ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/239943156

Anomaly Detection and Localization in Crowded Scenes

Article *in* IEEE Transactions on Pattern Analysis and Machine Intelligence · June 2013

DOI: 10.1109/TPAMI.2013.111 · Source: PubMed

CITATIONS
268

READS
2,319

3 authors, including:

Wei-Xin LI
Facebook
8 PUBLICATIONS **958** CITATIONS

SEE PROFILE

Nuno Vasconcelos
University of California, San Diego
234 PUBLICATIONS **11,152** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Guided Augmentation for One shot and few shot learning View project

Project Bag-of-Words Image Classification View project

All content following this page was uploaded by Nuno Vasconcelos on 15 April 2015.

The user has requested enhancement of the downloaded file.

Anomaly Detection and Localization in Crowded Scenes

Weixin Li, *Student Member, IEEE,* Vijay Mahadevan, *Member, IEEE,*
and Nuno Vasconcelos, *Senior Member, IEEE*

Abstract—The detection and localization of anomalous behaviors in crowded scenes is considered, and a joint detector of temporal and spatial anomalies is proposed. The proposed detector is based on a video representation that accounts for both appearance and dynamics, using a set of mixture of dynamic textures models. These models are used to implement 1) a center-surround discriminant saliency detector that produces spatial saliency scores, and 2) a model of normal behavior that is learned from training data and produces temporal saliency scores. Spatial and temporal anomaly maps are then defined at multiple spatial scales, by considering the scores of these operators at progressively larger regions of support. The multi-scale scores act as potentials of a conditional random field that guarantees global consistency of the anomaly judgments. A dataset of densely crowded pedestrian walkways is introduced and used to evaluate the proposed anomaly detector. Experiments on this and other datasets show that the latter achieves state-of-the-art anomaly detection results.

Index Terms—Video Analysis, Surveillance, Anomaly Detection, Crowded Scene, Dynamic Texture, Center-surround Saliency.

1 Introduction

Surveillance video is extremely tedious to monitor when events that require follow-up have very low probability. For crowded scenes, this difficulty is compounded by the complexity of normal crowd behaviors. This has motivated a surge of interest in anomaly detection in computer vision [1], [2], [3], [4], [5], [6], [7], [8], [9]. However, this effort is hampered by general difficulties of the anomaly detection problem [10]. One fundamental limitation is the lack of a universal definition of anomaly. For crowds, it is also infeasible to enumerate the set of anomalies that are possible in a given surveillance scenario. This is compounded by the sparseness, rarity, and discontinuity of anomalous events, which limit the number of examples available to train an anomaly detection system.

One common solution to these problems is to define anomalies as events of low probability with respect to a probabilistic model of normal behavior. This enables a statistical treatment of anomaly detection, which conforms with the intuition of anomalies as events that deviate from the expected [10]. However, it introduces a number of challenges. First, it makes anomalies dependent on the *scale* at which normalcy is defined. A normal behavior at a fine visual scale may be perceived as highly anomalous when a larger scale is considered, or vice-versa. Hence, normalcy models must be defined at multiple scales. Second, different tasks may require *different models of normalcy.* For instance, a detector of freeway speed limit violations will rely on normalcy models based on speed features. On the other hand, appearance is more important for the detection of carpool lane violators, i.e., single-passenger vehicles in carpool lanes. Third, crowded scenes require normalcy models robust to *complex scene dynamics*, involving many independently moving objects that occlude each other in complex ways, and can have low-resolution.

In result, anomaly detection can be extremely challenging. While this has motivated a great diversity of solutions, it is usually quite difficult to objectively compare different methods. Typically, these combine different representations of motion and appearance with different graphical models of normalcy, which are usually tailored to specific scene domains. Abnormalities are themselves defined in a somewhat subjective form, sometimes according to what the algorithms can detect. In some cases, different authors even define different anomalies on common datasets. Finally, experimental results can be presented on datasets of very different characteristics (*e.g.*, traffic intersection *v.s.* subway entrance), frequently proprietary, and with widely varying levels of crowd density.

In this work, we propose an integrated solution to all these problems. We start by introducing normalcy models that *jointly account for the appearance and dynamics of complex crowd scenes.* This is done by resorting to a video representation based on dynamic textures (DTs) [11]. This representation is then used to design models of normalcy over both space and time. *Temporal normalcy* is modeled with a mixture of DTs [12] (MDT) and enables the detection of behaviors that deviate from those observed in the past. *Spatial normalcy* is measured with a discriminant saliency detector [13] based on MDTs, enabling the detection of behaviors that deviate from those of the surrounding crowd. The integration of spatial and temporal normalcy with respect to either appearance or dynamics leads to a flexible model of normalcy, applicable to the detection of

- *W. Li and N. Vasconcelos are with the Electrical and Computer Engineering Department, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093. E-mail: {wel017, nvasconcelos}@ucsd.edu.*
- *V. Mahadevan is with Yahoo! Labs, Embassy Golf Links Business Park, Bangalore 560071, India. E-mail: vijay.mahadevan@gmail.com.*

anomalies of relevance to various surveillance tasks.

To address the *scale* problem, MDTs are learned at multiple spatial scales. This is done with an efficient hierarchical model, where layers of MDTs with successively larger regions of video support are learned recursively. The local measures of spatial and temporal abnormality are then integrated into a globally coherent anomaly map, by probabilistic inference. This is implemented with a conditional random field (CRF), whose single-node potentials are classifiers of local measures of spatial and temporal abnormality, collected over a range of spatial scales. They are complemented by a novel set of interaction potentials, which account for spatial and temporal context, and integrate anomaly information across the visual field.

Finally, to address the difficulties of empirical evaluation of anomaly detectors on crowded scenes, we introduce a *dataset* of video from walkways in the campus of University of California, San Diego (UCSD), depicting crowds of varying densities. The dataset contains 98 video sequences, and five well defined abnormal categories. These are not "synthetic", or "staged", but abnormal events that occur naturally, *e.g.*, bicycle riders that cross pedestrian walkways. Ground truth is provided for abnormal events, as well as a protocol to evaluate detection performance.

The remainder of the paper is organized as follows. Section 2 reviews previous work on anomaly detection in computer vision. The problems of temporal and spatial anomaly detection in crowded scenes are discussed in Section 3. This is followed by the mathematical characterization of multi-scale anomaly maps, in Section 4, and the proposed CRF for integration of spatial and temporal anomalies across different spatial scales, in Section 5. Finally, an extensive experimental evaluation is discussed in Section 6 and some conclusions are presented in Section 7.

2 Prior Work

Recent advances in anomaly detection address event representation and globally consistent statistical inference. Contributions of the first type define features and models for the discrimination of normal and anomalous patterns. Models of normal and abnormal behavior are then learned from training data, and anomalies detected with a minimum probability of error decision rule. Although there are some exceptions [5], the distribution of abnormal patterns is usually assumed uniform, and abnormal events formulated as events of low probability under the model of normalcy.

One intuitive representation for event modeling is based on object trajectories. It comprises of either explicitly or implicitly segmenting and tracking each object in the scene, and fitting models to the resulting object tracks [14], [15], [16], [6], [17], [18]. While capable of identifying abnormal behaviors of high-level semantics (*e.g.*, unusual long-term trajectories), these procedures are both difficult and computationally expensive for crowded or cluttered scenes. A number of promising alternatives, which avoid processing individual objects, have been recently proposed. These include the modeling of motion patterns with histograms of pixel change [5], histograms of optical flow [19], [8], [20], or optical flow measures [3], [4], [17], [1]. Among these, [3] models local optical flow with a mixture of probabilistic principal component analysis (PCA) models, [4] and [17] draw inspiration from classical studies of crowd behavior [21] to characterize flow with interaction features (*e.g.*, social force model), and [1] learns the representative flow of groups by clustering optical flow-based particle trajectories.

These approaches emphasize dynamics, ignoring anomalies of object appearance and thus anomalous behavior without outlying motion. Optical flow, pixel change histograms, or other classical background subtraction features, are also difficult to extract from crowded scenes, where the background is by definition dynamic, there are lots of clutter, and occlusions. More complete representations account for both appearance and motion. For example, [2] models temporal sequences of spatio-temporal gradients to detect anomalies in densely crowded scenes, [22] declares as abnormal spatio-temporal patches that cannot be reconstructed from previous frames, and [23] pools appearance and motion features over spatial neighborhoods, using a distance to the nearest spatially co-located feature vector among all training video clips, to quantify abnormality.

Object-based representations, based on location, blob shape, and motion [7] or optical flow magnitude, gradients, location, and scale [9] have also been proposed. Other representations include a bag-of-words over a set of manually annotated event classes [24]. Various methods have also been used to produce anomaly scores. While simple spatial filtering suffices for some applications [19], crowded scenes require more sophisticated graphical models and inference. For example, [6] and [1] adopt Gaussian mixture models (GMM) to represent trajectories of normal behavior. [8] and [20] learn a sparse basis and define unusual events as those that can only be reconstructed with either large error or the combination of a large number of basis vectors.

Contributions of the second type address the integration of local anomaly scores, which can be noisy, into a globally consistent anomaly map. [2], [25] and [7] guarantee temporally consistent inference by modeling normal temporal sequences with hidden Markov models (HMMs). While this enforces consistency along the temporal dimension, there have also been efforts to produce spatially consistent anomaly maps. For example, latent Dirichlet allocation (LDA) has been applied to force flow features, in the model of spatial crowd interactions of [4]. On the other hand, [5] and [3] rely on Markov random fields (MRF) to enforce global spatial consistency. In the realm of sparse representations, [20] guarantees consistency of reconstruction coefficients over space and time by inclusion of smoothness terms in the underlying optimization problem. Finally, [9] models object relationships, using Bayesian networks to implement occlusion reasoning.

It should be noted that most of these methods have not been tested on the densely crowded scenes considered in this work. It is unclear that many of them could deal with the complex motion and object interactions prevalent in such scenes. Furthermore, while most methods

include some mechanism to encourage spatial and temporal consistency of anomaly judgements (MRF, LDA, *etc*) the underlying decision rule tends to be either predominantly temporal (*e.g.*, trajectories, GMMs, HMMs or sparse representations learned over time) or spatial (*e.g.*, interaction models) but is rarely discriminant with respect to both space and time. This makes it difficult to infer whether spatial or temporal modeling are critically important by themselves, or what benefits are gained from their joint modeling. Furthermore, the role of scale is rarely considered. These issues motivate the contributions of the following sections.

3 Anomaly Detection

We start by proposing an anomaly detector that accounts for scene appearance and dynamics, spatial and temporal context, and multiple spatial scales.

3.1 Mathematical Formulation

A classical formulation of anomaly detection, which we adopt in this work, equates anomalies to outliers. A statistical model $p_{\mathbf{X}}(\boldsymbol{x})$ is postulated for the distribution of a measurement $\boldsymbol{X}$ under *normal* conditions. Abnormalities are defined as measurements whose probability is below a threshold under this model. This is equivalent to a statistical test of hypotheses

- $\mathcal{H}_0$: $\boldsymbol{x}$ is drawn from $p_{\mathbf{X}}(\boldsymbol{x})$;
- $\mathcal{H}_1$: $\boldsymbol{x}$ is drawn from an uninformative distribution $p_{\mathbf{X}}(\boldsymbol{x}) \propto 1$.

The minimum probability of error rule for this test is to reject the null hypothesis $\mathcal{H}_0$ if $p_{\mathbf{X}}(\boldsymbol{x}) < \nu$, where ν is the normalization constant of the uninformative distribution. As usual in the literature, we consider the problem of anomaly detection from *localized* video measurements $\boldsymbol{x}$, where $\boldsymbol{x}$ is a spatio-temporal patch of small dimensions.

3.2 Spatial *v.s.* Temporal Anomalies

The normalcy model $p_{\mathbf{X}}(\boldsymbol{x})$ can have both a *temporal* and a *spatial* component. Temporal normalcy reflects the intuition that normal events are *recurrent* over time, *i.e.*, previous observations establish a contextual reference for normalcy judgements. Consider a highway lane where cars move with a certain orientation and speed. Bicycles or cars heading in the opposite direction are easily identified as abnormal because they give rise to observations $\boldsymbol{x}$ substantially different from those collected in the past. In this sense, temporal normalcy detection is similar to *background subtraction* [26]. A model of normal behavior is learned over time, and measurements that it cannot explain are denoted *temporal anomalies*.

Spatial normalcy reflects the intuition that some events which would not be abnormal per se are abnormal *within* a crowd. Since the crowd places physical or psychological constraints on individual behavior, behaviors feasible in isolation can have low probability in a crowd context. For example, while there is nothing abnormal about an ambulance that rides at 50 mph in a stretch of highway, the same observation within a highly *congested* highway is abnormal. Note that the only indication of abnormality is the *difference* between the crowd and the object at the time of the observation, not that the ambulance moves at 50 mph. Since the detection of such abnormalities is mostly based on spatial context, they are denoted *spatial anomalies*. Their detection does not depend on memory. Instead, it is based on a continuously evolving, instantaneously adaptive, definition of normalcy. In this sense, the detection of spatial anomalies can be equated to *saliency detection* [27].

3.3 Roles of Crowds and Scale

Most available background subtraction and saliency detection solutions are not applicable to crowded scenes, where backgrounds can be highly dynamic. In this case, it is not sufficient to detect variations of image intensity, or even optical flow, to detect anomalous events. Instead, normalcy models must rely on sophisticated *joint representations of appearance and dynamics*. In fact, even such models can be ineffective. Since crowds frequently contain distinct sub-entities, *e.g.*, vehicles or groups of people moving in different directions, anomaly detection requires modeling *multiple video components of different appearance and dynamics*. A model that has been shown successful in this context is the mixture of DTs (MDT) [12]. This is the representation adopted in this work.

Another challenging aspect of anomaly detection within crowds is scale. Spatial anomalies are usually detected at the scale of the smallest scene entities, typically people. However, a normal event at this scale may be anomalous at a larger scale, and vice-versa. For example, while a child that rides a bicycle appears normal within a group of bicycle riding children, the group is itself anomalous in a crowded pedestrian sidewalk. Local anomaly detectors, with small regions of interest, cannot detect such anomalies. To address this, we represent crowded scenes with a hierarchy of MDTs that cover successively larger regions. This is done with a computationally efficient hierarchical model, where MDT layers are estimated recursively.

A similar challenge holds for temporal anomalies. While their detection is usually based on a small number of video frames, certain anomalies can only be detected over long time spans. For example, while it is normal for two pedestrian trajectories to converge or diverge at any point in time, a cyclical convergence and divergence is probably abnormal. Anomaly detection across time scales is, however, more complex than across spatial scales, due to constraints of instantaneous detection and implementation complexity. Since video has to be buffered before anomalies can be detected, large temporal windows imply long detection delays and storage of many video frames. Due to this, we do not consider multiple temporal scales in this work. A single scale is chosen, using acceptable values of delay and storage complexity, and used throughout our experiments. Note that, like their spatial counterparts, temporal anomaly maps are computed at multiple spatial scales. Hence, in what follows, the term "scale" refers to the spatial support of anomaly detection, for *both* spatial and temporal anomalies.

Fig. 1. Temporal anomaly detection. An MDT is learned per scene subregion, at training time. A temporal anomaly map is produced by measuring the negative log probability of each video patch under the MDT of the corresponding region.

4 Normalcy and Anomaly Modeling

In this section we review the MDT model, discuss the design of temporal and spatial models of normalcy, and formulate the computation of anomaly maps.

4.1 Mixture of Dynamic Textures

The MDT models a sequence of τ video frames $\boldsymbol{x}_{1:\tau} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_\tau]$ as a sample from one of K dynamic textures (DTs) [11]

$$p(\boldsymbol{x}_{1:\tau}) = \sum_{i=1}^{K} \pi_i p(\boldsymbol{x}_{1:\tau}|z=i). \tag{1}$$

The mixture components $p(\boldsymbol{x}_{1:\tau}|z=i)$ are linear dynamic systems (LDS) defined by

$$\begin{cases} \boldsymbol{s}_{t+1} &= A_z \boldsymbol{s}_t + \boldsymbol{n}_t, \quad (2a) \\ \boldsymbol{x}_t &= C_z \boldsymbol{s}_t + \boldsymbol{m}_t, \quad (2b) \end{cases}$$

where Z is a multinomial random variable of parameters $\boldsymbol{\pi}$ ($\pi_i \geqslant 0, \sum_i \pi_i = 1$), which indexes the mixture component from which $\boldsymbol{x}_t$ is drawn. $\boldsymbol{s}_t$ is a hidden state variable that encodes scene dynamics, and $\boldsymbol{x}_t$ the vector of pixels in video frame t. A_z, C_z are the transition and observation matrices of component z, whose initial condition is $\boldsymbol{s}_1 \sim \mathcal{N}(\boldsymbol{\mu}_z, S_z)$, and noise processes are defined by $\boldsymbol{n}_t \sim \mathcal{N}(\mathbf{0}, Q_z)$ and $\boldsymbol{m}_t \sim \mathcal{N}(\mathbf{0}, R_z)$. The model parameters are learned by maximum likelihood estimation (MLE) from a collection of video patches, with the expectation-maximization (EM) algorithm of [12], which is reviewed in Appendix A.1.

4.2 Temporal Anomaly Detection

Temporal anomaly detection is inspired by the popular background subtraction method of [26]. This uses a GMM per image location to model the distribution of image intensities. Observations of low probability under these GMMs are declared foreground. For anomaly detection in crowds, the GMM is replaced by an MDT, and the pixel grid replaced by one of preset displacement. Grid locations define the center of video cells, from which video patches are extracted. The patches extracted from a subregion (group of cells) are used to learn an MDT, during a training phase, as illustrated in Fig. 1. After this phase, subregion patches of low probability under the associated MDT are considered anomalies. Given patch $\boldsymbol{x}_{1:\tau}$, the distribution of the hidden state sequence $\boldsymbol{s}_1^\tau$ under the ith DT component, $p_{\mathbf{S}|\mathbf{X}}(\boldsymbol{s}_{1:\tau}|\boldsymbol{x}_{1:\tau}, z=i)$, is estimated with a Kalman filter and smoother [28], [29], as discussed in Appendix A.2. The value of the temporal anomaly map at location l is the negative-log probability of the most-likely state sequence for the patch at l

$$\mathcal{T}(l) = -\log\left[\sum_{i=1}^{K} \pi_i p(\boldsymbol{s}_{1:\tau}^{\{i\}}(l)|z=i)\right], \tag{3}$$

where $\boldsymbol{s}_{1:\tau}^{\{i\}}(l) = \operatorname{argmax}_{\boldsymbol{s}_{1:\tau}} p(\boldsymbol{s}_{1:\tau}|\boldsymbol{x}^\tau(l), z=i)$. We note that this generalizes the mixture of PCA models of optical flow [3]. The matrix C_z of (2b) is a PCA basis for patches drawn from mixture component z, but the PCA decomposition reports to *patch appearance*, not optical flow. Patch dynamics are captured by the hidden state sequence $\boldsymbol{s}_{1:\tau}$, which is a trajectory in PCA space. Hence, unlike mixtures of optical flow, the representation is temporally *smooth*. The joint representation of appearance and dynamics makes the MDT a better representation for crowd video than the mixture of PCA.

4.3 Spatial Anomaly Detection

Spatial anomaly detection is inspired by previous work in saliency detection [27], [13]. Saliency is defined in a center-surround manner. Given a set of features, salient locations are those of substantial feature contrast with their immediate surround. Spatial anomalies are then defined as locations whose saliency is above some threshold. In this work, we rely on the discriminant saliency criterion of [13].

4.3.1 *Discriminant Saliency*

Discriminant saliency formulates the saliency problem as an hypothesis test between two classes: a class of *salient stimuli*, and a *background* class of stimuli that are not salient. Two windows are defined at each scene location l: a *center window* $\mathcal{W}_l^1$, with label $\mathcal{C}(l)=1$, containing the location, and a surrounding annular window $\mathcal{W}_l^0$, with label $\mathcal{C}(l)=0$, containing *background*. A set of feature responses $\mathbf{X}$ are computed for each of the windows $\mathcal{W}_l^c, c \in \{0,1\}$ and $\mathcal{S}(l)$, the saliency of location l, defined as the extent to which they discriminate between the two classes. This is quantified by the mutual information (MI) between feature responses and class label [13]

$$\mathcal{S}(l) = \sum_{c=0}^{1} \left\{ p_{\mathcal{C}(l)}(c) \mathrm{KL}\left[p_{\mathbf{X}|\mathcal{C}(l)}(\boldsymbol{x}|c) \| p_{\mathbf{X}}(\boldsymbol{x})\right] \right\}, \tag{4}$$

where $p_{\mathbf{X}|\mathcal{C}(l)}(\boldsymbol{x}|c)$ are class-conditional densities and $\mathrm{KL}(p\|q) = \int_{\mathcal{X}} p_{\mathbf{X}}(\boldsymbol{x}) \log \frac{p_{\mathbf{X}}(\boldsymbol{x})}{q_{\mathbf{X}}(\boldsymbol{x})} d\boldsymbol{x}$ the Kullback-Leibler (KL) divergence between $p_{\mathbf{X}}(\boldsymbol{x})$ and $q_{\mathbf{X}}(\boldsymbol{x})$ [30].

Locations of maximal saliency are those where the discrimination between center and surround can be made with highest confidence, *i.e.*, where (4) is maximal. The discriminant saliency principle can be applied to many features [31]. When $\mathbf{X}$ consists of optical flow, it generalizes

Fig. 2. Spatial anomaly detection using center-surround saliency with MDT models.

the force flow model of [4], where saliency is defined as the difference between the optical flow at l and the average flow in its neighborhood (see (8) in [4]). This is a simplified form of discriminant saliency, which replaces the MI of (4) by a difference to the mean background response.

4.3.2 Center-Surround Saliency with MDTs

Optical flow methods provide a coarse representation of dynamics and ignore appearance. For background subtraction, this problem has been addressed with the combination of DTs and discriminant saliency [32]. While using a more powerful representation than force flow, this method learns a single DT from both center and surround windows. This assumes an homogeneity of appearance and dynamics within the two windows that does not hold for crowds, where foregrounds and backgrounds can be quite diverse.

In this work, we adopt the MDT as the probability distribution $p_{\mathbf{X}|\mathcal{C}(l)}(\boldsymbol{x}_{1:\tau}|c)$ from which spatio-temporal patches $\boldsymbol{x}_1^\tau$ are drawn. We note that, under assumptions of Gaussian initial conditions and noise, patches $\boldsymbol{x}_{1:\tau}$ drawn from a DT have a Gaussian probability distribution [33],

$$\boldsymbol{x}_{1:\tau} \sim \mathcal{N}(\boldsymbol{\gamma}, \Phi), \tag{5}$$

whose parameters follow from those of the LDS (2). When the class-conditional distributions of the center and surround classes, $c \in \{0,1\}$, at location l are mixtures of K_c DTs, it follows that

$$\begin{aligned} p_{\mathbf{X}|\mathcal{C}(l)}(\boldsymbol{x}_{1:\tau}|c) &= \sum_{i=1}^{K_c} \pi_i^c \mathcal{N}(\boldsymbol{x}_{1:\tau}; \boldsymbol{\gamma}_i^c, \Phi_i^c) \\ &= \sum_{i=1}^{K_c} \pi_i^c p_{\mathbf{X}|\mathcal{C}(l)}^i(\boldsymbol{x}_{1:\tau}|c), \end{aligned} \tag{6}$$

for $c \in \{0,1\}$. The marginal distribution is then

$$\begin{aligned} p_{\mathbf{X}}(\boldsymbol{x}_{1:\tau}) &= \sum_{c=0}^{1} \left[p_{\mathcal{C}(l)}(c) p_{\mathbf{X}|\mathcal{C}(l)}(\boldsymbol{x}_{1:\tau}|c) \right] \\ &= \sum_{c=0}^{1} \left[p_{\mathcal{C}(l)}(c) \sum_{i=1}^{K_c} \pi_i^c \mathcal{N}(\boldsymbol{x}_{1:\tau}; \boldsymbol{\gamma}_i^c, \Phi_i^c) \right] \\ &= \sum_{i=1}^{K_0+K_1} \omega_i \mathcal{N}(\boldsymbol{x}_{1:\tau}; \boldsymbol{\gamma}_i, \Phi_i) \\ &= \sum_{i=1}^{K_0+K_1} \omega_i p_{\mathbf{X}}^i(\boldsymbol{x}_{1:\tau}), \end{aligned} \tag{7}$$

and the saliency measure of (4) requires the KL divergence between (6) and (7). This is problematic, because there is no closed form solution for the KL divergence between two MDTs. However, since the MDT components are Gaussian, it is possible to rely on popular approximations to the KL divergence between Gaussian mixtures. We adopt the variational approximation of [34]

$$\mathrm{KL}\left(p_{\mathbf{X}|\mathcal{C}} \| p_{\mathbf{X}}\right) \approx \sum_i \left\{ \pi_i^{\mathcal{C}} \log \frac{\sum_j^{K_{\mathcal{C}}} \pi_j^{\mathcal{C}} \exp\left(-\mathrm{KL}\left(p_{\mathbf{X}|\mathcal{C}}^i \middle\| p_{\mathbf{X}|\mathcal{C}}^j\right)\right)}{\sum_j^{K_0+K_1} \omega_j \exp\left(-\mathrm{KL}\left(p_{\mathbf{X}|\mathcal{C}}^i \| p_{\mathbf{X}}^j\right)\right)} \right\}. \tag{8}$$

Each term of (8) contains a KL divergence between DTs, which can be computed in closed-form [35]. For example, for the terms in the denominator

$$\begin{aligned} &\mathrm{KL}\left(p_{\mathbf{X}|\mathcal{C}}^i \middle\| p_{\mathbf{X}}^j\right) = \\ &\frac{1}{2}\left[\log \frac{|\Phi_j|}{|\Phi_i^{\mathcal{C}}|} + \mathrm{Tr}\left(\Phi_j^{-1}\Phi_i^{\mathcal{C}}\right) + \left\|\boldsymbol{\gamma}_i^{\mathcal{C}} - \boldsymbol{\gamma}_j\right\|_{\Phi_j}^2 - m\tau \right], \end{aligned} \tag{9}$$

where m is the number of pixels per frame, and $||\boldsymbol{z}||_\Phi = \boldsymbol{z}^T \Phi^{-1} \boldsymbol{z}$. Numerator terms are computed similarly. All computations can be performed recursively [35].

4.3.3 Spatial Anomaly Map

The spatial anomaly map is a map of the saliency $\mathcal{S}(l)$ at locations l. Given a location, this requires 1) learning MDTs from center and surround windows, and 2) computing a weighted average of these mixtures to obtain (7). Since learning MDTs per location is computationally prohibitive, we resort to the following approximation. A dense collection of overlapping spatio-temporal patches are first extracted from $\mathcal{V}(t)$, a 3D video volume temporally centered at the current frame. A single MDT with K^g mixture components, denoted $\{\boldsymbol{\gamma}_i^g, \Phi_i^g\}_{i=1}^{K^g}$, is learned from this patch collection. Each patch is then assigned to the mixture component of largest posterior probability. This segments the volume into super-pixels, as shown in Fig. 2.

At location l, the MDTs of (6) and (7) are derived from the global mixture model. The DT components are assumed equal to those of the latter and only the mixing proportions are recomputed, using the ratio of pixels assigned to each component in the respective windows

$$p_{\mathbf{X}|\mathcal{C}(l)}(\boldsymbol{x}_{1:\tau}|c) = \sum_{i=1}^{K^g} \frac{\sum_{l \in \mathcal{W}_l^c} \mathcal{M}_{il}}{\sum_{l \in \mathcal{W}_l^c} 1} \mathcal{N}(\boldsymbol{x}_{1:\tau}; \boldsymbol{\gamma}_i^g, \Phi_i^g), \tag{10}$$

for $c \in \{0,1\}$. $\mathcal{M}_{il} = 1$ if l is assigned to mixture component i and 0 otherwise. The prior probabilities for center and surround, $p_{\mathcal{C}}(c)$, are proportional to the ratio of volumes of center and surround windows. $\mathcal{S}(l)$ is computed with (4), using (8) and (9). Note that the KL divergence terms in (8) only require the computation of $\binom{K^g}{2}$ KL divergences between the K^g mixture components, and these are computed only once per frame, since all mixture components are shared (*i.e.*, the terms $\exp\left(-\mathrm{KL}\left(p \| q\right)\right)$ in (8) are fixed per frame). This procedure is repeated for every frame in the test video, as illustrated in Fig. 2.

Fig. 3. Computation of temporal anomaly maps with multi-scale spatial supports using the H-MDT. MDTs of increasingly larger spatial support are estimated recursively, with the H-EM algorithm. Their application to a query video produces temporal anomaly maps based on supports of various spatial scales.

Fig. 4. CRF filter. Top: factor graph. Bottom: spatial and temporal neighborhoods.

4.4 Multi-scale Anomaly Maps

To account for anomalies at multiple spatial scales, we rely on a hierarchical mixture of dynamic textures (H-MDT). This is a model with various MDT layers, learned from regions of different spatial support. At the finest scale, a video sequence is divided into n_L subregions (*e.g.*, 5×8 subregions). n_L MDT models $\{\boldsymbol{M}_i\}_{i=1}^{n_L}$ are then learned from patches extracted from each of the subregions. At the coarsest scale, the whole visual field is represented with a global MDT. This results in a hierarchy of MDT models $\{\{\boldsymbol{M}_i^1\}_{i=1}^{n_1}, \cdots, \boldsymbol{M}_1^L\}$ where $\boldsymbol{M}_j^s$, the jth model at scale s, is learned from subregion $\mathcal{R}_j^s$. The hierarchy of support windows $\{\{\mathcal{R}_i^1\}_{i=1}^{n_1}, \cdots, \mathcal{R}^L\}$ resembles the spatial pyramid structure of [36]. H-MDT models can be learnt efficiently with the hierarchical expectation-maximization (H-EM) algorithm of [37]. Rather than collecting patches anew from larger regions, it estimates the models at a given layer directly from the parameters of the MDT models at the layer of immediately higher resolution.

For anomaly detection, each model is applied to the corresponding window. This produces L anomaly maps, $\{\mathcal{T}^1, \cdots, \mathcal{T}^L\}$, as illustrated in Fig. 3. A hierarchy of spatial anomaly maps, $\{\mathcal{S}^1, \cdots, \mathcal{S}^L\}$ is also computed. For all s, the computation of $\mathcal{S}^s$ relies on a global mixture model $\boldsymbol{M}$. The mixing proportions of (10) are computed using surround windows of size identical to $\{\mathcal{R}_i^s\}$ and center windows of constant size, as summarized in Algorithm 1 (see Appendix B for all algorithms).

5 Globally Consistent Anomaly Maps

In this section, we introduce a layer of statistical inference to fuse anomaly information across time, space, and scale in a globally consistent manner.

5.1 Discriminative Model

The anomaly maps of the previous section span space, time, and spatial scale. Being derived from local measurements, they can be noisy. A principled framework is required to 1) integrate anomaly scores from the individual maps, 2) eliminate noise, and 3) guarantee spatio-temporal consistency of anomaly judgements throughout the visual field. For this, we rely on a conditional random field (CRF) [38] inspired by the discriminative random field (DRF) of [39]. An anomaly label $y_i \in \{-1, 1\}$ is defined at each location i in a set S of observation sites. Given a video clip $\boldsymbol{x}$, the conditional likelihood of observing a configuration of anomaly labels $\{\boldsymbol{y}, i \in \mathsf{S}\}$ is

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z} \exp\Big\{ \sum_{i \in \mathsf{S}} \mathcal{A}(y_i, \boldsymbol{x}) + \sum_{i \in \mathsf{S}} \Big[\frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathcal{I}(y_i, y_j, \boldsymbol{x}, i, j) \Big] \Big\}, \tag{11}$$

where Z is a partition function and $\mathcal{N}_i$ the neighborhood of site i. The single site and interaction potentials of (11),

$$\mathcal{A}(y_i, \boldsymbol{x}) = \log \sigma(y_i \boldsymbol{w}^T \boldsymbol{f}_i), \tag{12}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function, and

$$\mathcal{I}(y_i, y_j, \boldsymbol{x}, i, j) = y_i y_j \cdot \boldsymbol{v}^T \boldsymbol{\mu}(\boldsymbol{f}_i, \boldsymbol{f}_j, i, j), \tag{13}$$

are based on a feature vector $\boldsymbol{f}_i$ that concatenates the spatial and temporal anomaly scores of site i at the L spatial scales, plus a bias term (set to 1)

$$\boldsymbol{f}_i = [1, \mathcal{T}^1(i), \cdots, \mathcal{T}^L(i), \mathcal{S}^1(i), \cdots, \mathcal{S}^L(i)]^T. \tag{14}$$

$\boldsymbol{w}, \boldsymbol{v}$ are parameter vectors and $\boldsymbol{\mu}$ a compound feature

$$\boldsymbol{\mu}(\boldsymbol{f}_i, \boldsymbol{f}_j, i, j) = e^{-\alpha|i-j|} \exp(-\boldsymbol{h}_{i,j}), \tag{15}$$

where $|i-j|$ is the Euclidean distance between sites i, j, and $\exp(-\boldsymbol{h}_{i,j})$ the entry-wise exponential of $-\boldsymbol{h}_{i,j}$. The vector $\boldsymbol{h}_{i,j}$ contains the diagonal entries of $(\boldsymbol{f}_i - \boldsymbol{f}_j)(\boldsymbol{f}_i - \boldsymbol{f}_j)^T$.

The single site potential of (12) reflects the anomaly belief at site i. Using it alone, *i.e.*, without (13), (11) is a logistic regression model. In this case, the detection of each anomaly is based on information from site i exclusively. The addition of the interaction potential of (13) enables

the model to take into account information from site i's neighborhood $\mathcal{N}_i$. This smoothes the single site prediction, encouraging consistency of neighboring labels. The interaction potential can be interpreted as a classifier that predicts whether two neighboring sites have the same label. Note that, since $\boldsymbol{f}$ contains anomaly scores at different spatial scales, $\boldsymbol{h}_{i,j}$ (or $\boldsymbol{\mu}_{i,j}$) accounts for the similarity between the two observations in anomaly spaces of different scale (*i.e.*, under different spatial normalcy contexts). The interaction potentials adaptively modulate the intensity of inter-site smoothing according to these similarity measures (and how they are weighted by $\boldsymbol{v}$). The parameters $\boldsymbol{w}$ and $\boldsymbol{v}$ encode the relative importance of different features.

5.2 Online CRF Filter

The model of (11) requires inference over the entire video sequence. This is not suitable for online applications. An online version can be implemented by conditioning the anomaly label $\boldsymbol{y}^{(\tau)}$ at time τ on 1) observations for $t \leqslant \tau$, and 2) anomaly labels for $t < \tau$, leading to

$$P(\boldsymbol{y}^{(\tau)}|\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1},\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau};\boldsymbol{\Theta}) = \frac{1}{\mathcal{Z}}\exp\Big\{\sum_{i\in\mathsf{S}^\tau}\Big[\mathcal{A}(y_i^{(\tau)},\boldsymbol{x}^{(\tau)}) + \frac{1}{|\mathcal{N}_i^{\mathbb{S}}|}\sum\nolimits_{j\in\mathcal{N}_i^{\mathbb{S}}}\mathcal{I}_{\mathbb{S}}(y_i^{(\tau)},y_j,\boldsymbol{x}^{(\tau)},i,j) + \frac{1}{|\mathcal{N}_i^{\mathbb{T}}|}\sum\nolimits_{k\in\mathcal{N}_i^{\mathbb{T}}}\mathcal{I}_{\mathbb{T}}(y_i^{(\tau)},y_k,\boldsymbol{x},i,k)\Big]\Big\}, \tag{16}$$

where S^τ is the set of observations at time τ (pixels of the current frame). Two neighborhoods are defined per location i: spatial $\mathcal{N}_i^{\mathbb{S}}$ ($\mathcal{N}_i^{\mathbb{S}} \subseteq \mathsf{S}^\tau$) and temporal $\mathcal{N}_i^{\mathbb{T}}$ ($\mathcal{N}_i^{\mathbb{T}} \subseteq \{\mathsf{S}^t\}_{t=1}^{\tau-1}$). The graphical model is shown at the top of Fig. 4, and these neighborhoods at the bottom. The parameters $\boldsymbol{\Theta} = \{\boldsymbol{w},\boldsymbol{v}_{\mathbb{T}},\boldsymbol{v}_{\mathbb{S}},\alpha_{\mathbb{T}},\alpha_{\mathbb{S}}\}$ are estimated during training.

5.2.1 Learning

Both (11) and (16) can be learned with standard optimization techniques, such as gradient descent or the *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) method. To improve generalization, the model is regularized with a Gaussian prior of standard deviation ϵ, for all parameters. Given N independent training samples $\{\boldsymbol{x}^{(n)},\boldsymbol{y}^{(n)}\}_{n=1}^N$, the gradients of the regularized log-likelihood with respect to $\boldsymbol{w}$, $\boldsymbol{v}$ and α are

$$\frac{\partial}{\partial\boldsymbol{w}}\log p(\{\boldsymbol{y}^{(n)}\}_{n=1}^N|\{\boldsymbol{x}^{(n)}\}_{n=1}^N) = -\frac{1}{\epsilon_{\boldsymbol{w}}^2}\boldsymbol{w} + \sum_{n=1}^N\Big\{\sum_{i\in\mathsf{S}}\sigma(-y_i^{(n)}\boldsymbol{w}^T\boldsymbol{f}_i^{(n)})y_i^{(n)}\boldsymbol{f}_i^{(n)} - \mathbb{E}\Big[\sum_{i\in\mathsf{S}}\sigma(-y_i\boldsymbol{w}^T\boldsymbol{f}_i^{(n)})y_i\boldsymbol{f}_i^{(n)}\Big]\Big\}, \tag{17}$$

$$\frac{\partial}{\partial\boldsymbol{v}}\log P(\{\boldsymbol{y}^{(n)}\}_{n=1}^N|\{\boldsymbol{x}^{(n)}\}_{n=1}^N) = -\frac{1}{\epsilon_{\boldsymbol{v}}^2}\boldsymbol{v} + \sum_{n=1}^N\Big\{\sum_{i\in\mathsf{S}}\Big[\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\Big(e^{-\alpha|i-j|}y_i^{(n)}y_j^{(n)}\exp(-\boldsymbol{h}_{i,j}^{(n)})\Big)\Big] - \mathbb{E}\Big[\sum_{i\in\mathsf{S}}\Big(\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\Big(e^{-\alpha|i-j|}y_iy_j\exp(-\boldsymbol{h}_{i,j}^{(n)})\Big)\Big)\Big]\Big\}, \tag{18}$$

and

$$\frac{\partial}{\partial\alpha}\log P(\{\boldsymbol{y}^{(n)}\}_{n=1}^N|\{\boldsymbol{x}^{(n)}\}_{n=1}^N) = -\frac{1}{\epsilon_\alpha^2}\alpha + \sum_{n=1}^N\Big\{\sum_{i\in\mathsf{S}}\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\Big(-\mathcal{I}(y_i^{(n)},y_j^{(n)},\boldsymbol{x}^{(n)},i,j)|i-j|\Big) + \mathbb{E}\Big[\sum_{i\in\mathsf{S}}\Big(\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\Big(\mathcal{I}(y_i,y_j,\boldsymbol{x}^{(n)},i,j)|i-j|\Big)\Big)\Big]\Big\}, \tag{19}$$

where the expectation is evaluated with distribution $p(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta})$. The conditional expectations of (17)-(19) require evaluation of the partition function $\mathcal{Z}$, a problem known to be NP-hard. As is common in the literature, this difficulty is avoided by estimating expectations through sampling. Although sampling methods such as *Markov chain Monte Carlo* (MCMC) can converge to the true distribution, this usually requires many iterations. Since the procedure must be repeated per gradient ascent step, these methods are impractical. On the other hand, approximations such as contrastive divergence minimization (which runs MCMC a limited number of times with specific starting points) have been shown successful for vision applications [40], [41]. We adopt these approximations for CRF learning.

This leverages the fact that, denoting any of the parameters $\boldsymbol{w},\boldsymbol{v}_{\mathbb{T}},\boldsymbol{v}_{\mathbb{S}},\alpha_{\mathbb{T}},\alpha_{\mathbb{S}}$ by $\boldsymbol{\theta}$, the partial gradients of (17)-(19) are

$$\frac{\partial}{\partial\boldsymbol{\theta}}\log p(\{\boldsymbol{y}^{(n)}\}_{n=1}^N|\{\boldsymbol{x}^{(n)}\}_{n=1}^N;\boldsymbol{\Theta}) = -\frac{1}{\epsilon_{\boldsymbol{\theta}}^2}\boldsymbol{\theta} + \sum_{n=1}^N\Big\{F_{\partial\boldsymbol{\theta}}(\boldsymbol{y}^{(n)},\boldsymbol{x}^{(n)}) - \mathbb{E}_{(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta})}\Big[F_{\partial\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{x}^{(n)})\Big]\Big\}, \tag{20}$$

where $F_{\partial\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{x})$ is the sum of the terms in the summations of (17), (18), or (19) that depend on $\boldsymbol{\theta}$. Contrastive divergence approximates the intractable conditional expectation $\mathbb{E}_{(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta})}\big[F_{\partial\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{x}^{(n)})\big]$ by $F_{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{y}},\boldsymbol{x}^{(n)})$, where $\hat{\boldsymbol{y}}$ is the "evil twin" of the ground-truth label field $\boldsymbol{y}^{(n)}$ [41]. $\hat{\boldsymbol{y}}$ is drawn by MCMC, using the inference procedure discussed in Section 5.2.2, the current parameter estimates, and the ground-truth labels $\boldsymbol{y}^{(n)}$ as a starting point.

Given the estimate of the partial gradients, the gradient ascent rule for parameter updates reduces to

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta\Big[\sum_{n=1}^N\Big(F_{\partial\boldsymbol{\theta}}(\boldsymbol{y}^{(n)},\boldsymbol{x}^{(n)}) - F_{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{y}}^{(n)},\boldsymbol{x}^{(n)})\Big) - \frac{1}{\epsilon_{\boldsymbol{\theta}}^2}\boldsymbol{\theta}\Big], \tag{21}$$

where η is a learning rate. In our implementation, this rule is initialized with $\boldsymbol{v}_{\mathbb{T}} = \boldsymbol{v}_{\mathbb{S}} = \mathbf{1}$ and $\alpha_{\mathbb{T}} = \alpha_{\mathbb{S}} = \mathbf{0}$. The initial value of $\boldsymbol{w}$ is learned, assuming a logistic regression model ($\boldsymbol{v}_{\mathbb{T}} = \boldsymbol{v}_{\mathbb{S}} = 0$ in (16)), with the procedure of [42].

5.2.2 Inference

The inference problem is to determine the most likely anomaly prediction $\boldsymbol{y}^\star$ for a query frame $\boldsymbol{x}^{(\tau)}$, given pre-

Fig. 5. Exemplar normal/ abnormal frames in Ped1 (top) and Ped2 (bottom). Anomalies (red boxes) include bikes, skaters, carts and wheelchairs.

vious predictions $\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1}$, and observations $\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau}$:

$$\begin{aligned}\boldsymbol{y}^{\star} &= \underset{\boldsymbol{y}}{\operatorname{argmax}}\ \log p(\boldsymbol{y}|\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1},\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau};\boldsymbol{\Theta})\\ &= \underset{\boldsymbol{y}}{\operatorname{argmax}} \sum_{i\in\mathsf{S}}\left[\mathcal{A}(y_i,\boldsymbol{x}^{(\tau)}) + \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\mathcal{I}(y_i,y_j,\boldsymbol{x},i,j)\right].\end{aligned} \tag{22}$$

Again, exact inference is intractable. We rely on Gibbs sampling to approximate the optimal prediction. This consists of drawing labels from the conditional distribution

$$\begin{aligned}&p(y_i|\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau},\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1},\boldsymbol{y}_{-i};\boldsymbol{\Theta})\\ &= \frac{p(y_i,\boldsymbol{y}_{-i}|\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau},\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1};\boldsymbol{\Theta})}{p(\boldsymbol{y}_{-i}|\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau},\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1};\boldsymbol{\Theta})}\\ &= \frac{1}{\mathcal{Z}_{-i}}\exp\left[F_{y_i}(\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau},\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1},\boldsymbol{y}_{-i},y_i;\boldsymbol{\Theta})\right],\end{aligned} \tag{23}$$

where $\boldsymbol{y}_{-i} = \{y_1,\cdots,y_{i-1},y_{i+1},\cdots,y_{|\mathsf{S}|}\}$, $F_{y_i}(\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau},\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1},\boldsymbol{y}_{-i},y_i;\boldsymbol{\Theta})$ is the sum of potential functions that depend on y_i (*i.e.*, its "Markov blanket"),

$$\begin{aligned}&F_{y_i}\left(\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau},\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1},\boldsymbol{y}_{-i},y_i;\boldsymbol{\Theta}\right) =\\ &\quad\mathcal{A}(y_i^{\tau},\boldsymbol{x}^{\tau}) + \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\mathcal{I}(y_i,y_j,\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau},i,j)\\ &\quad+\sum_{j:i\in\mathcal{N}_j}\frac{1}{|\mathcal{N}_j|}\mathcal{I}(y_j,y_i,\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau},j,i)\end{aligned} \tag{24}$$

and $\mathcal{Z}_{-i}$ the normalization constant

$$\mathcal{Z}_{-i} = \sum\nolimits_{y_i'}\exp\left[F_{y_i}\left(\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau},\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1},\boldsymbol{y}_{-i},y_i';\boldsymbol{\Theta}\right)\right]. \tag{25}$$

The procedure is detailed in Algorithm 2 and Algorithm 3, where we present the online CRF filter used to estimate the label field. During learning, the filter is initialized with the ground truth labels ($\boldsymbol{y}_0 = \boldsymbol{y}^{(\tau)}$). During testing, this initialization relies on the predictions of the single-site classifiers ($\boldsymbol{v}_{\mathbb{T}} = \boldsymbol{v}_{\mathbb{S}} = \boldsymbol{0}$). In our implementation, the filter is run for $N_s = 10$ iterations. The complete anomaly detection procedure is summarized in Algorithm 4.

6 Experiments

In this section, we introduce a new dataset and an experimental protocol for evaluation of anomaly detection in crowded environments and use it to evaluate the proposed anomaly detector.

TABLE 1
Composition of UCSD Anomaly Dataset

Scene	Nor.	Abnormal[a]					
		Bike	*Skater*	*Cart*	*Walk Across*	*Other*	Total[b]
Ped1	34	19/28	13/13	6/6	3/4	3/3	36/54
Ped2	16	11/19	3/3	1/1	0/0	0/0	12/23

[a] number of clips / number of anomaly instances.
[b] some clips contain more than one type of anomaly.

6.1 UCSD Pedestrian Anomaly Dataset

In the literature, anomaly detection has frequently been evaluated by visual inspection [19], [7], [3], or with coarse ground-truth, *e.g.*, frame-level annotation of abnormal events [4], [1]. This does not completely address the anomaly detection problem, where it is usually desired to *localize* anomalies in both *space and time*. To enable this, we introduce a dataset[1] of crowd scenes with precisely localized anomalies and metrics for the evaluation of their detection. The dataset consists of video clips recorded with a stationary camera mounted at an elevation, overlooking pedestrian walkways on the UCSD campus. The crowd density in the walkways is variable, ranging from sparse to very crowded. In the normal setting, the video contains only pedestrians. Abnormal events are due to either 1) the circulation of non-pedestrian entities in the walkways, or 2) anomalous pedestrian motion patterns. Commonly occurring anomalies include bikers, skaters, small carts, and people walking across a walkway or in the surrounding grass. A few instances of wheelchairs are also recorded. All abnormalities occur naturally, *i.e.*, they were not staged or synthesized for dataset collection.

The dataset is organized into two subsets, corresponding to the two scenes of Fig. 5. The first, denoted "Ped1", contains clips of 158×238 pixels, which depict groups of people walking towards and away from the camera, and some amount of perspective distortion. The second, denoted "Ped2", has spatial resolution of 240×360 pixels and depicts a scene where most pedestrians move horizontally. The video footage of each scene is sliced into clips of 120-200 frames. A number of these (34 in Ped1 and 16 in Ped2) are to be used as training set for the condition of normalcy. The test set contains clips (36 for Ped1 and 12 for Ped2) with both normal (around 5500) and abnormal (around 3400) frames. The abnormalities of each set are summarized in Table 1.

Frame-level ground-truth annotation, indicating whether anomalies occur within each frame, and manually collected pixel-level binary anomaly masks, which identify the pixels containing anomalies, are available per test clip. We note that this includes ground-truth on Ped1 contributed by [9], and supersedes the ground truth available on a earlier version of this work [43]. We denote the current ground-truth by "full annotation", and the previous one by "partial annotation". Unless otherwise noted, the results of the subsequent sections correspond to the full annotation.

[1] available from http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

6.2 Evaluation Methodology

Two criteria are used to evaluate anomaly detection accuracy: a *frame-level criterion* and a *pixel-level criterion*. Both are based on true- and false-positive rates, denoting "an anomalous event" as "positive" and "the absence of anomalous events" as "negative". A frame containing anomalies is denoted a positive, otherwise a negative. The true- and false-positive under the two criteria are:

- **Frame-level criterion** An algorithm predicts which *frames* contain anomalous events. This is compared to the clip's frame-level ground-truth anomaly annotations to determine the number of true- and false-positive frames.
- **Pixel-level criterion** An algorithm predicts which *pixels* are related to anomalous events. This is compared to the pixel-level ground-truth anomaly annotation to determine the number of true positive and false positive frames: a frame is a true-positive if 1) it is positive and 2) at least 40% of its anomalous pixels are identified; a frame is a false-positive if it is negative and any of its pixels are predicated as anomalous.

The two measures are combined into a receiver operating characteristic (ROC) curve of *true positive rate* (TPR) *v.s. false positive rate* (FPR):

$$\text{TPR} = \frac{\text{\# of true positive frame}}{\text{\# of positive frame}}, \ \text{FPR} = \frac{\text{\# of false positive frame}}{\text{\# of negative frame}}.$$

Performance is also summarized by the *equal error rate* (EER), the ratio of misclassified frames at which FPR $= 1-$TPR, for the frame-level criterion, or *rate of detection* (RD), *i.e.*,1-EER, for the pixel-level criterion.

Note that, although widely-used in the literature, the frame-level criterion only measures temporal localization accuracy. This enables errors due to "lucky co-occurrences" of prediction errors and true abnormalities. For example, it assigns a perfect score to an algorithm that identifies a single anomaly at a random location of a frame with anomalies. The pixel-level criterion is much stricter and more rigorous. By evaluating both the temporal and spatial accuracy of the anomaly predictions it rules out these "lucky co-occurrences". We believe that the pixel-level criterion should be the predominant criterion for evaluation of anomaly detection algorithms.

6.3 Experimental Set-up

Unless otherwise noted, observation sites are a video sublattice with spatial interval of 4 pixels and temporal interval of 5 frames. Temporal anomaly maps rely on patches of $13 \times 13 \times 15$ pixels. The temporal extent of 15 frames provides a reasonable compromise between the ability to detect anomalies and the delay (1.5s) and storage (15 video frames) required for anomaly detection. To minimize computation, patches of variance smaller than 500 are discarded[2]. Temporal H-MDT models are learned from fine to coarse scale. At the finer scale, there are 6×10 windows $\mathcal{R}_i^1$ on Ped1 (8×11 for Ped2), each covering a 41×41 pixel area and overlapping by 25% with each of its 4-neighbors. An MDT of 5 components is learned per window. At coarser spatial scales, an MDT is estimated from the MDTs of the four regions that it covers at the immediately finer resolution. Each estimated MDT has one more component than its ancestor MDTs. Overall, there are 10-scales in Ped1 and 11 in Ped2. Spatial anomaly maps use a 31×31 center window and surround windows of size equivalent to $\mathcal{R}_i^s$. For segmentation, $7 \times 7 \times 10$ patches are extracted from the 40 frames surrounding that under analysis. There are 5 DT components at all levels of the spatial hierarchy. Both temporal and spatial MDTs have a 8-dimensional state space. The sensitivity of the proposed detector to some of these parameters is discussed in Appendix C.2.

[2]This variance threshold is quite conservative, only eliminating regions of very little motion. For the datasets used in our experiments, this has not lead to the elimination of any objects from further consideration. In other contexts, *e.g.*, scenes where objects are static for periods of time, this could happen. In this case, the threshold should be set to zero.

TABLE 2
Performance on UCSD Anomaly Dataset

Descriptor	Criterion			
	EER		RD	
	Ped1	Ped2	Ped1*	Ped2
MDT-temp.	**22.9%**	**27.9%**	**59.3%** (48.2%)	56.8%
MDT-spat.	43.8%	28.7%	50.8% (**54.2%**)	**63.4%**
MPPCA	35.6%	35.8%	23.2% (27.0%)	22.4%
force flow	36.5%	35.0%	40.9% (38.8%)	27.6%
LMH	38.9%	45.8%	32.6% (35.3%)	22.4%

* numbers outside/inside parentheses are results by full/partial annotation (same for the rest of the paper).

6.4 Descriptor Comparison

The first experiment evaluated the benefits of MDT-based over optical flow descriptors. The optical flow descriptors considered were the local motion histogram (LMH) of [19], the force flow descriptor of [4] and the mixture of optical flow models (MPPCA) of [3]. LMH uses statistics of local motion, and is representative of traditional background subtraction representations, force flow is a descriptor for spatial anomaly detection, and MPPCA a temporal anomaly detector. For the MDT, only the anomaly maps of finest temporal and coarsest spatial scale were considered here. Since the goal was to compare descriptors, the high-level components of the models in which they were proposed, *e.g.*, the LDA of [4], the MRF of [3] and the proposed CRF, were not used. Instead, anomaly predictions were smoothed with a simple $20 \times 20 \times 10$ Gaussian filter. Anomaly predictions were generated by thresholding the filtered anomaly maps, and ROC curves by varying thresholds.

The performance of the different descriptors, under both the frame-level (EER) and pixel-level (RD) criteria (using both full and partial annotation in Ped1), is summarized in Table 2. The corresponding ROC curves are presented in Appendix C.1 (Fig. 13). Examples of detected anomalies are shown in Fig. 6. Under the frame-level criterion, temporal MDT has the best performance in both scenes. Spatial MDT performs worse than others in Ped1 but ranks

Fig. 6. Anomaly predictions of temporal MDT, spatial MDT, MPPCA, force flow and LMH (from left to right). Red regions are abnormal pixels. All predictions generated with thresholds such that the different approaches have similar FPR under frame-level protocol (these settings apply to all the subsequent figures unless otherwise stated).

Fig. 7. Examples of anomaly localization with Gaussian smoothing (in blue) and CRF filter (red). The latter predicts more accurately the spatio-temporal support of anomalies in crowded regions, where occlusion is prevalent.

second in Ped2. However, for the more precise pixel-level criterion, spatial MDT is the top or second best performer. In this case, both MDTs significantly outperform all optical flow descriptors. The gap between corresponding competitors (*e.g.*, temporal MDT *v.s.* MPPCA or LMH, spatial MDT *v.s.* force flow) is of at least 10% RD. These results show that there is a definite benefit to the joint representation of appearance and dynamics of the MDT.

This is not totally surprising, given the limitations of optical flow. First, the brightness constancy assumption is easily violated in crowded scenes, where stochastic motion and occlusions prevail. Second, optical flow measures instantaneous displacement, while the DT is a smooth motion representation with extended temporal support. Finally, while optical flow is a bandpass measure, which eliminates most of the appearance information, the DT models both appearance and dynamics. The last two properties are particularly important for crowded scenes, where objects occlude and interact in complicated manners. Overall, although optical flow can signal fast moving anomalous subjects, it leads to too many false positives in regions of complex motion, occlusion, *etc*. More interesting is the lack of advantage for either spatial or temporal anomaly detection, both among MDT maps and prior techniques (no clear advantage to either force flow or MPPCA). In fact, as shown in Fig. 6, temporal and spatial anomalies tend to be different objects. This suggests the combination of the two strategies.

TABLE 3
Performance on UCSD Anomaly Dataset

Criterion	Scene	Method		
		S-MDT w/ smoothing	H-MDT w/ smoothing	H-MDT w/ CRF-filtering
Frame-Level (EER)	Ped1	21.3%	21.6%	**17.8%**
	Ped2	23.9%	22.3%	**18.5%**
Pixel-Level (RD)	Ped1	59.7% (56.7%)	71.2% (60.1%)	**74.5% (64.8%)**
	Ped2	65.0%	69.6%	**70.1%**

6.5 Scale and Globally Consistent Prediction

We next investigated the benefits of information fusion across space and scale, with the proposed CRF. We started with a *single-scale* description (S-MDT), using only the anomaly maps at finest temporal and coarsest spatial scales, *i.e.*, a 3-D feature per site. We next considered a *multi-scale* description, using the whole H-MDT. In both cases, inference was performed with logistic regression, *i.e.*, the interaction term of (16) turned off, and the Gaussian filter of the previous section. In each trial, the logistic classifier was trained by Newton's method [42]. Finally, we considered the full blown CRF, denoted *CRF filter*. The dimensions of the spatial and temporal CRF neighborhoods were set to $|\mathcal{N}^{\mathbb{S}}| = 6$, $|\mathcal{N}^{\mathbb{T}}| = 3$. ROC curves were generated by varying the threshold for prediction.

Table 3 presents a comparison of the three approaches. The corresponding ROC curves are shown in Appendix C.1 (Fig. 14). Under the pixel-level criterion, the multi-scale maps have higher accuracy than their single-scale counterparts, demonstrating the benefits of modeling anomalies in scale space (improvement of RD by as much as 11%). The CRF filter further improves performance (improvement of RD by as much as 3%), demonstrating the gains of globally consistent inference. As shown in Fig. 7, the visual improvements are even more substantial[3]. Simple filtering does not take into account interactions

[3]more results at http://www.svcl.ucsd.edu/projects/anomaly/results.html

TABLE 4
Performance (RD/seconds per frame) by Pixel-level Criterion on UCSD Anomaly Dataset

	MDT (temporal)	MDT (spatial)	H-MDT-temp. w/ CRF-filtering	H-MDT-spat. w/ CRF-filtering	H-MDT w/ CRF-filtering$^{\diamond}$	sparse reconstruction [8]$^{\natural}$	Bayesian video parsing [9]$^{\#}$
Ped1	- / 0.52	- / 0.57	65% (52%) / 0.61	57% (58%) / 0.65	75% (65%) / 1.11	- (46%) / 3.8	**77% (68%) / 5∼10**
Ped2	- / 0.64	- / 0.69	55% / 0.76	68% / 0.80	**70% / 1.38**	-	-

implementation: $^{\diamond}$C/2.8GHz CPU/2GB RAM; $^{\natural}$C++&Matlab (feature extraction/model inference)/2.6GHz CPU/2GB RAM; $^{\#}$Matlab/dual-core 2.7GHz CPU/8GB RAM.

Fig. 8. ROC curves of pixel-level criterion on Ped1.

between neighboring sites and smooths the anomaly maps uniformly. On the other hand, the CRF adapts the degree of smoothing to the spatiotemporal structure of the anomalies, increasing the precision of anomaly localization. Note how, in Fig. 7, the CRF-filter successfully excludes occluded but normally behaving pedestrians from anomaly regions. These improvements are not always captured by the frame-level criterion. In fact, there is little EER difference between S-MDT and H-MDT. The inconsistency between frame- and pixel-level results in Tables 2-3 shows that the former is not a good measure of anomaly detection performance. Henceforth, only the pixel-level criterion is used in the remaining experiments on this dataset.

6.6 Anomaly Detection Performance

We next evaluated the performance of the complete anomaly detector. For this, we selected two detectors from the recent literature, with state-of-the-art performance for temporal [8] and combined spatial and temporal anomaly detection [9]. The RD of the various methods is summarized in Table 4, for both partial and full annotation. The corresponding ROC curves are shown in Fig. 8. Table 4 also presents the processing time per video frame of each method. Missing entries indicate unavailable results for the particular dataset and/or annotation type. A discussion of the detection errors made by the detector is given in Appendix C.3.

On Ped1, the temporal component of the proposed detector substantially outperforms the temporal detector of [8]. A multiple-scale temporal anomaly map with CRF filtering increases the 46% RD[4] of [8] to 52%. A similar implementation of the spatial anomaly detector (a multiple-scale map plus CRF filtering) achieves 58%. Combining both maps and multiple spatial scales, further improves the RD to 65%. Computationally, the proposed detector is also much more efficient. For implementations on similar hardware (see footnotes of Table 4), it requires 1.11 s/frame, as compared to the 3.8 s/frame reported for [8].

Like the proposed detector, the Bayesian video parsing (BVP) of [9] combines spatial and temporal anomaly detection, using a more complex video representation, parsing of the video to extract all the objects in the scene, a support vector machine classifier for detection of temporal anomalies, a graphical model with 7 nodes per site (and multiple nonparametric models for location, scale, and velocity) for detection of spatial anomalies, and occlusion reasoning. This is an elegant solution, which achieves slightly better RD than the proposed detector (2% for full and 3% for partial annotation), but at substantially higher computational cost (5 to 10 times slower). We believe that, when both accuracy and computation are considered, the proposed detector is a more effective solution. However, these results suggest that gains could be achieved by expanding the proposed CRF, as [9] trades a much simpler representation of video dynamics (optical flow *v.s.* MDT) for more sophisticated inference. It would be interesting to consider CRF extensions with some of the properties of the graphical model of [9], namely explicit occlusion reasoning. This is left for subsequent research.

6.7 Role of Context in Anomaly Judgments

We next investigated the impact of normalcy context in anomaly judgements. For temporal anomalies, context is determined by the subregion size: as the latter increases, temporal models become more global. Fig. 9 shows that the scale of normalcy context significantly impacts anomaly scores. For example, the two cyclists on the left-most columns of the figure are missed at small scales but detected by the more global models. On the other hand, a leftwards heading pedestrian in the third column has high anomaly score at the finest scale but is not anomalous in larger contexts. In summary, no single context is effective for all scenes. Due to the stochastic arrangements of people within crowds, two crowds of the same size can require different context sizes. In general, the optimal size depends on the crowd configuration and the anomalous event.

A similar observation holds for spatial anomalies, where context is set by the size of the surround window. For example, in the fourth column of Fig. 9, the subject walking on the grass is very salient when compared to her

[4]these numbers refer to partial annotation, the only available for [8].

Fig. 9. Impact of context on anomaly maps. First three columns: temporal anomalies, cell coverage at different HMDT layers shown in blue. Last two columns: spatial anomalies, example center (surround) windows shown in blue (light yellow).

immediate neighbors, and anomaly detection benefits from a narrower context. For larger contexts, she becomes less unique than a man that walks in the direction opposite to his neighbors. On the other hand, the cart and bike of the last column only pop out when the surround window is large enough to cover some pedestrians. In summary, anomalies depend strongly on scene context, and this dependence can vary substantially from scene to scene. It is, thus, important to fuse anomaly information across spatial scales.

6.8 Performance on Other Benchmark Datasets

The detection of anomalous events in crowded scenes can be evaluated in a few datasets other than UCSD. These have various limitations in terms of size, saliency of the anomalies, evaluation criteria, *etc*. They are discussed in this section where, for completeness, we also present the results of the proposed anomaly detector.

UMN: The UMN dataset[5] contains three escape scenes. Normal events depict individuals wandering around or organized in groups. Abnormal events depict a crowd escaping in panic. Each scene contains several normal-abnormal events (*e.g.*, seconds of normalcy followed by a short abnormal event). The main limitations of this dataset are that 1) it is relatively small (scenes 1, 2 and 3 contain 2, 6, and 3 anomaly instances), 2) has no pixel-level ground-truth, 3) the anomalies are staged, and 4) produce very salient changes in the average motion intensity of the scene. In result, several methods achieve near perfect detection.

The proposed detector was based on 3×3 sub-regions of size 180×180 at the finest spatial scale and a 3-scale anomaly map for both the temporal and spatial components. One normal-abnormal instance of each scene was used to train the temporal normalcy model and CRF filter, and the remaining instances for testing. A comparison to previous results in the literature, under the frame-level criterion, is presented in Table 5 and Fig. 11. Due to the salient motion discontinuities, the temporal component (99.2% AUC) substantially outperforms the spatial component (97.9%). Nevertheless, the complete detector achieves the best performance (99.5%). This is nearly perfect, and comparable to the previous best results in the literature.

Subway: The Subway dataset [19] consists of two sequences recorded from the entrance (1h and 36 min, 144249 frames) and exit (43 min, 64900 frames) of a subway station. Normal behaviors include people entering and exiting the station; abnormal consist of people moving in the wrong direction (exiting the entrance or entering the exit) or avoiding payment. The main limitations of this dataset are 1) reduced number of anomalies, and 2) predictable spatial localization (entrance and exit regions). The original 512×384 frames were down sampled to 320×240, and 2×3 sub-regions of size 90×90, covering either the entrance or exit regions, were used at the finest spatial scale. A 3-scale anomaly map was computed for both spatial and temporal anomalies. Video patches were of size $15 \times 15 \times 15$, and 10 min of video from each sequence were used to train the temporal normalcy model and CRF filters, while the remaining video was used for testing. Table 5 and Fig. 11 present a comparison of the proposed detector against recently published results on this dataset. Again, the temporal component outperforms its spatial counterpart, but the best performance is obtained by combination of both temporal and spatial anomaly maps (H-MDT CRF). This achieves the best result among all methods, outperforming the sparse reconstruction of [8] and the local statistical aggregates of [23]. Note that, for this dataset, the gains in both AUC and EER are substantial.

U-turn: The U-turn dataset [5] consists of one video sequence (roughly 6000 frames of size 360×240) recorded by a static camera overlooking the traffic at a road intersection. The video is split into two clips of equal length for cross-validation and anomalies consist of illegal vehicle motion

[5] http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi

TABLE 5
Anomaly Detection Performance in AUC/ERR (%).

method	UMN	Subway		U-turn
		entrance	exit	
chaotic invariants [1]	99.4/5.3	-/-	-/-	-/-
social force [4]	94.9/12.6	-/-	-/-	-/-
sparse [8]	**99.6**/**2.8**	80.2/26.4	83.3/24.4	-/-
local stat. aggr. [23]	99.5/3.4	88.4/17.9	-/-	94.7/**8.3**
H-MDT-spat. CRF	97.9/7.8	68.2/37.0	67.0/34.1	83.9/23.4
H-MDT-temp. CRF	99.2/5.3	88.9/18.6	87.5/17.9	92.9/14.6
H-MDT CRF	99.5/3.7	**89.7**/**16.4**	**90.8**/**16.7**	**95.2**/12.8

Fig. 10. Anomalies detected by H-MDT CRF on UMN (left), Subway (center), and U-turn (right) datasets.

Fig. 11. ROC curves of frame-level criterion on UMN (left), Subway (center), and U-turn (right) datasets.

at the intersection. The main limitations of this dataset are 1) the limited size, 2) absence of pixel-level ground-truth, and 3) sparseness of the scenes. The latter enables the use of object-based operations, *e.g.*, tracking, and analysis of object trajectories [5], which we do not exploit.

For temporal anomaly detection, MDTs were learned using $20 \times 20 \times 30$ patches from 3×4 sub-regions covering the intersection. This was the finest level of a 3-scale hierarchical model. For spatial anomaly detection, segmentation was computed with a 5-component MDT learnt from $15 \times 15 \times 30$ patches extracted from 45 consecutive frames. An observation lattice of step $15 \times 15 \times 10$ was used to evaluate anomaly scores, and the neighborhood size of the CRF filter was 2. The performance of the detector is summarized in Table 5 and Fig. 11. Due to the sparsity of the scenes (not enough spatial context around cars making illegal turns to establish them as anomalous) the performance of the spatial anomaly detector is quite weak. However, the combination of the spatial and temporal anomaly maps again outperforms the temporal channel, achieving the best performance. Overall, the proposed detector has the best AUC on this dataset. Examples of detected anomalies, for this and the other two datasets, are shown in Fig. 10.

7 Conclusion

In this work, we proposed an anomaly detector that spans time, space, and spatial scale, using a joint representation of video appearance and dynamics and globally consistent inference. For this, we modeled crowded scenes with a hierarchy of MDT models, equated temporal anomalies to background subtraction, spatial anomalies to discriminant saliency, and integrated anomaly scores across time, space, and scale with a CRF. It was shown that the MDT representation substantially outperforms classical optical flow descriptors, that spatial and temporal anomaly detection are complementary processes, that there is a benefit to defining anomalies with respect to various normalcy contexts, *i.e.*, in anomaly scale space, and that it is important to guarantee globally consistent inference across space, time and scale. We have also introduced a challenging anomaly detection dataset, composed of complex scenes of pedestrian crowds, involving stochastic motion, complex occlusions and object interactions. This dataset provides both frame-level and pixel-level ground truth, and a protocol for the evaluation of anomaly detection algorithms. The proposed anomaly detector was shown effective on both this and a number of previous datasets. When compared to previous methods, it outperformed various state-of-the-art approaches, either in absolute performance or in terms of the trade-off between anomaly detection accuracy and complexity.

References

[1] S. Wu, B. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," *IEEE Conf. Computer Vision and Pattern Recognition*, 2010. 1, 2, 8, 13

[2] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. 1, 2

[3] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates," *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. 1, 2, 4, 8, 9

[4] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. 1, 2, 5, 8, 9, 13

[5] Y. Benezeth, P. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurences," *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. 1, 2, 12, 13

[6] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," *IEEE Conf. Computer Vision and Pattern Recognition*, 2008. 1, 2

[7] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, p. 893, 2008. 1, 2, 8

[8] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," *IEEE Conf. Computer Vision and Pattern Recognition*, 2011. 1, 2, 11, 12, 13

[9] B. Antić and B. Ommer, "Video parsing for abnormality detection," *IEEE Int'l Conf. Computer Vision*, 2011. 1, 2, 8, 11

[10] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, p. 15, 2009. 1

[11] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto, "Dynamic textures," *Int'l J. Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003. 1, 4

[12] A. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 909–926, 2008. 1, 3, 4

[13] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: Computational principles, biological plausibility, and implications for neurophysiology and psychophysics," *Neural Computation*, vol. 21, no. 1, pp. 239–271, 2009. 1, 4

[14] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000. 2

[15] T. Zhang, H. Lu, and S. Li, "Learning semantic scene models by object classification and trajectory clustering," *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. 2

[16] N. Siebel and S. Maybank, "Fusion of multiple tracking algorithms for robust people tracking," *European Conf. Computer Vision*, 2006. 2

[17] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," *IEEE Conf. Computer Vision and Pattern Recognition*, 2011. 2

[18] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011. 2

[19] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, p. 555, 2008. 2, 8, 9, 12

[20] B. Zhao, L. Fei-Fei, and E. Xing, "Online detection of unusual events in videos via dynamic sparse coding," *IEEE Conf. Computer Vision and Pattern Recognition*, 2011. 2

[21] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, pp. 4282–4286, 1995. 2

[22] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int'l J. Computer Vision*, vol. 74, no. 1, pp. 17–31, 2007. 2

[23] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," *CVPR*, 2012. 2, 12, 13

[24] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman, "Detection and explanation of anomalous activities: Representing activities as bags of event n-grams," *IEEE Conf. Computer Vision and Pattern Recognition*, 2005. 2

[25] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted hmms for unusual event detection," *IEEE Conf. Computer Vision and Pattern Recognition*, 2005. 2

[26] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," *IEEE Conf. Computer Vision and Pattern Recognition*, 1999. 3, 4

[27] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998. 3, 4

[28] R. Shumway and D. Stoffer, "An approach to time series smoothing and forecasting using the em algorithm," *J. Time Series Analysis*, vol. 3, no. 4, pp. 253–264, 1982. 4, 15, 16

[29] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural computation*, vol. 11, no. 2, pp. 305–345, 1999. 4, 16

[30] S. Kullback, *Information Theory and Statistics*. Dover Publications, 1968. 4

[31] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp. 1–18, 2008. 4

[32] V. Mahadevan and N. Vasconcelos, "Background subtraction in highly dynamic scenes," *IEEE Conf. Computer Vision and Pattern Recognition*, 2008. 5

[33] A. Chan and N. Vasconcelos, "Probabilistic kernels for the classification of auto-regressive visual processes," *IEEE Conf. Computer Vision and Pattern Recognition*, 2005. 5

[34] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," *IEEE Inte'l Conf. Acoustics, Speech and Signal Processing*, 2007. 5

[35] A. B. Chan and N. Vasconcelos, "Efficient computation of the kl divergence between dynamic textures," Tech. Rep. SVCL-TR-2004-02, 2004. 5

[36] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *IEEE Conf. Computer Vision and Pattern Recognition*, 2006. 6

[37] A. Chan, E. Coviello, and G. Lanckriet, "Clustering dynamic textures with the hierarchical em algorithm," *IEEE Conf. Computer Vision and Pattern Recognition*, 2010. 6

[38] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proc. 18th Int'l Conf. Machine Learning*, 2001. 6

[39] S. Kumar and M. Hebert, "Discriminative fields for modeling spatial dependencies in natural images," *Advances in Neural Information Processing Systems*, 2004. 6

[40] X. He, R. Zemel, and M. Carreira-Perpinán, "Multiscale conditional random fields for image labeling," *IEEE Conf. Computer Vision and Pattern Recognition*, 2004. 7

[41] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, pp. 1771–1800, 2002. 7

[42] T. Minka, "A comparison of numerical optimizers for logistic regression," Microsoft Research, Tech. Rep., 2003. 7, 10

[43] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," *IEEE Conf. Computer Vision and Pattern Recognition*, 2010. 8

[44] T. P. Kah-Kay Sung, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 39–51, 1998. 18

Weixin Li received his bachelor degree from Department of Automation at Tsinghua University, Beijing, China in 2008, the M.Sc. in electrical engineering at University of California, San Diego in 2011, and is now a Ph.D. candidate. His research interests mainly consist of computational vision and machine learning, with specific focus on visual analysis of human behavior, activity and event, and models with latent variables and their applications.

Vijay Mahadevan received the B.Tech degree from Indian Institute of Technology, Madras, in 2002, and M.S. from Rensselaer Polytechnic Institute, Troy, NY, in 2003 and Ph.D from University of California, San Diego in 2011, all in electrical engineering. From 2004-2006, he was with the Multimedia group at Qualcomm Inc., San Diego, CA. He is currently with Yahoo! Labs, Bangalore. His interests are in computer vision and machine learning.

Nuno Vasconcelos received the licenciatura in electrical engineering and computer science from the Universidade do Porto, Portugal, and the MS and PhD degrees from the Massachusetts Institute of Technology. He is a Professor in the Electrical and Computer Engineering Department at the University of California, San Diego, where he heads the Statistical Visual Computing Laboratory. He has received a NSF CAREER award, a Hellman Fellowship, and has authored more than 150 peer-reviewed publications. He is a senior member of the IEEE.

APPENDIX A
MIXTURE OF DYNAMIC TEXTURES: LEARNING AND INFERENCE

A.1 The EM Algorithm for Learning Mixture of Dynamic Textures

Given a set of N independent and identically distributed (*i.i.d.*) samples $\mathcal{D}_i = \{\boldsymbol{x}^{(i)}\}_{i=1}^N$ (incomplete data), maximum likelihood estimates (MLE) of the parameters of an MDT $p(\boldsymbol{x};\boldsymbol{\Theta})$ of K components

$$\begin{aligned}\boldsymbol{\Theta}^* &= \operatorname*{argmax}_{\boldsymbol{\Theta}} p(\mathcal{D}_i;\boldsymbol{\Theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\Theta}} \sum_{i=1}^N \log p(\boldsymbol{x}^{(i)};\boldsymbol{\Theta})\end{aligned} \tag{26}$$

are learned with the EM algorithm. There are two types of hidden variables in the MDT model: 1) the hidden state sequence $\boldsymbol{s}$ and 2) the assignment Z of each sequence to a mixture component. EM iterates between two steps

E-Step:

$$\mathcal{Q}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k)}) = \mathbb{E}_{\mathcal{D}_h|\mathcal{D}_i;\boldsymbol{\Theta}^{(k)}}\Big[\log p(\mathcal{D}_c;\boldsymbol{\Theta})\Big], \tag{27}$$

M-Step:

$$\boldsymbol{\Theta}^{(k+1)} = \operatorname{argmax}_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k)}), \tag{28}$$

where the hidden data $\mathcal{D}_h$ consists of the hidden variables $\{\boldsymbol{s}^{(i)}\}_{i=1}^N$ and $\{z^{(i)}\}_{i=1}^N$, and the complete data $\mathcal{D}_c = \mathcal{D}_i \cup \mathcal{D}_h$. The assignment variable $z^{(i)}$ is represented by a vector $\mathbf{z}_i \in \{0,1\}^K$, such that $\mathbf{z}_{i,j} = 1$ if and only if $z^{(i)} = j$.

A.1.1 E-step

The log-likelihood of the complete data is (up to a constant)

$$\begin{aligned}\log p(\mathcal{D}_c;\boldsymbol{\Theta}) = &\sum_{i,j}\mathbf{z}_{i,j}\log\pi_j - \frac{1}{2}\sum_{i,j}\mathbf{z}_{i,j}\Bigg\{\log|S_j| \\ &+ \operatorname{Tr}\Big[S_j^{-1}\Big(P_{1,1}^{(i)} - 2\boldsymbol{s}_1^{(i)}\boldsymbol{\mu}_j^T + \boldsymbol{\mu}_j\boldsymbol{\mu}_j^T\Big)\Big]\Bigg\} \\ &- \frac{1}{2}\sum_{i,j}\mathbf{z}_{i,j}\Bigg\{(\tau-1)\log|Q_j| + \sum_{t=2}^{\tau}\operatorname{Tr}\Big[Q_j^{-1}\Big(P_{t,t}^{(i)} \\ &- 2P_{t,t-1}^{(i)}A_j^T + A_jP_{t-1,t-1}^{(i)}A_j^T\Big)\Big]\Bigg\} \\ &- \frac{1}{2}\sum_{i,j}\mathbf{z}_{i,j}\Bigg\{\tau\log|R_j| + \sum_{t=1}^{\tau}\operatorname{Tr}\Big[R_j^{-1}\Big(\boldsymbol{x}_t^{(i)}\boldsymbol{x}_t^{(i)T} \\ &- 2\boldsymbol{x}_t^{(i)}(\boldsymbol{s}_t^{(i)})^TC_j^T + C_jP_{t,t}^{(i)}C_j^T\Big)\Big]\Bigg\},\end{aligned} \tag{29}$$

where $P_{t,r}^{(i)} = \boldsymbol{s}_t^{(i)}(\boldsymbol{s}_r^{(i)})^T$. The $\mathcal{Q}$-function is then

$$\begin{aligned}&\mathcal{Q}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)}) = \\ &\sum_j \hat{N}_j\Big[\log\pi_j - \frac{1}{2}\log|S_j| - \frac{\tau-1}{2}\log|Q_j| - \frac{\tau}{2}\log|R_j|\Big] \\ &- \frac{1}{2}\sum_j\operatorname{Tr}\Big[S_j^{-1}\Big(\eta_j - \xi_j\boldsymbol{\mu}_j^T - \boldsymbol{\mu}_j\xi_j^T + \hat{N}_j\boldsymbol{\mu}_j\boldsymbol{\mu}_j^T\Big)\Big] \\ &- \frac{1}{2}\sum_j\operatorname{Tr}\Big[Q_j^{-1}\Big(\varphi_j - \Psi_jA_j^T - A_j\Psi_j^T + A_j\phi_jA_j^T\Big)\Big] \\ &- \frac{1}{2}\sum_j\operatorname{Tr}\Big[R_j^{-1}\Big(\Lambda_j - \Gamma_jC_j^T - C_j\Gamma_j^T + C_j\Phi_jC_j^T\Big)\Big],\end{aligned} \tag{30}$$

with

$$\begin{aligned}&\hat{N}_j = \textstyle\sum_i\hat{\mathbf{z}}_{i,j} && \Phi_j = \textstyle\sum_i\hat{\mathbf{z}}_{i,j}\sum_{t=1}^{\tau}\hat{P}_{t,t|j}^{(i)} \\ &\xi_j = \textstyle\sum_i\hat{\mathbf{z}}_{i,j}\hat{\boldsymbol{s}}_{1|j}^{(i)} && \varphi_j = \textstyle\sum_i\hat{\mathbf{z}}_{i,j}\sum_{t=2}^{\tau}\hat{P}_{t,t|j}^{(i)} \\ &\eta_j = \textstyle\sum_i\hat{\mathbf{z}}_{i,j}\hat{P}_{1,1|j}^{(i)} && \phi_j = \textstyle\sum_i\hat{\mathbf{z}}_{i,j}\sum_{t=2}^{\tau}\hat{P}_{t-1,t-1|j}^{(i)} \\ &\Psi_j = \textstyle\sum_i\hat{\mathbf{z}}_{i,j}\sum_{t=2}^{\tau}\hat{P}_{t,t-1|j}^{(i)} \\ &\Lambda_j = \textstyle\sum_i\hat{\mathbf{z}}_{i,j}\sum_{t=1}^{\tau}\boldsymbol{x}_t^{(i)}(\boldsymbol{x}_t^{(i)})^T \\ &\Gamma_j = \textstyle\sum_i\hat{\mathbf{z}}_{i,j}\sum_{t=1}^{\tau}\boldsymbol{x}_t^{(i)}(\hat{\boldsymbol{s}}_{t|j}^{(i)})^T,\end{aligned} \tag{31}$$

where

$$\begin{aligned}\hat{\mathbf{z}}_{i,j} &= p(z^{(i)} = j|\boldsymbol{x}^{(i)};\boldsymbol{\Theta}^{(k-1)}) \\ &= \frac{\pi_jp(\boldsymbol{x}^{(i)}|z^{(i)}=j;\boldsymbol{\Theta}^{(k-1)})}{\sum_{k=1}^K\pi_kp(\boldsymbol{x}^{(i)}|z^{(i)}=k;\boldsymbol{\Theta}^{(k-1)})}\end{aligned} \tag{32}$$

are the posterior assignment probabilities, which can be calculated as discussed in Appendix A.3; and the statistics are aggregates of the expectations

$$\hat{\boldsymbol{s}}_{t|j}^{(i)} = \mathbb{E}_{\boldsymbol{s}^{(i)}|\boldsymbol{x}^{(i)},z^{(i)}=j;\boldsymbol{\Theta}^{(k-1)}}\Big(\boldsymbol{s}_t^{(i)}\Big), \tag{33}$$

$$\hat{P}_{t,r|j}^{(i)} = \mathbb{E}_{\boldsymbol{s}^{(i)}|\boldsymbol{x}^{(i)},z^{(i)}=j;\boldsymbol{\Theta}^{(k-1)}}\Big(P_{t,r}^{(i)}\Big). \tag{34}$$

The conditional expectation of (33) and (34) can be efficiently computed via the Kalman smoothing filter [28] (see Appendix A.2), *e.g.*, $\hat{P}_{t,r|j}^{(i)} = \hat{V}_{t,r|j}^{(i)} + \hat{\boldsymbol{s}}_{t|j}^{(i)}(\hat{\boldsymbol{s}}_{r|j}^{(i)})^T$, where

$$\hat{V}_{t,r|j}^{(i)} = \operatorname{cov}_{\boldsymbol{s}^{(i)}|\boldsymbol{x}^{(i)},z^{(i)}=j;\boldsymbol{\Theta}^{(k-1)}}(\boldsymbol{s}_t^{(i)},\boldsymbol{s}_r^{(i)}) \tag{35}$$

for $r = t, t-1$.

A.1.2 M-step

In the M-step, the new parameter estimate $\boldsymbol{\Theta}^{(k)}$ is computed by maximizing the $\mathcal{Q}$-function: $\boldsymbol{\Theta}^{(k)} = \operatorname{argmax}_{\boldsymbol{\Theta}}\mathcal{Q}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)})$. This leads to the updates

$$\begin{aligned}&A_j^{(k)} = \Psi_j(\phi_j)^{-1}, Q_j^{(k)} = \tfrac{1}{(\tau-1)\hat{N}_j}\Big(\varphi_j - A_j^{(k)}\Psi_j^T\Big), \\ &C_j^{(k)} = \Gamma_j(\Phi_j)^{-1}, R_j^{(k)} = \tfrac{1}{\tau\hat{N}_j}\Big(\Lambda_j - C_j^{(k)}\Gamma_j\Big), \\ &\boldsymbol{\mu}_j^{(k)} = \tfrac{1}{\hat{N}_j}\xi_j, \qquad S_j^{(k)} = \tfrac{1}{\hat{N}_j}\eta_j - \boldsymbol{\mu}_j^{(k)}(\boldsymbol{\mu}_j^{(k)})^T, \\ &\pi_j^{(k)} = \tfrac{\hat{N}_j}{N}.\end{aligned} \tag{36}$$

A.2 Kalman Smoothing Filter

The mean and covariance of the state sequence $\{\boldsymbol{s}_t\}_{t=1}^{\tau}$, conditioned on the entire observed sequence $\{\boldsymbol{x}_t\}_{t=1}^{\tau}$, can be estimated by the Kalman smoothing filter [28], [29]. Defining the expectations conditioned on the observed sequence from time $t=1$ to $t=r$ as

$$\hat{\boldsymbol{s}}_t^r = \mathbb{E}_{\boldsymbol{s}_t|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_r}\left[\boldsymbol{s}_t\right], \quad (37)$$

$$\hat{V}_{t,k}^r = \mathbb{E}_{\boldsymbol{s}_t|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_r}\left[(\boldsymbol{s}_t-\hat{\boldsymbol{s}}_t^r)(\boldsymbol{s}_k-\hat{\boldsymbol{s}}_k^r)^T\right], \quad (38)$$

the estimates are calculated with the following recursion: for $t=1,\ldots,\tau$, compute

$$\hat{V}_{t,t}^{t-1} = A\hat{V}_{t-1,t-1}^{t-1}A^T+Q, \quad (39)$$

$$K_t = \hat{V}_t^{t-1}C^T(C\hat{V}_{t,t}^{t-1}C^T+R)^{-1} \quad (40)$$

$$\hat{V}_{t,t}^{t} = \hat{V}_{t,t}^{t-1}-K_tC\hat{V}_{t,t}^{t-1} \quad (41)$$

$$\hat{\boldsymbol{s}}_t^{t-1} = A\hat{\boldsymbol{s}}_{t-1}^{t-1} \quad (42)$$

$$\hat{\boldsymbol{s}}_t^{t} = \hat{\boldsymbol{s}}_t^{t-1}+K_t(\boldsymbol{x}_t-C\hat{\boldsymbol{s}}_t^{t-1}), \quad (43)$$

where the initial conditions are $\hat{\boldsymbol{s}}_1^0=\boldsymbol{\mu}$ and $\hat{V}_1^0=S$. The estimates $\hat{\boldsymbol{x}}_t^{\tau}$, $\hat{V}_{t,t}^{\tau}$ and $\hat{V}_{t,t-1}^{\tau}$ are then computed with the following backward recursion: for $t=\tau,\ldots,1$,

$$J_{t-1} = \hat{V}_{t-1,t-1}^{t-1}A^T(\hat{V}_{t,t}^{t-1})^{-1} \quad (44)$$

$$\hat{\boldsymbol{s}}_{t-1}^{\tau} = \hat{\boldsymbol{s}}_{t-1}^{t-1}+J_{t-1}(\hat{\boldsymbol{s}}_t^{\tau}-A\hat{\boldsymbol{s}}_{t-1}^{t-1}) \quad (45)$$

$$\hat{V}_{t-1,t-1}^{\tau} = \hat{V}_{t-1,t-1}^{t-1}+J_{t-1}(\hat{V}_{t,t}^{\tau}-\hat{V}_{t,t}^{t-1})J_{t-1}^T, \quad (46)$$

and for $t=\tau,\ldots,2$,

$$\hat{V}_{t-1,t-2}^{\tau} = \hat{V}_{t-1,t-1}^{t-1}J_{t-2}^T + J_{t-1}(\hat{V}_{t,t-1}^{\tau}-A\hat{V}_{t-1,t-1}^{t-1})J_{t-2}^T \quad (47)$$

with initial condition $\hat{V}_{\tau,\tau-1}^{\tau}=(I-K_{\tau}C)A\hat{V}_{\tau-1,\tau-1}^{\tau-1}$.

A.3 Probabilistic Models for the DT

The conditional distribution $p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1})$ of the DT of (2) is

$$p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}) = \mathcal{N}(\boldsymbol{s}_t;A\boldsymbol{s}_{t-1},Q) = \frac{1}{\sqrt{(2\pi)^n|Q|}}\exp\left\{-\frac{1}{2}\|\boldsymbol{s}_t-A\boldsymbol{s}_{t-1}\|_Q^2\right\}. \quad (48)$$

Using the Markov property, for Gaussian initial conditions $\boldsymbol{s}_1\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{S})$, the distribution of state sequence $\boldsymbol{s}_1^{\tau}=\left[\ \boldsymbol{s}_1^T,\ \cdots,\ \boldsymbol{s}_{\tau}^T\ \right]^T$ is Gaussian

$$p(\boldsymbol{s}_1^{\tau}) = p(\boldsymbol{s}_1)\prod_{t=2}^{\tau}p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}) = \mathcal{N}(\boldsymbol{s}_1^{\tau};\boldsymbol{\mu}_1^{\tau},\Sigma). \quad (49)$$

Since

$$\boldsymbol{s}_t = A^{t-1}\boldsymbol{s}_1+\sum_{i=1}^{t-1}A^{t-1-i}\boldsymbol{n}_i, \quad (50)$$

it follows that

$$\boldsymbol{\mu}_t = \mathbb{E}\left[\boldsymbol{s}_t\right] = A^{t-1}\boldsymbol{\mu}_1, \quad (51)$$

$$\Sigma_{n,m} = \mathbb{E}\left[(\boldsymbol{s}_n-\boldsymbol{\mu}_n)(\boldsymbol{s}_m-\boldsymbol{\mu}_m)^T\right] = A^{n-1}S(A^{m-1})^T+\sum_{i=1}^{k-1}A^{n-1-i}Q(A^{m-1-i})^T, \quad (52)$$

where $k=\min(n,m)$. The covariance can be computed recursively, given $\Sigma_{1,1}=S$,

$$\Sigma_{n,m} = \begin{cases}(A^{m-n}\Sigma_{n,n})^T, & \text{if } m>n,\\ A(\Sigma_{n-1,n-1})A^T+Q, & \text{if } m=n,\\ A^{n-m}\Sigma_{m,m}, & \text{if } m<n.\end{cases} \quad (53)$$

Hence, the sufficient statistics of $p(\boldsymbol{s}_1^{\tau})$ are

$$\boldsymbol{\mu}_1^{\tau} = \left[\ \boldsymbol{\mu}_1^T,\ \cdots,\ \boldsymbol{\mu}_{\tau}^T\ \right]^T \quad (54)$$

and

$$\Sigma = \begin{bmatrix}\Sigma_{1,1} & (A\Sigma_{1,1})^T & \cdots & (A^{\tau-1}\Sigma_{1,1})^T\\ A\Sigma_{1,1} & \Sigma_{2,2} & \cdots & (A^{\tau-2}\Sigma_{2,2})^T\\ \vdots & \vdots & \ddots & \vdots\\ A^{\tau-1}\Sigma_{1,1} & A^{\tau-2}\Sigma_{2,2} & \cdots & \Sigma_{\tau,\tau}\end{bmatrix}. \quad (55)$$

Since the image sequence $\boldsymbol{x}_1^{\tau}$ is a linear transformation of the state sequence $\boldsymbol{s}_1^{\tau}$, it has distribution

$$p(\boldsymbol{x}_1^{\tau}) = \mathcal{N}(\boldsymbol{x}_1^{\tau};\tilde{\boldsymbol{\gamma}},\tilde{\Phi}), \quad (56)$$

where $\tilde{\boldsymbol{\gamma}}=\tilde{C}\boldsymbol{\mu}_1^{\tau}$ and $\tilde{\Phi}=\tilde{C}\Sigma\tilde{C}^T+\tilde{R}$, and $\tilde{C}$ and $\tilde{R}$ are block diagonal matrices where each diagonal block is equal to C and R respectively.

Appendix B
Algorithms

In this appendix, we summarize the algorithms used to compute the hierarchical spatial anomaly maps (Algorithm 1) of Section 4.4, the inference procedure of the proposed CRF filter (Algorithm 2 and Algorithm 3) discussed in Section 5.2.2, and the procedure used to predict anomalies by combining hierarchical anomaly maps and the CRF filter (Algorithm 4), discussed in the same section.

Algorithm 1: spatial_anomaly

Input : a video $\boldsymbol{x}$, a set of frames $\{t_i\}_{i=1}^{n_f}$, observation sites S, multi-scale spatial supports $\{\mathcal{R}^k\}_{k=1}^{L}$
Output: spatial anomaly maps $\{\mathcal{S}^k\}_{k=1}^{L}$ with multi-scale spatial surrounds

```
foreach frame t_i do
    extract spatio-temporal patches around t_i:
        {pat_i} = extract_patch(x(t_i));
    learn a MDT: {DT_i} = clustering({pat_i}) with (31)
    and (36);
    compute intra-component KLs:
        KL(i,j) = KL(DT_i, DT_j), ∀i ≠ j, with (9);
    foreach observation site S_j^{t_i} in t_i do
        foreach spatial support with the size of R_{i,j}^k do
            compute spatial anomaly S^k(i,j) using
            KL(i,j) and the segmentation maps of (4) (8)
            and (10);
        end
    end
end
```

Algorithm 2: sample_label_field

Input : previous prediction $\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1}$; initialization $\boldsymbol{y}_0$; number of iterations T.
Output: predicted label field $\boldsymbol{y}'$

$\boldsymbol{y}' \leftarrow \boldsymbol{y}_0$;
for $i \leftarrow 1$ **to** T **do**
  foreach $j \in \mathsf{S}^{(\tau)}$ **do**
    draw y'_j from
      $y'_j \backsim p(y_j|\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau}, \{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1}, \boldsymbol{y}'_{-j}; \boldsymbol{\Theta})$,
    using (23)-(25) and (12)-(15), where
    $\boldsymbol{y}'_{-j} = \{y'_1, \cdots, y'_{j-1}, y'_{j+1}, \cdots, y'_{|\mathsf{S}|}\}$;
  end
end

Algorithm 3: CRF_inference

Input : previous prediction $\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1}$; initialization $\boldsymbol{y}_0 = \{y_{0,1}, \cdots, y_{0,|\mathsf{S}|}\}$; observation up to current frame τ $\{\boldsymbol{x}^{(t)}\}_{t=1}^{\tau}$, cooling time T_c, sampling period T_S, number of samples N_S, threshold γ.
Output: predicted anomaly labels for current frame $\boldsymbol{y}^{(\tau)}$

$\boldsymbol{y} \leftarrow \mathbf{0}$, $\boldsymbol{y}' \leftarrow$ *sample_label_field*$(\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1}, \boldsymbol{y}_0, T_c)$;
for $n \leftarrow 1$ **to** N_S **do**
  $\boldsymbol{y}' \leftarrow$ *sample_label_field*$(\{\boldsymbol{y}^{(t)}\}_{t=1}^{\tau-1}, \boldsymbol{y}', T_s)$;
  $\boldsymbol{y} \leftarrow \boldsymbol{y} + \frac{1}{N_S}\boldsymbol{y}'$;
end
$\boldsymbol{y}^{(\tau)} = I(\boldsymbol{y} \geqslant \gamma\mathbf{1})$, where $I(\cdot)$ is the element-wise indicator function.

Algorithm 4: anomaly_detector

Input : a query video clip $\boldsymbol{x}$, a set of frames $\{t_i\}_{i=1}^{n_f}$, observation sites S, CRF filter $\boldsymbol{\Theta}$, multi-scale spatial supports $\{\{\mathcal{R}_i^k\}_{i=1}^{n_k}\}_{k=1}^{L}$ and associated temporal MDTs $\{\{\boldsymbol{M}_i^{(k)}\}_{i=1}^{n_k}\}_{k=1}^{L}$.
Output: predicted anomaly labels for each site of each frame $\{\boldsymbol{y}^{(t_i)}\}_{i=1}^{n_{kf}}$

foreach *frame* t_i **do**
  compute spatial anomaly maps:
  $\{\mathcal{S}^k\}_{k=1}^{L} \leftarrow$ *spatial_anomaly*$(\boldsymbol{x}, \{t_i\}, \mathsf{S}, \{\mathcal{R}\})$;
  foreach *observation site* S_j **do**
    foreach *spatial scale* k **do**
      compute hidden state sequence $\boldsymbol{s}_1^{\tau}$ for each MDT component using (37)-(47);
      compute temporal anomaly maps $\mathcal{T}_j^k$ with (3) and (49);
    end
  end
  draw the initial label field by logistic regression, using (22): $\boldsymbol{y}_{(0)}^{(t_i)} \sim p(\boldsymbol{y}|\boldsymbol{x}^{(t_i)}; \boldsymbol{w})$;
  infer labels using $\boldsymbol{y}_{(0)}^{(t_i)}$ as the starting point:
  $\boldsymbol{y}^{(t_i)} \leftarrow$ *CRF_inference*$(\{\mathcal{S}^k\}, \{\mathcal{T}^k\}, \{\boldsymbol{y}^{(t_j)}\}_{(j<i)}, \boldsymbol{y}_{(0)}^{(t_i)})$;
end

APPENDIX C EXPERIMENT

C.1 Descriptor Comparison

In this appendix, we present ROC curves corresponding to the comparisons of Table 2-3. These provide a more detailed picture of the results presented in the tables and may be useful for performance comparison with future methods. Fig. 13 presents ROC curves for the various descriptors of Table 2. Fig. 14 presents ROC curves for the filters of Table 3. In general, the ROC figures confirm the conclusions derived from the tables.

C.2 Parameter Sensitivity

The performance of the proposed anomaly detector depends on a few parameters. These include descriptor settings, such as the number of MDT components, and the size of the training data. In particular, temporal anomaly maps degrade when MDTs are learned from small training samples. Spatial anomaly detection is more flexible, as it has no memory. This is illustrated in the second row of Fig. 6, for an anomalous cart at the bottom of the frame. Since no events have been previously observed in this region, there is no training data for the temporal MDTs. Some time is thus required to learn these models, and the temporal map does not capture this anomaly. This is unlike the spatial map, where the cart is robustly detected. On the other hand, the segmentation required for spatial anomaly detection can be computationally more intensive than the detection of temporal anomalies, depending on how many video frames it requires.

Several experiments were performed to evaluate the impact of these parameters on anomaly detection accuracy. Fig. 12 characterizes the performance of one-layer temporal/spatial anomaly detection under different parameter settings. The figure on the left shows that both temporal and spatial anomaly detection improve with the number of DT components, with best performance for $K \in \{5, 6, 8\}$. Note, in particular, the significant improvement over the DT ($K = 1$). Above $K = 8$ there is some potential for overfitting and performance can degrade. Since more components imply more computation, we use $K = 5$ in all our experiments. The center figure presents ROC curves for temporal anomalies on Ped1, as a function of the number of MDT training clips. Performance increases quickly from 1 to 15 clips (200 to 3000 frames), saturating after 25. The right figure characterizes the trade-off between the efficiency and accuracy of spatial anomaly detection, as more frames are considered in the segmentation process. Increasing the number of frames from 10 to 40 improves RD by more than 10%. Beyond that, performance saturates.

C.3 Error Analysis

In this appendix, we briefly discuss the errors made by the different detector components. Some of these turned

Fig. 12. Anomaly detection performance (pixel-level criterion) on Ped1. Left: RD *v.s.* number of MDT components. Middle: ROC curves of temporal anomaly detection *v.s.* number of training clips. Right: RD of spatial anomaly detection *v.s.* number of frames used for segmentation.

out to be mislabeled instances on the two datasets. For example, the first column of Fig. 15 depicts a pedestrian that suddenly redirects her route, unexpectedly moving across the walkway. Similarly, the third column of the figure depicts a pedestrian who takes a very unorthodox route, so as to clear the way for an incoming cart. It is, in our opinion, positive that the detector flags these events, demonstrating ability to detect subtle anomalies that would not even be necessarily detected by a human without close inspection. This also confirms the well known fact that anomalies are, by definition, difficult to define a priori.

A second type of false-positives, which are technically incorrect detections, arise due to normal events that are either unusual or occur in unusual scenes. For example, in the second column of Fig. 15, a person walking leftwards at the bottom of the scene is identified as an anomaly by the temporal detector. This is because the overwhelming majority of the training events in this region are of vertically moving pedestrians (the south-north walkway leads to a much busier area of the campus than the east-west one). A more careful training set collection, using standard bootstrap procedures [44], would eliminate these false positives. With regards to spatial anomalies, unusually sparse scenes can be a source of concern. For example, in the fourth column of Fig. 15, a pedestrian entering a very sparsely populated walkway is denoted a spatial anomaly. These errors are not very serious, since spatial anomaly detection could simply be disabled for sparse scenes, or the anomaly detector could be complemented by vision techniques that perform well in these scenes (*e.g.*, pedestrian detection).

More problematic are errors due to pedestrians that move "against the flow" of the surrounding crow. This is the case of the fifth column of Fig. 15, where a left moving pedestrian enters the walkway when all other pedestrians are moving right. This behavior could be considered anomalous in some cases but not in others, depending on the scene context. For example, if the crowd was fleeing from a dangerous occurrence on the left of the scene (*e.g.*, fire) the pedestrian should be stopped. Otherwise, there is no anomaly. Again, we believe that these errors are acceptable in principle, although further studies would be required to verify that they do not overwhelm the operator of the surveillance system when there are no anomalies.

Fig. 13. Descriptor ROC curves on the UCSD dataset. Plots relative to frame-level criterion are shown on the top row, pixel-level criterion on bottom row. Left: Ped 1 dataset. Right: Ped 2 dataset. Shown in brackets are the areas under the curve (AUC). For frame-level, chance performance is the diagonal from $(0, 0)$ to $(1, 1)$. For pixel-level, it is close to a line at 0.

Fig. 14. Filter ROC curves (pixel-level) on the UCSD dataset. Left: Ped 1 dataset. Right: Ped 2 dataset.

Fig. 15. False positive anomalies. Left two columns: temporal anomaly detector. Top row shows anomaly predictions in red, the bottom the temporal anomaly maps in “jet” colormap. Right three columns: spatial anomaly detector. Top rows shows the crowd segmentations, bottom row shows the spatial anomaly maps.

View publication stats