# Understanding Transit Scenes: A Survey on Human Behavior-Recognition Algorithms

Joshua Candamo, Matthew Shreve, *Member, IEEE*, Dmitry B. Goldgof, *Fellow, IEEE*, Deborah B. Sapper, and Rangachar Kasturi, *Fellow, IEEE*

***Abstract*—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. This paper describes the current state-of-the-art image-processing methods for automatic-behavior-recognition techniques, with focus on the surveillance of human activities in the context of transit applications. The main goal of this survey is to provide researchers in the field with a summary of progress achieved to date and to help identify areas where further research is needed. This paper provides a thorough description of the research on relevant human behavior-recognition methods for transit surveillance. Recognition methods include single person (e.g., loitering), multiple-person interactions (e.g., fighting and personal attacks), person–vehicle interactions (e.g., vehicle vandalism), and person–facility/location interactions (e.g., object left behind and trespassing). A list of relevant behavior-recognition papers is presented, including behaviors, data sets, implementation details, and results. In addition, algorithm's weaknesses, potential research directions, and contrast with commercial capabilities as advertised by manufacturers are discussed. This paper also provides a summary of literature surveys and developments of the core technologies (i.e., low-level processing techniques) used in visual surveillance systems, including motion detection, classification of moving objects, and tracking.**

***Index Terms*—Anomaly detection, event detection, human behavior recognition, smart transit system, video analytics, visual surveillance.**

## I. Introduction

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 doorstep. However, 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 City Transit System [1] is the busiest metro system in the U.S. (based on 2006 statistics), with a total of 468 stations and 1.49 billion riders a year, that is, 4.9 million riders a day. Moscow metro [2] is the busiest metro in Europe and, as of 2007, has 176 stations with 2.52 billion riders annually, that is, 9.55 million daily riders. This ridership represents a 9.53% growth since 1995. Transit systems are spread through hundreds of kilometers and already require several tens of thousands of employees for daily operations. A complete deployment of visual surveillance to cover a system of this magnitude requires thousands of cameras, which makes human-based/dependent surveillance unfeasible 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 exclusively leaving the task of threat detection to human operators. Detecting specific activities in a live feed or searching in video archives (i.e., video analytics) almost completely relies on costly and scarce human resources. Detecting multiple activities in real-time video feeds is currently performed by assigning multiple analysts to simultaneously watch the same video stream. Each analyst is assigned a portion of the video and is given a list of events (behaviors) and objects for which to look. 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 [3]. Thus, 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. It is our goal 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 [4], [5], [93], [123]. Previous surveys

Manuscript received September 18, 2008; revised April 28, 2009 and July 15, 2009. First published October 2, 2009; current version published March 3, 2010. This work was supported in part by the Center of Urban Transportation and Research, University of South Florida, and the Florida Department of Transportation under Grant BD549-49. The Associate Editor for this paper was R. I. Hammoud.

J. Candamo was with the Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620 USA. He is now with K9 Bytes, Inc., Tampa, FL 33617 USA (e-mail: candamo@cse.usf.edu).

M. Shreve, D. B. Goldgof, and R. Kasturi are with the Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620 USA (e-mail: mshreve@cse.usf.edu; goldgof@cse.usf.edu; r1k@cse.usf.edu).

D. B. Sapper is with the Center for Urban Transportation Research, University of South Florida, Tampa, FL 33620 USA (e-mail: sapper@cutr.usf.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2009.2030963

TABLE I
RELATED LITERATURE SURVEY SUMMARY

| First Author | Yr | Topic | Ref |
|---|---|---|---|
| Zhan | 08 | Crowd analysis | [123] |
| Kang | 07 | Intelligent visual surveillance | [93] |
| Stoykova | 07 | 3D scene capture | [39] |
| Sun | 06 | On-road vehicle detection systems | [152] |
| Forsyth | 06 | Human Motion | [6] |
| Yilmaz | 06 | Object tracking | [73] |
| Radke | 05 | Image change detection | [45] |
| Valera | 05 | Intelligent distributed surveillance systems | [5] |
| Haykin | 04 | Object tracking | [72] |
| Foresti | 04 | Multi-sensor tracking | [84] |
| Weiming | 04 | Motion and tracking for surveillance | [4] |
| Fasel | 03 | Facial expressions | [35] |
| Moeslund | 00 | Human motion capture (large scale body movements) | [7] |
| Aggarwal | 99 | Motion analysis of the human body | [94] |
| Gavrila | 99 | Human movement | [64] |
| Pavlovic | 97 | Hand gestures | [34] |
| Ju | 96 | Human Motion Estimation and Recognition | [66] |
| Cedras | 95 | Motion-based classification | [65] |
| Aggarwal | 94 | Elastic non-rigid motion | [95] |
| Cedras | 94 | Motion detection | [33] |
| Barron | 92 | Optical flow | [55] |

have emphasized low-level processing techniques used in visual surveillance (what we will refer to as "core technologies," e.g., motion detection and tracking). In contrast, we focus on human behavior recognition topics, drawing special attention to transit system applications. However, 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 I).

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, multimodal 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 Visual Surveillance and Monitoring (VSAM, 1997) [8] and Human Identification at a Distance (HID, 2000). Recently, the Video and Image Retrieval Analysis Tool (VIRAT, 2008) project has been 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 increasingly become 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. However, it is unclear how well systems are able to cope with crowds of people, which is 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 [9]–[17], [19]. Additionally, there are methods for evaluating the performance of the evaluators [18]. Despite the large number of existing evaluation techniques, a robust study that experimentally compares algorithms for human activity recognition is still lacking.

In the last decade, there have been many conferences and workshops dedicated to visual surveillance, including the IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS) 2005 challenge, which focused on real-time event detection solutions. The Challenge for Real-time Events Detection Solutions (CREDS) [19] defined by the needs of the public transportation network of Paris (RATP, the second busiest metro system in Europe) focused on proximity warning, dropping objects on tracks, launching objects across platforms, persons trapped by the door of a moving train, walking on rails, falling on the track, and crossing the rails. Several CREDS proposals can be found in [20]–[23]. The Performance Evaluation of Tracking and Surveillance (PETS) [24] workshops started with the goal of evaluating visual tracking and surveillance algorithms. The initiative provides standard data sets, with available ground truth, to evaluate object tracking and segmentation. Recently, a metric to evaluate surveillance results has also been introduced [25]. Some PETS data sets contain relevant information closely related to transit systems. Data sets include single-camera outdoor people and vehicle tracking (PETS, 2000); multicamera 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 metro networks (e.g., France's RATP, the U.K.'s LUL and BAA, and Italy's ATM) have deployed and tested large real-time transit visual-surveillance systems that include human-behavior recognition. There have been several transit surveillance projects that have been funded by the European Union. The Proactive Integrated Systems for Security Management by Technological, Institutional, and Communication Assistance (PRISMATICA) [26] has deployed video analytic systems in France. The Content Analysis and Retrieval Technologies to Apply Knowledge Extraction to Massive Recording (CARETAKER) [27] project was deployed in Italy. The Annotated Digital Video for Intelligent Surveillance and Optimized Retrieval (ADVISOR) [28] was successfully deployed and tested in Spain and Belgium, including previous work from the Crowd Management with Telematic Imaging and Communication Assistance (CROMATICA) project [29]–[32].

### A. Paper Overview

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 preprocessing steps prior to behavior recognition are referred to in this paper as "core technologies." Human behavior recognition

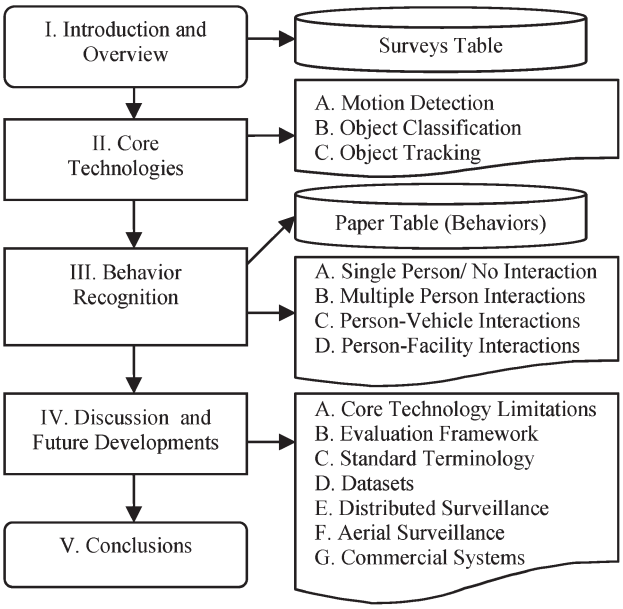



Fig. 1. Paper organization flowchart.

using video starts with the detection of foreground objects, which is 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 spatiotemporal relationship between the objects and the scene. The organization of this paper is depicted in Fig. 1. We begin Section II with a brief glance on the core technologies, to facilitate the understanding of the later sections of this paper. For organization purposes, all pertinent surveys dealing with core technologies are identified and summarized in Table I. Behavior-recognition strategies are discussed in Section III. Section IV elaborates on many important topics describing the current state-of-the-art strengths, weaknesses, and future research directions. Section V summarizes the contents of this paper.

## II. Core Technologies

### A. 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 [33]. In transit-surveillance applications, motion detection typically refers to movement of objects as a whole, e.g., movement of pedestrians or vehicles. However, 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: 1) large-scale body movements like movements of the head, arms, torso, and legs [7]; and 2) small-scale body movements like hand gestures and facial expressions [34], [35]. 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 2-D data for motion processing; thus, the focus of this survey will be on this domain. However, advances in image sensors and evolution of digital computation are leading to creation of new sophisticated methods for capturing, processing, and analyzing 3-D data from dynamic scenes. Recent developments include 3-D environmental modeling reconstructed using the shape-from-motion technique [36] and 3-D imagery from a moving monocular camera [37]. Most 3-D approaches require landmarks to be present in the scene [38] 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 3-D scene capture technologies can be found in [39].

*1) 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 a significant difference between the observed and the estimated image. Common algorithms include methods by Heikkila and Silven [40], Stauffer and Grimson (adaptive Gaussian mixture model or GMM) [41], Halevy and Weinshall [42], Cutler and Davis [43], and Toyama *et al.* (Wallflower) [44]. A detailed general survey of image change algorithms can be found in [45]. 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 goes on, 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 movements, and noise. Nevertheless, transit surveillance researchers continue to emphasize the importance of robust background subtraction methods [47] and online construction and adaptive background models [46]. 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 [47], 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, authors of [38] state that, for a GMM, speed can be improved by a factor of 8 with an image-size of $640 \times 480$ 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. However, an

image-stabilization algorithm is required when there is a significant movement of the camera [48]. 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 [49], [50] often combine both background subtraction and temporal differencing methods to provide more robust segmentation strategies.

*2) 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 [51], Horn and Schunck [52], Lucas and Kanade [53], and Szeliski and Couglan [54]. A comparison of methods for calculating optical flow can be found in [55].

### B. 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. Furthermore, 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 here 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 [56]. Some of these properties are also used in postobject classification to keep track of the object in sequential frames or separate cameras. In general, for object classification in surveillance video, there are shape-based, motion-based, and feature-based classification methods.

*1) Shape-Based Classification:* The geometry of the extracted regions (boxes, silhouettes, blobs) containing motion are often used to classify objects in video surveillance. Some common classifications in transit system surveillance are humans, crowds, vehicles, and clutter [8]. For transit applications, particularly 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 [57]–[59]. In general, 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 [60], 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 partially be occluded most of the time, in which case, the head [61] could be the only salient feature in the scene.

*2) 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 for human identification. Starting with the HumanID Gait Challenge [62], image-processing researchers actively proposed gait-based methods [63] for human identification at a distance. For more information on motion extraction and motion-based classification, see [64] and [65]. For an overview of motion estimation and recognition, with focus on optical flow techniques, see [66].

*3) Other Classification Methods:* Skin color [67] 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 successfully been combined with other descriptors [68] for classification purposes. In [60], the authors describe a method that consists of three parts: First, a red–green–blue normalization procedure was adopted to get the pure color components. A color transform is 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 shows the region of the image that has closely matched with skin color, indicating human motion. This method has also been extended in [69] and fused with other methods, including depth analysis using binocular imaging. Fusion of methods has been shown to be very effective when combining shape- and motion-based methods [70], [71].

### C. 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 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, the reader is referred to detailed object tracking surveys [72], [73]. Specific issues that arise within the transit domain include dealing with multiple persons in complex scenarios [74], tracking across large-scale distributed camera systems [75], tracking in highly congested areas with crowds of people [76] (e.g., near ticket offices, metro, or bus waiting areas at rush hour), or tracking using mobile platforms [77]. Extremely frequent occlusions are typical; consequently, the traditional localization and tracking of individuals is not sufficiently reliable. Furthermore, surveillance inside transit vehicles often only allows parts of individuals to be captured by the sensors (e.g., common occlusions from seats and other passengers often expose 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. In addition, feature uniqueness plays an important role. Some common features used in image-processing applications are color, edges, motion, and texture. In [78], researchers describe a system that monitors suspicious human activities around bus stops, in which tracking of pedestrians is performed using a kernel-based method proposed in [79]. This tracker is based on the color distribution of previously detected targets. The 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 [60], the shirt color is used as the main feature for tracking purposes, and kernel-based tracking is dropped in favor of blob-based tracking. Blob-based tracking offers a computational advantage over kernel search since the latter 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, [20] considers the use of a long-memory matching algorithm [80] using the blob's area, perimeter, and color histogram, and [22] performs 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, whereas the other body parts are occluded, particularly by the seats. Therefore, tracking is performed using faces with a color particle filter [81] similar to [82]. Tracking is based on the likelihood from the Bhattacharyya distance between color histograms in the hue-saturation-value color space. Color-based tracking is robust against vibration of the moving vehicles like trains and buses. However, it is sensitive to extreme changes in lighting conditions, such as a train entering a tunnel. Many multisensor approaches [83], [84], algorithm-fusion techniques [85], and integrating features over time [86] have been proposed to overcome many of the mentioned tracking difficulties and to generate robust tracking performance in transit-surveillance applications.

## III. Human-Behavior Recognition

In this survey, the terminology and classification strategy for the human behavior is similar to that used by the VIRAT project. VIRAT divides the human behavior into two categories, namely, "events" and "activities." An event refers to a single low-level spatiotemporal entity that cannot be further decomposed (e.g., person standing and person walking). An activity refers to a composition of multiple events (e.g., a person loitering). Across the literature, the term "event" is often interchangeably used to describe "events" or "activities," as defined by VIRAT. For clarity, we use the term "behavior" to include both "events" and "activities" and do not worry about inconsistencies of technical definitions. For organizational purposes, transit surveillance operationally relevant behaviors are divided into four general groups: 1) single person or no interaction; 2) multiple-person interactions; 3) person–vehicle interactions; and 4) person–facility/location interactions. Next, we provide some examples for each of these groups.

Fig. 2. Sample single-person or no interaction behavior. Suspicious person (marked with an ellipse) loitering for a long period of time without leaving in a bus.[1]

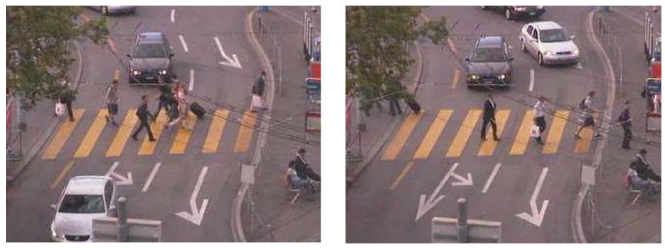


Fig. 3. Sample multiple-person interaction behavior. Pedestrians on a crosswalk.[2]

1) *Single person or no interaction* (see Fig. 2) consists of behaviors that can be defined only by considering person(s), which are not interacting with any other person or vehicle. For example, loitering, people (crowd) counting, crowd flow (behavior) analysis, and person talking on a cell phone.
2) *Multiple-person interactions* (see Fig. 3) consist of behaviors that involve persons interacting with each other. For example, following, tailgating, meeting, gathering, moving as a group, dispersing, shaking hands, kissing, exchanging objects, and kicking.
3) *Person–vehicle interactions* (see Fig. 4) consist of behaviors that are defined through interactions with persons and vehicles. For example, driving, getting in (out), loading (unloading), opening (closing) trunk, crawling under car, breaking window, dropping off, and picking up.

[1]Images courtesy of the Center for Distributed Robotics, University of Minnesota. Images are part of the data set used in [60].

[2]Images courtesy of the Computer Vision Laboratory, ETH Zurich. Images are part of the data set used in [86].

Fig. 4. Sample person–vehicle interaction. Person being run over by a vehicle.[3]

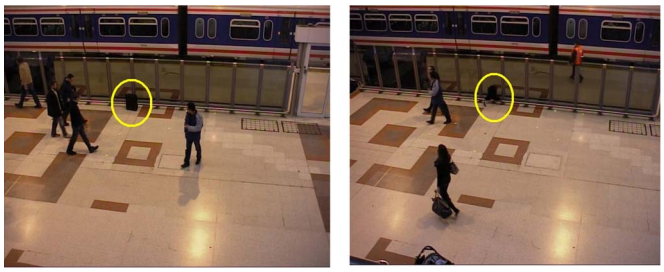


Fig. 5. Sample person–facility/location interaction. Object left behind in a train station.[4]

4) *Person–facility/location interactions* (see Fig. 5) are behaviors defined through interactions with persons and facilities/locations. For example, entering (exiting), standing, waiting at checkpoint, evading checkpoint, passing through gate, object left behind, and vandalism.

In surveillance systems, behavior recognition can be ambiguous, depending on the scene context. The same behavior may have several different meanings depending on 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) [87], [88], Event Mining (2003) [89], [90], and Event Detection and Recognition (2004) [91], [92]. See [93], where a brief background review of advances in intelligent visual surveillance is presented, and [94] and [95] 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 with other methods, such as the Dempster–Shafter theory [96]. Uncertainty handling can improve visual attention schemes [97]. Various other models have been used in surveillance-related applications, including classifying human motion and simple human interactions using a small belief network [98], human postures using belief networks [99], description of traffic scenes using a dynamic Bayes network [100], human activity recognition using a hierarchical Bayes network [101], and anomalous behavior detection using trajectory learning with hidden Markov models [102], [103].

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

[4]Object left behind sample images from PETS 2006 data set [24].

### A. Single Person or No Interaction

*1) 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, graffiti vandals, among others. In this survey, loitering refers to a behavior that exclusively involves a human. It is not to be confused with stationarity of objects (e.g., object left behind), which in our 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 [104].

In general, the literature for loitering detection in transit system applications mostly consists of tracking using indoor video (see Table II). However, publications often lack of implementation and technical details [21], [105], [106]. The technical literature exclusively to outdoor loitering detection is scarce. In [60], outdoor loitering is used as a cue to detect potential drug-dealing operations in bus stations. Often, drug dealers wait for their clients to come by bus, buy drugs, and leave. Consequently, a 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 [60] 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. However, in transit scenes, it is often impractical to manually measure camera parameters on site and almost impossible when working only with prerecorded examples [107].

*2) 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. Moreover, 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 (e.g., camera on a bus) [108] that analyze the optic flow generated from the moving objects and the moving platform.

Researchers have identified crowd counting to be often highly sensitive to training data [109], and in these cases, algorithms or

TABLE II
EXPERIMENTAL RESULTS AS STATED IN THEIR RESPECTIVE PUBLICATIONS (TP: TRUE POSITIVES; FP: FALSE POSITIVES; ROC: RECEIVER OPERATING CHARACTERISTIC CURVE)

| Paper | Feature | Results |
|---|---|---|
| [21] | Blobs | 7%-100% TP depending on configuration. 0%-25% FP. |
| [22] | Blobs | Only qualitative results given, no quantitative empirical analysis. |
| [23] | Blobs, motion characteristics | 64%-100% TP depending on event. 0%-29% FP. |
| [26] | Edges, Motion, blob's position, shape, and trajectory | 87.5%-100% TP depending on event. 0%-4% FP. |
| [29] | Motion and intensity | 5.9 average. queue length error in pixels over 255 measurements. Robust low contrast, illumination changes, and crowded scenes. |
| [30] | Level-lines | 98% TP, 2% FP |
| [32] | Intensity texture | 53.85-94.44% accuracy depending on type of crowd. Provides an output in terms of a range of densities |
| [60] | Blob's size, shape and clothing's color | 100% TP and 11%FP with 66% tracking accuracy |
| [104] | Motion | Results shown using penalized log-likelihood by the activity type |
| [107] | 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 |
| [108] | Optic flow | No empirical analysis |
| [109] | 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. |
| [110] | Statistical Methods, (Grey Level Dependency Matrix) GLDM | Total error is less than 12%. No FP rate is reported. |
| [116] | Motion, Blob's color, position, shape, and trajectory | Only visual sample results, no empirical analysis |
| [118] | 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 |
| [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. |
| [191] | 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 |
| [192] | 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% |
| [193] | 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%. |
| [194] | 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 |
| [195] | Motion, blob's position | Diurnal: 65-97% TP, 5% FP. Nocturnal: 0-91% TP, 6% FP |
| [196] | Motion, blob's color, centroid, position, height, and width | Only visual sample results, no empirical analysis |
| [197] | Motion | Results shown through polar plots where the direction angles are divided into a discrete range |

crowd density classifiers [110] will greatly benefit from having a realistic and robust training data set. However, new techniques for creating human crowd scenes are continuously being developed, particularly due to the growing demand from the motion picture industry [111]. Simulated crowds have widely been studied in many application domains, including emergency response [112] and large-scale panic situation modeling [113], [114]; perhaps, simulated crowds [115] 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: The first type uses an overhead camera, which contains "virtual gaits" that count the number of people crossing a predetermined area. Clearly, segmentation of a group of people into individuals is necessary for this purpose [116]. 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 a space between each person, whereas the same scenario from a side-view angle could incorrectly be segmented as one continuous object. When strictly people counting, some surveillance cameras are placed at bottlenecked entrance points, where at most one person at any given time is crossing some predetermined boundary (such as a security checkpoint or an access gate at a subway terminal). However, 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 [117]. Hence, using multiple cameras may further complicate crowd counting. In cases where overhead surveillance views are not available, side-view cameras must be used to count people, and the multiple problems associated with this view (e.g., 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 [118], [119]. However, most of these methods heavily rely on image quality and frame rate for accurate results. Shape indexing and skin colors are considered robust to poor video quality, whereas 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].

*3) Crowd Behavior:* Crowd behavior analysis has drawn significant interest from researchers closely working with the transit domain [122]. A recent survey [123] focused on crowd analysis methods employed in image processing. The flow of large human crowds [107] 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]–[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 (e.g., panic, fighting, and vandalism) is a new research direction for projects like SEcuRization KEeps Threats (SERKET) [131], which has recently been 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 [125]–[127]. Behavior classification is often based on the vector fields generated by crowd motion instead of individual person tracking.

*4) Human Pose Estimation (Stance Change):* In transit surveillance applications, human pose estimation refers to the pose of the entire human body (e.g., going from standing to lying down in a metro is an indication of pedestrian collapse), and not a pose related to a single body part, such as a head pose, that can be used in applications such as driving monitoring [132]. However, keeping track of multiple body parts is often useful to estimate the global body poses. In fact, there are two main approaches to estimating the body pose. The first approach calculates ratios between the height and the width of the bounding box around a detected human. In [133], 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 [134], [135], both because they are useful for indicating the human pose and because, when accurately modeled, they can be used to recover the pose even after occlusion and other common tracking failures [136]. Due to self-occlusion and background clutter, some approaches also use the motion generated from each body part as a feature for pose change [137], since movements from each joint are shown to be interdependent. In [138], the observed motion is compared with registered motion exemplars, whereas action models are used to estimate possible future poses.

### *B. Multiple-Person Interactions*

Multiple-person interactions have largely been motivated by the growing demand for recognizing suspicious activities in security and surveillance applications. In [139], the behavior detection process consists of foreground segmentation, blob detection, and tracking. Semantic descriptions of suspicious human behaviors are defined through groups of low-level blob-based events. For example, fights are defined as many blobs' centroid 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 Computer-assisted Prescreening of Video Streams or Unusual Activities (BEHAVE) (2004–2007) [140] and Context Aware Vision using Image-based Active Recognition (2002–2005) [141] have each produced several publications focusing on multiple-person interactions. Algorithms include the use of a nearest neighbor classifier based on

trajectory information [142] to detect human interactions such as walking together, approaching, ignoring, meeting, splitting, and fighting; Bayesian networks [143]; and moment-invariant feature descriptions [144] 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 a two-person interaction in [145], which used a hierarchical Bayesian network to recognize human behaviors based on body part segmentation and motion. This work was extended [146] to track multiple body parts of multiple people. Processing at three levels (pixel, blob, and object) was used to distinguish punching, handshaking, pushing, and hugging. A technique that does not use temporal motion information but instead uses pose is discussed in [147]. By using a string matching method using a $K$-nearest neighbor 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 [148], 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 pickpocketing is discussed in [149], and a real-time implementation of this behavior can be found in [150]. Other methods have extended the concept of "objects left behind" to analyze higher level information [190] of objects being "switched," i.e., changing hands. A noncontact hand gesture between people such as waiving was studied in [130]. This event was based on the localization of spatiotemporal patterns of each human motion and uses a shape-and-flow matching algorithm.

### C. Person–Vehicle Interactions

In general, transit systems involve surveillance of motorized vehicles and humans. Spatiotemporal relationships between people and vehicles for situational awareness [151] are the basis for analysis of "the big picture." However, operationally relevant behavior detection (e.g., human breaking in or vandalizing a car) has yet to be addressed in the research literature. As mentioned before, the focus of interest for this survey is human behavior recognition; however, 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 [152].

Most existing automated vehicle surveillance systems are based on trajectory analysis. Detected events are abnormal low-frequency ones (e.g., U-turns, sudden brake, and pedestrians trespassing the street) [153], [154] or a small group of predefined events, such as accidents [155], [156], illegal parking [157], congestion status [158], illegal turns, or lane driving [159]. Events of interest are commonly learned using expectation–maximization [160] or modeled using semantic rules [161] similar to the human interpretation of such events and validated using existing data. Trajectory-based approaches have been the subject of significant study, particularly in the traffic analysis domain. Common approaches to trajectory analysis are based on Kalman filter [162], [163], dynamic programming [164], and hidden Markov models [160]. Discrete behavior profiling has been proposed [165] to avoid tracking difficulties associated with occlusion and noise. There is significant research done in domain-independent anomaly behavior detection [166], [167], as well as events based on group activities [165]. Transit surveillance involves many subproblems, including classification of different types of vehicles [168]–[170], vehicle recognition [171], or discrimination between vehicles and other frequent objects [172], such as pedestrians, bicycles, buses, cars, pickups, trucks, and vans.

### D. Person–Facility/Location Interactions

*1) 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 (e.g., after hours) or spatial relationships (e.g., 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 "trip wire." A trip wire is typically a line drawn over the image, which separates regions into "allow" and "do not allow" areas. In [20], [21], and [23], 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 behaviors, 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 [173]. Trespasser hiding [139] 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 [174].

Another security-sensitive activity similar to intrusion is tailgating (i.e., illegal piggyback entry). Tailgating is a topic that has not received much attention in research but has been implemented in many commercial systems. 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 him, sometimes without the knowledge of the authorized person.

*2) 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 heavily rely on a tracking algorithm, since successful tracking allows the movement of the object to easily be estimated and later compared with acceptable motion vectors [175]. In some scenarios, the overall crowd characteristics, which do not

rely on the tracking of individual objects, may be sufficient [107]. For instance, the movement of large groups of people in an uncommon direction may indicate panic or danger. To entirely automate the process, motion vectors can be calculated in conjunction with a GMM to learn the correct directional patterns of traffic in the scene [176].

*3) Vandalism:* Vandalism is defined in [139] as irregular centroid motion of a blob, combined with detected changes in the background. This definition is also implemented in [177] when a blob enters a scene and causes changes in the background or predefined "vandalisable" areas. In [178], vandalism is detected in unmanned railway environments using a neural net by detecting erratic or strange behaviors of a single person or a group.

*4) Object Stationarity (Object Removal and Object Left Behind):* In this survey, object stationarity exclusively refers to nonanimated objects. In transit surveillance systems, objects left behind usually represent suspicious or potentially dangerous elements (e.g., a suitcase and a backpack). 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 PETS. Most algorithms presented a simple background subtraction method to find stationary objects that were not present before. Many other methods have been proposed to deal with objects left behind or removed. In [179], an edge matching algorithm is used, which compares the current frame to the background model to detect objects removed or left behind. In [21], 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 to 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 researchers [180] have admitted this weakness and mentioned ways to include crowd segmentation algorithms to improve stationarity detection performance.

## IV. State-of-the-Art Discussion and Future Developments

Future developments mentioned in the previous survey [4] include multimodal data fusion, robust occlusion handling, usage of 3-D data, and usage of personal identification. In this section, additional potential directions of work are explored. In addition, 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 papers in Table III offer an indication to the level of interest in this research area. As shown in Fig. 6, it is clear that behavior recognition is an active research topic. In fact, there have been three times as many publications in the last three years than the number of all publications found before 2005.

### A. Core Technology Limitations

Human behavior algorithms heavily rely on the available core technology. 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 available algorithms. 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 [4]. In this survey, we separate algorithms in terms of processing speed into two groups, namely, real time and offline processing (see Table III). Nevertheless, in the last decade, the image-processing community in this context agrees that the definition of real time is not clear, although many researchers use it in their systems [7]. This point brings the biggest concern to create 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 data set 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 overly simplify the real problem. Assumptions such as smoothness of motion, limited occlusion, illumination constancy, and high contrast with respect to background [73] effectively limit the algorithms' usability in real scenarios within the transit-surveillance domain.

### B. 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. However, there are no accepted baseline algorithms in behavior recognition for transit applications. Surprisingly, a few papers in Table III formally compare performance against any other related work, making behavior-detection algorithm comparison scarce in the literature. Dealing with new detection tasks that have not previously been studied 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 with 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 particularly difficult problems, including poor illumination with drastic lighting changes (e.g., 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 papers, found algorithms to be heavily constrained to assumptions [7] 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 concerns are assumptions of camera motion, camera parameters, field of view, background complexity, landmarks, lighting and weather conditions, crowd density,

TABLE III

PUBLICATIONS OF BEHAVIOR RECOGNITION ALGORITHMS APPLICABLE TO TRANSIT SURVEILLANCE SYSTEMS (O: DATA SET INCLUDES OUTDOOR DATA SETS; R: MENTIONS A REAL-TIME IMPLEMENTATION; C: DATA SET INCLUDES CROWDED SCENES)

| 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 | [144] |
| Bissacco | 08 | Human Pose | 2950 images of human walking in circle, unspecified resolution. | N | N | N | [136] |
| Jang | 08 | Human Pose | 600 images of unspecified resolution | N | N | N | [138] |
| 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 | [142] |
| Dong | 07 | People counting, crowd density | 2 videos | Y | Y | Y | [118] |
| Fathi | 07 | Human Pose | 1008 images (divided into 4 subjects), unspecified resolution | N | N | N | [137] |
| Ghazal | 07 | Theft, graffiti, defacing | 3 videos | Y | Y | N | [177] |
| 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 | [134] |
| Monteiro | 07 | Wrong direction | Unspecified number of 320x240 px images | Y | Y | N | [176] |
| Park | 07 | Human-vehicle situational awareness | 30 minute video | Y | N | Y | [151] |
| Park | 07 | Person – Person interaction, shaking hands, pointing, standing hand-in-hand | Train 30 images, test 38 images | N | N | N | [147] |
| Ribnick | 07 | Thrown Objects | Unspecified indoor and outdoor videos | Y | Y | N | [188] |
| 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 | [189] |
| Ferrando | 06 | Object left behind, object switching | 800 images | N | Y | N | [190] |
| Park | 06 | approaching, departing, handshaking, pointing, pushing, hugging | Unspecified number of sequences, 320x240px | N | N | N | [146] |
| 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 | [109] |
| Ribnick | 06 | Camera Tampering | Unspecified indoor and outdoor videos | Y | Y | Y | [191] |
| Sijun | 06 | Object ownership , Object stationarity | 92 training and 45 testing videos | N | N | N | [180] |
| Velastin | 06 | Circular and diverging flows, obstacle detection | Unspecified number of 512x512px grayscale videos | N | Y | Y | [107] |
| Wu | 06 | People counting, crowd density | 70 320x240px images | Y | N | Y | [110] |
| Angiati | 05 | Vandalism | 2 videos (diurnal and nocturnal) with 7 graffiti drawn | Y | N | N | [192] |
| Bird | 05 | Loitering | Train 205 images. Test 30 minutes 720x480px video | N | N | N | [60] |
| Black | 05 | Crossing , falling on, proximity, throwing objects to, walking on tracks | Entire CREDS dataset | N | Y | Y | [21] |
| Fuentes | 05 | Unattended luggage, intrusion into forbidden areas, falls onto tracks, People hiding, vandalism, fights | Unspecified number of 384x288px color images | N | Y | Y | [139] |
| Lee | 05 | Human Pose | PETS 2003 Smart meeting video | N | N | N | [135] |
| Liu | 05 | Virtual gate crowd counting, proximity to tracks | 1 10 minute video | N | N | Y | [116] |
| 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 | [104] |
| 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 | [22] |
| Seyve | 05 | Crossing, dropping, falling, proximity, throwing object, walking on tracks, trap by train door | Unspecified dataset from CREDS | N | Y | N | [23] |
| 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 | [26] |
| Aubert | 04 | Loitering, objects left behind | 436 stationary situations test cases on gray level 256x256px images | N | N | N | [30] |
| Fuentes | 04 | Objects left behind , intrusion, falls, hiding, vandalism/Graffiti, fights, attacks | Unspecified number of 384x288px color images | N | Y | N | [193] |
| Kang | 04 | Security breaches (i.e. wrong direction) | Dataset not specified | N | Y | N | [175] |
| Reisman | 04 | Crowd detection | 320x240px video from mobile platform | Y | Y | Y | [108] |
| 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 | [174] |
| Park | 03 | Approaching, departing, pointing, standing hand-in-hand, shaking hands, hugging, punching, kicking, and pushing | 56 320x240px sequences | N | N | N | [145] |
| Cupillard | 02 | Fighting, Blocking, Forbidden Zone, Pickpocket | 20 sequences | N | N | N | [149] |
| Sacchi | 01 | Graffiti, gang behavior: “agitated” and “calm” behavior | 270 frames for training, 118 image sequences for testing | N | N | N | [178] |
| Aubert | 99 | Queue length estimation | 255 measurements from 2 hours of video of airport scenes | N | Y | Y | [29] |
| Haritaoglu | 99 | Handbag detection, object exchange | 100 320x240px videos | Y | Y | N | [148] |
| Marana | 97 | Crowd density estimation | 151 train and 149 test images | N | Y | Y | [32] |
| 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 | [194] |

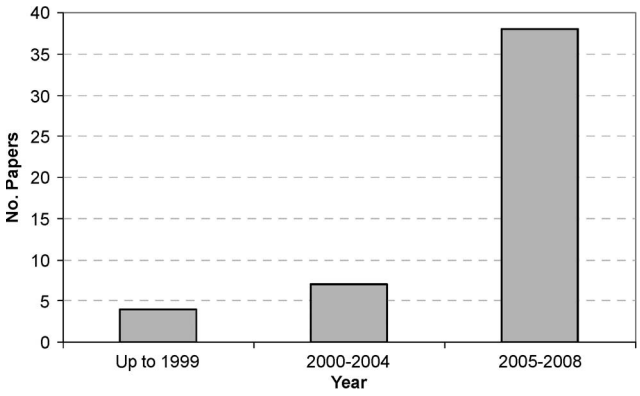



Fig. 6. Increasing interest on human behavior recognition research is shown through a comparison of a number of papers published up to 2008.

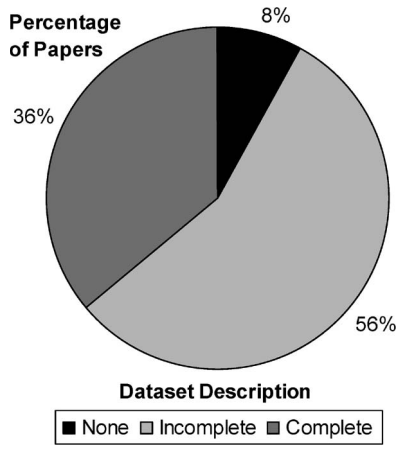



Fig. 7. Data set description analysis based on 52 transit surveillance-related papers surveyed in this paper. "None" refers to papers that include no reference to the data set used. "Complete" indicates that a full description is included, i.e., quantity and pixel resolution for both training and testing data. "Incomplete" indicates that there is some description but not enough to account for "Complete."

number and severity of occlusions, subject initialization or *a priori* information (e.g., known pose, movement, and tight-fitting clothes), 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.

### C. Standard Terminology

It is often assumed that crowds will be evenly distributed 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 (e.g., fire or police department and emergency agency). For example, in U.K., the Communities and Local Government regulations set the limit occupancy for a bar [181] to 0.3–0.5 $m^2$ 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 [109] and [182], definitions based on current practical safety guidelines are used. For example, very low density is defined as people$/m^2 < 0.5$, whereas very high density is defined as people$/m^2 > 2$. Other studies use less mathematically precise definitions such as "Overcrowding occurs when too many people congregate within a certain location. Congestion is a situation where it becomes difficult for an individual to move within a crowded area" [21]. A common approach is to describe a crowd in terms of the number of individuals in it, like in [32], where 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 way to define crowd levels in the literature. Additionally, it is hard to identify methods that directly refer to similar data sets in terms of crowd density.

### D. Data Sets

This survey has found across the literature the tendency to not fully specify the data set used. As shown in Fig. 7, most papers, regardless of the review process, chose to not completely disclose the data set description of their work. Clearly, this information is necessary when showing the significance of an algorithm and understanding their results. Moreover, relative improvements over other previously reviewed publications may be hard to quantify since a comparison of the data sets cannot be made. Consequently, it is often unclear what level of empirical validation is behind published techniques. An advantage of using similar or common data sets is that performance scores from different algorithms can directly be compared, as long as the evaluation framework is comparable. However, in general, transit security data are hard to come by, due to the difficulty of gathering an adequate supply of valid video sequences containing operationally relevant events [139], and overcome privacy and security concerns [193]. Initiatives like TREC Video Retrieval Evaluation [183] encourage research by providing large data set 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 data sets report concrete results only on very small portions of the data set 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. In this survey, we found these to be common problems in the literature.

### E. 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

TABLE IV
BEHAVIOR RECOGNITION SUMMARY ADVERTISED BY COMMERCIAL PROVIDERS IN THEIR WEBSITES

| Ref. | Manufacturer | Object Tracking | Breach | Loiter | Crowd Analysis | Stance Change | Object Left | Object Removal |
|---|---|---|---|---|---|---|---|---|
| [200] | Agent Vi | ✔ | ✔ | ✔ | ✔ | | ✔ | |
| [201] | Aimetis Corp | ✔ | ✔ | ✔ | ✔ | | ✔ | ✔ |
| [202] | Cernium Corp | ✔ | ✔ | ✔ | ✔ | | ✔ | |
| [203] | Eptascape, Inc | ✔ | ✔ | ✔ | ✔ | | ✔ | ✔ |
| [204] | Honeywell International, Inc | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| [205] | Indigo Vision | ✔ | ✔ | | ✔ | | ✔ | ✔ |
| [206] | Intelliview Technologies Inc | ✔ | ✔ | | | | | |
| [207] | Intellivision | ✔ | ✔ | | ✔ | | ✔ | ✔ |
| [208] | IPSOTEK Ltd | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| [209] | March Networks | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | |
| [210] | Mate Intelligent Video | ✔ | ✔ | ✔ | ✔ | | ✔ | ✔ |
| [211] | Object Video | ✔ | ✔ | ✔ | ✔ | | ✔ | ✔ |
| [212] | SightLogix Inc | ✔ | ✔ | | | | | |
| [213] | Verint | ✔ | ✔ | ✔ | ✔ | | ✔ | ✔ |
| [214] | Vidient | ✔ | ✔ | ✔ | ✔ | | ✔ | ✔ |
| [215] | Nice Systems | ✔ | ✔ | | ✔ | | ✔ | |
| [106] | TrueSentry, Inc. | ✔ | ✔ | ✔ | ✔ | | ✔ | ✔ |
| [216] | Ioimage, Ltd | ✔ | ✔ | ✔ | | | ✔ | ✔ |

interact to form a "bigger picture," which can potentially offer a viable solution to complex problems. Many transit systems have large sensor networks (e.g., audio, video, motion sensors, and smoke detectors) already in place. Under 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 [184]. In addition, sensors can often overcome weaknesses of other sensors; for example, fusing color and infrared video can be used to improve tracking through occlusions [185]. However, there is not much work reported on the integration of different types of sensors in automated video surveillance systems [5]. Multimodal fusion, such as audio and video [186] or infrared and stereovision [187], can potentially offer better scene understanding, thereby improving situational awareness and response time. For general distributed surveillance, see a detailed survey in [5] for more information.

### F. Aerial Surveillance

Moving cameras and mobile surveillance platforms have yet to become an important player in transit surveillance. With much research and commercial interest in UAVs and mobile surveillance platforms, current solutions are not far from being usable as an efficient surveillance platform for transit networks. Early works using surveillance video from UAVs [195], [196] describe behavior analysis algorithms for low-resolution vehicles to monitor roadblock checkpoints (e.g., avoiding, passing through, and getting closer). As aerial surveillance has gained increased interest within the research community, authors have proposed techniques to detect low-resolution vehicles [197] and buildings [198] 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 [199] from aerial video. As resolution and video quality increases, transit surveillance, including people, vehicles, and behavior analysis, is logically the next step.

### G. Commercial Systems

There are many commercial system providers who offer visual surveillance solutions for residential, commercial, government, and law enforcement agencies. Most modern systems now include video analytics software useful for automatic behavior recognition. Moreover, the actual storage of the video is significantly reduced if recording is performed only when an alarm is triggered by one of such behaviors. Customers are also able to specify defining attributes of moving objects, such as shape and speed, to provide means to efficiently search large video databases. Along with a visual cue on a security monitor, other common features include automatic e-mail or text messaging to cell phones or personal display assistants when alarms are triggered. In addition, geocoded mapping tools combined with real-time communication allow key personnel to further investigate the alarms. Table IV shows a summary of existing commercial systems, including current general behavior detection capabilities. Capabilities are based on information advertised by vendors in their websites as in April 2009. Due to the broad terminology used by different providers, we based behavior labels on the most common names found. In Table IV,

crowd analysis refers to any analytics that targets crowds in general; thus, it would include events like people counting, crowd density, and queue length monitoring.

As capabilities advertised by commercial providers increase, the necessity for an independent evaluation of such capabilities becomes increasingly 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, and therefore, we could not compare performance across platforms. In other words, we have no idea which system (vendor) will perform better using a given set of requirements. Let us consider detection of the loitering behavior as an example. Looking at Table IV, almost two thirds of vendors advertised loitering detection capabilities. Let us take into account that, as discussed earlier in this paper, we describe that, in [60], loitering is detected over long periods of time, including likely situations of subjects leaving the scene or being frequently occluded. However, it is unclear that any of the systems listed in this table can achieve the same results as in [60]. 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.

## V. Conclusion

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 research papers published in the last three years outnumbers the rest of the previous related literature threefold. This survey presented an overview of the state-of-the-art developments on behavior recognition algorithms for transit visual surveillance applications. A literature sample of 52 papers was used to study state-of-the-art strengths and weaknesses. Analysis includes behaviors, data sets, and implementation details. A classification strategy is presented that separates these papers by the targeted human behavior. The behavior groups are as follows: 1) single person or no interaction (i.e., behaviors exhibited by a single person that does not interact with any other person or vehicles); 2) multiple-person interactions; 3) person–vehicle interactions; and 4) person–facility/location interactions.

In this survey, a brief overview of the core technologies (i.e., all preprocessing steps before behavior recognition) has been included. There are many well-known limitations in the core technologies that need to be addressed. Techniques are often sensitivity to poor resolution, frame rate, drastic illumination changes, detrimental weather effects, and frequent occlusions, among other common problems prevalent in transit surveillance systems. Consequently, improved core technology algorithms are needed to increase the reliability of human behavior recognition. Over 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 set of terminology and generating operationally similar data sets. For example, a bus and a metro can both be "crowded." However, operationally, the "crowds" in both situations are very different. Thus, without a standard precise definition of "crowd," formal comparisons become a very difficult task.

There are vast amounts of untapped information present in surveillance video footage, which can be exploited for automatic behavior detection. However, there is still a big gap in analytical skills between a typical 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. Many think of computer technology as the only solution that is able to close that gap.

## Acknowledgment

The authors would like to thank D. Kelsey (Hart) and S. Godavarthy and W. Cheng (University of South Florida) for their involvement and support in the completion of this paper.

## References

[1] *Official website for Metropolitan Transportation Authority*. [Online]. Available: http://www.mta.info

[2] *Official website for Moscow Metro*. [Online]. Available: http://www.mosmetro.ru

[3] N. Sulman, T. Sanocki, D. Goldgof, and R. Kasturi, "How effective is human video surveillance performance?" in *Proc. Int. Conf. Pattern Recog*., 2008, pp. 1–3.

[4] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 3, pp. 334–352, Aug. 2004.

[5] M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems: A review," *Proc. Inst. Elect. Eng.—Vision, Image Signal Process.*, vol. 152, no. 2, pp. 192–204, Apr. 2005.

[6] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, "Computational studies of human motion: Part 1, tracking and motion synthesis," *Found. Trends Comput. Graph. Vis.*, vol. 1, no. 2/3, pp. 77–254, 2005.

[7] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Comput. Vis. Image Underst.*, vol. 81, no. 3, pp. 231–268, Mar. 2001.

[8] 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 Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-00-12, 2000.

[9] 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. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 319–336, Feb. 2009.

[10] F. Yin, D. Makris, and S. A. Velastin, "Performance evaluation of object tracking algorithms," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveillance*, 2007, pp. 733–736.

[11] J. Black, T. Ellis, and P. Rosin, "A novel method for video tracking performance evaluation," in *Proc. IEEE Int. Workshop Perform. Anal. Video Surveillance Tracking*, 2003, pp. 125–132.

[12] C. Erdem and B. Sankur, "Performance evaluation metrics for object-based video segmentation," in *Proc. X Eur. Signal Process. Conf.*, 2000, pp. 917–920.
[13] B. Georis, F. Bremond, M. Thonnat, and B. Macq, "Use of an evaluation and diagnosis method to improve tracking performances," in *Proc. IASTED Int. Conf. Vis., Imaging Image Process.*, 2003, pp. 827–832.
[14] V. Y. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, H. Li, D. Doermann, and T. Drayer, "Performance evaluation of object detection algorithms," in *Proc. Int. Conf. Pattern Recog.*, 2002, pp. 965–969.
[15] 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," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveillance*, 2005, pp. 1–8.
[16] D. Doermann and D. Mihalcik, "Tools and techniques for video performances evaluation," in *Proc. Int. Conf. Pattern Recog.*, 2000, pp. 167–170.
[17] S. Muller-Schneiders, T. Jager, H. S. Loos, and W. Niem, "Performance evaluation of a real time video surveillance system," in *Proc. IEEE Int. Workshop Visual Surveillance Perform. Eval. Tracking Surveillance*, 2005, pp. 137–143.
[18] T. List, J. Bins, J. Vazquez, and R. B. Fisher, "Performance evaluating the evaluator," in *Proc. IEEE Int. Workshop Visual Surveillance Perform. Eval. Tracking Surveillance*, 2005, pp. 129–136.
[19] 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," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2005, pp. 201–206.
[20] M. Spirito, C. S. Regazzoni, and L. Marcenaro, "Automatic detection of dangerous events for underground surveillance," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2005, pp. 195–200.
[21] J. Black, S. A. Velastin, and B. Boghossian, "A real time surveillance system for metropolitan railways," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2005, pp. 189–194.
[22] K. Schwerdt, D. Maman, P. Bernas, and E. Paul, "Target segmentation and event detection at videorate: The EAGLE project," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2005, pp. 183–188.
[23] C. Seyve, "Metro railway security algorithms with real world experience adapted to the RATP dataset," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2005, pp. 177–182.
[24] *Performance Evaluation of Tracking and Surveillance official website*. [Online]. Available: http://www.cvg.rdg.ac.uk/slides/pets.html
[25] J. Aguilera, H. Wildenauer, M. Kampel, M. Borg, D. Thirde, and J. Ferryman, "Evaluation of motion segmentation quality for aircraft activity surveillance," in *Proc. IEEE Int. Workshop Visual Surveillance Perform. Eval. Tracking Surveillance*, 2005, pp. 293–300.
[26] 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. Syst., Man, Cybern. A, Syst., Humans*, vol. 35, no. 1, pp. 164–182, Jan. 2005.
[27] 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," in *Proc. Inst. Eng. Technol. Conf. Crime Security*, 2006, pp. 470–475.
[28] C. I. Attwood and D. A. Watson, "Advisor-socket and see: Lessons learnt in building a real-time distributed surveillance system," *IEEE Intell. Distrib. Surveillance Syst.*, pp. 6–11, Feb. 2004.
[29] D. Aubert, "Passengers queue length measurement," in *Proc. IEEE Int. Conf. Image Anal. Process.*, 1999, pp. 1132–1135.
[30] 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, Feb. 2004.
[31] 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," in *Proc. IEEE Int. Conf. Image Anal. Process.*, 1997, pp. 757–764.
[32] A. N. Marana, L. F. Costa, S. A. Velastin, and R. A. Lotufo, "Estimation of crowd density using image processing," in *Proc. IEEE Colloq. Image Process. Security Appl.*, 1997, pp. 11/1–11/8.
[33] C. Cedras and M. Shah, "A survey of motion analysis from moving light displays," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 1994, pp. 214–221.
[34] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human–computer interaction: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 677–695, Jul. 1997.
[35] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, Jan. 2003.
[36] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, "Visual modeling with a hand-held camera," *Int. J. Comput. Vis.*, vol. 59, no. 3, pp. 207–232, Sep./Oct. 2004.
[37] T. Osawa, X. Wu, K. Wakabayashi, and T. Yasuno, "Human tracking by particle filtering using full 3D model of both target and environment," in *Proc. Int. Conf. Pattern Recog.*, 2006, vol. 2, pp. 25–28.
[38] A. Dominguez-Caneda, C. Urdiales, and F. Sandoval, "Dynamic background subtraction for object extraction using virtual reality based prediction," in *Proc. MELECON*, 2006, pp. 466–469.
[39] 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 Syst. Video Technol.*, vol. 17, no. 11, pp. 1568–1586, Nov. 2007.
[40] J. Heikkila and O. Silven, "A real-time system for monitoring of cyclists and pedestrians," in *Proc. IEEE Workshop Visual Surveillance*, 1999, pp. 74–81.
[41] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 1999, vol. 2, pp. 246–252.
[42] G. Halevy and D. Weinshall, "Motion of disturbances: Detection and tracking of multi-body non-rigid motion," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 1997, pp. 897–902.
[43] R. Cutler and L. Davis, "View-based detection and analysis of periodic motion," in *Proc. Int. Conf. Pattern Recog.*, 1998, pp. 495–500.
[44] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1999, pp. 255–261.
[45] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, Mar. 2005.
[46] V. Jain, B. B. Kimia, and J. L. Mundy, "Background modeling based on subpixel edges," in *Proc. IEEE Int. Conf. Image Process.*, 2007, vol. 6, pp. VI-321–VI-324.
[47] S.-C. Cheung and C. Kamath, "Robust background subtraction with foreground validation for urban traffic video," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 14, pp. 2330–2340, 2005.
[48] M. Hansen, P. Anandan, K. Dana, G. Van Der Wal, and P. Burt, "Real-time scene stabilization and mosaic construction," in *Proc. DARPA Image Underst. Workshop*, 1994, pp. 54–62.
[49] F.-Y. Hu, Y.-N. Zhang, and L. Yao, "An effective detection algorithm for moving object with complex background," in *Proc. IEEE Int. Conf. Mach. Learn. Cybern.*, 2005, vol. 8, pp. 5011–5015.
[50] 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," in *Proc. Hybrid Inf. Technol.*, 2006, vol. 1, pp. 263–269.
[51] M. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise smooth flow fields," *Comput. Vis. Image Underst.*, vol. 63, no. 1, pp. 75–104, Jan. 1996.
[52] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1–3, pp. 185–203, Aug. 1981.
[53] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. DARPA Image Underst. Workshop*, 1981, pp. 121–130.
[54] R. Szeliski and J. Coughlan, "Spline-based image registration," *Int. J. Comput. Vis.*, vol. 22, no. 3, pp. 199–218, Mar./Apr. 1997.
[55] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt, "Performance of optical flow techniques," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 1992, pp. 236–242.
[56] R. N. Hota, V. Venkoparao, and A. Rajagopal, "Shape based object classification for automated video surveillance with feature selection," in *Proc. IEEE Int. Conf. Inf. Technol.*, 2007, pp. 97–99.
[57] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 886–893.
[58] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 878–885.
[59] Q. Zhu, M. Yeh, K. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 1491–1498.
[60] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs, "Detection of loitering individuals in public transportation areas," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 167–177, Jun. 2005.
[61] B. C. Chee, M. Lazarescu, and T. Tan, "Detection and monitoring of passengers on a bus by video surveillance," in *Proc. IEEE Int. Conf. Image Anal. Process.*, 2007, pp. 143–148.

[62] 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. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[63] Y.-B. Li, T.-X. Jiang, Z.-H. Qiao, and H.-J. Qian, "General methods and development actuality of gait recognition," in *Proc. IEEE Int. Conf. Wavelet Anal. Pattern Recog.*, 2007, vol. 3, pp. 1333–1340.

[64] D. M. Gavrila, "The visual analysis of human movement: A survey," *Comput. Vis. Image Underst.*, vol. 73, no. 1, pp. 82–98, Jan. 1999.

[65] C. Cedras and M. Shah, "Motion-based recognition, A survey," *Image Vis. Comput.*, vol. 13, no. 2, pp. 129–155, Mar. 1995.

[66] S. Ju, "Human motion estimation and recognition (depth oral report)," Univ. Toronto, Toronto, ON, Canada, 1996.

[67] S.-H. Kim and H.-G. Kim, "Face detection using multi-modal information," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, 2000, pp. 14–19.

[68] S. Harasse, L. Bonnaud, and M. Desvignes, "Human model for people detection in dynamic scenes," in *Proc. Int. Conf. Pattern Recog.*, 2006, vol. 1, pp. 335–354.

[69] M.-T. Yang, Y.-C. Shih, and S.-C. Wang, "People tracking by integrating multiple features," in *Proc. Int. Conf. Pattern Recog.*, 2004, vol. 4, pp. 929–932.

[70] M. J. Jones and D. Snow, "Pedestrian detection using boosted features over many frames," in *Proc. Int. Conf. Pattern Recog.*, 2008, pp. 1–4.

[71] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 428–441.

[72] S. Haykin and N. DeFreitas, "Special issue on: Sequential state estimation: From Kalman filters to particle filters," *Proc. IEEE*, vol. 92, no. 3, pp. 399–400, 2004.

[73] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, pp. 1–45, 2006.

[74] L. M. Fuentes and S. A. Velastin, "People tracking in surveillance applications," *Image Vis. Comput.*, vol. 24, no. 11, pp. 1165–1171, Nov. 2006.

[75] N. Ning and T. Tan, "A framework for tracking moving target in a heterogeneous camera suite," in *Proc. IEEE Int. Conf. Control, Autom., Robot. Vis.*, 2006, pp. 1–5.

[76] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[77] A. Ess, B. Leibe, K. Schindler, and K. L. Van Gool, "A mobile vision system for robust multi-person tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[78] G. Gasser, N. Bird, O. Masoud, and N. Papanikolopoulos, "Human activities monitoring at bus stops," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2004, vol. 1, pp. 90–95.

[79] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.

[80] A. Tesei, A. Teschioni, C. S. Regazzoni, and G. Vernazza, "Long memory matching of interacting complex objects from real image sequences," in *Proc. Conf. Time Varying Image Process. Moving Objects Recog.*, 1996, pp. 283–286.

[81] M. Isard and A. Blake, "Condensation conditional density propagation for visual tracking," *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 5–28, Aug. 1998.

[82] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 661–675.

[83] U. Scheunert, H. Cramer, B. Fardi, and G. Wanielik, "Multi sensor based tracking of pedestrians: A survey of suitable movement models," in *Proc. Intell. Vehicles Symp.*, 2004, pp. 774–778.

[84] G. L. Foresti, C. S. Regazzoni, and P. K. Varshney, *Multisensor Surveillance Systems: The Fusion Perspective*. Norwell, MA: Kluwer, 2003.

[85] N. T. Siebel and S. Maybank, "Fusion of multiple tracking algorithms for robust people tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 373–382.

[86] B. Leibe, K. Schindler, and L. Van Gool, "Coupled detection and trajectory estimation for multi-object tracking," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[87] R. J. Morris and D. C. Hogg, "Statistical models of object interaction," in *Proc. IEEE Workshop Visual Surveillance*, 1998, pp. 81–85.

[88] 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," in *Proc. IEEE Workshop Visual Surveillance*, 1998, pp. 26–32.

[89] R. Nevatia, T. Zhao, and S. Hongeng, "Hierarchical language based representation of events in video steams," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshop*, 2003, vol. 4, pp. 39–46.

[90] R. Hamid, Y. Huang, and I. Essa, "ARGMode—Activity recognition using graphical models," in *Proc. IEEE Comput. Vis. Pattern Recog. Workshop*, 2003, vol. 4, pp. 38–44.

[91] R. Nevatia, J. Hobbs, and B. Bolles, "An ontology for video event representation," in *Proc. IEEE Workshop Event Detection Recog.*, 2004, p. 119.

[92] C. Rao and M. Shah, "View-invariant representation and learning of human action," in *Proc. IEEE Workshop Detection Recog. Events Video*, 2001, pp. 55–63.

[93] W. Kang and F. Deng, "Research on intelligent visual surveillance for public security," in *Proc. IEEE/ACIS Int. Conf. Comput. Inf. Sci.*, 2007, pp. 824–829.

[94] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Comput. Vis. Image Underst.*, vol. 73, no. 3, pp. 428–440, Mar. 1999.

[95] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata, "Articulated and elastic non-rigid motion: A review," in *Proc. Workshop Motion Non-Rigid Articulated Objects*, 1994, pp. 2–14.

[96] G. Shaffer, *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton Univ. Press, 1976.

[97] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Handling uncertainty in video analysis with spatiotemporal visual attention," in *Proc. Fuzzy Syst.*, 2005, pp. 213–217.

[98] P. Remagnino, T. Tan, and K. Baker, "Agent-oriented annotation in model based visual surveillance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1998, pp. 857–862.

[99] V. Girondel, A. Caplier, and L. Bonnaud, "A belief theory-based static posture recognition systems for real-time video surveillance applications," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2005, pp. 10–15.

[100] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russell, and J. Weber, "Automatic symbolic traffic scene analysis using belief networks," in *Proc. Nat. Conf. Artif. Intell.*, 1994, pp. 966–972.

[101] K. M. Kitani, Y. Sato, and A. Sugimoto, "Deleted interpolation using a hierarchical Bayesian grammar network for recognizing human activity," in *Proc. IEEE Int. Workshop Visual Surveillance Perform. Eval. Tracking Surveillance*, 2005, pp. 239–246.

[102] M. Brand and V. M. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 844–851, Aug. 2000.

[103] B. Morris and M. Trivedi, "An adaptive scene description for activity analysis in surveillance video," in *Proc. Int. Conf. Pattern Recog.*, 2008, pp. 1–4.

[104] J. Nascimento, M. Figueiredo, and J. S. Marques, "Segmentation and classification of human activities," in *Proc. Workshop Human Activity Recog. Modeling*, 2005, pp. 79–86.

[105] N. Haering and K. Shafique, "Automatic visual analysis for transportation security," in *Proc. IEEE Conf. Technol. Homeland Security*, 2007, pp. 13–18.

[106] D. Abrams and S. McDowall, "Video content analysis with effective response," in *Proc. IEEE Conf. Technol. Homeland Security*, 2007, pp. 57–63.

[107] 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," *Transp. Res. Part C: Emerging Technol.*, vol. 14, no. 2, pp. 96–113, Apr. 2006.

[108] P. Reisman, O. Mano, S. Avidan, and A. Shashua, "Crowd detection in video sequences," in *Proc. Intell. Vehicles Symp.*, 2004, pp. 66–71.

[109] H. Rahmalan, M. S. Nixon, and J. N. Carter, "On crowd density estimation for surveillance," in *Proc. Inst. Eng. Technol. Conf. Crime Security*, 2006, pp. 540–545.

[110] X. Wu, G. Liang, K. Lee, and Y. Xu, "Crowd density estimation using texture analysis and learning," in *Proc. IEEE Int. Conf. Robot. Biometrics*, 2006, pp. 214–219.

[111] S. Baek, I.-K. Jeong, and I.-H. Lee, "Implementation of crowd system in Maya," in *Proc. Int. Joint Conf. SICE-ICASE*, 2006, pp. 2713–2716.

[112] A. Shendarkar, K. Vasudevan, S. Lee, and Y.-J. Son, "Crowd simulation for emergency response using BDI agent based on virtual reality," in *Proc. Winter Simul. Conf.*, 2006, pp. 545–553.

[113] S. Banarjee, C. Grosan, and A. Abraham, "Emotional ant based modeling of crowd dynamics," in *Proc. Symbolic Numeric Algorithms Sci. Comput.*, 2005, pp. 279–286.

[114] N. Courty and S. R. Musse, "Simulation of large crowds in emergency situations including gaseous phenomena," in *Proc. Int. Conf. Comput. Graph.*, 2005, pp. 206–212.

[115] Y.-Y. Lin and Y.-P. Chen, "Crowd control with swarm intelligence," in *Proc. Evol. Comput.*, 2007, pp. 3321–3328.
[116] X. Liu, P. H. Tu, J. Rittscher, A. Perera, and N. Krahnstoever, "Detecting and counting people in surveillance applications," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2005, pp. 306–311.
[117] I. Cohen, A. Garg, and T. S. Huang, "Vision-based overhead view person recognition," in *Proc. Int. Conf. Pattern Recog.*, 2000, vol. 1, pp. 1119–1124.
[118] L. Dong, V. Parameswaran, V. Ramesh, and I. Zoghlami, "Fast crowd segmentation using shape indexing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[119] H. Wang and D. Suter, "Tracking and segmenting people with occlusions by a sample consensus based method," in *Proc. IEEE Int. Conf. Image Process.*, 2005, vol. 2, pp. 410–413.
[120] S. Lin, J. Chen, and H. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 31, no. 6, pp. 645–654, Nov. 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," in *Proc. Int. Conf. Pattern Recog.*, 2008, pp. 1–4.
[122] B. Maurin, O. Masoud, and N. P. Papanikolopoulos, "Tracking all traffic: Computer vision algorithms for monitoring vehicles, individuals, and crowds," *IEEE Robot. Autom. Mag.*, vol. 12, no. 1, pp. 29–36, Mar. 2005.
[123] B. Zhan, N. D. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, "Crowd analysis: A survey," *Mach. Vis. Appl.*, vol. 19, no. 5/6, pp. 345–357, Sep. 2008.
[124] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2006, vol. 1, pp. 705–711.
[125] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modeling crowd scenes for event detection," in *Proc. Int. Conf. Pattern Recog.*, 2006, vol. 1, pp. 175–178.
[126] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Hidden Markov models for optical flow analysis in crowds," in *Proc. Int. Conf. Pattern Recog.*, 2006, vol. 1, pp. 460–463.
[127] E. L. Andrade, R. B. Fisher, and S. Blunsden, "Detection of emergency events in crowded scenes," in *Proc. Inst. Eng. Technol. Conf. Crime Security*, 2006, pp. 528–533.
[128] E. Andrade, S. Blunsden, and R. Fisher, "Performance analysis of event detection models in crowded scenes," in *Proc. Workshop Towards Robust Visual Surveillance Techn. Syst. Visual Inf. Eng.*, 2006, pp. 427–432.
[129] J. H. Yin, S. A. Velastin, and A. C. Davies, "Image processing techniques for crowd density estimation using a reference image," in *Proc. Asian Conf. Comput. Vis.*, 1995, pp. 489–498.
[130] Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[131] S. Antipolis, "Intelligent environments for problem solving by autonomous systems," Institut National de Recherche en Informatique et en Automatique, Rocquencourt, France, p. 41, 2007.
[132] Y. Zhu and K. Fujimura, "Head pose estimation for driver monitoring," in *Proc. IEEE Intell. Vehicles Symp.*, 2004, pp. 501–506.
[133] 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," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[134] M. W. Lee and R. Nevatia, "Body part detection for human pose estimation and tracking," in *Proc. Motion Video Comput.*, 2007, p. 23.
[135] M. W. Lee and R. Nevatia, "Dynamic human pose estimation using Markov chain Monte Carlo approach," in *Proc. Motion Video Comput.*, 2005, vol. 2, pp. 168–175.
[136] A. Bissacco, M. H. Yang, and S. Soatto, "Fast human pose estimation using appearance and motion via multi-dimensional boosting regression," in *Proc. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
[137] A. Fathi and G. Mori, "Human pose estimation using motion exemplars," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[138] A. Baumberg and D. Hogg, "An efficient method for contour tracking using active shape models," in *Proc. IEEE Workshop Motion Non-Rigid Articulated Objects*, 1994, pp. 194–199.
[139] L. M. Fuentes and S. A. Velastin, "Tracking-based event detection for CCTV systems," *Pattern Anal. Appl.*, vol. 7, no. 4, pp. 356–364, Dec. 2004.
[140] *BEHAVE official website*. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/BEHAVE/
[141] *CAVIAR Project dataset*. [Online]. Available: http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/
[142] S. Blunsden, E. Andrade, and R. Fisher, "Non parametric classification of human interaction," in *Proc. 3rd Iberian Conf. Pattern Recog. Image Anal.*, 2007, pp. 347–354.
[143] A. Madabhushi and J. K. Aggarwal, "A Bayesian approach to human activity recognition," in *Proc. IEEE Workshop Visual Surveillance*, 1999, pp. 25–32.
[144] H. Yasin and S. A. Khan, "Moment invariants based human mistrustful and suspicious motion detection, recognition and classification," in *Proc. Comput. Modeling Simul.*, 2008, pp. 734–739.
[145] S. Park and J. K. Aggarwal, "Recognition of two-person interactions using a hierarchical Bayesian network," in *Proc. Int. Workshop Video Surveillance*, 2003, pp. 65–76.
[146] S. Park and J. K. Aggarwal, "Simultaneous tracking of multiple body parts of interacting persons," *Comput. Vis. Image Underst.*, vol. 102, no. 1, pp. 1–21, Apr. 2006.
[147] S. Park and J. K. Aggarwal, "Recognition of human interaction using multiple features in gray scale images," in *Proc. Int. Conf. Pattern Recog.*, 2000, vol. 1, pp. 51–54.
[148] I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis, "Backpack: Detection of people carrying objects using silhouettes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1999, vol. 1, pp. 102–107.
[149] F. Cupillard, F. Bremond, and M. Thonnat, "Group behavior recognition with multiple cameras," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 2002, pp. 177–183.
[150] A. Avanzi, F. Brémond, C. Tornieri, and M. Thonnat, "Design and assessment of an intelligent activity monitoring platform," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 14, pp. 2359–2374, 2005.
[151] S. Park and M. M. Trivedi, "Homography-based analysis of people and vehicle activities in crowded scenes," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 2007, p. 51.
[152] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 694–711, May 2006.
[153] F. Jiang, Y. Wu, and A. K. Katsaggelos, "Abnormal event detection from surveillance video by dynamic hierarchical clustering," in *Proc. IEEE Int. Conf. Image Process.*, 2007, vol. 5, pp. V-145–V-148.
[154] F. Jiang, Y. Wu, and A. K. Katsaggelos, "Abnormal event detection based on trajectory clustering by 2-depth greedy search," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 2129–2132.
[155] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi, "Traffic monitoring and accident detection at intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 2, pp. 108–118, Jun. 2000.
[156] X. Chen and C. Zhang, "Incident retrieval in transportation surveillance videos—An interactive framework," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2007, pp. 2186–2189.
[157] J. T. Lee, M. S. Ryoo, M. Riley, and J. K. Aggarwal, "Real-time detection of illegally parked vehicles using 1-D transformation," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2007, pp. 254–259.
[158] 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," in *Proc. IEEE Int. Conf. Mach. Learn. Cybern.*, 2002, vol. 2, pp. 934–938.
[159] V. Kettnaker and M. Brand, "Minimum-entropy models of scene activity," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 1999, vol. 1, pp. 281–286.
[160] X. Li and F. M. Porikli, "A hidden Markov model framework for traffic event detection using video features," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 24–27, 2004, vol. 5, pp. 2901–2904.
[161] S. Kamijo, M. Harada, and M. Sakauchi, "An incident detection system based on semantic hierarchy," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, Oct. 3–6, 2004, pp. 853–858.
[162] H. Veeraraghavan, P. Schrater, and N. Papanikolopoulos, "Switching Kalman filter-based approach for tracking and event detection at traffic intersections," in *Proc. Intell. Control, Med. Conf. Control Autom.*, Jun. 27–29, 2005, pp. 1167–1172.
[163] H. Y. Cheng and J. N. Hwang, "Multiple-target tracking for crossroad traffic utilizing modified probabilistic data association," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 1, pp. I:921–I:924.
[164] P. Kumar, S. Ranganath, H. Weimin, and K. Sengupta, "Framework for real-time behavior interpretation from traffic video," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 1, pp. 43–53, Mar. 2005.
[165] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, vol. 2, pp. 742–749.
[166] P. G. Raeth and D. A. Bertke, "Finding events automatically in continuously sampled data streams via anomaly detection," in *Proc. Nat. Aerosp. Electron. Conf.*, 2000, pp. 580–587.

[167] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 893–908, May 2008.
[168] A. J. Lipton and N. Haering, "Commode: An algorithm for video background modeling and object segmentation," in *Proc. IEEE Int. Conf. Control, Autom., Robot. Vis.*, 2002, vol. 3, pp. 1603–1608.
[169] H. Tao, H. S. Sawhney, and R. Kumar, "Object tracking with Bayesian estimation of dynamic layer representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 75–89, Jan. 2002.
[170] N. Buch, J. Orwell, and S. A. Velastin, "Detection and classification of vehicles for urban traffic scenes," in *Proc. Int. Conf. Visual Inf. Eng.*, 2008, pp. 182–187.
[171] O. Sidla, L. Paletta, Y. Lypetskyy, and C. Janner, "Vehicle recognition for highway lane survey," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2004, pp. 531–536.
[172] J. A. Vijverberg, N. A. H. M. De Koning, J. Han, P. H. N. De With, and D. Cornelissen, "High-level traffic-violation detection for embedded traffic analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 2, pp. 793–796.
[173] A. Makarov, J.-M. Vesin, and M. Kunt, "Intrusion detection using extraction of moving edges," in *Proc. Int. Conf. Pattern Recog.*, 1994, vol. 1, pp. 804–807.
[174] V. Kettnaker, "Time-dependent HMMs for visual intrusion detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog. Workshop*, 2003, vol. 4, p. 34.
[175] S. Kang, B. Abidi, and M. Abidi, "Integration of color and shape for detecting and tracking security breaches in airports," in *Proc. 38th Annu. Int. Carnahan Conf. Security Technol.*, 2004, pp. 289–294.
[176] G. Monteiro, M. Ribeiro, J. Marcos, and J. Batista, "Wrongway drivers detection based on optical flow," in *Proc. IEEE Int. Conf. Image Process.*, 2007, vol. 5, pp. V-141–V-144.
[177] M. Ghazal, C. Vazquez, and A. Amer, "Real-time automatic detection of vandalism behavior in video sequences," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2007, pp. 1056–1060.
[178] C. Sacchi, C. S. Regazzoni, and G. Vernazza, "A neural network-based image processing system for detection of vandal acts in unmanned railway environments," in *Proc. IEEE Int. Conf. Image Anal. Process.*, 2001, pp. 529–534.
[179] P. Spagnolo, A. Caroppo, M. Leo, T. Martiriggiano, and T. D'Orazio, "An abandoned/removed objects detection algorithm and its evaluation on PETS datasets," in *Proc. IEEE Int. Conf. Video Signal Based Surveillance*, 2006, p. 17.
[180] S. Lu, J. Zhang, and D. Feng, "A knowledge-based approach for detecting unattended packages in surveillance video," in *Proc. IEEE Int. Conf. Video Signal Based Surveillance*, 2006, p. 110.
[181] *Fire Safety Risk Assessment—Small and Medium Places of Assembly*, The Stationary Office, Edinburg, U.K., Jun. 5, 2006.
[182] G. K. Still, "Crowd dynamics," Ph.D. dissertation, Math. Dept., Warwick Univ., Coventry, U.K., 2000.
[183] *TREC Video Retrieval Evaluation official website*. [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/
[184] T. Ahmedali and J. J. Clark, "Collaborative multi-camera surveillance with automated person detection," in *Proc. IEEE Can. Conf. Comput. Robot Vis.*, 2006, p. 39.
[185] F. Bunyak, K. Palaniappan, S. K. Nath, and G. Seetharaman, "Geodesic active contour based fusion of visible and infrared video for persistent object tracking," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 2007, p. 35.
[186] B. P. L. Lo, J. Sun, and S. A. Velastin, "Fusing visual and audio information in a distributed intelligent surveillance system for public transport systems," *Acta Autom. Sin.*, vol. 29, no. 3, pp. 393–407, May 2003.
[187] S. J. Krotosky and M. M. Trivedi, "Person surveillance using visual and infrared imagery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1096–1105, Aug. 2008.
[188] E. Ribnick, S. Atev, N. Papanikolopoulos, O. Masoud, and R. Voyles, "Detection of thrown objects in indoor and outdoor scenes," in *Proc. Intell. Robots Syst.*, 2007, pp. 979–984.
[189] N. Bird, S. Atev, N. Caramelli, R. Martin, O. Masoud, and N. Papanikolopoulos, "Real time, online detection of abandoned objects in public areas," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2006, pp. 3775–3780.
[190] S. Ferrando, G. Gera, M. Massa, and C. S. Regazzoni, "A new method for real time abandoned object detection and owner tracking," in *Proc. IEEE Int. Conf. Image Process.*, 2006, pp. 3329–3332.
[191] E. Ribnick, S. Atev, O. Masoud, N. Papanikolopoulos, and R. Voyles, "Real-time detection of camera tampering," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2006, p. 10.
[192] D. Angiati, G. Gera, S. Piva, and C. S. Regazzoni, "A novel method for graffiti detection using change detection algorithm," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2005, pp. 242–246.
[193] L. Fuentes and S. A. Velastin, "Advanced surveillance: From tracking to event detection," *IEEE Latin America Trans.*, vol. 2, no. 3, pp. 206–211, Sep. 2004.
[194] 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," in *Proc. Int. Conf. Road Traffic Monitoring Control*, 1994, pp. 127–132.
[195] R. Nevatia, G. Medioni, and I. Cohen, "Event detection and analysis from video streams," in *Proc. IUW*, 1998, pp. 63–72.
[196] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 873–889, Aug. 2001.
[197] T. Zhao and R. Nevatia, "Car detection in low resolution aerial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, vol. 1, pp. 710–717.
[198] Z. W. Kim and R. Nevatia, "Automatic description of complex buildings from multiple images," *Comput. Vis. Image Underst.*, vol. 96, no. 1, pp. 60–95, Oct. 2004.
[199] M. Hu, S. Ali, and M. Shah, "Detecting global motion patterns in complex videos," in *Proc. Int. Conf. Pattern Recog.*, 2008, pp. 1–5.
[200] *The official website for AgentVi*. [Online]. Available: http://www.agentvi.com/
[201] *The official website for Aimetis Corp*. [Online]. Available: http://www.aimetis.com
[202] *The official website for Cernium*. [Online]. Available: http://www.cernium.com
[203] *The official website for Eptascape*. [Online]. Available: http://www.eptascape.com
[204] *The official website for Honeywell*. [Online]. Available: http://www51.honeywell.com
[205] *The official website for Indigo Vision*. [Online]. Available: http://www.indigovision.com
[206] *The official website for Intelliview*. [Online]. Available: http://www.intelliview.ca/
[207] *The official website for Intellivision*. [Online]. Available: http://www.intelli-vision.com
[208] *The official website for Ipsotek*. [Online]. Available: http://www.ipsotek.com
[209] March Networks. [Online]. Available: http://www.marchnetworks.com
[210] MATE Intelligent Video. [Online]. Available: http://www.mate.co.il
[211] *The official website for Object Video*. [Online]. Available: http://www.objectvideo.com
[212] *The official website for Sightlogix*. [Online]. Available: http://www.sightlogix.com/
[213] *The official website Verint*. [Online]. Available: http://verint.com
[214] *The official website for Vidient*. [Online]. Available: http://www.vidient.com/
[215] *The official website for Nice*. [Online]. Available: http://www.nice.com/products/video/index.php
[216] *The official website of IoImage*. [Online]. Available: http://www.ioimage.com

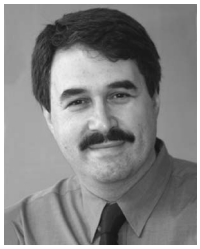
**Joshua Candamo** received the Ph.D. degree in computer science from the University of South Florida, Tampa, in 2009.

He is currently the Chief Executive Officer of K9 Bytes, Inc., Tampa, which is a software solution provider for the pet care industry. In K9 Bytes, he has taken the pet care industry to a new level of technology innovation. His main areas of research and technical publications are in the fields of image processing and pattern recognition.

**Matthew Shreve** (M'06) received the B.S. degree in computer science and the M.S. degree in mathematics from Youngstown State University, Youngstown, OH, in 2004 and 2006, respectively. He is currently working toward the Ph.D. degree in computer science and engineering with the University of South Florida, Tampa.

His research interests include computer vision, image processing, pattern recognition, and artificial intelligence applied to facial motion analysis and surveillance applications.

Mr. Shreve has served on organizing committees for international conferences on both abstract mathematics and computer science.



**Dmitry B. Goldgof** (M'83–SM'90–F'07) received the Ph.D. degree from the University of Illinois at Urbana-Champaign.

He is currently a Professor and an Associate Chair of the Department of Computer Science and Engineering, University of South Florida, Tampa. He has graduated 14 Ph.D. students, has published more than 65 journal and 145 conference papers and 16 books chapters, and has edited four books. His research interests include image and video analysis, pattern recognition, and bioengineering.

Prof. Goldgof is an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS and the *International Journal of Pattern Recognition and Artificial Intelligence*.

**Deborah B. Sapper** received the B.S. degree in geography from Michigan State University, East Lansing.

She is currently a Senior Researcher Associate with the Center for Urban Transportation Research, University of South Florida, Tampa. She specializes in transit safety and security, bus system safety and security program reviews, transit event investigation, transit planning, and policy analysis. She has 20 years of experience in the transit and the transportation field, working for the Fairfax County Transit Division, the Southeast Michigan Council of Governments, and the American Association of State Highway Transportation Officials.

**Rangachar Kasturi** (S'82–M'82–SM'88–F'96) received the B.E. (electrical) degree from Bangalore University, Bangalore, India, in 1968 and the M.S.E.E. and Ph.D. degrees from Texas Tech University, Lubbock, in 1980 and 1982, respectively.

He was a Professor of computer science and engineering and electrical engineering with the Pennsylvania State University, University Park, during 1982–2003 and was a Fulbright Scholar during 1999. He is currently with the Department of Computer Science and Engineering, University of South Florida, Tampa. He has coauthored the textbook *Machine Vision* (McGraw-Hill, 1995). His research interests are in document image analysis, video sequence analysis, and biometrics.

Dr. Kasturi is a Fellow of the International Association for Pattern Recognition (IAPR). He served as the President of the IEEE Computer Society in 2008 and as the President of the IAPR during 2002–2004. He served as the Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE during 1995–1998 and *Machine Vision and Applications* during 1993–1994.