PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

RECONHECIMENTO DE VIOLÊNCIA EM VÍDEO UTILIZANDO DEEP LEARNING

ARIEL REICHELT NESSI

Proposta de Trabalho de Conclusão apresentada como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Rodrigo C. Barros

Porto Alegre
2018

RECONHECIMENTO DE VIOLÊNCIA EM VÍDEO UTILIZANDO DEEP LEARNING

RESUMO

A presença ubíqua de câmeras de segurança pelas capitais e a implementação de grandes sistemas de monitoramento já são uma realidade. No entanto é humana impossível controlar de modo eficiente o que ocorre em um sistema com mais de 1.000 câmaras simultâneas. Ao mesmo tempo, a escalada da violência nos últimos anos demanda uma solução inteligente para este problema público. Dado esta problemática, este trabalho visa realizar um estudo de técnicas de *deep learning* a fim gerar um modelo capaz de detectar crimes em andamento em sistemas de monitoramento por videocâmara.

Palavras-Chave: Deep Learning, Reconhecimento de Ações, Detecção de Crime.

VIOLENCE RECOGNITION IN VIDEO WITH DEEP LEARNING

ABSTRACT

The ubiquitous presence of security cameras throughout capitals and the implementation of big monitoring centers are a set reality. However it is not feasible for a human operator to deal efficiently with grids with well over 1.000 simultaneous devices. Meanwhile the crescent rise in crime rates during the lasts years demands for an intelligent solution for this public issue. Given this scenario this jobs sets out to study deep learning techniques that would enable a system to recognize ongoing crimes captured in monitoring systems.

Keywords: Action Recognition, Crime Detection, Deep Learnig.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

CCTV – Circuito Fechado de televisão

DL – Deep Learning

GPU – Graphical Processing Unit

IVS – Intelligent Visual Surveillance

ONU – Organização das Nações Unidas

ONUDC – United Nations Office on Drugs and Crime

SVM – Support Vector Machines

SUMÁRIO

1. INTRODUÇÃO

1.1 Motivação

A violência no Brasil é uma matéria de interesse público, alvo de diversos estudos por instituições nacionais e internacionais, entre as organizações que monitoram a criminalidade a *United Nations Office on Drugs and Crime* (UNODC), entidade ligada à ONU, realiza uma avaliação anual em diversos países, compilando índices em escala internacional. A Tabela 1.1 evidencia a posição do Brasil para cada tipo de crime monitorado no último ano em que este participou da avaliação.

Tabela 1.1: Posição do Brasil por crime

Crime	Ano	Posição	Total de Países
Agressão	2013	2	98
Estupro	2013	2	109
Furto	2013	2	111
Furto em Domicílio	2012	24	82
Furto em Estabelecimento	2013	6	101
Homicidio	2015	1	122
Roubo	2013	1	101
Roubo de carro	2013	2	95
Sequestro	2013	19	86
Sexual	2013	3	108

O país se encontra entre os três países de maior incidência para sete dos dez crimes monitorados. Em uma pesquisa de vitimização em escala nacional [3] 43.9% dos brasileiros relatam temer ser assaltados no próprio bairro, 43.5% tem medo de agressão física e 45.7% se sentem inseguros ao andar na rua de modo geral. Em uma pesquisa na capital gaúcha [18] 50.2% dos entrevistados reporta se sentir muito inseguro à noite no bairro em que reside e 82.5% consideram Porto Alegre uma cidade bastante violenta.

Dado este cenário, sistemas de vigilância inteligentes (IVS) se tornam aliados interessantes tanto para população quanto para o governo, e existem esforços notáveis por parte de prefeituras em direção à uma vigilância digitalizada, dando origem à centrais de monitoria por videocâmara [16] [13]. No âmbito da pesquisa diversas aplicações de IVS vêm sendo exploradas dentro do domínio de segurança, dentre elas, controle de acesso, identificação de pessoas, fluxo de multidões e detecção de anomalias [9] [10].

Não obstante o esforço despendido nesta área a detecção de anomalias e a categorização de ações humanas, ainda apresenta desafios consideráveis para a visão compu-

tacional. Enquanto existem resultados sólidos para classificação de diversas ações simples como sentar, caminhar e correr [1], a identificação de violência em vídeo, devido uma escassez de dados anotados, se dá majoritariamente na literatura em trechos de filmes. Trabalhos recentes introduzem conjuntos de dados mais realísticos como brigas em partidas de Hockey [1] e ocorrências de diversos crimes capturados por câmeras de circuito fechado de televisão (CCTV) [19]. Conquistas obtidas no cenário de hardware em conjunto com bibliotecas robustas de paralelização vêm viabilizando também a utilização de arquiteturas de redes neurais profundas, trazendo progressos consistentes em diversas tarefas dentro e fora da visão computacional[11].

1.2 Objetivos

O objetivo deste trabalho é estudar arquiteturas de Aprendizado Profundo na detecção de atividades criminosas capturadas por câmeras de vigilância e oferecer um *feedback* em tempo-real para operadores de segurança.

Como objetivos é possível pontuar :

1. Experimentar soluções existentes na detecção de anomalias e ou violência em ambientes de CCTV.
2. Estabelecer um baseline sólido para o problema.
3. Desenvolver uma técnica de identificação de crime em *streaming*.

1.3 Organização do texto

Este documento se encontra organizado em 4 capítulos, no capitulo 2, Referencial Teórico, são introduzidos os conceitos básicos de *deep learning*, visão computacional e uma breve explicação e contextualização da tarefa de reconhecimento de ações. No capítulo 3 é feita uma revisão dos trabalhos relacionados que serão utilizados como blocos de construção no processo de solução do problema. No capítulo 4 é apresentada a abordagem pela qual este trabalho visa detectar violência em vídeos, se trata efetivamente da proposta do TCC onde consta os resultados finais que se esperam obter e um cronograma de execução.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo visa apresentar a área de visão computacional, em um segundo momento é feita uma introdução à conceitos básicos de Aprendizado Profundo (*Deep Learning*, DL), por final definimos do que se tratam as tarefas de reconhecimento de ações e detecção de anomalias dentro do contexto deste trabalho.

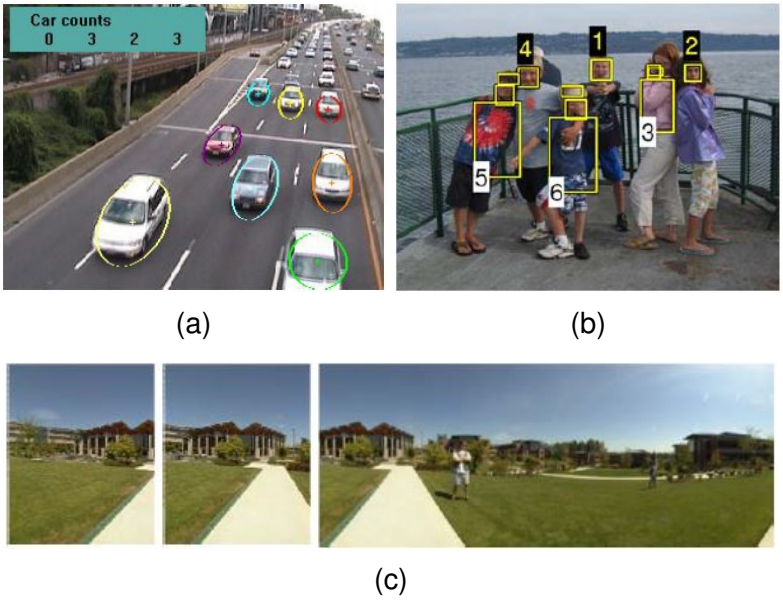
2.1 Visão Computacional

A visão computacional é uma área da ciência da computação que estuda a implementação da percepção visual do ser humano através de algoritmos, buscando reconhecer propriedades de elementos em uma imagem, como cor, formato, tamanho e iluminação [20]. Algumas aplicações notáveis da área que podem ser comumente observadas nas novas gerações de celulares e dispositivos móveis são, reconhecimento biométrico e facial, estabilização de imagens trêmulas e fotografias panorâmicas. A visão computacional no entanto não se limita à imagens estáticas, com os recentes desenvolvimentos em Unidades de Processamento Gráfico (*Graphical Processing Unit*, GPU) nas últimas duas décadas [14] soluções para problemas em vídeo vêm se tornando tratáveis.

Embora existam diversas abordagens para problemas de visão computacional, as tarefas podem ser agrupadas em três níveis genéricos [6]:

- **Nível Baixo**: Não requer capacidades de inteligência, como aplicação de um filtro ou eliminação de ruído.
- **Nível Médio**: Consistem em identificar pontos de interesse e elementos em uma imagem realizando, por exemplo, segmentação e detecção de objetos.
- **Nível Alto**: Tarefas que simulem a cognição humana e busquem interpretar uma imagem, como reconhecimento de ações e descrição de cenas.

Esta hierarquia se organiza de modo incremental, de modo que abordagens para soluções de alto nível comumente se utilizam de construções dos níveis inferiores. Na Figura 2.1 apresentamos alguns exemplo de tarefas de cada um dos níveis descritos acima.



(a) (b)

(c)

Figura 2.1: Exemplos de tarefas de alto, médio e baixo nível, respectivamente, (a) Reconhecimento de tráfego, (b) Segmentação por pessoas e (c) Processamento de foto panorâmica. Imagens retiradas de Szeliski [20].

2.2 Deep Learning

Aprendizado Profundo, melhor conhecido como *Deep Learning*, consiste de um método de aprendizado de máquina que busca aprender representações através de um conjunto de dados. Nesta abordagem os dados atravessam por uma rede de camadas interconectadas, a intuição do método é de que cada uma destas camadas aprenda atributos incrementalmente mais complexos. A Figura 2.2 ilustra uma arquitetura de deep learning teórica que busca aprender representações para pessoas, carros e animais, neste exemplo as camadas podem ser descritas como:

- **Camada Visível**: Canais de cores visíveis da imagem são dados como entrada
- **Primeira Camada**: Busca aprender representações para linhas através da identificação de padrões de alteração de cores fornecidos pela camada anterior.
- **Segunda Camada**: Identifica bordas e contornos através da utilização das diversas linhas identificadas na Primeira Camada.
- **Terceira Camada**: Assimila objetos distintos através da combinação de diferentes bordas e contornos extraídos.
- **Quarta Camada**: Retorna entre os objetos reconhecidos pela rede, aquele de maior semelhança com a imagem de entrada.

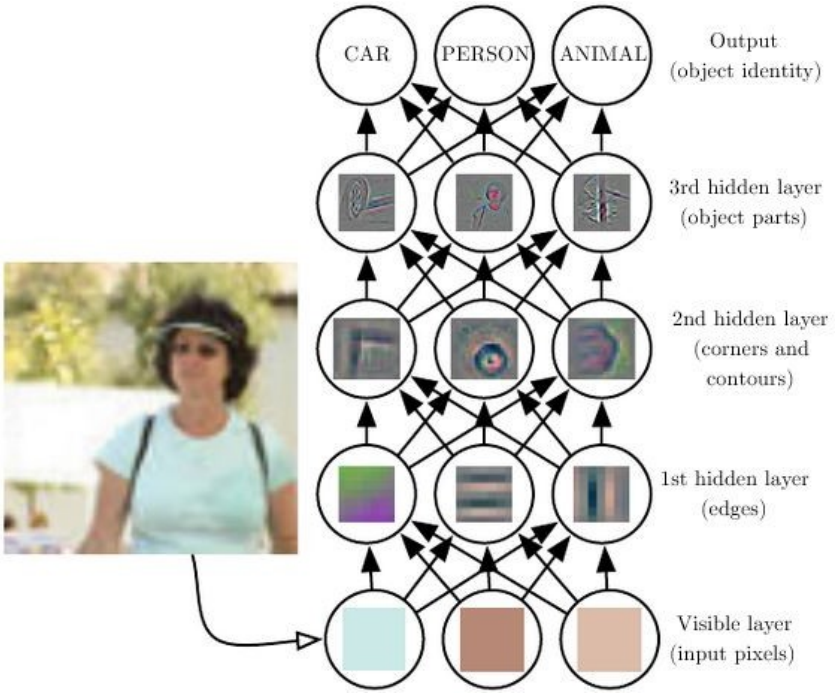


Figura 2.2: Exemplo de um modelo de *Deep Learning* [7].

2.3 Reconhecimento de ações

A tarefa de reconhecimento de ações consiste em categorizar ou ainda, descrever, uma ação observada. A definição do que consiste uma ação varia na literatura entre, um padrão de movimentos simples executados por um único indivíduo [21], até definições mais formais como a de Pope *et al* [15] que estabelece unidades atômicas primitivas, cuja sequência de interações configura uma ação. Podemos utilizar também as ações contidas em *datasets* disponíveis na área [2] como base para definição da tarefa, instâncias de ações incluem, caminhar, pular, acenar, bater palmas, sentar e deitar. Alguns autores fazem distinção entre ações e atividades, como movimentos de um único individuo e uma ação complexa conjunta entre um grupo de pessoas, respectivamente, exemplos desta última são lutar, apertar mãos, empurrar, abraçar e beijar.

As aplicações desta tarefa são diversas, especialmente para sistemas de segurança pública, que contam com uma baixíssima razão câmera-operador em um cenário de baixa ocorrência de eventos. Soluções para o reconhecimento de ações dependem de dois grandes componentes, os atributos extraídos em tempo de pré-processamento, geralmente consistindo em informações dos nível baixo e médio, conforme descritos na Sessão 2.1, e um algoritmo de aprendizado de máquina ou *deep learning* que usa estes dados como entrada para seu processamento.

Apesar de possuir aplicações promissora esta tarefa ainda apresenta desafios bem conhecidos da visão computacional, entre eles os mais notáveis são [21]:

- **Variação do ponto de vista**: Obter soluções robustas à alterações do ponto de vista é um desafio por si só. Considere que a mesma ação pode ser observada por uma infinidade de ângulos e é desejável que o sistema seja robusto à estas alterações.
- **Variação de execução**: A mesma ação pode ser executada em diferentes velocidades e de diferentes maneiras, gerando uma infinidade de variações intra-classe.
- **Antropometria**: Humanos vêm em diferentes tamanhos, sexo e cores. No entanto se espera que sistemas de reconhecimento de ação sejam capazes de identificar ações invariavelmente dos aspectos físicos envolvidos.

3. TRABALHOS RELACIONADOS

Neste capítulo é feito uma revisão da literatura de visão computacional em trabalhos relevantes para a detecção de crime, enquanto existem tópicos que oferecem contribuições diretas à essa tarefa, como reconhecimento de violência e de anomalias, é importante frisar que para o melhor conhecimento do autor, a detecção de crimes em vídeo não configura uma área ativamente pesquisada.

3.1 Reconhecimento de violência

A detecção de violência pode ser considerado uma tarefa de reconhecimento de ações, descrita na Seção 2.3 em que o objetivo é identificar um comportamento agressivo ou um evento de luta [1]. Devido a natureza da tarefa, diversas abordagens exploram capturar as variações de movimentação entre frames.

Datta *et al* [4] em 2002, propõe um método baseado em *hand-crafted features* para detecção da cabeça e dos membros, o método então calcula vetores de aceleração e orientação dos membros, no entanto o método é extremamente sensível à mudanças de perspectiva da câmera, a proposta depende também que os eventos ocorra com os participantes na horizontal. Nievas *et al* [1] em 2011 introduz um conjunto de dados que consiste de brigas de Hockey, para a classificação dos *frames* em que ocorrem eventos violentos, o artigo propõe a utilização de uma máquina de vetores de suporte (do inglês, *Support Vector Machines*, SVM), tratando a tarefa como uma classificação binária de violento ou não-violento. A SVM é alimentada com um vetor de descritores de movimentação extraída em tempo de pré-processamento, apesar da abordagem se mostrar mais robusta à variação de perspectiva, os métodos aplicados são computacionalmente custosos reduzindo sua usabilidade em cenários de tempo real.

Hassner *et al* [8] em 2012 propõe uma abordagem de detecção para áreas densamente populadas em **tempo real**, utilizando-se também de uma SVM a contribuição deste trabalho se dá na introdução de um novo descritor de vídeo, que identifica movimentações bruscas comparando a variação média de magnitude entre vetores de fluxo. Esta solução se mostra robusta a mudanças de perspectiva e apresenta resultados promissores em dados reais, introduzidos no mesmo artigo. Gao *et al* [5] apresenta um novo descritor desenvolvido como um incremento ao trabalho de Hassner [8] incorporando nos vetores de deslocamento a informação de orientação, o trabalho alcança resultados superiores ao antecessor com uma combinação do novo descritor, aliada à utilização de um classificador AdaBoost para selecionar os atributos relevantes e uma SVM para classificação do vídeo.

3.2 Reconhecimento de anomalias

O reconhecimento de anomalias consiste na identificação de uma atividade atípica, no entanto, enumerar todos comportamentos que configuram um desvio do normal não é factível, a literatura de modo geral considera anômalo todo comportamento de possível interesse no monitoramento de sistemas de CCTV [22][19][12][23]. Genericamente, as abordagens para reconhecimento de anomalias buscam criar um modelo de normalidade e então classificam como anormal frames que apresentam um desvio deste.

Xiang *et al*[22] em 2008 propôs um classificador não supervisionado que gera um agrupamento para um número de categorias desconhecido e identifica anomalias acumulando a distancia entre comportamento observado e conhecido como normal. Experimentos são realizados na detecção de fluxos incomuns de pessoas em um cenário de câmera fixa.

Em 2012 Lu *et al* [12] aplica um modelo de aprendizado esparso no qual cada frame é analisado em várias escalas diferentes, aprendendo *features* em diferentes níveis de agrupamento de pixels. A solução proposta é aplicável em tempo real e identifica ações como fluxo em direções anormais e movimentações bruscas.

Trabalhos mais recentes na área incluem, Xu *et al* [23] 2015, que propõe o uso de uma rede profunda não-supervisionada para extrair representações de aparência, movimentação e uma combinação destes dois fatores. Após a aplicação da rede, três classificadores SVM são alimentados com as representações extraídas e combinados para fornecer a classificação final da instância. Em 2018 Chen *et al* propõe uma abordagem supervisionada baseada em redes profundas e traz o primeiro conjunto de dados que envolve não apenas movimentações anômalas, mas situações de risco como acidentes de trânsito, roubo e brigas. O trabalho apresenta resultados competitivos no seu próprio *dataset*, no entanto o autor não reporta seus resultados nos *datasets* utilizados até então por outros métodos.

4. METODOLOGIA

4.1 Abordagem

Este trabalho busca desenvolver uma arquitetura de rede neural profunda capaz de realizar a detecção de eventos que ameacem a segurança pública. A solução deve executar em tempo real e levaremos em consideração a realidade do Brasil, desconsiderando eventos como abandono de objetos e demais preocupações ligadas ao terrorismo. Se possível, o modelo será refinado para incluir a detecção do tipo de crime em andamento e a segmentação para facilitar a visualização dos operadores.

4.2 Datasets

Para o desenvolvimento deste trabalho foi feita uma analise dos *datasets* existentes em trabalhos relacionados. Para a finalidade de detecção de crimes 3 datasets foram separados para utilização na modelagem da solução.

- **Hockey**: Vídeos de brigas em partidas de Hockey com movimentos de câmera e close-ups.
- **UCF Crime**: Vídeos coletados do *YouTube*, gravados por câmeras de CCTV capturando a ocorrência de diversos crimes.
- **Violent Flows**: Vídeos coletados do youtube apartir de diferentes meio de gravação como CCTV e câmera de celular .

4.2.1 Hockey

Este *dataset* foi construído como parte do trabalho de Nievas *et al* [1] para detecção de violência. A base consiste de 1.000 clipes de ação em partidas de Hockey da Liga Nacional de Hockey, cada clipe consiste de 50 frames com uma resolução de 720x576 pixels e manualmente anotados como "briga"ou "não-briga", a anotação é feita apenas em nível do vídeo, no entanto dado o curto tamanho dos clipes, as ações de briga geralmente se encontram inteiramente dentro dos 50 frames.

O trabalho inclui também um conjunto de dados para o teste, consistindo de 200 clipes, 100 contendo violência e outros 100 que apresentam atividades normais.

4.2.2 UCF Crime

Introduzido por Chen *et al* [19] este *dataset* conta com a captura de 13 anomalias dentre elas, abuso, acidente automotivo, agressão, arresto, arrombamento, briga, explosão, furto, furto em loja, incêndio, roubo, tiroteio e vandalismo. Este conjunto de dados possuí um total de 1900 vídeos, 950 contendo anomalias e 950 vídeos de atividades normais, na Tabela 4.1 a quantidade para cada tipo de anomalia se encontra anotada.

Os dados se encontram anotados em dois níveis diferentes de especificidade, um nível semi-anotado em que os vídeos são apenas identificados como possuindo uma anomalia ou não, e outro nível mais especifico nos quais os vídeos também possuem os frames de inicio, fim e tipo de anomalia. A quantidade média de quadros por instância são 7.247 em um total de 128 horas de vídeo. Todos vídeos consistem de capturas de CCTV, adquiridas por buscas no *YouTube*.

Tabela 4.1: Número de vídeos por anomalia

Anomalia	# De Vídeos
Abuso	50
Acidente Automotivo	150
Agressão	50
Arresto	50
Arrombamento	100
Briga	50
Explosão	50
Furto	100
Furto em Loja	50
Incêndio	50
Roubo	150
Tiroteio	50
Vandalismo	50

4.2.3 Violent Flows

Coletado por Hassner *et al* [8] este conjunto consiste majoritariamente de brigas entre torcidas de futebol, capturadas no *YouTube*. As cenas de agressão deste *dataset* têm como característica ocorrências em lugares com multidões. Contêm uma média de 3.6 segundos de duração por vídeo e conta com 246 instâncias.

4.3 Ambiente de Desenvolvimento

Para o desenvolvimento deste projeto utilizaremos o ambiente de desenvolvimento *PyTorch* [17], que consiste de uma biblioteca robusta de desenvolvimento em *Python*, fornecendo funcionalidades de aceleração em GPU bem como, facilitando as implementações de algoritmos de redes neurais através de um sistema de auto-diferenciação. A biblioteca é capaz de gerar automaticamente o grafo de diferenciação da rede em tempo de execução e efetua o calculo dos gradientes correspondentes.

4.4 Cronograma

As atividades programadas durante o periodo de TC I são:

1. Definição da tarefa a ser executada em conjunto com o orientador;
2. Elaboração da Proposta de TCC;
3. Entrega da proposta;
4. Revisão da proposta após avaliação pelo Orientador e Avaliador;
5. Estudo das técnicas de extração de *features* utilizadas na detecção de violência e anomalia;
6. Estudo de arquiteturas de Redes Neurais Profundas
7. Configuração e obtenção de recursos computacionais necessários.
8. Redigir documento Final do TCC I
9. Entrega final do documento de TCC I

Tabela 4.2: Cronograma de Atividades do TCI

Atividade	Março	Abril	Maio	Junho
1	X			
2	X	X		
3		X		
4		X		
5		X	X	X
6		X	X	X
7				X
8			X	
9				X

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Bermejo Nievas, E.; Deniz Suarez, O.; Bueno García, G.; Sukthankar, R. “Violence detection in video using computer vision techniques”. In: Computer Analysis of Images and Patterns, Real, P.; Diaz-Pernil, D.; Molina-Abril, H.; Berciano, A.; Kropatsch, W. (Editores), 2011, pp. 332–339.

[2] Chaquet, J. M.; Carmona, E. J.; Fernández-Caballero, A. “A survey of video datasets for human action and activity recognition”, *Computer Vision and Image Understanding*, vol. 117–6, 2013, pp. 633–659.

[3] CRISP; DATAFOLHA. “Pesquisa nacional de vitimização”, Relatório Técnico, CRISP, Av. Presidente Antônio Carlos, Pampulha - Unidade Administrativa III, 2013, 43p.

[4] Datta, A.; Shah, M.; Lobo, N. D. V. “Person-on-person violence detection in video data”. In: Pattern Recognition, 2002. Proceedings. 16th International Conference on, 2002, pp. 433–438.

[5] Gao, Y.; Liu, H.; Sun, X.; Wang, C.; Liu, Y. “Violence detection using oriented violent flows”, *Image and vision computing*, vol. 48, 2016, pp. 37–41.

[6] Gonzalez, R. C.; Woods, R. E.; Eddins, S. L. “Digital Image Processing using MATLAB”. Pearson, 2004.

KEY: GONZALEZCV
ANNOTATION: SIGNATUR = 000.000

[7] Goodfellow, I.; Bengio, Y.; Courville, A. “Deep Learning”. MIT Press, 2016, http://www.deeplearningbook.org.

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

[9] Hu, W.; Tan, T.; Wang, L.; Maybank, S. “A survey on visual surveillance of object motion and behaviors”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34–3, Aug 2004, pp. 334–352.

[10] Kim, I. S.; Choi, H. S.; Yi, K. M.; Choi, J. Y.; Kong, S. G. “Intelligent visual surveillance — a survey”, *International Journal of Control, Automation and Systems*, vol. 8–5, Oct 2010, pp. 926–939.

[11] LeCun, Y.; Bengio, Y.; Hinton, G. “Deep learning”, *Nature*, vol. 521–7553, 5 2015, pp. 436–444.

[12] Lu, C.; Shi, J.; Jia, J. “Abnormal event detection at 150 fps in matlab”. In: Computer Vision (ICCV), 2013 IEEE International Conference on, 2013, pp. 2720–2727.

[13] Luiz, M. “Número de câmeras em porto alegre vai quase dobrar até a copa”. Capturado em: http://g1.globo.com/rs/rio-grande-do-sul/noticia/2013/06/numero-de-cameras-em-porto-alegre-vai-quase-dobrar-ate-copa.html, Mar 2018.

[14] McClanahan, C. “History and evolution of gpu architecture”, *A Survey Paper*, 2010, pp. 9.

[15] Poppe, R. “A survey on vision-based human action recognition”, *Image and vision computing*, vol. 28–6, 2010, pp. 976–990.

[16] Prefeitura de, S. P. “City cameras”. Capturado em: https://www.citycameras.prefeitura.sp.gov.br/howworks, Mar 2018.

[17] Pytorch. “Pytoch”. Capturado em: http://pytorch.org/, Mar 2018.

[18] Radmann, E.; Orso, M.; Rodrigues, G.; Velho, E.; Massaú, E.; de Nascimento, M.; Mello, D.; D’avila, F.; da Silva, I. M.; Fernandez, M.; Silva, C.; Munchow, A. C.; Klain, G.; Rolim, M. “Primeira pesquisa de vitimização de porto alegre”, Relatório Técnico, Instituto Cidade Segura, Av. Presidente Antônio Carlos, Pampulha - Unidade Administrativa III, 2017, 50p.

[19] Sultani, W.; Chen, C.; Shah, M. “Real-world Anomaly Detection in Surveillance Videos”, *ArXiv e-prints*, Jan 2018, 1801.04264.

[20] Szeliski, R. “Computer Vision: Algorithms and Applications”. New York, NY, USA: Springer-Verlag New York, Inc., 2010, 1st ed..

[21] Turaga, P.; Chellappa, R.; Subrahmanian, V. S.; Udrea, O. “Machine recognition of human activities: A survey”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18–11, Nov 2008, pp. 1473–1488.

[22] Xiang, T.; Gong, S. “Video behavior profiling for anomaly detection”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 30–5, 2008, pp. 893–908.

[23] Xu, D.; Ricci, E.; Yan, Y.; Song, J.; Sebe, N. “Learning deep representations of appearance and motion for anomalous event detection”, *arXiv preprint arXiv:1510.01553*, 2015.