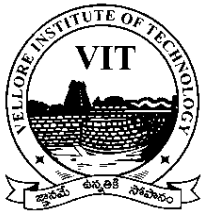


VIT-AP UNIVERSITY

Computer Vision

(Course Code: 4047)

Module-4: Computer Vision Applications

Lecture-5: Vehicle Anomaly Detection in Video Surveillance

Gundimeda Venugopal, Professor of Practice, SCOPE

Spatio-Temporal Analysis

- Spatiotemporal models arise when data are collected across time as well as space and has at least one spatial and one temporal property.
- An event in a spatiotemporal dataset describes a spatial and temporal phenomenon that exists at a certain time t and location x.
- Typical examples of spatiotemporal data mining include
  - Video Sequence Analysis
  - Action Detection
- Other examples
  - Discovering the evolutionary history of cities and lands
  - Uncovering weather patterns
  - Predicting earthquakes and hurricanes and determining global warming trends.

Video Anomaly Detection (VAD)

- The goal of a practical anomaly detection system is to timely signal an activity that deviates normal patterns and identify the time window of the occurring anomaly. It can be considered as coarse level video understanding, which filters out anomalies from normal patterns.
- A critical task in video surveillance is detecting anomalous events such as traffic accidents, crimes or illegal activities.
- Anomalous events rarely occur as compared to normal activities. Hence the application of this task is to "alleviate the waste of labor and time, developing intelligent computer vision algorithms for automatic video anomaly detection".

Anomaly Detection (Difference Domains)

Anomaly detection is utilized across many domains.

1- Fraud Detection in Finance

2- Intrusion Detection in Cyber Security

3- Fault Detection in Industrial Systems

Video Anomaly Detection

Anomaly detection is utilized across many domains.

surveillance monitoring where anomalies like fighting, stealing, accidents can be detected.

Anomaly Detection: Approaches

How Video anomaly detection (VAD) models function?

1) one-class classification approaches

2) weakly-supervised learning approaches.

VAD: One Class Classification approach

How Video anomaly detection (VAD) models function?

1) one-class classification approaches

- dataset (Normal videos only)

- feature learning (feature extraction)

- model training to reconstruct these videos based on those features

Goal: Minimize the reconstruction error

Inference: During inference, the trained model extracts features and reconstructs input videos. If the reconstruction error exceeds a predefined threshold, the input is flagged as anomalous, based on the assumption that anomalies will have significantly different features compared to normal videos.

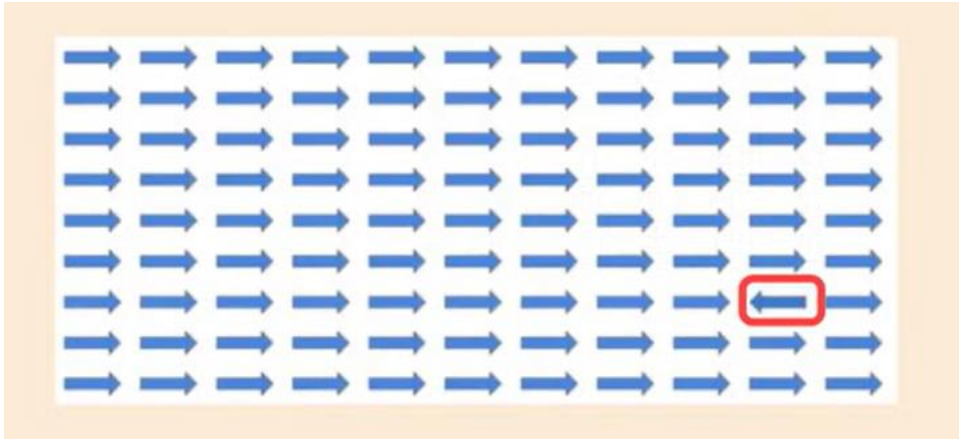
VAD: One Class Classification approach

- Trained with normal videos
  - To understand how the normal behavior looks like
  - Anomalies are rare (can be treated as outliers)
- Feature Learning (Feature Extraction)
  - Motion Patterns
  - Colour Distributions
  - Texture
- Model for Feature Learning: e.g., Auto Encoders

- Training the Network: Minimize Reconstruction Error
  - If the Generated Video from the features learnt is highly similar to the original video, the loss would be minimum. Otherwise, it would be high
  - The goal is to minimize the reconstruction loss and train the network with all the normal videos

- During Inference Time, the network would features from the input video and the network reconstructs the video. Reconstruction Error is calculated
- If Reconstruction > Threshold (say 0.6),
  - the video is an anomalous video
- Else
  - The video is a normal video

Example: Parked cars in a Parking Lot



Auto Encoders

Anomaly Detection: Weakly supervised approach

2) weakly-supervised learning approaches.

- dataset: Normal and Anomalous Videos

Anomalous

Anomalous

Normal

Anomaly Detection: Weakly supervised approach

Real-world Anomaly Detection in Surveillance Videos

Waqas Sultani[1], Chen Chen[2], Mubarak Shah[2]
[1]Department of Computer Science, Information Technology University, Pakistan
[2]Center for Research in Computer Vision (CRCV), University of Central Florida (UCF)
waqas5163@gmail.com, chenchen870713@gmail.com, shah@crcv.ucf.edu

Abstract

Surveillance videos are able to capture a variety of realistic anomalies. In this paper, we propose to learn anomalies by exploiting both normal and anomalous videos. To avoid annotating the anomalous segments or clips in training videos, which is very time consuming, we propose to learn anomaly through the deep multiple instance ranking framework by leveraging weakly labeled training videos, i.e. *the training labels (anomalous or normal) are at video-level instead of clip-level. In our approach, we consider normal and anomalous videos as bags and video segments as instances in multiple instance learning (MIL), and automatically learn a deep anomaly ranking model that predicts high anomaly scores for anomalous video segments. Furthermore, we introduce sparsity and temporal smoothness constraints in the ranking loss function to better localize anomaly during training.*

We also introduce a new large-scale first of its kind dataset of 128 hours of videos. It consists of 1900 long and untrimmed real-world surveillance videos, with 13 realistic anomalies such as fighting, road accident, burglary, robbery, etc. as well as normal activities. This dataset can be used for two tasks. First, general anomaly detection considering all anomalies in one group and all normal activities in another group. Second, for recognizing each of 13 anomalous activities. Our experimental results show that our MIL method for anomaly detection achieves significant improvement on anomaly detection performance as compared to the state-of-the-art approaches. We provide the results of several recent deep learning baselines on anomalous activity recognition. The low recognition performance of these baselines reveals

1. Introduction

Surveillance cameras are increasingly being used in public places *e.g.* streets, intersections, banks, shopping malls, etc. to increase public safety. However, the monitoring capability of law enforcement agencies has not kept pace. The result is that there is a glaring deficiency in the utilization of surveillance cameras and an unworkable ratio of cameras to human monitors. One critical task in video surveillance is detecting anomalous events such as traffic accidents, crimes or illegal activities. Generally, anomalous events rarely occur as compared to normal activities. Therefore, to alleviate the waste of labor and time, developing intelligent computer vision algorithms for automatic video anomaly detection is a pressing need. The goal of a practical anomaly detection system is to timely signal an activity that deviates normal patterns and identify the time window of the occurring anomaly. Therefore, anomaly detection can be considered as coarse level video understanding, which filters out anomalies from normal patterns. Once an anomaly is detected, it can further be categorized into one of the specific activities using classification techniques.

A small step towards addressing anomaly detection is to develop algorithms to detect a specific anomalous event, for example violence detector [30] and traffic accident detector [23, 34]. However, it is obvious that such solutions cannot be generalized to detect other anomalous events, therefore they render a limited use in practice.

Real-world anomalous events are complicated and diverse. It is difficult to list all of the possible anomalous events. Therefore, it is desirable that the anomaly detection algorithm does not rely on any prior information about the events. In other words, anomaly detection should be done with minimum supervision. Sparse-coding based approaches [28, 41] are considered as representative methods that achieve state-of-the-art anomaly detection results. These methods assume that only a small initial portion of a

Motivation and contributions. Although the above-mentioned approaches are appealing, they are based on the assumption that any pattern that deviates from the learned normal patterns would be considered as an anomaly. However, this assumption may not hold true because *it is very difficult or impossible to define a normal event which takes all possible normal patterns/behaviors into account* [9]. More importantly, the boundary between normal and anomalous behaviors is often ambiguous. In addition, under realistic conditions, the same behavior could be a normal or an anomalous behavior under different conditions. Therefore, it is argued that the training data of normal and anomalous events can help an anomaly detection system learn better. In this paper, we propose an anomaly detection algorithm using weakly labeled training videos. That is we only know the video-level labels, *i.e. a video is normal or contains anomaly somewhere, but we do not know where.* This is intriguing because we can easily annotate a large number of videos by only assigning video-level labels. To formulate a weakly-supervised learning approach, we resort to multiple instance learning (MIL) [12, 4]. Specifically, we propose to learn anomaly through a deep MIL framework by treating normal and anomalous surveillance videos as bags and short segments/clips of each video as instances in a bag. Based on training videos, we automatically learn an anomaly *ranking model* that predicts high anomaly scores for anomalous segments in a video. During testing, a long-untrimmed video is divided into segments and fed into our deep network which assigns anomaly score for each video segment such that an anomaly can be detected. In summary, this paper makes the following contributions.

Anomaly Detection: Weakly supervised approach

2) weakly-supervised learning approaches.

Multiple Instance Learning (MIL):

Bags Formations. We divide each video into the equal number of non-overlapping temporal segments and use these video segments as bag instances. Given each video segment, we extract the 3D convolution features [36]. We use this feature representation due to its computational efficiency, the evident capability of capturing appearance and motion dynamics in video action recognition.

Anomalous (positive bag)

Normal (negative bag)

- Each video is treated as a bag. The video contains multiple instances/segments. A global Video Label is assigned to the Video (Normal or Anomalous).
- If the video is labelled as anomalous (a positive bag), anomalies may be present in any one segment or in multiple segments. Where in the video is anomalous activity is not specified.
- No labels are assigned to the individual segments/ instances. This is makes the dataset a weakly labelled dataset.
- This technique is called Multiple Instance Learning.
- Labeling work is minimal for this approach. If we want to label each segment, it would be huge amount of work.

Anomaly Detection: Weakly supervised approach

Anomalous video:
(positive bag)

Example- Divided in 40 segments where each segment have 4 frames. First 3 segments displayed below:

Normal video:
(negative bag)

Example- Divided in 40 segments where each segment have 4 frames. First 3 segments displayed below:

Multiple Instance Learning (in the context of a video)

- Multiple Instance Learning(MIL) is a Machine Learning(ML) framework where each video is associated with a bag. The video contains multiple instances / segments.
- If the bag contains an anomaly (it may be in any segment/instance of the video), it is labelled as a positive bag
  - If the video is labelled as anomalous (a positive bag), anomalies may be present in any one segment or in multiple segments. Where in the video is anomalous activity is not specified.
  - No labels are assigned to the individual segments/instances. This is makes the dataset a weakly labelled dataset.
- If the bag does not contain any anomaly, it is labelled as a negative bag

Anomaly Detection: Weakly supervised approach

3. Proposed Anomaly Detection Method

The proposed approach (summarized in Figure 1) begins with dividing surveillance videos into a fixed number of segments during training. These segments make instances in a bag. Using both positive (anomalous) and negative (normal) bags, we train the anomaly detection model using the proposed deep MIL ranking loss.

3.1. Multiple Instance Learning

In standard supervised classification problems using support vector machine, the labels of all positive and negative examples are available and the classifier is learned using the following optimization function:

$$\min_{\mathbf{w}} \left[\frac{1}{k} \sum_{i=1}^{k} \overbrace{max(0, 1 - y_i(\mathbf{w}.\phi(x) - b))}^{①} \right] + \frac{1}{2} \|\mathbf{w}\|^2, \quad (1)$$

where ① is the hinge loss, y_i represents the label of each example, $\phi(x)$ denotes feature representation of an image patch or a video segment, b is a bias, k is the total number of training examples and $\mathbf{w}$ is the classifier to be learned. To learn a robust classifier, accurate annotations of positive and negative examples are needed. In the context of supervised anomaly detection, a classifier needs temporal annotations of each segment in videos. However, obtaining temporal annotations for videos is time consuming and laborious.

MIL relaxes the assumption of having these accurate temporal annotations. In MIL, precise temporal locations of anomalous events in videos are unknown. Instead, only video-level labels indicating the presence of an anomaly in the *whole* video is needed. A video containing anomalies

Anomalous behavior is difficult to define accurately [9], since it is quite subjective and can vary largely from person to person. Further, it is not obvious how to assign 1/0 labels to anomalies. Moreover, due to the unavailability of sufficient examples of anomaly, anomaly detection is usually treated as low likelihood pattern detection instead of classification problem [10, 5, 20, 26, 28, 42, 18, 26].

In our proposed approach, we pose anomaly detection as a regression problem. We want the anomalous video segments to have higher anomaly scores than the normal segments. The straightforward approach would be to use a ranking loss which encourages high scores for anomalous video segments as compared to normal segments, such as:

$$f(\mathcal{V}_a) > f(\mathcal{V}_n), \quad (3)$$

where $\mathcal{V}_a$ and $\mathcal{V}_n$ represent anomalous and normal video segments, $f(\mathcal{V}_a)$ and $f(\mathcal{V}_n)$ represent the corresponding predicted scores, respectively. The above ranking function should work well if the segment-level annotations are known during training.

However, in the absence of video segment level annotations, it is not possible to use Eq. 3. Instead, we propose the following multiple instance ranking objective function:

$$\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i), \quad (4)$$

where *max* is taken over all video segments in each bag. Instead of enforcing ranking on every instance of the bag, we enforce ranking only on the two instances having the highest anomaly score respectively in the positive and negative bags. The segment corresponding to the highest anomaly score in the positive bag is most likely to be the true positive

Ranking Loss + Training + Inference

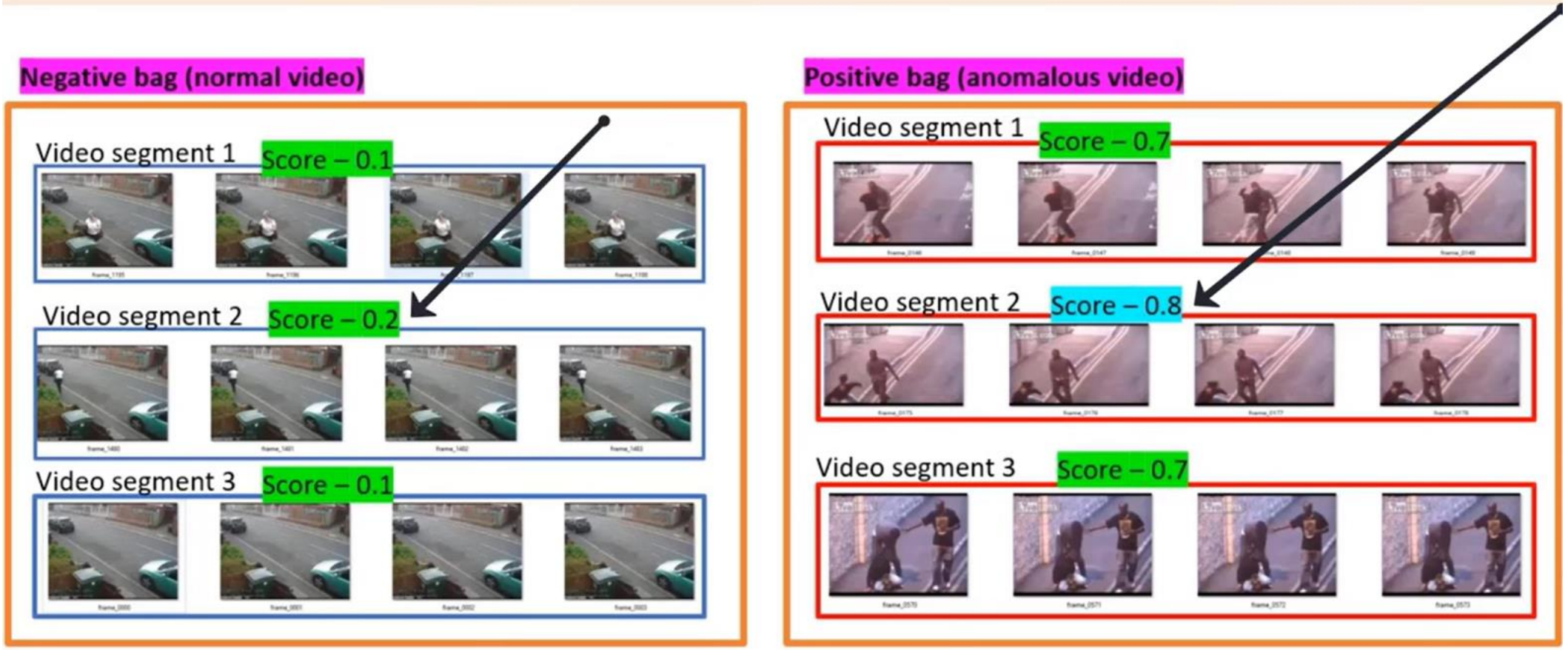
- A regression approach is used for Anomaly Detection
- Instance Anomaly Score = How likely the instance is anomalous
- Bag Anomaly Score = How likely the bag is anomalous.
- The Goal is to assign higher anomaly score for positive (anomalous) bag segments/instances compared to negative (normal) bag segments/instances.
- Highest anomalous segment score of a positive bag should be higher than higher anomaly score of a segment in negative bag. There may be multiple anomalous segment in a positive video with different anomaly scores and many segments of the anomalous video may not be anomalous with low segment anomalous scores.
- During training, the model processes both positive bags and negative bags. For each bag, the model calculates the Bag anomalous Score.
- Bag's overall anomalous score = Maximum anomalous score among all segment
- Calculate the Ranking loss between highest score instances in the positive and negative bags
- If the maximum anomaly score of negative bag > maximum of positive bag, the model considers a normal video is more anomalous than the anomalous video. To correct this, The model is penalized and loss is increased.
- Using this approach, during training, the model learns to recognize normal and anomalous segments thereby classifying a video as Anomalous or Normal.
- During inference, the trained model processes is used to process each video segment one by one to give anomaly score which is then compared to a threshold to decide if the segment is anomalous or not. If a segment is anomalous, the video is categorized as anomalous.

Note: Root Mean Square Error or Cross Entropy loss is not used for individual segment/instance loss

Anomaly Detection: Weakly supervised approach

2) weakly-supervised learning approaches.

Multiple Instance Learning (MIL):



2) weakly-supervised learning approaches.

Multiple Instance Learning (MIL):

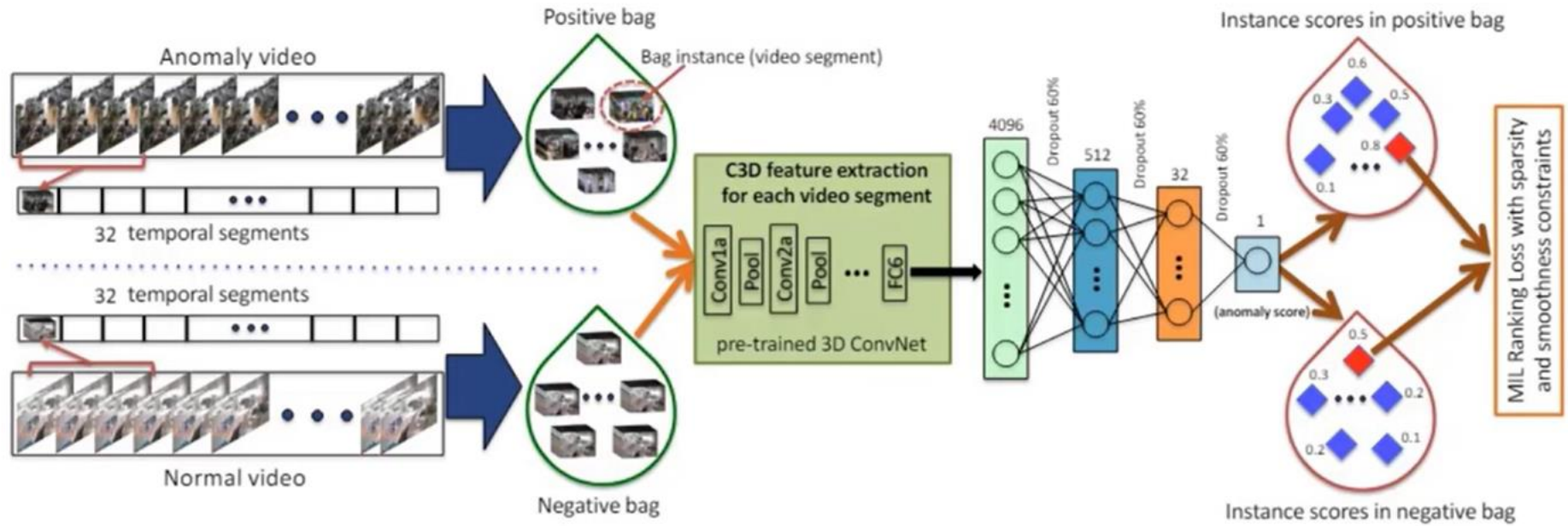


Figure 1. The flow diagram of the proposed anomaly detection approach. Given the positive (containing anomaly somewhere) and negative (containing no anomaly) videos, we divide each of them into multiple temporal video segments. Then, each video is represented as a bag and each temporal segment represents an instance in the bag. After extracting C3D features [36] for video segments, we train a fully connected neural network by utilizing a novel ranking loss function which computes the ranking loss between the highest scored instances (shown in red) in the positive bag and the negative bag.

2) weakly-supervised learning approaches.

Multiple Instance Learning (MIL):

MIL-based approaches usually consist of at least two modules:

1) a video processing backbone

- extract features (C3D or I3D models)
- convert features into embeddings

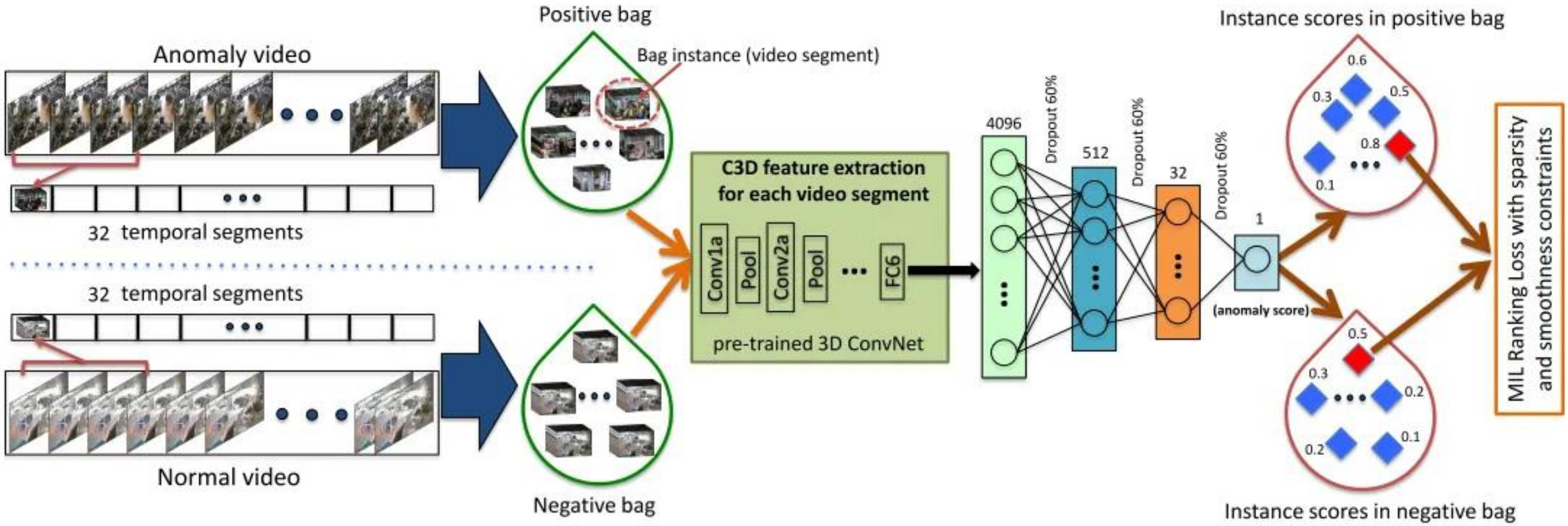
2) a prediction head.

- embedding from video processing backbone is the input to prediction head
- prediction head predicts an anomaly score

Vehicle Anomaly Detection

- Anomaly detection is such an important problem that many Intelligent Transportation Systems (ITS) are facing with it.
- Detecting anomalies in vehicles direction of movement is a subset of this complex problem.
- Example 1: Vehicles moving in the wrong direction pose a major risk for other drivers. Without doubt, if anomalies in vehicles direction of movement are detected accurately in real-time; risk of accidents can be decreased significantly.
- Example 2: Vehicles not Parked at right parking slots
- Example 3: Accidents on the road. Vehicles in irregular directions

Multiple instance learning (MIL) framework



Multiple instance learning (MIL) framework (image from Sultani et al., 2018)

References

- Real-world Anomaly Detection in Surveillance Videos CVPR 2018
- https://sertiscorp.medium.com/video-anomaly-detection-an-introduction-232bf48c9a8d