A comprehensive review on intelligent surveillance systems

Sutrisno Ibrahim*

Electrical Engineering Department, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia

Article history:
Received: 28 April 2016 / Received in revised form: 11 May 2016 / Accepted: 17 May 2016

Abstract

Intelligent surveillance system (ISS) has received growing attention due to the increasing demand on security and safety. ISS is able to automatically analyze image, video, audio or other type of surveillance data without or with limited human intervention. The recent developments in sensor devices, computer vision, and machine learning have an important role in enabling such intelligent system. This paper aims to provide general overview of intelligent surveillance system and discuss some possible sensor modalities and their fusion scenarios such as visible camera (CCTV), infrared camera, thermal camera and radar. This paper also discusses main processing steps in ISS: background-foreground segmentation, object detection and classification, tracking, and behavioral analysis.

Keywords: Intelligent surveillance system (ISS), object detection, human detection, moving object detection, object tracking, object recognition, behavioral analysis, CCTV.

1. Introduction

Massive amount of security cameras, along with other sensors, have been deployed to monitor critical infrastructure such as: military bases, airport, power plant, banking, campuses, etc. Manual monitoring by human operator is inefficient solution or even unpractical because human resource is expensive and has limited ability [1]. Intelligent surveillance system (ISS) is envisioned to automatically monitor the environment or infrastructure with less or without human intervention. Such monitoring tasks include automatically detecting and tracking object (like human or vehicle) and performing further analysis and actions. Signal processing, image processing, and artificial intelligence (machine learning) techniques play important role to develop such intelligent system.

Visible camera such as CCTV is the most common modalities (device) for surveillance system. It has long been in use to monitor environments, people, events and activities. Extensive studies have been conducted to automatically analyze data (image or video) from surveillance camera. Much of these studies have been discussed in several focused review papers: background-foreground segmentation [2-7], objects detection and classification [8-10], tracking [11-14], and behavioral analysis [15-17]. Different sensor modalities other than visible camera have been explored also for surveillance system such as infrared camera and thermal camera [18], radar (radio detection and ranging) [19, 20], lidar (light detection and ranging) [21-23], audio sensor [24], etc. Several review papers have discussed also different techniques for sensor fusion [25-29] to improve the system performance. However, there is still lack of comprehensive paper that discusses general overview of intelligent surveillance system.

The main objective of this paper is to provide general overview of intelligent surveillance system and review the existing methods for each its processing steps. The rest of this paper is organized as follows: Section II presents an overview of intelligent surveillance systems. Followed by Section III discusses some possible sensor modalities and different fusion scenarios. Section IV reviews the existing methods for background-foreground segmentation, object detection, classification, tracking, and behavioral analysis. Section V concludes the paper and highlights future research direction in this field.

2. Intelligent surveillance system (ISS) overview

Huge amount of security cameras, intelligent surveillance system (ISS) is a surveillance system that has intelligent capability to automatically analyze surveillance data and perform necessary actions such as generating alarm or warning. ISS is interdisciplinary topic that involves electronic (sensing device), computer vision and pattern recognition, artificial intelligence (machine learning), networking, communication and other areas. Intelligent surveillance system is promising to be implemented in various environments and applications. Some typical applications are listed as follows.

* Corresponding author.
Email: suibr@ksu.edu.sa

© 2016 KIPMI
- Home security [30, 31] and intrusion detection [32].
- Home care and safety [33].
- Public transport area such as airport, seaport, bus/train terminal [24].
- Public area [34] such as colleges, campuses, governmental building.
- Traffic monitoring [35].
- Crown management and analysis [36].
- Pedestrian detection and autonomous car [9, 10].
- Remote military surveillance, border monitoring, perimeter surveillance for power plant, company, etc.

Example of surveillance systems that have been previously studied or developed to have automation or intelligent capabilities: VSAM (video surveillance and monitoring) [37], W4 [38], PRISMATICA (pro-active integrated systems for security management by technological institutional and communication assistance) [24, 39], ADVISOR (annotated digital video for intelligent surveillance and optimized retrieval) [40]. Fig. 1 shows the overview of PRISMATICA system that has been proposed to improve passenger security and safety in the public transport system. It contains several main components: camera network (existing CCTV), intelligent camera system, transmission system, audio surveillance, operator, and also the main server (MIFSA).

Another impressive surveillance system is DARPA ARGUS-IS (autonomous real-time ground ubiquitous surveillance imaging system) [41]. With 1.8 Gigapixels video system, ARGUS-IS is able to auto-track every moving object within a 40 square kilometers (size of small city) using single platform. Such commercially available ISS products: DETEC AS (www.detec.no) and DETER (detection of events for threat evaluation and recognition) [42]. Intelligent surveillance system may play a significant role in security and safety in public, as well as in private domain. However, it is highly challenging due to some practical issues, such as:

- Performance: such as system accuracy
- Robustness: the systems should be robust against real word issues such as illumination variation, clutter, occlusion, weather change, camouflage, etc.
- Reliability
- Real time constraint: the system should fast enough
- Cost effective

3. Possible Sensor Modalities and Fusion Methods

3.1. Visible Camera

Visible (Video) camera is common sensor modalities for surveillance system. It has long been in use to monitor environments, people, events and activities. It is the most commercially available surveillance sensors starting from low cost IP camera to high performance professional CCTV. Security cameras have been placed in everywhere, from private homes, streets, public buildings, as well as in border between countries. Extensive research has been conducted for visible or video surveillance system [43]. Different types of visible cameras have been investigated for surveillance system such as color (or RGB) camera, monocular, stereo, omnidirectional camera, etc. Valera et al [44] divided the technological evolution of visual surveillance systems into three generations: Analog CCTV systems (1st generation), automated visual surveillance by combining computer vision technology with CCTV systems (2nd) and automated wide-area surveillance system (3rd).

3.2. Infrared (IR) and Thermal Camera

Visible camera is working well only in the environment that has enough illumination or light intensity, for example during in daytime. In the environment with low light intensity or during the night, visible camera cannot capture the scene effectively. In this case, there are two possible solutions: using infrared camera or thermal camera. Object (like human) that has contrast temperature with the surrounding environment is much easier to distinguish in the thermal or infrared camera image than in the visible camera image.

Both cameras capture infrared radiation that is invisible for human eye, therefore the “infrared camera” and “thermal camera” terms are usually interchangeable. However, infrared camera usually referred to a camera that captures near-infrared (NIR) or short-wavelength infrared (SWIR) emissions to increase the visibility. Infrared cameras are suitable for environments with a low illumination level. While thermal camera is referred to a camera that is able to capture long-wave or far-infrared (FIR) radiation emitted or reflected by objects. Thermal camera is useful if the scene is completely dark. Thermal camera can be divided into two types: cooled and uncooled. Cooled thermal camera provides higher resolution and image quality, but generally more expensive and consumes more power. Examples of infrared and thermal camera are FLIR cameras, produced by FLIR System (www.flir.com), and AXIS Q19 camera series. Fig. 2 shows a scene that is captured using visible camera and thermal camera in the same time [18].

Fig. 1. An overview of PRISMATICA system [24]

Fig. 2. A scene captured using visible and thermal camera [18]
3.3. Radar and lidar

Range sensing is an interesting sensor modality due to its accuracy, large field of view and robustness with respect to illumination changes. Such range sensing includes radar (radio detection and ranging) and lidar (light detection and ranging). Radar uses uses radio waves for sensing, while lidar uses light or laser. In range data, changes in the background can be easily filtered out by excluding all data outside of the tracking area. One drawback is that range data is generally less informative than vision data for person or object recognition. Spinello et al [21-22] proposed people tracking using 3D lidar. Recently, Banedek [23] also proposed 3D people surveillance using rotating multi-beam (RBM) lidar (as shown in Fig. 3). Javed et al [19] developed an embedded surveillance robot.

3.4. Other sensor (Audio, Ultrasonic, etc)

There are many sensor modalities have been explored to improve or assist surveillance system, such as: audio [24], ultrasonic [30], passive infrared (PIR) [31] pressure sensor, etc. Environmental sound like breaking of glass, dog’s barking, people screaming, fire alarm, gun firing and similar kind of sounds, may give a reasonable degree of confidence in making a decision about ‘secure’ or ‘insecure’ state [45]. Some sensors maybe used for alerting; ones they detect an object, visible camera (or other sensor) is activated for more reliable recognition. Bai et al [30] developed an embedded home surveillance system based on multiple ultrasonic sensors. In their other work [31], they used pyroelectric infrared sensors (PIR) and pressure sensors as an alert system to save the power.

3.5. Sensor Fusion

Intuitively, combining multiple sensors will provide more accurate information about the targeted object. Multiple sensors might be homogeneous (same modality, such as multiple cameras) or heterogeneous (different modalities). Some sensor modalities are intuitively closed and complementary, such as visible, infrared and thermal camera, since they capture information in 2D image perspective. Similarly, both lidar and radar capture information in range domain (2D or 3D). Sensor fusion may happen at low-level (data fusion), high-level (decision fusion) or in between. In data fusion, each sensor sends its original measurement to the fusion center, and then the center makes the decision about the event. In decision fusion, each sensor makes its own decision based on its own measurement, and then the fusion center makes the final decision based on all individual decisions (for example using majority voting). Each fusion scenario has its own advantages and drawbacks. Challenges in sensor fusion: how to handle different data modalities (visual, audio, radio signal, etc), data imperfection, conflicting data, sensor topology etc [28]. More information about basic sensor fusion may refer to [25-29].

Extensive works have been done on visual surveillance system using multiple cameras [46-48]. Reference [29] reviews recent progress in intelligent video surveillance using multiple cameras that include multi-camera calibration, computing the topology of camera networks, multi-camera tracking, object re-identification, and also multi-camera activity analysis. Robertson et al [49] combine visible, infrared and thermal camera for outdoor people detection for moving platform (vehicle). Premebida et al [50] proposed pedestrian detection combining RGB camera and dense LIDAR data.

4. Data Processing Techniques for ISS

4.1. Foreground-Background Segmentation

Foreground-background segmentation is the first important step for intelligent surveillance system. The goal is to separate the object or moving object (foreground) and the environment (background). It commonly referred also as background modeling, background subtraction, or change detection. Many foreground-background segmentation techniques have been proposed, especially for visible/video surveillance. Several review paper also available that focused discuss foreground-background segmentation methods [2-7], Bouwmans [5] discussed and provide a comprehensive list most of available techniques (see Table 1). Fig. 4 shows foreground-background segmentation using different methods: generalized mixture of Gaussians (MOG) [51], non parametric kernel density estimation [52], and codebook [53].
Obviously any possible techniques for foreground-background segmentation depend on the corresponding sensor modality. Recently, Sobral et al [7] compared 29 methods using BMC (Background Models Challenge) dataset [68]. Top five promising methods based on this experimental work are the methods that proposed by Wen et al [57], Kaewtrakulpong et al [69], Yao et al [70], Maddalena et al [63], Hofmann et al [71]. Cristani et al [3] discussed also other sensing modalities (such as audio, infrared and thermal camera) in their survey paper. Most of the proposed method in background-foreground segmentations employed only single sensor modality, and particularly using visible camera. Obviously, combining different sensor modality would make the system more robust or simplify the processing process for segmentation. For example, by combining visible camera and range data the background-foreground segmentation task become easier. Changes in the background can be easily filtered out by excluding all data outside of the observed area in the range data and the visible/image data is used for fine segmentation.

<table>
<thead>
<tr>
<th>Category</th>
<th>Methods</th>
<th>Main Contributor (Author, year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>Mean</td>
<td>Lee et al. (2002)[54]</td>
</tr>
<tr>
<td>Background Modeling</td>
<td>Median</td>
<td>Mac Farlane et al. (1995)[55]</td>
</tr>
<tr>
<td></td>
<td>Histogram over time</td>
<td>Zheng et al. (2006)[56]</td>
</tr>
<tr>
<td>Statistical</td>
<td>Single Gaussian</td>
<td>Wren et al. (1997)[57]</td>
</tr>
<tr>
<td>Background Modeling</td>
<td>Mixture of Gaussians</td>
<td>Stauffer and Grimson (1999)[51]</td>
</tr>
<tr>
<td></td>
<td>Kernel Density Estimation</td>
<td>Elgammal et al. (2000,2002)[52,58]</td>
</tr>
<tr>
<td>Fuzzy</td>
<td>Fuzzy Running Average</td>
<td>Sigi et al. (2008)[59]</td>
</tr>
<tr>
<td>Background Modeling</td>
<td>Type-2 Fuzzy Mixture of Gaussians</td>
<td>El Baf et al. (2008)[60]</td>
</tr>
<tr>
<td>Background Clustering</td>
<td>K-Means Codebook</td>
<td>Butler et al. (2003)[61]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kim et al. (2005)[53]</td>
</tr>
<tr>
<td>Neural Network</td>
<td>General Regression</td>
<td>Cullbirk et al. (2006)[62]</td>
</tr>
<tr>
<td>Background Modeling</td>
<td>Neural Network Self Organizing</td>
<td>Maddalena and Petrosino (2007)[63]</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td></td>
</tr>
<tr>
<td>Wavelet</td>
<td>Discrete Transform Wavelet</td>
<td>Biswas et al. [64]</td>
</tr>
<tr>
<td>Background Modeling</td>
<td>Estimation</td>
<td></td>
</tr>
<tr>
<td>Background Estimation</td>
<td>Wiener Filter</td>
<td>Toyama et al. (1999)[65]</td>
</tr>
<tr>
<td></td>
<td>Kalman Filter</td>
<td>Messelodi et al. (2005)[66]</td>
</tr>
<tr>
<td></td>
<td>Tchebychev Filter</td>
<td>Chang et al. (2004)[67]</td>
</tr>
</tbody>
</table>

4.2. Object Detection and Classification

The ability to automatically detect and classify object (such as human and vehicle) is one of key component in intelligent surveillance system (ISS). For a machine (computer), detecting object like human is a hard job due to wide range of possible appearance as result of changing articulated pose, clothing, lighting and background [8]. Huge methods have been proposed for people detection based on visual camera. In their experimental survey, Enzweiler et al [9] showed an advantage of HOG/linSVM [72] at higher image resolutions and lower processing speeds, and a superiority of the wavelet-based AdaBoost [73] cascade approach at lower image resolutions and closed real-time processing speeds. In the more recent benchmarking effort, Dollar et al [10] show that FPDW [74] has the best overall performance, but if computational cost is not a consideration, then MULTIFTR+MOTION [75] is the best choice.

Spinello et al [21, 22] proposed people detection using a bottom-up top-down detector, based on lidar data. The bottom-up detector learns a layered person model from a bank of specialized classifiers for different height levels of people that collectively vote into a continuous space. In the top-down step, the candidates are classified using features that are computed in voxels of a boosted volume tessellation. While in [23] Banedek et al map the 3D lidar point data into depth-image, and performing people detection in 2D. Spinello et al in [22] presented a people detection approach based on RGB-Depth sensors that provide both image and range data. Fig. 5 shows example of their result for people detection. Most of the proposed method in objects detection and classification focusing only for couple of object types for example human and car. In fact, in the real setting, there are a lot of object that should be considered also for example different type of animal or other subject that have potential threat for security or safety.

Fig. 5 people detection using RGB-D: color image (left), depth image (right)

4.3. Object Tracking and Re-Identification

After object detection, surveillance systems generally track the object in the spatiotemporal domain. Object tracking in are realistic scene is a challenging problem due to illumination changes, occlusion, clutter, sensor motion, and other issues. A large number of visual tracking algorithms (based on visible camera) have been proposed in recent years. Object tracking methods based on visual camera can be classified into five groups: model-based, appearance-based, contour- and mesh-based, feature-based, and hybrid methods [76]. Several review papers that focused on visual tracking problem are available such as [11-14]. Recently, Smeluders et al [12] performed experimental survey based on Amsterdam Library of Ordinary Videos (ALOV) for 19 online trackers. Another effort for benchmarking visual object trackers was proposed by Wu et al [14]. According to the Visual Object Tracking challenge (VOT2014) result, the best tracker (combined accuracy and
4.4. Behavioral Analysis

There is an increasing interest to automatically analyze surveillance scene not only in the “object level” (such as detecting, tracking), but also further into “event level”. Particular interests such are automated human behavior analysis [80, 81], group behavior analysis [82], crown analysis [36, 83, 84], and event analysis. Some review papers have been devoted to this topic [15-17]. Human behavior analysis can play a significant role in security by decreasing the time taken to thwart unwanted events and picking them up during the suspicion stage itself [17]. Analysis of human behavior although crucial, is highly challenging. Basic component in human behavior analysis is classifying the human behavior. Different ways have been proposed to classify human behavior. Kiryati et al [85] proposed simple classification: normal and abnormal. Foroughi et al [86] expand the classification into normal, unusual and abnormal. Previously, Park and Aggarwal [87] classified the activities as positive, neutral and negative activities.

Human can be monitored as isolated individuals, groups of people, or crowds. Examples of group events are people fighting, people being followed, people walking together, terrorists launching attacks in groups, etc [82]. Solmaz et al [83] proposed a method for identifying five crowd behaviors (bottlenecks, fountainheads, lanes, arches, and blocking) in visual scenes. Bremond et al [88] proposed an activity-monitoring framework for recognizing behaviors, involving either isolated individuals, groups of people, or crowds, in the context of visual monitoring of metro scenes, using multiple cameras. For example, Fig. 7 shows their result to recognize the “fighting behavior” in a metro station. The combined four methods or descriptions to recognize fighting behavior such as: (A) a group of people gathering around a lying person, (B) group width varied significantly, (C) people inside a group separate quickly, and (D) group trajectory changes very fast [88].

However, it should be noted that current research in the behavior analysis are still considering simple or simplified scene. More realistic and complex scene should be investigated. For example, in real “fighting behavior” it may be involving also the use of weapon such as knife or gun, so there will be less contact between fighting groups or persons.

The four descriptions proposed by Bremond et al above may fails to characterize this fighting behavior.

5. Conclusion and Future Direction

In this paper, general overview of intelligent surveillance systems has been presented. Such intelligent systems are promising to be implemented in various environments and applications. This paper also has discussed some possible sensor modalities and their fusion scenarios to improve the system performance. Numerous techniques have been proposed to tackle several main processing steps: background-foreground segmentation, object detection and classification,
tracking, and behavioral analysis. Although several promising results have been obtained, further studies are needed for real implementation with more complex settings. For example, in the background-foreground segmentation process, different combination of sensor modality should be explored to make the system robust or to simplify the processing process. Current studies in behavior analysis are still considering simplified scene, and thus more realistic and complex scene should be investigated. With decreasing price in sensor and processing devices, researchers should also consider developing and a low cost intelligent surveillance system.

References

7. A. Sobral, A. Vacavant, A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos, Computer Vision and Image Understanding 122 (2014) 4-21.
34. F. Porikli, Y. Ivanov and T. Haga, Robust abandoned object detection


