. However, research in real-time processing methods for multimodal data of mobile multi-agents is still in its early stages. Consequently, we present a data framework capable of providing an efficient and integrated process for large multimodal data sets collected from multi-agents in smart cities.
3 | MULTI-AGENT BASED MULTI-LAYERED SURVEILLANCE MAP

3.1 | Overview of map-based surveillance

In a smart city, for an efficient and integrated multi-modal data process for the multi-agent system, we suggest a surveillance map data structure comprising two parts: an observation map and a probability map. The observation map is generated in two stages: (1) each agent converts large multi-modal data in a multi-layered local map and transmits it to the server, and (2) the server integrates each local observation map to develop a global multi-layered observation map. Using the previous observation map DB, the server computes the global probability map for the current observation map. This surveillance map can be developed by stacking multiple pieces of information based on the user requirements. In this study, as shown in Figure 1, information such as the behavior patterns of people and vehicles, elevation changes, and temperature distribution, were converted in a surveillance map.

3.2 | Multi-modal data collection and processing agent

3.2.1 | Multi-modal data collection from heterogeneous multiple agents

To monitor large areas at a low cost, we used two types of agents: moving agents and fixed agents, as shown in Figure 2. Monitoring efficiency can be increased by running a fixed agent and a mobile agent. The mobile agent collects multimodal data using an RGB camera, a thermal camera, and 3D Light Detection and Ranging, and it estimates the global position and orientation using Global Positioning System, Inertial Measurement Unit, and 3D LiDAR. The fixed agent is a CCTV-type agent mounted on a fixed pole or the wall of a building in an area that requires extensive monitoring. The fixed agent has only an RGB camera to recognize objects such as a human and a vehicle. It can estimate the actual position of objects in the image via calibration with the marker on the ground [19]. Moreover, the global positions of objects can be estimated using the marker position and orientation.

3.2.2 | Local multi-layered observation map building

Each agent can develop a local multi-layered observation map centered on itself using multi-modal sensors such as an RGB camera, a thermal camera, and 3D LiDAR. The local multi-layered local observation map is sent to the data analysis server to be integrated in a large global observation map. In this study, the multi-layered observation map comprises an object map, a thermal map, and an elevation map. In this study, people and vehicles were used as objects. The object map represents the moving direction, speed, and density of objects such as humans and vehicles. In this study, YOLOv3 [20] detector performs object detection, and the fast tracker [21] estimates the motion of objects such as speed and moving direction. Density is the number of objects in a given radius. For mobile agents, the position of detected objects is calculated from 3D LiDAR points on the object. The thermal map converts the temperature of the thermal image in a temperature distribution in 3D space using the point cloud of 3D LiDAR calibrated with the thermal camera. The thermal map is generated by projecting points with the temperature on the grid map. The elevation map is developed by projecting the height $z$ of 3D points on the grid map as per the position $(x, y)$.

**FIGURE 1** Concept of multi-layered surveillance map based security monitoring from multi-agent in smart city

**FIGURE 2** Data processing architecture in fixed agents, moving agents, data analysis server, and agent control server
Each local map layer generated in different frequencies should be merged in a multi-layered local map and transmitted to the surveillance map server with minimal projection error and data loss. The local observation map aligned with the global coordinates is generated to minimize position and orientation error in projection as shown in Figure 3A. In other words, the local coordinates for map building are aligned to the global coordinates; when projecting the local map onto the global map, this method can eliminate the projection orientation error.

To transmit the most recent observations having a small data size and minimal loss, different map layers are merged using asynchronous local multi-layered observation map integration. Each map layer generated at a different frequency has the global positions at the same time of sensor input and is projected on the integrated multi-layered global observation map via the global position, as shown in Figure 3B. In addition, when finally transmitting a local observation map to the surveillance map server, a map of a specified size is extracted and transmitted based on the latest position of mobile agents.

3.3 | Global multi-layered observation and probability map building

3.3.1 | Multi-layered global observation map

The data analysis server builds a multi-layered global observation map by gathering and merging multi-layered local observation maps from heterogeneous multi-agents. Because each agent asynchronously transmits the local observation map, each local observation map is projected on the global observation map based on its center position. Moreover, the map server updates the global observation map at a slower cycle compared with the agent transmission cycle. To prevent map data loss, the local observation maps with the center position are stored in the buffer, and a set number of recent observation maps are projected onto the global observation map. All map layers (object, thermal, and elevation maps) included in local observation maps are included in the global observation map, and the map layers are updated with the same map layers in the local observation map. Each map layer’s update is used to update the global observation map. The update map layer displays the object map’s sensing range and the detected position for the thermal and elevation maps.

3.3.2 | Probability map generation from the observation map

As shown in Figure 4, the probability of map generation is the process of calculating the probability score that the observation value will occur at each point. The probability scores are calculated based on the global observation map DB. The training and reasoning inference proceed as shown in Figure 5, and the conditional probability can be calculated by combining each unit probability score [22].

In this study, the probability scores for the presence of humans and vehicles, the density of crowds, the moving direction, and their speed were used. Moreover, the probability scores for the elevation and temperature of each observation point were calculated based on the observation map. The probability score normalized to 0–1 is converted into an image-type probability map by assigning the probability scores to the RGB channel depending on the situation. The object observation map shows the situation concerning humans and vehicles, as shown in Figure 4A. When a human or vehicle is detected, the probability score value is calculated at the corresponding location on the object observation map. For elevation and thermal observation maps, the observed values and probability scores for the observed values at the related point are expressed; if the observation value has a low probability, it is displayed as a dark value on the probability map.
3.4 Anomaly detection using a global surveillance map

3.4.1 Anomaly classification in a probability map

Using the deep learning algorithm, it is possible to determine whether the current situation is normal [23]. The proposed surveillance map comprises an observation map and a probability map, and unusual situations can be estimated [24]. The surveillance map-based anomaly detection is composed of (1) learning DB (Database) generation, (2) training phase, and (3) inference phase and each process is processed in parallel on the server, as shown in Figure 5.

**DB generation.** The training DB comprises four types of DBs as follows. Initially, an unlabeled DB is created using continuous observation map input. Then, the user labels some samples of unlabeled DB with normal and abnormal labels. If necessary, the user creates a synthetic abnormal DB for situations such as one-way streets and no-entry zones. Because a synthetic anomaly DB is expressed in the form of a surveillance map, it is relatively simple to define synthetic anomalies.

**Training phase.** Three types of trainings are conducted based on the created DB. First, the weight for the probability map is calculated using the unlabeled DB. Second, [25] auto-encoder training is performed on random samples of unlabeled and labeled normal DB. For the auto-encoder, a nine-layer network composed of convolution, batch-normalization, max pooling, and other techniques was employed. Finally, we train binary classification [26] for labeled normal and the combination of labeled abnormal and synthetic abnormal DB using a five-layer CNN network.

**Inference phase.** Abnormal candidates are identified using the probability map score and auto-encoder reconstruction error. On the extracted anomaly candidates, the CNN network is used for the second anomaly candidate filtering. If both steps determine that there is an anomaly, the user is notified that an abnormal situation has occurred.
4 | EXPERIMENTS

4.1 | Experimental settings

As shown in Figure 6, five fixed agents and three mobile agents are deployed in the surveillance area. The size of the entire monitoring area is about 170 m × 140 m, and the moving agent patrol path is set as per the number of mobile agents operating at the same time.

The moving agent is equipped with Real Sense D435, FLIR’s A65 thermal imaging camera, Velodyne’s VLP16 3D LiDAR, and Xsens’s MTi-G 710 IMU, and the fixed agent is equipped with only an RGB camera. Agents were developed based on a robot operation system (ROS). The multiple sensors mounted on the mobile agent and calibration results are shown in Figure 7. For multimodal sensor calibration, data coordinates are converted into LiDAR-RGB and LiDAR-thermal based on LiDAR [27]. The object recognition results and thermal maps are calculated based on the LiDAR coordinate system.

For global map building, robot position and heading information is necessary, and it can be obtained from the robot navigation process. For robot navigation, GPS, IMU, LiDAR, and visual odometry information are used to proceed, and an algorithm composed of weak/strong localizer and covariance matrix fusion modules is used, as shown in Figure 8 [28].

For the mobile agent, an average of 30 MB/s of multimodal data was generated. By surveillance map converting, it was reduced into 30 kB/s on average. There was a significant improvement in efficiency for data transmission, processing, and storage using the presented method. However, an information loss may occur during the conversion process as compensated by allowing the user to verify the corresponding RGB image. It is possible to add image processing algorithms on each agent and include the results in the additional layers of the surveillance map.

4.2 | Long-term operation and data collection

4.2.1 | Flow analysis of humans and vehicles

The system was built in the testbed, operated for 4 months, and the surveillance map DB was collected. The testbed that operates this system is jointly used by a research institute and a number of venture companies; it is an environment in which 60 employees are working. In such an environment, the maximum crowd that occurred in 4 months and the total activity score of people on that day are shown in Figure 9. Typically, the number of people who are outdoors is usually about four to eight people. However, a crowd of up to 17 people was temporarily observed because of an institutional visit and project demonstrations. However, the largest crowd typically occurs during lunch time, as shown in Figure 10.
4.2.2 | Spatial analysis of humans and vehicles

In Figure 11, the existence probability of humans and cars are shown as the probability map of the testbed. Moreover, the probability of crowd number in 5 m when there is a human, and the probability of parking and moving when there is a car are shown. The probability distribution map shows that the brighter the probability, the higher the probability, and each class is color-coded. People tend to park their car in the parking lot and move to the entrance of the building. There is a high probability that the cars are parked in the parking lot.

4.3 | Multi-layered environment map-based anomaly classification results

4.3.1 | Algorithms for abnormal candidate and filtering

The proposed anomaly detection method uses the probability score and reconstruction error of autoencoder [29] for abnormal candidates. The probability score as shown in Figure 4 was calculated using the event histogram weight. In this study, an autoencoder with 9-layer and 89k-parameter was used as optimization of server computing power, network training cost, and real-time inference. For the abnormal candidate filtering, a supervised CNN algorithm was used. There are excellent algorithms such as ResNet [30] and DenseNet [31]; however, when applied to our surveillance map data, there was no significant difference considering the computational cost; therefore, we applied 5-layer and 56k parameter general CNN network.

4.3.2 | DB composition and classification results for humans and vehicles

As explained in the previous system structure, the DB structure of this study comprises unlabeled, labeled normal, labeled abnormal, and synthetic abnormal. For human and vehicle DB, unlabeled 80k cases, labeled normal 9.3k cases, labeled the abnormal of 0.7k cases in such as unusual crowd, pedestrian, and parking. Moreover, the synthetic random motion of abnormal 5.0k cases was determined via the testbed system operation. Figure 13 shows an example of labeled abnormal surveillance map and detected abnormal cases.

For first candidate extraction, a train DB was developed by combining random sampling from unlabeled DB
and labeled normal. The test DB was constructed by labeling normal, labeled abnormal, and synthetic abnormal. When 50% of the constructed DB was trained using the autoencoder, normal 94.2% and abnormal 94.1% detection rates were achieved for test DB. For second anomaly candidate filtering, the normal situation was defined with labeled normal, and the abnormal situation is composed of labeled abnormal and synthetic abnormal. When 20% of the constructed DB was trained, detection performance of normal 99.7% and abnormal 98.4% was obtained for the remaining 80%. For the combined performance of steps 1 and 2, 92.6% of true positives and 0.3% of false positives were obtained based on the labeled DB (Figure 14).

4.3.3 | DB composition and anomaly classification results for elevation and temperature

The elevation and temperature DB comprises unlabeled 10k cases, labeled normal 9.2k cases, and labeled abnormal 2.3k cases. When 50% of constructed DB was trained using fist stage autoencoder, normal 58.4% and abnormal 98.4% was obtained for the remaining 80%. For combined performance of steps 1 and 2, 92.6% of true positives and 0.3% of false positives were obtained based on the labeled DB (Figure 14).
was composed of labeled normal DB and labeled abnormal DB. When 20% of constructed DB was trained, the detection performance of normal 98.2% and abnormal 96.6% was obtained for the remaining 80%. For the combined performance of steps 1 and 2, 92.9% of true positives and 1.8% of false positives were obtained based on the labeled DB.

5 | CONCLUSIONS

This study presents a data framework that efficiently processes multimodal data generated from multi-agents in a smart city environment. The proposed surveillance map, which comprises an observation map and a probability map, can express environmental information in the area where the agent is active. It allows for efficient data transmission and processing via multi-agent activity. This method transforms multimodal data in a local observation map in each agent and a global surveillance map in the server. The observation data generated by each agent can be modified based on the sensor installed in the agent and the needs of the user. The global observation map DB is used to generate a probability map. In this study, we used observation data for human and vehicle behaviors, in addition to elevation and temperature distributions. However, it is possible to provide a security service that detects unusual situations of the monitoring area using the proposed data framework. This study was supported by the ICT R&D program of MSIP/IITP (2017-0-00306, Development of Multimodal Sensor-based Intelligent Systems for Outdoor Surveillance Robots).

ACKNOWLEDGEMENTS

This work was supported by the ICT R&D program of MSIP/IITP (2017-0-00306, Development of Multimodal Sensor-based Intelligent Systems for Outdoor Surveillance Robots).

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

ORCID

Hochul Shin https://orcid.org/0000-0001-6987-3406
Ki-In Na https://orcid.org/0000-0002-4229-4786
Taeyoung Uhm https://orcid.org/0000-0002-2289-0521

REFERENCES


FIGURE 14  Examples of abnormal elevation and temperature surveillance map and detected abnormal cases. Fallen trees and the appearance of large objects were detected as abnormal elevation, and abnormal high temperature were mainly caused by vehicle exhaust pipes, brakes, and excavator


### AUTHOR BIOGRAPHIES

**Hochul Shin** received his ME and PhD in Mechanical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea. Since 2005, he has been working at the Intelligent Robotics Division of the Electronics and Telecommunications Research Institute (ETRI). His research interests include artificial intelligence and robotic applications.
Ki-In Na received a BS degree in Mechanical Engineering from the Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea, in 2009 and an MS degree in Robotics from KAIST, Daejeon, Republic of Korea, in 2011, where he is currently pursuing a PhD degree. Since 2011, he has been a Research Scientist with the Electronics and Telecommunication Research Institute (ETRI), Daejeon, Republic of Korea. His current research interests include detection and tracking of moving objects, social navigation, human–robot interaction, and artificial intelligence for real applications.

Jiho Chang received his BS (2004) and MS (2006) degrees in Electronics Engineering from Chung-Ang University, Republic of Korea. Since 2006, he has been a research scientist at the Electronics and Telecommunication Research Institute (ETRI). His primary research interests include computer vision, intelligent robots, and artificial intelligence.

Taeyoung Uhm received his BS degree in Electrical Engineering from Kyonggi University, Suwon, Republic of Korea, in 2004 and his MS and PhD degrees in Electronics and Computer Engineering from Hanyang University, Seoul, Republic of Korea, in 2014. From 2014 to 2017, he was a researcher with the Korea Institute of Science and Technology. In 2017, he joined the Smart Mobility Research Center at the Korea Institute of Robotics and Technology Convergence, Republic of Korea. His current research interests include multimodal 3-D sensors, robotics, autonomous navigation, visual slams, recognition, and human–robot interaction.

How to cite this article: H. Shin, K.-I. Na, J. Chang, and T. Uhm, Multimodal layer surveillance map based on anomaly detection using multi-agents for smart city security, ETRI Journal 44 (2022), 183–193. https://doi.org/10.4218/etrij.2021-0395