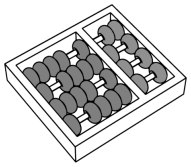
Universidade Estadual de Campinas

Instituto de Computação



Proposta de Dissertação de Mestrado

Localização de cenas pornográficas usando informações temporais e técnicas de *Deep Learning*

Candidato: João Paulo P. Martin

Orientador: Prof. Dr. Zanoni Dias

Coorientadora: Prof[a]. Dr[a]. Sandra Avila

8 de maio de 2017

Sumário

Resumo

A produção e o consumo de vídeo em tempo real ou sob demanda alcançaram um volume que torna infactível a classificação e curadoria humanas. A identificação e filtragem de conteúdos como violência e pornografia, hoje disponíveis irrestritamente, é um problema atual importante. Vários trabalhos foram realizados abordando a detecção de conteúdo pornográfico em imagens e vídeos usando técnicas como detecção de pele, descritores locais, dicionários de palavras visuais e aprendizagem profunda.

O problema abordado neste projeto refere-se à *localização* de conteúdo pornográfico, uma tarefa em que se deseja encontrar um tipo específico de conteúdo e retornar seus limites temporais dentro de um vídeo. Pouquíssimos trabalhos foram realizados sobre esta tarefa, entretanto a literatura apresenta algumas técnicas como Redes Neurais Convolucionais 3D (3D-CNNs, do inglês *3D-Convolutional Neural Networks*) e Redes Neurais Recorrentes (RNNs, do inglês *Recurrent Neural Networks*) que encontraram bons resultados quando aplicadas a outros contextos (como classificação de atividades e geração automática de legendas) e que podem ser úteis nessa tarefa.

Neste projeto, deseja-se explorar técnicas de aprendizado de máquina e visão computacional para extrair informações temporais e de movimento presentes nos vídeos, que sejam úteis para a localização de conteúdo pornográfico.

1 Introdução

O aumento de banda e velocidade para o acesso à internet, aliado à queda de preços de armazenamento de dados em nuvem, possibilitou o crescimento exponencial em produção e consumo de vídeo online. Hoje, a maior parte do tráfego de rede em horário de pico é relacionado a entretenimento em tempo real[1]. Ao percorrer redes sociais, é possível encontrar vários tipos de conteúdo transmitido em tempo real como shows, atividades diárias, cultos religiosos, protestos, e até mesmo crimes e suicídios.

É importante classificar, descrever e legendar estes conteúdos de forma automática para que eles sejam melhor encontrados pelas pessoas interessadas. Esta análise de conteúdo também é necessária para avisar, proteger ou mesmo censurar a apresentação de conteúdo impróprio a um determinado público (por exemplo, crianças). Caso o conteúdo identificado exija uma ação da polícia, também é possível desencadear automaticamente a requisição. A análise de conteúdo também pode gerar ganho de tempo e abrangência em atividades forenses e de persecução criminal. Assim, depender de supervisão humana para endereçar estas tarefas pode ser prejudicial aos indivíduos envolvidos por causa dos conteúdos analisados, além de ser também uma tarefa infactível devido ao volume de dados produzidos diariamente.

Entre as tarefas que podem ser realizadas sobre imagens e vídeos, a *classificação* de conteúdo pornográfico é um caso bastante estudado na literatura [18, 19, 26, 27]. Por outro lado, a tarefa de *localização* de pornografia em vídeo tem, até onde se pode verificar, pouquíssimos trabalhos. Localização é a detecção de um conteúdo procurado e a determinação de seu início e fim dentro de um *streaming* de vídeo. Por exemplo, Moreira [31] encontrou bons resultados para localização de pornografia em vídeo utilizando a técnica de *Bag-of-Visual-Words* (BoVW) [41] (abordagem de dicionários de palavras visuais). Este projeto pretende contribuir na tarefa de localização de pornografia em vídeo.

Usualmente, no contexto de detecção de conteúdo pornográfico em vídeo, as técnicas aplicadas são adaptações das técnicas inicialmente desenvolvidas para processamento de imagens [3, 19, 31, 38]. Uma técnica que tem se destacado na literatura é *Deep Learning*, que pode ser descrita como uma classe de técnicas de aprendizado de máquina que explora muitas camadas de processamento não linear de informação para extração e transformação

[1] https://www.sandvine.com/downloads/general/global-internet-phenomena/2012/1h-2012-global-internet-phenomena-report.pdf

de características, e para classificação e análise de padrões [9].

Deep Learning, mais especificamente através de Redes Neurais Convolucionais (CNNs, do inglês *Convolutional Neural Networks*) [23, 25], representa o estado da arte em classificação de imagens [20, 22, 45]. Para o processamento de dados com forte relação temporal, Redes Neurais Recorrentes (RNNs, do inglês *Recurrent Neural Networks*) [6] e suas variantes como Memória de Curto e Longo Prazo (LSTM, do inglês *Long Short-Term Memory*) [15, 17] têm alcançado bons resultados. Sua aplicação na localização de conteúdo pornográfico pode alavancar a exploração de informação temporal presente nos vídeos.

Neste projeto, deseja-se explorar técnicas de aprendizado de máquina e visão computacional para extrair informações temporais e de movimento presentes nos vídeos, que sejam úteis para a localização de conteúdo pornográfico. O restante do texto está organizado da seguinte forma. A Seção 2 apresenta explanação sobre as abordagens de detecção de pornografia em imagens e vídeos, e a localização de pornografia em vídeo. Discorre-se também sobre algumas técnicas encontradas na literatura para a realização de tarefas de processamento de vídeo não relacionadas à detecção de pornografia (por exemplo, geração automática de legendas). A Seção 3 introduz brevemente os conceitos e as arquiteturas úteis a este projeto, relacionadas às redes neurais como CNNs e RNNs. A Seção 4 apresenta a metodologia de pesquisa proposta, a base de dados utilizada, as métricas de validação dos experimentos e o cronograma do projeto.

2 Trabalhos Relacionados

Os trabalhos referentes à detecção de conteúdo pornográfico evoluíram juntamente com os algoritmos e técnicas pesquisados. Em princípio, muito da pesquisa foi realizada sobre imagens [12, 13, 21], mas com o crescimento da disponibilidade de vídeos, passamos a tratar problemas de detecção nesse universo [31, 33, 35]. As subseções abaixo apresentam as seguintes discussões: a Subseção 2.1 descreve trabalhos relacionados à detecção de pornografia em imagem, a Subseção 2.2 relata trabalhos relacionados à detecção de pornografia em vídeo, a Subseção 2.3 apresenta trabalhos da literatura relacionados à localização de pornografia em vídeo e a Subseção 2.4 apresenta trabalhos relacionados às tarefas de processamento de vídeo não relacionados à detecção de conteúdo pornográfico.

2.1 Detecção de Pornografia em Imagem

Os primeiros trabalhos que abordaram a detecção de conteúdo pornográfico em imagens usaram algoritmos capazes de analisar características locais dos *pixels*, como intensidade e cor, ou características de micro-regiões da imagem, como textura. Usando essas ferramentas, a estratégia para detecção de pornografia foi a detecção de pele. Por exemplo, Fleck *et al.* [12] conseguiram identificar tons e cores de pele. Essas possíveis regiões de pele encontradas eram submetidas a restrições geométricas buscando encontrar estruturas corporais humanas. Da mesma forma, Forsyth e Fleck [13] buscavam regiões compostas de *pixels* relacionados à pele usando propriedades de textura e cor combinadas. Esse conteúdo era submetido a um agrupador geométrico que indicava se havia encontrado uma estrutura suficientemente complexa que poderia ser associada a uma parte do corpo humano. Jones e Rehg [21] construíram modelos para atributos de baixo nível como cor, gerando histogramas baseados em uma grande base de dados com *pixels* classificados entre as classes pele e não-pele. Estas abordagens apresentaram bons resultados para a detecção de nudez. Entretanto, a definição de conteúdo pornográfico é mais complexa que a identificação de pele humana. Quando aplicadas à detecção de pornografia, essas ferramentas apresentavam alto índice de falso positivo em atividades como natação ou lutas, e também índice elevado de falso negativo quando analisavam conteúdos com oclusão parcial ou conteúdos em que atividades sexuais eram realizadas com indivíduos não totalmente despidos.

Avila [1] relata que podemos classificar as características extraídas das imagens como locais ou globais, devido à região da imagem utilizada para a extração. Uma característica global é extraída através do processamento da imagem completa enquanto uma característica local é extraída usando-se pequena porção da imagem. Ao processarmos as características extraídas, podemos converter estas informações em descritores. Alguns exemplos de descritores locais são SIFT (*Scale Invariant Feature Transformation*) [30] e o SURF (*Speeded Up Robust Features*) [5], bastante utilizados na literatura de visão computacional.

Descritores globais podem ser submetidos diretamente a classificadores (classificação será abordada logo adiante neste texto), enquanto que descritores locais podem precisar de um método para agregá-los para que, em seguida, sejam submetidos à classificação. Após a geração de descritores de baixo nível, o passo seguinte no processo mais comum

de classificação de imagem é a extração de características intermediárias com o intuito de identificar conceitos de nível intermediário através do processamento dos descritores de baixo nível. Um exemplo desse tipo de técnica é o *Bag-of-Visual-Words* (BoVW) [41]. Deselaers *et al.* [10] propuseram um modelo BoVW que aumentou a identificação de imagens pornográficas em comparação com métodos de detecção de nudez. Steel [43] propôs um método de reconhecimento de imagens pornográficas baseado em BoVW usando *mask*-SIFT em um sistema de classificação em cascata.

Descritores podem ser úteis para relatar o que se procura nos dados analisados. Entretanto, em sua concepção, é necessário que se faça um trabalho manual de análise dos dados e do conteúdo que se deseja extrair deles. Uma outra abordagem para atacar problemas de visão computacional se desenvolveu com *Deep Learning* [25]. Dentro das arquiteturas de redes neurais disponíveis, as Redes Neurais Convolucionais (CNNs) obtiveram sucesso no tratamento de tarefas referentes à visão computacional. Moustafa *et al.* [33] foi um dos primeiros a aplicar *Deep Learning* à detecção de pornografia em imagens. Nian *et al.* [34] apresentaram um método rápido de detecção de imagens pornográficas que utiliza uma abordagem de janelas deslizantes. Wang *et al.* [52] apresentaram uma definição de imagem pornográfica caracterizada como a exposição de partes íntimas do corpo. Como estas partes posicionam-se em regiões específicas, o trabalho modelou cada imagem como um *bag* de fragmentos de instâncias e assumiu que, para cada imagem pornográfica pelo menos uma instância representa o conteúdo pornográfico dentro dele. Li *et al.* [26] propuseram uma arquitetura construída sobre CNNs que aceita imagens de entrada de diferentes tamanhos e incorpora características extraídas em diferentes níveis hierárquicos da rede para realizar a previsão.

Como último passo, após extração de características de baixo e de médio nível efetua-se a classificação supervisionada. Neste momento, utilizam-se algoritmos de aprendizado de máquina para efetuar a classificação das imagens. Para isso, um algoritmo de aprendizado de máquina encontra uma função preditiva usando um conjunto de exemplos usualmente chamado *conjunto de treinamento*. Visando dimensionar a performance de algoritmos desse tipo, Frondana [14] testou 16 algoritmos em 59 bases de dados reais visando encontrar quais geram melhores resultados de regressão. Entre os algoritmos testados estavam *Support Vector Machines* (SVMs) [49] e *k-Nearest Neighbor* [16] que, entre outros, são comumente utilizados em problemas de classificação de imagens. Quando usamos *Deep*

Learning, a rede montada pode incluir uma camada de classificação que, quando treinada com sucesso, pode substituir um classificador externo.

2.2 Detecção de Pornografia em Vídeo

Grande parte das técnicas usadas para detecção de pornografia em vídeo vêm do uso de técnicas desenvolvidas para imagens. A metodologia básica consiste na extração de quadros a uma determinada taxa seguida do processamento de cada quadro usando as técnicas já desenvolvidas para imagens e uma posterior agregação dos resultados para definição da resposta global de detecção do vídeo. A classificação de conteúdo pornográfico baseado em detecção de pele também foi explorada em trabalhos que abordaram vídeo. Polastro e Eleuterio [37] propuseram uma detecção estatística da quantidade de pele identificada nos quadros para uma classificação da presença de nudez no vídeo. A principal contribuição deste trabalho foi a eficiência na classificação a fim de permitir o uso em possíveis cenas de crimes. Uke e Thool [47] procuraram aliar a detecção de nudez à extração e processamento de informações de áudio.

Usando técnicas baseadas em BoVW, Ulges *et al.* [48] estudaram um descritor de atividade baseado em vetores de compensação de movimento MPEG. O trabalho concluiu que é necessário o processamento de múltiplas informações do *streaming* de vídeo para compensar as limitações de cada técnica individual. Liu *et al.* [28, 29] incluíram informações de áudio para complementar a análise baseada em BoVW. Avila *et al.* [2] estenderam o formalismo do BoVW introduzindo uma estratégia de agregação baseada na função densidade em lugar das estratégias clássicas de agregação de soma ou de máximo. Esta mudança permitiu representar melhor a ligação entre o dicionário de palavras e os descritores locais.

Usando *Deep Learning*, Moustafa [33] propôs um método que processa os quadros dos vídeos com duas CNNs diferentes, criando uma fusão dos resultados (*late fusion*) para a classificação final. Também usando CNNs, Perez *et al.* [35] propuseram diferentes maneiras para combinar informações estáticas dos quadros e informações dinâmicas, extraídas usando a variação de fluxo óptico entre quadros adjacentes, obtendo os melhores resultados conhecidos em detecção de pornografia em vídeo.

2.3 Localização de Pornografia em Vídeo

O problema de localização de pornografia em vídeo tem, até onde se pode verificar, um único trabalho na literatura. Moreira [31] desenvolveu a tarefa de localização de conteúdo aplicado aos problemas de pornografia e de violência. Ele percebeu que em problemas de classificação de vídeos, a técnica BoVW gera uma classificação para o vídeo inteiro. Ao separar o vídeo em partes de interesse e estabelecendo *bags* para cada uma das partes, isto faria a técnica de BoVW diretamente aplicável ao problema da localização de vídeo. Para a extração das características dos fragmentos de vídeos foram empregados os descritores de áudio *Mel-Frequency Cepstral Coefficients* (MFCC) e *Prosodic Features* (PROS), o descritor de imagens *Histograms of Oriented Gradients* (HOG), e o descritor de características temporais *Temporal Robust Features* (TRoF) foi desenvolvido. Em seu trabalho são testados três métodos de aprendizagem para a classificação: *Score Thresholding* (THR), *Naïve Bayes Classifier* (NBC) e *Support Vector Machines* (SVM). O trabalho obteve bons resultados na manipulação de dados temporais e na tarefa de classificação de conteúdo pornográfico. Entretanto, os resultados da tarefa de localização (acurácia de 90.7%) não alcançaram a mesma performance da classificação (acurácia de 95.8%). Além disso, o descritor TRoF pode ser dispendioso para processamento em tempo real. O trabalho também introduziu a base de dados *Pornography-2k*, que será usada neste projeto e é descrita na Subseção 4.1.

2.4 *Deep Learning* em Tarefas de Processamento de Vídeo

Quando pesquisamos outras tarefas de processamento de vídeo não relacionadas à pornografia como, por exemplo, reconhecimento de atividade humana e geração automática de legendas, vemos que abordagens baseadas em *Deep Learning* têm sido usadas em variadas atividades, produzindo bons resultados [22, 40, 46, 51]. Abordaremos com mais detalhes arquiteturas de redes neurais como CNN, RNN e LSTM na Seção 3, entretanto descrevemos aqui alguns trabalhos.

Karpathy *et al.* [22] usaram uma rede CNN-3D para processar grandes conjuntos de treinamento como o *Sport 1 Million* (introduzido por eles mesmos) e o *UCF-101* [42]. Seus resultados apresentaram melhora significativa comparados a trabalhos baseados em descritores, mas apenas uma melhora modesta quando comparados a CNNs comuns. Si-

monyan e Zisserman [39] apresentaram uma sequência de processamento onde duas CNNs independentemente treinadas processam informações diferentes extraídas do mesmo vídeo. Uma CNN é treinada com fluxo óptico [8, 36] e tenta aprender características ligadas a informações de movimento. A outra CNN é treinada com a entrada RGB dos quadros extraídos. Após o processamento nas duas CNNs, os resultados são agregados para gerar uma classificação global para o vídeo. Donahue *et al.* [11] aplicaram um tipo de RNN sobre uma estrutura parecida com a proposta por Simonyan e Zisserman [39] e, com isso, conseguiram abordar tarefas de geração de legenda, reconhecimento de ações e descrição de vídeo. Ballas *et al.* [4] extraíram entradas de diferentes camadas de uma CNN comum e aplicaram como entrada para uma outra configuração de RNN. Esta coleta de informações de várias camadas permitiu fornecer à parte recorrente da rede características de baixo e médio nível, além das características de alto nível que são a saída comum dos últimos níveis da CNN utilizada.

3 Conceitos Relacionados

LeCun *et al.* [25] descrevem os métodos de *Deep Learning* como uma rede ou conjunto de métodos de aprendizagem de representação dos dados com múltiplos níveis. Essa rede é obtida pelo empilhamento ou sequenciamento de módulos simples, porém não lineares. No processamento de uma imagem, geralmente temos como entrada o conjunto de *pixels*. As características aprendidas na primeira camada de representação usualmente retratam a presença ou ausência de arestas (Figura 1a). A segunda camada comumente realça arranjos particulares de arestas (Figura 1b). A terceira camada pode agregar os arranjos anteriores e descrever partes de objetos (Figura 1c). Até que camadas superiores possam detectar objetos como a combinação dessas partes.

A aprendizagem ocorre durante o treinamento da rede. O método mais comum nesta classe de problemas é uma aprendizagem supervisionada. Durante o treinamento os dados e os rótulos são apresentados à rede. Através do algoritmo de *back-propagation*, o erro entre a resposta encontrada e a esperada é usado para modificação dos pesos das conexões entre os neurônios da rede. Essa mudança é proporcional à contribuição para o erro verificado. Esse processo é realizado com uma grande quantidade de dados e tem seu fim quando o erro para a classificação no conjunto de treinamento encontra-se abaixo de um limiar ou

(a) (b) (c)

Figura 1: Exemplo de extração hierárquica de características em três camadas de rede neural[2].

[2]https://devblogs.nvidia.com/wp-content/uploads/2015/11/hierarchical_features.png

um número máximo de execuções é alcançado. A verificação dos resultados é realizada em um conjunto de teste que contém dados que não fizeram parte do treinamento.

Quando deseja-se realizar o treinamento da rede mas não existe base de dados suficiente para alcançar boa acurácia, é possível construir um conhecimento novo sobre uma base de dados que já foi utilizada anteriormente para outro propósito, possivelmente mais geral. Partindo-se de uma rede já treinada para um problema (mesmo que com outras classes de treinamento), efetua-se o *fine-tuning* das últimas camadas da rede usando os dados do problema em foco e as classes requeridas como saída.

Em seguida descrevemos o funcionamento das CNNs na Subseção 3.1 e a arquitetura de RNNs e suas evoluções na Subseção 3.2.

3.1 Redes Neurais Convolucionais (CNNs)

CNNs foram projetadas com o objetivo de emular o córtex visual humano [24]. Elas reforçam a correlação espacial através da conexão esparsa de neurônios de camadas adjacentes. Então, por construção, uma camada captura informação de pequenas regiões da imagem (campos receptivos) e a camada subsequente processa uma região mais ampla. Dessa forma, a pilha de camadas pode extrair desde informação geral como cor e brilho até informações mais elaboradas como cantos e bordas. A profundidade da pilha pode permitir o aprendizado de conceitos ainda mais complexos, como partes do corpo humano [55]. A Figura 2 apresenta uma configuração simplificada de exemplo para uma CNN.

As camadas da rede podem ser repetidas e conectadas de diversas formas. Algumas

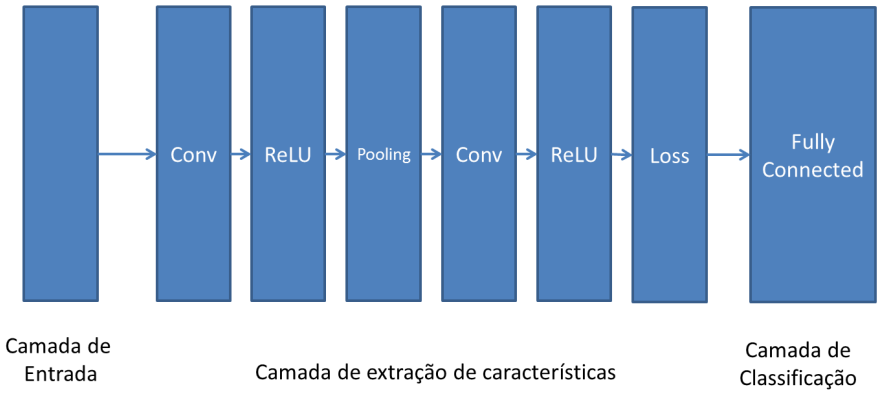


Figura 2: Exemplo de configuração de uma CNN. As camadas são conectadas gerando um caminho por onde as representações dos dados trafegam até culminar na classificação realizada pela última camada.

das camadas usualmente incluídas em CNNs são:

- Camada de convolução: A convolução é a operação principal de uma CNN. Os parâmetros (por exemplo, máscara de convolução e passo) constituem um conjunto de filtros (ou *kernels*) que podem ser aprendidos.
- Camada de agregação (*Pooling*): É uma forma não linear de redução de amostragem.
- Camada ReLU (*Rectified Linear Units*): É capaz de aumentar as propriedades não lineares da função de decisão e da rede como um todo sem afetar os campos receptivos das camadas de convolução.
- Camada de perda (*Loss*): Esta camada especifica como a rede penaliza o desvio entre as classes preditas e realmente computadas pela rede.
- Camada totalmente conectada (*Fully Connected*): Normalmente usadas nas camadas finais da rede, cada um de seus neurônios tem conexão com todas as ativações da camada anterior. A classificação é realizada nesta camada.

Além da rede CNN comum (2D), que tem duas dimensões na camada de entrada referentes às coordenadas cartesianas x e y das imagens, podemos encontrar a CNN-3D que é relacionada a uma quantidade específica e finita de quadros contíguos que podem ser processados. A fim de capturar informações temporais presentes em vídeo, Tran *et al.* [46] construíram uma CNN-3D. Em seu processamento é possível perceber que a rede foca

num objeto de interesse e, em seguida, percebe os movimentos que ocorrem próximos ao ponto focal. Com isso ele obteve bons resultados em tarefas de classificação de atividades esportivas, reconhecimento de ações, identificação de similaridade de ação e reconhecimento de objetos e cenas. Karpathy *et al.* [22] também obtiveram bons resultados em tarefas de reconhecimento de atividades em vídeo usando essa arquitetura.

3.2 Redes Neurais Recorrentes (RNNs)

Bengio *et al.* [6] relatam que uma rede recorrente tem ciclos em seu grafo que lhe permitem guardar informação sobre entradas passadas por um tempo não previamente fixado, dependendo de seus pesos e dos dados de entrada. Em contraste, redes estáticas (não recorrentes) não conseguem guardar informação por um tempo indefinido. RNNs apresentaram sucesso em tarefas como geração de texto [44] e reconhecimento de fala [50]. Entretanto, foi observado que capturar dependências em um longo intervalo era difícil, tornando a tarefa de minimização do erro no treinamento quase impossível para algumas tarefas quando o horizonte temporal cresce suficientemente [7, 17].

Visando solucionar este problema, Hochreiter e Schmidhuber [17] introduziram modificações em redes recorrentes criando a arquitetura de rede Memória de Curto e Longo Prazo (LSTM, do inglês *Long Short-Term Memory*). Essa rede pode ser treinada para aprender ou para esquecer memórias prévias, considerando a entrada corrente e o estado atual. É possível também aprender quanto de memória deveria passar aos estados escondidos do próximo módulo LSTM. Isto permite que LSTM aprenda sequências temporais longas e complexas para problemas de processamento de vídeo como reconhecimento de atividades [53, 54], descrição de conteúdo [11] e criação de legendas [11].

Em resumo, a comparação de CNNs e RNNs evidencia diferentes características que podem ser utilizadas para abordar o problema de localização de pornografia. Enquanto CNNs têm capacidade especializada para tratar problemas de processamento de imagens, com possível modificação para tratar de uma sequência limitada de quadros, RNNs e suas variações têm a capacidade de processar sequências longas, permitindo que se construa conhecimento ao processar séries de tamanho indeterminado. A natureza diferente, e possivelmente complementar, dessas redes poderá permitir a construção de pipelines com uma composição de ambas [11, 54].

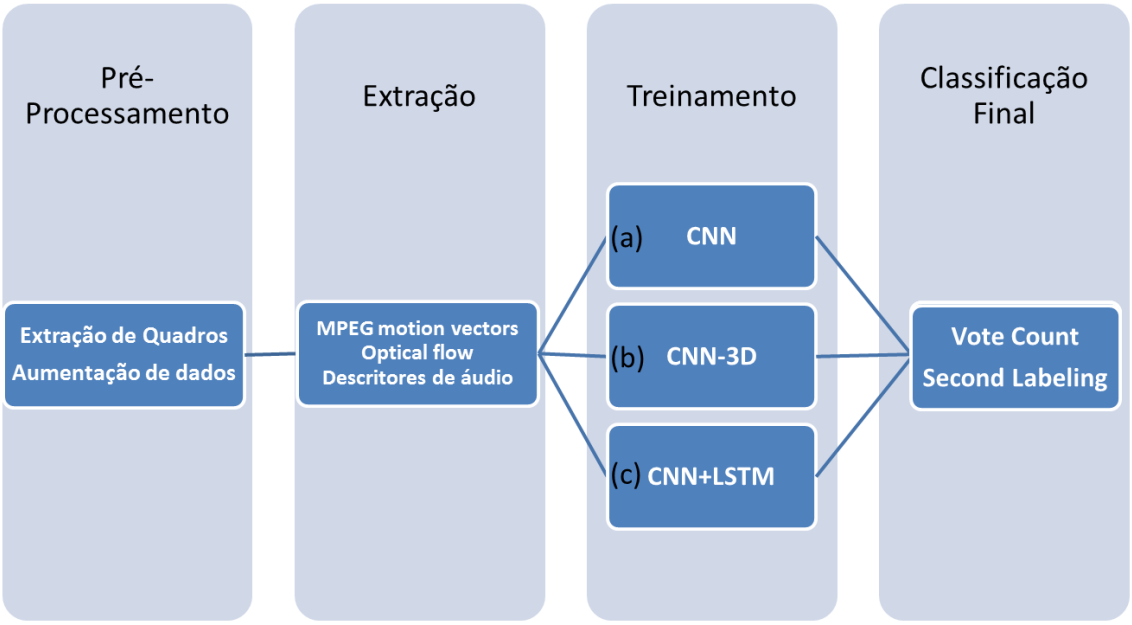


Figura 3: Três possíveis abordagens: a) utilizando uma CNN estática (*baseline*); b) utilizando uma CNN-3D e c) utilizando uma combinação de CNN e LSTM.

4 Metodologia Proposta

Durante o desenvolvimento deste projeto, diferentes abordagens podem ser avaliadas. As propostas iniciais estão retratadas na Figura 3. Em cada uma das possíveis abordagens existem etapas similares que são descritas a seguir.

Pré-processamento: Nesta etapa realizaremos a extração de quadros dos vídeos para submeter à rede neural. Podemos submeter todos os quadros do vídeo à rede ou extrair uma quantidade menor de quadros. A quantidade de quadros extraídos a cada segundo define a taxa de amostragem. Pretendemos encontrar experimentalmente uma taxa de amostragem que traga uma boa relação de compromisso entre os resultados e a quantidade de processamento requerido. Ainda nessa etapa, podemos realizar a **aumentação de dados**, uma forma de ampliação da base de dados usando um conjunto de transformações geométricas (por exemplo, escala, rotação e recorte) ou fotométricas (por exemplo, contraste e brilho) realizadas sobre os quadros e preservando a marcação temporal (rótulos) do vídeo original.

Extração: Nesta etapa podemos extrair informações a partir dos quadros ou do vídeo (por exemplo, MPEG *motion vectors*, fluxos ópticos ou características de áudio) com o objetivo de fornecer informações complementares às informações visuais obtidas

pela rede através do processamento dos quadros. Será necessário avaliar a utilidade do uso de informações de áudio tendo em vista que a marcação da base de dados foi realizada apenas para informações visuais, não existindo marcação da informação de áudio para conteúdo pornográfico.

Treinamento: Nesta etapa podem ser utilizadas diferentes configurações de redes neurais profundas. O *baseline* deve ser a utilização de uma CNN estática para que seja possível mensurar os ganhos referentes à captura e processamento das características temporais dos vídeos. Para as outras abordagens, pretendemos avaliar variações de redes CNN-3D e CNN em conjunto com LSTM. Devido ao limite temporal da entrada de dados em redes CNN-3D, é esperado que ela consiga extrair informações temporais locais. As redes LSTM por outro lado podem, em tese, indicar o quanto informações temporais globais podem auxiliar na tarefa de localização. Cada uma das redes indicadas (CNN, CNN-3D e LSTM) podem ter variações em sua arquitetura e, por isso, a investigação da opção mais adequada para a tarefa de localização será necessária. Além disso, é possível avaliar a complementariedade de uma possível fusão das arquiteturas CNN-3D e CNN+LSTM, verificando se seu uso em único pipeline pode trazer ganhos aos resultados da localização.

Classificação Final: Devido à característica de processamento em janela temporal da CNN-3D, é possível efetuar a classificação através de uma votação, onde cada segundo analisado possui a quantidade de votos que a profundidade temporal da CNN-3D apresentar. Para a rede estática é possível extrair apenas um quadro intermediário para a representação do segundo analisado ou também pode-se utilizar uma amostragem de quadros e realizar uma agregação ou votação para definir a classe final para aquele segundo.

4.1 Base de Dados

Para validar os métodos propostos, consideraremos a base de dados *Pornography-2k* proposta por Moreira *et al.* [32], que é uma extensão do conjunto de vídeos originalmente proposto por Avila *et al.* [3]. O *dataset* possui cerca de 140 horas em 1000 vídeos pornográficos e 1000 vídeos não pornográficos, cuja duração varia entre 6 segundos e 33 minutos. Os vídeos com conteúdo pornográfico somam 99h32min, sendo que deste tempo 91h43min são referentes a conteúdo pornográfico. A aquisição de conteúdo foi realizada em sites de propósito geral e também em sites especializados em pornografia de forma a contemplar

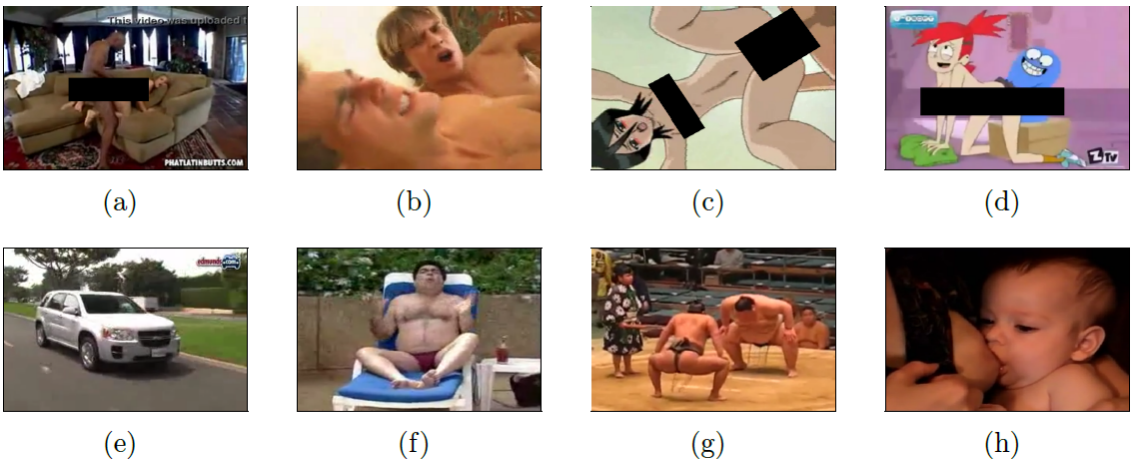


Figura 4: Amostras de quadros da base de dados *Pornography-2k*. Os quadros da linha superior representam conteúdo sensível. Na linha de baixo temos conteúdo não pornográfico, enfatizando exemplos com exposição de pele. Imagem extraída de Moreira *et al.* [32].

vários gêneros pornográficos, incluindo diversidade étnica e de comportamento. Entre o conteúdo não pornográfico, além de amostras que não possuem ambiguidade de classificação, existem amostras coletadas buscando-se por conteúdo associado à exposição de pele (por exemplo, lutas, natação e praia) e que pode causar mais dificuldade na classificação. Os vídeos que contêm conteúdo pornográfico foram rotulados quadro a quadro de forma a ser possível realizar a localização com granularidade ou segmentação a nível de quadros. A Figura 4 mostra exemplos de quadros de vídeos presentes na base de dados.

4.2 Validação

Para validação da abordagem proposta, em princípio, serão utilizadas as mesmas métricas referentes à tarefa de localização definidas e utilizadas em Moreira *et al.* [32]. Entretanto, com o andamento do projeto é possível a definição ou uso de métricas adicionais.

As métricas usadas em Moreira *et al.* [32] (Taxa de Verdadeiros Positivos (TVP) e Taxa de Verdadeiros Negativos (TVN)) serão coletadas para cada segundo dos vídeos permitindo verificar a correção das respostas a cada segundo. As métricas Acurácia normalizada (ACC) e medida F_2 (Equação 1, com $\beta = 2$) serão apresentadas como médias de todos os segundos que compõem todos os vídeos analisados.

$$F_\beta = (1 + \beta^2) \times \frac{precisão \times revocacão}{\beta^2 \times precisão + revocacão} \tag{1}$$

4.3 Cronograma

O cronograma apresentado na Tabela 1 pode ser adaptado, tendo em vista que novas percepções e oportunidades podem trazer resultados melhores que os previamente previstos.

Tabela 1: Cronograma de pesquisa proposto para o projeto.

Cronograma	**2016**		**2017**				**2018**	
Atividades \ Trimestres	3	4	1	2	3	4	1	2
Convalidação de créditos obrigatórios[3]	•							
Revisão bibliográfica		•	•	•	•	•	•	
Escrita da proposta de mestrado		•	•					
Avaliação de arquiteturas disponíveis	•	•	•	•				
Exame de Qualificação de Mestrado (EQM)				•				
Experimentos				•	•	•	•	
Escrita da dissertação					•	•	•	•
Construção e avaliação de arquiteturas propostas					•	•	•	•
Publicação de resultados						•	•	•

[3] Os créditos obrigatórios ao programa foram cursados antes da realização da matrícula como aluno regular.

Referências

[1] Sandra Avila. *Extended Bag-of-Words Formalism for Image Classification.* PhD thesis, Federal University of Minas Gerais and Pierre and Marie Curie University, 2013. 3

[2] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de A. Araújo. Bossa: Extended bow formalism for image classification. In *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP'2011)*, pages 2909–2912, 2011. 5

[3] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de A. Araújo. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, 117(5):453–465, 2013. 1, 12

[4] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv:1511.06432*, 2015. 7

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *Proceedings of the European Conference on Computer Vision (ECCV'2006)*, pages 404–417. Springer, 2006. 3

[6] Yoshua Bengio, Paolo Frasconi, and Patrice Simard. The problem of learning long-term dependencies in recurrent networks. In *Proceedings of the IEEE International Conference on Neural Networks (IJCNN'1993)*, pages 1183–1188. IEEE, 1993. 2, 10

[7] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994. 10

[8] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, 2011. 7

[9] Li Deng, Dong Yu, *et al.* Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014. 2

[10] Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. Bag-of-visual-words models for adult image classification and filtering. In *Proceedings of the 19th International Conference on Pattern Recognition (ICPR'2008)*, pages 1–4, 2008. 4

[11] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2015)*, pages 2625–2634, 2015. 7, 10

[12] Margaret M. Fleck, David A. Forsyth, and Chris Bregler. Finding naked people. In *Proceedings of the European Conference on Computer Vision (ECCV'1996)*, volume 1065, pages 593–602, 1996. 2, 3

[13] David A. Forsyth and Margaret M. Fleck. Automatic Detection of Human Nudes. *International Journal on Computer Vision*, 32(1):63–77, 1999. 2, 3

[14] Giovani Frondana. Empirical Comparison of 16 Regression Algorithms in 59 Datasets. Master's thesis, Unicamp, 2017. 4

[15] Klaus Greff, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 2016. 2

[16] Peter E. Hart, David G. Stork, and Richard O. Duda. Pattern classification. *John Willey & Sons*, 2001. 4

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. 2, 10

[18] Weiming Hu, Haiqiang Zuo, Ou Wu, Yunfei Chen, Zhongfei Zhang, and David Suter. Recognition of adult images, videos, and web page bags. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 7S(1):1–24, 2011. 1

[19] Christian Jansohn, Adrian Ulges, and Thomas M. Breuel. Detecting pornographic video content by combining image features with motion information. In *Proceedings of the 17th ACM International Conference on Multimedia - (ACMMM'2009)*, page 601, 2009. 1

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. 2

[21] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002. 2, 3

[22] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with Convolutional Neural Networks. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR'2014)*, pages 1725–1732, 2014. 2, 6, 10

[23] Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 2

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 8

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, 2015. 2, 4, 7

[26] Kai Li, Junliang Xing, Bing Li, and Weiming Hu. Bootstrapping deep feature hierarchy for pornographic image recognition. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'2016)*, pages 4423–4427. IEEE, 2016. 1, 4

[27] Bei-bei Liu, Jing-yong Su, Zhe-ming Lu, and Zhen Li. Pornographic Images Detection Based on CBIR and Skin Analysis. In *Proceedings of the 4th International Conference on Semantics, Knowledge and Grid (SKG'2008)*, pages 487–488, 2008. 1

[28] Yizhi Liu, Xiangdong Wang, Yongdong Zhang, and Sheng Tang. Fusing audio-words with visual features for pornographic video detection. In *Proceedings of the 10th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom'2011)*, pages 1488–1493. IEEE, 2011. 5

[29] Yizhi Liu, Ying Yang, Hongtao Xie, and Sheng Tang. Fusing audio vocabulary with visual features for pornographic video detection. *Future Generation Computer Systems*, 31:69–76, 2014. 5

[30] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'1999)*, volume 2, pages 1150–1157, 1999. 3

[31] Daniel Moreira. *Sensitive-Video Analysis*. PhD thesis, Unicamp, 2016. 1, 2, 6

[32] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Pornography classification: The hidden clues in video space–time. *Forensic Science International*, 268:46–61, 2016. 12, 13

[33] Mohamed Moustafa. Applying deep learning to classify pornographic images and videos. *arXiv:1511.08899*, 2015. 2, 4, 5

[34] Fudong Nian, Teng Li, Yan Wang, Mingliang Xu, and Jun Wu. Pornographic image detection utilizing deep convolutional neural networks. *Neurocomputing*, 210:283–293, 2016. 4

[35] Mauricio Perez, Sandra Avila, Daniel Moreira, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, 230: 279–293, 2017. 2, 5

[36] Aurélien Plyer, Guy Le Besnerais, and Frédéric Champagnat. Massively parallel Lucas Kanade optical flow for real-time video processing applications. *Journal of Real-Time Image Processing*, 11(4):713–730, 2016. 7

[37] Mateus de C. Polastro and Pedro M. da S. Eleuterio. A statistical approach for identifying videos of child pornography at crime scenes. In *Proceedings of the 7th International Conference on Availability, Reliability and Security (ARES'2012)*, pages 604–612, 2012. 5

[38] Mateus de C. Polastro, Pedro M. da S. Eleuterio, and Pedro Monteiro. Quick identification of child pornography in digital videos. *The International Journal of Forensic Computer Science*, 2:21–32, 2012. 1

[39] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS'2014)*, pages 568–576, 2014. 7

[40] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*, 2014. 6

[41] Josef Sivic and Andrew Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings 9th IEEE International Conference on Computer Vision (ICCV'2003)*, pages 1470–1477 vol.2, 2003. 1, 4

[42] Khurram Soomro, Amir R. Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. 6

[43] Chad M. S. Steel. The Mask-SIFT cascading classifier for pornography detection. In *Proceedings of the World Congress on Internet Security (WorldCIS'2012)*, pages 139–142, 2012. 4

[44] Ilya Sutskever, James Martens, and Geoffrey E. Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML'2011)*, pages 1017–1024, 2011. 10

[45] Graham W. Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *Proceedings of the 11th European Conference on Computer Vision (ECCV'2010)*, pages 140–153, 2010. 2

[46] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'2015)*, pages 4489–4497, 2015. 6, 9

[47] Nilesh J. Uke and Ravindra C. Thool. Detecting pornography on web to prevent child abuse–a computer vision approach. *International Journal of Scientific and Engineering Research*, 3(4):1–3, 2012. 5

[48] Adrian Ulges, Christian Schulze, Damian Borth, and Armin Stahl. Pornography detection in video benefits (a lot) from a multi-modal approach. In *Proceedings of the 2012 ACM International Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis (AMVA'2012)*, pages 21–26. ACM, 2012. 5

[49] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical Learning Theory*, volume 1. Wiley New York, 1998. 4

[50] Oriol Vinyals, Suman V. Ravuri, and Daniel Povey. Revisiting recurrent neural networks for robust asr. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2012)*, pages 4085–4088. IEEE, 2012. 10

[51] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *arXiv:1507.02159*, 2015. 6

[52] Yuhui Wang, Xin Jin, and Xiaoyang Tan. Pornographic image recognition by strongly-supervised deep multiple instance learning. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'2016)*, pages 4418–4422. IEEE, 2016. 4

[53] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM International Conference on Multimedia (ACMMM'2015)*, pages 461–470. ACM, 2015. 10

[54] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2015)*, pages 4694–4702, 2015. 10

[55] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV'2014)*, pages 818–833, 2014. 8