Online Dominant and Anomalous Behavior Detection in Videos

Mehrsan Javan Roshtkhari, Martin D. Levine
Center for Intelligent Machines, McGill University
Montreal, QC., Canada
javan@cim.mcgill.ca, levine@cim.mcgill.ca

Abstract

We present a novel approach for video parsing and simultaneous *online learning of* dominant *and* anomalous *behaviors in surveillance videos. Dominant behaviors are those occurring frequently in videos and hence, usually do not attract much attention. They can be characterized by different complexities in space and time, ranging from a scene background to human activities. In contrast, an anomalous behavior is defined as having a low likelihood of occurrence. We do not employ any models of the entities in the scene in order to detect these two kinds of behaviors. In this paper, video events are learnt at each pixel without supervision using densely constructed spatio-temporal video volumes. Furthermore, the volumes are organized into large contextual graphs. These compositions are employed to construct a hierarchical codebook model for the dominant behaviors. By* decomposing *spatio-temporal contextual information into unique spatial and temporal contexts, the proposed framework learns the models of the dominant spatial and temporal events. Thus, it is ultimately capable of simultaneously modeling high-level behaviors as well as low-level spatial, temporal and spatio-temporal pixel level changes.*

1. Introduction

In this paper, we seek to simultaneously parse an *entire video* into local spatio-temporal regions in order to detect *all activities, anomalies and objects* using *unsupervised learning*. In addition, we will show that this can be achieved using a single unified formalism without possessing any models of the contents beforehand. Normal events observed in a scene will be referred to as the "dominant" behavior. These are events that have a higher probability of occurrence than others in the video and hence generally do not attract much attention. We can further categorize dominant behavior into two classes. In the literature on attention, one usually deals with foreground activities in space and time[3, 2, 8, 14, 13] while the other describes the scene background. Typically, the latter is more restrictively referred to as background subtraction, which is the building block of almost all computer vision algorithms. However, dominant behavior detection is more general and more complicated than background subtraction, since it includes the scene background while not being limited to it. The manner in which these two differ is the way that they use the scene information. Most background subtraction methods are based on the principle that the photometric properties of the scene in the video, such as luminance and color, are stationary. In contrast, dominant behavior understanding can be seen as a generalization of this in which all of the dynamic contents (foreground) of the video come into play.

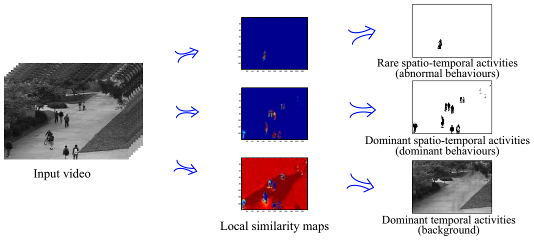

Input video
Local similarity maps
Rare spatio-temporal activities
(abnormal behaviours)
Dominant spatio-temporal activities
(dominant behaviours)
Dominant temporal activities
(background)

Figure 1: Video parsing. The input video is parsed into three meaningful components: background, dominant activities (walking pedestrians), and rare activities (the bicyclist).

Here we concentrate on detecting two of the elements in Figure 1, that is dominant spatio-temporal activities and abnormal behavior in a video. As opposed to trajectory-based methods for behavior understanding [20, 22], our approach is grounded on a pixel-by-pixel analysis. Using *densely sampled* spatio-temporal video volumes (STVs), we create both local and global compositional graphs of volumes at each pixel. Although employing STVs in the context of bag of video words (BOV) has been extensively studied for the well-known problem of activity recognition, generally it involves supervised training. Here we do not use any training sets at all but continuously update time-varying BOV lookup tables. Therefore, our approach has the ability to learn newly observed behaviors without any offline or supervised training. After initializing the algorithm, typically using one or two seconds of video, the system builds an adaptive model of the dominant behavior while simultane-

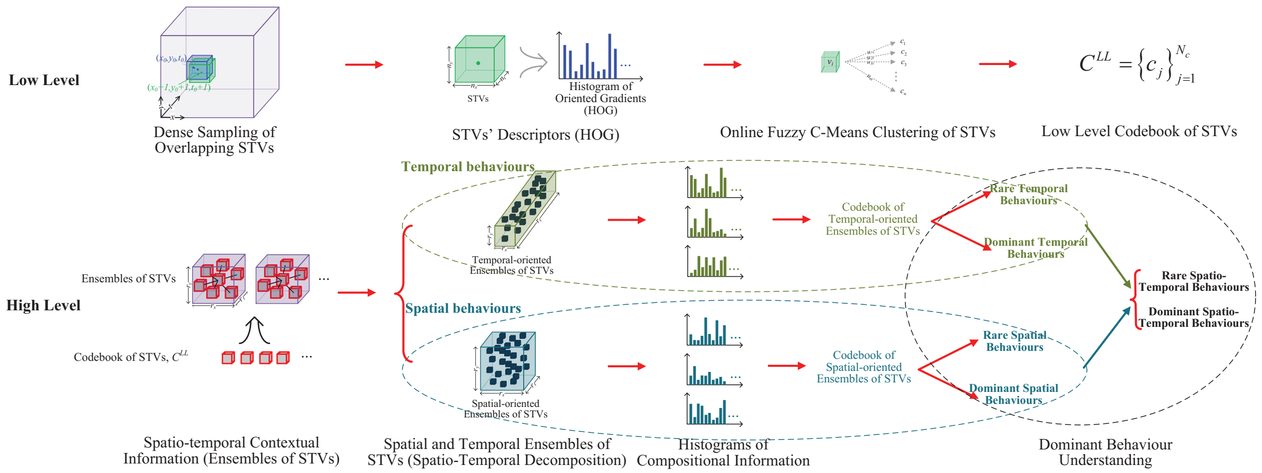

Low Level
(x_0,y_0,t_0)
(x_0+1,y_0+1,t_0+1)
Dense Sampling of Overlapping STVs
STVs
Histogram of Oriented Gradients (HOG)
STVs' Descriptors (HOG)
Online Fuzzy C-Means Clustering of STVs
$C^{LL}=\{c_j\}_{j=1}^{N_c}$
Low Level Codebook of STVs
Temporal behaviours
Temporal-oriented Ensembles of STVs
Codebook of Temporal-oriented Ensembles of STVs
Rare Temporal Behaviours
Dominant Temporal Behaviours
Rare Spatio-Temporal Behaviours
Dominant Spatio-Temporal Behaviours
High Level
Ensembles of STVs
Codebook of STVs, C^{LL}
Spatial behaviours
Spatial-oriented Ensembles of STVs
Codebook of Spatial-oriented Ensembles of STVs
Rare Spatial Behaviours
Dominant Spatial Behaviours
Spatio-temporal Contextual Information (Ensembles of STVs)
Spatial and Temporal Ensembles of STVs (Spatio-Temporal Decomposition)
Histograms of Compositional Information
Dominant Behaviour Understanding

Figure 2: Algorithm overview: behavior understanding. Behaviors are learnt from local low-level visual information, which is achieved by constructing a hierarchical codebook of the STVs. To capture spatio-temporal configurations of video volumes, a probabilistic framework is employed by estimating probability density functions of the arrangements of video volumes. The uncertainty in the codeword construction of STVs and contextual regions is considered, which makes the final decision more reliable. The high-level output can be employed to simultaneously model normal and abnormal behaviors.

ously detecting anomalies.

Consider the structure of the algorithm in Figure 2. Initially, the video is densely sampled, STVs are constructed, and similar ones are grouped to reduce the dimensions of the search space. Codebook construction of STVs is performed in an online manner while considering uncertainties in the codeword assignment. Then, a large contextual region containing many STVs (in space and time) around each pixel is examined and the compositional relationships between STVs are approximated using a probabilistic framework. We are interested in detecting different kinds of behavior in the spatial and temporal domains. To achieve this, we cluster all of the STVs that constitute all of the compositional graphs obtained during a time period in the near past. We use a modified version of online fuzzy clustering and thereby track the dominant spatio-temporal activities (clusters). These clusters of STVs provide concurrent distinctive spatial and temporal *models* of the scene. For example, we can determine all of the abnormal ("anomalous") spatial and temporal behaviors in a video.

The main contribution of this paper is an approach capable of learning both dominant and anomalous behaviors in videos of different spatio-temporal complexity. This makes it possible to construct a hierarchical layered model of the scene to understand different behaviors. Thus, the algorithm can simultaneously model *high level behaviors* and detect *abnormalities* by considering both spatial and temporal contextual information while also performing *temporal pixel level change detection* and background subtraction. This characteristic makes our algorithm more general than both abnormality detection and background subtraction methods on their own. More precisely, the main characteristic of our approach and also the contributions of the paper are as follows: *I*- The spatio-temporal contextual information in a scene is decomposed into separate spatial and temporal *contexts*, which make the algorithm capable of detecting purely spatial or temporal activities, as well as spatio-temporal abnormalities. *II*- High level activity modeling and low level pixel change detection are performed simultaneously by a single algorithm. Thus the computational cost is reduced since the need for a separate background subtraction algorithm is eliminated. This makes the algorithm capable of understanding behaviors of different complexity. *III*- The algorithm adaptively learns the behavior patterns in the scene in an online manner. This makes it a preferable choice for visual surveillance systems. *IV*- The major benefit of the algorithm is its *extendibility*, achieved by a hierarchical clustering.

In order to evaluate capabilities of our approach we have conducted experiments using different datasets with different dominant behavior patterns. The results indicate that our approach is comparable to the state-of-the-art, while it can be extended to more difficult problems[1].

2. Related work

To date, most of the reported approaches for behavior understanding that are not based on a priori models are grounded on trajectory analysis of the objects, which requires precise tracking methods [20, 22]. On the other hand, techniques that do not require object detection followed by tracking focus on *local* spatio-temporal behaviors in videos and have recently gained increased popularity [2, 11]. Most

[1] All videos and additional results are available at: http://www.cim.mcgill.ca/~javan/index_files/Dominant_behavior.html

of these methods rely mainly on extracting and analyzing low-level visual features, such as color, motion and texture in local regions in space and time. This is achieved either by constructing a pixel-level background model and behavior template [16, 14, 3, 8, 19] or by employing spatio-temporal video volumes [6, 4, 15, 29]. The recent trend in video analysis is to use spatio-temporal video volumes in the context of BOV models[2]. The classical BOV and probabilistic topic models often ignore the spatio-temporal relationships between video volumes. However, this is crucial for accurate scene understanding [25, 24]. Although there have been some efforts to incorporate either spatial or temporal compositions of the video volumes into the probabilistic topic models, they suffer from high computational complexity. Therefore, they cannot be employed for online behavior understanding and real-time scene monitoring [12].

More closely related to our proposed approach are those methods that construct a spatio-temporal behavioral model of the scene [14, 2, 13, 8]. To date, these have focussed on detecting low-level local anomalies in a video by analyzing the activity pattern of each pixel as a function of time. In [14], each pixel is processed independently and the relationships between the pixels in space and time are ignored, thereby making such methods too local. In an improved version of [14], the spatial dependencies between pixels are taken as a function of pixel location by constructing a co-occurrence frequency matrix [2]. Although the latter has achieved good results for abnormality detection, the method requires that the activity pattern of each pixel be constructed by employing a conventional method for background subtraction. These are known to be deficient for non-stationary situations.

In contrast to the aforementioned approaches that attempt to model either local spatio-temporal activity patterns of a pixel or trajectories of moving objects, our goal is to construct a hierarchical model for all of the activities in a scene. We present a novel method for inference of motion patterns, which overcomes the drawbacks and limitations of the current methods, while employing simple yet powerful hierarchical methodologies.

3. Behavior Understanding

Consider the structure presented in Figure 2. We use densely sampled videos and construct a hierarchy of spatio-temporal regions in the video to model dominant local activity patterns. The proposed hierarchical codebook structure has two important characteristics: it codes the compositional information of the video volumes and analyzes the spatial and temporal information independently, thereby making it capable of detecting purely spatial or temporal abnormalities. Moreover, the uncertainty in the codebook construction process is considered in the hierarchical structure.

3.1. Low level scene representation

The first stage of the algorithm is to represent a surveillance video by meaningful spatio-temporal descriptors. This is achieved by dense sampling, thereby producing STVs, and then clustering similar video volumes.

3.1.1 Spatio-temporal video volume descriptors

The 3D STVs, $v_i \in \mathbb{R}^{n_x \times n_y \times n_t}$ are constructed by assuming a volume of size $n_x \times n_y \times n_t$ (typically $5 \times 5 \times 5$) around each pixel (in which $n_x \times n_y$ is the size of the spatial (image) window and n_t is the depth of the video volume in time). These volumes are then characterized by the histogram of the spatio-temporal gradient of the video in polar coordinates [4, 27]. Assume that $G_x\left(x,y,t\right)$ and $G_y\left(x,y,t\right)$ are spatial gradients and $G_t\left(x,y,t\right)$ is the temporal gradient for each pixel at (x,y,t). The spatial gradient used to calculate the 3D gradient magnitude is normalized to reduce the effect of local texture and contrast. Hence, let:

$$\tilde{G}_s = \frac{\sqrt{G_x^2\left(x,y,t\right) + G_y^2\left(x,y,t\right)}}{\sum\limits_{(x,y,t)\in v_i} \sqrt{G_x^2\left(x,y,t\right) + G_y^2\left(x,y,t\right)} + \epsilon_{\max}} \quad (1)$$

where $\tilde{G}_s$ is the normalized spatial gradient and $\epsilon_{\max}$ is a constant, set to 1% of the maximum spatial gradient magnitude in order to avoid numerical instabilities. Thus the 3D normalized gradient is represented in polar coordinates:

$$[M,\theta,\phi] = \left[\sqrt{\tilde{G}_s^2 + G_t^2}, \tan^{-1}\left(\frac{G_y}{G_x}\right), \tan^{-1}\left(\frac{G_t}{\tilde{G}_s}\right)\right] \quad (2)$$

where M is the 3D gradient magnitude, and ϕ and θ are the orientations within $\left[\frac{-\pi}{2}, \frac{\pi}{2}\right]$ and $[-\pi, \pi]$, respectively. The descriptor vector for each video volume, taken as a histogram of oriented gradients (HOG), is constructed using the quantized gradients of all pixels (into $n_\theta + n_\phi$ bins) in each video volume, and will be referred to as $h_i \in \mathbb{R}^{n_\theta + n_\phi}$. This descriptor represents both motion and appearance and possesses some degree of robustness to unimportant variations in the data, such as illumination changes [4, 27]. Notwithstanding its simplicity, the results obtained are very promising. However, it should be noted that our algorithm does not rely on a specific descriptor for the video volumes, so that other more complex descriptors might enhance the performance of the approach.

[2] Essentially, the probabilistic topic models, such as the Latent Dirichlet Allocation (LDA) and its variations [29, 12], can also be considered as BOV approaches since they ignore the spatio-temporal order of the local features [17].

3.1.2 Online clustering of video volumes

In the previous section, a set of spatio-temporal volumes, v_i, was constructed using dense sampling and represented by a descriptor vector, h_i. As the number of these volumes is extremely large, it is advantageous to group similar spatio-temporal volumes to reduce the dimensions of the search space, as commonly performed in “bag of video words” approaches [4, 25]. To be capable of handling large amounts of data, and also considering the sequential nature of the video frames, the clustering strategy needs to be capable of limiting the amount of memory used for data storage and computations. Thus, we adopt an online fuzzy clustering approach for very large datasets, which is capable of incrementally updating the cluster centers as new data are observed [9]. The basic idea is to consider a chunk of data, cluster it, and then construct another chunk of data using the new observations. The clusters are then updated [9]. Here we adopt the online single-pass fuzzy clustering algorithm of [10].

Let N_d denote the number of feature vectors in the dth chunk of data and N_C the number of cluster centroids (codewords). These are represented by a set of vectors, $C = \{c_n\}_{n=1}^{N_C}$. We modify the objective function (J) [10] for fuzzy probabilistic clustering as follows:

$$J = \sum_{i=1}^{N_C} \sum_{j=1}^{N_d} u_{i,j}^m w_j d_{ij} \left(h_j, c_i\right) \tag{3}$$

where the parameter w_j is the weight of the jth sample. Note that in the original version, $w_j = 1, \forall j$ [10]. Using the Euclidean distance as the similarity measurement between STVs descriptors, we define the update rule for the cluster center, similarity matrix and the weights w_i as follows:

$$u_{n,j} = \left(\sum_{i=1}^{N_C} \left(\frac{\|h_j - c_n\|}{\|h_j - c_i\|} \right)^{\frac{2}{m-1}} \right)^{-1} \tag{4}$$

$$c_n = \frac{\sum\limits_{j=1}^{N_d} w_j u_{n,j}^m h_j}{\sum\limits_{j=1}^{N_d} w_j u_{n,j}^m}, \qquad w_i = \sum_{j=1}^{N_d+N_C} u_{i,j} w_j \tag{5}$$

Employing this clustering procedure, a set of clusters is formed for the STVs. These are used to produce is a codebook of STVs and sets of similarity values for every STV. Ultimately, each STV, h_i, will be represented by a set of similarity values: $\{u_{j,i}\}_{j=1}^{N_C}$.

3.2. Contextual information: Ensembles of volumes

As indicated earlier, in order to understand the scene background and make the correct decision regarding normal and suspicious (foreground) events, it is necessary to analyze the spatio-temporal arrangements of volumes [6, 25] in the clusters determined in section 3.1. The main drawback of many previously reported approaches is that they do not consider the context (spatio-temporal composition of the STVs) at each pixel in the video. In this paper, we present a probabilistic framework for capturing these arrangements. Instead of a single video volume, we consider a large region R around each pixel. R contains many video volumes and thereby captures both local and more distant information in the video frames. Such a set is called an *ensemble* of volumes around the particular pixel in the video (Figure 3).

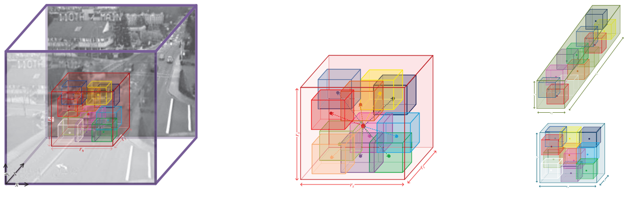

(a) (b) (c)

Figure 3: Ensembles of video volumes. (a) An ensemble of STVs. (b) Spatio-temporal contextual information. (c) Spatial and temporal oriented ensembles.

The ensemble of volumes ($E_{s,t}$) surrounding each pixel s in the video at time t, is defined as:

$$E_{s,t} = \left\{ v_i^{E_{s,t}} \right\}_{i=1}^{I} \triangleq \{v_i : v_i \in R_{s,t}\}_{i=1}^{I} \tag{6}$$

where $R_{s,t}$ is a region with pre-defined spatial and temporal radii centered at point (s,t) in the video (*e.g.*, $r_x \times r_y \times r_t$), and I indicates the total number of volumes in the ensemble. To capture the spatio-temporal compositions of the video volumes, we use the *relative* spatio-temporal coordinates of the volume in each ensemble [25]. Thus, $x_{v_i}^{E_{s,t}} \in \mathbb{R}^3$ is the relative position of the ith video volume, v_i(in space and time), inside the ensemble of volumes, $E_{s,t}$, for a given point (s,t) in the video (Figure 3b). During the codeword assignment process described in the previous section, each volume v_i inside each ensemble was assigned to all labels c_j with weights of $u_{j,i}$ using (4). Let the central volume of $E_{s,t}$ be given by v_c. Therefore, the ensemble is characterized by a set of volume position vectors, codewords and their related weights:

$$E_{s,t} = \left\{ x_{v_i^{E_{s,t}}}, u_{ji} \right\}_{i=1:I, j=1:N_C} \tag{7}$$

A common approach for calculating similarity between ensembles of volumes is to use the star graph model [6, 21, 4]. This model uses the joint probability between a database and a query ensemble to decouple the similarity of the topologies of the ensembles and that of the actual video volumes [21]. To avoid such a decomposition, we estimate the *pdf* of the volume composition in an ensemble. Thus, the probability of a particular arrangement of volumes v inside the ensemble of $E_{s,t}$ is given by:

$$\begin{aligned} P_{E_{s,t}}(v) &= P(x_v, c_1, c_2, ..., c_n) \\ &= \sum_{i=1}^{n} P(x_v | v = c_i) P(v = c_i) \end{aligned} \tag{8}$$

The first term in the summation in (8), $P(x_v|v = c_i)$, expresses the topology of the ensembles, while the second,$P(v = c_i)$, expresses the similarity of their descriptors (*i.e*. the weights for the codeword assignments at the first level). We would like to represent each ensemble of volumes by its *pdf*, $P_{E_{s,t}}(v)$. Therefore, given the set of volume positions and their assigned codewords, the probability density function (*pdf*) of each ensemble can be formed using either a parametric model or non-parametric estimation. Here, we approximate the *pdf*s describing each ensemble using (nonparametric) histograms.

3.3. Space/Time decomposition of ensembles

As stated previously, we are interested in detecting *normal* spatial and temporal activities to ultimately distinguish them from both spatial (shape and texture changes) and temporal abnormalities. These are typically foreground regions, and so our approach can also be considered as performing a *focus of attention* task. In order to individually characterize the different behaviors in the video, two sets of ensembles of spatio-temporal volumes are formed, one for the spatially oriented ensembles of volumes and the other, for the temporally oriented ones.

$$\begin{aligned} \mathbf{D}^S &= \{E_{s,t} | r_t \ll min\{r_x, r_y\}\} \\ \mathbf{D}^T &= \{E_{s,t} | r_t \gg max\{r_x, r_y\}\} \end{aligned} \tag{9}$$

where $\mathbf{D}^S$ and $\mathbf{D}^T$ represent the sets of spatially- and temporally-oriented ensembles, respectively, and $(r_x \times r_y \times r_t)$ is the size of the ensembles in (6). The spatial and temporal decomposition of ensembles of STVs is illustrated in Figure 3c.

3.4. Clustering ensembles of STVs

Once a video clip has been processed by the first level of BOV clustering in section 3.1.2, each ensemble of spatio-temporal volumes has been represented by a *pdf* of its spatio-temporal volume distribution, as described in 3.2. Note that such an ensemble *pdf* represents a moving foreground object in the video. The histogram of each ensemble, as obtained from (8), is employed as the feature vector to cluster the ensembles. This will then permit us to construct a behavioral model for the video as well as infer the dominant behavior. Using the *pdf* to represent each ensemble of volumes makes it possible to use a divergence function from statistics and information theory as the dissimilarity measure. Here we use the symmetric Kullback-Leibler (KL) divergence to measure the difference between the two *pdf*s [5]. Therefore the distance between two ensembles of volumes, E_{s_i,t_i} and E_{s_j,t_j}, is defined as:

$$\begin{aligned} d\left(P_{E_{s_i,t_i}}, P_{E_{s_j,t_j}}\right) =& KL\left(P_{E_{s_i},t_i} || P_{E_{s_j,t_j}}\right) \\ &+ KL\left(P_{E_{s_j,t_j}} || P_{E_{s_i,t_i}}\right) \end{aligned} \tag{10}$$

where $P_{E_{s_i,t_i}}$ and $P_{E_{s_j,t_j}}$ are the *pdf*s of the ensembles E_{s_i,t_i} and E_{s_j,t_j}, respectively, and d is the symmetric KL divergence between the two *pdf*s in (10). The next step is to apply online fuzzy single-pass clustering, as described in section 3.1.2, thereby, producing a set of membership values for each pixel. The clustering is performed independently for the two sets of ensembles, $\mathbf{D}^S$ and $\mathbf{D}^T$, obtained from (9). The resulting two codebooks are then represented by $\mathbf{C}^S = \{c^S_{k_S}\}_{k_S=1}^{N_S}$ and $\mathbf{C}^T = \{c^T_{k_T}\}_{k_T=1}^{N_T}$, respectively.

4. Behavior analysis

The result of the processing in section 3 permits us to construct a set of behavior patterns for each pixel. As stated previously, we are interested in detecting *dominant* spatial and temporal activities as an ultimate means of determining both spatial (shape and texture changes) and temporal abnormalities (foreground regions). Next, we consider the scenario of a continuously operating surveillance system. At each temporal sample *t*, a single image is added to the already observed frames and a new video sequence, the *query*, *Q*, is formed. The query is densely sampled in order to construct the video volumes and thereby, the ensembles of STVs, as described in section 3.

Given the already existing codebooks of ensembles constructed in 3.4, each pixel in the query, q_i is characterized by a set of similarity matrices, $\mathbf{U}^S_{q_i} = \left\{u^S_{k_S,i}\right\}_{k_S=1}^{N_S}$ and $\mathbf{U}^T_{q_i} = \left\{u^T_{k_T,i}\right\}_{k_T=1}^{N_T}$. We note that $u^S_{k_S,i}$ and $u^T_{k_T,i}$, respectively, are the similarity of the observation to the k_S spatial and k_T temporal cluster of ensembles. Then the description that best describes a new observation is given by:

$$(k^*_S, k^*_T) = \arg\left(\max_{k_S}\left\{u^S_{k_S,i}\right\}, \max_{k_T}\left\{u^T_{k_T,i}\right\}\right) \tag{11}$$

To infer normality or abnormality of the query, q_i, two similarity thresholds, Θ_{k_S} and Θ_{k_T}, are employed:

$$\left(\alpha u^S_{k^*_S,i} + \beta u^T_{k^*_T,i}\right) \overset{dominant}{\underset{rare}{\gtrless}} \left(\alpha\Theta_{k^*_T} + \beta\Theta_{k^*_S}\right) \tag{12}$$

where α and β are preselected weights for the spatial and temporal codebooks, respectively and Θ_{k_S} and Θ_{k_T} are the *learnt* likelihood thresholds for the kth codeword of the spatial and temporal codebooks, respectively. To determine these, we employ the set of previously observed pixels, $\mathbf{D} = \{p_i\}$, as represented by the two cluster similarity matrices obtained in section 3.4, $\mathbf{U}_{p_i}^{S} = \left\{ u_{k_S,i}^{S} \right\}_{k_S=1}^{N_S}$ and $\mathbf{U}_{p_i}^{T} = \left\{ u_{k_T,i}^{T} \right\}_{k_T=1}^{N_T}$. Thus, the previous observations can be divided into N_S and N_T disjoint subsets:

$$\mathbf{D}_{k_S} = \{p_i | u_{k_S,i}^{S} > \varepsilon\}_{p_i \in \mathbf{D}}, \bigcup_{k_S=1}^{N_S} \mathbf{D}_{k_S} = \mathbf{D}$$
$$\mathbf{D}_{k_T} = \{p_i | u_{k_T,i}^{T} > \varepsilon\}_{p_i \in \mathbf{D}}, \bigcup_{k_T=1}^{N_T} \mathbf{D}_{k_T} = \mathbf{D} \qquad (13)$$

where $\mathbf{D}_{k_S}$ and $\mathbf{D}_{k_T}$ contain only the most representative examples of each cluster, k_S and k_T respectively. Clearly, representativeness is governed by the parameter ε. Then, similar to [20], we construct the likelihood thresholds as follows:

$$\begin{aligned}\Theta_{k_S} =& \frac{\gamma}{|\mathbf{D}_{k_S}|} \sum_{i \in \mathbf{D}_{k_S}} \log u_{k_S,i}^{S} \\ &+ \frac{1-\gamma}{|\mathbf{D}| - |\mathbf{D}_{k_S}|} \sum_{i \notin \mathbf{D}_{k_S}} \log u_{k_S,i}^{S} \\ \Theta_{k_T} =& \frac{\gamma}{|\mathbf{D}_{k_T}|} \sum_{i \in \mathbf{D}_{k_T}} \log u_{k_T,i}^{T} \\ &+ \frac{1-\gamma}{|\mathbf{D}| - |\mathbf{D}_{k_T}|} \sum_{i \notin \mathbf{D}_{k_T}} \log u_{k_T,i}^{T}\end{aligned} \qquad (14)$$

where the parameter $\gamma \in [0, 1]$ controls the abnormality/normality detection rate and $|\mathbf{D}|$ indicates the number of members of $\mathbf{D}$. Returning to (12), the parameters α and β are seen to control the balance between spatial and temporal abnormalities based on the ultimate objective of the abnormality detection. As an example, if the objective is to detect the temporal abnormality in the scene (background/foreground segmentation), then one can assume that $\alpha = 0$.

5. Online model updating

In this section we describe how the algorithm is updated in an online manner. The scenario we have considered implies on-line and continuous surveillance of a particular scene in order to simultaneously detect dominant and anomalous patterns. As described in section 3, the algorithm only requires the first N frames of the video stream to initiate the process. This is achieved by constructing the codebook of STVs (section 3.1.2), ensembles of volumes (section 3.2) and finally the codebook of ensembles (section 3.3).

When new data are observed, the past N_d frames are always employed to update the learnt codebooks, *i.e.* the clusters of both STVs and ensembles of STVs. This process is performed continuously and the detection thresholds, Θ_{k_S} and Θ_{k_T} are updated in an ongoing manner as described in (14) based on the previously learnt codebooks.

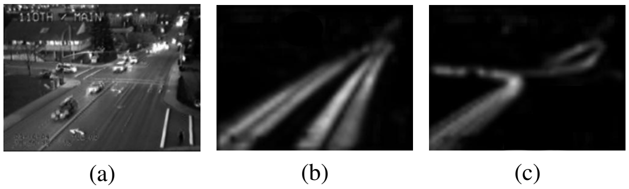

(a) (b) (c)

Figure 4: Dominant behavior understanding on data captured by a camera during different times of the day. The lighting conditions change gradually from daylight to night. a) A sample frame. b) The dominant behaviors are produced by the cars passing through the lanes running from top to bottom and vise versa. c) The abnormalities are those cars entering the intersection from the left.

6. Experiments

The algorithm has been tested using the following datasets: the dominant behavior understanding dataset in [28][3], UCSD pedestrian dataset [18][4], and subway surveillance videos [1][5]. In all cases, we have assumed that local video volumes are of size 5×5×5 and the HOG is calculated assuming $n_\theta = 16$, $n_\phi = 8$ and $N_d = 50$ frames. Parameters α and β were selected depending on the desired goal of the abnormality detection. These were set empirically to 0.1 and 0.9 for motion detection and to 0.5 for abnormal activity detection. Quantitative evaluation and comparison of different approaches are presented in terms of precision-recall and ROC curves, obtained by varying the parameter γ in (14)[6].

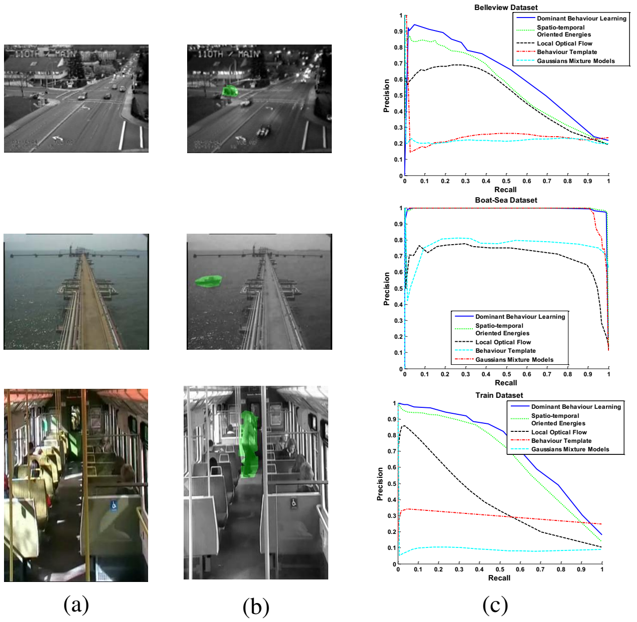
The first dataset consists of three videos sequences. The first one, *Belleview*, is a traffic scene in which lighting conditions gradually change during different times of the day. The dominant behaviors are either the static background or the dynamic cars passing through the lanes running from top to bottom. Thus, the rare events ("abnormalities") are the cars entering the intersection from the left. Figure 4 (a), (b), and (c) illustrate a sample frame, and the dominant and abnormal behavior maps, respectively. In the *Boat-Sea* video sequence, the dominant behavior is the waves while the abnormalities are the passing boats since they are newly observed objects in the scene. The *Train* sequence, is one of the most challenging videos available [28] due to drastically varying illumination and camera jitter. The background changes rapidly as the train passes through tunnels. In this sequence the abnormality relates to people movement. Figure 5 shows a sample video frame of each video sequence, the detected abnormal regions and the precision/recall curves. We followed the same initialization strategy as [28] and compared the results with two alternative pixel-level anomaly detection methods: spatio-

[3] http://www.cse.yorku.ca/vision/research/spatiotemporal-anomalous-behavior.shtml

[4] http://www.svcl.ucsd.edu/projects/anomaly

[5] Obtained from the authors of [1]

[6] To make a quantitative comparison possible, the algorithm is evaluated for abnormality detection and compared to the state-of-the-art.

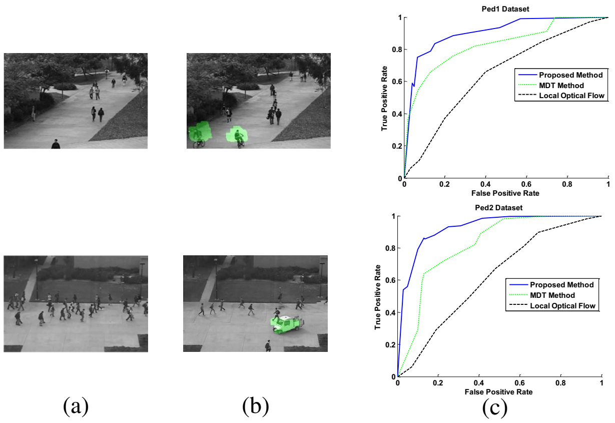


(a) (b) (c)

Figure 5: Dominant behavior understanding and abnormality detection. Experiments with three videos are illustrated from top to bottom in the figure: *Belleview*, *Boat-Sea* and *Train*. The first experiment (first row) is concerned with detecting dominant and abnormal behavior in a busy traffic scene. The second and third experiments were conducted on videos in which the abnormalities were defined as being rare but nevertheless acceptable foreground motions. The anomalous regions are highlighted in green. Column a) Sample frames from the three videos. Column b) The detected anomalous regions are cars moving from right to left (top), a boat moving to the right (middle), and a moving person (bottom). Column c) Precision/recall curves.

temporal oriented energies in [28] and local optical flow in [1]. As the abnormalities in this dataset are low level motions, we also include the pixel-level background models (Gaussians Mixture Models [30]) and the behavior template approaches in [14] for comparison.

Comparing the performance of the different approaches in Figure 5c, we observe that, in general, our method was comparable or superior to the others shown. In particular, the method based on spatio-temporal oriented energy filters [28] produced results comparable to ours, but might not be useful for more complex behaviors for two reasons: it is too local and does not consider contextual information. It is also clear that conventional methods for background subtraction (GMM) fail to detect dominant behaviors in scenes containing complicated behaviors, such as the *Train* and *Belleview* video sequences. However, they still do produce good results for background subtraction in a scene with a stationary background (*Boat-Sea* video sequences). In the latter case, the so-called abnormality (the appearance of the boat)is sufficiently different from the scene model. Thus, GMM seems promising for this video. On the other hand, we observe that simple local optical flow features, combined with online learning [1], do not yield acceptable results in the scenes with dynamic backgrounds. It appears that the optical flow approach has difficulty capturing temporal flicker and dynamic textures.



(a) (b) (c)

Figure 6: Frame level abnormality detection using the UCSD pedestrian datasets. Top: Ped1 dataset, Bottom: Ped2 dataset. a) Sample frames. b) Detected anomalous regions: bicyclist (top), a car (bottom). c) ROC curves for the proposed approach and alternatives (MDT [18], Local optical flow [1]).

Table 1: Quantitative comparison of the proposed method and the state-of-the-art for anomaly detection using the Ped1 dataset. (* indicates that the method is claimed to have real time performance).

Algorithm	EER (frame-level)	EER (pixel-level)
*Proposed algorithm	15%	29%
MDT (Mahadevan *et al.*, 2010, [18])	25%	58%
Sparse Reconstruction (Cong *et al.* 2011 [7])	19%	-
*Bertini *et al.*, 2012, [4]	31%	70%
*Reddy *et al.*, 2011, [23]	22.5%	32%
ST-MRF (Kim and Grauman, 2009, [15])	40%	82%
*Local optical flow, (Adam *et al.* 2008 [1])	38%	76%
Saligrama and Chen, 2012, [26]	16%	-

We also conducted experiments with the UCSD pedestrian dataset[7]. It contains video sequences from two pedestrian walkways where abnormal events occur. The dataset exhibits different crowd densities, and the anomalous patterns are the presence of non-pedestrians on a walkway (bikers, skaters, small carts, and people in wheelchairs). Figure 6 contains samples of two videos with the detected suspicious regions as well as the ROC curves for different methods (Figure 6c). In order to make a quantitative comparison the equal error rate (EER) was also calculated for both pixel and frame level detection as suggested by [18][8].

The results in Table 1 indicate that the proposed al-

[7] This dataset was employed as it includes pixel level ground truth showing the exact location of the abnormal regions in each frame.

[8] Frame level detection implies that a frame is marked as suspicious if it contains any abnormal pixel, regardless of its location. On the other hand, pixel level detection attempts to measure the localization ability of an algorithm. This requires that the detected pixels in each video frame be compared to a pixel level ground truth map.

gorithm outperformed all other *real-time algorithms* and achieved the best results for the UCSD pedestrian dataset at both frame level detection and pixel level localization. Furthermore, the number of *initialization frames* required by the proposed algorithm is significantly lower than the alternatives (200 frames compared to 6400 frames). This is a major advantage of the proposed method that can also learn dominant and abnormal behaviors on the fly. Moreover the computational time required by the method described in this paper is significantly lower than others in the literature. In summary, our experiments signify that our approach is capable of reasonably handling drastically and gradually changing backgrounds and illumination conditions, as well as detecting abnormal events with different spatial and temporal complexities, ranging from the scene background to human activities. Furthermore, the algorithm is adaptive. It does not require a long training video and updates itself after observing a small number of initialization frames.

7. Conclusions and future work

This paper presents a novel approach for simultaneously learning dominant behaviors and detecting anomalous patterns in videos. The algorithm is centered on three main ideas: hierarchical analysis of multi-scalar visual features; accounting for their spatio-temporal compositional information; and spatial and temporal decomposition of the behaviors in order to learn dominant spatial and temporal activities. A limitation of the current approach is that it does not account for trajectories and hence, long term behaviors are not learnt. Future research will extend the approach by adding another level of analysis in the hierarchical structure to model the spatial and temporal connectivity of the learnt behaviors.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):555–560, 2008.

[2] Y. Benezeth, P.-M. Jodoin, and V. Saligrama. Abnormality detection using low-level co-occurring events. *Pattern Recogn. Lett.*, 32(3):423–431, 2011.

[3] Y. Benezeth, P. M. Jodoin, V. Saligrama, and C. Rosenberger. Abnormal events detection based on spatio-temporal co-occurences. In *CVPR*, pages 2458–2465, 2009.

[4] M. Bertini, A. Del Bimbo, and L. Seidenari. Multi-scale and real-time non-parametric approach for anomaly detection and localization. *Compt. Vis. Image Und.*, 116(3):320–329, 2012.

[5] C. M. Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006.

[6] O. Boiman and M. Irani. Detecting irregularities in images and in video. *Int. J. Comput. Vision*, 74(1):17–31, 2007.

[7] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, pages 3449–3456, 2011.

[8] E. B. Ermis, V. Saligrama, P. M. Jodoin, and J. Konrad. Motion segmentation and abnormal behavior detection via behavior clustering. In *ICIP*, pages 769–772, 2008.

[9] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami. Fuzzy c-means for very large data. *IEEE Trans. Fuzzy Syst.*, PP(99):1–1, 2012.

[10] P. Hore, L. Hall, D. Goldgof, Y. Gu, A. Maudsley, and A. Darkazanli. A scalable framework for segmenting magnetic resonance images. *Journal of Signal Processing Systems*, 54(1):183–203, 2009.

[11] T. Hospedales, S. Gong, and T. Xiang. Video behaviour mining using a dynamic topic model. *Int. J. Comput. Vision*, pages 1–21, 2012.

[12] T. M. Hospedales, L. Jian, G. Shaogang, and X. Tao. Identifying rare and subtle behaviors: A weakly supervised joint topic model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2451–2464, 2011.

[13] P. Jodoin, V. Saligrama, and J. Konrad. Behavior subtraction. *IEEE Trans. Image. Proc.*, 21(9):4244–4255, 2012.

[14] P. M. Jodoin, J. Konrad, and V. Saligrama. Modeling background activity for behavior subtraction. In *Int. Conf. Distributed Smart Cameras*, pages 1–10, 2008.

[15] J. Kim and K. Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In *CVPR*, pages 2921–2928, 2009.

[16] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis. Real-time foregroundbackground segmentation using codebook model. *Real-Time Imaging*, 11(3):172–185, 2005.

[17] J. Li, S. Gong, and T. Xiang. Learning behavioural context. *Int. J. Comput. Vision*, 97(3):276–304, 2012.

[18] V. Mahadevan, L. Weixin, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, pages 1975–1981, 2010.

[19] A. Mittal, A. Monnet, and N. Paragios. Scene modeling and change detection in dynamic scenes: A subspace approach. *Compt. Vis. Image Und.*, 113(1):63–79, 2009.

[20] B. T. Morris and M. M. Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(11):2287–2301, 2011.

[21] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal localization and categorization of human actions in unsegmented image sequences. *IEEE Trans. Image Process.*, 20(4):1126–1140, 2011.

[22] K. Ouivirach, S. Gharti, and M. N. Dailey. Incremental behavior modeling and suspicious activity detection. *Pattern Recognition*, 46(3):671–680, 2013.

[23] V. Reddy, C. Sanderson, and B. C. Lovell. Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In *CVPR Workshops*, pages 55–61, 2011.

[24] E. Ricci, G. Zen, N. Sebe, and S. Messelodi. A prototype learning framework using emd: Application to complex scenes analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP(99):1–1, 2012.

[25] M. J. Roshtkhari and M. D. Levine. A multi-scale hierarchical codebook method for human action recognition in videos using a single example. In *Conf. Computer and Robot Vision*, pages 182–189, 2012.

[26] V. Saligrama and C. Zhu. Video anomaly detection based on local statistical aggregates. In *CVPR*, pages 2112–2119, 2012.

[27] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *International conference on Multimedia*, pages 357–360, Augsburg, Germany, 2007. ACM.

[28] A. Zaharescu and R. Wildes. Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In *ECCV*, pages 563–576, 2010.

[29] X. Zhu and Z. Liu. Human behavior clustering for anomaly detection. *Frontiers of Computer Science in China*, 5(3):279–289, 2011.

[30] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *ICPR*, pages 28–31, 2004.