Towards Zero Shot Learning of Geometry of Motion Streams and Its Application to Anomaly Recognition

Himanshu Buckchash [1] and Balasubramanian Raman [2]

[1]Indian Institute of Technology Roorkee
[2]Affiliation not available

October 30, 2023

Towards Zero Shot Learning of Geometry of Motion Streams and Its Application to Anomaly Recognition

Himanshu Buckchash*, Balasubramanian Raman

Department of Computer Science and Engineering
Indian Institute of Technology Roorkee

Abstract

Visual anomaly recognition (VAR) is the core part of many intelligent systems. However, vagueness in definitions and lack of a priori knowledge about the distribution of anomalies, makes VAR a challenging problem. Supervised solutions often fail to work in such scenarios due to lack of ability to adapt with concept drifts. To this end, we have studied the effect of temporal derivatives over differential manifolds for designing a zero-shot (label agnostic) VAR solution. Rationale behind this work is leveraging the genericity and discriminative representation available in the geometric-structure of motion-tensors. Our approach proceeds by drawing segments of temoral-derivatives from raw image-sequences and projecting them over Grassmann product space before clustering. Suitability of the proposed approach is corroborated with extensive experiments and comparisons with other arts.

Keywords: Unsupervised learning, Temporal derivatives, Multilinear algebra, Visual anomaly recognition

1. Introduction

Large number of CCTV cameras are being deployed at public places and transport networks. Camera market is forecasted to grow at a compound rate

*Corresponding author
Email address: hbuckchash@cs.iitr.ac.in, balarfma@iitr.ac.in (Himanshu Buckchash*, Balasubramanian Raman)

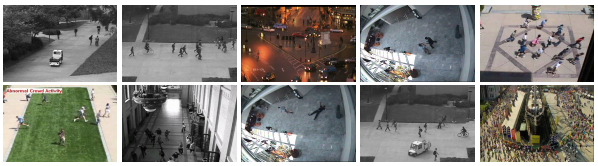


Figure 1: Figure shows sample anomalies from five different anomaly datasets.

of 16.6% annually during the period 2017 – 2025 [1]. According to Security Info Watch, in just last five years data generated by security cameras has increased from 566 petabytes to 2,500 petabytes per day [2]. Another report by Google and Smart Insights indicates, approximately one million hour of video (mostly unlabeled) is uploaded on YouTube every day [3]. This indicates about the availability of large chunks of unlabeled space-time data about crowd behavior. This can be leveraged (for surveillance purposes) by means of low-shot learning systems that do not require any supervision. Present surveillance systems have human-in-the-loop and are limited by bio-mechanical constraints such as fatigue and personal biases. For these reasons, Intelligent Surveillance Systems (ISS) are required that can handle unstructured data, such as surveillance camera streams or offline security scans, at high accuracy and reliability. Video Anomaly Recognition (VAR) is an integral task of any ISS and requires swift recognition of mistrustful events [4, 5]. The purpose of a VAR model is to recognize outlier patterns like errant behavior or event (in videos) which do not conform to normal pattern. An anomalous pattern is generally a rare event, it may include (but not limited to) – accidents, stampedes, wrong driving, physical fighting, other abnormal behaviors in public places (few samples are shown in Fig. 1). Recognizing anomalies may have potential ramifications such as preventing stampedes, intelligent human-machine interaction, old age health care, defense reconnaissance, transportation systems and so forth.

There are no set definitions for anomalies. The propensity of anomalies being sparse and inauthentic, makes the annotation of anomalies challenging. Same event might behave contradictorily under different contexts. For instance, rapid traffic might be a normal event at cross-roads, however, same is not true for a

peaceful public place. Owing to this, an unsupervised approach is more suitable for handling anomalies. According to no-free-lunch theorem the teacher based techniques trained on a labeled dataset may not perform well on data from an unseen distribution [6]. Suitability of unsupervised algorithms for unstructured data is propelled by the existence of a huge variation in spatio-temporal data, the space of labeling anomalies is immense, and an exhaustive training of a supervised model is not possible. Also, the performance of an unsupervised system increases as more unlabeled data is supplied [7].

Conventional methods mainly capture the likelihood of stationarity of active objects in the scene. For human feature extraction, it is observed that the focus should be on essential elements in image sequences rather than on RGB data, since it easily overfits to unessential elements [8]. We have observed that anomalies do not always come from distributions modeling humans rather they can be due to other objects in the scene (as is evident in the experimented datasets). Hence, a few object related representations cannot account for all kinds of irregularities. On the other hand, the clustering algorithms cannot be blamed for unaligned clusters, the major credit goes to the lack of suitable feature representations for the latent space. Here, motion can be used as a strong prior for recognition tasks like segmentation, as shown by [9, 10]. Taking cue from these observations, we pose VAR as finding irregularities based on the geometry of motion streams. For this we have employed temporal derivatives (which are indicators of temporal change in a scene) since they are sparse and locally continuous, and can be seen as trajectories in Riemannian space. Unlike the conventional approaches for anomaly recognition, our approach does not entirely depend on the dominant motion, which may suffer from perspective distortions.

Our algorithm begins by pooling all image sequences. These are then processed by SuBSENSE foreground segmentation [11]. It uses local feedback loops and adaptive sensitivity towards illumination variation. In another approach, rather than extracting foreground we use deep optical flow from FlowNet2 [12]. Both, SuBSENSE and FlowNet2 provide the required temporal derivatives with

respect to each image sequence from the pool. These are then split into smaller temporal segments. This process is visually explained in Fig. 3 in section 3. Each of these segments can be treated as a tensor of order three. These tensors are then decomposed using factor-k flattening. Each such factor can be represented as a point on a Grassmannian. Next, we take the product of Grassmannians rather than using each Grassmannian in isolation. It has been observed that the product of Grassmannians provide better results than individual factor manifolds. Chordal distance is then employed to measure the geodesic distances amongst different points on the product of Grassmannians [13]. Following this the similarity matrix, obtained from the pool of geodesic distances amongst temporal segments, is then clustered using Minimum-Cluster-Variance (MCV) based Agglomerative Hierarchical Clustering (AHC). It produces separate clusters of anomalies from non-anomalies. This works good for offline scenarios, where the entire data is available before clustering, however, in case of responsive surveillance it is crucial to have an online processing system. To this end, we present an unsupervised active learning approach, where we have weak-oracle which works on the basis of two parameters – β and γ as its confidence measures. The key idea is to delay the clustering as long as possible without compromising the confidence in clustering. Using this approach the data is processed as it arrives. Based on separate approaches for extracting temporal derivatives, we call SuBSENSE based approach – Unsupervised Segmentation (US), and FlowNet2 approach – Unsupervised Flow (UF). The results are compared with state-of-the-art deep models and unsupervised approaches for anomaly recognition. The main contribution of this work is listed below:

- To the best of our knowledge, we are the first to analyze the space-time manifolds purely on the basis of temporal derivatives with multi-linear motion representations. It allowed us to assert that the quality of anomaly recognition is not due to appearance or illumination rather owing to the inherent motion biases.

- We have studied the challenges of anomaly recognition and formulated

a simple yet generic approach for offline zero-shot anomaly recognition. Additionally, a novel unsupervised active learning approach is presented, which takes help of weak oracles in order to lay down the context. This extends our framework to online learning regime.

- We have conducted extensive empirical study with five publicly available anomaly recognition benchmarks, having coarse-to-fine level of anomalies. Further, the robustness and genericity of the proposed technique is examined under variety of problem domains such as action recognition and gesture analysis.

This paper is organized as follows – in section two, related work is provided. Following this, section three presents methodology, where we discuss the details of the proposed offline and online approaches. Section four, covers the experiments, results, ablations and analysis. Lastly, section five concludes this paper along with future research directions.

2. Related Work

The features learned by unsupervised techniques are more generalizable [14]. Bag-of-Visual Words (BoW) is a famous model and has surfaced in many zero-shot classification works [14, 15, 16]. Wang *et al.* have used spatio-temporal local features and have clustered them with k-means algorithm [14]. Similarly, Niebles *et al.* have used generative modeling of spatio-temporal features [16]. Chen *et al.* have used force fields to model crowd behavior in terms of size, position and orientation [17].

The problem of abnormality recognition (VAR) is usually formalized as an outlier recognition problem. An outlier can be detected based on the temporal or spatial data. Some prior arts like [18, 19] have used raw optical flow, [20, 17] have used pixel based approach, [21, 22, 23] have used particle based approach, [24, 25] have employed trajectory based representation to provide parametric and non-parametric solutions to VAR. These are sophisticated algorithms

which are sensitive to either local-fluctuations (in appearance) or dominant motion. However, deep learning (DL) approaches overcome these limitations by automating feature discovery and transfer learning [26, 27]. Nonetheless, it is noticed that feature transferability does not often lead to improved performance unless a model incorporates essential elements of representation [8]. Semi-supervised techniques have been proposed to leverage large unlabeled data [5, 28, 29, 30, 31, 32], however, they still depend on large labeled datasets and easily deviate with a concept drift [33, 34]. Additionally, they are a little tricky to train, which is not in the spirit of a generic solution. Our approach avoids this by employing genericity of Riemannian structure without needing any labeled data.

Few unsupervised non-visual anomaly recognition arts [35, 36, 28] have used autoencoders for feature extraction. However, these cannot directly fit on to spatio-temporal data. Moreover, these do not define any trainable objective and thus fail to extract differential representation for anomalies. To avoid this, weak-supervision models based on representation learning have been proposed [5, 30]. Representation learning assumes that if the set of regular events is known a priori then a generative or discriminative model can be trained. Generally, these models learn the latent space distribution of the regular events by minimizing a reconstruction loss. Higher approximation error values indicate anomalous event. These approaches get self-constrained by the a priori assumption on the event distribution.

Sequential DL models like RNNs have been leveraged in an encoder-decoder fashion for learning representations and frame prediction [30, 31]. Yuan *et al.* have shown how multiple LSTMs can be spatially stretched for analysis of heterogeneous input data [29]. Cho *et al.* have shown how a variational autoencoder can model video latent space [31]. Many deep learning methods have been proposed for solving VAR related problems [5, 37, 30, 4, 31, 32, 28]. However, the deep models suffer from class-imbalance since anomalies are very sparse and spurious, and the use of binary labels for weak-supervision makes the system not-fully-automated which leads to increased label bias [38].

On the other hand, it has been proved that data for many vision problems lies on low-dimensional manifolds [39]. For example, covariance matrix of an image lies as a point on a manifold of symmetric positive definite matrix. Similarly, an image set can be considered as a point on a Grassmannian [40]. These manifolds are generally structured using multilinear algebra. Multilinear algebra is used in tensor decomposition for subspace analysis of factors and their interactions. In this regard, many manifold learning strategies have been explored in the recent past for spectral decomposition of mode-2 or mode-3 tensors [41, 42, 43, 44].

Motivated by the above discussion, we pose VAR as finding irregularities based on the geometry of motion streams. Our approach employs a Riemannian metric for projecting the raw space-time data onto a manifold of temporal derivatives that leverages the temporal shape of the objects which does not get captured by trajectory based approaches or raw flow analysis. This is explained in detail in next section.

3. Approach

Our approach tries to solve the problem of video anomaly recognition (VAR) in a zero-shot way. Given the openness of criteria for defining anomalies, it is generally harder to get appropriate representations. Even the supervised models can not be trained exhaustively. In this situation one option is to use either few-shot or a zero-shot learning model. This work proposes a data driven zero-shot learning solution whose performance goes up as more and more data is added. With the zero-shot models, representation of data is the key attribute in the context of learning. We have noticed that, clustering algorithms alone cannot be blamed for unaligned clusters.

Anomalies are often a rare event. Anomaly agents leave peculiar marks in the spatio-temporal space. Our idea is to discriminatively capture these marks by leveraging the Riemannian structure present in their geometry. In such a setting the space-time anomalies can be seen as a trajectory on a non-linear manifold. An image sequence can be seen as a 3D hyperplane with H, W, T

(a) Non-anomalous instance (b) Anomalous instance

Figure 2: Energy diagram showing 2D projection of temporal segments corresponding to two different anomalous and non-anomalous instances.

representing the height, width, time axes respectively. The displacement of each agent in the image sequence leads to the evolution of different geometric structures in the hyperplane. In such a setting, each anomaly can be perceived to have individual geometry. This geometry is effected by the attributes of the anomaly agent on the H, W, T axes in the hyperplane. For visually motivating the reader, we have employed energy diagram (in Fig. 2) to capture the 2D projection of the temporal-derivatives of the anomalous and non-anomalous events in the 3D hyperplane. Fig. 2(a) and (b) illustrate the instance of non-anomalous and anomalous (a person riding a skating-board) events respectively. The anomalous event has a different projection in comparison to the normal event. The stretched out geometry of the person can be noticed in the 2D projection of the anomalous event.

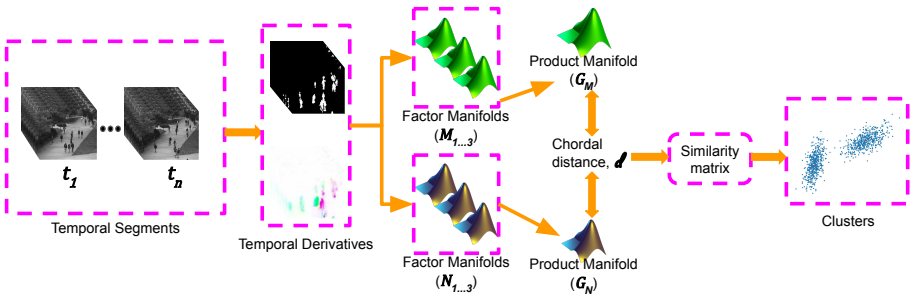


Figure 3: Framework for offline anomaly recognition.

Fig. 3 explains a high level working of the proposed approach for VAR. This approach requires the entire dataset to be available in advance for processing it all in a single pass. Hence, it is termed offline-VAR approach. It starts by creating a pool of image-sequences which are further processed to ex-

tract the corresponding temporal derivatives with reduced spatial clutter. These temporal-sequences are then broken down into smaller parts called temporal-segments or simply segments. Each segment is a mode-3 tensor. Corresponding factors are obtained from the factor-3 flattenings of these tensors. These factor manifolds ($M_{1\cdots3}$) are multiplied to achieve a Product Manifold (PM), G_M, which performs better than the individual factor manifolds. This process transforms each segment into a point on a PM. Chordal distance [13], $dist_{chordal}$ (in Eqn. (4)), is then employed as a geodesic measure between any two points to form a similarity matrix. This similarity matrix is then clustered into groups, each having points belonging to anomalous or normal event distributions. This process is explained in detail in the next few sections. Proposed approach is generic in the sense that it is data driven and has multiple application as demonstrated under section 4.5.

3.1. Grassmann manifold

Unlike many other manifolds which have intrinsic Riemannian structure, Grassmann manifold (or Grassmannian) has been found to be the most suitable representation for 3D tensors [45, 46]. Grassmannian is an abstract quotient manifold derived from Stiefel manifold, and is used to fit the orthogonality constraints. A Grassmannian $\mathcal{G}_{n,p}$ with non-zero p and n such that $n \leq p$ is the set of all n-dimensional linear subspaces of real p-dimensional space in $\mathbb{R}^p$. This forms a compact Riemannian manifold of $n(p-n)$ dimension with a homogeneous space isomorphism to $O(p)/(O(n) \times O(p-n))$. Each point on a Grassmannian $\mathcal{G}_{n,p}$ is an n-dimensional linear subspace of $\mathbb{R}^p$. It is spanned by the linked orthogonal basis Y with dimensions $p \times n$ such that $Y^T Y = I_n$, with I_n being an identity matrix with size $n \times n$. There can be other orthogonal basis, however, the selected basis matrix Y acts as the representative of the subspace span (Y). Any two points on a Grassmannian are considered equal if a $p \times p$ orthogonal linear transformation $\mathcal{R}$ maps one of these two points to the other i.e. $\lfloor \mathcal{R} \rfloor = \mathcal{R}Q_p : Q_p \in O_p$, where the element $\lfloor \mathcal{R} \rfloor$ lies on a Grassmannian $\mathcal{G}_{n,p}$ and O_p is an orthogonal group.

3.2. Geodesic similarity measure

Geodesic distance (GD) acts as the inter segment similarity measure between any two points on a Grassmannian. GD on a Grassmannian $\mathcal{G}_{n,p}$ between two p-dimensional linear subspaces – $\mathcal{P}$ and $\mathcal{Q}$ in $\mathbb{R}^n$ can be characterized in multiple ways, however, a well accepted norm is to use the canonical angles $\theta_1, \ldots, \theta_m$, between the canonical vectors of the two subspaces. It can be computed recursively as shown below:

$$\theta_k = \max_{x \in \lfloor \mathcal{P} \rfloor, y \in \lfloor \mathcal{Q} \rfloor} \cos^{-1}(\langle x, y \rangle) = \cos^{-1}(\langle x_k, y_k \rangle) \tag{1}$$

where $\langle x, y \rangle$ is the inner product between x and y. x_k is the k^{th} canonical vector of the n-plane $\mathcal{P}$ and similarly y_k is the k^{th} canonical vector of the n-plane $\mathcal{Q}$ subject to

$$\langle x, x \rangle = \langle y, y \rangle = 1$$
$$\langle x, x_j \rangle = 0, \quad \langle y, y_j \rangle = 0, \quad \text{for} \quad 1 \leq j \leq k-1$$

Chordal distance is used as GD, since it is differential everywhere and works best for Grassmannians [13]. It is defined as the $L2$-norm of the sine(s) of the angles between the corresponding canonical vectors of the two points on a Grassmannian as shown below:

$$dist_{chordal}(\mathcal{P}, \mathcal{Q}) \;\; = \;\; \| \sin\theta \|_2 \tag{2}$$

Product manifold (PM). PM is a compound object in high dimensional space which is composed of factor manifolds (FM). To understand it better, lets consider an example where we have two FMs. One is a line in $\mathbb{R}^1$ another is a circle in $\mathbb{R}^2$. The PM of these two FM is an infinite cylinder in $\mathbb{R}^3$. In a similar way, a PM represents the cross-section of its constituent FMs. A GPM is a PM of Grassmann FMs. Lets consider $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_j$ be a set of Grassmann FMs. When the topology of this set is same as the product topology then it is called

a PM. For this set, the PM can be defined as:

$$\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_j \tag{3}$$

where $\times$ represents the cartesian product. Our experiments revealed that PM yields better performance than FMs. For this reason, all experiments (in this work) have been carried out on GPM. GD on a PM $\mathcal{M}$ can be calculated as the cartesian product of the GDs on the FMs $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_j$ [45, 47]. For two mode-n tensors $\mathcal{S}$ and $\mathcal{T}$, GD on a PM can be computed as the chordal distance with $L2$-norm of the component-wise sine function as shown below:

$$dist_{chordal}(\mathcal{S}, \mathcal{T}) \;\; = \;\; \| \sin \Phi \|_2 \tag{4}$$

where $\Phi = (\Theta_1, \Theta_2, \ldots, \Theta_n)$, and the set of canonical angle $\Theta_k \in \mathcal{M}_k$ is separately calculated on each FM.

Algorithm 1: Clustering anomalies on a Grassmannian

Input: Set $\mathbb{D}$ of all temporal segments
Output: Clusterer Ψ
foreach $\mathcal{T} \in \mathbb{D}$ **do**
  $\check{\mathcal{T}} \leftarrow \texttt{temporal\_derivative}(\mathcal{T})$
  $A_{(1)}, A_{(2)}, A_{(3)} \leftarrow \texttt{unfold}_3(\check{\mathcal{T}})$
  $A_{(1)}, A_{(2)}, A_{(3)} \leftarrow A_{(1)}^T, A_{(2)}^T, A_{(3)}^T$
  $\check{\mathcal{S}} \times_1 V^{(1)} \times_2 V^{(2)} \times_3 V^{(3)} \leftarrow \texttt{hosvd}(A_{(1)}, A_{(2)}, A_{(3)})$
  $\mathcal{P} \leftarrow \{V^{(1)}, V^{(2)}, V^{(3)}\}$
  $\mathbb{E} \leftarrow \{\mathcal{P}\} \cup \mathbb{E}$
end
$\Sigma_{i,j} \leftarrow \texttt{dist}_{\texttt{chordal}}(\mathcal{P}_i, \mathcal{P}_j) \mid \forall (\mathcal{P}_i, \mathcal{P}_j) \in \mathbb{E} \times \mathbb{E}$
$\Psi \leftarrow \texttt{cluster}(\Sigma)$

3.3. Subspace representation of anomalies

As explained in Algorithm 1, the i^{th} image-sequence IS_i having frames $\{f_1, f_2, \ldots, f_n\}$ is divided into segment set $\mathbb{D} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_m\} | m < n$; with each segment $\mathcal{T}_i$ having length 10 and overlap of 30% with $\mathcal{T}_{i-1}$. Respective temporal derivative $\check{\mathcal{T}}_i$ is obtained for each segment $\mathcal{T}_i \in \mathbb{D}$. Temporal derivatives are obtained from SuBSENSE [11] and FlowNet2 [12] in the form of foreground segmentation and deep optical flow respectively.

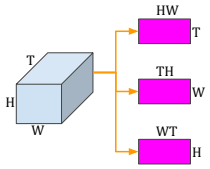


Figure 4: Factor-k flattening of a three dimensional tensor.

$\check{\mathcal{T}}_i$ is a mode-3 tensor, with dimensions H, W, T for height, width and time respectively. It is decomposed into mode-k matrices – $A_{(1)}, A_{(2)}, A_{(3)}$, with dimensions $(T, HW), (W, TH), (H, WT)$ respectively, using factor-k flattening as illustrated in Fig. 4. These matrices are decomposed into orthogonal matrices – $V^{(1)}, V^{(2)}, V^{(3)}$ along the k^{th}-axis, and one all-orthogonal core tensor $\check{\mathcal{S}}$, using a higher order singular value decomposition (HOSVD), as described in Algorithm 1. The subspaces defined by these orthogonal matrices can be treated as a point on a Grassmannian $\mathcal{G}_{n,p}$. Thus, three points are obtained on three FMs, each corresponding to an orthogonal matrix $V^{(k)}$. These FMs constitute a joint representation, point $\mathcal{P}$, on a GPM. All GPM points are collected in a set $\mathbb{E}$. A similarity matrix Σ is constructed with one entry per ordered pair of set $\mathbb{E} \times \mathbb{E}$.

Distance between two segments $(\mathcal{T}_i, \mathcal{T}_j)$ is equivalent to the chordal distance $dist_{chordal}(\mathcal{P}_i, \mathcal{P}_j)$, between two corresponding points $(\mathcal{P}_i, \mathcal{P}_j)$ on GPM. It is formulated as the $L2$-norm of the component-wise sine(s) of the principal angles between the column spaces spanned by the orthogonal matrices of the two points, as laid out by Eqn. (4). All points in the set $\mathbb{E}$ are then clustered using pairwise chordal distance and similarity matrix Σ.

3.4. Clustering

For measuring the clustering accuracy, we have employed the concept of cluster purity (accuracy). Conventionally the objective of a clustering algorithm is to increase inter-class variance while keeping the intra-class variance as low as possible. A transparent measure of clustering quality is cluster accuracy. Cluster accuracy $\eta_{(\psi_k \in \Psi)}$ of the cluster ψ_k belonging to clustered set (clusterer) Ψ, having elements with different labels from label set $\mathbb{L}$, is defined as the

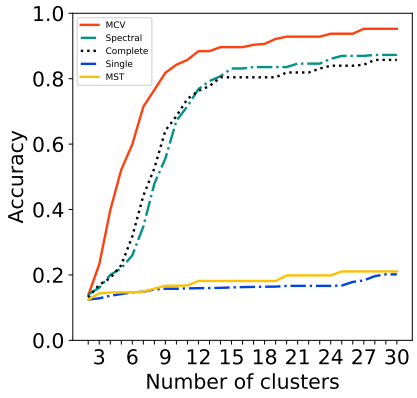


Figure 5: Effect of variation in cluster count on the performance of different clustering algorithms. It can be seen that MCV criterion based approach performs best.

ratio of number of instances of the label in majority to the total number of instances in that cluster, as shown by the Eqn. (5). Here, K is the maximum number of clusters and J is the number of unique labels. The minimum accuracy happens when $K = 1$, in that case accuracy is the ratio of label having maximum number of instances to the total number of instances. The maximum accuracy is achieved when $K = N$, where N is the total number of instances.

$$\mathbb{L} = \{\ell_1, \ell_2, \ldots, \ell_J\}; \qquad \Psi = \{\psi_1, \psi_2, \ldots, \psi_K\}$$

$$\eta_{(\Psi \in \psi_k)} = \max_j |\psi_k \cap \ell_j| / |\psi_k| \tag{5}$$

where $j = 1, 2, \ldots, J$ and $|\cdot|$ denotes set cardinality.

Fig. 5 presents the comparison among the cluster accuracies of spectral, minimum spanning tree (MST), and three HCA algorithms (each having different criteria *viz.* single connectivity, complete connectivity and minimum cluster variance (MCV)). The fluctuation in cluster accuracy is plotted (on Y-axis) with respect to variation in number of clusters used during clustering. Of all the compared algorithms, MCV has shown the best results and hence it has been used for all further experiments.

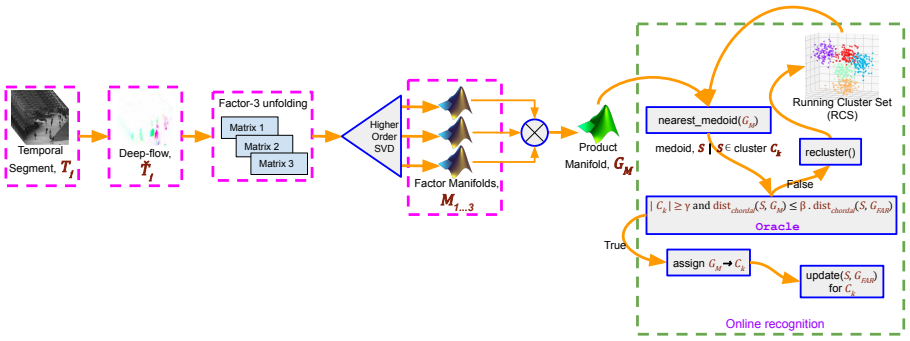


Figure 6: Framework for online anomaly recognition.

3.5. Online VAR

Thus far we have seen the working of offline VAR approach. It assumes entire data to be available beforehand, however, this assumption cannot be satisfied in situations where the data generation is a function of time. For example, in case of live surveillance, continuous steam of data is broadcasted and anomalies are required to be detected as they appear in the scene. To this end, we have designed an unsupervised active learning algorithm (Fig. 6) which leverages the existing clustered data for assigning an incoming segment to a cluster.

Algorithm 2: Online anomaly recognition

Input: Point set $\mathbb{E}$, GPM Point $\mathcal{P}_i \in \mathbb{E} \mid \mathcal{P}_i \widehat{=}$ Temporal segment $\check{\mathcal{T}}_i$
`initialize_params()`
$m \leftarrow$ `nearest_medoid`$(\mathcal{P}_i) \mid m \in$ cluster ψ_k
if $|\psi_k| \geq \gamma$ and `dist`$_{\text{chordal}}(\mathcal{P}_i, m) \leq \beta *$ `dist`$_{\text{chordal}}(m, \mathcal{P}_{far})$ **then**
  `assign`$(\mathcal{P}_i \rightarrow \psi_k)$
  `find_medoid`(ψ_k)
  `update`$(\mathcal{P}_{far}, \psi_k)$
  `update`$(\Sigma, \mathcal{P}_i, \check{\mathcal{P}}) \; \forall \; \check{\mathcal{P}} \in \psi_k$
else
  `recluster`$(\mathbb{E})$
end

Algorithm 2 describes the steps involved in online VAR. It begins by one-time initialization with offline approach, on segment set $\tilde{\mathbb{D}} \mid \tilde{\mathbb{D}} \subset \mathbb{D}$. This gives a point set $\mathbb{E}$, a similarity matrix Σ and clustered set or clusterer Ψ. After initialization, medoids are found for each cluster in Ψ. Upon arrival of a new segment $\mathcal{T}_i$, its temporal derivative $\check{\mathcal{T}}_i$ is extracted. Tensor $\check{\mathcal{T}}_i$ is then transformed into three

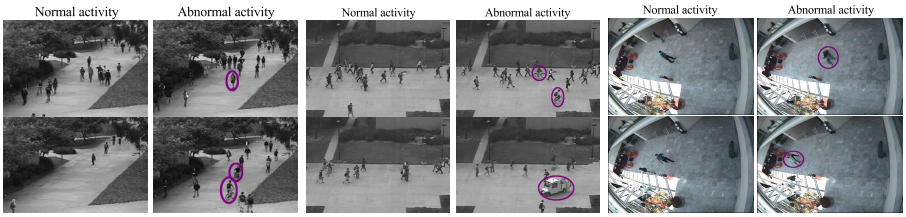
mode-k matrices with factor-k flattening. After decomposing these matrices with HOSVD, a GPM point $\mathcal{P}_i$ is obtained using Eqn. (3). Two constants – γ and β are maintained for every cluster in Ψ. These act as confidence measures for the weak-oracle. γ is the minimum number of elements to be maintained by a cluster ψ_k. $\beta \in (0, 1)$ is used for finding the maximum allowed chordal distance between medoid $m \in \psi_k$ and the GPM point $\mathcal{P}_i$, beyond this distance $\mathcal{P}_i$ is not assigned to ψ_k. If the cardinality of ψ_k, containing the nearest medoid m of $\mathcal{P}_i$, is not less than γ, and distance between $\mathcal{P}_i$ and m is not greater than β times the distance between m and the farthest element $\mathcal{P}_{far} \in \psi_k$, then the oracle assigns the point $\mathcal{P}_i$ to ψ_k and updates the similarity matrix Σ and $\mathcal{P}_{far}$, otherwise reclustering is done over the entire point set $\mathbb{E}$ including $\mathcal{P}_i$. The effect of choice of γ and β has been discussed in section 4.3.

4. Experiments

For comparison of our work with recent arts, we have selected three widely used deep-learning paradigms for video processing *viz.* Convolutional LSTM (ConvLSTM) [5], 3D Convolutional Auto-Encoders (3DConvAE) [30] and Variational Auto-Encoders (VAE) [31]. These are selected for their novelty and support for weak-supervision. Two methods with full supervision have also been selected *viz.* MRF based probabilistic inference framework called MPPCA [19] and mixture of dynamic textures (MDT) [22]. The other two works – AMC [17] and OADC [24] have been selected for their global feature based unsupervised approach. Due to unavailability of code for these methods, we have used an in-house implementation. In this section, we proceed by first presenting VAR datasets, offline and online VAR results and related ablations, followed by results analysis and few applications.

4.1. Datasets

We have carefully selected five publicly available datasets in a fine to coarser way, such that they cover both global and local anomalies. Additionally, we have



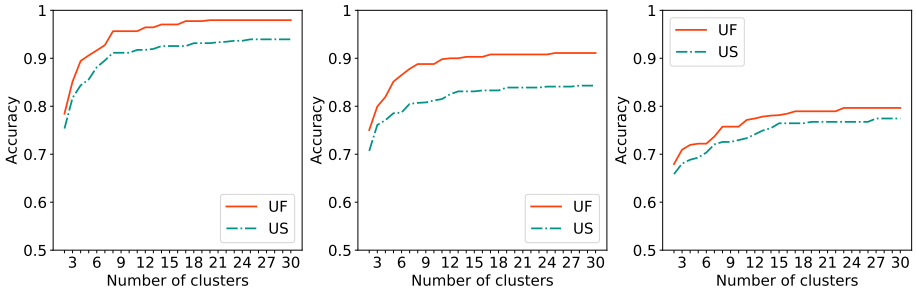
(a) UCSD Ped 1 (b) UCSD Ped 2 (c) Caviar

(d) UMN Crowd (e) UMN Web

Figure 7: Anomaly recognition standard datasets.

modified the caviar dataset by keeping anomaly related data such as fighting, slouching on the floor, idling, leaving objects and so on. Datasets used in our experiments are presented in Fig. 7, each with few samples.

UMN Crowd dataset has three crowd escape scenes comprising of a total 7740 frames, with 1–3 scene-wise frame count as 1450, 4145, and 2145 respectively. The videos start with a normal crowded situation and progresses into an abnormal crowd behavior in which the crowd escapes. UMN Web dataset comprises of videos (collected from internet) corresponding to crowd in different urban scenes. It contains 20 video clips from real life scenarios in which there are eight panic escape scenes representing abnormal behavior like people clashing, crowd fights and violent protests etc. and twelve video clips depicting normal crowd behavior scenes such as people walking or running. UMN datasets were presented by Mehran *et al.* [48]. UCSD pedestrian dataset was contributed by [21]. Normal behavior includes – pedestrians walking on the pathways and no unusual activity is happening. Abnormal behavior includes – pedestrians walking on surrounding grass, on wheelchairs, non-pedestrians such



(a) UMN Crowd (b) UMN Web (c) Caviar

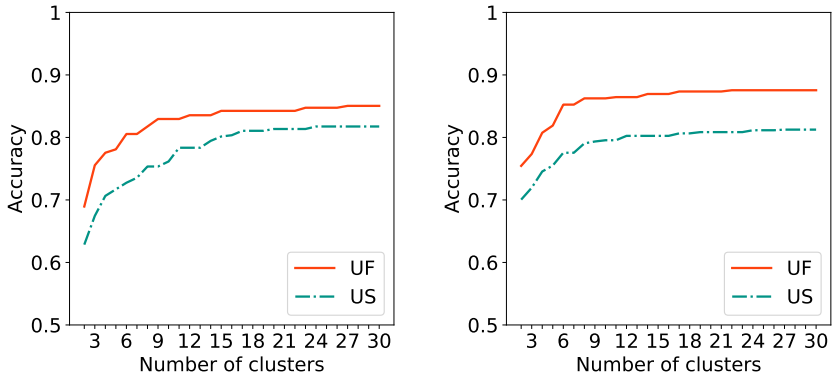
Figure 8: Figure describes the effects on performance of the UF and US approach with respect to variation in cluster count for the UMN Crowd, Web and Caviar datasets

as carts, bikers or skaters. Depending on the sites of recording, the dataset is divided into – 'Ped 1' and 'Ped 2'. Both jointly contain 50 training and 48 testing image-sequences of length 120-200, however, for our purpose we do not require training set, hence, we have combined them. Each clip is 120-200 frames long. Caviar dataset contains clips such as people meeting-and-splitting, people fighting, people slumping on ground.

4.2. Results: offline VAR

Cluster accuracy, as defined in section 3.4, refers to segment level accuracy. Nonetheless, we have used frame level accuracy for comparison of the proposed approach with other algorithms. Based on the kind of temporal approach used *viz.* foreground segmentation or deep optical flow, we have dubbed the two proposed variants as US and UF respectively. All experiments have used images of size 96×64 and a segment length of 10 with 30% overlap. TensorFlow parallel processing framework has been used with a 1080Ti hardware.

Variation in cluster count. Although there are fixed anomaly kinds in the investigated datasets which vary in the range of three to seven, we have considered cluster accuracy plots to show – the effect of variation in number of clusters used for clustering, on the cluster accuracy. Cluster accuracy plots provide useful information about the sensitivity of the model to number of clusters. Cluster

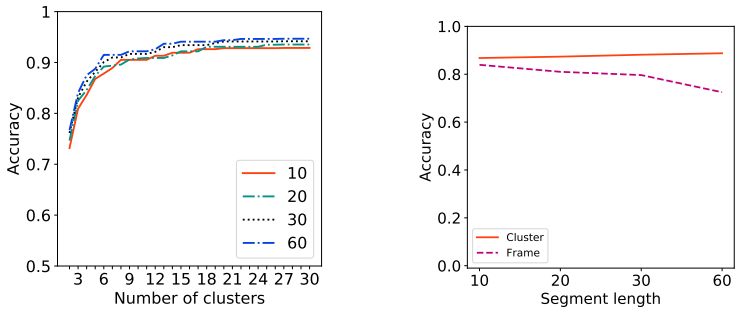


(a) UCSD Ped 1 (b) UCSD Ped 2

Figure 9: Figure describes the effects on performance of the proposed approaches with respect to variation in cluster count for UCSD Ped1 and Ped2 datasets

accuracy plots for the US and UF approaches are reported in Fig. 8, 9. Segment length for these experiments was kept ten. We can observe that in all of the plots, performance of the UF approach is better than US. One reason for this can be that the flow contains variable magnitude at each point in spatio-temporal space than segmentation. It is also evident from the plots that the clustering performance starts saturating as we increase the number of clusters on X-axis. In general, it can be observed that cluster accuracies are higher for UMN dataset which contains global anomalies in comparison to UCSD and Caviar datasets.

Variation in segment length. For the proposed approach, segment length plays a significant role in controlling the temporal resolution. A smaller value of segment length may result in a shorter temporal context while local events might get squished at larger scales. It was observed under the previous ablation on cluster count and accuracy, UF approach works better than US. Keeping this in mind we performed segment length ablations with UF approach. UMN crowd dataset contains only three occasions of anomalous behavior, hence, it was merged with UMN web dataset for augmenting the overall size. Results on segment length ablations are reported in Fig. 10a. Cluster accuracies of segment



(a) Variation in segment length

(b) Effects on prediction accuracy

Figure 10: Experiments for ablation of segment length were carried on the whole UMN dataset (crowd, web). Fig. (a) shows the performance of UF approach corresponding to segment lengths – 10, 20, 30 and 60, along with variation in cluster count on X-axis. Fig. (b) shows the effect of change in segment length on the prediction accuracies. It is observed that increase in segment length results in marginal increment in cluster accuracy, on the contrary it leads to significant decrease in frame level accuracy. This shows the suitability of segment length ten in comparison to others.

lengths – 10, 20, 30 and 60 are plotted against different cluster counts for the UF approach. It can be noticed that better cluster accuracy is achieved for a higher segment length, however, the variance in cluster accuracies amongst different segment lengths is not very high. To investigate further, Fig. 10b reports the effect of segment length variation on cluster and frame level accuracy for UF approach at cluster count five. Here, it is clearly visible that as the segment length increases, the cluster accuracy does not increase in-proportion. However, the frame-level accuracy decreases significantly with increase in segment-length. This suggests that fine level discrimination reduces as we increase the segment length.

FlowNet captures variable spatio-temporal magnitude of flow which is not very good at the edges of a moving object, on the other side, SuBSENSE based segmentation has crisp edges. This suggests that the information captured by the two approaches can be fused together. To explore this idea, we have employed conjunctive and disjunctive late-fusion approach. In the conjunctive fusion, a frame is considered anomalous if both UF and US assign the corresponding segment to a cluster with anomalous segments in majority, otherwise

Table 1: This table presents the accuracy (%) of proposed US and UF algorithms along with their combinations (to observe their supplementary effects) on the anomaly recognition task using the five datasets. Unsupervised learning types are marked with 'U'. The conjunctive and disjunctive compositions are marked with $\wedge$ and $\vee$ respectively. Highest scores are marked in bold.

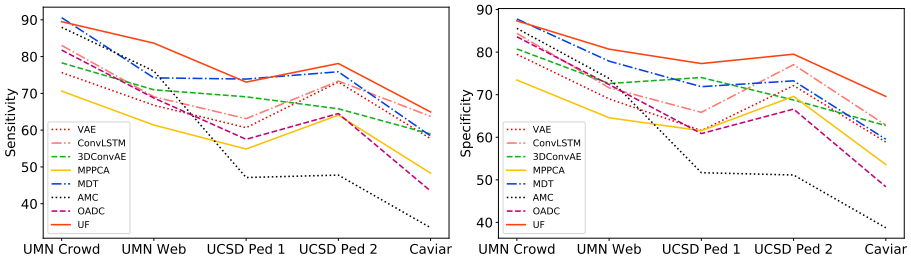
Approach	Learning	UMN crowd	UMN web	UCSD Ped1	UCSD Ped2	Caviar
US $\wedge$ UF	U	46.20	40.51	33.93	34.61	26.43
US $\vee$ UF	U	74.77	65.03	57.53	59.73	48.64
US	U	82.72	74.75	69.18	72.67	65.86
UF	U	**87.81**	**81.35**	**75.54**	**79.16**	**67.73**

Table 2: Table below summarizes the anomaly recognition performance of proposed UF method in comparison to other arts, over five anomaly datasets in terms of accuracy (%). Supervised, Weakly-Supervised and Unsupervised learning types are marked with 'S', 'WS' and 'U' respectively. Highest scores are marked in bold.

Approach	Learning	UMN crowd	UMN web	UCSD Ped1	UCSD Ped2	Caviar
VAE [31]	WS	78.63	68.33	61.21	72.47	58.45
ConvLSTM [5]	WS	84.14	70.81	64.73	75.56	63.17
3DConvAE [30]	WS	80.23	72.14	71.87	67.58	61.21
MPPCA [19]	S	72.83	63.36	58.55	67.40	51.42
MDT [22]	S	**88.59**	76.71	72.69	74.15	59.09
AMC [17]	U	86.06	74.51	49.73	53.32	36.52
OADC [24]	U	83.19	71.16	59.43	65.78	46.26
UF	U	87.81	**81.35**	**75.54**	**79.16**	**67.73**

it is considered normal. Contrary to this, disjunctive fusion means that a frame is considered anomalous if anyone among US or UF assigns the segment, corresponding to the frame, to an anomalous cluster. The frame-level accuracies for US and UF along with their late fusion schemes are summarized under Table 1. It can be noted that conjunctive scheme performs significantly low, whereas disjunctive scheme performs slightly better. However, this plain fusion strategy does not explain the semantics behind low performance, our plan to extend fusion is discussed in section 4.4 and 5.

Table 2 presents the frame level accuracy results of UF algorithm (at cluster count five) in comparison to other arts. All compared algorithms have shown roughly the same trend in their sensitivity towards the level of difficulty of the



(a) Sensitivity (b) Specificity

Figure 11: Sensitivity and specificity plots for comparison of different approaches. Experiments for UF approach were carried out with five clusters.

considered datasets. It can be observed in Table 2, in general, all methods perform better on the UMN dataset than UCSD or Caviar dataset. Weakly supervised deep learning methods (VAE, ConvLSTM, 3DConvAE) have shown comparative performance across datasets. Of these three, ConvLSTM has shown marginally better accuracy. MDT has shown better results than MPPCA, and has attained highest score on the UMN crowd dataset. OADC is implemented without motion saliency as we think it interferes with natural motion by increasing motion contrast irrespective of knowledge of kind of anomaly it handles. Due to this OADC has performed better than AMC on local events and comparatively on global events. Amongst the unsupervised approaches UF performs best followed by OADC. AMC shows worst performance for the UCSD and Caviar datasets due to its biases towards global events. Performance of the proposed UF approach has been better than others over all datasets.

Fig. 11 reports the specificity and sensitivity plots of different algorithms. Sensitivity measures the probability of an anomalous event being recognized as anomalous whereas specificity measures the probability of a normal event being recognized as normal. It is evident from Fig. 11 that UF has better recognition rates for both anomalies and normal events, whereas MPPCA and AMC have the lowest average recognition rates for anomalies and normal events.

Fig. 12 presents UF clustering results on the UCSD Ped1 dataset, using five clusters. Each row belongs to a cluster and contains images corresponding

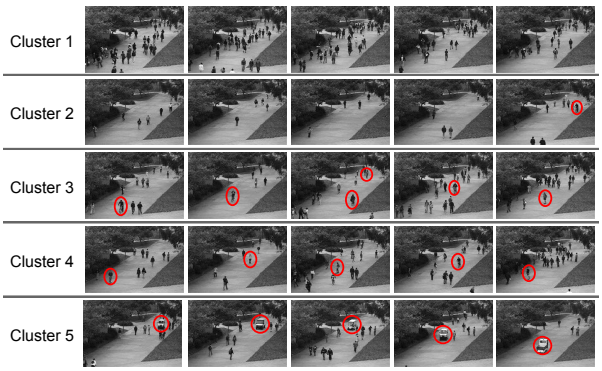
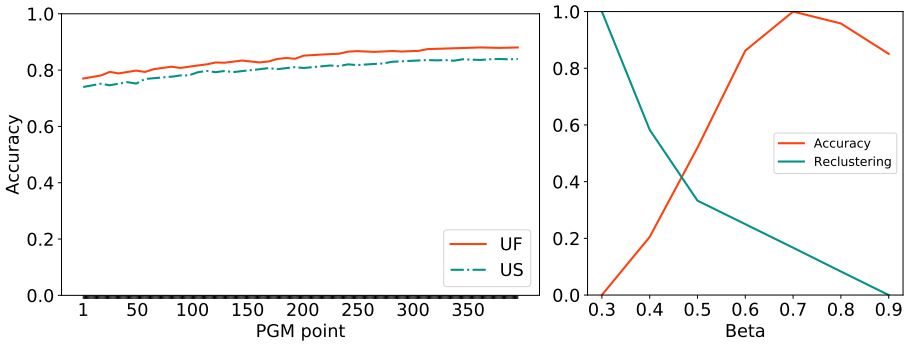


Figure 12: Qualitative result of clustering with UF algorithm, for five number of clusters, is shown for the UCSD Ped1 dataset. Five medoids are selected from each cluster. Anomalies are encircled in red. It can be noticed that each cluster tries to capture a different situation in the scene.

to five medoids from that cluster. Anomalies are marked in red. One can observe that each cluster tries to identify different aspect of the scene. Anomalies are concentrated towards the last three clusters while the first two clusters capture the density of pedestrians. Cluster one has high pedestrian density and some anomalies as well. Cluster two has low pedestrian density a few anomalies. Cluster three and four have bike and skating anomalies. Cluster five has anomalies involving large size vehicle movement. Fig. 12 shows that each cluster captures some specific kind of information about the scene.

4.3. Results: online VAR

Online approach is suitable for streamed data. Out of the five anomaly datasets, only UMN Crowd contains a long duration, continuous image-sequence with anomaly instances. Online VAR cluster accuracy results, for US and UF with five clusters on UMN Crowd dataset, are displayed in the Fig. 13a. For online VAR experiment, the dataset was split with 70:30 proportion, where 70% of the data was used for initializing the online approach with offline VAR (as explained by Algorithm 2), and remaining 30% data was used to test the performance of online approach. Segment length ten, with 30% overlap, was used during all experiments.



(a) Online anomaly recognition (b) Variation in β

Figure 13: Fig. (a) shows how the clustering accuracy changes as new segments are added for UMN crowd dataset. Fig. (b) reveals how the accuracy and the tendency to recluster varies with modulation in β.

By the end of image-sequence stream, the online variant – US and UF reach almost same performance as their offline variants. Another important observation about the Fig. 13a hints at the data-driven nature of proposed approach as the cluster accuracy increases with increase in data. Two important confidence measures of the online approach are γ and β. γ is empirically determined as ten. Ablations on β are explained next.

Relation between accuracy and beta. Oracle uses β as a confidence measure to decide the maximum allowed distance from the medoid of a cluster to a new point, before assigning it to that cluster. The relationship of cluster accuracy and frequency/tendency to recluster with respect to β is reported in Fig. 13b. The plot contains normalized reclustering and accuracy scores. One can observe that frequency of reclustering is inversely proportional to β, which means that for smaller values of β, reclustering often happens. Accuracy on the other hand varies proportionately with β and beyond a point it varies inversely. This fluctuation in accuracy creates a trade-off. We expect our approach to have higher accuracy even if it incurs some reclustering. Owing to this reason β is set to 0.7 for all our experiments with online VAR.

4.4. Results analysis and discussion

Through above experiments we found that in the absence of any supervision, inherent biases in the data samples can lead to semantically pronounced clusters. This makes the proposed approach very suitable for anomaly recognition and for other tasks that do not require any supervision. It was observed that anomalies like people slumping on a floor have negligible temporal signatures and can easily qualify as a normal event. We think that this can be mitigated by either jointly modeling the spatial context with temporal derivatives or by introducing weak binary supervision. It can be addressed in the future works.

Performance scores from Table 2 indicate that datasets with global scale anomalies have mean-accuracies – 82.68, 72.29; though, datasets with local level anomalies have mean accuracies –64.21, 69.42, 55.48. It can be observed that usually the anomalies at global scale are identified better than the anomalies at local scale. This is also corroborated by sensitivity rates under Fig. 11a. We found that unlike methods which work well for global level anomalies such as AMC, the proposed approach works well for both global and local kind of anomalies. We found that the proposed approach has best sensitivity and specificity in comparison to other approaches. The sensitivity scores are slightly higher than corresponding specificity scores. This implies that the anomalies have higher recognition rates and are better discriminated than non-anomalous data.

Qualitative clustering results are being reported in Fig. 12. One can notice that the anomalies are clustered towards the end while the first two clusters contain non-anomalous pedestrian movements. However, it should be noted that cluster one and two differ from each other in terms of population density. In Fig. 12, one may examine that the UF approach sometimes fails to distinguish between the kind of anomaly such as bike or skating.

Due to uncertainty in the types and count of anomalies, we have assessed the performance of proposed approach by clustering with different number of clusters. It was discovered that the performance in cluster accuracy curves (Fig. 8, 9) normally saturates towards the end; this suggests that if we take large enough clusters then most of the events can be categorized well. It is

positive from the cluster accuracy curves that UF performs better than the US approach. One reason for this could be the additional directional information captured by the UF approach than US.

It was observed in section 4.2, variation in segment length does not affect the cluster accuracy, however, frame level accuracy is affected by it. Owing to this deduction, segment length of ten was used across all experiments. In order to enable the proposed approach for streamed data, online UF approach was proposed. Performance of the online approach was found equivalent to the offline approach. Online approach was found to improve its performance when number of segments were increased with time. These observations make the online approach a suitable candidate for VAR in online scenarios. Online approach depends on its confidence measure β. We observed that higher values of beta resulted in low reclustering rates and better accuracies.

US and UF variants of the proposed approach were combined in a late fusion manner to gauge their effectiveness to complement each other. However, as the results revealed in Table 1, this strategy was not successful. Other direction could be to have feature level fusion of the two variants, however, this may prove challenging due to zero-shot learning nature of the algorithms. Furthermore, the performance of the proposed technique can be enhanced by having a hierarchical deep learning model for fusion of manifolds with different segment lengths. Such a model can leverage a variational auto-encoder approach in hierarchical manner for combining long and short term temporal contexts. This can help in alleviation of decline in frame-level performance by providing finer temporal resolutions.

For the sake of analyzing the generalization ability of the proposed approach, we have experimented on a few similar problems like nearest-neighbor retrieval, action and gesture recognition. This is covered in the next section.

4.5. Applications

In the previous sections we have seen that the proposed approach works well for VAR. In this section we evaluate the performance of the proposed approach,

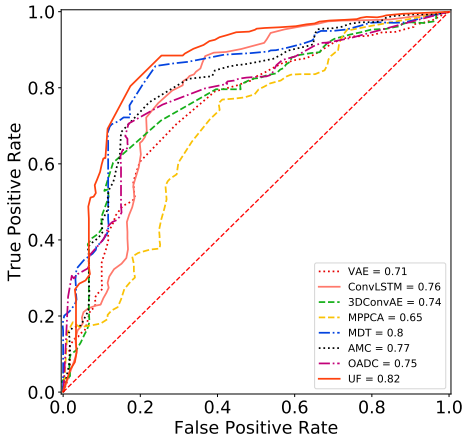


Figure 14: ROC curve for nearest-neighbor based anomalous event retrieval on the whole UMN dataset (crowd, web). AUC values are listed in the legend for each algorithm.

in the context of VAR related problems, to test its suitability for other applications. More specifically, we have considered the task of media retrieval, action recognition and gesture recognition.

4.5.1. Nearest-neighbor retrieval of events

With a view to measure the quality of representations learned by different techniques we have designed a label based, query-by-example nearest-neighbor retrieval (NNR) task. Given a query frame f_i, at time step i, with database index $\texttt{id}(f_i)$ which belongs to class $\texttt{class}(f_i)$ (either normal or anomalous), the task of NNR is to retrieve another frame f_j from the retrieval database $\texttt{D}_{\texttt{base}}$ such that $\texttt{id}(f_j) = argmin_{f_k \in \texttt{D}_{\texttt{base}}}(\texttt{dist}(f_i, f_k))$. If $\texttt{class}(f_i) = \texttt{class}(f_j)$, the retrieval is considered as true positive. Whole UMN dataset (crowd, web) is considered for NNR task. The joint dataset is split into two disjoint sets $\mathbb{S}_{90}$ and $\mathbb{S}_{10}$ with 90:10 proportion. Set $\mathbb{S}_{90}$ is used for constructing $\texttt{D}_{\texttt{base}}$, set $\mathbb{S}_{10}$ is used for query generation. None of the query frames $f_i \in \mathbb{S}_{10}$ is indexed in database $\texttt{D}_{\texttt{base}}$.

During NNR task, frame level accuracies are considered and results are evaluated with qualitative and quantitative approach. Fig. 14 shows the quantitative results with Receiver Operating Characteristic (ROC) curves along with Area

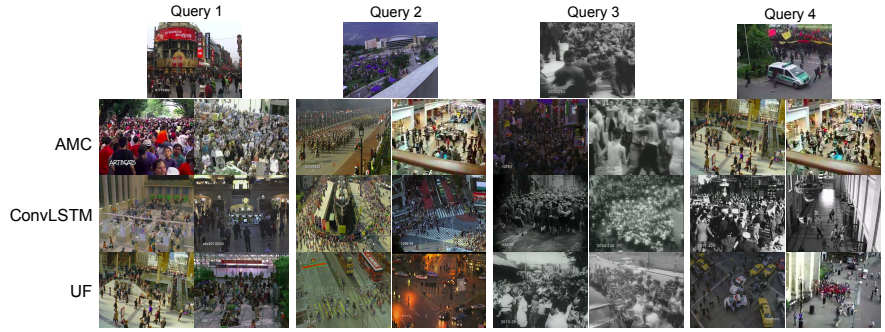


Figure 15: Retrieval results corresponding to query by example queries. Top row contains four different queries, other rows contain unsupervised and semi-supervised retrieval results. Each query is complemented by two adjacent results, displayed in front of the corresponding algorithm name.

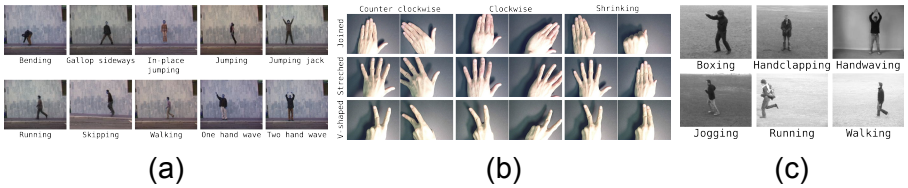


Figure 16: Dataset samples are shown in this figure. (a) Weizmann and (c) KTH are used for action recognition task. (b) Cambridge gesture dataset contains nine classes.

Under Curve (AUC) measure. Proposed approach has the highest AUC value, followed by MDT. The qualitative results are presented in Fig. 15. Each row shows the two retrieved nearest-neighbors of a query. We find that our method outperforms the other arts due to the GPM representation.

4.5.2. Action and gesture recognition

Other two closely related tasks considered under applications are action recognition and gesture recognition. We have used KTH and Weizmann datasets, having six and ten human action classes respectively. Classwise dataset samples are shown in the Fig. 16. Cambridge gesture data set has nine hand gesture types depending on rotation and finger pose.

Experiments have been performed with the proposed offline approach and cluster accuracy was observed with respect to given number of classes per dataset. Recognition accuracy values for action dataset are reported through

Table 3: Accuracy of action recognition on KTH dataset, comparison with other arts. Supervised and Unsupervised learning types are marked with 'S' and 'U' respectively.

Approach	Learning	Accuracy (%)
Dollar *et al.* [49]	S	81.1
Schuldt *et al.* [50]	S	71.7
Ke *et al.* [51]	S	63.0
Kim *et al.* [52]	S	90.0
STIP [53]	S	87.0
STW [16]	U	83.0
Niebles *et al.* [54]	U	83.3
UF	U	**93.6**

Table 4: Accuracy of action recognition on Weizmann dataset, comparison with other arts. Supervised and Unsupervised learning types are marked with 'S' and 'U' respectively..

Approach	Learning	Accuracy (%)
Scovanner *et al.* [55]	S	82.8
Kellokumpu *et al.* [56]	S	95.7
Wang *et al.* [57]	S	97.8
Niebles *et al.* [54]	U	72.0
UF	U	**94.6**

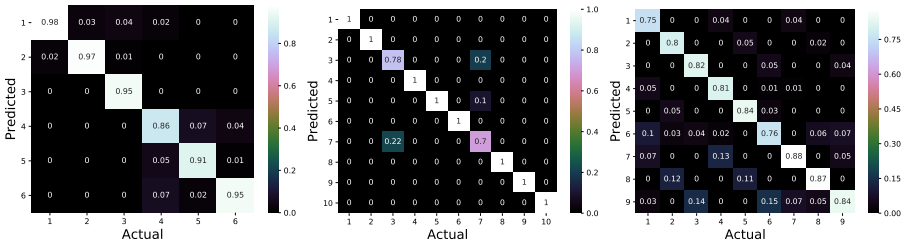
Table 3 and 4. Gesture recognition accuracies are reported in Table 5. Confusion matrices are provided for classwise recognition scores, for each of the three datasets in Fig. 17. We find that the performance of proposed approach is comparable to other supervised and unsupervised works. Consistent performance of the proposed approach across different multiclass applications demonstrates that the proposed approach is not biased towards anomaly related binary recognition tasks.

5. Conclusions and future work

In this work, a zero-shot learning approach for anomaly recognition is proposed by modeling the temporal derivatives as trajectory on Grassmann Product Manifold (GPM). GPM is leveraged for discriminative representations and less room for design choices. Video anomaly recognition problem is comprehensively studied in terms of coarse-to-fine scale anomalies on five publicly available

Table 5: Accuracy of gesture recognition on Cambridge gesture dataset, comparison with other arts. Supervised and Unsupervised learning types are marked with 'S' and 'U' respectively..

Approach	Learning	Accuracy (%)
Kim *et al.* [58]	S	81.5
Hu *et al.* [59]	S	80.8
Yan *et al.* [60]	S	78.8
Gu *et al.* [61]	S	77.6
Qiao *et al.* [62]	S	76.0
Wong *et al.* [63]	S	83.5
STW [16]	U	70.2
UF	U	**81.9**



(a) KTH (b) Weizmann (c) Cambridge gesture

Figure 17: Per class recognition scores are reported with the help of confusion matrix for action and gesture datasets.

datasets.

Additionally, an online variant is proposed for adapting the offline model to streamed data. For this, a modified version of active learning is presented where we have a weak oracle which uses confidence measures to take decisions without any help from a strong learner. Performance of the online approach is found comparable to the offline approach. We found that the ability to leverage the inherent bias in the data samples makes the proposed approach very suitable for anomaly recognition task. The genericity of the proposed approach is further validated over other multiclass recognition tasks. Despite using any label information, the overall performance of the proposed zero-shot approach is found comparable to other supervised or weakly-supervised works.

We observed that most of the approaches performed well for the UMN

dataset, however, all approaches had significantly low performance on Caviar dataset. One reason for this could be the inclusion of spatio-temporal stagnation of objects in the anomaly events and complex multiperson interactions. Few late fusion strategies have also been explored, however, the fusion showed significant drop in performance. This revealed the need for further exploration in the direction of early and feature level fusion schemes. We plan to address these issues in future. We also plan to adapt the proposed work with other applications such as video summarization, content based spatio-temporal search, automatic concept discovery. Our future work involves exploring metrics on GPMs with emphasis on large-scale complex human to human interactions.

References

[1] Video Surveillance Market, 2017 (accessed June 25, 2019). URL: https://www.transparencymarketresearch.com/video-surveillance-vsaas-market.html/.

[2] Data Generated by Surveillance Cameras, 2016 (accessed June 25, 2019). URL: https://www.securityinfowatch.com/video-surveillance/news/12160483/data-generated-by-new-surveillance-cameras-to-increase-exponentially-in-the-coming-yea

[3] R. Allen, What Happens Online in 60 Seconds, 2017 (accessed June 25, 2019). URL: https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/.

[4] Y. S. Chong, Y. H. Tay, Abnormal event detection in videos using spatiotemporal autoencoder, in: International Symposium on Neural Networks, Springer, 2017, pp. 189–196.

[5] W. Luo, W. Liu, S. Gao, Remembering history with convolutional lstm for anomaly detection, in: Multimedia and Expo (ICME), 2017 IEEE International Conference on, IEEE, 2017, pp. 439–444.

[6] D. H. Wolpert, The supervised learning no-free-lunch theorems, in: Soft computing and industry, Springer, 2002, pp. 25–42.

[7] N. Weber, A. Härmä, E. P. D. T. Heskes, Unsupervised learning in human activity recognition: A first foray into clustering data gathered from wearable sensors (2017).

[8] Y. Li, Y. Liu, C. Zhang, What elements are essential to recognize human actions, in: 2019 IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2019.

[9] D. Pathak, R. B. Girshick, P. Dollár, T. Darrell, B. Hariharan, Learning features by watching objects move, in: CVPR, volume 1, 2017, p. 7.

[10] P. Agrawal, J. Carreira, J. Malik, Learning to see by moving, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 37–45.

[11] P.-L. St-Charles, G.-A. Bilodeau, R. Bergevin, Subsense: A universal change detection method with local adaptive sensitivity, IEEE Transactions on Image Processing 24 (2014) 359–373.

[12] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2462–2470.

[13] J. H. Conway, R. H. Hardin, N. J. Sloane, Packing lines, planes, etc.: Packings in grassmannian spaces, Experimental mathematics 5 (1996) 139–159.

[14] D. Wang, Q. Shao, X. Li, A new unsupervised model of action recognition, in: Image Processing (ICIP), 2015 IEEE International Conference on, IEEE, 2015, pp. 1160–1164.

[15] A. Coates, A. Y. Ng, Learning feature representations with k-means, in: Neural networks: Tricks of the trade, Springer, 2012, pp. 561–580.

[16] J. C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, International journal of computer vision 79 (2008) 299–318.

[17] D.-Y. Chen, P.-C. Huang, Motion-based unusual event detection in human crowds, Journal of Visual Communication and Image Representation 22 (2011) 178–186.

[18] Y. Cong, J. Yuan, J. Liu, Abnormal event detection in crowded scenes using sparse representation, Pattern Recognition 46 (2013) 1851–1864.

[19] J. Kim, K. Grauman, Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 2921–2928.

[20] Y. Cong, J. Yuan, Y. Tang, Video anomaly search in crowded scenes via spatio-temporal motion context, IEEE transactions on information forensics and security 8 (2013) 1590–1599.

[21] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE transactions on pattern analysis and machine intelligence 36 (2014) 18–32.

[22] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 1975–1981.

[23] Y. Ma, P. Cisar, Event detection using local binary pattern based dynamic textures, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2009, pp. 38–44.

[24] Y. Yuan, J. Fang, Q. Wang, Online anomaly detection in crowd scenes via structure analysis, IEEE transactions on cybernetics 45 (2014) 548–561.

[25] H. Mousavi, M. Nabi, H. Kiani, A. Perina, V. Murino, Crowd motion monitoring using tracklet-based commotion measure, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 2354–2358.

[26] X. Hu, S. Hu, Y. Huang, H. Zhang, H. Wu, Video anomaly detection using deep incremental slow feature analysis network, IET Computer Vision 10 (2016) 258–267.

[27] S. Ma, S. A. Bargal, J. Zhang, L. Sigal, S. Sclaroff, Do less and achieve more: Training cnns for action recognition utilizing action images from the web, Pattern Recognition 68 (2017) 334–345.

[28] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, et al., Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications, in: Proceedings of the 2018 World Wide Web Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2018, pp. 187–196.

[29] Z. Yuan, X. Zhou, T. Yang, Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 984–992.

[30] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, X.-S. Hua, Spatio-temporal autoencoder for video anomaly detection, in: Proceedings of the 2017 ACM on Multimedia Conference, MM '17, ACM, 2017, pp. 1933–1941.

[31] J. An, S. Cho, Variational autoencoder based anomaly detection using reconstruction probability, Special Lecture on IE 2 (2015) 1–18.

[32] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional lstm network: A machine learning approach for precipitation

nowcasting, in: Advances in neural information processing systems, 2015, pp. 802–810.

[33] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, L. S. Davis, Learning temporal regularity in video sequences, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 733–742.

[34] H. He, S. Chen, K. Li, X. Xu, Incremental learning from stream data, IEEE Transactions on Neural Networks 22 (2011) 1901–1914.

[35] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360 (2016).

[36] T. Ergen, A. H. Mirza, S. S. Kozat, Unsupervised and semi-supervised anomaly detection with lstm neural networks, arXiv preprint arXiv:1710.09207 (2017).

[37] D. Balderas, P. Ponce, A. Molina, Convolutional long short term memory deep neural networks for image sequence prediction, Expert Systems with Applications 122 (2019) 152–162.

[38] H. Jiang, O. Nachum, Identifying and correcting label bias in machine learning, arXiv preprint arXiv:1901.04966 (2019).

[39] O. Tuzel, F. Porikli, P. Meer, Region covariance: A fast descriptor for detection and classification, in: European conference on computer vision, Springer, 2006, pp. 589–600.

[40] M. Harandi, C. Sanderson, C. Shen, B. C. Lovell, Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3120–3127.

[41] C. C. Olson, K. P. Judd, J. M. Nichols, Manifold learning techniques for unsupervised anomaly detection, Expert Systems with Applications 91 (2018) 374–385.

[42] B. Wang, Y. Hu, J. Gao, Y. Sun, B. Yin, Product grassmann manifold representation and its lrr models, in: AAAI, 2016, pp. 2122–2129.

[43] S. O'Hara, Y. M. Lui, B. A. Draper, Using a product manifold distance for unsupervised action recognition, Image and vision computing 30 (2012) 206–216.

[44] B. Vandereycken, Low-rank matrix completion by riemannian optimization, SIAM Journal on Optimization 23 (2013) 1214–1236.

[45] Y. Ma, J. Kosecka, S. Sastry, Optimal motion from image sequences: A riemannian viewpoint, in: In Proceeding of the Conference on Mathematical Theory of Networks and Systems, Citeseer, 1998.

[46] P. Turaga, A. Veeraraghavan, R. Chellappa, Statistical analysis on stiefel and grassmann manifolds with applications in computer vision, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[47] E. Begelfor, M. Werman, Affine invariance revisited, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 2, IEEE, 2006, pp. 2087–2094.

[48] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 935–942.

[49] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, IEEE, 2005, pp. 65–72.

[50] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: A local svm approach, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, volume 3, IEEE, 2004, pp. 32–36.

[51] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, volume 1, IEEE, 2005, pp. 166–173.

[52] T.-K. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 1005–1018.

[53] S.-F. Wong, R. Cipolla, Extracting spatiotemporal interest points using global information, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE, 2007, pp. 1–8.

[54] J. C. Niebles, L. Fei-Fei, A hierarchical model of shape and appearance for human action classification, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, pp. 1–8.

[55] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th ACM international conference on Multimedia, ACM, 2007, pp. 357–360.

[56] V. Kellokumpu, G. Zhao, M. Pietikäinen, Human activity recognition using a dynamic texture based method, in: BMVC, volume 1, 2008, p. 2.

[57] L. Wang, D. Suter, Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, pp. 1–8.

[58] T.-K. Kim, R. Cipolla, Gesture recognition under small sample size, in: Asian conference on computer vision, Springer, 2007, pp. 335–344.

[59] R.-X. Hu, W. Jia, D.-S. Huang, Y.-K. Lei, Maximum margin criterion with tensor representation, Neurocomputing 73 (2010) 1541–1549.

[60] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, H.-J. Zhang, Multilinear discriminant analysis for face recognition, IEEE Transactions on Image Processing 16 (2006) 212–220.

[61] Q. Gu, Z. Li, J. Han, Joint feature selection and subspace learning, in: Twenty-Second International Joint Conference on Artificial Intelligence, 2011.

[62] Z. Qiao, L. Zhou, J. Z. Huang, Sparse linear discriminant analysis with applications to high dimensional low sample size data., International Journal of Applied Mathematics 39 (2009).

[63] W. K. Wong, Z. Lai, Y. Xu, J. Wen, C. P. Ho, Joint tensor feature analysis for visual object recognition, IEEE transactions on cybernetics 45 (2014) 2425–2436.