CHAPTER

VIOLENCE DETECTION IN AUTOMATED VIDEO SURVEILLANCE: RECENT TRENDS AND COMPARATIVE STUDIES 11

S. Roshan, G. Srivathsan, K. Deepak and S. Chandrakala

Department of Computer Science, School of Computing, SASTRA Deemed to be University, Thanjavur, India

CHAPTER OUTLINE

11.1 Introduction 157
11.2 Feature Descriptors 159
11.2.1 Histogram of Oriented Gradients 159
11.2.2 Space–Time Interest Points 159
11.2.3 Histogram of Oriented Optical Flow 159
11.2.4 Violence Flow Descriptor 160
11.3 Modeling Techniques 160
11.3.1 Supervised Models 160
11.3.2 Unsupervised Models 162
11.4 Experimental Study and Result Analysis 165
11.4.1 Data Sets 166
11.4.2 Comparative Study on Related Work 166
11.4.3 Our Baseline Study 168
11.5 Conclusion 169
References 170

11.1 INTRODUCTION

The rapid growth in the amount of video data has led to the increasing need for surveillance and anomaly detection. Such anomalous events rarely occur as compared with normal activities. Therefore to lessen the waste of labor and time, developing automated video surveillance systems for anomaly detection has become the need of the hour. Detection of abnormalities in videos is a challenging task as the definition of anomaly can be ambiguous and vaguely defined. They vary widely based on the circumstances and the situations in which they occur. For example, riding a bicycle in a regular pathway is a normal activity, but doing the same in a walk-only lane should be flagged as anomalous. The irregular internal occlusion is a notable yet challenging feature to

The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems.
DOI: https://doi.org/10.1016/B978-0-12-816385-6.00011-8

describe the anomalous behavior. In addition, representation of video data and its modeling induce more difficulty due to its high dimensionality, noise, and highly varying events and interactions. Other challenges include illumination variations, viewpoint changes, camera motions, and so on.

One of the significant aspects of anomaly detection includes violence recognition and detection. The increase in threats to security around the world makes the use of video cameras to monitor people necessary, and thereby early detection and recognition of these violent activities could greatly reduce these risks. The modeling techniques used for anomaly or violent detection can be broadly classified as shallow and deep models. The main objective of our paper is to perform a comparative study of the above-mentioned models.

Shallow modeling techniques are those that are not capable of learning features on their own but rather features extracted using handcrafted methods must be provided to a shallow network for their classification. A shallow network can be classifier models like support vector machine (SVM), artificial neural network (ANN) with one hidden layer, and so on. These models are best suited for supervised learning, which in the given data should be well labeled. The main drawback of this modeling technique is that they do not adapt to pattern changes automatically. Also the labeling process can be manually intensive. Lloyd et al. [1] have proposed a real-time descriptor that models crowd dynamics for anomaly detection by encoding changes in crowd texture using temporal summaries of gray-level cooccurrence matrix features, in which k-fold cross validation was performed for training a random forest classifier. Their proposed method outperforms the state-of-the-art results over the UMN, UCF, and Violent Flows (ViF) data sets. Similarly, Bilinski and Bremond [2] have used an extension of improved Fisher vectors (IFVs), which allows the videos to be represented using both local features and their spatio-temporal positions for violence recognition and detection. Their results have shown significant improvement in four publicly available standard benchmark data sets.

In contrast to shallow models, most of the deep models do not require a separate feature extractor, as they are based on the feature learning technique, which is that they learn their own features from the given data and classify based on them. In addition, apart from end-to-end learning, the above extracted features can be given as input to the SVMs and other shallow model classifiers. Another way to implement deep models is by using the features from handcrafted feature descriptors and providing it to a deep classifier. These models work on both supervised and unsupervised learning-based methods but are better suited for the latter. Even though they work with unlabeled data, they require high volumes of data and computational power. Chong and Tay [3] propose a convolutional spatio-temporal autoencoder to learn the regular patterns in the training videos for anomaly detection. Even though the model can detect abnormal events and is robust to noise, depending on how complex the activity is, more false alarms may occur. One other work on this model is proposed by Sudhakaran and Lanz [4], in which a convolutional long short-term memory (CLSTM) is used to train a model for violence detection. On comparing this method with other state-of-the-art techniques, their proposed method shows a promising result on the used data sets. A general system for abnormality or violent detection is shown in Fig. 11.1.

In this chapter, we wish to compare and analyze the above shallow and deep models based on their performance. The content of the chapter is as follows: Section 11.2 presents the recent and promising feature detectors used in anomaly and violence detection tasks, Section 11.3 discusses recent works in anomaly detections and a few methods that have been proven to be promising for our task, and the experimentation and analysis part of this chapter is dealt in Section 11.4.

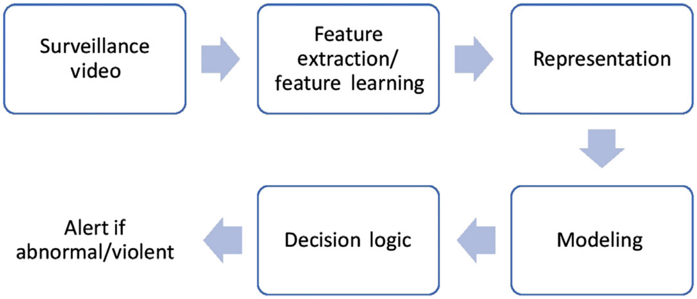
Surveillance video
Feature extraction/ feature learning
Representation
Modeling
Decision logic
Alert if abnormal/violent

FIGURE 11.1

Overview of violence detection system.

11.2 FEATURE DESCRIPTORS

This section discusses about the feature descriptors used in our studies and also other recent state-of-the-art descriptors.

11.2.1 HISTOGRAM OF ORIENTED GRADIENTS

Histogram of oriented gradients (HOGs) is a feature descriptor for object detection and localization, which can compete with the performance given by deep neural networks. In HOG, the distribution of the directions of the gradients is used as features. This is due to the fact that edges and corners have high variations in intensities, and hence, calculating the gradient along with the directions can help detect this information from the image.

11.2.2 SPACE—TIME INTEREST POINTS

By extending the Harris detector, Laptev and Lindeberg [5] and Laptev [6] proposed the space—time interest point (STIP) detector. After extracting the points with large gradient magnitude with the help of a 3D Harris corner detector, a second-moment matrix is computed for each spatio-temporal interest points. The features obtained from this descriptor are used to characterize the spatio-temporal, local motion, and appearance information in volumes.

11.2.3 HISTOGRAM OF ORIENTED OPTICAL FLOW

Due to the relative motion between an observer (camera) and a scene (image, video), a pattern of apparent motion of objects, surfaces, and edges is created. This is called as *Optical Flow*. Histogram of oriented optical flow (HOF) [7] is a feature based on the optical flow that represents the sequence of actions at each instance of time. It is scale-invariant and independent of the direction of motion.

11.2.4 VIOLENCE FLOW DESCRIPTOR

One important feature descriptor is the violence flow, which uses the frequencies of quantized values in a vectorized form. This is different from other descriptors in a way that, rather than considering magnitudes of temporal information, the comparison of the magnitudes is taken for each as it gives much more meaningful measures in terms of the predecessor frame [8]. Instead of using local appearances, the similarities between flow-magnitudes with respect to time are considered.

11.3 MODELING TECHNIQUES

We divide the modeling techniques as supervised and unsupervised. In supervised, the training data contain both normal and anomalous videos, while unsupervised training data contain only normal videos.

11.3.1 SUPERVISED MODELS

11.3.1.1 Shallow models

There were many works carried out based on shallow models with simple handcrafted features given as input to a classifier. One such work was done by Wang and Snoussi [9], in which a histogram of optical flow orientation was introduced as a descriptor that was then fed to a one-class SVM for classification. Further, Zhang et al. [10] proposed an algorithm that used motion-improved Weber local descriptor (MoIWLD) for capturing low-level features and then gave it to a sparse-representation-based classifier. The proposed approach showed superior performance on three benchmark data sets for violence detection.

11.3.1.1.1 Support vector machine

All the data points that are nearest to the hyperplane, which on altering changes the position of the dividing hyperplane, are called *support vector.* A hyperplane is a plane of dimension one less than the dimension of data space, which divides the classes of data. SVM is a learning algorithm mainly used on classification problems, which considers the data as support vectors and generates a hyperplane to classify them. There are three major kernels used in an SVM: linear, polynomial, radial basis function (RBF). The linear kernel is useful when the data are linearly separable, whereas the polynomial kernel is more suitable for data that can be separated by a curve of polynomial degree. The RBF kernel is the one that uses the squared Euclidean distance between two vectors to generate the hyperplane. Hassner et al. [8] have represented the change in flow-vector magnitudes using the ViF descriptor and detected violence using a linear SVM.

11.3.1.2 Deep models

Recently, the approach of deep learning models in computer vision and anomaly detection has been of great significance. In a work done by Ionescu et al. [11], they have differentiated two consecutive video sequences by using a binary classifier, which is trained iteratively. At each step, the classifier removes the most discriminant features, thereby helping the classifier to discriminate them more effectively. Another work done by Tran and Hogg [12] uses a convolutional autoencoder

(CAE), which extracts motion-feature, encodes it, and provides as input to a one-class SVM. To obtain a sparsity of higher degree, a winner-take-all step is brought in after the encoding layer. Further inspired by the strong feature learning ability of the convolutional neural networks (CNNs), Smeureanu et al. [13] extracted deep learning features using a pretrained CNN and an SVM for classification.

11.3.1.2.1 Artificial neural networks

ANNs [14] or simply neural networks are one of the main methods used for classification and are inspired from the working of the human brain. An ANN has one input layer, one output layer, and one or more hidden layers. More and more hidden layers are used for learning more complex features. The architecture of an ANN is shown in Fig. 11.2. Each node in a layer has a vector of weights and an activation function through which data are transmitted for further learning. There are two main phases in the learning process of a neural network: forward and backward propagation. When the training data are fed into the network, it calculates the predicted output and compares it with the true output. An error is generated in the output layer based on this comparison, which is transmitted to the previous layer. With respect to this error received by each layer, the weights of each node are tuned.

11.3.1.2.2 Convolutional neural networks

Similar to a neural network, CNN [15] also receives inputs through layers and has nodes through which this information is passed through. But its layers are more specialized and can accept volumes of data (image and video) unlike a simple neural network. This network comprises of four types of layers: convolution, ReLu, pooling, and fully connected (FC). In the convolution layer, a filter or a kernel is slid over the volume and convolution operation is applied to obtain an activation

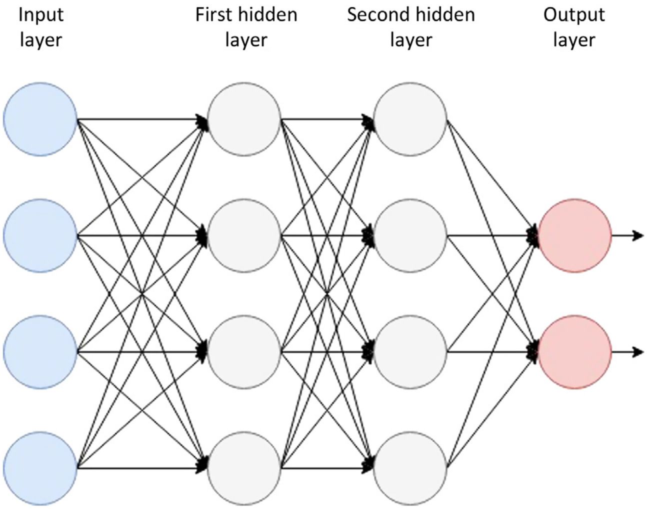

FIGURE 11.2

A simple ANN architecture.

map. Then this map is passed to the ReLu layer to increase nonlinearity. The resultant volume is fed to the pooling layer, which is used to capture the important features from the previous layers. The pooling layer can be either a maxed-pooling or an averaged pooling layer. The matrix is then flattened to obtain a one-dimensional column vector, which is then fed to the FC layer. The FC layer is used for classification purpose. One of the most significant CNNs is the AlexNet whose architecture is given in Fig. 11.3.

11.3.1.2.3 Long short-term memory

Long short-term memory (LSTM) [16] networks are a special kind of recurrent neural networks that are capable of selectively remembering patterns for long duration of time. It is an ideal choice to model sequential data and hence used to learn complex dynamics of human activity. The long-term memory is called the cell state. Due to the recursive nature of the cells, previous information is stored within it. The forget gate placed below the cell state is used to modify the cell states. The forget gate outputs values saying which information to forget by multiplying 0 to a position in the matrix. If the output of the forget gate is 1, the information is kept in the cell. The input gates determine which information should enter the cell states. Finally, the output gate tells which information should be passed on to the next hidden state.

Two of the important variations for the LSTM model are deep LSTM (DLSTM) and CLSTM. DLSTM differs from the general LSTM in the number of layers the model contains. A single-layer LSTM will not be able to obtain well-defined temporal information. However, when more layers are stacked in the LSTM model, it will be able to acquire better temporal features, and hence will be more suitable in capturing motion in the time dimension [16]. In CLSTM, the data are first passed through convolutional layers, which ensure in capturing the spatial features, as shown in Fig. 11.4. The output from the CNN is provided to the LSTM, which will get the temporal features, and hence, the model will capture a motion with respect to both space and time [4]. These two variants can also be combined to give a convolutional deep LSTM, where the outputs from a CNN are given to a multilayer stacked LSTM [16], which is guaranteed to provide a better result at the cost of increased computational complexity.

11.3.2 UNSUPERVISED MODELS

11.3.2.1 Shallow models

A work done by Xiao et al. [18] employed a spatio-temporal pyramid, which captured the spatial and temporal continuities and also used a local coordinate factorization to tell whether a video is anomalous. Cheng et al. [19] presented a method with hierarchical feature representation to detect

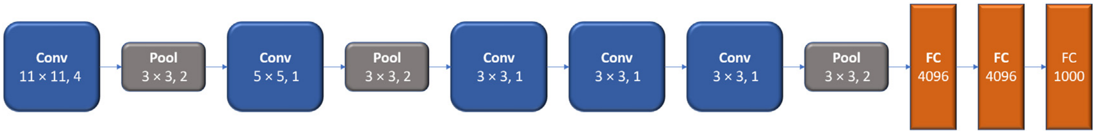

FIGURE 11.3

Illustration of the AlexNet architecture used for image recognition [15].

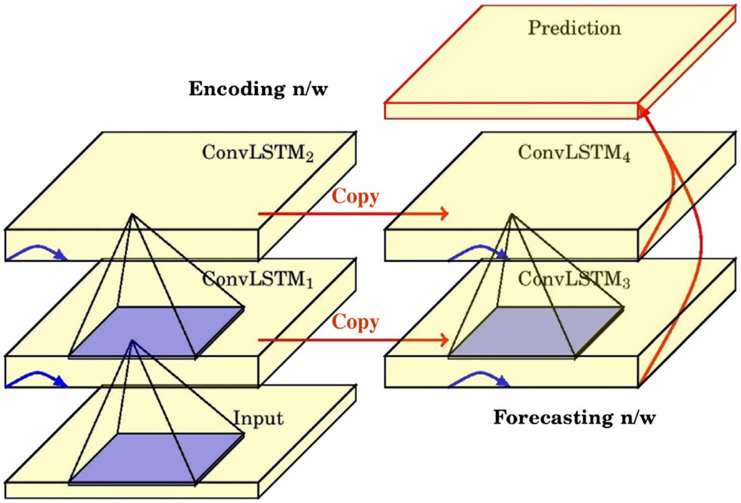

FIGURE 11.4

Structure of a convolutional long short-term memory [17].

local and global anomalies simultaneously by finding the relations between nearby sparse spatio-temporal interest points that were modeled by a Gaussian process regression. A popular work by Bermejo et al. [20] introduced the Hockey Fight data set, where violence detection was performed by taking spatio-temporal interest points (STIP) and motion scale invariant feature transform (MoSIFT) as action descriptors and bag-of-words (BoW) for fight detection. Besides, Leyva et al. [21] used the Gaussian mixture model (GMM), Markov chains, and BoW to prepare a compact feature set through which abnormal events are detected.

11.3.2.1.1 Principal component analysis

Principle component analysis is a form of representation learning model, which reconstructs the data to a lower dimension from the given training data and learns by reducing the reconstructed error. This is also used for dimensionality reductions and feature extractions, since principal component analysis (PCA) is known for its ability to extract the important features and still maintain the integrity of the original data. It computes eigenvectors by finding the covariance matrix of the standardized data. These vectors provide the variance directions and help in reconstruction.

For videos, PCA is used for modeling the spatial correlations between each pixel of a frame from its corresponding vector. In the case of anomaly or violence detection, the vector obtained will be of lower dimension and this captures the anomalous behavior. As each frame is associated with an optical flow value, this can be used for evaluating the reconstruction error. Kim and Grauman [22] used a probabilistic principal component analyzer, which captured the typical optical flow and, thereby, also learning normal patterns. The complex and costly step in this model is the optical flow estimation.

11.3.2.2 Deep models

Huang et al. [23] extracted low-level features including visual, motion-map, and energy features. Also mid-level features were extracted using a restricted Boltzmann machine (RBM) and deep representations of the crowd patterns were learned for the detection of unusual events. Recently, Sultani et al. [24] have introduced a new large data set comprising of surveillance videos. They have performed anomaly detection on this data set by segregating the normal and abnormal videos into bags and considering video sequences as instances for multiple instance learning. Another recent paper includes a method proposed by Ravanbakhsh et al. [25] uses generative adversarial nets (GANs) trained on normal frames and then used for abnormality detection. Likewise, Vu et al. [26] proposed a method where data representation was learned using an RBM followed by the reconstruction of the data. Based on the reconstruction errors, abnormal events were detected. In addition to the above methods, considering the importance of violence detection in video surveillance, Zhou et al. [27] trained FightNet by using image acceleration field as their input modal, which helps in capturing better motion features.

11.3.2.2.1 Generative adversarial network

GAN is a model that is generative in nature as it uses joint probability distribution. GAN comprises of a generator and a discriminator. A generator constructs a fake sample from the given noisy training data. This fake sample is fed along with the stream of other training samples to the discriminator. The discriminator is similar to that of a binary classifier and classifies the training data as real or fake by assigning a probability to it. The generator is said to train on mapping the training data distribution and the discriminator trains on maximizing probability of assigning "real" label to the real training samples.

GANs can be easily used on videos for anomaly and violence detection by using the frames as training data. They evaluate a probability density distribution on the training set, which contains no anomalies, and provide an anomaly score that is the probability whether the sample is from the generator and thereby classifying it as an anomaly. GANs achieve this implicitly by minimizing the distance between the generative model and the training data distribution without the use of a parametric loss function. The mapping in the generator is done by transforming the image domain of the frames to a latent representation space. The loss from the discriminator is used in the backpropagation process of both the generator to generate images similar to the training samples and in the discriminator to classify the samples better. Ravanbakhsh et al. [28] used a modified version of GAN to produce the state-of-the-art results. They proposed a cross-channel GAN, as shown in Fig. 11.5, where the generator network is split into two parts: one to generate optical flow from frames and another to generate frames from the optical flow. The discriminator trains on both the generations, and hence, their method modeled a spatio-temporal correlation among the channels for better predictions.

11.3.2.2.2 Autoencoders

Autoencoders are alternatives to PCA used for the purpose of dimensionality reduction by decreasing the reconstruction error on the training data. It is a neural network, which is trained by backward propagation. It performs a linear pointwise transform of the input using transformation

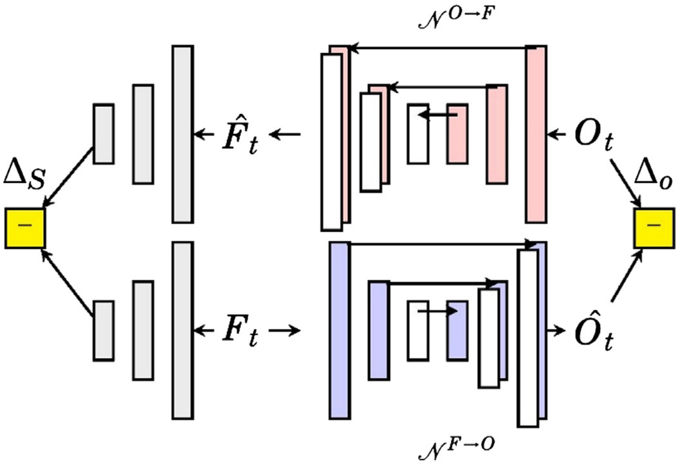

FIGURE 11.5

Cross-channel mechanism used in generative adversarial net for abnormality detection [17].

functions like ReLU or Sigmoid. Two of the main types of autoencoders are CAEs and 3D autoencoder.

Convolutional autoencoder. The input signal is viewed as a signal that is decomposed as the sum of other signals by a normal autoencoder. This decomposition is made explicit by CAEs. CAEs are a type of CNNs. However, instead of manually assigning filters, we let the model learn optimal filters that minimize the reconstruction error. These filters can then be used to extract features from any input. Therefore CAEs are general-purpose feature extractors, which are trained only to learn filters capable of extracting features that can be used to reconstruct the input.

In a work done by Hasan et al. [29], an input sequence of frames from a trained video set was reconstructed by using a deep CAE. This is otherwise called as a spatio-temporal stacked frame autoencoder (STSAE). The STSAE stacks the frame sequence with each frame treated as a different channel in the input layer to a CAE. The architecture of the CAE and that of a stacked autoencoder is depicted in Fig. 11.6 and Fig. 11.7).

3D Autoencoder. As discussed in [30], while 2D ConvNets are appropriate for image recognition and detection tasks, they are incapable of capturing the temporal features of consecutive frames for video analysis tasks. For this purpose, 3D convolutional architectures, depicted in Fig. 11.8, are used in the form of autoencoders. The 3D convolutional feature maps are encoded by the 3D autoencoder to obtain representations, which are invariant to spatio-temporal changes.

11.4 EXPERIMENTAL STUDY AND RESULT ANALYSIS

In this chapter, having security as a concern, we have studied extensively on violence recognition as it is regarded to be the most important section in anomaly detection. Although appearance features are prominently used, motion features have proven to be more effective in violence detection

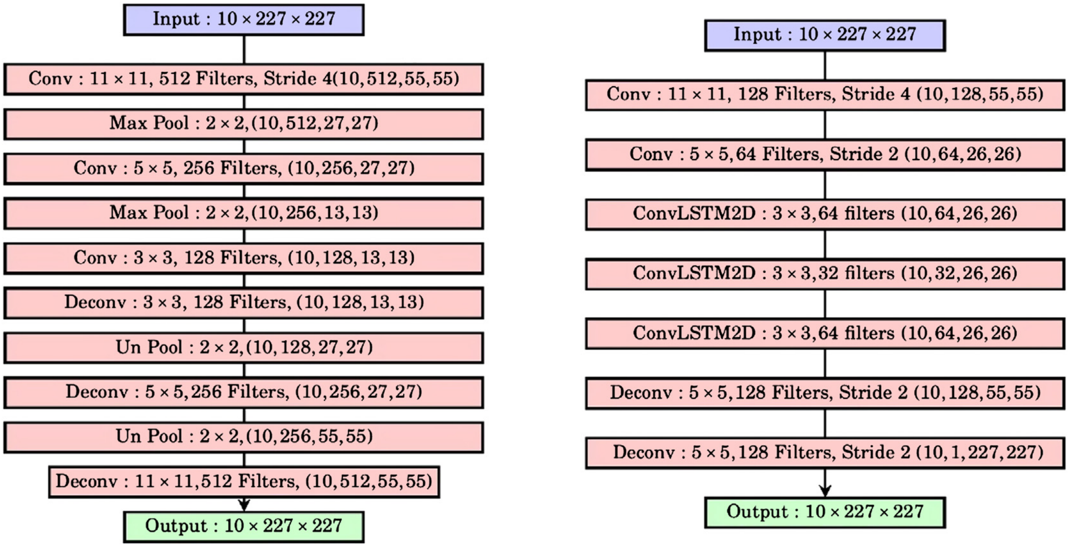
Input : 10 × 227 × 227
Conv : 11 × 11, 512 Filters, Stride 4(10,512,55,55)
Max Pool : 2 × 2,(10,512,27,27)
Conv : 5 × 5, 256 Filters, (10,256,27,27)
Max Pool : 2 × 2,(10,256,13,13)
Conv : 3 × 3, 128 Filters, (10,128,13,13)
Deconv : 3 × 3, 128 Filters, (10,128,13,13)
Un Pool : 2 × 2,(10,128,27,27)
Deconv : 5 × 5,256 Filters, (10,256,27,27)
Un Pool : 2 × 2,(10,256,55,55)
Deconv : 11 × 11,512 Filters, (10,512,55,55)
Output : 10 × 227 × 227
Input : 10 × 227 × 227
Conv : 11 × 11, 128 Filters, Stride 4 (10,128,55,55)
Conv : 5 × 5,64 Filters, Stride 2 (10,64,26,26)
ConvLSTM2D : 3 × 3,64 filters (10,64,26,26)
ConvLSTM2D : 3 × 3,32 filters (10,32,26,26)
ConvLSTM2D : 3 × 3,64 filters (10,64,26,26)
Deconv : 5 × 5,128 Filters, Stride 2 (10,128,55,55)
Deconv : 5 × 5,128 Filters, Stride 2 (10,1,227,227)
Output : 10 × 227 × 227

FIGURE 11.6

A stacked spatio-temporal autoencoder (left) and a convolutional long short-term memory autoencoder (right) for abnormal event detection [17].

task as appearance features sometimes might degrade the performance of the classifier. So in this study, we have focused on HOF feature along with SVM and ANN as classifiers for our experimentation.

11.4.1 DATA SETS

Our study is conducted on two standard benchmark challenging data sets: Hockey Fight and Crowd Violence data sets. The Hockey fight data set comprises of a total of 1000 video clips categorized as fight and no fight from the National Hockey Leagues. Each category consists of 500 video clips and thereby having 500 violent and 500 nonviolent clips. Each clip exactly consists of 50 frames having resolution of 360 × 288 pixels for each frame.

The Crowd Violence data set is specialized to test violence detection based on a crowd behavior. These videos characterize the violent and nonviolent behavior of crowd in public places, making it suitable for surveillance task. Crowd Violence has a total of 246 real video clips, of which 123 are violent and 123 are nonviolent with each frame having a resolution of 320 × 240 pixels.

11.4.2 COMPARATIVE STUDY ON RELATED WORK

The results of various recent state-of-the-art methods along with our basic study applied over Crowd violence and Hockey Fight data sets are shown in Tables 11.1 and 11.2. Shallow modeling techniques have proven to be effective in Crowd Violence data set. Due to the sparsely represented MoIWLD approach by Zhang et al. [10], there is minimal reconstruction and classification error,

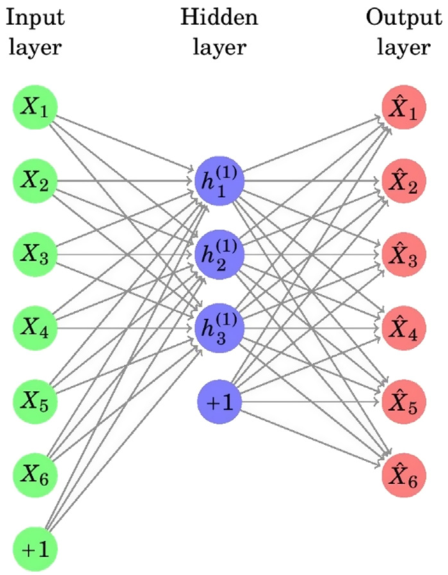

FIGURE 11.7

A simple autoencoder [17].

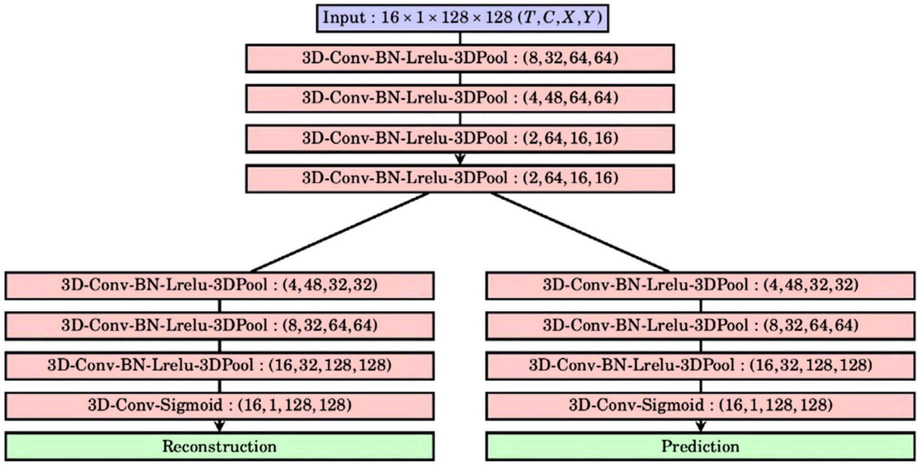
Input : $16 \times 1 \times 128 \times 128$ (T,C,X,Y)
3D-Conv-BN-Lrelu-3DPool : (8, 32, 64, 64)
3D-Conv-BN-Lrelu-3DPool : (4, 48, 64, 64)
3D-Conv-BN-Lrelu-3DPool : (2, 64, 16, 16)
3D-Conv-BN-Lrelu-3DPool : (2, 64, 16, 16)
3D-Conv-BN-Lrelu-3DPool : (4, 48, 32, 32)
3D-Conv-BN-Lrelu-3DPool : (8, 32, 64, 64)
3D-Conv-BN-Lrelu-3DPool : (16, 32, 128, 128)
3D-Conv-Sigmoid : (16, 1, 128, 128)
Reconstruction
3D-Conv-BN-Lrelu-3DPool : (4, 48, 32, 32)
3D-Conv-BN-Lrelu-3DPool : (8, 32, 64, 64)
3D-Conv-BN-Lrelu-3DPool : (16, 32, 128, 128)
3D-Conv-Sigmoid : (16, 1, 128, 128)
Prediction

FIGURE 11.8

3D autoencoder architecture for video anomaly detection [17].

which provides a significant result in Crowd Violence. By capturing local and spatio-temporal features through IFV, Bilinski and Bremond [2] produced a state-of-the-art result in the Hockey Fight data set.

Table 11.1 Results on Crowd Violence Data Set

Model	Method	ACC (%)
Shallow	IWLD [10]	88.16
	VIF + SVM [8]	82.13
	HOF + SVM	**83.37**
	HOF + ANN	**79.32**
Deep	GoogleNet + DLSTM [16]	93.59
	HOF + ANN (dp = 0.1)	**78.47**
	HOF + ANN (dp = 0.13)	**78.92**

ANN, *Artificial neural network;* DLSTM, *deep long short-term memory;* dp, *dropout level;* HOF, *histogram of oriented optical flow;* IWLD*; improved Weber local descriptor;* SVM, *support vector machine;* ViF, *Violent Flows.*

Table 11.2 Results on Hockey Fight Data Set

Model	Method	ACC (%)
Shallow	STIFV [2]	93.40
	STIP-HOG + HIK [20]	91.70
	HOF + SVM	**87.40**
	HOF + ANN	**87.13**
Deep	CLSTM [4]	97.10
	HOF + ANN (dp = 0.1)	**87.25**
	HOF + ANN (dp = 0.13)	**87.75**

ANN, *artificial neural network;* CLSTM, *convolutional long short-term memory;* dp, *dropout level;* HOF, *histogram of oriented optical flow;* HOG, *histogram of oriented gradients;* STIP, *space—time interest point;* SVM, *support vector machine.*

We studied a baseline approach by using the HOF feature descriptor on both SVM and ANN models, where the results are obtained as shown in Table 11.1. Deep features even though being less explored have proven to show promising results on these data sets. Through the combination of GoogleNet Inception V3 CNN and stacked LSTM methods, Zhuang et al. [16] outperforms certain state-of-the-art results for Crowd Violence data set. Sudhakaran and Lanz [4] proposed a CLSTM which was capable of extracting low-level localized features and thereby reducing false alarm rate to a great extent.

11.4.3 OUR BASELINE STUDY

Some of the handcrafted feature descriptors and the classifiers mentioned in Section 11.3 are used for conducting this study. The accuracies attained for these models, for each data set, are shown in Tables 11.1 and 11.2. The results of our experiments are shown in bold. For both the models, features extracted remain the same, that is, the features extracted from HOF are given to a classifier. For modeling, we have considered SVM and ANN with one hidden layer as shallow, while the ANN with two hidden layers as deep. As mentioned earlier, if handcrafted features are given to a shallow network classifier, it represents a shallow model, while when provided to a deep network, it can be said to represent a deep model. We have conducted our study solely based on the above statement. HOF features are used to obtain the optical flow information. These features extracted are provided to an SVM and also to an ANN, whose parameters are determined by hyperparameter tuning, for classification based on shallow model representation. The features extracted for shallow models above are also used for studying deep models by providing them to an ANN consisting of two hidden layers, which is treated as a deep network. This network was trained on two different dropout levels to study the impact of it on the model.

We conducted experiments by using the method of *k*-fold cross-validation, where *k* is five, that is, each data set is divided into five divisions each containing both the violent and nonviolent video clips. Features are extracted for each fold separately from each feature descriptor.

Training is performed by considering 80% as training set and the other 20% as test set. The average of all the accuracy in the fivefold validation is said to be the accuracy of the model.

Confusion Matrix for the second fold:

$$\text{Crowd Violence:} \quad \underset{\text{(Shallow)}}{\begin{bmatrix} 23 & 2 \\ 4 & 21 \end{bmatrix}} \quad \underset{\text{(Deep)}}{\begin{bmatrix} 22 & 3 \\ 7 & 18 \end{bmatrix}}$$

$$\text{Hockey Fight:} \quad \underset{\text{(Shallow)}}{\begin{bmatrix} 90 & 10 \\ 8 & 92 \end{bmatrix}} \quad \underset{\text{(Deep)}}{\begin{bmatrix} 97 & 3 \\ 6 & 94 \end{bmatrix}}$$

From the above obtained confusion matrix from one of the folds, it could be inferred that the Crowd Violence data set works better with shallow models as the false alarm rate is bound to be higher in deep models. This might be due to the fact that Crowd Violence being a small data set does not work well with deep networks. In contrast to the above, the Hockey Fight data set proves to work well with deep networks, since it has lesser false alarm rate. This is because of the large volume of data available in this data set compared with Crowd Violence. Our experimental study on deep networks was done by providing HOF features to a deep ANN model with two variations in dropout level. It can be seen that this baseline study with an HOF descriptor is more effective with Hockey Fight than that with Crowd Violence on both the methods. On further tuning, ANN may produce better results than other shallow methods.

11.5 CONCLUSION

Violence detection is one of the most important and essential tasks of video surveillance. In this chapter, we have focused on shallow and deep modeling techniques on two standard benchmark data sets such as Crowd Violence and Hockey Fight. We have done a comparative study of shallow and deep models on the above data sets and also on other state-of-the-art approaches for different feature descriptors and analyzed their results. In this chapter, we have inferred that for a small data set, shallow models perform well, but for a large data set, deep models give comparatively better performance at the cost of training time complexity.

REFERENCES

[1] K. Lloyd, P. Rosin, D. Marshall, S. Moore, Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures, Mach. Vis. Appl. (2017) 28, pp.361-371.

[2] P. Bilinski, F. Bremond, Human violence recognition and detection in surveillance videos, in: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Colorado Springs, CO, 2016, pp. 30−36.

[3] Y.S. Chong, Y.H. Tay, Abnormal event detection in videos using spatiotemporal autoencoder, in: F. Cong, A. Leung, Q. Wei (Eds.), Advances in Neural Networks - ISNN 2017. ISNN 2017. Lecture Notes in Computer Science, vol. 10262, Springer, Cham, 2017, pp.189−196.

[4] S. Sudhakaran, O. Lanz, August. Learning to detect violent videos using convolutional long short-term memory. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* 2017 (pp. 1−6). IEEE.

[5] I. Laptev, T. Lindeberg, Space-time interest points, Int. J. Comput. Vis.- IJCV 64 (2003) 432−439. Available from: https://doi.org/10.1109/ICCV.2003.1238378.

[6] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2005) 107−123. nos. 2−3.

[7] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, CVPR (2008).

[8] T. Hassner, Y. Itcher, O. Kliper-Gross, Violent flows: real-time detection of violent crowd behavior, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, 2012, pp. 1−6.

[9] T. Wang, H. Snoussi, Detection of abnormal visual events via global optical flow orientation histogram, IEEE Trans. Inf. Forensics Security 9 (2014) 988−998.

[10] T. Zhang, W. Jia, X. He, J. Yang, Discriminative dictionary learning with motion weber local descriptor for violence detection, IEEE Trans. Circuits Syst. Video Technol. 27 (3) (2017) 696−709.

[11] R.T. Ionescu, S. Smeureanu, B. Alexe, M. Popescu, Unmasking the abnormal events in video. In *Proceedings of the IEEE International Conference on Computer Vision* 2017. (pp. 2895−2903).

[12] H.T. Tran, D. Hogg, September. Anomaly detection using a convolutional winner-take-all autoencoder. In *Proceedings of the British Machine Vision Conference 2017*. British Machine Vision Association, 2017.

[13] S. Smeureanu, R.T. Ionescu, M. Popescu, B. Alexe, Deep appearance features for abnormal behavior detection in video, in: S. Battiato, G. Gallo, R. Schettini, F. Stanco (Eds.), Image Analysis and Processing—ICIAP 2017. Lecture Notes in Computer Science, vol. 10485, Springer, Cham, 2017.

[14] M. Mishra, M. Srivastava, A view of artificial neural network, in: 2014 International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), Unnao, 2014, pp. 1−3.

[15] H. Shin, et al., Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imag. 35, 2016, pp. 1285−1298.

[16] N. Zhuang, J. Ye, K.A. Hua, Convolutional DLSTM for crowd scene understanding, in: 2017 IEEE International Symposium on Multimedia (ISM), Taichung, 2017, pp. 61−68.

[17] B.R. Kiran, D.M. Thomas, R. Parakkal, An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos, J. Imaging 4 (2018) 36.

[18] T. Xiao, C. Zhang, H. Zha, F. Wei, Anomaly detection via local coordinate factorization and spatio-temporal pyramid, in: D. Cremers, I. Reid, H. Saito, M.H. Yang (Eds.), Computer Vision—ACCV 2014. ACCV 2014. Lecture Notes in Computer Science, vol. 9007, Springer, Cham, 2015, pp. 66−82.

[19] K. Cheng, Y. Chen, W. Fang, Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 2909–2917.

[20] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, R. Sukthankar, Violence detection in video using computer vision techniques, in: P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, W. Kropatsch (Eds.), Computer Analysis of Images and Patterns. CAIP 2011. Lecture Notes in Computer Science, vol. 6855, Springer, Berlin, Heidelberg, 2011, pp. 332–339.

[21] R. Leyva, V. Sanchez, C. Li, Video anomaly detection with compact feature sets for online performance, IEEE Trans. Image Process. 26 (2017) 3463–3478.

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

[23] S. Huang, D. Huang, X. Zhou, Learning multimodal deep representations for crowd anomaly event detection, Math. Probl. Eng. 2018 (2018) 13.

[24] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, arXiv:1801.04264 [cs.CV], 2018.

[25] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, N. Sebe, Abnormal event detection in videos using generative adversarial nets, in: 2017 IEEE International Conference on Image Processing (ICIP), Beijing, 2017, pp. 1577–1581.

[26] H. Vu, T.D. Nguyen, A. Travers, S. Venkatesh, D. Phung, Energy-based localized anomaly detection in video surveillance, in: J. Kim, K. Shim, L. Cao, J.G. Lee, X. Lin, Y.S. Moon (Eds.), Advances in Knowledge Discovery and Data Mining. PAKDD 2017. Lecture Notes in Computer Science, vol. 10234, Springer, Cham, 2017, pp. 641–653.

[27] P. Zhou, Q. Ding, H. Luo, X. Hou, Violence detection in surveillance video using low-level features, PLoS one 13 (10) (2018) e0203668.

[28] M. Ravanbakhsh, E. Sangineto, M. Nabi, N. Sebe, Training adversarial discriminators for cross-channel abnormal event detection in crowds, CoRR, vol. abs/1706.07680, 2017.

[29] 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.

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