Um Método Para Detecção de Ataques a Sistemas Biométricos Faciais Baseado em *Deep Learning*

Rodrigo Bresan[1], Carlos Beluzo[1] e Tiago Carvalho[1]

[1]Instituto Federal de São Paulo
rcbresan@gmail.com, {cbeluzo, tiagojc}@ifsp.edu.br

Abstract

The usage of biometric devices in order to authenticate users into systems or devices such as smartphones or ATMs is an increasingly adopted practice in people's lives. The objective of this work is to develop a methodology for Presentation Attack Detection (PAD) using Deep Learning techniques. Using Convolutional Neural Networks (CNN), the methodology proposed in this work extracts features associated with lighting, depth and segmentation of the videos, and uses a Support Vector Machine (SVM) classifier, in order to distinguish between a real biometric sample and a synthetic one.

Resumo

A utilização de dispositivos biométricos para a autenticação de usuários em sistemas ou dispositivos como, por exemplo, smartphones ou caixas eletrônicos é uma prática adotada cada vez mais na vida das pessoas. O objetivo deste trabalho é desenvolver uma metodologia para a Detecção de Ataques de Apresentação – PAD (Presentation Attack Detection) utilizando técnicas de Deep Learning. Utilizando Redes Neurais Convolucionais (CNN), o modelo proposto neste trabalho extrai características associadas a iluminação, profundidade e segmentação dos vídeos, e utiliza um classificador do tipo Máquina de Vetor de Suporte – SVM (Support Vector Machine), de modo a permitir a distinção entre uma amostra biométrica real e uma sintética.

1.1. Introdução

O uso de sistemas biométricos se dá de maneira cada vez mais comum no dia a dia das pessoas, como por exemplo para a autenticação de usuários em sistemas ou dispositivos, bem como também para o acesso a áreas restritas em organizações. Neste cenário, surgem também técnicas que visam fraudar estes mecanismos de autenticação. Ataques em sistemas de reconhecimento facial são comumente realizados por meio da apresentação de elementos que sintetizem a fisionomia do usuário à câmera de captura do sistema de autenticação. Isto pode ser realizado pela reprodução de um vídeo real do usuário, exibido através de um *tablet* ou um *smartphone*, ou ainda pela simples apresentação de uma foto impressa do usuário em alta qualidade [d. S. Pinto et al., 2012].

Assim, faz-se necessária a criação de métodos capazes de detectar e prevenir tais tipos de ataques, conhecidos na literatura como Ataques de Apresentação (do inglês *Pre-*

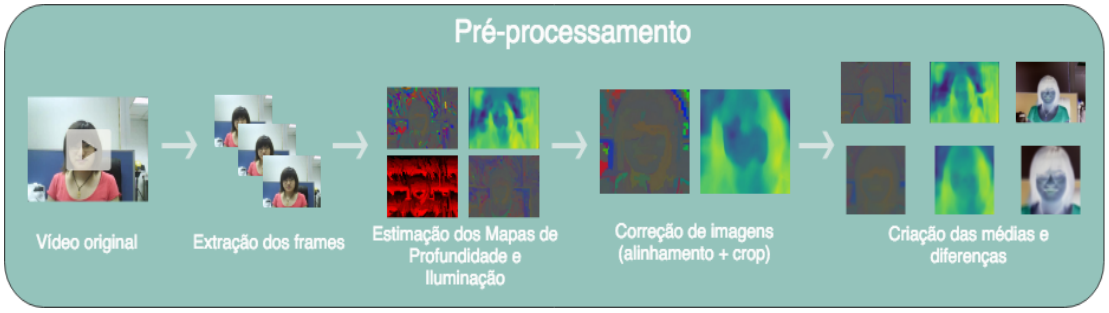

Figure 1.1. Etapas do Pré-processamento dos *datasets*

sentation Attacks). A solução proposta neste trabalho propõe a utilização de características intrínsecas ao ambiente, como profundidade e iluminação, juntamente com o uso de Redes Neurais Convolucionais (CNN) e de classificadores robustos como SVM (*Support Vector Machines*) para a detecção destes tipos de ataque.

1.2. Metodologia

O método proposto neste trabalho pode ser dividido em quatro etapas principais: (1) pré-processamento, (2) extração de características, (3) classificação das amostras e, (4) validação do modelo.

1.2.1. Pré-processamento das amostras de vídeo

A etapa de pré-processamento consiste em realizar a padronização dos dados, para que estes possam ser utilizados nas etapas seguintes, possibilitando a extração de características que possuam relevância no contexto de técnicas de ataque, viabilizando então a criação de um modelo que seja capaz de realizar tal distinção com eficácia. Na Figura 1.1, temos um diagrama ilustrando o fluxo de pré-processamento dos dados.

1.2.1.1. Mapas de profundidade

Uma das etapas fundamentais que constituem o pré-processamento da nossa base de dados é a inferência da profundidade de cada imagem existente, uma vez que esta serve de característica para a criação de padrões que levem à identificação de tentativas de ataque, visto que em uma situação de ataque, o dispositivo de reprodução de ataque muitas vezes é um *tablet* ou uma folha impressa, este que após a inferência de profundidade, resulta em um mapa no qual todos os pontos ao redor da região da face apresentam o mesmo nível de profundidade (*z-index*).

Uma das implementações utilizadas para realizar tal estimativa se dá através do uso do projeto *Monodepth* (*Monocular Depth Prediction*), no qual através de um modelo já pré treinado, realiza a inferência de profundidade de uma determinada imagem, através do aprendizado não supervisionado, portanto sem haver a necessidade do mapa de profundidade real (*Ground Truth*) [Godard et al., 2017].

1.2.1.2. Mapas de iluminação

Também são gerados durante a etapa de pré-processamento os mapas de iluminação de uma determinada imagem. Isso porque outra característica que será levada em consideração para a criação de nosso modelo é o modo como a luz incide sobre os objetos presentes na cena, uma vez que durante uma tentativa de ataque, seja através de uma impressão em alta qualidade do rosto da pessoa, ou através da exibição de um vídeo via *tablet*, tal reflexão se apresenta de maneira divergente das demais presentes no resto do ambiente [d. Carvalho et al., 2013].

1.2.1.3. Correção das imagens

Considerando que algumas imagens apresentam inclinações diferentes entre si, quando tomadas como base o nível dos olhos, se faz necessária a correção destas, para que assim fiquem todas devidamente alinhadas. Nesta fase, se faz necessário também realizar um recorte na região da face, de modo a eliminar atributos que sejam irrelevantes à elaboração do modelo. Ambas as transformações serão realizadas através de bibliotecas como o *OpenCV*.

1.2.1.4. Criação das médias e diferenças

Após estimados todos os mapas necessários, é realizada a etapa final do pré-processamento, no qual é gerado um novo artefato resultante da média aritmética de todas as imagens de um determinado vídeo, para posteriormente passar pelo processo no qual há a extração das características e finalmente ser utilizado em um classificador do tipo SVM.

1.2.2. Extração de Características e Classificação

As etapas de extração de características e da criação do modelo se darão através da utilização de:

- Redes Neurais Convolucionais (CNN) associadas a abordagens de *transfer learning* para a extração do vetor de características;
- SVM (*Support Vector Machines*) para a etapa de classificação.

1.2.3. Métricas de Validação

O uso da métrica *Half Total Error Rate* (HTER) se faz altamente relevante no que diz respeito à classificação de sistemas de biometria em aprendizado de máquina, uma vez que estes muitas das vezes apresentam uma quantidade de amostras muito inferiores em quantidade de tentativas de acesso fraudulentas. Tal desbalanceamento é uma das principais razões para o uso deste cálculo para a acurácia do sistema [Bengio and Mariéthoz, 2004]. A métrica HTER servirá como um indicador da performance do modelo proposto neste trabalho.

A validação do método se dará através de dois protocolos: *intra-dataset* e *inter-dataset*. No primeiro, serão utilizados os protocolos definidos pelos próprios autores da

base de dados, enquanto que no segundo, será realizado uso de um dataset para a fase de treino e outro distinto para a fase de teste.

1.3. Bases de dados

Neste trabalho, duas bases de dados de acesso público foram selecionadas para serem utilizadas nos experimentos. Uma breve descrição de cada uma delas pode ser encontrada nas seções a seguir:

1.3.1. Dataset CSBR

Contendo um total de 600 vídeos, estes subdivididos em 12 cenários, onde cada cenário é composto por 3 vídeos reais e 9 vídeos falsos de um mesmo indivíduo. Os vídeos reais foram coletados em 3 diferentes resoluções, classificadas em baixa, média e alta. Os vídeos falsos foram criados a partir dos vídeos reais em alta resolução e foram criados simulando 3 tipos de ataques: ataque com foto distorcida, ataque de foto cortada e ataque com vídeo [Zhang et al., 2012].

1.3.2. Dataset NUAA

O *dataset NUAA Photograph Imposter Database* - NUAA, é uma coleção de dados contendo imagens geradas a partir de *frames* de vídeos capturados por *webcams* simples e genéricas, sem nenhuma característica específica relevante. Os vídeos foram capturados a partir de 15 diferentes indivíduos, os quais estão distribuídos aleatoriamente em ambos os sexos, assim como utilizando óculos ou não. Para cada indivíduo foram capturados vídeos reais, e vídeos falsos, onde era apresentada à webcam uma foto colorida do indivíduo, todas com a mesma resolução (640 x 480 pixels). O *dataset* possui 5105 imagens gerados de frames de vídeos reais, e 7509 imagens geradas de frames de vídeos falsos [Tan et al., 2010].

1.4. Resultados Preliminares

O trabalho encontra-se em um estágio intermediário, sendo as etapas de pré-processamento já concluídas. Abaixo descrevemos de maneira breve as bases de dados escolhidas para a avaliação do método às quais a etapas de pré-processamento já foram aplicadas.

1.5. Considerações Finais

Ao final deste trabalho espera-se obter uma solução que consiga realizar a detecção de ataques de apresentação com uma eficiência aceitável, e de baixo custo, já que não serão necessários componentes de hardware adicionais, tais como sensores de profundidade e iluminação. Além disso, espera-se que a técnica que será proposta, o modelo implementado, e a respectiva avaliação de seu desempenho nas bases selecionadas, possam trazer contribuições relevantes à área de reconhecimento de padrões. Por fim, a solução implementada poderá ainda ser avaliada utilizando em sua execução outras bases de dados.

1.6. Agradecimentos

Gostaríamos de agradecer o suporte financeiro do projeto Capes DeepEyes, ao Programa Institucional Voluntário de Iniciação Científica e/ou Tecnológica do IFSP/Campinas -

PIVICT (Processo SUAP 23305.004801.2018-52), à Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP De ja'Vu (Concessão 2017/12646-3), FAPESP (Concessão 2017/12631-6, 2018/00858- 9), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (423797/2016-6), e à NVIDIA pela doação das GPUs utilizadas nesta pesquisa.

Referências

[Bengio and Mariéthoz, 2004] Bengio, S. and Mariéthoz, J. (2004). A statistical significance test for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*.

[d. Carvalho et al., 2013] d. Carvalho, T. J., Riess, C., Angelopoulou, E., Pedrini, H., and d. R. Rocha, A. (2013). Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security*, 8(7):1182–1194.

[d. S. Pinto et al., 2012] d. S. Pinto, A., Pedrini, H., Schwartz, W., and Rocha, A. (2012). Video-based face spoofing detection through visual rhythm analysis. In *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 221–228.

[Godard et al., 2017] Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *CVPR*.

[Tan et al., 2010] Tan, X., Li, Y., Liu, J., and Jiang, L. (2010). Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *Proceedings of the 11th European Conference on Computer Vision: Part VI*, ECCV'10, pages 504–517, Berlin, Heidelberg. Springer-Verlag.

[Zhang et al., 2012] Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., and Li, S. Z. (2012). A face antispoofing database with diverse attacks. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 26–31.