IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 October 2023, accepted 23 October 2023, date of publication 1 November 2023, date of current version 8 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329062

RESEARCH ARTICLE

Augmenting the Robustness and Efficiency of Violence Detection Systems for Surveillance and Non-Surveillance Scenarios

MUHAMMAD SHOAIB[1], ASAD ULLAH[1], IRSHAD AHMED ABBASI[2,3,4], (Member, IEEE), FAHAD ALGARNI[4], AND ADNAN SHAHID KHAN[2], (Senior Member, IEEE)

[1]Department of Computer Science and Information Technology, Sarhad University of Science and Information Technology, Peshawar 25000, Pakistan
[2]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan 94300, Malaysia
[3]Department of Computer Science, Faculty of Science and Arts Belqarn, University of Bisha, Sabt Al-Alaya 61985, Saudi Arabia
[4]Faculty of Computing and Information Technology, University of Bisha, Bisha 67714, Saudi Arabia

Corresponding author: Irshad Ahmed Abbasi (aabasy@ub.edu.sa)

This work was supported in part by the Deanship of Scientific Research at the University of Bisha through the Fast-Track Research Support Program; and in part by Universiti Malaysia Sarawak, Malaysia.

ABSTRACT Violence detection holds immense significance in ensuring public safety, security, and law enforcement in various domains. With the increasing availability of video data from surveillance cameras and social media platforms, the need for accurate and efficient violence detection algorithms has become paramount. Automated violence detection systems can aid law enforcement agencies in identifying and responding to violent incidents promptly, thereby preventing potential threats and ensuring public protection. This research focuses on violence detection in large video databases, proposing two keyframe-based models named DeepkeyFrm and AreaDiffKey. The keyframes selection process is critical in violence detection systems, as it reduces computational complexity and enhances accuracy. EvoKeyNet and KFCRNet are the proposed classification models that leverage feature extraction from optimal keyframes. EvoKeyNet utilizes an evolutionary algorithm to select optimal feature attributes, while KFCRNet employs an ensemble of LSTM, Bi-LSTM, and GRU models with a voting scheme. Our key contributions include the development of efficient keyframes selection methods and classification models, addressing the challenge of violence detection in dynamic surveillance scenarios. The proposed models outperform existing methods in terms of accuracy and computational efficiency, with accuracy results as follows: 98.98% (Hockey Fight), 99.29% (Violent Flow), 99% (RLVS), 91% (UCF-Crime), and 91% (ShanghaiTech). The ANOVA and Tukey tests were performed to validate the statistical significance of the differences among all models. The proposed approaches, supported by the statistical tests, pave the way for more effective violence detection systems, holding immense promise for a safer and secure future. As violence detection technology continues to evolve, our research stands as a crucial stepping stone towards achieving improved public safety and security in the face of dynamic challenges.

INDEX TERMS Violence detection, key frames, evolutionary search, statistical test, multimodal CNN.

I. INTRODUCTION

Increased criminal activity in the 21st century has resulted in more loss of life and property [1] compared to other human-centered issues. Intelligent surveillance systems are among the most important methods for detecting abnormal human activities at early stages. These systems can have the ability to automatically detect and generate reports of anomalous human activities, which is vital in maintaining public safety indoors and outdoors. Over the past few decades, many distributed surveillance cameras have been installed in public spaces such as hospitals, prisons, airports, and public parks to guarantee national security [2]. Manually analyzing the large amounts of surveillance footage produced by these

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

cameras is complicated, tedious, error-prone, and costly. Therefore, using computer vision technology to detect abnormal events automatically is effective and efficient. However, it presents many challenges, the most notable of which are different lighting levels, the appearance of the person being photographed, and the distance of the viewpoint from the camera. Therefore, in today's age of technology, intelligent surveillance technology is needed. These technologies can detect abnormal events quickly and accurately and generate a report to alert the concerned authorities [3], [4]. Detecting abnormal scenes in moving or static camera-based recorded surveillance video comprises target overlap, cluttered images, partial or complete occlusion, fixed-pattern noise, low video pixel density, and change management under poor solidity and lighting conditions. Vision sensors that cover many entities, including desired targets acting abnormally and person without disability, make it more difficult to detect anomalies using visual inputs. For example, a runner is considered normal in one scenario (a soccer field) but ambiguous in another (a shopping mall). This makes it difficult to explicitly collect the exceptional samples and define exceptions for the AI model, as exception events are unlimited, rarely occur, and are not well defined. Normal events in surveillance videos are easier to manage than collecting data on abnormal events. Detecting data anomalies involves looking for data points that do not conform to the typical patterns. Pervasive anomalies can be detected by anomalies [5], [6], [7]. The detection of anomalies reveals a wide range of anomalies. Many traditional image descriptor-based approaches [8], [9] have been developed for violent scene detection. However, due to the non-invariant nature of these feature descriptors, such as illumination, translation, rotation, and scale, such systems fail to perform better in challenging environments. Nevertheless, today's systems tend to have tendencies in particular areas, resulting in a limited ability to classify various surveillance anomalies. Anomaly detection, such as the detection of violence in images and videos, is a hot topic of research in AI and is of interest to people with diverse research backgrounds. The AI model capable of learning can be divided into three categories: supervised learning, which requires labeled features, and semi-supervised learning, which requires Partially Labeled features. Clustering, frame reconstruction, and techniques based on future predictions are used for unsupervised anomaly detection [5], [10], [11] to locate anomalies when labels are not present in the training data [5], [10]. These methods do not work well against complex real-world surveillance video data. In this paper, we present a supervised anomaly detection technique as a way to overcome these limitations.

The technique's training set contains both normal and abnormal data to identify both types of anomalies. Weak supervision techniques, in particular, provide video-level labels for normal and abnormal events only in the training sets, trying to solve the anomaly detection problems relatively better. This is so because weak regulation requires less information. The researchers demonstrated in paper [3] that an unsupervised classification is a viable approach to handling an anomaly detection system using weak learners, which can be found here. In multiple instances learning system, a collection of video frames is often called a package, and the individual frame features are considered instances. The next step in the system is to learn the instance-level exception labels based on the bag-level annotations. The concept of anomalies as events that do not conform to the normal behavior of predictions [12] has been the basis for the successful application of semi-supervised techniques, which have been proven to be successful. In running data, events that do not follow a typical pattern are called exceptions.

Many previous studies have made semi-supervised anomaly detection the main research topic. The primary goal of such models is the development of a system or representation which can capture the object's normal motion patterns and visual appearance [13], e.g., the authors of articles [8], [14], [15] used the trajectory of movable objects to represent the hidden patterns of a certain object of interest. Patterns that deviate significantly from the specification are outliers, also known as outliers. Because they only consider visual patterns and ignore the importance of the objective, trajectory-based techniques perform poorly when encountering challenging environments, such as those involving crowded environments. This is because these techniques focus solely on visual patterns. Dictionary learning and sparse encoding are two other video advertising methods that have gained notoriety in recent years [16], [17]. In the proposed schemes, normal scenes are usually encoded into a vocabulary called the input features and treated as exception events. Upon the completion of model training, the model is evaluated with the test data; the events are classified as normal or abnormal events with a minimum false classification rate. The most notable disadvantage of such methods is time complexity for optimally calculating the sparse coefficient with other factors such as weather, lights, etc.

In this research, two keyframe-based models have been proposed for the task of violence detection in large video databases. These models are DeepkeyFrm and AreaDiffKey. The proposed models aim to address the challenge of selecting keyframes, which is a critical step in violence detection systems. The keyframes selection process is necessary to reduce the computational complexity of the system and improve the accuracy of violence detection. The proposed models have been designed to tackle this challenge in different ways. EvoKeyNet is a proposed deep learning-based model that uses an evolutionary algorithm to select optimal features attributes of the keyframes learnable features. The algorithm is based on a fitness function that considers both the quality and diversity of the CNN (Convolutional Neural Network) learnable features attribute. The selected optimal features are then fed into the deep neural network to perform violence classification.

KFCRNet is another proposed deep learning-based model that uses a multimodel CNN to extract features from the keyframes, the features are passed from the evolutionary