Joint Position-based Anomaly Detection using Graph Convolution Network

Prepared By
Md. Wasiul Haque 180042110
Mohammed Afzal Siddique 180042112
Md. Hasan Saju 180042113

Under the Supervision of
Md. Bakhtiar Hasan
Assistant Professor
Department of Computer Science and Engineering
Islamic University of Technology

Dr. Md. Hasanul Kabir
Professor
Department of Computer Science and Engineering
Islamic University of Technology

May 20, 2023

Department of Computer Science and Engineering
Islamic University of Technology

Declaration of Authorship

This is to certify that the work presented in this thesis represents our own research and intellectual contributions of Md. Wasiul Haque, Mohammed Afzal Siddique and Md. Hasan Saju under the supervision of Dr. Md. Hasanul Kabir, Professor, Department of Computer Science and Engineering, Islamic University of Technology and Md. Bakhtiar Hasan, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology. It is also declared that this thesis has not been submitted in its whole or in part for any academic qualification. All the sources of information used and referenced from published or unpublished work of others have been duly acknowledged and cited.

Authors:

Wasiul 28-5-23	Afzal 28-5-23	Saju 28-5-23
Md. Wasiul Haque	Mohammed Afzal Siddique	Md. Hasan Saju
Student ID: 180042110	Student ID: 180042112	Student ID: 180042113

Supervisors:

Dr. Md. Hasanul Kabir
Professor
Department of Computer Science and Engineering
Islamic University of Technology

Bakhtiar

Md. Bakhtiar Hasan
Assistant Professor
Department of Computer Science and Engineering
Islamic University of Technology

Contents

List of Figures

List of Tables

Acknowledgment

In the name of Allah, the Most Merciful and Most Gracious.

We are extremely grateful to our thesis advisors, Md. Bakhtiar Hasan and Dr. Md. Hasanul Kabir, for their exceptional guidance, encouragement, and knowledge throughout our research. Their unwavering support, extensive knowledge, and commitment to academic excellence were essential in the formation of this thesis. Their mentorship, invaluable insights, and constructive feedback have significantly contributed to the quality and profundity of this work, for which we are extremely grateful.

We are indebted to Md. Bakhtiar Hasan for his unwavering support, tolerance, and guidance. His extensive experience, meticulous attention to detail, and commitment to research have been an endless source of motivation. We are extremely appreciative of his advice, which has helped us overcome obstacles and refine our ideas, methodology, and analysis.

We would also like to thank Dr. Md. Hasanul Kabir for his unwavering support, intellectual direction, and scholarly contributions. This thesis's quality and rigor have been significantly enhanced by his knowledge, profound subject-matter comprehension, and insightful comments. We are appreciative of his encouragement, which inspired me to investigate new perspectives and approach difficult problems with assurance.

The completion of this undergraduate thesis would not have been feasible without Md. Bakhtiar Hasan and Dr. Md. Hasanul Kabir's unwavering support and guidance. We are extremely appreciative of their contributions and the chance to learn from them.

Abstract

The conventional approach to detecting anomalies in videos involves utilizing convolutional neural networks (CNNs) for pixel-level computations, which can be resource-intensive. The consideration of background information in frames by CNN-based approaches can lead to a reduction in the accuracy of anomaly predictions. The proposed solution for addressing the issues related to anomaly prediction is the Double-feature Double-motion Network (DD-Net), which is an architecture that is both lightweight and efficient. The utilization of human joint positions extracted from video frames as input to DD-Net is proposed as an alternative to processing frames pixel by pixel. Significant reduction in computational overhead can be achieved by adopting this approach as opposed to pixel-level computations. The utilization of human joint positions serves as a means to reduce the impact of extraneous factors on the forecasting of anomalies. The thesis demonstrates that DD-Net is capable of achieving quicker anomaly prediction without compromising on the accuracy of the predictions, as evidenced by the results of the experiments conducted. DD-Net's simplicity enables possible improvements and expansions, including the investigation of alternative frame sampling techniques, and integration of RGB or depth data to enhance performance. The effectiveness and potential of utilizing DD-Net for precise and efficient detection of joint-based anomalies are underscored by our findings.

Chapter 1

Introduction

The exponential growth of information volume and complexity in today's data driven world is unprecedented. The generation of vast amounts of data across various domains, from financial transactions and network traffic to healthcare records and industrial sensors, is a prevalent phenomenon. The vast amount of data available contains untapped potential in the form of valuable insights and concealed patterns. The impact of anomalies on datasets can lead to distorted understanding and analysis. The identification and separation of aberrant instances from normal behavior is the primary objective of anomaly detection, which is a crucial task in data analytics. Anomaly detection is a critical component in maintaining data integrity, improving decision-making capabilities, and identifying possible threats or fraudulent activities.

The manifestation of anomalies can take on various forms, such as sudden fluctuations in stock prices, fraudulent credit card transactions, network intrusions, manufacturing defects, and atypical medical symptoms. The presence of anomalies can signal the occurrence of crucial incidents or malfunctioning systems that demand prompt action. The real-time detection of these entities is crucial for mitigating harm, reducing losses, and maintaining the optimal operation of diverse applications and systems.

Anomaly detection is a methodology that utilizes statistical, machine learning, and data mining techniques to detect patterns and deviations from anticipated behavior in datasets. The limitations of traditional anomaly detection methods, which rely on domain-specific rules or thresholds, are evident in their inability to adapt to changing data patterns or identify novel anomalies. This thesis discusses the emergence of sophisticated and automated approaches that utilize advanced algorithms and techniques, including clustering, classification, regression, and deep learning, to address the challenges presented by anomalies in complex datasets.

The detection and interpretation of anomalies in visual data pose a complex challenge due to their ability to manifest in various forms. The presence of anomalies can encompass a variety of phenomena such as uncommon entities or occurrences, unanticipated visual configurations, irregularities in structure, or atypical conduct. Anomaly detection in surveillance systems can be utilized to identify suspicious activities, such as trespassing or theft, by recognizing unusual movements or objects in real-time video streams. Anomaly detection in medical imaging can aid in the timely identification of illnesses by identifying abnormal characteristics or areas in X-rays, MRIs, or histopathological slides. Anomaly detection is a crucial tool in manufacturing processes as it enables the identification of defects or deviations from quality standards, thereby ensuring that only products that meet the desired specifications are released into the market. The limitations of the methods used were observed in their inability to generalize across various datasets and effectively process intricate visual patterns. The utilization of deep learning, specifically convolutional neural networks (CNNs), has transformed the field of anomaly detection through the facilitation of automated feature extraction and end-to-end learning from extensive visual data. The utilization of deep learning has resulted in notable progress in anomaly detection, leading to improved precision and resilience in detecting anomalies across various visual domains.

1.1 Motivation

This thesis research on joint position-based anomaly detection was motivated by the need for a rapid and efficient method to detect anomalies in video sequences. In conventional approaches, anomaly detection is typically conducted with convolutional neural networks (CNNs), where each pixel of the video frames is analyzed individually to predict anomalies. This pixel-by-pixel computation can be computationally costly and time-consuming, however.

To address this difficulty, our motivation is to reduce the computational burden by leveraging joint position information. By concentrating on the positions and movements of various body parts rather than pixel-level analysis, we hope to expedite the process of anomaly detection. By providing joint position data as input, we anticipate a substantial reduction in computation requirements, allowing for quicker anomaly detection.

By employing a joint position-based strategy, we aim to improve the efficacy and speed of anomaly detection, thereby enabling prompt responses to anomalous events and facilitating the prompt resolution of their repercussions. This study seeks to investigate the viability of joint position information as a lightweight alternative to pixel-level analysis, providing a more effective method for anomaly detection in video sequences.

1.2 Scope

The scope of our thesis research is the creation and evaluation of a joint position-based anomaly detection technique. The primary objective is to investigate the feasibility of utilizing joint position information to reduce the computational burden of anomaly detection processes. By utilizing joint position data as opposed to pixel-level analysis, we hope to accelerate the process of anomaly detection, allowing for rapid responses to anomalous events.

Our research extends to real-world situations requiring immediate action, such as emergency situations involving gunfire or security intrusions. By employing a joint position-based approach, we hope to provide a more effective and timely solution for anomaly detection that can assist in expeditious decision-making and efficient resolution of anomalous incidents.

In addition, within the scope of our thesis, we will compare the performance of the joint position-based anomaly detection method to that of conventional CNN-based methods. Extensive experiments and comparative analyses will be conducted to evaluate the computational efficiency, detection accuracy, and reactivity of the proposed method.

In addition to the above-mentioned primary scope, we may investigate additional aspects, such as the influence of various joint position representations or the incorporation of temporal information for improved anomaly detection. These potential extensions will depend on the progress of the research and the availability of resources throughout the duration of the thesis.

1.3 Research Challenges

In our dissertation on joint position-based anomaly detection, we encountered difficulties regarding the availability and suitability of data sets. First, there are insufficient datasets designed specifically for joint position-based anomaly detection. Therefore, extant datasets must be modified to facilitate the extraction and utilization of simultaneous position information. This procedure requires overcoming a number of technical and methodological obstacles, such as accurately capturing and representing joint positions in the modified dataset.

The large volume of extant datasets for anomaly detection poses a resource constraint in terms of storage capacity, computational power, and processing time. These datasets, which are frequently collected by scanning YouTube videos, may not adequately represent real-world circumstances and may lack a diversity of anomaly types and scenarios. As a result, assuring the applicability and generalizability of our joint position-based anomaly detection method to real-world environments becomes a formidable challenge.

In addition, addressing the efficacy and accuracy of the joint position-based approach in comparison to traditional pixel-based CNN methods may present additional challenges. Evaluating the efficacy and robustness of the proposed method against different types of anomalies, varying levels of occlusions or noise, and variations in joint position representations may present additional research challenges.

In addition, contemplating the temporal aspect of anomaly detection and incorporating it into the joint position-based approach could represent an additional challenge in the field of research. Exploring methods to capture and utilize temporal dynamics in joint position sequences for enhanced accuracy and responsiveness in anomaly detection may necessitate the development of novel techniques and algorithmic modifications.

Overall, overcoming these research obstacles associated with dataset modification, resource constraints, dataset representativeness, and addressing the performance and temporal aspects will contribute to the advancement and efficacy of joint position-based anomaly detection methods.

1.4 Research Contributions

Our contribution to the field of joint position-based anomaly detection encompasses a number of important aspects.

Initially, we modified the existing DD-Net model's architecture by altering parameters such as attrition rate, tolerance level, and learning rate. These modifications were intended to improve the model's accuracy and efficacy in detecting anomalies using joint position data.

Second, we modified an existing dataset to address the dearth of dedicated datasets for joint position-based anomaly detection. This process of modifying the dataset included integrating and processing joint position data to assure its compatibility with our proposed method. By providing a modified dataset designed particularly for joint position-based anomaly detection, we improve the accessibility and utility of resources in this field.

In addition, we conducted extensive experiments to evaluate and compare the efficacy of conventional anomaly detection methods (e.g., pixel-based CNN) with our joint position-based anomaly detection method. We quantitatively and qualitatively evaluated the efficacy, efficiency, and overall superiority of our proposed method through extensive experimentation and analysis. This comparative analysis contributes to the advancement of anomaly detection techniques and reveals the potential advantages of leveraging joint position information.

In addition to the previously mentioned contributions, our dissertation may also introduce novel insights and perspectives regarding the application of joint position-based anomaly detection in real-world scenarios. By investigating and highlighting the benefits and limitations of this method, we contribute to the field of anomaly detection research and lay the groundwork for future studies and advancements in this area.

Overall, our dissertation on joint position-based anomaly detection contributes to the refinement of architectural models, the creation of modified datasets, the empirical evaluation of detection performance, and the investigation of novel applications. Through these contributions, we hope to advance the discipline and encourage additional research in the area of anomaly detection based on joint position information.

1.5 Organization

The remaining sections of the dissertation are structured as follows. Background and motivation for joint position-based anomaly detection are discussed in the second chapter. It also identifies the issues that continue to plague the existing literature. The third chapter presents a novel method for detecting anomalies using joint position. The fourth chapter evaluates the performance of the proposed pipeline and compared it to the performance of other existing systems. The fifth chapter concludes our discussion and outlines the purview of future research.

Chapter 2

Literature Review

Detecting anomalies from a given video has been researched for years [8, 9, 10, 11, 12, 13]. Some early works of Anomaly Detection treat it as a multiple-instance learning (MIL) problem [1]. Scores were given to video segments named "Anomaly Score", and to determine the score they proposed a MIL ranking loss with sparsity and smoothness constraints for a deep learning network [1].

Anomaly detection by self-training has also been widely used in semi-supervised learning [14, 15]. This method increases the labeled data through pseudo-label generation. Recent self-learning approaches involve representation learning of the feature encoders [15].

Here, our architecture will use pose estimation to self-train clip-based pseudo labels to all clips in abnormal videos using a feature extractor that uses GCN.

Convolutional neural networks (CNNs) [16, 17, 18] and recurrent neural networks (RNNs) [19, 20, 21] were the go-to models in the early phases of deep learning-based techniques for skeleton-based action recognition. However, as none of these methods did explicitly exploit the structural topology of the joint positions, their capability was limited.

2.1 Real-World Anomaly Detection in Surveillance Videos [1]

Sultani et al. [1] have proposed a method for detecting anomalies in surveillance video using multiple instance learning (MIL). They suggested that by using MIL, they can avoid the time-consuming process of annotating each anomalous segment in the training data and instead only label the entire video as anomalous or normal. They also propose introducing sparsity and temporal smoothness constraints in the ranking loss function to better localize anomalies during training.

They reported that their MIL-based anomaly detection method achieved significant improvement in anomaly detection performance compared to state-of-the-art approaches.

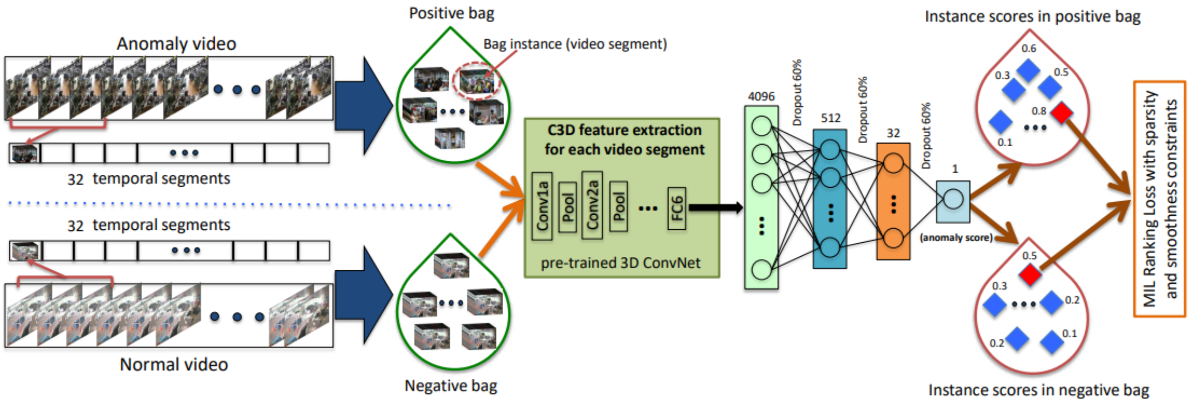

Figure 2.1: The Flow Diagram of the MIL-based Anomaly Detection Approach

The challenges they faced include the need for a practical anomaly detection system to be able to signal deviations from normal patterns in a timely manner, the difficulty of developing algorithms that can detect a wide range of anomalous events without prior information about those events, and the issue of false alarms due to normal behaviors that may appear anomalous due to changes in the environment over time.[1]

In this approach, they divided all the videos into positive and negative sections, labeling the videos that contain anomalies somewhere as positive and the ones that do not have any anomalies as negative. Each of the videos is represented as a bag and each temporal section of the video is described as an instance in the bag. After extracting features from the video segments, a fully connected neural network is trained. This training utilizes a novel ranking loss function that computes the ranking loss between the highest-ranked instances in the positive and negative bags [1]. But as

However, as Sultani et al. noted, these approaches can still produce high false alarm rates due to the complexity and diversity of real-world anomalous events. [1]

Their approach uses CNN to extract the features from the videos, it is very computationally expensive and can be hindered in complicated and complex situations. Using a graph to extract features instead of CNN's pixel-by-pixel approach is expected to lessen the computational intensity and separate features more efficiently in situations with significant background noises.

2.2 Video Anomaly Detection System using Deep Convolutional and Recurrent Models [2]

Qasim et al. have suggested three different models for effective anomaly detection from surveillance videos. For extracting features from the sample videos, they have used a combination of a deep convolutional neural (CNN) network and a simple recurrent unit (SRU). They have built an automated system that can detect anomalous activities from video footage [2]. And the dataset they have used for this is UCF-Crime [1]. The models that they have tested are:

1. ResNet18 + SRU
2. ResNet34 + SRU
3. ResNet50 + SRU

2.2.1 Proposed System by Qasim et al. [2]

Residual Networks (ResNet) are deployed first to get the spatial features from the videos. And to get the temporal features, Simple Recurrent Model is used. A combination of these two is used to extract the spatiotemporal features from the video frames, and then to identify anomalous videos from the extracted frames, the attributes are processed through the method of maxpooling and over fully connected layers [2].

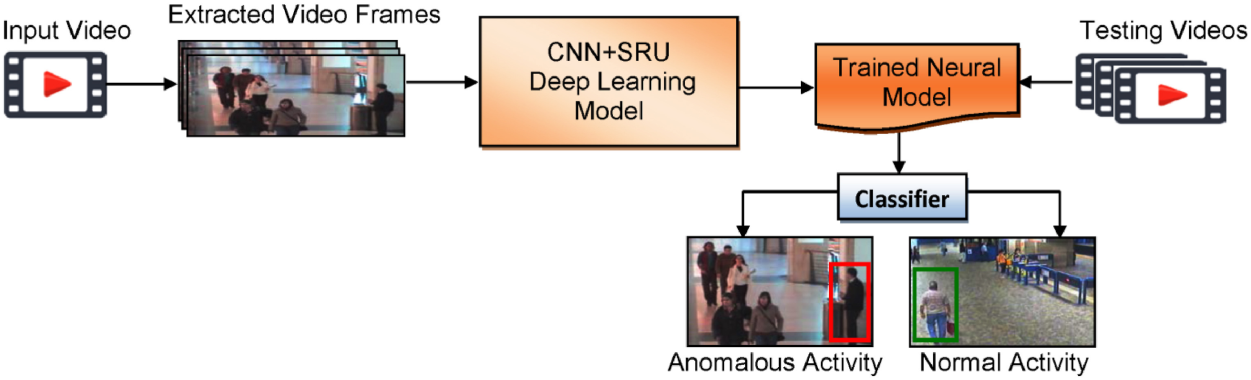

Figure 2.2: Proposed mechanism for Qasim et al. [2]

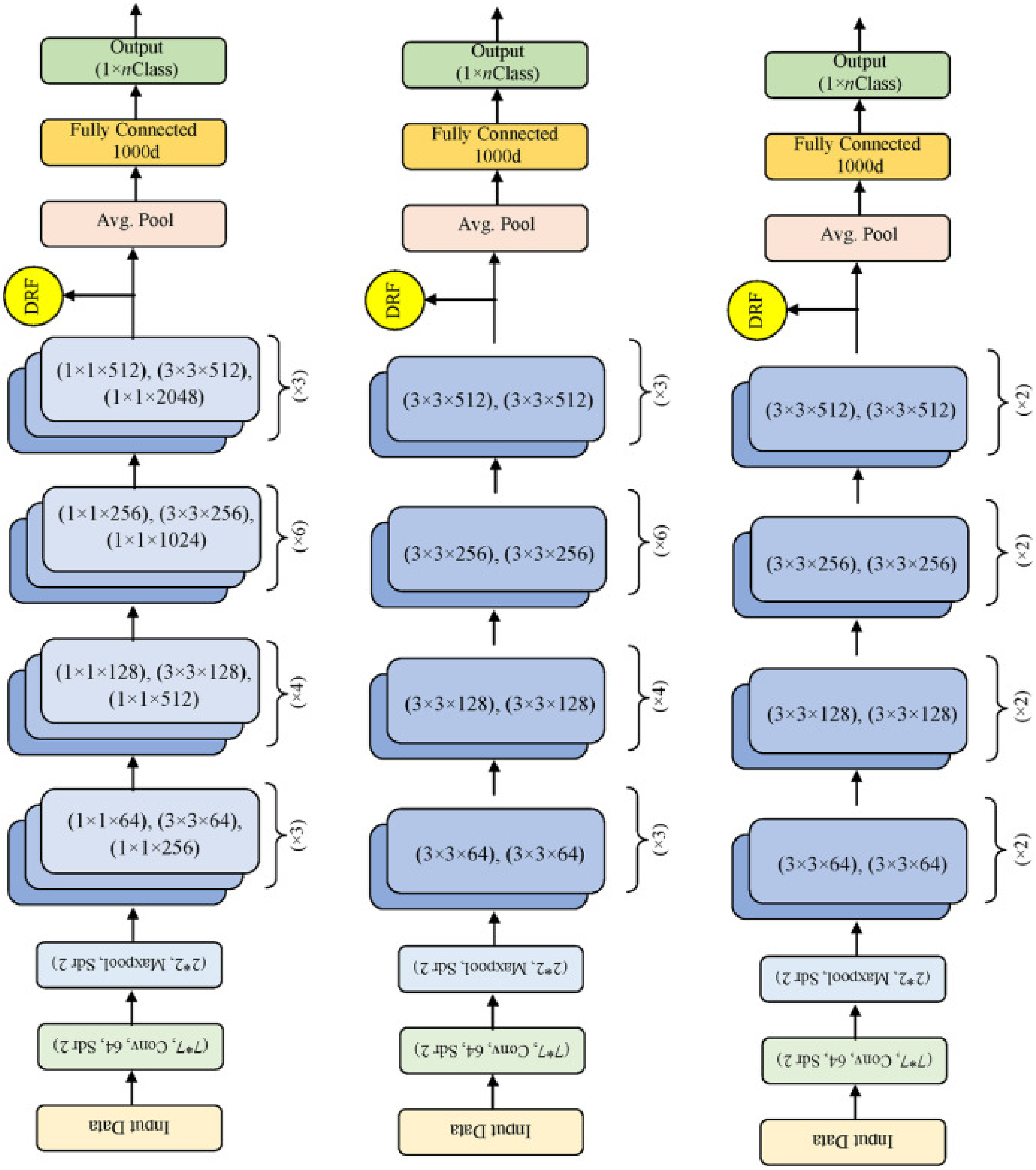

Figure 2.3: Architecture of ResNet50, ResNet34 and ResNet18 [2]

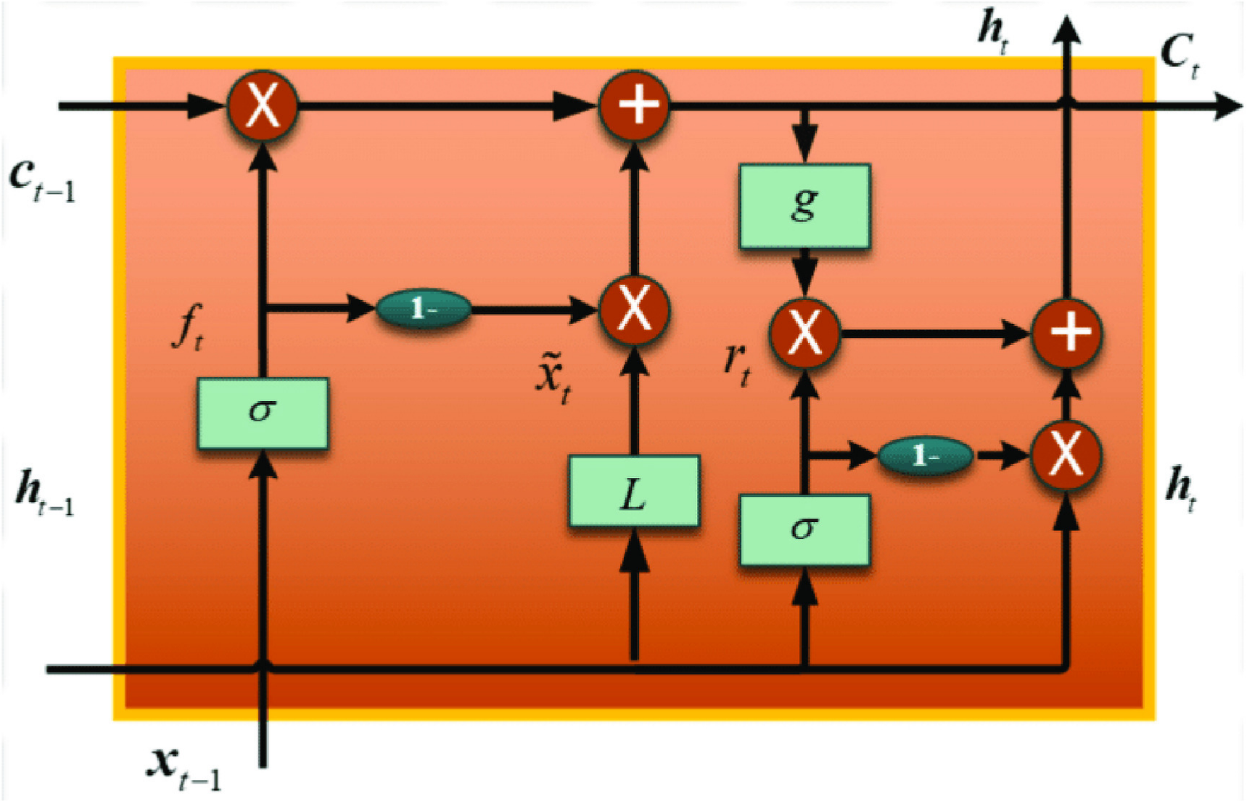

Figure 2.4: Internal Configuration of the Simple Recurrent Unit [2]

2.2.2 Setting the Model

Qasim et al. have performed the experiments by using the ResNet architectures from the Keras library. Several hyper-parameters were used to tune the models and achieve the highest result. Table 2.1 shows their hyper-parameter setup and the accuracy measures they have got for each of the tunings.

Hyper Parameters	Tuning	Accuracy in %
Weights Intialization	Glorot-Uniform (Xavier)	91.11%
Weights Intialization	Random-Uniform	89.23%
Weights Intialization	He-Uniform	89.07%
Optimizer	Adam	91.43%
Optimizer	RMSprop	90.30%

Table 2.1: Hyper Parameters values by Qasim et al. [2]

2.2.3 Contribution of Qasem et al. [2]

Their whole contribution can be summarized into the following points:

1. A novel framework for identifying anomalies in the binary dataset. This framework uses a combination of different ResNet architectures along with SRU models.

2. Their approach surpassed the accuracy of all other models in the same dataset.

2.3 Multiple Instance Self-Training Framework for Video Anomaly Detection [3]

Feng et al. have proposed a multiple-instance self-training framework for weakly supervised video anomaly detection (WS-VAD). WS-VAD refers to the task of distinguishing anomalous events from normal events in a video using only video-level annotations, rather than detailed annotations for each anomalous event. [3]

The proposed framework, called MIST, is designed to efficiently refine task-specific discriminative representations for WS-VAD. MIST consists of two components: a multiple instances pseudo-label generator and a self-guided attention-boosted feature encoder. The pseudo label generator uses a methodology to sparse continuous sampling to generate more credible clip-level pseudo labels, while the feature encoder aims to automatically focus on anomalous regions in the video frames while extracting task-specific representations. The authors also adopt a self-training scheme to optimize both components and obtain a task-specific feature encoder. [3]

The existing WS-VAD methods can be divided into two categories: encoder-agnostic methods [1, 22], which use task-agnostic features extracted from a vanilla feature encoder to estimate anomaly scores, and encoder-based methods, which train both the feature encoder and the classifier simultaneously.[3]

2.3.1 Pseudo Label Generation

Extracts features of abnormal videos and normal videos from a pre-trained encoder. It trains a pseudo-label generator with extracted features. The parameters of the pseudo-label generator are updated by means of the deep MIL ranking loss. By feeding the extracted features to the pseudo-label generator, the anomaly scores of the clips are estimated. [3]

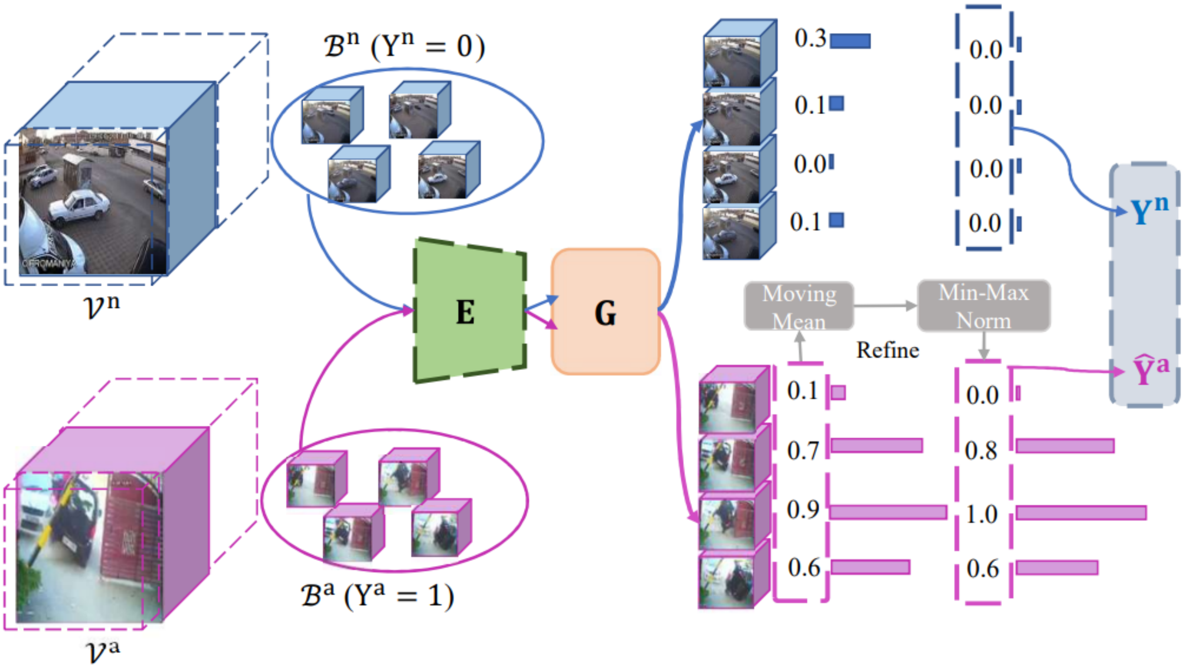

Figure 2.5: Pseudo Labels Generation

2.3.2 Feature Encoder Fine-tuning

The pre-trained encoder is task-agnostic. To better distinguish anomalous clips from normal ones, a self-guided attention module in the feature encoder is introduced. This encoder is task aware and can produce more discriminative representations.[3]

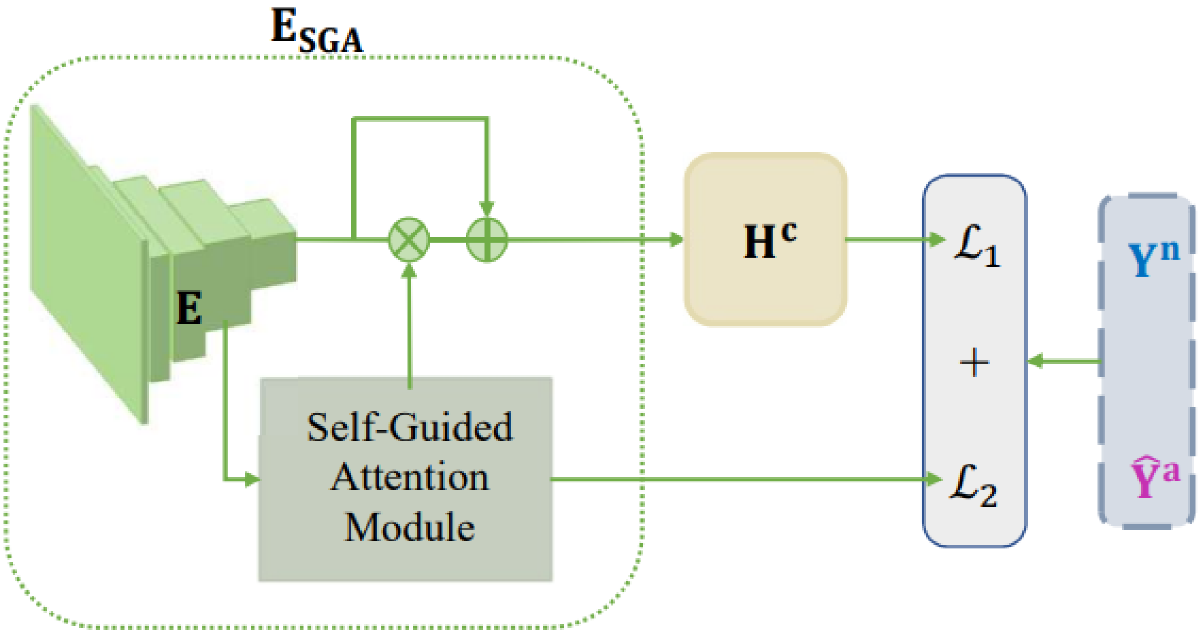

Figure 2.6: Feature Encoder Fine Tuning

Feng et al. [3] have proposed a new WS-VAD method called Multiple Instance Self-Training (MIST), which consists of a multiple instances pseudo label generator and a self-guided attention-boosted feature encoder. The pseudo label generator uses a MIL framework and a sparse continuous sampling strategy to produce more accurate pseudo labels, while the feature encoder uses a self-guided attention module to emphasize anomalous regions in the video. They also introduced a deep MIL ranking loss and an efficient two-stage self-training scheme to optimize MIST [3].

2.4 Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection [4]

Ristea et. al have proposed a new approach for anomaly detection that involves integrating a reconstruction-based functionality into a self-supervised predictive architectural building block. This block is designed to be generic and can be easily incorporated into existing state-of-the-art anomaly detection methods. [4]

The block comprises of a channel attention module and a convolutional layer with dilated filters. The activation maps produced by masking the center of the receptive field in the filters are then sent via the channel attention module. The block has a loss function that reduces the reconstruction error relative to the receptive field's masked region [4].

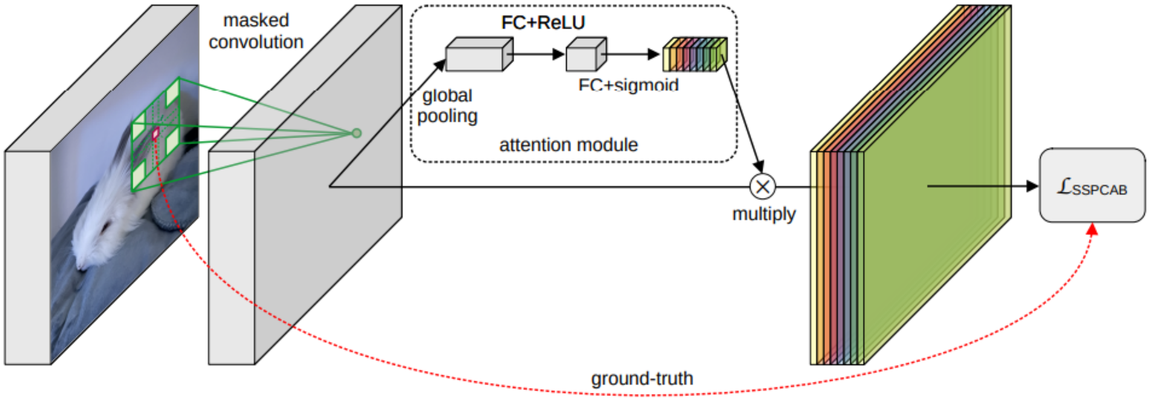

Figure 2.7: Self-supervised Predictive Convolutional Attentive Block (SSPCAB) [4]

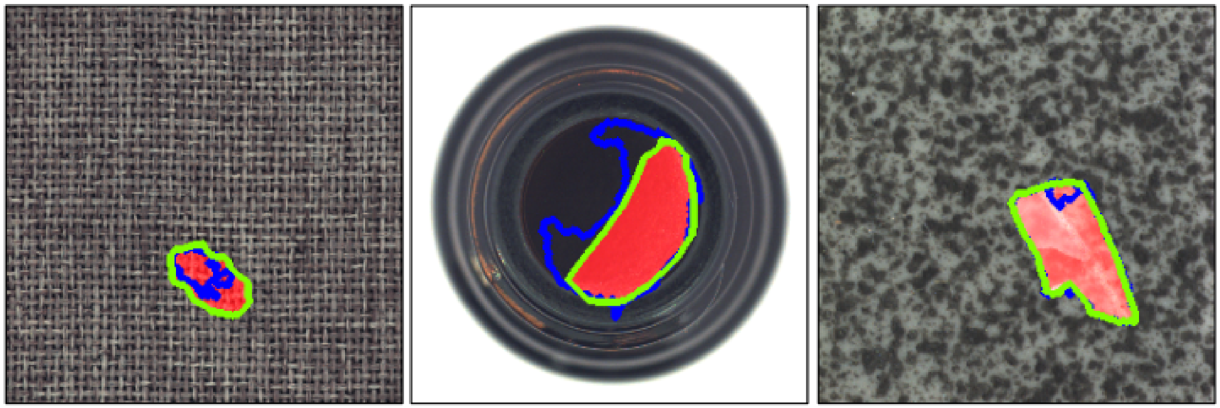

Figure 2.8: Anomaly localization examples. The ground truth anomalies are marked with a red mask. [4]

2.5 Self-Supervised Sparse Representation for Video Anomaly Detection [5]

Wu et al. have proposed a self-supervised sparse representation (S3R) framework for video anomaly detection (VAD). The S3R framework is designed to be able to handle both one-class VAD, where the model can only learn from standard training samples, and weakly-supervised VAD, where the model is provided with video-level normal/anomaly labels.

The S3R framework models the concept of the anomaly at the feature level using a combination of dictionary-based representation and self-supervised learning. It consists of two coupled modules, en-Normal and de-Normal, reconstructing snippet-level features and filtering out normal-event features using a learned dictionary. The self-supervised techniques used in S3R also allow the generation of pseudo-normal/anomaly samples to train the anomaly detector. [5]

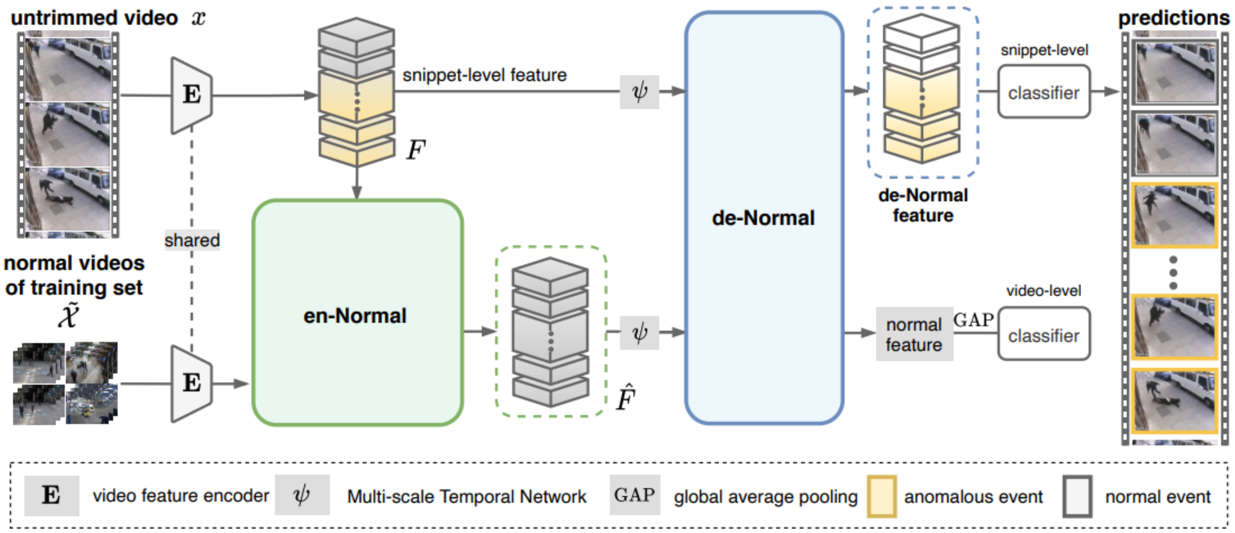

Figure 2.9: S3R framework [5] couples dictionary learning with self-supervised techniques to model the concept of feature-level anomaly

2.6 Make Skeleton-based Action Recognition Model Smaller, Faster and Better [6]

One drawback of Skeleton-based Action Recognition is that the model size is quite heavy, hence the execution speed is very slow [23, 24, 25]. This heaviness and slow execution made it difficult to apply these skeleton features to detect action in real-world scenarios. From that problem, Yang et al. have proposed a lightweight architecture called Double-feature Double-motion Network (DD-Net) [6]. This network architecture contains a Joint Collection Distance (JCD) feature and a two-scale global motion feature. Their model also has the capability to adapt in different scenarios [6].

One limitation to their approach can be due to the lightness of their model and the usage of fewer parameters might limit its ability to capture complex and complicated patterns present in some challenging action recognition tasks.

They used both SHREC and JHMDB datasets for evaluating their results [6].

	SHREC Dataset	**JHMDB Dataset**
Number of Samples	2800	928
Training / Testing Setup	1 Training Set 1 Testing Set	3 Split Training/ Testing Sets
Dimention of skeletons	3	2
Subject	hand	body
Number of actions	14 and 28	21
Actions are strongly correlated to global trajectories	Yes	No

Table 2.2: Dataset setup for DD-Net [6]

But as their model is lightweight enough to process skeleton-based data fast and efficiently, there is potential for this to be used in detecting anomalies, as this is already being used for recognizing actions.

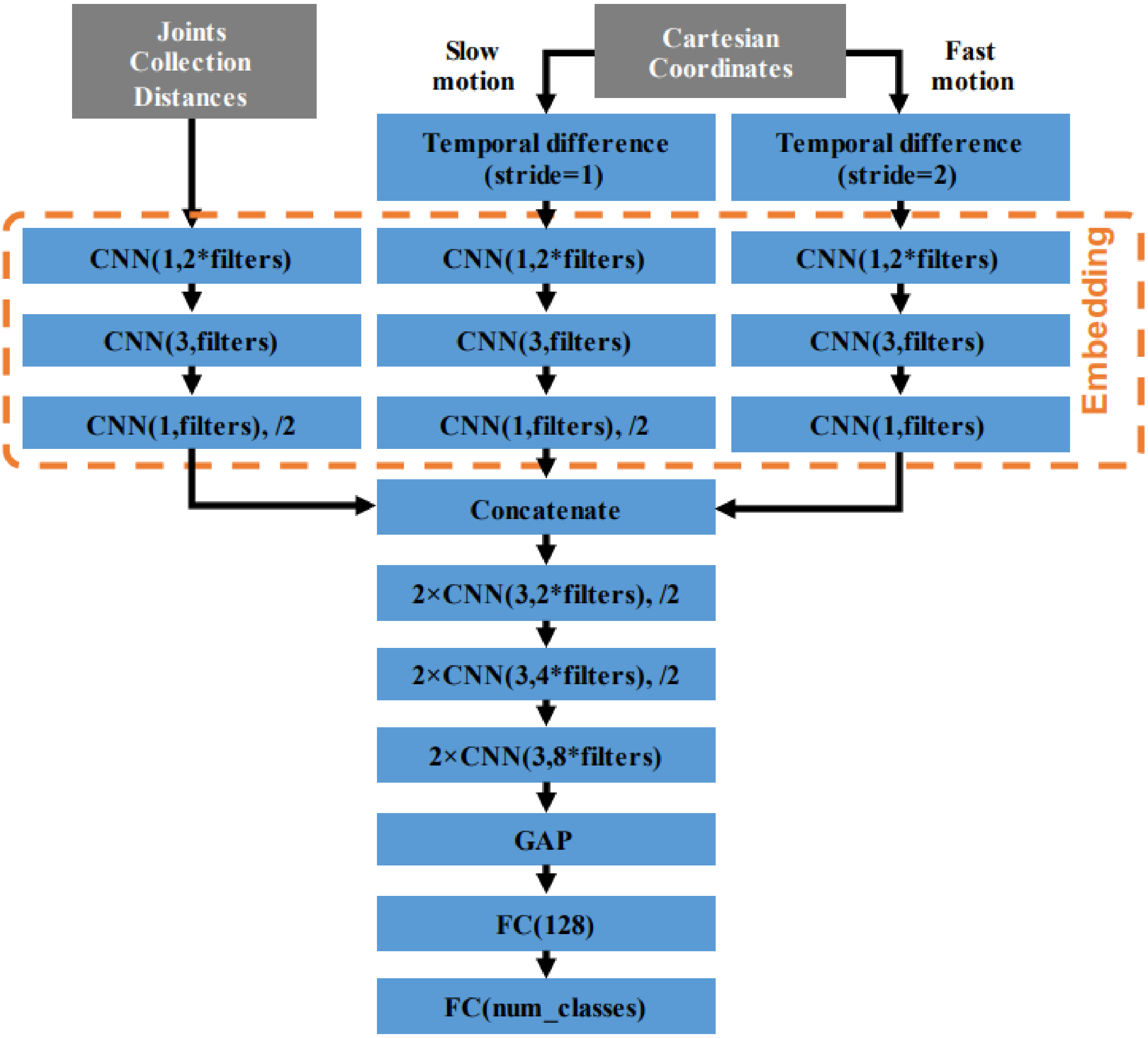

Figure 2.10: Architecture of DD-Net [6]

2.7 InfoGCN: Representation Learning for Human Skeleton-based Action Recognition [7]

Chi et al. have proposed a learning framework for human skeleton-based action recognition called InfoGCN [7]. The framework combines a novel learning objective and an encoding method to learn informative but compact latent representations of human action. [7]

The model is directed to concentrate on the most instructive features of the data by the learning target, which is based on an information bottleneck [26]. The authors offer attention-based graph convolution, which captures the context-dependent intrinsic topology of human activity, to give discriminative information for action categorization. Additionally, they give a multi-modal depiction of the skeleton that incorporates additional spatial information from the relative joint locations. [7]

Human action recognition based on the skeleton has been a popular topic in recent times in the field of computer vision due to its resilience against a noisy background [27, 28]. And a significant achievement in this field is the introduction of the graph convolution network (GCN) based approach [29].

Traditional graph representation using bone connectivity has a limitation that it can ignore the indirect joint relations during various activities, which are only visible from the intrinsic topological perspective. In this regard, InfoGCN [7] develops a novel graph convolution module to capture the underlying graph structure at the time of constructing the skeleton sequence. According to the behavioral settings of identical postures that arise in various activities, as indicated in Figure 8, the inferred topologies may change.

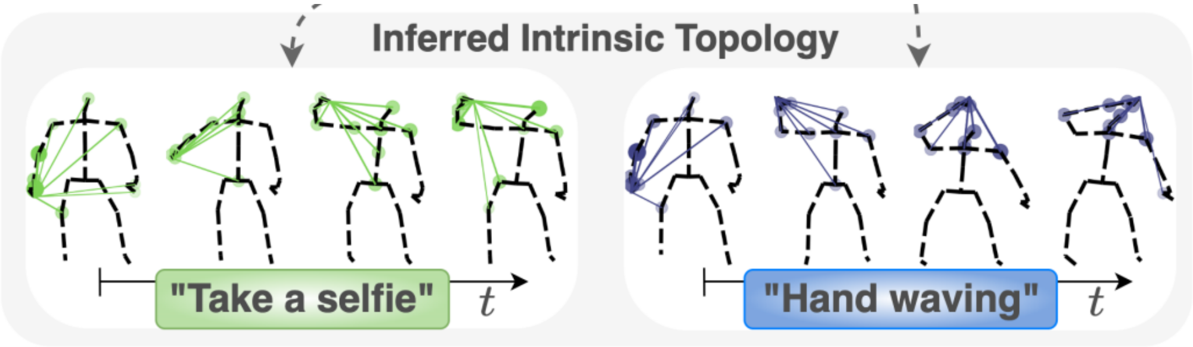

Figure 2.11: Intrinsic topology of two different actions [7]

The architecture of InfoGCN is divided into these three parts: [7]

1. **Embedding Block:**

   Transforms a sequence of the skeleton to initial joint representations

2. **Encoding Block:**

   Extracts spatio-temporal features from the initial joint representations

3. **Softmax:**

   For classifying 60 action classes

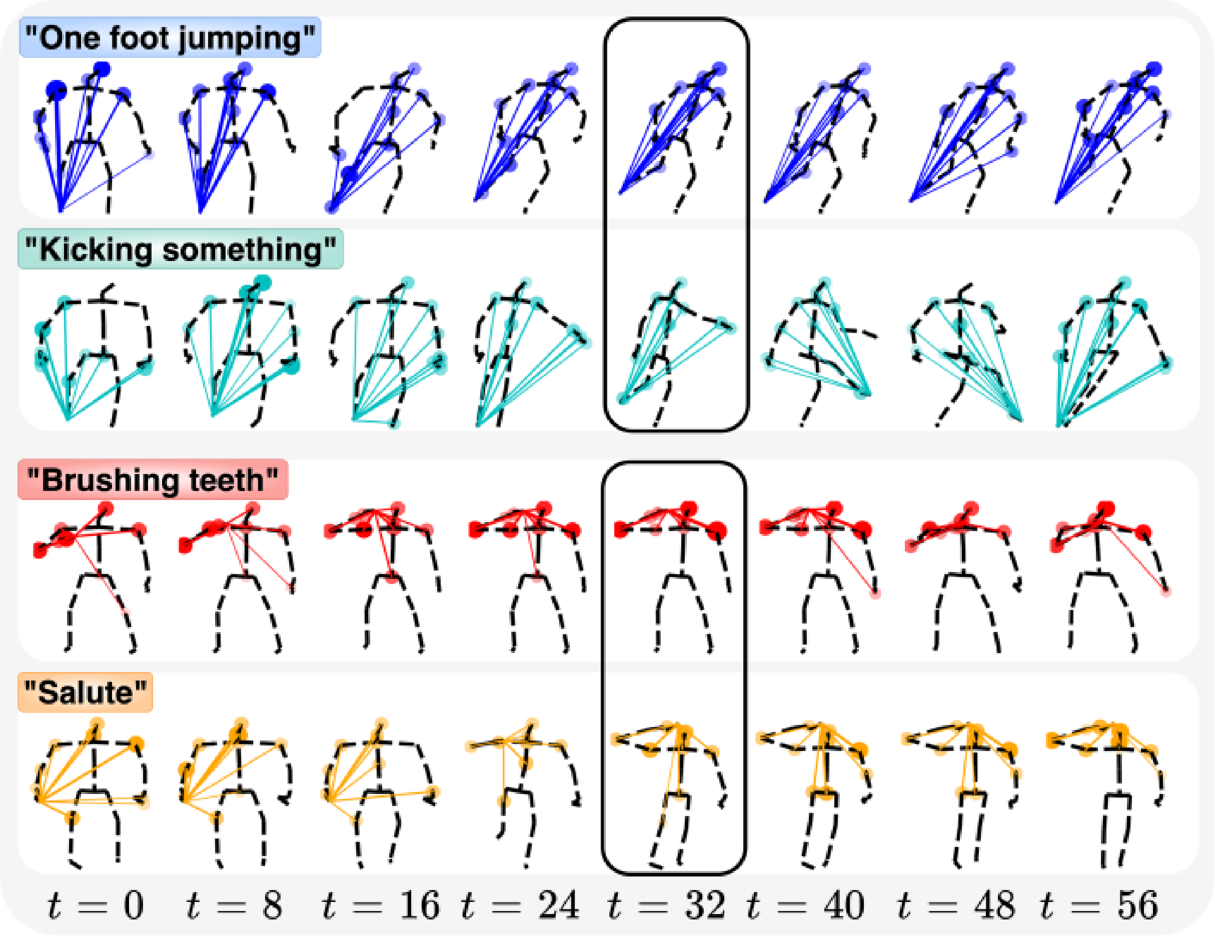

Figure 2.12: Examples of the context-dependent intrinsic topology of self-attention-based graph-convolution. Colored lines show the topology that may be derived from a certain joint (such as the hand or foot) to all the other joints. The intensity of the implied relationship is inversely correlated with the thickness of the colored lines and the size of the circles on joints [7]

Finally, Chi et al. has proposed a method for improving recognition performance in a computer vision system that involves using multiple modalities, specifically the relative positions of joints in a skeleton representation, in order to provide complementary spatial information about the joints. Using an ensemble of models trained with these representations has further improved the recognition performance. [7]

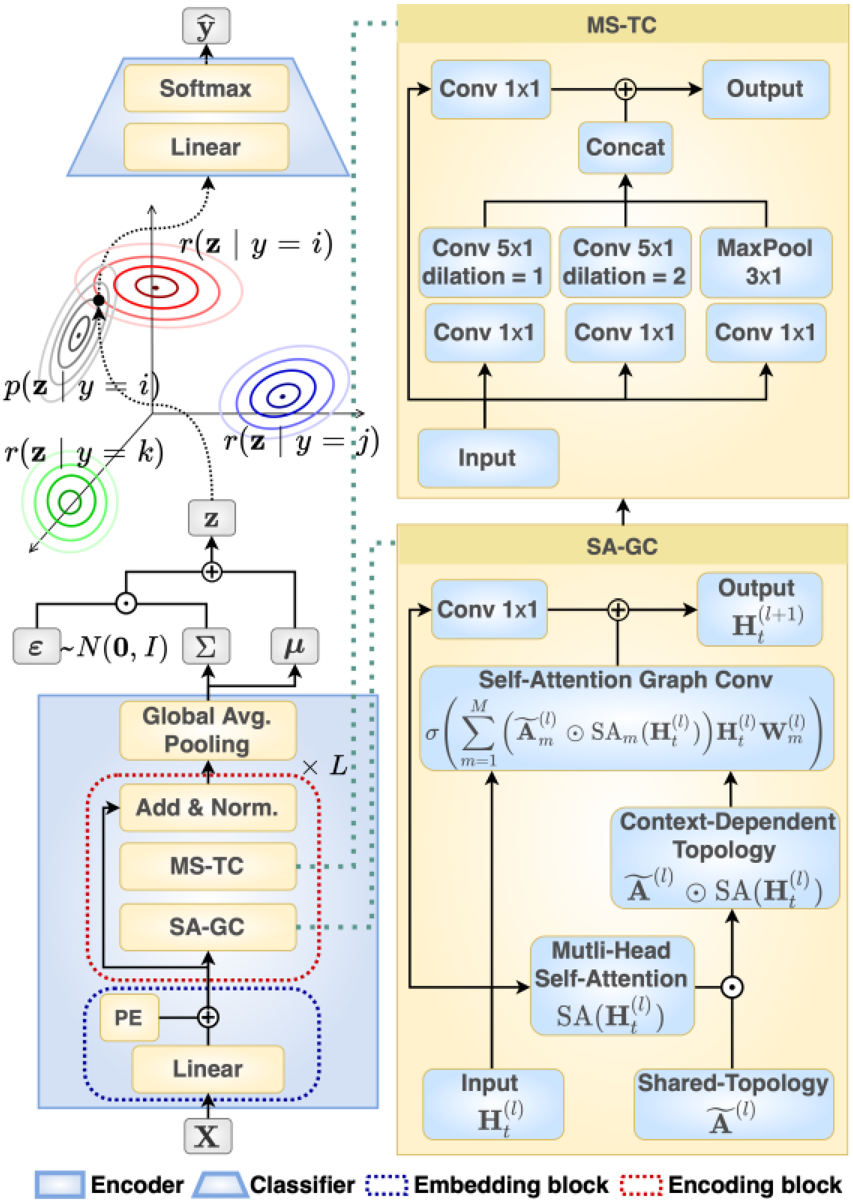
MS-TC
Conv 1x1
Output
Concat
Conv 5x1 dilation = 1
Conv 5x1 dilation = 2
MaxPool 3x1
Conv 1x1
Conv 1x1
Conv 1x1
Input
SA-GC
Conv 1x1
Output $\mathbf{H}_t^{(l+1)}$
Self-Attention Graph Conv
$\sigma\left(\sum_{m=1}^{M}\left(\widetilde{\mathbf{A}}_m^{(l)} \odot \mathrm{SA}_m(\mathbf{H}_t^{(l)})\right)\mathbf{H}_t^{(l)}\mathbf{W}_m^{(l)}\right)$
Context-Dependent Topology
$\widetilde{\mathbf{A}}^{(l)} \odot \mathrm{SA}(\mathbf{H}_t^{(l)})$
Mutli-Head Self-Attention
$\mathrm{SA}(\mathbf{H}_t^{(l)})$
Input $\mathbf{H}_t^{(l)}$
Shared-Topology $\widetilde{\mathbf{A}}^{(l)}$
$\hat{\mathbf{y}}$
Softmax
Linear
$r(\mathbf{z} \mid y = i)$
$p(\mathbf{z} \mid y = i)$
$r(\mathbf{z} \mid y = k)$
$r(\mathbf{z} \mid y = j)$
$\mathbf{z}$
ε ~$N(\mathbf{0}, I)$
Σ
μ
Global Avg. Pooling
× L
Add & Norm.
MS-TC
SA-GC
PE
Linear
$\mathbf{X}$
Encoder
Classifier
Embedding block
Encoding block

Figure 2.13: InfoGCN Architecture [7]

The whole contribution of Chi et. al can be summarized as:[7]

- A novel learning objective based on information bottleneck [7]

- A self-attention based graph-convolution module while spatially modeling a skeleton that extracts the context-dependent intrinsic topology [7]
- A multi-modal representation of a skeleton to improve action recognition [7]
- State-of-the-art performance on three benchmark skeleton datasets [7]:

    1. NTU RGB+D 60 [30]
    2. NTU RGB+D 120 [31]
    3. NWUCLA [32]

Chapter 3

Methodoloy

The focus of this study is on the Double-feature Double-motion Network (DD-Net), which is a deep learning architecture that has been developed with the aim of improving human action recognition performance. And we use this architecture to detect anomaly using joint position. The objective of this study is to improve the accuracy of the model by enhancing the discriminative power of the model through the effective capture of both spatial and temporal information from human joint sequences.

The Double-feature Network (DFN) and the Double-motion Network (DMN) are DD-Net's two basic building blocks.

1. **Double-feature Network (DFN)**: The DFN specializes in processing spatial information. Convolutional neural networks (CNNs) are used to extract visual information from a stack of successive frames from a video stream. By separating the feature maps into areas of various sizes and pooling them individually, the spatial pyramid pooling method used by the DFN gathers multi-scale information. Because of this, the network can record both local information and a wider context, producing complex spatial representations.

2. **Double-motion Network (DMN)**: The DMN is in charge of accumulating temporal data. It accepts optical flow data, which describes the motion between successive frames, as

input. The DMN uses CNNs to extract motion characteristics in a manner similar to the DFN. The model may take into account various temporal scales since it also makes use of spatial pyramid pooling to extract multi-scale temporal information from the motion data.

To classify anomaly, the DFN and DMN outputs are merged and fed into fully linked layers. For anomaly detection tasks, the integration of spatial and temporal networks in DD-Net enables the extraction of complementing information from both appearance and motion signals.

The capacity of DD-Net to concurrently collect and use spatial and temporal information is what gives it its efficacy. The design can capture fine-grained spatial features and temporal dynamics in video sequences by integrating the DFN and DMN. This integration aids the model in developing discriminative representations that include both visual and motion signals, which enhances its performance in detecting anomalies.

In addition, DD-Net can benefit from transfer learning by pre-training networks on massive datasets for related tasks, such as image classification or optical flow estimation. This pre-training enables the model to utilize general knowledge about visual patterns and motion dynamics, which can then be refined using labeled data for anomaly detection tasks.

Nevertheless, the efficacy of DD-Net or any other architecture depends on a number of variables, including the quality and diversity of the training data, the architectural design choices, the optimization and regularization techniques employed, and the specific requirements and characteristics of the anomaly actions in the video. Training and evaluation are required to ensure that the model's performance is optimal for the intended application.

Architecture:

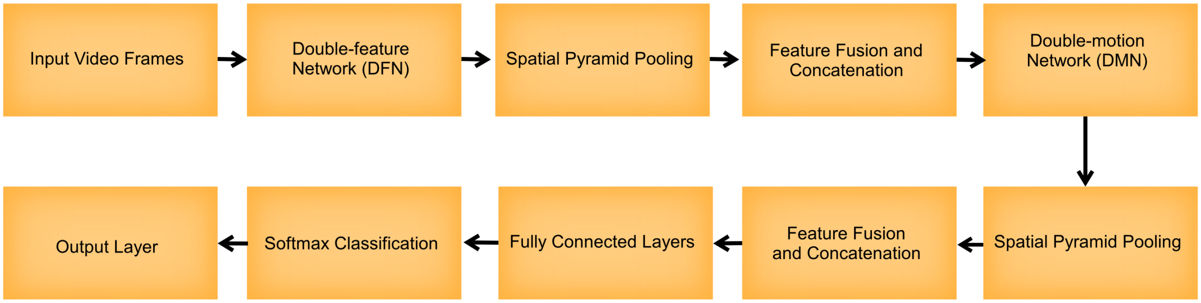

Figure 3.1: Generalized Architecture of DD-Net

Input in this architecture consists of successive video frames depicting human action joints. Both the Double-feature Network (DFN) and the Double-motion Network (DMN) analyze the frames in parallel.

The DFN extracts visual characteristics from the input frames in order to capture spatial information. The DFN-obtained feature maps are then input into a spatial pyramid pooling layer, which divides the feature maps into regions of varying proportions and conducts pooling independently for each region. The combined features capture information at multiple scales.

Similarly, the DMN accepts as input optical flow data representing the motion between consecutive frames. The DMN extracts motion characteristics from the optical flow and transmits them to a spatial pyramid pooling layer, which captures multiscale temporal information.

The outputs of the spatial pyramid aggregating layers of both DFN and DMN are combined and concatenated to capture both spatial and temporal data. This amalgamation assists in combining the complementary signals from appearance and motion.

The fused features are then passed through layers with complete connectivity, which can learn representations at a high level. Finally, a classification layer is employed for predicting anomaly.

This is a comprehensive illustration of the DD-Net architecture. Depending on the specifics of the network, such as the number of layers, varieties of convolutional or pooling operations, and precise configuration of fully connected layers, the implementation may vary.

Data Preprocessing: So for using DD-net, we first preprocessed the data. Here is the description of data preprocessing:

(a) zoom(p, $target\_l = 64$, $joints\_num = 25$, $joints\_dim = 3$): Resizes a sequence of joint positions to a specified length. It iterates over the joints and dimensions, applies a median filter to normalize the joint coordinates, and resizes the sequence using interpolation. The coordinates of the resized joints are returned.

(b) $norm\_scale(x)$: This function normalizes the array x passed as input. It calculates the array's mean and subtracts it from each element in x. Then, each element is divided by the mean, effectively normalizing the array's scale. Returns the normalized array.

(c) $get\_CG(p, C)$: This function computes the Joint Collection Distances (JCD) attribute from a sequence of joint positions p. It computes the Euclidean distances between each frame's joint pairings. It extracts the upper triangular portion (excluding the diagonal) of the distance matrix and applies the norm_scale function to normalize the distances. The returned value is the normalized JCD feature matrix.

(d) $data\_generator(T, C, le)$: This function generates the dataset required for training or testing. In addition to configuration parameters C and a label encoder le, it requires a dictionary T containing pose data, labels, and other information, as well as a label encoder le. It iterates over the pose data and preprocesses them. It applies the zoom function to each sample to resize the joint positions, thereby obtaining the normalized joint positions. The JCD feature matrix is then computed utilizing the get_CG function. The identifiers are encoded with a one-hot algorithm. The function returns the generated features X_0 (the JCD features) and X_1 (the normalized joint positions) as well as their corresponding designations Y.

In conclusion, these functions provide the functionality essential for preprocessing and generating the input data required for a joint-position based anomaly detection model.

Code: In our Double-feature Double-motion Network (DD-Net) architecture implementation, we build it to work with anomaly detection. Here is a description of the model:

(a) '$poses\_diff(x)$' function:

- This function accepts a tensor 'x' representing human joint positions as an input parameter. Motion information is captured by calculating the differences between successive frames of 'x'. The result is then resized to match 'x's' original proportions.

(b) '$pose\_motion(P, frame\_l)$' function:

- This function accepts 'P' (representing human joint positions) and 'frame_l' (the number of frames in the input sequence) as inputs. The 'poses_diff' function computes the motion differences between consecutive frames for both sluggish and rapid motion. Slow motion differences are reshaped to match the frame length, while rapid motion differences are downsampled to half the frame length. The resultant discrepancies between sluggish and rapid motion are returned.

(c) ‘$c1D(x, filters, kernel)$‘ function:

  - This function executes a 1D convolution on the tensor ’x’ that it receives as input. Batch normalization and LeakyReLU activation are applied. The output is the tensor after processing.

(d) ‘$block(x, filters)$‘ function:

  - This function represents a network construction element. The input tensor ’x’ undergoes two consecutive ’c1D’ operations with the specified number of filters.

(e) ‘$d1D(x, filters)$‘ function:

  - This function applies a dense layer (layer with all edges connected) to the tensor ’x’. Batch normalization and LeakyReLU activation are included.

(f) ‘$build\_FM(frame\_l, joint\_n, joint\_d, feat\_d, filters)$‘ function:

  - This function constructs the Feature Module (FM) for feature extraction in DD-Net. It accepts the frame length (’frame_l’), number of joints (’joint_n’), joint dimensions (’joint_d’), feature dimensions (’feat_d’), and number of filters (’filters’) as input parameters. Using the defined ’c1D’, ’block’, and pooling operations, the function constructs the FM. FM paradigm is the outcome.

(g) ‘$build\_DD\_Net(C)$‘ function:

  - This function constructs the entire DD-Net model by combining the FM and additional layers. It requires a configuration object ’C’ containing various parameters such as frame length, joint dimensions, filter count, etc. The ’build_FM’ function is used to instantiate the FM model. The motion (’M’) and joint position (’P’) input tensors are transmitted through the FM model. For classification, the output of the FM is further processed using dense layers and dropout. The self-supervised portion of the model is also provided but is not explicated explicitly in the provided code. The final model is returned, including the self-supervised and classification components.

Overall, the code represents the development of the DD-Net architecture, which consists of the FM for feature extraction and additional classification layers.

In our study, we conducted experiments to determine the effect of varying the dropout rate, the patience value, and the learning rate (LR) in the DD-Net model for joint-based anomaly detection. Initial model parameters included an dropout rate of 0.5, a patience value of 6, and a learning rate of 1e-4. However, after modifying these hyperparameters, namely by decreasing the dropout rate to 0.2, the patience value to 4, and the LR to 1e-2, we observed enhanced results.

The reduction of the dropout rate from 0.5 to 0.2 led to a decrease in regularization during training. Dropout is a regularization technique that assigns a random fraction of input units to zero during each training phase to prevent overfitting. By reducing the dropout rate, we enabled the model to retain more information during training, which may have resulted in enhanced data-driven learning and the identification of crucial characteristics.

Additionally, the reduction of the patience value from 6 to 4 altered the early ceasing mechanism during training. The patience value determines the number of epochs to wait before contemplating a non-improving validation loss as an indication to terminate training. By decreasing the value of forbearance, we permitted the model to terminate earlier if the validation loss did not improve substantially within a shortened time period. This modification prevented the model from overfitting or becoming trapped in suboptimal solutions, which could have resulted in enhanced generalization and performance.

In addition, the reduction of the learning rate from 1e-4 to 1e-2 had an effect on the model's optimization procedure. A delayed learning rate results in fewer modifications to the model's weights during training, resulting in a slowed convergence but potentially enhanced accuracy. By decreasing the learning rate, the model was able to make more refined weight updates, allowing it to better navigate the optimization landscape and potentially locate better minima.

The efficacy of the DD-Net model was enhanced as a consequence of these modifications. Reducing the dropout rate increased the model's ability to learn and capture pertinent features, whereas reducing the patience value and learning rate improved optimization and prevented overfitting. By adjusting these hyperparameters, we improved the performance of joint position-based anomaly detection.

Chapter 4

Experiments

On the basis of our initial hypothesis, we have carried out a number of early experiments. Our initial plan included overcoming the drawbacks of existing anomaly detection architectures and improving the result of the detection. Details of the experiments are discussed below:

4.1 Datasets

There are a number of datasets available for training and evaluating anomaly detection models on video data. These datasets often contain a variety of anomalous events, such as accidents, thefts, and other unusual occurrences. The specific classes of anomalies included in a dataset can vary, depending on the dataset and the purpose for which it was created. Some examples of those datasets are given here.

4.2 CUHK Avenue

This dataset has 15 sequences, and each sequence is about 2 minutes long. It contains 16 training and 21 testing video clips. The videos are captured in CUHK Campus Avenue and contain 30652 frames in total. [9]

There are 14 unusual events including:

- Running
- Throwing Objects
- Loitering

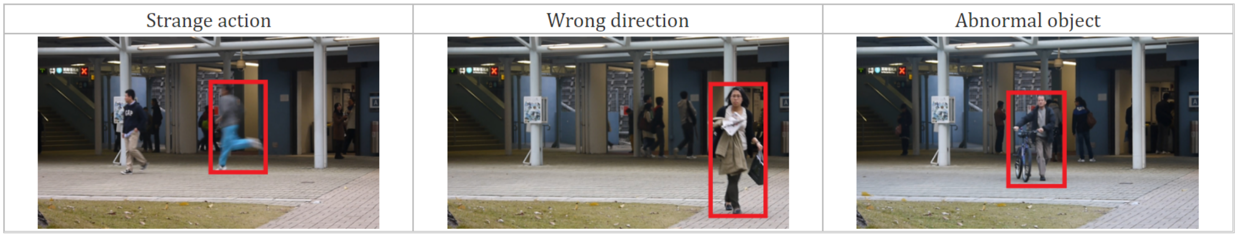

Figure 4.1: Sample of the videos in CUHK Avenue dataset

4.3 ShanghaiTech

This is a large-scale crowd-counting dataset. It only contains images. The images are annotated and the no. of images is 1198 containing a total of 330,165 people with the center of their heads annotated.[33]

The dataset is divided into two parts: Part_A and Part_B. Part_A contains 482 images, and these images are randomly parsed from the internet. Part_B contains 716 images, and these were captured in Shanghai's crowded city streets. [33] The crowd density greatly differs between two of the subsets.

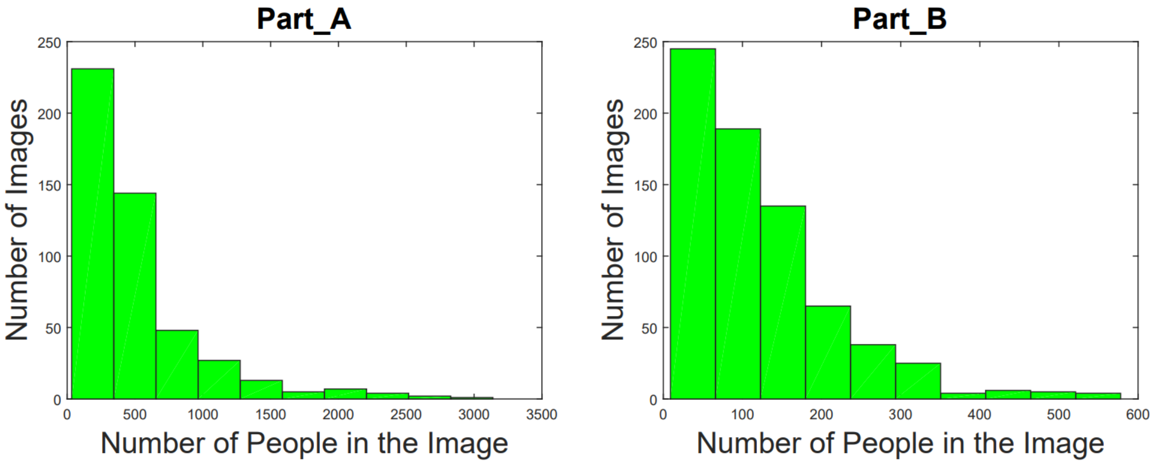

Figure 4.2: Histograms of crowd counts in ShanghaiTech Dataset

4.4 UCF Crime Dataset

This is a large-scale dataset that contains 128 hours of videos [**?**]. The videos are unmodified and untrimmed and collected from real-world surveillance systems. Some of the videos have multiple anomalies. The videos cover 13 real-world anomalies[**?**]:

Anomaly	**No. of Videos**
Abuse	50
Arrest	50
Arson	50
Assault	50
Burglary	100
Explosion	50
Fighting	50
Road Accidents	150
Robbery	150
Shooting	50
Shoplifting	50
Stealing	100
Vandalism	50
Normal Events	950

Table 4.1: Total number of videos of each anomaly in UCF Crime Dataset

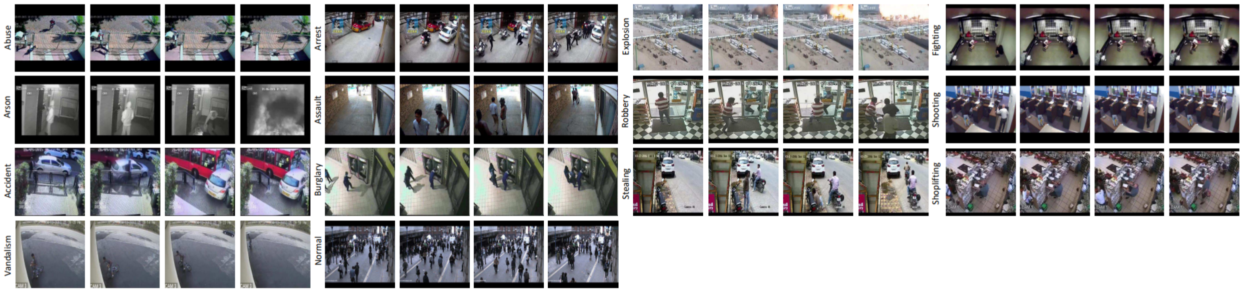

Figure 4.3: Example of different types of anomaly snippets in UFC Crime Dataset

4.5 XD-Violence

This dataset has a total of 4754 videos. It contains 2405 violent videos and 2349 non-violent videos [34]. The sources of these videos are:

- Live Scenes captured by CCTV cameras
- Hand-Held cameras
- Car-driving recorders
- Movies
- News etc.

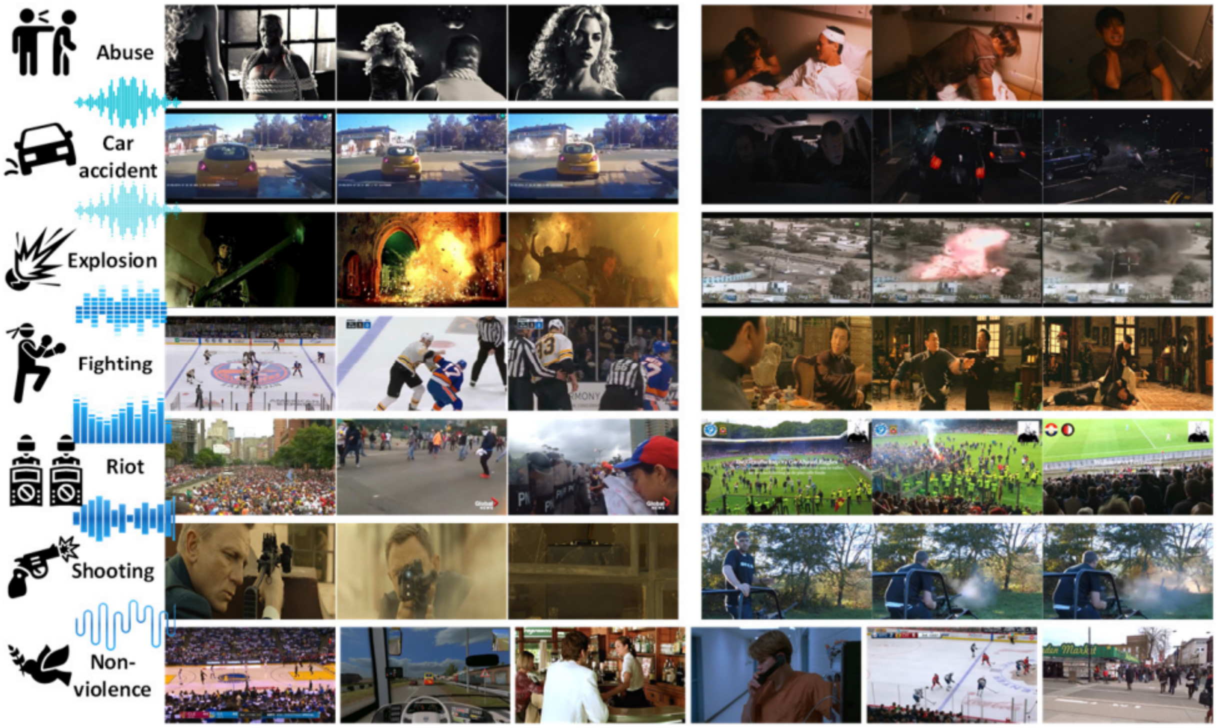

Figure 4.4: Sample videos from the XD-Violence Dataset

The classification of anomalous activities that this dataset covers are:[34]

- Fighting

- Shooting
- Car Accident
- Explosion
- Riot

The videos in this dataset also contain audio signals, which can be helpful for algorithms to leverage multi-modal information and gain more confidence.

4.6 SHREC dataset

The dataset comprises 2800 sequences of hand gestures performed by 28 right-handed subjects. These gestures are made up of fourteen separate motions that may be done with one finger or the complete hand. The gesture, finger count, performer, and trial are all assigned to each sequence. The sequence frames include depth pictures as well as the coordinates of 22 joints in both 2D depth image space and 3D world space, resulting in a full hand skeleton.

The dataset was captured using the Intel RealSense short-range depth camera, which captured depth pictures and hand skeletons at 30 frames per second with a resolution of 640x480 pixels. The duration of the example motions ranges from 20 to 50 frames.

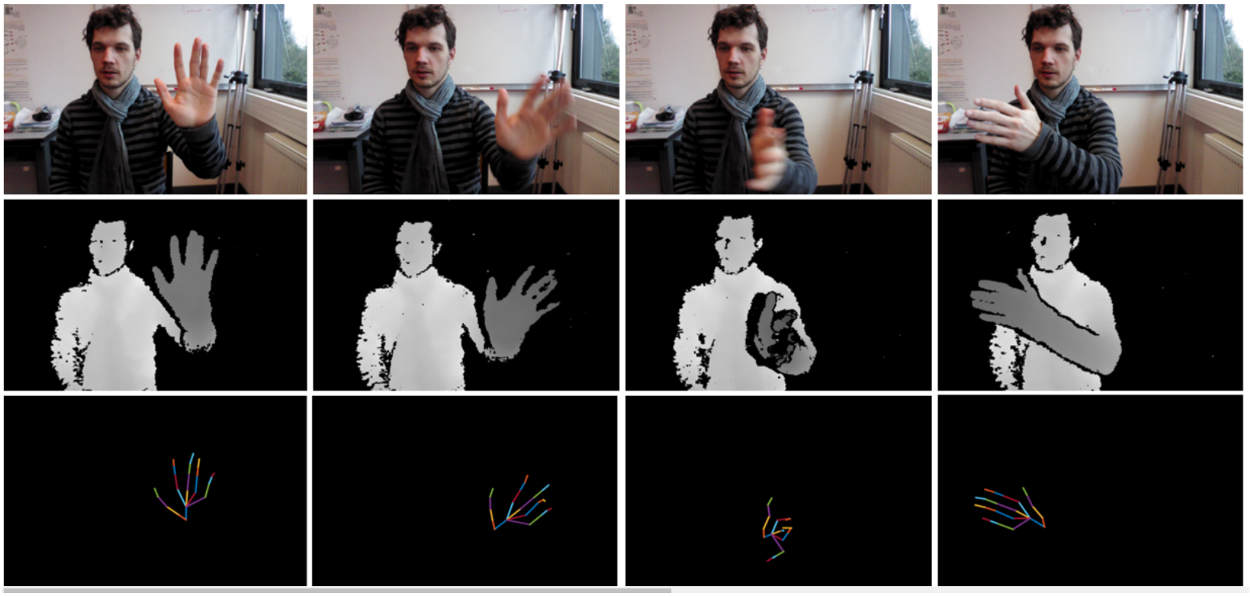

Figure 4.5: Illustration showcasing a swipe left gesture depicted in color (top), depth map representation (middle), and skeletal data visualization (bottom).

The bulk of these movements were selected to be cutting-edge. The authors did, however, eliminate the difference between conventional and scroll swipes since it was handled by their technique based on the amount of digits used. The pinch-and-expand and open-and-close movements were treated in the same way.

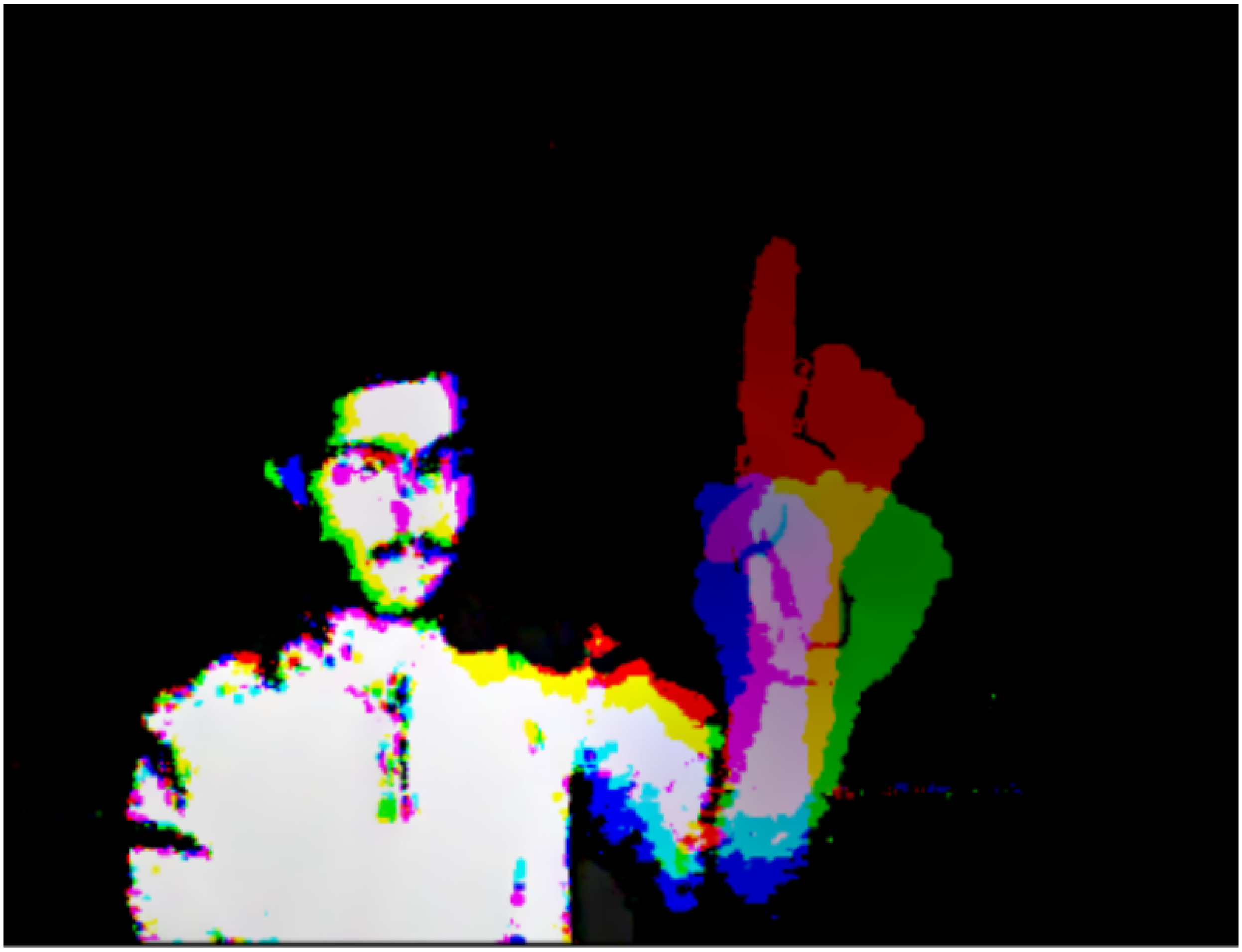

Figure 4.6: Concatenation of three keyframes illustrating a grab gesture.

Furthermore, the Grab gesture was added due to its value in augmented reality applications as well as the scientific problems posed by the possibility of performance variation. They also added the Shake gesture, which may be understood as a repeat of opposing swipe movements, to allow identification engines to discriminate between gestures composed of other gestures.

Gesture	**Label**	**Tag name**
Grab	Fine	G
Expand	Fine	E
Pinch	Fine	P
Rotation CW	Fine	R-CW
Rotation CCW	Fine	R-CCW
Tap	Coarse	T
Swipe Right	Coarse	S-R
Swipe Left	Coarse	S-L
Swipe Up	Coarse	S-U
Swipe Down	Coarse	S-D
Swipe X	Coarse	S-X
Swipe V	Coarse	S-V
Swipe +	Coarse	S-+
Shake	Coarse	Sh

Table 4.2: List of Gestures

4.7 JHMDB dataset

The HMDB51 database is a large collection of over 5,100 video segments depicting 51 distinct human actions from films and the Internet. Due to the impracticality of annotating the entire dataset, the J-HMDB subset, which focuses on fewer categories, was created. Excluded were categories predominantly involving facial expressions, interactions with others, and actions specific to particular situations.

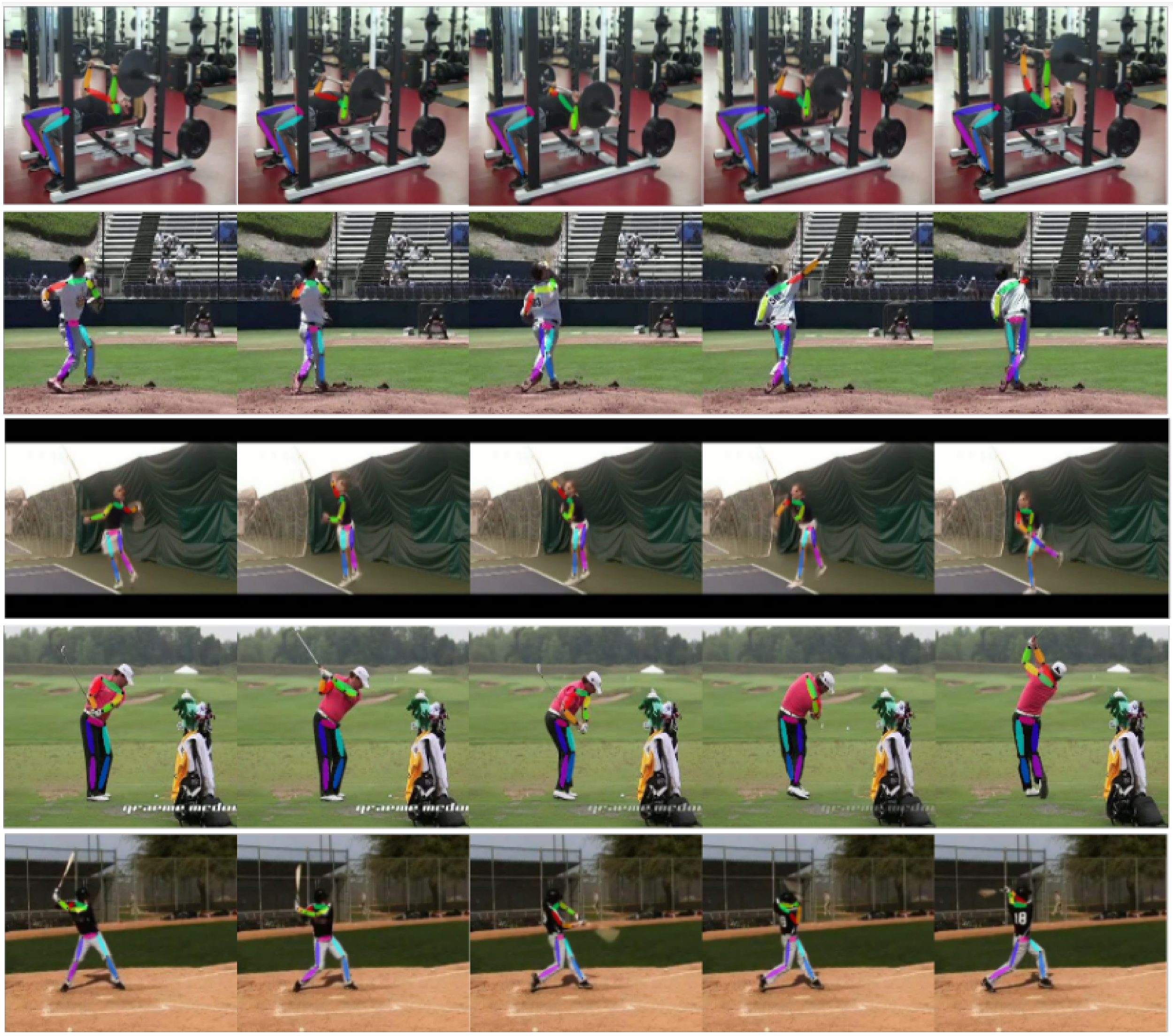

Figure 4.7: Examples of some classes

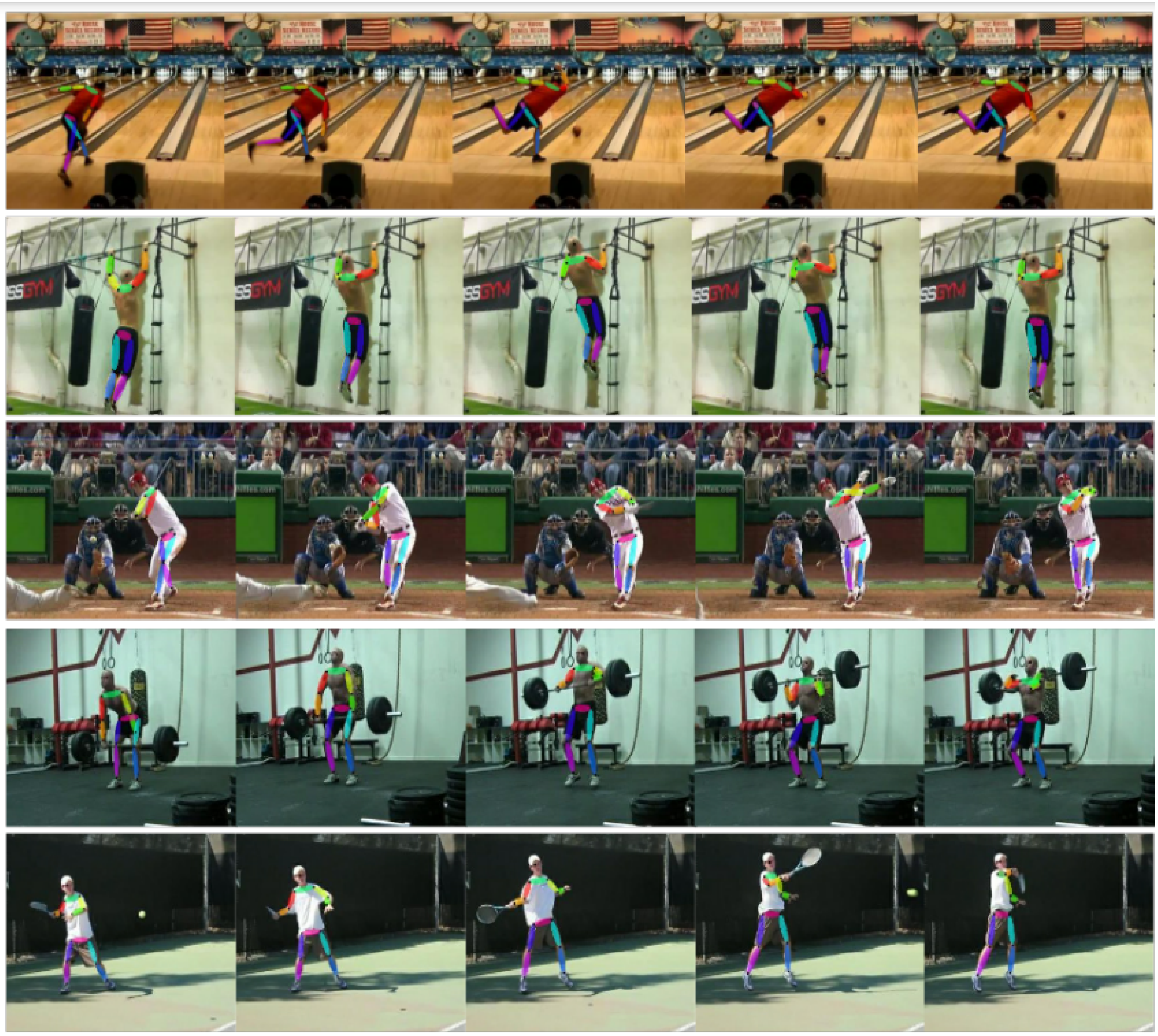

Figure 4.8: Examples of some classes

The resultant subset of the J-HMDB contains 21 categories of single-person actions, including catching, jumping, shooting a ball, pouring, swinging a baseball bat, throwing, sitting, running, kicking a ball, standing, brushing hair, climbing stairs, shooting a gun, clapping, picking, waving, push-ups, walking, shooting a bow, golfing. To ensure that the focus is on the individual conducting the action, segments in which the actor is obscured have been eliminated.

In addition, the remaining segments were time-stripped to capture the initial and final frames corresponding to the beginning and end of each action. As a result of this selection and cleansing procedure, there are now 36-55 segments per action class, each comprising 15-40 frames. The dataset contains 31,838 annotated frames in total. You can access the J-HMDB dataset at http://jhmdb.is.tu.mpg.de.

4.8 Reproducing Real-world Anomaly Detection in Surveillance Videos [1]

We tried to reproduce the work of Sultani et al. [**?**]. The AUC values are:

Anomaly	**From Paper[1]**	**DenseNet121**	**ResNet152v2**
Abuse	0.57	0.52	0.57
Arrest	0.48	0.47	0.47
Arson	0.85	0.87	0.87
Assault	0.75	0.76	0.73
Burglary	0.76	0.71	0.77
Fighting	0.40	0.43	0.44
Normal	075	0.76	0.77
Shooting	0.60	0.65	0.63
Shoplifting	0.52	0.56	0.6
Stealing	0.60	0.66	0.59

Table 4.3: Reproducing Real-world Anomaly Detection in Surveillance Videos[1]

The paper originally used DenseNet121. Our reproduction using the DenseNet121 also produced similar results. Although using ResNet152v2 has given better AUC values in some of the classifications i.e. for arson, fighting, shoplifting etc.

4.9 Reproducing InfoGCN [7]

We have also reproduced InfoGCN [7] and got an accuracy of 90.30%. The actual result is 93.6%.

As we have got 90.30% in action recognition, which suggests a very good action recognition approach, we can expect it to perform better as well in anomaly detection as well.

Chapter 5

Result Analysis

5.1 Result of Real World Anomaly Detection [1]

As we have reproduced the result of Real World Anomaly Detection [1]:

Anomaly	**From Paper[1]**	**DenseNet121**	**ResNet152v2**
Abuse	0.57	0.52	0.57
Arrest	0.48	0.47	0.47
Arson	0.85	0.87	0.87
Assault	0.75	0.76	0.73
Burglary	0.76	0.71	0.77
Fighting	0.40	0.43	0.44
Normal	075	0.76	0.77
Shooting	0.60	0.65	0.63
Shoplifting	0.52	0.56	0.6
Stealing	0.60	0.66	0.59

Table 5.1: Reproducing Real-world Anomaly Detection in Surveillance Videos, AUC values [1]

The paper originally used DenseNet121. We have used ResNet152v2 and initially got better results. For example, in Arson, the AUC value got increased by a bit. For fighting and shooting, there is some improvement and a significant increment in AUC values are being seen for Shoplifting. These improvements can be due to the usage of a different model.

5.2 Result of InfoGCN [7]

Traditionally, a graph convolution network is used to recognize human actions. InfoGCN [7] is a state-of-the-art algorithm for detecting human activities. We are proposing to use the encoder of InfoGCN to extract the features and this feature will be used to detect anomalies.

We tried to reproduce the result of InfoGCN [7]:

Dataset	**Accuracy**	**Reproduced Accuracy**
NTU RGB+D 60	93%	90.1%
NTU RGB+D 120	89.7%	87.3%
NW-UCLA	97%	94.7%

Table 5.2: Reproducing InfoGCN [7]

5.3 Reproducing Make Skeleton-based Action Recognition Model Smaller, Faster and Better [6]

5.3.1 Initial Approach

In our research, we sought to replicate the architecture proposed in Yang et al.'s "Reproducing Make Skeleton-based Action Recognition Model Smaller, Faster, and Better" [6] paper. This paper presented a method for enhancing the performance of skeleton-based action recognition models by reducing their size, increasing their speed, and obtaining greater precision. By recreating their architecture and methodology, we aimed to validate their findings and contribute to the comprehension of effective action recognition models.

During our endeavors to reproduce, we discovered a significant disparity between the training accuracy and the test accuracy. Despite the paper's remarkable results, we discovered that our training accuracy was consistently 10-15% lower than the corresponding test accuracy. This discrepancy indicates that our replicated model had difficulty generalizing to unobserved data, implying possible overfitting during training. When a model becomes overly specialized in capturing the training data's patterns and fails to generalize to new instances, this is known as overfitting.

Moreover, despite our best efforts to replicate the architecture and methodology specified in the paper, we found that the overall accuracy of our replicated model was significantly lower than the reported results. Different implementation details, hyperparameter settings, or data preprocessing procedures may have contributed to this discrepancy. Reproducing research findings can be difficult due to the subtle details and nuances that are frequently omitted from research publications.

Dataset	**Accuracy**	**Reproduced Accuracy**
JHMDB	82.5%	74.1%
SHREC	94.7%	82.7%

Table 5.3: Result of reproducing Make Skeleton-based Action Recognition Model Smaller, Faster and Better [6]

5.3.2 Changing Parameters

In the DD-Net model for joint-based anomaly detection, we experimented with the dropout rate, the tolerance value, and the learning rate (LR). The initial parameters of the model were 0.5 dropout, 6 tolerance, and 1e-4 learning. However, outcomes were enhanced by decreasing the attrition rate to 0.2, the perseverance value to 4, and the likelihood ratio to 1e-2.

The regularization of training decreased as the attrition rate decreased from 0.5 to 0.2 percent. By setting a random percentage of input units to zero during training, dropout prevents overfitting. We decreased the attrition rate to assist the model in retaining more information during training, which may have enhanced data-driven learning and helped identify key characteristics.

Reducing patience from 6 to 4 also had an impact on the early training cease mechanism.

The patience parameter regulates the number of epochs to wait for a non-improving validation loss before terminating training. If the validation loss did not improve within a shortened period of time, we permitted the model to complete early by decreasing forbearance. This modification prevented the model from becoming overfit or trapped in subpar solutions, thereby enhancing generalization and performance.

The decrease in the learning rate from 1e-4 to 1e-2 impacted the model's optimization strategy. Convergence is slowed by delaying learning, but accuracy may be improved by reducing model weight changes throughout training. By reducing the learning rate, the model may be able to enhance weight updates, traverse the optimization terrain more effectively, and discover better minima.

These modifications enhanced the DD-Net model. Reducing the attrition rate, patience value, and learning rate facilitated model learning and reduced overfitting. These hyperparameters improved the joint position-based detection of anomalies.

Description	**Prev Value**	**New Value**
Dropout	0.5	0.2
Patience Value	6	4
LR	1e-4	1e-2

Table 5.4: Changes of the parameter values

After testing with the parameter values, our experiment result got improved while testing with both the JHMDB and SHREC datasets.

Dataset	**Accuracy Before Change**	**Accuracy After Change**
JHMDB	74.1%	81.1%
SHREC	82.7%	87.3%

Table 5.5: Improvements in the result after tuning the parameters

5.4 Result Comparison

The aim of this study was to evaluate the accuracy of different anomaly detection models and compare their performance. The JHMDB dataset was utilized for the purpose of conducting a comparison, as it is a widely recognized benchmark dataset that is frequently employed for tasks related to action recognition and anomaly detection. The objective of this study was to determine the optimal method for detecting anomalies in video data by assessing the precision of various models on the given dataset.

The study evaluated various models and found that ResNet50 achieved the highest accuracy, with a notable rate of 90.14%. The excellent performance of ResNet50 in image and video-related tasks has made it a widely adopted deep convolutional neural network architecture. ResNet50 demonstrates its effectiveness in distinguishing abnormal events or patterns from normal ones within the JHMDB dataset, as evidenced by its high accuracy.

The accuracy rate of 85.32% was achieved by InceptionV3, which is a well-known deep learning model, as observed in our comparison. The performance of InceptionV3 in detecting anomalies within the JHMDB dataset was found to be satisfactory, albeit slightly lower than that of ResNet50. The effectiveness and simplicity of VGG19, a deep convolutional neural network, is demonstrated by its achievement of an accuracy rate of 87.70%. The suitability of InceptionV3 and VGG19 for anomaly detection tasks is demonstrated by the results, despite their slightly lower accuracy compared to ResNet50.

Incorporating DD-Net, as suggested by Yang et al. [6], as the final step in our methodology has been done to improve the precision of our anomaly detection approach. The purpose of this text is to present DD-Net, a model architecture that has been specifically developed for the purpose of detecting anomalies in videos. The integration of DD-Net into our framework resulted in an accuracy of 88.61%, which is on par with the top accuracy achieved by the previous three models. The effectiveness of DD-Net in capturing and detecting anomalies within the JHMDB dataset is demonstrated.

The variability in training time among different models was a noteworthy aspect observed during the experiments. The study found that DD-Net achieved comparable accuracy rates to ResNet50, InceptionV3, and VGG19, while requiring significantly less training time. The DD-Net's architectural design and computational efficiency make it a compelling choice for anomaly detection tasks, especially when handling extensive video datasets, resulting in a reduction in training time.

Models	**Accuracy (%)**	**Training Time (For 100 epoch)**
ResNet50	90.14%	12.5 min
InceptionV3	85.32%	11 min
VGG19	87.70%	14 min
DD-Net	**88.61%**	**9 min**

Table 5.6: Result comparison of different models

Using the JHMDB dataset, we compared the precision of various models for anomaly detection. The results demonstrated ResNet50's superiority while also highlighting the efficacy of InceptionV3, VGG19, and DD-Net. These results contribute to the comprehension of model selection for anomaly detection in computer vision by shedding light on accuracy and training time considerations.

Chapter 6

Conclusion

Joint position-based information improves anomaly detection efficiency and accuracy in our thesis. CNN pixel-by-pixel analysis for anomaly detection is computationally intensive and prone to background interference. We suggest using combined position data to minimize computation and improve anomaly detection.

We study anomalies to respond quickly. We intend to accelerate anomaly identification and limit their effects by switching from pixel-level analysis to joint position-based analysis. In emergencies like shootings, anomaly detection must be fast and precise.

We struggled without combined position-based anomaly detection datasets. To address this obstacle, we changed existing datasets by combining and analyzing joint position information. This dataset update makes combined position-based anomaly detection resources available for study.

We addressed resource constraints associated with YouTube-scale anomaly detection datasets. We provide a more realistic assessment of our technique by altering the dataset for joint position-based detection.

We adjusted dropout rates, patience, and learning rates in the Double-feature Double-motion Network (DD-Net) model in our study. These changes improved the model's joint position-based anomaly detection accuracy and performance.

We ran extensive trials to test our combined position-based anomaly detection system. We showed that combined position information improves anomaly detection performance, efficiency, and potential.

Our thesis improves anomaly detection by using joint position-based information. We improve detection efficiency and accuracy to respond faster to aberrant occurrences and deliver insights for real-world applications. Our results provide the groundwork for combined position-based anomaly detection research and implementation.

Joint position anomaly detection has several benefits. First, it requires far less computing than pixel-level analysis. Traditional approaches process each frame's pixels, which is computationally costly. The calculation is simplified by concentrating on human joint locations, speeding anomaly detection. This advantage is critical when fast reaction times are needed to reduce aberrant situations.

Second, combined position-based anomaly detection removes noise from analysis. Traditional approaches might be hampered by background noise and clutter, making abnormalities hard to spot. Directly analyzing body component locations and motions makes identification more robust and less sensitive to background fluctuations. This enhances anomaly detection.

Joint position-based anomaly detection is also human-centric. Body component abnormalities frequently cause human behavior anomalies. The detection procedure captures human behavior by concentrating on joint positions, making suspicious or anomalous activity easier to understand. This human-centric method improves anomaly identification in public safety, healthcare monitoring, and human-computer interaction.

Joint position-based anomaly detection is useful in real-world situations that need immediate action. To avoid attacks, security surveillance must quickly discover anomalies. The approach efficiently detects abnormalities in real time by evaluating joint locations, which indicate human activities. In emergency scenarios, joint position-based detection helps identify aberrant motions or postures for quick reaction and intervention.

Finally, combined position-based detection yields interpretable findings. Specific joint locations or atypical motion patterns explain the abnormalities, offering actionable information for future research. This interpretability helps spot irregularities and make decisions quickly.

Joint position anomaly detection enables decreased processing, background independence, human-centric analysis, real-world application, and interpretable findings. These benefits increase anomaly detection, enable rapid replies, and improve anomaly detection systems in numerous domains and applications.

Joint position-based anomaly detection enables new research avenues. Two crucial sectors have untapped potential.

First, combined position-based anomaly detection-specific deep learning architectures may be improved. Our thesis improved the DD-Net design to include joint position information, but more advanced network topologies are possible. Attention methods may help the network concentrate on key joint locations and motions, enhancing anomaly detection accuracy and efficiency. Recurrent neural networks (RNNs) and graph neural networks (GNNs) can record temporal and spatial connections among joint locations to better analyze human motion patterns. Investigating these sophisticated designs may enhance anomaly detection.

Second, real-world deployment and assessment of combined position-based anomaly detection systems are key research areas. Our thesis involves tests and comparisons with standard anomaly detection approaches, but implementing the system in real-world circumstances may reveal its efficacy and viability. Field testing in surveillance, smart, or industrial settings may assess the system's real-time and complicated performance. User studies with domain experts may assess system usability and identify issues. This empirical review will refine the system and prove its practicality, enabling its real-world adoption.

Addressing future scopes is beneficial. Deep learning architectures for combined position-based anomaly detection will improve accuracy, robustness, and efficiency. Complex models may reveal novel ways to use joint position information. Real-world deployment and assessment will bridge the gap between research and practical application, testing and validating the system in actual scenarios. This will inform future advances and allow combined position-based anomaly detection in surveillance, robotics, healthcare, and human-computer interaction.

Finally, collaborative position-based anomaly detection research should concentrate on deep learning architectures and real-world assessments. These efforts will improve anomaly detection systems and promote collaborative position-based techniques in practice.

REFERENCES

[1] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

[2] M. Qasim and E. Verdu, "Video anomaly detection system using deep convolutional and recurrent models," *Results in Engineering*, vol. 18, p. 101026, 2023.

[3] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "Mist: Multiple instance self-training framework for video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14009–14018, 2021.

[4] N.-C. Ristea, N. Madan, R. T. Ionescu, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah, "Self-supervised predictive convolutional attentive block for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13576–13586, 2022.

[5] J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, "Self-supervised sparse representation for video anomaly detection," in *European Conference on Computer Vision*, pp. 729–745, Springer, 2022.

[6] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proceedings of the ACM multimedia asia*, pp. 1–6, 2019.

[7] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, “Infogcn: Representation learning for human skeleton-based action recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20186–20196, 2022.

[8] B. Zhao, L. Fei-Fei, and E. P. Xing, “Online detection of unusual events in videos via dynamic sparse coding,” in *CVPR 2011*, pp. 3313–3320, 2011.

[9] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 fps in matlab,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2720–2727, 2013.

[10] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, “Learning temporal regularity in video sequences,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742, 2016.

[11] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked rnn framework,” in *Proceedings of the IEEE international conference on computer vision*, pp. 341–349, 2017.

[12] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection–a new baseline,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

[13] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, “Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, 2019.

[14] M.-R. Amini and P. Gallinari, “Semi-supervised logistic regression,” in *ECAI*, vol. 2, p. 11, 2002.

[15] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, p. 896, 2013.

[16] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 3218–3226, 2015.

[17] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[18] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[19] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1110–1118, 2015.

[20] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "Rnn fisher vectors for action recognition and image annotation," in *European Conference on Computer Vision*, pp. 833–850, Springer, 2016.

[21] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1647–1656, 2017.

[22] J. Zhang, L. Qing, and J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4030–4034, IEEE, 2019.

[23] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. Le Saux, and D. Filliat, "Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset," in *3DOR-10th Eurographics Workshop on 3D Object Retrieval*, pp. 1–6, 2017.

[24] G. Devineau, W. Xi, F. Moutarde, and J. Yang, “Convolutional neural networks for multivariate time series classification using both inter-and intra-channel parallel convolutions,” in *Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP'2018)*, 2018.

[25] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, “Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition,” in *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[26] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, “An attention enhanced graph convolutional lstm network for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1227–1236, 2019.

[27] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, “Channel-wise topology refinement graph convolution for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13359–13368, 2021.

[28] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, “Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1113–1122, 2021.

[29] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

[30] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “Ntu rgb+ d: A large scale dataset for 3d human activity analysis,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

[31] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, “Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[32] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, “Cross-view action modeling, learning and recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2649–2656, 2014.

[33] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-image crowd counting via multi-column convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 589–597, 2016.

[34] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, “Not only look, but also listen: Learning multimodal violence detection under weak supervision,” in *European conference on computer vision*, pp. 322–339, Springer, 2020.