# Crime Prediction from Digital Videos Using Deep Learning

Jialiang Liu

A thesis submitted to the Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2021

School of Engineering, Computer & Mathematical Sciences

# Abstract

In the surge of intelligent surveillance, surveillance alarming has been demanded never before, which makes people aware of their ordinary security. Traditional surveillance models are relatively simple, which cannot be applied to detect real-time crimes automatically and undoubtedly waste social resources. Intelligent monitoring approaches make use of pattern recognition and machine learning to analyze and tackle video footages. If abnormal behaviors are detected, an alarm will be triggered timely.

In this thesis, we train our models based on Spatial and Temporal Graph Convolutional Networks (ST-GCN) as well as Temporal Relational Networks (TRN) for detection and recognition of human behaviors from digital videos. The skeleton sequences of human behaviors are extracted from surveillance videos. The risk level is determined by setting the corresponding thresholds. The TRN networks are compared with the ST-GCN which are based on optical flow to combine with the temporal relationship of video frames. The main purpose of this TRN method is to extract a spate of frames from the given videos in a random way. The results show that human behavior recognition method based on skeleton and optical flow outperforms than other algorithms in deep learning.

For the identification of human dangerous behaviors, in this thesis, we train model based on spatial temporal graph convolutional neural networks and time series relational networks, respectively, for the detection and identification of human criminal behaviors. The key to the recognition method based on a spatial temporal graph convolutional neural network is the extraction of the human skeleton.

Taken into consideration of the skeleton sequence of human behavior, the skeleton of each frame contains 18 joint points of human skeleton and the estimated confidence value of the skeleton of each frame. According to the obtained skeleton feature information, combined with the time vector in the skeleton sequence, a spatial temporal map is established. The network model classifies the criminal behavior and determines the criminal behavior of the behavior by setting the corresponding threshold. Compared with the spatial temporal graph convolutional network, the time-series relational

network uses different feature information. The relational network establishes a time-series relational network model based on the human optical flow information and the relational reasoning of video frames. The key to the identification method based on a temporal relation network is to extract several frames of input temporal relation network from the videos in an order or randomly.

The experimental results based on the collected dataset show that the recognition result is better than the single feature algorithm, the recognition accuracy is higher, and the robustness is better. The network equipped with the time series relationship module effectively improves the recognition accuracy in the detection and recognition of criminal behavior. In this thesis, we take use of a variety of methods to conduct comparative experiments, the results show that the recognition method based on skeleton and optical flow features is significantly better than the manual feature extraction algorithm.

**Keywords**: Deep learning, risky behavior detection, crime prediction, ST-GCN, TRN

# Table of Contents

# List of Figure

# List of Table

# Attestation of Authorship

This Master's thesis is the result of research undertaken and achieved under the supervision of my supervisor. To the best of my knowledge, no other published or authored research is included in the thesis, except where specifically indicated and acknowledged in the article. Individuals and groups who have made significant contributions to this research are clearly identified in the text.

Signature: Date: 18 Nov 2021

# Acknowledgement

I would like to thank my parents first. I received financial support from my parents so that I can study at the Auckland University of Technology (AUT). Owning to this support, I completed my thesis project successfully at AUT. At the same time, I want to thank both my friends and teachers for their support and care during my study life.

My supervisor Wei Qi Yan is an amiable friend, and he has deep academic knowledge in deep learning. I am grateful for his support throughout this research project. He helped me grasp deep learning the knowledge of visual object detection, and how to finish my master's degree on time.

Jialiang Liu

Auckland, New Zealand

November 2021

# Chapter 1

# Introduction

The identification of human behavior is a principal issue in intelligent surveillance which is essential for crime prevention. Automatic detection and identification of high-risk behaviors in video surveillance enable potential hazards to be identified and be handled immediately. Creating a comprehensive and immediate alarming method that can effectively reduces or eliminates public safety problems and ensure the safety of people and their properties.

## 1.1 Background and Significance

In recent years, with the frequent occurrence of violent attacks, public safety has also become a growing concern, with global crime incidents showing a year-on-year growth trend. According to incomplete statistics, in the past five years, the average number of unnatural deaths and disabilities caused by public safety incidents exceeded 300,000 and 2.5 million respectively, with economic losses amounting to 700 billion yuan. Statistics on actual incidents show that crime is on the rise year on year.

Faced with the rising trend of crimes and shift of crimes from monolithic to diversified, advanced video surveillance, human behavior identification and crime prediction are demanded (Altamimi et al. 2018). It is clear that conventional video surveillance could not meet the needs of practical applications, meanwhile intelligent surveillance systems incorporate deep learning, pattern recognition, image recognition and computer vision. Our survey through literature review shows that deep learning-based human behavior recognition has better research outputs and significantly reduces social costs (Bengio, Simard, & Frasconi, 1994).

The frequency of crime incidents has risen continuously in recent years. The use of surveillance equipment has been taken globally, with the main purpose of detecting and analyzing people in public places. Traditional video surveillance is grouped into two categories: One is operated with real-time identification through human naked eyes. This method leaves the police susceptible to visual fatigues and has low detection accuracy and efficiency. Another method is to examine and search for videos as evidence after an abnormal event, but there is an information lag that does not allow a quick real-time response. It is clear that traditional video surveillance systems no longer meet the needs of today's practical needs, the intelligent surveillance systems are a new generation of video surveillance technology that incorporates with deep learning, pattern recognition, and computer vision.

The vision-based abnormal behavior analysis aids to automatically detect abnormal events in crime scenes without the needs of human judgment based on the existence of abnormalities, provide timely responses and alarming to those unexpected

events, truly achieve all-weather, fully automatic, real-time monitoring, and improve efficiency while save social resources.

Intelligent monitoring better reflects the development of intelligent video surveillance, compared with traditional surveillance systems, intelligent systems have a powerful capability for image and video analytics, in the event of dangerous situation, the more accurate and timely prediction of alarms is needed to provide better protection for people and save their lives and security resources.

Deep learning has yielded a plethora of research outcomes in human behavior recognition in recent years. Both 3D convolutional neural network (CNN) for human behavior recognition and dual-stream CNN for crime action analysis perform well based on the UCF101 dataset. The study of this thesis through the literature shows that deep-learning-based behavior recognition has better outcomes and significantly reduces personnel costs.

## 1.2 Research Content and Difficulties

### 1.2.1 Content

In a vast variety of scenarios, the detection and identification of hazard behaviors need to be carried out rapidly. The quality of videos and the occurrence of dangerous behaviors make it impossible for intelligent surveillance to deal with and solve security problems timely. In a surge of crimes, we need a fast and accurate method for detecting hazardous behaviors.

A Spatial Temporal Graph Convolutional Network (ST-GCN) and TRN methods are employed for the detection and recognition of dangerous human behavior. Extracting human behavioral skeleton sequences from surveillance video assists us to construct a spatial temporal relationship network to classify criminal behavior by setting the level of danger of the corresponding line of criminal behavior. In the spatial temporal relationship network using visual feature information, a spatial temporal network model is established based on optical flow, combined with the spatial temporal

relationship existing in video frames.

In this thesis, we propose two methods of abnormal behavior recognition, obtain video data through video surveillance, establish behavior recognition models, and classify the behaviors and hazardous level of the given behavior. We implement the real-time prediction of crime occurrence through the recognition of human behaviors under the realistic scenarios to verify the accuracy and efficiency of our methods in this thesis for human behavioral detection in realistic scenarios.

### 1.2.2 Research Difficulties

Deep learning has the availability for crime prediction. However, it requires a large amount of data to be labelled as a training dataset. The unavailability of crime data, due to the fact that crime data is not publicly available, limits research outcomes in crime prediction and becomes a challenge.

The definition of crimes is overly complicated which is difficult to be interpreted. Human behavior prediction has a vast range of real-world applications and has been applied to predict abnormal behaviors before it occurs. However, the complexity of human behaviors as well as the effects of environment, luminance, and occlusion, have led to complex research outcomes. Behavioral prediction currently faces several challenges.

In real scenes, luminance changes of moving targets due to light source variations easily cause serious deviations between the extracted visual features and the real existing features, result in a low accuracy rate of human behavior recognition.

Occlusion is happened usually in crowds, which results in the loss of visual information and leads to a significant gap between the extracted features as well as the implemented algorithm that has a particularly critical impact on the accuracy of crowd anomaly prediction.

If crime scene has a cluttered background, visual feature extraction is easily influenced by background, the extraction of exact visual features becomes much tough. These problems are also hindering the development of human behavior prediction,

result in a low accuracy rate of behavior detection. In summary, the existing research work is far away from satisfying the practical needs, more extensive and in-depth research outcome is needed in behavioral representation and model design.

### 1.2.3 Main Contributions

In this thesis, a method to extend spatial graph convolutional neural network to spatial temporal CNN for abnormal behavior recognition is proposed, using SoftMax classifier for abnormal behavior recognition, which has high performance in real time and strong robustness. In order to explore human behavior recognition, RGB-color videos and 3D skeleton information are fused together with attentional models.

In this thesis, a hazard recognition method based on TRN is proposed for reasoning based on the temporal relationship in the video frames by using the TRN network to infer the temporal relationship of the frames before and after the behavior, in recognizing human actions, improving the effectiveness, and speeding up the hazard behavior recognition.

## 1.3 Structure of This Thesis

In Chapter 1, we provide an introduction. Firstly, the background of this thesis, the contents including challenges of the project are presented. Finally, two methods for crime behavior recognition are presented.

In Chapter 2, we present a review of literature. Firstly, a summary outlines the state-of-the-art research work in human behavior detection and abnormal behavior alarming. Then, the fundamentals and databases for human behavior recognition are introduced. Finally, the difficulties of current recognition behavior recognition algorithms are stated.

In Chapter 3, we deal with the methodology. Firstly, the collected dataset is processed. Secondly, hazard behavior recognition based on ST-CNN is identified, a method for human dangerous behavior recognition based on ST-GCN method is proposed. ST-GCN is constructed by using a spatial temporal graph convolutional

network for human behavior recognition that extracts and fuses human information like skeleton and time vectors. Then, human behavior recognition based on a temporal relational network is presented, a human body recognition method based on a temporal relational network is proposed. The temporal relational network model is established through the reasoning of temporal relationships, the optical flow and dual-stream network are combined for hazard identification.

In Chapter 4, our experimental results are presented for the use of ST-GCN and TRN, the human behavior identification methods are evaluated by using two models.

In Chapter 5, we detail the discussion. Our focus is on analyzing the experimental results and comparing the effectiveness and accuracy of the proposed models.

In Chapter 6, we present the conclusions of this thesis and looks forward to our future work. Thus, a review of the main studies is presented. The shortcomings of the research in this thesis are pointed out along with our future work that needs to be improved.

# Chapter 2

# Literature Review

In this chapter, we describe the state-of-the-art research work on human abnormal behavior recognition. The basis of our human behavior recognition and the currently available datasets are introduced, finally, the methods for evaluating human behavior recognition are presented.

## 2.1 Current Status of Behavior Recognition

Behavior prediction differs from behavior recognition because human behavior prediction is the observation of human actions that exist in the given videos, in which the video frames are provided as streaming data, whereas human action recognition takes the entire sequence of observations as the input. The main prediction methods include logistic regression (Martinezetal, 2018), support vector machine (Luetal, 2014; Candelieri, 2017), artificial neural networks (Gunay, 2016; Chaawlaetal, 2019), gradient boosted decision trees (GB-DT) (Martinezetal, 2018), etc.

Law enforces have been using crime mapping and forecasting for the consideration of social security. The traditional methods are greatly updated by using intelligent surveillance. Usually, crime analysts would be use of massive amounts of data to determine if a type of crimes were committed. The information could be applied to prevent similar happenings in the future.

As a strategy for reducing crimes, predictive policies have been developed. Using data-driven approaches improves crime prediction and management. Reddy *et al.* explored the evolution of crime forecasting through the use of machine learning algorithms (Shah, Reddy, 2013). The concept of machine learning allows us to visualize a vast amount of data and provide predictions with the relationships between various patterns of criminal cases.

A probabilistic model was proposed with a bag-of-word method for video representation by using a sparse representation to reconstruct the training set. The collected surveillance videos are applied to combine the training sets.

The difficulty of predicting human behavior was to investigate an action in long videos based on the detection of actions, an offence is computationally intensive to deal with human behaviors in complex outdoor environments. A multiscale discriminative prediction model was proffered based on support vector machine. Human behavior prediction requires much extensive and in-depth research work in behavioral representation, model design, etc.

The first problem to be resolved in the prediction problem is to make decision in

a scenario where the accuracy of the prediction is guaranteed before the action is taken effect, the accuracy needs to be as high as possible as well as real time.

Currently, there are a slew of research directions in human behavior recognition. The basic process of human behavior recognition is to firstly tackle the input video, then extract features using image processing algorithms, and finally conduct classification. The aim is to reduce the content of the video data that is not relevant to the behavior, to alleviate the redundancy and obtain valid information. Human behavior detection is the process of extraction features of visual objects from frames in the given video, which involves human detection, salient object detection, and others.

After the detection, the features of the frame sequence need to be extracted. The feature extraction includes static features and dynamic features. The static features are extracted by 2D images, feature extraction methods include Histogram of Oriented Gradient (HOG) (Samuel, Manogaran 2019), Scale Invariant Feature Transform (SIFT) (Laptev, 2005), Local Binary Patterns (LBP), etc. According to the existing work, the most effective features are those that blend temporal and spatial features together. The methods of feature extraction for dynamic information include optical flow. The optical flow method takes use of the changing illumination of the pixels in motion picture sequences to obtain the motion of the visual object target based on the speed at which the pixels are moving from one time to another.

The main problem of human behavior recognition is how to represent human behaviors in each video. The behaviors of people in digital videos are influenced by many factors, thus behavioral representation is a challenging problem. Extracting visual features to be effective from the video is a major step in human behavior analysis. The same kind of features describe different classes of behaviors to a variety of degrees, the efficient features need to be extracted as a key step in human behavior recognition.

Up to date, there are two types of human behavior recognition methods. Traditional human behavior recognition is based on preprocessing of behavior recognition datasets to feature extraction and feature modelling. Deep learning, on the other hand, has more steps than traditional methods, saves much time and storage space but needs to compute visual features as shown in Figure 2.1.

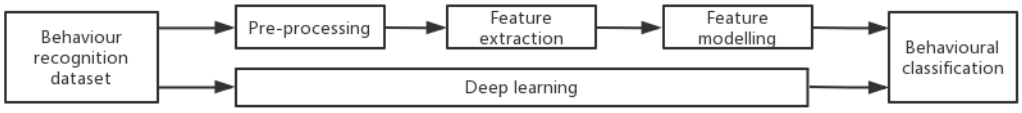


Figure 2.1: The pipeline of human behavior analysis

Since artificial intelligence has only been around for a few decades, we see that police departments still figure out how to predict crime using surveillance systems. As recent criminology, information technology such as Geographic Information Systems (GIS), crime prevention and management that utilize technological alternatives, crime forecasting has become a viable alternative to curbing crime.

Crime forecasting has made progress surprisingly, especially using visual surveillance with biometrics. Because of the needs to eliminate all human errors, surveillance has developed rapidly in the industry. Crime forecasting focuses on the application of biometrics for surveillance to detect suspects using human bodies like face, iris, voice, hands, and fingerprints. For identifying body parts, we analyze individual's unique characteristics based on Euclidean distances, hamming distances, hidden Markova models, and string matching. Human behavior has been proven to be a powerful way for automatic crime forecasting.

Usually, crime analysts take use of a large amount of crime prediction and identification data to determine whether a crime have been committed. This information has been applied to prevent similar crimes in the future. The methods of human behavior and action analysis include the use of evolutionary algorithms and gradient descent algorithms to initialize and train Convergent Neural Networks (CNN). However, the conventional behavior detection is based on the extraction of global performance of human body structure, shape, and motion, or based on local feature extraction and local feature description.

The approaches taken in human behavior analysis have involved initializing and training CNN by utilizing evolutionary algorithms and gradient descent algorithms.

Due to the challenges and complexities involved in predicting crime with our naked eyes, the work related to how machine learning is to enhance crime prediction by the police has become even more important. In addition, machine learning is crucial when comparing historical, spatial, and temporal crime-related data. If this is accomplished, the analysis will be much accurate. Therefore, the focus of this project is on surveillance, identification, and crime forecasting so as to develop a definitive approach for crime forecasting in surveillance.

## 2.2 Traditional Behavior Recognition Methods

Traditional behavior recognition methods are based on the extraction of global representations of human skeleton, shape, and movement, local feature extraction and feature description, a global representation-based approach for human behavior recognition.

The global feature-based human behavior recognition takes the entire samples into account, fully takes advantage of a top-down approach to describe visual features with strong capabilities of human behavior recognition. The difficulty in human behavior recognition is the extraction of global feature from digital videos.

The recognition rate of global feature-based behavior recognition methods is related to the completeness of human regions that is also interfered by external factors such as camera movement, low resolution of the videos, and the obscuration of moving objects. Therefore, in practice, the recognition accuracy of global feature-based recognition methods is generally low and unsuitable for practical applications.

The local feature-based human behavior method extracts visual features that can describe our objects directly from the video scenes. The steps such as detection and segmentation of the object region as well as object tracking are generally not required. Local features are robust to complex video backgrounds, partial occlusion of visual objects and changes in viewpoint, but they lack semantics.

There are five steps in the local feature-based behavior recognition method: Video input, local feature extraction, feature coding, feature regularization, visual object

classification and recognition.

Traditional human behavior recognition has a plenty of drawbacks compared to deep learning-based human behavior recognition methods, it is specific and will take a long time. Deep learning, on the other hand, has the ability to automatically carry out feature extraction from video frames.

## 2.3 Deep Learning Behavior Detection Methods

The concept of deep learning was firstly introduced in 2006 by using unsupervised training methods to initialize the weights of the network and then fine-tune the parameters for model training purposes.

In the context of deep learning, joint point behavior recognition is the extraction of features from edited video clips. The deep learning methods to deal with articulation data are CNN, Recurrent Neural Network (RNN) and GCN, which correspond to the representation of articulation data as pseudo images. The networks include RNN-based Joint Point Behavior Recognition (JPBR), GCN-based JPBR, and Hybrid Network (HN)-based JPBR.

At present, basic network models have achieved excellent results in computer vision such as visual object classification, object segmentation, pose estimation, and pedestrian detection. In the area of human behavior recognition, the existing methods for deep learning-based behavior recognition were basically from digital image processing and natural language processing.

Deep learning and neural networks are introduced to crime forecasting. There are four types of neural networks in crime forecasting. An example is a feedforward network, this model consists of three units: Input, hidden, and output. Based on the input and the output, this model is essentially straightforward.

The second model is based on CNN. Surveillance images were the input data of this model. The importance of real-time crime surveillance as the shortfalls is shown as predicting crime. This is mitigated by using a spatial temporal residual network to predict crime occurrence. The ST-ResNet structure and cuDNN are introduced to the

research work so as to train models and deal with time series for each grid. Because convolution operations are excluded from the models, transitions between grids are not considered. The DNN-based prediction model incorporates feature-level data fusion. In fact, DNNs have the ability to combine various features into a single one by using the knowledge how to integrate them.

In addition to illuminations, static foreground objects, dynamic backgrounds, and camouflage, traditional surveillance systems have many constraints. Despite these problems are still the challenges of modern surveillance systems, CNN is the solution which gives the remarkable advances in computer vision. This work presents a foreground segmentation algorithm based on VGG-16. A blur in the foreground may be caused by a feature extracted by using the algorithm. In order to ensure that all losses computed from CNN, the output vector components are not affected by other relative losses, a sigmoid function is employed for an error function in the form of binary cross-entropy.

Our network can capture both temporal and spatial information if CNNs and RNNs are combined. Thus, our spatial temporal crime prediction relies on DNN networks. In crime reduction of visual surveillance, the proposed STCN takes effect. In STCN, CNN is incorporated, which has an automatic crime reference function. Based on large amounts of historical data, STCN predicted crime for the following day by using historical data. An image of a knife or a gun is employed for crime forecasting by using CNN. Several functions of CNN were modeled to ensure the accuracy is improved. These include dropout, Rectified Linear Units (ReLUs), fully connected layers, and convolutional layers. CNN is the method that identifies the rearing images, which can predict potential crimes. It would allow law enforcement agents to prevent crimes to be happened.

By using binary cross-entropy as an error function, a sigmoid function ensures that all losses computed from CNN output vector components are unaffected by other relative components. A network that captures spatial temporal information is created by combining CNNs and RNNs. Its significance lies in the inclusion of CNN, which automatically infers crimes. By analyzing large amounts of historical data, the STCN

can predict the next crime. A CNN-based crime prediction has been proposed which detects knives and guns in digital images. Various components of the model are applied to ensure the accuracy of CNN-based behavior recognition. Law enforcement is employed to predict potential crimes in advance by detecting the extracted images, which is an integral part of visual surveillance.

Topology-based Convolutional Networks (TCNs) were proposed. Afterward, ST-ResNet and GCN are combined to develop a ST-GCN and apply it to the field of human behavior recognition, which compensates for the lack of traditional spatial temporal models in topological space by capturing transmission effects of distinct types of crime events.

Human action recognition quickly takes effects after its introduction. Kong, for example, proposed a dynamic skeleton model based on ST-GCN and a module for analyzing attention. Gene proposed the ST-MGCN based on the generalized network. The topology-based crime is predictable, such as Weber's use of GCN in combating money laundering. In the crime prediction model, a T-GCN model is applied to combine GCN and RNN. DT-MGCN models that combine the GCN, and a multigraph convolutional network (MCN) are treated as the state-of-the-art method for human behavior recognition.

### 2.3.1 Convolutional Neural Networks

The convolutional network consists of four components, each with a different function, namely, convolutional layer, pooling layer, classification layer, and fully connected layer. The convolutional layer is mainly harnessed for feature extraction, convolution operation is a weighted summation of the input data. The pooling layer samples the input data, compresses the amount of data and parameters and reducing overfitting. The fully connected layer acts as a classifier, turning the input data into feature vectors that are classified by the classifier.

The convolutional layer contains convolution kernels. The convolution operation of CNN is to take a box of convolutional kernels on the original image and slide a

window over it, each time the elements of the convolutional kernels are multiplied by corresponding elements to sum up the value of the last element of the output feature, the higher the number of convolutional layers in CNN. In the convolutional layer, the first step is to slide a smaller filter over the selected image. A product calculation is generated by using the weights created by the convolution process. Consequently, there is a relationship between the weights and the filtered region. In the output layer, the nodes are connected from the input nodes through the hidden nodes to the output nodes. As a result, the output layer is like the layer of an artificial neural network (ANN). The size of neurons differs from two aspects, despite the similarities between them. Convolution does not use the entire neuron set, but rather than only a subset of its neurons.

ReLU layer is the activation layer, which is generally employed as the activation function in CNN, whose role is to make a nonlinear mapping of convolutional layer as the output. A few activation functions are listed such as sigmoid, tanh, ReLU, and max-out. Sigmoid function outputs a range of values between 0 and 1, the input is any function whose mathematical expression takes the form of Equation (2.1).

$$\sigma(x) = \frac{1}{(1+e^{\{-x\}})} \tag{2.1}$$

where tanh function differs from sigmoid function, its output is between -1 and 1, it is faster than sigmoid function, but from the function image, the gradients at both ends are also almost 0, it does not avoid the problem of gradient vanishing.

ReLU function is a piecewise function, from the image of this function, we see that if the input is greater than 0, its gradient is constant at 1.0, which avoids the problem of slow parameter updates; if the input is less than 0, the gradient is 0. The aim is to make the weights sparse and reduce computational costs.

In CNN, pooling layer is generally located between two convolutional layers. Its primary purpose is to reduce the dimensionality of the data by mimicking human visual system, with the pooling layer running independently on each feature map. Data and parameters are downsampled in the pooling layers to reduce overfitting. As a result of the pooling layer downsampling the input image, the input is resized. The main

strategies for downsampling in the pooling layer are max pooling, which is the most useful pooling operation, it selects the maximum value in the image region as the desired feature pixel, meanwhile, average pooling selects the average value in the image region as the desired feature pixel.

All local features are combined into a global feature, which is used to calculate the probability of human behavior recognition for each type of actions or to feed deep features into a deep network for model training. By means of a fully connected structure, the previous output features are reassembled into a complete image.

Utilizing convolutional neural networks is the goal to achieve human behavior recognition. Convolutional layers produce feature vectors, which are put into the pooling layer to reduce dimensionality. The output information is then processed by using convolutional layer, then the classification layer. After the sampling, the dimensionality of the image is reduced.

The advantage of CNN is that it takes use of local perception and parameter sharing which has a high processing power for large datasets. There are no difficulties on the processing of high-dimensional data. It is possible to extract deeper information from the image without manual feature selection, just train the convolution kernel and bias term to obtain the feature values.

The disadvantages of CNNs are that it takes longer time to train deep learning models and requires much large training data samples, it is generally recommended to use GPUs for model training. The results in each layer cannot be interpreted, which is also a disadvantage of neural networks.

### 2.3.2 Recurrent Neural Networks (RNNs)

RNNs are used to analyze datasets, convert them into computer-recognizable vectors, and commence feature extraction. An algorithm for feature analysis was constructed by training RNNs to produce the desired model. Compared to traditional methods, it is much accurate and subject to less influenced. RNNs are mainly employed for the data from processing and prediction sequences, which influence the output of later nodes by

using information from memory of the previous ones. RNNs are essentially replicated neural networks whose structures were replicated over time.

The main structure of RNN is replicated in time series analysis, the structure is also called a cyclic body. How to design the network structure of the cyclic body is the key to solve practical problems in RNN. In RNN, the parameters in the loop are also shared at multiple moments. RNN undergoes multiple mappings from the input state to the hidden state and finally to the output with the aim of building a dynamic time model.

Regarding the simplest RNN using a single fully connected layer as the loop, the small yellow tanh box in the figure indicates a fully connected layer using tanh as the activation function.

The inputs to cycle at time $t$ consist of $X_t$ and the hidden state $h_{t\text{-}1}$ transmitted from time $t$-1. $X_t$ and $h_{t\text{-}1}$ are directly spliced into a larger matrix-vector [$X_t$, $h_{t\text{-}1}$] [$Xt$, $h_{t\text{-}1}$]. We assume that the shapes of $X_t$ and $h_t$-1 are [1, 3] and [1, 4], respectively, the final input vector for all connected layers in the loop has the shape [1, 7]. Once the stitching is completed, we simply follow the fully connected layers. In order to transform the implicit state $h_t$ at the current moment into the final output $y_t$, the RNN needs another fully connected layer to complete the process. This leads to the same fully connected layer in CNN.

There are four types of RNNs: (1) One to one: Both input and output are not sequence. (2) One to many: The input is not a sequence; the output is a sequence. (3) Many to one: The input is a sequence; the output is not a sequence. (4) Many to many: The input and output are both sequences, but they have different lengths.

RNN is a dynamic network with drawbacks, such as the difficulty of training the network due to long-term dependencies, which has the problem of vanishing gradient due to the propagation of gradients through the unfolding CNN.

### 2.3.3 Graph Convolutional Network

The convolution in CNN is a kind of discrete operations, which essentially uses a kernel with shared parameters to compute a weighted sum of central pixel points and

neighboring pixel points to form a feature map for spatial feature extraction, the weighting factor is of course the weight of the convolution kernel.

Most of data does not have a regular spatial structure, called non-Euclidean data, e.g., graphs abstracted from recommendation systems, electronic transactions, etc. Each node in the graph is connected differently, the nodes have three connections and a plenty of nodes have only one connection. For these irregular data, ordinary convolutional networks do not work well. If non-Euclidean spaces are considered, it is difficult to select a fixed convolution kernel to accommodate the irregularities of the whole graph, such as the uncertainty in the number of neighboring nodes and the uncertainty in the order of nodes.

GNN is a network structure that combines graph theory and deep learning. The main ones currently include GCN, Graph Attention Networks (GAN), Graph Auto-encoder (GAE), Graph Generative Networks (GGN), Graph Spatial Temporal Networks (GST). Pertaining to original GNN networks, the features of points and edges are passed into the network together.

GCN is a combination of graph and convolution. The difference between the two is that spectral method is based on Laplacian matrix, which is closely related to the graph and has weak generalizability, while the spatial method defines the convolution directly on the graph and operates on nodes that are closely related, divided into point classification and graph classification.

GCN is a feature extractor that is similar to CNN, but it is used to better extract features from graphs so that they are used for node classification, graph classification, and prediction purposes.

In general, GCN falls into two categories: A spectral perspective, e.g., a spectral approach analyzes the local nature of graph convolution and examines convolutional filters directly in graph nodes. The proposed ST-GCN models adhere to the second principle and construct a GCN based on the neighborhood distance of each node and then construct the CNN kernel in the spatial domain.

### 2.3.4 Criminal Behavior Identification Using YOLO

YOLO has been employed for crime recognition, which is necessary to use labeling to define the criminal behaviors in the video to distinguish between normal and criminal behavior, and later to extract and classify human behaviors through the YOLO network.

YOLO as one of the more popular networks has been employed for human behavior analysis. The advantage is that YOLO is fast, while the accuracy is high, and the false alarm rate is low. The main idea of YOLO network is to treat object detection as a regression problem, using a neural network to calculate the output from the input to the class probability of the entire graph, enabling direct end-to-end prediction (Bengio & Glorot, 2010). Moreover, any crime is caused by the interaction of three elements: Crime, criminal target, and criminal environment (Billings & Yang, 2006).

The network structure of YOLOv5 is divided into four parts: Input, backbone, neck, prediction. The backbone includes focus structure, CSP structure; the heck has FPN+PAN structure; the prediction encapsulates GIOU Loss, Mosaic data is enhanced with random scaling, random cropping, and random arrangement. The mosaic data is enhanced with random scaling, random cropping, and random rows to improve detection of small objects.

In this thesis, we apply the network structure YOLO to the classification problem and solve the detection of abnormal behaviors as a regression problem. We install the necessary environment dependencies, create a data folder for the customized datasets, and structure the directory as follows: Under the annotations folder, the xml files are associated with each image, the mages have JPEG type in the VOC dataset format, the software labelImage is used as the annotation tool.

There is a subfolder under the Imagesets folder. In the directory, the training set, validation set, and test set are stored, generated by creating the script file. After running the code, the .txt documents are generated under the main folder.

The next step is to prepare for the labels by converting the dataset format to yolo_txt format, i.e., extracting the bounding box information into XML txt format. Each image corresponds to a .txt file, each row of the file contains the information of one object, including the coordinates $x_{min}$, $x_{max}$, $y_{min}$, and $y_{max}$.

We set the path, training classes, and the number of training classes for the training

set. We modify the number of training classes, select the training model, the number and size of images in the weights file and the training file, respectively. After the training, two pre-trained models will be generated.

By training YOLOv5x and YOLOv5s in YOLOv5, we achieve human behavior recognition with five classes: Abuse, assault, burglary, shooting, and shoplifting. Because our initial training dataset has 300 images, YOLOv5s model has been trained for four hours, YOLOv5x model for 10 hours, respectively. Among them, 225 training samples were generated for training and 15 validation samples were generated for verification.

### 2.3.5 Analysis of Crime Prediction Models

Any crime is caused by the interaction of three elements: The crime, the crime target and the crime environment, that crime will only occur within the triangle formed by using the three lines representing the three elements. The predictive model was built by using the crime environment as *z*-axis, the crime target as *x*-axis and the crime as the *y*-axis to create a spatially right-angled coordinate system. In the coordinate system, *x*-axis represents the characteristics of the crime target, and together with the crime environment represented by using *z*-axis, *x*-axis and *z*-axis, forms the plane *zox* showing the daily behavior pattern of crimes. The same *y*-axis represents the characteristics of the crime and *zoy* plane constructs the daily behavioral model of the crime.

In *zoy*, the target shares a spatial temporal intersection *z*-axis, which consists of a suitable crime environment. If $x=0$, no crime will occur in this plane due to the absence of the prime target. Similarly, in the planes of *zox* and *yox*, when *y* and *z* are 0 respectively, the planes will be missing the target and the spatial temporal environment of the crime, so there will be no crime in either plane. In the space enclosed by the positive direction of *x*, *y* and z-axe, the crime target and the crime interact in the same spatial temporal environment in order for a crime to occur. This forms a crime cube, where the disagreement of the coordinates is the effect of the three elements on the degree of crime occurrence, the likelihood of crime occurrence is expressed by the

distance from the point to the origin.

The coordinates *x*, *y,* and *z* of any point **P** (*x, y, z*) within the crime cube represent the crime target, the crime environment and the three influencing factors, the likelihood of a crime occurring is expressed in terms of the length of **PO**.

We see that as the three elements continue to get larger, the influence on the occurrence of crime increases and the likelihood of crime becomes larger.

It should be the case that there are three faces of the square and three faces of the coordinates that overlap, the points on the ADOD' face, DCCD' face and ACC'A' face of the positive and negative are missing, therefore no crime will occur. Crime targets and the influence of the crime environment on crime are all abstract concepts and the coordinates all need to be represented by data, we assign values to the three factors in order to actually calculate and compare them to predict the likelihood of crime occurring. We have used the British "5x5x5" model, which rates crime, crime targets and crime environments, the three factors affecting crimes are split into five levels, with the levels represented by the numbers 1~5, so that the coordinates of the crime points can be assigned a value.

### 2.3.6 Long Short-Term Memory

CNN networks cannot capture logical order and relationships between the information in each video and cannot learn from inter-sequence dependencies within each video. Better identification of action videos is needed especially for those that have long time spans. Deep learning models which introduce correlation information before and after time sequences are Recurrent Neural Networks (RNNs), which are mostly applied to deal with time series problem. The gradient loss of early RNNs is not able to learn a remote temporal structure over long training periods. Long short-term memory (LSTM) is based on the RNN structure in order to address the research problem.

An LSTM feeds the output of the implicit layer at the past moment into the network together with the input at present to determine what will happen at the next moment, thus preserve the temporal dimension of information. LSTM is initially

applied to obtain the features of image sequences, which are input into the CNN in the order for which they appear in the original image sequence. This effectively improves the recognition accuracy of the model by obtaining the logical order and dependencies across image sequences.

LSTM cells are composed of input gates and output gates, where the input gates determine whether the current input information is retained in the state and the output gates determine the final output based on the combination of current input, previous input, and state information. As a result of the forgetting gate, the data is selectively remembered or forgotten from prior state information and the gradient can be controlled during training while maintaining long-term memory.

### 2.3.7 Combining CNN and LSTM

Combining CNN and LSTM yields LRCN (Long-Term Recurrent Convolution Networks) based on the output of CNN FC6 layer, the LSTM unit is used to determine correlations between video sequences using the LRCN visual appearance and motion information. Streaming networks are combined with scores from spatial and temporal networks to produce the final recognition results. We extracted the dependence between implicit states by using LSTM cross recursion based on RGB images and optical flow, then we employed a SoftMax layer to classify all video frames within a time window based on the average probability. Using the highest probability label, the action label is selected.

### 2.3.8 ST-GCN Combined with LSTM

In crime prediction, the long short-term memory network (LSTM) based on RNN structure is employed, the output is combined with CNN. Because crime often shows spatiotemporal relationships, the prediction algorithms based on topology structure can solve the problems. In this chapter, the spatiotemporal feature extraction algorithm was proposed by using ST-GCN. Since historical cases may have affected future

occurrences (i.e., time-self-relationship), the gradient boosting decision tree (GBDT) is applied to combine the two extracted features and reveal the final crime prediction. GBDT is used to finally predict crimes. Therefore, in this thesis, we propose a crime prediction model based on the fusion of LSTM and ST-GCN.

A brief outline of the basic framework of the model is followed by an explanation of how LSTM and ST-GCN which were employed to construct the model. Hereinafter, we discuss the principles of LSTMs and ST-GCNs (as they are used in the spatial temporal features extraction module and the temporal features extraction module respectively).

The LSTM algorithm is based on RNNs, a multilayer perceptron network specifically was created for the process of sequential data. It is important to remember that the late outputs of the LSTM are highly correlated with the earlier inputs and components of the outputs, the long-term 'memory' and the inputs are related.

In addition to forwarding propagation calculations, backpropagation through time (BPTT), Adam parameter optimization, LSTM were developed as an improved version of the RNN. The RNN is modified by using LSTM because it filters the memorized information and only transmits the information which is needed for rememorisation. It avoids the gradient disappearance or gradient explosion that can occur during the backpropagation of a model when the dependency sequence or multiplicative terms increase. LSTMs are frequently employed in applications for crime prediction because of upon combining new input crime data with historical crime data, the LSTM retains the memory of both crime data sets. In addition, LSTM can reduce the loss of historical crime information due to long-term memory failure.

A crime prediction model based upon LSTM and ST-GCN fusion is proposed in this thesis. This fusion is applied to extract both temporal and spatial temporal features, and finally fuse these features for predicting criminal behavior. Figure 2.2 shows the proposed LSTM and ST-GCN fusion of prediction model. Model components are separated into three modules: Module-I involves temporal feature extraction, module-II has spatial temporal feature extraction, and module-III includes feature fusion. The temporal feature extraction takes use of an LSTM network to extract the trend of crimes

over time. This module was implemented by using ST-GCN. A GBDT model is employed as one part of the feature fusion module to combine spatial temporal features.

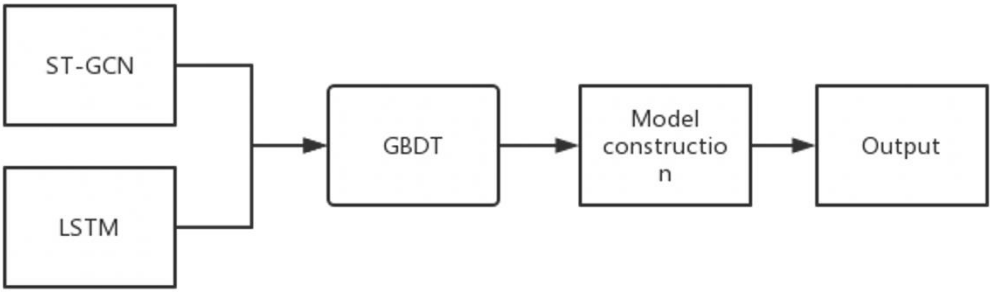


Figure 2.2: ST-GCN+LSTM framework

Gradient Boosting Decision Tree (GBDT) is a modified boosting tree structure designed to be used in feature fusion. GBDT relies on the residuals between the predictions from previous decision trees and actual data to complement the output from each subsequent decision tree, with the exception of the first decision tree. GBDT combines the predictions from each of the decision trees as the final prediction after all trees have been trained.

Followed the calculation of the graph convolution and residuals, the output of the ST-GCN model is fused. Graph convolution data gathered from the external layers are fused with the external fusion data from the external layer through a fully connected layer. As the activation function for the experiments, ReLU was employed and was offered to obtain the final output layer of one node. The loss function and the training method for the ST-GCN model.

The ST-GCN model was trained, loss function was as same as the LSTM model. The regressor GBDT is applied to create the final fusion of the predictions from the two groups of models. Depended on the actual usage, a sliding window training method is used in the training process, i.e., each trained model is only utilized to predict one day data, the model is cyclically trained and predicted by using a sliding window, the prediction results are stored for the final model validation.

## 2.4 Behavioral Feature Extraction

Dependent on the behavioral recognition needs, there are certain differences in the points of focus on human behavior. The extraction of valid features from the video is a major step in behavioral analysis, which affects the accuracy and robustness of the subsequent behavioral analysis. There is also a link between the selection of features and behavioral classes.

There are three layers based on feature behavior extraction: Bottom layer, middle layer, and top layer. The bottom-level features are use of pixel points and prime points as sewing analysis objects. The bottom-level features of videos are extracted based on the bag-of-words model to construct the person behavior module; The scene based on the stream learning framework is applied to explore human movement patterns by mapping spatial temporal features to the scenes.

Middle-level features are mainly a motion pattern as the analysis of visual object, the dense trajectory method based on the probability map of the model to establish the behaviors of the person, a new narrative is proposed based on motion boundary histogram; for calculating the domain eigenvalues of the region of interest in the scene, the stability of the eigenvalues assists to detect the behavioral state of the object; high-level features have a larger spatial scale of the motion trajectory as the point. The direction and velocity magnitude of human motion are based on the optical flow, and clustering algorithms for regions are taken actions to classify of motions and detect human behaviors.

There are three types of algorithms for feature extraction in the video. The algorithms based on tracking or pose estimation, the features extracted by using the algorithms are mainly static features and motion-based dynamic features, the effectiveness of the extraction algorithm is related to the accuracy of pose estimation and tracking. In real environments, the background information is very confusing and there are many moving objects, accurate tracking and pose estimation are not possible, the robustness of such methods is low. The algorithms based on image processing to extract visual features from digital images, usually based on dynamic and spatial

temporal features of the optical flow, which allows the scan of video frames and spatial temporal cube of local movements, are computationally intensive and vulnerable to noises.

The algorithm obtains descriptive attributes by extracting the mid-level semantic features of human body, posture, and scene, which are much accurate for behavior recognition in specific scenes because the "action attribute space" is artificially defined, the uncertainty of human behavior in real scenes makes the accuracy of behavior recognition. The accuracy of behavioral recognition in real scenarios is low.

A summary and comparison of the characteristics in the three behavioral identifications: The static features are represented by the size, color, contour, shape, and depth of the target, the action information is represented by using edge data obtained with the Canny edge detector, which extracts keyframes from the video and compares them with the action. This method has the robustness and is able to accurately recognize similar shapes with multiple poses. However, it is not easy to extract edge and contour information in a complex environment. Dynamic features in human behavior analysis are trajectory, the direction of movement, and speed of movement. The motion trajectory represents the path of the object in space, the speed and direction of movement of the object are calculated based on the trajectory. However, its method is less effective in object detection, tracking, and recognition, and has a high error rate in complex environments.

Spatiotemporal characterization is the attribute of a video because a spatiotemporal body in 3D space such as spatial temporal cubes and spatial temporal shapes to describe the action. The benefits of spatial temporal features are that dynamic changes over longer periods are captured effectively, the continuity on both spatial temporal scales is integrated, the complexity of feature matching is reduced, and robustness to factors such as occlusion is rejected. Descriptive features and behavior recognition algorithms are based on underlying visual features of static, dynamic, and spatial temporal features, which are employed to describe a behavioral class throughout underlying features which is classified by using visual features.

## 2.5 Behavioral Models and Classifications

Human behavior is recognized by using machine learning methods after the visual features have been extracted and represented by using statistical methods. Based on the different descriptions of behavioral patterns, behavioral models are classified into four types. Amid dealing with simple behavior, template matching or judgment is used for recognition. While dealing with complex behavior, the behavior is represented by using a state-space model and decomposing it into atomic actions with a spatial temporal order. When dealing with multi-person interaction behavior or collective behavior, the spatial representation of states is insufficient, a syntactic model is required to represent complex structural relationships.

The template matching method calculates the distance between the template and the candidate video region on the video. If the distance is less than a threshold, it means that the target is detected, a method can perform object recognition on a single frame or sequence of images. The template matching algorithm does not specify that the template features must satisfy the conditions, the features are employed in the feature model. Whilst calculating the distance between the template and the candidate video region, the Euclidean distance is harnessed. The difficulty with the template matching method is the choice of the time interval, but the smaller the time interval, the greater the number of models to be stored, the smaller the difference between the original target features and the template, the higher the accuracy. Conversely, the larger the time interval, the less effective the recognition will be, and the lower the accuracy will be.

The goal of discriminant models is that the features are known, and the probability distribution is computed directly for the corresponding classification. The main discriminant models include SVN, boosting (Boosting), conditional random field (CRF), etc. Random forest algorithm in machine learning has been applied in action recognition models, which contain a classifier with multiple decision trees, the output category depends on the category of each tree. This algorithm can handle large amount of data with high accuracy. Therefore, a balance of computation, storage, and accuracy is needed.

The core of the state space model is the description of motion through a parametric time-series model, where the dynamics of the motion are obtained through changes of state variables. Dynamic Bayesian Networks (DBN) and Hidden Markov Models (HMM) are the most broadly adopted temporal models in state-space-based approaches, which treat the execution of a behavior as a process of implicit state transitions. In the underlying model of the HMM, $t$ represents the current moment, where $y_{t-2}$, $y_{t-1}$ and $y_t$, are the observed variables, $x_{t-2}$, $x_{t-1}$ and $x_t$, show the corresponding implied variables.

A deterministic Markov model, using which action segmentation and recognition was performed simultaneously. In this model, a Viterbi dynamic programming algorithm was designed to perform real-time operations, the sequence of actions needs to follow Markov model by using this type of method for action recognition. Whereas DBN approach is scalable compared to the HMM approach, complex behaviors are decomposed into the combination of atomic actions. A dynamic Bayesian network is applied to describe the sub-actions and obtain the transformation relationships between the atomic actions.

Although these methods are effective in human behavior recognition, the prediction is still high dimensional, the features are sparse and discontinuous, these methods are not suitable for solving the complicated problems (Azizpour, Razavian, Sullivan, Carlsson, 2014).

## 2.6 Abnormal Behavior

Most of the current research work on human abnormal behavior detection is about the detection of abnormal behaviors in crowds or multiple people. The literature takes use of optical flow to describe crowds, extract visual features from optical flow in an unsupervised manner, and create models to predict abnormal behavior in crowds. Although anomalous behavior detection has now been successfully applied to intelligent monitoring public places, it still encounters a lot of challenges.

The basic difficulties are as follows: It is difficult to define normal areas in the video stream. The boundary between normal and abnormal objects is blurred. The

concept of abnormality is quite subjective and changes from scene-to-scene applications. Marking abnormal behavior in the video is also difficult. Normal behavior tends to be evolved.

Most main assumptions to detect anomalies are: Anomalous events rarely occur in video sequences (Hu & Davis, 2018) and anomalous events have low similarity to normal events (Ullah & Uzair, 2009). However, most of these methods are employed for specific scenes (Wen-Cui, 2014), different models are adopted to analyze motion and use different behavioral features.

## 2.7 Behavioral Model Assessment

In comparisons, dataset D consists of two sets, one serves as the training set **S** and the other presents as the test set **T**, i.e., $\mathbf{D} = \mathbf{S} \cup \mathbf{T}$, $\mathbf{S} \cap \mathbf{T} = \varnothing$. After trained the model using the training set **S**, **T** is applied to evaluate the test error as an evaluation of the generalization error.

Hence, firstly, the partitioning of the training and test sets needs to be as consistent as possible in terms of data distribution in order to avoid additional bias being introduced into the data partitioning process, which may have impact on the results. Given the sample proportions of the training and test sets, there are also multiple ways of partitioning the initial dataset, which also have an impact on the evaluation results of the model. The results obtained from a single use of the leave-out method are therefore often less stable and reliable. Using the leave-out method, several random divisions are generally used, repeated, and then averaged as the evaluation result of the leave-out method. By evaluating the performance of a model, the leave-out method requires a division of the training and test set, which leads to a problem: If the training set contains most of the samples, the trained model may be equivalent to the model, but the evaluation results may be unstable and inaccurate due to the small size. If the alternative test set contains more samples, the difference between the training set will be even larger. If the other test set contains more samples, the training set **S** will be more different from **D**, the results of the evaluated model may have errors compared to

the model trained with **D**, thus reducing the evaluation results. Therefore, about 2/3 to 4/5 of the samples are employed for model training, the rest samples are employed for testing.

Cross-validation is a process in which the original dataset is applied to train the model and evaluate it. Training sets are applied to train a model, validation sets are employed to select and configure model parameters, the test sets are offered to determine whether the model is generalizable.

In *k*-fold cross-validation, the training set is split into *k* identical subsets, the number of training samples is assumed to be *n*. Each subset has *n/k* training samples, and the corresponding subsets are $\{s_1, s_2, ..., s_k\}$. Once a subset is employed as the test set, another *k*-1 dataset will be employed as the training set. The classification rate is obtained by placing the trained model or hypothesis function on the test set. The average of *k* times, the classification rate is calculated, which is utilized as the final classification rate of the model or hypothesis function. This method makes use of all the samples. However, the calculation is tedious and requires *k* training sessions and *k* testing sessions.

Bootstrapping is another method of model validation (evaluation), based on bootstrap sampling, i.e., sampling with return or repeated sampling. In a dataset with *m* samples, one sample is randomly selected at a time as a training sample, then put back into the dataset, thus sampling back and forth for *m* times to produce a dataset of the same size as the original dataset. In this way, a plethora of samples may appear in the training set several times, while others may never appear. The new dataset will not contain approximately 36.8% samples from the original set. In a dataset with *m* samples, it is 1/*m* each time a sample is drawn, it is (1-1/*m*) that no sample is selected. As the value of *m* tends to infinity, the probability of a sample not being drawn is the negative power of *e*, which approximately equals to 0.368.

$$\lim_{m\to\infty}(1-\frac{1}{m})^m \to \frac{1}{e} \approx 0.368 \quad . \tag{2.2}$$

We take use of these samples, which do not appear in the new dataset as the

validation set. The previous steps are repeated several times so that several models are trained, and the validation errors are obtained, and averaged as the validation error of the model. The advantage of this method is that the total number of samples in the training set is *m*, as in the original set, that about 1/3 of the data does not appear in the training set that can be used as the validation set. The disadvantage is that this produces a training set with a different data distribution than the original data set, which introduces estimation bias.

## 2.8 Datasets for Human Behavior Recognition

Tagging human behavior on an appropriate time is difficult. Most publicly available human behaviors are from video datasets, such as UCF101, HMDB51, Kinetics, contain plentiful behaviors that can be recognized with only still frames and a small number of video frames. The diversity of data types, the irregularity of video clip duration.

UCF101 is a collection of real-time action videos collected from YouTube that are classified into 101 action categories. It has the most diverse collection of 13,320 videos by 101 action classes, as well as a broad range of camera movements, object looks, and poses, points of view, cluttered backgrounds, lighting conditions, and more. It has so far been proven to be a challenging dataset. There are 101 action videos grouped into 25 classes, each of which contains four to seven action videos. In the same group of videos, backgrounds and perspectives are the same. A class of actions is human interactions, body movements, playing musical instruments, and movement.

HMDB51 was released by Brown University in 2011, with videos mostly from films, and some from public databases and online video repositories such as YouTube. The database contains 6,849 samples, grouped into 51 classes, each class contains at least 101 samples, with a resolution 320×240. The actions mainly include: 1) General facial actions smiling, laughing, chewing, talking. 2) Facial actions with objects: Smoking, eating, drinking. 3) General body actions: Cartwheeling, clapping, climbing, climbing stairs, jumping, landing on the floor, backhand flip, handstand, jumping,

pulling, pushing, running, sitting down, sitting up, heeling, standing up, turning, walking, wave. 4) Actions interacting with objects. 5) Human action.

Kinetics videos are sourced from YouTube and are currently available in three versions, including 400, 600, and 700 categories, containing 200,000, 500,000 and 650,000 videos, respectively. Each clip has a behavioral class, annotated by hand that is around 10 seconds in duration. The classes in the dataset fall into three main categories: Single interaction, such as playing a musical instrument; person interaction, such as shaking hands and hugging; and sports.

## 2.9 Problems of The Existing Methods

With the in-depth study of human behavior recognition, the recognition has achieved particularly satisfactory results, but there are still a spate of problems in practical. The problem of multiple behavior recognition in video exists. The individual samples in the existing set are segmented, but there are multiple behavioral categories and unpredictable interference for a given video in a real scene. The large randomness of human behavior leads to a large variation in the features of the same class of actions. The scale of the data is insufficient, the behavior recognition methods in deep learning are use of complex neural networks that require large samples of video data sets.

In practice, collecting and labeling the dataset become exceedingly difficult. Moreover, training deep neural networks takes a lot of time, while behavior recognition by deep learning has high computer requirements.

# Chapter 3

# Methodology

For the methodology, firstly, the dataset is collected. Secondly, a method for human behavior recognition based on ST-GCN is proposed. Then, abnormal behavior recognition based on TRN is introduced, a method for abnormal behavior recognition based on temporal relational networks is proposed.

## 3.1 Data Preprocessing

We have chosen dataset UCF101 with 40 classes. Around 2,400 (80%) samples are selected for training, another 300 video clips are chosen for testing. The remaining samples are defined as non-criminal cases. The clips have resolution 340×256 with the frame rate at 30 FPS.

The main steps of data preprocessing are spit into fourfold: Data cleaning, data integration, data statistics, and data conversion. The main purpose of data cleaning is to resolve problems existing in the data by adding missing values, smoothing noisy data, or removing contours.

The reasons of data loss are various due to the objectives. The approach to deal with these missing values is primarily based on the distribution of the variables and the significance of the variables. Depending on the rate of missing variables, deletion or padding methods are employed. If the variables to be populated are continuous, they are populated by using the mean and random difference methods; if the variable is discrete, the median or dummy variable is applied.

Outlines are treated as anomalies that affect the quality of the data amid the data processing, using a relatively simple and intuitive method of identifying outlines for variables in conjunction with box plots and MAD statistical methods. In determining whether to remove the data, the number and impact are considered.

Noise is the random variance of a variable which is the error between the observed and actual data points. The method is to manipulate the data into containers, separate the frequency which has equal width of containers through using the mean, central, or boundary values of each container instead of all the numbers in the container, which acts as a smoothing filter of the data. Another approach is to construct a regression model of that variable and predictor variable by using an approximation based on the regression coefficients and the predictor variable.

## 3.2 Criminal Behaviors

There are three categories of criminal behaviors: (1) Vandalism; (2) Abuse, assault, burglary, shoplifting; (3) Shooting and arrest. The definition of these human behaviors is listed in Table 3.1.

The dataset UCF101, including 540 crime classes, is taken into account in this book chapter, 240 videos were employed for training, the other 300 clips were for testing purposes. The remaining clips are defined as non-criminal cases. The format of the clips is with the resolution of 340x256, the frame rate is 30 fps. As shown in Table 3.1, in this book chapter, human behaviors in the dataset are classified as Grade I, Grade II, and Grade III. By grading the criminal behaviors, alarming treatments are carried out according to the levels of criminal behaviors, which reduces the probability of false alarms.

Table 3.1: The levels of criminal behaviors

| **Grades of criminal behaviours** | **Behaviour performance** | **Basis of classification** |
|---|---|---|
| Grade I | Vandalism | Destruction of public goods with purpose and intent, using violent means. |
| Grade II | Abuse, Assault, Burglary, Shoplifting | Using violent means with purpose and intent to harm another person behaviour. |
| Grade III | Shooting, Arrest | Transgressions, which are acts that may result in the safety of another person's life. |

## 3.3 Human Pose Estimation

Human pose estimation plays a vital role in computer vision. It is employed extensively in motion recognition, autonomous driving, and abnormal behavior recognition, while human pose estimation algorithms are also employed extensively in criminal prevention. Human pose estimation has evolved from traditional methods based on holistic features and human models to the algorithms based on depth estimation. The algorithms for human pose estimation have been evolved from single-person human pose estimation to 2D human pose estimation, and finally refined to obtain multi-person human pose estimation. However, the accuracy of human pose estimation is often less than ideal one due to occlusion, view angle, and the randomness of human movement.

Human posture assessment is mainly based on human behavioral movements. The sequence data of the key 18 skeletal points of human body is acquired by using human posture estimation algorithm, the sequence data of skeletal points are applied as the input to the ST-GCN and predict the actions by using pattern classification with softmax function. Accurate human pose estimation is beneficial to improve the accuracy of action recognition. In this thesis, two different methods of human pose evaluation were utilized to test the ST-GCN. A bottom-up human pose estimation method OpenPose, and a top-down human pose evaluation method AIphaPose, were commenced to compare the effectiveness of OpenPose and AIphaPose in the recognition of abnormal behavior.

In human pose estimation, the main objective is to locate the skeletal nodes of human body in a video frame and connect them together. In order to accomplish human pose estimation, there are two general approaches. One is to categorize the coordinates of the skeletal points so as to locate the skeletal nodes. Another way is to determine the probability of a skeletal node in every pixel of each node in the human pose.

OpenPose is a bottom-up algorithm for human pose estimation. With this algorithm, the key nodes of all the bones in the image are firstly obtained, then the key points are classified by clustering, and finally, these key points are connected to obtain human skeleton.

The OpenPose network is grouped into two stages, the first stage is the top half of the image, which is used to predict the confidence map of the human joints. OpenPose takes use of two loss functions to predict the confidence map, the part affinity fields (PAFs) of the skeletal joints. OpenPose also has a spatial weighting mechanism in the loss functions, which is robust to human postures by using the training set. The two loss functions are represented as Figure 3.1.

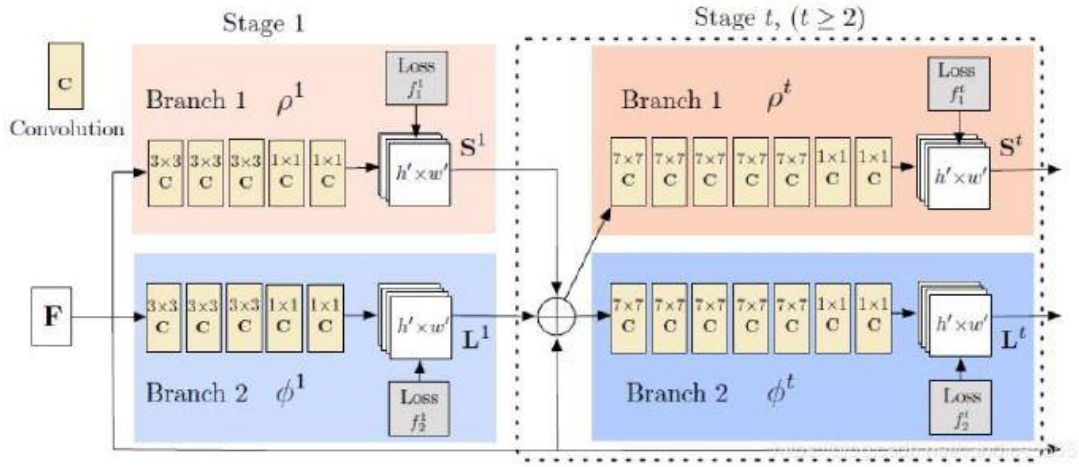


Figure 3.1: OpenPose structural diagram

$$f_s^t = \sum_{j=1}^{J} \sum_{p} w(p). \| S_j^t(p) - S_j^*(p) \|_2^2 \tag{3.1}$$

$$f_L^t = \sum_{c=1}^{c} \sum_{p} w(p). \| L_c^t(p) - L_c^*(p) \|_2^2 \tag{3.2}$$

For the detected key points connected to the human skeleton, OpenPose algorithm generates a confidence map based on the image of the annotated 2D key points for calculating the likelihood of a particular part of human body occurred at a point on the image in the confidence map.

The AlphaPose-based pose estimation method is a top-down approach, the algorithm investigates the extraction of human skeletons in complex environments. The main problem of AlphaPose is to handle errors and redundant data with human behavior recognition. In human behavior detection, AlphaPose adopts an asymmetric spatial temporal variation network to improve the accuracy of behavior detection and a parametric pose non-maximal suppression method to deal with redundant data. A pose-

guided region generator is also designed.

Asymmetric spatial-temporal variation network starts with the RGB images, after labelled the frames with human body, a spatial transformation is conducted based on each frame region to finally obtain the results of pose estimation. The requirements for the frames in the videos are high in the existing pose estimation algorithms, any problems with the over detection frames have a significant impact on the accuracy of object detection.

In order to avoid redundant data in human detection, AlphaPose takes use of a parametric pose non-maximum suppression method. The method of parametric pose non-maximum suppression consists of the predicted pose with a high confidence level in the pose set by comparing it with other poses one by one and eliminating the redundant data according to the elimination criteria.

## 3.4 ST-GCN-Based Behavior Identification

ST-GCN is a combination of TCN and GCN. TCN performs convolutional operations on the visual data in the temporal dimension and GCN performs convolutional operations on data in the spatial dimension. GCN belongs to GNN, which is based on graph theory. The traditional data handled by using neural networks are all structured data reserving Euclidean distance, such as 2D images, 1D sounds, and so on. For non-Euclidean distance data, such as social networks, transportation networks, etc. the traditional network structure cannot be directly processed, GNN is applied to tackle this type of data. The current algorithm for target skeletal behavior recognition adds the connectivity of adjacent key points to improve accuracy and add spatial structural features of human body.

It generally needs three parts to develop a spatiotemporal map of a skeleton sequence. The first part is to construct a spatial map of natural joints of human body together for each frame of a video image. Additionally, we connect the same points from two adjacent frames so as to form the temporal edges. Finally, all the key points in the input video frames form the set of nodes, all the edges in Step 1 and Step 2 form

the set of edges **E**, which comprise of the desired spatiotemporal map. The edge set consists of two subsets, the first one of which is the concatenation subset within the human skeleton, which is denoted as Eq. (3.3).

$$E = \{V_{ti}V_{tj} \mid (i, j) \in H\} \tag{3.3}$$

where $H$ is the set of lines between joints.

The second subset is the set of concatenations between frames, obtained by joining the same key points of two adjacent frames, which is represented as Equation (3.4).

$$E_F = \{V_{ti}V_{(t+1)}i\} \tag{3.4}$$

The skeleton data is usually a sequence of video images, if the edges between the skeleton nodes are considered as a 2D grid, like a 2D image, the output after the convolution operation is also a 2D grid. Because convolutions do not change the original data structure. Thus, the coordinate vectors of the ST-GCN input graph nodes are like the vectors of pixels located at a 2D image grid as input to an image-based convolutional network.

The skeletal sequence is represented by using the coordinates of key points of human body, the spatiotemporal map $G = (V, E)$ represents the relationship between the temporal and spatial connections between $N$ joints and the skeletal sequence of $T$ frames, while each key point has the coordinates and confidence level of that point constituted. As the human skeleton is based on local space, a specific space allocation strategy is needed in the partitioning process. In this thesis, ST-GCN features were extracted to perform criminal behavior recognition. The method firstly extracts the behavioral skeleton sequence of the target from the videos, identifies it by using ST-GCN algorithm, and finally classifies the criminal behaviors of the target by using softmax classifier.

The GCN aids us to extract the local features of adjacent skeletal nodes in human skeleton space. It is not enough to analyze the features in space, we also need to have the local features of the skeletal nodes as they are changing in time. Temporal Convolutional Networks (TCN) and Long Short-Term Memory (LSTM) are the two main research areas in this area.

ST-GCN takes use of TCN to analyze local features of joint points that change over time. Since the shape is fixed, temporal convolutions are fulfilled by using regular layers. It is akin to the convolution operation for images. The shape of the last three dimensions of the ST-GCN feature map corresponds to the shape of the image feature map.

TCN is relatively easy to be understood. The input to ST-GCN is (**C**, **V**, **T**). With the input images, the channel also has an initial value 3, but the difference is that the two dimensions of the graph matrix **T** and **V** are independent of each other, **T** represents time and **V** refers to the number of nodes, a node corresponds to a pixel on the image. Each row in the matrix represents the same node, with a state in different time, each column represents all nodes at the same time, which is the graphical structure.

After the convolution of all ST-GCN units, the number of joint feature dimensions was increased to 256 and the number of keyframe dimensions was reduced to 38. Finally, the average pool and FCN were applied to classify the features. Human skeleton map is obtained to tackle the raw videos by using OpenPose and AlphaPose. The algorithm takes on various joint points of human body. The video is split into frames as the inputs for OpenPose and AlphaPose, the feature vectors are obtained after a 10-layer VGG-19. The feature vectors are fed into two branches so as to predict the vectors of confidence and affinity for each joint point. Finally, the clustering of the key points of the features and the assembly of the skeleton is output. As a preprocessing tool for the video, OpenPose as a software tool for processing videos, OpenPose has a higher confidence level and model robustness. The accuracy of skeleton detection of the videos after OpenPose processing is higher. In the case of a single frame image, the output of ST-GCN, according to the spatial partitioning strategy, is shown in Equation (3.5).

$$f_{out} = \Lambda^{-\frac{1}{2}}(A+I)\Lambda^{-\frac{1}{2}} f_{in} w \tag{3.5}$$

where weight matrix **W** is composed of the weight vectors of multiple input channels. The input feature map is represented as a tensor of (**C**, **V**, **T**) in spatiotemporal way.

In a 2D convolution, the grid naturally exists around the center. Therefore, the pixels in the domain have a fixed spatial order. The pixels are then indexed by using

tensors according to the spatial order. We construct our weighting function based on this idea. The process was accommodated to simplify the process by dividing the set of neighbor skeleton points into a fixed number of subsets by splitting, each subset has a number of labels as shown in Equation (3.6).

$$w(v_{ti}, v_{tj}) = w'(l_{ti}(v_{tj})). \tag{3.6}$$

As ST-GCN shares weights between nodes, it is important to maintain a consistent scale of input data of nodes, therefore the data needs to be normalized before it is taken into account. In ST-GCN network modeling, there are nine layers of spatiotemporal graph convolution. In general, the first three layers have 64 output channels, the middle layers have 128 output channels, the last three layers have 256 output channels. Nine spatial temporal dimensions are contained in each layer and ResNet was applied to connect each ST-GCN. To avoid overfitting, the features from the last 50% of each ST-GCN cell are randomly selected, a global pool at the top of the network can handle input sequences of uncertain length.

In the process of identifying behavior, the data is firstly normalized by feeding the skeletal sequences into the normalization layer. The tensor is then globally aggregated to produce a 256D feature vector for each sequence. In addition, the ST-GCN network is trained by using stochastic gradient descent with a learning rate of 0.01. Each skeleton sequence is transformed affinely to simulate the motion of the camera. That is, from the first frame to the last frame, fixed angle, translation, and scaling parameters are selected as candidates, then a combination of these three factors is randomly chosen to generate the affine transform. Secondly, a random sample of the original skeleton sequence segments is selected in training, while all video frames are taken into account for testing.

# 3.5 Behavior Identification Based on TRN

Temporal segmentation network requires dense sampling of video frames, a sparse temporal sampling strategy is proposed, which is to divide the input video into *K* segments, regardless of the length of the video, then a random time segment is found in

each segment by using CNN to extract the spatial features respectively and perform the feature-level fusion. Finally, SoftMax classification is performed as shown in Equation (3.7)

$$TSN(T_1, T_2, ...T_K) = H(G(F(T_1; W), F(T_2; W), ...F(T_K; W))). \quad (3.7)$$

Temporal relational reasoning is the ability to understand the relationships of people or visual objects in the temporal domain. TRN is a real-time temporal relational reasoning framework for video-level on top of the TSN framework for learning and reasoning about temporal dependencies between video frames. The main contributions of TRN are that the design of new fusion functions characterizes the relation of different temporal segments, the improvement of video-level robustness through multiscale feature fusion in the temporal dimension.

This chapter presents a TRN approach to human behavior recognition. The method utilities a TRN network to analyze the temporal relationships in the frames before and after the video by learning and reasoning about the temporal relationships of all frames in the video. By sparsely sampling the video frames, the TRN network can accurately detect human interactions and has significant performance in recognizing human postures. Our analysis shows that TRN-based network models typically yield intuitive visual perception capabilities, better compared to dual-stream networks and 3D convolutional networks.

In real reality, people analyze and deduce what will be happened in the future by looking at the past and the present, this is conducted by reasoning temporal relationships.

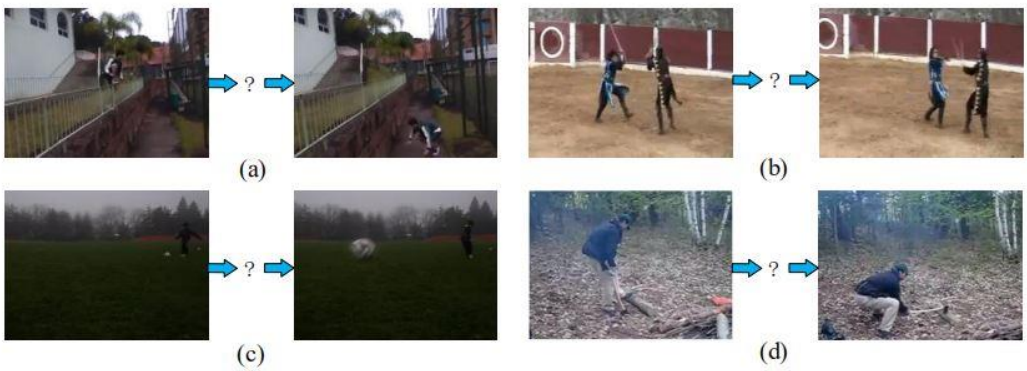


Figure 3.2: Temporal relationships in a video

In Figure 3.2, the temporal relationship between two states in a realistic situation is reflected by using the poses of human bodies, the temporal relationship between temporal states is inferred by using the relationship between two frames in the video. A single human behavior contains multiple temporal relationships, such as long and short temporal relationships.

Human behavior recognition in digital videos is an important topic in computer vision, but the ambiguity of describing behaviors on a temporal scale poses a significant challenge to neural networks.

In terms of both data and observations, CNNs are extremely limited, the basic structure is characterized by using temporal relationships. The goal of TRN is not to model space, but to describe temporal relationships in videos, while TRNs tackle temporal relationships over multiple time scales. CNN is mixed with TRN because it is a general scalable model. In recent years, numerous powerful CNNs for human behavior recognition have emerged because of visual object recognition.

The following examples to show how to combine RGB frames in the time dimension by using 3D CNN networks to extract feature information from the frames and integrate the feature information by using LSTM models.

The existing CNN networks for behavioral recognition suffer from a great deal of problems at the same time, for example, optical flow can extract features and reduces redundant values in consecutive frames and the difficulty to reveal long-term temporal relationships. To address these issues, TRN takes use of a sparse sampling of video frames. The TRN with temporal relationship inference can recognize human interaction, while the TRN has a strong predictive power over other networks for video frames, with a significant improvement in accuracy. The composite function for time series analysis is defined by Equation (3.8).

$$T_2(V) = h_{\varphi}(\sum_{i<j} g_{\theta}(f_i, f_j) \tag{3.8}$$

An ordered frame of a video **V** is specified by **V**= $\{f_1, f_2, f_3, ..., f_n\}$, where $f_i$ represents the representation of the $i$-th frame of a video. A number of frames are

combined by $h$ (·) and $g$ (·). The multilayer perceptron (MLP) is implemented in this case with the parameters $\varphi$ and $\theta$. To speed up the computation, a pair of frames $i$ and $j$ are sampled and ordered. The composite function for the 2-frame relationship is further extended to a higher relationship, e.g., a 3-frame relationship function as shown in Equation (3.9)

$$T_3(V) = h'_\phi \left( \sum_{i<j<k} g_\theta'(f_i, f_j, f_k) \right) \tag{3.9}$$

where the sum of the frames exceeds the sampled and ordered frames $i$, $j$, and $k$. Using Equation (3.10) as a composite function, we represent frame relationships on different time scales and capture temporal relationships.

$$MT_N(V) = T_2(V) + T_3(V) + \ldots + T_N(V) \tag{3.10}$$

Each relational term $T_d$ captures the temporal relationship between $d$ ordered frames. Each $T_d$ has two independent functions $h$ (·) and $g$ (·). All the temporal relationship functions are distinguishable end-to-end, so they all have been used in CNN to extract video features. The entire network is shown in Figure 3.3, which illustrates the TRN network for 2, 3, and 4 frames.

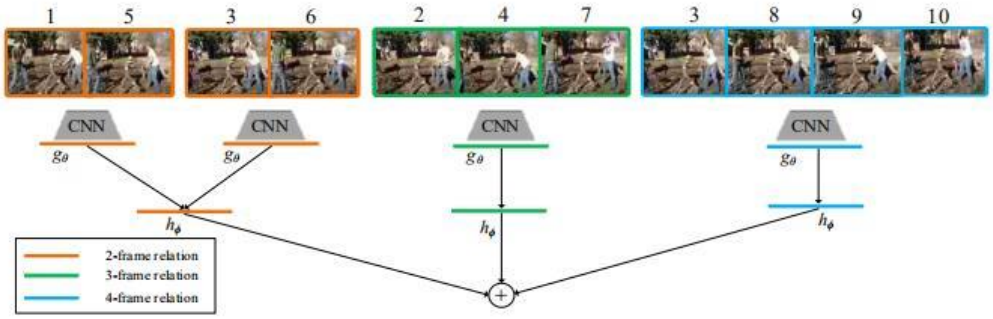


Figure 3.3: TRN Network

As shown in Figure 3.3, the video is firstly segmented. It is necessary to ensure that the sampled keyframes are ordered so that no errors occur in the process of fusing local timing and spatial information, the network structure is based on 3D CNN. After trained a multiscale timing network, we sample the sum by choosing a different $d$-frame for each relational term $T_d$ of the video. A reduced sampling scheme is then employed.

Firstly, we sample a set of $N$ frames, $\mathbf{V}_N \subset \mathbf{V}$, uniformly from $N$ segments of the video and use $V_N$ to compute $T_N(V)$. Then, for each $T_d$<N, $k$ random down-samples of $d$ frames $V_{kd} \subset V_N$ are chosen, these are applied to compute $d$-frame relations for each $T_d(\mathbf{V})$. CNN is run based on $N$ frames while sampling the temporal relations for $K_N$, all parts are trained together by using end-to-end way.

For the sake of tests, the networks equipped with TRN were combined with feature sequences to efficiently process streaming videos. The videos are sampled to obtain isometric frames, the features of the isometric frames are extracted by using CNN. Distinctive features are combined into the relationship maps to further predict the activity at each stage. Visual features are extracted from key frames for prediction by using CNN. The TRN network is run in real time that has the ability to deal with live videos sensed from web cams.

The perceptual knowledge in TRN allows for more explainable structure than in C3D and dual-stream convolutional neural networks. The visual generic knowledge from TRN can be better analyzed through temporal inference. The nature of human observation behavior is predicted and inferred by acquiring representative frames in a video clip. In order to obtain a sequence of representative frames in the TRN, it is necessary to compute features of equal-length frames in the video and input the relational tuples of different frames into the TRN after randomly combining these equal-length get features to generate the relational tuples of different frames. Finally, the relationship tuples of different frames are ranked.

Amid identifying the same behavior with TRN model for different video frames, for complex behaviors, two frames are no longer able to identify the behavior using the TRN. Therefore, additional frames are added to the TRN to accurately identify the behaviors, which improves the accuracy of the predicted behaviors.

TSN and TRN are two models, TRN takes use of pooling operations on video frames, the temporal relationship (TR) is applied to emphasize on the temporal dependence of frames, while average pooling is applied to ignore temporal order. We evaluate both pooling operations on details. The usage of averaging pools and TR pools leads to different experimental results and illustrates the importance of temporal order

in the video dataset.

## 3.6 C3DP-LA + ST-GCN

The deep fusion of C3DP-LA + ST-GCN is to combine the intermedia representations of base networks as the input of the rest of each base network, and then take use of the depth of several intermediate representations. This deep fusion network has the following advantages:

(1) It can learn multi-scale representations because it can have the advantages of more base networks.

(2) It is composed of a deep base network and a shallow base network. The information flow from the early intermediate layer of the deep base network to the output, from the input of the deep base network to the later intermediate layer is all. improved.

(3) Joint learning of deep base network and shallow base network can benefit each other.

An algorithm framework for multimodal feature fusion based on human behavior recognition is shown in Figure 3.5. The C3DP-LA network and the ST-GCN network are two main models in visual feature extraction, where C3DP-LA network is an improved 3D CNN and a spatial temporal attention. The skeletal sequences of human behavioral samples are obtained from video frame sequences based on the inputs (Aggarwal, Xia, Chen, 2012). In C3DP-LA network, 3D CNN with spatial temporal pyramidal pooling can automatically process RGB video frames of arbitrary size and obtain preliminary spatial temporal features by using fast convolution, followed by LSTM module with its memory function to pass information from the current or even earlier moment to the next moment, further extracting temporal features and enhancing key information with spatial temporal attention mechanism. The final RGB video features are obtained.

By constructing a multilayer spatial temporal map over the skeletal sequence and performing a multilayer spatial temporal convolution operation, ST-GCN can generate more advanced skeletal features. Lastly, the two extracted modal features are fused together by using a standard SoftMax model so as to achieve action classification. The

detailed descriptions of each module are presented in Figure 3.4.

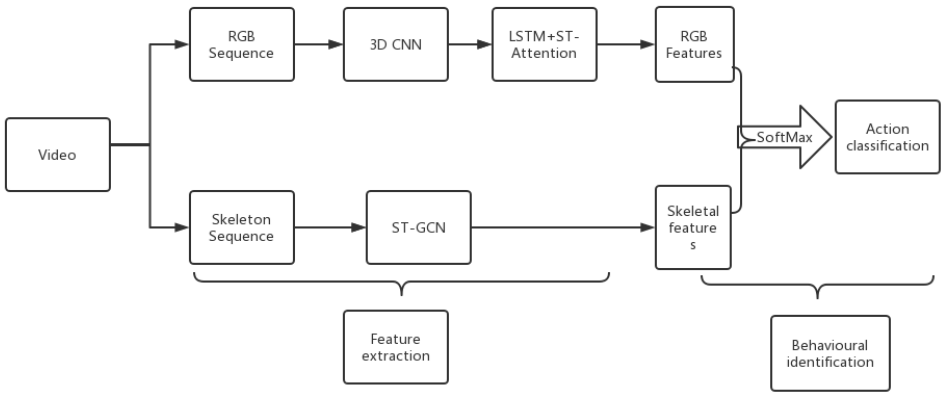


Figure 3.4: C3DP-LA + ST-GCN framework

The 3D convolutional networks not only capture the appearance of the target but also acquire the motion information, they are working together for human behavior recognition. The structures are simpler and faster than current behavior recognition models. Multiple consecutive frames are stacked together to form a cube, different kernels are applied to perform 3D convolution on each channel, the feature maps are concatenated to adjacent frames, allow for spatial information to be extracted while temporal information is obtained. In order to compute various types of visual features, all channel information is combined.

3D convolution was applied to extract the temporal features of the video. Multiple video frame sets are applied to 3D convolution, where the number of channels d is smaller than the number of frames *L*, and the convolution results in a 3D feature map that retains temporal information. Although 3D convolution is more suitable for video data, it is difficult to build deep convolutional networks due to the additional depth dimension of its convolution kernel, which adds many more parameters than 2D convolutional networks.

Three convolutional layers, a hardline layer, and two down sampling layers compose a 3D CNN structure, which takes as input seven 60×40 frames. The hardline layer extracts the five channels of grayscale, horizontal gradient, vertical gradient, optical flow, and optical flow from each frame to generate 33 feature maps. The C2

convolution layer takes use of two different 3D kernels to convolve each of the five channels of information output from the previous layer, C4 convolution layer is use of three different convolution kernels to convolve each of the feature maps so as to obtain more feature maps with both spatial temporal dimensions. Convolution layer C6 convolutions each feature map with a 7×4 2D kernel to generate 128 feature maps, i.e., 128D feature vectors of behavior information in the input frame. As the final layer of convolutions, layer C6, convolve each feature map by using 7×4 2D kernels to acquire 128 feature maps, that is, 128D features from the input frame. The feature vectors are sent to the fully connected layer for human behavior recognition.

In order to train and test the 3D CNN, the fully connected layers within the network must have a fixed size and frame size. If the videos with any resolutions are input to the 3D CNN, it will crop or scale down to create fixed size samples. This will create distortion, as well as reduce the amount of useful information. In this thesis, we replace the last pooling layer in the 3D CNN with a spatial temporal pyramidal pooling layer to tackle the input of arbitrary sizes, transform them into fixed-length feature vectors and extract more features from temporal viewpoint.

It receives inputs of any resolutions and generate outputs of any size, as the convolutional layer can receive inputs. The feature map of the last convolutional layer is taken as the pooling cube, the width of the frame, given an RGB video sequence as the input to a 3D CNN. While conventional sliding windows pooling in 3D CNNs uses a fixed size of the sliding window to adjust for the number of features that are generated by the pooling layer, STPP dynamically changes the size of the sliding window. In order to compute the size of each pooling cube, we define **P** ($p_t$, $p_s$) as the temporal pooling, where $p_{t.}$ indicates the temporal and $s_{pa}$ shows the spatial pooling. Using $p_s$=4,2,1 and $p_t$=1, the output of different size convolutions is converted into a fixed-dimensional feature vector and fed into the fully connected layer. The responses are pooled by using the max pooling of the spatial temporal cubes. Using STPP, we configure the improved 3D CNN so as to support video frames of any scales.

Human behavior recognition using 3D CNNs is based on appearance of the target, and its motion information. Convolution kernels are employed on each channel to create

a cube, and successive frames are stacked onto the cube to obtain 3D maps linked to adjacent frames in order to extract spatial information and temporal information simultaneously. Combining all channel information enables the computation of distinct types of features.

An eight-layer CNN structure allows input from seven frames with 60×40 dimensions and outputs three layers of convolutions, one layer of hardline, and two layers of down sampling. The hardline layer extracts the five channels of grayscale, horizontal gradient, vertical gradient, optical flow from each frame to generate 33 features maps; C2 convolution layer takes use of two 3D kernels so as to convolve each of the five channels of information output from the previous layer, the C4 convolution layers use three convolution kernels to convolve each of the feature maps so as to obtain more feature maps. C6 is the convolution layer that requires each file to be convoluted with a 7×4 2D kernel to generate 128 feature maps, i.e., 128D feature vectors based on the action information in the input file. In order to obtain 128 feature maps from the input frame, 7×4 2D kernels are applied to a final layer of convolution, layer C6. A fully connected layer receives the feature vectors.

# Chapter 4

# Results and Analysis

The main part of this chapter presents the crime prediction model and the results of human behavior detection of our experiments. We will also evaluate the behavior model. Finally, we will demonstrate the experimental results of the proposed model.

At present, there is little research outputs based on deep learning for crime prediction and recognition. Most of the literature investigated human abnormal behaviors of crowds and the recognition of a single abnormal behavior, but there are truly little research outcomes on the prediction of criminal behaviors. In this thesis, we take advantage of the ST-GCN model and TRN algorithm to identify criminal behaviors.

# 4.1 ST-GCN Results

The ST-GCN takes use of the position of skeleton joints as an input parameter. In order to obtain the joint nodes, all video resolutions were firstly resized as 340×256, the video frame rate was normalized to 30 FPS. The positions of 18 joints on each frame of the clip were then evaluated, where the positions of human joints are represented by using 2D coordinates in the coordinate system. Thus, a node is applied to represent each joint, an array of 18 nodes is combined to form human skeleton. Pertaining to multiperson detection and skeleton extraction, only two joints with the highest confidence in the human skeleton were selected in each video clip. In this way, the frames of video clips were converted into a skeleton sequence.

In this thesis, we take use of UCF101 crime dataset which contains a variety of crimes, starting with the configuration of the ST-GCN environment. Whilst configuring the ST-GCN model, we need to use Cmake to compile OpenPose and AlphaPose as the operating system is Microsoft Window 10. While downloading and installing Cmake, the data in the target build folder has not been working properly. Also, the bins under the build file do not produce any of the required files.

There is a slew of advantages of Google Colab. Firstly, there are free GUP available on cloud, but it is important to note that the use of GUP on Colab is limited; Secondly, there are fewer problems with environment configuration on Colab; Finally, it is faster. We think Colab as an Ubuntu virtual machine with GPUs, except that we only operate them from the command line. After mounting Google drive, a drive folder will be created in the virtual machine, we treat the Google drive as a hard disk. Colab was used continuously for up to 12 hours, the system will cut off the running program

and take back the virtual machine.

Figure 4.1 shows the ST-GCN algorithm for human behavior recognition. The white dots and white lines represent the skeletal outline of human body in each image, the text of the video represents the result of human behavior detection and recognition. From the experimental results, we see the algorithm can detect the human body accurately and make a correct estimation of human behavior, thus we complete the behavior detection.

In this thesis, we use UCF101 dataset, on which OpenPose and AlphaPose were compared for the recognition of abnormal behavior. The evaluation criteria against human pose with three parameters in the equations (4.1~4.3), namely, prediction accuracy, recall, and accuracy. The metrics take values ranging from 0 to 1, which include.

$$a = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{4.1}$$

$$p = \frac{Tp}{Tp + Fp} \tag{4.2}$$

$$r = \frac{Tp}{Tp + Fn} \tag{4.3}$$

Figure 4.1: Identification of risky behavior

Table 4.1: Evaluation results from OpenPose

| | Grade I | Grade II | Grade III |
|---|---|---|---|
| Prevision | 60.16% | 61.91% | 58.13% |
| Recall | 73.52% | 72.43% | 79.62% |
| Accuracy | 81.24% | 72.34% | 81.17% |

Table 4.2: Evaluation results from AlphaPose

| | Grade I | Grade II | Grade III |
|---|---|---|---|
| Prevision | 73.21% | 75.32% | 71.26% |
| Recall | 64.13% | 61.71% | 61.01% |
| Accuracy | 83.05% | 84.53% | 82.84% |

Based on our experiments, we see that OpenpPose is not exactly accurate in recognizing human actions, which results in errors at key points in the recognition of human behaviors. AIphaPose, on the other hand, can accurately import the computed sequences into the behavior recognition model for classification. In the ST-GCN algorithm, there is no need to obtain the background in the video in advance and the recognition accuracy is high. Also, ST-GCN performs well in multiperson behavior recognition, but the time required for recognition is longer as shown in Figure 4.2.

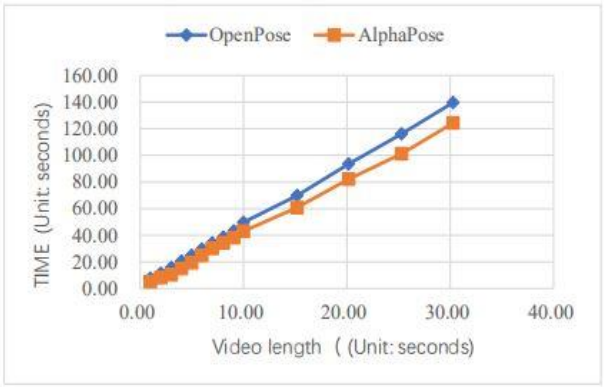


Figure 4.2: The timelines

In this thesis, video data from real situations was employed to recognize behavior in the video, the result of behavior recognition consists of four components, the original video frames (original video), the pose map (pose estimation), an attention model, a prediction graph (attention + prediction), an attention mechanism, and RGB

(attention+RGB).

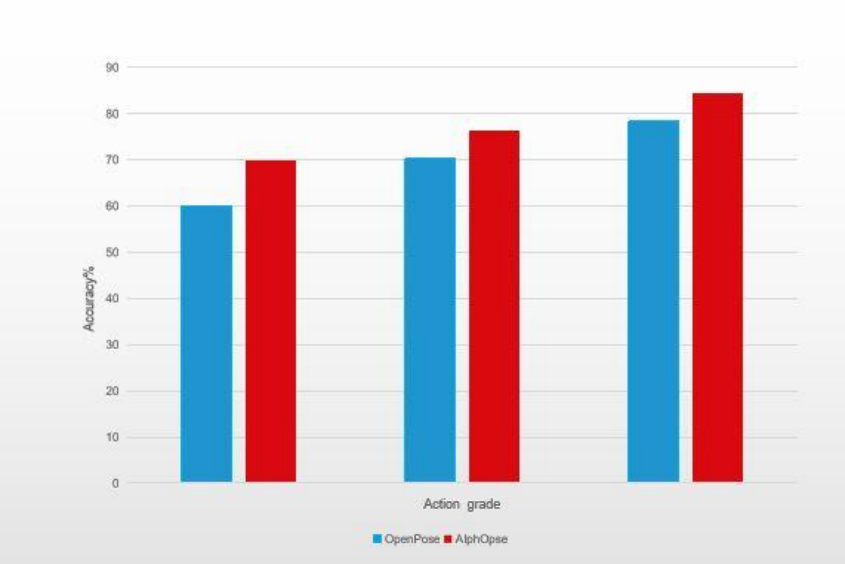


Figure 4.3: Experimental results of ST-GCN method

In this thesis, we are use of OpenPose algorithm to extract human skeletal sequences for human pose estimation, classify the behavioral actions by using ST-GCN, add the attention mechanism, and output real-time action categories so as to classify the whole video as shown in Figure 4.4.

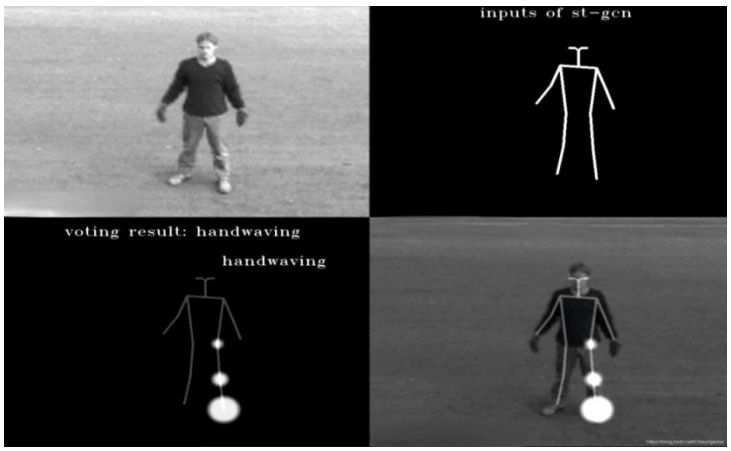


Figure 4.4: Human behavior identification results

The focus of ST-GCN is on designing network models for human behavior recognition based on skeletal information of human body. The models are trained and tested in real scenarios. The accuracy and robustness of the ST-GCN model are observed. OpenPose runs much fast in the experiments.

## 4.2 Early Behavior Recognition

Identifying behaviors or predicting behaviors as early as possible before they occur or occur is challenging in human behavior recognition. As shown in Table 4.3. the TRN model extracts top 25% and 50% video frames of each validation video for action recognition, respectively. Throughout the comparisons, we get that TRN is able to predict human behavior by using temporal relationships. The more ordered frames received, the higher the accuracy is. A qualitative evaluation of the examples shows that the model predictions made on the initial frames indeed was used as reasonable predictions.

Table 4.3: TRN predicted behaviors

| | FPS | TRN |
|---|---|---|
| TOP 25% | 22.36 | 77.64% |
| TOP 50% | 41.29 | 80.61% |
| ALL | 53.79 | 81.43% |

In this chapter, the dataset selected for the experiments of this thesis, the method of processing the data, the analysis of the data and the evaluation of the relevant models are presented. TRN takes use of temporal relation to detect recognition behavior. As more ordered frames are entered, the accuracy rate increases significantly.

## 4.3 TRN Results

Understanding the temporal and causal relationships of events is an important task in predicting behavior in behavioral literacy. There must be a reason for an event to be happened, so there is a need to find a connection between temporal and causal relationships. TRN uses temporal relationships to infer behavior in the input video, the network is highly accurate.

In this thesis, we employ a TRN model and a CNN algorithm to identify abnormal human behavior. The model represents an accurate point of introduction in biometric

surveillance to further tackle crime prediction. The TRN model annotates the spatial relationships between human bodies in the images, the positional ordering between the targets in the images is analyzed and the temporal relationships between human bodies in the images are inferred. The use of a recurrent network in combination with a TRN network is intended to address the problem of long-time dependence in the network, allows the model to better focus on the temporal relationships between images.

Temporal segmentation networks are similar to traditional two-stream neural networks and are trained by using CNN. TSN improves discrimination by exploring more input patterns. In addition to spatial flow convolutional networks that manipulate a single RGB image as in two-stream, temporal flow convolutional networks take continuous optical flow as input.

TSN encloses the segments of the given video into $k$ segments according to equal time divisions and takes a video segment at random in each video segment, models the time series of the segments in time by using the TSN network in Equation (3.12).

- $(T_1, T_2, \cdots, T_K)$ represents the sequence of segments, with $T_k$ all drawn randomly from the corresponding $S_k$ segments.
- $F$ $(T_k, W)$ function shows a convolutional network using $W$ as a parameter acting on a short segment $T_k$, the function returns the score of $T_k$ related to all classes.
- $G$ (The segmental consensus function) is the consensus function of the segment.
- The prediction function $H$ (•) takes use of a SoftMax function that incorporates a standard classifiable cross-entropy loss.

In human behavior prediction, the image frame information varies in speed on the time axis as the duration of the different actions varies, $N$-relation is calculated separately.

The fusion effect varies from frame to frame, with the larger scale, the more information is fused. After 5 frames, the effect does not change significantly. In this thesis, we compare three good network architectures: BN-Inception, GoogLeNet, and VGG-16. Among these architectures, BN-Inception performs the best, which is chosen as CNN architecture for TSN.

The network is normalized and initialized with BN-Inception. TRN network is

modeled and trained with the same parameters following BN's training strategy. In the TRN network, the number of triples in each relational network layer is $k$ that is set to 3. $g_\varphi$ is only two layers of MLP with 256 units each, $h_\varphi$ is a single layer of MLP with the number of units being the numbered value of the corresponding behavior. TRN takes 1,000 hours to complete 100 training epochs on a single GPU, with a multiscale relation, by experimentally comparing a 2-frame TRN to an 8-frame. In the multitime scale relationship, the network models with TRN of 2 to 8 frames show that TRN with more than 8 frames does not improve accuracy and take longer to train.

In order to demonstrate the effectiveness of time series analysis in behavior recognition, we conducted an experiment in which the input frames were temporally disordered during the training of the TRN. The frames in the relationship module were also randomly disrupted during the training of TRN. The experiments conducted on the UCF101 dataset observed the differences between ordered and disordered frames. The ordered temporal frames are needed in behavior recognition for temporal inference and thus lead to better recognition of human behaviors.

To further investigate how temporal order affects activity recognition in the TRN, as shown in Figure 4.5, there is a difference between extracting ordered and unordered inputs in the UCF101 dataset. The results tell us that inputting temporal frames in the TRN or disrupting frames in the relationship module may have effects on the results of human behavior recognition.

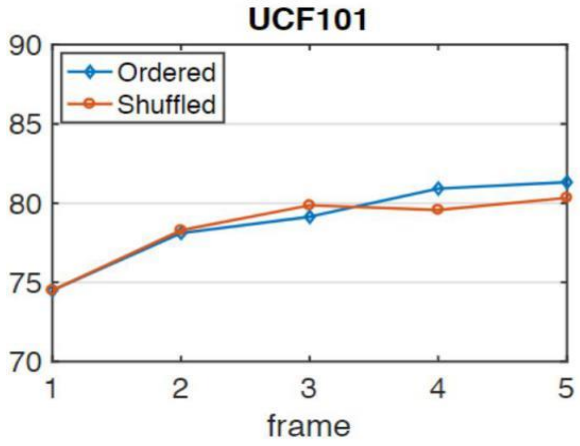


Figure 4.5: The accuracy obtained by using ordered and unordered video frames from UCF101 dataset.

TRNs are employed for human behavior recognition by using temporal relationships. With TRN, our attention needs to be paid to input ordered frames for human behavior recognition, but the input unordered frames may affect the results.

Table 4.4: The results of the validation and test sets

| | Train | Test |
|---|---|---|
| 3D CNN | 70.42% | 71.64% |
| TSN | 72.97% | 73.71% |
| TRN | 89.15% | 81.43% |

In Table 4.4, the results for the validation set and the data set are presented in the proposed networks. The validity of temporal relationships in behavioral recognition is demonstrated by comparing the TRN with TSN, which take use of feature information at average depth, where the model only extracts the same information from the features and does not reason temporal relationships as shown in Table 4.5.

Table 4.5: TRN *vs* TSN

| | TRN | TSN |
|---|---|---|
| 2 Frames | 61.83% | 67.49% |
| 3 Frames | 73.54% | 78.86% |
| 4 Frames | 86.15% | 84.61% |
| 5 Frames | 89.26% | 88.32% |
| 6 Frames | 92.51% | 90.46% |
| 7 Frames | 90.49% | 90.16% |

Table 4.5 shows TRN has a better performance than TSN, while Table 4.4 shows, the higher the frame rate is, the better TRN will recognize. According to the results, the TRN model had higher recognition accuracy than the TSN model, the level of accuracy of the TRN model was dependent on the number of frames. Thus, human behavior recognition depends on temporal relationships.

TRN networks outperform other network models significantly. TRN makes use of the pooling of temporal relationships to pool video frames, but TSN uses an averaging pooling technique for video frames. For temporal relationship inference, video frame

sequences play a key role in the experimental results. Compared to the network model without TRN temporal relations, the TRN network demonstrated significantly higher recognition accuracy. It was found that TRN improved with increasing frames, indicating the importance of temporal relationships, as well as the TRN-based technique outperformed that of dual-stream, demonstrating its utility based on temporal relationship inference.

## 4.4 C3DP-LA + ST-GCN Results

A 3D convolutional network performs a combination of convolution and pooling in the spatial-temporal dimension and is more suitable for learning spatial-temporal features of videos than a 2D convolutional network. The video loses the temporal information after each convolutional operation with 2D convolution. Video segments are convoluted in 3D to preserve the temporal information from the input video.

In C3DP-LA + ST-GCN, three LSTM-based algorithms are used: (1) Adding STPP to the 3D CNN, (2) integrating RGB video features with skeleton features, and (3) fusing RGB video features with the LSTM. In this chapter, we examine the effects of the first two modules on recognition and compare the final recognition model with approaches. In this thesis, we use the UCF101 dataset, each module is added separately, and the recognition performance is shown in Table 4.6.

Table 4.6: The accuracy of human behavior recognition

| | Accuracy |
|---|---|
| Two streams | 88.00% |
| 3D CNN | 82.00% |
| C3DP-LA+ST-GCN | 89.20% |
| ST-GCN | 86.32% |
| TRN | 83.30% |

On UCF101, the C3D network achieved 85.2% recognition rate. Even though this is low in comparison to dual-stream networks, C3D networks remain a hot research topic. It is primarily since C3D networks are faster than other methods and do not require the extraction of optical flow features, allowing for real-time processing.

STPP allows for resizing video input sizes with the improved 3D CNN, unlike the original 3D CNN. The module in this thesis is trained by using the UCF101 dataset with multisize videos, while 3D CNN is trained using fixed-size videos. The multiple sizes training of the improved 3D CNN model is shown to be better than that of the original 3D CNN. The increased recognition accuracy is promising because multiple sizes training can prevent the network from overfitting.

In human behavior recognition, the performance of LSTMs and spatial temporal attention mechanisms was improved by adding a CNN-connected 3D LSTM model, which improved the accuracy of the model by 4.5%. The spatial temporal attention can improve the expressive abilities of the model because it enhances key features of spatial temporal information and filters out more complicated information.

The module in our thesis is trained by using UCF101 dataset with multiresolution videos, while 3D CNN is trained by using fixed-size videos. The improved 3D CNN is shown to be better than that of the original 3D CNN. The increased recognition accuracy is due because the training can prevent the network from overfitting.

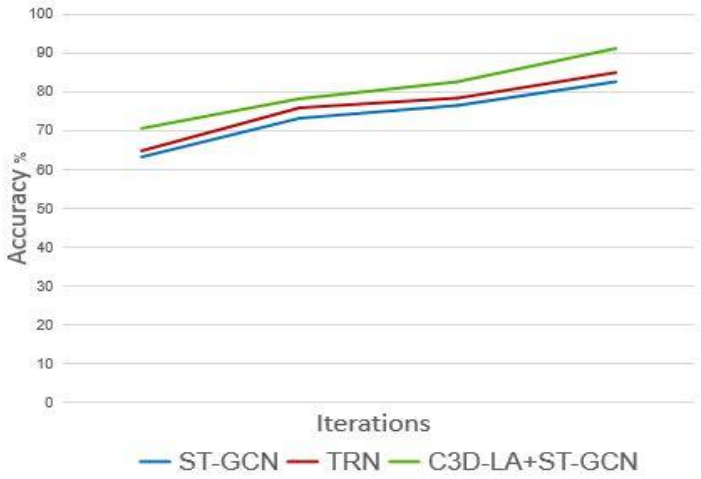


Figure 4.6: Results of different methods with iterations on UCF101 dataset.

Due to the experimental data, it takes a long time to train the proposed model. At the same time, the noises in the video data are processed, which affects the accuracy of human behavior recognition.

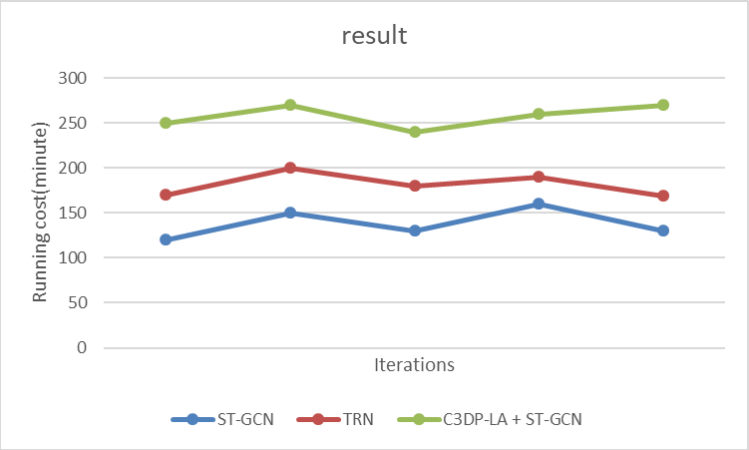


Figure 4.7: Running cost of all methods on UCF101 dataset.

The module in our thesis is trained by using UCF101 dataset with multiresolution videos, while the 3D CNN is trained by using fixed-size videos. The improved 3D CNN is shown to be better than that of the original 3D CNN. The increased recognition accuracy is owing to the training process that prevents the network from overfitting.


Figure 4.8: Comparison of accuracy of different methods

In video behavior recognition, the performance of LSTM and spatial temporal attention mechanisms was improved by adding a CNN-connected 3D LSTM model,

which improved the accuracy of the model by 4.5%. spatial temporal attention can improve the expressive abilities of the model because it enhances the key features of spatial temporal information and filters out more additional information.

# Chapter 5

# Discussion

Our experimental results have been given without much analysis and comparison. In this chapter, we will focus on analyzing the experimental results and compare the effectiveness and accuracy of the different models. Finally, we will explain the reasons for using this model and the contributions we have made in this thesis.

The existing research methods for the recognition of human abnormal behaviors in videos focus on spatial temporal features after the behaviors occurred. In contrast, in real scenarios, we need an early prediction of behaviors if the behavior occurs or when the behavior is not completed. The prediction of human behavior has a high value such as crime prediction.

Human behavior prediction requires early inference and judgment of human behaviors, unlike human behavior detection, none of the existing models for frontal behavior recognition is well suited to solve the problem of crime prediction. In terms of hardware, human abnormal behavior prediction has a high demand for real-time videos, which needs to quickly identify the abnormal behaviors in digital videos and make early predictions. The computational requirements for computing are high.

After overview the existing literature, the use of computer vision for video processing and effective video feature information has become one of the priorities for the recognition of abnormal human behaviors in surveillance videos. In intelligent surveillance for crime prediction, the identification of human behaviors is becoming increasingly important.

The purpose of criminal behavior recognition is to detect and identify criminal behaviors of human bodies in unknown videos. However, the variability and high complexity of human motion makes recognition tasks increasingly difficult. Considering the complexity of space and time, the skeleton feature information and optical flow information of human behavior are extracted from the video. In this thesis, based on a 2D graph neural network, a crime recognition method based on ST-GCN is proposed. The experimental results show that, compared with OpenPose, the recognition accuracy and efficiency of the method based on AlphaPose are greatly improved.

Due to the limitation of the computer configuration used in the experiment, the model training in the experiment takes a long time. Moreover, due to the large dataset used, it took nearly 6 months to extract the skeleton and optical flow feature information. Second, the classification of crimes and grades is subjective, and there is currently no uniform definition. Furthermore, the accuracy and efficiency drop due to the lack of

video and noise in the dataset. There is also a need to find better dataset preprocessing algorithms.

In this thesis, we discuss the results of behavior identification and prediction of criminal behaviors, evaluate the performance of ST-GCN model and TRN network based on our collected data. The prediction results are also compared with those of other models, the results show that the identification method for human behavior anomalies proposed in this thesis is much effective.

In this thesis, we took use of the ST-GCN model and TRN for the detection of human abnormal behavior. In the experimental results, we see that ST-GCN has a better performance for the recognition of human abnormal behaviors. The experiments with TRN using temporal relations had a higher performance, nearly 70% overall for the recognition of human abnormal behaviors. Conversely, in TSN, because no temporal relationships were used, the performance of the records was relatively low.

According to the literature review in Chapter 2, there are a number of reasons for human behaviors. Video surveillance met challenges if luminance changes, for instance. For this reason, ST-GCN are used (Shahbaz, Hoang, & Jo, 2019), CNNs are good enough for video surveillance. ST-GCN and TRN show the excellence, importance, and popularity in tackling various challenges. We extract visual features from the input according to our methodology. In this way, the human abnormal behaviors are recognized and applied to develop patterns for crime monitoring without changes of the data itself.

At present, a vast majority of algorithms for the recognition of abnormal human behavior are based on static images or the detection and recognition of abnormal behavior of human in videos. These algorithms are not very efficient and only identify abnormal behaviors of crowds, but not specific behaviors, the main idea is to classify human behaviors. This has failed to meet the current needs, the accuracy and efficiency of the recognition are not high, and the prediction of human criminal behavior is still a challenge. In this thesis, two crime identification methods are effective in detecting and identifying human criminal behaviors that identify multiple crimes without the limitations of public places.

Deep learning is usually performed by combining features at various scales to improve performance. The features at bottom layer have higher resolution and more positional details, but the semantic information is less detailed due to the less convolutions applied. The higher-level features, have stronger semantic information but have low resolution, and poor details. The key to improve the model is to effectively fuse the two parts. Depending on the order of fusion and prediction, they are grouped into early and late fusions.

Firstly, the features from multiple layers are fused, the predictive machine is trained by using the fused features. The skip connection approach also takes use of *concat* and *add* operations. These operations are represented by using Inside-Outside Network (ION) and HyperNet.

The detection performance is improved by combining the results of different layers, starting the detection on part of the fused data before the final fusion is completed, there will be multiple layers of detection, eventually fusing the results of several detections.

In this thesis, we propose a deep learning-based method for the recognition of human behaviors, which effectively detects abnormal behaviors and is no longer limited to scenarios for the recognition of multiple behaviors. The recognition method for human behaviors by using ST-GCN and interpretable TRN.

In the recognition of human behaviors due to the randomness and complexity of human behavior, it also makes the recognition of human criminal behavior with variability. For the complexity of human behavior in space, in this thesis, we propose a method based on human skeleton extraction, through the extraction of skeletal features, the identification of human criminal behavior in space and time based on ST-GCN is proposed.

In this thesis, we compare the top-down AlphaPose pose evaluation method with the bottom-up OpenPose pose evaluation method by evaluating the human pose evaluation algorithm. It is shown experimentally that the accuracy and effectiveness of the method b using AlphaPose are better than that of OpenPose. AlphaPose takes longer time to compute than OpenPose. In this thesis, we harnessed the relatively fast OpenPose method for pose evaluation.

In this thesis, ST-GCN is applied to recognize human behavior in multiple scenarios by testing it based on digital videos and standard datasets. The algorithm is robust on human behavior recognition. By evaluating the model, the algorithm meets the needs.

A TRN-based relational network was employed in combination with the optical flow information of the skeleton while ensuring the accuracy of the crime recognition. Firstly, the tested data were preprocessed, after training the model, the classification of human behaviors was performed by using SoftMax classification. Also, the experimental results were described in the previous chapter, the TRN-based approach improved the accuracy of human behavior identification due to the average pooling provided by the TRN.

Crime is on the rise; this project aims to provide the police with an advantage by using intelligent surveillance to reduce crimes. Despite the police having tried everything they can, the problem cannot be solved by the police alone. There are also challenges associated with visual surveillance that traditional surveillance has. It is challenging for surveillance systems to accurately predict crimes because of these challenges. Thus, we used the ST-GCN model and TRN model to detect abnormal behaviors, thereby improve the accuracy of human abnormal behavior prediction. Based on this study, we see that incorporating ST-GCN and TRN model can improve the performance and enhance the predictive ability of our surveillance systems.

To address the problem that single features cannot adequately represent complicated human actions, resulting in low accuracy of behavioral recognition, a human behavior recognition algorithm based on multimodal feature learning extract RGB features and skeletal features of the video separately, we fuse them to take advantage of the complementarity of the visual features to improve the recognition rate. This method has been demonstrated to have higher recognition accuracy than many other algorithms and recognize human actions more effectively.

# Chapter 6

# Conclusion and Future Work

We summarize the shortcomings and limitations of our methods in this chapter, the work needs to be improved further in the future.

Human behavior recognition has long been an important research topic in the field of computer vision and pattern recognition. One of the most popular problems is the recognition of human behavior in video. With the rapid development of deep learning in the field of artificial intelligence in recent years theory has developed rapidly in recent years, providing more new solutions to this problem. The core of this thesis is how to apply deep learning theory and related techniques to the task of recognizing human behavior in video in a more effective way.

In this thesis, we focus on deep learning-based video human behavior recognition and achieve some research results: to address the problems of overfitting and slow convergence when using dual-stream convolutional neural networks for human behavior on a limited set of human behavior video samples, we design a temporal and spatial feature fusion method is designed to address the problems of overfitting and slow convergence of training on a limited set of human behavior video samples. Although the network model and feature fusion method designed in this thesis can effectively perform the human behavior recognition task and significantly improve the recognition ability of the network, there are still some shortcomings to be remedied and improved.

From the literature that the prediction of human criminal behavior in the ST-GCN, TRN models play a pivotal role in security prevention and control, as well as provides aids in the allocation of police resources and other issues, but there are a few shortcomings in the literature.

In the current behavior recognition methods, the prediction of human behavior is basically attributed to the problem of human behavior classification. Correctly distinguishing between behavioral classification and behavioral prediction is a problem that we need to challenge.

Similarly, it is much difficult to accurately segment the behavior of human mentions in a video for prediction. A single human behavior may be accompanied by many sub-behaviors, the various sub-behaviors constitute the randomness of the human behavior, these affect the accuracy of models in classifying the behaviors. In realistic scenarios, the duration of the video footages is unknown, how to ensure the long-term

memorability of models becomes a principal issue. For complex behaviors with long time spans, the predictive behavior is much urgent for the semantic level. Relational inference networks will be used to predict human intentions. Reasoning what the next behavior of a human will be.

Due to the data and computer configuration of our experiments, the time spent in model training was long. At the same time, the extraction of skeleton and optical flow-based information took longer time because there was insufficient data, which had an impact on the results of the detection. At the same time, the noises in the videos are not handled well, result in less precise accuracy of object recognition. The classification of human crime in this thesis is subjective and has not standard definition, in specific scenarios, normal behaviors may also be classified as an abnormal.

In this thesis, we propose a method for classifying criminal behavior based on skeletal sequences, but skeletal sequences are difficult to obtain in practice and the data is in the form of surveillance videos. The obtained skeletal sequences will become even much tough in our cases. Therefore, the next step of this project is to improve the accuracy and computational speed of the OpenPose skeletal sequences.

Current human behavior recognition is based on datasets with single and two-person behavior, whereas in real-life events, the models need to be detected with multiple people or more complex crowds. Therefore, more complex human behavior datasets are needed for our experiments, the proposed model needs to be trained and extended to allow for the accurate recognition soon.

On feature fusion, the method proposed in this thesis focuses on video score features and performs video behavioral feature fusion based on the convolutional layer in a convolutional neural network, which can theoretically yield deep feature information.

The scenarios such as video surveillance of special people are an area that needs to be studied in depth at a later stage. Our experiments show that the proposed method performed well in detecting and recognizing dangerous human behaviors in videos, but the running speed still needs to be improved. Since the complexity of the video collected in the actual situation is much larger than the set used in this thesis, and the

types of criminal behaviors are not limited to the types summarized in this thesis, the algorithm proposed in this thesis has a few limitations.

The current recognition lacks reasoning ability. One of the goals of artificial intelligence is to develop a machine having logical reasoning capabilities like our human beings. Therefore, it is necessary to have a deep understanding of learning and reasoning in deep learning.

# References


Al-Sarayreh, M., Reis, M., Yan, W., Klette, R. (2019) A sequential CNN approach for foreign object detection in hyperspectral images. Computer Analysis of Images and Patterns.

An, N., Yan, W. (2021) Multitarget tracking using Siamese neural networks. ACM Transactions on Multimedia Computing, Communications and Applications.

Aggarwal, J., Ryoo M. (2011) Human activity analysis: A review. ACM Computing Surveys, 43(3): 16.

Aggarwal, J., Xia, K. (2014) Human activity recognition from 3D data: A review. Pattern Recognition Letters, 48, 70-80.

Aggarwal, J., Xia L., Chen, C. (2012). View invariant human action recognition using histograms of 3D joint. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 20-27.

Agrawal, P., Nair, A.V., Abbeel, P., Malik, J., Levine, S. (2016). Learning to poke by poking: Experiential learning of intuitive physics. Advances in Neural Information Processing Systems, pp. 5074–5082.

Ahmed, J., Rodriguez, M., Shah, M. (2008). Action matches a spatial temporal maximum average correlation height filter for action recognition. IEEE CVPR.

Alice, G., Lai, G. (2014). A survey on still image based human action recognition. Pattern Recognition, 47(10), 3343-3361.

Altamimi, A., Ullah, H., Uzair, M., et al. (2018). Anomalous entities detection and localization in pedestrian flows. Neurocomputing, 290, pp.74-86.

Andonian, A., Zhou, B., Oliva, A., et al. (2017). Temporal relational reasoning in videos. IEEE CVPR.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B. (2014). 2D human pose estimation: new benchmark and state of the art analysis. IEEE CVPR.

Anguelov, D., Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Erhan, D., Vanhoucke, V., Rabinovich, As. (2015). Going deeper with convolutions. IEEE CVPR.

Azizpour, H., Razavian, A., Sullivan, J., Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. IEEE CVPR Workshops.

Badii, C., Bellini, P., Ddfino, A., et al. (2019). Smart city IoT platform respecting GDPR privacy and security aspects. IEEE Access.

Bakshi, S. Guo, G. Proença, H. & Tistarelli, M. (2020) Visual surveillance, biometrics: Practices, challenges, and possibilities. IEEE Access.

Baradel, F., Wolf, C., Mille, J. (2017). Pose-conditioned spatial temporal attention for human action recognition. CoRR, abs/1703.10106.

Baradel F, Wolf C, Mille J. (2017). Human action recognition: Pose-based attention draws focus to hands. IEEE International Conference on Computer Vision Workshops (ICCVW), pp.604–613.

Begleiter, R., El-yaniv, R., Yona., G. (2004). On prediction using variable order Markov models. Journal of Artificial Intelligence Research, vol. 22, pp. 385-421.

Bengio, Y., Simard, P., Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166.

Bengio, Y., Ducharme, R., Vincent, P., et al. (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3(2): 1137-1155.

Bengio, Y., Glorot, X. (2010). Understanding the difficulty of training deep feedforward neural networks. AISTATS.

Billings, D., Yang, J. (2006). Application of the ARIMA models to urban roadway travel time prediction-A case study, IEEE SMC, pp. 2529–2534.

Bischof, H., Zach, C., Pock, T. (2007). A duality-based approach for realtime TV-L1 optical flow. Joint Pattern Recognition Symposium, pp.214-223.

Blank, M., Gorelick, L., Shechtman, E., et al. (2005). Actions as space-time shapes. IEEE International Conference on Computer Vision (ICCV'05), pp. 1395-1402.

Bobick, A., Davis, J. (2001). The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis & Machine Intelligence, 23(3):257-267.

Boser, B., LeCun, Y., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.

(1989). Backpropagation applied to handwritten zip code recognition. Neural Computation.

Bouchrika, I. (2018). A survey of using biometrics for smart visual surveillance: Gait recognition. Surveillance in Action (pp. 3-23). Springer, Cham.

Bourdev, L., Tran, D., Fergus, R., et al. (2015). Learning spatial temporal features with 3D convolutional networks. IEEE International Conference on Computer Vision, pp. 4489-4497.

Boureau, Y., Bach, F., Lecun, Y., et al. (2010). Learning mid-level features for recognition. IEEE CVPR, pp.2559-2566.

Boulay, B., Francois, B., Thonnat, M. (2006). Applying 3D human model in a posture recognition system. Pattern Recognition Letters, 27(15):1788-1796.

Brantingham, P. J., Wang, B., Yin, P., Bertozzi, A. L., Osher, S. J., & Xin, J. (2019). Deep learning for real-time crime forecasting and its internalization. Chinese Annals of Mathematics, Series B, 40(6), 949-966.

Brown, P., Desouza, P., Mercer, L., et al. (1992). Class-based *n*-gram models of natural language. Computational Linguistics, 18(4): 467-479.

Buades, A., Coll, B., Morel, J. (2005). A non-local algorithm for image denoising. IEEE CVPR.

Caputo, B., Laptev, I., Schuldt, C. (2004). Recognizing human actions: A local SVM approach. International Conference on Pattern Recognition, pp. 32-36.

Carreira, J., Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. arXiv:1705.07750.

Chaib-Draa, B., Giguere, P., Trottier, L. (2016). Parametric exponential linear unit for deep convolutional neural networks, arXiv, pp. 1–16.

Chen, J., Weber, M., Suzumura, T., et al. (2018). Scalable graph learning for anti-money laundering: A first look. Arxiv:1812.00076.

Chen, H., Chen, H., Chen, Y., et al. (2006). Human action recognition using star skeleton. ACM International Workshop on Video Surveillance & Sensor Networks, 171.

Chen, H., Wang, G., Xue, J., et al. (2006). A novel hierarchical framework for human

action recognition. Pattern Recognition, 148-159.

Chen, S., Wang, X., Tang, Y., et al. (2017). Aggregating frame-level features for large-scale video classification. arXiv:1707.00803.

Chen, Z., Shi, X., Wang, H., et al. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. International Conference on Neural Information Processing Systems, pp. 802–810.

Cheng, E., Duan, L., Hu, T., Zhu, J., & Gao, C. (2017). Deep convolutional neural networks for spatial temporal crime prediction. International Conference on Information and Knowledge Engineering (IKE), pp. 61-67.

Choi, J., Ng, Y., Neumann, J., et al. (2016). ActionFlowNet: Learning motion representation for action recognition.

Cleary, J., Witten, I. (1984). Data compression using adaptive coding and partial string matching. IEEE Trans. on Inform. Theory, vol. 24, no. 4, pp. 413–421.

Collins, R., Lipton, A., Kanade, T., et al. (2000). A System for Video Surveillance and Monitoring. Carnegie Mellon University.

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. International Journal of Digital Crime and Forensics (IJDCF) 8 (1), 26-36.

Cui, Z., Chen, W., Chen, Y. (2016). Multi-scale convolutional neural networks for time series classification, arXiv:1603.06995.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control Signals and Systems, 2(4):303-314.

David G. Lowe.(2010). Computer Science Department，University of British Columbia .

Deng, J., Dong, W., Socher, R., et al. (2009). ImageNet: A large-scale hierarchical image database. IEEE CVPR, pp. 248-255.

Donahue, J., Hendricks, L., Guadarrama, S., et al. (2015). Long-term recurrent convolutional networks for visual recognition and description. IEEE CVPR, 2625-2634.

Efros, Berg, Mori, et al. (2003). Recognizing action at a distance. IEEE International Conference on Computer Vision, pp.726-733.

Evans, M., Osborne, C., Ferryman, J., et al. (2013). Multicamera object detection and tracking with object size estimation. Advanced Video and Signal based Surveillance, 177-182.

Feichtenhofer, C., Pinz, A., Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. IEEE CVPR, pp. 1933-1941

Fernando, B., Gavves, E., et al. (2015). Modeling video evolution for action recognition. IEEE CVPR, pp. 5378–5387.

Fournier-Viger, P., Gueniche, T., Tseng, V.S. (2012). Using partially ordered sequential rules to generate more accurate sequence prediction. Intern. Conf. Advanced Data Mining and Applications, Springer LNAI 7713, pp. 431-442.

Fraley, C. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal, 41(8):578-588.

Gao, X., Nguyen, M., Yan, W. (2021) Face image inpainting based on generative adversarial network. International Conference on Image and Vision Computing New Zealand.

Gopalratnam, K., Cook, D. (2007) Online sequential prediction via incremental parsing: The active LEZI algorithm. IEEE Intelligent Systems, 22(1): 52-58.

Gorr, W., & Harries, R. (2003). Introduction to crime forecasting. International Journal of Forecasting, 19(4), 551-555.

Gowdra, N., Sinha, R., MacDonell, S., Yan, W. (2021) Maximum categorical cross entropy (MCCE): A noise-robust alternative loss function to mitigate racial bias in convolutional neural networks (CNNs) by reducing overfitting. Pattern Recognition.

Gu, D., Nguyen, M., Yan, W. (2016) Cross models for twin recognition. International Journal of Digital Crime and Forensics 8 (4), 26-36

Gu, Q., Yang, J., Yan, W., Klette, R. (2017) Embedded and real-time vehicle detection system for challenging on-road scenes. Optical Engineering 56 (6).

Gu, Q., Yang, J., Yan, W., Klette, R. (2017) Integrated multi-scale event verification in an augmented foreground motion space. Pacific-Rim Symposium on Image and Video Technology, 488-500.

Gu, Q., Yang, J., Yan, W., Li, Y., Klette, R. (2017) Local Fast R-CNN flow for object-centric event recognition in complex traffic scenes. Pacific-Rim Symposium on Image and Video Technology, 439-452.

Gueniche, T., Fournier-Viger, P., Tseng, V.-S. (2013). Compact prediction tree: A lossless model for accurate sequence prediction. Intern. Conf. Advanced Data Mining and Application, Springer LNAI 8347, pp. 177–188.

Guha, S., Rastogi, R., Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. Information Systems Volume 25(5), 345-366.

Haghani, A., Zhang, Y. (2015). A gradient boosting method to improve travel time prediction. Transportation Research Part C-Emerging Technologies, vol. 58, pp. 308–324.

Halkidi, M., Batistakis, Y., Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems: Integrating Artificial, Intelligence and Database Technologies, 17(2-3):107-145.

Han, J., Pei, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M. (2004). Mining sequential patterns by pattern-growth: The prefix span approach. IEEE Trans. Known. Data Engin. 16(11): 1424–1440.

Han, X., Hu, X., Wu, H., et al. (2020). Risk prediction of theft crimes in urban communities: An integrated model of LSTM and ST-GCN. IEEE Access, 8: 217222-217230.

Han, Y., Pinto, L., Gandhi, D., Park, Y.L., Gupta, A. (2016). The curious robot: Learning visual representations via physical interactions. European Conference on Computer Vision, Springer, pp. 3–18.

Harchaoui, Z., Gaidon, A., Schmid, C. (2013). Temporal localization of actions with ac-toms. IEEE Transactions on Pattern Analysis & Machine Intelligence, pp. 2782-2795.

Harchaoui, Z., Gaidon, A., Schmid, C. (2014). Activity representation with motion hier-archies. International Journal of Computer Vision, 107(3): 219–238.

Hasan, M., Nakib, M., Khan, R. T., & Uddin, J. (2018). Crime scene prediction by detecting threatening objects using convolutional neural network. IEEE

International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), pp. 1-4.

Hausknecht, M., Ng, J., Vijayanarasimhan, S., et al. (2015). Beyond short snippets: Deep networks for video classification. IEEE CVPR, pp. 4694-4702.

Haykin, S., Kosko, B. (2009). Gradient based learning applied to document recognition. Wiley-IEEE Press.

He, K., Zhang, X., Ren, S., Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. IEEE ICCV.

He, K, Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In IEEE CVPR.

Hebert M, Sukthankar R, Ke Y. (2007). Spatial temporal shape and flow correlation for action recognition. IEEE Conference on Computer Vision and Pattern Recognition.

Horn, B., Schunck, B. (1981). Determining optical flow. Artificial Intelligence, 17(1-3):185-203.

Huang, G., Liu, Z., Der Maaten, L., et al. (2017). Densely connected convolutional networks. IEEE Computer Vision and Pattern Recognition, pp. 2261-2269.

Ji, H., Liu, Z., Yan, W. Yan, Klette, R., (2019) Early diagnosis of Alzheimer's disease based on selective kernel network with spatial attention. Asian Conference on Pattern Recognition 2 (1), 503-515.

Ji, H., Yan, W., Klette, R. (2019) Early diagnosis of Alzheimer's disease using deep learning. International Conference on Control and Computer Vision.

Ji, S., Xu, W., Yang, M., et al. (2013). 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 221-231.

Kang, H. W., & Kang, H. B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. PloS One, 12(4): e0176244.

Karpathy, A., Toderici, G., Shetty, S., et al. (2014). Large-scale video classification with convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725-1732.

Ke, Y., Sukthankar, R., Hebert, M., et al., (2007). Spatial temporal shape and flow correlation for action recognition. IEEE Conference on Computer Vision and Pattern Recognition.

Khan, A., Javed, K., Saba, T. (2020). Human action recognition using fusion of multiview and deep features: An application to video surveillance. Multimedia Tools and Applications.

Kieran, D., Yan, W. (2010) A framework for an event-driven video surveillance system. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS).

Kipf, T., Welling, M. (2016). Semi-supervised classification with graph convolutional networks. Arxiv:1609.02907.

Klaser, A., Marszalek, M., Schmid, C., et al. (2008). A spatial temporal descriptor based on3D-gradients. British Machine Vision Conference.

Krizhevsky, A., Sutskever, I., Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, pp.1097-1105.

Kuehne, H., Jhuang, H., Garrote, E., et al. (2011). HMDB: A large video database for human motion recognition. IEEE International Conference on Computer Vision.

Kuo, C. (2016). Understanding convolutional neural networks with a mathematical model. Journal of Visual Communication and Image Representation, (41):406-413.

Lan, C., Zhu, W., Xing, J., et al. (2017). Co-occurrence feature learning for skeleton-based action recognition using regularized deep LSTM networks. AAAI Conference on Artificial Intelligence.

Laptev, I. (2005). On space-time interest points. International Journal of Computer Vision, 64(2):107-123.

Le, R., Nguyen, M., Yan, W. (2021) Augmented reality and machine learning incorporation using YOLOv3 and ARKit. Applied Sciences

Le, R., Nguyen, M., Yan, W. (2021) Training a convolutional neural network for transportation sign detection using synthetic dataset. International Conference on Image and Vision Computing New Zealand.

Lea, C., Vidal, R., Reiter, A., Hager, C. (2016). Temporal convolutional networks: A unified approach to action segmentation, ECCV Workshop, pp. 47–54.

Lee, M., Lee, S., Son, S., et al. (2018) Motion feature network: Fixed motion filter for action recognition, ECCV.

Lefèvre, S., Vasquez, D., Laugier, C., (2014). A survey on motion prediction and risk assessment for intelligent vehicles, ROBOMECH Journal, 1(1): 1.

Li, C., Xu, K., Tian, Y., et al. (2018). Representation learning on graphs with jumping knowledge networks. arXiv:1806.03536.

Li, C., Zhong, Q., Xie, D. (2018). Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. arXiv:1804.06055.

Li, F., Gan, C., Liu, X., et al. (2017) Temporal modeling approaches for large-scale YouTube-8M video understanding. arXiv:1707.04555

Li, F. and Yao, B. (2012). Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),1691-1703.

Li, L., Kong, Y., Zhang, K., et al. (2019). Attention module-based spatial temporal graph convolutional networks for skeleton-based action recognition. Journal of Electronic Imaging, 28(4): 43032.

Li, L., Su, X., Zhang, Y., Lin, Y., Li, Z. (2015). Trend modeling for traffic time series analysis: An integrated study. IEEE Transactions on Intelligent Transportation Systems, 16, (6): 3430–3439.

Li, F., Zhang, Y., Yan, W., Klette, R. (2016) Adaptive and compressive target tracking based on feature point matching. International Conference on Pattern Recognition (ICPR), 2734-2739

Li, R. (2013). Space, disorder, and crime: A summary of western environmental criminology research. Journal of Xinjiang University of Finance and Economics (3), 43-48.

Li, S., Zhang, X., Sha, L., et al. (2017). Attentive interactive neural networks for answer selection in community question answering. AAAI Conference on Artificial

Intelligence.

Li S. (2011) Biometrics in Video Surveillance. In: van Tilborg H.C.A., Jajodia S. (eds) Encyclopedia of Cryptography and Security. Springer, Boston, MA.

Li, Y., Geng, X., Wang, L., et al. (2019). Spatial temporal multi-graph convolution network for ride-hailing demand forecasting. AAAI Conference on Artificial Intelligence, pp. 3656-3663.

Li, R., Nguyen, M., Yan, W. (2017) Morse codes enter using finger gesture recognition. International Conference on Digital Image Computing: Techniques and Applications.

Liang, S., Yan, W. (2022) Multilingual speech recognition based on the end-to-end framework. Multimedia Tools and Applications.

Lin, J., Keogh, E., Wei, L., (2007). Lonardi, Experiencing SAX: A novel symbolic representation of time series. Data Mining and Knowledge Discovery, 15(2): 107–144.

Liu, C., Yan, W. (2020) Gait recognition using deep learning. Handbook of Research on Multimedia Cyber Security, pp.214-226.

Liu, J., Shahroudy, A., Xu, D., et al. (2016). Spatial temporal LSTM with trust gates for 3D human action recognition. European Conference on Computer Vision (ECCV), pp. 816–833.

Liu, M., Yuan, J. (2018) Recognizing human actions as the evolution of pose estimation maps. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.1159–1168.

Liu, X., Nguyen, M., Yan, W. (2019) Vehicle-related scene understanding using deep learning. Asian Conference on Pattern Recognition.

Liu, X., & Sun, S. (2018). Research on abnormal behavior detection method based on YOLO network model. Electronic Design Engineering.

Liu, X., Yan, W., Kasabov, N. (2020) Vehicle-related scene segmentation using CapsNets. International Conference on Image and Vision Computing New Zealand.

Liu, X., Yan, W. (2021) Traffic-light sign recognition using Capsule network.

Springer Multimedia Tools and Applications.

Liu, Z., Yan, W., Yang, B. (2018) Image denoising based on a CNN model. International Conference on Control, Automation and Robotics.

Lu, J., Nguyen, M., Yan, W. (2020) Comparative evaluations of human behavior recognition using deep learning. Handbook of Research on Multimedia Cyber Security, pp.176-189.

Lu, J., Nguyen, M., Yan, W. (2020) Human behavior recognition using deep learning. International Conference on Image and Vision Computing New Zealand.

Lu, J., Nguyen, M., Yan, W. (2021) Sign language recognition from digital videos using deep learning methods. International Symposium on Geometry and Vision.

Luo, Z., Nguyen, M., Yan, W. (2021) Sailboat detection based on automated search attention mechanism and deep learning models. International Conference on Image and Vision Computing New Zealand.

Lu, J. Shen, J., Yan, W., Boris, B. (2017) An empirical study for human behaviour analysis. International Journal of Digital Crime and Forensics 9 (3), 11-17.

Lu, J. Yan, W., Nguyen, M. (2018) Human behavior recognition using deep learning. IEEE International Conference on Advanced Video and Signal Based Surveillance.

Ma, X., Yan, W. (2021) Banknote serial number recognition using deep learning. Multimedia Tools and Applications.

Mahajan, D., Girshick, R., Ramanathan, V., et al. (2018). Exploring the limits of weakly supervised pretraining. European Conference on Computer Vision.

Mahdisoltani, F., Berger, G., Gharbieh, W., Fleet, D., Memisevic, R. (2018).: Fine-grainedvideo classification and captioning. arXiv:1804.09235.

Marszalek, M., Laptev, I., Schmid, C., et al. (2009). Actions in context. IEEE Conference on Computer Vision & Pattern Recognition, pp.2929-2936.

Mehtab, S., Yan, W. (2021) FlexiNet: Fast and accurate vehicle detection for autonomous vehicles-2D vehicle detection using deep neural network. International Conference on Control and Computer Vision.

Mehtab, S., Yan, W. (2022) 3D vehicle detection using cheap LiDAR and camera

sensors. International Conference on Image and Vision Computing New Zealand.

Mehtab, S., Yan, W. (2022) Flexible neural network for fast and accurate road scene perception. Multimedia Tools and Applications.

Meng, L., Zhao, B., Chang, B., et al. (2018). Interpretable spatial temporal attention for video action recognition. arXiv: 1810.04511.

Meng, S., Wang, T., Liu, L. (2010). Monitoring continuous state violation in datacenters: Exploring the time dimension. International Conference on Data Engineering (ICDE 2010), pp. 968 – 979.

Milone, M. (2001). Biometric surveillance: Searching for identity. Bus. Law., 57: 497.

Monfort, M., Andonian, A., Zhou, B., et al. (2019) Moments in time dataset: One million videos for event understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Ng, J., Hausknecht, M., Vijayanarasimhan, S., et al. (2015). Beyond short snippets: Deep networks for video classification. IEEE CVPR, pp. 4694-4702.

Niebles, J., Wang, H., Li, F., et al. (2008). Unsupervised learning of human action categories using spatial temporal words. International Journal of Computer Vision, 79(3):299-318.

Padmanabhan, V., Mogul, J. (1998). Using prefetching to improve World Wide Web latency. Computer Communications, 16: 358–368.

Pan, C., Liu, J., Yan, W., Zhou, Y. (2021) Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing.

Pan, C., Yan, W. (2018) A learning-based positive feedback in salient object detection. International Conference on Image and Vision Computing New Zealand.

Pan, C., Yan, W. (2020) Object detection based on saturation of visual perception. Multimedia Tools and Applications, 79 (27-28), 19925-19944.

Pantic, Maja Rothkrantz, Jm L. (2003). Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE, 91(9): 1370-1390.

Parisi, G. (2020). Human action recognition and assessment via deep neural network

self-organization. ArXiv, abs/2001.05837.

Paul, I., E., & Krishna, M., (2017). Human behavioral analysis using evolutionary algorithms and deep learning. Hybrid Intelligence for Image Analysis and Understanding, pp.165-186.

Piccardi, M. (2004). Background subtraction techniques: A review. IEEE International Conference on Systems, pp. 3099-3104.

Pirsiavash, H., Ramanan, D. (2014). Parsing videos of actions with segmental grammars. IEEE Conference on Computer Vision and Pattern Recognition, pp. 612–619.

Qin, Z., Yan, W. (2021) Traffic-sign recognition using deep learning. International Symposium on Geometry and Vision.

Remagnino, P., Tan, T., Baker, K. (1998). Multi-agent visual surveillance of dynamic scenes. Image and Vision Computing, 16(8):529-532.

Ren, Y., Nguyen, M., Yan, W. (2018) Real-time recognition of series seven New Zealand banknotes. International Journal of Digital Crime and Forensics (IJDCF) 10 (3), 50-66.

Rudin, C. (2013). Predictive policing using machine learning to detect patterns of crime. Wired Magazine.

Sabour, S., Frosst, N., Hinton, G. et al. (2017). Dynamic routing between capsules. Neural Information Processing Systems, pp. 3856-3866.

Samuel, J., Manogaran, G., et al. (2019). Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM. Computer Networks. 151: 191-200.

Santoro, A., Raposo, D., Barrett, D., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T. (2017). A simple neural network module for relational reasoning. Neural Information Processing Systems (NIPS).

Schafer, P., Leser, U. (2017). Fast and accurate time series classification with WEASEL, arXiv:1701.07681.

Schafer, P. (2014). The BOSS is concerned with time series classification in the presence of noise. Data Mining and Knowledge Discovery, 29(6): 1505–1530.

Schmid, C., Jégou, H., Douze, M., et al. (2010). Aggregating local descriptors into a compact image representation. IEEE CVPR, pp. 3304-3311.

Sebe, A., Nicu, J. (2007). Multimodal human–computer interaction: A survey. Computer Vision Image Understanding, 108(1-2): 116-134.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y. (2014). OverFeat: Integrated recognition, localization and detection using convolutional networks. ICLR.

Sermanet, P., Lynch, C., Hsu, J., Levine, S. (2017). Time-contrastive networks: Self-supervised learning from multi-view observation. arXiv:1704.06888.

Shah, M., Reddy, K. (2013). Recognizing 50 human action categories of web videos. Machine Vision and Applications, 24(5): 971-981.

Shahbaz, A., Hoang, T., & Jo, H. (2019). Convolutional neural network-based foreground segmentation for video surveillance systems. IEEE IECON, pp. 86-89.

Shahroudy, A., Ng, T., Yang, Q., et al. (2015). Multimodal multipart learning for action recognition in depth videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 2123- 2129.

Shahroudy, A., Liu, J., Ng, T., et al. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1010–1019.

Shen, D., Xin, C., Nguyen, M., Yan, W. (2018) Flame detection using deep learning. International Conference on Control, Automation and Robotics (ICCAR).

Shen, Y., Yan, W. (2018) Blind spot monitoring using deep learning. International Conference on Image and Vision Computing New Zealand (IVCNZ)

Shetty, S., Karpathy, A., Toderici, G., Leung, T., Sukthankar, R., Fei-Fei, L. (2014). Largescale video classification with convolutional neural networks. IEEE CVPR.

Shun, L., Xing, G., JianGuo, W. (2019). Human action recognition method based on key frame and skeleton information. Transducer and Micro system Technologies, pp.26 – 30.

Sigurdsson, A., Divvala, S., Farhadi, A., Gupta, A. (2017). Asynchronous temporal fields for action recognition. IEEE CVPR.

Simonyan, K., Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, pp.568–576

Simonyan, K., Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. ICLR.

Song, C., He, L., Yan, W., Nand, P. (2019) An improved selective facial extraction model for age estimation. International Conference on Image and Vision Computing New Zealand (IVCNZ)

Song, S., Lan, C., Xing, J., et al. (2017). An end-to-end spatial-temporal attention model for human action recognition from skeleton data. AAAI Conference on Artificial Intelligence, pp. 4263–4270.

Soomro, K., Zamir, A., Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. Computer Science, 1212.0402.

Savarese, S., Delpozo, A., Niebles, J. et al. (2008). Spatial-temporal correlations for unsupervised action classification. IEEE Workshop on Motion and Video Computing.

Stec, A., & Klabjan, D. (2018). Forecasting crime with deep learning. arXiv:1806.01486.

Su, H., Russakovsky, O., Deng, J., et al. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252.

Sun, L., Jia, K., Yeung, D., et al. (2015). Human action recognition using factorized spatial-temporal convolutional networks. IEEE International Conference on Computer Vision, pp. 4597-4605.

Sutskever, I., Krizhevsky, A., Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. NIPS.

Szegedy, C, Ioffe S. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning, pp. 448-456.

Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going deeper with convolutions.

Tan, M., Le, Q. (2019). Efficient net: Rethinking Model Scaling for Convolutional Neural Networks. ICML.

Tibshirani, R., Hastie, W. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society. Series B (Statistical Methodology),63(2):411-423.

Tomasi, C., Manduchi, R. (1998). Bilateral filtering for grayscale and color images. IEEE ICCV.

ToppiReddy, R., Saini, B., & Mahajan, G. (2018). Crime prediction & monitoring framework based on spatial analysis. Procedia Computer Science, 132: 696-705.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M. (2015). Learning spatial temporal features with 3D convolutional networks. IEEE CVPR.

Umakanthan, S., Denman, S., Sridharan, S., et al. (2012). Spatio temporal feature evaluation for action recognition. Digital Image Computing Techniques and Applications.

Wang, G., Wu, X., Yan, W. (2017) The state-of-the-art technology of currency identification: A comparative study. International Journal of Digital Crime and Forensics 9 (3), 58-72.

Wang, J., Nguyen, M., Yan, W. (2017) A framework of event-driven traffic ticketing system. International Journal of Digital Crime and Forensics (IJDCF) 9 (1), 39-50.

Wang, H., Schmid, C. (2013). Action recognition with improved trajectories. IEEE ICCV, pp. 3551-3558.

Wang, H., Wang, et al. (2014). Clustering by pattern similarity in large data sets. SIGMOD, 23(4):394-405.

Wang, H., Zhang, T., Wu, J. (2017). The monkey typing solution to the YouTube-8M video understanding challenge. arXiv:1706.05150.

Wang, J., Bacic, B., Yan, W. (2018) An effective method for plate number recognition. Multimedia Tools and Applications 77 (2), 1679-1692

Wang, J., Liu, Z., Wu, Y., et al. (2004). Learning actionlet ensemble for 3D human

action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 914-927.

Wang, J., Yan, W. (2016) BP-neural network for plate number recognition. International Journal of Digital Crime and Forensics (IJDCF) 8 (3), 34-45.

Wang, L., Qiao, Y., Tang, X. (2016). MoFAP: A multi-level representation for action recognition. International Journal of Computer Vision, 119(3): 254–271.

Wang, L., Qiao, Y., Tang, X. (2015). Action recognition with trajectory-pooled deep-convectional descriptors. IEEE CVPR, pp. 4305-4314.

Wang, L., Qiao, Y., Tang, X. (2014). Video action detection with relational dynamic poselets. European Conference on Computer Vision (ECCV), pp.565–580.

Wang, L., Xiong, Y., Wang, Z., et al. (2016). Temporal segment networks: Towards good practices for deep action recognition. European Conference on Computer Vision, pp.20–36.

Wang, L., Yan, W. (2021) Tree leaves detection based on deep learning. International Symposium on Geometry and Vision.

Wang, P., Li, W., Wan, J., et al. (2018). Cooperative training of deep aggregation networks for RGB-D action recognition. AAAI Conference on Artificial Intelligence (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), pp. 7404–7411.

Wang, X., Yan, W. (2019) Cross-view gait recognition through ensemble learning. Neural Computing and Applications.

Wang, X., Yan, W. (2019) Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. International Journal of Neural Systems 29 (12).

Wang, X., Yan, W. (2019) Multi-perspective gait recognition based on ensemble learning. Springer Neural Computing and Applications.

Wang, X., Yan, W. (2019) Gait recognition using multichannel convolutional neural networks. Neural Computing and Applications.

Wang, X., Yan, W. (2019) Human gait recognition based on self-adaptive hidden

Markov model. IEEE/ACM Transactions on Biology and Bioinformatics.

Wang, X., Yan, W. (2020) Cross-view gait recognition through ensemble learning. Neural computing and applications 32 (11), 7275-7287.

Wang, X., Yan, W. (2020) Non-local gait feature extraction and human identification. Multimedia Tools and Applications.

Wang, Z., Yan, W., Oates, T., (2017). Time series classification from scratch with deep neural networks: A strong baseline, IEEE IJCNN, pp. 1578–1585.

Willems, G., Tuytelaars, T., Van Gool, L., et al. (2008) An efficient dense and scale-invariant spatial-temporal interest point detector. European Conference on Computer Vision, pp. 650-663.

Wong, Y., Mooney, R. (2006). Learning for semantic parsing with statistical machine translation. Association for Computational Linguistics, pp. 439-446.

Wu, J., Zhou, D., Xiao, G., et al. (2013). A hierarchical bag-of-words model based on local space-time features for human action recognition. IEEE International Conference on IT Convergence and Security.

Wu, Z., Wang, X., Jiang, Y., et al. (2015). Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. ACM International Conference on Multimedia, pp. 461-470.

Xing, J., Yan, W. (2021) Traffic sign recognition using guided image filtering. International Symposium on Geometry and Vision

Xiang, Y., Yan, W. (2021) Fast-moving coin recognition using deep learning. Multimedia Tools and Applications.

Xiao, B., Nguyen, M., Yan, W. (2021) Apple ripeness identification using deep learning. International Symposium on Geometry and Vision

Xing, J., Nguyen, M., Yan, W. (2021) The improved framework of traffic sign recognition by using guided image filtering. Springer Nature Computer Science.

Xu, Z., Yang, Y., Hauptmann, A. (2015). A discriminative CNN video representation for event detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1798-1807.

Yamato, J., Ohya, J., Ishii, K. (1993) Recognizing human action in time-sequential

images using hidden Markov model. Transactions of the Institute of Electronics Information & Communication Engineers, 76(9):379-385.

Yan, W., Kankanhalli, M., (2016) Face search in encrypted domain. Pacific-Rim Symposium on Image and Video Technology, 775-790.

Yan, S., Xiong, Y., Lin, D., et al. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. AAAI Conference on Artificial Intelligence, Innovative Applications of Artificial Intelligence (IAAI-18), and the Symposium on Educational Advances in Artificial Intelligence (EAAI-18), pp. 7444–7452.

Yan, W., Jain, R. (2008) Event detection from picture observations. International Workshop on Advanced Image Technology.

Yan, W., Weir, J. (2010) Fundamentals of Media Security. Bookboon

Yan, W., Chambers, J., Garhwal, A. (2014) An empirical approach for currency identification. Multimedia Tools and Applications 74 (7)

Yan, W. (2021) Computational Methods for Deep Learning: Theoretic, Practice and Applications, Springer.

Yan, W. (2019) Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics, Springer.

Ye, H., Wu, Z., Zhao, R., et al. (2015). Evaluating two-stream CNN for video classification. ACM on International Conference on Multimedia Retrieval, pp. 435-442.

Yeung, S., Russakovsky, O., Jin, N., et al. (2015). Every moment counts: Dense detailed labeling of actions in complex videos. International Journal of Computer Vision, pp. 375-389.

Yeung, S., Russakovsky, O., Mori, G., et al. (2016). End-to-end learning of action detection from frame glimpses in videos. IEEE Conference on Computer Vision and Pattern Recognition, pp.2678–2687.

Yu, Z., Yan, W. (2020) Human action recognition using deep learning methods. International Conference on Image and Vision Computing New Zealand.

Zeiler, M., Fergus, R. (2014). Visualizing and understanding convolutional neural

networks. ECCV.

Zeiler, D., Fergus, R. (2014). Visualizing and understanding convolutional networks. European Conference on Computer Vision. Springer, pp. 818-833.

Zhang, B., Wang, L., Wang, Z., et al. (2016). Real-time action recognition with enhanced motion vector CNNs. IEEE CVPR, pp. 2718-2726.

Zhang, J., Zheng, Y., Qi, D. (2017). Deep spatial-temporal residual networks for citywide crowd flows prediction. AAAI Conference on Artificial Intelligence.

Zhang, L., Yan, W. (2020) Deep learning methods for virus identification from digital images. International Conference on Image and Vision Computing New Zealand.

Zhang, Q., Yan, W., Kankanhalli, M. (2019) Overview of currency recognition using deep learning. Journal of Banking and Financial Technology 3 (1), 59–69.

Zhang, Q., Yan, W. (2018) Currency detection and recognition based on deep learning. IEEE International Conference on Advanced Video and Signal Based Surveillance.

Zhang, Y., Yan, W., Narayanan, A. (2017) A virtual keyboard implementation using finger recognition. International Conference on Image and Vision Computing New Zealand.

Zhang, T., Yang Z, Jia W, et al. (2015). A new method for violence detection in surveillance scenes. Multimedia Tools and Applications, pp. 75.

Zhao, K., Yan, W. (2021) Fruit detection from digital images using CenterNet. International Symposium on Geometry and Vision.

Zheng, K., Yan, W., Nand, P. (2017) Video dynamics detection using deep neural networks. IEEE Transactions on Emerging Topics in Computational Intelligence.

Zhou, L., Yan, W., Shu, Y., Yu, J. (2018) CVSS: A cloud-based visual surveillance system. International Journal of Digital Crime and Forensics (IJDCF) 10 (1), 79-91.

Zhou W., Bovik, A., Sheikh, H., Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612.

Zhu, L., Liu, Y., Yang, Y. (2017). UTS submission to Google YouTube-8M Challenge.

arXiv:1707.04143.

Zhu, Y., Yan, W. (2022) Traffic sign recognition based on deep learning. Multimedia Tools and Applications.

Zhu, W., Hu, J., Sun, G., et al. (2016). A key volume mining deep framework for action recognition. IEEE CVPR, pp. 1991-1999.

Zhu, W., Lan, C., Xing, J., et al. (2016). Co-occurrence feature learning for skeleton-based action recognition using regularized deep LSTM networks. National Conference on Artificial Intelligence, pp. 3697-3703.