Faster abnormality localization and recognition in a secured video bitstream by implementation of video encryption techniques

Mallepogu Sivalakshmi 1, K R Rajendra Prasad 2*, Chigarapalle Shoba Bindu 1

1 Department of Computer Science and Engineering, JNTUA Ananthapur, Andhra Pradesh, India
2 Department of Computer Science and Engineering (CS), Institute of Aeronautical Engineering, Hyderabad, Telangana, India.

Abstract. Cloud video storage uses an encrypted format to protect user data. It means encrypted video processing is an essential part of secured cloud storage. In order to detect suspicious or anomalous behavior, video surveillance must have encrypted cloud access. The primary goals of this research are to estimate parameters and detect abnormalities in an encrypted video bitstream. Various typical properties of video encoding frameworks and format-compliant encryption algorithms are investigated to identify abnormalities in an encrypted video bitstream using format-compliant encryption. The encrypted bitstream is decrypted to get three different kinds of enhancement features: the sizes of macroblocks, partitions of macroblocks, and the magnitude of the motion vector. The identification and localization methods now do not include video decryption or complete decompression. The proposed strategy has been created to implement the video encryption scheme efficiently and is compatible with various video encryption techniques. The experimental findings demonstrate that, in comparison to other methods, the proposed approach provides optimal running time and detection rate performance.

Keywords. Encryption, Video Processing, format-compliant Encryption, Video-Decryption, Anomaly Detection.

1 Introduction

Video surveillance, fraud detection, healthcare, transportation, and industrial automation [1], [2] are just some of the areas that could benefit from efficient abnormality recognition and localization [3] in a bitstream video in the cloud computing area [4]. Other potential applications include industrial automation and healthcare [5]. The method identifies anomalies or odd occurrences within a video stream while simultaneously maintaining the confidentiality of the individuals depicted in the video. It makes it possible to discover anomalies in real-time while at the same time protecting the privacy of individuals whose images are being streamed. The ability to recognize and localize abnormalities in a bitstream...

---

*Corresponding author: krprgm@gmail.com, 01147 (2024) MATEC Web of Conferences https://doi.org/10.1051/matecconf/202439201147
Video is essential for several causes, including the early detection of anomalies, increased security, improved resource management, cost savings, and greater efficiency. In order to extract strange information from encrypted bitstreams, a study of the use of format-compliant video encryption [6] is carried out.

This study uses three different sorts of predicted values, all of which are obtained in some way from the encrypted bitstream and codeword structures. Specifically, the magnitude of the motion vector difference (MVD) [7], the number of bits that make up the macroblock's contents, and the macroblock's partitioning method.

Multiple methods are used to compress a video frame. Independent of other frames, an I-frame (intra-coded image) [8] can perform its intended task. Because it only includes the differences between the two frames, the P-frame (predicted picture) relies on referencing the prior I- or P-frame. The video frame is partitioned into non-overlapping segments for H.264/AVC encoding. The term "slice" refers to a frame segment that is physically isolated from the remainder of the frame and encoded independently from the rest of the frame. A compressed video bitstream will be generated using the entropy-encoded data from the slices.

The information encoded in the slice's entropy can be compared to a bitstream. Determine the location of any particular bit in the order of bits as a function of the series' initial bit. Prediction and correction are commonly used for compression within video encoding frameworks. Microblocks are very small (only a few bits) because most of the background and typical contents can be accurately predicted. The term "anomalous" motion refers to a "unexpected" motion and usually indicates a high rate of motion, which requires more bits in the video bitstream than normal motion. This knowledge is being put to full use in the technique that we have developed for identifying anomalies.

Additionally, the encrypted video bitstream can be utilized to derive information regarding the macroblock partitions. Divining a 16 × 16 MB H.264/AVC file into many sub-MBs is possible. When a microblocks (MB) has more details, it will comprise more sub-MBs. Because of the characteristics of video coding, it is more likely that the sub-MBs that make up an abnormal region in the P-frame will have a small size. MV is a component that can be found in various anomaly detection approaches. However, conventional MV feature extraction algorithms cannot be used with our strategy because the MVs themselves will be encrypted within the encrypted motion picture.

Fig. 1: Proposed Framework Video Learning Model
The video bitstream consists of two components: the MVD (Motion Vector Deviation) and the expected MV (Motion Vector). The MV is calculated by subtracting the observed MV from the expected MV. The MVD magnitude represents the motion in the video. Despite encryption, the video decoder can still read the encrypted bitstream format due to the preservation of codeword length. This is done for security reasons.

Learning requires accurate location or frame-level labels, which are not currently accessible. The technique currently being used employs supervised learning. Motion pictures of varied lengths are split into a predefined number of fragments, each containing the same number of frames. These fragments are then combined to create the final product.

The recommended method assumes that fragments generated from anomaly films contain at least one anomalous fragment, unlike the fragments from regular videos, which only contain typical fragments.

The proposed framework, illustrated in Fig. 1, comprises three stages. In stage 1, features are extracted from the clips using a pre-trained hybrid video model. In stage 2, an attention layer records both long- and short-term dependence. In stage 3, the Efficient Magnitude Video Feature Model (EMVFM) model is used in the proposed anomaly detection model. The temporal feature magnitude, also known as the $l_2$ norm of video snippets, is utilized for anomaly identification. Features with a low magnitude represent typical footage clips.

2 Literature Review

The literature reviews the problem of secured video analytics towards efficient abnormality detection in video surveillance. High-Efficiency Video Coding (HEVC) was developed by Liang Zhao et al. [9]. Segmenting and classifying moving objects from compressed data is essential for intelligent video monitoring [10]. In contrast to H.264/AVC, HEVC possesses more technical coding characteristics that can be applied to a broader range of situations; for example, in the segmentation and classification of moving objects. In this study, authors presented a real-time technique for video surveillance that divides and classifies moving objects using unique features directly obtained from the HEVC compressed domain.

The motion vector (MV) [11] and outlier elimination interpolation are employed as preprocessing procedures. Non-zero MV blocks are clustered together in the connected foreground regions using a four-connectivity component labeling technique [12]. The methodology for the detection of video anomalies based on local statistical aggregates was proposed by V.S. and Chen [13]. Within this methodology, the techniques of the hidden Markov model (HMM) [14] and k-nearest Neighbours (kNN) [15] are utilized. It primarily clarifies that many applications for video surveillance can discover anomalies based on their local spatial-temporal signatures or the fact that the events in the issue take place within a limited amount of time or area. Activities usually function outside this anomalous spatiotemporal zone, allowing us to differentiate these circumstances. They develop a probabilistic framework to explain the local spatiotemporal anomalies we see. The work has the benefit of producing a framework that may be expanded to include other helpful statistics. The limitation of this work is that it is not appropriate for databases that can mislead datasets, which can increase the likelihood of making incorrect predictions.

Another method, Scalable Video Coding (SVC) proposed by Yifeng Zheng et al. [16]. They primarily described a secure system architecture that merges secure deduplication with efficient compression.
3 Proposed Work

The frame-level approach and the localization method make up the second and third levels of the proposed methodology for identifying anomalies. After employing the approach at the frame level used to locate abnormal frames, it is possible to locate abnormal areas by using the localization methodology. Due to the lack of an accurate position or frame-level labeling, the proposed approach relies on supervised learning. Instead, films of varied lengths are divided into an equal number of frames-per-second fragments. The proposed approach was based on the presumption that, unlike snippets generated from conventional videos, which comprised only typical snippets, snippets generated from weird videos featured at least one anomaly snippet. Stage 1 involves feature extraction from the clips using a pre-trained hybrid model. The second stage includes adding an attention layer to the feature in order to capture relevant long-range and short-range temporal relationships. In step 3, the EMVFM model is finally used to find abnormalities in the features acquired in stage 2.
3.1 Detecting anomalies at the frame level

Each frame should be analyzed for further examination of any anomalies. A group of frames can define a video stream. A hybrid swing transformation model is utilized and applied to frame-level images for the training phase. However, it isn't easy to compute self-attention for images compared to other elements in conventional transformer models. In order to tackle this problem, the swing transformer slices images into frames and then only calculates an estimate of the amount of self-attention occurring within each of these frames. Now we will slide the window over the images to more accurately establish the self-attention value of the complete collection of images. In another stage, learning the discriminative representation of normal and abnormal snippets is the primary objective of this stage. It will be accomplished by improving the quality of the feature map that was constructed in an earlier stage. For achieving this objective, an attention layer is utilized since it encodes long-range and short-range dependencies in the temporal domain on the feature map while also focusing the attention of the model on the qualities that are considered to be the most significant. The proposed anomaly detection model uses the Efficient Magnitude Video Feature Model (EMVFM) model. This is a model in which the temporal feature magnitude, also known as the L2 norm of video snippets, is used for anomaly detection. Low-magnitude features represent typical snippets, while high-magnitude features represent abnormal snippets. The model that has been developed assumes that abnormal snippets have a mean feature magnitude that is greater than that of regular snippets.

3.2 Anomaly Region Localization

Anomalies can impact a variety of realms of influence. Some just influence localized regions, while others affect the entire frame. When an irregularity is only present in a limited area, it would be more helpful to identify it at the MB level rather than at the frame level. More processing time is required because a single video can include a significant number of megabytes. On the other hand, the MBs that make up a single frame have a high probability of having unique characteristics when discovered in different regions and having statistical characteristics in common with one another when discovered in the same area.

![Frame segment and Block sequence diagram](image-url)
3.3 Working Principles

3.3.1 Feature Extraction

In order to get started with the feature extraction process, the initial step that needs to be performed is to split the videos up into individual frames (of a size that is, let’s say, M by N).

3.3.2 Attention Layer

It generates output attention feature maps with the format $F \in \mathbb{R}^{A \times B}$ whenever it is given an input feature map $I \in \mathbb{R}^{A \times B}$. The module on the left is called a short-range module, whose purpose is to record temporal and spatial dependencies over short periods. The module on the right is referred to as a long-range module, which is employed in calculating global temporal context.

3.3.3 Anomaly detection

Let $x$ represent the characteristic size of the clips, where $x^+$ indicates an abnormal fragment and $x^-$ indicates a standard fragment. These snippets are acquired by standard videos ($X^+$) and deviant videos ($X^-$). The number of abnormal fragments in the deviant video is $k$, and the model learns by optimizing a score called $score(X^+, X^-)$ that represents the difference between the mean of L2 norm of the top K features from the abnormal and normal bag.

4 Experimental Study

This section illustrates the efficient detection of abnormalities using the proposed real-time video datasets techniques. Benchmarked video snaps are collected from [22], on which experiments are conducted to find the abnormalities.

In this experimental study, two different classes of video are used: video without any abnormality and video with abnormality. These sample videos are shown in Fig. 3 and Fig. 4.

Fig. 3. Video Snap without any Abnormality
Fig. 4. Video Snap with Abnormality

Fig. 5. Graphs that show the accuracy of finding abnormalities in a particular period of time

Fig. 6. Graphs that show the overall accuracy of finding abnormalities in all videos

Fig. 5. shows the accuracy obtained during finding abnormalities at a particular period of time in a video that is considered as a sequence of frames in the experimental of the proposed work. It shows the sequence of a set of video snaps used to find the abnormalities in the video that is taken under testing of abnormalities. Fig. 6 shows the accuracy we got in finding the abnormalities in whole videos that we considered under the training and testing of 40 abnormalities in 40 videos with the proposed work.

5 Conclusion

This paper proposes a method for getting characteristics extraction of a higher quality. It is carried out with a hybrid videoswin transformer model and an attention layer. It is utilized to represent the long- and short-range correlations in the spatial video realm. The proposed research implements an efficient magnitude video feature model (EMVFM) model, and it trains more discriminating characteristics than other machine learning paradigms and is more effectively capable of using aberrant data, making it the best technique for efficient abnormal classifications.
6 Future Scope

Future trials may look at other techniques to minimize the noise in video with excellent features. The work cannot deliver the abnormal classification activities with the typical size of videos. Large-scale videos tend to be accurate training models. In the future, scalable video modeling will be developed to detect anomalies over a large scale of high-quality videos.

References


