Detecção de Mídias Pornográficas em Dispositivos com Recursos Limitados para Controle Parental

Jhonatan Geremias[1], **Eduardo K. Viegas**[1], **Altair O. Santin**[1], **Jackson Mallmann**[1,2]

[1]Programa de Pós-Graduação em Informática (PPGIa)
Pontifícia Universidade Católica do Paraná (PUCPR)
80.215-901 - Curitiba - PR

[2]Instituto Federal Catarinense – Jardim Maluche
88.354-300 - Brusque - SC

Abstract. *Mobile devices are currently widely used by underaged persons. This kind of device has Internet access, allowing it to be used for streaming pornographic content. In light of this, this paper proposes a context-based approach for real-time detection of pornographic videos for parental control purposes. According to the video frames, motion-based descriptors are extracted and fed to a CNN model, providing resources to a shallow classifier. Experiments have shown the proposal feasibility, reaching 93.62% of accuracy while being executed in a resource-constrained device.*

Resumo. *Dispositivos móveis, atualmente, são amplamente utilizados por menores de idade. Este tipo de dispositivo possui acesso a Internet, permitindo assim o seu uso para a visualização de conteúdos pornográficos. Dado este contexto, este artigo propõe uma nova abordagem baseada em contexto para a detecção em tempo real de vídeos pornográficos para controle parental. A partir da sequência de frames de um vídeo, descritores de movimento extraem informação para alimentar um modelo de CNN, fornecendo subsídios para o classificador raso. Resultados experimentais demonstram que a abordagem proposta obteve 93,62% de acurácia enquanto executada em dispositivo com recursos limitados.*

1. Introdução

Estudos apontam que uma quantia significativa de usuários utilizam dispositivos com recursos computacionais limitados para navegarem na Internet [Clement 2019]. Os objetivos são vários, e dentre eles, a visualização de vídeos pornográficos[1]. Entre os usuários, estão as crianças e adolescentes, que por questões legais[2], necessitam de atenção especial de seus responsáveis. Ou seja, monitorar/controlar crianças e adolescentes para que não acessem conteúdos impróprios. Isto pode ser realizado mediante uso de software de controle paterno[3], em que é necessário informar classificações estáticas. Entretanto, existem conteúdos na Internet sem nenhuma classificação [Amato et al. 2009].

[1]Enough is Enough - https://enough.org/stats_porn_industry

[2]Questões legais - Prevista no estatuto da criança e do adolescente, BRASIL. Lei n.° 8.069, de 13 de julho de 1990.

[3]Controle paterno - Mecanismo utilizado pelos pais/responsáveis para controlar o acesso que as crianças podem ter na Internet.

Muitos vídeos pornográficos são alocados na Internet sem que haja classificação correta[4]. Para sanar este problema, o software de controle paterno deve realizar a classificação automática de vídeos, assim crianças e adolescentes não seriam expostos a conteúdos pornográficos. Na literatura, este tipo de classificação tem se utilizado de diversas abordagens [Ji et al. 2013][Horchulhack et al. 2024b][dos Santos et al. 2021]. Destaca-se as Redes Neurais Convolucionais (CNN), que tem comprovado sua eficiência principalmente em termos de acurácia. Entretanto, é uma tarefa complexa que demanda recursos computacionais significativos [Viegas et al. 2020].

Este artigo propõe uma abordagem baseada no uso de CNN. Combinam-se diferentes características de movimento para a detecção em tempo real de imagem pornográfica em vídeo. A abordagem é direcionada a dispositivos com recursos computacionais limitados. Primeiramente, extraímos as caracteríticas de informações de movimento que são obtidas entre dois frames adjacentes. Nesta extração são utilizadas técnicas de fluxo óptico e mapas de similaridade estrutural. As informações de movimento entre os frames permitem determinar a direção do movimento, região de deslocamento e similaridade entre os frames. Como resultado, a abordagem proposta fornece características adicionais para a criação de modelos para classificação do contexto da cena do vídeo, levando em consideração informações de movimento entre frames adjacentes.

Em resumo, entre as principais contribuições desse trabalho destacam-se:

- Caracterização do problema de contexto em vídeos. Na detecção de conteúdo pornográfico em vídeo muitas vezes as características contidas em um único frame não são suficientes para determinar seu contexto. Por outro lado, uma sequência de frames de um vídeo pode conferir informações adjacentes para interpretação e análise do vídeo;
- Propomos uma nova representação baseada na informação de movimento para extrair o contexto dos frames de vídeos em tempo real. Nossa abordagem extraí as características de movimento de frames adjacentes baseadas em fluxo óptico;
- Propomos uma abordagem de baixo custo computacional para detectar em tempo real pornografia em vídeos, baseada na utilização de um classificador raso. Abordagem projetada para auxiliar softwares de controle paterno destinada a dispositivos com limitações de recursos computacionais.

O artigo está organizado da seguinte forma. A Seção 2 descreve a fundamentação teórica. A Seção 3 discute os trabalhos relacionados. Na Seção 4 apresenta-se a proposta. Na Seção 5 são abordados os aspectos técnicos que proporcionam a reprodutibilidade do trabalho, assim como o dataset formalizado, os resultados, e a discussão. Finalmente, na Seção 6 apresenta-se as conclusões.

2. Estado da Arte

Nesta seção apresentam-se conceitos teóricos que são aplicados na proposta. Dentre eles, CNN e a classificação de vídeos.

2.1. Redes Neurais Convolucionais

As Redes Neurais Convolucionais (CNNs) têm proporcionado resultados impressionantes em várias áreas, como em reconhecimento de imagem e detecção de obje-

[4]Classificação correta - Classificação do teor do conteúdo do vídeo por idade.

tos [Katta et al. 2022] [dos Santos et al. 2023]. Elas visam aprender a classificar um problema, ou seja, descobrir um modelo. Para isto, utiliza-se de uma grande quantidade de amostras, que em nosso caso, são frames extraídos de vídeos. As CNNs são constituídas por camadas: *Convolutional*, *Pooling* e *Fully-Connected Layers* [Szegedy et al. 2015]. Sendo que, a estruturação e interposição das camadas é denominado de arquitetura CNN [Simonyan and Zisserman 2014, Horchulhack et al. 2024a].

A sua execução ocorre em duas etapas: treinamento e teste [Gu et al. 2018]. A diminuição do tempo de processamento é obtido com a adoção das GPUs (*Graphics Processing Unit*). Contudo, existem situações onde não se dispõem de tal recurso ou o recurso computacional é limitado. Por exemplo, alguns smartphones e tablets que não possuem recursos suficientes para executar uma arquitetura CNN tradicional. Diante disto, alguns autores propuseram arquiteturas CNN que exigem pouco recurso computacional. Por exemplo, o Mobilenet [Howard et al. 2017] é uma proposta do Google. Este aplica funções convolucionais em profundidade e separáveis para reduzir o número de parâmetros usados, diminuindo também a necessidade do uso da memória. Da mesma forma, outra arquitetura CNN, denominada Squeezenet [Iandola et al. 2017], também visa diminuir as necessidades de memória, reduzindo o número de parâmetros necessários.

2.2. Classificação de Vídeos

De maneira geral, as tarefas que envolvem a classificação de vídeo possuem alguns aspectos fundamentais, tais como: resolução do vídeo, espaço de cor e taxa de frames [Kuroki et al. 2007]. A resolução é uma característica que define as dimensões do vídeo, partindo de um plano cartesiano onde pixels são estruturados. O espaço de cor refere-se à quantidade de bits necessários para representar cada uma das cores básicas no espaço de cor. Por padrão as CNNs comumente utilizam o espaço de cor RGB (*Red*, *Green* e *Blue*). E por fim, a taxa de frames que corresponde ao número de frames que são exibidos por segundo. É necessário uma taxa de 23,97 FPS (*Frames Per Second*) para se adequar a percepção humana [Geremias et al. 2022] .

Tais características são abstraídas pela CNN [Perez et al. 2017], que recebem como entrada os frames dos vídeos como se fossem imagens. Portanto, para permitir o processamento computacional dos vídeos utilizando a CNN, torna-se necessário a extração dos frames. Na extração de frames, o vídeo é dividido em um conjunto de frames. O número de frames é definido de acordo com o processo de codificação do vídeo.

No processo de classificação do frame, cada frame extraído é classificado individualmente [Karpathy et al. 2014]. Em um cenário direcionado a classificação do conteúdo de vídeo em tempo real, os frames devem ser classificados na velocidade em que o conteúdo do vídeo é transmitido. Visando executar a tarefa de detecção de conteúdo de vídeo em dispositivos com recursos computacionais limitados, será necessário classificar cada frame individualmente, além de executar outras tarefas em paralelo, como reproduzir o vídeo ou executar outros aplicativos em segundo plano. Consequentemente, a tarefa de detecção de conteúdo de vídeo deve utilizar o mínimo de recursos possível.

3. Trabalhos Relacionados

Na literatura referente a detecção de conteúdo pornográfico em vídeos, uma das abordagens mais exploradas tem sido a análise de movimento, tal como disposto

em [Endeshaw et al. 2008] que extrai informações de movimentos mediante análise de frames adjacentes, gerando um vetor de movimentos para cada frame. Experimentalmente, os autores examinaram a frequência do conteúdo e então definiram a região de janelas baseando-se na frequência. Posteriormente, avaliaram o desempenho da detecção de movimento repetitivo. Finalmente, efetuaram a detecção de pornografia em vídeos a partir das características de movimento baseadas na correlação temporal.

Outro trabalho, focado na detecção de conteúdo pornográfico em vídeo propõem um método a partir da combinação de características visuais associadas a características de áudio [Rea et al. 2006]. Entre as características visuais, os autores utilizam a detecção de regiões de pele por meio de histograma de referência, obtendo indícios de conteúdo pornográfico. Entre as características visuais, foram utilizadas informações de movimento, advindas da extração dos vídeos MPEG e construindo vetores de movimento. Na sequência são utilizadas as características de áudio, tais como métrica de periodicidade, energia do sinal e autocorreção no áudio. Os autores ressaltam que a combinação de diferentes características visuais em combinação com características de áudio são ideais para serem aplicadas em abordagens de detecção de conteúdo pornográfico em tempo real, principalmente quando disponibilizadas em vídeo stream.

Em [Lee et al. 2009], os autores propuseram um sistema hierárquico multinível baseado em várias características de cor, texto e forma. Características extraídas de diferentes domínios temporais, organizadas para determinar se o conteúdo do vídeo é pornográfico. Os autores destacam a utilização do espaço de cor HSV como sendo um modelo de cor robusto quando aplicado no reconhecimento do componente de pele. Após a extração das características, um classificador SVM foi aplicado, possibilitando o desenvolvimento de um sistema para detecção de pornografia em tempo real.

Moreira et al. propõem análise do contexto de vídeos para detecção de pornografia, utilizando características de áudio e conteúdo visual associadas a abordagens espaço-temporal para classificação dos vídeos [Moreira et al. 2019]. Nesta abordagem, efetuam a fusão multimodal para detecção de cena com conteúdo sensível. Os autores efetuam a combinação de diferentes classificadores destinados a tarefas de classificação específicas, tais como a de frames estáticos, fluxo de áudio e movimento de vídeo.

Diversos autores propuseram a utilização de CNNs voltadas especificamente para a classificação de vídeos. Comumente tais abordagens analisam o fator do movimento, que é uma importante propriedade para diferenciar um frame de outro. Existem diferentes propostas, tais como em [Karpathy et al. 2014], onde os autores estendem as camadas convolucionais para trabalhar sob o domínio do tempo utilizando informações de espaço e tempo. Os autores utilizaram uma arquitetura CNN denominada Slow Fusion e a destacaram como uma rede bem representativa no reconhecimento do movimento.

Li et al. propõem uma representação do espaço temporal multigranular para reconhecimento de ações realizada em vídeos, abordagem que utiliza as convoluções da CNN em diferentes planos (2D e 3D) associadas ao LSTM para modelagem temporal [Li et al. 2016]. Entre as arquiteturas CNN utilizadas, os autores destacam o uso do VGG-19 e Alexnet. Para lidar com as características de movimento stream, os autores calcularam o fluxo óptico com a arquitetura VGG-19 em conjunto com o LSTM.

Outras abordagens da combinação da CNN com o

LSTM [Yue-Hei Ng et al. 2015] [Wu et al. 2015]. A arquitetura LSTM utiliza o conceito de célula de memória [Jones 2020]. Essa permite manter um valor por um curto ou longo prazo dependendo da sua importância. Devido tal características, as abordagens da combinação CNN com o LSTM utilizadas em tarefas voltadas para classificação de vídeos estão se tornando reincidentes na literatura, onde existem indícios de que a célula de memória do LSTM seja capaz de armazenar informações de movimento.

Existem abordagens que utilizam convoluções 3D para tratar o problema de contexto em vídeos tal como em [Ji et al. 2013]. Nessa abordagem os autores extraem características das dimensões espaciais e temporais a fim de executar convoluções 3D, capturando informações de movimento codificadas em múltiplos frames adjacentes. Tal abordagem realmente é eficaz para tratar o problema do contexto dos vídeos, entretanto é necessário ressaltar a complexidade e o custo computacional necessário para utilizar tal estratégia, o que acaba inviabilizando essa abordagem em dispositivos com limitação de recurso computacional.

Conforme os trabalhos dispostos, a utilização de arquiteturas CNN para tratar o problema de detecção de pornografia são essenciais. Tais abordagens mostram ser bastante efetivas. Porém para classificação de um vídeo, utilizar apenas a saída da CNN não é suficiente, pois para determinar o contexto do vídeo muitas vezes é necessário a analise de uma sequência de frames, nos quais estão inseridas informações de movimento. Assim, estratégias suplementares são necessárias. Descartamos a utilização de arquiteturas LSTM e convoluções 3D para estruturar nossa abordagem secundária, apesar tratar o problema do contexto em vídeo são arquiteturas complexas que exigem uma grande quantidade de recurso computacional, não sendo adequadas para serem utilizadas em dispositivos com restrição de recurso computacional. Visando tais limitações, a abordagem proposta foi estruturada de forma que a etapa secundária analíse as informações de movimento, mas não necessite de tanto recurso computacional.

4. Proposta

Nossa abordagem foi estruturada em três módulos: extrator de frames, extrator de movimento, e classificação de cena, conforme se mostra na Figura 1. Objetiva-se a detecção em tempo real de vídeo pornográfico, auxiliando softwares de controle paterno que são executados em dispositivos que possuam limitação de recursos computacionais, como smartphones e tablets.

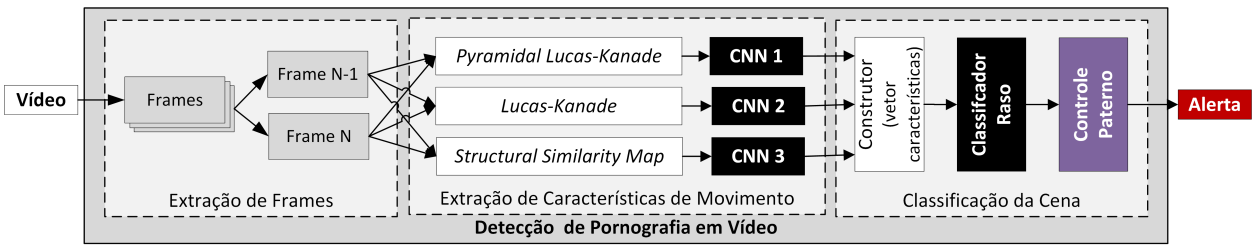

Figura 1. Abordagem proposta baseada em informação de movimento para detecção em tempo real de pornografia em vídeos sob granularidade fina (granularidade de frames).

Efetua-se a análise do contexto do vídeo levando em consideração uma granularidade em nível de frames, sendo uma abordagem de detecção sensível ao contexto

dos vídeos. Para isso estruturamos um modelo de detecção de pornografia baseada em informação de movimento. O objetivo é permitir a detecção em tempo real de conteúdo pornográfico dependente do contexto em vídeos, sem haver a necessitade de ter todos os frames do vídeo para o processo de classificação.

Após frames serem extraídos do vídeo, esses são organizados em sequência para aplicação de estimativas de movimento por meio de fluxo óptico e mapa de similaridade. E também, servem como entrada para um modelo CNN que analisa o contexto do frame que informa valores correspondentes a predição e probabilidade de cada uma das técnicas de movimento. Em seguida, estrutura-se um vetor de características baseado em informação de movimento que é responsável pela classificação.

Finalmente o software de controle paterno recebe a classificação da cena. Caso seja verificado a existência de conteúdo pornográfico na cena, realiza-se uma ação ou gera-se um alerta. Na sequência apresenta-se detalhadamente os três módulos.

4.1. Módulo Extrator de Frames

Este módulo recebe os frames de vídeos a medida que são capturados. Por exemplo, isto pode ocorrer em tempo real em transissões de vídeos, ou ainda como recurso para verificar o que está sendo gravado pela câmera de um celular.

Para efetuar a captura dos frames, rotinas implementadas na linguagem Python utilizando a biblioteca OpenCV na versão 2.4 [OpenCV 2020] foram utilizadas. Os frames são armazenados em uma janela deslizante, organizados em sequência a medida que são recebidos. Por sua vez, a janela deslizante possui um tamanho fixo e segue um comportamento de fila. Assim que um novo frame é capturado, a janela desloca um frame para a posição do frame corrente (Frame N, Figura 1).

O frame corrente é diponibilizado em conjunto com o frame antecessor (Frame N-1, Figura 1), armazenado na última posição da fila na janela de deslizante de frames. Permanecendo nessa posição até haver um novo deslocamente na janela de frames assumindo o papel do frame corrente. Nesse sentido, sempre que a janela de frames é deslocada, o frame corrente e o antecessor são repassados para o módulo de extração movimento.

(a) 1ºF **(b)** 2ºF **(c)** PLK **(d)** LK **(e)** SSM

Figura 2. Exemplo de frames de um vídeo pornográfico, e a aplicação de descritores de movimento relacionado entre o primeiro e o segundo frame do vídeo. (a) 1º frame, classe normal; (b) 2º frame, classe pornográfica; (c) Pyramidal Lucas-Kanade(PLK); (d) Lucas-Kanade(LK); (e) Structural Similarity Map (SSM).

Tabela 1. Desempenho da acurácia das arquiteturas CNN do estado da arte projetadas para dispositivos com recursos limitados sobre o dataset FPD. A acurácia normal e pornô denota a proporção de frames normais e pornográficos classificados corretamente como tais. A acurácia dependente de contexto indica a proporção de frames normais em vídeos pornográficos, classificados corretamente como frames normais.

CNN	Acurácia por Frame (%)			
	Normal	**Pornô**	**Dependente de Contexto**	**Geral**
Mobilenet	99,37	42,35	82,54	74,75
Resnet	99,49	30,21	80,56	70,09
Squeezenet	99,05	62,51	71,65	77,74

4.2. Módulo Extrator de Movimento

Este módulo efetua a extração da informação de movimento utilizando três descritores de movimento amplamente utilizados na literatura. Os fluxos ópticos obtidos com os algoritmos Lucas-Kanade [Lucas and Kanade 1981] e PLK (Pyramidal Lucas-Kanade) [Bouguet et al. 2001] e também um algoritmo de similaridade estrutural entre frames adjacentes, o SSM (Structural Similarity Map) [Zhang 2004].

Um exemplo da aplicação dos descritores baseados em movimento é mostrado na Figura 2. Os descritores de movimento são aplicados sobre dois frames: o atual e o anterior. Após a aplicação dos descritores de movimento são obtidas novas amostras de frames, sendo uma amostra para cada um dos descritores. As novas amostras de frames contém características adicionais de movimento. Características que fornecem uma noção da direção do movimento, região de alteração na cena e similaridade entre os frames adjacentes.

Por fim, as novas amostras de frames são encaminhadas para o modelo CNN treinado respectivamente para cada um dos descritores. Visando a aplicação em dispositivos com recursos limitados, nossa abordagem foi estruturada utilizando arquiteturas CNN específicas para os dispositivos alvo.

Foram realizados testes experimentais para selecionar a melhor arquitetura CNN para dispositivos móveis, conforme disposto na Tabela 1. Para realização dos experimentos, 60%, 20% e 20% dos vídeos respectivamente foram utilizados para fins de treinamento, validação e teste. As arquiteturas CNN utilizadas foram o Mobilenet [Howard et al. 2017], Squeezenet [Iandola et al. 2017] e Resnet [Niu et al. 2019]. No qual um modelo CNN foi previamente treinado para cada um dos descritores de movimento. Os modelos CNN classificam as novas amostras, onde são obtidos o valor da predição e a distribuição da probabilidade do frame avaliado ser pornográfico (Pornô) ou normal.

Posteriormente o resultado dos valores obtidos da CNN são repassados para o módulo de classificação da cena.

4.3. Módulo de Classificação de Cena

Por fim, apresenta-se o módulo responsável por produzir o resultado final da classificação do frame do vídeo. Isso ocorre mediante o recebimento do conjunto de classificações

individuais da CNN de cada um dos descritores de movimento.

O resultado é composto pelos rótulos e valores de confiança da distribuição das classes (normal e pornô) que foram atribuídos pela CNN. Consequentemente, cada descritor de movimento utilizado produz três valores, que são usados como atributos para compor um único vetor de características para cada frame do vídeo. Este vetor de caracteristisca é estruturado no construtor disposto no módulo de classificação de cena (Construtor, Figura,1).

O vetor de características construído é encaminhado para um classificador raso aplicado na plataforma de mineração de dados WEKA[5]. Assim, realiza-se a classificação final, onde é determinado se o frame é normal ou pornográfico.

Caso o frame seja caracterizado como pornográfico, o software de controle paterno é acionado e pode gerar um alerta ou realizar determinada ação, e.g., enviar uma mensagem aos ”pais/responsáveis”, fechar determinados aplicativos, desabilitar recursos entre outras. As ações do controle paterno são diversas sendo inerente a cada aplicação.

Portanto, o modelo proposto consegue avaliar conjuntamente vários descritores de movimento, que atuam como uma representação do contexto dos frames dos vídeos, permitindo classificar os frames dos vídeos adequadamente, uma análise realizada em tempo real e de acordo com seu contexto.

5. Experimentos

Na sequência, apresentam-se informações relativas ao dataset formalizado, as avaliações, e a discussão.

5.1. Dataset

Para avaliações, foi formalizado o dataset FPD através da análise manual de 14.671 vídeos, que extraídos totalizam 476.482 frames. O dataset foi rotulado em uma granularidade fina. Em outras palavras, esse dataset permite analisar o vídeo em uma granularidade de frames, onde todos os frames foram rotulados manualmente como pornografia ou normal, de acordo com o contexto do frame do vídeo. Os vídeos do dataset foram coletados de domínio público, como sites pornográficos e plataformas públicas de compartilhamento de vídeos. A Figura 3 mostra exemplos de frames de um mesmo vídeo pertencente ao dataset FPD e seus respectivos rótulos atribuídos manualmente.

Os datasets públicos utilizados na literatura para detecção de pornografia, não conseguem fornecer o nível esperado de granularidade. Em geral, apenas os vídeos são rotulados nas classes pornográfico ou normal. Ou seja, tais datasets acabam desprezando o contexto dos vídeos, pois vídeos pornográficos podem conter frames que não tenham teor pornográfico. Consequentemente, as técnicas propostas construídas com esses datasets, embora apresentem uma alta taxa de precisão, não operaram bem em condições de ambiente de produção.

5.2. Avaliação

A avaliação concentra-se em responder quatro perguntas de pesquisa: (Q1) A classificação baseada em movimento proposta permite a detecção de conteúdo pornográfico em vídeos? (Q2) O módulo de classificação de cena proposto ajuda na detecção

[5]WEKA - https://www.cs.waikato.ac.nz/ml/weka/

Figura 3. Amostras de frames do dataset FPD rotulado em uma granularidade de frames. As amostras (a),(b) são amostras de frames dependente de contexto, classificadas manualmente como normais. As amostras (c),(d),(e) são amostras classificadas como pornográficas.

de conteúdo pornográfico em tempo real? (Q3) A abordagem proposta é adequada para utilização em dispositivos com recursos computacionais limitados? (Q4) A abordagem proposta pode ser aplicada para detectar conteúdo pornográfico em uma granularidade de vídeo?

Sendo assim, nas próximas subseções descrevem-se as respostas.

5.3. Construção do Modelo

O modelo foi construído para avaliar nossa abordagem. Na seção 4 foram selecionadas as arquiteturas CNN e utilizamos o mesmo protocolo (60%, 20% e 20%). Também, três descritores de movimento: os fluxos ópticos obtidos com os algoritmos Lucas-Kanade [Lucas and Kanade 1981] e PLK [Bouguet et al. 2001] e um mapa de similaridade estrutural [Zhang 2004] de frames adjacentes. Conforme descrito na subseção 4.2, cada um dos descritores de movimento efetuam a extração das características a partir da análise do frame atual e o anterior. Para cada descritor de movimento utilizado, um modelo CNN foi construído e avaliado.

5.4. Classificação de Movimento

Q1 tem como objetivo avaliar se cada descritor de movimento pode ser aplicado na classificação. Construímos e avaliamos no mesmo conjunto as arquiteturas da CNN (subseção 4.2) para cada um dos descritores de movimento. A Tabela 2 mostra a acurácia obtida pelos descritores de movimento, apuradas para cada classe (normal, pornô e dependente de contexto) e de maneira geral.

É possível observar que as características baseadas em movimento utilizadas apresentaram uma acurácia semelhante aos resultados obtidos com a CNN baseada em imagem. Na maioria dos casos, a proposta apresenta taxas de detecção mais estáveis do que aquelas obtidas em arquiteturas CNN utilizando apenas imagem. Em outras palavras, a acurácia das arquiteturas CNN construídas a partir de descritores de movimento, não é significativamente degradada quando o conteúdo pornográfico depende do contexto.

Por exemplo, a arquitetura CNN Mobilenet, onde o descritor de vídeo de Lucas-Kanade (Mobilenet, Tabela 2), consegue classificar uma quantidade maior de frames pornográficos com um *tradeoff* de apenas 0,62% em termos de acurácia quando comparadas

Tabela 2. Desempenho da acurácia das arquiteturas CNN baseadas em movimento sobre o dataset FPD, considerando as seguintes abordagens de movimento: Pyramidal Lucas-Kanade (PLK), Lucas-Kanede(LK) e Structural Similarity Map (SSM).

Descritor de Movimento	CNN	Acurácia (%)			
		Normal	Pornô	Dependente de Contexto	Geral
PLK	Mobilenet	99,39	20,06	89,19	69,55
	Resnet	***97,87***	***23,98***	***80,26***	***67,37***
	Squeezenett	99,90	0,11	99,98	66,67
LK	***Mobilenet***	***99,11***	***42,52***	***80,75***	***74,13***
	Resnet	97,62	7,30	94,28	66,40
	Squeezenet	99,41	25,50	81,34	68,75
SSM	***Mobilenet***	***97,99***	***38,29***	***78,68***	***71,65***
	Resnet	86,40	25,23	74,67	62,10
	Squeezenet	99,35	12,79	85,31	65,82

a abordagem tradicional. Demostrando que a CNN que utiliza descritor de movimento é capaz de classificar pornografia nos vídeos levando em consideração os frames que são dependentes do contexto. Como resultado, o esquema de detecção proposto baseado em movimento pode permanecer confiável para o usuário, levando em consideração que ela apresentará taxas de precisão semelhantes à fase de teste quando usada na produção.

Para responder Q3, se a abordagem proposta é adequada para utilização em dispositivos com recursos limitados, realizamos um conjunto de experimentos de *benchmark*. Os experimentos foram realizados na placa desenvolvimento da NVIDIA Jetson TK1[6], selecionada devido a semelhança entre sua arquitetura de hardware e alguns dispositivos limitados. Para os experimentos, selecionamos a melhor arquitetura CNN (Mobilenet, Tabela 2) e o melhor classificador raso (REPTree, Tabela 3).

Replicamos o processo de classificação da CNN utilizando os descritores de movimento. Na sequência verificamos o tempo (unidade em milissegundos) que é necessário para execução de cada módulo da proposta, conforme disposto na Tabela 4. Extraímos o tempo necessário que a abordagem leva para carregar em memória a arquitetura CNN e o modelo para cada um dos descritores de movimento. Enfim, efetuamos uma avaliação do módulo de classificação de cena, para isso verificamos o tempo de predição necessário do classificador raso REPTree na placa Jetson TK1. Por meio dos experimentos de benchmark demostramos que a abordagem proposta é promissora para aplicação em dispositivos com recursos limitados, possuindo um tempo aceitável na detecção em tempo real.

5.5. Classificação de Cena

Para responder Q2, selecionamos os modelos CNN mais precisos para cada um dos descritores de movimento (mostrados em negrito na Tabela 2), para criar o vetor de características de entrada do classificador raso (Módulo Classificação de Cena, Figura 1).

Os seguintes classificadores rasos foram utizados: Adaboost, Bagging com o REPTree sendo utilizado como classificador base, BayesNet como sendo um classificador

[6]NVIDIA Jetson TK1 - https://docs.nvidia.com/jetpack-tk1/

Tabela 3. Desempenho final da acurácia da proposta sobre o dataset FPD (análise por frame).

Classificador Raso	**Acurácia (%)**			
	Normal	**Pornô**	**Dependente de Contexto**	**Geral**
AdaBoost	92,03	93,97	59,39	81,80
Bagging (REPTree)	87,52	93,37	75,01	85,30
BayesNet	88,28	92,81	74,21	85,11
J48	87,87	93,56	66,42	82,62
RandomForest	83,82	92,90	74,80	83,84
REPTree	***88,31***	***93,62***	***74,09***	***85,35***

Tabela 4. Desempenho da proposta aplicada dispositivos com recursos limitados.

CNN + Descritor de Movimento	**Benchmark Jetson TK1 (ms.)**	
	CPU	**GPU**
Pyramidal Lucas-Kanade	1572.02	1170
Lucas-Kanade	1550.95	1168
Structural Similarity Map (SSM)	1548.54	1109
Classificador Raso		
REPTree	0.7631	

probabilístico, o J48 que é uma implementação Java do algoritmo C4.5, RandomForest com 100 árvores de decisão utilizadas como base de aprendizagen, e árvore de decisão REPTree. Estes classificadores estão disponíveis na API Weka, versão 3.8.

Para os experimentos foram mantidos os parâmetros *default*. A Tabela 3 mostra a precisão da classificação obtida para cada um dos classificadores rasos avaliados. Nesse caso, é possível observar uma melhora significativa na acurácia quando comparada às CNNs baseadas em imagem tradicionais. Em outras palavras, a abordagem proposta foi capaz de melhorar significativamente a acurácia da detecção de conteúdo pornográfico dependente do contexto. Por exemplo, em relação ao classificador raso mais preciso, REPTree, a abordagem proposta foi capaz de melhorar a precisão da detecção em 7,61% até 15,26% na avaliação geral da proposta, levando assim em consideração informações pornográficas e de contexto. Sendo evidenciado pelo aumento das taxas de acurácia para a classe de frames pornográficos e dependente de contexto.

5.6. Classificação em Granularidade de Vídeo

Por fim, para responder a questão Q4, aplicamos a abordagem proposta para a classificação de vídeo. Avaliamos todos os frames do vídeo e consideramos um vídeo pornográfico de acordo com determinado limiar de pornografia. O limiar de pornografia estabelece a proporção dos frames do vídeo que devem ser classificados como pornográficos para classificar o vídeo como pornográfico. A Figura 4 mostra a relação entre o limite pornográfico e a precisão do vídeo obtido. É possível notar que nossa proposta é capaz de atingir uma taxa de Falso-Negativos (FN) de 6% e uma taxa de Falso-Positivos (FP) de 15% ao usar um limite pornográfico de 50%. No entanto, ao variar o limite por-

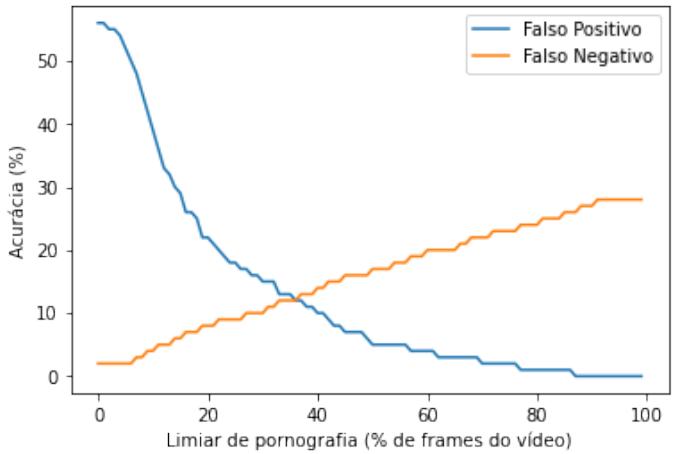

Figura 4. Proposta implementada utilizando o algoritmo REPTree, taxas de FP e FN quando usado para fins de classificação de vídeo.

nográfico usado, é possível diminuir ainda mais a taxa de FP de acordo com sua discrição. Além de ser aplicável à classificação em tempo real em granularidade de frames, nossa proposta também pode ser aplicada à classificação *offline* de vídeos pornográficos, com altas taxas de precisão para os dois casos.

6. Conclusão

As abordagens atuais para detecção de pornografia em vídeos são limitadas. Elas não foram projetadas para detecção em tempo real, tão pouco voltadas a dispositivos que possuam limitações de recursos computacionais. Para detecção de conteúdo pornografico em vídeo, atualmente as abordagens mais promissoras são as abordagens que utilizam convoluções 3D ou abordagens que partem da combinação das arquiteuras de aprendizagem profunda, a CNN e o LSTM. Tais abordagens se destacam por tratar o problema do contexto nesse tipo de mídia. Contudo, aplicar esse tipo de abordagem em dispositivos como alguns smartphones e tablets é inviável devido as restrições de recursos computacionais dos dispositivos, como limitação de memória, CPU ou disco.

Observando tais limitações, propusemos um esquema de classificação baseado em informação de movimento que é capaz de melhorar significativamente a precisão da detecção em tempo real na classificação de conteúdo pornográfico dependente do contexto em vídeos. O *insight* da proposta utiliza descritores baseados em informação de movimento para aumentar as características de vídeo a serem utilizadas pelo modelo subjacente da CNN. Além disso, para permitir o uso de vários descritores de movimento, aplicamos um classificador raso para avaliar o resultado final da classificação dos frames dos vídeos sobre um conjunto características de movimento.

Como resultado, nossa proposta foi capaz de melhorar significativamente a acurácia na tarefa de detecção de pornografia em tempo real, mas principalmente fornecendo recursos para auxiliar o controle paterno. Além da abordagem ser leve, pode ser implementada em dispositivos com limitação de recurso computacional. Para trabalhos futuros, avaliaremos nossa proposta em outros campos dependentes do contexto, como cenas violentas e reconhecimento de atividades humanas. Além do que, pretende-se aperfeiçoar o dataset FPD para disponibiliza-lo publicamente.

Agradecimentos

Os autores agradecem ao CNPq pelo apoio financeiro parcial ao projeto, processos 304990/2021-3, 302937/2023-4, e 407879/2023-4.

Referências

Amato, G., Bolettieri, P., Costa, G., la Torre, F., and Martinelli, F. (2009). Detection of images with adult content for parental control on mobile devices? In *Proceedings of the 6th International Conference on Mobile Technology, Application & Systems - Mobility '09*. ACM Press.

Bouguet, J.-Y. et al. (2001). Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel corporation*, 5(1-10):4.

Clement, J. (2019). Mobile internet usage worldwide - statistics facts. https://www.statista.com/topics/779/mobile-internet/ Accessed: Julho 28, 2020.

dos Santos, R. R., Viegas, E. K., and Santin, A. O. (2021). A reminiscent intrusion detection model based on deep autoencoders and transfer learning. In *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE.

dos Santos, R. R., Viegas, E. K., Santin, A. O., and Tedeschi, P. (2023). Federated learning for reliable model updates in network-based intrusion detection. *Computers amp; Security*, 133:103413.

Endeshaw, T., Garcia, J., and Jakobsson, A. (2008). Classification of indecent videos by low complexity repetitive motion detection. In *2008 37th IEEE Applied Imagery Pattern Recognition Workshop*. IEEE.

Geremias, J., Viegas, E. K., Santin, A. O., Britto, A., and Horchulhack, P. (2022). Towards multi-view android malware detection through image-based deep learning. In *2022 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377.

Horchulhack, P., Viegas, E. K., Santin, A. O., Ramos, F. V., and Tedeschi, P. (2024a). Detection of quality of service degradation on multi-tenant containerized services. *Journal of Network and Computer Applications*, 224:103839.

Horchulhack, P., Viegas, E. K., Santin, A. O., and Simioni, J. A. (2024b). Network-based intrusion detection through image-based cnn and transfer learning. In *2024 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861.

Iandola, F. N., Moskewicz, M. W., Ashraf, K., Han, S., Dally, W. J., and Keutzer, K. (2017). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡1mb model size. *ArXiv*, abs/1602.07360.

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.

Jones, T. (2017 (https://www.ibm.com/developerworks/br/library/cc-machine-learning-deep-learning-architectures/index.html Acesso em 2020 Julho 23, 2020)). *Deep learning architectures*.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Katta, S. S., Nandyala, S., Viegas, E. K., and AlMahmoud, A. (2022). Benchmarking audio-based deep learning models for detection and identification of unmanned aerial vehicles. In *2022 Workshop on Benchmarking Cyber-Physical Systems and Internet of Things (CPS-IoTBench)*. IEEE.

Kuroki, Y., Nishi, T., Kobayashi, S., Oyaizu, H., and Yoshimura, S. (2007). A psychophysical study of improvements in motion-image quality by using high frame rates. *Journal of the Society for Information Display*, 15(1):61.

Lee, S., Shim, W., and Kim, S. (2009). Hierarchical system for objectionable video detection. *IEEE Transactions on Consumer Electronics*, 55(2):677–684.

Li, Q., Qiu, Z., Yao, T., Mei, T., Rui, Y., and Luo, J. (2016). Action recognition by learning deep multi-granular spatio-temporal video representation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval - ICMR '16*. ACM Press.

Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision (ijcai). volume 81.

Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2019). Multimodal data fusion for sensitive scene localization. *Information Fusion*, 45:307–323.

Niu, W., Ma, X., Wang, Y., and Rén, B. (2019). 26ms inference time for resnet-50: Towards real-time execution of all dnns on smartphone. *ArXiv*, abs/1905.00571.

Perez, M., Avila, S., Moreira, D., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2017). Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, 230:279–293.

Rea, N., Lacey, G., Dahyot, R., and Lambe, C. (2006). Multimodal periodicity analysis for illicit content detection in videos. In *3rd European Conference on Visual Media Production (CVMP 2006). Part of the 2nd Multimedia Conference 2006*. IEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Viegas, E. K., Santin, A. O., Cogo, V. V., and Abreu, V. (2020). *Facing the Unknown: A Stream Learning Intrusion Detection System for Reliable Model Updates*, page 898–909. Springer International Publishing.

Wu, Z., Wang, X., Jiang, Y.-G., Ye, H., and Xue, X. (2015). Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia - MM '15*. ACM Press.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, H. (2004). The optimality of naive bayes. In *American Association for Artificial Intelligence*. FLAIRS Conference.