Estruturas de aprendizagem supervisionada para pesquisa multimédia

Nzakiese Mbongo

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

Orientador: Prof. Doutor Hugo Pedro Martins Carriço Proença

Covilhã, Junho de 2020

Dissertação elaborada no Departamento de Informática da Universidade da Beira Interior e submetida à Universidade da Beira Interior para discussão em provas públicas.

Dedicatória

Dedico este trabalho à minha esposa e aos meus filhos!

Agradecimentos

Agradeço, infinitamente ao Deus Pai Todo-Poderoso pela graça que me concedeu.

Sincero agradecimento, especialmente, ao meu orientador, Professor Doutor Hugo Proença, por ter acreditado em mim, pelo seu conhecimento e pelo apoio incondicional em todos os momentos deste trabalho.

Agradeço aos meus colegas do laboratório ”Socialab” e aos da turma da visão computacional da UBI por toda a ajuda, colaboração e pelos incentivos.

A toda comunidade angolana e conhecidos.

À minha família que me deu força apesar das muitas dificuldades sempre acreditou em mim.

À minha esposa e meus filhos pelo apoio, compreensão e dedicação, em especial à minha esposa, mesmo de pois de perdemos a nossa filha esta etapa de aprendizagem, não me abandonou sozinho.

À Universidade Kimpavita, à Escola Superior Politécnica do Uìge e particularmente à Faculdade de Direito da universidade Kimpavita pelo todo apoio, aos colegas de serviços e amigos.

A todos, meus muitos Obrigados!

Resumo

Tal como escrito no título “Estruturas de aprendizagem supervisionada para pesquisa multimédia”. Este trabalho, destina-se à pesquisa e deteção de uma determinada identidade numa sequência de video. É apresentado também como a técnica de aprendizagem profundo por redes neuronais artificiais consegue detetar um personagem através dos vídeos, essa técnica de deteção de um personagem em um vídeo é um campo de estudo muito interessante já pela sua complexidade em múltiplos objetos presentes nos dados em análise, a partir de vídeos de desenhos animados, usamos um banco de dados separado por dois agrupamentos de dados, incluindo o conjunto de quadros que contém a personagem a ser detetada e outro conjunto que não contém a personagem, ou seja, contém outros objetos e entidades que não é a personagem ou pessoa que queremos localizar sua presença.

Nesse caso, os rótulos ou etiquetas de formação dos nossos dois estados específicos de frames podem ser ”personagem ausente ou presente” no vídeo, no entanto o 0(zero) representa a ausência da personagem e o 1(um) marcamos sua existência na sequência dos frames do vídeo; isto significa que temos um conjunto de dados brutos de 20h et 20 minutos contendo movimentos rápidos e normal e, todos eles são vídeos de desenhos animados, isto é, o caso-de-uso foi a pesquisa de personagem animadas em videos de desenhos animados, por uma questão de facilidade em arranjar conjuntos de dados. Devemos, a todo custo, reconhecer uma personagem (o nosso caso de uso foi o avatar ”korra”) que imaginamos detetar nas sequências de vídeos usando assim a rede neural convolucional 3D, treinado para a deteção de ações anormais nos vídeos de vigilâncias. O algoritmo de aprendizagem supervisionado permitiu finalmente neste trabalho para a deteção do indivíduo. É uma tarefa difícil que requer a extração de características espaço-temporais significativas para capturar as informações necessárias. A Rede Convolucional 3D (C3D) codifica fluxos de vídeo usando uma rede tridimensional totalmente convolucional e, esta rede gera regiões de temporárias candidatas contendo objetos nas regiões selecionadas para atividades específicas. De seguida, analisaram-se os casos de erro observados, procurando perceber os padrões nos dados que levaram a tais conclusões erradas por parte dos modelos.

Palavras-chave

Rede neuronal convolucional, Aprendizagem profunda,multimedia.

Abstract

As it was written in the title ”Supervised learning structures for multimedia research”. This work is aimed at researching and detecting a specific person within the videos. We also present how the technique of deep learning by artificial neural networks can detect a character through videos, this technique of detecting a character in a video is a very interesting field of study because of its complexity in multiple objects present in the data under analysis , from cartoon videos, we used a database separated by two data groups, including the set of frames that contains the character to be detected and another set that does not contain the character, that is, it contains other objects and entities that it is not the character or person that we want to locate your presence.

In this case, the formation labels or tags of our two specific frame states can be ”missing or present character” in the video, however 0 (zero) represents the absence of the character and 1 (one) marks its existence in the sequence video frames; this means that we have a set of raw data of 20h and 20 minutes containing fast and normal movements and they are all cartoon videos, that is, the use case was the search for animated characters in cartoon videos, for the sake of ease in arranging data sets. We must, at all costs, recognize a character (in our case it was The avatar korra) that we imagine to detect in the video sequences, thus using the 3D convolutional neural network, trained to detect abnormal actions in the surveillance videos. The supervised learning algorithm finally allowed this work to detect the individual. It is a difficult task that requires the extraction of significant spatiotemporal characteristics to capture the necessary information. The 3D Convolutional Network (C3D) encodes video streams using a fully convolutional three-dimensional network, and this network generates regions of temporary candidates containing objects in the regions selected for specific activities. Finally, we analyzed the error case observed, as an attempt to perceive the data patterns that led the networks to such erroneous conclusions.

Keywords

Convolutional neural networks, Deep learning, multimedia.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

2D 2 (Duas) Dimensões

3D 3 (Três) Dimensões

asf Advanced Streaming Format File

AUC Area Under the ROC Curve

avi Audio Video Interleave

bpg BatchPhoto Profile Format

C3D Convolution three-dimensional

CNN Convolutional Neural Network

drw Vector Drawing Format

EUA Estado Unidos De América

F-Perso Frames com Personagem a classificar

F-S/Perso Frames sem Personagem

flv Flash Video Encoder

FPS Frames Por Segundo

gif graphic interchange format

img Macintosh Disk Image File

JPEG Joint Photographics Experts Group

max Maximum

MIL Multiple Instance Learning

min Minimum

mov Apple QuickTime Movie File

mp4 Moving Picture Experts Group-4

mpg Moving Picture Experts Group

mvk Matroska Media Container

png Portable Network Graphics

psd PhotoShop Document

ROC Receiver Operating Characteristic

Socialab Soft Computing and Image Analysis Laboratory

SVM Support Vector Machine

rmvb Real Media Variable Bitrate

UBI Universidade da Beira Interior

VC Visão Computacional

vob Video Object

webm Mashable Web Video

wmv Windows Media Video

Capítulo 1

Introdução

Este trabalho destina-se à pesquisa de uma determinada pessoa em sequência de vídeo. Uma das habilidades fundamentais dos seres humanos é analisar seu ambiente. Na maioria dos casos, isso envolve o reconhecimento dos elementos do nosso campo de visão: encontrar outras pessoas, identificar carros, animais e, ...etc.

A Visão Computacional é um ramo da inteligência artificial, ela é o processo de modelagem e replicação da visão humana usando software e hardware. A visão computacional é uma disciplina que estuda como reconstruir, interromper e compreender, detectar, analisar uma cena 3d a partir de suas imagens 2d em termos das propriedades ou características, assim consideramos que ela é uma forma através de qual um agente inteligente(maquina ou robô, etc...) pode perceber, compreender etc.. o que ele está observando por intermédio dos olhos artificiais que nós aqui chamamos de câmeras, que nelas o ambiente observado é composto pelos frames que representam as atividades e mudança dos estados do ambiente que as câmeras capturam. Existe diferentes tipos de formatos de frames ou imagens que provem certamente das câmeras e que são resultado do que elas viram. Em relação a imagens, podem estar codificadas numa variedade de formatos, tais como ".bpg, .deep, .drw, .img, .jpeg, .gif, .psd" ou no caso de um vídeo ".avi, .mp4, .mov, .mpg, .asf, .vob, .webm, .mvk, .flv, .wmv, .rmvbe".

Em última análise, os próprios vídeos são compostos por imagens com uma certa determinadas frequência de leitura dos frames por segundo, isto é chamado de FPS "Frames per second ou Quadros por segundo em português", a maioria das câmeras de video-vigilância tem uma velocidade entre 16fps e 48fps, isto é, 16 fotos por segundo e 48 fotos por segundo. As câmeras dos smartphones actuais têm uma velocidade entre 24fps e 240 fps isto é entre 24 fotos por segundo e 240 fotos por segundo, mas quando as câmeras profissionais existe câmeras a une frequência de gravação até muito mais que possamos imaginar,[01] Há pouco mais de um ano, a equipe de Lihong Wang, pesquisadora do Instituto de Tecnologia da Califórnia (Caltech, EUA), apresentou a câmera mais rápida do mundo. É capaz de capturar 10.000 bilhões de imagens por segundo. O que até consegue capturar o percurso da luz. Hoje, a mesma equipe nos convida a descobrir uma nova tecnologia chamada fotografia sensível à fase compactada ultra-rápida (pCUP). Uma tecnologia capaz de filmar objetos transparentes. Ou fenómenos muito efémeros, como uma onda de choque ou os sinais transmitidos pelos neurónios.

São procurados as possibilidades para entender as informações contendo num frame. Essa possibilidade de um estado do ambiente capturado pelas câmeras seja compreen-

dido pela máquina foi possível graças as técnicas dos algoritmo utilizados em Machine-Learning(aprendizagem automática) também em Deep-Learning(Aprendizagem profunda). [02] Acontece que a Visão computacional e reconhecimento de imagem são termos frequentemente usados como sinónimos, mas o primeiro abrange mais do que apenas analisar imagens. Isso porque, mesmo para os seres humanos, ver? também envolve muitas coisas mais além de uma simples percepção, juntamente com uma série de análises. Cada ser humano usa cerca de dois terços do seu cérebro para o processamento visual, por isso não é nenhuma surpresa que os computadores precisariam usar mais do que apenas o reconhecimento de imagem para obter sua visão de forma realista e correta.

Com a generalização das diferentes formas e maneiras de utilização dos vídeos para análise e identificação dos objetos em movimento nos vídeos, torna cada vez mais usado e o processo demonstra a sua capacidade e ainda por muitos aplicativo torna muito relevante e indispensável, para os sistema de segurança automobilista, sistema de segurança aeroportuário, na precisão para uma intervenção cirúrgica, para deteção de fraude nos diferentes jogos(desporto) ...etc. geralmente a tecnologia é mais desejado nos ambientes ou sistemas onde a compreensão da visão é mais indispensável para analisar a existência, ausência, mudança do estado de um determinado ambiente.

As empresas precisam das informações a partir dos estados dos ambientes que foram capturados pelas câmeras, Mas, com a multiplicidade de informações que as câmeras podem capturarem ao longo da sua utilização, a deteção de um determinado objetos fique ainda mais difícil vendo o volume dos dados a processar ou a identificar, mas graças aos algoritmos de machine-learning à deteção de um objetos a partir das informações que o mesmo disponibilizam nos dá com uma certa precisão aonde se situa o objeto no nosso caso, o caso-de-uso foi a pesquisa de personagem animadas em videos de desenhos animados, a escolha foi estes video por uma questão de facilidade em arranjar conjuntos de dados. Esta forma de obter a informação que precisamos a partir das imagens ou de um vídeo que é chamado da Deteção que pode ser de (um objeto, uma ação, uma pessoa, ...etc.).

O processo de deteção pode ser feito para detetar os objetos nas feeling das vozes vozes(das crianças, bebes, adultos) também pode detetar os sentimentos das pessoas, animais por intermédio das suas vozes ..etc., objetos nas fotos(animais, casas, carros, ...etc.), como também pode até detetar as letras, caracteres manuscritos das diferentes pessoas. Mas no caso do estudo que estamos a realizar, queremos que a deteção seja feita nos objetos em movimente nos vídeos que também podem parecer ser estáticos significa que as câmeras estão a se movendo enquanto os objetos estão fixados num sítio sem ter a possibilidade de se mexer ou bem como dinâmicos, isto é, as câmeras são estáticas bem fixadas sem sendo mexidas e os objetos a detetar estão a se mexendo mudando o estado dos ambientes capturados ou num outro caso os objetos e as câmeras estão estáticos ou dinâmicos enquanto também os objetos podem ser estáticos ou dinâmicos para a pesquisa da personagem nos diferentes vídeos.

1.1 Motivação

Um dos tópicos mais interessantes na VC ”visão computacional” é o reconhecimento automático de identidades em sequências de vídeo, com varadas aplicações, tal como no campo da cinematografia, video-vigilância no caso de criminosos, em locais sensíveis (estações, aeroportos, portos, ...etc.) A motivação para a realização deste tema deve-se relativamente ao aumento do interesse das instituições de segurança pública e privada, governamentais ou não governamentais na procura de diversas soluções de que modo encontrar um indivíduo no meio de uma multidão, isto solucione problemas reais do quotidiano dos cidadãos.

1.2 Definição do Problema a Resolver e Objetivos da Investigação

1.2.1 Problema a Resolver

O desenvolvimento de métodos automatizados para aplicativos de vídeo quer seja video-vigilância, câmeras de telefone, vídeo cinematográfico, ...etc., tem concentrado cada vez mais esforços de pesquisas e, devido às preocupações atuais com a segurança governamental e a segurança das sociedades modernas, as quantias consideráveis de recursos económicos e humanos estão sendo investidas nesse tipo de tarefas. No âmbito dos sistemas de vigilância automatizada, uma das áreas emergentes é a chamada ”biometria comportamental”, que visa reconhecer ações e intenções num grupo de sujeitos. Estruturas de aprendizagem supervisionada para pesquisa multimédia pela sua pesquisa na identificação ou deteção da personagem a partir dos vídeos de desenhos animados, vai nos permitir partindo da sua resolução dar uma via aberta de varias outras maneira que possam ajudar na resolução dos demais tópicos relacionados com este trabalho.

1.2.2 Objetivos

Portanto, o principal objetivo deste trabalho de pesquisa, é planear e desenvolver um sistema de visão computacional que destina-se à pesquisa de uma determinada personagem que seja capaz de detetar a presença desta determinada entidade em dados em proveniência de vídeos dos desenhos animados, o mesmo pode ser usado com vídeos dos seres humanos neste casos pedem ser vídeos filmando com câmera telefónica, câmera privada, câmera de vigilância ou outra fonte de vídeos, mas nos escolhemos os vídeos dos desenhos animados para facilidade em encontrar e arranjar vídeos para o nosso conjunto de dados. isto é, realizar uma deteção em vídeos que contem existências de várias outras entidades ou personagens, então o nosso terá de ser capaz de detetar dentro deste inúmeras personagens, qual é a aquela que interesse na nossa pesquisa.

A resposta desta preocupação irá responder em muitas questões que muitas empresas até hoje precisam para o avanço do sector da segurança no caso por exemplo da criminologia onde o serviço da polícia criminal, quer saber onde encontrar um prisioneiro altamente perigoso que já fez muitas cenas filmadas que acabou de fugir na cadeia ou tanto outros. Na área da cinematografia onde é que a personagem foi vista ...etc. Então a partir desta podemos confrontar os diferentes vídeos que recolhemos nos diferentes lugares onde teve antes da fuga e as diferente imagens dos seus crimes cometidos e construir uma dataset através de qual as técnicas de machine learning com o deep-learning pode lhe encontrar numa determinada zona onde ele passou, se encontra ou até esta a operar os seus actos criminosos logo passar por exemplo nas câmeras de vigilância controladas pelas agências da polícia ou outros serviços de segurança.

As técnicas de deteção usando a visão computacional são hoje uma resposta a estes diversos problemas da sociedade atual e futura.

1.3 Organização da Dissertação

O presente trabalho de pesquisa de uma determinada entidade está estruturado da forma a seguir:

- Introdução: O primeiro capítulo apresenta de forma sintética o projeto. Descreve as estruturas de aprendizagem supervisionada para pesquisa multimédia com o caso à pesquisa de uma determinada pessoa a partir dos videos de desenhos animados. Compõe-se das motivações, problemáticas e respetiva resolução, bem como uma descrição dos objetivos a atingir e seus benefícios para a sociedade presente e futura.
- Trabalhos Relacionados: O segundo capítulo apresenta um estudo das tecnologias relacionadas ao nosso trabalho e disponíveis, assim como descreve os avanços científicos relacionados com o tema em questão. Este capítulo nos ajudou a ter uma visão muito aberta para a realização do mesmo.
- Abordagem selecionada: Neste capítulo descrevemos a nossa abordagem ao problema, que visa minimizar as limitações identificadas no método de referência. que nos permitiu ter de forma estruturado o nosso trabalho e os passos a seguir para realizar a parte pratica do trabalho.
- Experiências e Resultados: Aqui são apresentados os resultados alcançados pelo nosso método e nos permitiu ganhar a experiência, perceber na realidade como o modelo funciona, bem como antecipar os sucessos e falhas do sistema.
- Conclusões e Trabalho Futuro: No capítulo final são apresentadas as conclusões do projeto, é identificado o trabalho futuro, bem como as melhorias que podem ser implementadas.

- Referências: As referências dos trabalhos analisados com vista ao desenvolvimento desta dissertação.

Capítulo 2

Trabalhos Relacionados

Neste capítulo iremos abordar o estado da arte, isto é, os diferentes trabalhos existentes relacionados ao nosso estudo.

2.1 Deteção

O objetivo da fase de deteção é localizar as regiões de interesse (RoI) nas quais o objeto provavelmente será encontrado. Durante esta fase, a imagem é tipicamente segmentada [18], em que a parameterização do objeto é obtida, de acordo com características pré-estabelecidas, baseadas na cor ou na forma.

Esta fase é especialmente sensível em função da variabilidade existente no meio ambiente, e da forma como as características do objeto a detectar podem variar, em função do tamanho, da forma, luminosidade. Uma vez que é a fase que tipicamente inicia o processo de reconhecimento de imagem, é também aqui que se lida de forma mais imediata com a inconsistência da informação recebida, sendo por isso necessário o desenvolvimento de métodos especialmente capazes em termos de robustez, isto é, da forma como se adaptam aos diferentes tipos de dados recebidos a partir dos sensores (cameras).

2.2 Reconhecimento

A fase de reconhecimento é a principal, na qual o sinal previamente detectado é etiquetado (classificação), ou é obtido um valor numérico correspondente (regressão). As técnicas de reconhecimento de objetos mais clássicas podem incluir métodos estatísticos, tais como *Support Vector Machines*, *Adaboost* ou Redes Neuronais (MLPs).

2.3 Abordagens relacionadas

Os parágrafos seguintes descrevem as diferentes abordagens com vista à fase de deteção já abordada, bem como ao reconhecimento das ações, pessoas, personagens em vídeos, quer seja filmes, vídeo vigilância, dispositivos móveis, etc... Sendo verdade que existem muitos

bons algoritmos já publicados sobre detecções e reconhecimento das ações, pessoas, face, carros, acidentes, após a nossa análise concluímos que os principais módulos deste tipo de sistemas envolvem a deteção de objetos, a classificação dos objetos, reconhecimento de padrões e a compreensão do que foi detetado. Focando em concreto na deteção automática de uma personagem (identidade) em sequências multimédia, os métodos que iremos de seguida analisar focam-se em uma ou várias análises detalhadas de imagem ou vídeo, desde o tratamento ou organização do conjunto de dados até à obtenção dos resultados.

Nos últimos anos, as abordagens de aprendizagem profunda tornaram-se cada vez mais populares e eficientes. As Redes Neuronais Convolucionais obtiveram grande sucesso no campo da classificação de imagens e reconhecimento de objetos. Diferentemente dos métodos tradicionais, as CNNs podem ser treinadas para extrair automaticamente as características importantes de uma classe de objetos, bem como detectar os objetos desejados de maneira significativamente mais rápida e confiável. De seguida, serão apresentados alguns exemplos:

2.3.1 Real-World Anomaly Detection in Surveillance Videos

[14]Os vídeos de vigilância são capazes de capturar uma variedade de anomalias realistas. Neste artigo, foi proposta aprendizagem automática para deteção de anomalias explorando vídeos normais e anómalos. Para evitar a anotação dos segmentos ou clipes anómalos nos vídeos de treino, o que consome muito tempo, ele propõe aprender anomalias através da estrutura profunda de classificação de várias instâncias, aproveitando vídeos de treino com etiquetas fracas, ou seja, os rótulos de treino (anómalos ou normais) estão em vídeo- nível em vez do nível do clipe. nesta abordagem, estão considerado vídeos normais e anómalos como malas e segmentos de vídeo como instâncias na aprendizagem de várias instâncias (MIL) e o sistema aprenda automaticamente um modelo de classificação profunda de anomalias que prevê altas pontuações de anomalia para segmentos de vídeo anómalos. Além disso, introduzimos restrições de esparsidade e suavidade temporal na função de função de custo de classificação para localizar melhor a anomalia durante o treino. Também introduzimos um novo conjunto de dados em grande escala, com 128 horas de vídeos. Consiste em 1900 vídeos de vigilância do mundo real, longos e sem cortes, com 13 anomalias realistas, como brigas, acidentes de viação, roubo, roubo, além de atividades consideradas ”normais”.

Eles usaram um conjunto de dados que pode ser usado para duas tarefas. Primeiro, deteção geral de anomalias, considerando todas as anomalias como um grupo único e todas as atividades normais em outro grupo. Segundo, por reconhecer cada uma das 13 atividades anómalas.

Os resultados experimentais mostram que o método MIL para deteção de anomalias alcança uma melhoria significativa no desempenho da deteção de anomalias, em compara-

ção com as abordagens estado-da-arte. Também eles forneceram os resultados de várias linhas de base recentes de aprendizagem profunda sobre o reconhecimento de atividades anómalas. O baixo desempenho de reconhecimento dessas linhas de base revela que que o conjunto de dados é muito desafiante e de elevada dificuldade, em função da heterogeneidade dos ambientes e das muito diferentes características dos videos existentes no conjunto de dados (interior/exterior, diurnos/nocturnos,...)

1. Resultado.

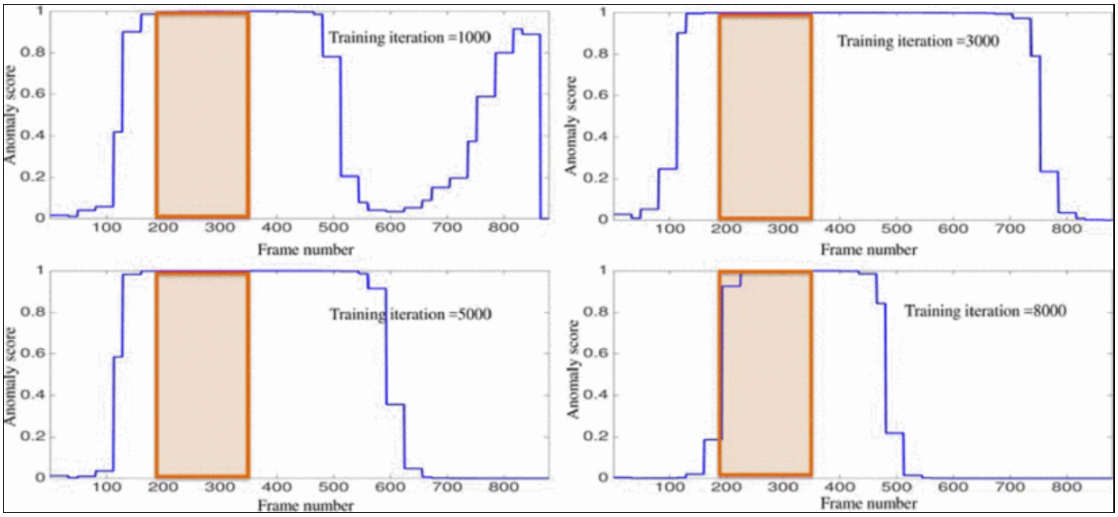


Figura 2.1: Evolução da pontuação em um vídeo de treinamento sobre iterações. Janela colorida representa a verdade do solo (região anômala). À medida que a iteração aumenta, nosso método gera altas pontuações de anomalia em segmentos de vídeo anômalos e pontuações baixas em segmentos normais.

2.3.2 *Détection de Personnes dans les Vidéos pour Leur Immersion dans un Espace Virtuel*

Em [03], apresenta-se uma proposta de um aplicativo permitindo assim a extração de pessoas nas sequências filmadas no interior como também no exterior via uma câmera fixa, Essas pessoas são incorporadas em tempo real em um ambiente virtual de desenho animado, onde podem interagir como atores e espectadores. A originalidade deste método baseia-se em dois pontos principais; Primeiro, a deteção de alterações é realizada combinando dois tipos de máscaras: uma máscara de região e uma máscara de contorno. Em seguida, essas máscaras são obtidas usando uma imagem benchmark, que é construído e atualizado regularmente. Este último leva em consideração as diferentes mudanças que aparecem frequentemente quando a sequência dura várias horas. Os resultados mostram que as pessoas são extraídas corretamente e rapidamente desenho animados.

1. Resultado.

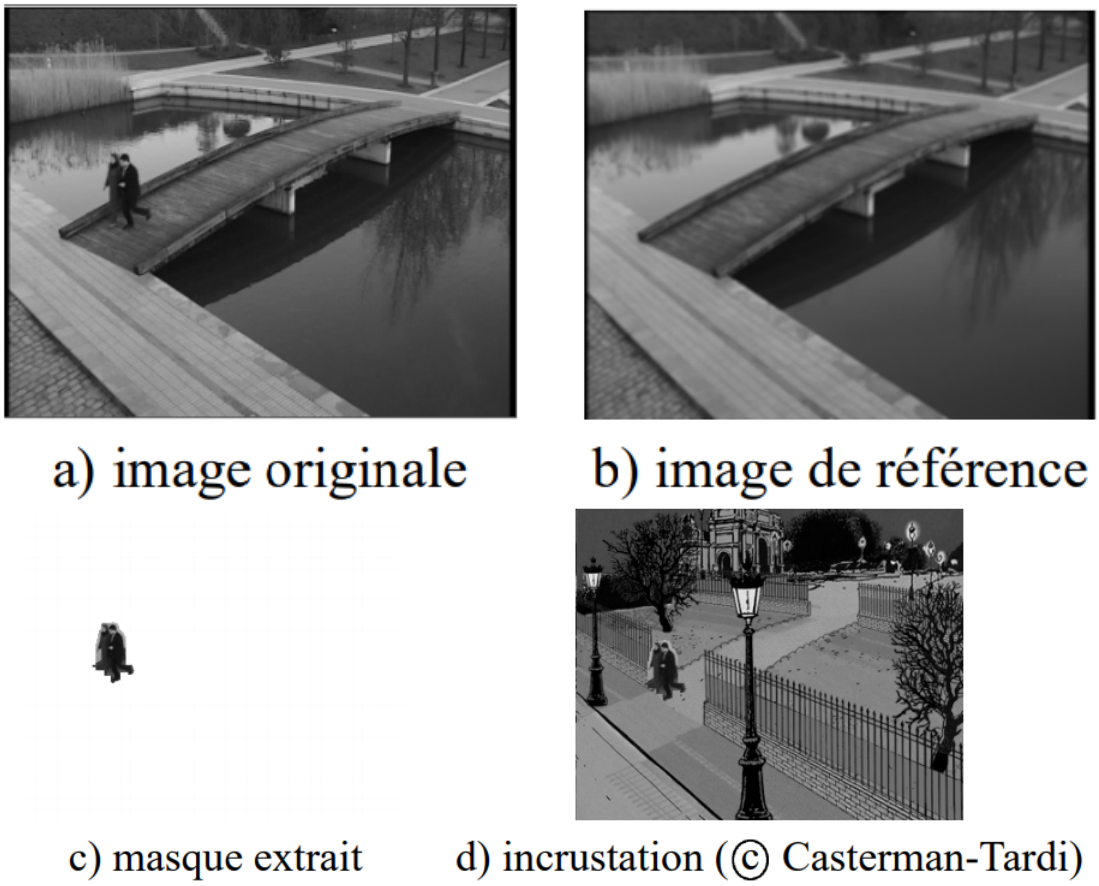


Figura 2.2: 6 - Imagem 1000: A variação de iluminação da revista é premiada com base na referência.

2.3.3 Detecting Moving People in Video Streams

[04] A deteção de pessoas em movimento é uma tarefa importante para os sistemas de vigilância por vídeo. Esta referência apresenta um algoritmo de segmentação de movimento para detetar pessoas que se deslocam em ambientes internos. O algoritmo proposto trabalha com câmeras móveis e é composto de duas partes principais. Na primeira parte, um procedimento quadro a quadro é aplicado para calcular a imagem da diferença e uma rede neuronal é usada para classificar se a imagem resultante representa uma cena estática ou uma cena contendo objetos móveis. A segunda parte tenta reduzir os erros de deteção em termos de alarmes falsos ou perdidos. Um autómato de estado finito foi projectado para fornecer uma classificação robusta e reduzir o número de blobs falsos ou perdidos. Finalmente, uma elipse delimitadora é calculada para cada blob detetado, a fim de isolar pessoas em movimento.

Os resultados da deteção de alterações, adotando limiares adaptativos, às vezes podem ser imprevisíveis. Observando dois casos típicos, uma cena vazia e uma cena com um objeto em movimento, o problema com o qual queremos lidar fica claro: as técnicas de limiarização podem classificar como pixeis em movimento o ruído de aquisição . Para resolver

esse problema, uma nova técnica para classificar corretamente a imagem atual foi desenvolvida.

1. Resultado de Detecting moving people in video streams.

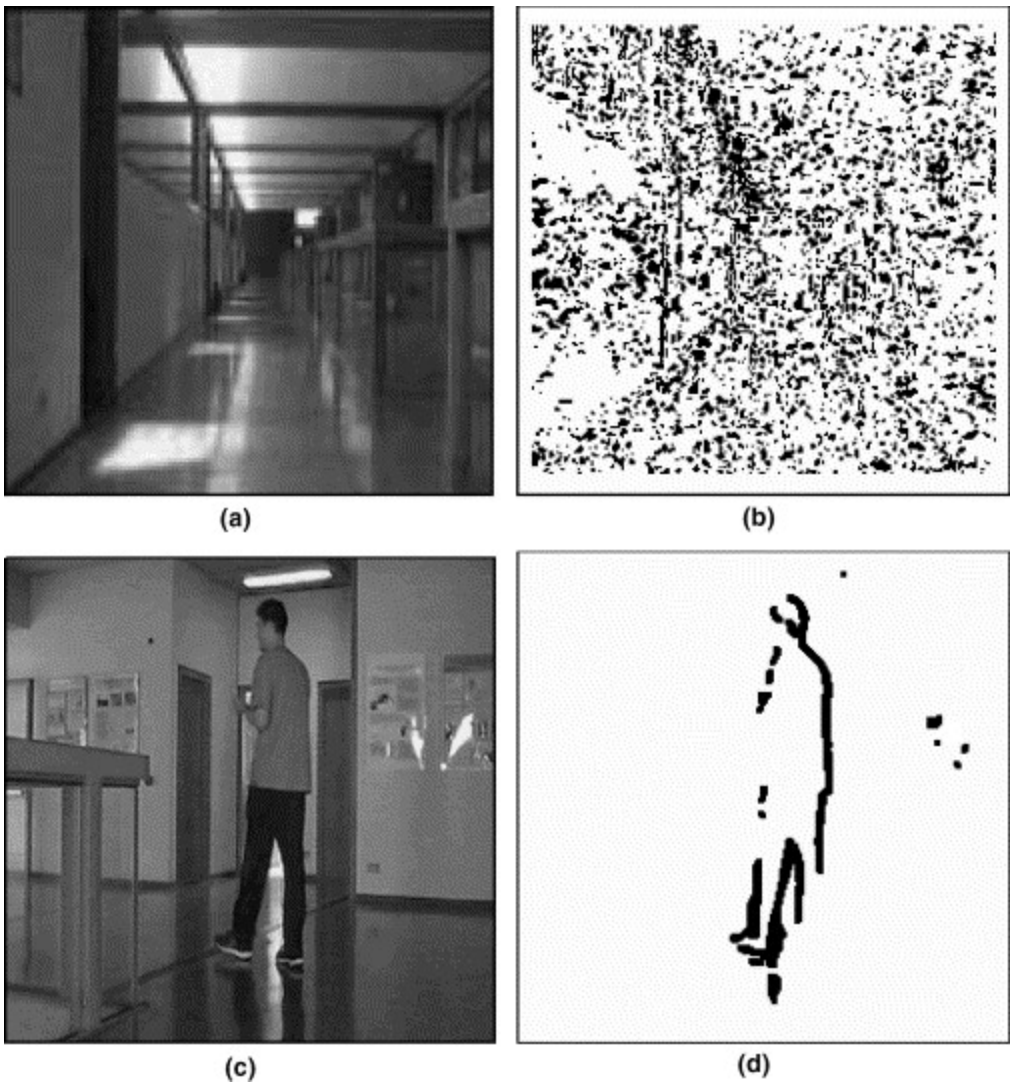


Figura 2.3: Exemplos de detecção de alterações com o método proposto em (Ridler e Calvard, 1978). (a) e (b) representam uma cena vazia enquanto (c) e (d) mostram um objeto em movimento.

1. Figura 2.3

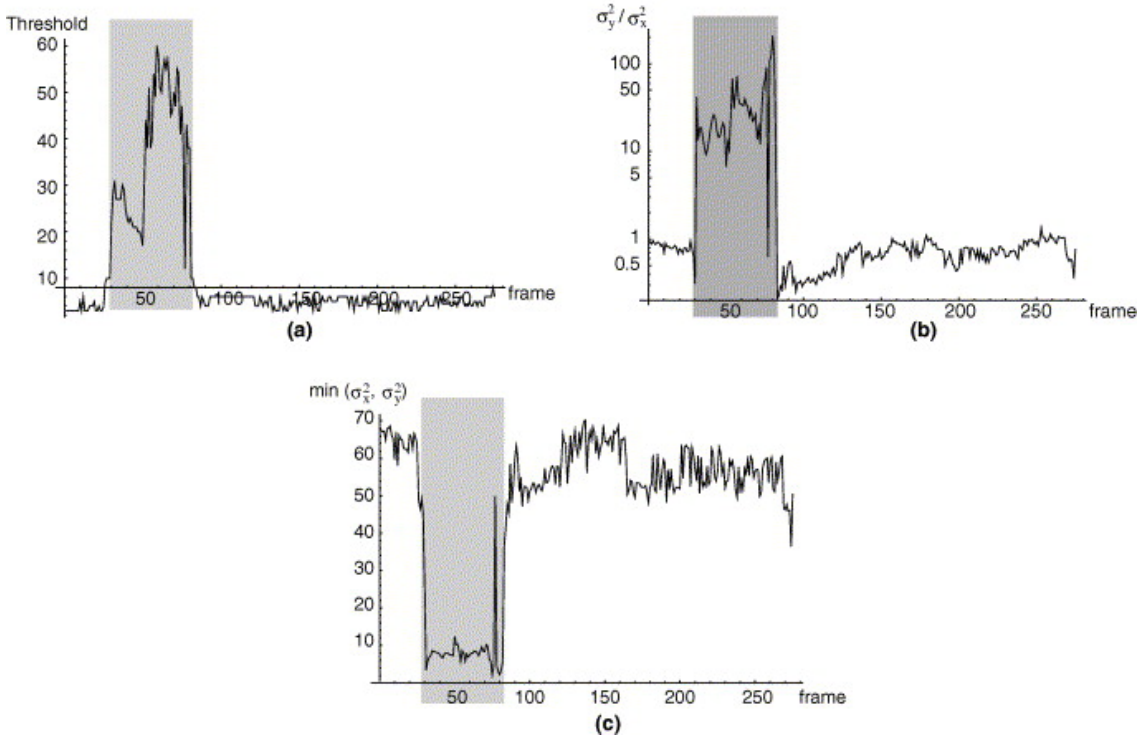


Figura 2.4: Os gráficos mostram os valores dos parâmetros de detecção de pessoas em uma sequência com uma pessoa em movimento. A área cinza delimita os quadros em que a pessoa está visível.

Na figura 2.3 é ilustrado um exemplo do cálculo dos três parâmetros. Os três gráficos das Fig. 2.3(a)-(c) correspondem a uma sequência na qual uma pessoa está se movendo dentro da cena do quadro 30 para o quadro 83. Nesses quadros, temos um valor alto do limiar *Th*, um valor alto da taxa de variação e um valor baixo de pelo menos uma das duas variações.

A sequência capturada com uma câmera fixa não apresentou problemas particulares, executando uma classificação incorreta em 5% dos casos, causada em grande parte por uma má deteção de objeto quando o objeto está entrando na cena e é apenas parcialmente visível. Nos quadros com um objeto totalmente visível, a classificação está correta em 98% dos casos.

1. Figura 2.4 ilustra quando o objeto esta parcialmente ou totalmente visível.

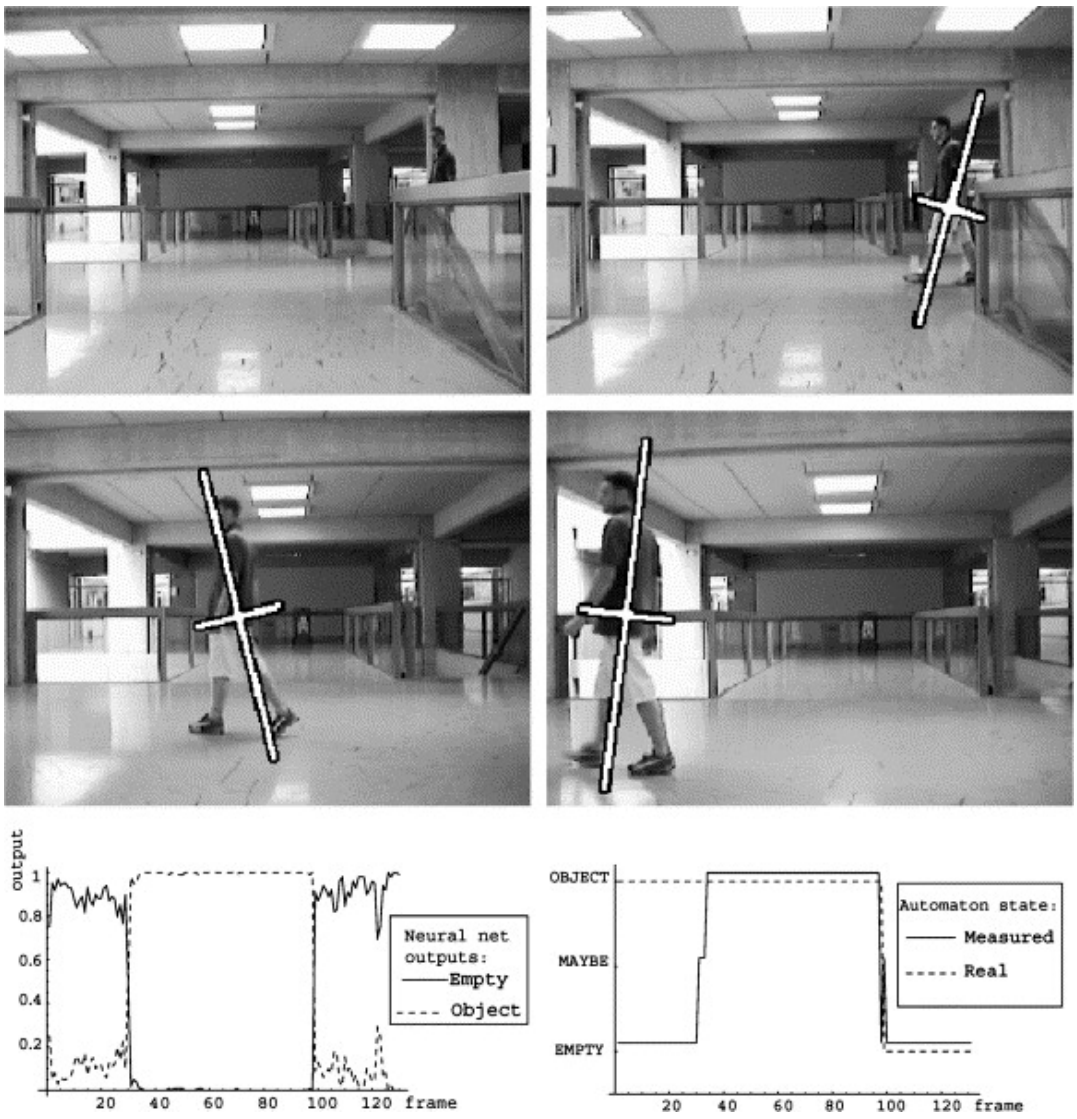


Figura 2.5: A primeira linha mostra alguns quadros da sequência de tradução com um objeto em movimento. Na segunda linha, são mostrados os resultados da rede neuronal e do autômato.

2.3.4 A Localized Approach to Abandoned Luggage Detection with Foreground-Mask Sampling

Em [06], os autores sugerem uma nova abordagem para a detecção de bagagem abandonada em video-vigilância. Os candidatos a itens de bagagem abandonados que podem representar possíveis ameaças à segurança são identificados e localizados pela nossa técnica de amostragem de máscara de primeiro plano proposta. a abordagem pode lidar com peças de bagagem de forma e cor arbitrárias, sem a necessidade de aprendizado prévio, e funciona bem em situações de muita gente e com muita confusão. Essa localização de itens suspeitos de bagagem em cena permite concentrar a atenção e o processamento subsequente apenas em seus bairros. O proprietário da bagagem é então localizado e rastreado para determinar se a bagagem foi ou não abandonada deliberadamente. Um modelo de probabilidade usando o princípio MAP é empregado para calcular uma pontuação de confiança a posteriori para o evento de abandono de bagagem, e um alarme será acionado

automaticamente se a certeza do abandono de bagagem for maior que um limite predefinido.

1. Figura 2.5 apresenta os objetos abandonados.

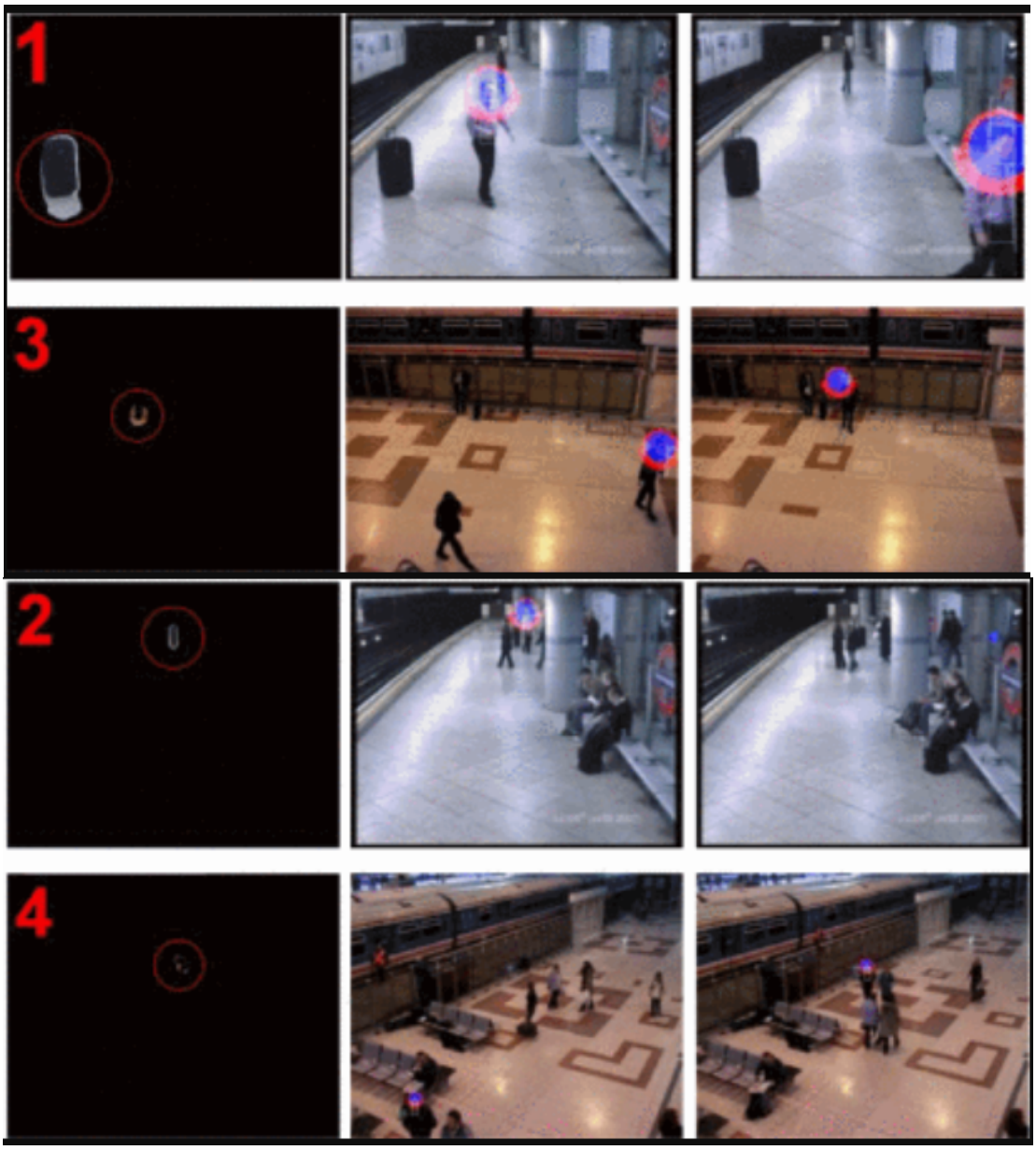


Figura 2.6: Sequência 1, 3 e 4: bagagem estática (da esquerda) detectada; o rastreamento do proprietário é iniciado; proprietário sai de cena, alarme disparado. Sequência 2: bagagem estática detectada; o rastreamento do proprietário é iniciado; proprietário perdido devido à oclusão, alarme disparado. A sequência 1 e 2 são do conjunto de dados AVSS 2007; Sequência 3 e 4 do conjunto de dados PETS 2006.

2.3.5 Person Detection in Video Surveillance

Em [07], apresentaram um método para deteção e reconhecimento de pessoas em movimento a partir de um fluxo de vídeo gerado por um sistema de vigilância. O objetivo deste trabalho foi implementar um algoritmo que diferencia uma pessoa de outros tipos de formas ou outros objetos presentes no vídeo. A deteção de um objeto em movimento é realizada com base na comparação de dois quadros extraídos em um intervalo de tempo específico do fluxo de vídeo em tempo real. Com base na comparação resultante, um algoritmo de binarização e um algoritmo de crescimento de região são usados para extrair todos os objetos móveis independentes no caso de movimento. Para cada região extraída, uma série de recursos geométricos são calculados para diferenciar a forma de uma pessoa das outras formas detetadas.

O algoritmo de detecção de pessoas proposto foi testado em mais de 400 quadros extraídos de um vídeo gravado na primeira zona, mais de 500 quadros na segunda zona e mais de 900 quadros na terceira zona. Ao diferenciar os quadros para detectar movimento, o algoritmo proposto detectou o movimento da pessoa em todos os casos. A única desvantagem observada durante os testes é que a pele de uma pessoa detectada nem sempre é marcada como parte da região detectada que gerou o movimento. Assim, a precisão da metodologia proposta mostrou-se muito alta, pois 100% das imagens testadas foram classificadas corretamente, embora o algoritmo tenha sido testado em diferentes ambientes. Em relação à complexidade do algoritmo proposto, o tempo de processamento de dois quadros consecutivos para detecção de pessoas é menor que 400 ms, o que o torna adequado para qualquer sistema de vigilância.

No primeiro exemplo (Fig. 2.6), após a aplicação da metodologia proposta, a pessoa é detectada e destacada. Além disso, várias regiões classificadas como pessoas pelo algoritmo proposto são apresentadas na Fig. 2.7.

1. Figura 2.6 detecção e reconhecimento de pessoas em movimento a partir de um fluxo de vídeo.

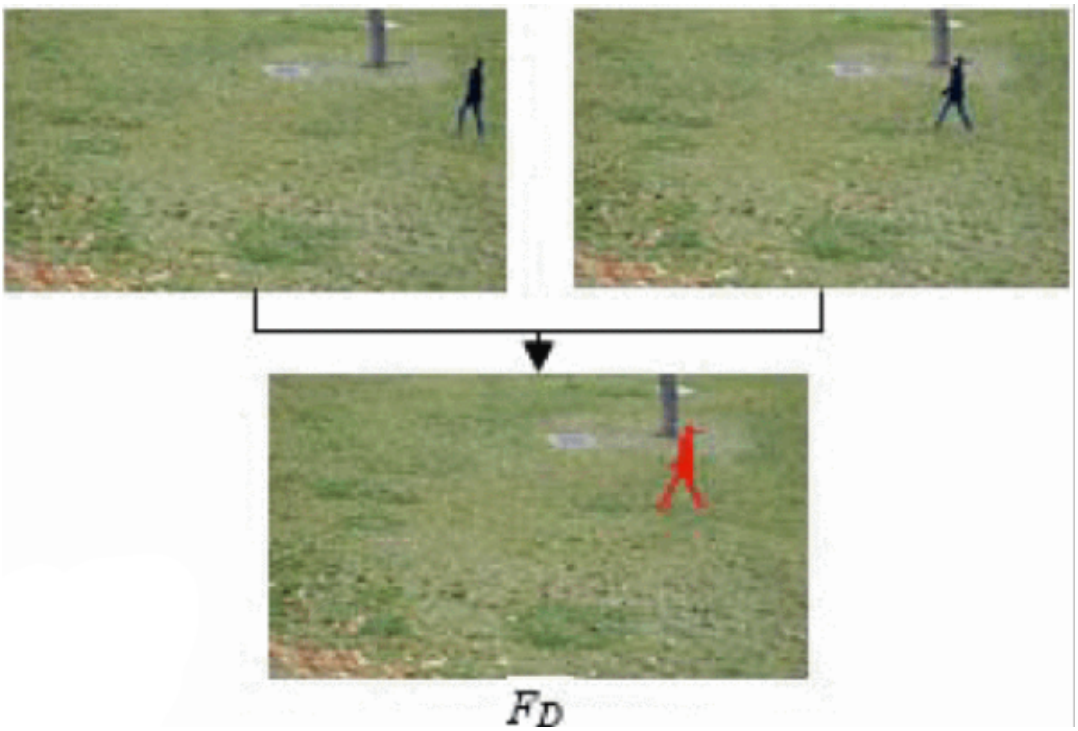


Figura 2.7: Detecção de movimento no exemplo 1.

1. Figura 2.7 extração.

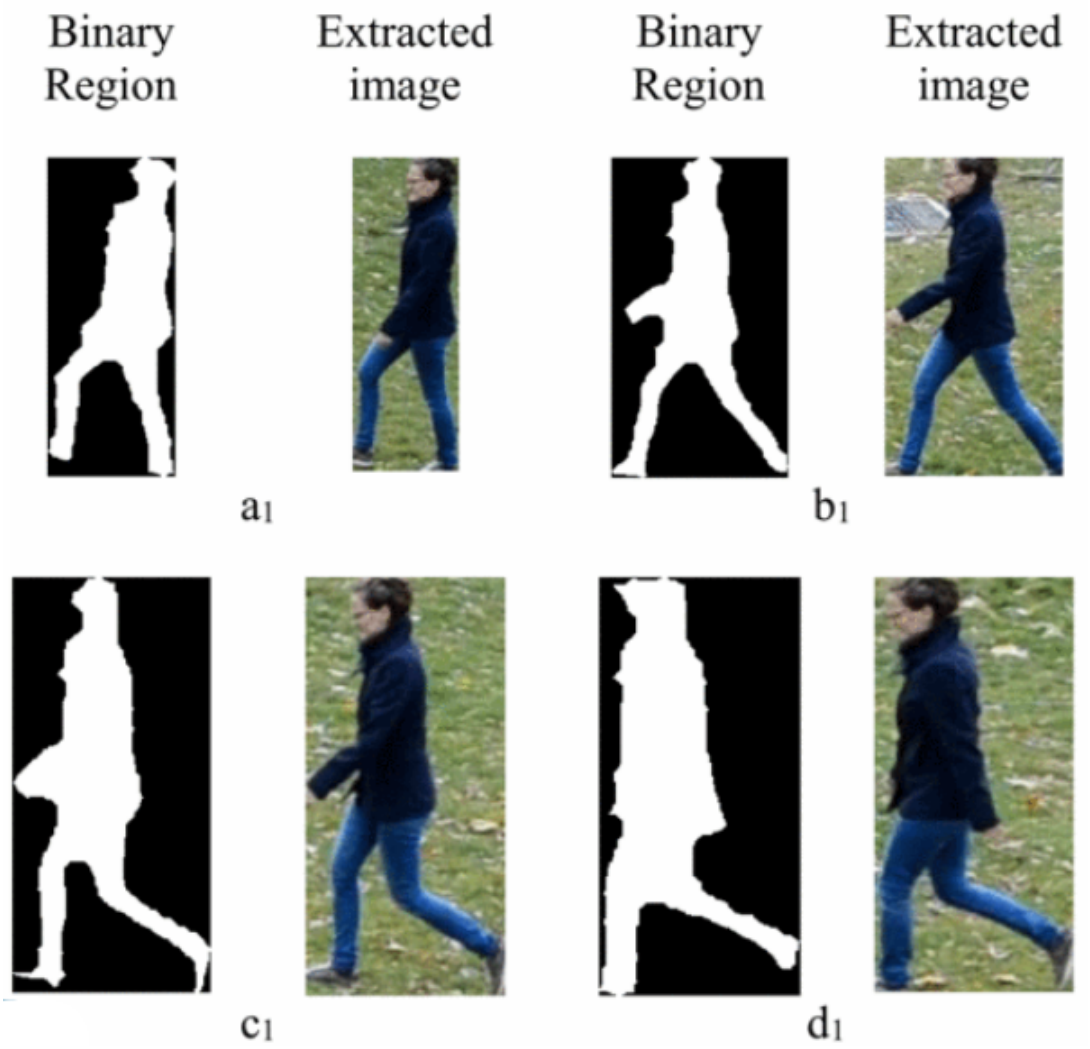


Figura 2.8: Regiões detectadas e imagens extraídas, por exemplo 1.

2.3.6 Overhead View Person Detection Using YOLO

[08] Num sistema de vigilância por vídeo, uma das tarefas importantes é detectar pessoas. Nos últimos anos, diferentes algoritmos de visão computacional e aprendizagem profunda (*Deep Learning*) foram desenvolvidos para este efeito, o que fornece resultados robustos de detecção de pessoas. A maioria dessas técnicas desenvolvidas focou-se em vistas frontais e assimétricas. Portanto, neste artigo, a detecção de pessoas foi realizada de uma perspectiva significativamente alterada, ou seja, visão aérea. Um modelo de aprendizagem profundo, ou seja, o YOLO (You Look Only Once), foi explorado no contexto da detecção de pessoas da visão aérea. O modelo é treinado no conjunto de dados da vista frontal e testado no conjunto de dados da pessoa de visão aérea. Além disso, a contagem de pessoas de visão geral foi realizada usando informações da caixa delimitadora classificada. O modelo YOLO produz resultados significativamente bons com TPR de 95% e FPR de até 0,2%.

1. Figura 2.8 resultado da visão aérea.

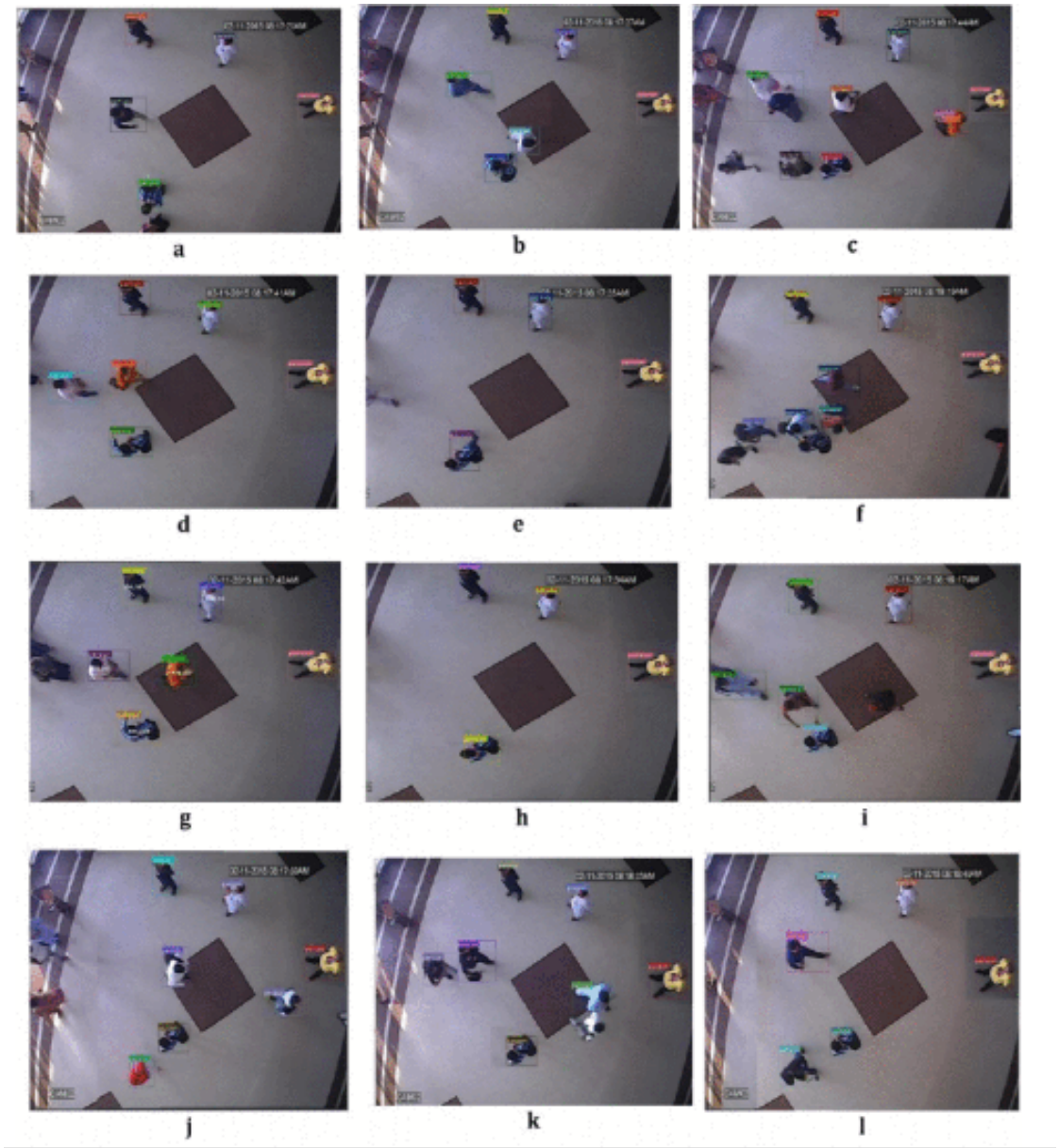


Figura 2.9: Resultados da detecção e contagem de pessoas de visão geral, os detalhes de cada subfig são fornecidos na seção experimental. Os resultados da detecção de pessoas podem ser vistos com várias caixas delimitadoras detectadas em cores. Os resultados não detectados também podem ser visto no subfig (g, fe i).

Para fins de avaliação, são utilizados diferentes parâmetros, incluindo True Positivo (TPR) e Falso positivo (FPR). Para avaliar o desempenho dos propósitos do modelo acima, a verdade básica foi criada manualmente. A Tabela I. apresentou os resultados do modelo baseado em aprendizagem profunda usado acima para detecção e contagem de pessoas em visão aérea. Pode-se observar na Tabela I. que o TPR e o FPR da contagem e detecção de pessoas variam muito com o número de pessoas presentes na cena. Como neste trabalho, o modelo de vista frontal pré-treinado é usado para detecção de pessoas, mas ainda assim a precisão do método é boa em comparação com outros métodos tradicionais. A curva para o modelo de contagem e detecção de pessoas é mostrada na Fig. 2.9.

Tabela 2.1: Detecção de pessoas e contagem de resultados.

No of Person	Person Detection and Counting	
	TPR	FPR
2	95%	0.14%
3	95%	0.14%
4	94%	0.16%
5	94.5%	0.17%
6	93.6%	0.17%
7	92.7%	0.18%
8	91.5%	0.19%
9	91.2%	0.2%
10	90.7%	0.2%

1. Figura 2.9 - Gráfico de TPR e FPR.

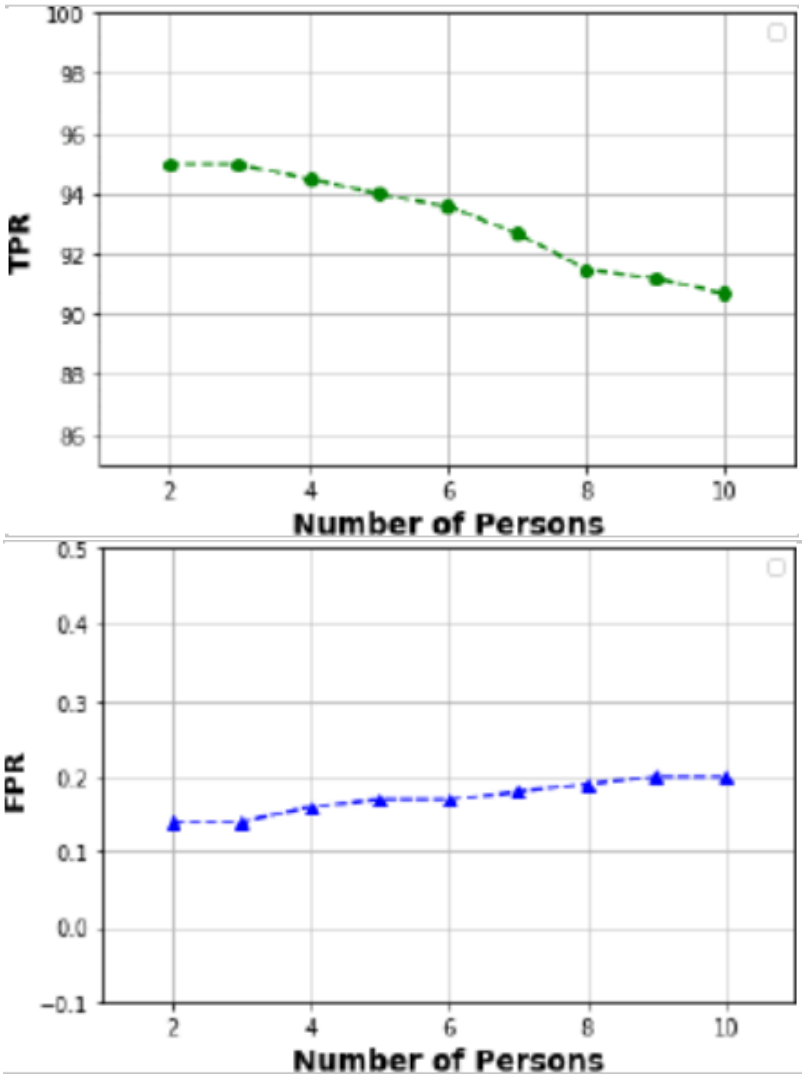


Figura 2.10: TPR e FPR para detecção e contagem de pessoas na visão aérea.

2.3.7 Video-based Action Detection and Recognition of Humans using Optical Flow and an SVM Classifier

[09] com amplas aplicações em vigilância por vídeo, recuperação e análise de vídeo e interação homem-computador. A detecção e o reconhecimento da ação humana a partir dos bancos de dados de vídeo é realmente uma tarefa difícil e desafiadora. Neste artigo, a detecção e o reconhecimento de ações humanas baseadas em vídeo são abordados e realizados no conjunto de dados KTH e em vídeos em tempo real. Inicialmente, cem quadros são extraídos de cada sequência de vídeo e o fluxo óptico entre os quadros é calculado. Os dados extraídos usando fluxo óptico são convertidos em imagem binária. Em seguida, o descritor Histogram of Oriented Gradient (HOG) é usado para extrair o vetor de recursos das imagens binárias. Esses vetores de recursos são fornecidos como recursos de treino para o classificador SVM (Support Vector Machine) para preparar um modelo treinado. Para teste, criamos nossos próprios vídeos em tempo real, que consistem em ações: caminhada, corrida, corrida, boxe, acenar com as mãos e bater palmas. Para cada vídeo, o mesmo tipo de recurso é extraído e fornecido ao classificador SVM para classificação de ações.

1. Figura 2.10 - Gráfico de TPR e FPR.

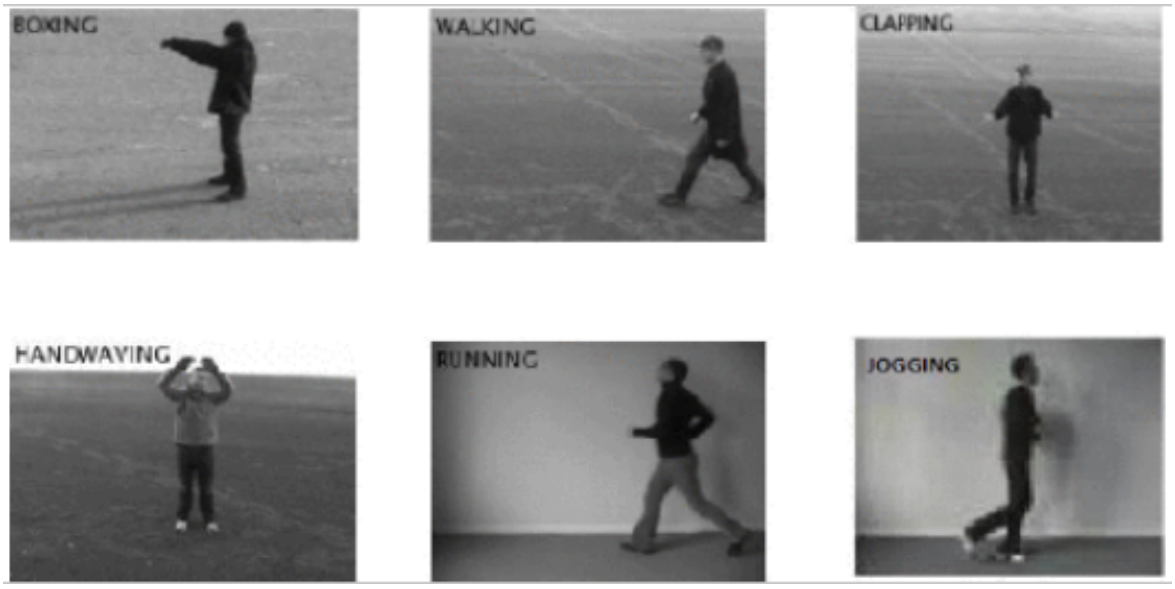


Figura 2.11: Ações reconhecidas com anotação.

Tabela 2.2: Precisão na classificação de ação em vídeos treinados.

Action/Resultado	Running	Jogging	Walking	Boxing	Clapping	Handwaving
Running	85	10	5	0	0	0
Jogging	8	87	5	0	0	0
Walking	4	8	8	0	0	0
Boxing	0	0	0	93	3	4
Clapping	0	0	0	3	91	6
Handwaving	0	0	0	1	4	95

2.3.8 Pedestrian Fall Action Detection and Alarm in Video Surveillance

Em [10], foi apresentado um novo método para detectar e alarmar a ação de queda de pedestres no vídeo-vigilância usando o método de diferença de quadros e o quadro de marcadores de pedestres. Primeiramente, foi usado a câmera fixa para capturar informações de imagens de vídeo e foi aplicado o método de erosão e dilatação morfológica para reduzir os ruídos. Em seguida, combinou o método de diferença de quadros e o método de diferença de fundo para marcar o contorno de destino e analisou as informações de contorno para corrigir a localização do pedestre. Finalmente, analisando a posição do movimento e a trajetória do centróide do pedestre e analisando a relação comprimento/largura da estrutura do marcador de pedestres e o tempo de permanência para detectar a ação de queda de pedestres e outros movimentos anormais e acionar o alarme a tempo. As experiências mostram que o método proposto captura a ação de queda de pedestres com sucesso.

1. Figura 2.11. Mostra o movimento normal dos pedestres

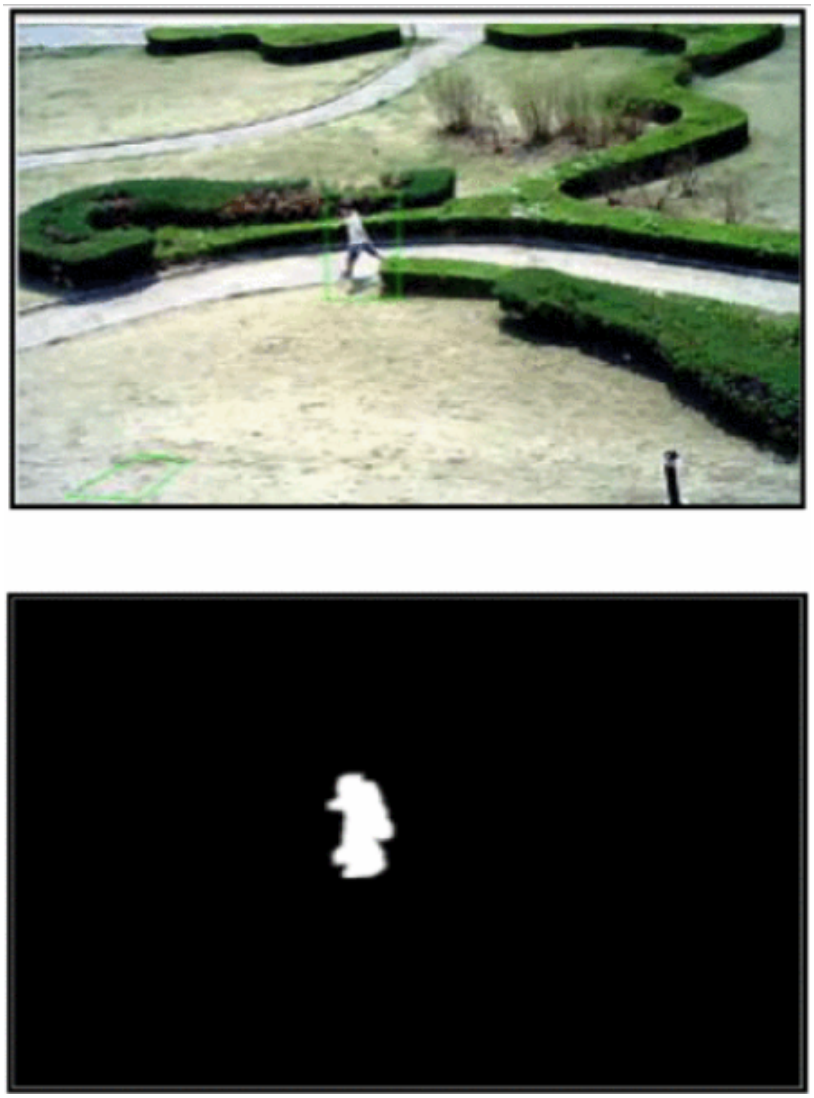


Figura 2.12: Detectar movimento de caminhada.

1. Figura 2.11. A imagem colorida mostra dois homens, um deitado no chão, alarmado com a borda vermelha e as palavras escritas, e o outro caminhando. A imagem binária também mostra duas figuras retangulares dos pedestres

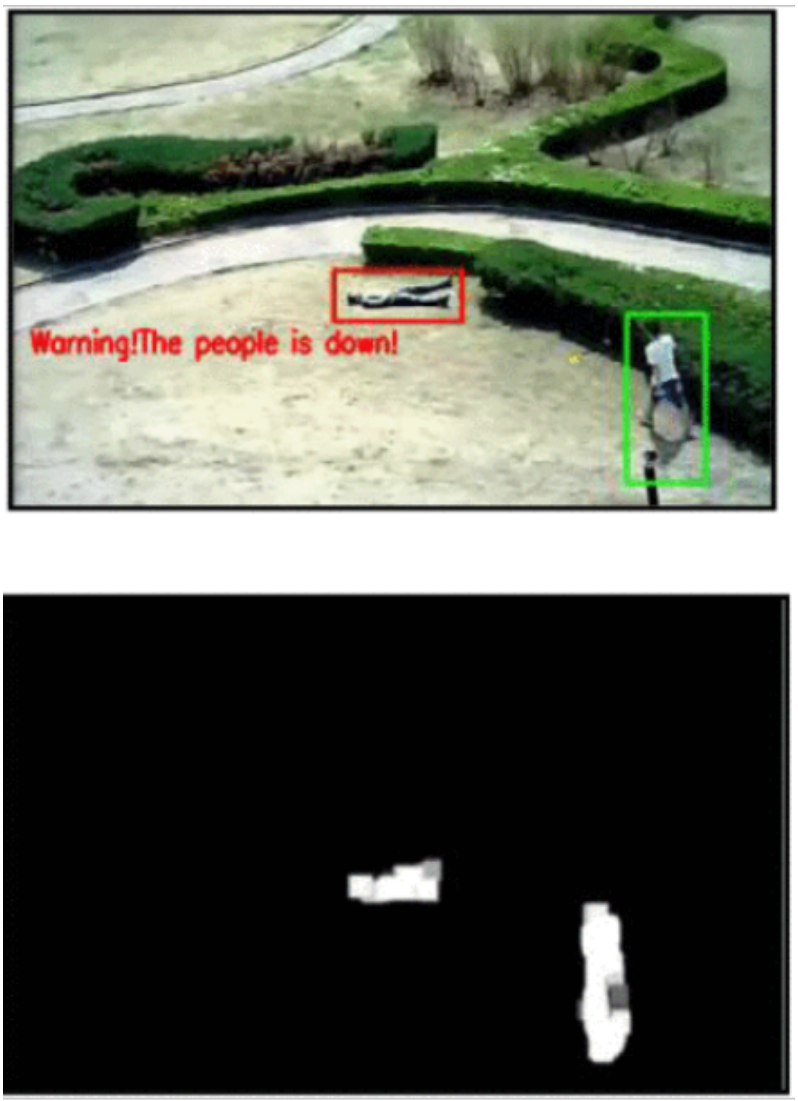


Figura 2.13: Détecte les mouvements normaux et anormaux.

Com as duas primeiras imagens, a imagem colorida marca o contorno do pedestre no vídeo; a imagem binária é o pedestre selecionado usando o método da diferença de fundo. No entanto, devido às operações de dilatação e erosão, a imagem correta apresenta uma leve distorção.

Com as segundas imagens, a imagem colorida mostra dois homens, um deitado no chão, alarmado com a borda vermelha e as palavras escritas, e o outro caminhando. A imagem binária também mostra duas figuras retangulares dos pedestres.

Capítulo 3

Abordagem Selecionada

Neste trabalho, o objetivo consistia em propor um modelo para a pesquisa de uma determinada entidade (pessoa) em sequência de vídeo de personagens animadas em videos. Neste capitulo é apresentada a metodologia descrita no artigo de referência utilizado "Real-world Anomaly Detection in Surveillance Videos" [14], que serviu como modelo de base para este projeto.

Para a realização deste trabalho, criamos duas pastas em que cada uma dela tem vários vídeos, a primeira pasta também considerado de saco (*bag*) contém os vários vídeos onde há presença da pessoa a detetar (identidade), enquanto a outra pasta ou saco contém os exemplos onde não temos a pessoa de interesse (normal). Em termos de aprendizagem automática, o objetivo é maximizar essa separação entre os dois sacos (positivos e negativos), obtendo um modelo capaz de produzir scores diferentes consoante determinada personagem apareça (ou não) num determinado segmento de video. No próximo capítulo será descrito de uma forma mais detalhada o conjunto de dados usado.

$$\mathrm{X_m} = \{x_1, x_2, \ldots, x_n\}$$

Onde $\mathrm{x_i}$ é o i-ésimo saco e m é o número de instâncias de cada saco.

3.1 Aprendizagem em múltiplas instâncias

O MIL generaliza a aprendizagem supervisionada. Na aprendizagem supervisionada, cada instância é descrita por um vetor de características e associada a um rótulo de classe que identifica em que família de objetos este pertence. No MIL, cada instância também é descrita por um vetor de características. Mas, neste caso, a etiqueta da classe não está associado a uma instância, mas sim associada ao sacos em que as imagens ou vídeos pertencem. O saco positivo e o saco negativo [21]. Um saco é negativo se e somente se todas as suas instâncias forem negativas (ou seja, não contem pessoa a detetar) e é positivo se pelo menos uma das suas instâncias for positiva (isto é, contém a pessoa a detetar).

Consideramos o conjunto de dados de treino $\mathcal{B} = \{B_1, B_2, \ldots, B_Z\}$ em que Z é um conjunto de sacos, em que cada saco está associado a uma etiqueta $Y_i \in \{-1, 1\}$ que ele contem N_i instancias; então $B_i = \{x_i1, x_i2, \ldots, x_{iN_i}\}$ uma instancia X_{ij} pode considerada ser etiquetada positiva ou negativa dependendo do saco em que ela pertencia, assim dizemos que $y_{ij} \in \{-1, 1\}$ [22].

$$Y_i = \begin{cases} +1, & \text{if} \quad \exists y \in B_i : \ y_{ij} = +1 \\ -1, & \text{if} \quad \forall y \in B_i : \ y_{ij} = -1. \end{cases} \tag{1}$$

3.2 Otimização

Para a optimização foi usada a função de custo de ”hinge loss”. Esta permite classificar o modelo entre os valores {-1,1} a *hinge* loss funciona muito bem quando as etiquetas dos sacos das instancias são bem conhecidas para optimizar a sair vendo entre valores das classes reais e os que foram previstos. Por fim, é conhecido que a camada de saída da rede de SVM deve ser configurada para ter um único neurónio com uma função de ativação de tangente hiperbólica capaz de gerar um único valor no intervalo de [-1,1].

1. Figura 3.1 Otimizador.

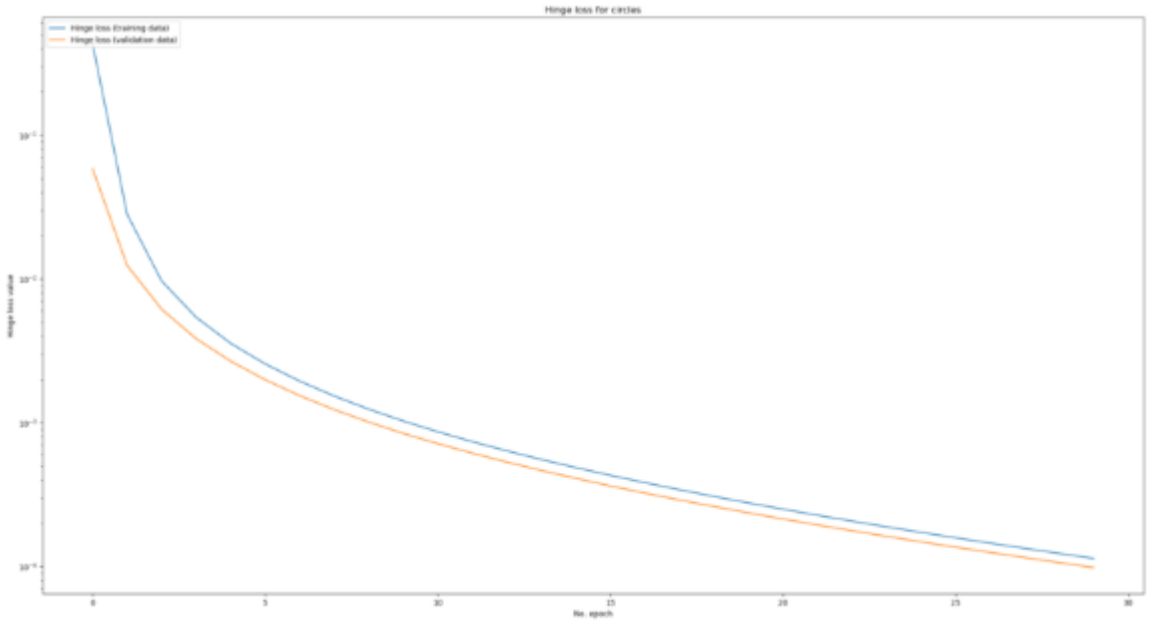


Figura 3.1: Função de custo Hinge loss [11].

$$\min_{\mathbf{w}} \frac{1}{z} \sum_{j=1}^{z} \overbrace{max(0, 1 - Y_{\mathcal{B}_j}(\max_{i \in \mathcal{B}_j}(\mathbf{w}.\phi(x_i)) - b))}^{\text{❍}} + \frac{1}{2}\|\mathbf{w}\|^2, \quad (2)$$

[14] Onde z é o número total de saco que são exemplos de treino ❍é o loss de dobradiça, Y_{Bi} representa a etiqueta ao nível do saco e B_p representa um vídeo positivo como saco positivo, on os segmentos temporários criam instancias individuais positivas $(p^1, p^2, ...p^m)$, B_n denota etiqueta negativa, assim também $(n^1, n^2, ...n^m)$, $\phi(x)$ designa a representação de recursos de um patch de imagem ou segmento de vídeo, b é um bias e w é o classificador.

3.3 O Modelo Deep Ranking MIL

No campo da deteção de pessoa com base dos videos de desenhos animado, é difícil capturar características eficazes e discriminatórias como resultado das variações do corpo, mudança do estilo. As variações são causadas principalmente por escala, roupas, posturas, estar das costas ou de lado, fixação de vista e, cenas dinâmicas.

Como mencionado por Wan et al. [16], quando a entrada é vídeo em vez de imagem estática, essa tarefa precisa de mais esforços já que o modelo precisa aprender a partir das características temporais. No processo da aprendizagem profunda, os recursos podem ser aprendidos automaticamente dos domínios temporal e espacial simultaneamente.

A figura 3.2: apresenta e descreve um fluxograma que ilustra como que o processo é feito; Primeiro temos em entradas uma dataset em formato vídeo de diferentes minutos que vão ser generalizados a um tamanho uniforme para todos de 32 frames por segundo e obtendo assim um segmento de video temporal, colocamos os vídeos positivos e negativo na entrada da rede C3D e, depois obtemos na saída da rede C3D as features(características) como resultado da extração. os recursos extraídos desta rede convolucional tridimensionais em formato de ficheiro fc6-1 são colocadas na entrada na entrada da nossa rede neuronal fully Connected para obtenção do nosso modelo treinado para a Estrutura de aprendizagem supervisionada para pesquisa multimédia, no casa da deteção de personagem em sequência de video de desenhos animados.

1. Figura 3.2 fluxograma do modelo proposto.

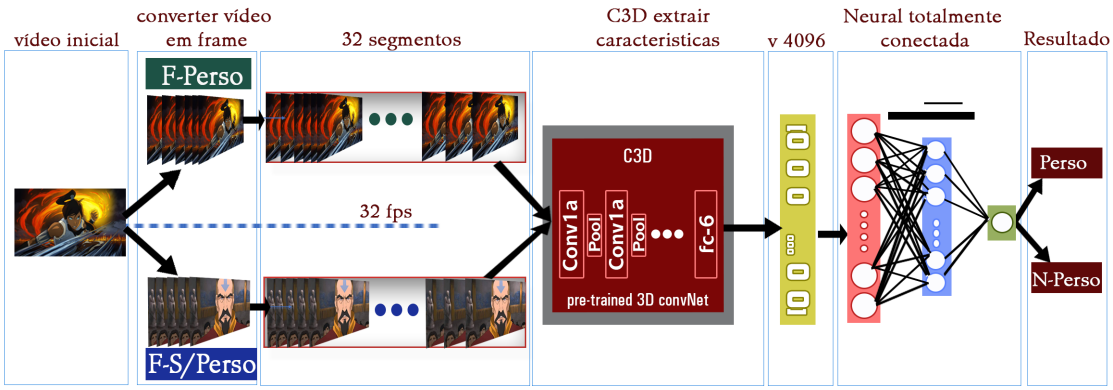


Figura 3.2: Fluxograma do modelo proposto para Estrutura de aprendizagem supervisionada para pesquisa multimédia no caso da deteção de uma personagem na sequencias dos vídeos de desenhos animados. (Adaptada de [14].).

Regressão de várias instâncias. Aqui, cada saco é associado a um único número real, como na regressão standard. Na regressão de MI considera-se que há uma instância em cada saco, chamada de "instância principal", que este determina a etiqueta do saco [17]. O objetivo ideal da regressão de MI seria encontrar um hiper-plano que minimizasse a função de custo quadrada das instâncias primárias em cada saco.

Dum lado por falta dos exemplos de diferentes tipos de abordagens realizadas para Estrutura de aprendizagem supervisionada para pesquisa multimédia no caso da deteção de uma personagem na sequências dos vídeos de desenhos animados, isto deixa muitas equações pela a forma à detetar um personagem nos vídeo pois a situação se torne subjectiva, pelo que diferentes coisas pode influenciar na sua identificação, duas pessoas com a mesma alturas, mesmo tamanho, roupas, estilo de cabelo etc... pode influenciar muito para realizar esta deteção. [14] Queremos que os segmentos de vídeo dos sacos positivos(vídeos com personagem a identificar) tenham pontuações de identificação ou deteção mais altas que os segmentos de vídeo dos sacos negativos onde não há existência da entidade. A abordagem direta seria usar uma função de custo de classificação para obter na saída do modelo uma pontuações altas para segmentos de vídeo positivo (da personagem a detetar) e penalizar a pontuação dos segmentos de vídeo negativo. mas como os segmentos de vídeo que temos não tem são anotados, é neste ponto que propomos a função de objetivo para a classificação das instâncias Múltiplos.

A função objetivo representa a meta a atingir para o problema de Múltiplos instancias. O conjunto de restrições define condições sobre o espaço de estado que as variáveis devem satisfazer. Estes condicionantes são frequentemente de desigualdade ou de igualdade e limitam geralmente o espaço de procura (soluções viáveis). A solução ideal para o problema é encontrar o ponto ou conjunto de pontos no espaço de busca que melhor satisfaça a função objectiva. O resultado é chamado de valor óptimo.

O objetivo principal durante a aprendizagem do modelo, é que a Loss de ”hinge” se minimize, evitando também o modelo a cair na overfitting. Ao treinar um grande número de sacos positivos e negativos, esperamos que durante esta fase o nosso modelo seja o mais generalizado possível para permitir todos os sacos positivos tenham scores mais elevados, enquanto os segmentos negativos tenham os scores mais baixos possíveis.

Capítulo 4

Experiências e Resultados

No capítulo anterior vimos o modelo e a forma em que este esta a funcionar e, neste capítulo nós apresentamos a experiência que tivemos quando usamos o modelo, o que foi entendido durante a fase de aprendizagem, assim como é que o conjunto de dados que temos foi interpretado durante o processo e, por fim os resultados que tivemos utilizando à abordagem escolhida e o conjunto de dados.

4.1 Descrição do conjunto de dados

O recurso mais precioso que nós dispomos é este chamado conjunto de dados (em inglês ”Dataset”), ele constitua o mais precioso de tudo o que devemos necessitar para a realização de um bom estudo nas diversas áreas tal como: inteligência artificial, big data, data ciência, e tantas outras áreas, sem ele nenhum estudo levara numa solução. Temos de utiliza-lo da boa maneira para no final saber qual será o bom método a escolher e de que modo o treino do modelo deveria ser executado e, por fim de poder testar a eficiência e a eficaz do modelo treinado.

O conjunto de dados dados utilizado contém uma grande variação de luminosidade, cores, espaços, plano de fundo não são uniformes. Quando temos dados do mundo real a tratar no caso vídeo, fotos ou informações à proveniência das cameras, smartphone, internet etc... estes dados do mundo real, na maioria dos casos são dados brutos que necessitam de um tratamento antes da sua utilização, e também da forma como estes mesmos serão usadas. Também da mesma forma disponibilizar dois grupos de dados, um grupo de dados para o treino do modelo e o outro para o teste e validação do modelo.

4.1.1 Coleção de dados(vídeos)

A constituição do conjunto de dados foi realizada a partir do site [18], onde os vídeos de desenho animado da Lenda de avatar ”korra” foram baixados de uma forma gratuita. Enquanto estávamos na procura dos vídeos para realizar o nosso estudo sobre Estrutura de aprendizagem supervisionada para pesquisa multimédia, no seu caso-de-uso a pesquisa de personagem animadas em sequências de vídeos de desenhos animados, por uma questão de facilidade em arranjar dados para a formação da nossa dataset. Tentamos arranjar também outros vídeos no YouTube mas a maioria dos vídeos que estavam disponíveis no YouTube os cenários estavam muitos repetidos e misturados, isto rendavam

o processo de criação da dataset muito longo por fator primeiramente deveríamos acompanhar(assistindo) cada vídeo enquanto anotar as história e cenários par depois verificar nos demais vídeos se a mesma história ou o mesmo cenário está sendo repetido, numa série de mais 400 vídeos dum tamanho em tempo variado de 2 minutos a 10h. No final, foi muito trabalhoso usar estes vídeos do YouTube.

Por este motivo que procuramos nos outros sites onde conseguimos os vídeos de series(temporada) "A lenda de avatar korra" num total quase de 20 horas e 20 minutos Conforma a descrição da tabela abaixo por temporada.

Tabela 4.1: Descrição de vídeos por livro (Série ou Temporada).

Books	Números dos vídeos por serie	T_min-T_max	Tot_t/livros
Book_1_Ar	12	22':40" até 25':25"	287'
Book_2_Espirito	14	23':02" até 23':57"	333'
Book_3_Mudança	13	22':02" até 22':58"	294'
Book_4_Equilíbrio	13	22':48" até 23':00"	298?
-	-	-	-
TOTAL	52	22':02" até 25":25"	1212'

4.1.1.1 Descrição da Tabela

A coluna books (ou livro em português) representa o conjunto das temporadas, cada linha desta coluna representa uma temporada da Lenda de Avatar Korra, então Book_1_Ar contem 12 episódios, em que o vídeo com o tempo mínimo(T_min) tem 22 minutos e 40 segundos e o vídeo com tempo máximo(T_max) tem 25 minutos e 25 segundos, assim o Tot_t/livros 1(Book_1_Ar) é de 287 minutos. Book_2_Espirito contém 14 episódios, em que o vídeo com o tempo mínimo(T_min) tem 23 minutos e 02 segundos e o vídeo com tempo máximo(T_max) tem 23 minutos e 57 segundos, assim o Tot_t/livros 2(Book_2_Espirito) é de 333 minutos. Book_3_Mudança contém 13 episódios, em que o vídeo com o tempo mínimo(T_min) tem 22 minutos e 02 segundos e o vídeo com tempo máximo(T_max) tem 22 minutos e 58 segundos, assim o Tot_t/livros 3(Book_3_Mudança) é de 294 minutos. Book_4_Equilíbrio contém 13 episódios, em que o vídeo com o tempo mínimo(T_min) tem 22 minutos e 48 segundos e o vídeo com tempo máximo(T_max) tem 23 minutos e 00 segundos, assim o Tot_t/livros 4(Book_4_Equilíbrio) é de 298 minutos.

4.1.2 Organização e Anotação

Aqui esta descrita como é que o nosso conjunto de dados foi organizado e de mondo que anotamos-lo.

4.1.2.1 Organização do conjunto de dados

Para organizar o nosso conjunto de dados, precisou-se a partir dos 52 vídeos brutos que baixamos disjuntar dos videos que temos as imagens e som (audios associados a cada

video obtido), e depois eliminamos sons porque na fase da deteção não precisamos de ficheiro áudio associado as imagens do vídeo e depois convertemos os vídeos para frames(fotos); assim foi o processo, selecionamos todos os vídeos e convertemos estes para frames. Isto para nos facilitar durante a fase da eliminação das sequências de video sem desnecessários para o estudo, conforme será detalhado na figura abaixa; os frames provenientes dum determinado vídeo ficam agrupados numa só pasta e assim também para os restantes de frames de restantes vídeos. Assim cada pasta representa um vídeo, é as nomeações das pastas são feitas por ordem de extração de frames.

Durante a constituição do nosso conjunto de dados, a organização do nosso conjunto de vídeo foi da seguinte forma; eliminamos as partes que não são importantes para a formação do nosso modelo estas parte eliminadas são as sequências das imagens que não representam nada, elas não fornecem nenhuma informação útil, isto são, as características extraídas nessas sequencias não fazem parte dos recursos que o modelo precisa exemplo as escritas e comentários no início e final do filme e até isto irá aumentar um pouco de dificuldade durante a fase de aprendizagem do nosso modelo o algoritmo não vai achar nessas imagens o padrão que se encontra nas características das outras imagens quer sejam as que tem a nossa personagem(imagens positivas) a detetar ou as que não tem personagem(imagens negativas. Então para não sobrecarregar o sistema com os dados não necessários enquanto ele está aprendendo. as restante são sequencias das imagens que nos guardamos para compor o nosso conjunto de dados.

4.1.2.2 Anotação

Como já foi dito nos pontos acima e capítulo precedente, não classificámos os vídeos detalhadamente em classes. A única maneira de anotar os vídeos foi colocar todos os vídeos positivos em uma pasta única e os que são negativos numa outra pasta.

4.1.2.3 Redimensionamento do vídeos a um único tamanho

Como os vídeos foram obtidos a partir de duas fontes diferentes, estes não tinham a mesma dimensão em termos espaciais. A primeira fontes disponibilizou vídeos com tamanho de 640 x 360 pixeis e a segunda de 1920 x 1080 pixeis. Todas as imagens do nosso conjunto de dados foram redimensionadas para 640 x 360 pixeis, pois essa resolução é necessária para a extração das características C3D e teste do modelo.

4.1.3 Normalização

A normalização consistiu na reconstruir dos vídeos temporários constituídos por 32 frames por segundos a partir das frames que consideramos importante ”X” no processo da organização da dataset, então cada vídeo tem uma duração de 16 segundos que equivalem

a 480 frames. Depois de tudo, separamos os que tenham personagem(positiva) colocamos numa pasta e as que não tinham personagem(negativa) numa outra pasta diferente como ilustra as figuras abaixas.

Figura 4.1: Sequências de vídeos com pessoa em que pesquisamos para detetar dento da pasta positiva.

Figura 4.2: Sequências dos vídeos sem pessoa em que pesquisamos para detetar dento da pasta negativa.

Por aqui teremos as diferentes sequências de vídeos com e sem pessoa a pesquisar nas diferentes pastas que são duas, uma colocamos todos os vídeos que contem o avatar ”Korra” que é a nossa personagem a detetar e consideramos a pasta como positivo e outra contem

diversas personagens mas nós não precisamos pesquisar elas.

4.1.4 Conjuntos de treino e teste

Primeiramente o erro que vamos evitar é, portanto, avaliar a qualidade do nosso modelo usando os mesmos dados que vão ser usados durante a fase do treino. De fato o conjunto de dados para o treino é um conjunto de exemplos utilizados para ajustar os parâmetros do modelo(por exemplo pesos das ligações da rede), o modelo é totalmente otimizado para os dados com os quais ele foi traindo. Por isso é preso avaliar o modelo por um conjunto de dados que o modelo nunca viu! Daí que surgiu o conceito de dividir o conjunto de dados para o treino e outros para avaliar ou testar a performance do modelo.

Tabela 4.2: Divisão do conjunto de dados(conjunto de treino e conjunto de teste.

Conjunto de dados	**Qtd.**	**Com Avatar Korra**	**Outras pessoas**	**%**
Training	3225	953	2272	80
Testing	806	238	568	20
-	-	-	-	-
Total	4061	1191	2840	100

4.1.4.1 Descrição da tabela 4.2

Logo depois de processo de normalização, tivemos um total de 4061 vídeos pela qual 1191 são vídeos com a personagem Avatar Korra e 2840 são vídeos normais sem personagem Avatar Korra(1191 vídeos positivos e 2840 vídeos negativos). Então colocar 80% dos dados na pasta para o training set(conjunto para o treino) e 20% para o testing set(conjunto para o teste) isto é, o conjunto de dados para o treino são 953 vídeos com a personagem Avatar Korra e são 2272 vídeos sem personagem Avatar Korra e, o conjunto para o teste consiste em 238 vídeos com a personagem Avatar Korra e 568 vídeos sem personagem Avatar Korra. Cada conjunto que seja de treino ou teste foram colocados nas pastas diferente.

4.1.5 Configuração

O experimento é realizado em uma máquina com a seguinte configuração, maquina i7, com:

- uma unidade de processamento gráfica GPU GeForce serie 10 (GTX-1080) com uma memória de vídeo de 8G;
- memória RAM de 16Go;
- sistema operativo linux 18.04 de 64-bit;

- CUDA 10.1;
- cuDNN 7.6.5
- Python 3.6;
- PyCharm 2019.3

4.2 Implementação

A implementação iniciou-se logo depois termos um conjunto de dados pronto depois da normalização nós extraímos as características. as características das imagens determinam ou revelam os diferentes pontos(valores) comuns que existem entre diferentes imagens e, também os diferentes pontos(valores) que não existem entre elas. [12] usamos o modelo convolution 3D(C3D) para extração das características de dos os vídeos(treino e teste). Extraímos características a partir do modelo C3D desenvolvido pelo facebook que baixamos no [20], exatamente como se refere na figura 3.2 colocamos na entrada da rede C3D o nosso conjunto de dados que contem todos vídeos temporais de 16 segundos e obtemos na saída um do modelo C3D un vetor de 4096 dimensões.

1. Figura 4.4 Extração das características do nosso conjunto de dados com o modelo C3D.

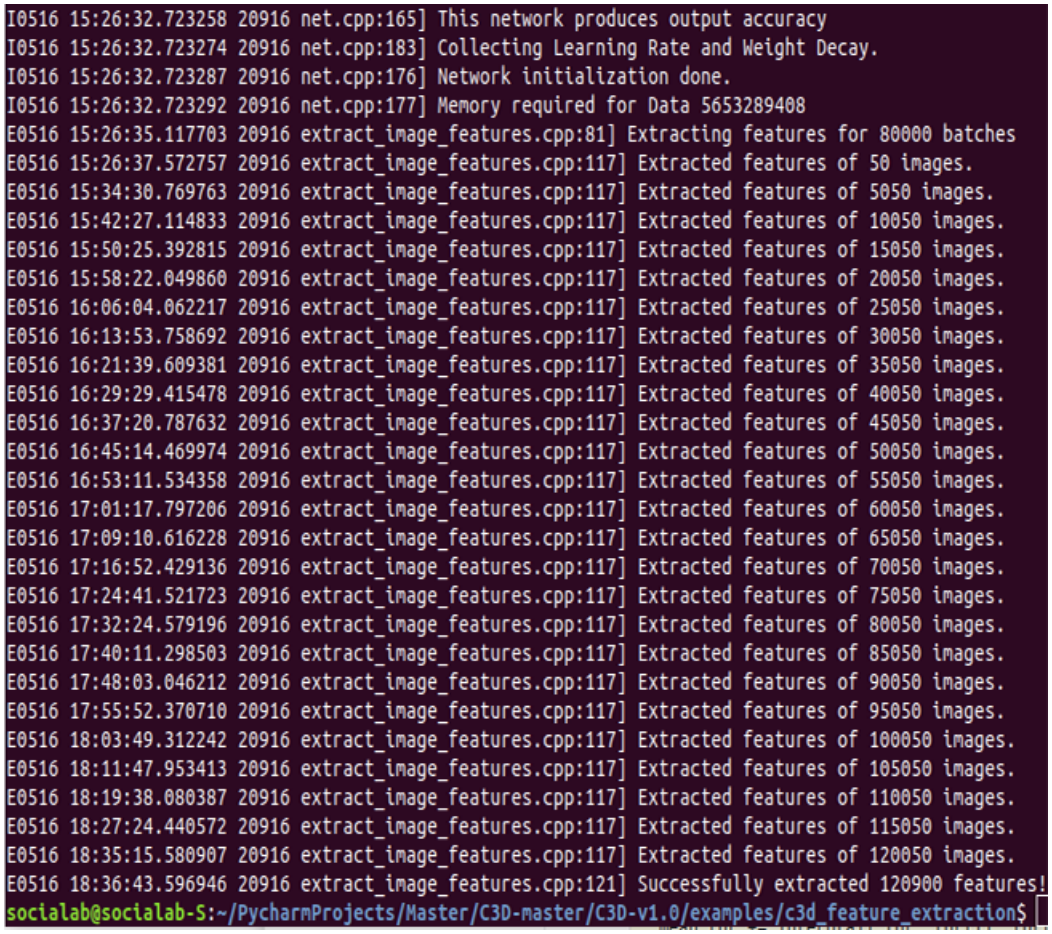


Figura 4.3: O Modelo C3D extraindo as características para cada pequenino clip de 16 quadros, o total dos clipes extraídos são 120900, sabendo que um clip equivale a 16 quadros e a extração é feita com um tamanho de bloco (*batch size*) de 50 blocos de 16 quadros. De forma explicita um vídeo é igual a um segmento, então um bloco de 16 quadros de um vídeo representem neste caso cada instância de um segmento. na saída do modelo obtemos um recurso(características) em forma de um vetor de 4096 numero que ele represente um segmento do nosso conjunto de dados para a deteção de Avatar Korra.

O resultado do modelo C3D que é um vetor de 4096 numeros será usado para representar os recursos complexos de um determinado vídeo, este recurso (vetor de 4096) vai servir na entrada do nosso classificador de arquitetura Fully connected ”FC” com 3 (três) camdas totalmente conectadas com os seguintes parâmetros 512 neurónios na primeiras camada totalmente conectados com 32 neurónios da segunda camada e estes 32 são totalmente conectados com 1 neurónio da ultima camada. [23] para evitar que o sistema caísse em sobre-ajustamento muito cedo antes que o processo de aprendizagem concluísse usamos DropOut de 0.6(60%) para regularização do modelo; foi depois usado o inicializador de peso glorot_normal, primeira camada usamos função de activação Relu e na ultima camada usamos o Sigmoid.

4.2.1 Otimização

Treinamos a rede, otimizando as tarefas de classificação, nesta experiência, comparamos 4 (quatro) diferentes otimizadores e depois da criação dos 4 (quatros) modelos, nós testamos a eficácia de cada um deles separadamente. Pode-se descobrir que entre os 4(quatros), a precisão do otimizador SGD ”Stochastic Gradient Descent” é relativamente alta comparando com os outros. Portanto, foi por este motivo adotamos por este otimizador nos testes de modelo subsequentes. Durante o processo do treino constatamos que os otimizadores Adam e Adagrad que têm a velocidade de convergência mais rápida e podem até economizar o tempo de treino, em comparação com o SGD. Portanto os resultados obtidos depois de teste são muito próximos as diferencias en accuracy só são vistas depois das virgulas. Mas como é de habito; sempre o melhor quem ganha. A tabela abaixo mostra os resultados experimentais.

Tabela 4.3: Tabela comparativa dos otimizadores testados.

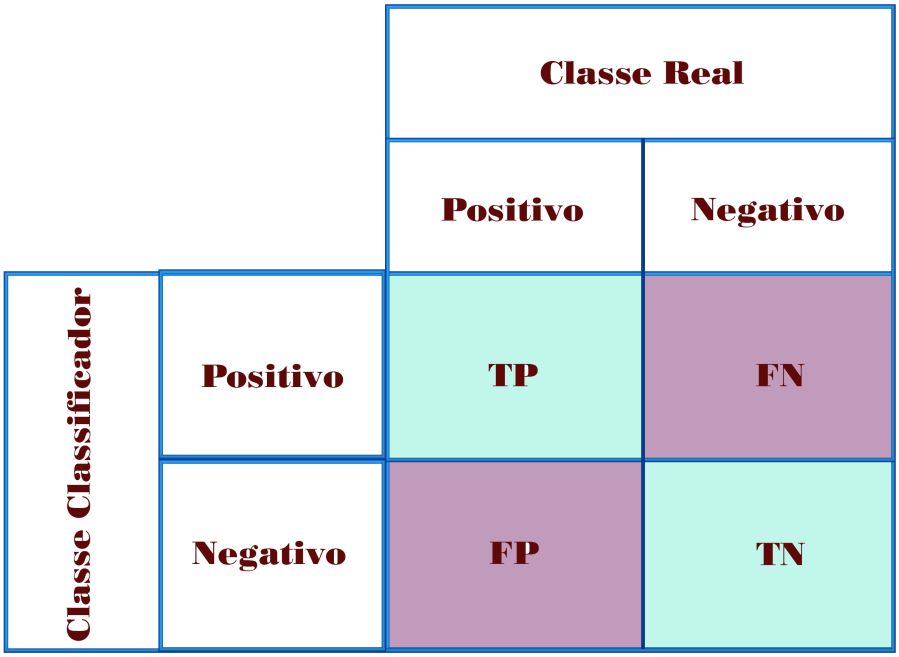
Otimizadores	**AUC (%)**
RMSprop	66,00%
Adagrad	67,00%
Adam	67,26%
SGD	67,68%

4.2.2 Métricas e Avaliações

O principal objetivo de um sistema baseado na Estrutura de aprendizagem supervisionada em multimédia no seu caso de uso de deteção de uma determinada pessoa no caso de Avatar Korra é ter a capacidade de distinguir o Avatar Korra com as restantes pessoas desconhecidas ou não importante do nosso método de pesquisa. um sistema modelo daria as respostas certas durante o processo de deteção, então para avaliar a performance do nosso modelo é inevitável que o modelo verifica os 4(quatro) fundamentais métricas:

- Taxa de Verdadeiros Positivos; - O padrão é verdadeiro porque tem o Avatar Korra nas sequências do vídeo e o classificador o classifica como tal.
- Taxa de Verdadeiros Negativos; - O padrão é falso porque não tem o Avatar Korra nas sequências do vídeo e o classificador o classifica como tal.
- Taxa de Falsos Positivos; - o padrão é Falso porque não tem o Avatar Korra nas sequências do vídeo, mas o classificador o classifica verdadeiro como se havia Avatar Korra nas sequências do vídeo.
- Taxa Falsos Negativos. o padrão é Verdadeiro porque tem o Avatar Korra nas sequências do vídeo, mas o classificador o classifica falso como se não havia Avatar Korra nas sequências do vídeo.

Tabela 4.4: A Matriz de Confusão.



Para avaliação do nosso modelo treinado as técnicas da Machine Learning a necessidade de utilizar a curva ROC ”Operating Operating Characteristic” foi inevitável, ela mostra o desempenho de un modelo de classificação em seu limites para classificar, No entanto a sua Taxa Positiva verdadeira assim que a Taxa Falsa Positiva são necessariamente requeridas para visualizar a curva ROC[24].

True Positive Rate

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR \tag{3}$$

.

False Positive Rate

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR \tag{4}$$

.

True Negative Rate

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR \tag{5}$$

False Negative Rate

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} = 1 - TPR \tag{6}$$

A precisão padrão e a área sob a curva (AUC) são usadas como métricas para a classificação. Na classificação binária, a precisão padrão pode ser escrita como:

$$Accuracy = \frac{TP + TN}{P + N} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

1. A Curava ROC para a classificação do modelo.

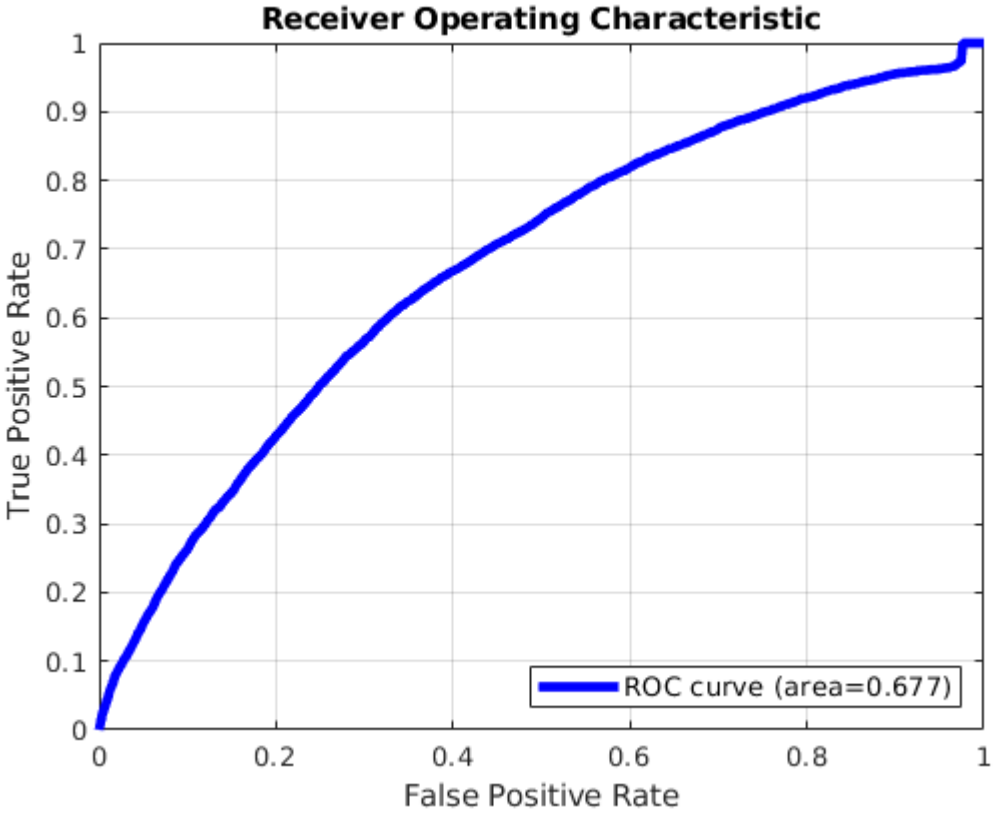


Figura 4.4: A curva ROC obtida utilizando o otimizador SGD com a area da curva ROC de 67,7%.

Portanto como já foi esclarecido nos pontos acima, que o conjunto de dados que nós utilizamos no caso da nossa pesquisa é tem muitas variações, assim com AUC da curava ROC que obtemos a partir do modelo treinado com a otimização em SGD, supera as áreas da curva ROC que os modelos treinados com as redes que utilizaram otimizadores diferentes do SGD. Com SGD alcançamos 67,7 d'area da curva de ROC(AUC).

4.3 Resultados

aqui nós apresentamos os resultados da aplicação do nosso método em alguns vídeos do nosso conjunto de dados, objetivo é mostrar os casos em que o nosso modelo conseguiu dar os resultados esperados e nos casos em que ele não nos deu os resultados de sucesso.

4.3.1 Casos de sucesso

Verificamos neste ponto todos os casos em que o modelo treinado deu sucesso, isto é verificamos nas sequências do video que temos onde é que o modelo previu que foi identificada a nossa entidade ”Avatar Korra” e isto é verdadeiro que ela esta mesmo numa parte das sequências do vídeo. e o modelo disse que a pessoa em que pesquisa mos não esta no vídeo e é verdadeiramente que ela não esta presente em nenhuma das sequências do vídeo.

4.3.1.1 Sucesso 1

A figura em baixo nos mostra o Avatar Korra dentro de um carro, e o modelo conseguiu lhe detetar muito bem mesmo que ele está numa pose lateral (perfil). A certeza de sensibilidade durante a classificação foi no mínimo 88

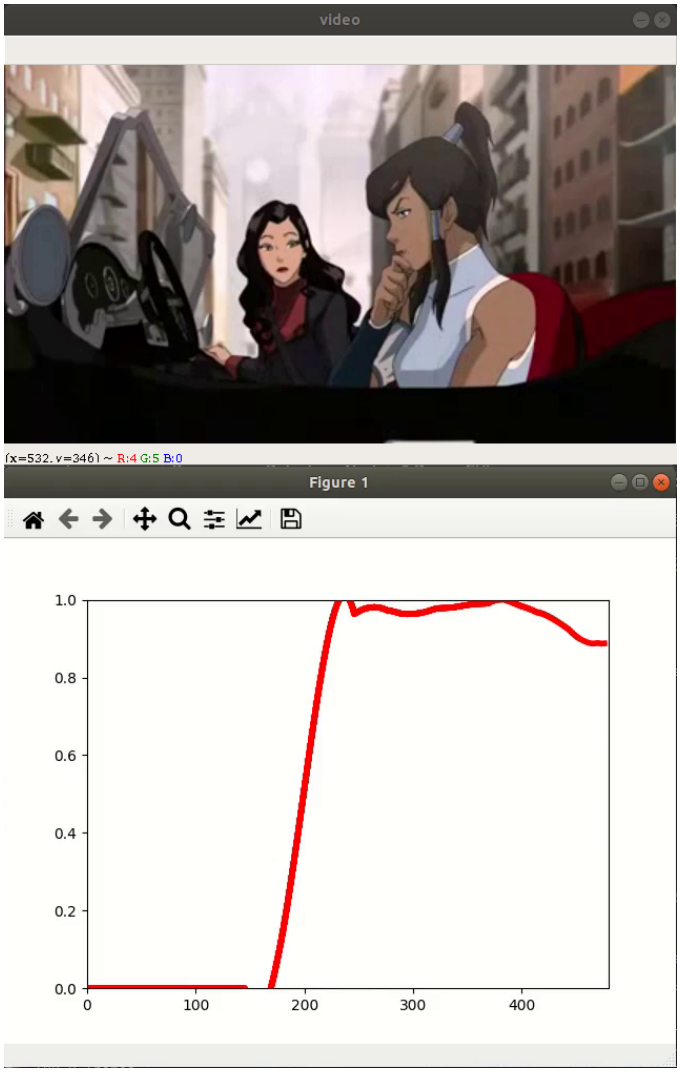


Figura 4.5: Caso de sucesso Avatar Korra na postura de Lado.

4.3.1.2 Sucesso 2

Aqui na figura em baixo, o modelo conseguiu detetar o Avatar Korra estando na postura de frente como ilustra a image 1 e gráfico 2 e, na image 3 com o seu gráfico 4 o modelo detetou o Avatar Korra mesmo estando das costas, isto nas mesmas sequências de um video. O sistema conseguiu classificar as duas posições numa sensibilidade superior a 87%.

Figura 4.6: Caso de sucesso Avatar Korra na postura de frente e das costas.

4.3.1.3 Sucesso 3

A figura seguinte nos mostra uma coisa muito diferente de habito do Avatar Korra, o modelo classificou ela estando na postura de lado, com uma roupa bem diferente das roupas que o modelo ja detectou nas figuras precedentes, e a mesma roupa com uma cor diferente da cor das roupas precedente.

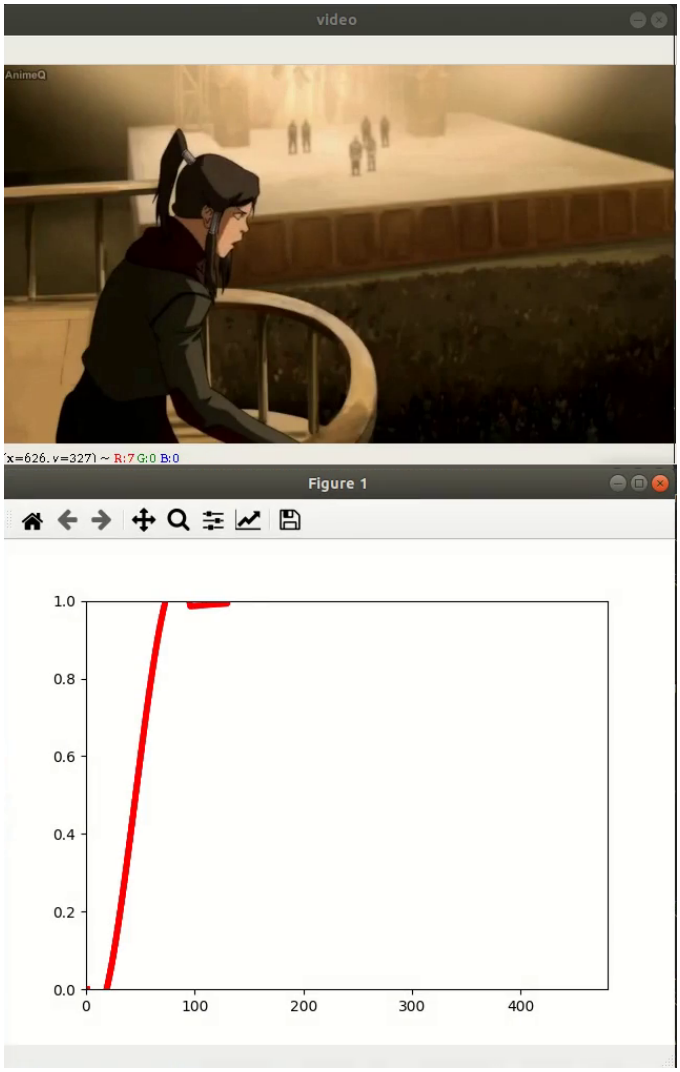


Figura 4.7: Caso de sucesso Avatar Korra na postura de de lado com roupa diferente e de cor diferente.

4.3.1.4 Sucesso 4

Um outro exemplo na figura a seguir que demonstrou a eficácia do método escolhido, o modelo classifica o Avatar Korra desta vez cobrindo uma parte da cabeça dela usando assim uma roupa de capuz muito mais diferente das roupas que já usadas nas figuras precedentes. assim o sistema demonstra que apesar da forma do seu cabelo seja também uma das características que lhe identifica, o modelo pesquisa muitas mais outras características importante que lhe classifica com tal.

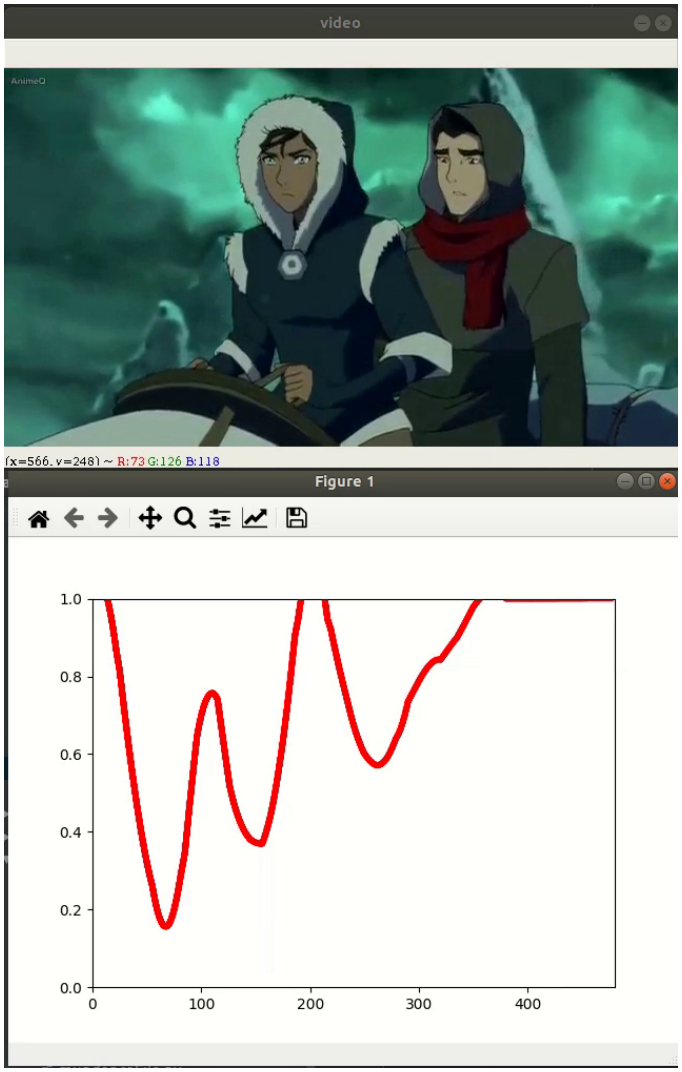


Figura 4.8: Exemplo de um caso de sucesso, em que o Avatar Korra com roupa de capuz a cobrir a sua cabeça e a cor da roupa diferente.

4.3.1.5 Sucesso 5

Durante um período no máximo de 15 segundos, as sequências do mesmo vídeo foi alteradas, capturando assim na diferentes frequências e posturas as pessoas filmadas, mas o que mostra a figura abaixo é que, apesar de todas essas mudança de perceção do camera, o modelo conseguiu detetar com uma sensibilidade de mais de 93% mantendo a especificidade do processo da deteção a cima desta mesma percentagem. sabendo que o Avatar Korra esta num estilo diferente de cabelo de costume e, esta com cabelo cortado.

Figura 4.9: Exemplo de caso de sucesso Avatar Korra nas diferentes posturas com o estilo de cabelo diferente e cortado.

4.3.1.6 Sucesso 6

Na figura a baixo apresentamos caso de sucesso em que o sistema não identificou o Avatar Korra e na realidade, ela não estava também presente nas diferentes sequências de videos, então as características que identifica ela, não foi encontrada nos videos durante o processo de teste ou validação do modelo. Portanto o modelo de treino foi treinado para logo que ele identifica as características do Avatar Korra ele possa lhe classificar como entidade detetada. mas nas imagens que estão representada na figura a seguir não ela não foi detetada e o modelo então disse a verdade.

Figura 4.10: Exemplo de caso de sucesso de não classificação do Avatar Korra nas diferentes sequências de videos.

4.3.2 Falhas no modelo

Aqui apresentemos as falhas que houve no método, como já é conhecido que o conjunto de dados foi um dos conjuntos mais difíceis. o conjunto contem sequências de imagens tal como corrida, luta, e sorriso(sabemos que quando os desenhos animados sorriam, as bocas deles ficam muito grande etc...), mudanças das roupas, mudança de estilo de cabelo, fundo da imagens tem mesmo aspecto com as roupas do Avatar isto influenciou negativamente durante o processo de aprendizagem.

4.3.2.1 Falha 1

Na figura a seguir notamos que quanto mais a distancias mais difícil fica para o modelo detetar o Avatar Korra.

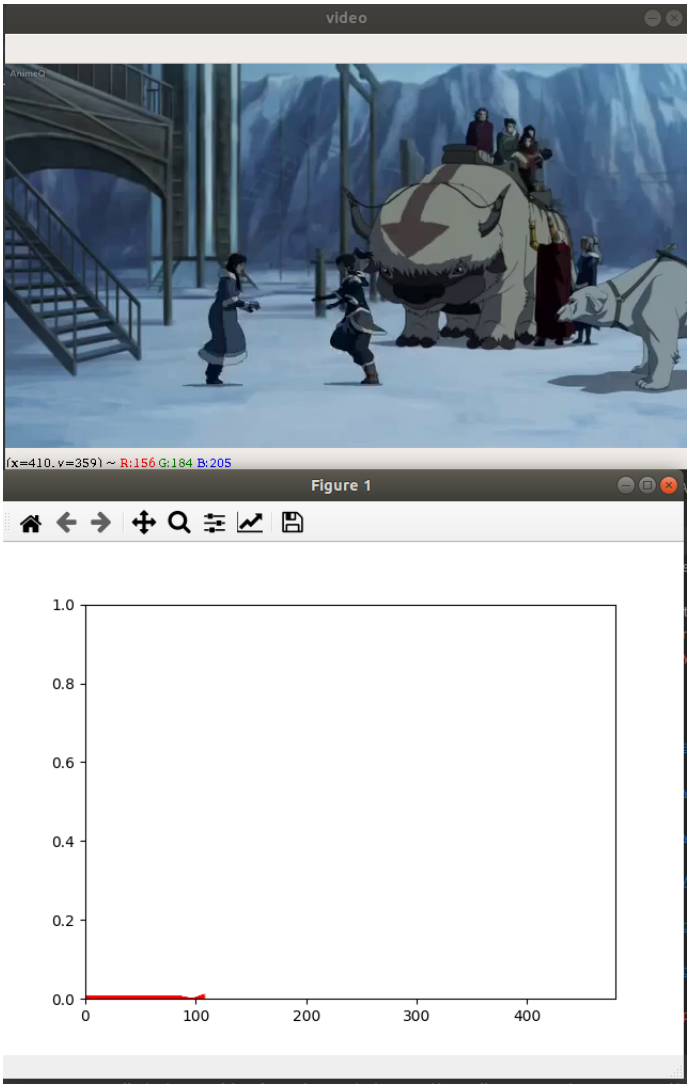


Figura 4.11: O método tem dificuldade de identificar o Avatar Korra quando ela está distante e aparece pequena da sequência de vídeo.

Como podemos observar que o método mostra logo que ela foi mais próxima à personagem, após alguns mili-segundos já o modelo conseguiu deteta-la.

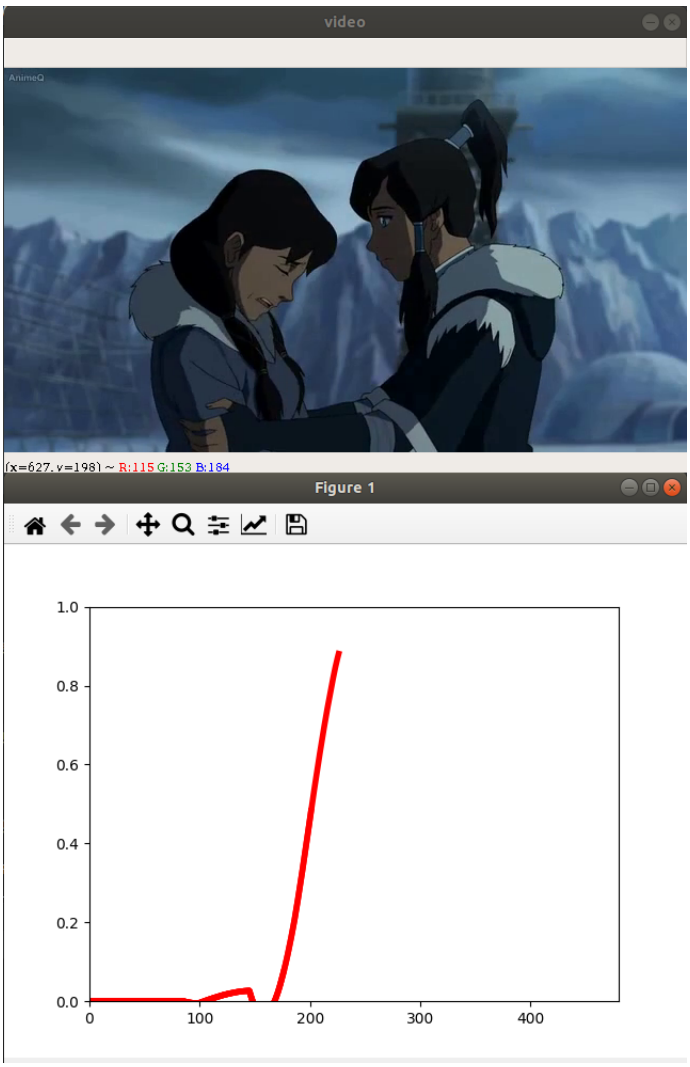


Figura 4.12: O método consegue identificar o Avatar Korra no mesmo video logo depois dele ser aproximado.

4.3.2.2 Falha 2

A figura a seguir apresenta uma falsa alerta de deteção de Avatar Korra no sistema, mas ao ver bem não há nenhuma presença da pessoa Avatar Korra. mas a senhora vestido em vermelho tem uma semelhança na cara com Avatar Korra e na sua traz tem uma outra senhora que tem a postura, cor da pele e estilo de vestuário um pouco parecido com Avatar Korra. Em tudo o modelo encontro um grande numero das características muito proximo das características que apresenta o Avatar Korra.

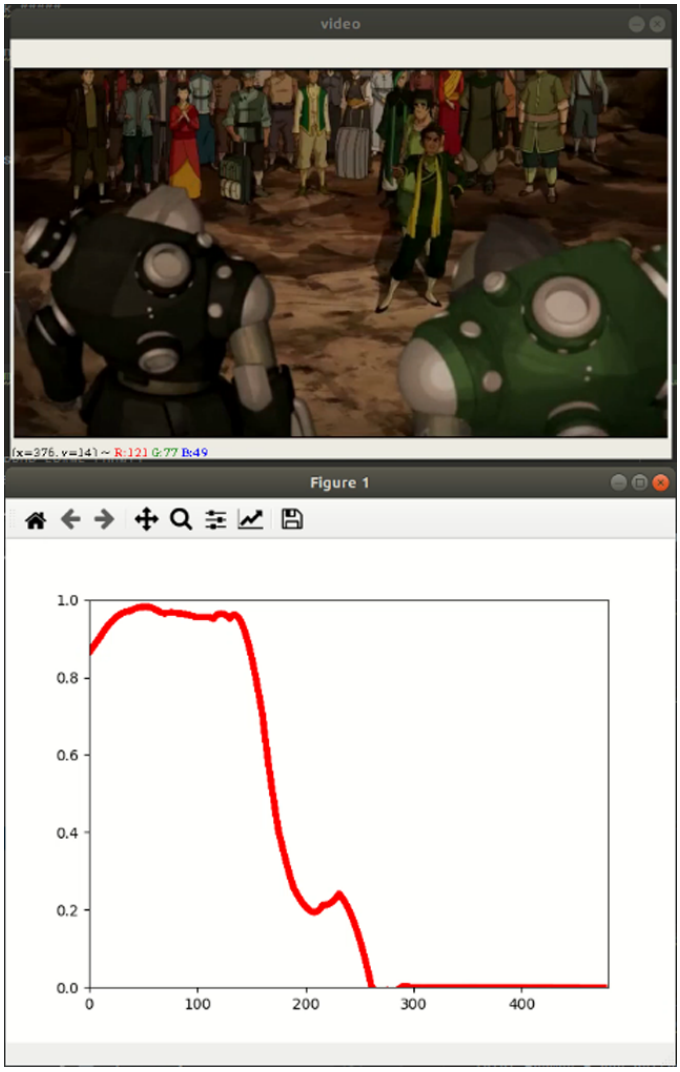


Figura 4.13: Caso de falha no modelo com um Falso Positivo.

4.3.2.3 Falha 3

A presente figura ilustre uma falha no sistema, na figura com sucesso vimos com o modelo depois de treinar, consegue identificar o Avatar, mesmo com roupas diferente, estilo de cabelo, a cor da roupa também diferente, mas nesta presente figura, o modelo não conseguiu classificar o Avatar Korra, um dos primeiro motivo ela usou uma roupa que nunca foi usada nas figuras precedente, alem disso, a proteção da cabeça que ela usou não permitiu o modelo encontrar pelo menos um numero reduzido das características que representa Avatar Korra. O sistema falho, por ele disse que não encontrou enquanto que ele estava lá presente, o sistema considera isso como um Falso Negativo.

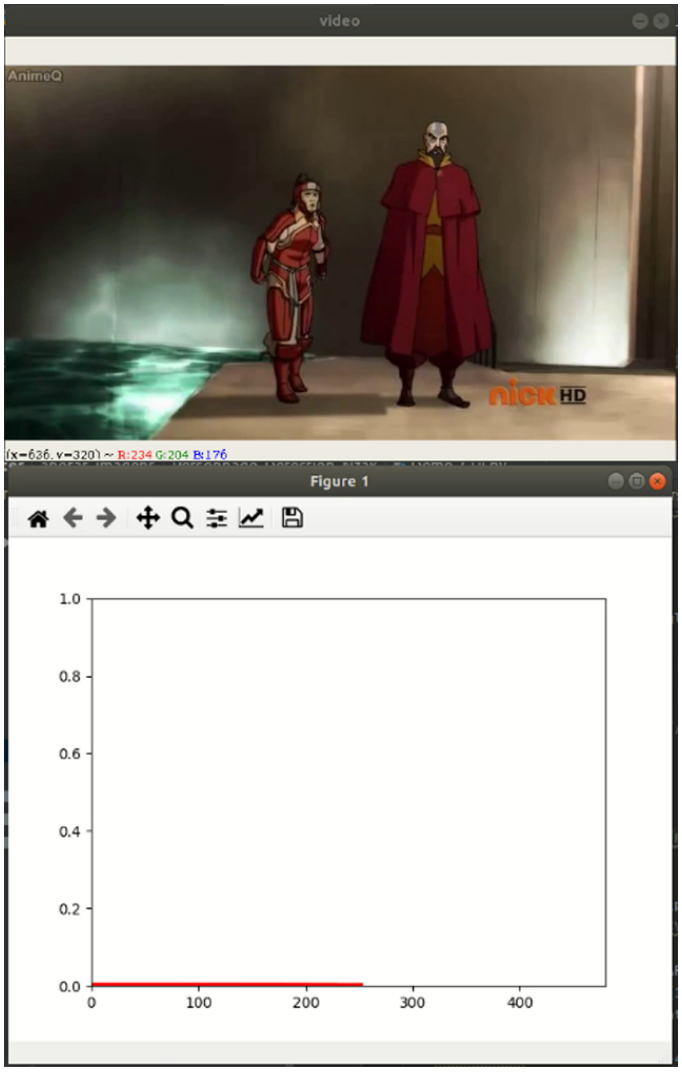


Figura 4.14: Caso de falha no modelo com um Falso Positivo.

4.3.2.4 Falha 4

Observamos uma falha no sistema quando o Avatar Korra esta deitada no chão, o sistema não consegue lhe detetar, na maioria dos caso durante o treino do modelo, o algoritmo não encontrou muitos exemplos com esta posição. então ao pesquisar o Avatar Korra com esta posição, o sistema lhe considera como um outro objeto diferente do Avatar Korra, além disso também houve problema do seu cabelo que também estava espalhado. Portanto conjunto de tudo isso que representam um fracasso durante a fase da deteção da personagem.

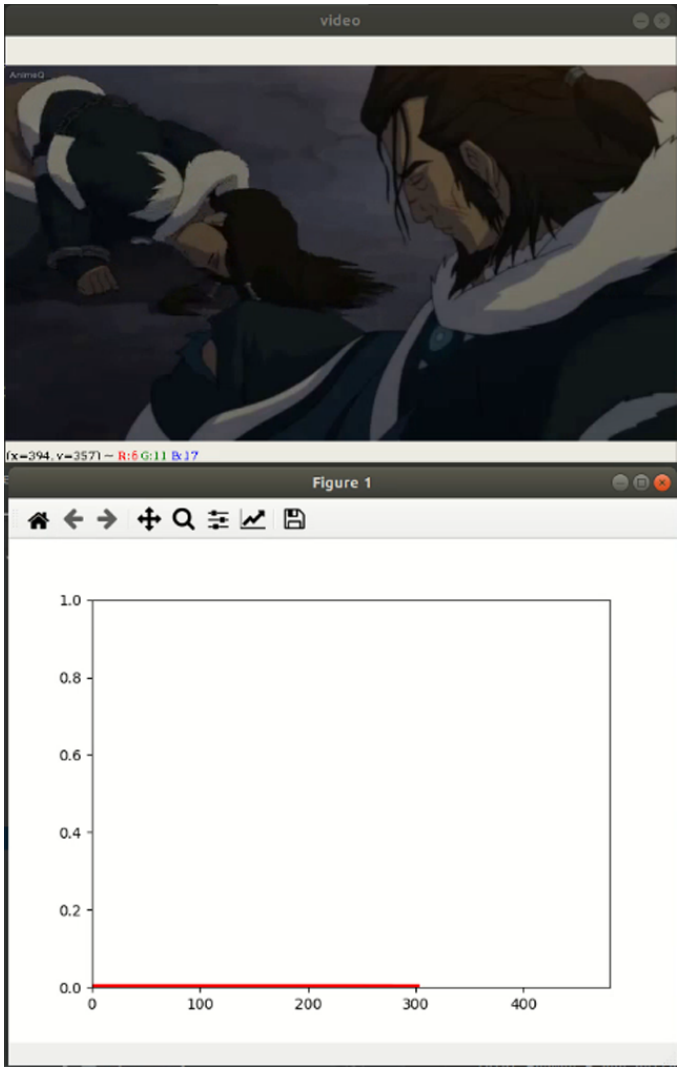


Figura 4.15: Caso de falha no modelo quando o Avatar Korra esta na posição deitada no chão.

4.3.2.5 Falha 5

Uma outra falha que nos observamos é durante as atividades muito rápido por exemplo da Luta, sabendo como se comportam os desenhos animados, as posturas que eles tomam durante a corrida, luta, eles tomam uma forma diferente de habito, isso faz com que as sua características muda e torna ser muito difícil classificar o Avatar Korra.

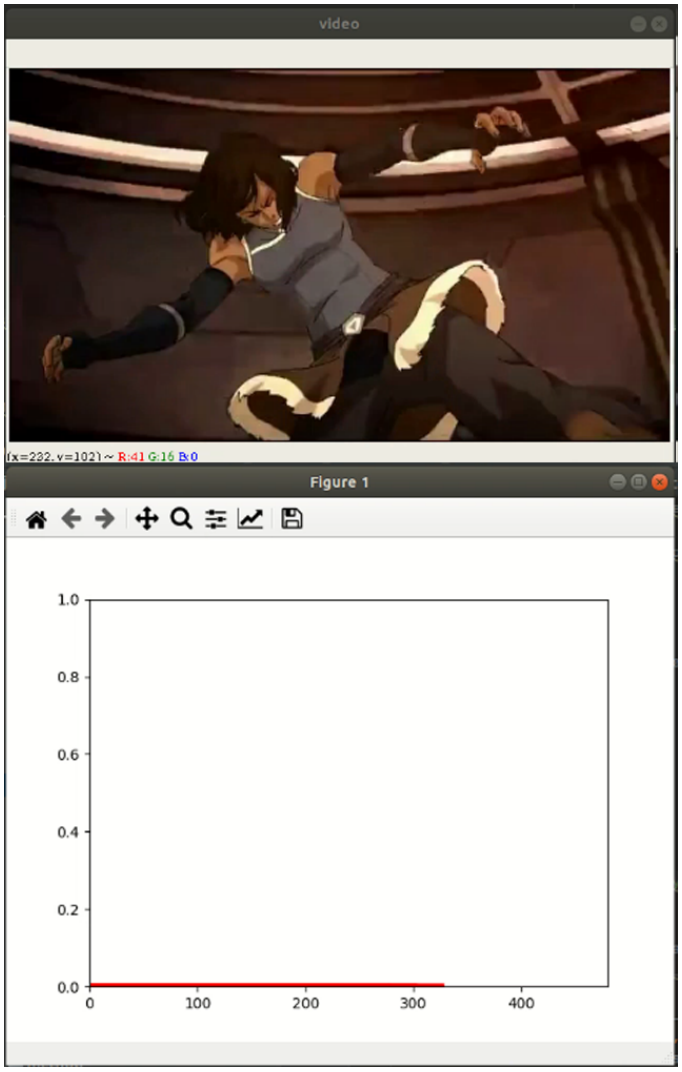


Figura 4.16: Caso de falha no modelo com um Falso Positivo.

Capítulo 5

Conclusões e trabalhos futuros

O presente capítulo concluí-se no cumprimento de todos os objetivos que nós propomos, falaremos a cercas das técnicas usadas, assim como das capacidades e aplicabilidades das mesmas técnicas desenvolvidas e também das limitações do método escolhido. Depois deste resumo conclusivo são também apresentadas no final algumas sugestões para trabalhos futuros.

5.1 Conclusão

Neste trabalho foi apresentada uma abordagem da Inteligência Artificial no ramo da Visão Computacional baseada em técnicas de aprendizagem supervisionada para pesquisa multimédia. Em particular, o seu caso de estudo centrou-se na ”pesquisa de personagem animadas em vídeos de desenhos animados”. Esta pesquisa possibilitou a compreensão sobre as técnicas de deteção e reconhecimento de um indivíduo em sequências dos vídeos tipicamente utilizadas.

O trabalho realizado ao longo desta dissertação permitiu alargar os conhecimentos teóricos, teórico-pratico e práticos. A abordagem escolhida consistiu em ter duas estruturas da redes neuronais artificiais para deteção e reconhecimento de uma entidade. A primeira rede Convolution3D receba na sua entrada os vídeos separados em dois sub-conjuntos: 1)numa parte os vídeos contendo o Avatar Korra, e 2) na segunda parte os restantes vídeos. Esta rede convolucional permitiu extrair as características dos videos e obter um vetor de 4096 características. Estes vetores são depois divididos em 32 segmentos, e usados como entrada de uma rede de classificação para que o modelo possa treinar e somente durante este processo de aprendizagem que conseguimos ter no final um modelo treinado com o optimizador SGD.

Para superar o problema de perda de tempo da anotação e manter a alta precisão na deteção, foi usada uma abordagem de aprendizagem baseada em várias instâncias (MIL). No entanto, na fase de avaliação dos modelos, todos os dados estavam anotados ao nível da imagem (frame), por forma permitir obter resultados fiáveis sobre as taxas de sucesso dos modelos desenvolvidos.

5.1.1 Casos de sucessos e falhas

Em função das experiências realizadas e dos resultados obtidos, concluímos que os modelos desenvolvidos são capazes de obter uma taxa de sucesso - apenas - relativamente alta na deteção da personagem de interesse. Após a fase de aprendizagem, a fase de teste e validação do modelo, permitiu identificar alguns casos típicos que conduzem ao sucesso/falha nos resultados. Quando testamos o modelo treinado ele nos deu alguns resultados que são satisfatórios e outros que necessita de uma melhoria nos trabalhos futuros. Mais quando a taxa de classificação atinge 50% para cima, o sistema deve considerar que a pessoa foi detetada e quando for inferior ele responde que não foi detetado.

Em relação aos casos de sucesso, o modelo treinado conseguiu identificar e classificar o Avatar Korra nas varias posturas, o que foi considerado um sucesso evidente. O sistema conseguiu a detetação mesmo quando ela está de costas, de perfil, ou com diferentes estilos de cabelo, quando está completa ou parcialmente visível, também quando ela tem vestidos diferente até mesmo de diferentes cores. Globalmente, os casos de sucesso foram em muito maior número que os casos de insucesso. No geral, as probabilidades de deteção do Avatar Korra é muito maior na varias posturas, nas diversas visões nas percepções das cameras ele esta sendo classificada.

No entanto, como todo trabalho humano tem falhas e que essas falhas podem se observadas, o nosso também não escapou desta realidade (vários casos de falsos positivos foram também observados, apesar de em quantidades relativamente reduzidas). O modelo não conseguiu reconhecer o Avatar Korra quanto ela estava com um ou outro vestido em particular, apesar de nesses casos o sistema ja ter feito deteções com sucesso em situações muito parecidas. Um outro caso problemático observado está relacionado com oclusões parciais da região da cabeça a partir de 60% da cabeça, em que praticamente todos os traços distintivos da personagem estão ocultos. Finalmente, o factor final para o insucesso foi o problema da escala, isto é, os casos em que a personagem de interesse aparece muito pequena na imagem, o que não permite que o extrator de características obtenha descrições minimamente fiáveis e condizente com o conjunto de treino recebido.

5.1.2 Algumas dificuldades

Houve muitas dificuldades durante a realização deste trabalho, como também durante o processo de aprendizagem, mas a maior dificuldade encontrada foi o conjunto de dados utilizado. Este conjunto apresenta uma elevada variabilidade e complexidade, com variações em nível da personagem, ambiente/luminosidade, tempos (dia, noite), e diferentes atividades(luta, corrida) etc...

Noutra perspetiva, a escolha do algoritmo de aprendizagem foi também problemática, tendo os primeiros otimizadores utilizado conduzido a faces resultados em termos da ca-

pacidade de generalização dos modelos inferidos. Posteriormente, só através da utilização do Stochastic Gradient Descend (SGD), foi possível obter uma redução significativa do número de falsos positivos. De forma que sabemos que os desenhos animados são, no poucos dos casos quando o Avatar Korra falava ou sorria com a boca extremamente aberta, o modelo não reconhecia o objeto de interesse, por que este perdia muitas das características mais importantes para a sua definição.

No geral, conclui-se que apesar de o valor global de AUC poder ser considerado relativamente fraco, deve-se considerar que o modelo conseguiu obter múltiplas situações em que o Avatar Korra foi corretamente detetado. Ainda, espera-se que com base num conjunto de dados real, se possa obter um valor de AUC bem mais satisfatório.

5.2 Trabalhos futuros

Na sequência do presente trabalho sugerimos alguns pontos para eventuais trabalhos futuros, entre os quais os seguintes:

- Melhorar a precisão do modelo, reduzindo assim a taxa dos erros, quer em termos de falsos positivos, ou falsos negativos.
- Realizar o projeto com um conjunto de dados muito maior que este que usamos e desta fez arranjar o conjunto de dados dos humanos e não desenhos animados.
- Usar o projeto para a supervisão de todos criminosos do monde real, contendo as diferentes imagens(vídeos) de cada criminosos conhecidos polos sistemas de seguranças nacional e privadas.
- Realizar o projeto utilizando todas as cameras das cidades ou país para pesquisas e controlo de algumas entidades.

Bibliografia

[01] Nathalie Mayer, “Un appareil photo ultrarapide filme des objets transparents”, Futura-Science, le 21-01-2020.
disponível em: https://www.futura-sciences.com/sciences/actualites/matiere-appareil-photo-ultrarapide-filme-objets-transparents-35315/

[02] http://datascienceacademy.com.br/blog/o-que-e-visao-computacional/, Postado em janeiro 24, 2018

[03] P.Bertolino, G.Foret, D.Pellerin, ”Détection de personnes dans les vidéos pour leur immersion dans un espace virtuel”, GRETSI, Groupe d’Etudes du Traitement du Signal et des Images, 2001.

[04] G.L. ForestiC. MicheloniC. Piciarelli, ”Detecting moving people in video streams ”, Pattern Recognition Letters 26 (2005) 2232–2243, 1 February 2005

[05] S. A. N. C. R. M. O. M. e. N. P. N. Bird, “Real time, online detection of abandoned objects in public areas” Proceedings 2006 IEEE International Conference on Robotics and Automation, pp. 3775-3780, 15-19 May 2006.

[06] J.-Y. C. e. L.-G. C. H.-H. Liao, “A Localized Approach to Abandoned Luggage Detection with Foreground-Mask Sampling,” IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance, pp. 132-139, 1- 3 Sept. 2008.

[07] A. L. Sumalan, L. Ichim and D. Popescu, ”Person Detection in Video Surveillance”, 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Iasi, Romania, pp. 1-6, 2018.

[08] M. Ahmad, I. Ahmed and A. Adnan, ”Overhead View Person Detection Using YOLO,” 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, USA, 2019, pp. 0627-0633.

[09] B. Jagadeesh and C. M. Patil, ”Video based action detection and recognition human using optical flow and SVM classifier”, 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2016, pp. 1761-1765.

[10] J. Wang and Y. Hou, ”Pedestrian Fall Action Detection and Alarm in Video Surveillance,” 2016 3rd International Conference on Information Science and Control Engineering (ICISCE), Beijing, 2016, pp. 502-505.

[11]https://www.machinecurve.com/index.php/2019/10/15/how-to-use-hinge-squared-hinge-loss-with-keras/#.

[12]https://docs.google.com/document/d/1-QqZ3JHd76JfimY4QKqOojcEaf5g3JSolNh-FHTxLag/edit.

[13] http://www.nick.co.uk/shows/legend-of-korra/videos/air-bending-basics/0sxstw,

[14] W. Sultani, C. Chen and M. Shah, ”Real-World Anomaly Detection in Surveillance Videos,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 6479-6488,2018.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama e T. Darrell, Caffe: Arquitetura Convolucional para Incorporação Rápida de Recursos , arXiv 2014.

[16] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera, “Chalearn looking at people rgb-d isolated and continuous datasets for gesturerecognition,” in IEEE CVPR Workshop, 2016.

[17] https://en.wikipedia.org/wiki/Multiple_instance_learning.

[17] https://en.wikipedia.org/wiki/Multiple_instance_learning.

[18] Lu Ming, ”Pesquisa e Melhoria de Algoritmos de Segmentação de Imagens”, 3ª Conferência Internacional sobre Teoria e Engenharia Avançadas de Computadores (ICACTE) , 2010.

[19] https://www.animeshd.biz/legendado/1367.

[20] https://github.com/facebookarchive/C3D.

[21] J. Amores, ”Multiple instance classification: Review taxonomy and comparative study”, Artif. Intell., vol. 201, pp. 81-105, Aug. 2013.

[22] T. G. Dietterich, R. H. Lathrop and T. Lozano-Pérez, ”Solving the multiple instance problem with axis-parallel rectangles”, Artif. Intell., vol. 89, no. 1, pp. 31-71, Jan. 1997.

[23] KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. ImageNet Classification with Deep Convolutional Neural Networks. Working Paper, 2012.

[24]C. Dhiman and D. K. Vishwakarma, ”A Robust Framework for Abnormal Human Ac-

tion Recognition Using $\mathcal{R}$ -Transform and Zernike Moments in Depth Videos," in IEEE Sensors Journal, vol. 19, no. 13, pp. 5195-5203, 1 July1, 2019.