Evaluation of Smart Video for Transit Event Detection

Project #BD549-49

FINAL REPORT

Prepared for the

Florida Department of Transportation
Research Center

Prepared by the

National Center for Transit Research
Center for Urban Transportation Research
and
Department of Computer Science and Engineering
University of South Florida

June 2009

Disclaimer

The opinions, findings, and conclusions expressed in this publication are those of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the views and policies of the Florida Department of Transportation or the Research and Innovative Technology Administration. This report does not constitute a standard, specification, or regulation.

Aknowledgements

We would like to thank David Kelsey at the Hillsborough Area Regional Transit (HART) agency for their generous support, allowing us to use their resources for testing surveillance equipment, and providing us with information and perspectives from the transit community. We would also like to thank Gary O'Neal of Verint and Mark Steffler of Vidient for allowing us to demonstrate commercial technology, and providing us with information on the state of the art.

1. Report No.

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Evaluation of Smart Video for Transit Event Detection

5. Report Date
June 2009

6. Performing Organization Code

7. Author(s)
Dmitry B. Goldgof, Deborah Sapper, Joshua Candamo, and Matthew Shreve

8. Performing Organization Report No.
2117-7807-00

9. Performing Organization Name and Address

National Center For Transit Research (NCTR)
Center for Urban Transportation Research
University of South Florida – CUT100
4202 East Fowler Avenue, Tampa, FL 33620

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
BD 549-49

12. Sponsoring Agency Name and Address
Office of Research and Special Programs (RSPA)
U.S. Department of Transportation, Washington, DC 20590

Florida Department of Transportation
605 Suwannee Street, MS 26, Tallahassee, FL 32399

13. Type of Report and Period Covered
Final Report

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Transit agencies are increasingly using video cameras to fight crime and terrorism. As the volume of video data increases, the existing digital video surveillance systems provide the infrastructure only to capture, store and distribute video, while leaving the task of threat detection exclusively to human operators.

The objective of this research project was to study and develop an evaluation framework for commercial video analytics systems. A state-of-the-art research literature survey was conducted. Identified strengths, weaknesses, future directions of research and state-of-the-art commercial video analytics products were surveyed. Product capabilities were identified by working together with vendors and analyzing the available literature offered by the providers. Use of analytic technology in transit agencies in Florida was analyzed. A technology survey among the largest agencies in the state indicates very low use of video analytics, significant skepticism, and poor general knowledge of the technology and its capabilities. Based on existing general evaluation frameworks, an evaluation framework for video analytics technology was developed, including annotation guidelines, scoring metrics, and implementation of the scoring metrics in the scoring software.

17. Key Word
Transit Security, Video Cameras, Video Analytics, Anomaly Detection Systems

18. Distribution Statement
Available to the public through the National Technical Information Service (NTIS), 5285 Port Royal Road, Springfield, VA 22161, (703) 487-4650, http://www.ntis.gov/ , and through the NCTR web site at http://www.cutr.usf.edu/

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
76

22. Price
No Cost

EXECUTIVE SUMMARY

Transit agencies are increasingly using video cameras to fight crime and terrorism. As the volume of video data increases, the existing digital video surveillance systems provide the infrastructure only to capture, store, and distribute video, while leaving the task of threat detection exclusively to human operators. Studies were done by Sandia National Laboratories for the U.S. Department of Energy to test the effectiveness of an individual whose task was to sit in front of a video monitor(s) for several hours a day and watch for particular events. The studies showed that even when assigned to a person who is dedicated and well-intentioned, this method of using technology will not support an effective security system. After only 20 minutes of watching and evaluating monitor screens, the attention of most individuals has degenerated to well below acceptable levels. Monitoring video screens is boring, mesmerizing, and has no intellectually engaging stimuli. To address this problem, New Jersey Transit has connected over 1,400 of its cameras to computers that can automatically detect suspicious activity, using complex vision-based algorithms. Any abnormal behavior detected will set off an alarm or a pager or give a call to whoever is responsible for that camera. Other types of smart video surveillance that can be used by transit agencies include:

1) The ability to preempt incidents - through real time alarms for suspicious behaviors.
2) Enhanced forensic capabilities - through content based video retrieval.
3) Situational awareness - through joint awareness of location, identity, and activity of objects in the monitored space.

In transit scenarios, an increase in situational awareness would directly benefit the safety and efficiency of both the passengers and the security personnel on the ground. Early warnings also can be issued before events occur. Decision making also becomes easier since the event can be replayed immediately on command, rather than second-guessing what may have been seen, and unnoticed behavior that is a concern becomes less common.

When criminal activity or a threat is detected, security personnel and the proper authorities can be provided with real-time information when assisting the situation. Various alerts can be set up, triggered by pre-defined operationally relevant events. Information can be disseminated using text messaging, on-screen alerts, email, geo-coded maps, pictures, and video. The faces of detected criminals can help pinpoint further appearances in past, current, or future video data. Attention-intensive activities such as object removal or objects left behind will be detected by the system immediately instead of possibly being unnoticed, resulting in a delayed reaction by a surveillance operator.

Some drawbacks of video analytic systems are their vulnerability to environmental variables, such as detrimental lighting conditions and weather. These adverse conditions can trigger false alarms, which may become a source of frustration for the user. Another drawback with video analytics is that events must be pre-defined, so events that have not been defined will not be detected. Conversely, a human analyst may use judgment and training to determine if an alarm should be raised for a wider range of scenarios. Video analytic algorithms often are sensitive to parameters and initial calibration. Event detection performance typically depends on this calibration process. It is difficult to achieve a good balance between event detection and false alarms.

Typically, a higher detection rate produces a higher false alarm rate, and vice-versa. Additionally, some video analytic implementations may require the system to be re-calibrated over time. Improved core technology algorithms are needed to increase the reliability of human behavior recognition.

During the last decade, numerous methods for evaluating core technologies have been proposed. However, there are no standard evaluation methods for human behavior recognition. Creating standard evaluation tools includes defining a common terminology and generating operationally similar datasets. For example, a bus and a metro station can both be "crowded." But operationally, the "crowds" in both situations are very different. Thus, without a standard, precise definition of "crowd," formal comparisons become a very difficult task.

The cost-effectiveness of these systems to transit agencies will depend on independent verification of the systems' performance against the task(s) deemed most important by the transit agencies for the application.

CONTENTS

FIGURES

TABLES

CHAPTER 1
BASIC TERMINOLOGY AND CONCEPTS

CLOSED-CIRCUIT TELEVISION (CCTV)

Closed-circuit television (CCTV) is the use of video cameras to transmit video signals to a given set of monitors. CCTV is commonly used in public transit surveillance as well as other applications that may need surveillance monitoring such as banks, casinos, military installations, among others. The total number of CCTV systems has increased rapidly over the last few decades [1]. Today, transit networks have large CCTV traffic-monitoring systems, which are used to review accidents and detect congestion status. For example, in England, the number of CCTV systems is estimated to have surpassed four billion [1].

BASIC VIDEO PROCESSING SOFTWARE

Basic video processing software refers to application software that can perform basic video processing techniques, such as video resizing (zooming), format conversion, jitter removal, noise removal, filtering, scene/cut segmentation, etc. Basic video processing capabilities are included in most CCTV packages. Basic video processing does not include video analytics, which can be used to automatically detect complex behaviors and events that are operationally relevant in public transit surveillance, such as suspicious baggage left behind, object exchange, loitering, intrusion, vandalism, etc. Basic video processing can also include smart recording technology. Smart recording refers to systems that record video when something of interest occurs. For example, a camera pointed towards a door needs to record only when people moving through it.

MANUAL VIDEO ANALYTICS

Manual video analytics is defined as a labor-intensive task where human analysts scan video data looking for operationally-relevant behaviors and events (suspicious behavior, accidents, etc.), without the aid of video analytics software. Manual video analytics can be done pro-actively (live/streaming feed) or reactively (after-the-fact) by reviewing previously-archived video data. After-the-fact video analysis is referred to as video forensics.

SOFTWARE VIDEO ANALYTICS

Software analytics is computer software technology that automatically detects pre-defined behaviors and events in video. Similar to its manual counterpart, software analytics products can be used pro-actively (real-time automated monitoring) or re-actively (video forensics). In the literature, software video analytics is also referred to as Intelligent Video Surveillance (IVS), or anomaly / event detection systems. In general, software analytic capabilities increase security, reduce shrinkage, and increase operational awareness and efficiency.

The analytics software is responsible for detecting pre-defined events of interest, which are instances of actions or behaviors of objects in a scene. The simplest form of video analytics is to detect motion in videos. A more complex example would be detecting a trespasser crossing the rails in a subway station or loitering behavior at a bus stop, typical of drug dealers and beggars. Smart recording is commonly used in conjunction with analytic software, which will considerably reduce the amount of video data that needs to be stored. The analytics software can either run on remote computers or be embedded within the surveillance equipment. Researchers distinguish between real-time and forensic analytics, since the variation between the two will impact the capabilities of the system. Some systems may process pre-recorded data to detect events of interest but cannot detect events as they occur. Conversely, some systems may be designed to detect events only in real time and will not be able to analyze archived data from cameras outside the network.

OBJECTS

An object is defined as anything that is of interest for further analysis [2]. In transit systems, common objects include humans, vehicles, bags, briefcases, backpacks, etc. Distinguishing characteristics of different objects are used to separate them (size, shape, motion, speed). Most behavior recognition algorithms rely on tracking objects over time. A behavior of interest may be a pedestrian loitering at a bus stop. For this, the object (person) must be tracked across the entire video, where occlusion (such as walking behind another object) is often a problem. Another example is if an object (person) suspiciously leaves behind an unattended object (such as a bag), in which case there are two very different objects of interest that must be tracked over time to establish a meaningful spatio-temporal relationship.

EVENTS

The most general definition of an event is something that happens at a given place and time. Commercial surveillance system providers might use slightly different definitions for the term event. However, the general definition in transit surveillance is widely accepted by the industry. To clarify further, event can refer to a single, low-level spatiotemporal entity that cannot be further decomposed (such as a person walking) or to a composition of multiple of these entities (such as loitering). Also, manufacturers will often refer to anomaly detection rather than event detection to emphasize the general purpose of video analytics. The general goal of analytic software systems is to accurately pinpoint occurrences of certain events of interest, which are often deviations of normal behavior.

SYSTEM ARCHITECTURE

Modern analytic surveillance systems usually consist of three components, as shown in Figure 1. The first component is a set of cameras, which collect and broadcast video through a data channel. Depending on the provider, previously-mounted cameras may be compatible with newer surveillance equipment; however, the specifications of each camera will directly influence the overall capabilities of the system. For example, accurate object detection can depend on the level of detail acquired in the video (resolution), so older cameras may offer reduced quality compared to newer ones.

The next component is the processing box. The processing box processes each channel from the cameras. The number of channels required depends on the size of the area to be surveyed, could range from a dozen for a small facility to tens of thousands for a large one.

The third and final component is the video analytics software, which may be partially embedded into the prior components using specialized hardware or run remotely on a computer server.

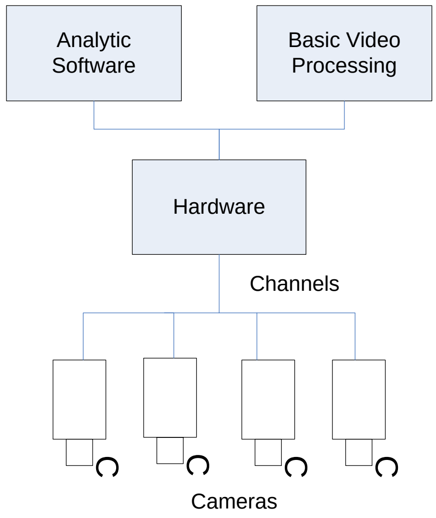


Figure 1 Typical Setup for Transit Surveillance Systems that Include Video Analytics

CHAPTER 2
RESEARCH LITERATURE SURVEY

Visual surveillance is an active research topic in image processing. Transit systems are actively seeking new or improved ways to use technology to deter and respond to accidents, crime, suspicious activities, terrorism, and vandalism. Human behavior recognition algorithms can be used proactively for prevention of incidents or reactively for investigation after the fact. In this section, the current state-of-the-art image processing methods for automatic behavior recognition techniques are described, with a focus on the surveillance of human activities in the context of transit applications.

This survey provides a summary of progress achieved to date and helps identify areas where further research is needed. A thorough description of the research on relevant human behavior recognition methods for transit surveillance is presented. Recognition methods include single person actions (such as loitering), multiple person interactions (such as fighting, personal attacks), person-vehicle interactions (such as vehicle vandalism), and person-facility/location interactions (such as objects left behind, trespassing). A list of relevant behavior recognition studies is presented, including behaviors, datasets, implementation details, and results. Also, algorithm weaknesses, potential research directions, and contrast with commercial capabilities as advertised by manufacturers are discussed. A summary of literature surveys and developments of the core technologies (low-level processing techniques) used in visual surveillance systems, including motion detection, classification of moving objects, and tracking, is also presented.

INTRODUCTION TO VIDEO ANALYTICS TECHNOLOGY

Military, intelligence, and mass transit agencies are increasingly using video cameras to fight crime and terrorism. Due to hardware and storage improvements during the last decade, a collection of continuous surveillance video is already at our doorsteps, while the means to continuously process it are not. To illustrate the scope and scale of large surveillance transit systems, consider the following examples. The New York Metro [3] is the busiest metro system in the United States (based on 2006 statistics), with a total of 468 stations and 1.49 billion riders per year, 4.9 million per day. Moscow Metro [4] is the busiest metro in Europe, and as of 2007 has 176 stations with 2.52 billion riders annually, 9.55 million per day. This ridership represents a 9.53 percent growth since 1995. Transit systems are spread across hundreds of kilometers and already require several tens of thousands employees for daily operations. A complete deployment of visual surveillance to cover systems of this magnitude requires thousands of cameras, which makes human-based/dependant surveillance infeasible for all practical purposes.

As the volume of video data increases, most existing digital video surveillance systems provide the infrastructure only to capture, store, and distribute video, while leaving the task of threat detection exclusively to human operators. Detecting specific activities in a live feed or searching in video archives (video analytics) almost completely relies upon costly and scarce human resources. Detecting multiple activities in real-time video feeds is currently performed by assigning multiple analysts to watch the same video stream simultaneously. Each analyst is assigned a portion of the video and is given a list of events (behaviors) and objects to look for. The analyst issues an alert to the proper authorities if any of the given events or objects are spotted. Manual analysis of video is labor-intensive, fatiguing, and prone to errors. Additionally, psychophysical research indicates that there are severe limitations in the ability of humans to monitor simultaneous signals [5]. It is clear that there is a fundamental contradiction between the current surveillance model and human surveillance capabilities.

The ability to quickly search large volumes of existing video or monitor real-time footage will provide dramatic capabilities to transit agencies. Software-aided real-time video analytics or forensics would considerably alleviate the human constraints, which currently are the main handicap for analyzing continuous surveillance data. The idea of creating a virtual analyst or software tools for video analytics has become of great importance to the research community. The purpose of this study is to review the state-of-the-art methods for automatic video analytic techniques, with focus on surveillance of human activities in transit systems. Human and vehicle behavior recognition has become one of the most active research topics in image processing and pattern recognition [6, 7, 94, 123]. Previous surveys have emphasized low-level processing techniques used in visual surveillance ("core technologies," such as motion detection, tracking, etc). In contrast, this research focuses on human behavior recognition topics, drawing special attention to transit system applications. For clarity, a brief review of the state-of-the-art core technologies is offered, and previous surveys in related areas are identified (see Table 1).

Table 1 Related Literature Survey Summary

First Author	*Yr*	*Topic*	*Ref #*
Zhan	2008	Crowd analysis	[123]
Kang	2007	Intelligent visual surveillance	[94]
Stoykova	2007	3D scene capture	[40]
Sun	2006	On-road vehicle detection systems	[154]
Forsyth	2006	Human motion	[8]
Yilmaz	2006	Object tracking	[74]
Radke	2005	Image change detection	[46]
Valera	2005	Intelligent distributed surveillance systems	[7]
Haykin	2004	Object tracking	[73]
Foresti	2004	Multi-sensor tracking	[85]
Weiming	2004	Motion and tracking for surveillance	[6]
Fasel	2003	Facial expressions (small-scale body movements)	[36]
Moeslund	2000	Human motion capture (large-scale body movements)	[9]
Aggarwal	1999	Motion analysis of the human body	[95]
Gavrila	1999	Human movement	[65]
Pavlovic	1997	Hand gestures (small-scale body movements)	[35]
Ju	1996	Human motion estimation and recognition	[67]
Cedras	1995	Motion-based classification	[66]
Aggarwal	1994	Elastic non-rigid motion	[96]
Cedras	1994	Motion detection	[67]
Barron	1992	Optical flow	[56]

Video analytics gained significant research momentum in 2000, when the Advanced Research and Development Activity (ARDA) started sponsoring detection, recognition, and understanding of moving object events. Research focused on news broadcast video, meeting/conference video, Unmanned Aerial Vehicle (UAV) motion imagery and ground reconnaissance video, and surveillance video. The Video Analysis and Content Extraction (VACE) project focused on automatic video content extraction, multi-modal fusion, event recognition, and understanding. The Defense Advanced Research Projection Agency (DARPA) has also supported several large research projects involving visual surveillance and related topics. Projects include the Visual Surveillance and Monitoring (VSAM, 1997) project [10] and the Human Identification at a Distance (HID, 2000). Recently, the Video and Image Retrieval Analysis Tool (VIRAT, 2008) project was announced. VIRAT's purpose is to develop and demonstrate a system for UAV video data exploitation, which would enable analysts to efficiently provide alerts of events of interest during live operations, and retrieve video content of interest from archives.

Video analytics have become increasingly popular in commercial systems. Later in this survey, a summary of some of the existing commercial systems is provided. The list includes advertised capabilities for human behavior recognition. It is unclear how well systems are able to cope with crowds of people, typical of mass transit systems. The cost-effectiveness of behavior detection systems to transit agencies depends on independent verification. Verification of the systems' performance is based on the tasks deemed most important by the transit agencies for the application. Efforts to create standard evaluation frameworks (methodologies to quantify and qualify performance) have been of increasing interest to the research surveillance community [11, 12, 13, 14, 15, 16, 17, 18, 19, 21]. Additionally, there are methods for evaluating the performance of the evaluators [20]. Despite the large number of existing evaluation techniques, a robust study that experimentally compares algorithms for human activity recognition is still missing.

In the last decade, there have been many conferences and workshops dedicated to visual surveillance, among them the IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) 2005 challenge, which focused on real-time event detection solutions. CREDS [21], defined by the needs of the public transportation network of Paris (RATP, the second busiest metro in Europe), focused on proximity warning, dropping objects on tracks, launching objects across platforms, and persons trapped by the door of a moving train, walking on rails, falling on the track, and crossing the rails. Several CREDS solution proposals can be found in the References section of this report [22, 23, 24, 25]. The Performance Evaluation of Tracking and Surveillance (PETS) [26] workshops started with the goal of evaluating visual tracking and surveillance algorithms. The initiative provides standard datasets, with available ground truth, for evaluating object tracking and segmentation. Recently, a metric to evaluate surveillance results also was introduced [27]. Some PETS datasets contain relevant information closely related to transit systems. Datasets include single-camera outdoor people and vehicle tracking (PETS 2000); multi-camera camera outdoor people and vehicle tracking (2001); diverse surveillance-related events including people walking alone, meeting with others, window shopping, fighting, passing out, and leaving a package in a public place (2004); and images containing left-luggage scenarios (2006).

Around the world, large underground transit networks (such as France's RATP, United Kingdom's LUL and BAA, Italy's ATM, etc.) have deployed and tested large real-time transit visual surveillance systems that include human behavior recognition. Several transit surveillance projects have been funded by the European Union. The Pro-active Integrated Systems for Security Management by Technological, Institutional and Communication Assistance (PRISMATICA) [28] has deployed video analytic systems in France. The Content Analysis and Retrieval Technologies to Apply Knowledge Extraction to Massive Recording (CARETAKER) [29] project was deployed in Italy. The Annotated Digital Video for Intelligent Surveillance and Optimized Retrieval (ADVISOR) [30] was successfully deployed and tested in Spain and Belgium, including previous work from the Crowd Management with Telematic Imaging and Communication Assistance (CROMATICA) project [31, 32, 33, 34].

LITERATURE SURVEY ORGANIZATION

The main focus of this survey is to offer a comprehensive survey of image processing human behavior recognition algorithms in the context of transit applications. All the pre-processing steps prior to behavior recognition are referred to in this study as "core technologies." Human behavior recognition using video starts with the detection of foreground objects are commonly achieved through environmental modeling or motion-based segmentation. Subsequently, foreground objects are classified depending on the application as humans or vehicles. Object classification can be shape-based, motion-based, or based on a particular descriptor suitable for a specific application. Finally, tracking establishes the spatio-temporal relationship between the objects and the scene.

The organization of this report is shown in Figure 2. The report begins with a brief glance of the core technologies to facilitate the understanding of the later sections of the paper. For organization purposes, all pertinent surveys dealing with core technologies are identified and summarized in Table 1. In Chapter 3, behavior recognition strategies are discussed. Chapter 4 elaborates on many important topics describing the current state-of-the-art, strengths, weaknesses, and future research directions. Chapter 5 summarizes the report.

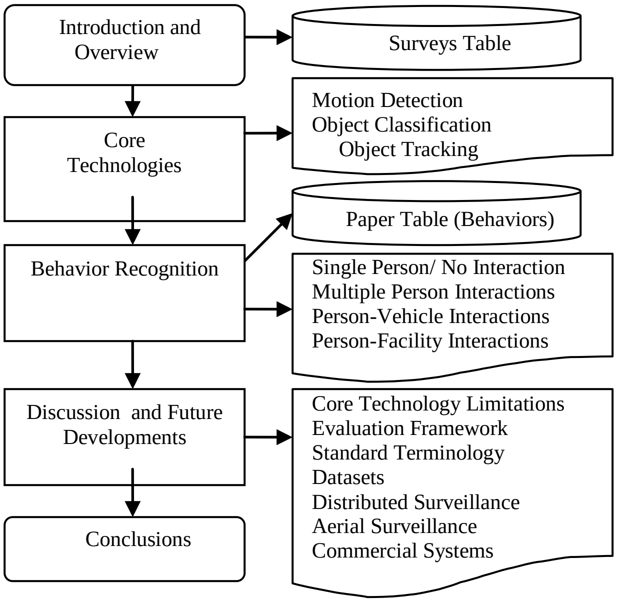


Figure 2 Study Organization Flowchart

CORE TECHNOLOGIES

Motion Detection

Visual surveillance systems for fixed cameras traditionally include some sort of motion detection. Motion detection is used to segment moving objects from the rest of the image. Knowledge about the motion of objects is useful in both the object and behavior recognition processes. A survey on early work in motion detection can be found in a 1994 study. In transit surveillance applications, motion detection typically refers to movement of objects as a whole (movement of pedestrians or vehicles). Human motion can also be referred to articulated motion of the human body, such as the motion of certain body-parts like legs or arms. There are two types of articulated motion: large-scale body movements like movements of the head, arms, torso, and legs [9], and small-scale body movements like hand gestures and facial expressions [35, 36]. In general, motion detection can be subdivided into environment modeling, motion segmentation, and object classification. All three often overlap during processing. Nearly all current surveillance systems rely on 2D data for motion processing; thus, the focus of this study will be on this domain.

Advances in image sensors and the evolution of digital computation is leading to creation of new sophisticated methods for capturing, processing, and analyzing 3D data from dynamic scenes. Recent developments include 3D environmental modeling reconstructed using the shape-from-motion technique [37] and 3D imagery from a moving monocular camera [38]. Most 3D approaches require landmarks to be present in the scene [39] in order to accurately estimate the required extrinsic parameters of the camera, which sets an additional set of practical constraints for deployment of systems. A survey on emerging perspective time-varying 3D scene capture technologies can be found in Stoykova et al. [40].

Background Subtraction and Temporal Differencing

A popular object segmentation strategy is background subtraction. Background subtraction compares an image with an estimate of the image as if it contained no objects of interest. It extracts foreground objects from regions where there is significant difference between the observed and the estimated image. Common algorithms include methods by Heikkila and Olli [41], Stauffer and Grimson (Adaptive Gaussian Mixture Model or GMM) [42], Halevy [43], Cutler [44], and Toyama (WALLFLOWER) [45]. A detailed general survey of image change algorithms can be found in Radke et al. [46].

The GMM is one of the most commonly-used methods for background subtraction in visual surveillance applications for fixed cameras. A mixture of Gaussians is maintained for each pixel in the image. As time passes, new pixel values update the mixture of Gaussians using an online K-means approach. The estimation update is used to account for illumination changes, slight sensor movement, and noise. Nevertheless, transit surveillance researchers continue to emphasize the importance of robust background subtraction methods [48] and online construction and adaptive background models [47]. A large number of recent background subtraction methods improve on prior existing methods by modeling the statistical behavior of a particular domain or by using a combination of methods. For example in Cheung and Kamath [48], a slow adapting Kalman filter was used to model the background over time in conjunction with statistics based on an elliptical moving object model. Robust background subtraction is typically computationally expensive; thus, methods to improve standard algorithms are becoming increasingly popular [31]. For example, Dominguez-Caneda et al. [40] state that for a GMM, speed can be improved by a factor of 8 with an image size of 640 by 480 pixels.

Another common object segmentation method is temporal differencing. In temporal differencing, video frames are separated by a constant time and compared to find regions that have changed. Unlike background subtraction, temporal differencing is based on local events with respect to time and does not use a model of the background to separate motion. Typically, two or three frames are used as separation time intervals, depending on the approach. A small time interval provides robustness to lighting conditions and complex backgrounds, since illumination changes and objects in the scene are more likely to be similar over short periods of time and a image stabilization algorithm is required when there is significant movement of the camera [49]. Temporal differencing is usually computationally inexpensive, but it regularly fails at properly extracting the shape of the object in motion and can cause small holes to appear. For these reasons, hybrid approaches [50, 51] often combine both background subtraction and temporal differencing methods in order to provide more robust segmentation strategies.

Optical Flow

Optical flow is a vector-based approach that estimates motion in video by matching points on objects over multiple frames. A moderately high frame rate is required for accurate measurements. It should be noted that a real-time implementation of optical flow will often require specialized hardware, due to the complexity of the algorithm. A benefit of using optical flow is that it is robust to multiple and simultaneous camera and object motions, making it ideal for crowd analysis and conditions that contain dense motion. Popular techniques to compute optical flow include methods by Black and Anandan [52], Horn and Schunck [53], Lucas and Kanade [54], and Szeliski and Couglan [55]. A comparison of methods for calculating optical flow can be found in Barron et al. [56].

Object Classification

After finding moving regions or objects in an image, the next step in the behavior recognition process is object classification. For example, a pedestrian crossing a street and a vehicle running a red light can be similar if there is no knowledge of the object causing the motion. Object classification could distinguish interesting motion from those caused by moving clouds, specular reflections, swaying trees, or other dynamic occurrences common in transit videos. It is important to note that there are multiple possible representations of objects before and after classification. Common geometric or topological properties used include height/width ratio, fill ratio, perimeter, area, compactness, convex hull, and histogram projection. (For detailed definitions of these properties, see [57]). Some of these properties are also used in post-object classification to keep track of the object in sequential frames or separate cameras. In general, for object classification in surveillance video, classification methods are shape-based, motion-based, and feature-based.

Shape-Based Classification

The geometry of the extracted regions (boxes, silhouettes, blobs) containing motion often are used to classify objects in video surveillance. Some common classifications in transit system surveillance are humans, crowds, vehicles, and clutter [10]. For transit applications, especially those oriented to human behavior recognition, appearance features extracted from static images have been proven effective in segmenting pedestrians without the use of motion or tracking [58, 59, 60]. Shape-based recognition methods find the best match between comparisons of these properties in association with a-priori statistics about the objects of interest. For example, in Bird et al. [61], blobs are first extracted and classified based on the calculated human height/width ratio based on data from the National Center for Health Statistics. Shape-based classification is particularly useful in certain transit systems when only certain parts of the objects are fully visible; for instance, in buses and metros, objects will be partially occluded most of the time, in which case the head [62] could be the only salient feature in the scene.

Motion-Based Classification

This classification method is based on the idea that object motion characteristics and patterns are unique enough to distinguish between objects. Humans have been shown to have distinct types of motion. Motion can be used to recognize “types” of human movements such as walking, running, or skipping, as well as used for human identification, Starting with the HumanID Gait Challenge [63], image processing researchers actively proposed gait-based methods [64] for people identification at a distance. (For more information on motion-extraction and motion-based classification, see [65] and [66]; for an overview of motion estimation and recognition with focus on optical flow techniques, see [67]).

Other Classification Methods

Skin color [68] has proved to be an important feature that can be used for the classification of humans in video, as it is relatively robust to changes in illumination, viewpoint, scale, shading, and occlusion. Skin color has also been successfully combined with other descriptors [69] for classification purposes. In Bird et al. [61], the authors describe a method that consists of three parts. First, a red-green-blue (RGB) normalization procedure was adopted to get the pure color components. A color transform was then applied which correlates each pixel to that of its Gaussian distribution of the skin color, higher intensities being closer to the center. Hence, the output showed the region of the image that closely matched with skin color indicating human motion. This method was extended in Yang et al. [70] and fused with other methods, including depth analysis using binocular imaging. The fusion of methods has shown to be very effective when combining shape and motion-based methods [71, 72].

Object Tracking

In the context of transit systems, tracking is defined as the problem of estimating the trajectory of a pedestrian in the image plane while he/she is in the transit station or vehicle. The increasing need for automated video analysis has motivated researchers to explore tracking techniques, particularly for surveillance applications. Object tracking in general is a difficult task. Many problems that come from general object tracking are the same as those for human and vehicle tracking, among them multiple moving objects, noise, occlusions, object complexity, scene illumination variations, and sensor artifacts. (For additional information on tracking, see detailed object tracking surveys [73, 74]). Specific issues that arise within the transit domain include dealing with multiple persons in complex scenarios [75], tracking across large-scale distributed camera systems [76], tracking in highly-congested areas with crowds of people [77] (such as near ticket offices, metro, or buses waiting areas at rush hour, etc.), or tracking using mobile platforms [78]. Extremely frequent occlusions are typical; consequently, the traditional localization and tracking of individuals is not sufficiently reliable. Surveillance inside transit vehicles often allows only parts of individuals to be captured by the sensors (such as common occlusions from seats and other passengers often exposes only faces inside buses and metros).

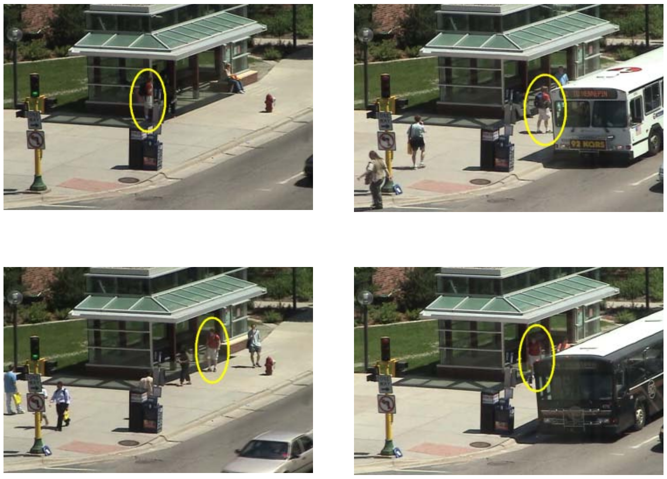
Tracking systems assign persistent identification tags to tracked pedestrians in different frames of a video. Depending on the application requirements, it is common for the system to also maintain other subject characteristics, such as aspect ratio, area, shape, color information, etc. Selecting good features that can be used for future tracking or identification is a necessity, since the object's appearance in a later frame may vary due to orientation, scale, or other natural changes. Also, feature uniqueness plays an important role. Some common features used in image processing applications are color, edges, motion, and texture. In Gasser et al. [79], researchers describe a system that monitors suspicious human activity around bus stops, in which tracking of pedestrians is performed using a kernel-based method proposed in Comanciu et al. [80]. This tracker is based on the color distribution of previously-detected targets. Current position is found by searching the neighborhood around the previously found target and computing a Bhattacharyya coefficient, which is used as a correlation score. In Bird et al. [61], the shirt color is used as the main feature for tracking purposes, and kernel-based tracking is dropped in favor of a blob-based tracking. Blob-based tracking offers a computational advantage over kernel-search since the latest has to be first initialized, which would redundantly require blob-extraction to be performed. Blob-based methods are extremely popular in the literature; for example, in proposed solutions to the CREDS challenge, Spirito et al. [22] considers the use of a long-memory matching algorithm [81] using the blob's area, perimeter, and color histogram, and another [24] performs a blob-based color histogram tracking. The French project Système d'Analyse de Médias pour une Sécurité

Intelligente dans les Transports publics (SAMSIT) focuses on automatic surveillance in public transport vehicles by analyzing human behaviors. Inside metros and buses, faces are the only body part mostly captured by surveillance cameras, while the other body parts are occluded, especially by the seats. Therefore, tracking is performed using faces with a color particle filter [82], similar to [83]. The tracking is based on the likelihood from the Bhattacharyya distance between color histograms in the Hue-Saturation-Value (HSV) color space. Color-based tracking is robust against vibration of the moving vehicles like trains and buses and is sensitive to extreme changes in lighting conditions, such as a train entering a tunnel. Many multi-sensor approaches [84, 85], algorithm fusion techniques [86], and integrating features over time [87] have been proposed to overcome many of the mentioned tracking difficulties, and to generate robust tracking performance in transit surveillance applications.

TYPES OF HUMAN BEHAVIOR RECOGNITION

In this survey, the terminology and classification strategy for human behavior are similar to those used by the VIRAT project. VIRAT divides human behavior in two categories: “events” and “activities.” An event refers to a single low-level spatiotemporal entity that cannot be further decomposed (such as a person standing, a person walking). An activity refers to a composition of multiple events (such as a person loitering). Across the literature, the term “event” is often used interchangeably to describe “events” or “activities” as defined by VIRAT. For clarity, in this study the term “behavior” includes both “events” and “activities.” For organizational purposes, transit surveillance operationally-relevant behaviors are divided into four general groups: (A) Single Person or No Interaction, (B) Multiple Person Interactions, (C) Person-Vehicle Interactions, and (D) Person-Facility/Location Interactions. Provided below are examples of each of these groups:

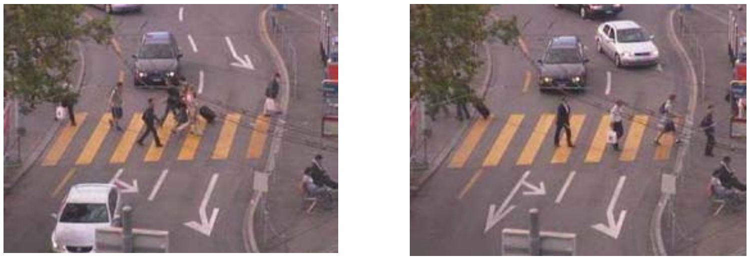
- *Single Person or No Interaction* (Figure 3) consists of behaviors that can be defined only by considering person(s) who are not interacting with any other person or vehicle, such as loitering, people-counting (crowd–counting), crowd flow (behavior) analysis, person talking on a cell phone, etc.



(Suspicious person marked with an ellipse loitering for a long period of time without leaving in a bus stop)
Images courtesy of the Center for Distributed Robotics, University of Minnesota. Images are part of the dataset used in [61].

Figure 3. Sample Single Person or No Interaction Behavior

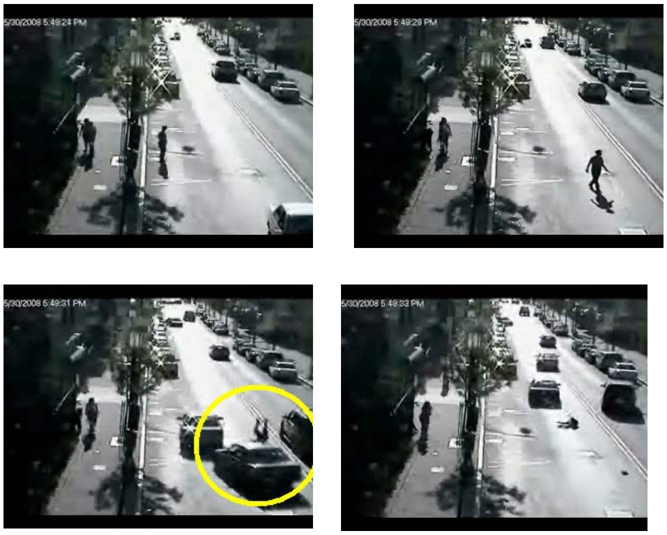
- *Multiple Person Interactions* (Figure 4) are behaviors that involve persons interacting with each other. An example of the behavior includes: following, tailgating, meeting, gathering, moving as a group, dispersing, shaking hands, kissing, exchanging objects, kicking, etc. breaking window, dropping off, picking up, etc.



Images courtesy of the Computer Vision Laboratory, ETH Zurich.
Images are part of the dataset used in [87].

Figure 4 Sample Multiple Person Interaction Behavior: Pedestrians on a Crosswalk

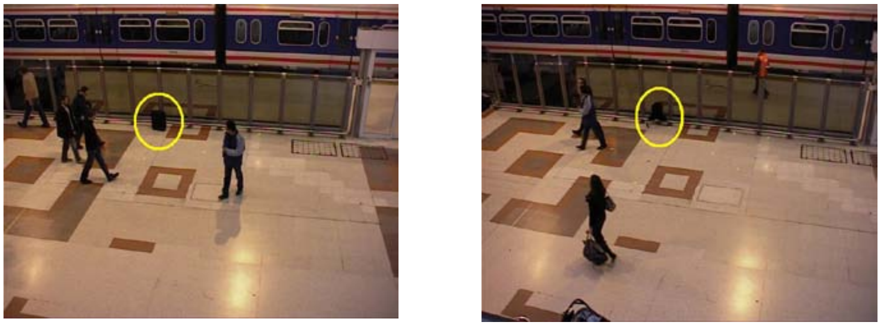
- *Person-Vehicle Interactions* (Figure 5) consist of behaviors that are defined through interactions with persons and vehicles, for example, driving, getting in or out of the car, loading or unloading, opening or closing the trunk, crawling under the car, etc.



Crime solver public video release from
Hartford Police Department in Connecticut.

Figure 5 Sample Person-Vehicle Interaction: Person Being Run Over by Vehicle

- *Person-Facility/Location Interactions* (Figure 6) are behaviors defined through interactions with persons and facilities/locations. An example of this behavior would include entering or exiting), standing, waiting at checkpoint, evading checkpoint, passing through gate, object left behind, vandalism, etc.



Object left behind sample images from PETS 2006 dataset [26]

Figure 6 Sample Person-Facility/Location Interaction: Person Leaving a Bag in a Train Station

In surveillance systems, behavior recognition can be ambiguous depending on the scene context. The same behavior may have several different meanings, depending upon the environment and task context in which it is performed. Human behavior recognition has been the focus of several workshops such as Visual Surveillance (1998) [88, 89], Event Mining (2003) [90, 91], and Event Detection and Recognition (2004) [92, 93]. (See [94] for a brief background review of advances in intelligent visual surveillance and [95, 96] for a review on studies of motion of the human body.)

Any reliable behavior recognition strategy must be able to handle uncertainty. Many uncertainty-reasoning models have been proposed by the artificial intelligence and image understanding community and already have been used in visual surveillance applications. The Bayesian approach is perhaps the most common model due its robustness and relatively low computational complexity as compared to other methods, such as the Demptster-Shafter theory [97]. Uncertainty handling can improve visual attention schemes [98]. Various other models have been used in surveillance-related applications, including classifying human motion and simple human interactions using a small belief network [99], human postures using belief networks [100], description of traffic scenes using a dynamic Bayes network [101], human activity recognition using a hierarchal Bayes network [102], and anomalous behavior detection using trajectory learning with Hidden Markov Models [103,104].

SINGLE PERSON OR NO INTERACTION

Loitering

Loitering is defined as the presence of an individual in an area for a period of time longer than a given time threshold. Methods for automatically detecting loitering in real-time would enable deployed security to investigate suspicious individuals or to target loitering stations for future investigation. Loitering is of special interest to public transit systems since it is a common practice of drug dealers, beggars, muggers, and graffiti vandals, among others. In this study, loitering refers to behavior that involves a human exclusively. It is not to be confused with stationary objects (such as objects left behind), which in this classification falls under Person-Facility Interaction behaviors. Before a loitering activity is detected, individuals can be engaged in other activities like browsing, entering, leaving, and passing through [105].

In general, literature for loitering detection in transit system applications consists primarily in tracking using indoor video (see Table 2). However, publications often lack of implementation and technical details [23, 106, 107]. The technical literature exclusively to outdoor loitering detection is scarce. In Bird et al. [61], outdoor loitering is used as a cue to detect potential drug-dealing operations in bus stations. Drug dealers often wait for their clients to come by bus, buy drugs, and leave. Consequently, suspicious activity is defined as individuals loitering, using a time threshold longer than the maximum time that it would typically take to catch a bus. The technique proposed in Bird et al. [61] uses a refined Gaussian Mixture background subtraction algorithm to detect motion blobs in a calibrated scene. Blobs are classified as humans using size and shape descriptors, and a short-term biometric based on the color of clothing is used for tracking purposes. A calibrated scene is used to calculate the effect of distortions in the pedestrian's size due to the perspective projection. In transit scenes it is often impractical to manually measure camera parameters on site and almost impossible when working only with pre-recorded examples [108].

Crowd Counting

Accurate people detection can increase management efficiency in public transportation by marking areas with high congestion or signaling areas that need more attention. Estimation of crowds in underground transit systems can be used to give passengers a good estimate of the waiting time in a queue. Multiple solutions to automate the crowd-counting process have been proposed, including solutions from a moving platform (such as a camera on a bus) [109] that analyze the optic flow generated from the moving objects as well as the moving platform.

Researchers have identified crowd counting to be often highly-sensitive to training data [110], in which cases algorithms or crowd density classifiers [111] will greatly benefit from having a realistic and robust training dataset. New techniques for creating human crowd scenes are continuously being developed, especially due to the growing demand from the motion picture industry [112]. Simulated crowds have been widely studied in many application domains, including emergency response [113] and large-scale panic situation modeling [114, 115]; perhaps simulated crowds [116] or flow models could also potentially offer visual surveillance researchers a new way to efficiently generate training data.

Solutions using fixed cameras that use standard image processing techniques can be separated into two types. In the first, an overhead camera that contains “virtual gaits” that counts the number of people crossing a pre-determined area is used. Clearly, segmentation of a group of people into individuals is necessary for this purpose [117]. The second type attempts to count pedestrians using people detection and crowd segmentation algorithms. In the overhead camera scenario, many difficulties that arise with traditional side-view surveillance systems are rarely present. For example, overhead views of crowds are more easily segmented, since there is likely space between each person, whereas the same scenario from a side-view angle could be incorrectly segmented as one continuous object. When strictly counting people, some surveillance cameras are placed at bottlenecked entrance points where, at most one person at any given time, is crossing some pre-determined boundary (such as a security checkpoint or an access gate at a subway terminal). A potential drawback is that overhead views are prone to tracking errors across several cameras (unless two cameras are operating in stereo), since human descriptors for overhead views are only reliable for a small number of pedestrians [118], using multiple cameras may further complicate crowd counting. In the cases where over-head surveillance views are not available, side-view cameras must be used to count people, and the multiple problems associated with this view (such as crowd segmentation and occlusion) come into play. In the case of crowd segmentation, some solutions that have been proposed include shape indexing, face detection, skin color, and motion [119, 121].

Most of these methods rely heavily on image quality and frame rate for accurate results. Shape indexing and skin colors are considered robust to poor video quality, while motion and face detection are most dependent on video quality. Occlusion is another problem, since all or part of a person may be hidden from view. Some techniques try to mitigate this issue by detecting only heads [120] or omega-shaped regions formed by heads and shoulders [121].

Crowd Behavior

Crowd behavior analysis has drawn significant interest from researchers working closely to the transit domain [122]. A recent survey [123] focused on crowd analysis methods employed in image processing. The flow of large human crowds [108] is a useful cue for human operators in real-time behavior detection, such as diverging crowd flow and obstacles. Flow cues can be used reactively by human operators to efficiently deal with accidents or preventively to timely control situations that potentially could lead to graver incidents. Recent crowd behavior analysis methods include tracking of moving objects [124], motion models using optical flow [125, 126, 127, 128] and crowd density measurement using background reference images [129]. A related surveillance problem consists of identifying specific individual events in crowded areas [130], in which motion from other objects in the scene will cause significant clutter under which algorithms might fail. Detecting particular behaviors based on crowd analysis (such as panic, fighting, vandalism) is a new research direction for projects like SERKET [131], recently funded by the European Union to create methods to analyze crowd behaviors and aid in the fight against terrorism. Common abnormal crowd characteristics that have been researched are fallen person, blocked exit, and escape panic [127, 132, 133]. Behavior classification is often based on the vector fields generated by crowd motion instead of individual person tracking.

Human Pose Estimation (Stance Change)

In transit surveillance applications, human pose estimation refers to the pose of the entire human body (for example, going from standing to lying down in a metro is an indication of pedestrian collapse) and not pose-related to a single body part, such as a head pose that can be used in applications such as driving monitoring [134]. Keeping track of multiple body parts is often useful to estimate the global body poses. There are two main approaches to estimating body pose. The first approach calculates ratios between the height and width of the bounding box around a detected human. In Balan et al. [135], vertical and horizontal projection templates are used to detect standing, crawling/bending, lying down, and sitting. The second approach attempts to track specific joints and body parts [136, 137], both because they are useful for indicating human pose and also because when accurately modeled, they can be used to recover the pose even after occlusion and other common tracking failures [138]. Due to self occlusion and background clutter, some approaches also use the motion generated from each body part as a feature for pose change [139], since movements from each joint are shown to be inter-dependent. In a study by Baumberg and Hogg [140], the observed motion is compared with registered motion exemplars, while action models are used to estimate possible future poses.

Multiple Person Interactions

Multiple person interactions have largely been motivated by the growing demand for recognizing suspicious activity in security and surveillance applications. In [141], the behavior detection process consists of foreground segmentation, blob detection and tracking. Semantic descriptions of suspicious human behavior are defined through groups of low-level blob-based events. For example, fights are defined as many blobs' centric moving together, merging and splitting, and overall fast changes in the blobs' characteristics. Attacks are defined as one blob getting too close to another blob, with one blob perhaps being initially static, and one blob erratically moving apart. Large projects like BEHAVE (years 2004-2007) [142] and CAVIAR (years 2002-2005) [143] have each produced several publications focusing on multiple person interactions. Algorithms include the use of a nearest neighbor classifier based on trajectory information [144] in order to detect human interactions such as walking together, approaching, ignoring, meeting, splitting, and fighting, Bayesian networks [145] and moment Invariant feature descriptions [146] to detect events including sitting down, standing up, bending over, getting up, walking, hugging, bending sideways, squatting, rising from a squatting position, falling down, jumping, punching, and kicking. Often, performance relies on the ability to accurately segment and separate multiple human motions. Multiple free-form blobs and course models of the human body were used in two person interaction in [147], which used a hierarchal Bayesian Network to recognize human behaviors based on body part segmentation and motion. This work was extended [148] to track multiple body parts of multiple people. Processing at three levels (pixel, blob, and object) was used to distinguish punching, hand-shaking, pushing, and hugging. A technique that does not use temporal motion information but instead uses pose is discussed in study by Park and Aggarwal [149]. By using a string matching method using a K-nearest neighbors approach, the authors were able to classify shaking hands, pointing, standing hand-in-hand, and the intermediate transitional states between these events.

Exchanging objects between persons is a common security concern in airports and other transit scenarios. In Haritaoglu et al. [150], backpack exchanging is detected based on the shape analysis of each person. First, a person is detected to be carrying or not carrying a backpack or any other object. Then, the object is segmented and tracked for possible future exchanges between people. The involuntary exchanging of objects such as pick-pocketing is discussed in Cupillard et al. [151] and a real-time implementation of this behavior can be found in Alberto et al. [152]. Other methods have extended the concept of “objects left behind” to analyze higher-level information of objects being “switched,” such as changing hands. A non-contact hand-gesture between people such as waiving was studied in Ke et al. [130]. This event was based on the localization of patio-temporal patterns of each human motion, and uses a shape and flow matching algorithm.

Person – Vehicle Interactions

In general, transit systems involve surveillance of motorized vehicles as well as humans. Spatiotemporal relationships between people and vehicles for situational awareness [153] are the basis for analysis of "the big picture." Operationally-relevant behavior detection (such as human breaking-in or vandalizing a car) has not yet been addressed in the research literature. As mentioned before, the focus of interest for this survey is human behavior recognition; for completeness this following section provides a short general overview on vehicle visual surveillance. (For a complete review of on-road vehicle detection systems, see Sun et al. [154].)

Most existing automated vehicle surveillance systems are based on trajectory analysis. Detected events are abnormal-low-frequency ones (such as U-turns, sudden braking, pedestrians trespassing the street, etc.) [155, 156], or a small group of pre-defined events, such as accidents [157, 158], illegal parking [159], congestion status [160], illegal turns, or lane-driving [161]. Events of interest are commonly learned using Expectation Maximization [162] or modeled using semantic rules [163] similar to the human interpretation of such events and validated using existing data. Trajectory-based approaches have been the subject of significant study, especially in the traffic analysis domain. Common approaches to trajectory analysis are based on Kaman filter [164] [165], dynamic programming [166], and Hidden Markov Models [162]. Discrete behavior profiling has been proposed [167]to avoid tracking difficulties associated with occlusion and noise. There is significant research done in domain-independent anomaly behavior detection [168, 169], as well as events based on group activities [170]. Transit surveillance involves many sub-problems, including classification of different types of vehicles [171, 172, 173], vehicle recognition [174], or discrimination between vehicles and other frequent objects [175], such as pedestrian, bicycles, buses, cars, pickups, trucks, and vans.

PERSON – FACILITY/LOCATION INTERACTIONS

Intrusion or Trespassing

Intrusion or trespassing is defined as the presence of people in a forbidden area. A forbidden area can also be defined in terms of time (such as after hours) or spatial relationships (such as a pedestrian walking close to the train platform edge or walking on the rails). A large number of intrusion-detection algorithms rely on the use of a digital "tripwire." A tripwire typically is a line drawn over the image that separates regions into "allow" and "don't allow" areas. In Spirito et al. [22], Black et al. [23], and Seyve [25], whenever a bottom corner of the bounding rectangle of an object intersects this line (rails in a subway), an intrusion is detected and a warning is given. The warning stops when both corners of the rectangle come back to the allowed area. Intrusion detection is necessary to detect suicidal behavior, such as people jumping on the train tracks. To reduce false positives, often the blob needs to be tracked over time for a given number of frames after intrusion. To mitigate strong illumination changes, edges can be used in the motion extraction process [176]. Trespasser hiding [141] can be defined as a blob disappearing in many consecutive frames, with the blob's last centroid position not close to an area previously defined as a possible "exit area." Access time and motion trajectory have also been shown to be useful for intrusion violation detection using Hidden Markov Models [177].

Another security-sensitive activity similar to intrusion is tailgating (illegal piggy-back entry). Tailgating is a topic that has not received much attention in research but has been implemented in many commercial systems (see Table 2). Rather than strictly detecting an intrusion past a trip wire, illegal entry can occur when a human gains access through a door or gate by staying close to the person or car in front of them, sometimes without the knowledge of the authorized person.

Wrong Direction

Wrong direction occurs when an object is moving in a restricted direction. Typical examples of this behavior are people or crowds breaching security checkpoints at airports and subways or cars driving in wrong traffic lanes. In general, algorithms used to detect wrong direction rely heavily on a tracking algorithm, since successful tracking allows the movement of the object to be easily estimated and later compared with acceptable motion vectors [178]. In some scenarios, the overall crowd characteristics, which do not rely on the tracking of individual objects, may be sufficient [108]. For instance, the movement of large groups of people in an uncommon direction may indicate panic or danger. To automate the process entirely, motion vectors can be calculated in conjunction with a GMM to learn the correct directional patterns of traffic in the scene [179].

Vandalism

Vandalism is defined in Fuentes and Velastin [141] as irregular centroid motion of a blob, combined with detected changes in the background. This definition is also implemented in Ghazal et al. [180] when a blob enters a scene and causes changes in the background or predefined “vandalisable” areas. In Sacchi et al. [181], vandalism is detected in unmanned railway environments using a neural net by detecting erratic or strange behavior of a single person or a group.

Object Stationarity (Object Removal and Object Left Behind)

In this survey, object stationarity refers exclusively to non-animated objects. In transit surveillance systems, objects left behind usually represent suspicious or potentially dangerous elements (such as a suitcase, backpack, etc). Detection of dangerous objects is a critical task that leads to safety and security of the passengers. In 2004 and 2006, object stationarity was one of the events targeted by the Workshop on Performance Evaluation of Tracking and Surveillance (PETS). Most algorithms presented a simple background subtraction to find stationary objects that were not present before. Many other methods have been proposed to deal with objects left behind or removed. In Spagnolo et al. [182], an edge-matching algorithm is used, which compares the current frame to the background model in order to detect objects removed or left behind. In Black et al. [23], a block-based matching algorithm is used to detect stationarity. Each video frame is separated into blocks and classified as background or foreground using frame differences with respect of the training phase. If at any given time a foreground block is not moving, it is then considered to be stationary. There is still quite a lack of research in terms of object stationarity in the context of crowded areas, but Sijun et al. [183] have admitted this weakness and mentioned ways to include crowd segmentation algorithms to improve stationarity detection performance.

STATE-OF-THE-ART DISCUSSION AND FUTURE DEVELOPMENTS

Future developments mentioned in the previous survey [6] include multi-modal data fusion, robust occlusion handling, usage of 3D data, and use of personal identification. In this section, additional potential directions of work are explored. Also, an analysis of the current state-of-the-art behavior understanding algorithms is presented. Research weaknesses are identified, and possible solutions are discussed. The surveyed studies in Table 2 offer an indication to the level of interest in this research area. As shown in Figure 7, it is clear that behavior recognition is an active research topic. In fact, there are three times as many publications in the last three years than the number of all publications found before 2005.

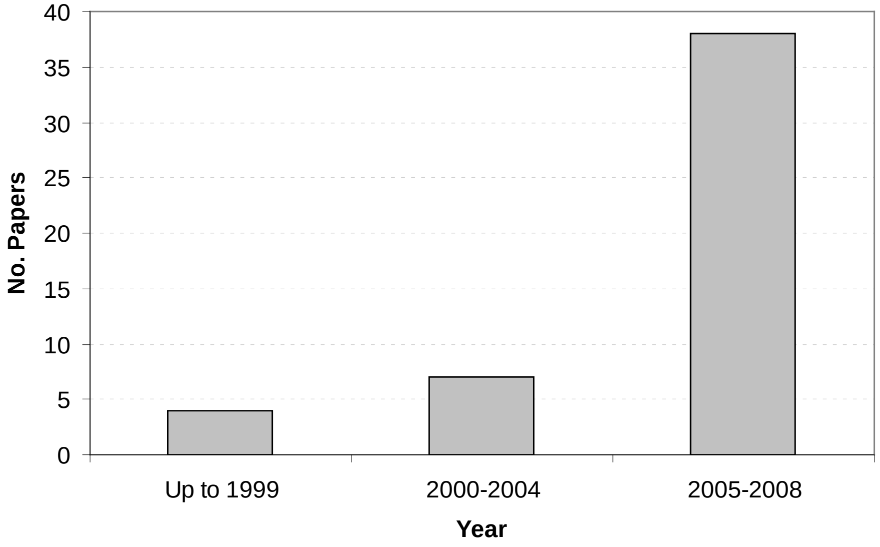


Figure 7 Increasing Interest in Human Behavior Recognition Research

CORE TECHNOLOGY LIMITATIONS

Human behavior algorithms rely heavily on the core technology available. There are many limiting factors to the usability of these core technologies in real transit systems. Implementing analytics on some videos may not be feasible or could be restricted to only a subset of the algorithms available. There are many hardware-related problems such as poor resolution, low frame-rates, or insufficient processing hardware. For instance, crowd monitoring algorithms usually rely on the calculation of optical flow, which requires a moderately high frame-rate and significant processing power. In fact, optical flow often requires special hardware if a real-time solution is needed [6]. In this study, algorithms are separated in terms of processing speed into two groups: real-time and offline processing (Table 2). Nevertheless, in the last decade the image processing community in this context agrees that the definition of real time is not clear even though many researchers use it in their systems [9]. This point brings the biggest concern for creating an accurate assessment of core technology limitations: the lack of independent studies that compares behavior detection performance in transit environments with a common set of dataset and metrics. For instance, although significant progress has been made in object tracking in the last decade, tracking methods usually rely on assumptions that often over-simplify the real problem. Assumptions such as smoothness of motion, limited occlusion, illumination constancy, and high contrast with respect of background [74] effectively limit the algorithms usability in real scenarios within the transit surveillance domain.

EVALUATION FRAMEWORK

Robust evaluation of automatic computer-vision methods is a complicated task. Standard baseline algorithms are required for comparison purposes. These baseline algorithms are usually well known to computer scientists working in related areas of research, but there are no accepted baseline algorithms in behavior recognition for transit applications. Surprisingly, few studies in Table 2 formally compare performance against any other related work, making behavior detection algorithms comparison scarce in the literature. Dealing with new detection tasks that have not been studied previously will clearly require baselines to be developed. In any case, the use of well-known and standard low-level processing techniques is a must. A meaningful study must compare performance with techniques that are likely to work under most circumstances, rather than compare to techniques likely to fail under the scope of interest. Transit data are far from common as are the problems that come along with them. On top of typical problems faced in vision-based surveillance applications, the transit domain faces especially difficult problems, including poor illumination with drastic lighting changes (such as underground stations and tunnels) and heavily crowded scenes. In outdoor transit, weather can also have a significant impact on the quality of the data. A previous study on capturing human motion, which compares over 130 studies, found algorithms to be heavily constrained to assumptions [9] related to movement, environments, and subjects. Nearly a decade later, algorithms still rely on many of the same assumptions. The problem is that performance under these situations is not well specified in the literature. In transit environments, particular concerning are assumptions of camera motion, camera parameters, field of view, background complexity, landmarks, lighting and weather conditions, crowd density, number and severity of occlusions, subject initialization or a-priori information (such as known pose, movement, tight-fitting clothes, etc.), and variability of motion patterns. Going back to a point made earlier, there is a lack of independent studies that attempt to describe the effect of these problems in different transit scenarios; therefore, it is unclear how behavior detection algorithms and commonly used low-level processing methods are affected by some of these domain-specific problems.

STANDARD TERMINOLOGY

It is often assumed that crowds will distribute evenly across the available space. However, that is not necessarily the case in transit areas such as a metro platform, where people are "competing" for space to ensure they get on the next train. The occupancy capacity of a given area depends on the pertinent licensing authority, such as fire or police department, emergency agency, etc. For example, in England, the Communities and Local Government regulations set the limit occupancy for a bar [184] to 0.3 to 0.5m2 per person, but the same regulations do not apply to shopping malls. In image processing, to find a common ground for publications and experimental results, sometimes it is necessary to use standard operational definitions. In Still [185] and Rahmalan et al. [110], definitions based on current practical safety guidelines are used. For example, very low density is defined as people/m^2<0.5, while very high density when people/m^2>2. Other studies use less mathematically-precise definitions such as "overcrowding occurs when too many people congregate within certain location and. congestion is a situation where it becomes difficult for an individual to move within a crowded area" [23]. A common approach is to describe a crowd in terms of the number of individuals in it, as in Marana et al. [34], where the authors define "very low density (0-15 people), low density (16-30 people), moderate density (31-45 people), high density (46-60 people) and very high density (more than 60 people)." Clearly, comparing related work dealing with "crowds" becomes extremely complicated, since there is no widely-accepted standard for defining crowd levels in the literature. Additionally, it is difficult to identify methods that refer directly to similar datasets in terms of crowd density.

DATASETS

This study found across the literature the tendency to not fully specify the dataset used. As shown in Figure 8, most studies, regardless of the review process, chose to not completely disclose the dataset description of their work. This information is necessary when showing the significance of an algorithm and understanding results. Relative improvements over other previously-reviewed publications may be difficult to quantify since a comparison of the datasets cannot be made. It is often unclear what level of empirical validation is behind published techniques. An advantage of using similar or common datasets is that performance scores from different algorithms can be compared directly, as long as the evaluation framework is comparable. In general, transit security data is difficult to come by, due to the difficulty of gathering an adequate supply of valid video sequences containing operationally relevant events [141] and overcoming privacy and security concerns. Initiatives like TRECVID [186] encourage research by providing large dataset collections and uniform scoring procedures. Efforts like this will be required as organizations become interested in comparing behavior detection reliability and results. Nevertheless, some authors using available datasets report concrete results only on very small portions of the dataset, but make reference of general testing on the entire data. Other authors refer to algorithms being able to work without any level of detail on performance, which does not offer researchers in the field with any meaningful performance information. This study found these to be common problems in the literature.

In Figure 8, the dataset description analysis based on 52 transit surveillance-related studies surveyed in this work is shown. "None" refers to studies that do not include any reference to the datataset used. "Complete" indicates a full description is included, that is, quantity and pixel resolution for both training and testing data. "Incomplete" indicates some description but not enough to account for "Complete."

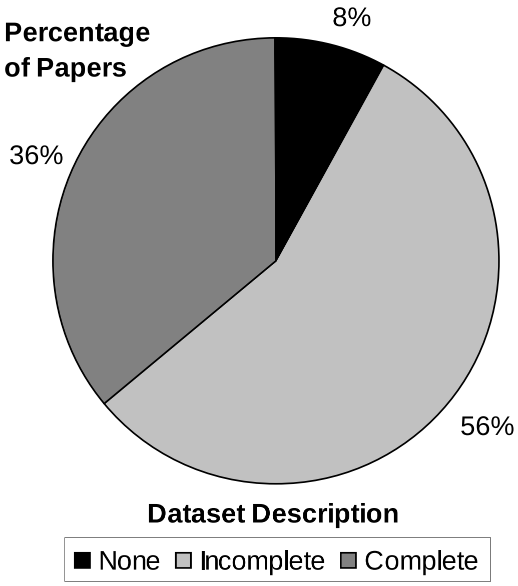


Figure 8 Dataset Description

DISTRIBUTED SURVEILLANCE

Distributed surveillance systems are networks of sensors that can be spread over large regions. Often, a single view of a transit scene could be insufficient to determine certain complex human behaviors. Large networks of cameras and other sensors could interact to form a "bigger picture," which can potentially offer a viable solution to complex problems. Many transit systems have large sensor networks (such as audio, video, motion sensors, smoke detectors, etc.) already in place. In such scenarios, multiple sensors can be used to generate more accurate, complete, and dependable data. For example, camera networks can be used to provide multiple views of a scene, which might diminish the number of tracking occlusions [187]. Also, sensors can often overcome weaknesses of other sensors; for example, fusing color and infrared video can be used to improve tracking through occlusions [188]. There is not much work reported on the integration of different types of sensors in automated video surveillance systems [7]. Multi-modal fusion, such as audio and video [189] or infrared and stereo-vision [190], can potentially offer better scene understanding, thereby improving situational awareness and response time. (For general distributed surveillance, see a detailed survey [7] for more information.)

AERIAL SURVEILLANCE

Moving cameras and mobile surveillance platforms are yet to become an important player in transit surveillance. With much research and commercial interest in unmanned aerial vehicles (UAV) and mobile surveillance platforms, current solutions are not far from being usable as an efficient surveillance platform for transit networks. Early work using surveillance video from UAV [191, 192] describe behavior analysis algorithms for low resolution vehicles to monitor road-block checkpoints (such as avoiding, passing-thru, getting closer, etc.). As aerial surveillance has gained increased interest within the research community, authors have proposed techniques to detect low resolution vehicles [193] and buildings [194] from aerial images. As surveillance techniques using image processing algorithms are created to be used on aerial platforms, tracking-based methods often used in current transit applications will likely have problems with aerial video. Tracking systems have problems with objects following broken trajectories resulting from limited field of view and occlusion due to terrain features. Recent work is being driven by these problems, leading to solutions for problems such as the study of global motion patterns [195] from aerial video. As resolution and video quality increases, transit surveillance including people, vehicles, and behavior analysis is logically the next step.

Table 2 Publications on Behavior Recognition Algorithms Applicable to Transit Surveillance Systems

First Author	*Yr*	*Behaviors*	*Dataset*	*O*	*R*	*C*	*Ref #*
Yasin	08	Bending down, gun shot, jumping up, kicking front, and punching forward	185 videos containing 5 types of motion	N	N	N	[146]
Bissacco	08	Human pose	2950 images of human walking in circle, unspecified resolution.	N	N	N	[138]
Jang	08	Human pose	600 images of unspecified resolution	N	N	N	[140]
Li	08	Crowd counting	Classifier training 1755 positive samples of 32x32px, and 906 for testing. Counting testing 12 minutes of video	Y	N	Y	[121]
Blunsden	07	Walking together, approaching, ignoring, meeting, splitting, and fighting	Unspecified number of videos from CAVIAR. Data described using number of activity points and sequences	N	N	N	[144]
Dong	07	People counting, crowd density	2 videos	Y	Y	Y	[119]
Fathi	07	Human Pose	1008 images (divided into 4 subjects), unspecified resolution	N	N	N	[139]
Ghazal	07	Theft, graffiti, defacing	3 videos	Y	Y	N	[180]
Ke	07	Picking up object, waiving, pushing elevator button	20 minutes of video, 160x120px	Y	Y	Y	[130]
Lee	07	Human pose	Unspecified number of videos, including indoor and outdoor scenes	Y	N	N	[136]
Monteiro	07	Wrong direction	Unspecified number of 320x240 px images	Y	Y	N	[179]
Park	07	Human-vehicle situational awareness	30 minute video	Y	N	Y	[153]
Park	07	Person – person interaction, shaking hands, pointing, standing hand-in-hand	Train 30 images, test 38 images	N	N	N	[149]
Ribnick	07	Thrown objects	Unspecified indoor and outdoor videos	Y	Y	N	[196]
Andrade	06	Crowd behavior: normal, blocked exit, and fallen person	6000 384x288px images	N	N	Y	[126]
Andrade	06	Crowd behavior: normal, blocked exit, and fallen person	Unspecified number of 384x288px images	Y	N	Y	[127]
Andrade	06	Blocked exit	3 simulated 384x288px datasets, train 1 sequence with 2000 frames	N	N	Y	[128]
Bird	06	Abandoned object	3 hours and 36 minutes, 4 videos, 320x240px	Y	Y	Y	[197]
Ferrando	06	Object left behind, object switching	800 images	N	Y	N	[198]
Park	06	approaching, departing, handshaking, pointing, pushing, hugging	Unspecified number of sequences, 320x240px	N	N	N	[148]
Rabaud	06	Crowd density	900 320x240px images, and 1000 640x480px images	Y	N	Y	[124]
Rahmalan	06	Crowd counting	150 200x200px training and 75 testing images	Y	N	Y	[110]
Ribnick	06	Camera tampering	Unspecified indoor and outdoor videos	Y	Y	Y	[199]
Sijun	06	Object ownership , object stationarity	92 training and 45 testing videos	N	N	N	[183]
Velastin	06	Circular and diverging flows, obstacle detection	Unspecified number of 512x512px grayscale videos	N	Y	Y	[108]
Wu	06	People counting, crowd density	70 320x240px images	Y	N	Y	[111]

Angiati	05	Vandalism	2 videos (diurnal and nocturnal) with 7 graffiti drawn	Y	N	N	[200]
Bird	05	Loitering	Train 205 images. Test 30 minutes 720x480px video	N	N	N	[61]
Black	05	Crossing , falling on, proximity, throwing objects to, walking on tracks	Entire CREDS dataset	N	Y	Y	[23]
Fuentes	05	Unattended luggage, intrusion into forbidden areas, falls onto tracks, People hiding, vandalism, fights	Unspecified number of 384x288px color images	N	Y	Y	[141]
Lee	05	Human Pose	PETS 2003 Smart meeting video	N	N	N	[137]
Liu	05	Virtual gate crowd counting, proximity to tracks	1 10 minute video	N	N	Y	[117]
Nascimento	05	Passing, entering, and leaving a storefront in a public area	40 trajectories from 25 movies of about 5 minutes. each	N	N	N	[105]
Schwerdt	05	Abnormal direction of motion, loitering , objects left behind, train presence, and crossing , proximity, walking on tracks	Camera C sequences from CREDS dataset	N	Y	N	[24]
Seyve	05	Crossing, dropping, falling, proximity, throwing object, walking on tracks, trap by train door	Unspecified dataset from CREDS	N	Y	N	[25]
Velastin	05	Overcrowding/congestion, Abnormal direction of motion, loitering, objects left behind, train presence	PRISMATICA live test. Validation of results with at least 200 activity samples	N	Y	Y	[28]
Aubert	04	Loitering, objects left behind	436 stationary situations test cases on gray level 256x256px images	N	N	N	[32]
Fuentes	04	Objects left behind , intrusion, falls, hiding, vandalism/graffiti, fights, attacks	Unspecified number of 384x288px color images	N	Y	N	[201]
Kang	04	Security breaches (i.e., wrong direction)	Dataset not specified	N	Y	N	[178]
Reisman	04	Crowd detection	320x240px video from mobile platform	Y	Y	Y	[109]
Kettnaker	03	Intrusion detection	Training 18 security officer sequences, and 9 cleaning sequences. Testing 15 sequences of normal and 12 of illegitimate behavior on 120x160px color images	N	N	N	[177]
Park	03	Approaching, departing, pointing, standing hand-in-hand, shaking hands, hugging, punching, kicking, and pushing	56 320x240px sequences	N	N	N	[147]
Cupillard	02	Fighting, blocking, forbidden zone, pickpocket	20 sequences	N	N	N	[151]
Sacchi	01	Graffiti, gang behavior: “agitated” and “calm” behavior	270 frames for training, 118 image sequences for testing	N	N	N	[181]
Aubert	99	Queue length estimation	255 measurements from 2 hours of video of airport scenes	N	Y	Y	[31]
Haritaoglu	99	Handbag detection, object exchange	100 320x240px videos	Y	Y	N	[150]
Marana	97	Crowd density estimation	151 train and 149 test images	N	Y	Y	[34]
Yin	95	Crowd density estimation	1 training and 2 testing train station sites	N	N	Y	[129]
Velastin	94	Crowd detection	100 512x512px gray level images	N	Y	N	[202]

(O = Dataset includes outdoor dataset, R = Mentions a real-time implementation, C = Dataset includes crowded scenes)

Table 3 Experimental Results as Stated in Their Respective Publications

Ref #	*Feature*	*Results*
[23]	Blobs	7%-100% TP depending on configuration. 0%-25% FP
[24]	Blobs	Only qualitative results given, no quantitative empirical analysis
[25]	Blobs, motion characteristics	64%-100% TP depending on event. 0%-29% FP
[28]	Edges, motion, blob's position, shape, and trajectory	87.5%-100% TP depending on event. 0%-4% FP
[31]	Motion and intensity	5.9 average. Queue length error in pixels over 255 measurements. Robust low contrast, illumination changes, and crowded scenes
[32]	Level-lines	98% TP, 2% FP
[34]	Intensity texture	53.85-94.44% accuracy depending on type of crowd. Provides an output in terms of a range of densities
[61]	Blob's size, shape and clothing's color	100% TP and 11%FP with 66% tracking accuracy
[105]	Motion	Results shown using penalized log-likelihood by the activity type
[108]	Motion	Overcrowding estimates 95.62% TP and 4% FP. Congestion 98.51% TP and 0.28% FP. Object stationarity 87.5-100% TP and 0-12.5% FP for different conditions including occlusions and pose/position variations
[109]	Optic flow	No empirical analysis
[110]	Grey Level Dependency Matrix (GLDM), Minkowsky Fractal Dimensions (MFD), Translation Invariant Orthonormal Chebyshev Moments (TIOCM)	TIOCM (novel) is compared with MFD and GLDM (see right). Accuracy for TIOCM reported as approx. 86% (based on chart), compared to approx. 35% for MFD, and approx. 80% for GLDM. Results based on morning and afternoon conditions. One operating point is used, and no false alarm rates given.
[111]	Statistical methods (Grey Level Dependency Matrix, GLDM)	Total error is less than 12%. No FP rate is reported
[117]	Motion, Blob's color, position, shape, and trajectory	Only visual sample results, no empirical analysis
[119]	Silhouettes of connected blobs, Fourier descriptors,	Confusion matrix and ROC given. Overall accuracy reported as 94.25%
[121]	Histogram of oriented gradients	Shown by ROC analysis
[124]	Feature tracking based on KLT, connectivity graphs	Average error ranges from 6.3% to 22%. No FP rates reported.
[126]	Features extracted from optical flow	Results shown using the log-likelihood mean and standard deviation, before and after an event has occurred
[127]	Features extracted from optical flow	Results shown using the log-likelihood mean and standard deviation, before and after an event has occurred
[128]	Features extracted from optical flow	Results shown using the log-likelihood mean and standard deviation, before and after an event has occurred
[129]	Number of pixels classified as pedestrian	1-2 difference (in persons) between manual and automatic pedestrian count
[130]	Spatiotemporal shape contours, optical flow	Shown by Precision and Recall graph, one for each event detected
[136]	Motion	Results are given based on the error between detected joints and actual joints (in pixels). Average error reported (per joint) is 9.86 pixels (which translates to 7-12 cm away from actual joint position for their dataset)
[137]	Regions, texture, skin color	Results are given based on the error between detected joints and actual joints (in pixels). Average error is 24.99 pixels
[138]	Silhouette, Harr, edges and lines	Accuracy reported as mean error, which ranges from 0.27 to 0.30
[139]	Motion	Results are given based on the error between detected joints and actual joints (in pixels). The mean error reported ranges from 15 pixels to 30 pixels depending on joint
[140]	Motion, landmarks	Results shown as the proportion of the principle axis
[141]	Motion, blob's color, centroid, position, height, and width	Only visual sample results, no empirical analysis
[144]	Eight motion features based on speed and direction between two people	Overall accuracy given for two scenarios: frame based, and sequence based. Frame based results are as follows: Walk Together – 100%, Approach – 46.9%, Ignore 85.1%, Meet – 100%, Split – 100%, Fight – 57.1%. Total accuracy is 90.8%. Sequence based as follows: Walk Together – 100%, Approach – 50%, Ignore 100%, Meet – 100%, Split – 100%, Fight – 100%. Total accuracy is 90.4%. Results are based on one operating point, with no FP rates reported

Ref #	***Feature***	***Results***
[146]	Calculates Hu Moment Invariants (7), and Euclidean distance from Binary image	Approach is compared to four other classifiers (Fuzzy-K Nearest Neighbor, Mahalanobis Distance, Quadratic Bayes Gaussian, and Linear Bayes Gaussian). Accuracy rate is 87.6%. Algorithm is the fastest running compared to all other classifiers
[147]	Individual body-part motion	50 to 100% (78% average) TP depending on the event, no FP rate reported
[148]	Blobs, contours, intensity	Overall score for one operating point (given) is 86%. Individual accuracy range 68%-100% depending on event. No FP rate is reported
[149]	Blobs, individual body part motion, normalized feature vector which is based on body part distances.	Overall accuracy rate of 86%. Shaking hands and standing hand-in-hand detected 100%, pointing 74%. No FP rates reported
[150]	Motion periodicity and silhouette symmetry	Shown by ROC analysis. Approximated operating point at 90% detection, 20% FA
[151]	Motion, blob's centroid, position, height, and width	70-95%, 3% FP
[153]	Motion features generated from planar homography using 4-point algorithm.	Precision and Recall rates given. One operating point approximated at 93% precision and 95% recall
[177]	Access time and motion trajectories	Normal behavior 100% detection, Unusual behavior 75% detection at regular "business hours" and 100% detection at "unusual hours." No FP rates reported
[178]	Motion, color and shape	Only qualitative results, no quantitative empirical analysis
[179]	Motion calculated from optical flow, Harr-like features used to distinguish motion	Only qualitative results given, no quantitative empirical analysis
[180]	Features are generated by motion	Only qualitative results given, no quantitative empirical analysis
[181]	Motion, blob's area, perimeter, centroid, and speed	About 84% TP, 9% FP
[183]	Eigenfeatures	Accuracy ranges from 78% to 93.7%, depending on event. Misclassification rates are given
[196]	Motion history, blobs, compactness, density	Accuracy ranges from 68 to 85% depending on size of object thrown relative to camera. Overall (average) accuracy is 74%. Results are based on one operating point
[197]	Blobs	Evaluation based on Percent Events Detected (PED) and Percent Alarms True (PAT), analysis of PED/PAT results with respect to time given. Overall score for one operating point (given) ranges from 42% to 67%
[198]	Motion history, blobs, color, Hu-moments	Results are given based on low and medium scene complexity. Low scene complexity detection rate ranges from 75% to over 99%, with a FP rate that ranges from less than 0.05% to 8.3%. Medium scene complexity TP rate ranges from 83% to 98.6%, with a FP rate ranging from 1.5% to 9.5%.
[199]	Image dissimilarity based on RGB and gradient histogram	Evaluation based on Percent Events Detected (PED) and Percent Alarms True (PAT). Overall accuracy reported as PAT 79.2% and PED 95% with 5 FP and 3 missed events
[200]	Motion, blob's position	Diurnal: 65-97% TP, 5% FP. Nocturnal: 0-91% TP, 6% FP
[201]	Motion, blob's color, centroid, position, height, and width	Only visual sample results, no empirical analysis
[202]	Motion	Results shown through polar plots where the direction angles are divided into a discrete range

(TP = True Positives, FP = False Positives, ROC = Receiver Operating Characteristic Curve)

CHAPTER 3
EVALUATION FRAMEWORK

The purpose of an evaluation framework is to statistically present a meaningful and objective comparison of different techniques used in surveillance applications. For this purpose, the Detection and Tracking Evaluation (DATE) software [11] is used to evaluate tracking algorithms in transit scenes. The performance measures are generated from the spatial overlap between the ground truth and the output of the tracking algorithm. These measures can be generated from a rigid or course level of overlap; at the rigid level, a one-to-one mapping is required from the ground truth annotation and the system output, while the course level will use a weight or threshold to determine a satisfactory level of overlap. Both the Video Analysis and Content Extraction (VACE) and the Performance Evaluation of Tracking and Surveillance (PETS) [12] metrics can be computed with this software. A comparative study of these metrics can be found in Manohar et al. [203].

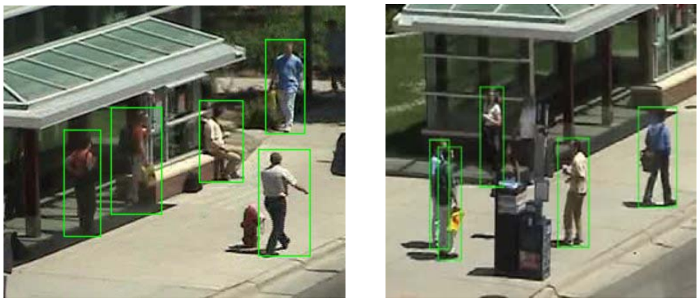


Figure 9   Example Ground Truth Images Used for Tracking

The first step for computing the evaluation scores is to annotate the ground truth. This was done using the VIPER [204,205] ground truth annotation tool. Some example ground truth images are given in Figure 9. Next, using the USF DATE (version 5) software, performance evaluation scores are computed between the ground truth images and the system output. The scores in Table 4 are based on the Sequence Frame Detection Accuracy (SFDA), which is a rigid frame level measurement that accounts for number of aligned, mal-aligned and missed tracking boxes, false alarms and spatial fragmentation. It is done separately for each frame in the sequence, and the scores are then summed and normalized.

The bounding box around each object in a scene can be different for two different outputs, yet be equally accurate. Some algorithms are entirely dependent on the object being detected and not concerned with the spatial coordinates of the objects. In these cases, the alignment can be relaxed to generate a more realistic measure of performance. The general idea is that if a portion of the tracking boxes overlap, then it is fully accurate. The exact portion of overlap can be defined by the user in the software. In this sample results, 25 percent is used.

Table 4 Performance Scores between Annotation and Ground Truth
(OLB=Object Left Behind)

	OLB (Ground Truth)	*Breach (Ground Truth)*
OLB (annotator)	100% (CLEAR)	n/a
Breach (annotator)	n/a	85% (CLEAR)

Other than overlap based performance scores, the USF DATE (version 5.2.0) software also provides other useful information about the event detection performance as well. Instead of using the overlap between bounding boxes used in tracking, it is also possible to use other properties such as the centroid of either the object being tracked or the bounding box. Diagnostic measures are also available for pinpointing areas that were missed or where false alarms were given.

PERFORMANCE MEASURES

The measures used to generate the performance scores between the ground truth and the system results were proposed and discussed in detail in Yin et al. [12]. NCTR researchers slightly changed the definitions to fit the needed requirements. For instance, objects are now referred to as events. To be clear, the following notion are used:

- G_i denotes the i^{th} ground truth event and G_i^t denotes the i^{th} ground truth event in t^{th} frame.
- D_i denotes the i^{th} detected event and D_i^t denotes the i^{th} detected event in t^{th} frame.
- $N_G^{(t)}$ and $N_D^{(t)}$ denotes the number of ground truth events and the number of detected events in the frame *t* respectively.
- N_G and N_D denotes the number of unique ground truth events and the number of unique detected events in the given sequence respectively. Uniqueness is defined by the object IDs.
- N_{frames} is the number of frames in the sequence.

- N^{i}_{frames}, depending on the context, is the number of frames the ground truth event (G_i) or the detected event (D_i) existed in the sequence.
- N_{mapped} is the number of mapped ground truth and detected events in a frame or whole sequence, depending on the context (detection / tracking).

The Sequence Frame Detection Accuracy (SFDA) is frame-based measure based on the principle that the two corresponding objects (G_i and D_i) should overlap. Any fragmentation caused by spatial alignment, missed objects, or false alarms will reduce the accuracy of the measure. First, the measure used for a single frame (FDA) is addressed.

$$FDA(t) = \frac{Overlap\_Ratio}{\left[\frac{N_G^{(t)} + N_D^{(t)}}{2}\right]}$$

Where, $Overlap\_Ratio = \sum_{i=1}^{N_{mapped}} \frac{\left|G_i^{(t)} \cap D_i^{(t)}\right|}{\left|G_i^{(t)} \cup D_i^{(t)}\right|}$

Hence, N^{t}_{mapped} is the number of mapped events, with minimal special overlap. To measure the entire sequence (SFDA), the FDA is normalized using the number of frames where the events were detected. And so,

$$SFDA = \frac{\sum_{t=1}^{t=N_{frames}} FDA(t)}{\sum_{t=1}^{t=N_{frames}} \exists(N_G^t \vee N_D^t)}$$

CHAPTER 4
COMMERCIAL SYSTEMS

There are many professional-grade surveillance systems that can be used by residential, commercial, government, and law enforcement agencies. Many of these systems now include analytic software capable of some level of event detection. The capabilities of commercial surveillance systems have increased significantly over the last decade. Early systems allowed clients to record only when motion was detected in regions of interest or when an external sensory device was triggered. Such technology was often limited to indoor scenes, as different weather conditions would frequently trigger false alarms. More recently, newer and more powerful analytic systems include environmental modeling, which have helped resolve such limitations. For instance, instead of triggering an alarm that is based only on motion within a user specified region of interest, the client is able to specify defining attributes of the object creating the motion, such as dimensions and shape. This, in turn, allows efficient retrieval of pre-defined events from large amounts of video. Moreover, the actual storage of the video can be reduced significantly if the client chooses to record only during such events.

Surveillance footage can be used proactively to detect suspicious events in real-time or reactively used to review archived data. Clearly, manual real-time surveillance of large transit systems is not usually possible. For example, New York metro, according to 2006 statistics [206], is the busiest metro in the United States and second busiest in the world. It has a total of 468 stations and 1.49 billion riders per year, 4.9 million per day.

Monitoring *objects left behind* (left baggage, briefcase, purse, etc.) in real-time footage would require thousands of analysts, a scarce and costly resource. Clearly, the ability to monitor real-time video for specific events would provide dramatic surveillance capabilities to transit agencies, which would become a great asset technology to deter and respond to accidents, crime, suspicious activities, terrorism, and vandalism. This technology is not limited to visual cues on security monitors; other common features include automatic messaging to Personal Display Assistants (PDAs) or other devices when an event has been detected. This would, in turn, allow key personnel in close proximity to further investigate the situation where the event took place.

During the last decade, human event detection has become one of the most active research topics in computer vision. After the catalytic terrorist attacks of September 11, 2001, against the United States, technology to automate surveillance security has grown exponentially. Recent reports from market researchers in the global technology industry [207] have shown a massive increase in the market, from 67.7 million in 2004 to 839.2 million in 2009. Surveys of state-of-the-art research dealing with event detection for transit systems [2] have also emphasized the importance of this topic among computer scientists. As public transit agencies are under mounting pressure to provide a safe and secure environment for passengers and staff, they are likely to start embracing this new generation of technology. As capabilities advertised by commercial providers increase, the necessity for an independent evaluation of such capabilities becomes more and more prominent. Currently, there are no published efforts in the literature or independent data that can sustain the providers' claims. Furthermore, it is not clear how typical problematic conditions of mass transit systems, such as heavy traffic, crowded areas, detrimental weather effects, and drastic illumination changes, could affect performance. Additionally, without independent verification studies, there is no way to determine strict technical terminology commonality; therefore, comparing performance across platforms was not possible in this study. For example, regarding the detection of loitering behavior, Table 5, which is based on the information available on each of the respective vendor's websites as of March 2009, indicates that almost 2/3 of vendors advertised loitering detection capabilities. Only software products offering software analytics are listed. Taking into account that, as discussed earlier, loitering is detected over long periods of time, including likely situations of subjects leaving the scene or being frequently occluded, it is unclear if any of the systems listed in Table 5 can achieve the same results as in Bird et al. [61]. In fact, based on direct discussions with some vendors, it was made clear that systems in general have significant limitations with respect to camera placement, image quality and resolution, lighting conditions, occlusions, object contrast and stationarity, and weather.

COST

Cameras can be analog or digital. Analog cameras are less expensive (around $250 each), but they tend to incur in higher deployment labor and infrastructure costs, since they use a technology that is already becoming obsolete. Digital cameras (IP cameras) are more expensive (around $1,800), but they are more flexible, manageable over networks, durable, and easier to deploy, which greatly reduces installation costs. The general approach is to use the cameras to gather video, send the raw data across a network, then store and process the data on a server.

Table 5 Behavior Recognition Summary Advertised by Commercial Providers on Their Websites

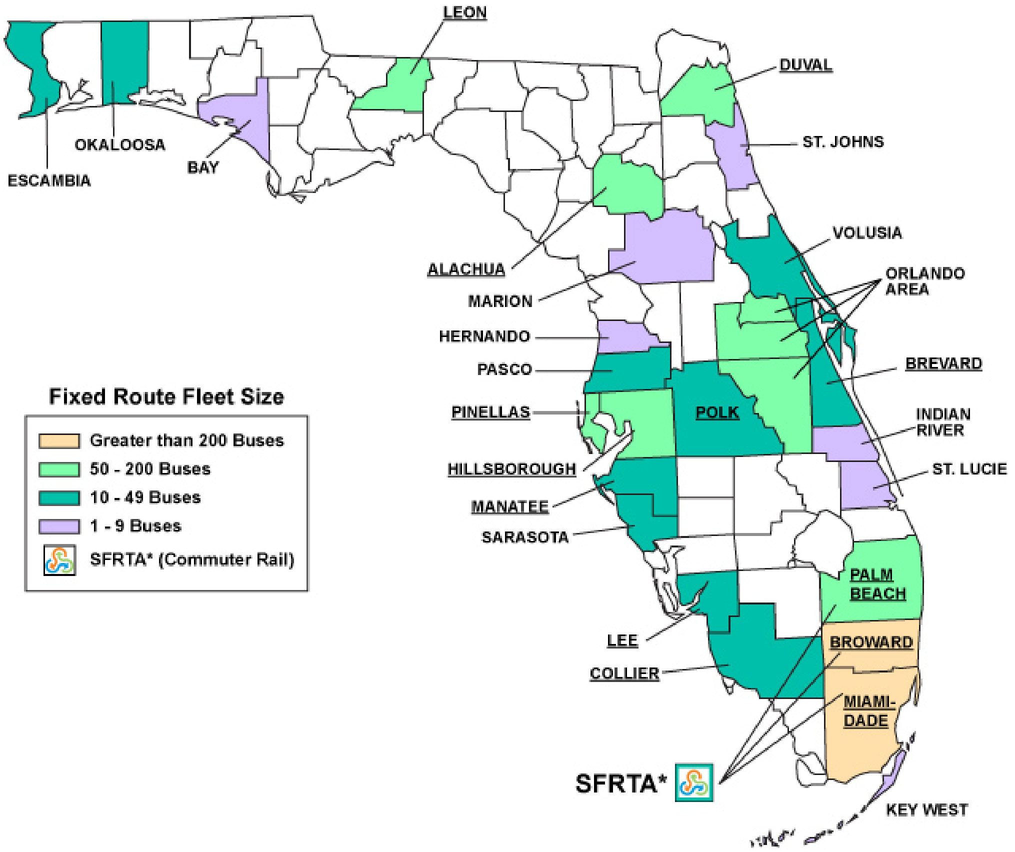
Ref #	Manufacturer	Object Tracking	Breach	Loiter	Crowd Analysis	Stance Change	Object Left	Object Removal
[208]	Agent Vi	✔	✔	✔	✔		✔	
[209]	Aimetis Corp	✔	✔	✔	✔		✔	✔
[210]	Cernium Corp	✔	✔	✔	✔		✔	
[211]	Eptascape, Inc	✔	✔	✔	✔		✔	✔
[212]	Honeywell International, Inc	✔	✔	✔	✔	✔	✔	✔
[213]	Indigo Vision	✔	✔		✔		✔	✔
[214]	Intelliview Technologies Inc	✔	✔					
[215]	Intellivision	✔	✔		✔		✔	✔
[216]	IPSOTEK Ltd	✔	✔	✔	✔	✔	✔	✔
[217]	March Networks	✔	✔	✔	✔	✔	✔	
[218]	Mate Intelligent Video	✔	✔	✔	✔		✔	✔
[219]	Object Video	✔	✔	✔	✔		✔	✔
[220]	SightLogix Inc	✔	✔					
[221]	Verint	✔	✔	✔	✔		✔	✔
[222]	Vidient	✔	✔	✔	✔		✔	✔
[223]	Nice Systems	✔	✔		✔		✔	
[224]	TrueSentry, Inc.	✔	✔	✔	✔		✔	✔
[225]	Ioimage, Ltd	✔	✔	✔			✔	✔

Clearly, analytics software prices will vary, depending on the vendor. The number of events to be detected will also affect the price. The system integrator will likely determine the final selling price based on a competitive bid. Assuming cameras are already installed (either using existing CCTV or acquiring new equipment), the overall cost per channel (for each analytics-capable camera) is roughly $1,700 - $2,100 for analog cameras and $1,900 - $2,300 for digital cameras (based on a small survey of commercial providers in Florida). Additionally, a server is required to host the data and will cost around $5,000. All prices provided thus far include installation fees, and it is worth noting that discounts and bulk rates will most likely apply.

CHAPTER 5
SURVEY OF SOFTWARE ANALYTICS USE IN FLORIDA

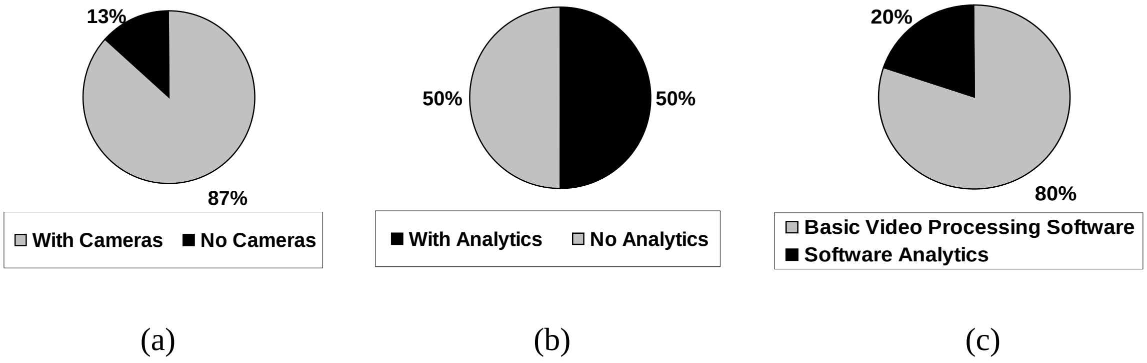
To get a clear picture of the use of video analytics, a survey of the largest transit agencies in Florida was conducted by NCTR researchers. All transit agencies involved in this survey are shown in Figure 10. The survey includes the largest transit agencies in the state based on classification by the Florida Department of Transportation [226], which is based on the agencies' fixed-route fleet size, from largest to smallest. The survey includes only agencies with a fixed-route fleet size of more than 9 buses. The response rate of the survey was as follows: large - 100% (2/2), medium - 100% (7/7), and small - 55% (6/11). The purpose of this study was to relate the state-of-the-art and the current effective use of the analytics technology. Complete data for this survey cannot be released due to the safety sensitivity of the data.

Most large transit agencies in Florida already have CCTV systems available for surveillance monitoring purposes (Figure 11). Only labor-expensive manual forensics is used on archived video to review reported incidents. New Jersey Transit, the largest statewide transit agency in the United States, currently uses real-time video analytics in conjunction with its CCTV systems to detect unattended packages [227] in its facilities. Only 20 percent of Florida agencies are agencies using any form of video analytics (forensic or real-time) for surveillance purposes. At the time of the survey, no agency in Florida was considering evaluating or deploying analytic systems, reporting that budget constraint was a limiting factor. Existing CCTV systems can potentially be used to



Underlined agency names correspond to those included in the security survey.
Map Source: Florida Department of Transportation, "Trends and Conditions, Pocket Guide 2007."

Figure 10. Transit Agencies in Florida



(a) (b) (c)

a) Agencies currently using CCTV systems for surveillance. (b) Transit agencies' responses "Does your camera system include video analytics (i.e., software to automatically detect accidents, theft, or any other suspicious event)?" (c) Reported CCTV systems that currently include some form of video analytics capabilities.

Figure 11. Ownership of CCTV Systems and Video Analytics in Transit Agencies in Florida

deploy analytics software solutions, significantly reducing the investment cost. Another reason for not evaluating analytics software is the misconception that there are no previous incidents that would have benefited from analytics software. But, as shown in Table 1, there are many suspicious behaviors that current analytic systems can detect in real-time, which are most likely being missed in day-to-day operations.

The security survey was distributed to the person responsible for safety and security at the transit agencies. As shown in Figure 11, agencies confuse manual video analytics (human operators manually review archived video) with software analytics (software that automatically detect pre-defined events) when asked “Does your camera system include video analytics (software to automatically detect accidents, theft, or any other suspicious event)? Fewer than 20 percent of the agencies in Florida currently have some sort of software analytic capabilities, and it is unclear to what extent they are being used or if they are being used at all.

CHAPTER 6
IMPACT OF ANALYTIC SOFTWARE

A surveillance systems equipped with analytic software has many benefits; primarily, resource and manual personnel intensive work becomes automated. This directly leads to potential decreases in the resources such as cost and labor and an increase in awareness (safety and efficiency). Post-event detection also becomes available and useful for finding evidence in forensic investigations.

LABOR AND COST REDUCTIONS

With analytic software, the necessity for continually monitoring video feeds can be reduced significantly. It may also provide a solution for transit scenarios that are far too large or busy to be completely monitored by human operators. Analytic software can be used to assist a single operator when searching for evidence in large amounts of previously-recorded video data. Previously, this would have required many operators working in parallel. Also, human-prone errors and false alarms are minimized since alarms would be triggered only by the continuous automatic surveillance system.

EFFICIENCY

In transit scenarios, increases in situational awareness would directly benefit the safety and efficiency of both the passengers and the security personnel on the ground.

For instance, alerts can be provided if long queues at a ticket booth are detected or if crowds become too heavy or show irregular behavior. Early warnings can also be issued before events occur. For example, if someone is heading towards a prohibited location, an alert can be provided before the subject actually reaches his/her destination. Furthermore, a single operator can monitor larger areas by taking the appropriate action when a suspicious behavior or alarm is triggered. Decision making also becomes easier since the event can be replayed immediately on command, rather than second-guessing what may have been seen, and unnoticed behavior of concern becomes less common.

Security

When criminal activity or a threat is detected, security personnel and the proper authorities can be provided with real-time information when assisting the situation. Various alerts can be set up, triggered by pre-defined, operationally-relevant events. Information can be disseminated using text messaging, on-screen alerts, email, geo-coded maps, pictures, and video. The faces of detected criminals can help pinpoint further appearances in past, current, or future video data. Attention-intensive activities such as object removal or object left behind will be detected by the system immediately instead of possibly being unnoticed, resulting in a delayed reaction by a surveillance operator.

Drawbacks

While it is clear that video analytics can offer many advantages over traditional CCTV systems, there are some concerns that should be addressed [22]. Video analytic systems may be vulnerable to environmental variables, such as detrimental lighting conditions and weather (see next section). These adverse conditions can trigger false alarms, which may become a source of frustration for the user. Another drawback with video analytics is that events must be pre-defined, so events that have not been defined will not be detected. Conversely, a human analyst may use judgment and training to determine if an alarm should be raised for a wider range of scenarios. Video analytic algorithms are often sensitive to parameters and initial calibration. Event detection performance typically depends on this calibration process. It is difficult to achieve a good balance between event detection and false alarms. Typically, a higher detection rate produces a higher false alarm rate, and vice-versa. Additionally, some video analytic implementations may require the system to be re-calibrated over time. For example, outdoor scenarios can change drastically depending on seasonal effects (such as leaves, rain, snow) or even the time of day (such as the shadow of a building being present in the afternoon but not in the morning). Hence, the initially high deployment cost and additional recurring costs to maintain and support the system over time may deter many potential users. This becomes even more valid since only sparse research is available that compares actual capabilities with advertised capabilities. The lack of independent verification of commercial products represents a great liability for transit agencies. Agencies like the Metropolitan Transit Authority have attempted to deploy camera systems costing over $300 million, as reported in the *New York Times* on May 28, 2009 [228]. Transit authority officials indicated that the system was not living up to its promise. This situation is likely to recur since there is no formal independent evaluation of such systems. In simple terms, no studies corroborate the vendor’s performance claims or indicate a relative performance comparison across different available products.

CHAPTER 7
CONCLUSIONS

Public transit agencies are under mounting pressure to provide a safe and secure environment for their passengers and staff on their buses, light-rail, subway systems, and transit facilities. Transit agencies are increasingly using video surveillance as a tool to fight crime, prevent terrorism, and increase the personal safety of passengers and staff. Visual surveillance for transit systems is currently a highly active research area in image processing and pattern recognition. The number of studies published in the last three years outnumbers all previous related literature three-fold.

Included in this report are an overview of state-of-the-art developments on behavior recognition algorithms for transit visual surveillance applications and a literature sample of 52 studies on state-of-the-art strengths and weaknesses. Analysis includes behaviors, datasets, and implementation details. A strategy is presented that classifies these studies by the targeted human behavior, including single person or no interaction, multiple person interactions, person-vehicle interactions, and person-facility/location interactions.

In this report, a brief overview of the core technologies (all pre-processing steps before behavior recognition) has been provided. There are many well-known limitations in the core technologies that should be addressed, including sensitivity to poor resolution, frame-rate, drastic illumination changes, detrimental weather effects, frequent occlusions, and other common problems prevalent in transit surveillance systems. Consequently, improved core technology algorithms are needed to increase the reliability of human behavior recognition. During the last decade, numerous methods for evaluating core technologies have been proposed. There are no standard evaluation methods for human behavior recognition. Creating standard evaluation tools includes defining a common terminology and generating operationally similar datasets. For example, a bus and a metro station can both be "crowded." But operationally, the "crowds" in both situations are very different. Thus, without a standard precise definition of "crowd," formal comparisons become a very difficult task.

A comparison of event detection capabilities across commercial providers is presented in this report. A survey of the largest transit agencies in Florida is used to identify the current use of analytic software in public transit. Data suggest that fewer than 20 percent of agencies have some sort of software analytics capabilities. Furthermore, there is no indication that any of these agencies are using the software to its full extent. A formal, independent evaluation of commercially-available systems for event detection currently does not exist. However, the means for performing such an evaluation do exist in the research literature. The evaluation framework used in academic research could be used to evaluate commercial systems at the event level. A meaningful and robust evaluation would allow public transit agencies to objectively compare commercial systems and evaluate product capabilities for their specific needs.

Vast amounts of untapped information are present in surveillance video footage, which can be exploited for automatic behavior detection, and a large gap exists between the analytical skills of a security guard and state-of-the-art image processing algorithms. On the other hand, there is a never-ending struggle to increase security personnel effectiveness over long periods of time while reducing labor costs.

REFERENCES

[1] Official website for UrbanEye. Available at: http://www.urbaneye.net/

[2] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: a survey,” ACM journal of computing surveys, vol. 38, no. 4, 2006.

[3] Official website for Metropolitan Transportation Authority. Available at: http://www.mta.info

[4] Official website for Moscow Metro. Available at: http://www.mosmetro.ru

[5] N. Sulman, T. Sanocki, D. Goldgof, and R. Kasturi, “How effective is human video surveillance performance?,” Int. Conference Pattern Recognition, pp. 1-3, 2008.

[6] W. Hu, T. Tan, L. Wang, and S. Maybank “A survey on visual surveillance of object motion and behaviors,” IEEE Trans. Systems, Man, and Cybernetics Part C, vol. 34, no. 3, pp. 334-352, 2004.

[7] M. Valera and S.A. Velastin, “Intelligent distributed surveillance systems: a review.” IEEE Proc. Vision, Image and Signal Processing, vol. 2, pp. 192-204, 2005.

[8] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O’Brien, and D. Ramanan. “Computational studies of human motion: Part 1, Tracking and Motion Synthesis,” Foundations and trends in computer graphics and vision, vol. 1, no. 2/3, 2005.

[9] T.B. Moeslund and E. Granum, “A survey of computer vision-based human motion capture,” Computer Vision and Image Understanding, vol. 81, pp. 231-268, 2001.

[10] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L.Wixson, “A system for video surveillance and monitoring,” Carnegie Mellon University Technical Report, CMU-RI-TR-00-12, 2000.

[11] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, “Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 319-336, 2009.

[12] F. Yin, D. Makris, and S.A. Velastin, “Performance evaluation of object tracking algorithms,” IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, 2007.

[13] J. Black, T. Ellis, and P. Rosin, “A novel method for video tracking performance evaluation,” IEEE Int. Workshop on Performance Analysis of Video Surveillance and Tracking, pp. 125-132, 2003.

[14] C. Erdem and B. Sanku, “Performance evaluation metrics for object-based video segmentation,” X European Signal Processing Conference, 2000.

[15] B. Georis, F. Bremond, M. Thonnat, and B. Macq, “Use of an evaluation and diagnosis method to improve tracking performances,” IASTED Int. Conference on Visualization, Imaging and Image Processing, 2003.

[16] V.Y. Mariano, J. Min, J-H. Park, R. Kasturi, D. Mihalcik, D. Doermann, and T. Drayer, “Performance evaluation of object detection algorithms,” IEEE Int. Conference on Pattern Recognition, pp. 965-969, 2002.

[17] L.M. Brown, A.W. Senior, Y-L. Tian, J. Connell, A. Hampapur, C-F. Shu, H. Merkl, and M. Lu, “Performance evaluation of surveillance systems under varying conditions,” IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, 2005.

[18] D. Doermann and D. Mihalcik, “Tools and techniques for video performances evaluation,” IEEE Int. Conference on Pattern Recognition, pp. 167-170, 2000.

[19] S. Muller-Schneiders, T. Jager, H.S. Loos, and W. Niem, “Performance evaluation of a real time video surveillance system,” IEEE Int. Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 137-143, 2005.

[20] T. List, J. Bins, J. Vazquez, and R.B. Fisher, “Performance evaluating the evaluator,” IEEE Int. Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 129-136, 2005.

[21] F. Ziliani, S. A.Velastin, F. Porikli, L. Marcenaro, T. Kelliher, A. Cavallaro, and P. Bruneaut, “Performance evaluation of event detection solutions: the CREDS experience,” IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 201-206, 2005.

[22] M. Spirito, C. S. Regazzoni, and L. Marcenaro, “Automatic detection of dangerous events for underground surveillance,” IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 195-200, 2005.

[23] J. Black, S. A. Velastin, and B. Boghossian, “A real time surveillance system for metropolitan railways,” IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 189-194, 2005.

[24] K. Schwerdt, D. Maman, P. Bernas, and E. Paul, “Target segmentation and event detection at videorate: the EAGLE project,” IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 183-188, 2005.

[25] C. Seyve, “metro railway security algorithms with real world experience adapted to the RATP dataset,” IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 177-182, 2005.

[26] Performance Evaluation of Tracking and Surveillance official website. Available at: http://www.cvg.rdg.ac.uk/slides/pets.html

[27] J. Aguilera, H. Wildenauer, M. Kampel, M. Borg, D. Thirde, and J. Ferryman, “Evaluation of motion segmentation quality for aircraft activity surveillance,” IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 293-300, 2005.

[28] S.A. Velastin, B.A. Boghossian, B.P.L. Lo, J. Sun, and M.A. Vicencio-Silva, “PRISMATICA: toward ambient intelligence in public transport environments,” IEEE Trans. Systems, Man, and Cybernetics Part A, vol. 35, no. 1, pp.164-182. 2005.

[29] C. Carincotte, X. Desurmont, B. Ravera, F. Bremond, J. Orwell, S.A. Velastin, J.M. Odobez, B. Corbucci, J. Palo, and J. Cernocky, “Toward generic intelligent knowledge extraction from video and audio: The EU-funded caretaker project,” The Institution of Engineering and Technology Conference on Crime and Security, pp. 470-475, 2006.

[30] C.I. Attwood and D.A. Watson, “Advisor-socket and see: lessons learnt in building a real-time distributed surveillance system,” IEEE Intelligent Distributed Surveillance Systems, pp. 6-11, 2004.

[31] D. Aubert, “Passengers queue length measurement,” IEEE Int. Conference Image Analysis and Processing, pp.1132–1135, 1999.

[32] D. Aubert, F. Guichard, and S. Bouchafa, “Time-scale change detection applied to real-time abnormal stationarity monitoring,” Real-Time Imaging, vol. 10, no. 1, pp. 9-22, 2004.

[33] L. Khoudour, J.P. Deparis, J.L. Bruyelle, F. Cabestaing, D. Aubert, S. Bouchafa, S.A. Velastin, M.A. Vicencio-silva, and M. Wherett, “Project cromatica,” IEEE Int. Conference on Image Analysis and Processing, 1997.

[34] A.N. Marana, L.F.Costa, S.A.Velastin, and R.A. Lotufo, “Estimation of crowd density using image processing,” IEEE Colloquium on Image Processing for Security Applications, 1997.

[35] V.I. Pavlovic, R. Sharma, and T. S. Huang, “Visual interpretation of hand gestures for human-computer interaction: a review,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 677-695, 1997.

[36] B. Fasel and J. Luettin, “Automatic facial expression analysis: a survey,” Pattern Recognition, vol. 36, no. 1, pp. 259-275, 2003.

[37] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, “Visual modeling with a hand-held camera,” Int. Journal of Computer Vision, vol. 59, no. 3, pp. 207-232, 2004.

[38] T. Osawa, W. Xiaojun, K. Wakabayashi, and T. Yasuno, “Human tracking by particle filtering using full 3D model of both target and environment,” Int. Conference on Pattern Recognition, vol. 2, pp. 25-28, 2006.

[39] A. Dominguez-Caneda, C. Urdiales, and F. Sandoval, “Dynamic background subtraction for object extraction using virtual reality based prediction,” Electrotechnical Conference (MELECON), pp. 466-469, 2006.

[40] E. Stoykova, A.A. Alatan, P. Benzie, N. Grammalidis, S. Malassiotis, J. Ostermann, S. Piekh, V. Sainov, C. Theobalt, T. Thevar, and X. Zabulis, “3-D time-varying scene capture technologies—A Survey,” IEEE Trans. Circuits and Systems for Video Technology, vol. 17, no. 11, pp. 1568-1586, 2007.

[41] J. Heikkila and O. Silven, “A real-time system for monitoring of cyclists and pedestrians,” IEEE Workshop on Visual Surveillance, pp. 74-81, 1999.

[42] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” IEEE Int. Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 246-252, 1999.

[43] G. Halevy and D.Weinshall, “Motion of disturbances: detection and tracking of multibody non-rigid motion,” IEEE Int. Conference on Computer Vision and Patter Recognition, pp. 897-902, 1997.

[44] R. Cutler and L. Davis, “View-based detection and analysis of periodic motion,” Int. Conference on Pattern Recognition, pp. 495-500, 1998.

[45] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, “Wallflower: Principles and practice of background maintenance,” IEEE Int. Conference on Computer Vision, pp. 255-261, 1999.

[46] R.J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, “Image change detection algorithms: a systematic survey,” IEEE Trans. on Image Processing, vol. 14, no. 3, pp. 294-307, 2005.

[47] V. Jain, B.B. Kimia, and J.L. Mundy, “Background modeling based on subpixel edges,” IEEE Int. Conference on Image Processing, vol. 6, pp. VI 321-324, 2007.

[48] S-C. Cheung and C. Kamath, “Robust background subtraction with foreground validation for Urban Traffic Video,” EURASIP Journal on Applied Signal Processing, vol. 14, pp. 1-11, 2005.

[49] M. Hansen, P. Anandan, K. Dana, G. Van Der Wal, and P. Burt. “Real-time scene stabilization and mosaic construction,” Proc. of DARPA Image Understanding Workshop, pp. 54-62, 1994.

[50] F-Y. Hu, Y-N. Zhang, and L. Yao, “An effective detection algorithm for moving object with complex background,” IEEE Int. Conference on Machine Learning and Cybernetics, vol.8, pp. 5011-5015, 2005.

[51] Y-S. Choi, P. Zaijun, S-W. Kim, T-H. Kim, and C-B. Park, “Salient motion information detection technique using weighted subtraction image and motion vector,” Hybrid Information Technology, vol. 1, pp. 263-269, 2006.

[52] M. Black and P. Anandan, “The robust estimation of multiple motions: Parametric and piecewise smooth flow fields,” Computer Vision and Image Understanding, vol. 63, no. 1, pp. 75-104, 1996.

[53] B.K.P. Horn and B.G. Schunk, “Determining optical flow,” Artificial Intelligence, vol. 17, pp. 185-203, 1981.

[54] B.D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” Proc. of the 1981 DARPA Image Understanding Workshop, pp. 121-130, 1981.

[55] R. Szeliski and J. Coughlan, “Spline-based image registration,” Int. Journal of Computer Vision, vol. 22, no. 3, pp. 199-218, 1997.

[56] J.L. Barron, D.J. Fleet, S.S. Beauchemin, and T.A. Burkitt, “Performance of optical flow techniques,” IEEE Int. Conference on Computer Vision and Pattern Recognition, pp. 236-242, 1992.

[57] R.N. Hota, V. Venkoparao, and A. Rajagopal, “Shape based object classification for automated video surveillance with feature selection,” IEEE Int. Conference on Information Technology, pp. 97-99, 2007.

[58] N. Dalal and B. Triggs. “Histograms of oriented gradients for human detection,” IEEE Int. Conference on Computer Vision and Pattern Recognition, pp. 886-893, 2005.

[59] B. Leibe, E. Seemann, and B. Schiele, “Pedestrian detection in crowded scenes,” IEEE Int. Conference on Computer Vision and Pattern Recognition, pp. 878-885, 2005.

[60] Q. Zhu, M. Yeh, K. Cheng, and S. Avidan. “Fast human detection using a cascade of histograms of oriented gradients,” IEEE Int. Conference on Computer Vision and Pattern Recognition, pp. 1491-1498, 2006.

[61] N.D. Bird, O. Masoud, N.P. Papanikolopoulos, and A. Isaacs, “Detection of loitering individuals in public transportation areas,” IEEE Trans. on Intelligent Transportation Systems, vol. 6, no. 2, pp. 167-177, 2005.

[62] B.C. Chee, M. Lazarescu, and T. Tan, “Detection and monitoring of passengers on a bus by video surveillance,” IEEE Int. Conference on Image Analysis and Processing, pp. 143-148, 2007.

[63] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K.W. Bowyer, “The HumanID gait challenge problem: data sets, performances, and analysis,” IEEE Trans. on Pattern Analysis and Machine Intelligence Conference, vol. 27, no. 2, pp. 162-177, 2005.

[64] Y-B. Li, T-X. Jiang, Z-H. Qiao, and H-J. Qian, “General methods and development actuality of gait recognition,” IEEE Int. Conference on Wavelet Analysis and Pattern Recognition, vol.3, pp. 1333-1340, 2007.

[65] D.M. Gavrila “The visual analysis of human movement: A survey,” Computer Vision and Image Understanding, vol. 73, no. 1, pp. 82–98, 1999.

[66] C. Cedras and M. Shah, “Motion-based recognition, A survey,” Image and Vision Computing, vol. 13, no. 2, pp. 129-155, 1995.

[67] S. Ju, “Human motion estimation and recognition (depth oral report),” University of Toronto Technical Report, 1996.

[68] S-H. Kim and H-G. Kim, “Face detection using multi-modal information,” IEEE Int. Conference on Automatic Face and Gesture Recognition, pp. 14-19, 2000.

[69] S. Harasse, L. Bonnaud, and M. Desvignes, “Human model for people detection in dynamic scenes,” Int. Conference on Pattern Recognition, vol. 1, pp. 335-354, 2006.

[70] M-T. Yang, Y-C. Shih, and S-C. Wang, “People tracking by integrating multiple features,” Int. Conference on Pattern Recognition, vol. 4, pp. 929-932, 2004.

[71] M.J. Jones and D. Snow, “Pedestrian detection using boosted features over many frames,” Int. Conference Pattern Recognition, pp. 1-4, 2008.

[72] N. Dalal, B. Triggs, and C. Schmid “Human detection using oriented histograms of flow and appearance,” Proc. European Conference Computer Vision, pp. 428-441, 2006.

[73] S. Haykin and N. DeFreitas, “Special issue on: Sequential state estimation: From Kalman filters to particle filters,” Proc. of the IEEE, vol. 92, no. 3, 2004.

[74] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” ACM Journal of Computing Surveys, vol. 38, no. 4, 2006.

[75] L.M. Fuentes and S.A. Velastin, “People tracking in surveillance applications,” Image and Vision Computing, vol. 24, no. 11, pp. 1165-1171, 2006.

[76] N. Ning and T. Tan, “A framework for tracking moving target in a heterogeneous camera suite,” IEEE Int. Conference on Control, Automation, Robotics and Vision, pp. 1-5, 2006.

[77] R. Eshel and Y. Moses, “Homography based multiple camera detection and tracking of people in a dense crowd,” IEEE. Int. Conference Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[78] A. Ess, B. Leibe, K. Schindler, and K. L. Van Gool, “A mobile vision system for robust multi-person tracking,” IEEE Int. Conference Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[79] G. Gasser, N. Bird, O. Masoud, and N. Papanikolopoulos, “Human activities monitoring at bus stops,” IEEE Int. Conference Robotics and Automation, vol. 1, pp. 90-95, 2004.

[80] D. Comanciu, V. Ramesh, and P. Meer, “Kernel-based object tracking,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 25, no. 5, pp. 564-577, 2003.

[81] A. Tesei, A. Teschioni, C.S. Regazzoni, and G. Vernazza, “Long memory matching of interacting complex objects from real image sequences,” in Proc. of Conference on Time Varying Image Processing and Moving Objects Recognition, pp. 283–286, 1996.

[82] M. Isard and A. Blake, “Condensation conditional density propagation for visual tracking,” Int. Journal of Computer Vision, vol. 29, no. 1, pp. 5-28, 1998.

[83] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, “Color-based probabilistic tracking,” Proc. European Conference on Computer Vision, pp. 661-675, 2002.

[84] U. Scheunert, H. Cramer, B. Fardi, and G. Wanielik, “Multi sensor based tracking of pedestrians: a survey of suitable movement models,” Intelligent Vehicles Symposium, pp. 774-778, 2004.

[85] G.L. Foresti, C.S. Regazzoni, and P.K. Varshney “Multisensor surveillance systems: The fusion perspective,” Kluwer Academic Publisher, 2003.

[86] N.T. Siebel and S. Maybank , “Fusion of multiple tracking algorithms for robust people tracking”, Proc. European Conference on Computer Vision, pp. 373-382, 2002.

[87] B. Leibe, K. Schindler, and L. Van Gool, “Coupled detection and trajectory estimation for multi-object tracking,” Int. Conference Computer Vision, pp. 1-8, 2007.

[88] R.J. Morris and D.C. Hogg, “Statistical models of object interaction,” IEEE Workshop on Visual Surveillance, pp. 81-85, 1998.

[89] T. Darrell, G. Gordon, J. Woodfill, H. Baker, and M. Harville, “Robust, real-time people tracking in open environments using integrated stereo, color, and face detection,” IEEE Workshop on Visual Surveillance, pp. 26-32, 1998.

[90] R. Nevatia, T. Zhao, and S. Hongeng “Hierarchical language based representation of events in video steams,” IEEE Conference on Computer Vision and Pattern Recognition Workshop, vol. 4, pp. 39, 2003.

[91] R. Hamid, Y. Huang, and I. Essa. “ARGMode - activity recognition using graphical models,” IEEE Workshop on Computer Vision and Pattern Recognition Workshop, vol. 4, pp. 38-44, 2003.

[92] R. Nevatia, J. Hobbs, and B. Bolles, “An ontology for video event representation,” IEEE Workshop on Event Detection and Recognition, pp. 119, 2004.

[93] C. Rao and M. Shah, “View-invariant representation and learning of human action,” IEEE Workshop on Detection and Recognition of Events in Video, pp. 55-63, 2001.

[94] W. Kang and F. Deng, “Research on intelligent visual surveillance for public security,” IEEE/ACIS Int. Conference on Computer and Information Science, pp. 824-829, 2007.

[95] J. K. Aggarwal and Q. Cai. “Human motion analysis: A review,” Computer Vision and Image Understanding, pp. 428-440, 1999.

[96] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata, “Articulated and elastic non-rigid motion: a review,” Workshop on Motion of Non-Rigid and Articulated Objects, pp. 2-14, 1994.

[97] G. Shaffer. “A Mathematical Theory of Evidence,” Princeton University Press, 1976.

[98] K. Rapantzikos, Y. Avrithis, and S. Kollias, “Handling uncertainty in video analysis with spatiotemporal visual attention,” Fuzzy Systems, pp. 213-217, 2005.

[99] P. Remagnini, T. Tan, and K. Baker, “Agent-oriented annotation in model based visual surveillance,” IEEE Int. Conference on Computer Vision, pp. 857-862, 1998.

[100] V. Girondel, A. Caplier, and L. Bonnaud, “A belief theory-based static posture recognition systems for real-time video surveillance applications,” IEEE Conference Advanced Video and Signal Based Surveillance, pp. 10-15, 2005.

[101] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russell, and J. Weber, “Automatic symbolic traffic scene analysis using belief networks,” Proc. National Conference on Artificial intelligence, pp. 966-972, 1994.

[102] K.M. Kitani, Y. Sato, and A. Sugimoto, “Deleted interpolation using a hierarchical Bayesian grammar network for recognizing human activity,” IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 239-246, 2005.

[103] M. Brand and V.M. Kettnaker, “Discovery and segmentation of activities in video,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 844-851, 2000.

[104] B. Morris and M. Trivedi, “An adaptive scene description for activity analysis in surveillance video,” Int. Conference Pattern Recognition, pp. 1-4, 2008.

[105] J. Nascimento, M. Figueiredo, and J. S. Marques. “Segmentation and classification of human activities,” Workshop on Human Activity Recognition and Modeling, pp. 79-86, 2005.

[106] N. Haering and K. Shafique, “Automatic visual analysis for transportation security,” IEEE Conference on Technologies for Homeland Security, pp. 13-18, 2007.

[107] D. Abrams and S. McDowall, “Video content analysis with effective response,” IEEE Conference on Technologies for Homeland Security, pp. 57-63, 2007.

[108] S.A. Velastin, B.A. Boghossian, and M.A. Vicencio-Silva, “A motion-based image processing system for detecting potentially dangerous situations in underground railway stations,” Transportation Research Part C: Emerging Technologies, vol. 14, no. 2, pp. 96-113, 2006.

[109] P. Reisman, O. Mano, S. Avidan, and A. Shashua, “Crowd detection in video sequences,” Intelligent Vehicles Symposium, pp. 66-71, 2004

[110] H. Rahmalan, M.S. Nixon, and J.N. Carter, “On crowd density estimation for surveillance,” The Institution of Engineering and Technology Conference on Crime and Security, pp. 540–545, 2006.

[111] X. Wu, G. Liang, K. Lee, and Y. Xu, “Crowd density estimation using texture analysis and learning,” IEEE Int. Conference on Robotics and Biometrics, pp.214–219, 2006.

[112] B. Seongmin, I-K. Jeong, and I-H Lee, “Implementation of crowd system in Maya,” Int. Joint Conference on SICE-ICASE, pp. 2713-2716, 2006.

[113] A. Shendarkar, K. Vasudevan, S. Lee, and Y-J. Son, “Crowd simulation for emergency response using BDI agent based on virtual reality,” Proc. of Winter Simulation Conference, pp. 545–553, 2006.

[114] S. Banarjee, C. Grosan, and A. Abraham, "Emotional ant based modeling of crowd dynamics," Symbolic and Numeric Algorithms for Scientific Computing, pp.8, 2005.

[115] N. Courty and S.R. Musse, "Simulation of large crowds in emergency situations including gaseous phenomena," Int. Conference on Computer Graphics, pp. 206-212, 2005.

[116] Y-Y. Lin and Y-P. Chen, "Crowd control with swarm intelligence," Evolutionary Computation, pp. 3321-3328, 2007.

[117] X. Liu, P.H. Tu, J. Rittscher, A. Perera, and N. Krahnstoever, "Detecting and counting people in surveillance applications," IEEE Conference Advanced Video and Signal Based Surveillance, pp. 306-311, 2005.

[118] I. Cohen, A. Garg, and T.S. Huang, "Vision-based overhead view person recognition," IEEE Int. Conference on Pattern Recognition, vol. 1, pp. 1119-1124, 2000.

[119] L. Dong, V. Parameswaran, V. Ramesh, and I. Zoghlami, "Fast crowd segmentation using shape indexing," IEEE Int. Conference on Computer Vision, pp. 1-8, 2007.

[120] S. Lin, J. Chen, and H. Chao, "Estimation of number of people in crowded scenes using perspective transformation," IEEE Trans. Systems, Man, and Cybernetics Part A, vol. 31, no. 6, pp. 645-654, 2001.

[121] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection," Int. Conference Pattern Recognition, pp. 1-4, 2008.

[122] B. Maurin, O. Masoud, and N.P. Papanikolopoulos, "Tracking all traffic: computer vision algorithms for monitoring vehicles, individuals, and crowds," Robotics & Automation Magazine, vol. 12, no. 1, pp. 29-36, 2005.

[123] B. Zhan, N.D. Monekosso, P. Remagnino, S.A. Velastin, and L-Q. Xu, "Crowd analysis: a Survey," Machine Vision and Applications, vol.19, no. 5-6, pp. 345-357, 2008.

[124] V. Rabaud and S. Belongie, "Counting crowded moving objects," IEEE Int. Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 705-711, 2006.

[125] E.L. Andrade, S. Blunsden, and R.B. Fisher, "Modeling crowd scenes for event detection," IEEE Int. Conference on Pattern Recognition, vol. 1, pp. 175-178, 2006.

[126] E.L. Andrade, S. Blunsden, and R.B. Fisher, "Hidden markov models for optical flow analysis in crowds," IEEE Int. Conference on Pattern Recognition, vol. 1, pp. 460-463, 2006.

[127] E.L. Andrade, R.B. Fisher, and S. Blunsden, "Detection of emergency events in crowded scenes," The Institution of Engineering and Technology Conference on Crime and Security, pp. 528-533, 2006.

[128] E. Andrade, S. Blunsden, and R. Fisher. "Performance analysis of event detection models in crowded scenes," Proc. Workshop on Towards Robust Visual Surveillance Techniques and Systems at Visual Information Engineering, pp. 427-432, 2006.

[129] J.H. Yin, S.A. Velastin, and A.C. Davies, "Image processing techniques for crowd density estimation using a reference image," Asian Conference on Computer Vision, pp. 489-498, 1995.

[130] Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," IEEE Int. Conference on Computer Vision, pp. 1–8, 2007.

[131] S. Antipolis, "Intelligent environments for problem solving by autonomous systems," Institut National De Recherche En Informatique Et En Automatique Activity Report, section 6.2.13, pp. 41, 2007.

[132] E.L. Andrade, S. Blunsden, and R.B. Fisher, "Modeling crowd scenes for event detection," IEEE Int. Conference on Pattern Recognition, vol.1, pp. 175-178, 2006.

[133] E.L. Andrade, S. Blunsden, and R.B. Fisher, “Hidden markov models for optical flow analysis in crowds,” IEEE Int. Conference on Pattern Recognition, vol.1, pp. 460-463, 2006.

[134] Z. Youding and K. Fujimura, “Head pose estimation for driver monitoring,” IEEE Intelligent Vehicles Symposium, pp. 501-506, 2004.

[135] A.O. Balan, M.J. Black, H. Haussecker, and L. Sigal, “Shining a light on human pose: On shadows, shading and the estimation of pose and shape,” IEEE Int. Conference Computer Vision, pp. 1-8, 2007.

[136] M.W. Lee and R. Nevatia, “Body part detection for human pose estimation and tracking,” Motion and Video Computing, pp. 23-23, 2007.

[137] M.W. Lee and R. Nevatia, “Dynamic human pose estimation using markov chain monte carlo approach,” Motion and Video Computing vol.2, pp. 168-175, 2005.

[138] A. Bissacco, M.H. Yang, and S. Soatto, “Fast human pose estimation using appearance and motion via multi-dimensional boosting regression,” Computer Vision and Pattern Recognition, pp. 1-8, 2007.

[139] A.Fathi and G. Mori, “Human pose estimation using motion exemplars,” Computer Vision, pp. 1-8, 2007.

[140] A. Baumberg and D. Hogg, “An efficient method for contour tracking using active shape models,” Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects, pp. 194-199.

[141] L.M. Fuentes and S.A. Velastin, “Tracking-based event detection for CCTV systems,” Pattern Analysis and Applications, vol. 7, no. 4, 2005.

[142] BEHAVE official website. Available at: http://homepages.inf.ed.ac.uk/rbf/BEHAVE/

[143] CAVIAR Project dataset. Available: http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/

[144] S. Blunsden, E. Andrade, and R. Fisher. “Non parametric classification of human interaction,” Proc. 3rd Iberian Conference on Pattern Recog. and Image Analysis, pp. 347-354, 2007.

[145] A. Madabhushi and J.K. Aggarwal, “A Bayesian approach to human activity recognition,” IEEE Workshop on Visual Surveillance, pp. 25–32, 1999.

[146] H. Yasin and S.A. Khan, “Moment invariants based human mistrustful and suspicious motion detection, recognition and classification,” Computer Modeling and Simulation, pp. 734-739 2008.

[147] S. Park and J.K Aggarwal, “Recognition of two-person interactions using a hierarchical Bayesian network,” Int. Workshop on Video Surveillance, 2003.

[148] S. Park and J.K. Aggarwal, “Simultaneous tracking of multiple body parts of interacting persons,” Computer Vision Image Understanding, pp. 1-21, 2006.

[149] S. Park and J.K. Aggarwal, “Recognition of human interaction using multiple features in gray scale images,” IEEE Int. Conference on Pattern Recognition, vol. 1, pp. 51-54, 2000.

[150] I. Haritaoglu, R. Cutler, D. Harwood, and L.S. Davis, “Backpack: detection of people carrying objects using silhouettes,” IEEE Int. Conference on Computer Vision, vol. 1, pp. 102-107, 1999.

[151] F. Cupillard, F. Bremond, and M. Thonnat, “Group behavior recognition with multiple cameras,” Applications of Computer Vision, pp. 177-183, 2002.

[152] A. Alberto, B. Francois, T. Christophe, and T. Monique, “Design and assessment of an intelligent activity monitoring platform” EURASIP Journal on Applied Signal Processing, no. 14, pp. 2359-2374, 2005.

[153] S.Park and M.M. Trivedi, “Homography-based analysis of people and vehicle activities in crowded scenes,” IEEE Workshop on Applications of Computer Vision, pp. 51, 2007.

[154] Z. Sun , G. Bebis, and R. Miller, “On-road vehicle detection: a review,” Trans. on Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 694-711, 2006.

[155] F. Jiang, Y. Wu, and A.K. Katsaggelos, “Abnormal event detection from surveillance video by dynamic hierarchical clustering,” IEEE Int. Conference on Image Processing, vol. 5, pp. V145-V148, 2007

[156] F. Jiang, Y. Wu, and A.K. Katsaggelos, “Abnormal event detection based on trajectory clustering by 2-depth greedy search,” Acoustics, Speech and Signal Processing, pp. 2129-2132, 2008.

[157] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi, “Traffic monitoring and accident detection at intersections,” IEEE Trans. on Intelligent Transportation Systems, vol. 1, no. 2, pp. 108-118, 2000.

[158] X. Chen and C. Zhang “Incident retrieval in transportation surveillance videos - An interactive framework” Multimedia and Expo, 2007 IEEE Int. Conference pp. 2186-2189, 2007.

[159] L. Jong Taek, M.S. Ryoo, M. Riley, and J.K. Aggarwal, “Real-time detection of illegally parked vehicles using 1-D transformation,” IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 254-259, 2007.

[160] C. Zhang, Z. Zhang, B. Zhang, S. Hao, M. Wu, and J. Guo, “A real-time vehicle flow-measuring algorithm for complex urban intersection in the daytime,” IEEE Int. Conference on Machine Learning and Cybernetics, vol. 2, pp. 934-938, 2002.

[161] V. Kettnaker and M. Brand, “Minimum-entropy models of scene activity,” IEEE Int. Conference on Computer Vision and Pattern Recognition, vol. 1, 1999.

[162] L. Xiaokun and F.M. Porikli, “A hidden Markov model framework for traffic event detection using video features” IEEE Int. Conference on Image Processing, vol. 5, no. 24-27, pp. 2901-2904, 2004.

[163] S. Kamijo, M. Harada, and M. Sakauchi, “An incident detection system based on semantic hierarchy,” IEEE Int. Conference on Intelligent Transportation Systems, no. 3-6, pp. 853-858, 2004.

[164] H. Veeraraghavan, P. Schrater, and N. Papanikolopoulos, “Switching kalman filter-based approach for tracking and event detection at traffic intersections,” Proc. of Intelligent Control, Medical Conference on Control and Automation, no. 27-29, pp. 1167 – 1172, 2005.

[165] H.Y. Cheng and J.N. Hwang, “Multiple-target tracking for crossroad traffic utilizing modified probabilistic data association,” Acoustics Speech and Signal Processing, vol. 1, pp. I: 921-924, 2007.

[166] P. Kumar, S. Ranganath, H. Weimin, and K. Sengupta, “Framework for real-time behavior interpretation from traffic video,” IEEE Trans. on Intelligent Transportation Systems, vol. 6, no. 1, pp. 43-53, 2005.

[167] S. Gong and T. Xiang. “Recognition of group activities using dynamic probabilistic networks,” IEEE Int. Conference on Computer Vision, pp. 742-749, 2003.

[168] P.G. Raeth and D.A.Bertke, “Finding events automatically in continuously sampled data streams via anomaly detection,”
National Aerospace and Electronics Conference, pp. 580-587, 2000.

[169] T. Xia and S. Gong, “Video behavior profiling for anomaly detection,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30, no. 5, pp. 893-908, 2008.

[170] S. Gong and T. Xiang, “Recognition of group activities using dynamic probabilistic networks,” IEEE Int. Conference Computer Vision, vol. 2, pp. 742-749, 2003.

[171] A. J. Lipton and N. Haering, “Commode: An algorithm for video background modeling and object segmentation,” IEEE Int. Conference on Control, Automation, Robotics and Vision, vol. 3, pp. 1603-1608, 2002.

[172] H. Tao, H. S. Sawhney, and R. Kumar, “Object tracking with Bayesian estimation of dynamic layer representations,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 75-89, 2002.

[173] N. Buch, J. Orwell, and S.A. Velastin, “Detection and classification of vehicles for urban traffic scenes,” Int. Conference Visual Information Engineering, pp. 182-187, 2008.

[174] O. Sidla, L. Paletta, Y. Lypetskyy, and C. Janner, “Vehicle recognition for highway lane survey,” IEEE Int. Conference. on Intelligent Transportation Systems, 3-6 pp. 531-536, 2004.

[175] J.A.Vijverberg, N.A.H.M. Koning, J. Han, P.H.N. With, and D. Cornelissen, “High-level traffic-violation detection,” Int. Conference on Embedded Traffic Analysis, vol. 2, pp. 793-796, 2007.

[176] A. Makarov, J-M. Vesin, and M. Kunt, “Intrusion detection using extraction of moving edges,” IEEE Int. Conference on Pattern Recognition, vol. 1, pp. 804-807, 1994.

[177] V. Kettnaker, “Time-dependent HMMs for visual intrusion detection,” IEEE Int. Conference Computer Vision and Pattern Recognition Workshop, vol. 4, pp. 34-34, 2003.

[178] S. Kang, B. Abidi, and M. Abidi, “Integration of color and shape for detecting and tracking security breaches in airports,” Security Technology, 38th Annual 2004 Int. Carnahan Conference on, pp. 289-294, 2004.

[179] G. Monteiro, M. Ribeiro, J. Marcos, and J. Batista, “Wrongway drivers detection based on optical flow,” IEEE Int. Conference on Image Processing, vol.5, pp. 141-144, 2007.

[180] M. Ghazal, C. Vazquez, and A. Amer, “Real-time automatic detection of vandalism behavior in video sequences,” IEEE Int. Conference on Systems, Man, and Cybernetic, pp. 1056-1060, 2007.

[181] C. Sacchi, C. Regazzoni, and G. Vernazza, “A neural network-based image processing system for detection of vandal acts in unmanned railway environments,” IEEE Int. Conference Image Analysis and Processing, pp. 529-534, 2001.

[182] P. Spagnolo, A. Caroppo, M. Leo, T. Martiriggiano, and T. D'Orazio, “An abandoned/removed objects detection algorithm and its evaluation on PETS datasets,” Video and Signal Based Surveillance, 2006. IEEE Int. Conference, pp. 17-17, 2006.

[183] L. Sijun, J. Zhang, and D. Feng, “A knowledge-based approach for detecting unattended packages in surveillance video,” IEEE Int. Conference on Video and Signal Based Surveillance, pp. 110-110, 2006.

[184] Fire safety risk assessment - small and medium places of assembly, ISBN 978 1 85112 820 4, 5 June 2006.

[185] G.K. Still, “PhD thesis : Crowd dynamics,” Mathematics Department, Warwick University, 2000.

[186] TREC Video Retrieval Evaluation Official Website. Available at: http://www-nlpir.nist.gov/projects/trecvid/

[187] T. Ahmedali and J.J. Clark, “Collaborative multi-camera surveillance with automated person detection,” IEEE Canadian Conference Computer and Robot Vision, pp. 39-39, 2006.

[188] F. Bunyak, K. Palaniappan, S.K. Nath, and G. Seetharaman, “Geodesic active contour based fusion of visible and infrared video for persistent object tracking,” IEEE Workshop on Applications of Computer Vision, pp. 35-35, 2007.

[189] B. Ping Lai Lo, J. Sun, and S.A. Velastin, “Fusing visual and audio information in a distributed intelligent surveillance system for public transport systems,” Acta Automatica Sinica, vol. 29, no. 3, pp. 393-407, 2003.

[190] S.J. Krotosky and M.M. Trivedi, “Person surveillance using visual and infrared imagery,” IEEE Trans. Circuits and Systems for Video Technology, vol. 18, no. 8, pp. 1096-1105, 2008.

[191] R. Nevatia, G. Medioni and I. Cohen, “Event detection and analysis from video streams,” IUW, pp. 63-72, 1998.

[192] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, “Event detection and analysis from video streams,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 8, pp. 873-889, 2001.

[193] T. Zhao and R. Nevatia, “Car detection in low resolution aerial images,” IEEE Int. Conference on Computer Vision, vol. 1, pp. 710-717, 2001.

[194] Z.W. Kim and R. Nevatia, “Automatic description of complex buildings from multiple images,” Computer Vision and Image Understanding, vol. 96, no. 1, pp. 60-95, 2004.

[195] M. Hu, S. Ali, and M. Shah, “Detecting global motion patterns in complex videos,” Int. Conference Pattern Recognition, pp. 1-5, 2008.

[196] E. Ribnick, S. Atev, N. Papanikolopoulos, O. Masoud, and R. Voyles, “Detection of thrown objects in indoor and outdoor scenes,” Intelligent Robots and Systems, pp. 979-984, 2007.

[197] N. Bird, S. Atev, N. Caramelli, R. Martin, O. Masoud, and N. Papanikolopoulos, “Real time, online detection of abandoned objects in public areas,” IEEE Int. Conference Robotics and Automation, pp. 3775-3780, May 2006.

[198] S. Ferrando, G. Gera, M. Massa, and C. Regazzoni, “A new method for real time abandoned object detection and owner tracking,” IEEE Int. Conference on Image Processing, pp. 3329-3332, 2006.

[199] E. Ribnick, S. Atev, O. Masoud, N. Papanikolopoulos, and R. Voyles, “Real-time detection of camera tampering,” IEEE Conference on Advanced Video and Signal Based Surveillance, 2006.

[200] D. Angiati, G. Gera, S. Piva, and C.S. Regazzoni, “A novel method for graffiti detection using change detection algorithm,” IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 242-246, 2005.

[201] L. Fuentes and S. Velastin, “Advanced surveillance: From tracking to event detection,” IEEE. Latin America Transactions, vol.2, no.3, pp. 1-1, 2004.

[202] S.A.Velastin, J.H. Yin, A.C. Davies, M.A. Vicencio-Silva, R.E. Allsop, and A. Penn, “Automatic measurement of crowd density and motion using image processing,” Int. Conference on Road Traffic Monitoring and Control, pp. 127-132, 1994.

[203] V. Manohar, M. Boonstra, V. Korzhova, P. Soundararajan, D. Goldgof, R. Kasturi, S. Prasad, H. Raju, R. Bowers, and J. Garofolo, "PETS vs. VACE evaluation programs: A comparative study", In the Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1422-0, Pages: 1-6, In Conjunction with IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2006

[204] http://www.umiacs.umd.edu/lamp/media/research/viper/

[205] Doermann, D.; Mihalcik, D., "Tools and techniques for video performance evaluation," Pattern Recognition, 2000. Proceedings. 15th International Conference on , vol.4, no., pp.167-170 vol.4, 2000

[206] Official website for Metropolitan Transportation Authority. Available at: http://www.mta.info

[207] http://www.imsresearch.com/press_release_details.html&press_id=700

[208] Official website for AgentVi. Available at: http://www.agentvi.com/

[209] Official website for Aimetis Corporation. Available at: http://www.aimetis.com
[210] Official website for Cernium. Available at: http://www.cernium.com
[211] Official website for Eptascape. Available at: http://www.eptascape.com
[212] Official website for Honeywell. Available at: http://www51.honeywell.com
[213] Official website for Indigo Vision. Available at: http://www.indigovision.com
[214] Official website for Intelliview. Available at: http://www.intelliview.ca/
[215] Official website for Intellivision. Available at: http://www.intelli-vision.com
[216] Official website for Ipsotek. Available at: http://www.ipsotek.com
[217] March Networks. Available at: http://www.marchnetworks.com
[218] MATE Intelligent Video. Available at: http://www.mate.co.il
[219] Official website for Object Video. Available at: http://www.objectvideo.com
[220] Official website for Sightlogix. Available at: http://www.sightlogix.com/
[221] Official website Verint. Available at: http://verint.com
[222] Official website for Vidient. Available at: http://www.vidient.com/
[223] Official website for Nice. Available at: http://www.nice.com/products/video/index.php
[224] D. Abrams and S. McDowall, "Video Content Analysis with Effective Response," IEEE Conference on Technologies for Homeland Security, pp. 57-63, 2007.
[225] Official website of IoImage. Available at: http://www.ioimage.com
[226] Florida Department of Transportation. "Florida Transportation Trends & Conditions Report," 2007. Available at: http://www.dot.state.fl.us/planning/trends/tc-report/
[227] S. Chan, "U.S. Transit Agencies Turn to Cameras in Terror Fight, but Systems Vary in Effectiveness," New York Times, July 14, 2005.
[228] W. Neuman, "Lockheed Sued to Pull Out of Security Contract With Transit Agency," New York Times, April 28, 2009.