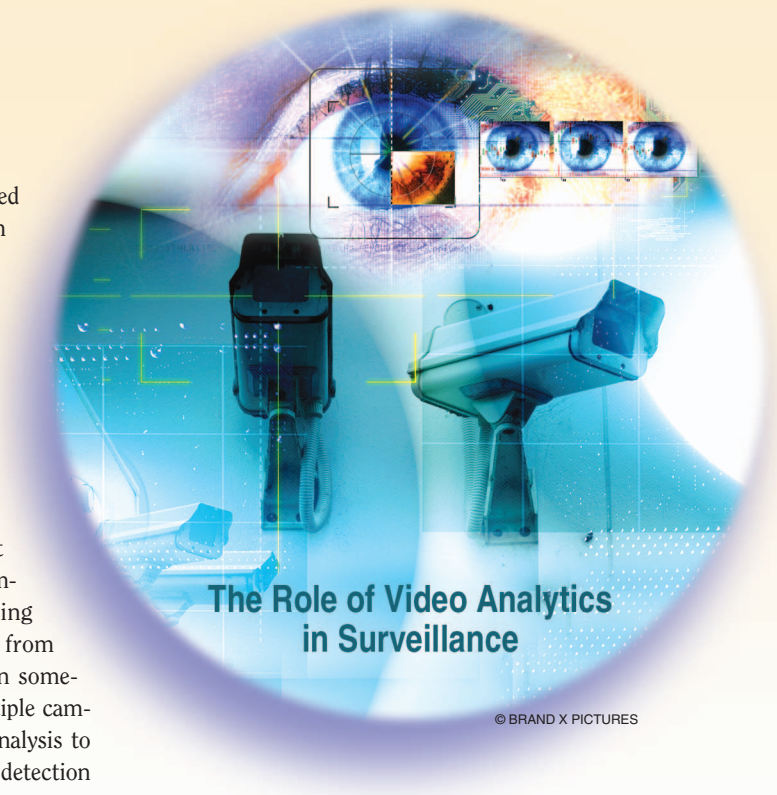
Venkatesh Saligrama, Janusz Konrad, and Pierre-Marc Jodoin

Video Anomaly Identification

A statistical approach



This article describes a family of unsupervised approaches to video anomaly detection based on statistical activity analysis. Approaches based on activity analysis provide intriguing possibilities for region-of-interest (ROI) processing since relevant activities and their locations are detected prior to higher-level processing such as object tracking, tagging, and classification. This strategy is essential for scalability of video analysis to cluttered environments with a multitude of objects and activities. Activity analysis approaches typically do not involve object tracking, and yet they inherently account for spatiotemporal dependencies. They are robust to clutter arising from multiple activities and contamination arising from poor background subtraction or occlusions. They can sometimes also be employed for fusing activities from multiple cameras. We illustrate successful application of activity analysis to anomaly detection in various scenarios, including the detection of abandoned objects, crowds of people, and illegal U-turns.

INTRODUCTION

Video camera networks are ubiquitous. Over 30 million cameras produce close to 4 billion hours of video footage per week in the United States alone [1]. This proliferation is taking place since, unlike other sensing modalities, visible-light cameras provide excellent space-time resolution, long capture range, wide field of view, and low latency. Consequently, large-scale camera networks can provide pervasive, wide-area monitoring to protect national infrastructures. Such a protection necessitates careful video analysis in suitable data space (such as photometric, behavior, and attitude) and with appropriate objectives, from counting objects and object localization to identifying suspicious movement of people or assets.

Our focus in this article is on identifying anomalous activity in space and time. Currently, this type of video analysis requires significant human supervision. However, since human supervision is not scalable, most of the video data from surveillance cameras is usually stored and rarely processed [1]. Clearly, autonomous analysis algorithms for networked camera systems operating in unstructured and highly cluttered indoor and outdoor environments are of interest. Furthermore, as computing power available in cameras increases, the computational intelligence is expected to move to the network edge, i.e., individual camera nodes, thus calling for distributed solutions.

There exist fundamental technical challenges to realizing this goal. First, urban scenarios provide a deluge of dynamic data. Identifying relevant information, such as meaningful change detection, in urban clutter is not easy. Second, doing so reliably, i.e., with small false alarm and miss rates, is difficult and perhaps impossible in harsh sensing environments (e.g., camera jitter). Third, combining views from multiple cameras for dynamic scene characterization and to improve detection, localization, and tracking performance is challenging.

Digital Object Identifier 10.1109/MSP.2010.937393

1053-5888/10/$26.00©2010IEEE

We describe recent developments in statistical, event-ensemble framework for dynamic scene modeling and anomaly detection and localization. This framework, which has been evolving over the past decade, has an interesting ROI interpretation. The main idea is to detect relevant activities and corresponding locations prior to higher-level analysis, such as object or activity tracking, tagging, and classification. This framework places its emphasis on modeling contextual and behavioral attributes at different locations and times rather than the conventional emphasis on object tracking and classification. One of the advantages of this framework is its robustness to pose and illumination changes. Some of these contextual attributes are geometrically invariant to camera zoom and orientation, leading to new concepts for characterizing spatial correspondences between different locations in single as well as multiple cameras. This is particularly valuable in a heterogeneous camera network wherein the cameras have different locations (wide-baseline scenarios), orientations with respect to the scene, and zoom levels and in which the inherent topology of the network is dynamic, i.e., the cameras can often change their orientations and zoom levels. The material for this article has been largely drawn from our earlier work [2]–[6].

VIDEO ANOMALY DETECTION: CHALLENGES

Anomalies or outliers, as described in one of the survey papers [7], are observations that do not conform to a well-defined notion of normal behavior. For example, if most observations associated with nominal behavior belong to some set, then observations outside (or sufficiently far away from) this set can be considered as outliers or anomalies. A statistical framework is often used to describe anomaly detection. In this setting, we are given features, ℓ, in a suitably high-dimensional space. Feature instances are distributed with a probability density function (pdf), $g_0(\cdot)$, if they come from a nominal distribution. Anomalous instances are distributed with pdf, $g_1(\ell)$. In the statistical framework, the anomaly detection problem amounts to predicting whether an instance ℓ is distributed according to nominal or anomalous pdf. Thus, the detection problem can be stated as follows:

$$H_0 : \ell \overset{\text{dist.}}{\sim} g_0(\cdot) \text{ versus the alternative (anomaly) } H_1 : \ell \overset{\text{dist.}}{\sim} g_1(\cdot). \tag{1}$$

If both pdfs are either known or can be estimated from training data this task reduces to the well-known likelihood ratio test (LRT) [8].

Nevertheless, video anomaly detection is generally difficult because of the following issues:

- The nominal and anomalous distributions are generally unknown and difficult to estimate even when the training data is given. This is because the boundary between normal and anomalous behavior depends on the choice of feature descriptors and distance metrics used that, in turn, significantly affect the performance. Anomalous behavior is typically not apparent from raw video (pixel intensities) and the video data needs to be transformed to a feature space where anomalies may become apparent.
- Both normality and abnormality are diverse since a typical urban video contains a multitude of activities. Therefore, there is no single distribution that captures either the nominal or anomalous activity.
- A critical issue is the general lack of labeled data for training/validation. This is particularly acute for anomalous data, since it is difficult to stage a rich set of anomalies, which is representative of a real-world scenario.
- Computational speed, in addition to statistical performance, is also an important performance metric for an anomaly detection system. For instance, there maybe a need for real-time detection in surveillance situations.
- Finally, the nominal background activity is nonstationary over a long time scale, e.g., activities during the day are significantly different from those at night. Consequently, the nonstationarity has to be accounted for.

There are many types of video anomalies, and it is generally difficult to group them. One possible classification is based on the fundamental dynamical nature of video data. In this perspective, video anomalies can be thought of as the presence/absence of usual (or unusual) objects or motion attributes in unusual (or usual) locations or times. For instance, an abandoned-object anomaly occurs when an object is present at an unusual location for an extended period of time. A trespassing anomaly occurs when motion is present at unusual locations. An illegal U-turn anomaly occurs when an unusual motion attribute appears at the usual locations. Clearly, video anomalies can either be localized, as in the case of abandoned objects, or spatially distributed, as in the case of illegal U-turns.

A significant effort has been devoted to video anomaly detection in surveillance applications over the last decade. The general problem, however, is still open because of the wide variety of anomalies; most of the existing anomaly detection techniques solve the problem in a specific scenario. In the following, we review some major approaches to video anomaly detection. These techniques differ both in terms of what is known about the training data as well as the different transformations and metrics used for anomaly detection.

VIDEO ANOMALY DETECTION: STATE OF THE ART

The existing techniques to video anomaly detection can be broadly classified into supervised and unsupervised approaches.

SUPERVISED APPROACHES

In the supervised video anomaly detection problem, the collection of anomalies are assumed to be known. One first constructs a dictionary of anomalies, and then, for each observed video, one checks if a match can be found in the dictionary. From the statistical perspective, the problem reduces to a conventional classification problem in machine learning. Once a set of features ℓ is selected, pdfs under nominal and anomalous distribution can be estimated (or implicitly characterized) and, finally, an LRT (which usually can be reduced to the

evaluation of a distance metric) can be applied to detect anomalous behavior. This approach has been used for many interesting scenarios, including detecting people carrying boxes, suitcases, and handbags [9] or detecting abandoned objects [10]. The main drawback of the supervised approach is that anomalous instances are far fewer compared to the normal instances in the training data, and obtaining representative instances of anomalies is generally difficult. Also, the method does not generalize to new anomalous patterns.

THE MAIN IDEA IS TO DETECT RELEVANT ACTIVITIES AND CORRESPONDING LOCATIONS PRIOR TO HIGHER-LEVEL ANALYSIS.

UNSUPERVISED APPROACHES

In the past decade, several researchers have focused on an alternative unsupervised approach. Unsupervised approaches generally imply that labels are unknown. However, there is an implicit assumption that the number of nominal instances far outnumber the anomalous ones. For this reason, there is an overlap between unsupervised and semisupervised approaches where the nominal instances are provided as a training set. In these approaches, the central aspect is the modeling of nominal activity either automatically or based on a training set of nominal video patterns.

OBJECT-TRACKING-BASED METHODS

A number of methods have been developed for learning two-dimensional (2-D) motion paths [11], [12] resulting from tracking of objects or people [13]. Here, a large number of normal individuals or objects are tracked over time during the training phase. The resulting paths are then summarized by a set of motion trajectories, often translated into a symbolic representation of the background activity. In the detection phase, paths extracted from the monitored video are compared against those extracted in the training phase. Tracking is generally performed by means of graphical state-based representations, such as hidden Markov models or Bayesian networks [13]–[17]. Johnson and Hogg [18] consider human trajectories in this context. The method begins by vector-quantizing tracks and clustering the result into a predetermined number of pdfs using a neural network. Based on the training data, the method predicts trajectory of a pedestrian and decides if it is anomalous or not. This approach was subsequently improved by simplifying the training step [19] and embedding it into a hierarchical structure based on co-occurrence statistics [20]. More recently, Saleemi et al. [21] proposed a stochastic, nonparametric method for modeling scene tracks. The authors claim that the use of predicted trajectories and tracking method robust to occlusions jointly permit the analysis of more general scenes, unlike other methods that are limited to roads and walkways.

From a statistical perspective, the tracks amount to features ℓ here and a nominal distribution of tracks, $g_0(\ell)$, is obtained in the training phase. For the anomalous tracks, the implicit assumption is that the tracks are uniformly distributed over the set of all tracks. In this statistical perspective, the optimal detection rule is a thresholding strategy, whereby outliers with respect to the nominal tracks are declared abnormal. In particular, one can define a negative log-likelihood function as follows:

$$\Lambda(\ell) = -\log(g_0(\ell)) \underset{\text{nominal}}{\overset{\text{anomalous}}{\gtrless}} \theta. \tag{2}$$

Although there are advantages to using paths as motion features, there are clear disadvantages as well. First, tracking is a difficult task, especially in real time and in urban scenarios where often a large number of objects are present. Since the anomaly detection is directly related to the quality of tracking, a tracking error will inevitably bias the detection step. Second, since each individual or object monitored is related to a single path, it is hard to deal with people occluding each other.

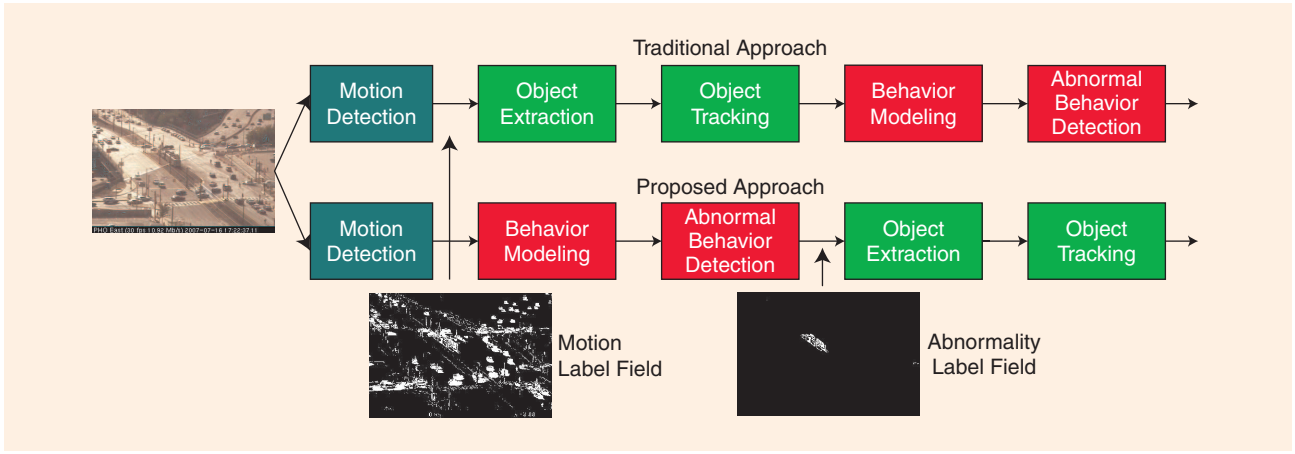
ANALYSIS-BY-SYNTHESIS METHODS

One alternative approach to track-based approaches is that proposed by Boiman and Irani [22]. This approach is based on spatiotemporal intensity correlation among different snippets of video. An observed sequence is built from spatiotemporal segments extracted from a training sequence. In this analysis-by-synthesis method, only regions that can be built from large contiguous chunks of the training data are considered normal. Malinici et al. [23] propose a related method for grasping behavior without having to extract motion features. This is done by extracting features from the raw video with invariant subspace analysis. An infinite hidden Markov model is then used to model the sequential characteristics of typical video. Methods by Yilmaz et al. [24] and Gorelick et al. [25] also use spatiotemporal chunks of video to characterize behavior. In this case, the silhouette of a moving object is concatenated into a three-dimensional (3-D) volume whose shape is used to recognize specific types of behavior. Unfortunately, these methods seem only effective when very few moving objects are seen simultaneously. Also, in many of these approaches, either a video clip or a large part of the field of view is treated as an integral entity. Consequently, the whole video clip or a large part of the field of view is labeled as either nominal or anomalous.

CONTEXTUAL AND BEHAVIORAL METHODS

These approaches model the contexts and behavioral attributes of an object rather than the object itself. Context in a video generally means the location and time of an object passing through the field of view. Behavioral attribute refers to the noncontextual attributes such as the size, speed, direction, and color of the object passing by a specific location.

These methods can work at a pixel level or more generally on larger blocks of pixels. Some methods use summary of



[FIG1] Traditional versus proposed approach to abnormal behavior detection. While in the traditional approach, object extraction has to deal with clutter, in the proposed approach clutter is removed due to the focus on abnormalities.

motion vectors [26]–[30] or motion labels [2], [31]–[34] for each location to describe activity in the scene. Consequently, an image-like 2-D structure provides a summary of activities over a large time window, thus easing memory and processing requirements. At the core of many of these approaches is a departure from traditional object-based tracking perspective in favor of a new perspective that relies on location-based statistical modeling.

Our perspective on many of these approaches is illustrated in Figure 1. As described earlier, the main problem with an object-track-based approach is that, in case of complex environments, such as those depicted in Figure 1, object extraction and tracking are performed directly on cluttered raw video or motion labels. In contrast, in the more recent approaches, it is possible to envisage a paradigm where anomaly detection is performed first, followed by object extraction and tracking, as shown in the lower branch of the block diagram in Figure 1. If the anomalous activity is reliably identified (tram in Figure 1), then object extraction and tracking focus on a ROI, and thus are relatively straightforward, both in terms of difficulty and computational complexity, on account of sparsity and absence of clutter.

In this line of research Xiang et al. [32] (and later Li et al. [34]) represent each moving blob by a seven-dimensional (7-D) vector that accounts for the blob's position, shape features, and motion vectors. An expectation maximization (EM)-based algorithm is used to cluster these 7-D points from training data to form the so-called behavior profiles. These profiles serve as a model for background activity. In the testing phase, the moving blobs in each frame are assigned to one or more of the behavior profiles. Those that are left unassigned (or have a low likelihood of assignment) can be thought of as anomalies. Wang et al. [29] present a related approach for background activity modeling. In their so-called bag-of-words approach, they divide the video into short clips, which they call "documents." Then, for each document, they extract visual features (usually motion vectors) for each moving blob. These features are then quantized into "visual words," which they cluster into the so-called topics. In this way, documents with similar topics are grouped together. Again, anomaly detection can be performed by looking at distances from clusters. The main difference between the two approaches appears to be the features chosen and the adaptive methods used for choosing the number of clusters. Simon et al. [35] also present modeling of background activity, which is aimed at recognizing high-level visual events. To do so, they implement a three-stage procedure in which they 1) gather spatiotemporal motion features for each moving object, 2) cluster the moving objects, and 3) recognize events with an ensemble of randomized decision trees made of those clusters.

While these attempts at modeling background activity appear promising for anomaly detection, they suffer from the following deficiencies:

- *Unstructured Activity:* In many situations, the background activity is made up of unstructured events such as random motion of water surface, trees shaking in the background, water fountains, or activities on a plaza observed from sufficiently far away. In these cases, motion is usually random, both spatially and in time, and, possibly, uncorrelated. In these situations, it is unclear whether background models based on clustering motion vectors, shape and position descriptors can provide meaningful characterization of the background activity. For instance, motion estimated on shimmering water surface may have quite random magnitudes and orientations, and it is unclear what would be the impact of clustering in this scenario. Both the bag-of-words approach of [29] and the behavior profiling approach of [32] have only been tested on a limited set of highly structured scenarios, and further work is necessary to understand performance of these approaches in unstructured situations.
- *Collective Anomalies*: The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection may be anomalous. This may happen, for instance, when an abandoned object is placed at a specific location. At any instant the presence of an object at a specific location is not anomalous but rather the existence of an object over a longer period of time is what makes the object anomalous.

Depending on how motion vectors are clustered, it can result in poor detection in such instances.

- *Clutter*: Spatially localized anomalies can suffer from contamination caused by several other nominal activities occurring in the vicinity. This phenomenon can be attributed to poor signal-to-noise ratio (SNR). In such cases, the anomalous signal is small relative to the noise associated with surrounding activity. Rather than form global spatial clusters and test those for anomalies, this scenario requires focusing attention on appropriate spatial scale to ensure that the anomalous signal is detected.
- *Feature Statistics*: Frames in a video are highly correlated both spatially and temporally. Therefore, the collection of features such as motion vectors, shape parameters, and positions are all highly correlated in time. In general, it is difficult to estimate feature statistics, such as its pdf, from correlated samples.
- *Multicamera Fusion*: Most features, including motion vectors and shape parameters, strongly depend on geometry. Therefore, fusing multiple camera views to characterize background activity based on these features appears to be difficult. Consequently, other methods are required to establish correspondence between the different cameras. This can be an issue, particularly in low-power, low-bandwidth environments.

VIDEO ANOMALIES CAN BE THOUGHT OF AS THE PRESENCE/ABSENCE OF USUAL (OR UNUSUAL) OBJECTS OR MOTION ATTRIBUTES IN UNUSUAL (OR USUAL) LOCATIONS OR TIMES.

In the remainder of the article, we build upon the existing work on contextual and behavioral paradigm for video analysis in surveillance scenarios. We develop feature descriptors that have a concrete statistical interpretation. These feature descriptors amount to an event-ensemble framework for dynamic scene characterization. Our event-based framework enjoys certain geometric invariance properties that lead to new concepts for characterizing spatial correspondences between different locations in single as well as multiple cameras. This is particularly valuable in heterogeneous camera networks, wherein the cameras have different locations (wide-baseline scenarios), orientations with respect to the scene, and zoom levels and in which the inherent topology of the network is dynamic.

NOMINAL ACTIVITY MAPS: A STATISTICAL APPROACH

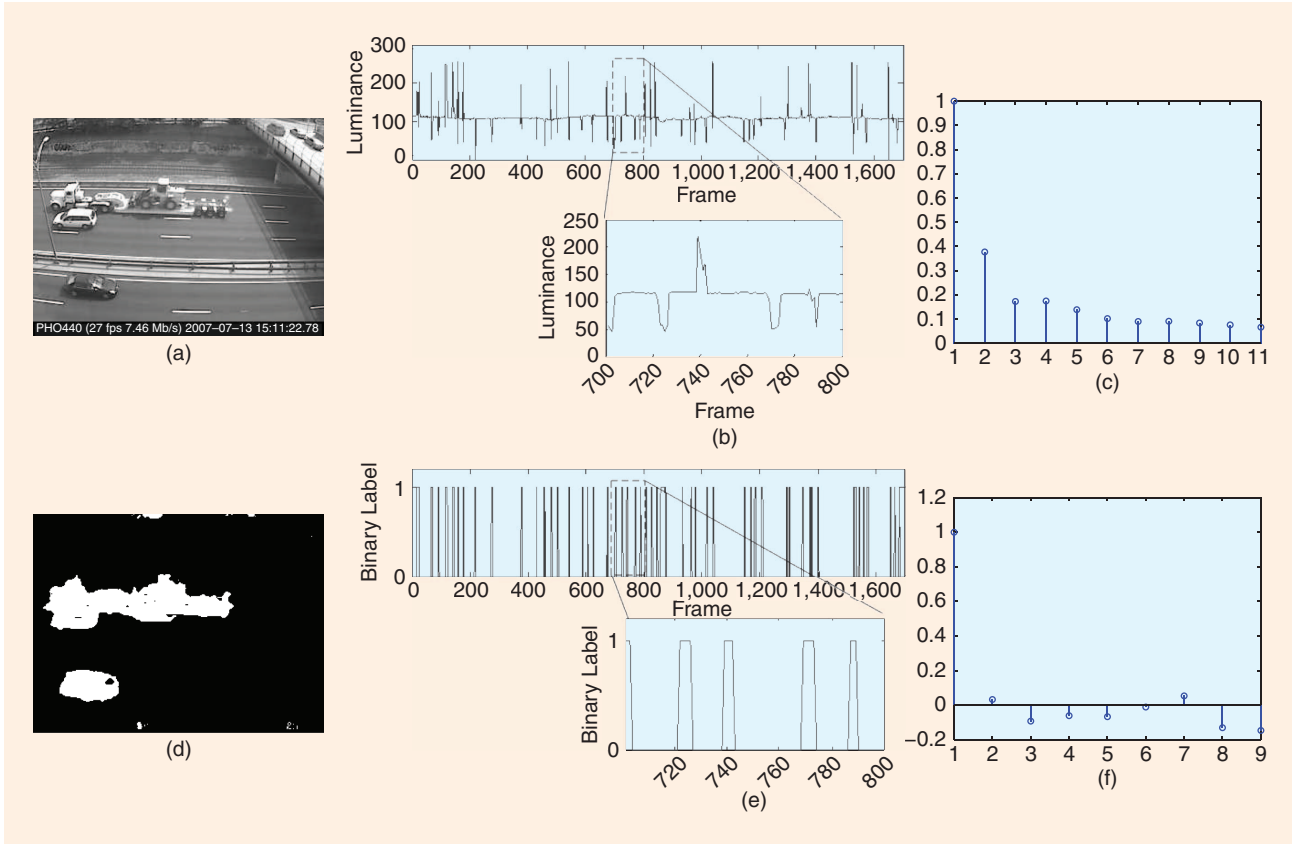
Raw video can be thought of as a sequence of frames $I_t, t = 1, 2, \ldots$. This representation, although common and useful for many tasks, is redundant from the standpoint of anomaly detection. First, for a static camera, the case considered here, there is a significant correlation between frames; little innovation is captured by a new frame. Second, luminance and color in a single frame, in principle, contain no information on scene dynamics; two or more frames need to be considered jointly to infer dynamics in the scene. Third, although frame-by-frame representation is inherently dense, activity in a typical surveillance scenario is inherently sparse, both spatially and temporally. Motivated by these issues, we develop a new event-based representation, where we characterize activity at each pixel captured by the camera by means of a motion label with two possible states: "moving" or "static." Furthermore, to allow a richer description, we augment this dynamic representation with additional features, such as size, shape, speed and direction of movement, or even photometric properties.

More formally, let $L_t(\vec{x})$ denote a motion label at location $\vec{x} = (x, y)$ and time t. Clearly, $L_t, t = 1, 2, \ldots$ denotes a sequence of motion label fields associated with the sequence of frames $I_t, t = 1, 2, \ldots$. There exist many approaches to computing labels L_t. Robust background subtraction methods [20], [36], [37] are preferable, although computationally demanding. Figure 2(d) shows a typical motion label field with black denoting a static ($L_t(\vec{x}) = 0$) and white denoting a moving ($L_t(\vec{x}) = 1$) state. It is essential to realize that there exists an alternative interpretation of motion labels: a sequence of consecutive 1s can be considered a busy period and a sequence of 0s—an idle period. In this case, at each location $\vec{x}$ the motion label sequence $L_t(\vec{x}), t = 1, 2, \ldots$ can be considered as subsequent busy and idle periods (Figure 3). As discussed earlier, luminance exhibits strong correlation temporally. On the other hand, the idle period lengths are largely independent across time, as are the busy period lengths. Consider, for example, a traffic stream on a busy road; one vehicle and its attributes (size, shape, and color) generally are not predictive either of the presence of preceding/subsequent vehicles or of their corresponding attributes. We confirm this experimentally by computing autocorrelation at a specific pixel of traffic pattern shown in Figure 2(a); the idle period length is practically an uncorrelated random variable [Figure 2(f)], while the luminance is not [Figure 2(c)]. The decomposition of video into busy and idle periods leads to an interesting model that we discuss next.

MODELING PIXEL ACTIVITY USING MARKOV CHAINS

This busy-idle, or event-based, video representation places activity analysis firmly within the conventional statistical learning theory, which relies on independent samples. Furthermore, as suggested previously, each event, defined as one busy period followed by one idle period, can be augmented with various features to further enrich the description. This alternative representation has two immediate benefits: computational efficiency, due to the binary nature of labels, and mutual independence not only among busy and idle periods, but also independence among features belonging to different busy or idle periods. Consequently, by the nature of this construction, events and their attributes can be viewed as independent samples in a suitable space, and advanced learning techniques can be readily applied.

These ideas and the empirical evidence from Figure 2 suggest a two-state Markov chain model with busy and idle states shown in Figure 3. Additionally, feature descriptors, such as



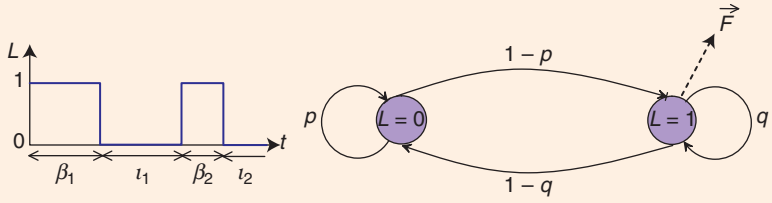
[FIG2] Comparison of frame- and event-based video representation: (a) original image, (b) luminance evolution, (c) luminance autocorrelation function at a pixel, (d) motion label field obtained using method from [36], (e) label evolution, and (f) autocorrelation function of idle period length (label = 0) at the same pixel. Note the Dirac-like shape of the bottom autocorrelation, suggesting independence of idle period lengths.

size, shape, color, texture, and velocity, can be modeled at each location $\vec{x}$ and time t as a random vector $\vec{F}_t(\vec{x})$ conditioned on the underlying state. For example, if only size were to be used as a feature, then the random vector $\vec{F}$ would become a scalar, whereas should, additionally, speed and direction be used, then $\vec{F}$ would be a 3-D vector. These descriptors are not different from those employed for describing contextual and behavioral attributes (see, for instance, [29], [32], and [35]). However, our feature descriptors have a statistical interpretation. At each location $\vec{x}$ our model implicitly assumes not only independence among the busy and idle periods but also a conditional independence of the feature descriptors when conditioned on the state (busy/idle), e.g., the color of a car is independent of the color of subsequent cars. Different types of models for feature descriptors can be incorporated. For instance, a Gibbs-Markov model with suitable potential function can be applied to provide a measure of temporal variability of descriptor $\vec{F}$ within a busy period.

Based on the Markov chain model from Figure 3, we can now describe the pdf of features ℓ, namely $g_0(\ell)$ from (1), for a sequence of frames I_t observed at pixel $\vec{x}$ over a w-frame time window $\mathcal{W} = [t - w + 1, t]$. So far, the set of features ℓ was not precisely defined. Now, let ℓ combine all motion labels L_t and feature descriptors $\vec{F}_t$ within time window $\mathcal{W}$ as follows: $\{\ell_i\}_{\mathcal{W}} = \{L_{t-w+1}(\vec{x}), \ldots, L_t(\vec{x}), \vec{F}_{t-w+1}(\vec{x}), \ldots, \vec{F}_t(\vec{x})\}$. Note that, to simplify notation, we omit location $\vec{x}$ from $\{\ell_i\}_{\mathcal{W}}$, assuming it implicitly. It can be shown that the negative log-likelihood of the pdf has the following linear form:

$$\Lambda_{\vec{x}}(\{\ell_i\}_{\mathcal{W}}) = -\log(g_0(\{\ell_i\}_W)) = \left(\sum_{k=t-\mathcal{W}+1}^{t} L_k(\vec{x})[A_1 + A_2\mathcal{V}_k(\vec{x})]\right) + A_3\mathcal{K}_t(\vec{x}), \quad (3)$$

where A_j are constants, $\mathcal{V}_k(\vec{x})$ is a potential function from the Gibbs-Markov model of the feature descriptor $\vec{F}$, and $\mathcal{K}_t(\vec{x})$ is proportional to the total number of busy-idle transitions in time window $\mathcal{W}$ for pixel at $\vec{x}$. The expression in (3) is not immediately obvious, but its derivation is outside of the scope of this overview article. For derivation details and discussion, see [6]. Intuitively, however, if $A_2 = 0$, i.e., no additional features are considered, then this statistic combines two quantities at location $\vec{x}$ within time window $\mathcal{W}$: the total busy time (the summation) and the average number of busy-idle transitions, which jointly characterize how often and for how long an object passes through pixel at $\vec{x}$. For $A_2 \neq 0$, the additional term expresses the total energy of the descriptor $\vec{F}$ and captures characteristics of an object passing through $\vec{x}$. For example, it can account for size and speed or both. The form of this statistic remains valid



[FIG3] Markov chain model for a dynamic event: p, q are state probabilities (static and moving, respectively), and $1-p$, $1-q$ are transition probabilities. β_1, ι_1, β_2, ι_2 denote consecutive busy and idle intervals. A feature descriptor $\vec{F}$, such as its size, speed/direction of motion, color, luminance, etc., is associated with each busy interval.

for various descriptors $\vec{F}$, including a temporal Gibbs-Markov model that can be used to measure the variability of the descriptor in time.

As a concrete example, we present a specific feature descriptor for situations when object size is of importance. Let $S_t(\vec{x})$ be a scalar value describing size at location $\vec{x}$ and time t as follows:

$$S_t(\vec{x}) = \frac{1}{|\mathcal{N}|} \sum_{\vec{y} \in \mathcal{N}(\vec{x}), \vec{y} \bowtie \vec{x}} \delta(L_t(\vec{x}), L_t(\vec{y})), \tag{4}$$

where $\mathcal{N}(\vec{x})$ is a square window centered at $\vec{x}$ and $\vec{y} \bowtie \vec{x}$ means that $\vec{y}$ and $\vec{x}$ are connected. Connectedness here is in the sense of graph connectivity. At each time t we can associate a graph with vertices corresponding to the pixels; an edge exists between any two vertices if they are spatial neighbors and also share the same motion labels. Two pixels $\vec{x}$ and $\vec{y}$ are connected if a path exists from $\vec{x}$ to $\vec{y}$. Also, $\delta(\cdot) = 1$ if and only if $L_t(\vec{x}) = L_t(\vec{y}) = 1$, i.e., if both $\vec{x}$ and $\vec{y}$ are deemed moving, otherwise $\delta(\cdot) = 0$. Although this descriptor is not a direct measure of the size of an object passing through pixel at $\vec{x}$, it is related to object size; objects smaller than the window $\mathcal{N}$ will produce smaller values of S_t, whereas larger objects will produce descriptors $S_t(\vec{x}) = 1$ for locations $\vec{x}$ in a close spatial proximity. If S_t is the only descriptor used in $\vec{F}_t$, then, since size is nonnegative, the Gibbs-Markov potential takes a particularly simple form, namely $\mathcal{V}_k(\vec{x}) = S_k(\vec{x})$. Should one be interested in measuring the variability of size across time, another potential function could be used: $\mathcal{V}_k(\vec{x}) = |S_k(\vec{x}) - S_{k-1}(\vec{x})|$.

Note that (3) takes a specific linear form, namely, a weighted linear combination of motion labels and feature descriptors. Alternatively, it may be advantageous in certain cases to retain these features as a vector. Indeed, a closer examination of (3) reveals that a feature vector may be composed of the average activity level, average number of transitions, and average feature descriptor.

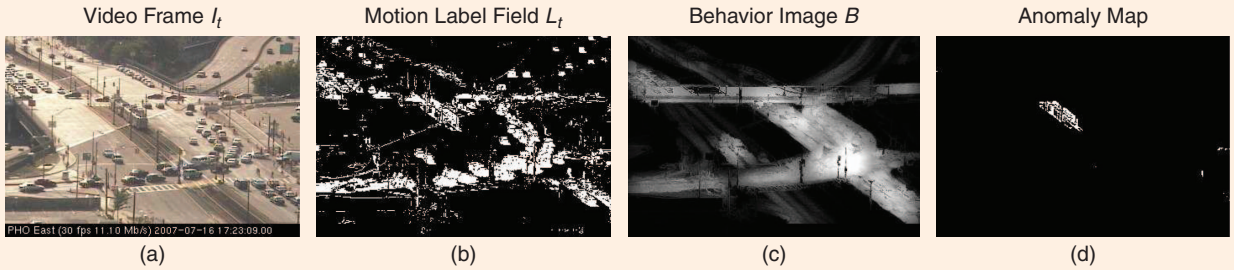
LOCALIZED NOMINAL MAPS

In many applications, the task is to localize video anomalies both spatially and temporally. This is often accomplished by means of localized nominal maps, i.e., spatially and temporally localized maps that capture nominal activity. These types of maps can be used for detecting spatiotemporally localized anomalies. Several approaches are possible based on features $\{\ell_i\}_{\mathcal{W}}$ defined at each location $\vec{x}$, and this will be described in the section "Anomaly Detection Results." Here, we graphically illustrate localized nominal maps to underline the fact that features are limited to a specific spatial location.

To illustrate the implications of this localization, consider (3) and, in particular, consider the maximum of $\Lambda_{\vec{x}}(\{\ell_i\}_{\mathcal{W}})$ as a localized nominal map. This quantity is spatially variant and can be captured by a 2-D array

$$B_{\max}(\vec{x}) = \max_{t \in [1, M]} \Lambda_{\vec{x}}(\{\ell_i\}_{\mathcal{W}}) \tag{5}$$

over M frames of the training data, where $M >> w$. Recall from the section "Modeling Pixel Activity Using Markov Chains," that the window $\mathcal{W}$ implicitly depends on the time t. We call $B_{\max}$ the background behavior image [2], [6], as it captures the background activity, in this case, peak activity, in the training data in a low-dimensional representation (one scalar per location $\vec{x}$). As shown in Figure 4(c), behavior image succinctly synthesizes the ongoing activity in a training sequence. It



[FIG4] (a)–(d) Behavior subtraction results for the maximum-activity nominal map (see text) on data captured by a stationary, although vibrating, camera. This is a highly cluttered intersection of two streets and interstate highway. Although the jitter induces false positives during background subtraction (L_t), only the tramway is detected by behavior subtraction; the rest of the scene is considered normal.

implicitly includes the paths followed by moving objects as well as the amount of activity registered at every point in the training sequence. Also, note that the behavior image is robust to moderate false alarms and misses arising from errors in motion label estimation since the nominal map is derived from activity over the temporal window $\mathcal{W}$. Clearly, perfect background subtraction to compute motion labels is not essential. In the section "Anomaly Detection Results," we will describe how such behavior images can be used to detect localized anomalies in a wide range of scenarios. We emphasize that the maximum activity is an instantiation for the purpose of illustration, and, in general, one could instead represent a multivariate feature vector by its p-value, which is statistically more meaningful.

WHILE PHOTOMETRIC SCENE PROPERTIES ARE HIGHLY CORRELATED IN TIME, DYNAMICS OCCURRING IN THE SCENE REMAIN LARGELY UNCORRELATED TEMPORALLY THUS FACILITATING THE USE OF STATISTICAL LEARNING METHODS.

ROBUSTNESS TO ERRORS IN BACKGROUND SUBTRACTION

The feature vectors as well as the scalar statistic in (3) are based on binary random variables $L_t(\vec{x})$ whose realizations are computed using background subtraction. Since the computed labels will be necessarily noisy, i.e., will include false positives and misses, a positive bias will be introduced into the event model [even if the noise process can be considered consisting of independent, identically distributed random variables (IID), i.e., its mean is positive since labels are either 0 or 1]. The simplest method of noise suppression is by means of low-pass filtering. Thus, in scenarios with severe event noise (e.g., unstable camera, unreliable background subtraction) its impact can be mitigated by using a simple averaging filter to compute the background behavior image [38]

$$B_{\text{avg}}(\vec{x}) = \frac{1}{M}\sum_{t=1}^{M} \Lambda_{\vec{x}}(\{\ell_i\}_{\mathcal{W}}). \tag{6}$$

Note that this background behavior image estimate provides a space-variant bias derived from the training data. This bias arises from errors in background subtraction but is temporally stationary. Consequently, it captures the stationary aspects of the scene and can be reliably estimated.

SPATIAL CO-OCCURRENCES

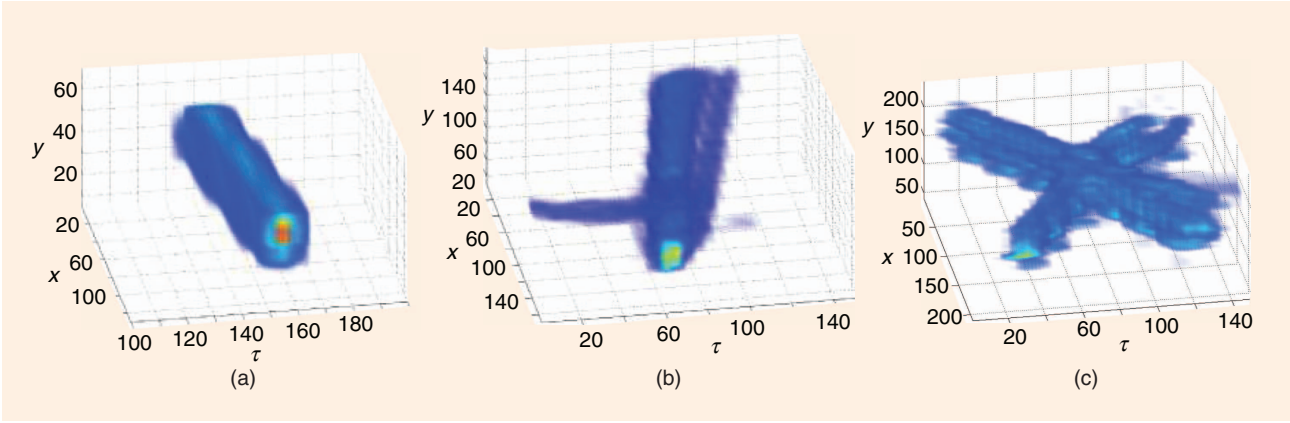
The patterns of activity described in the "Modeling Pixel Activity Using Markov Chains" section primarily dealt with spatially localized time series, whereas in reality, activities are spatially correlated. For instance, vehicles traveling on a highway are characterized by spatial paths they traverse. In general, a pure pixel-by-pixel approach is insufficient in applications where anomaly is manifested spatially, e.g., cars running against traffic flow, and cars making illegal U-turns. Consequently, a strategy is needed for incorporating spatial patterns in addition to the temporal patterns of motion label sequences. The shortcomings of characterizing purely temporal behavior become clearer when we consider two pixels with identical signatures arising from opposite directions of motion (left to right versus right to left). Obviously, normal/abnormal behavior arising from the pattern of activity between the two pixels cannot be captured through a purely pixel-by-pixel analysis. One possible direction in solving this issue is to cluster the features for each location and follow along the lines described in [32] and [29]. Here we present an alternative approach based on Markov random fields (MRFs). To this end we develop co-occurrence models as a function of location [4].

Consider two neighboring pixels $\vec{x}$ and $\vec{y}$ with associated motion label time series $L_t(\vec{x})$, $L_t(\vec{y})$. Suppose the two pixels are sufficiently close that they lie on a path traversed by a moving object. While the correlation between $L_t(\vec{x})$ and $L_t(\vec{y})$ is small, the correlation between $L_t(\vec{x})$ and $L_{t+\tau}(\vec{y})$, i.e., time-shifted series at $\vec{y}$, should be significant for some time delay τ (the same object induces changes at $\vec{x}$ and $\vec{y}$ except for a time shift). This implies that one should be able to determine spatial relationships by matching temporal signatures, i.e., through activity matching.

In the simplest case, two spatiotemporal sites $(\vec{x}, t)$ and $(\vec{y}, t+\tau)$ are considered as co-occurring if they are both occupied, namely, $L_t(\vec{x}) = L_{t+\tau}(\vec{y}) = 1$. Clearly, other features, such as similarity between feature vectors measured at $(\vec{x}, t)$ and $(\vec{y}, t+\tau)$, could also be incorporated into this definition. Consequently, in this generalized perspective $(\vec{x}, t)$ and $(\vec{y}, t+\tau)$ co-occur not only due to the position and orientation of the camera in the scene but also due to the shape, velocity, and direction of the moving objects. Nevertheless, to keep the development simple, we describe the simplest case of co-occurrence here. Interestingly, several moving objects exhibiting regular behavior (think of cars on a highway going in the same direction) leave, after a while, similar traces in the spatial neighborhood of $\vec{x}$. Our method encapsulates such traces in terms of a co-occurrence frequency matrix, which captures the frequency with which two spatiotemporal neighbors co-occur. The co-occurrence matrix for locations $\vec{x}$ and $\vec{y}$ is defined as follows [4]:

$$\alpha_{\vec{x}\vec{y}}(\tau) = \frac{1}{M}\sum_{t=t_0}^{t_0+M-1} \delta(L_t(\vec{x}), L_{t+\tau}(\vec{y})), \tag{7}$$

where t_0 is a time instant of interest, δ is the same indicator function as in (4), and the range of time delays is limited by $\tau_{\max}$, i.e., $-\tau_{\max} \le \tau \le \tau_{\max}$. From (7), $\alpha_{\vec{x}\vec{y}}(\tau)$ can be understood as a 3-D matrix storing the number of times a pixel $\vec{x}$ at time t co-occurred with a pixel $\vec{y}$ at time $t+\tau$ within an M-frame time window. This matrix can serve as the basis for characterizing nominal activity as follows: Let $\mathcal{M}_{\vec{x},t}$ be a small spatiotemporal window centered at $(\vec{x}, t)$, and let $\{\ell_i\}_{\mathcal{M}_{\vec{x},t}}$ be the set of features restricted to $\mathcal{M}_{\vec{x},t}$. Our probabilistic model for co-occurrences is an MRF parametrized through $\alpha_{\vec{x}\vec{y}}$ (7) as follows:



[FIG5] **Co-occurrence matrix $\alpha_{\vec{x}\vec{y}}(\tau)$ for a specific location $\vec{x} = \vec{x}_0$ depicted as a function of $\vec{y} = (x, y)$ and τ with color representing the degree of co-occurrence (red = high, blue = low) for: (a) regular traffic flow, (b) regular traffic flow and pedestrians crossing the street, and (c) pedestrians walking from left to right and from right to left.**

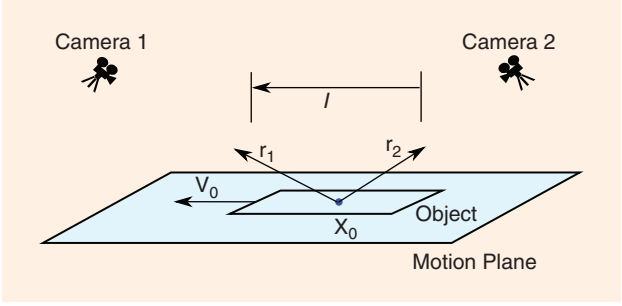
$$g_0(\{\ell_i\}_{\mathcal{M}_{\vec{x},t}}) = \frac{1}{Z}\exp\left(\sum_{(\vec{y},t+\tau)\in\mathcal{M}_{\vec{x},t}} \alpha_{\vec{x}\vec{y}}(\tau)\delta(L_t(\vec{x}), L_{t+\tau}(\vec{y}))\right), \quad (8)$$

where Z is a partition function. We assume that the temporal size of window $\mathcal{M}_{\vec{x},t}$ is less than w, so that the summation is well defined. Clearly, $g_0(\cdot)$ describes a joint probability distribution on the observed features in the window $\mathcal{M}_{\vec{x},t}$ centered at location $\vec{x}$ and time t. Note that this probabilistic model accounts for co-occurrences, and we use it here separately from the Markov chain model (3). It would be interesting to explore how to combine these two models into one representation.

Typical co-occurrence matrices that explicitly account for spatial paths around a specific pixel at $\vec{x}$ are shown in Figure 5. For example, the plot in Figure 5(a) shows co-occurrences for an object that moves linearly from left to right. Other plots show more complex scenarios. Note that the co-occurrence matrix can be updated in time to account for changes in the behavior. This can be done by simply adding, in a linear fashion, new traces as they appear in a streaming video.

GEOMETRIC INVARIANCE AND MULTICAMERA NOMINAL MAPS

We also describe the basis for deriving statistical models of multicamera nominal activity. Our description will be based on the concept of geometric invariance (see [5]). As described in the "Introduction," in heterogeneous camera networks, uncalibrated cameras are common. Nevertheless, one would like to benefit from combining observations of the same activity captured by multiple cameras. The main issue is how to ensure that activity seen at a specific location(s) in one camera is the same activity seen in other cameras. Clearly, the geometric properties change based on orientation and zoom levels. While a significant effort has been devoted in the literature to deriving correspondences across cameras based on feature detection methods [39] such as scale invariant feature transform (SIFT) [40], we explore a different idea based on activity correspondences.



[FIG6] **Illustration of geometry independence. The projection of l and v_0 scale with the same factor for each camera. This cancels the impact of camera viewing angle.**

To build intuition, we consider a hypothetical situation of two cameras with infinite resolution pointed at a flat object moving on a plane. Suppose the object moves through a physical point $\mathbf{x}_0$ in 3-D space and is observed by two cameras as depicted in Figure 6. Let $\mathbf{x}_0$ project onto locations $\vec{x}_1$ and $\vec{x}_2$ in the imaging planes of Camera 1 and Camera 2, respectively. The property of geometric invariance asserts that irrespective of how the cameras are placed (as long as they lie outside the plane of motion), the motion labels at $\vec{x}_1$ and $\vec{x}_2$ are busy for the same duration and at the same physical times. To understand why this is true, note that the busy interval at Camera 1 (Camera 2) is equal to the ratio of the effective length of the object over its velocity, both projected onto Camera 1 (Camera 2). While the individual projected lengths and projected velocities are different, their ratios are identical. This is because the projection operation is identical for both the object length and velocity. These ideas can also be extended to 3-D objects, cameras with finite resolution, and, essentially, arbitrary orientations as described in [5].

The geometry independence principle immediately lends itself to a correspondence algorithm. For each pixel $\vec{x}_1$ in Camera 1 with an associated motion label sequence, one locates the corresponding pixel $\vec{x}_2$ in Camera 2

whose motion label sequence is the closest in the normalized Hamming distance sense. This normalized Hamming distance metric has a statistical basis, namely, two pixels at $\vec{x}_1$ and $\vec{x}_2$ are matched if the difference between the corresponding motion label sequences can be explained by errors due to background subtraction. On the other hand, the two pixels do not match if their corresponding motion label sequences are significantly uncorrelated. This hypothesis test leads to the following metric:

$$d(\vec{x}_1,\vec{x}_2) \doteq \frac{1}{\eta}\sum_{t=1}^{M}|L_t(\vec{x}_1) - L_t(\vec{x}_2)| \underset{\text{match}}{\overset{\text{no match}}{\gtrless}} \theta, \qquad (9)$$

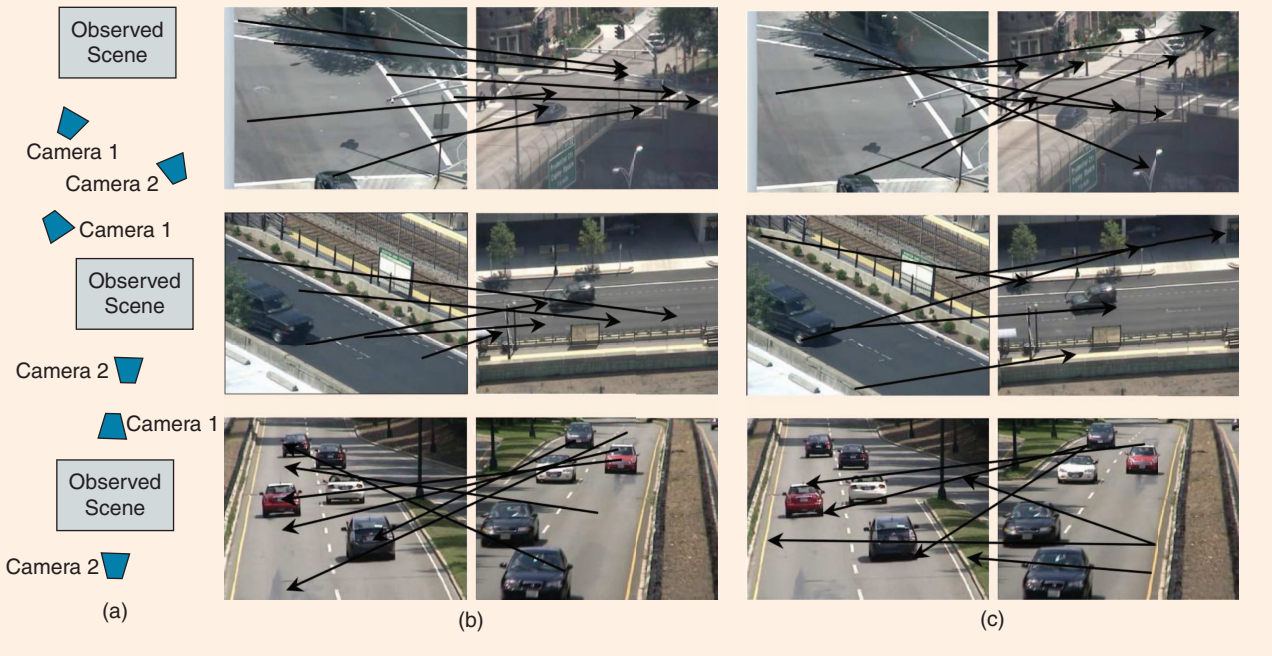
where θ is a suitable threshold and $\eta = \max(\|L_t(\vec{x}_1)\|_1, \|L_t(\vec{x}_2)\|_1)$ corresponds to the maximum activity. Note that the significance of the metric is that if pixels have low-level activity, the Hamming distance has to be relatively small for a potential match. Experiments conducted on a wide range of scenarios, as shown in Figure 7, provide an empirical evidence for the efficacy of the method. A similar strategy has been used by Hengel et al. [41] to recover the vision graph of a camera network, i.e., determine which cameras have overlapping fields of views. However, since that method has been geared toward sparse intercamera correspondences, it isn't clear how well it would perform for pixel-to-pixel correspondences. The principle underlying the correspondence problem can be utilized further to develop nominal maps of activity for multiple cameras (see [5] for details). Furthermore, note that the geometric independence extends to co-occurrence maps as well, namely, that the co-occurrence potentials are identical when viewed from different cameras. Consequently, one can describe a general MRF model that integrates the activities from multiple cameras.

IN MANY APPLICATIONS, THE TASK IS TO LOCALIZE VIDEO ANOMALIES BOTH SPATIALLY AND TEMPORALLY, AND THIS CAN BE OFTEN ACCOMPLISHED BY USING LOCAL FEATURES.

ANOMALY DETECTION RESULTS

To summarize our earlier discussion, we are given w video frames, $I_{t-w+1}, I_{t-w+2}, \ldots, I_t$ and a specific location $\vec{x}$, and our task is to determine whether this sequence is consistent with nominal activity or, alternatively, it is anomalous. We also have training data that describes the nominal activity. In this context, our feature vectors described in the section "Nominal Activity Maps: A Statistical Approach" provide a representation for the observed video frames.

In the following sections, we consider three realistic scenarios: 1) spatially localized anomalies, such as abandoned objects, sudden suspicious activity at a location, and unanticipated activity at unanticipated times, 2) structured anomalies in unstructured environments, and 3) spatially distributed anomalies, such as illegal U-turns.



[FIG7] Intercamera matching based on temporal signatures: (a) camera setup; (b) matching results from 90 s of video using temporal signatures as proposed here; (c) matching results using spatial signatures from SIFT [40]. Note that the cameras have different zoom levels. The proposed approach is able to match the corresponding pixels with a small error, whereas the method based on SIFT fails to find correct matches. Although our method provides a dense mapping, we show only a few correspondences for legibility.

SPATIALLY LOCALIZED ANOMALIES

EVENT-BASED FRAMEWORK HAS LOW MEMORY REQUIREMENTS AND IS OFTEN STRAIGHTFORWARD ENOUGH TO BE CONSIDERED FOR EMBEDDED PLATFORMS AT THE CAMERA NETWORK EDGE.

The features described in the section "Modeling Pixel Activity Using Markov Chains" will now be used for spatially localized anomaly detection. There are a number of different statistical learning algorithms for anomaly detection. We point out a few popular ones, which implicitly assume that the anomalous distribution is uniform over the feature vectors. One approach is to empirically estimate the nominal pdf of the feature vector. Given this nominal pdf, a threshold rule, such as the one shown in (2), can be applied where the threshold is chosen to ensure a desired false-alarm level. Another approach is based on K-nearest neighbors. Here, a test vector is declared anomalous if the distance between the test feature and its Kth nearest neighbor is above a certain threshold level. A third approach is based on one-class support vector machine (SVM) for each location [42]. The main point to note is that, for spatially localized anomalies, the tests are conducted at each location. Each location has different underlying statistics, which are either learned or incorporated into the anomaly detection algorithm. Consequently, one can obtain a uniform false-alarm rate across all the locations. This idea is reminiscent of the well-known constant false-alarm rate detector [8], where the detector is adapted to the location to compensate for the spatial variation of the noise statistics.

The performance is clearly dependent on the algorithm employed. Nevertheless, for the purpose of illustration, we present results in this section based on a simple algorithm. Specifically, we consider localized nominal maps: $B_{\max}$ (5) or B_{avg} (6), that are derived during training from the statistic (3) while ignoring contributions arising from the number of transitions ($A_3 = 0$) and assuming $A_1 = A_2 = 1$. We mainly use the maximum-activity nominal map $B_{\max}$ (5) except for situations when background subtraction is unreliable, in which case we use the average-activity map B_{avg} (6). Once the nominal map is known, we test the observed feature $\{\ell_i\}_{\mathcal{W}}$ at location $\vec{x}$ and declare it as an anomaly if its statistic $\Lambda_{\vec{x}}(\{\ell_i\}_{\mathcal{W}})$ is larger, within some tolerance, than the nominal activity map at the same location, i.e., $\Lambda_{\vec{x}}(\{\ell_i\}_{\mathcal{W}}) - B_{\max}(\vec{x}) > \theta$, with θ being a tunable threshold. We call this approach behavior subtraction [2], [6], as it is analogous to background subtraction. However, unlike background subtraction, which operates on photometric quantities, behavior subtraction operates on dynamic features; photometric stationarity is replaced by dynamic stationarity thus treating nominal (regular) motion as a background activity that needs to be ignored.

We demonstrate the powerful characteristics of behavior subtraction for both the maximum- and average-activity nominal maps using the size descriptor S_t (4). Thus, we use $\{\ell_i\}_{\mathcal{W}} = \{L_{t-w+1}(\vec{x}), \ldots, L_t(\vec{x}), S_{t-w+1}(\vec{x}), \ldots, S_t(\vec{x})\}$ as our feature vector. We use training sequences of length $M \in [10^3, 5 \times 10^3]$, a temporal window $w \in [100, 200]$, and a threshold $\theta \in [0.5, 0.7]$. Note that a departure from these values does not change the results dramatically. As demonstrated in Figures 4 and 8, despite its simplicity both in formulation and computation, the size descriptor S_t seems to be a sufficiently discriminative characteristic for the detection of structured anomalies in unstructured environments. In fact, the method's robustness to inaccuracies in motion labels L_t is quite remarkable. Even if moving objects are not precisely detected, the resulting anomaly maps are surprisingly precise. This is especially striking in Figure 4, where a highly cluttered environment results in high density of motion labels while camera jitter corrupts many of those labels, and yet the anomaly map is quite coherent.

Figure 8 demonstrates the effectiveness of behavior subtraction for different nominal maps in different scenarios. When used with the maximum-activity nominal map (5), behavior subtraction removes unstructured, parasitic motion, such as that due to water animation (fountain, shimmering surface) or fluttering leaves [Figure 8(a)–(c)]; while motion label fields L_t include such unstructured detections, only excessive motion is accurately captured by the anomaly maps. The same algorithm is also capable of detecting abandoned objects [Figure 8(d)–(e)] and people lingering [Figure 8(f)]. Figure 8(g) demonstrates a remarkable robustness of behavior subtraction based on the average-activity nominal map (6) to camera jitter; although the detected motion labels L_t are extremely noisy, the detected anomaly is relatively precise. In fact, it can serve as a jitter-resilient background subtraction algorithm.

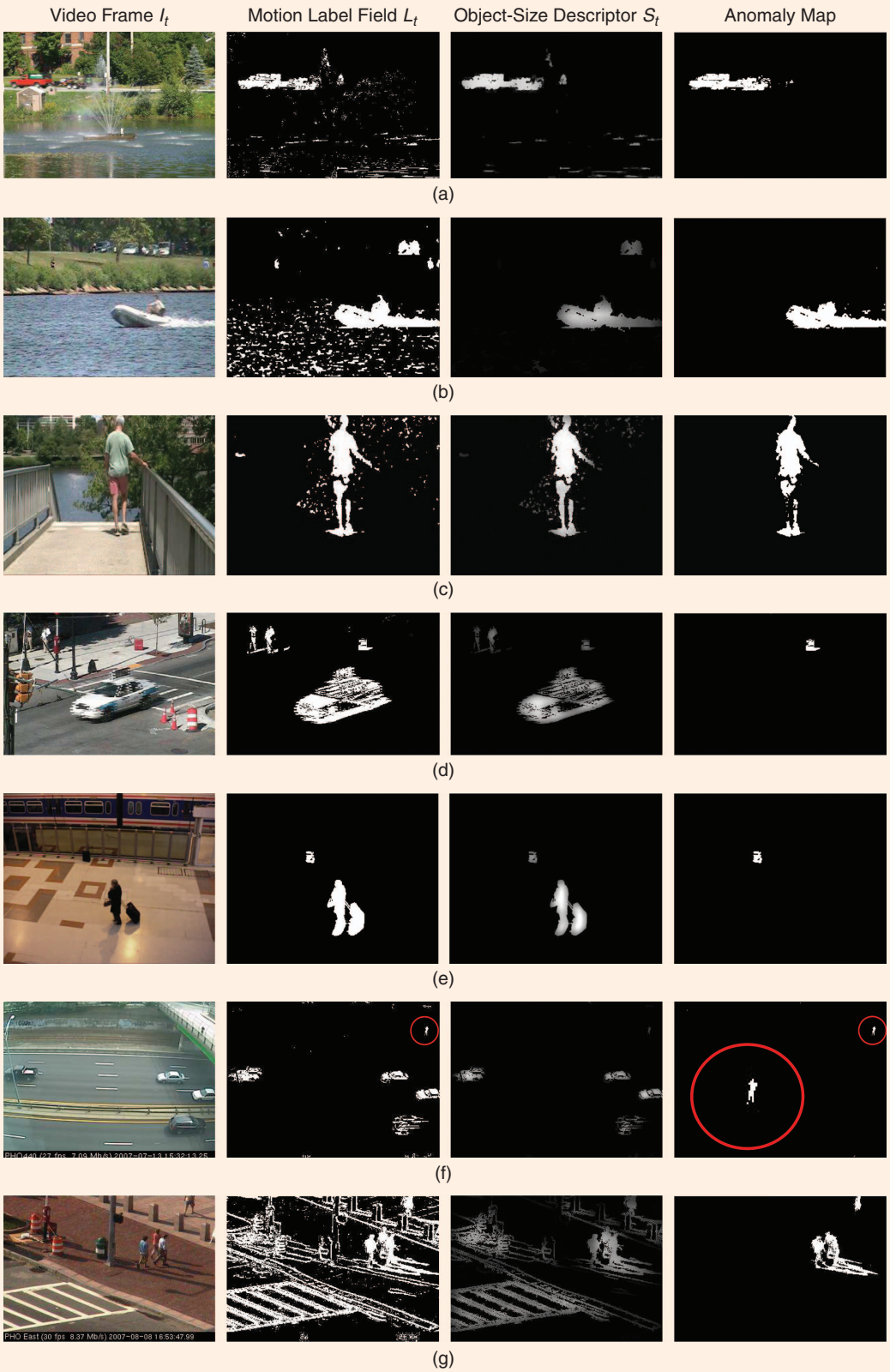
Figure 9 shows yet another interesting outcome of behavior subtraction. In this case, the maximum-activity nominal map (5) was trained on video with a single pedestrian. While the object-size descriptor (middle column) captures both individual pedestrians and groups thereof, anomalies are detected only when a large group of pedestrians passes in front of the camera. In this case, behavior subtraction serves as a "crowd detector."

To validate behavior subtraction quantitatively, we have compared it with a tracking-based method proposed by Chen et al. [12]. Table 1 shows a

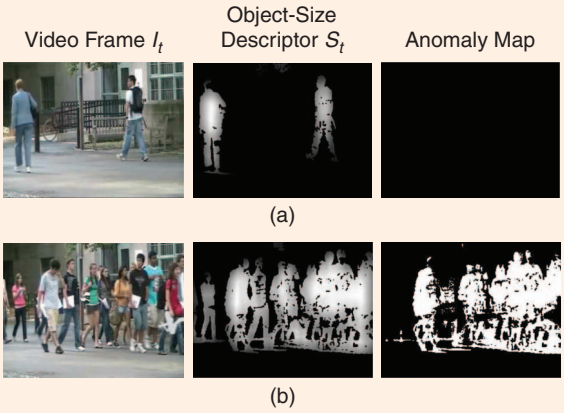
[TABLE 1] CONFUSION MATRICES FOR BEHAVIOR SUBTRACTION USING THE MAXIMUM-ACTIVITY NOMINAL MAP BASED ON SIZE DESCRIPTOR AND THE METHOD OF CHEN ET AL. [12] FOR VIDEO SEQUENCES FROM FIGURE 8(D)–(E).

BEHAVIOR SUBTRACTION		
	TRUE OCCURRENCE	
	NORMAL	ABNORMAL
NORMAL	93.2%	4.9%
ABNORMAL	6.8%	95.1%

METHOD OF CHEN ET AL. [12]		
	TRUE OCCURRENCE	
	NORMAL	ABNORMAL
NORMAL	73.9%	2%
ABNORMAL	26.1%	98%



[FIG8] Behavior subtraction results for the maximum-activity nominal map (5) in presence of: (a)–(b) shimmering water surface, (c) fluttering leaves, or containing (d)–(e) small abandoned object [(e) is from PETS data set], or (f) small erratic object in presence of steady traffic, as well as (g) for the average-activity nominal map (6) in presence of camera jitter.

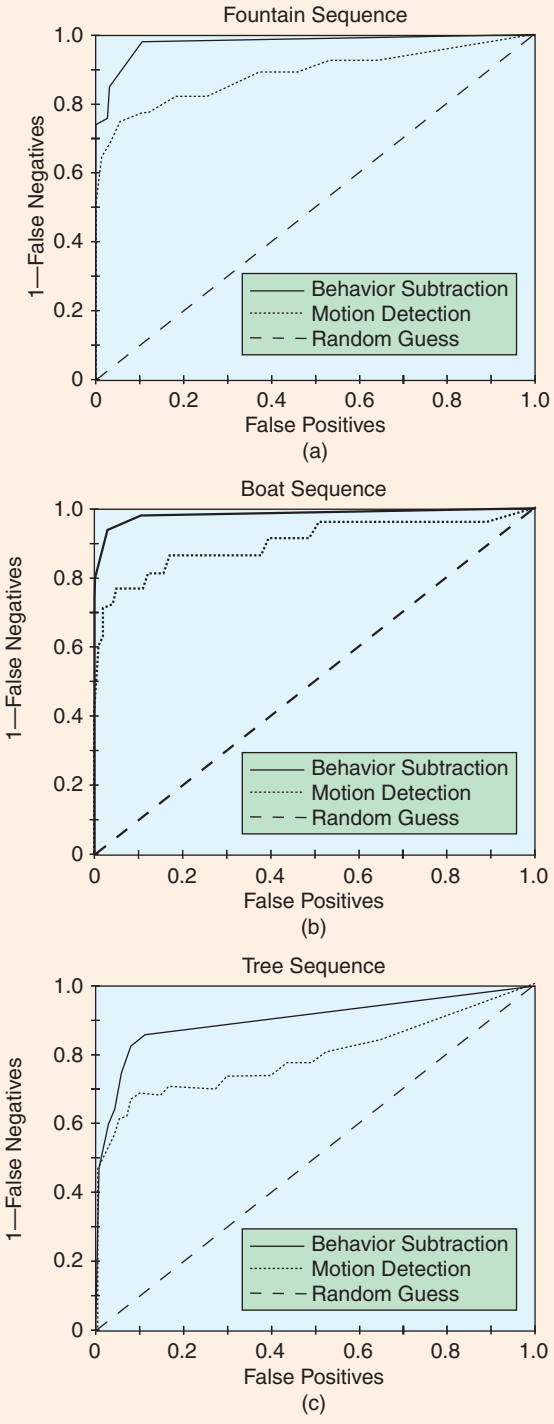


[FIG9] Results of behavior subtraction for the maximum-activity nominal map (5) with training performed on a video containing single pedestrian. (a) Two isolated pedestrians are considered to be normality. (b) A group of pedestrians (later in the same sequence) is associated with a large amount of activity and thus detected as anomaly.

quantitative comparison between the two methods tested in terms of confusion matrices. Clearly, behavior subtraction results in far fewer false positives than the method of Chen et al. while only slightly increasing the miss rate.

In another quantitative test, we compared behavior subtraction with robust background subtraction (a variant of approach from [37]). In this test, behavior subtraction serves the purpose of a robust motion detector. Consequently, anomalies correspond to any real moving object whose signature is different from background signatures. We used three sequences, each containing animated background: a fountain [Figure 8(a)], shimmering water surface [Figure 8(b)], and fluttering leaves [Figure 8(c)]. First, we manually outlined the moving objects in a number of frames. Then, we applied both methods to each of the videos, and, finally, we computed the number of correct detections (moving or stationary) as well as false positives and misses by comparing each result with the manually derived ground truth. By tuning the threshold θ in behavior subtraction $(\Lambda_{\vec{x}}(\{\ell_i\}_{\mathcal{W}}) - B(\vec{x}) > \theta)$ and the corresponding threshold in background subtraction, we produced receiver operator characteristics (ROCs) curves shown in Figure 10. Clearly, behavior subtraction outperforms robust background subtraction [37] by a wide margin.

One should note that behavior subtraction is efficient in terms of processing power and memory use, and thus can be implemented on modest-power processors (e.g., embedded architectures) for edge deployment in camera networks. It requires one floating-point number for each pixel of statistic (3), for example, $B_{\max\,(5)}$ or B_{avg} (6). This corresponds to the total of 11 bytes per pixel for $w = 24$. This is significantly less than, for example, 12 floating-point numbers per pixel needed when working in photometric space using a trivariate Gaussian model for color video data (three floating-point numbers for R, G, B means and nine numbers for the covariance matrix).



[FIG10] ROC curves obtained for three sequences exhibiting a large number of false positives: (a) cars moving in front of a fountain [Figure 8(a)], (b) boats moving in wavy water [Figure 8(b)], and (c) people walking in front of a tree shaken by the wind [Figure 8(c)].

SPATIALLY DISTRIBUTED ANOMALIES

So far, we presented results for anomaly detection at a single pixel. However, as we pointed out in the section "Spatial

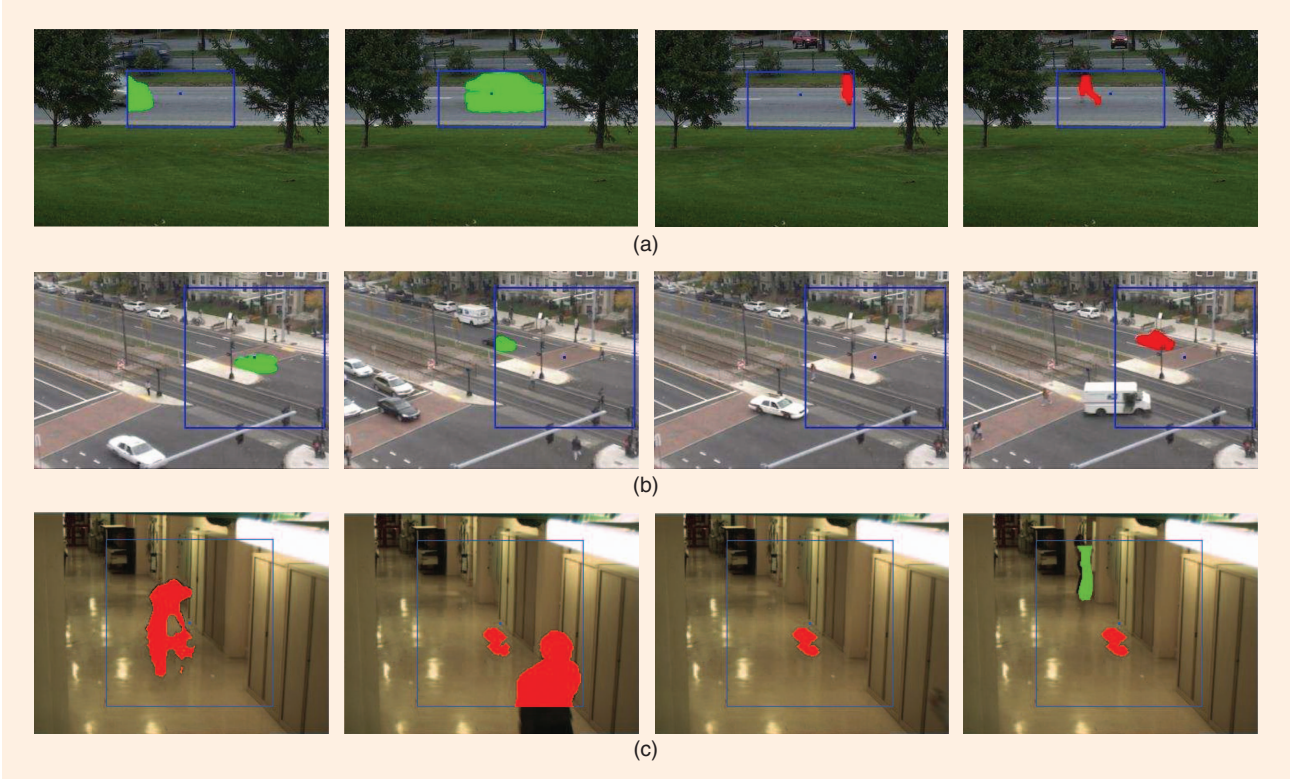
Co-Occurrences," to detect spatially distributed anomaly patterns, co-occurrence models are needed. The co-occurrence matrix $\alpha_{\vec{x}\vec{y}}(\tau)$ (7) contains a summary of every trace left by nominally moving objects in the training sequence (examples of such a trace are shown in Figure 5). The goal is to locate every spatiotemporal window $\mathcal{M}_{\vec{x},t}$ in the observed video in which the trace left by moving objects differs from the co-occurrence matrix obtained during training. We accomplish this by looking for outliers in the MRF model (8). For each sequence, a co-occurrence matrix of size ranging between $130 \times 70 \times 300$ and $210 \times 210 \times 150$ was used. The number of frames M used to estimate $\alpha_{\vec{x}\vec{y}}(\tau)$ (7) varies between 2,000 and 7,000 (i.e., from one to four minutes of video) depending on the sequence.

Figure 11 shows three examples of spatially distributed anomaly detection. One example shows left-to-right motor traffic. The pixel $\vec{x}$ is located in the traffic lane, and its co-occurrence matrix, shown in Figure 5(a), clearly exhibits a strong unimodality of motion. After a while, however, a pedestrian shows up, walking from right to left. Since the trace left by the pedestrian is significantly different from the co-occurrence histogram, the pedestrian is identified as an unusual moving object [red in Figure 11(a)]. The second example shows normal right-to-left motor traffic in the top lanes [Figure 11(b)], but, this time, a car making an illegal U-turn is labeled abnormal. The third example, from the performance evaluation of tracking and surveillance (PETS) data set, shows a person dropping a bag, which is an unusual event and thus colored in red. The most updated PETS-related Web site is http://www.cvg.rdg.ac.uk/PETS2009/. After a while, another person walks by the abandoned bag but does not stop or linger and thus is not labeled as anomalous.

To quantitatively validate our spatially distributed anomaly detection, we compared it with an object-based method using object path analysis [12]. The resulting confusion matrices are shown in Table 2. The test was carried out on two of our own videos and two PETS videos (see the caption of Table 2 for details) with the total of 80 normal and 12 abnormal events. While the co-occurrence-based method discovered all abnormal events with only 9% of false positives, the path-based method produced a 10% miss rate and 19.9% false positive rate. More results can be found in [4].

MULTICAMERA ANOMALIES

Matching activities in two cameras is usually possible if they are calibrated. However, in heterogeneous camera networks, calibration may not be possible. For instance, cameras have different locations, orientations, and zoom levels. Perhaps more importantly, the camera network topology can be dynamic, i.e., cameras can frequently change orientations and zoom levels. In such situations, the activity matching approach



[FIG11] Spatially distributed anomaly detection results: (a) left-to-right motor traffic labeled as normal (green) and right-to-left walking pedestrian labeled as anomaly (red); (b) normal motor traffic (green) and an illegal U-turn (red); (c) an abandoned bag and a person dropping it (red), and another person passing by but not stopping/lingering (green).

[TABLE 2] CONFUSION MATRICES FOR SPATIALLY DISTRIBUTED ANOMALY DETECTION BASED ON CO-OCCURRENCE MATRIX COMPARISON AND FOR CHEN ET AL.'S METHOD [12] ON VIDEOS FROM FIGURE 8(B) AND (E), AND FIGURE 11(B) AND (C).

CO-OCCURRENCE METHOD			METHOD OF CHEN ET AL. [12]		
	TRUE OCCURRENCE			TRUE OCCURRENCE	
	NORMAL	ABNORMAL		NORMAL	ABNORMAL
NORMAL	90.0%	0%	NORMAL	80.1%	10.0%
ABNORMAL	9.0%	100%	ABNORMAL	19.9%	90.0%

described in the section "Geometric Invariance and Multicamera Nominal Maps" can help establish correspondences and lead to nominal maps for multicamera activity. We have conducted a few experiments based on this idea [5]; these experiments have demonstrated the utility of this approach for multicamera anomaly detection. Nevertheless, more experiments are needed to fully understand the advantages and limitations of such an approach.

CONCLUSIONS

We have described a family of unsupervised approaches to video anomaly detection that are based on statistical activity analysis of background-subtracted video sequences. The described methods do not require identifying or tracking of objects, yet they can be effectively employed for discovering a variety of anomalies. This is particularly advantageous in highly cluttered urban scenarios.

We have provided a statistical framework for modeling activity and discovering anomalies. Central to our approach is the notion of an anomaly as the presence or absence of usual or unusual objects at unusual or usual locations and times. This formulation leads to various anomalies ranging from spatially localized and temporally persistent anomalies, to spatially distributed anomalies and to structured anomalies in unstructured scenarios.

In our framework, activity at each location is modeled by means of a binary state Markov chain that associates a feature descriptor, such as size, shape, and motion vector, with the moving state. The durations of moving and static states together with feature descriptors provide a statistical model for nominal activity. Computing statistics of these features at various spatial and/or temporal scales permits characterization of moving object behavior and the discovery of anomalies. Furthermore, the saliency of signatures permits spatial association within the camera's field of view, for example by means of co-occurrence analysis. Another crucial property of these signatures is their essential invariance to geometry (camera orientation). This provides a unique distinguishing attribute for pixel-level correspondences across different cameras.

The proposed methods are not without limitations, especially in cluttered environments. First, we assume that meaningful, though not necessarily perfect, motion labels can be computed. Second, all abnormalities, to be detected, need to have a different activity signature from that of the background. Third, to insure geometric invariance camera positioning in the scene and object height are both subject to restrictions (the objects cannot be too tall or camera positioned too low over the scene). Furthermore, for multicamera anomaly detection to be effective the occurrence of events needs to be independent or uncorrelated for pixels not in correspondence.

Working under these limitations, we demonstrated that our approach is capable of reliably discovering both localized and spatially distributed anomalies based on pixel-level analysis and without higher-level processing that involves explicit tracking of objects. Our framework requires storage of binary values only thus reducing memory requirements, and in some implementations it is straightforward enough to be considered for embedded platforms at camera network edge. It is our hope that this framework will open gates to many other exciting applications in the near future.

ACKNOWLEDGMENTS

This research was supported by the Department of Homeland Security under award number 2008-ST-061-ED0001, National Geospatial-Intelligence Agency (NGA) under award HM1582-09-1-0037, Presidential Early Career Award (PECASE) N00014-02-100362, NSF Career Award ECS 0449194, NSF Award CCF-0905541 and the NSERC Discovery Grant 371951. The authors would like to thank Erhan Ermis and Pierre Clarot for their help with preparation of the multicamera matching results and Yannick Benezeth for co-occurrence results.

AUTHORS

Venkatesh Saligrama (srv@bu.edu) is a faculty member in the Department of Electrical and Computer Engineering at Boston University. He holds a Ph.D. degree from Massachusetts Institute of Technology. His research interests are in statistical signal processing, information and control theory, and statistical learning theory and its applications to video analytics. He has edited a book on networked sensing, information, and control. He was an associate editor for *IEEE Transactions on Signal Processing* and is currently a member of the Signal Processing Theory and Methods Committee. He received the Presidential Early Career Award, ONR Young Investigator Award, and the NSF Career Award.

Janusz Konrad (jkonrad@bu.edu) received the M.Eng. degree from the Technical University of Szczecin, Poland, and the Ph.D. degree from McGill University, Montréal, Canada. From 1989 to 2000 he was with INRS-Télécommunications, Montréal. Since 2000, he has been with Boston University. He is an associate technical editor for *IEEE Communications Magazine* and associate editor for *EURASIP International Journal on Image and Video Processing*. He was an associate

editor for *IEEE Transactions on Image Processing* and *IEEE Signal Processing Letters*, a member of the IMDSP Technical Committee of the IEEE Signal Processing Society, as well as the Technical Program cochair of AVSS-2010 and ICIP-2000 and Tutorials cochair of ICASSP-2004. He is a corecipient of the 2001 IEEE Signal Processing Magazine Award and the 2004–2005 EURASIP Image Communications Best Paper Award. He is a Fellow of the IEEE.

Pierre-Marc Jodoin (pierre-marc.jodoin@usherbrooke.ca) studied physics at the Université de Montréal and then completed a B.S. degree in computer science at the École Polytechnique de Montréal in 2000. He then obtained an M.S. degree in computer graphics in 2002 and a Ph.D. degree in computer vision and video analysis in 2007, both from the Université de Montréal. He is an assistant professor at the Université de Sherbrooke, Canada. His research interests are in video surveillance, video analysis/processing, medical imaging, and computer vision.

REFERENCES

[1] J. Vlahos, "Surveillance society: New high-tech cameras are watching you," *Popular Mechanics*, Jan. 2008, pp. 64–69.

[2] P.-M. Jodoin, V. Saligrama, and J. Konrad, "Behavior subtraction," in *Proc. SPIE Visual Communications and Image Processing*, Jan. 2008, vol. 6822, pp. 10.1–10.12.

[3] E. Ermis, V. Saligrama, P.-M. Jodoin, and J. Konrad, "Motion segmentation and abnormal behavior detection via behavior clustering," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2008, pp. 769–772.

[4] Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurrences," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2009, pp. 2458–2465.

[5] E. Ermis, P. Clarot, P.-M. Jodoin, and V. Saligrama, "Activity based matching in distributed camera networks," *IEEE Trans. Image Processing,* to be published.

[6] P.-M. Jodoin, V. Saligrama, and J. Konrad. (2009). Behavior subtraction. CoRR. abs/0910.2917 [Online]. Available: http://arxiv.org/abs/0910.2917

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1:58, 2009.

[8] V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer-Verlag, 1994.

[9] C.-H. Chuang, J.-W. Hiseh, L.-W. Tsai, S.-Y. Chen, and K.-C. Fan, "Carried object detection using ratio histogram and its application to suspicious event analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 6, pp. 911–916, 2009.

[10] K. Smith, P. Quelhas, and D. Gatica-Perez, "Detecting abandoned luggage items in a public space," in *Proc. IEEE Performance Evaluation of Tracking and Surveillance Workshop (PETS)*, 2006, pp. 75–82.

[11] C. Piciarelli, C. Micheloni, and G. Foresti, "Trajectory-based anomalous event detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1544–1554, 2008.

[12] T. Chen, H. Haussecker, A. Bovyrin, R. Belenov, K. Rodyushkin, A. Kuranov, and V. Eruhimov, "Computer vision workload analysis: Case study of video surveillance systems," *Intell. Technol. J.*, vol. 9, no. 2, pp. 109–118, 2005.

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

[14] 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, 2005.

[15] M. Bennewitz, G. Cielniak, and W. Burgard, "Utilizing learned motion patterns to robustly track persons," in *Proc. IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Nice, France, 2003, pp. 102–109.

[16] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 831–843, 2000.

[17] N. Vaswani, A. Roy-Chowdhury, and R. Chellappa, "Shape activity: A continuous state hmm for moving/deforming shapes with application to abnormal activity detection," *IEEE Trans. Image Processing*, vol. 14, no. 10, pp. 1603–1616, 2005.

[18] N. Johnson and D. Hogg, "Learning the distribution of object trajectories for event recognition," *Image Vis. Comput.*, vol. 14, no. 8, pp. 609–615, 1996.

[19] N. Sumpter and A. Bulpitt, "Learning spatio-temporal patterns for predicting object behavior," *Image Vis. Comput.*, vol. 18, no. 9, pp. 697–704, 2000.

[20] C. Stauffer and E. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 747–757, 2000.

[21] I. Saleemi, K. Shafique, and M. Shah, "Probabilistic modeling of scene dynamics for applications in visual surveillance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 8, pp. 1472–1485, 2009.

[22] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 17–31, 2007.

[23] I. Pruteanu-Malinici and L. Carin, "Infinite hidden Markov models for unusual-event detection in video," *IEEE Trans. Image Processing*, vol. 17, no. 5, pp. 811–821, 2008.

[24] A. Yilmaz and M. Shah, "A differential geometric approach to representing the human actions," *Comput. Vis. Image Understand.*, vol. 109, no. 3, pp. 335–351, 2008.

[25] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.

[26] J. Kim and G. Kristen, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2009, pp. 2921–2928.

[27] Q. Dong, Y. Wu, and Z. Hu, "Pointwise motion image (PMI): A novel motion representation and its applications to abnormality detection and behavior recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 407–416, 2009.

[28] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 3, pp. 555–560, 2008.

[29] X. Wang, X. Ma, and E.Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 3, pp. 539–555, 2009.

[30] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2009, pp. 935–942.

[31] P. Cui, L.-F. Sun, Z.-Q. Liu, and S.-Q. Yang, "A sequential Monte Carlo approach to anomaly detection in tracking visual events," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2007, pp. 1–8.

[32] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 5, pp. 893–908, 2008.

[33] J. Oh, J. Lee, and S. Kote, "Real time video data mining for surveillance video streams," in *Proc. Conf. Knowledge Discovery and Data Mining*, 2003, pp. 222–233.

[34] J. Li, S. Gong, and T. Xiang, "Scene segmentation for behavior correlation," in *Proc. European Conf. Computer Vision*, 2008, pp. 383–395.

[35] C. Simon, J. Meessen, and C. DeVleeschouwer, "Visual event recognition using decision trees," *Multimedia Tools Applicat.*, vol. 50, no. 1, pp. 951–121, Oct. 2010.

[36] J. McHugh, J. Konrad, V. Saligrama, and P.-M. Jodoin, "Foreground-adaptive background subtraction," *IEEE Signal Process. Lett.*, vol. 16, no. 5, pp. 390–393, May 2009.

[37] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density for visual surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, 2002.

[38] P.-M. Jodoin, J. Konrad, V. Saligrama, and V. Veilleux-Gaboury, "Motion detection with an unstable camera," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2008, pp. 229–232.

[39] D. Devarajan, Z. Cheng, and R. Radke, "Calibrating distributed camera networks," *Proc. IEEE* (Special Issue on Distributed Smart Cameras), vol. 96, no. 10, pp. 1625–1639, 2008.

[40] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[41] A. van den Hengel, H. Detmold, C. Madden, A. Dick, A. Cichowski, and R. Hill, "A framework for determining overlap in large scale networks," in *Proc. IEEE Conf. Distributed Smart Cameras*, 2009, pp. 1–8.

[42] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.

[SP]