A hybrid filter based image classification framework for real-time anomaly detection video databases

Melam Nagaraju, M. Babu Rao

Research Scholar, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Kakinada (JNTUK), Kakinada, melam.nagaraju5810@gmail.com
Professor and Head, Department of Computer Science and Engineering in Gudlavalleru Engineering College, Gudlavalleru, Krishna District, Andhra Pradesh, baburaompd@gmail.com

Abstract— Human anomaly detection has been one of the most promising fields of study in the last few years. Auto-detection of multi-class human anomalies will make it easier to understand more complicated actions and their variations. It's because there are so many features and training images that most multi-class anomaly detection models don't need to deal with noise elimination or figure out how to separate features. These models, on the other hand, use only a few features to look for multi-class anomalies. There are more and more different types of human anomalies. It takes a lot of memory and time to find the multi-class anomaly because it takes a lot of memory and time. In order to improve the process of detecting multiple-class human anomalies, a hybrid multiple feature extraction method is proposed to find the important multiple features in the motion vectors for the classification problem. Non-linear SVM classification is used to make a hybrid convolution neural network framework even better. Using experiments, it was found that the proposed model has a better human anomaly detection rate than the traditional multi-class segmentation models that have been around for a long time.

Index Terms— Anomaly detection, feature extraction, classification algorithm, foreground objects, background objects.

I. Introduction

The segmentation of images is considered the initial and standard operation for the analysis and interpretation of images. In other words, the process of fragmentation of images can be defined as a particular type of fragmentation of images into front and background objects. A segmentation process is primarily aimed at discriminating against the homogeneous regions of a certain image. Let us consider one example where the segmentation process depends upon detecting the maximum homogeneity in the grey level inside the selected regions[1]. All the basic information is transmitted as the outcomes of a particular imaging sensor. The regions generated as outputs of the segmentation process must show standard behavior. The most important issue is associated with
the selection strategy of its foreground object from its background. Segmentation analysis is a method by which data with similar characteristics are grouped into larger analytical units. The first pixels usually have gray levels to distinguish the background pixels[2-4]. Two approaches can widely be classified as thresholding techniques: global approaches and local approaches. These are: In the process of image segmentation, different approaches have been introduced over the years. The basic concepts of adaptive Gaussian smoothing are involved in that approach. A Gaussian kernel variance function is initially constructed according to the sub-regions of the image. Here, the image smoothing occurs with the aid of the adaptive variance Gaussian filter. The level set model is implemented in the subsequent phase for the segmentation of the smoothing components. Finally, to avoid curve vanishing, a condition of convergence is formulated by taking into account the confidence of sub-regions in segmentation. The rest of the paper is organized as follows. Section 2 describes the literature work[ - 6]. Section 3, describes the proposed solution to the filtered based anomaly detection framework. Section 4 describes the experimental results of proposed model to the existing models. Finally, we conclude the paper in section 4.

II. RELATED WORKS

Local features have been used to make a hierarchical Segmentation model [7]. If you use global pattern matching in the hierarchical partitioning algorithm, you may get a noisy pattern in the results. To solve this problem, a partitioning-based hierarchical segmentation was created to look for anomaly patterns using the global pattern matching. This helped solve the problem. [8] said that removing the background is a common way to break up movement in still scenes. It tries to remove the present image pixel by pixel from a reference image to find moving parts. In this case, the first pixels are called. This is called "background evidence," and it can happen after an initial period when images are averaged together to make a new background. Also, morphological post-processing activities like breaking, expanding, and shutting are done after the foreground pixels are marked. This helps to reduce noise and improve the areas that are marked. After a certain amount of time, new images will change the background to match the changes in the scene. There has been a lot of research done on how to build accurate ensemble classifiers in machine learning models. Boosting, Bagging, and stacking are the three main machine learning models that are used to learn ensemble classifiers on very large datasets. Most traditional ensemble classifiers have a high error rate because they have a lot of data and have to be set up the right way [9]. They came up with a boosting learning model called Trada boost to improve the prediction rate on a small training feature space. There is a big problem with this model. As the number of dimensions grows, the true positive rate goes down. [10], an extra model called the TrBagging method is used to look at SCI data. During this method, the weak classifiers are learned based on the information that was used to train them. A forward feature selection method has been used to choose only a few features from the original feature space, instead of all of them. Adaboost (Adaptive Boosting) is a method from the ensemble learning group. It is based on meta learning. Use the base weak classifiers to make the strong classifier even stronger. AdaBoost is an iterative method. In each iteration, a weak base classifier is chosen to keep the model's error rate low. High dimensionality is a big problem
for machine learning models. In order to improve the accuracy of the classification algorithm, most of the approaches use feature selection measures like mutual information, correlation coefficient, rough-set, chi-square test, and so on to choose the best features for the algorithm to look at. [11] have said that the main goal of this algorithm has been to cut down on the amount of detection of stationary front floors. The system used a bootstrap filter to pick and match objects to targets so that it could track many people and find specific classifiers. The merits include effective handling, but at high computer complexity, of false positive detections. The serious disadvantage was that there was no background modeling. Particular grouping minimized the mapping complexity and models for the dynamic scene in the proposed system in order to avoid additional calculations. [12] used a temporal color model to trace the occlusion of several people. The time color-functions combined color values with corresponding weights were used. Size, duration, frequency and adjacency of a color object were determined by weight. This technique can only address partial occlusion in some instances. The proposed MLP approach can effectively deal with any type of occlusion. The proposed MLP-based procedure deals with occlusions using training and bus topology calculation, and results in greater exactness. Intuition and experimental validation have been provided to demonstrate high stability. The advantage of this method provided effective cross-visual awareness and superior public performance. The major downfall of a person was assuming that he was roughly located in the frame. Multiple objects can still be tracked using the proposed MLP-based method. Radial base feature networks have traditionally been utilized as an interpolation technique. RBF networks generally have three different levels: the input layer, a covered layer, and a linear output layer. The variable in the input layer consists of a single neuron and its value is transferred to the hidden layer. The hidden layer consists of a function of the radial basis that calculates the distance between the input neuron and the training neuron. This reduces the computing time dramatically with the additional benefit of avoiding the local minima problem, usually found when simulating standard multi-layer perceptron’s. Upon completion of the background estimate, the moving objects can be detected by background subtraction with the background image obtained. The image must be transformed into a grayed picture to make the separate picture (i.e. the foreground and background separation) a binary picture. A threshold should be compared for transforming a gray image into a binary image. The background of the scene consists of all pixel values below this threshold. Several approaches are now available to automatically detect, segment and track video sequences of objects. Fundamental movement detection algorithms compare static background with a pixel-by-pixel current video scene framework. In BS, a background model is created and this model is comparable with the existing framework to detect a zone where significant variation takes place. In short, the purpose of the subtraction is to separate moving object(s) in the foreground from the background [13]. They derived knowledge about space-time from the contours of images subtracted from the background. They identified the action by selecting the points of interest on the surface of the form, such as ridge, hill, valley, saddle, and pit points. In a video sequence, [14] implemented a motion context descriptor to identify motion area. Using a polar grid, it is computed over the moti.cm area and each cell in the grid is then represented with the
SIFT descriptor histogram. With the probabilistic latent semantic analysis model (PLSA), the author uses the help vector machine (SVM) for classification. In order to distinguish human activity, optical flow and volumetric are features were extracted and matched with the action template. For human action identification, the key downside of holistic representation is that it involves accurate silhouette extraction and shape, usually obtained through segmentation. The precision of the classification of these techniques depends on the outcome of segmentation and it is very difficult to achieve accurate segmentation of videos from the real world. They improved the 20-Harris corner detector to include a 30-Harris corner video sequence detector. By using the own values of the second order moment matrix at each point in a video series, it obtains corner points[15]. The final interest points of the second order moment matrix are obtained as the local limit. Furthermore, the local extreme of the spatio-temporal Laplacian operator selects the spatio-temporal scales. By adding the quadrature pair of separable linear filters, the interest points are extracted. They spatially and temporally applied the Gaussian smoothening filter and the one-dimensional Gabor filter. They expanded the Hessian salience detector to the spatio-temporal interest point detector Hessian3D. It makes use of the 30 Hessian matrix determinant to calculate saliency in a video. Using an integral video data structure, they speed up the calculations. Over space , time, and various scales, a non-maximum suppression algorithm chooses joint extremes. Instead of the local community, they used global knowledge to assess salient characteristics. The movement is synthesized by this model and the dominant regions in motion are established. As a linear transform, the dynamic model is approximated. Through non-negative matrix factorization, a sub-space representation is computed. A large number of points of interest and related video features are extracted by this detector. The performance of dense sampling on the KTH dataset, UCF dataset and Hollywood2 dataset was compared with the Gabor detector, Harris-3D and Hessian3D. Harris 3D has been found to perform better on the KTH dataset, while dense sampling on the UCF dataset offers the highest classification accuracy. Dense sampling offers best output with reference video for the Hollywood2 dataset, while Harris3D achieves best results using a full resolution video. The local feature descriptor uses motion and shape details to identify the local area surrounding the point of interest and paths. [16] suggested that local descriptors characterize the gradient, optical flow and normalized pixel value of the extracted cuboid. Using three different strategies, they developed a function vector: histogram the local neighborhood, flattering the local neighborhood and dividing the local area into grids, and creating a histogram for each grid cell. Then they applied PCA to all the techniques to reduce the final feature vector dimension. On the gradient-based descriptor, they obtained the highest classification precision[17]. They defined local motion information and PCA was applied to spatio-temporal gradient and optical flow. For video representation, they extended the HOG image descriptor. By splitting the local area into a spatio-temporal grid, the structural information was embedded in the descriptor, and then a histogram was computed for each grid. Finally, to form a final descriptor, histograms are normalized and concatenated. This describes the local appearance and form, measured and quantified by the gradient vectors based on orientation in histogram bins. By using optical flow, the HOF descriptor extracts the local motion information[18]. Based on the direction of the optical field into histogram bins, optical flow vectors are quantized.
Gaussian, based at a given key point, weights each pixel in the spatio-temporal region surrounding the key point and votes in a 3-dimensional grid of histograms of oriented gradients. For quantizing the oriented gradient, spherical coordinates are used and are split into 8X4 histogram bins[19]. Methods based on deep learning are designed to learn several layers of raw data representation and abstraction. It automates the process of extracting, representing and classifying image and video data for function length. For action representation and recognition, it makes use of trainable feature extractors and computational models with several processing layers. In order to minimize the search space by pooling and weight-sharing layers, the convolution networks use the image structure correctly and are therefore robust against spatial and scale variation. A 3D convolution network was proposed in [20], using 3D kernel filters to capture spatial and temporal information. During the network training, it utilizes optical flow as additional knowledge. They empirically demonstrates that 3D convolution networks perform better than 2D convolution networks. In general, the 3D convolutional network architecture has a firm temporal structure and is therefore unable to capture macro motion at various speeds in different behavior. It accepts consecutive video frames and processes them across the same collection of layers. All responses are collected from the temporal domain to acquire the video descriptor and fed to the fully connected network. For generic video descriptors called C3D, the 3D convolution network. It is accomplished by averaging the outputs of the C3D network's first fully connected layer. The authors argue that the network outperforms the variable temporal depth filter with 3X3X3 homogeneous filters. They also included temporal domain versatility by adding a layer of 3D pooling.

III. Proposed Model

Figure 1 shows the framework for the multi-class anomaly detection model that we've come up with. Multi-class feature extraction, multi-class feature ranking, multi-class segmentation, and multi-class classification model are some of the steps in the proposed model that are done in four stages. In the first step, a hybrid ensemble feature extraction model is made and used on the training data. Use a multi-class feature ranking method to find the feature ranking process for C3D in phase two. In the third phase, a multi-class feature segmentation is used to remove the noise from the edges of the picture. This segmentation process is used to remove the parts of the edges that are too many times divided. All of this is done in the fourth phase, when a multi-class SVM classification model is used on the best C3D features to predict anomalies.

Feature extraction for C3D model: Ensemble feature extraction models for multi-class anomaly detection
Gaussian Entropy Feature extraction Measure
Input: Training frames T, Features space FS.
1. For each training image frame F in the training data.
2. do
3. Compute the proposed gaussian probability measure to find the key object in the given training data.

\[
G_{Pr}(F) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{(F-I(i) - \mu(I(i)))^2}{2\sigma(I(i))^2}}
\]

Where \(G_{Pr}(F)\) defines the gaussian probability measure of each image block features \(F\).

\(F(I(i))\) defines histogram value of ith frame

**Non-linear exponential gaussian estimator**

In this filter, a non-linear exponential gaussian estimation is done on the training image frames that are used. A histogram is made for each block in the ith video frame in order to make the kernel filter in the C3D framework.

Non-linear gaussian estimation using histogram H is computed as:

\[
\text{Nonlinear gaussianEstimator} = NGE(F) = \sum_{i=0}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{H(I(i)) - \mu(H)}{2\sigma(H)^2}} \quad ----(1)
\]

**Figure 1: Proposed Feature extraction based C3D-SVM framework**
There are kernels in this framework that are made up of hybrid ensemble feature extraction models. They filter the motion vectors in the training frames by looking for certain things in them. In this framework, a 3x3x3 ensemble kernel estimator is used as a filter. It has 5 pooling layers and 2 fully connected layers, and it is called a filter. It is part of the C3D framework so that the noise in the process of finding features can be cut out. There are eight converting layers, five max pooling layers, two fully connected layers, and the softmax output layer at the end of C3D. Finally, in the C3D framework, an ensemble multi-class SVM is proposed to predict anomalies on the features, which is a good idea.

2. Normalized correlation based segmentation algorithm

In this work, a hybrid and faster method is used to find the edges that have been over-segmented during the feature ranking and classification process. Multi-class anomaly detection and masking is all about threshold-based segmentation, which is one of the most important things to know. Most of these methods work better when there is no noise distortion on the edges of the video frames you want to look at. These algorithms, on the other hand, can't deal with noisy images. Normalized A list of the feature selection measures used to find the ensemble ranking feature selection process is shown below

\[
\text{Normalized Correlated Noise Index} = \alpha = 2 \sum_{p,q=0}^{N-1} \frac{Pr(F_{p,q})}{MC_k} \frac{(p-\mu)(q-\mu)}{|\mu-\sigma|^2}
\]

\[
\text{Normalized Homogeneity Noise Index} = \beta = \frac{1}{N} \sum_{i=0}^{N-1} \frac{Pr(F_{i,j})}{MC_k} \frac{1}{1+(i-j)^2}
\]

\[
\text{Normalized Feature noise selection index} = \text{NFNSI} = \text{Max} \{\alpha, \beta\}
\]

An efficient algorithm that completely depends upon the motion similarity score in order to carry out the block wise graph segmentation process.

\[
\min H(V, C) = \sum_{i=1}^{M} \sum_{j=1}^{N} (\mu_{p_j}) \| X_q - C_p \|^2
\]

subject to \( \sum_{i=1}^{N} \mu_{p_i} = 1, 0 \leq \mu_{p_i} \leq 1 \)

where, \( V \) and \( C \) are graph vertices and central regions, \( X_q \) represents each pixel in the region, \( C_p \) is the central pixel of each region, \( \mu \) is the mean of the region.

3. Multi-class C3D-SVM Classification model

Convolution Neural Networks (CNN) is a deep machine learning model focused on enhanced artificial neural feed-forward networks. The CNN idea was inspired by the discovery of a visual system in the brain known as the visual cortex. The visual cortex these cells were constructed by several cells called receptive fields and responsible for deciding the light in a small, overlapping and sub-region of the visual field. Local filters represented by these cells, bigger receptive fields presented in the more complex cells, over the input space. One or more convolution layers concatenated with a regular neural network consisting of one or more hidden layers called
completely connected (FC) layers can be present in CNN architecture. The new FC layer is the output layer that determines the making of the diction. The convolution layers are normally accompanied by a sub-sampling layer and a nonlinear layer. The role of visual cortex cells in the brain corresponds to the function of the convolution layer in CNN. Convolution layers are used to create feature maps of the input image as a feature extractor. This means that the CNN convolution layer is synonymous with local filters added to the input space and with the coefficient of the filter kernel calculated during the training process. A series of primitive patterns; the low-level features represented in the input images, such as edges and lines; the first convolution layer can be removed. By integrating these primary features, such as corners, the second convolution layer defines patterns of patterns. By integrating these secondary features obtained from the previous layer, the third convolution layer extracts higher-level characteristics based on detecting patterns of certain patterns, and so on. A non-linear "trigger" function used to differentiate signal apart from useful features on each hidden layer in a general neural network as well as CNNs. The proposed model used a max pooling feature that intersected the previous layer with a neuron cluster and used the maximum value as the output. These neuron clusters are generated by dividing the input images into non-overlapping two-dimensional spaces and selecting each space considered to be the cluster and the maximum value. Two completely connected layers (dense layers) followed by one output layer (decision making) layer reflect the classification phase of the proposed model. Fully-connected layers are a typical neural network that links all neurons in one layer to all neurons in the next layer. As a weighting sum of the previous layer of features, a particular target output result can be interpreted mathematically. The input of each completely linked layer is a vector of values that can be calculated by multiplying the output of the height and width form of the last max pooling layer with the depth of the neurons (the number of filters used in the last layer of the convolution).

**IV. EXPERIMENTAL RESULTS**

In this work, different training anomaly datasets are used to simulate in Amazon AWS environment. In this study, different types of anomaly datasets are used as training videos for feature extraction and anomaly prediction, so they can learn how to do that. It looks at different video anomaly datasets, such as dataset1 from http://cvrc.ece.utexas.edu and dataset2 from https://svip-lab.github.io/datasets/campus_dataset.html. These datasets are used as training and testing evaluations, so they can help you learn how to do things. Video anomaly detection in real time is found with a new model used in this work. To predict multi-class test data, different types of anomaly classes are used. They are then used with the new filtering-based C3D framework to do so. All of the videos are first trained with the proposed C3D-based multi-class classification model. When you use this framework, you get rid of all the multi-class features and only keep the best ones so you can predict the class label with high computational accuracy and a low rate of mistakes. In this section, the runtime of the proposed models, feature extraction measures, classification accuracy, recall, precision, and error rates are calculated and compared to the conventional models, as well as how
well the proposed models work.

Figure 2: Sample input frames extracted from the source[1]
Figure 2, represents the sample input video frames of the data source[]. Initially, the input data source is used to extract the frames for the filtering process and anomaly classification prediction. In the figure 2, different types of class actions are observed for the anomaly detection in the realtime video.

Class type 1: Sample frames in Video

Figure 3: Anomaly detection of the proposed model on the data source[].
Figure 3, describes the anomaly detection of the filtered based multi-class C3D framework on the input data source[]. In this figure, all the input video frames are trained using the multi-class deep classification process for anomaly prediction.
Figure 4: Sample input frames extracted from the source[1]
Figure 4, represents the sample input video frames of the data source[]. Initially, the input data source is used to extract the frames for the filtering process and anomaly classification prediction. In the figure 4, different types of class actions are observed for the anomaly detection in the realtime video.

Figure 5: Anomaly detection of the proposed model on the data source[].
Figure 5, describes the anomaly detection of the filtered based multi-class C3D framework on the input data source[]. In this figure, all the input video frames are trained using the multi-class deep classification process for anomaly prediction.

Figure 6: Sample input frames extracted from the source[1]
Figure 6, represents the sample input video frames of the data source[]. Initially, the input data source are filtered using the proposed filtering methods for key frames extraction and anomaly classification prediction. In the figure 6, different types of class actions are observed for the anomaly detection in the realtime video.

Figure 7: Anomaly detection of the proposed model on the data source[]. Figure 7, describes the anomaly detection of the filtered based multi-class C3D framework on the input data source[]. In this figure, all the input video frames are trained using the multi-class deep classification process for anomaly prediction.

Figure 6: Comparative analysis of multi-class anomaly prediction model to the conventional models using feature extraction runtime(ms). Figure 6, represents the performance of the proposed multi-class anomaly detection to the conventional feature selection models using runtime analysis(ms). These classification models are trained and tested on C3D framework. As represented in the figure, proposed approach has better feature extraction runtime than the conventional approaches.

IV. Conclusion

Video anomaly detection is an important part of real-time surveillance systems, and it's very
important. They use fixed feature sizes and computational memory in most of the old-fashioned ways to find out about things that don't seem right. These models, on the other hand, use only a few features to look for multi-class anomalies. There are more and more different types of human anomalies. It takes a lot of memory and time to find the multi-class anomaly because it takes a lot of memory and time. An advanced multi-class segmentation-based classification model is made and used to improve the process of detecting multiple types of human anomalies. This model is used on the different databases that track different types of human anomaly actions. In the proposed model, a hybrid filter-based C3D framework is used to find the most important features from a lot of different types of human actions. Hybrid multi-class-based anomaly detection is proposed in this work to improve the rate at which anomalies are found in real-time video datasets, which can be hard to find. In terms of accuracy, precision, and recall, the new multi-class anomaly detection model is better than the old models. It also takes less time to run. A cloud computing environment could be used in the future to do more work on big video databases in parallel.

References


http://www.webology.org/


