UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PEDRO GABRIEL SANTOS DO COUTO SOARES

Um sistema para detecção de violência baseado em métodos de *Pose Track*

GARANHUNS

2019

Pedro Gabriel Santos do Couto Soares

Um sistema para detecção de violência baseado em métodos de *Pose Track*

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação da Unidade Acadêmica de Garanhuns da Universidade Federal Rural de Pernambuco.

Orientador:

Luis Filipe Alves Pereira

Garanhuns, Julho de 2019

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Ariano Suassuna, Garanhuns-PE, Brasil

S676s Soares, Pedro Gabriel Santos do Couto
Um sistema para detecção de violência baseado em métodos de Pose Track / Pedro Gabriel Santos do Couto Soares - 2019.
39 f. ; il.

Orientador: Luis Filipe Alves Pereira.
Trabalho de Conclusão de Curso (Graduação em Ciência da Computação)-Universidade Federal Rural de Pernambuco, Departamento de Ciência da Computação, Garanhuns, BR-PE, 2019.
Inclui referências.

1. Sistemas de reconhecimento de padrões 2. Visão por computador 3. Circuito fechado de televisão 4. Computação I. Pereira, Luis Filipe Alves, orient. II. Título

CDD 004

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação da Unidade Acadêmica de Garanhuns da Universidade Federal Rural de Pernambuco, aprovada pela comissão examinadora que abaixo assina.

Luis Filipe Alves Pereira - Orientador
UAG
UFRPE

Tiago Buarque Assunção de Carvalho - Examinador
UAG
UFRPE

Alixandre Thiago Ferreira Santana - Examinador
UAG
UFRPE

Garanhuns, Julho de 2019

Resumo

Devido a grandes índices de criminalidade no Brasil, mais especificamente em casas lotéricas, surge a necessidade de inovar com mecanismos de segurança que sejam eficazes em medidas contra ações criminosas. Este trabalho propõe uma nova arquitetura para detecção de violência em vídeos registrados por circuitos internos de TV de casas lotéricas. O método proposto tem como foco utilizar imagens de circuitos internos de TV para rastrear o posicionamento de partes do corpo de humanos durante a execução do vídeo, modelar seu comportamento e interpretar as ações realizadas. Além disso, apresentamos uma nova base de dados, com o foco inteiramente em cenas com violência registradas em casas lotéricas do Brasil que conta com aproximadamente 47.280 quadros com e sem violência. Por fim, é apresentada a performance do método baseado em *pose track* na base de dados construída, apresentando bons resultados para detecção de pessoas com mãos ao alto e deitadas.

Palavras-chave: Visão computacional. Reconhecimento de padrões. Detecção de violência

Sumário

Lista de Figuras **iii**

Lista de Tabelas **iv**

1 Introdução **1**
1.1 Objetivos 2
1.1.1 Objetivo principal 2
1.1.2 Objetivos específicos 2
1.2 Justificativa 2
1.3 Organização do trabalho 3

2 Fundamentação teórica **4**
2.1 Rastreamento em sinais de vídeo 4
2.1.1 Tracking 4
2.1.2 Pose Tracking 6
2.2 Reconhecimento de padrões 7

3 Trabalhos relacionados **9**
3.1 Base de dados 9
3.2 Compreensão de cenas 10
3.3 Considerações finais 13

4 Base de dados **14**

5 Método Proposto **17**

6 Experimentos e discussão **20**
6.1 Considerações finais 26

7 Conclusões **30**

Bibliografia **32**

Lista de Figuras

2.1 Exemplo de bounding box utilizada na detecção de humanos e objetos . . . 5
2.2 Demonstração de funcionamento do You Only Look Once(YOLO) 5
2.3 Exemplo de classificação 8
3.1 Representação de violência pelas bases de dados relacionadas 10
3.2 Fluxo comum de detecção de violência utilizando sistemas de vigilância . . 11
4.1 Trechos de vídeos com assalto em progresso 15
5.1 Diagrama representativo do método proposto 18
5.2 Classes extraídas da base 19
6.1 Sequencia de quadros com *tracking* 20
6.2 Sequencia de quadros com *pose tracking* 21
6.3 Gráfico em barras de média e variância, comparando as classes Parado e Mãos ao alto 23
6.4 Gráfico em barras de média e variância, comparando as classes Parado e Deitado 24
6.5 Sequencia de quadros com *pose tracking* 27
6.6 Sequencia de quadros com *pose tracking* 28
6.7 Matriz de confusão para a classe: Parado 29
6.8 Matriz de confusão para a classe: Deitado 29
6.9 Matriz de confusão para a classe: Mãos ao alto 29

Lista de Tabelas

4.1 Comparativo das bases de dados do estado da arte para detecção de violência. A base de dados proposta difere das outras. 16
6.1 Classes do vetor de médias e suas respectivas descrições 25
6.2 Classes do vetor de variância e suas respectivas descrições 25
6.3 Comparativo da taxa de acerto para cada classe, entre rede neural e AdaBoost 25

Capítulo 1

Introdução

O Brasil é um país conhecido pela extrema taxa de violência. A Organização Mundial de Saúde (OMS) em 2018, liberou um relatório [1] que traz alguns dados estatísticos sobre questões de saúde publica em vários países, dentre estes tópicos um que chama atenção é sobre a taxa de homicídio. O Brasil, portanto, é considerado o sétimo país mais violento das Américas, ficando atrás de países como Colômbia e Venezuela. Além de homicídios, o Brasil já foi considerado o terceiro colocado no ranking de países da América Latina com maiores taxas de roubos [2]. Tais índices são extremamente preocupantes para qualquer cidadão, independente do seu quadro social.

Em nosso país, casas lotéricas são um especial foco de violência, pois são locais onde existem uma alta movimentação de dinheiro em espécie. Estas, geralmente, servem de alvo para assaltantes, por guardar uma grande quantidade de dinheiro e por muitas vezes não possuir ao menos um segurança de prontidão. Hoje, no Brasil, casos de violência nesses estabelecimentos são cada vez mais comuns [6, 4, 3] e mesmo com a utilização de diversas medidas de segurança, parecem não cessar os ataques.

Uma das formas de proteger casas lotéricas é com a utilização de sistemas de vigilância, cujo uma equipe de segurança realiza o monitoramento em tempo real, verificando pessoas e atividades suspeitas para assim acionar as autoridades locais ou até a própria equipe. Contudo, sabemos que monitorar múltiplos monitores por horas, pode vir a ser extremamente desgastante, devido a necessidade de muita atenção. Ou por sua vez, não vale a pena atribuir uma pessoa para ficar o tempo todo monitorando, onde as câmeras se tornam apenas ferramenta de auxilio na apuração de crimes, posteriormente.

Hoje, graças aos avanços tecnológicos, tarefas como a automatização de sistemas de vigilância são uma realidade, detectar comportamentos anormais, violência e assaltos, começam a se tornar algo viável na prática. Entretanto, ainda é necessário resolver alguns problemas. Automatizar detecção de assaltos pode vir a ser um grande desafio, devido que assaltos não necessariamente acompanham atividades que demonstram panico, violência ou aparição de armas explicitas, se tornam um tanto quanto subjetivas e difíceis

de interpretar até para humanos.

Métodos de detecção de cenas, baseados em aprendizagem de maquina, precisam de uma base de dado especifica para poderem funcionar. Atualmente não existe uma base de dados especifica para detecção de assaltos. Além de que, as bases de dados atuais para detecção de violência não servem para este fim, devido que, ou utilizam de cenas de filmes ou são ambientes que não tem aspectos em comum a um assalto.

Este trabalho apresenta um estudo feito na área de visão computacional com objetivo de detectar assaltos em lotéricas. Para isso, nos propusemos a criar uma base de dados inteiramente com este foco, contendo vídeos de vigilância de lotéricas. Além disso, propomos um novo método para reconhecimento de atividades realizadas por humanos a partir de pontos *Pose Track*, onde a detecção de certas ações poderão ser utilizadas para identificar assaltos. Os resultados foram extraídos a partir da utilização do método na base de dados proposta.

1.1 Objetivos

1.1.1 Objetivo principal

Os principais objetivos deste estudo são detectar algumas posturas de humanos recorrentes em assalto, que através desta detecção seja possível interpretar que esteja ocorrendo alguma atividade suspeita. E construir uma base de dados inteiramente com foco em assaltos em casas lotéricas, possibilitando a utilização na validação do método proposto e de outros métodos que surjam com este foco.

1.1.2 Objetivos específicos

- Realizar um estudo sobre os métodos de Pose Estimation e Pose Tracking, com objetivo de detectar e rastrear partes do corpo de seres humanos em um vídeo;
- Criar uma base de dados contendo vídeos de vigilância em casas lotéricas. A base deve ser supervisionada com o registro dos vídeos que registraram assaltos ou situações comuns;
- Prototipação de métodos próprios para reconhecimento de atividades realizadas por pessoas no vídeo.

1.2 Justificativa

Devido as grandes taxas de assaltos e violência no Brasil, se faz necessário criar meios de identificar possíveis ameaças e combater-las de maneira rápida e eficaz. Sabemos que

nem todas nem todas as loterias possuem sistemas de vigilância monitorados integralmente ou até seguranças presentes, e mesmo assim as que possuem ainda continuam suscetíveis a ações de violência. Assaltos costumam acontecer de forma inesperada e acabam inibindo qualquer tipo de reação por parte dos funcionários, outras vezes podem acabar com a morte dos seguranças [5] . O intuito deste projeto é prover uma ferramenta inteligente de resposta ao crime, onde os responsáveis pela seguranças seriam alertados mediante uma situação potencialmente perigosa, possuiriam informações do número de assaltantes e sua localização no recinto. Dessa forma, possibilitando uma melhor tomada de decisão na proteção dos civis e ao patrimonio privado.

1.3 Organização do trabalho

Nos capítulos que se seguem, o leitor entenderá um pouco mais sobre a forma na qual este estudo foi conduzido. No Capítulo 2, falaremos um pouco mais sobre os métodos de visualização e rastreio de postura e como é possível extrair características para treinar um classificador. O Capítulo 3 apresentará os trabalhos relacionados tanto ao método proposto quanto a base de dados, extraídos a partir de um levantamento da literatura. No Capítulo 4, falaremos um pouco mais sobre a base de dados criada, suas diferenças entre as bases existentes e suas características mais detalhadas. No Capítulo 4, apresentaremos um novo método para a detecção de violência. O Capítulo 5 apresentará experimentos realizados utilizando o método de *Pose Track* e um método de *Tracking*, onde iremos avaliar os resultados de ambos utilizando a base de dados proposta. Assim, apresentaremos o resultado das classificações das ações humanas. Por fim, no Capítulo 6, serão apresentadas as considerações finais.

Capítulo 2

Fundamentação teórica

Nesta seção são apresentados os métodos de *tracking*, *pose track* e reconhecimento de padrões. Explicaremos como cada um deles fizeram parte do projeto e a justificativa para o uso dos mesmo.

2.1 Rastreamento em sinais de vídeo

Rastear pessoas é essencial na detecção de violência, pois, podemos distinguir a movimentação de pessoas em vídeo e assim prover uma interpretação sobre as ações desta no intervalo de tempo do rastreio. Então é necessário entender um pouco mais sobre os métodos existentes nesta área.

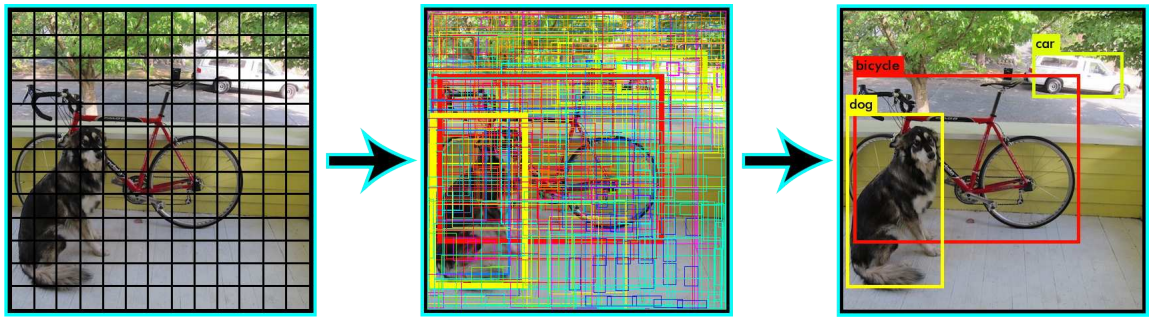
2.1.1 Tracking

Dado o estado inicial de um dado objeto na primeira imagem, o objetivo do rastreio é estimar o estado destas imagem nos quadros subsequentes [19]. Implicando dizer que as informações rastreadas no intervalo servem para categorizar tipos distintos de ações. É necessário detectar objetos entre os quadros para em seguida conectá-los afim de formar o rastreio. A detecção de um objeto é o ato de identificar a região da imagem que compõem um determinado objeto de uma cena e através disso apresentar os resultados visualmente com a utilização de *bounding boxes*, como apresentado na Figura 2.1. Estas, são formadas por quadriláteros que demarcam o espaço e localidade do objeto na imagem.

Dentre as técnicas mais utilizadas para detecção de objetos *You Only Look Once* (YOLO) é uma das que mais se destaca, por ser leve e preciso. Apresentado por Redmon em [16], YOLO utiliza algoritmos do estado da arte de detecção em tempo real de objetos. Este, consegue detectar qualquer tipo de objeto para qual tenha sido previamente treinado, e consegue realizar tal processamento de forma extremamente rápida e efetiva, viabilizando a sua utilização em vídeos em tempo real, ao invés de gravações. O YOLO, divide a imagem em uma grade de 13x13 células, demonstrado na Figura 2.2, e utiliza

Figura 2.1: Exemplo de bounding box utilizada na detecção de humanos e objetos

Figura 2.2: Demonstração de funcionamento do método You Only Look Once (YOLO)



Disponível em: `https://pjreddie.com/darknet/yolov2/`

apenas uma rede neural em todo este conjunto. Em cada célula é predito um total de cinco *bounding boxes*, havendo ou não um objeto de fato, uma pontuação para o nível de confiabilidade e sua classe. As pontuações e classes vizinhas são combinadas de diversas formas e apenas as que tiverem uma probabilidade maior que o limiar definido são utilizadas. Então, a partir do resultado da detecção entregue pelo YOLO, é necessário utilizar algum algoritmo a fim de rastrear todas as detecções percorrendo todos os quadros do vídeo.

Existem diversas técnicas para rastreio de objetos entre cenas, porém, por ser um processo usualmente custoso, se torna inviável a utilização de métodos mais robustos no processamento de vídeos em tempo real. O *Simple Online and Realtime Tracking* (SORT), descrito em [7], é um algorítimo de rastreio, que utiliza um vetor de *bounding boxes* como entrada e consegue rastrear estes múltiplos objetos em um vídeo em tempo real, alcançando bons resultados com baixo processamento, devido a utilização de métodos mais simples, porém ainda eficientes. Basicamente, o SORT utiliza todos os pontos detectados, constrói uma predição do posicionamento de cada ponto para os próximos quadros e assimila aquele ponto que tem maior proximidade a sua predição. É necessário utilizar uma matriz associativa de custos para manter o registro dos pontos, para que a partir disso seja possível obter o índice de menor custo.

O YOLO funcionando em conjunto com o SORT provem uma estrutura solida. Entretanto, devido a problemas como oclusão, o rastreio torna-se bastante confuso. Em ambientes com pessoas trafegando entre si e regiões com tumultos se torna uma tarefa extremamente complexa, rastrear uma pessoa através do vídeo.

2.1.2 Pose Tracking

Pose Tracking consiste em rastrear a pose de humanos quadro-a-quadro com intuito de construir uma visualização mais detalhada sobre a movimentação e posicionamento de vários membros do corpo humano durante um espaço de tempo. Para isso, primeiramente é necessário detectar humanos e suas partes do corpo, e.g. joelhos, cotovelos, calcanhares, para assim, para assim conseguir rastrear-las. Diferentemente do *tracking* utilizado no SORT, que apenas rastreia objetos, *pose tracking* rastreia cada parte do corpo de uma pessoa individualmente, onde é possível interpretar ações, postura e gestos. Para rastrear partes do corpo humano, é necessário anteriormente realizar a detecção de cada um destes membros, quadro a quadro, de forma que seja possível visualizar toda estrutura humana a cada etapa, o rastreio então é realizando para cada parte individualmente.

Pose Estimation é a tarefa de realizar detecção para localizar humanos em cenas e assim extrair pontos de interesse no espaço, que vêm a ser traduzidos como posicionamento de membros do corpo humano visíveis. Quanto aos não visíveis, são realizados predições realista de acordo com o que já foi detectado e então, todas as informações, explícitas ou

não, são armazenadas, como representação do corpo humano. Em [10], os autores propõe um algorítimo de *Pose Estimation* para múltiplas pessoas em uma cena, onde cada uma é detectada e interpretada como uma *bounding box* e o seu conteúdo é processado de forma a estimar membros do corpo.

Pose Flow é um método proposto em [20], com o intuito de rastrear *Pose Estimation* de várias pessoas em tempo real. Baseando-se em modelos matemáticos que utilizam níveis de confiança para avaliar todas as detecções de um quadro, este algorítimo possibilita a escolha da melhor amostra relativa a uma pessoa e através disso realiza uma associação da amostra passada com a seguinte.

2.2 Reconhecimento de padrões

O ato de reconhecer pessoas e objetos é uma habilidade inerente do humano, observar e interpretar ações, sons e contextos. Na computação, é possível implementar algoritmos que visam simular a capacidade humana em compreender e identificar padrões em imagens, textos, e ate sons. Em [8], os autores demonstram que o trabalho de reconhecimento de padrões acontece em duas etapas, sendo a primeira extrair características e a segunda em classificação. Extrair características é a etapa no processo de reconhecimento de padrões que tem o intuito de reduzir a dimensionalidade, preservando a semântica de interesse do problema de uma base de dados. Por exemplo, uma corretora de imoveis mantém registrado fotos de todos os cômodos e aspectos das casas as quais ela é responsável, para assim poder apresentar a melhor casa que se adéque a possíveis compradores. Para realizar uma classificação e decidir qual será a casa escolhida para um determinado cliente, todas as imagens das casas poderiam ser reduzidos em um vetor contendo, tamanho da casa, número de quartos, preço por metro quadrado, entre outras coisas. A dimensionalidade foi reduzida, porém cada dimensão ainda representa um valor semanticamente relevante para o problema.

Muitas vezes os dados presentes em bases são imensamente redundantes ou irrelevantes, o que torna o processo de classificação mais trabalhoso e portanto mais lento. Então, se faz necessário traçar abordagens mais robustas para tratar estas bases de dados. Algumas formas de reduzir a dimensionalidade consiste em escolher características mais representativas ou até combinar conjuntos menores em outros mais concisos. Não existe de fato uma forma para saber quais são as melhores características que resolvam um problema, no entretanto, deve ser levado em consideração o método e o contexto da questão em si, para saber qual a melhor abordagem.

Já a classificação, por exemplo, ocorre quando possuímos valores que já possuem com certos rótulos/classes, e diante do surgimento de novas instâncias é analisado suas características. Caso possuam atributos semelhantes, estes são associados a mesma classe/rotulo

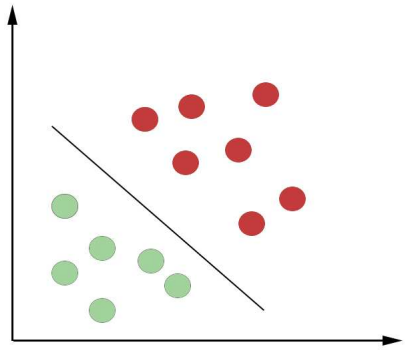


Figura 2.3: Exemplo de classificação

do outro valor. Para realizar processos de classificação é necessário utilizar um classificador, este é responsável por realizar cálculos de similaridade entre os vetores que compõem o espaço de características.

A Figura 2.3 demonstra o objetivo da classificação, que é encontrar uma função que descreve uma curva capaz de separar corretamente os padrões. Para este caso, seria basicamente encontrar uma função que distinguisse a classe dos círculos verdes e vermelhos. A função, neste contexto, seria a reta que corta os dois grupos.

Dizemos que um método de classificação em que é calculado a similidade entre um elemento desconhecido com outros já conhecidos é considerado como uma classificação supervisionada. Neste processo, a base é divida entre objetos de treino e teste, onde todos que compõem o conjunto de treino já possuem uma classe. Então, o classificador, para cada elemento de teste, realiza uma comparação com os valores de treino e assimila a classe de maior semelhança. Em casos em que não haja classes definidas para nenhum das amostras, são aplicados métodos não supervisionados, em que, são realizados agrupamentos entre as características semelhantes entre si, e ao final cada grupo resultante resulta em uma classe.

Capítulo 3

Trabalhos relacionados

Apresentaremos, neste capítulo, trabalhos relacionados que focam (i) na criação de uma nova base de dados para detecção de violência, e (ii) na proposição de métodos para reconhecimento de ações violentas em cenas de vídeo.

3.1 Base de dados

Para o desenvolvimento de métodos para reconhecimento de violência em vídeos, se faz necessário o acesso a bases de dados contendo imagens alvo com cenas de violência e não violência. Os trabalhos da literatura listados a seguir propõem novas bases de dados para detecção de violência em diferentes contextos.

Marszalek *et al.* [14] apresenta uma grande base de dados contendo uma variedade de ações humanas totalmente extraída de filmes, onde apenas 70 quadros são de violência. Representado pela Figura 3.1b. Blunsden *et al.* [9] propôs uma base de dados sobre comportamento entre interação de múltiplas pessoas em ambientes externos. Dentre dez tipos de comportamentos, existem três que indicam potencial assalto: Perseguição e luta, Figura 3.1d. Nievas *et al.* [15] apresentou uma base de dados composta de clipes capturados durante os jogos da liga nacional de hockey (NHL). Jogo que é bastante agressivo por natureza e que facilmente é possível ver uma briga. Cada clipe foi classificado manualmente como: luta ou não luta. Nievas *et al.* também, no mesmo trabalho, apresentou uma pequena base de dados composta de cenas de filmes em que apenas existem brigas. Figura 3.1c. Soomoro *et al.* [18] criou a base de dados *UCF101* que é composta de ações de clipes coletados do youtube. Existem cento e uma categorias de ações neste grupo, que incluem: flexões, corte de cabelo, tocando piano, entre outros. Nas classes, esmurrando e sumô, da base UCF101, existem cenas de luta adquirida de vídeos de lutas de box (imagem 3.1a) e partidas de sumô.

(a) luta de boxe (b) cena de luta em filme

(c) briga em partida de hockey (d) briga em rua

Figura 3.1: Representação de violência pelas bases de dados relacionadas

3.2 Compreensão de cenas

A arquitetura usual dos métodos computacionais para reconhecimento de cenas é apresentada na Figura 3.2, idealizada a priori por Amira Ben Mabrouk em [12]. Nela, percebe-se os módulos de extração de características de baixo nível (i), descrição das características (ii) e classificação (iii). Desenhado de forma crescente ao nível de semântica obtidos através dos módulos, onde ao final seja possível identificar se está ocorrendo ou não indícios de violência.

Primeiro, no modulo de extração de características de baixo nível (i), são extraídas características de baixo nível, podendo ser locais ou globais. Características locais são extraídas de partições do contexto global de uma cena, e que foram definidas através de pontos de interesse em um espaço bidimensional ou tridimensional, como uma imagem ou vídeo, respectivamente. As características globais são representações completas das cenas apresentadas, onde as mudanças nesse contexto são avaliadas e monitoradas nos espaço-tempo para identificar características. Em seguida, é necessário representar as características extraídas de alguma forma, seja através do seu formato, textura ou até da movimentação. Dependendo do formato de descrever características, podemos tratar-las de formas distintas.

No modulo de descrição das características (ii), podemos descrever estas de algumas formas distintas. Utilizar o sentido da movimentação como forma de detectar comporta-

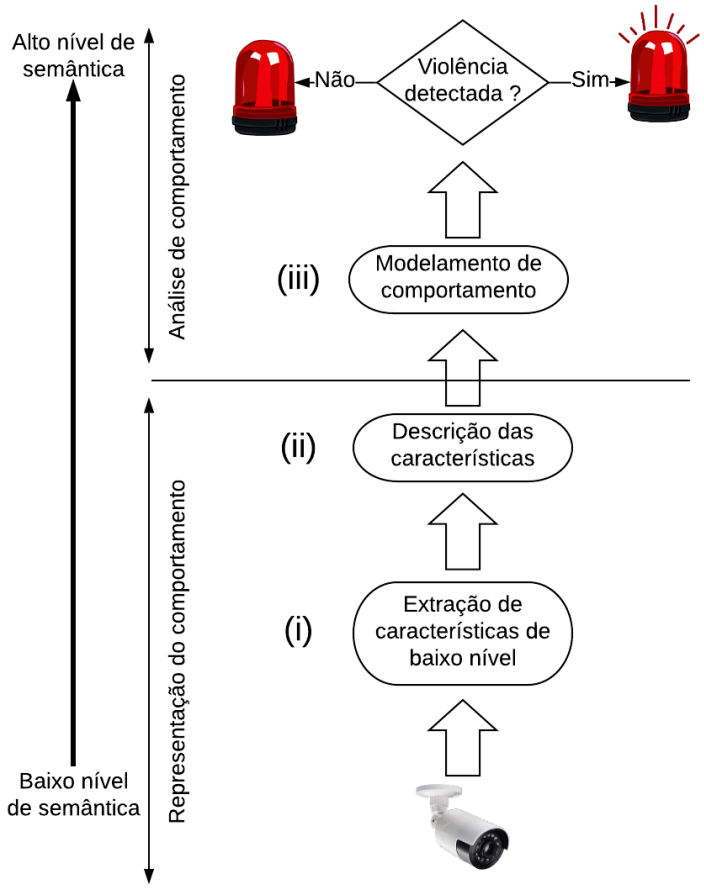


Figura 3.2: Fluxo comum de detecção de violência utilizando sistemas de vigilância

mentos lineares e mais comuns, ou desordenados e mais foras do comum. Interpretar que cada quadro de um vídeo é uma textura, e que a modificação desta ao decorrer do tempo traga informações relevantes, como por exemplo, em um primeiro quadro existe uma fila de pessoas e no quadro dez esta fila foi desfeita, ou seja, houve uma modificação na composição da imagem (textura) de forma abrupta. Ou até, compreender como se fossem formatos, possibilitando o calculo de aspectos mais geométricos, como, movimentação em cena, altura, largura e afins.

No modulo de classificação (iii), a modelagem de comportamento nos permite entender as ações realizadas por um objeto e determinar se seu comportamento é violento ou não. Para isto, precisamos classificar o comportamento das pessoas entre um conjunto de classes pré-definidas para que posteriormente seja possível detectar ações de forma automática. É necessário então, utilizar algorítimos de reconhecimento de padrão.

Em [15] ou autor, por exemplo, utiliza *Space-Time Interest Points* (STIP) para extrair características locais (i) a partir de pontos de interesse, tanto no espaço bidimensional quanto no tridimensional. Estes pontos são caracterizados pela alta variação de intensidade no espaço e baixa variação de intensidade no tempo. Através da utilização de *Histograms of Optical Flow*(HOF) combinado com *Histograms of Oriented Gradients*(HOG), resulta um vetor de características que vem a ser descrito como movimentação (ii). Estas extrações são utilizas em um classificador (iii) *Support Vector Machine* (SVM) treinado em identificar brigas.

Em [11], o autor utiliza de extração de características globais (i), onde o todo de uma cena é extraído sem distinção. Então, interpreta-se essa característica como uma textura (ii), onde ao decorrer do tempo esta sofre alterações, que dependendo do nível de alteração pode modelar comportamentos distintos. Este método é bastante utilizado em detecção de comportamento em multidões, onde é mais fácil interpretar grandes movimentações ao analisar a perturbação na textura em relação a quadros anteriores. E utiliza de uma SVM para classificar violência dentre este contexto (iii).

Em [17], o autor propõe uma extração de características locais no espaço-tempo (i) e através da utilização de HOF, estas são interpretadas como pontos de movimentação de interesse (ii). Estes pontos são então analisados e classificados conforte seu comportamento (iii), caso sejam movimentos lineares não é detectado nenhum tipo de violência, caso sejam movimentações desordenadas, já é possível assumir algo. Então, é utilizado uma SVM para treinar o método de detecção a partir das movimentações.

Em [13], o autor propõe a utilização da extração de características locais no espaço e tempo (i), entretanto é utilizado para detecção em locais com grandes quantidades de pessoas. Cada características é representada por um pedaço de corte em uma grade pré-determinada, estas então, são descritas como texturas (ii).Então é realizado uma classificação para padrões de textura extraídos (iii). Por fim, os resultados são analisados para identificar anomalias no espaço e caso existam áreas que se distinguam completamente do

seu arredor, são identificadas como anomalias.

3.3 Considerações finais

Nenhuma das bases de dados disponíveis na literatura seria suficiente para treinar um sistema computacional capaz de reconhecer e detectar assaltos reais em casas lotéricas. Mesmo com a utilização destas outras bases, nenhuma se atêm ao contexto real em locais fechados.

Capítulo 4

Base de dados

Devido a ausência de material na literatura, visto no capítulo anterior, que aborda atos violentos reais em locais fechados, foi necessário construir uma nova base de dados inteiramente com este foco, onde é possível visualizar diversos padrões e ações, que usualmente ocorrem durante assaltos, e através disso é possível aplicar métodos para detectar tais comportamentos.

A base de dados proposta contém diversas amostras, sendo divididas entre 145 vídeos com violência e 52 sem violência. Cada um destes vídeos possui taxa de atualização de vinte e quatro quadros por segundo e aproximadamente 10 segundos de duração, nunca mais do que isso. Todos estes foram coletados a partir do *Youtube* e são vídeos de vigilância em lotéricas, editados e escolhidos a mão para evitar qualquer tipo de exposição a pessoas, morte ou até elementos que pudessem infringir direitos autorais. A menor quantidade de vídeos sem violência é facilmente explicada pela publicação apenas das cenas com violência, que servem hoje para noticiar e identificar criminosos. Teoricamente, não existiriam motivos para a existência de vídeos contendo um cotidiano pacato em lotéricas. As amostras sem violência na base foram extraídas a partir do inicio de vídeos com violência, onde o assalto ainda não havia acontecido. Os vídeos contam com pessoas paradas, andando, correndo, levantando as mãos, deitadas, se dispersando de filas e assaltando. Algumas amostras contem vídeos em que se classificados como violência, é possível um humano, identificar de maneira inequívoca que está havendo um assalto, outras são menos perceptíveis, onde todas as pessoas estão paradas. E ainda existem as que são mais fáceis de visualizar, onde existem movimentos mais bruscos. Na Figura 4.1 podemos ver algumas sequências onde existe a presença de assalto, esta é apenas uma pequena parte de base de dados.

A base de dados proposta tem como foco disponibilizar uma série de novos segmentos para avaliar algorítimos de detecção de violência. Esta, possui como seu maior diferencial, cenas de violência em ambientes internos, todas extraídas do mundo real sem nenhum tipo de acontecimentos simulados. Em [9], os autores propõem algo similar ao proposto,

Figura 4.1: Trechos de vídeos com assalto em progresso

entretanto, todos seus acontecimentos ocorrem em ambiente externo e contam com diversos tipos de classes, onde apenas 1751 são quadros relativos a brigas. Enquanto isso, a base proposta por [14] se destaca pela quantidade de 600 mil quadros relativos a ações, porem apenas 70 amostras de um total de 884 descrevem violência. Além disso, todas as extrações foram realizadas em filmes de Hollywood e representam ações ensaiadas. Já os autores em [15] constroem uma base de dados com foco inteiramente em atos de violência em um ambiente real, porém utiliza como contexto violência ocorrida durante jogos de hockey, que vem a ser bem específico. [18] também utiliza competições esportivas como algumas amostras de sua base, como box e sumô, porem não chega a ser o foco da base. Algumas informações sobre as bases podem ser vistas através da Tabela 4.1. A base proposta tem como principal diferencial das outras, apresentar cenas de violência em ambientes distintos e reais, contendo aproximadamente 34.800 quadros com violência e 12.480 sem violência, totalizando 47.280 para a base total, superando o número de amostras das bases em questão.

Tabela 4.1: Comparativo das bases de dados do estado da arte para detecção de violência. A base de dados proposta difere das outras.

referencia	Nº quadros	Nº classes	contexto	ambiente
Marszalek *et al.* [14]	600.000	12	filme	externo
Blunsden *et al.* [9]	2.059	3	mundo real	externo
Nievas *et al.* [15]	5.000	2	esporte	interno
Soomoro *et al.* [18]	2.400.000	101	esporte	interno
Base proposta	47.280	2	mundo real	interno

Capítulo 5

Método Proposto

Este método propõe a utilização de *Pose Flow* associado com *Pose Tracking* para detectar e rastrear humanos e a posição dos membros de seus corpos durante o processamento do vídeo em tempo real (Figura 5.1). Assim, extraindo características sobre as ações realizadas por estas pessoas, sendo possível modelar um contexto sobre cena e identificar ações suspeitas, não necessariamente contendo violência. Na prática, utilizando estes algorítimos na base proposta, conseguiremos extrair pontos sobre a postura de várias pessoas em cena e rastreá-las pelo vídeo.

Esta extração será armazenada em um arquivo contendo todos os pontos do seus membros durante a execução do vídeo. Posteriormente, serão classificadas manualmente por pessoas, onde, para cada humano detectado no vídeo, existem pontos distinto.

As classes de ações foram definidas em sete, as quais podem ser vistas na Figura 5.2 como: parado, andando, correndo, deitado, dispersão (se afastando de uma fila), mãos ao alto e assaltando. Todos estes pontos são armazenados não levando em conta as coordenadas do vídeo a que foram extraídos. As coordenadas (x,y) relativas ao ponto de cada membro foram registradas sempre em relação ao ombro esquerdo do indivíduo para que seja possível extrair ações das pessoas em um contexto geral, de forma a fazer com que cada extração seja independente do seu vídeo de origem. Isto auxilia a classificação nas etapas seguintes, onde são ignoradas informações sobre localização global e podemos trabalhar com cada amostra de forma independente. Neste viés, é importante ressaltar que o vetor de características extraído a partir do *Pose Track* dispõe de dezessete dimensões, onde, cada uma representa partes do corpo, como, olhos, ombros, braços, pernas, entre outras. Todas estas dimensões são extraídas a cada quadro, gerando uma série de informações cruciais para o entendimento de uma cena, porém criando um enorme vetor de pontos.

O esperado é que, através destas extrações, seja possível visualizar uma certa diversidade entre um vetor de característica que representa ação e outro que não. Em um contexto real, seria possível distinguir as características de pessoas com mãos ao alto e

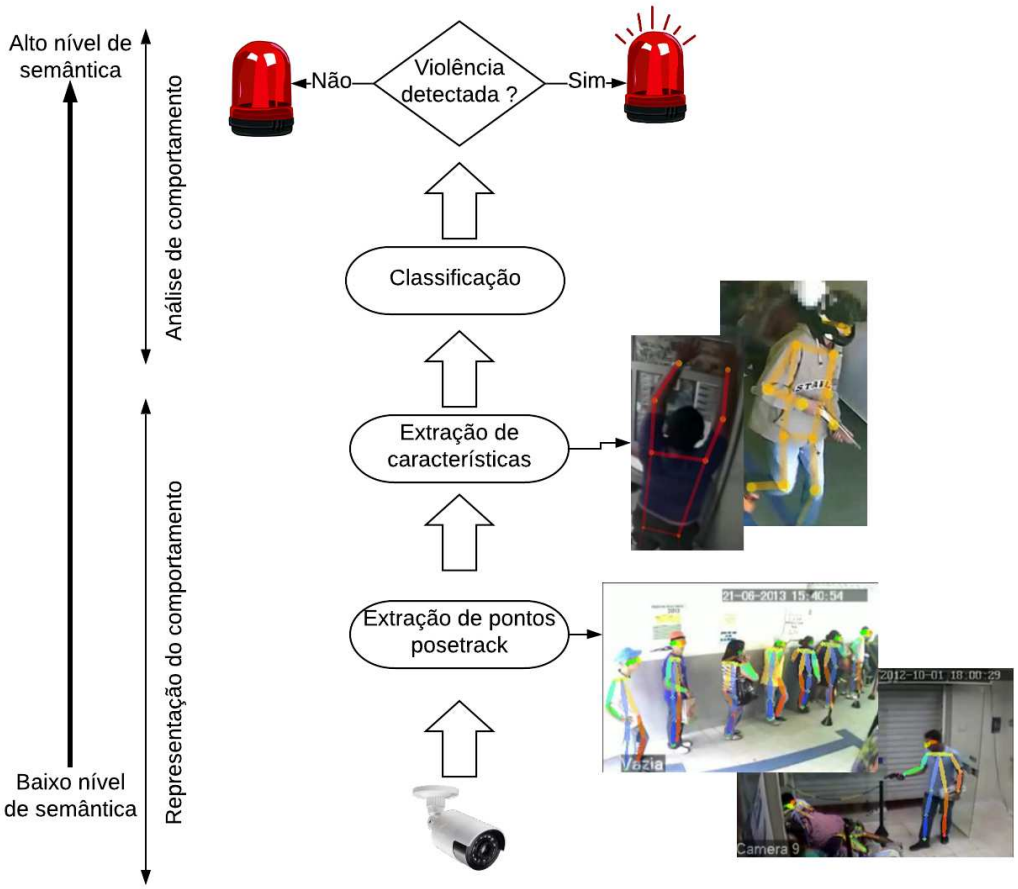


Figura 5.1: Diagrama representativo do método proposto

pessoas paradas, apenas olhando o posicionamento de seus pontos no espaço, como por exemplo, se o ponto da mão estiver maior no eixo Y que o ponto dos ombros e cotovelo, é bem provável que o indivíduo esteja com as mãos para o alto. Levando em conta que, em um intervalo de tempo, as ações realizadas por um indivíduo possam variar ou até mesmo que existam ruídos na detecção (pessoas se espreguiçando, acenando, fazendo movimentos bruscos, entre outros) podemos interpretar que a média e variância do conjunto de posicionamento para cada característica possam vir a ser úteis para generalizar o comportamento. Onde, por exemplo, em um caso de mãos ao alto, a média das mãos e cotovelos no intervalo de tempo do vídeo, seriam vistos, claramente, como pontos localizados específico da cabeça e a variância nos mostraria mais sobre a movimentação de cada parte do corpo, sendo possível distinguir se houveram variações no posicionamento. Mediante destas informações, é essencial extrair o vetor de média e variância, para cada uma das dimensões de um vídeo, com intuito de simplificar os dados de forma que continuem legíveis e tenham uma representação mais precisa.

Por fim, através da utilização deste vetor, podemos restringir a classificação para processar apenas 34 pontos por vídeo, requirindo bem menos custo computacional do que

(a) Parado (b) Andando (c) Correndo (d) Deitado

(e) Mãos ao alto (f) Dispersão (g) Assaltando

Figura 5.2: Classes extraídas da base

processar cada quadro deste. A classificação pode ser realizada utilizando redes neurais, *random forest*, ou *AdaBoost*, onde anteriormente os dados tenham sidos normalizados. O intuito é que ao final deste processo o classificador tenha uma boa acurácia e que seja possível identificar cada classe de forma independente.

Capítulo 6

Experimentos e discussão

A priori, foram realizados testes com o método de *tracking*, porém, como foi dito ao final do tópico 2.1.1, a oclusão prejudica muito o rastreio, tornando difícil modelar o comportamento de pessoas em cena, e ainda mais que as *bounding boxes* não conseguem transmitir a ideia da ação de forma eficaz, sendo necessário interpretar as dimensões desta para interpretar movimentações.Podemos ver claramente na Figura 6.1, quadros com poucos segundos de diferença. É possível ver que os identificadores de cada um se modificam drasticamente, sendo difícil identificar a mesma pessoa ao decorrer do vídeo.

Figura 6.1: Sequencia de quadros com *tracking*. Existe uma dificuldade para monitorar uma mesma pessoa durante o vídeo, além que as *bounding boxes* transmitem uma pequena representatividade das ações

O método de *pose flow* encontra a mesma barreira na sua detecção, contornando a dificuldade, devido a estimação de posicionamento de partes do corpo, se torna mais fácil rastrear a movimentação das pessoas, com a vantagem de, mesmo com trechos de poucos segundos de vídeo, é possível extrair uma representação clara sobre a ação realizada pela pessoa naquele momento. A Figura 6.2 realiza uma comparativo com a cena da Figura 6.1, sendo *pose estimation* a forma mais representativa neste caso.

Vemos na Figura 6.3 um gráfico contendo um vetor de médias e variâncias de partes

Figura 6.2: Sequencia de quadros com *pose tracking*. Os rastreios são mais representativos, é mais fácil visualizar as ações correspondentes

do corpo, extraídos a partir de dois vídeos da base de dados proposta. O primeiro vídeo representa a classe **parado** e o segundo a classe **mãos ao alto**. Com ajuda das Tabelas 6.1 e 6.2, conseguimos dar sentidos para os índices presentes na imagem. Cada um representa uma parte do corpo para ambos eixos X e Y, sendo 34 classes, médias para o vetor dessa característica no espaço de tempo do vídeo. Para este caso específico, é possível visualizar que, os pontos de cotovelo e punho para o eixo Y estão acima do ponto zero, que como dito anteriormente é a localização do ombro esquerdo, e que o punho está mais acima do ombro, logo, por intuição, se no período de tempo daquele vídeo, a media do posicionamento destas partes permaneceram assim, é certo afirmar que o ser humano rastreado está com as mãos para cima. Da mesma forma, para o caso na Figura 6.4, onde é comparado uma pessoa parada com outra deitada. O único diferencial se dá por conta da extração ocorrer no mesmo local de vídeo para as duas classes, onde, antes de ocorrer o assalto havia uma pessoa parada e durante o ato criminoso uma pessoa sentada no chão. Podemos visualizar através do gráfico, que a classe joelho, para o eixo Y, encontra-se no mesmo nível ou mais alto que o eixo Y do quadril, além que todas as classes de média estão bastante próximas ao ponto 0, em comparação com outra pessoa que estava em pé no mesmo vídeo, através dessas informações podemos chegar a um consenso que isto

representa características marcantes em humanos que estão deitados ou até sentados no chão.

Podemos utilizar do mesmo pressuposto para julgar a base de dados como um todo. Através da utilização de diagrama de caixa na Figura 6.5, visualizamos com mais cautela a distribuição dos pontos de cada dimensão para toda a classe normalizada, onde para cada diagrama, **1** representa a classe informada e **0** o oposto. Isto é, no diagrama 6.5a a classe 1 representa parado e 0 "não parado", no gráfico 6.5b a classe 1 representa mãos ao alto e a 0 "não mãos ao alto", esta mesma lógica é utilizada para todos os diagramas a seguir.

Então, podemos realizar a comparação das classes **parado** e **mãos ao alto** entre diagramas distintos, onde por exemplo, a distribuição geral para o cotovelo em **mãos ao alto** (Figura 6.5b) tende levemente a ser maior que o mesmo membro na classe **parado** (Figura 6.5a), e da mesma forma, a mão (Figura 6.5d) é levemente superior (Figura 6.5c), devido ao fato das mãos e ombros estarem dispostos de forma superior ao eixo Y no momento em que estiverem para o alto. A mesma lógica serve para a classe deitado, onde vemos que a distribuição para a cintura (Figura 6.6a) é superior a do joelho (Figura 6.6b) na classe **parado**, como qualquer pessoa em pé. Mas, para uma pessoa deitada, a cintura(Figura 6.6c) tem uma distribuição inferior ao joelho (Figura 6.6d)

Estas informações são uteis para entender que de fato existe uma lógica por trás do vetor de medias e variâncias, e que será uma das formas que o classificador possivelmente interpretará. Todavia, através dos diagramas de caixa, visualizamos que existe uma proximidade entre as classes, isto é refletido pela dificuldade de classificação da base de dados e pelas limitações do método de *posetrack* em ambientes que ocorrem excessivas oclusões.

Quanto à classificação, foram realizados dois procedimentos. O primeiro consiste em utilizar o vetor de médias e variância como parâmetro de uma rede neural e o segundo na utilização desta mesma entrada, só que, utilizando um algoritmo de floresta aleatória onde seus pesos são processados pelo *Adaptive Boosting* (AdaBoost) para melhorar a sua performance. Ambos, são métodos de classificação que realizam a separação de padrões conforme o exemplo mostrado na Figura 2.3. Na Figura 6.7 podemos ver o comparativo entre a aplicação de uma rede neural e *AdaBoost* para a classe **Parado**, onde o numeral "1"representa esta classe e "0"o que não representa, exemplo: parado e não parado. Vemos que a rede neural apresentou melhores resultado, onde obteve 376 acertos para **Parado** e 75 erros, de um total de 451 amostras. No exemplo seguinte, para a classe de deitado (classe "1"), pode ser visto na Figura 6.8, que o *AdaBoost* se saiu um pouco melhor, acertando 66 de um total de 88 elementos, porém a classe inversa teve resultados um pouco pior que a de redes neurais, acertando 56 elementos, contra 61. Por fim, Para a classe **Mãos ao alto**, Figura 6.9, o *AdaBoost* possui os melhores resultados, com 37 acertos de 42 amostras. A Figura 6.3 mostra a taxa de acerto para cada classe utilizando os dois classificadores.

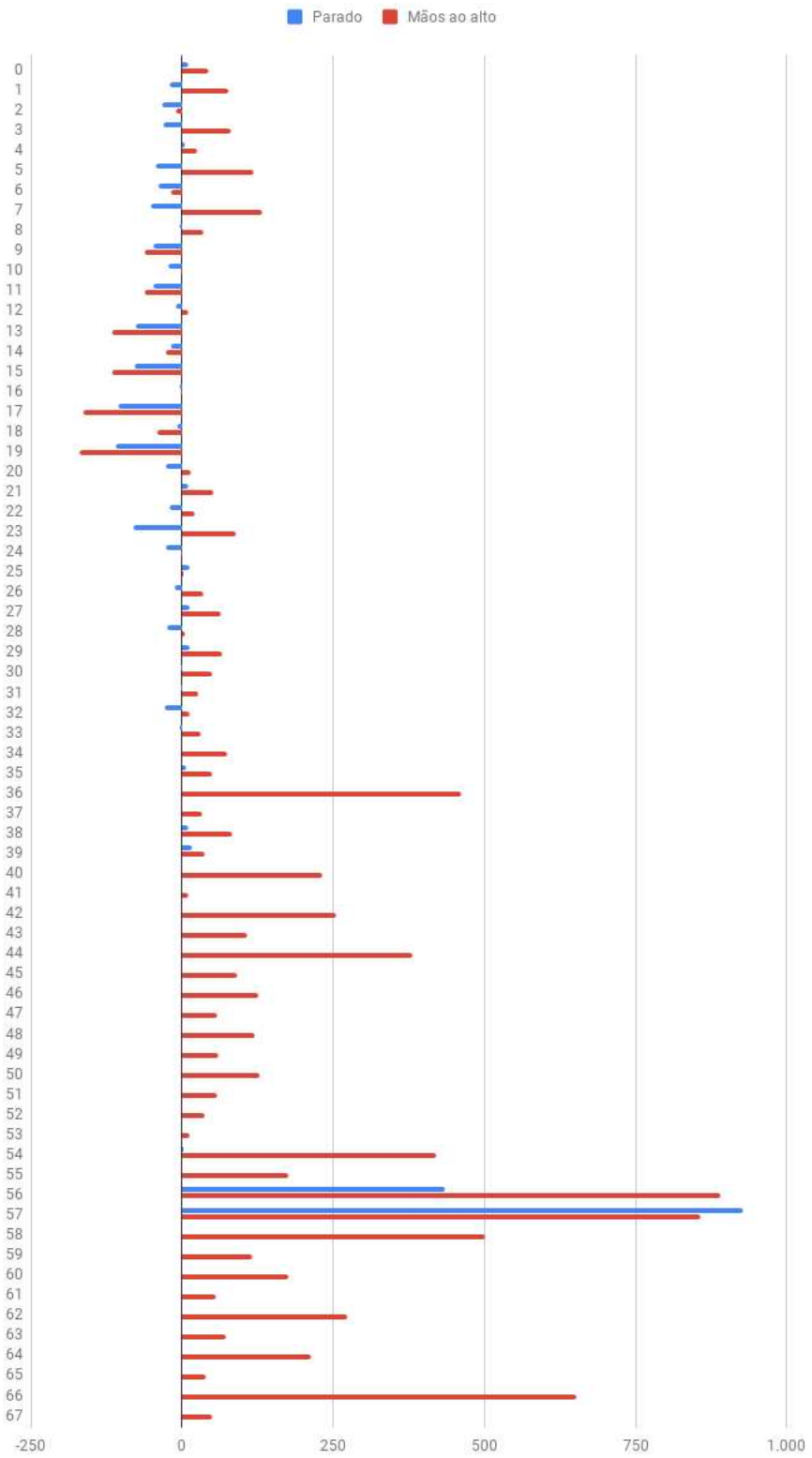


Figura 6.3: Gráfico em barras listando média e variância do vetor de características, comparando as classes Parado (azul) e Mãos ao alto (vermelho) em uma mesma cena

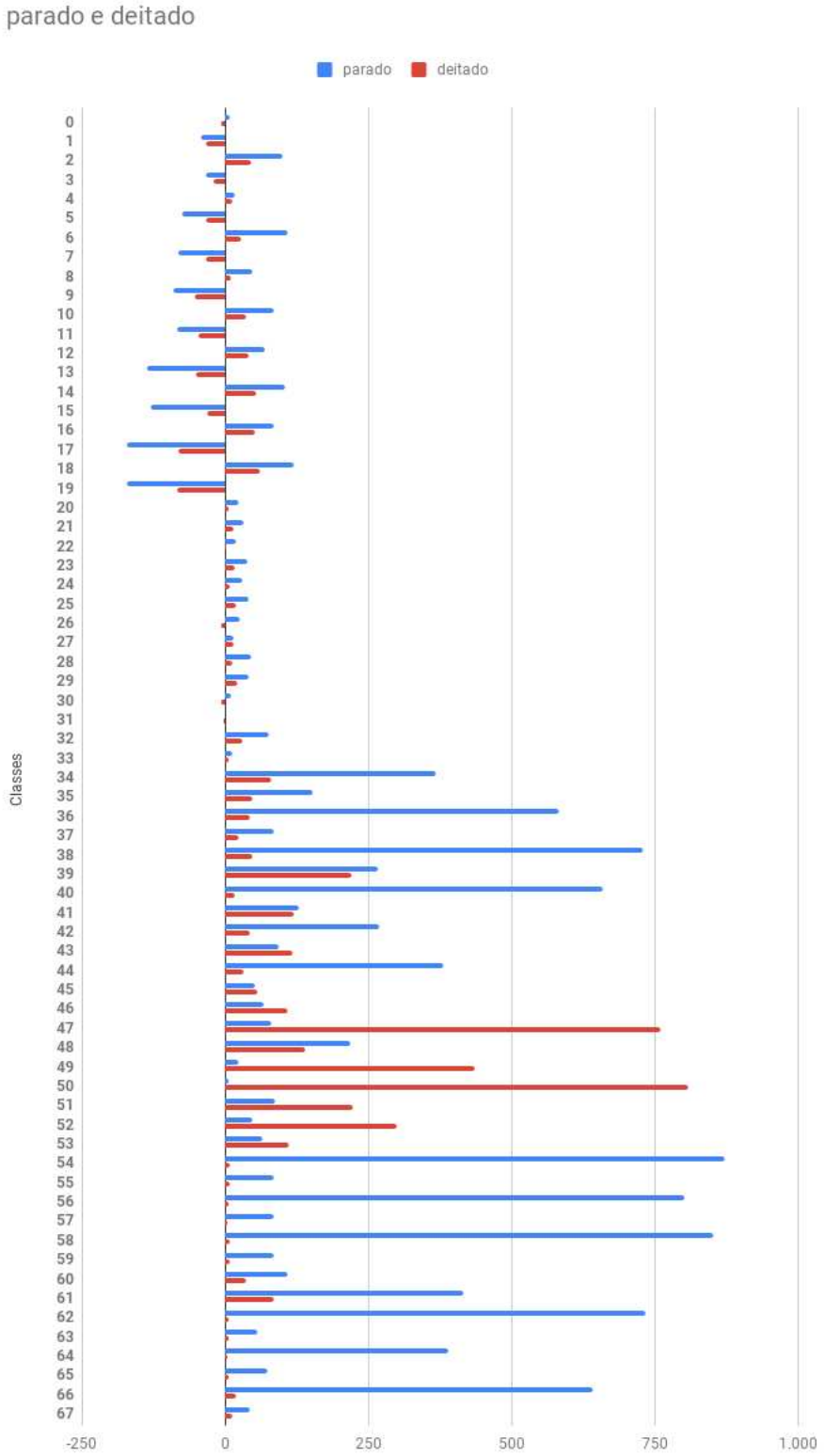


Figura 6.4: Gráfico em barras listando média e variância do vetor de características, comparando as classes Parado (azul) e Deitado (vermelho) em cenas distintas

Classe	Descrição
0	cotovelo esquerdo X
1	cotovelo esquerdo Y
2	cotovelo direito X
3	cotovelo direito Y
4	pulso esquerdo X
5	pulso esquerdo Y
6	pulso direito X
7	pulso direito Y
8	quadril esquerdo X
9	quadril esquerdo Y
10	quadril direito X
11	quadril direito Y
12	joelho esquerdo X
13	joelho esquerdo Y
14	joelho direito X
15	joelho direito Y
16	tornozelo esquerdo X
17	tornozelo esquerdo Y
18	tornozelo direito X
19	tornozelo direito Y
20	nariz X
21	nariz Y
22	olho esquerdo X
23	olho esquerdo Y
24	olho direito X
25	olho direito Y
26	orelha esquerda X
27	orelha esquerda Y
28	orelha direita X
29	orelha direita Y
30	ombro esquerdo X
31	ombro esquerdo Y
32	ombro direito X
33	ombro direito Y

Tabela 6.1: Classes do vetor de médias e suas respectivas descrições

Classe	Descrição
34	cotovelo esquerdo X
35	cotovelo esquerdo Y
36	cotovelo direito X
37	cotovelo direito Y
38	pulso esquerdo X
39	pulso esquerdo Y
40	pulso direito X
41	pulso direito Y
42	quadril esquerdo X
43	quadril esquerdo Y
44	quadril direito X
45	quadril direito Y
46	joelho esquerdo X
47	joelho esquerdo Y
48	joelho direito X
49	joelho direito Y
50	tornozelo esquerdo X
51	tornozelo esquerdo Y
52	tornozelo direito X
53	tornozelo direito Y
54	nariz X
55	nariz Y
56	olho esquerdo X
57	olho esquerdo Y
58	olho direito X
59	olho direito Y
60	orelha esquerda X
61	orelha esquerda Y
62	orelha direita X
63	orelha direita Y
64	ombro esquerdo X
65	ombro esquerdo Y
66	ombro direito X
67	ombro direito Y

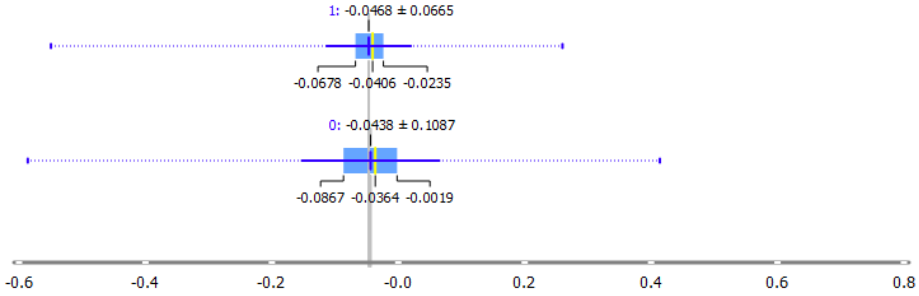
Tabela 6.2: Classes do vetor de variância e suas respectivas descrições

Classe	Rede Neural	AdaBoost
Parado	83%	78%
Deitado	74%	75%
Mãos ao alto	81%	88%

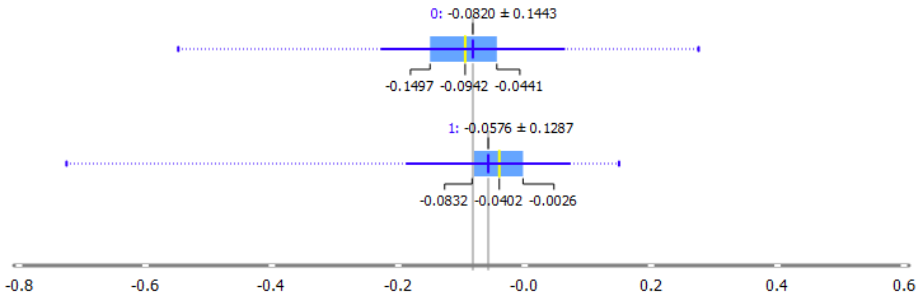
Tabela 6.3: Comparativo da taxa de acerto para cada classe, entre rede neural e AdaBoost

6.1 Considerações finais

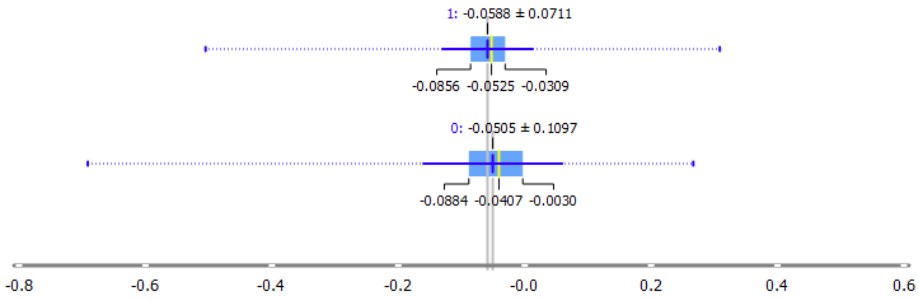
Os resultados exibidos pelas tabelas são bons e nos mostram que é possível, de fato, identificar estas ações em vídeos de vigilância em lotéricas. Através da detecção destas classes, podemos com a ajuda de um humano, identificar que está ocorrendo um assalto. Visto que, caso sejam identificadas algumas pessoas com mãos ao alto ou pessoas deitadas em uma casa lotérica, é intuitivo afirmar que algo fora do comum está ocorrendo. Ou seja, através destas informações, podemos criar mecanismos para soar algum tipo de alarme quando alguma destas ações forem realizadas por mais de uma pessoa.



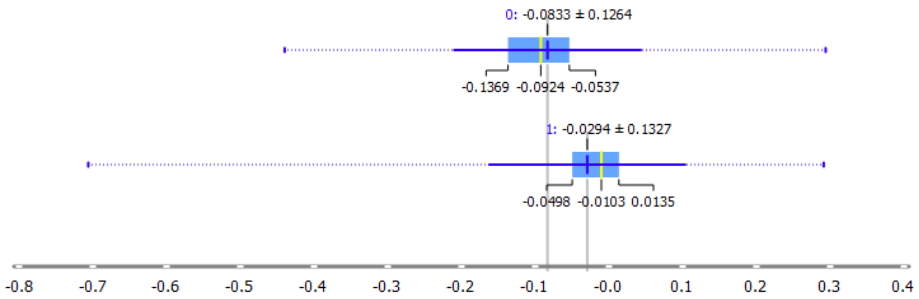
(a) Cotovelo direito para classe **Parado** vs **"não Parado"**



(b) Cotovelo direito para classe **Mãos ao alto** vs **"não Mãos ao alto"**

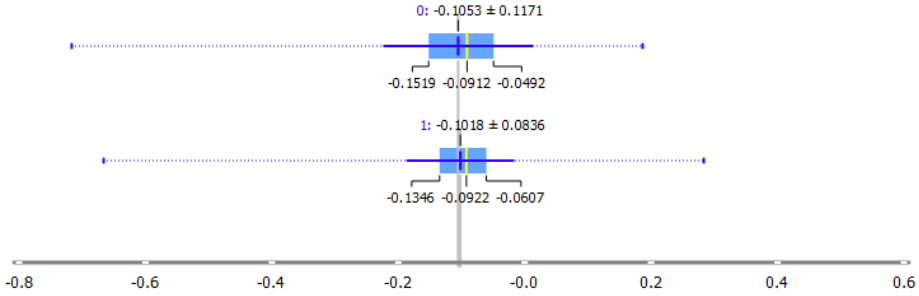


(c) Punho direito para classe **Parado** vs **"não Parado"**

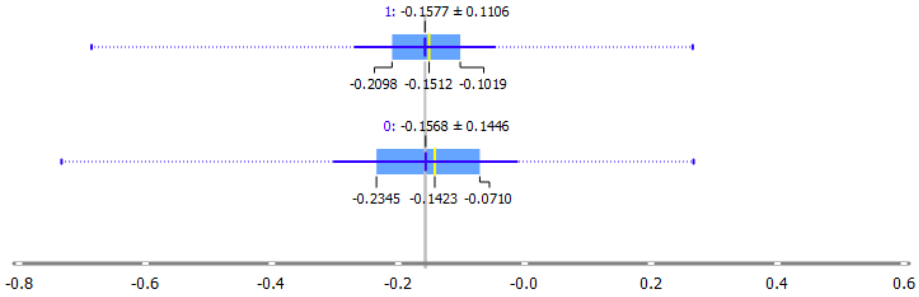


(d) Punho direito para classe **Mãos ao alto** vs **"não Mãos ao alto"**

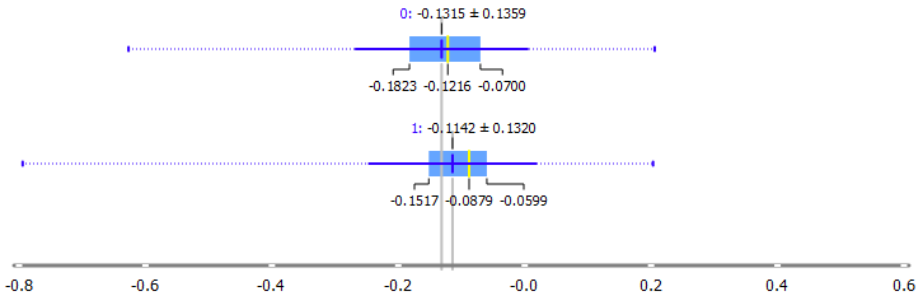
Figura 6.5: Diagrama de caixa representando uma dimensão de um conjunto de vídeos de uma classe inteira. Demonstrando comparativo entre pessoas com mãos ao alto **(classe 1)** e as que não estão **(classe 0)** e pessoas paradas **(classe 1)** e pessoas que não estão paradas **(classe 0)**.



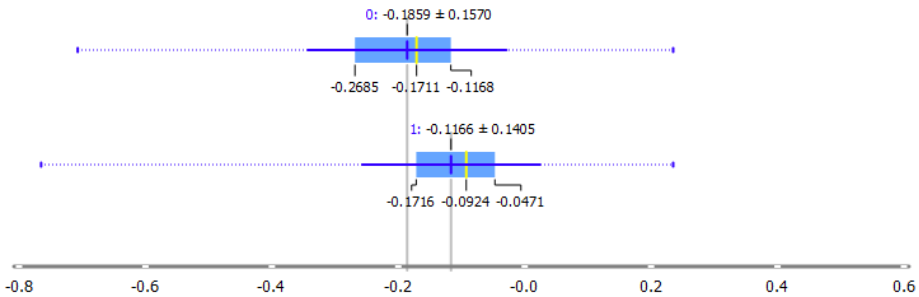
(a) Cintura direita para classe **Parado** vs **"não Parado"**



(b) Joelho direito para classe **Parado** vs **"não Parado"**

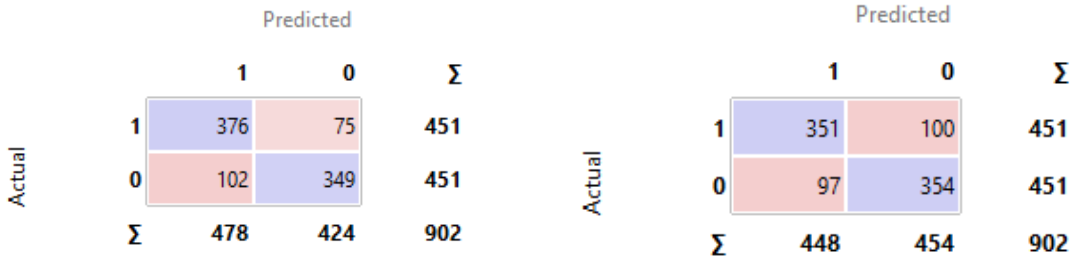


(c) Cintura direita para classe **Deitado** vs **"não Deitado"**



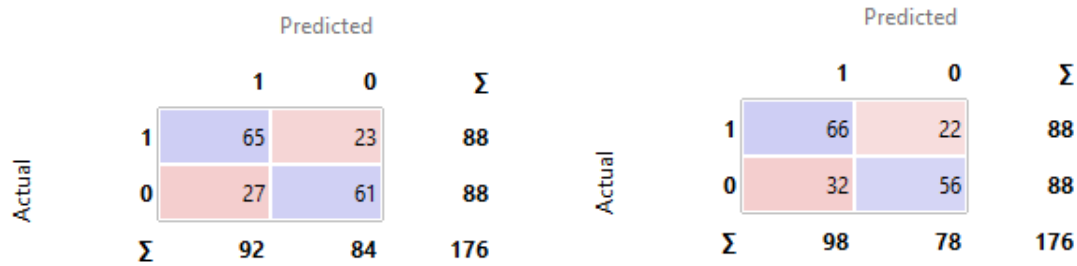
(d) Joelho direito para classe **Deitado** vs **"não Deitado"**

Figura 6.6: Diagrama de caixa representando uma dimensão de um conjunto de vídeos de uma classe inteira. Demonstrando comparativo entre pessoas deitadas **(classe 1)** e outras que não estão **(classe 0)** e pessoas paradas **(classe 1)** e pessoas que não estão paradas **(classe 0)**.



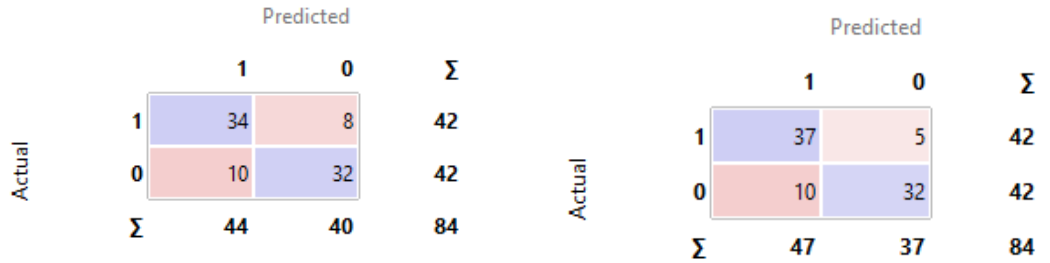
(a) Rede Neural (b) Floresta aleatória + AdaBoost

Figura 6.7: Matriz de confusão para a classe: Parado



(a) Rede Neural (b) Floresta aleatória + AdaBoost

Figura 6.8: Matriz de confusão para a classe: Deitado



(a) Rede Neural (b) Floresta aleatória + AdaBoost

Figura 6.9: Matriz de confusão para a classe: Mãos ao alto

Capítulo 7

Conclusões

Este trabalho apresenta resultado propostos por uma nova arquitetura na área de detecção de violência, onde apresentamos uma nova arquitetura para a detecção desta e uma base de dados inteiramente sobre violência nas casas lotéricas. Vimos, que existem diversos métodos para detecção e rastreio de humanos, como YOLO e *Pose Track*, onde, através da utilização e comparação, chegamos a conclusão que o mais eficaz para detectar assaltos seria através de *pose tracking*, que nos proporciona a visualização e rastreio das partes do corpo humano. Graças a isso, vimos que a classificação de algumas ações frequentes em investidas criminosas é possível, como mãos ao alto e pessoas deitadas. Sendo assim, podemos construir uma aplicação capaz de detectar estas investidas. Futuramente, conduziremos estudos para aperfeiçoar os métodos de detecção e rastreio de seres humanos, especificadamente em lotéricas, para assim podermos construir uma aplicação capaz de detectar precisamente características que juntas possam indicar um possível assalto. Possibilitando a utilização em conjunto a sistemas de monitoramento em tempo real.

Bibliografia

[1] BRASIL sobe duas posições e passa a ter 7ª maior taxa de homicídios das Américas, diz OMS. Nações Unidas Brasil, [S. l.], p. 1, 18 maio 2018. Disponível em: https://nacoesunidas.org/brasil-sobe-duas-posicoes-e-passa-a-ter-7a-maior-taxa-de-homicidios-das-americas-diz-oms/. Acesso em: 30 jun. 2019.

[2] BRASIL tem a terceira maior taxa de roubos da América Latina, diz Pnud. G1, São Paulo, p. 1, 12 nov. 2013. Disponível em: http://g1.globo.com/mundo/noticia/2013/11/brasil-tem-terceira-maior-taxa-de-roubos-da-america-latina-diz-pnud.html. Acesso em: 30 jun. 2019.

[3] LEAL, Catarina Costa. Vídeo mostra ação de criminosos em assalto a lotérica em Parnaíba. G1, Piauí, p. 1, 30 maio 2019. Disponível em: https://g1.globo.com/pi/piaui/noticia/2019/05/30/video-mostra-acao-de-criminosos-em-assalto-a-loterica-em-parnaiba.ghtml. Acesso em: 24 jun. 2019.

[4] LEAL, Rafaela. Empresário assaltado em lotérica diz que criminoso se passou por cliente para roubar dinheiro. G1, Piauí, p. 1, 19 jun. 2019. Disponível em: https://g1.globo.com/pi/piaui/noticia/2019/06/19/empresario-assaltado-em-loterica-diz-que-criminoso-se-passou-por-cliente-para-roubar-dinheiro.ghtml. Acesso em: 24 jun. 2019.

[5] SEGURANÇA é morto em tentativa de assalto à lotérica na Terra Firme. G1, Pará, p. 1, 30 abr. 2019. Disponível em: https://g1.globo.com/pa/para/noticia/2019/04/30/seguranca-e-morto-em-tentativa-de-assalto-a-loterica-na-terra-firme.ghtml. Acesso em: 24 jun. 2019.

[6] VÍDEO mostra assaltos a casas lotéricas no interior de Alagoas. G1, Alagoas, p. 1, 18 mar. 2019. Disponível em: https://g1.globo.com/al/alagoas/noticia/2019/03/18/video-mostra-assaltos-a-casas-lotericas-no-interior-de-alagoas.ghtml. Acesso em: 24 jun. 2019.

[7] Bewley, A. ; Ge, Z. ; Ott, L. ; Ramos, F. ; Upcroft, B. Simple Online and Realtime Tracking. *CoRR* abs/1602.00763 (2016).

[8] Bishop, C. M. *Neural Networks for Pattern Recognition*. New York, NY, USA : Oxford University Press, Inc., 1995. – ISBN 0198538642.

[9] Blunsden, S. ; Fisher, R. The BEHAVE video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA* 4, 1-12 (2010), 4.

[10] Fang, H.-S. ; Xie, S. ; Tai, Y.-W. ; Lu, C. RMPE: Regional Multi-person Pose Estimation. In *ICCV*, (2017), pp.

[11] Hassner, T. ; Itcher, Y. ; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, (2012), pp. 1–6.

[12] Mabrouk, A. B. ; Zagrouba, E. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications* 91 (2018), 480 - 491.

[13] Mahadevan, V. ; Li, W. ; Bhalodia, V. ; Vasconcelos, N. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (June 2010), pp. 1975–1981.

[14] Marszałek, M. ; Laptev, I. ; Schmid, C. Actions in context. In *CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition*, (2009), pp. 2929–2936.

[15] Nievas, E. B. ; Suarez, O. D. ; García, G. B. ; Sukthankar, R. Violence detection in video using computer vision techniques. In *International conference on Computer analysis of images and patterns*, (2011), pp. 332–339.

[16] Redmon, J. ; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* (2018).

[17] Ribeiro, P. C. ; Audigier, R. ; Pham, Q. C. RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance. *Computer Vision and Image Understanding* 144 (2016), 121 - 143. – Individual and Group Activities in Video Event Analysis.

[18] Soomro, K. ; Zamir, A. R. ; Shah, M. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision* (2012).

[19] Wu, Y. ; Lim, J. ; Yang, M. Object Tracking Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 9 (2015), 1834–1848.

[20] Xiu, Y. ; Li, J. ; Wang, H. ; Fang, Y. ; Lu, C. Pose Flow: Efficient Online Pose Tracking. In *BMVC*, (2018), pp.