A Two-tier Image Representation Approach to Detecting Child Pornography

Paulo Vitorino*†, Sandra Avila‡, Anderson Rocha§
*Departamento de Engenharia Elétrica, Universidade de Brasília, Brasil
†Unidade Técnico-Científica da Delegacia de Polícia Federal em Guaíra, Brasil
‡Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Brasil
§Instituto de Computação, Universidade Estadual de Campinas, Brasil
paulo.prrv@pf.gov.br, sandra@dca.fee.unicamp.br, anderson.rocha@ic.unicamp.br

***Abstract*—Sexual exploitation of children is without a doubt one of the most heinous crimes in modern society, prompting us to the need of responding firmly with the development and implementation of strict laws as well as the development of efficient and effective means to detect such activities. In this article, we propose an initial solution to the problem of child pornography detection in digital images aiming at assisting law-enforcement agents on investigating criminal activities related to the generation and distribution of this type of content. Our solution relies upon a two-tier representation of image features calculated through the concept of visual dictionaries with distance-preserving properties of low-level features enabled by the recently proposed BossaNova formalism. We validate it with thousands of images of real-world apprehensions of the Brazilian Federal Police. The results, although still far from solving the problem, show that this solution is promising.**

I. Introdução

Segundo o Fundo das Nações Unidas para a Infância (UNICEF)[1], a cada 15 segundos, uma criança é abusada no mundo e, a cada 8 minutos, uma criança é abusada sexualmente no Brasil. Crianças estão em situação de risco devido à produção, exibição, distribuição e consumo de pornografia infantil.

A questão do que vem a constituir pornografia infantil é um problema em si. Os conceitos de *criança* e *pornografia* diferem de país para país e referenciam convicções morais, culturais, sexuais, sociais e religiosas que nem sempre se traduzem nas respectivas legislações [1].

No Brasil, o Estatuto da Criança e do Adolescente (Lei 11.829/2008[2]), em seu art. 1º, preceitua que produzir, reproduzir, dirigir, fotografar, filmar ou registrar — por qualquer meio — cena de sexo explícito ou pornográfica envolvendo criança, ou adolescente, constitui crime.

Assim, para combater a exploração da criança em materiais pornográficos, a detecção automática de conteúdo de forma inteligente e contínua é primordial. Tipicamente, as soluções encontradas na literatura para a detecção automática de pornografia infantil são baseadas em: (1) comparação de assinaturas únicas (*hash*) dos arquivos [2–6]; (2) detecção de pele [2, 7]; (3) identificação de partes do corpo humano (por exemplo, faces, genitais) [2, 8]; e (4) extração de características locais e modelos Sacolas de Palavras Visuais [9–11].

[1]http://www.unicef.org/brazil
[2]http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11829.htm

Claramente, abordagens baseadas em (1) funcionam apenas para conteúdos com poucas modificações (dado que modificações mais complexas podem destruir a assinatura do objeto multimídia compartilhado) e necessitam que haja denúncia por parte de algum interessado apontando certo conteúdo como criminoso (e.g., ferramentas de denúncia de conteúdo impróprio em redes sociais). Abordagens baseadas em (2) naturalmente sofrem com o alto número de falsos positivos uma vez que há um grande fosso semântico entre o conceito de exposição de pele e a extrapolação de que isso está diretamente ligado à pornografia infantil. Complementarmente, abordagens baseadas em (3) não são eficazes quando não se tem exposição explícita de órgãos genitais ou da face, por exemplo. Nesse sentido, várias pesquisas têm sido feitas na direção (4) e esse artigo também vai ao encontro de tais abordagens.

Nesse contexto, nesse artigo apresentamos uma solução automática multi-nível para detecção de pornografia infantil em imagens digitais baseada no modelo de sacolas de palavras visuais. No primeiro nível, codificamos as principais informações presentes em uma imagem de entrada utilizando o conceito de descritores locais de baixo nível. Para isso, lançamos mão do descritor *Speeded-Up Robust Features* (SURF) [12]. Em seguida, mapeamos tais descritores de baixo nível em uma representação de médio nível utilizando o conceito de dicionários visuais mas com a preservação do conceito de distância entre as características de baixo nível mapeadas. Para isso, empregamos a técnica BossaNova [13]. Finalmente, os conceitos representados no médio nível são capturados em uma representação semântica mais sofisticada (alto nível) para o conceito pornografia infantil vs. não pornografia infantil. Esse último passo é implementado a partir de um classificador de padrões baseado em máquinas de vetores de suporte. As principais diferenças com outros trabalhos utilizando modelos de sacolas de palavras estão no fato de que todas as outras abordagens utilizam apenas o pipeline clássico (descrição de baixo nível e criação do dicionário sem preservação de distância), os códigos-fontes não estão diretamente disponíveis, além do fato de que suas validações são, normalmente, limitadas a poucos exemplos e não casos reais de apreensão. Diferentemente, nesse artigo utilizamos casos reais de apreensões da Polícia Federal do Brasil bem como propomos uma solução automática, em multi-nível, de código aberto.

Dentre as principais contribuições deste trabalho, destacam-se a representação com preservação de distância nas características de baixo nível, que conferem uma representação mais robusta no médio nível, bem como a avaliação consistente em um conjunto de imagens real em parceria com a Polícia Federal do Brasil contendo dezenas de milhares de imagens.

Este trabalho está organizado como segue: na Seção II são descritos os trabalhos relacionados; na Seção III é apresentada a solução proposta; na Seção IV são discutidos os resultados experimentais; por fim, as conclusões e trabalhos futuros são apresentados na Seção V.

II. Revisão da Literatura

Diversas abordagens foram propostas na literatura para detecção de imagens de pornografia infantil, algumas focadas diretamente no conteúdo das imagens e outras baseadas em informações associadas tais como textos, legendas e nomes de arquivos.

A abordagem textual busca identificar conteúdo pornográfico infantil através das palavras utilizadas na nomenclatura dos arquivos [2, 14], e nas descrições textuais neles contidas [15]. Claramente, as abordagens baseadas em texto não são apropriadas, uma vez que palavras inofensivas podem estar maliciosamente associadas a conteúdos impróprios [16].

Outro conjunto de técnicas explorado na literatura consiste na extração de assinaturas únicas (*hash*) de imagens de pornografia infantil [4]. Bancos de dados de assinaturas são utilizados pelas principais ferramentas forenses de análise de mídias (e.g. FTK[3], EnCase[4], Microsoft PhotoDNA [3]), e pelas ferramentas de monitoramento de redes ponto-a-ponto (e.g. EspiaMule [5], RoundUp [6]). Um problema evidente com essas técnicas está na busca por assinaturas exatas (i.e., imagens já registradas nos bancos de dados), o que impede a detecção de uma imagem que nunca foi analisada anteriormente. Algumas extensões para comparação aproximada foram propostas mas, ainda assim, alterações no objeto digital muitas vezes podem gerar uma assinatura *hash* completamente diferente.

As abordagens mais promissoras para detecção de pornografia infantil são baseadas no conteúdo visual das imagens. A maioria dos trabalhos encontrados na literatura, no entanto, propõe técnicas de detecção de pele para filtragem (ou pré-filtragem) de conteúdo pornográfico infantil [2, 8, 11, 17]. Isto se dá pelo fato de que a propriedade mais intuitiva em imagens de pornografia infantil é a grande fração de pixels relacionados à pele humana [18]. Apesar disso, essas abordagens geralmente têm como desvantagem a alta taxa de falsos positivos, dado que imagens com grandes áreas de exposição de pele não significam necessariamente conteúdo pornográfico infantil (por exemplo, imagens com pessoas usando roupas de banho, imagens relacionadas a esportes). Além disso, e mais importante, imagens de pornografia infantil podem apresentar pequenas áreas de exposição de pele.

Para contornar estes problemas, Ulges & Stahl [9] propuseram um método baseado no modelo Sacolas de Palavras Visuais (BoVW, do inglês *Bag of Visual Words*) [19], que surgiu como uma das abordagens mais promissoras para a classificação de imagens (ver Seção III). Para detectar pornografia infantil em imagens, os autores utilizaram coeficientes de baixa frequência da transformada discreta do cosseno como descritores locais, e Máquinas de Vetores de Suporte (SVM, do inglês *Support Vector Machines*) para a classificação. Apesar do método apresentar resultados melhores do que a técnica de detecção de pele, a taxa de erro para classificação de pornografia infantil foi de 24.0%.

Carvalho [10] apresentou uma análise comparativa de descritores locais — para o modelo BoVW — na classificação de conteúdo pornográfico e pornográfico infantil. Foram avaliados os descritores SIFT, OpponentSIFT e WSIFT, em conjunto com o classificador Naïve Bayes. Segundo o autor, para a comparação de imagens de pornografia feminina vs. pornografia infantil feminina, o melhor resultado foi obtido com o descritor OpponentSIFT (taxa de erro de 26%). Os experimentos, entretanto, foram realizados em um conjunto de dados com apenas 400 imagens.

Outro trabalho que merece destaque na literatura é o de Schulze et al. [11]. Em tal trabalho, os autores utilizam a fusão de diferentes características multimodais. A descrição do conteúdo visual de imagens e frames de vídeo é realizada com a extração de quatro características de baixo nível (correlogramas de cores, características de pele, pirâmides visuais e palavras visuais) e uma característica de médio nível (SentiBank). SentiBank é uma representação de conteúdo visual de médio nível, baseada na *Visual Sentiment Ontology*, e consiste em 1.200 conceitos semânticos e classificadores automáticos correspondentes, cada um sendo definido como um par substantivo-adjetivo (ANP, do inglês *Adjective Noun Pair*). Cada conjunto de características passa por um classificador SVM e os scores individuais são posteriormente combinados.

Além das abordagens encontradas na literatura, ferramentas forenses também têm sido propostas para detectar conteúdo pornográfico infantil. O Microsoft PhotoDNA [3], por exemplo, caracteriza imagens de pornografia infantil por meio da utilização de assinaturas únicas de *hash*. O NuDetective [2] — ferramenta desenvolvida pela Polícia Federal do Brasil — filtra imagens e vídeos por meio de informações textuais, assinaturas únicas e detectores de pele. O INACT [20], por sua vez, implementa um banco de dados de assinaturas únicas e descritores MPEG-7 de imagens de pornografia infantil.

Em resumo, podemos observar que, apesar dos recentes avanços no contexto de detecção automática de pornografia infantil, a tarefa continua sendo um desafio em aberto, visto que a utilização de técnicas tradicionais (por exemplo, detecção de pele), nem sempre são adequadas devido à alta taxa de falsos positivos. Ademais, em relação aos modelos BoVW, as abordagens propostas têm aplicado apenas o modelo mais clássico [19]. Este modelo tem se aprimorado e atualmente existem várias extensões (e.g., BossaNova [13], Fisher Vector [21]) que oferecem melhores taxas de classificação. Estas extensões ainda não foram utilizadas para detectar conteúdo pornográfico infantil, motivando assim o presente trabalho. Por fim, uma deficiência evidente nos trabalhos analisados é a falta da comparação com outras abordagens.

[3] http://accessdata.com/solutions/digital-forensics/forensic-toolkit-ftk
[4] https://www.guidancesoftware.com/encase-forensic

III. Abordagem Proposta

Nesse artigo, apresentamos uma solução para detecção automática de pornografia infantil baseada em uma representação multi-nível, consistindo em um mapeamento suave de características de baixo nível semântico para características de maior nível conceitual associado. Para essa representação, lançamos mãos do conceitos de descritores locais, dicionários visuais com preservação de informação de distância e classificação de padrões.

Na Figura 1, é apresentada uma visão geral da abordagem proposta para detecção de pornografia infantil, que consiste em duas fases: treinamento (*offline*) e teste (*online*). Na fase de treinamento, depois da extração dos descritores locais (passo A:1), o dicionário de palavras visuais é obtido aplicando uma abordagem de aprendizado não-supervisionado sobre uma amostragem de descritores locais (passo A:2.1). Em seguida, para cada imagem do conjunto de treinamento, os descritores locais são quantizados e agregados em uma representação global (passo A:2.2), que é dada como entrada para um algoritmo de aprendizagem de máquina para a construção do modelo de classificação (passo A:3). Na fase de teste, imagens desconhecidas são apresentadas ao sistema e os mesmos métodos da fase de treinamento são aplicados (passo B:1 $\Leftrightarrow$ A:1 e passo B:2 $\Leftrightarrow$ A:2.2). Neste caso, as imagens são representadas usando o dicionário visual previamente aprendido (passo B:2), e as classes das imagens são determinadas a partir do modelo de classificação (passo B:3). Cada um destes passos é detalhado a seguir.

A. Baixo Nível — Extração de Descritores Locais

Descritores locais de baixo nível extraem características visuais, de pequenas regiões das imagens, baseados em atributos perceptuais (como cor, forma, textura). Estas características são geralmente codificadas de modo que sejam invariantes a transformações das imagens, tais como a translação, rotação, mudanças de escala ou deformações [22].

Dentre os diversos tipos de descritores locais encontrados na literatura, o SIFT (*Scale-Invariant Feature Transform*) [23] e o SURF (*Speeded-Up Robust Features*) [12] são provavelmente os mais referenciados.

Para a solução proposta, optamos utilizar o descritor SURF (passos A:1 e B:1). Segundo Bay et al. [12], o SURF é mais rápido e mais robusto que o SIFT. Além disso, o SURF retorna um vetor com 64 dimensões, metade do tamanho do vetor retornado pelo algoritmo SIFT.

B. Médio Nível — Extração das Representações BoVW

Devido à significativa distância semântica entre a codificação de *baixo nível* das imagens e o que elas representam em *alto nível* (por exemplo, pornografia infantil), um único nível de descrição é quase sempre insuficiente para tarefas de reconhecimento de imagens. Nessa perspectiva, surgiram as representações de *médio nível*, tendo como principal objetivo transformar os descritores locais de baixo nível, previamente extraídos, em uma representação global e mais rica da imagem [24].

Tipicamente, a extração das representações de médio nível pode ser dividida em duas etapas [24]: *coding* e *pooling* (passos A:2.2 e B:2). A primeira etapa quantifica os descritores locais de baixo nível de acordo com um dicionário de palavras visuais, normalmente construído aplicando um algoritmo de agrupamento (por exemplo, k-means) em uma amostragem dos descritores locais extraídos. A segunda etapa, por sua vez, agrega os descritores quantizados em um único vetor.

A abordagem de representação de médio nível mais utilizada na literatura, proposta por Sivic & Zisserman [19], é conhecida como Sacolas de Palavras Visuais (BoVW, do inglês *Bag of Visual Words*). Tradicionalmente, na representação BoVW, os descritores locais da imagem são associados ao elemento mais próximo do dicionário visual (*hard coding*) e, em seguida, a média desses descritores locais codificados é calculada, compactando toda a informação em um único vetor (*average pooling*).

Várias extensões da representação BoVW foram propostas na literatura. Entre elas, ressalta-se a representação Bossa-Nova [13]. Em linhas gerais, esta representação segue o formalismo do modelo BoVW (*coding* & *pooling*), oferecendo um aprimoramento na etapa de *pooling*, a fim de preservar de uma maneira mais rica a informação obtida durante a etapa de *coding*. Assim, em vez de compactar toda a informação relacionada a uma palavra visual em um único valor escalar, a etapa de *pooling* resulta em uma distribuição de distâncias. Para isto, Avila et al. [13, 25] usaram uma estimação não-paramétrica da distribuição dos descritores, calculando um histograma de distâncias entre os descritores encontrados na imagem e cada palavra visual do dicionário.

Em suma, a representação BossaNova tem a vantagem de ser compacta, conceitualmente simples, e flexível. Por essas razões, para a solução proposta, utilizamos esta representação.

C. Alto Nível — Classificação Supervisionada

Finalmente, o objetivo da classificação supervisionada é aprender uma função de mapeamento que atribui rótulos (discretos) às imagens arbitrárias. Assim, na fase de treinamento, um classificador é treinado usando os vetores de médio nível obtidos (passo A:3). Uma vez que modelo de classificação é aprendido, ele pode ser utilizado para prever o rótulo de um novo exemplo (passo B:3).

Diversos algoritmos de aprendizado de máquina podem ser utilizados neste último passo. Na literatura BoVW [13], a técnica de maximização de margem usando Máquinas de Vetores de Suporte é a mais utilizada, e também considerada neste trabalho.

IV. Experimentos e Resultados

Nesta seção são apresentados os experimentos realizados e os resultados obtidos. Primeiramente, na Seção IV-A, é descrita a base de dados utilizada; em seguida, na Seção IV-B, são apresentadas as métricas de avaliação aplicadas; na Seção IV-C, são detalhados os protocolos experimentais; e por fim, na Seção IV-D, são reportados os experimentos realizados e são analisados os resultados alcançados.

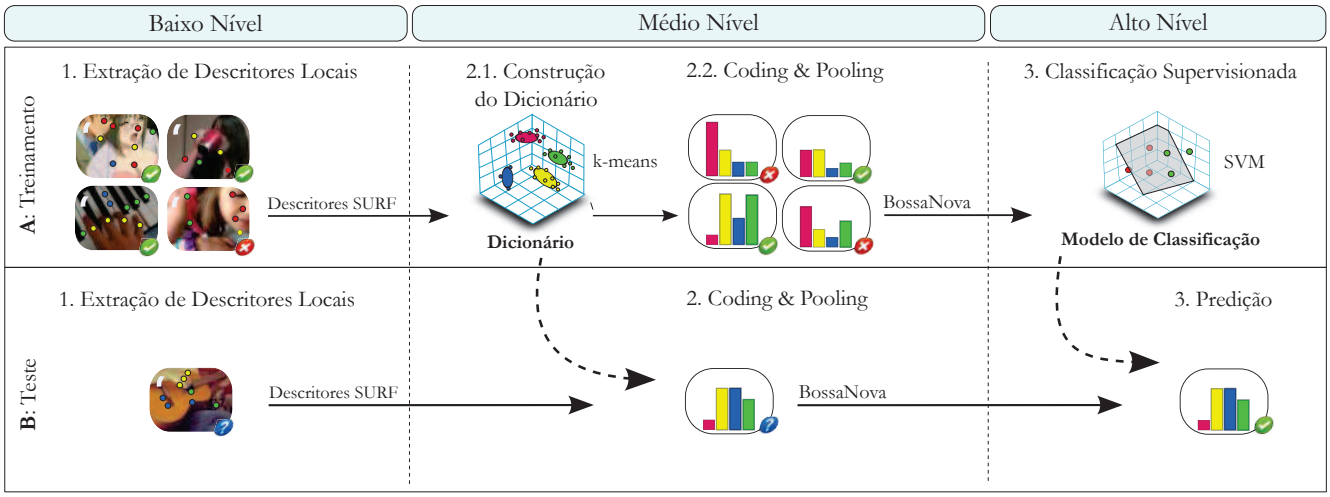

Figura 1. Visão geral da abordagem proposta para detecção de pornografia infantil em imagens.

A. Base de Dados

As imagens utilizadas nos experimentos foram obtidas do arquivo de laudos gerados por peritos criminais da Polícia Federal do Brasil. No total, a base de dados contém 9.987 imagens, sendo 4.998 imagens de pornografia infantil. É importante ressaltar que o conjunto de imagens que não apresenta conteúdo pornográfico infantil, é composto de imagens de conteúdo geral, inclusive imagens de nudez e pornografia, o que torna a tarefa de classificação de pornografia infantil, nesta base de dados, ainda mais desafiadora. Devido à vedação legal (Lei nº 11.829/2008), as imagens não são ilustradas neste artigo.

A base de dados é dividida em (i) treinamento, composto por 5.995 imagens; e (ii) teste, composto por 3.992 imagens.

B. Métricas de Avaliação

Para avaliar o desempenho dos classificadores de pornografia infantil, as métricas utilizadas foram a acurácia normalizada (ACC) e a medida F_2 (F_2).

ACC reporta a taxa de sucesso do classificador, independentemente dos rótulos de classe. Matematicamente, isto pode ser expresso como:

$$ACC = \frac{TVP + TVN}{2}, \quad (1)$$

onde TVP é a taxa de verdadeiros positivos e TVN é a taxa de verdadeiros negativos.

F_2, por sua vez, é a média harmônica ponderada da precisão e revocação, na qual a revocação tem o dobro do peso da precisão ($\beta = 2$). Para a detecção de pornografia infantil, a medida F_2 é crucial dado que falsos negativos são inaceitáveis, porque permitem que imagens com conteúdo pornográfico infantil não sejam detectadas. A medida F_β é definida como:

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

onde o parâmetro β denota a importância da revocação em relação à precisão. Para os nossos experimentos, empregamos $\beta = 2$, o que significa que a revocação é duas vezes mais relevante que a precisão.

É importante destacar que as duas métricas devem ser avaliadas em conjunto.

C. Protocolo Experimental

1) Detector de Pele Humana: Técnicas de detecção de pele humana são amplamente aplicadas para classificação de pornografia infantil [2, 7]. Diversas soluções para detecção de pele têm sido propostas na literatura [18], sendo que a forma mais comum e intuitiva utiliza características de cor [26].

Kovac et al. [27] propuseram um detector de pele humana que utiliza testes lógicos para determinar a classe que a imagem pertence, a partir das diferenças entre os valores do pixel no espaço de cor RGB (ver Eq. 3). Dessa forma, determina-se um subespaço de cores capaz de identificar a pele humana. Por ser uma técnica bastante utilizada e de fácil implementação, o detector de pele de Kovac et al. foi reimplementado neste trabalho para fins de comparação.

$$\begin{aligned} &\{[(R > 95) \wedge (G > 40) \wedge (B > 20)] \wedge \\ &[max(R,G,B) - min(R,G,B) > 15] \wedge \\ &[|R - G| > 15 \wedge (R > G) \wedge (R > B)]\}. \end{aligned} \quad (3)$$

2) Ferramentas Forenses: Apesar de existirem várias ferramentas forenses para combater a pornografia infantil, a maioria não está disponível para o público em geral — por motivos óbvios —, e poucas soluções analisam o conteúdo visual das imagens para classificá-las.

Assim, para avaliar o desempenho da solução proposta, foram selecionadas as seguintes ferramentas: NuDetective [2], Localizador de Evidências Digitais (LED), MediaDetective [28] e Snitch Plus [29]. O NuDetective pode ser adquirido gratuitamente por agências de aplicação da lei ou para fins de pesquisa, enquanto que a ferramenta LED está disponível apenas para os peritos criminais da Polícia Federal do Brasil. O MediaDetective e o Snitch Plus são soluções comerciais para filtragem de conteúdo pornográfico. Todos estes sistemas

analisam o conteúdo das imagens por meio de técnicas de detecção de pele, principalmente.

Além disso, para o MediaDetective e o Snitch Plus, as imagens são classificadas de acordo com o valor de probabilidade. Ou seja, a imagem tem conteúdo pornográfico infantil se a probabilidade for igual ou maior que 50%. De modo similar, para o LED, as imagens são classificadas de acordo com o valor do Detector de Imagens Explícitas (DIE), que varia de 0 a 1000. Para os nossos experimentos, a imagem tem conteúdo pornográfico infantil se o valor do DIE for igual ou maior que 500. O NuDetective, por outro lado, atribui valores binários para a imagem: positivo (pornografia infantil) ou negativo (não pornografia infantil).

Por fim, o MediaDetective e o Snitch Plus têm quatro modos de execução pré-definidos, que diferem principalmente quanto ao rigor do detector de pele. Em nossos experimentos, optamos pelo modo mais rigoroso. Em relação ao NuDetective e ao LED, empregamos suas configurações padrão.

3) Solução Proposta — Modelos BoVW: Para a etapa de baixo nível, primeiramente, as imagens são redimensionadas para no máximo 100k pixels, economizando o tempo de processamento das etapas posteriores [30]. Em seguida, descritores SURF [12] são extraídos usando uma amostragem densa com cinco escalas diferentes. Precisamente, usamos regiões de tamanhos 24×24, 32×32, 48×48, 68×68 e 96×96 pixels, com espaçamento de $4, 6, 8, 11$ e 16 pixels, respectivamente. Além disso, a dimensionalidade do descritor SURF é reduzida, de 64 para 32 dimensões, utilizando Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*). Os códigos do descritor SURF e do PCA foram obtidos a partir do repositório OpenCV[5], uma das bibliotecas mais populares de Visão Computacional.

Para a etapa de médio nível, são extraídos vetores BossaNova [13], mantendo os valores padrão[6], e vetores BoVW [19] (*hard coding* e *average pooling*), para fins de comparação. O tamanho dos dicionários é variado em $\{512, 1024, 2048\}$ palavras visuais.

Por fim, na etapa de alto nível, classificadores SVMs são aplicados, utilizando um kernel χ^2 implementado pela biblioteca PmSVM[7] (*Power Mean SVM*) [31].

D. Resultados

Na Tabela I são reportados os resultados obtidos pela abordagem proposta, ferramentas forenses e técnicas da literatura. Para dar ao leitor uma visão mais ampla dos resultados, além da medida F_2 (F_2) e da acurácia normalizada (ACC), também são reportadas a taxa de verdadeiros positivos (TVP) e a taxa de verdadeiros negativos (TVN).

Como podemos observar, as abordagens baseadas no modelo BoVW superam a solução do detector de pele e as ferramentas forenses analisadas. Não surpreendentemente, as soluções baseadas em detecção de pele apresentam elevado número de falsos positivos, variando de $55,1\%$ (Snitch-Plus [29]) a $83,7\%$ (Kovac et al. [27]). A robustez das técnicas baseadas em BoVW é ainda mais evidenciada quando comparada (SURF & BossaNova) com a melhor ferramenta forense (NuDetective [2]): uma redução de erro de $28,0\%$ em relação à ACC. No que concerne à F_2, como as soluções baseadas em detecção de pele reportam baixo número de falsos negativos — e alto número de falsos positivos —, os valores de F_2 são consequentemente mais altos. No entanto, para detecção de conteúdo pornográfico infantil, é essencial obter bons resultados em relação às duas métricas, ACC e F_2.

Tabela I
RESULTADOS OBTIDOS PELA ABORDAGEM PROPOSTA, FERRAMENTAS FORENSES E TÉCNICAS DA LITERATURA.

Solução	TVP (%)	TVN (%)	F_2 (%)	ACC (%)
Kovac et al. [27]	89,3	16,3	77,9	52,8
LED	93,1	16,4	80,7	54,7
MediaDetective [28]	70,8	16,5	63,9	43,6
NuDetective [2]	76,6	36,1	70,9	56,4
Snitch Plus [29]	70,9	44,9	67,4	57,9
SURF & BoVW	**65,1**	**67,2**	**65,4**	**66,2**
SURF & BossaNova	**66,4**	**70,2**	**66,9**	**68,3**

TVP: taxa de verdadeiros positivos — TVN: taxa de verdadeiros negativos
F_2: medida F_2 — ACC: acurácia normalizada

Entre as soluções baseadas em BoVW, a solução proposta fornece classificadores mais eficazes, tanto em termos de ACC ($68,3\%$) como de F_2 ($66,9\%$). A comparação das soluções baseadas em BoVW é particularmente relevante, porque, nos experimentos realizados, as etapas de baixo nível e alto nível foram fixadas para avaliar o médio nível. Os resultados confirmaram a importância do uso de uma representação mais robusta no médio nível para detecção de pornografia infantil. Para as duas soluções, o melhor resultado foi obtido com 2048 palavras visuais.

V. CONCLUSÃO

Neste artigo, foi apresentada uma solução automática para detecção de conteúdo pornográfico infantil em imagens digitais, baseada nos modelos avançados de sacolas de palavras visuais. Mais especificamente, foi utilizada a representação de imagem BossaNova, que preserva informações importantes sobre a distribuição dos descritores locais em relação às palavras visuais.

A solução proposta, avaliada em um conjunto de imagens real em parceria com a Polícia Federal do Brasil (PF), apresentou resultados com qualidade superior em relação às diversas abordagens encontradas na literatura, inclusive ferramentas forenses. No entanto, apesar da solução ter apresentado resultados promissores, a detecção automática de pornografia infantil continua sendo um problema em aberto e continuaremos nossos esforços nesse sentido em parceria com a PF.

Em trabalhos futuros, nosso primeiro objetivo é aumentar a base de dados coletada. Em uma primeira avaliação, já conseguimos acesso a 60.000 imagens mas que ainda precisam ser manualmente anotadas. Em seguida, visamos estender esse número para um milhão. Além disso, destaca-se que todo o processamento dessas imagens é feita nas premissas da PF por questões de segurança e privacidade. Além disso, iremos

[5]OpenCV: opencv.org
[6]BossaNova: https://sites.google.com/site/bossanovasite
[7]PmSVM: https://sites.google.com/site/wujx2001/home/power-mean-svm

investigar abordagens de representação de médio nível baseadas em vetores de Fisher [21] e explorar soluções de fusão de dados uma vez que diferentes abordagens podem explorar propriedades diferentes do problema. Finalmente, após obtermos dados de treinamento suficientes, também objetivamos explorar soluções baseadas em aprendizado em profundidade (*deep learning*) de modo a conseguirmos extrair nuances do problema diretamente dos dados aumentando, assim, nossas chances de resolver esse difícil problema.

Agradecimentos

Os autores agradecem ao CNPq, FAPESP — Projeto DéjàVu #2015/19222-9 — e CAPES — Projeto DeepEyes —, pelo suporte financeiro dessa pesquisa.

Referências

[1] K. Figueiredo and S. Bochi, "Violência sexual: Um fenômeno complexo," http://www.unicef.org/brazil/pt/Cap_03.pdf.

[2] M. Polastro and P. Eleuterio, "Nudetective: A forensic tool to help combat child pornography through automatic nudity detection," in *Workshops on Database and Expert Systems Applications*, 2010, pp. 349–353.

[3] Microsoft Inc., "Microsoft's photodna: Protecting children and businesses in the cloud," https://news.microsoft.com/features/microsofts-photodna-protecting-children-and-businesses-in-the-cloud, 2015.

[4] A. Vrubel, "Creation and maintenance of MD5 hash libraries, and their application in cases of child pornography," in *International Conference on Forensic Computer Science*, 2011, pp. 137–141.

[5] J. Oliveira and E. Silva, "Espiamule e wyoming toolkit: Ferramentas de repressão à exploração sexual infanto-juvenil em redes peer-to-peer," in *Conferência Internacional de Perícias em Crimes Cibernéticos*, 2009, pp. 108–113.

[6] R. Hurley, S. Prusty, H. Soroush, R. J. Walls, J. Albrecht, E. Cecchet, B. N. Levine, M. Liberatore, B. Lynn, and J. Wolak, "Measurement and analysis of child pornography trafficking on P2P networks," in *International Conference on World Wide Web*, 2013, pp. 631–641.

[7] M. Islam, P. A. Watters, and J. Yearwood, "Real-time detection of children's skin on social networking sites using markov random field modelling," *Information Security Technical Report*, vol. 16, no. 2, pp. 51–58, 2011.

[8] N. Sae-Bae, X. Sun, H. T. Sencar, and N. D. Memon, "Towards automatic detection of child pornography," in *IEEE International Conference on Image Processing*, 2014, pp. 5332–5336.

[9] A. Ulges, C. Schulze, and A. Stahl, "Automatic image and video understanding for investigations of child sexual abuse," in *European Academy of Forensic Science Conference*, 2011, pp. 1–2.

[10] I. A. de Carvalho, "Classificação de imagens de pornografia e pornografia infantil utilizando recuperação de imagens baseada em conteúdo," Master's thesis, Universidade de Brasília, 2012.

[11] C. Schulze, D. Henter, D. Borth, and A. Dengel, "Automatic detection of CSA media by multi-modal feature fusion for law enforcement support," in *ACM International Conference in Multimedia Retrieval*, 2014, pp. 353–360.

[12] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[13] S. Avila, N. Thome, M. Cord, E. Valle, and A. Araújo, "Pooling in image representation: the visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, pp. 453–465, 2013.

[14] L. Barreto, L. Nunes, and D. Cunha, "Beeapeer: Uma ferramenta para localização e monitoramento de material de abuso sexual infantil na rede Gnutella," in *International Conference on Forensic Computer Science*, 2010, pp. 33–41.

[15] A. Panchenko, R. Beaufort, H. Naets, and C. Fairon, "Towards detection of child sexual abuse media: Categorization of the associated filenames," in *European Conference on Advances in Information Retrieval*, 2013, pp. 776–779.

[16] A. Lopes, S. Avila, A. Peixoto, R. Oliveira, and A. Araújo, "A bag-of-features approach based on hue-SIFT descriptor for nude detection," in *European Signal Processing Conference*, 2009, pp. 1152–1156.

[17] N. Kawale and S. Patil, "An approach to maintain the stroage of contentious image in the form of descriptor," in *IEEE International Conference on Computational Intelligence and Computing Research*, 2014, pp. 1–6.

[18] W. Kelly, A. Donnellan, and D. Molloy, "Screening for objectionable images: A review of skin detection techniques," in *IEEE International Machine Vision and Image Processing Conference*, 2008, pp. 151–158.

[19] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, 2003, pp. 1470–1477.

[20] M. Grega, D. Bryk, and M. Napora, "INACT – INDECT Advanced Image Cataloguing Tool," *Multimedia Tools and Applications*, vol. 68, no. 1, pp. 95–110, 2014.

[21] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[22] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[23] D. Lowe, "Distinctive image features from scale-invariant keypoints," vol. 60, no. 2, 2004, pp. 91–110.

[24] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *IEEE Computer Vision and Pattern Recognition*, 2010, pp. 2559–2566.

[25] S. Avila, N. Thome, M. Cord, E. Valle, and A. Araújo, "BOSSA: Extended BoW formalism for image classification," in *IEEE International Conference on Image Processing*, 2011, pp. 2909–2912.

[26] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. 40, no. 3, pp. 1106–1122, 2007.

[27] J. Kovac, P. Peer, and F. Solina, "Human skin color clustering for face detection," in *IEEE International Conference on Computer as a Tool*, 2003, pp. 144–148.

[28] "Media Detective," www.mediadetective.com.

[29] "Snitch Plus," www.hyperdynesoftware.com.

[30] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Good practice in large-scale learning for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 507–520, 2014.

[31] J. Wu, "Power mean SVM for large scale visual classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2344–2351.