Weakly and Partially Supervised Learning Frameworks for Anomaly Detection

Bruno Manuel Degardin

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Dr. Hugo Pedro Martins Carriço Proença

junho de 2020

Acknowledgments

This dissertation is dedicated to everyone involved throughout the attainment of this work, letting my deepest sincere gratitude.

Primarily, I would like to express my gratefulness to my supervisor, Doctor Hugo Proença, for his continuous guidance, support, patience, and expertise, which encouraged me so much during the whole process of this research work. My most profound thank you for teaching me the knowledge that passionated and motivated me so much, making the conclusion of this dissertation possible. I am indebted to the trust and opportunity given to me for working with such a professional and demanding person, ending up being a real lifechanging experience. My deepest gratitude for your confidence and the possibility of working with you.

To the closest persons to me, I would like to thank them from the bottom of my heart. Despite my absent time and canceled plans, their absolute and unconditional support made not only the closure of this episode possible but also my entire educational and academic path, being an authentic cornerstone of my life.

Equally significant, I thank every friend of mine that supported me during this dissertation and academic journey. Especially my colleagues from SociaLab for providing an excellent working environment and where help was never denied, namely Vasco Lopes, Nuno Pereira, and Gaspar Ramôa, hopefully, to build a bright future by your side. I also want to thank my former colleagues not only for willingly following me to Lisbon and helping me capturing the test simulations for this work, but also to be always present throughout these five academic years, namely Miguel Azinheira, Marco Raposo, Nuno Aparício, Alexandre Carvalho, Francisco Nunes, Luís Falcão, Hugo Lopes, Gabriel Brito, and Fábio Ramalhos. I wish the best for each and every one of you.

Last but not least, I owe a particular thanks to TOMI WORLD and Instituto de Telecomunicações for the BIODI project's financial support, which made this research work possible. And also, to the project leader Professor João Neves for its willingness of cooperation.

Resumo

A detecção automática de eventos anómalos em imagens de videovigilância permanece uma inquietação por parte da comunidade científica. Sendo a proteção o principal propósito da instalação de sistemas de vigilância, a capacidade de monitorização da segurança pública, e a sua rápida resposta para satisfazer essa finalidade, é uma adversidade até para o ser humano. Nos dias de hoje, com o aumento do uso de sistemas de videovigilância, a capacidade humana não tem alcançado a cadência necessária, exigindo uma supervisão exorbitante para a identificação de acontecimentos invulgares que coloquem uma identidade ou sociedade em risco. O facto da probabilidade de se suceder um incidente ser extremamente reduzida comparada a eventualidades normais, existe um gasto substancial de tempo de ofício. Consequentemente, a necessidade para um algorítmo de detecção automática de incidentes tem vindo a ser crucial em videovigilância. Mesmo sendo alvo de vários trabalhos científicos publicados na última década, o desempenho do estado-da-arte continua insatisfatório e abaixo do requisitado para uma implementação eficiente deste tipo de tecnologias em ambientes e cenários totalmente espontâneos e incontinentes. Porém, apesar de toda a investigação realizada nesta área, a automatização de detecção de incidentes é um desafio que perdura por várias razões. Começando pela diversidade ambiental, a complexidade da semalhança entre movimentos de ações distintas, cenários de multidões, e ter em conta todos os padrões para definir uma ação normal, é indiscutivelmente difícil ou impossível. Não obstante a dificuldade de resolução destes problemas, o obstáculo fundamental consiste na obtenção de um número suficiente de instâncias classificadas anormais, considerando algoritmos de visão computacional é essencial. Mais importante ainda, obter um vasto conjunto de diferentes vídeos capazes de satisfazer as condições previamente mencionadas, não é uma tarefa simples. Em adição ao esforço e tempo despendido, estabelecer um limite entre ações normais e anormais é frequentemente indistinto.

Tendo estes aspetos em consideração, neste trabalho, o principal objetivo é providenciar diversas soluções para os problemas previamente mencionados, concentrando na análise de métodos do estado-da-arte e apresentando uma visão abrangente dos mesmos para clarificar os conceitos aplicados na captura de padrões normais e anormais. Inclusive, a exploração de diferentes estratégias habilitou-nos a desenvolver novas abordagens que aprimoram consistentemente o desempenho do estado-da-arte. Por último, anunciamos a disponibilidade de um novo conjunto de dados, em grande escala, totalmente anotado ao nível da frame em relação à detecção de anomalias em um evento específico com uma vasta diversidade em cenários de luta, podendo ser livremente utilizado pela comunidade científica. Neste documento, com o propósito de requerer o mínimo de supervisão, são descritas duas propostas diferentes; O primeiro método põe em prática a recente técnica de aprendizagem auto-supervisionada para evitar a árdua tarefa de anotação, onde o conjunto de treino é classificado autonomamente usando uma estrutura de aprendizagem iterativa composta por duas redes neuronais independentes que fornecem dados entre si

através de uma estrutura Bayesiana. A segunda proposta explora um novo método para aprender um modelo de classificação de anomalias no paradigma multiple-instance learning manuseando vídeos fracamente anotados, onde a classificação do conjunto de treino é feita ao nível do vídeo. As experiências foram concebidas em vários conjuntos de dados, e as nossas soluções superam consolidamente o estado-da-arte. Adicionalmente, como sistema de prova de conceito, apresentamos os resultados da execução do nosso modelo em simulações reais em diferentes ambientes.

Palavras-chave

Análise de Imagens de Vídeo, Aprendizagem Automática, Aprendizagem Auto-Supervisionada, Aprendizagem Não Supervisionada, Aprendizagem Supervisionada, Biometria, Detecção de Anomalias, Detecção de Eventos Anormais, Inteligência Artificial, Reconhecimento de Padrões, Vigilância Visual, Visão Computacional

Resumo Alargado

A detecção automática de eventos anómalos em imagens de videovigilância permanece uma inquietação por parte da comunidade científica. Sendo a proteção o principal propósito da instalação de sistemas de vigilância, a capacidade de monitorização da segurança pública, e a sua rápida resposta para satisfazer essa finalidade, é uma adversidade até para o ser humano. Nos dias de hoje, com o aumento do uso de sistemas de videovigilância, a capacidade humana não tem alcançado a cadência necessária, exigindo uma supervisão exorbitante para a identificação de acontecimentos invulgares que coloquem uma identidade ou sociedade em risco. O facto da probabilidade de se suceder um incidente ser extremamente reduzida comparada a eventualidades normais, existe um gasto substancial de tempo de ofício. Consequentemente, a necessidade para um algorítmo de detecção automática de incidentes tem vindo a ser crucial em videovigilância. Mesmo sendo alvo de vários trabalhos científicos publicados na última década, o desempenho do estado-da-arte continua insatisfatório e abaixo do requisitado para uma implementação eficiente deste tipo de tecnologias em ambientes e cenários totalmente espontâneos e incontinentes. Porém, apesar de toda a investigação realizada nesta área, a automatização de detecção de incidentes é um desafio que perdura por várias razões. Começando pela diversidade ambiental, a complexidade da semalhança entre movimentos de ações distintas, cenários de multidões, e ter em conta todos os padrões para definir uma ação normal, é indiscutivelmente difícil ou impossível. Não obstante a dificuldade de resolução destes problemas, o obstáculo fundamental consiste na obtenção de um número suficiente de instâncias classificadas anormais, considerando algoritmos de visão computacional é essencial. Mais importante ainda, obter um vasto conjunto de diferentes vídeos capazes de satisfazer as condições previamente mencionadas, não é uma tarefa simples. Em adição ao esforço e tempo despendido, estabelecer um limite entre ações normais e anormais é frequentemente indistinto.

Esta dissertação tem como principal finalidade apresentar conceitos e soluções de visão computacional em resposta ao desafio biométrico em detecção de humanos com comportamentos anormais. De forma mais inerente, pretende-se: 1) compreender como extrair características temporais de movimento e aparência em imagens de vídeo; 2) perceber como manusear essas características para diferentes tipos de aprendizagem na detecção de anomalias, com especial foco em aprendizagem fraca, incluindo uma detalhada descrição dos métodos do estado-da-arte; 3) desenvolver dois novo métodos diferentes para a detecção de anomalias, um operando num sistema de aprendizagem autosupervisionada, enquanto que o segundo opera no paradigma de aprendizagem supervisionada fraca; 4) apresentar um novo conjunto de dados público para uso científico na área de detecção de anomalias. De modo a alcançar estes objetivos, esta tese apresenta diversas contribuições pormenorizadas ao longo de sete capítulos.

O primeiro capítulo determina o âmbito e o contexto do problema em que se enquadra esta dissertação. Ademais, são descritos os principais objetivos e motivações do presente trabalho de investigação, bem como as principais contribuições da investigação realizadas nesta tese de forma a melhorar o estado-da-arte na área da detecção de anomalias. Por fim, é descrito a estrutura deste documento.

O segundo capítulo apresenta uma revisão detalhada da literatura nos domínios da investigação essencial ao desenvolvimento de um método para a detecção de eventos anormais. Este capítulo encontra-se organizado em secções com o objetivo de: 1) apresentar o estado-da-arte em extração de características espaço-temporal de imagens de vídeo; 2) explicar detalhadamente o uso dessas mesmas características de vídeos no paradigma da aprendizagem supervisionada fraca para a detecção de segmentos de anomalias, incluindo o atual método considerado estado-da-arte neste tipo de aprendizagem; e 3) expor com detalhe as diferentes propostas que abordam o problema com o tipo de aprendizagem não supervisionada. Considerando a tarefa extremamente difícil de obter um conjunto de dados robusto, que tenha a capacidade de apresentar as diversas complexidades dos movimentos anormais para que seja possível aprender um modelo eficiente em detectar diversas anomalias, é necessário ter em conta as diferentes abordagens ao problema para contornar esta dificuldade e contudo aprender um modelo que seja eficaz. O estado-da-arte atual foca-se principalemente em tentar minimizar ao máximo a supervisão deste tipo de conjunto de dados. Começando pelos metodos que aplicam aprendizagem supervisionada fraca, consistindo em providenciar vídeos, normais e anormais, ao modelo sem qualquer informação temporal e apenas dispor a informação se o modelo está a analisar um video normal, apenas contendo eventos do dia-a-dia, ou um vídeo anormal, contendo nalgum momento um evento incomum, sem que o modelo saiba exatamente quando. Finalmente, é também apresentado os diferentes métodos de aprendizagem não supervisionada que concentram-se em apenas analisar os padrões de comportamentos normais, isto é, o modelo apenas irá observar vídeos normais. Destes métodos são abordados diferentes propostas para captura de padrões normais: 1) através de um auto-encoder que aprende apenas características espaciais e temporais normais, sendo depois incapaz de reproduzir um evento anormal; 2) através de redes adversárias generativas apenas analisando o comportamento comum presente nos vídeos normais, desenvolvendo uma incapacidade de gerar eventos anormais; e 3) através da aprendizagem de um auto-encoder mediante a extração de características baseadas nos comportamentos normais do esqueleto humano, tornando-se incapaz de reproduzir um comportamento anormal do esqueleto humano. Desta leitura, foi possível concluir a incapacidade da aplicação dos métodos de aprendizagem não supervisionada em conjunto de dados com bastante heterogeneidade entre os vídeos, enquanto que os métodos de aprendizagem supervisionada, tendo a vantagem de analisar ambos os tipos de vídeo, apresentam níveis de desempenho mais elevados.

No terceiro capítulo apresentamos o primeiro método proposto neste documento, com o objetivo de detectar anomalias no paradigma da aprendizagem supervisionada fraca

e aplicando a recente técnica de aprendizagem auto-supervisionada. Este método é composto por duas redes neuronais independentes, contidas num sistema iterativo, que fornecem dados entre si através de uma estrutura Bayesiana. A base lógica desta proposta consiste no aumento do conjunto de treino de cada modelo com instâncias não classificadas. Esse aumento é gerado através do sistema iterativo que, baseado na filtração das instâncias incertas indicadas pelas estrutura Bayesiana, acrescenta apenas as instâncias negativas e positivas com maior confiabilidade, provocando assim um aumento de desempenho na iteração seguinte do modelo.

O quarto capítulo apresenta o segundo método proposto nesta dissertação, focando-se apenas no paradigma de aprendizagem supervisionada fraca na detecção de anomalias. Esta proposta aplica a técnica do modelo de mistura de Gaussianas para manipular as distribuições de pontuações do modelo consoante as características extraídas de ambos os vídeos normais e anormais. A base lógica deste método consiste nos diferentes supostos comportamentos de distribuições entre vídeos normais e anormais. Baseado nessa divergência de distribuições, e conjuntamente das aplicações de distribuições normais e mistura de Gaussianas, o objetivo é penalizar o nosso modelo caso não cumpra as regras impostas na nossa função de custo. Nesta proposta analisámos diferentes modos de abordar o problema aplicando a mistura de Gaussianas. Após o estudo matemático deste tipo de modelos, fomos capazes de desenvolver um modelo, aplicando apenas aprendizagem supervisionada fraca, e superando assim o estado-da-arte atual neste tipo de aprendizagem.

O quinto capítulo apresenta uma detalhada descrição dos conjuntos de dados em detecção de anomalias, atualmente públicos, assim como também anunciamos e apresentamos uma descrição pormenorizada de um novo conjunto de dados para detecção de anomalias, *UBI-Fights*, totalmente disponível. Com a análise prévia dos conjuntos já existentes, identificamos as características indispensáveis que não se encontravam contidas nestes conjuntos. Por conseguinte, desenvolvemos o nosso próprio conjunto de dados, de grande escala, apresentando características distintas dos existentes, e abrangendo os requisitos necessários, podendo ser livremente utilizado pela comunidade científica na área de detecção de anomalias. Este novo conjunto de dados é composto por uma vasta diversidade em cenários de luta, e encontra-se totalmente anotado a nível temporal.

O sexto capítulo consiste na apresentação dos resultados obtidos nas experiências realizadas em diversos conjuntos de dados, com o propósito de comparar os nossos métodos propostos com o atual estado-da-arte. Adicionalmente, como sistema de prova de conceito, apresentamos os resultados da execução do nosso modelo em simulações reais, em diferentes ambientes, pessoalmente recolhidas por nós através de uma parceria com a empresa de tecnologia *TOMI WORLD*. Com o objetivo de observar o comportamento do nosso modelo consoante os diferentes cenários simulados e supervisionados por nós.

Por fim, no sétimo capítulo é discutido as contribuições e conclusões deste trabalho de investigação para o desenvolvimento de um modelo de inteligência artificial para a detecção de anomalias em imagens de vídeos. Apresentando um resumo dos métodos propostos, as suas vantagens e desvantagens para a motivação de desenvolvimento futuro neste tipo de tecnologias. Por último, é feito uma perspetiva do trabalho futuro em diferentes tipos de aprendizagem na detecção automática de eventos anómalos em imagens de vídeo.

Abstract

The automatic detection of abnormal events in surveillance footage is still a concern of the research community. Since protection is the primary purpose of installing video surveillance systems, the monitoring capability to keep public safety, and its rapid response to satisfy this purpose, is a significant challenge even for humans. Nowadays, human capacity has not kept pace with the increased use of surveillance systems, requiring much supervision to identify unusual events that could put any person or company at risk, without ignoring the fact that there is a substantial waste of labor and time due to the extremely low likelihood of occurring anomalous events compared to normal ones. Consequently, the need for an automatic detection algorithm of abnormal events has become crucial in video surveillance. Even being in the scope of various research works published in the last decade, the state-of-the-art performance is still unsatisfactory and far below the required for an effective deployment of this kind of technology in fully unconstrained scenarios. Nevertheless, despite all the research done in this area, the automatic detection of abnormal events remains a challenge for many reasons. Starting by environmental diversity, the complexity of movements resemblance in different actions, crowded scenarios, and taking into account all possible standard patterns to define a normal action is undoubtedly difficult or impossible. Despite the difficulty of solving these problems, the substantive problem lies in obtaining sufficient amounts of labeled abnormal samples, which concerning computer vision algorithms, is fundamental. More importantly, obtaining an extensive set of different videos that satisfy the previously mentioned conditions is not a simple task. In addition to its effort and time-consuming, defining the boundary between normal and abnormal actions is usually unclear.

Henceforward, in this work, the main objective is to provide several solutions to the problems mentioned above, by focusing on analyzing previous state-of-the-art methods and presenting an extensive overview to clarify the concepts employed on capturing normal and abnormal patterns. Also, by exploring different strategies, we were able to develop new approaches that consistently advance the state-of-the-art performance. Moreover, we announce the availability of a new large-scale first of its kind dataset fully annotated at the frame level, concerning a specific anomaly detection event with a wide diversity in fighting scenarios, that can be freely used by the research community. Along with this document with the purpose of requiring minimal supervision, two different proposals are described; the first method employs the recent technique of self-supervised learning to avoid the laborious task of annotation, where the training set is autonomously labeled using an iterative learning framework composed of two independent experts that feed data to each other through a Bayesian framework. The second proposal explores a new method to learn an anomaly ranking model in the multiple instance learning paradigm by leveraging weakly labeled videos, where the training labels are done at the video-level. The experiments were conducted in several well-known datasets, and our solutions solidly outperform the state-of-the-art. Additionally, as a proof-of-concept system, we also present

the results of collected real-world simulations in different environments to perform a field test of our learned models.

Keywords

Abnormal Event Detection, Anomaly Detection, Artificial Intelligence, Biometrics, Computer Vision, Machine Learning, Pattern Recognition, Self-Supervised Learning, Supervised Learning, Unsupervised Learning, Video Image Analysis, Visual Surveillance

Contents

1 Introduction **1**

1.1 Motivation and Objectives . . . 2
1.2 Main Contributions . . . 2
1.3 Thesis Organization . . . 3

2 Related Work **5**

2.1 Spatiotemporal Features with 3D Convolutional Networks . . . 5
2.1.1 3D Convolution and Pooling . . . 5
2.1.2 Network Settings and Exploring Kernel Temporal Depth . . . 6
2.1.3 Spatiotemporal Feature Learning . . . 7
2.2 Weakly-supervised Learning . . . 7
2.2.1 Multiple Instance Learning and Ranking Model . . . 7
2.2.2 Feature Extraction . . . 9
2.3 Generative Adversarial Networks . . . 9
2.3.1 Conditional GANs . . . 9
2.3.2 Abnormality Detection . . . 9
2.4 Spatiotemporal Auto-encoders . . . 10
2.4.1 Variational Auto-encoders . . . 11
2.5 Skeleton-based Trajectories . . . 11
2.5.1 Crowd Pose Estimation . . . 12
2.5.2 Pose Flow Tracking . . . 12
2.5.3 Learning Regularity . . . 13
2.6 Conclusion . . . 13

3 Iterative Weak/Self-Supervised Learning Framework **15**

3.1 Weakly-supervised Network and Novel Term . . . 16
3.2 Bayesian Classifiers . . . 18
3.3 Self-supervised Network . . . 20

4 Proposed Distribution-Based Loss **21**

4.1 Definitions and Expectation-Maximization Algorithm . . . 22
4.1.1 Expectation-step . . . 24
4.1.2 Maximization-step . . . 24
4.2 Anomaly Detection and Ranking Model . . . 25
4.3 Ranking Model vs. Negative Log-Likelihood Model . . . 27

5 UBI-Fights Dataset **31**

5.1 Related Datasets . . . 31
5.1.1 UMN Dataset . . . 32
5.1.2 USCD Peds1 and Peds2 Datasets . . . 32

5.1.3 CUHK-Avenue Dataset 33
5.1.4 Street Scene Dataset 33
5.1.5 ShanghaiTech Campus Dataset 34
5.1.6 UCF-Crime Dataset 34
5.2 New Dataset 35
5.2.1 Collection and Preprocessing 36
5.2.2 Annotation Process 36

6 Experiments and Discussion **39**
6.1 Datasets, Baselines, and Empirical Evaluation Protocol 39
6.2 Implementation Details 41
6.3 Self-Supervised Learning Results Comparison 41
6.4 Weakly-Supervised Learning Results Comparison 43
6.5 Qualitative Results 44

7 Conclusions and Future Work **51**
7.1 Conclusions 51
7.2 Future Work 52

Bibliografia **55**

List of Figures

2.1 2D and 3D convolution operations. 6
2.2 Convultional 3-Dimensional network (C3D) architecture. 6
2.3 Auto-encoder architecture. 10
2.4 Variational auto-encoder architecture. 11
2.5 Skeleton pose estimation example. 12

3.1 Cohesive pictorial view of the iterative weak/self-supervised learning framework. 15
3.2 Comparison between histograms in relation to the sample entropy term. . 17
3.3 Comparison between the posteriors probabilities obtained by the Bayesian classifier. 19

4.1 Predicted scores and respective distribution of a video containing an anomaly. 22
4.2 The flow diagram of the proposed Gaussian Mixture Model-based approach. 26

5.1 Illustration images of the scenes from the *UMN* dataset. 32
5.2 Illustration images of the scenes from the *UCSD Pedestrians* dataset. . . . 33
5.3 Illustration images of the scenes from the *CUHK-Avenue* dataset. 33
5.4 Illustration images of the scenes from the *Street Scene* dataset. 34
5.5 Illustration images of the scenes from the *ShanghaiTech Campus* dataset. 34
5.6 Illustration images of the scenes from the *UCF-Crime* dataset. 35
5.7 Illustration images of the scenes from the *UBI-Fights* dataset. 36
5.8 Statistics of all video sequences in the *UBI-Fights* dataset. 37
5.9 New developed annotation aplication used for the *UBI-Fights* dataset. . . . 37

6.1 Comparison between the ROC curves obtained by the eight methods considered in comparison with the Weakly-Supervised model (WS)/Self-Supervised model (SS) framework, for the *UBI-Fights*, *UCF-Crime* [SCS18] and *UCSD-Peds1* [LMV13] datasets. 43
6.2 Comparison between the ROC curves obtained by the seven methods considered in comparison with our Gaussian Mixture Model (GMM)-based proposal, for the *UBI-Fights*, *UCF-Crime* [SCS18] and *UCSD-Peds1* [LMV13] datasets. 44
6.3 Distribution scores comparison between Sultani [SCS18]'s and our GMM-based approaches in the three datasets. 47
6.4 Performance evolution with respect to the WS/SS framework iteration, observed for the *UBI-Fights*, *UCF-Crime* [SCS18], and *UCSD-Peds1* [LMV13] sets. 49
6.5 Requested *TOMI* devices for the experiments in capturing the scenes illustrated in figure 6.6. 49

6.6 Qualitative results of our model performed in real-world scenarios. The red window corresponds to the ground-truth of the scene. 50

List of Tables

6.1 Performance summary, concerning the individual training in each scene of the *UCF-Crime* dataset [SCS18], of the proposed weak/self-supervised method with respect to the state-of-the-art. The result values correspond to the Area Under the Curve (AUC). 42
6.2 Performance summary of the GMM method with respect to the state-of-the-art. 46
6.3 Performance summary of the proposed method with respect to the state-of-the-art. The final row in each dataset provides the performance of the GMM approach incorporated in the WS/SS framework. 48

Acronyms

AE Auto-encoder

AI Artificial Intelligence

AUC Area Under the Curve

C3D Convultional 3-Dimensional network

cGANs Conditional Generative Adversarial Networks

EER Equal Error Rate

EM Expectation-Maximization

FCN Fully Connected Neural network

GANs Generative Adversarial Networks

GMM Gaussian Mixture Model

GOF Goodness of Fit

HOF Histogram of Optical-Flows

HOG Histogram of Oriented Gradients

IoU Intersection over Union

KDE Kernel Density Estimator

LSTM Long Short Term Memory

MIL Multiple Instance Learning

ML Machine Learning

MLE Maximum Likelihood Estimation

MPED Message-Passing Encoder-Decoder

MSE Mean Squared Error

NLL Negative Log-Likelihood

NMS Non-Maximum Suppression

PDF Probability Density Function

PF Pose Flow

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

SS Self-Supervised model

SVM Support Vector Machine

SPPE Single Person Pose Estimation

VAE Variational Auto-encoder

WS Weakly-Supervised model

Chapter 1

Introduction

Nowadays, the thriving field of Artificial Intelligence (AI) has many practical applications and research areas. The development and employment of intelligent systems have been crucial in the human's routine labor, biometric system, speech understanding, medical analyses, and big data processing, being the basis support in every scientific research domains. The difficulties faced by those systems rely on the AI systems' ability to extract patterns and acquire their own knowledge from the observed raw data, which is described as Machine Learning (ML). With the advances of computers' technology and, consequently, the ML domain, the improvements on artificial neural networks have allowed diverse research for hard-coded knowledge, leading to a field known as deep learning [LBH15, GBC16]. Deep learning solves the main problem of extracting high-level and abstract features, such as every individual pixel when analyzing images of persons, where the factors of variations are innumerous, including the person's pose, clothes, ethnicity, the image's angle, and brightness. Deep learning algorithms learn representations that are denoted from others with multiple levels of abstraction, i.e., patterns among data collection, enabling computers to develop complex solutions from simple concepts. These computational models that are composed of multiple processing layers, dramatically improved the state-of-the-art in solving these problems [KSH12, Sch15], resulting in its increased use in various scientific research domains, continuously, bringing breakthroughs in deep convolutional neural networks in processing images, video, speech, and audio.

With the advances in the deep learning field and increasing use of convolutional neural networks, video image analysis has become an active research topic in solving biometric challenges over the last decade. Due to the continuous advent of surveillance systems, the development of biometric technologies to satisfy the need for an automatic anomaly detection algorithm is still a concern in the computer vision research community.

Anomaly detection consists of distinguishing which instances are dissimilar in a data-driven fashion. Those instances are known as anomalies, abnormalities, or outliers, and they may have origin on errors, but depending on which concept we are analyzing, occasionally, it may indicate a previously unknown event. So, an anomaly is an occurrence that deviates significantly from previously observed occurrences [Haw80, CC19]. In the video image concept, the data has spatial and temporal characteristics, so normally the anomalies are caused by motion or appearance of foreign objects. Abnormal event detection, in the video image analysis, is a challenging task not only for its large input data but also for the intra-diversification in a dataset, complexity of movements, and resemblance behaviors.

This work presents a set of solutions to bypass the difficulties of video image analysis in detecting abnormal events, in the supervised learning paradigm, starting with an extensive overview of the current state-of-the-art methods, and presenting two new distinct approaches that consistently outperforms the state-of-the-art, and additionally, introducing a new freely available database for anomaly detection.

1.1 Motivation and Objectives

With the increasing use of video surveillance systems, we are specifically engaged with the biometric challenge of detecting abnormal human behavior. The primary purpose of this dissertation is to present computer vision concepts in response to this challenge and clarify how motion and appearance can be computed in order to detect anomalies in multiple supervision paradigms. These objectives aim to provide several answers for the current difficulties approaching the different supervised learning domains. The motivation behind these objectives is to show how our computer vision algorithms successfully detect abnormal events and, hopefully, the supplied solutions support further research in this field. In anomaly detection, considering the video image analysis domain, one of the key challenges is the difficult task of providing divergent labeled data to the algorithm. Concerning its effort and time-consuming, we are also interested in employing weakly and partially supervised learning to face this key challenge. Therefore, in contrast to unsupervised learning algorithms, we will provide not the exact information as strongly-supervised learning applies, but we still provide minimum information about the data fed to the algorithms, to side-step the difficult task of annotation. To attain this goal, several approaches that already avoid the strong annotation task will be analyzed and used as a baseline of our two new proposals.

1.2 Main Contributions

The main contributions of our research work are a set of novel approaches to detect abnormal events in surveillance footage. The results achieved by our proposals can be considered an accomplishment, considering that apart from being novel approaches and consistently outperform the state-of-the-art, they performed quite successfully when testing in real-world scenarios. The individual contributions of this thesis to improve the state-of-the-art in abnormal event detection are briefly described in the following paragraphs.

As the first contribution, the comparison and discussion among the detailed state-of-the-art methods in the anomaly detection field contained in chapter 2.

The second contribution regards the proposal of using weak/self-supervised learning frameworks for abnormal events detection, resulting in a paper published in the 2020 IEEE International Joint Conference on Biometrics (IJCB) [DP20].

For the third contribution, we present a novel approach in the weakly-supervised learning paradigm for abnormal events detection. The new method employs a Gaussian Mixture Model (GMM) to learn a new ranking model that predicts anomaly events.

As the fourth and final contribution, we introduce the construction of a new large-scale first-of-the-kind anomaly detection database, which is freely available through the web for research purposes. As described in chapter 5, this database presents the characteristics that distinguish from the remaining public and free anomaly detection datasets.

1.3 Thesis Organization

The remainder of this thesis is organized as follows: chapter 2 presents a detailed review of the state-of-the-art in motion and appearance extraction on video sequences. Furthermore, a detailed overview of the anomaly detection state-of-the-art is given, focusing on how supervised and unsupervised learning methods use motion and appearance to detect abnormal events, with a focal point in weakly-supervised learning. Chapter 3 describes our new approach employing an iterative framework composed of two independent experts that provide labeled data to each other through a Bayesian framework, working under the weak and strong supervision paradigms based on self-supervised learning. Chapter 4 describes the main concepts in Gaussian Mixtures, and a novel proposal working under the weakly-supervised learning paradigm is further detailed. A detailed description of the current publicly available anomaly detection datasets is given in chapter 5, and the new public and freely available *UBI-Fights* database is introduced and detailed. Chapter 6 describes all the experiments involving both the newly proposed approaches in this work with comparison to the considered state-of-the-art methods. Furthermore, the real-world simulation scenarios captured for field test purposes are summarized and discussed. Finally, chapter 7 presents the conclusions, summarizes our achievements, and outlines the future work.

Chapter 2

Related Work

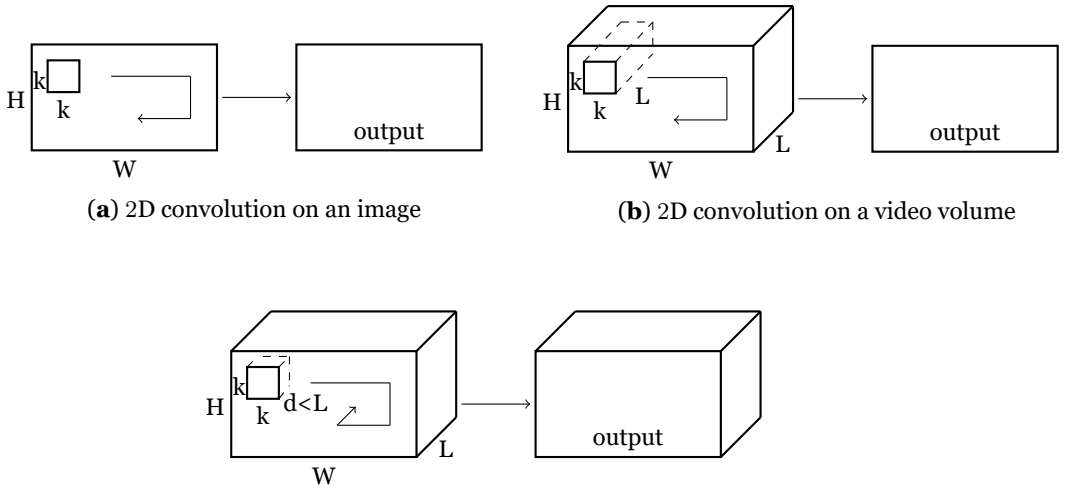
In this chapter, we review the fundamental concepts and techniques applied in the state-of-the-art methods related to abnormal events detection, how previous approaches capture motion and patterns to identify anomalies, and how they side-step the problem about the impossibility of having strongly-supervised training sets. Section 2.1 introduces the state-of-the-art method regarding the extraction of combined motion and appearance information through video volume data. Section 2.2 reviews the recent method, used as a baseline for this work, proposed in the weak supervision paradigm, where the instances are labeled at the video-level, instead of frame-level. Section 2.3 introduces the Generative Adversarial Networks (GANs) concept, and how they can be molded to be applied in the abnormal event detection problem. Section 2.4 summaries the Auto-encoder (AE) technique and how previous approaches employed them to detect anomalies in image analysis. Section 2.5 reviews the skeleton-based concept in extracting human motion trajectories to identify abnormal behaviors. Finally, section 2.6 presents the most relevant conclusions in the state-of-the-art of abnormal events detection.

2.1 Spatiotemporal Features with 3D Convolutional Networks

Inspired by the deep learning breakthroughs in the image domain, it is proposed by Tran *et al.* [TBF+15], a spatiotemporal feature learning by using deep Convultional 3-Dimensional networks (C3D) (3D ConvNets). In the context of large-scale supervised training datasets, they have created an effective video descriptor, that can be generic (representing diverse types of videos, and still being discriminative), compact (e.g., help to store, process, and retrieve tasks more scalable), efficient (in computing to process in real-world systems) and simple to implement.

2.1.1 3D Convolution and Pooling

3D ConvNets can model temporal information better [JXYY12, KTS+14, TBF+15], comparing to 2D ConvNets since these are done only spatially. In figure 2.1, it is illustrated the difference in convolutions, 2D convolution applied on an image and on multiple images (multiple frames as multiple channels) will output an image. Therefore, using 2D ConvNets, most of the networks lose their input's temporal signal after every convolution operation. 3D convolution is the only one that preserves the temporal information obtaining an output volume as a result. 2D and 3D pooling apply the same phenomena.



(**a**) 2D convolution on an image

(**b**) 2D convolution on a video volume

(**c**) 3D convolution

Figure 2.1: 2D and 3D convolution operations.

Karpathy *et al.* [KTS+14] share some similarities with Tran *et al.* [TBF+15] in terms of using full frames for training ConvNets. This approach [KTS+14] used fusion models with 2D convolutions, and those networks lost temporal information of the input signal after the first convolution layer. The slow Fusion model in [KTS+14], on the other hand, is the only one using 3D convolution, and for that reason, it performs better than the other networks studied in [KTS+14].

2.1.2 Network Settings and Exploring Kernel Temporal Depth

In this approach [TBF+15], they split videos into 16-frame clips with an 8-frame overlap between two consecutive clips and used them as input to the network. The final designed 3D ConvNet has 8 convolution layers, with 5 pooling layers, and 2 fully-connected layers, with a softmax loss layer to predict tasks tested. The convolution kernel size is represented by $d \times k \times k$, where d is kernel temporal depth, and k is spatial kernel size. Figure 2.4 represents the network architecture.

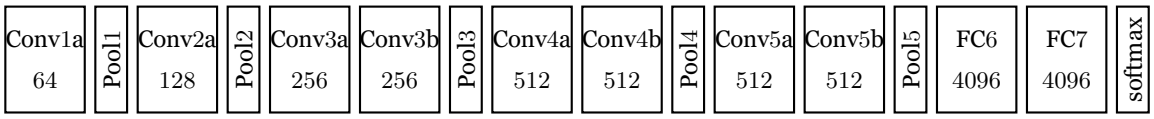


Figure 2.2: C3D architecture.

To aggregate temporal information through deep networks, they got the best results with constant kernel temporal depth of size 3 after experimented with two types of architectures concerning the temporal depth: homogeneous temporal depth (the kernel temporal depth is constant in all convolution layers) and varying temporal depth (where kernel temporal depth is changing across the layers).

2.1.3 Spatiotemporal Feature Learning

The training was done on the Sports-1M-dataset consisting of 1.1 million sports videos with 487 sports categories, obtaining a model from scratch, and other fine-tuned on Sports-1M from the model pre-trained on the I380K dataset. This last one performed better than the model trained from scratch, and both C3D models obtained better results compared to the state-of-the-art approaches, iDT [WS13] (improved dense trajectories) and Caffe's Imagenet pre-train model [JSD+14].

To understand what C3D was learning internally, they used the deconvolution method by Zeiler *et al.* [ZF14], using Deconvnet (applying unpool, rectification, and filtering operations to reverse a ConvNet). By visualizing the deconvolution of the *conv5b* layer, they concluded that C3D begins by focusing on the appearance in the first frames and starts tracking the motion in the following frames. Modeling not only appearance but also motion, this is where C3D differs from most of 2D ConvNets.
Extracting features every 16 frames, C3D can be used as a feature extractor to be applied for other video analysis tasks, by extracting the sixth fully-connected layer (*FC6*) activations and obtaining a 4096-dim video descriptor for every 16 frames.

2.2 Weakly-supervised Learning

Abnormal situation scenarios are very diverse, and since we are analyzing video images, the anomaly detection should be done with minimum supervision. This is due to the effort and time-consuming task of annotating the unpredictable number of hours of videos needed to train a model to reach satisfying performances. For this reason, recently, Sultani *et al.* [SCS18] proposes to learn anomalies through leveraging weakly labeled negative and positive videos, i.e., training labels are at video-level (weakly-supervised learning) instead of frame-level (strongly-supervised learning); therefore the anomaly detection model only knows if it is a normal video or if the video contains somewhere an anomaly.

2.2.1 Multiple Instance Learning and Ranking Model

[SCS18] propose to learn anomaly through a deep Multiple Instance Learning (MIL) framework, a type of supervised learning that receives a set of individually labeled instances, and all the instances with the same label are grouped, as bags.

By exploiting both normal and abnormal videos, each video is divided into non-overlapping temporal segments, where different segments make individual instances in a bag. Thus, a video is labeled as a positive bag if at least one of the instances contains an anomaly, and a video without any abnormal instance is labeled as a negative bag. As the exact information of the positive instances is not known, the implemented ranking function can be optimized concerning the maximum scored instance in each bag, as followed:

$$\min_{w} \left[\frac{1}{z} \sum_{j=1}^{z} \overbrace{\max\left(0, 1 - Y_{B_j}\left(\max_{i \in B_j}\left(\mathbf{w}.\phi\left(x_i\right)\right) - b\right)\right)}^{①} \right] + \|\mathbf{w}\|^2, \quad (2.1)$$

where ① is the hinge loss (for maximum-margin classification), Y_{B_j} is the bag-level label, z is the total number of bags, $\phi\left(x_i\right)$ denotes feature representation of the video segment i, b is a bias and $\mathbf{w}$ is the classifier to be learned (model weights).

In this approach [SCS18], the goal is to have high scores for abnormal videos and low scores for normal videos. So they propose, as the annotation is video-level, a multiple instance ranking objective function, used as loss function, based on ranking only the two instances having the highest anomaly score respectively in the positive and negative bag:

$$\max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) > \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right), \quad (2.2)$$

where max take over all video segments in each bag, $f\left(\mathcal{V}_a^i\right)$ and $f\left(\mathcal{V}_n^i\right)$ represent the prediction score of the abnormal and normal video segments i, respectively.

As a result of the highest anomaly score in the positive bag, the probability of being the true positive instance is very high. On the other hand, the highest score in the negative bag may look like an anomalous segment but is still a normal segment, which may generate a false positive. The purpose of the multiple instance ranking objective function 2.2 is to penalize both negative instances with high scores and positive instances with low scores.

Until now the underlying temporal structure of the anomalous video has been ignored, to solve that problem, the instances' scores in the positive bag should be sparse, meaning that only a few segments may contain the anomaly. And because one video is a sequence of instances, the anomaly score should vary smoothly between instances. This was done by incorporating the sparsity and smoothness constraints on the instances' scores:

$$\begin{aligned} l\left(\mathcal{B}_a, \mathcal{B}_n\right) = \max\left(0, 1 - \max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)\right) \\ + \lambda_1 \overbrace{\sum_{i}^{(n-1)} \left(f\left(\mathcal{V}_a^i\right) - f\left(\mathcal{V}_a^{i+1}\right)\right)^2}^{①} + \lambda_2 \overbrace{\sum_{i}^{n} f\left(\mathcal{V}_a^i\right)}^{②}, \end{aligned} \quad (2.3)$$

where ① denotes the temporal smoothness term, taking into account the next segment, and ② represents the sparsity term.

2.2.2 Feature Extraction

C3D video descriptor [TBF+15] was used as the feature extractor of their method, after computing C3D features for every 16-frame video clip to all the dataset and extracting the fully connected layer *FC*6 from the C3D network. This approach [SCS18] obtains one video segment's feature by splitting the video into 32 segments and taking the average of all 16-frame clip features contained in each segment. Thus a bag contains 32 instances.

2.3 Generative Adversarial Networks

Generative Adversarial Nets (GANs) [GPAM+14], are based on two networks competing with each other. One network is the generator (G), which has the purpose of generating realistic data. The second network is the discriminator (D), which has the purpose of discriminating generated data (from G) from real data. Both networks are trained with unsupervised data to learn a generative model of data.

These networks take as input an image r and generate a new image g. D will try to identify the real one and the fake one, while G will try to mislead D by producing even more realistic images that are harder to distinguish.

2.3.1 Conditional GANs

Ravanbakhsh *et al.* [RNS+17] use the method proposed by Isola *et al.* [IZZE17] based on Conditional Generative Adversarial Networks (cGANs), which aim to learn a conditional generative model. In cGANs, G and D receive some extra conditioning input information, and by adding a vector of features, it will control the output and guide G to figure out what to do.

This approach [RNS+17] compute optical-flow from consecutive frames and train two networks, $\mathcal{N}^{F\to O}$ to generate optical-flow from frames, and $\mathcal{N}^{O\to F}$ to generate frames from optical-flow. Thus by training networks with only normal videos, they will not be able to reconstruct/generate an abnormal event.

2.3.2 Abnormality Detection

For instances, by showing frames from a sidewalk where usually only people are allowed to pass by, and if for some reason a vehicle is found at the scene, the optical-flow generated by the frame with the vehicle, used as input to $G^{F\to O}$, will not be correctly reconstructed. As well as, the frame generated by the real optical-flow, containing the motion of the vehicle in the sidewalk, used as input to $G^{O\to F}$, the network will not be able to reconstruct correctly the area where the car is located.

From $G^{F \to O}$ they obtain an optical-flow reconstruction and compare with the real corresponding optical-flow (O), using a simple pixel-to-pixel difference, will highlight the differences between the real and the generated one. Likewise, they obtain the frame reconstruction from $G^{O \to F}(O)$ and compare with the real frame (F) by using a semantic difference computed using another network, AlexNet [KSH12] pre-trained on ImageNet Dataset [DDS+09] and extracting the fifth convolutional layer (which reproduces the information in a sufficiently abstract space). Fusing both normalized differences resulted from $G^{F \to O}$ and $G^{O \to F}(O)$ with the real O and F they are able to obtain the abnormality heatmap. The abnormality is detected by analyzing if the frame's heatmap contains at least one predicted abnormality pixel.

2.4 Spatiotemporal Auto-encoders

Auto-encoders (AEs) are deep learning algorithms that aim image reconstruction, meaning that whatever image we give to the input node, the network will generate the same result as the input.

AEs are composed of 3 main components: the encoder, responsible for compressing the input data, the bottleneck, representing a compressed low dimensional of the input, and the decoder, which will decompress the encoded data. The importance of image reconstruction is the compressed representation of the input data. For instance, if we have a feature vector of size 2048, learning an AE to be capable of reducing its size to 64 and reconstructing this feature vector from the bottleneck, means that we could transmit this compressed low dimensional representation through our network, reducing the cost of bandwidth and being able to reconstruct the image in the end with our decoder.

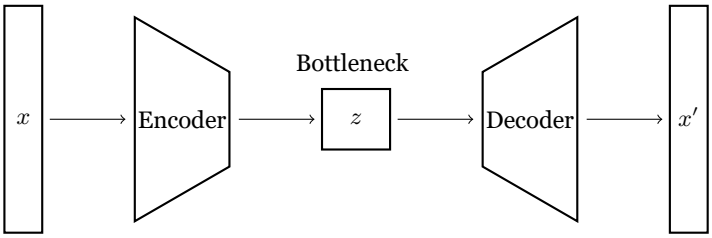


Figure 2.3: Auto-encoder architecture.

In real-world situations, when an abnormal event occurs, the most recent video frames will usually be much different from previous frames. Taking into account this principle, Chong and Tay [CT17], using only videos with normal situations scenarios, proposed to train a spatial feature extractor, extracting spatial features of every ten consecutive frames, and feed those features into a temporal AE using a non-linear activation function to minimize the reconstruction errors, from the decoder, in normal events. The hypothesis is that, since the AE was trained only with normal videos when an abnormal event occurs, it should not be able to reconstruct the scene.

Also, in the reconstruction proposal, Hasan *et al.* [HCN+16] extracted the motion features from the video frames consisting of Histogram of Oriented Gradients (HOG), and Histogram of Optical-Flows (HOF) fused both by employing as a baseline of improved trajectories the work by Wang and Schmid [WS13] and learning the regular motion signatures by feed those features to the fully-convolutional neural network based AE. They also approached the problem from the one-class classification perspective (a term coined by Moya and Hush [MH96]), training the AE with exclusively normal events, and assuming its incapability of abnormal events reconstruction, the irregular motions are detected from their high reconstruction errors.

2.4.1 Variational Auto-encoders

Instead of only learning a model to represent the input data (z, compressed data) as AEs, the goal of a Variational Auto-encoder (VAE) is to model a probability distribution of some latent variables (z) which we can sample from the probabilistic encoder/distribution $\mathcal{Z} \sim q_\phi(z|x)$ and decode/generate $\mathcal{X}' \sim p_\theta(x|z)$ new input samples. So it is designated as a generative model like, for instance, GANs.

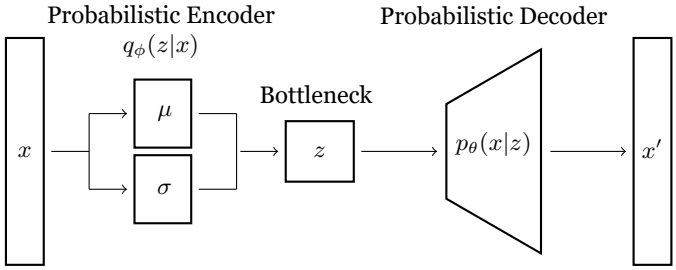


Figure 2.4: Variational auto-encoder architecture.

In the generative model paradigm, Wang *et al.* [WQL+18] proposed to extract the frame features using a pre-trained, on Pascal VOC [EVGW+], a fully convolutional neural network built based on VGG-16 [SZ14], obtaining motion features of two consecutive frames. Both motion and appearance features are used as input to a VAE removing unnecessary features. The filtered ones are fed into a second network combining U-net [RFB15] and VAE, adding skip connections across layers between low-level features and high-level features, linking the encoder's layers to decoder's layers to avoid information loss, which is inevitable in the decoding process. Since both networks are generative models, and by training them only with normal events, the anomaly is detected by the reconstruction error.

2.5 Skeleton-based Trajectories

The patterns of human movements and behaviors in surveillance videos can also be modeled through leveraging 2D human skeleton-based trajectories. Those features consist of a collection of human body joints' locations in the spatio-temporal field. This technique also avoids all appearance noises that images can contain, being modular, semantically rich,

and very descriptive, and consequently, concentrating the training model exclusively on human behavior.

2.5.1 Crowd Pose Estimation

Multi-person pose estimation lies in the extraction of skeleton pose features from images, consisting of capturing and estimating the locations of multiple joints in the human body (shoulders, elbows, hands, hips, knees, ...), as illustrated in Figure 2.5.

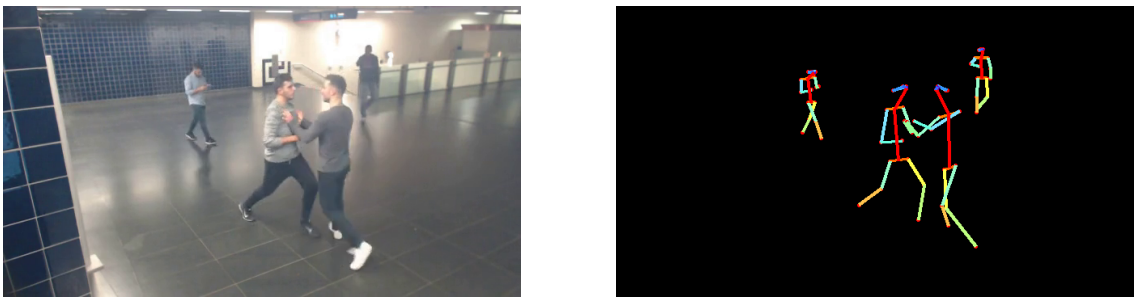


Figure 2.5: Skeleton pose estimation example.

Recent works [LWZ+19, FXTL17, XLW+18] use a pretrained object detector to detect humans and obtain the respective bounding boxes. Once obtained, the images are fed into a Single Person Pose Estimation (SPPE) [TS14, WRKS16], calculating the heatmap of the image proposal inside the bounding box to indicate human joint locations. In crowded scenarios, the problem lies in being highly likely to obtain other human parts inside the bounding box of another human since SPPE is limited to a human proposal. Li *et al.* [LWZ+19] propose a new heatmap loss that suppresses the interference joints that do not belong to the bounding box's target person, by considering the Mean Squared Error (MSE) between the output heatmaps of different persons and both heatmap target person's joints and interference joints. Therefore, each joint has a confidence associated with each bounding box.

2.5.2 Pose Flow Tracking

To obtain skeleton-based trajectories, not only we need to estimate the skeleton pose but also keeping track of that respective person. The most simply heuristics is the Intersection over Union (IoU) of boxes, which only by itself will most likely fail, for instance, if some person moves fast enough for boxes not overlapping over consecutive frames, and as abovementioned in crowded scenes where boxes may not be related with the respective skeleton pose.

Xiu *et al.* [XLW+18] proposed to use a pose tracker based on Pose Flow (PF), which consists of designing an optimization framework to build the relationship between cross-frame poses and build PFs. A PF Non-Maximum Suppression (NMS) (PF-NMS) is then

responsible for reducing redundant and reconnect temporal disjoint PFs. Analyzing similarities between poses in the same frame and distance between poses from consecutive frames, [XLW+18] proposes a PF-builder that focuses on the overall confidence score of a single PF, keeping track of the previous trajectories regarding the same PF to resist sudden occlusions and motion blur. Since there is a high likelihood of multiple PFs indicating the same person in multi-person scenarios, extracting the overlapping temporal sub-flows, allows them to perform an NMS scheme (IoU comparison between multiple PFs) to select the PF with maximum confidence score (after confidence summation of the PFs).

2.5.3 Learning Regularity

Once extracted the dynamic skeleton features, Morais *et al.* [MLT+19] propose a method to learn only the normal patterns of human behavior, learning its regularity through decomposing the skeleton trajectories into global body movement, tracking the dynamics of the whole body in the scene, and local body posture, describing the internal deformation of the skeleton in the respective bounding box.

To analyze both components, [MLT+19] developed a Message-Passing Encoder-Decoder (MPED) Recurrent Neural Network (RNN) (MPED-RNN), consisting of two branches, each one, responsible to the respective component, is composed of a single-encoder-dual-decoder architecture with three RNNs, and by employing a cross-branch message-passing mechanism, it does not only model each individual dynamics component but also interdependencies between them. For its similar performance to Long Short Term Memory (LSTM) [HS97], they used gated recurrent units in every segment of the MPED-RNN, where a segment from one branch receives the information from the other branch of its internal state at the previous time step.

With both single-encoder-dual-decoders training with only normal skeleton-based trajectories, the encoders are able to learn a compact representation to reconstruct normal human behavior. Therefore, in the testing phase, the anomaly is detected when the MPED-RNN cannot properly reconstruct the abnormal dynamics of the human body.

2.6 Conclusion

This chapter presented a comprehensive analysis of the concepts in anomaly detection and the state-of-the-art methods in this kind of technology. First, we reviewed in detail how the recent method of feature extraction in 3-dimensional volumes works, as the state-of-the-art in this feature extraction domain, we employed this feature descriptor for the developments of our proposals. Subsequently, we presented a detailed review of the state-of-the-art in weak supervision learning, and how the feature descriptor was applied in this type of learning. Afterward, we detailed the most important concepts regarding different methods of the state-of-the-art in the unsupervised learning domain. Despite the interesting approaches, there are some disadvantages in applying unsupervised learning methods

in this type of challenge, since every method only analyze normal patterns. We detected that in this type of learning, it is a fundamental request that the whole training set should be homogeneous, which means the videos are captured in invariant locations; otherwise, the unsupervised learning algorithms will fail in an almost catastrophic way, since training AEs or GANs to learn normal patterns in multi-variate scenarios is extremely difficult due to its heterogeneity of the inputs. For this reason, we will employ the state-of-the-art in weakly-supervised learning as the main baseline of our methods.

Chapter 3

Iterative Weak/Self-Supervised Learning Framework

In standard supervised classification problems, the labels of all negative and positive instances are given, and the classifier can be learned using maximum-margin classification based loss, used for Support Vector Machine (SVM) [CST99]. Since obtaining temporal annotations for videos is very effort and time-consuming, this chapter describes an iterative system composed of two independent networks based on the weakly and self-supervised learning paradigm. Regarding the principle that a model trained with videos labeled frame by frame, where the exact information is provided (strong supervision), will achieve better performance than training with weakly labeled videos, we try to attain, as much as possible, those performances as if it was a strongly-supervised model. By avoiding the annotation of anomalous segments or clips in training videos, these coupled deep-learning networks were designed to self-supervise non-annotated data at both weak and strong level annotations. It should be noted that the proposed method [DP20] relies on one condition: aside from having access to the weakly supervised set, an additional unlabelled set of videos should be available with the purpose of being annotated by the model itself.

Upon the work of Sultani *et al.* [SCS18] and to infer a model that distinguishes between normal and abnormal segments, a weakly supervised training set is used. Also, a novel term was introduced into the basis loss function (2.3) to fortify the response score's spread in the unit interval, and being able to discriminate the most reliable correctly classified instances.

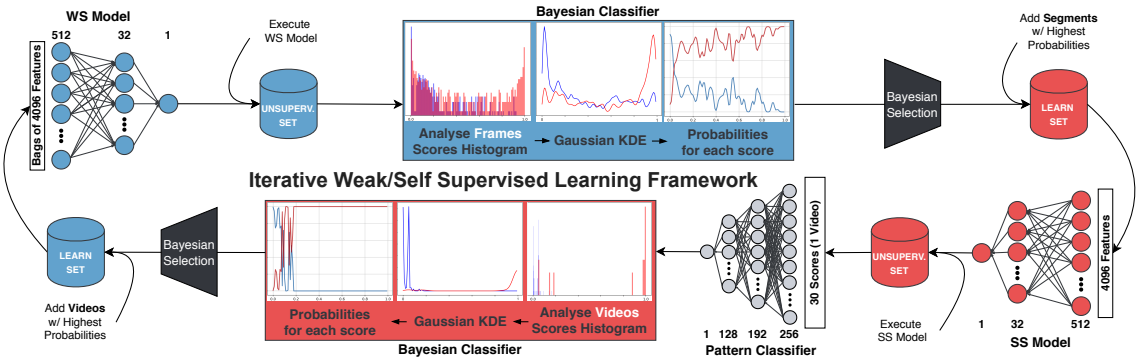


Figure 3.1: Cohesive pictorial view of the iterative weak/self-supervised learning framework.

Figure 3.1 presents the step-by-step diagram of the iterative learning framework proposed

in this work. The blue and red colors represent the two kinds of components used: at first, we adapted the Sultani [SCS18]'s method to spread its prediction scores in the unit interval as much as possible (WS). Next, a Bayesian framework is responsible for filtering out the instances coming from an unlabelled data set that is deemed to be the most reliably classified, which are used as input of the self-supervised learning network working at the frame level (SS). Once again, the responses predicted by this expert are used by a second Bayesian framework, that filters out the videos to be used in the next iteration of the first framework.

For consistency purposes, the notation adopted is as close as possible to the used by Sultani *et al.* [SCS18]. $\mathcal{V}_u^i$ denotes the i^{th} segment of a video (bag) $\mathcal{B}_u$, where $u \in \{a, n\}$ referring respectively videos with abnormal and only normal events. $\boldsymbol{\mathcal{B}}^{(t)} = \{\mathcal{B}_1, ..., \mathcal{B}_{n_w}\}$ denotes a learning set of n_w videos at iteration t (used as learning data by the WS model), and $\boldsymbol{\mathcal{V}}^{(t)} = \{\mathcal{V}_1, ..., \mathcal{V}_{n_s}\}$ is the learning set composed of n_s segments, used by the SS model.

3.1 Weakly-supervised Network and Novel Term

Based upon the work of Sultani [SCS18], the proposed approach learns anomalies through a deep MIL framework by considering the input videos as bags. In a binary way, a bag is annotated as a positive instance, if at some point, it contains an abnormal event, while a bag is annotated as a negative instance if it surely contains solemnly normal segments.

In spite of the network learning to predict high scores for anomalous segments applying the loss function described in (2.3), during our experiments, we noticed that typically it produces extremely peaked distributions for both the normal and abnormal scores, turning hard to perceive the most confidently correctly classified instance since there is still a large number of false-negative and false-positive instances. In order to optimize the performance of the Bayesian framework, it was essential to ensure that the response scores for the most evidently normal and abnormal segments were close to the extremes of the unit interval, while all the responses for the doubtful observations should spread as much as possible among the intermediary values. Therefore, to achieve a better distribution for our purposes, an entropy-based term was added to the loss function (2.3), given by:

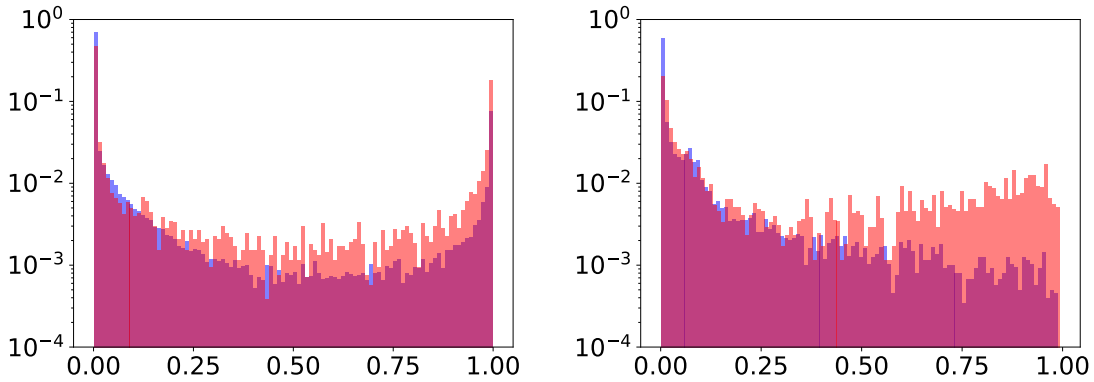
$$H(\mathcal{B}_u) = -\sum_{i=1}^{t} P\big(f(\mathcal{V}_u^i)\big)\ \log\Big(P\big(f(\mathcal{V}_u^i)\big)\Big), \quad (3.1)$$

where $\mathcal{B}_u \overset{\text{def}}{=} \{\mathcal{B}_a \cup \mathcal{B}_n\}$ denotes all the videos in the batch, $f\left(\mathcal{V}_u^i\right)$ represent the prediction scores of abnormal and normal segments in the batch, and $P\big(f(\mathcal{V}_u^i)\big)$ approximates the density of the $f(\mathcal{V}_u^i)$ scores. We divided the unit interval into bins of equal width and counted the number of values in each bin:

$$b_i = \sum_{j=1}^{n} \mathbb{1}_{\{\frac{i-1}{n} \leq f(\mathcal{V}_u^j) \leq \frac{i}{n}\}}, \forall i \in \{1, \ldots, t\}, \quad (3.2)$$

where $\mathbb{1}$ denotes the characteristic function. After obtaining the b_i statistic, values were normalized to have sum equal to one, i.e., to fit a probabilistic distribution: $b_i^* = \frac{b_i}{\sum_j b_j}$. In practice, the b_i^* values were used to approximate $P\big(f(\mathcal{V}_u^i)\big)$.

The main impact of the novel term is illustrated in Fig. 3.2, with the vertical axis providing the b_i^* values per bin, and the horizontal axis denoting the codomain of the classifier, where the blue and red bins represent the negative and positive scores, respectively. The left plot regards Sultani *et al.* [SCS18]'s loss function, while the right plot provides the values for the loss function of this proposal. For visualization purposes, the vertical axis is plotted in logarithmic scale.



(a) Sultani *et al.*'s loss. (b) Proposed loss with the novel term.

Figure 3.2: Comparison between histograms in relation to the sample entropy term.

By incorporating the entropy-based term (3.1) to the basis loss function (2.3) and also taking into account the regularization terms that minimize model weights, our complete objective function becomes:

$$\mathcal{L}(\mathcal{W}) = l\left(\mathcal{B}_a, \mathcal{B}_n\right) - \lambda_3 \overbrace{H(\mathcal{B}_u)}^{③} + \lambda_4 \|\mathcal{W}\|_F, \quad (3.3)$$

where $\mathcal{W}$ represents the model weights and ③ avoids peaked distributions for normal and abnormal events.

As stated above, the rationale of (3.3) is that less peaked distributions of scores will produce higher entropy values. Applying the subtraction of this term in the basis loss will reduce the cost function and, at the same time, making it harder to reach the extremes

values of the unit interval, even if it is less frequent to obtain scores close to those values, the important point is to ensure that only the most significant normal and abnormal instances attain values close to the extremes. On the other hand, the uncertain observations will be dispersed between intermediary values, consequently achieving a suitable distribution shape, which is the key feature for selecting the more likely correctly classified instances that are iteratively added to the next generation of the learning sets, by means of the Bayesian framework.

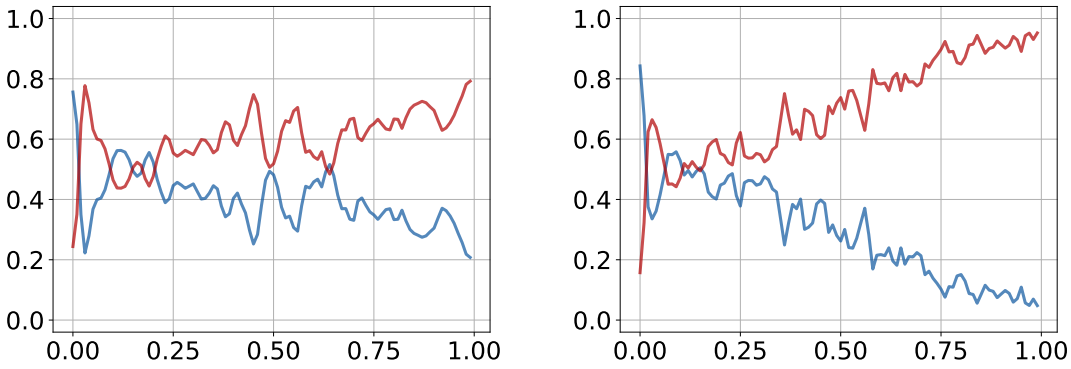
3.2 Bayesian Classifiers

The proposed approach is based on Bayesian classifiers to obtain, among the unsupervised data, the reliability for each classified instance. In a self-supervised fashion, both Bayesian classifiers were designed specifically to add to the next generation of the learning set, only the instances with extreme high beliefs, either for the WS and SS models. Since our coupled deep-learning networks are independent and working at distinct levels of supervision, two Bayesian classifiers are employed in our method. First, the weak one that receives the scores produced by the WS expert, filtering the video segments (labeling in a strongly-supervised manner) that should be used in the SS training. Second, the strong counterpart, that receives the scores from the SS expert (at the segment level), and based upon the scores' pattern of each unlabelled video, selects those that should be used in the next iteration of the WS model (labeling in a weakly-supervised fashion).

Every self-supervision technique requires a pre-task wielded as a baseline, which, in our case, both classifiers use the response scores from the validation set (where ground-truth is available), to estimate the exact degree of belief of a classified instance, used to obey the following probability calculus:

$$P_w\big(y|f(\mathcal{V}^i)\big) = \frac{P\big(f(\mathcal{V}^i)|y\big)\ P(y)}{P\big(f(\mathcal{V}^i)\big)}, \quad (3.4)$$

where $y \in \{a, n\}$ represent the abnormal and normal classes. A Gaussian Kernel Density Estimator (KDE), from two distributions (negative and positive scores), was used to more accurately approximate the conditional densities $P\big(f(\mathcal{V}^i)|y\big)$, with Scott's rule [Sco15] for bandwidth selection. Even though there is typically a significant imbalance between the number of normal and abnormal instances, we empirically adjusted the priors to $P(a) = P(n) = 0.5$, avoiding giving excessive weight to the normal instances.
Figure 3.3 compares the posteriors per class, regarding the histograms in Figure 3.2 and according to the Sultani *et al.* [SCS18]'s loss and ours, when the entropy term was also considered. The blue lines represent the normal events, and the red lines represent the abnormality scores. As it is notorious, our proposal's immediate effect is to obtain a more evident separation between the degrees of belief for both classes near the extremes of the unit interval (particularly in the upper extreme).



(a) Sultani *et al.*'s loss. (b) Proposed loss with the novel term.

Figure 3.3: Comparison between the posteriors probabilities obtained by the Bayesian classifier.

Now that our weak Bayesian classifier is able to discriminate the response scores qualitatively from the unsupervised set, formally, the ith segment is selected for the next generation of the SS learning data according to the rule:

$$\boldsymbol{\mathcal{V}}^{(t+1)} \stackrel{\text{def}}{=} \{\mathcal{V}^i \iff P_w\big(y|f(\mathcal{V}^i)\big) \geq \tau_1\}, y \in \{a, n\}, \quad (3.5)$$

i.e., if the posterior for either classes is higher than a threshold.

In the counterpart to also supply more data to the WS learning set, the strong Bayesian classifier is responsible for deciding at the bag (video) level, receiving the video segments scores and classifying that video globally as normal or abnormal (weak supervision):

$$P_s\big(y|f(\mathcal{B}^i)\big) = \frac{P\big(f(\mathcal{B}^i)|y\big)\ P(y)}{P\big(f(\mathcal{B}^i)\big)}, \quad (3.6)$$

where $y \in \{a, n\}$ represent the abnormal and normal classes. We initially considered the application of simple fusion rules (such as *max* or *prod*), on the response scores of each video, for obtaining the degree of belief per video. However, the poor levels of performance led to the choice of introducing a simple feed-forward network classification with four layers and $256 : 196 : 128 : 1$ architecture for the estimation of $P_s\big(y|f(\mathcal{B}^i)\big)$.

This network (described as Pattern Classifier in Figure 3.1) is trained from scratch with the result scores from the SS expert performed through the validation set, in each system's generation to be able to keep up accordingly to the behavior of the SS model's evolution. Finally, and in a way much similar to the weak Bayesian classifier, only the classified videos that provide the extreme degrees of belief are selected to the next iteration of the WS learning set:

$$\boldsymbol{\mathcal{B}}^{(t+1)} \stackrel{\text{def}}{=} \{\mathcal{B}^i \iff P_s\big(y|f(\mathcal{B}^i)\big) \geq \tau_2\}, \forall i \in \boldsymbol{\mathcal{B}}^{(t)}, y \in \{a, n\}, \quad (3.7)$$

The τ_i values were adjusted empirically, according to the performance observed in a validation set.

3.3 Self-supervised Network

Typically, a network trained with videos labeled at the frame-level, where we provide the exact information about the anomalies event, known as strongly-supervised learning, will achieve better performance than a model trained with weakly labeled videos. Since our weak Bayesian framework is now capable of providing that exact information in the most trusted instances, we developed a network to receive those instances.

Based upon the recent technique of self-supervised learning, the SS model is trained at the segment level (strong-supervised learning), where there was no human supervision in the learning set of this expert, being annotated only by the combination of the WS model and the weak Bayesian classifier with the most likely correctly classified segments. The main purpose of this model is to achieve high performances similar to a frame-level supervision done manually by a human.

This network receives C3D [TBF+15] feature vectors of 4096 components, describing video segments of each 16 frames of the respective video, and predicts the abnormality of that segment, representing the likelihood of containing an abnormal event. With a 3-layer fully connected architecture, this expert is composed of 512 units in the first layer, 32 units in the second one, and 1 unit in the output layer. We used the ReLU [NH10a] activation function for the two first layers and a sigmoid activation in the output neuron. A binary cross-entropy loss drove the learning process of this model.

Once our SS model is trained at the segment-level, we compute it through the unlabeled data and use our pattern classifier to analyze the response scores, from the validation set, and learn an efficient model to discriminate negative bags from positive bags based on the patterns of the SS model's scores in each video. Finally, our strong Bayesian framework will analyze the classified bags from the pattern classifier, providing additional weakly-labeled videos to the training set of our WS model. Since the WS model's performance is better than the previous iteration, the whole process is repeated until one of the networks' performance stabilizes.

Chapter 4

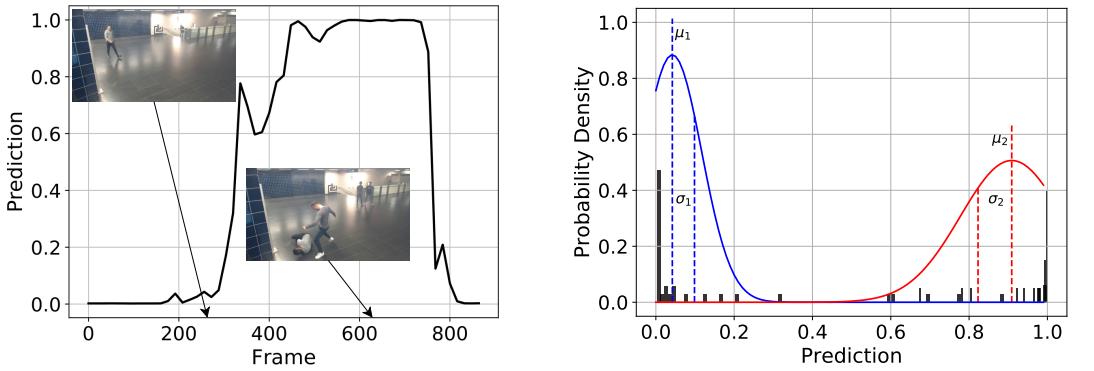
Proposed Distribution-Based Loss

As already observed in chapter 2, in the world of ML, we can distinguish two main fields: supervised and unsupervised learning. The main difference between these two types of learning resides in the nature of the data, and how approaches handle it. In supervised learning models, the algorithm learns based on a labeled dataset, providing its output that the model use to calculate its loss and evaluate its accuracy on the training set. On the other hand, computationally more complex, unsupervised learning algorithms learn with unlabeled data, which means only the input is given without any key answer to relate with. An important concept concerning unsupervised learning is clustering. It involves grouping a set of data points (denominated as a cluster), which are related to some features that distinguish them from other data points (consequently, other clusters), i.e., finding structures or patterns in a set of uncategorized instances.

A popular clustering algorithm, known as K-means, is a very powerful technique of unsupervised methods, consisting of updating the model's parameters of each cluster iteratively, by computing the mean (centroid) of each cluster regarding its distance to each data point and repeating the process until some convergence criterion is reached. Each data point is then labeled as part of one and only one cluster, concerning its closest centroid. Considered as a hard clustering technique, the K-means algorithm has some limitations regarding the miss information of an uncertainty measure or probability associated with a data point to each cluster. For this reason, we employed a Gaussian Mixture model (GMM), which is a unsupervised learning algorithm based on the principle that every finite number of data clusters can be generated from a mixture of a finite number of Gaussian distributions. Therefore, a collection of data can be represented by several Gaussians, each one identified by $k \in \{1, \ldots, \mathcal{K}\}$, where $\mathcal{K}$ denotes the number of clusters of our dataset, which means the number of kernels in our dataset's Gaussian mixture distribution. Consisting of a soft clustering technique for correlating data points to clusters, GMMs rely on a probabilistic estimation of a point belonging to a cluster; thus, each data point has a likelihood associated with each cluster.

This chapter introduces a GMM-based approach for anomaly detection in video surveillance footage, under the MIL paradigm similar to the WS expert described in chapter 3, that predicts high anomaly scores for anomalous video segments. The rationale is that, upon a video containing an anomaly, the predicted scores should be near the lower extreme of the unit interval when no abnormal event is observed, and when an anomaly event occurs, the predicted scores should be higher, near the upper extreme, following a two-kernel distribution-like comprised between the unit interval. On the other hand,

upon a normal video, the predicted scores should only concentrate near the lower extreme, meaning that using a two-kernel GMM, both kernels will achieve the same fitting as a normal distribution. For that reason, we introduce a new MIL solution by employing the estimated parameters, of both two-kernel GMM and normal distribution, in our novel loss function. As demonstrated in figure 4.1, the left plot regards the predicted scores upon an anomaly event in a real-case scenario, while the right plot illustrates the respective histogram of the predicted scores (higher the bin, higher the data points' density), and how a GMM algorithm, with $\mathcal{K} = 2$, adjusted its Gaussian functions to the predicted data points.



(a) Model's prediction scores in response to an anomaly event.

(b) Gaussian mixture distribution in relation to the predicted scores in figure 4.1a.

Figure 4.1: Predicted scores and respective distribution of a video containing an anomaly.

For consistency purposes, the annotation adopted is as close as possible to the used in chapter 3. $\mathcal{V}_u^i$ denotes the predicted score of the i^{th} segment of a video (bag) $\mathcal{B}_u$, where $u \in \{a, n\}$ referring respectively videos with abnormal and only normal events, and each $\mathcal{B}_u$ is composed of a fixed number of segments. $\boldsymbol{\mathcal{B}}^{(t)} = \{\mathcal{B}_1, ..., \mathcal{B}_{n_w}\}$ denotes a learning set of n_w videos at iteration t (used as learning data by the GMM-based approach).

4.1 Definitions and Expectation-Maximization Algorithm

As the name implies, a Gaussian Mixture is a distribution function comprised of $\mathcal{K}$ Gaussians, known as kernels. In the mixture, each Gaussian k is composed of the following parameters:

- The mean μ of the kernel, similarly to a centroid in other clustering algorithms.
- The standard deviation σ (for one-dimensional data) or covariance matrix Σ (for multivariate scenario) that defines its width.
- The mixing coefficient π that defines the weight of the distribution, i.e., the prior probability of a random variable associated to the kernel itself.

It can be observed from the histogram's bins, that the densities are higher in the extremes of the unit interval, and for that reason, the GMM means (μ_1 and μ_2) were adjusted to the

proximity of the extreme values. Likewise, σ_1 is smaller than σ_2 since, as also expressed by the histogram, the predicted scores are less scattered in the blue Gaussian, in comparison to the second Gaussian (red).

In our case, where we are only operating with a single variable (the prediction score $\mathcal{V}_u^i$ of a segment video), the Probability Density Function (PDF), in one-dimensional space, is given by:

$$\mathcal{N}(\mathcal{V}_u^i|\mu_k, \sigma_k^2) = \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mathcal{V}_u^i - \mu_k}{\sigma_k}\right)^2}, \quad (4.1)$$

which can be interpreted as the probability density of a predicted score falling within a specified Gaussian distribution. As abovementioned, figure 4.1b exemplifies a two-component mixture model, where each Gaussian describes the predicted scores of a video, $\mathcal{B}_u$, comprised in each of the two clusters. Therefore, the π parameters of both kernels must meet the following condition:

$$\sum_{k=1}^{\mathcal{K}} \pi_k = 1 \quad (4.2)$$

Accordingly to both equations (4.1) and (4.2), we can define a Gaussian mixture as the following summation of Gaussian distributions:

$$\mathcal{G}(\mathcal{V}_u^i|\gamma) = \sum_{k=1}^{\mathcal{K}} \pi_k \mathcal{N}(\mathcal{V}_u^i|\mu_k, \sigma_k^2), \quad (4.3)$$

which means the probability density of a random predicted score coming from a GMM configuration γ.

In order to ensure that each Gaussian fits the predicted points, defined as maximum likelihood, we need to determine the optimal configuration γ values for the GMM parameters to attain its maximum likelihood. In frequentist probability [Ven88], models are usually learned by employing the Maximum Likelihood Estimation (MLE) techniques, which maximizes the probability/likelihood of the observations given the parameters of the model by differentiating the logarithm of the joint probability of all observations, given by:

$$\ln\left(\prod_{i=1}^{N} \mathcal{G}(\mathcal{V}_u^i|\gamma)\right) = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{\mathcal{K}} \pi_k \mathcal{N}(\mathcal{V}_u^i|\mu_k, \sigma_k^2)\right), \quad (4.4)$$

where N represents the number of predicted scores in $\mathcal{B}_u$. Unfortunately, this solution for mixture models is usually analytically impossible, due to the derivative of the log-

likelihood of multiple Gaussians. To side-step this issue, we employ an iterative method, denominated as the Expectation-Maximization (EM) algorithm, which is typically used when (as shown below) we can calculate the closed-form expressions of the model parameters. Consisting of a numerical technique for maximum likelihood estimation, and since the EM algorithm strictly increases the maximum likelihood after every iteration, it is guaranteed to reach a local maximum.

4.1.1 Expectation-step

The expectation-step is responsible for calculating the expectation of the Gaussian kernel k assignment for every predicted score $\mathcal{V}_u^i \in \mathcal{B}_u$ given the mixture model parameters γ, which is similar to the evaluation of our model.

Simplifying, instead of estimating: "given a Gaussian k, what is the probability of generating a predicted score $\mathcal{V}_u^i$?" (4.1); we will need to estimate: "given a predicted score $\mathcal{V}_u^i$, what is the probability it has been generated by a Gaussian k ?", expressed as $P(g_{ik} = 1|\mathcal{V}_u^i, \gamma)$. g_{ik} would denote a latent variable that takes only two possible values, one if $\mathcal{V}_u^i$ came from kernel k, and zero otherwise. From the Bayesian inference, we know that:

$$P(g_{ik} = 1|\mathcal{V}_u^i, \gamma) = \frac{P(\mathcal{V}_u^i|g_k = 1)P(g_k = 1)}{\sum_{n=1}^{\mathcal{K}} P(\mathcal{V}_u^i|g_n = 1)P(g_n = 1)}, \forall i \in \mathcal{B}_u \quad (4.5)$$

Subsequently, as abovementioned and denoted from our previous equations (4.1) and (4.2), we already defined the following components:

$$P(\mathcal{V}_u^i|g_{ik} = 1) = \mathcal{N}\big(\mathcal{V}_u^i|\mu_k, \sigma_k^2\big) \quad \text{and} \quad P(g_{ik} = 1) = \pi_k \quad (4.6)$$

Replacing both components from (4.6), our final Bayes' theorem (4.5) becomes:

$$P(g_{ik} = 1|\mathcal{V}_u^i, \gamma) = \frac{\pi_k \mathcal{N}\big(\mathcal{V}_u^i|\mu_k, \sigma_k^2\big)}{\sum_{n=1}^{\mathcal{K}} \pi_n \mathcal{N}\big(\mathcal{V}_u^i|\mu_n, \sigma_n^2\big)} = \gamma(g_{ik}), \forall i \in \mathcal{B}_u \quad (4.7)$$

With equation (4.7) evaluating our actual GMM configuration, we are able to obtain the information about how good is this $\mathcal{K}$ mixtures configuration fitting the predicted data points.

4.1.2 Maximization-step

With the expectation calculated in the expectation-step, the maximization-step will maximize it with respect to the prior GMM parameters γ, which consists of updating, in each mixture component k, the new values for $\mu_k^{(t+1)}$, $\sigma_k^{(t+1)}$ and $\pi_k^{(t+1)}$ in order to achieve a better fitting in the next iteration, $t + 1$.

To achieve maximum likelihood, and by making use of the log of equation (4.3) with the expectation of equation (4.7), we obtain the following responsibilities from the differentiation with respect to each parameter:

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma(g_{ik})}{N}$$
$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma(g_{ik}) \mathcal{V}_u^i}{\sum_{i=1}^{N} \gamma(g_{ik})} \qquad \sigma_k^{2(t+1)} = \frac{\sum_{i=1}^{N} \gamma(g_{ik})(\mathcal{V}_u^i - \mu_k)^2}{\sum_{i=1}^{N} \gamma(g_{ik})} \tag{4.8}$$

Using these revised values to determine the new expectation in the next EM iteration, the entire iterative process repeats until some convergence criterion is reached, usually when the log-likelihood stabilizes, meaning that we have reached a local maximum.

4.2 Anomaly Detection and Ranking Model

The proposed Gaussian Mixture based approach, working under the same MIL paradigm as the WS model in section 3.1, is built upon the assumption that abnormal videos usually contain both normal and abnormal scenarios, while normal videos only show ordinary events, the purpose of this approach is to achieve a two-kernel distribution, in positive bags, and a single-kernel distribution (normal distribution), in negative videos. We introduce a novel loss function, which operates at the distribution-level of the GMM fitted to the predicted scores in each negative and positive bag, and by employing the configurations of the normal and mixture distribution in every bag, we will penalize the model if it does not shape a score distribution accordingly to the aforementioned. A cohesive pictorial view of the whole proposal's framework is given in figure 4.2, where the blue and red elements represent a negative scenario (containing no anomaly) and positive scenario (containing somewhere an anomaly), respectively—starting by dividing the training videos into a fixed number of temporal video segments, which make instances in a bag. After extracting the fully-connected (*FC6*) activations of the C3D [TBF+15] method for every video segment, a Fully Connected Neural network (FCN) is trained by making use of a novel cost function which computes the loss between the normal (black line) and Gaussian mixture (blue and red line) parameters in the negative and positive bags.
With the previously described EM algorithm, we are capable of fitting our two-kernel GMM to any predicted score distribution, intending to extract the mixture parameters, once fitted. Since abnormal videos should not have the same distribution scores as normal videos, we also gather information about the normal distribution of the instances' scores.

In this GMM-based proposal, we address the anomaly detection as a regression problem. We need the anomaly videos to have both low scores and high scores, while the normal videos should keep low scores. Since we are operating with a normal distribution and a two-kernel mixture distribution, it means that we end up with the parameters of

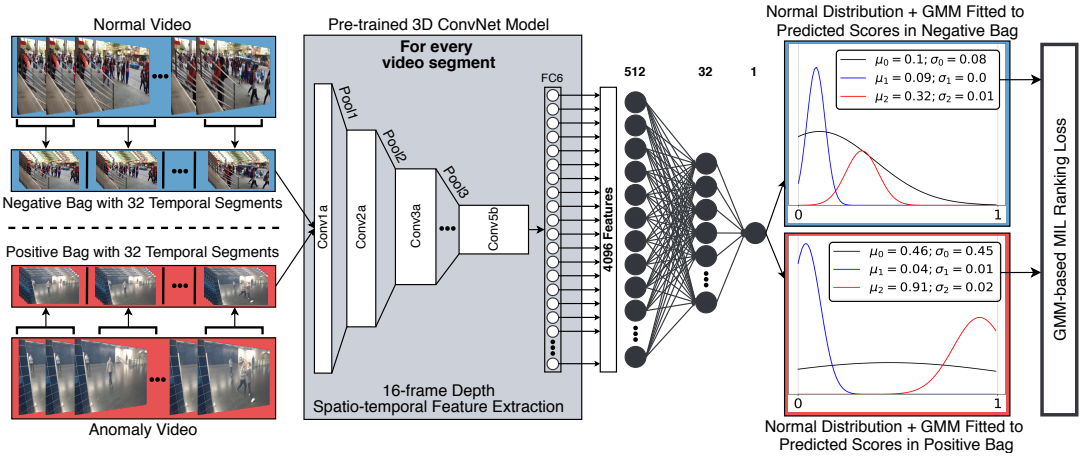


Figure 4.2: The flow diagram of the proposed Gaussian Mixture Model-based approach.

three Gaussians, $\{\mu_0, \sigma_0\}$ from the normal distribution, $\{\mu_1, \sigma_1\}$ from one kernel in the mixture, and $\{\mu_2, \sigma_2\}$ from the other kernel in the mixture, (as shown in figure 4.2).

The aim is to attain maximum divergency between the two kernels of the mixture in positive bags, and minimum divergency in negative bags. Therefore, the obvious conditions to satisfy are a ranking loss to encourage a higher mean value in one kernel compared to the other for anomaly videos, and equality between Gaussians for normal videos, such as:

$$\mathcal{G}(\mathcal{B}_a) \stackrel{\text{def}}{=} \{\mu_j < \mu_k\}, \forall j, k \in \{1, \ldots, \mathcal{K}\} \wedge j \neq k, \quad (4.9)$$

$$\mathcal{G}(\mathcal{B}_n) \stackrel{\text{def}}{=} \{\mu_j = \mu_k\}, \forall j, k \in \{1, \ldots, \mathcal{K}\} \wedge j \neq k, \quad (4.10)$$

where $\mathcal{G}(\mathcal{B}_a)$ and $\mathcal{G}(\mathcal{B}_n)$ represent the GMM in positive and negative bags, respectively, and $\mathcal{K}$ denotes the number of kernels in our mixture.

Consequently, condition (4.9) reaches its maximum divergency when $\mu_j = 0$ and $\mu_k = 1$, while condition (4.10) reaches its minimum divergency if both mean values are equal. Considering that we need to penalize the model, whenever the above conditions are not satisfied, the straightforward approach for the objective function would be to incorporate both conditions in a hinge loss, given as follows:

$$l\big(\mathcal{G}(\mathcal{B}_a), \mathcal{G}(\mathcal{B}_n)\big) = \max\big(0, 1 - \overbrace{\left|\mu_k^a - \mu_j^a\right|}^{①} + \overbrace{\left|\mu_k^n - \mu_j^n\right|}^{②}\big), \quad (4.11)$$

which means ① reaches its maximum value when each mean values of the kernels in $\mathcal{G}(\mathcal{B}_a)$ are in both extremes of the unit interval, maximizing the divergence of condition (4.9). On the counterpart, ② reaches its minimum value when both means in $\mathcal{G}(\mathcal{B}_n)$ have

the same value, minimizing the divergency of condition (4.10).

In spite of the above ranking function producing low and high scores for anomaly videos and produce minimum divergent scores in normal videos, until now the mean values of normal videos despite being equals, there is no constraint to which value reach. Therefore, since we need to enforce negative bags' scores only in the lower extreme of the unit interval and still satisfying condition (4.9) for positive bags, we measure the parameters of the normal distribution of the predicted scores, and by incorporating these parameters of the normal distribution, the loss function becomes:

$$l\big(\mathcal{N}(\mathcal{B}_u), \mathcal{G}(\mathcal{B}_u)\big) = \max\big(0, 1 - \overbrace{\max\left(\mu_0^a, \left|\mu_k^a - \mu_j^a\right|\right)}^{①} + \overbrace{\max\left(\mu_0^n, \left|\mu_k^n - \mu_j^n\right|\right)}^{②}\big), \quad (4.12)$$

where ① still reach its maximum value when the divergency of condition (4.9) is maximized, since $\mu_j < \mu_0 < \mu_k$. As well as, ② will only reach its minimum value when all mean values are in the lower extreme, minimizing the divergence of condition (4.10) and satisfying the condition of low scores on negative bags, meaning that if the mean values of the mixture only satisfy condition (4.10), the greater the mean value, the greater the penalty, due to the normal distribution.

With this MIL ranking loss, the error is back-propagated from the Gaussians divergences in both negative and positive bags. We presume that by training on a large training set, the FCN will learn a generalized model predicting high scores for anomalous segments in positive bags (as shown in figure 6.6). Furthermore, in comparison to the considered state-of-the-art method [SCS18], which only operates on the maximum scores level in each bag, we expect that a GMM can preserve better information regarding all scores in a bag, and consequently achieving a better distribution score in both negative and positive bags (see figure 6.3). Finally, our complete objective function is given by:

$$\mathcal{L}(\mathcal{W}) = l\big(\mathcal{N}(\mathcal{B}_u), \mathcal{G}(\mathcal{B}_u)\big) + \lambda_1 \|\mathcal{W}\|_F, \quad (4.13)$$

where $\mathcal{W}$ is the classifier to be learned.

4.3 Ranking Model vs. Negative Log-Likelihood Model

As previously stated, this GMM-based loss is based on the principle of the different divergences between the distribution scores of both normal and abnormal videos. On this basis, upon further research, we also studied a new alternative strategy for our ranking model. With the goal of dismissing the imposed conditions to mold distribution shapes, we employed the Negative Log-Likelihood (NLL) to attain a refined model fitting with any conditions engaged regarding the predicted scores directly.

The likelihood function consists of measuring the Goodness of Fit (GOF) of a distribution model to a sample of data points, higher the NLL value, better the fitting. In this strategy, the rationale is that a GMM fitted to a positive bag will result in a two-kernel mixture with drifted apart kernels and achieving a higher NLL value than a normal distribution. On the counterpart, a GMM adjusted to a negative bag will result in a two-kernel mixture with no divergence between the kernels and, consequently, achieving a similar NLL value compared to a normal distribution. The main idea is to shape two different distributions with minimized and maximized divergence for negative and positive bags, respectively, entirely by the model itself just by analyzing the likelihood of the outcome. Accordingly to equations (4.1), (4.3) and (4.4), we can define our negative log-likelihood calculations, which are given by:

$$nll\big(\mathcal{N}(\mathcal{B}_u)\big) = \ln\big(\prod_{i=1}^{N}\mathcal{N}(\mathcal{V}_u^i|\mu_0,\sigma_0^2)\big) \quad (4.14) \qquad nll\big(\mathcal{G}(\mathcal{B}_u)\big) = \ln\big(\prod_{i=1}^{N}\mathcal{G}(\mathcal{V}_u^i|\gamma)\big) \quad (4.15)$$

where N denotes the total number of predicted scores in our bag, $\mathcal{B}_u$. With the above equations, we are capable of comparing the GOF of our two models (normal and mixture), concerning each bag.

With the purpose of shaping a normal and two-kernel distribution for negative and positive bags, respectively. The straightforward approach would be the ratio between both NLL values in each bag, to encourage divergence in positive bags and similarity in negative bags, such as:

$$l\big(\mathcal{N}(\mathcal{B}_u),\mathcal{G}(\mathcal{B}_u)\big) = \max\big(0, 1 - \overbrace{\frac{nll\big(\mathcal{N}(\mathcal{B}_n)\big)}{nll\big(\mathcal{G}(\mathcal{B}_n)\big)}}^{③} + \overbrace{\frac{nll\big(\mathcal{N}(\mathcal{B}_a)\big)}{nll\big(\mathcal{G}(\mathcal{B}_a)\big)}}^{④}\big), \quad (4.16)$$

where ③ reaches its maximum value when the fitting of both normal and Gaussian mixture distributions are identical, meaning that all three Gaussians have the same parameters, thus, the same NLL. On the other hand, ④ reaches its minimum value when the fitting of a two-kernel distribution is better than a normal distribution, ending up with a lower NLL from the normal distribution and a greater NLL from the mixture.

This means, the better a two-kernel distribution fits the predicted scores, more the ratio between NLLs will tend to zero (minimizing the error in positive bags and maximizing it in negative bags), while the ratio between NLLs tends to one the more similar are the normal and mixture distributions (minimizing the error in negative bags and maximizing it in positive bags).

Unfortunately, there is a misconception with the described loss function regarding the maximum NLL value for the two-kernel mixture. During our experiments, we notice that

the NLL value, concerning the mixture, reached its maximum value with extremely unbalanced mixing probabilities between the two kernels, π_k. The unbalanced π_k verified was 97% to 3%, meaning that the NLL's mixture value reaches its maximum when 97% of the predicted scores fall in one of the Gaussians, and the remaining 3% in the second one.

In the context of optimization algorithms, neural networks seek to minimize the error; thus, this gap, in our previous objective function (4.16), creates an extremely unstable classifier. Since our model finds its minimum loss value in the NLLs ratio of the positive bags when the NLL's mixture is maximized, and considering that the Gaussians in negative bags should be identical, there is even more instability, pushing the predicted scores into one of the extreme values of the unit interval, without learning a good classifier.

Several GOF tests [S+78, Pea00, Aka74] of statistical models, rely its confidence on the quantity of the observations in the distribution; therefore, we also tried to side-step this issue with the extraction of 240 temporal segments instead of 32 in each video. However, the same unbalanced percentage values were observed.

We also tried to work around this issue by incorporating a constraint to the basis loss, to penalize the model exponentially whenever the difference between mixing probabilities π_k are greater than 40%, meaning that one of the Gaussians has a $\pi_k > 70\%$, consequently, the other will attain values lower than 30%.

By making use of a softplus activation function [DBB+01], also called as SmoothReLU, which has a similar behavior as the ReLU activation function [NH10b], but it is differentiable at the zero value, combining this constraint, our objective function becomes:

$$l\big(\mathcal{N}(\mathcal{B}_u), \mathcal{G}(\mathcal{B}_u)\big) = \max\big(0, 1 - \frac{nll\big(\mathcal{N}(\mathcal{B}_n)\big)}{nll\big(\mathcal{G}(\mathcal{B}_n)\big)} + \frac{nll\big(\mathcal{N}(\mathcal{B}_a)\big)}{nll\big(\mathcal{G}(\mathcal{B}_a)\big)}\big) + \ln\big(1 + e^{\lambda_1(d-0.4)}\big), \tag{4.17}$$

where λ_1 denotes an exponent constant to control the growth of the exponential, and d denotes the mixing probability difference between the kernels in the Gaussian mixture, such as: $d = \left|\pi_k^u - \pi_j^u\right|, \forall j, k \in \{1, \ldots, \mathcal{K}\} \wedge j \neq k$, and the negative exponent constant 0.4 is a shifting transformation to the right for the exponential growth starting around 0.4 instead of zero, alike the original softplus function.

Despite the exponential penalization constraint, the FCN still provides a bad classifier, predicting only extreme values, which means the network still finds its minimum value of error by maximizing the NLL value. For this reason, the misconception in the NLLs ratio requires a new approach to bypass this issue, and further research into the NLL measurement and GOF tests will be done. Besides the unsuccessful NLL-based loss, we

still provide results of our GMM-based approach using our MIL ranking loss that solidly outperforms the state-of-the-art methods.

Chapter 5

UBI-Fights Dataset

In this chapter, we present an overview of the existing public video abnormal event datasets on surveillance footage for anomaly detection purposes. We present the main characteristics and explore some advantages and limitations of the currently available datasets in addition to some image examples to illustrate each database. Due to those limitations, we present the availability [1] of a new large-scale database fully annotated at the frame-level, the *UBI-Fights* dataset constructed with the purpose of providing distinguishable solutions, in comparison with the existing ones, for further research in the anomaly detection field. Furthermore, we detail the collection process and a comparison between the previous datasets and ours.

5.1 Related Datasets

The research and developments in abnormal events detection require much video image analysis due to the extreme diversity of innumerous anomaly events. Therefore it is crucial to operate on the largest possible number of videos in a dataset, robust enough with respect to the learning domain employed. Regarding the detection algorithms of those events, different dataset scopes can be identified, such as databases staged in invariant locations, which means every video sequence is staged and captured only in one location always with the same camera angle, considering an abnormal event every motion that distinguishes from the usual flow of the entities present in the scenes, usually preferable in the unsupervised domain. On the other hand, for a more challenging approach, there are also databases staged in multivariate scenarios, meaning that in the same dataset, there are multiple different scenes with multiple camera angles, where a human manually states the abnormality event, and for this reason, it is more difficult to assemble and annotate such dataset. Currently, there are several publicly available video databases, presenting divergent solutions and detection protocols that could be used for both supervised and unsupervised learning domain algorithms.

In the following subsections, we describe the characteristics of the current most relevant anomaly event databases, commonly used as baselines for a fair result comparison between state-of-the-art detection methods in this field. With our attention turned to the scene's environment diversity, that each database contains for satisfying different purposes in the learning domains, along with its available ground-truth in each database, facilitating its use for further research. Additionally, we also pay attention to the natural-

[1] `http://socia-lab.di.ubi.pt/EventDetection`

ity of the captured scenes, i.e., if the entities present in the scenes are in normal situations scenarios, or if they are representing on purpose, meaning that it was synthetically made.

Since our current proposals are based on weak and self-supervised learning, we considered the analysis of these factors as the most relevant parameters, concerning the domain of supervised learning, for our new publicly available dataset.

5.1.1 UMN Dataset

The University of Minnesota (*UMN*) anomaly detection dataset [UMN] consists of a synthetic dataset to simulate crowded escape events, reproducing panic movement in populated areas. A normal crowd behavior, people walking around the scene, is observed at each video sequence beginning until at some point the behavior rapidly evolves into a panic crowded movement where all persons in the scene disperse from a central point. The *UMN* dataset contains 11 anomaly video sequences of three different scenes. The database includes colored and grayscale scenes, as well as indoor and outdoor scenarios (as shown in figure 5.1) however, the collected outdoor videos were only taken during the daytime. In spite of the small number of video sequences, any annotation is provided.

Figure 5.1: Illustration images of the scenes from the *UMN* dataset.

5.1.2 USCD Peds1 and Peds2 Datasets

The University of California San Diego has developed two anomaly detection subsets acquired from two stationaries cameras, in elevation, pointed out to pedestrians walkways in natural circumstances, *UCSD-Peds1* and *Peds2* [LMV13]. The main characteristics of both datasets are the circulations of non-pedestrians entities, some of which are small carts, skaters, and bikers, which causes an abnormal event among the pedestrians. Additionally, it is also considered as an anomaly, the abnormal pedestrian's motion patterns, such as people not walking in the same flow path as others, for instance, crossing ways or walking in the grass.

The two subsets correspond to two different scenes, and both datasets' video sequences were trimmed to approximately 200 frames. Both subsets include normal and abnormal situations; the subset *Peds1* contains 70 video sequences, while the *Peds2* only contains 28 videos. In each subset, the frame-level annotations of the whole dataset are provided; additionally, the test sets are also annotated at the pixel-level with binary masks to identify

the abnormal local regions. However, this dataset (as shown in figure 5.2) is only available in grayscale, and the video sequences were collected only during the daytime.

Figure 5.2: Illustration images of the scenes from the *UCSD Pedestrians* dataset.

5.1.3 CUHK-Avenue Dataset

The Chinese University of Hong Kong (*CUHK*) built a new univariate database for abnormal events detection in an outdoor environment, the *CUHK-Avenue* dataset [LSJ13], where the video sequences of normal situations scenarios, were captured from a static camera pointed to one avenue of the campus. The normal patterns of this dataset are pedestrians movement parallel to the camera plane, considering as outliers, every person that deviates from the normal path or motion (running), and if some abnormal object is present in the scene, such as bicycles, as shown in figure 5.3. The database includes normal and abnormal situations during the daytime, with 37 colored video sequences annotated at the frame-level and provides the bounding boxes of each object or person with an abnormal motion in the scene.

Figure 5.3: Illustration images of the scenes from the *CUHK-Avenue* dataset.

5.1.4 Street Scene Dataset

The North Carolina State University, in cooperation with Mitsubishi Electric Research Labs, developed a new univariate video dataset for anomaly detection, denominated as the *Street Scene* dataset [RJ20], containing video sequences of a two-lane street with bike lanes and pedestrian sidewalks in daily real-life scenarios. This database is composed of 81 videos of normal and abnormal events, captured from a fixed camera pointed down to the street, which is a challenging dataset presenting a variety of activities, such as daily traffic, biking, and pedestrians' flow. Despite the surveillance footage taken only during the daytime, it includes video sequences during different times of the day, presenting different lights and shadows. Ground-truth annotations are provided with this database, including bounding boxes, and pose tracking associated with each box.

Figure 5.4: Illustration images of the scenes from the *Street Scene* dataset.

5.1.5 ShanghaiTech Campus Dataset

The ShanghaiTech University presented a new, more challenging, and complex dataset in the anomaly detection field, designated as *ShanghaiTech Campus* Dataset [LLLG18], consisting of 130 video sequences across 13 different locations in their campus. Presenting a variety of real-life pedestrians' activity, different angles capturing different scenes, and with an outdoor environment in this database, it is considered as an abnormal event every entity that distinguishes from normal pedestrian's behavior, such as running, biking, or presenting abnormal objects into the scene like strollers and skaters. The *ShanghaiTech Campus* dataset is captured from a fixed camera in every different scene, providing colored images. The pixel-level ground-truth of the abnormal events is provided in the annotations of this dataset.

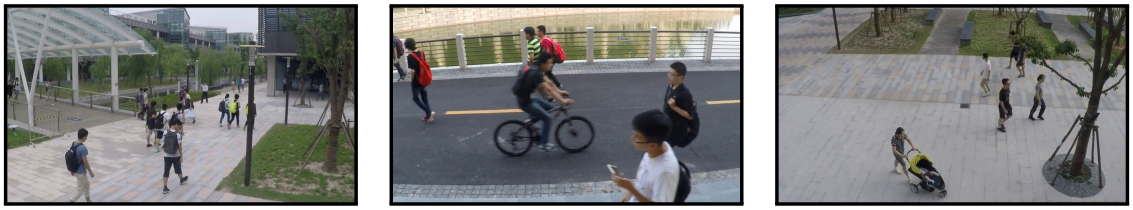


Figure 5.5: Illustration images of the scenes from the *ShanghaiTech Campus* dataset.

5.1.6 UCF-Crime Dataset

The University of Central California has developed the most challenging large-scale dataset in the abnormal event detection scope until now, the *UCF-Crime* dataset [SCS18]. Consisting of 1900 untrimmed video sequences completely wild, which means the videos in the database are not related in any way to any video within this dataset, and they are all captured from different scenes, the dataset length is comprised of 128 hours of video. It includes 13 different realistic anomalies, such as abuse, arrest, arson, assault, accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, and vandalism, in addition to normal situations scenarios. All selected to have a significant impact on public safety. This dataset has an enormous variety concerning the environment of different scenes, including captured indoor and outdoor scenarios, daytime and nighttime lighting effects, colored and grayscale images, fixed and rotated cameras, and divergent ethnicity among the humans, including only realistic scenarios. Such a dataset is not easy to annotate, and they only provide the frame-level ground-truth annotations of their testing set, while the remaining videos are annotated at the video-level.

(a) Explosion scenario. (b) Road accident scenario. (c) Vandalism scenario.

Figure 5.6: Illustration images of the scenes from the *UCF-Crime* dataset.

5.2 New Dataset

Considering the studied above described abnormal event databases, we came to the conclusion of the necessity of a large-scale dataset that satisfies all the gaps in the relevant parameters analyzed in the existing databases. Apart from the *UCF-Crime* dataset [SCS18], which presents a very rich and robust large-scale dataset, it does not provide a fully annotation at the frame-level in the training set, and despite the advantages of having 13 anomaly scenarios in both training and testing sets, there is a lack of unbalanced videos between individual anomaly scenarios and normal video sequences in their testing set. The remaining public datasets are considered not very divergent, and are more suitable to attain low levels of performance to the required for an effective deployment of this kind of technology in fully unconstrained scenarios that distinguishes from the learning videos within the dataset.

Based on these factors, we decided to create a new publicly available [2] large-scale anomaly detection dataset, the *UBI-Fights* [DP20], with the essential characteristics that distinguish from the existing public ones. It consists of a specific anomaly detection first of its kind dataset of only fighting scenarios and normal daily life situations.

The *UBI-Fights* dataset is composed of 80 hours of video fully annotated at the frame-level, consisting of 1000 wild video sequences, where 216 videos contain a real-life fighting event, and 784 videos contain only normal daily life scenarios. The videos were collected from *YouTube* and *LiveLeak* using a variety of text search queries for fight events, and in order to assemble as many videos as possible, we also used text searches in different languages (e.g., Portuguese, Spanish, English, French, Italian, Russian, German, Chinese) to obtain fighting events. Additionally, for consistency purposes, we removed all unnecessary video segments that could disturb the learning process of the methods executed on this database, giving us the advantage of taking profit of fight events taken from news, within compilation videos, or manually edited ones. We also included rotated cameras, as well as the most stable movable camera videos, which are slightly shaky, were also collected to reach the most divergence motions regarding this abnormal event. This dataset provides the frame-level ground-truth annotations of all video sequences in the database, as well as the environment and appearance flags, are contained in each video title, such

[2] `http://socia-lab.di.ubi.pt/EventDetection`

as environment (indoor or outdoor), camera type (fixed, rotated or movable), and color type (RGB or grayscale).

5.2.1 Collection and Preprocessing

As stated above, all fighting videos and some normal videos were collected from *YouTube* and *LiveLeak*, the remaining normal videos were collected from previous public databases [FBLF07, PMRC17, FSVC05] due to its very difficult search and acquisition.

Due to the variety of video types over the internet, preprocessing was required, resizing all videos to the dimension of 640×360 pixels. Every video with a different ratio was adjusted with white-colored margins with the required size to avoid distortions, and the frame-rate was normalized to 30 frames-per-second (fps). Figure 5.7 illustrates some images from the video sequences of the *UBI-Fights* dataset, and figure 5.8 presents the statistics of the durations regarding every video in the database.

(a) Indoor scenario. (b) Outdoor scenario. (c) Grayscale video sequence.

(d) Nighttime scenario. (e) Movable camera. (f) Normal daily situation.

Figure 5.7: Illustration images of the scenes from the *UBI-Fights* dataset.

5.2.2 Annotation Process

For uniformity purposes, the same abnormal boundary criteria was used through the annotations of every video sequences, considering the beginning of the fighting event, the moment when the aggressor(s) touch the victim(s), and the end of the event, the moment the aggressor(s) walk away from the victim(s).

The annotation of such a dataset requires much effort and time-consuming. To facilitate the whole process, we developed a new annotation application not only to annotate the abnormal events' ground-truth frames but also to trim video sequences to remove undesired frames or even to create subsequences from compilations, to take advantage of every video found in the collection process. Figure 5.9 illustrates the application used in the annotation process of the *UBI-Fights* dataset.

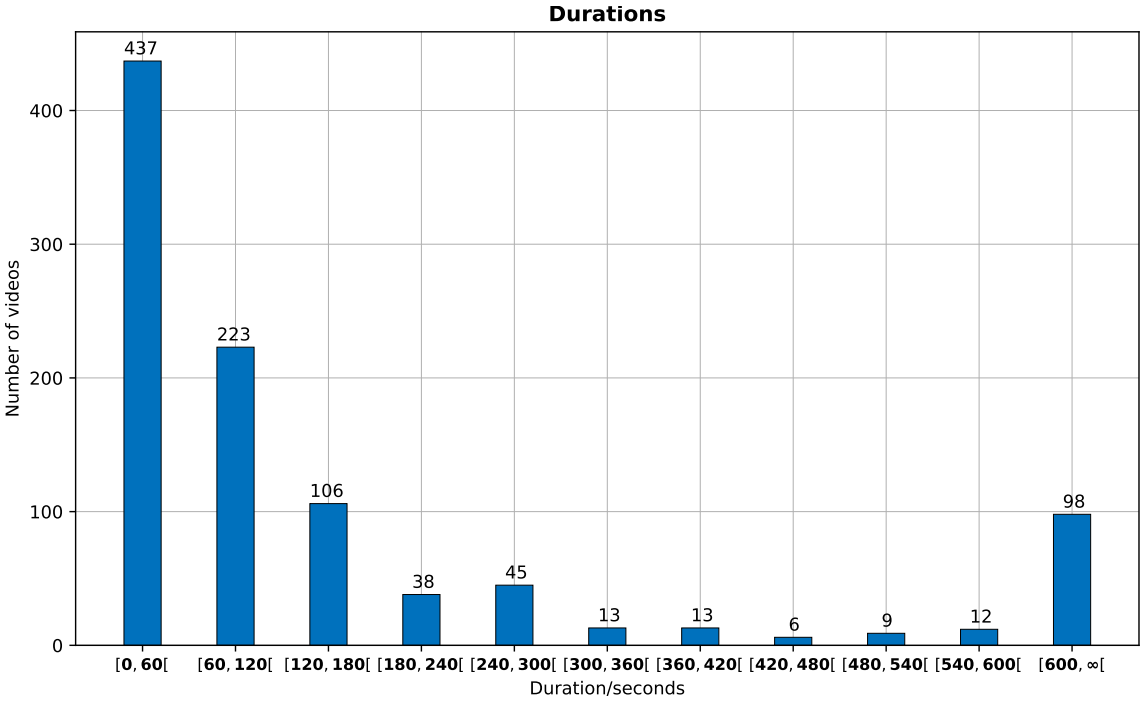


Figure 5.8: Statistics of all video sequences in the *UBI-Fights* dataset.

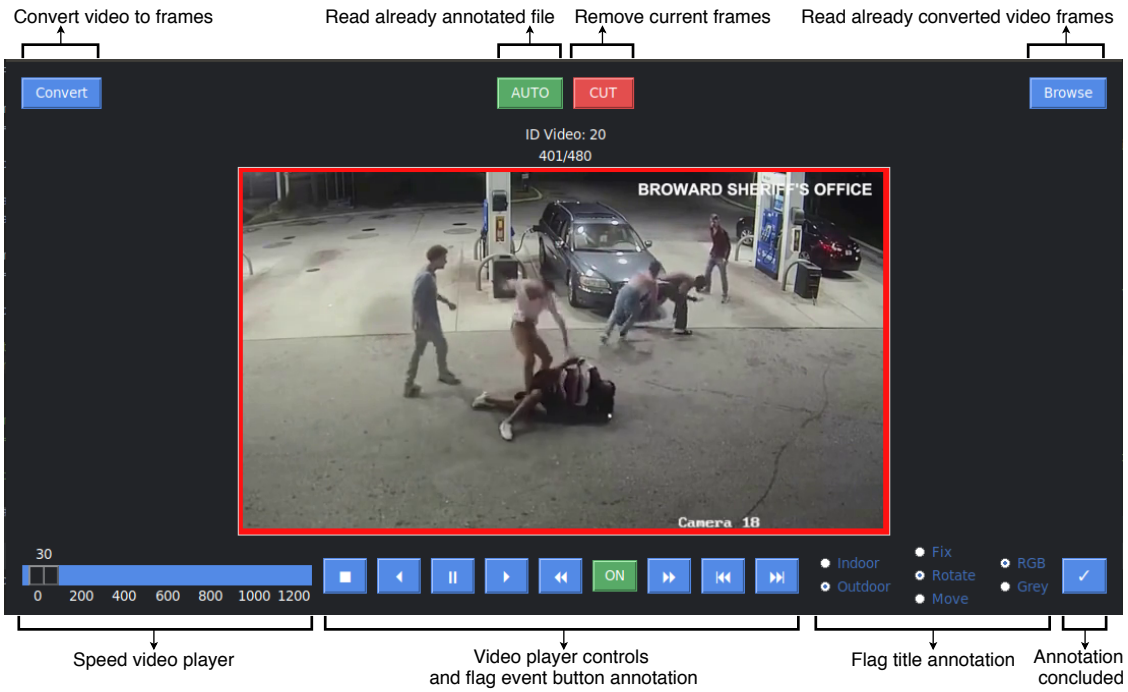


Figure 5.9: New developed annotation aplication used for the *UBI-Fights* dataset.

Chapter 6

Experiments and Discussion

In this chapter, we present the results of our experiments regarding both proposed approaches in this work. The experiments were engaged in two different ways: 1) in order to employ and compare the self-supervised learning experiments and taking into account the one premise that a set of unsupervised data is available, we disregarded the ground-truth labels of a subset in the learning sets, considering those subsets as unlabelled, and for a fair result comparison between the state-of-the-art and our WS/SS framework approach [DP20], all methods used rigorously the same manually annotated labels for learning purposes, therefore, in practice, each method starts training with the same learning set; 2) experiments on the Gaussian Mixture Model-based (GMM) approach were normally performed with all the available labels on the training set in each dataset, also using a fair result comparison using the same learning sets in the other methods. Hence, to guarantee a fair evaluation between the state-of-the-art approaches in abnormal detection and our proposals, all methods, in both ways of the experiments, used exactly identical testing sets within the respective experiment.

Finally, as our final experiment to examine the performance of our models in the real-world, we present the qualitative results consisting of executing our best model through some real-case fighting scenarios.

6.1 Datasets, Baselines, and Empirical Evaluation Protocol

With our two proposed methods, the experiments were conducted in three datasets (*UBI-Fights*, *UCF-Crime* [SCS18], and *UCSD-Peds1* [LMV13]). The *UBI-Fights* dataset was split into three disjoint subsets: 80% for the learning set, 5% for the validation set, and 15% for the testing phase. For the *UCF-Crime* [SCS18] dataset, we used the default learning data split, and for validation purposes, we use 50% of the testing set, randomly chosen. While for the *UCSD-Peds1* [LMV13], which the learning set only contains normal videos and the testing set only contains abnormal videos, we randomly move abnormal videos to the learning set, not belonging to the testing set anymore, and randomly move normal videos to the testing set, not including them in the learning set. Moreover, we use 50% of the final testing set only for validation purposes.

Essentially, in the WS/SS framework, the validation set is used by both Bayesian classifiers in order to obtain the distributions of both positive and negative instances, as well as the pattern classifier's learning operates over the SS expert's response scores in each system's iteration with the instance's ground-truth.

In order to employ the WS/SS framework, we trimmed the videos of the *UBI-Fights* and the *UCF-Crime* [SCS18] datasets into sub-videos with a fixed length (16 seconds), having manually adjusted the corresponding annotations. Since the *UCSD-Peds1* [LMV13] is already with a fixed length of 200 frames, no changes were made in those videos. In order to test the GMM-based approach, we use the original videos of each dataset, with no changes needed. Next, for every 16-frame clip (i.e., one segment) for the three datasets, the sixth fully-connected layer (*FC6*) of the C3D network [TBF+15] was employed to extract the spatiotemporal descriptors. Then, for the bag formation in our methods and for the Sultani *et al.*'s proposal [SCS18], we obtain the bag's features by taking the average of all 16-frame video clip features within that segment, using these features as input to the abovementioned methods.

We also converted the *UCSD-Peds1* [LMV13] to a frame rate of 30 fps since the C3D model was trained in videos with approximately 30 fps, for this reason, a lower frame rate will result in a higher motion captured, which will result in tremendous errors employing the SS expert, since it operates at the 16-frame clip level. On the other hand, the WS and GMM experts work at the bag level, where the temporal segments are extracted by the mean of each 16-frame clip, the errors are not so evident, but for precautions, the conversion to 30 fps was made in all experiments.

As baselines, five methods, regarding both one-class and binary classification, were considered to represent the state-of-the-art: 1) Sultani *et al.* [SCS18], analyzing both normal and abnormal videos, the model is trained with weakly-labeled videos producing high scores for anomaly events; 2) the work due to Ravanbakhsh *et al.* [RNS+17] training two networks with only normal events, one to generate optical-flow from frames and other to generate frames from optical-flow, detecting abnormal pixels for frame-level detection; 3) Chong and Tay [CT17] training a temporal AE with a spatial feature extractor, by using only normal videos, and detecting the anomaly by its reconstruction error; 4) Hasan *et al.* [HCN+16] by using improved trajectories features of only normal videos to train the AE, also using the reconstruction error to predict the anomaly; 5) Wang *et al.* [WQL+18], also under the one-class criterion, extracting foregrounds and optical-flow with a FCN, and filtering spatiotemporal features with a VAE, detecting the anomaly with the reconstruction error of a second VAE combined with a convolutional neural network. Each baseline is described and detailed in chapter 2. Apart these, we also tested the performance attained by a quadratic kernel SVM working under the single-class paradigm, fed by the same C3D feature vectors as our GMM, WS and Sultani *et al.* models.

The detection protocol used in each baseline was frame-level detection, and three performance measures are reported: the decidability index (d'), the AUC value, and the Equal Error Rate (EER), also providing the corresponding Receiver Operating Characteristic (ROC) curves.

6.2 Implementation Details

For both the WS and GMM-based experts, we split each video into 32 non-overlapping temporal segments, and the average is taken to all 16-frame features within those segments, using them as input to the FCN in both these approaches. While to our SS expert, the input instances used are directly the 16-frame video clip features, providing better information. Both networks in the WS/SS framework and the GMM method have the same network architecture, which is composed of a FCN with 512 units with ReLU as activation functions [NH10b], followed by 32 units in the second layer with linear activation, and one unit in the output layer with sigmoid as the activation function. Between FCN, the dropout regularization [SHK+14] used is 60% and applied AdaGrad [DHS11] with a initial learning rate of 0.01. In the three approaches, for a better comparison between them, the only difference lies in the applied ranking loss function, all described in chapters 3 and 4. The pattern classifier, as illustrated in figure 3.1, its architecture is composed of a 4-layer FCN with 256, 196, and 128 units for the first three layers with ReLU activations [NH10b] in each one, followed by 1 unit as the output layer applying sigmoid activation.

In [CT17], we use the temporal encoder-decoder with a 10-frame depth to encode its motion, with $128 : 64 : 32 : 64 : 128$ architecture. In Wang *et al.* [WQL+18], we train the first VAE, which is fed with a two-frame depth motion features extracted by a pre-trained fully convolutional neural network, the second VAE takes as input the filtered features, with the following architecture $256 : 128 : 6 : 128 : 256$. By using the fully convolutional network to feed a deep AE proposed by Hasan *et al.* [HCN+16], we fix a temporal window of 16 frames (near the recommended by the author), and the following AE architecture $512 : 256 : 128 : 64 : 128 : 256 : 512$. In [RNS+17] using cGANs, we use the proposed architecture by Isola *et al.* [IZZE17], training the networks until both generators' losses stop decreasing. Using the work by Sultani *et al.* [SCS18], we extract video features of every 16 consecutive frames (with C3D [TBF+15]) and convert them to 32 non-overlapped temporal segments, training the minimum number of epochs (considering the suggested by the author) and until the AUC value of the validation set is nearly constant.

6.3 Self-Supervised Learning Results Comparison

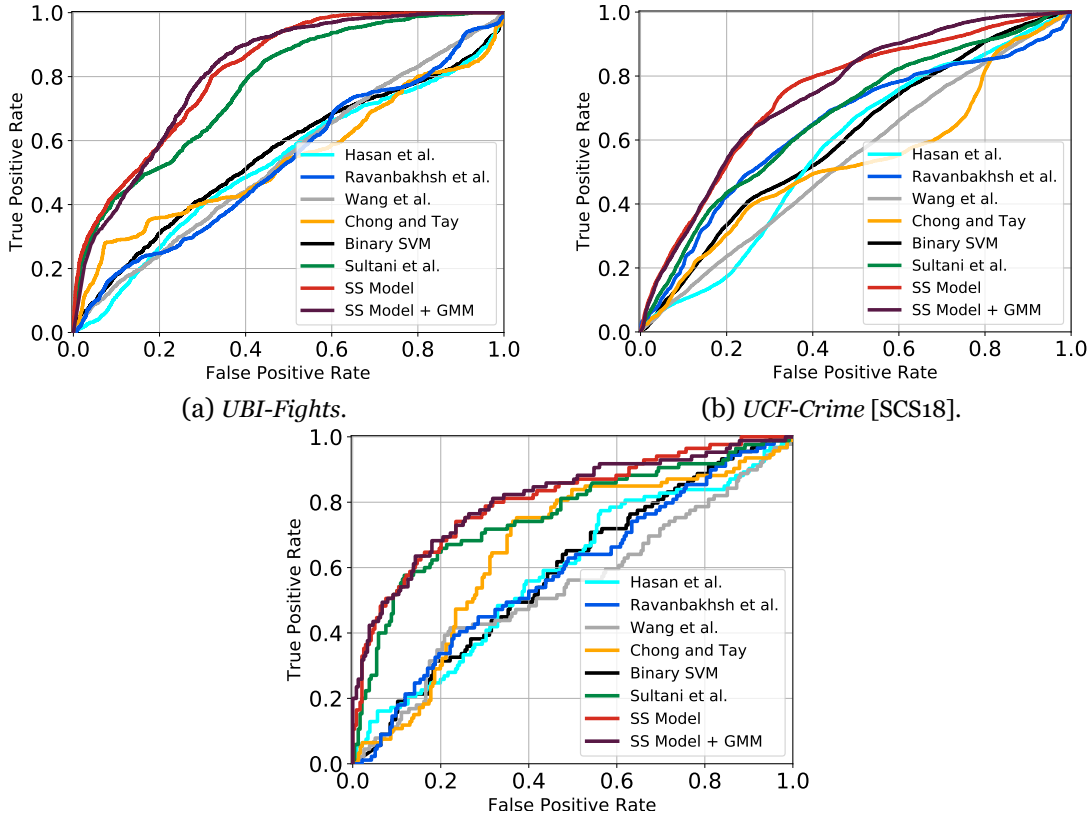
As stated above, we performed the WS/SS framework approach through the three datasets using a subset of our learning set as our unsupervised set for self-supervising purposes, using the same manually labeled learning set for every method. Regarding our WS/SS framework proposal, figure 6.4 provides the evolution in performance using only 30% of the ground-truth in the learning set. For each dataset, it can be observed that the AUC values of both experts in each iteration increase each other, attaining a clear margin from the Sultani's performance (continuous black lines), in exception of the *UCF-Crime* [SCS18], as it is considered a substantial harder task in finding simultaneously consistent patterns among 13 different types of anomalies, which justifies the performance of the WS expert.

We came to the conclusion that our GMM-based approach incorporated into the WS/SS framework results in slightly better performance, considering that the GMM proposal outperforms the Sultani *et al.* [SCS18] method in section 6.4. More importantly, since the Gaussian Mixture approach operates at the distribution level, instead of the maximum value in a bag, by difficulting the learning process with our novel term, the GMM achieves an even better distribution for the Bayesian framework, reducing the peaked distribution in both extremes of the unit interval. The reason to attain peaked distributions, in the lower extreme of the unit interval, in the standard WS/SS framework described in chapter 3, lies on its ranking loss function, which tries to separate the positive scores far apart from the lower values. So, by difficulting this process, the positive scores will be spread over the unit interval; however, the hardest positive scores will stay in the lower values. On the other hand, the GMM-based ranking loss, by working in the distribution level, and trying to separate the mean values, by difficulting this process, only the obvious normal instances will achieve lower values, and only the most reliable positive instances will attain higher values. However, in the *UBI-Fights*, the performance improvements are not so obvious due to the fact that the GMM and Sultani's approach are also quite similar in section 6.4. The final results are summarized in table 6.3 and its ROC curves in figure 6.1.

Additionally, as a proof-of-concept reinforcement of our coupled deep-learning networks system, we performed all the considered methods through each individual anomaly scenes in the *UCF-Crime* dataset [SCS18], i.e., we separated each anomaly scene from the remaining ones, so that each method can be trained with normal scenes and the corresponding scene individually, each approach was also evaluated with the normal and respective scene of the testing set provided by the *UCF-Crime* dataset [SCS18]. Each method rigorously starts its training with exactly the same 30% of the learning set in each scene. Table 6.1 provides the performance summary of every individual anomaly scene experiment. The first two scenes presented an extremely unbalance between normal and abnormal videos.

Individual UCF-Crime Scenes													
Method	**Abuse**	**Arrest**	**Arson**	**Assault**	**Burglary**	**Explosion**	**Fighting**	**Road Accident**	**Robbery**	**Shooting**	**Shop-lifting**	**Stealing**	**Vanda-lism**
[HCN$^+$16]	0.599	0.748	0.718	0.894	0.646	0.596	0.526	0.543	0.552	0.560	0.589	0.685	0.642
[RNS$^+$17]	0.558	0.617	0.764	0.633	0.576	0.545	0.536	0.549	0.609	0.677	0.633	0.533	0.522
[WQL$^+$18]	0.621	0.755	0.504	0.664	0.514	0.673	0.559	0.620	0.688	0.524	0.511	0.674	0.501
[CT17]	0.644	0.709	0.807	0.699	0.546	0.571	0.519	0.545	0.516	0.526	0.792	0.581	0.542
SVM	**0.702**	**0.865**	0.706	0.703	0.644	0.513	0.519	0.684	0.512	0.719	0.512	0.656	0.604
[SCS18]	0.589	0.581	0.921	0.688	0.843	0.819	0.805	0.944	0.638	0.642	0.665	0.881	0.842
SS Model	0.545	0.711	**0.955**	**0.902**	**0.915**	**0.878**	**0.864**	**0.962**	**0.695**	**0.746**	**0.873**	**0.937**	**0.896**

Table 6.1: Performance summary, concerning the individual training in each scene of the *UCF-Crime* dataset [SCS18], of the proposed weak/self-supervised method with respect to the state-of-the-art. The result values correspond to the AUC.



(a) *UBI-Fights*. (b) *UCF-Crime* [SCS18].

(c) *UCSD-Peds1* [LMV13].

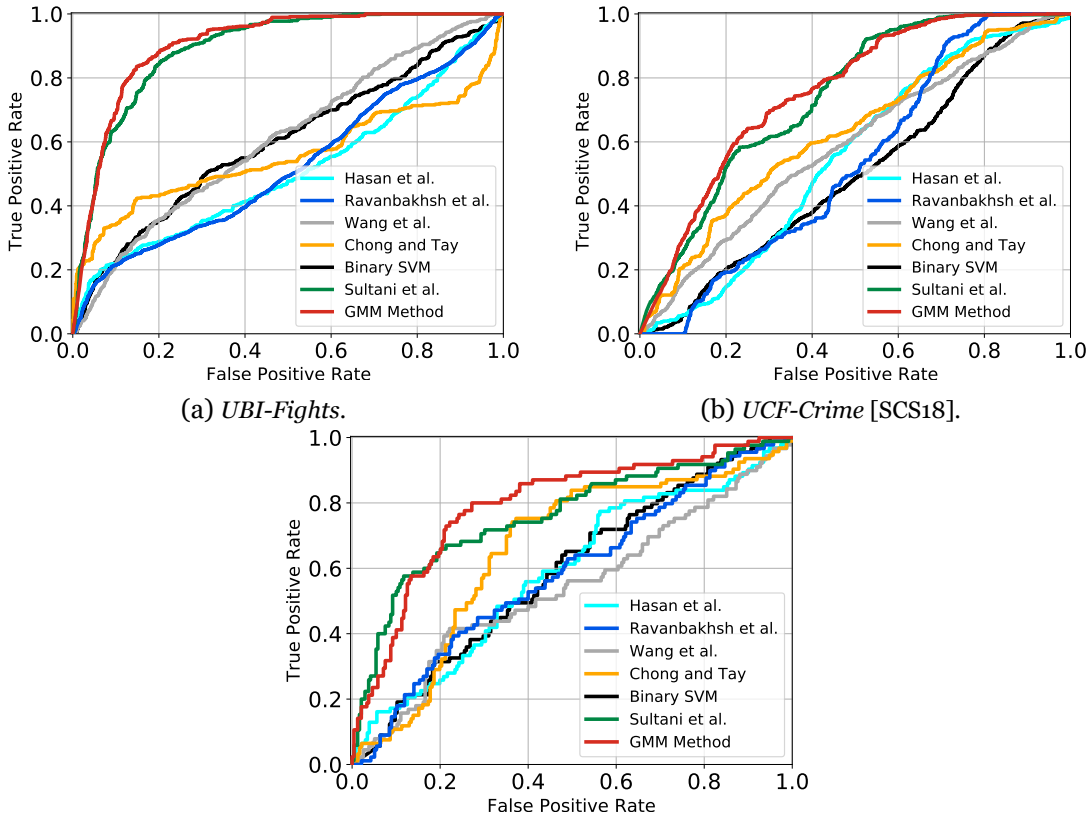
Figure 6.1: Comparison between the ROC curves obtained by the eight methods considered in comparison with the WS/SS framework, for the *UBI-Fights*, *UCF-Crime* [SCS18] and *UCSD-Peds1* [LMV13] datasets.

6.4 Weakly-Supervised Learning Results Comparison

As the second part of our experiments, and with the proposed GMM-based approach in the weakly-supervision paradigm, we also present the performance through the three datasets employed with each training set only weakly-labeled. Using the configuration splits stated above for the GMM experiments. Figure 6.2 illustrates the ROC curves from the summary of the performance presented in table 6.2.

The immediate observation was the extremely poor performance that single-class techniques attain in this type of problem, in accordance with an observation drawn by Sultani *et al.* [SCS18]. These techniques fail in a disastrous way to appropriately establish a boundary between normality and abnormality in test data. From our observations in the *UBI-Fights*, and *UCF-Crime* dataset [SCS18], this might occur due to the extreme heterogeneity of the expected inputs. On the other hand, in the *UCSD-Peds1* dataset [LMV13], despite its homogeneity of the inputs, the poor performance levels are due to the required additional (enormous) amounts of learning data to become minimally effective, since in this training set we reduce the normal instances, including the remaining ones in the test set, to split the abnormal instances in both training and test set for a fair comparison with the GMM and Sultani's approach. Concerning the Sultani *et al.* [SCS18]'s proposal,

their method performed well in all datasets, but there is still a very large amount of false positives, as shown in figure 6.3. Regarding our proposal, we were able to outperform the state-of-the-art in the weakly-supervised learning paradigm, reaching not only better AUC values, but also much better decidability index and EER essentially due to our better distribution scores in the test set, shown in figure 6.3.



(a) *UBI-Fights*. (b) *UCF-Crime* [SCS18].

(c) *UCSD-Peds1* [LMV13].

Figure 6.2: Comparison between the ROC curves obtained by the seven methods considered in comparison with our GMM-based proposal, for the *UBI-Fights*, *UCF-Crime* [SCS18] and *UCSD-Peds1* [LMV13] datasets.

6.5 Qualitative Results

For our third and final experiment, and as a proof-of-concept system with collaboration from the technological company *TOMI WORLD*, we have collected our own simulated real-case scenarios captured from *TOMI* devices located in Lisbon, Portugal. In order to perform the experiments in divergent conditions, we conducted the simulations in two *TOMI* devices, one in an indoor environment (located inside a subway station in Lisbon) and the second one in an outdoor environment (located in Lisbon's streets). Figure 6.5 illustrates the *TOMI* devices where the scenes were captured.

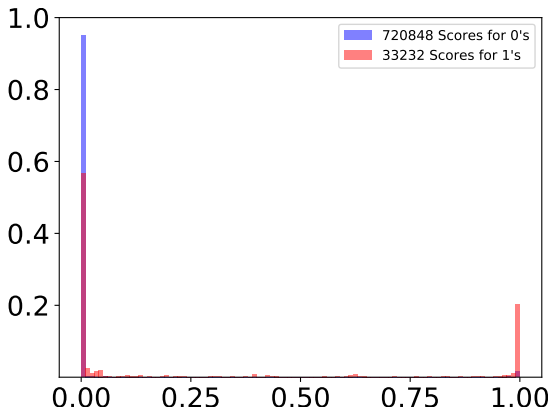
The simulations were captured with the normal crowd's flow around the environment in each *TOMI* device for natural behavior consistency. Nine persons were involved in the simulation of the fighting scenes, where we personally supervised the acting scenes to

provide different scenes, different clothes in the same actor, different numbers of characters involved in the scene, different distances from the *TOMI* device, and every detail was planned to provide as much divergence as possible between the captured scenes.

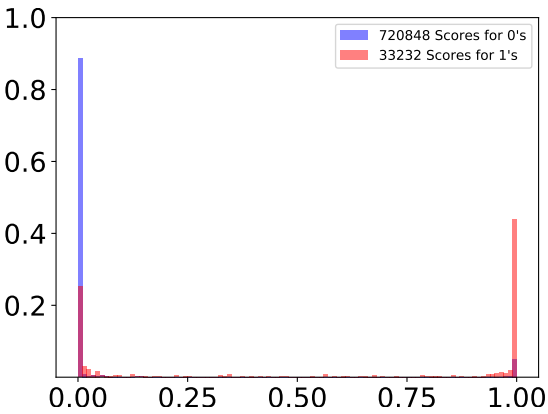
As a concluding remark, even though our proposed method provided - overall - satisfying results in each scene, there is still a gap between the performance attained in indoor and outdoor scenarios, where the reported scores typically are much more irregular than in the indoor settings. We also observed that by simulating different types of fights (standing and in the ground), our model identifies better the abnormal motions in standing fights in comparison to fights taken to the ground where the movement is not so radical. Consequently, in far distance fights, the motion is also more difficult to obtain. Concerning the number of persons present in the fight, obviously, when more people are involved in the event, the abnormal motions increase, being easier to identify.

Method	AUC	d'	EER
UBI-Fights			
Hasan *et al.* [HCN+16]	0.510 ±0.006	0.265 ±0.013	0.516 ±0.004
Ravanbakhsh *et al.* [RNS+17]	0.523 ±0.003	0.011 ±0.006	0.504 ±0.003
Wang *et al.* [WQL+18]	0.610 ±0.005	0.323±0.011	0.427 ±0.003
Chong and Tay [CT17]	0.557 ±0.006	0.540 ±0.18	0.469 ±0.006
Binary SVM Classifier	0.602 ±0.007	0.357 ±0.023	0.429 ±0.006
Sultani *et al.* [SCS18]	0.892 ±0.001	0.804 ±0.009	0.186 ±0.001
GMM Method	**0.906** ±0.001	**1.386** ±0.011	**0.160** ±0.001
UCF-Crime			
Hasan *et al.* [HCN+16]	0.552 ±0.002	0.041 ±0.015	0.448 ±0.003
Ravanbakhsh *et al.* [RNS+17]	0.548 ±0.003	0.063 ±0.021	0.451 ±0.004
Wang *et al.* [WQL+18]	0.587 ±0.004	0.064 ±0.008	0.437 ±0.003
Chong and Tay [CT17]	0.627 ±0.005	0.449 ±0.016	0.406 ±0.007
Binary SVM Classifier	0.504 ±0.004	0.067 ±0.014	0.508 ±0.005
Sultani *et al.* [SCS18]	0.743 ±0.001	0.613 ±0.006	0.353 ±0.002
GMM Method	**0.759** ±0.001	**0.885** ±0.005	**0.302** ±0.002
UCSD-Peds1			
Hasan *et al.* [HCN+16]	0.586 ± 0.011	0.289 ± 0.037	0.427 ± 0.011
Ravanbakhsh *et al.* [RNS+17]	0.583 ±0.016	0.159 ±0.006	0.445 ±0.013
Wang *et al.* [WQL+18]	0.543 ±0.013	0.101 ±0.048	0.473 ±0.012
Chong and Tay [CT17]	0.653 ±0.008	0.240 ±0.032	0.347 ±0.006
Binary SVM Classifier	0.588 ±0.009	0.307 ±0.028	0.439 ±0.009
Sultani *et al.* [SCS18]	0.759 ±0.009	0.794 ±0.042	0.297 ±0.011
GMM Method	**0.801** ±0.005	**1.001** ±0.019	**0.236** ±0.009

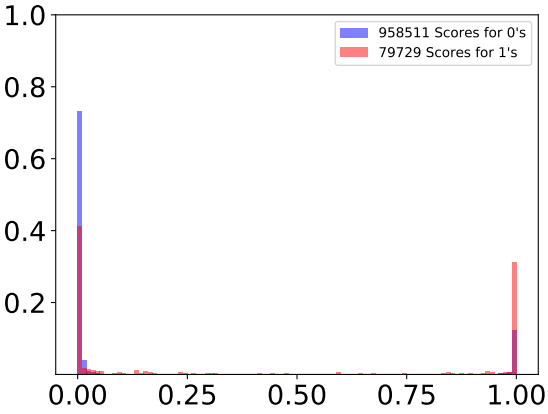
Table 6.2: Performance summary of the GMM method with respect to the state-of-the-art.



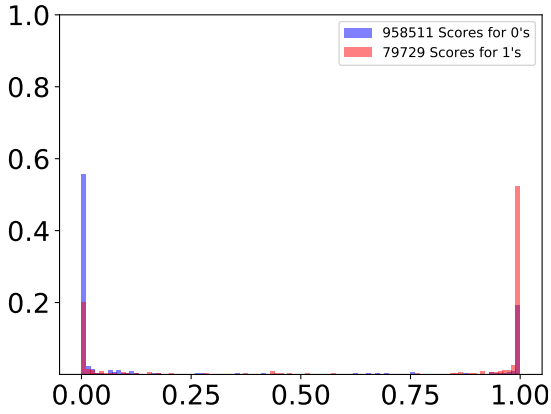
(a) Sultani's approach distribution scores on the *UBI-Fights* test set.



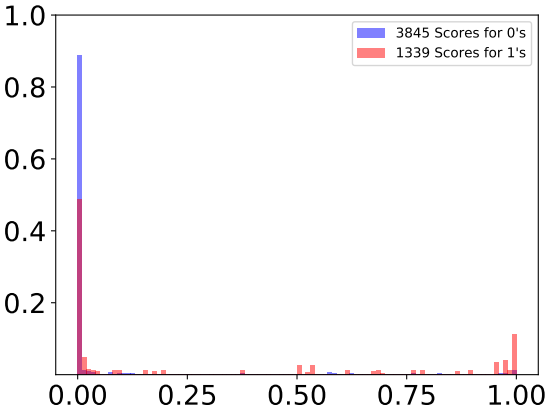
(b) GMM-based proposal distribution scores on the *UBI-Fights* test set.



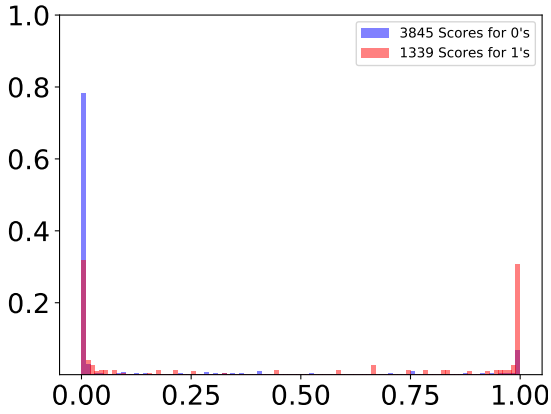
(c) Sultani's approach distribution scores on the *UCF-Crime* test set [SCS18].



(d) GMM-based proposal distribution scores on the *UCF-Crime* test set [SCS18].



(e) Sultani's approach distribution scores on the *UCSD-Peds1* test set [LMV13].

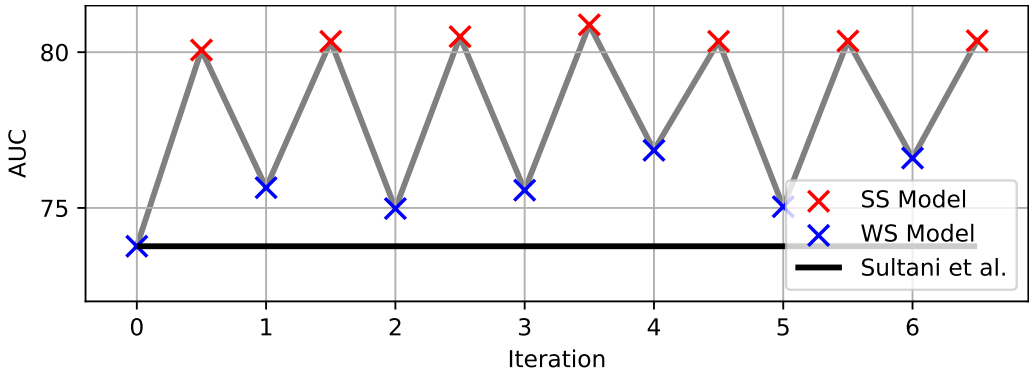


(f) GMM-based approach distribution scores on the *UCSD-Peds1* test set [LMV13].

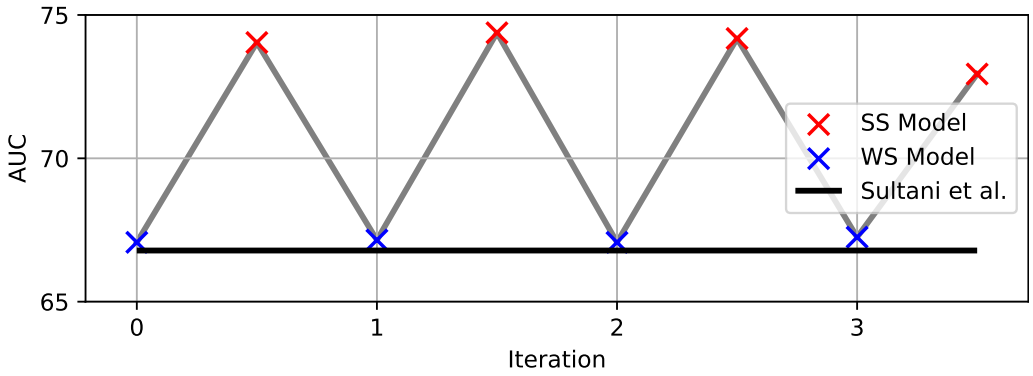
Figure 6.3: Distribution scores comparison between Sultani [SCS18]'s and our GMM-based approaches in the three datasets.

Method	AUC	d'	EER
UBI-Fights			
Hasan *et al.* [HCN+16]	0.528 ±0.003	0.194 ±0.007	0.466 ±0.002
Ravanbakhsh *et al.* [RNS+17]	0.533 ±0.003	0.147 ±0.001	0.484 ±0.002
Wang *et al.* [WQL+18]	0.540 ±0.002	0.164±0.008	0.475 ±0.002
Chong and Tay [CT17]	0.541±0.003	0.059 ±0.001	0.480 ±0.001
Binary SVM Classifier	0.556 ±0.003	0.128 ±0.012	0.443 ±0.003
Sultani *et al.* [SCS18]	0.787 ±0.002	0.738 ±0.011	0.294 ±0.002
SS Model	**0.819** ±0.001	**0.986** ±0.008	**0.284** ±0.001
SS Model w/ GMM Model	**0.820** ±0.001	**1.291** ±0.014	**0.262** ±0.001
UCF-Crime			
Hasan *et al.* [HCN+16]	0.573 ±0.001	0.167 ±0.006	0.424 ±0.001
Ravanbakhsh *et al.* [RNS+17]	0.642 ±0.001	0.423 ±0.005	0.376 ±0.001
Wang *et al.* [WQL+18]	0.539 ±0.001	0.104 ±0.004	0.472 ±0.001
Chong and Tay [CT17]	0.532 ±0.002	0.138 ±0.006	0.484 ±0.001
Binary SVM Classifier	0.604 ±0.001	0.377 ±0.003	0.441 ±0.001
Sultani *et al.* [SCS18]	0.668 ±0.001	0.523 ±0.004	0.375 ±0.001
SS Model	**0.744** ±0.001	**0.764** ±0.005	**0.305** ±0.001
SS Model w/ GMM Model	**0.758** ±0.003	**0.941** ±0.009	**0.304** ±0.004
UCSD-Peds1			
Hasan *et al.* [HCN+16]	0.586 ± 0.011	0.289 ± 0.037	0.427 ± 0.011
Ravanbakhsh *et al.* [RNS+17]	0.583 ±0.016	0.159 ±0.006	0.445 ±0.013
Wang *et al.* [WQL+18]	0.543 ±0.013	0.101 ±0.048	0.473 ±0.012
Chong and Tay [CT17]	0.653 ±0.008	0.240 ±0.032	0.347 ±0.006
Binary SVM Classifier	0.588 ±0.009	0.307 ±0.028	0.439 ±0.009
Sultani *et al.* [SCS18]	0.759 ±0.009	0.794 ±0.042	0.297 ±0.011
SS Model	**0.807** ±0.005	**1.119** ±0.015	**0.255** ±0.008
SS Model w/ GMM Model	**0.817** ±0.006	**1.218** ±0.051	**0.252** ±0.006

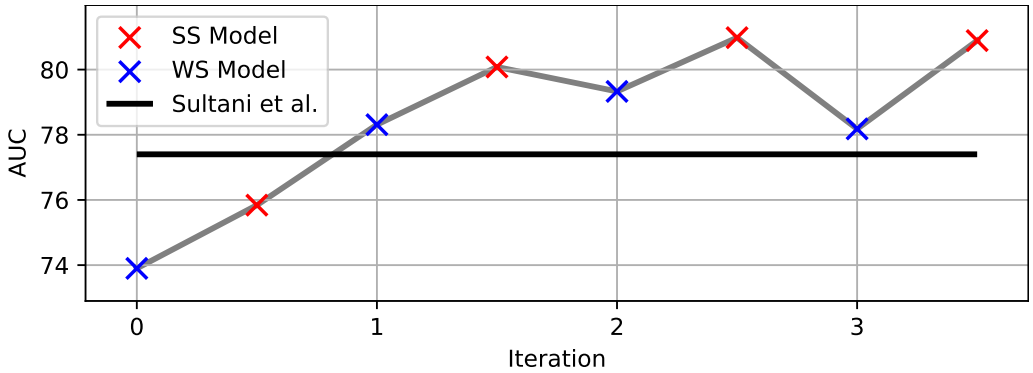
Table 6.3: Performance summary of the proposed method with respect to the state-of-the-art. The final row in each dataset provides the performance of the GMM approach incorporated in the WS/SS framework.



(a) *UBI-Fights*.



(b) *UCF-Crime* [SCS18].



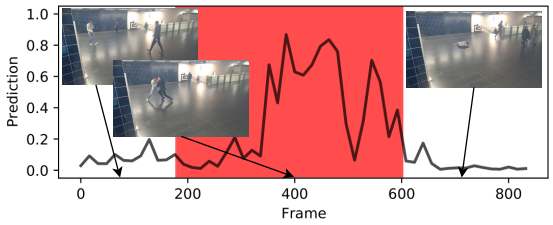
(c) *UCSD-Peds1* [LMV13].

Figure 6.4: Performance evolution with respect to the WS/SS framework iteration, observed for the *UBI-Fights*, *UCF-Crime* [SCS18], and *UCSD-Peds1* [LMV13] sets.

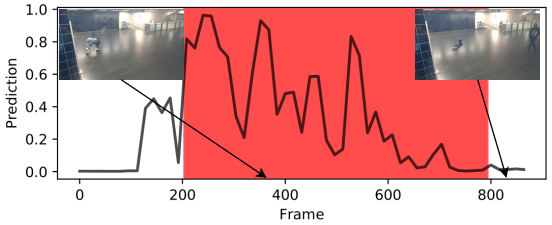
(a) *TOMI* indoor.

(b) *TOMI* outdoor.

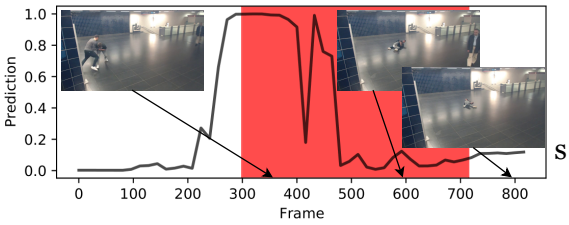
Figure 6.5: Requested *TOMI* devices for the experiments in capturing the scenes illustrated in figure 6.6.



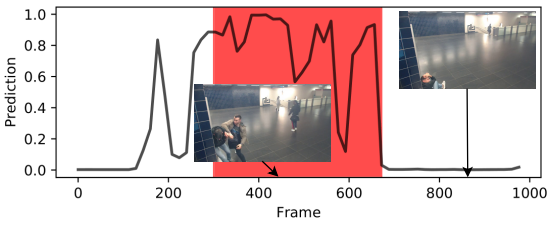
(a) Fight: standing; Persons: 2;
Distance: middle; Environment: indoor.



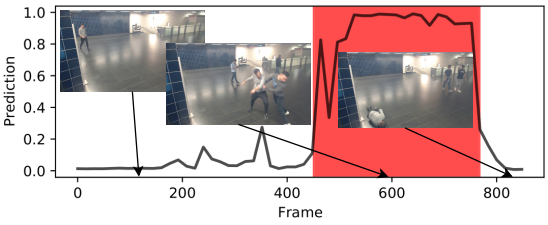
(b) Fight: ground; Persons: 2;
Distance: middle; Environment: indoor.



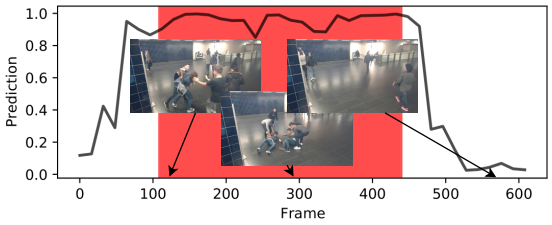
(c) Fight: standing + ground; Persons: 2;
Distance: middle; Environment: indoor.



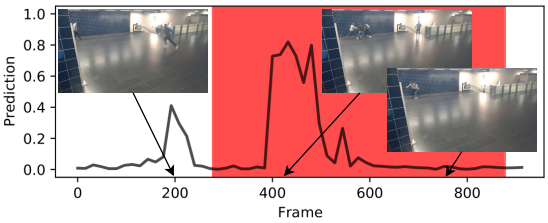
(d) Fight: against wall; Persons: 2;
Distance: up close; Environment: indoor.



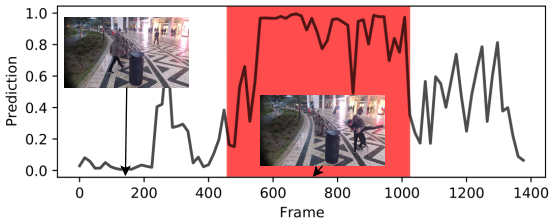
(e) Fight: standing; Persons: 2;
Distance: up close; Environment: indoor.



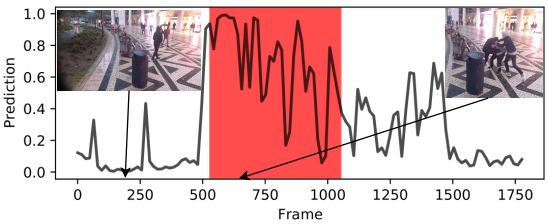
(f) Fight: standing + ground; Persons: 2 + 4;
Distance: up close; Environment: indoor.



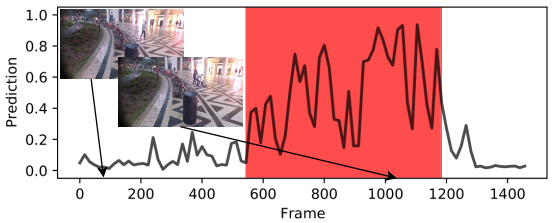
(g) Fight: run + standing; Persons: 2 + 2;
Distance: far; Environment: indoor.



(h) Fight: standing; Persons: 2;
Distance: up close; Environment: outdoor.



(i) Fight: standing + ground; Persons: 3;
Distance: middle; Environment: outdoor.



(j) Fight: standing + ground; Persons: 2 + 1;
Distance: far; Environment: outdoor.

Figure 6.6: Qualitative results of our model performed in real-world scenarios. The red window corresponds to the ground-truth of the scene.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In the presented work, the primary purpose of our research was the explanation and development of computer vision solutions in response to the biometric challenge of detecting abnormal events in surveillance footage under the weakly and partially supervised learning paradigm. In order to achieve this goal, it was essential to understand: 1) how to extract the human dynamics features and its surroundings from video sequences; 2) how the state-of-the-art methods currently handle those features to identify consistent patterns in different types of learning to detect the abnormal events, and 3) which type of learning employed in this field is more suitable considering the multiple obstacles in this kind of problems.

After presenting the video image analysis obstacles in this type of challenge and all the recent improvements in the computer vision domain, we conducted an extensive and detailed review of the state-of-the-art studied methods, in anomaly detection, to mastermind the concepts required to effectively develop this kind of technology in fully unconstrained scenarios with the minimum supervision needed. As described along this dissertation, we concluded that unsupervised learning methods in abnormal event detection were even more complex in learning a robust classifier capable of detecting anomalies in heterogeneous scenarios. Besides that, we have also identified its ineffectiveness in homogeneous scenarios when the amount of learning data is scarce. On the other hand, despite the fact that supervised learning methods need some information about the learning data, we face this problem by minimizing the manually labeling effort with the employment of weakly-supervised and self-supervised learning approaches.

This work introduces a new proposal with the objective of taking advantage of self-supervised algorithms supported by Bayesian classifiers. We show that this coupled deep learning networks were successfully capable of labeling unsupervised datasets and improve the performance of each other consistently over the framework's iterations by means of a Bayesian framework, in multiple experiments as a proof-of-concept. This new method has the benefit of taking advantage of small weakly-supervised datasets and labeling at the strong-supervised level without any additional human supervision. This recent self-supervision technique is still not being extensively used in computer vision, and we expect that this work will encourage further research in this domain, due to the continuously substantial problem in ML algorithms of having sufficient amounts of labeled data to learn a robust model effectively. By also taking this into account, we propose a second

approach within the weakly-supervised learning paradigm where the learning data is labeled at the video-level. With the application of the Gaussian Mixture models (GMM), and assuming the distinctiveness between the expected distribution scores in normal videos and anomaly videos, we were capable of learning a robust model based on the fitting of our GMM in those videos. This model was able to solidly outperform the current methods and prove to achieve better distribution scores compared to the state-of-the-art in weak supervision learning. This new proposal also has the advantage of requiring minimum supervision to learn an effective classifier to detect abnormal events. Moreover, we also prove the possible incorporation of other approaches into the WS/SS framework to achieve even better results on its own, which must be considered as an achievement.

Besides those proposed methods, and with the goal of supporting further research in this field, we constructed a new unique large-scale anomaly detection database, fulfilling the requirements that existent abnormal event datasets did not satisfied. Concerning a specific anomaly detection and still providing a wide diversity in fighting scenarios, fully annotated at the frame level, the *UBI-Fights* [DP20] database is publicly available [1] and free to use by the research community.

In conclusion, the development of self-supervised learning algorithms in computer vision challenges is still in its commencement, but it will surely start to receive even more attention due to the recurrent problem of data-hungry deep learning algorithms. At the moment, weakly and partially supervised algorithms are getting its consideration for avoiding the difficulties of the manual labeling of learning data, known as strong supervision, and being robust enough to face problems in real-world scenarios, as we have shown. Moreover, we successfully proposed two different approaches that employ different levels of supervised learning in the anomaly detection challenge. We also explained and showed how the self-supervised learning technique can be applied in computer vision problems and still capable of being integrated into other methods, as well as, how to apply Gaussian Mixture models in the weakly-supervised learning for abnormal event detection. Therefore, we consider the main purposes of our research work as accomplished.

7.2 Future Work

Inside the scope of this work, further research will be done in the proposed GMM-based loss approach, as we expect to bypass the misconception in our NLL ratio and build a GOF-based model to self-supervise the abnormal scores in the network's response distribution. Furthermore, the application of the proposed self-supervised learning framework to a multi-class classification paradigm would also be of great interest, considering our Bayesian classifiers are designed upon a Gaussian KDE, this kind of technique can also be scalable to multiple kernels representing the different classes.

[1] `http://socia-lab.di.ubi.pt/EventDetection`

In addition and as future work, it would be interesting to extend our work by merging unsupervised methods, like VAEs, with self-supervising learning, without the dependency of the previous learning data to learn a probabilistic encoder-decoder. This could be applied in any kind of environment, as the unsupervised learning method would self-adapt to its conditions. Based on the principle that most already deployed automated systems only use the captured scenes for the final step of classification purposes, the rationale of this future work is to take advantage of those captured images and not only classify those scenes but also continuously learn a better model through the analysis of the most common behavior within the respective environment, as soon as we are able to bypass the limited processing power needed for the deployment of this kind of technology.

Bibliography

[Aka74] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974. 29

[CC19] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019. 1

[CST99] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*. Cambridge University Press, USA, 1999. 15

[CT17] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*, pages 189–196. Springer, 2017. 10, 40, 41, 42, 46, 48

[DBB+01] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In *Advances in neural information processing systems*, pages 472–478, 2001. 29

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 10

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159, 2011. 41

[DP20] Bruno Degardin and Hugo Proença. Iterative weak/self-supervised learning frameworks for abnormal events detection. In *IEEE International Joint Conference on Biometrics*. IEEE, 2020. 2, 15, 35, 39, 52

[EVGW+] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 11

[FBLF07] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multi-camera people tracking with a probabilistic occupancy map. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):267–282, 2007. 36

[FSVC05] Robert Fisher, Jose Santos-Victor, and James Crowley. Caviar: Context aware vision using image-based active recognition, 2005. 36

[FXTL17] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 12

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 1

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 9

[Haw80] Douglas M Hawkins. *Identification of outliers*, volume 11. Springer, 1980. 1

[HCN+16] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 11, 40, 41, 42, 46, 48

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 13

[IZZE17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 9, 41

[JSD+14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014. 7

[JXYY12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 5

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 10

[KTS+14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 5, 6

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[LLLG18] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018. 34

[LMV13] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013. xv, 32, 39, 40, 43, 44, 47, 49

[LSJ13] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 33

[LWZ+19] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019. 12

[MH96] Mary M Moya and Don R Hush. Network constraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3):463–474, 1996. 11

[MLT+19] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 13

[NH10a] Vinod Nair and Geoffrey Hinton. Rectified linear units improve restricted boltzmann machines vinod nai. In *Proceedings of the IEEE international conference on machine learning*, pages 807–814, 2010. 20

[NH10b] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 29, 41

[Pea00] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900. 29

[PMRC17] Rameswar Panda, Niluthpol Chowdhury Mithun, and Amit K Roy-Chowdhury. Diversity-aware multi-video summarization. *IEEE Transactions on Image Processing*, 26(10):4712–4724, 2017. 36

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 11

[RJ20] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *The IEEE*

Winter Conference on Applications of Computer Vision, pages 2569–2578, 2020. 33

[RNS+17] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017. 9, 40, 41, 42, 46, 48

[S+78] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978. 29

[Sch15] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015. 1

[Sco15] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015. 18

[SCS18] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. xv, xvii, 7, 8, 9, 15, 16, 17, 18, 27, 34, 35, 39, 40, 41, 42, 43, 44, 46, 47, 48, 49

[SHK+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 41

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 11

[TBF+15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 5, 6, 9, 20, 25, 40, 41

[TS14] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 12

[UMN] Unusual crowd activity dataset of university of minnesota. Available from: `http://mha.cs.umn.edu/proj_events.shtml#crowd`. 32

[Ven88] John Venn. *The logic of chance: an essay on the foundations and province of the theory of probability, with especial reference to its logical bearings and its application to moral and social science, and to statistics*. Macmillan, 1888. 23

[WQL+18] Tian Wang, Meina Qiao, Zhiwei Lin, Ce Li, Hichem Snoussi, Zhe Liu, and Chang Choi. Generative neural networks for anomaly detection in crowded scenes. *IEEE Transactions on Information Forensics and Security*, 14(5):1390–1399, 2018. 11, 40, 41, 42, 46, 48

[WRKS16] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 12

[WS13] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 7, 11

[XLW+18] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977*, 2018. 12, 13

[ZF14] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 7

Glossary

Anomaly Detection	The process of distinguishing dissimilar instances in a data-driven fashion. Anomalies may have origin on errors or unknown events, depending on which concept being applied.
Biometrics	The study of techniques dealing with people's recognition based on physiological characteristics, such as the face, iris, fingerprint, and behavioral traits, such as motion or speech. It combines Computer Vision with knowledge of human physiology and behavior.
Computer Vision	A scientific field of the study concerning how computers can achieve high-level understanding from processing digital images and videos.
Deep Learning	A class of ML algorithms inspired by the structure of a human brain, neural networks. Using complex multi-layered neural networks, these architectures are able to extract high-level and abstract features.
Goodness of Fit	A set of hypothesis tests to measure how well a statistical model fits/represents the data observations. Those measures usually estimate the discrepancy between the observed values and expected values under the respective model.
Machine Learning	A subset of the larger field of Artificial Intelligence that focuses on learning from experience to automatically improve computer algorithms.
Neural Networks	A type of algorithms designed to recognize patterns through the underlying relationships of features in the training process, molded by trying to reproduce the human brain behavior.
Self-supervised Learning	Types of algorithms training without humans providing labeled data, but the model itself. Considered as an autonomous structure of supervised learning, it is applied in multiple domains, such as AE, auxiliary losses, data augmentations.

Supervised Learning	A type of ML algorithms to learn a predictive model using training data with known outputs, referred to as labeled data. Correlating features with the outcome, it is typically applied in the context of regression or classification problems.
Unsupervised Learning	A type of ML algorithms to learn an underlying model, where no output is given with the training data, referred to as unlabeled data. Discovering hidden patterns in data, it is typically applied in the context of clustering problems.