

Tiago de Freitas Pereira

ESTUDO COMPARATIVO DE CONTRAMEDIDAS PARA DETECÇÃO DE ATAQUES DE SPOOFING
A SISTEMAS DE AUTENTICAÇÃO FACIAL

Campinas
2012

Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Tiago de Freitas Pereira

ESTUDO COMPARATIVO DE CONTRAMEDIDAS PARA DETECÇÃO DE ATAQUES DE SPOOFING
A SISTEMAS DE AUTENTICAÇÃO FACIAL

Qualificação de Mestrado apresentada na Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.
Área de concentração: Computação

Orientador: Professor Doutor José Mario De Martino

Este exemplar corresponde a versão final do exame de qualificação apresentado pelo aluno, e orientado pelo Prof. Dr. José Mario De Martino

Campinas
2012

Resumo

Autenticação de usuários é uma tarefa crucial para proteger informações e neste área a biometria de face apresenta algumas vantagens. A biometria de face é natural, fácil de interagir e é uma das biometrias que possui um processo de coleta menos invasiva. Trabalhos recentes tem revelado que a biometria de face é vulnerável a ataques de *spoofing* utilizando que equipamentos baratos. Contramedidas tem sido propostas para mitigar este tipo de vulnerabilidade. Porém uma boa parte das contramedidas apresentadas na literatura são avaliadas utilizando métricas distintas e muitas vezes em bases de dados privadas impossibilitando uma comparação honesta das mesmas. Este projeto de mestrado tem como objetivo prover uma metodologia para avaliação de contramedidas para ataques de *spoofing* para sistemas de autenticação facial.

Palavras-chave: Antispoofing, Detecção de vitalidade, Contramedidas, Reconhecimento Facial, Biometria

Abstract

User authentication is an important step to protect information and in this field face biometrics is advantageous. Face biometrics is natural, easy to use and less human-invasive. Unfortunately, recent work has revealed that face biometrics is vulnerable to spoofing attacks using low-tech equipments. Countermeasures have been proposed in order to mitigate this vulnerabilities. However several works in the literature present evaluations using different metrics and in private database making the comparison of countermeasures a difficult task. The main goal of this masters project is to provide a clean methodology to evaluate countermeasures to face spoofing attacks.

Key-words: Antispoofing, Liveness detection, Countermeasure, Face Recognition, Biometrics

Lista de Figuras

1.1	Token em formato de chaveiro	18
1.2	Fluxograma básico de um sistema de autenticação biométrica e os seus possíveis pontos de ataque.	20
2.1	Fluxo dos dados no processo de autenticação de faces	24
2.2	Ataques efetuados com fotografias na base de dados NUAA	24
2.3	Algumas capturas da base de dados Replay Attack	26
2.4	Exemplos de acessos reais e ataques da base de dados CASIA FASD	27
2.5	Seleção de partes faciais	29
2.6	Fluxo dos dados da contramedida baseada em LBP	30
2.7	Representação da extração de parâmetros utilizando <i>LBP – TOP</i>	31
2.8	Fluxo dos dados da contramedida baseada em filtros DoG	32
4.1	Cuvas ROC de cada contra medida	39

Lista de Tabelas

1.1	Comparação das características biométricas mais utilizadas	19
2.1	Número de vídeos em cada subconjunto da base de dados	26
2.2	Performance em termos de <i>HTER(%)</i> da contramedida proposta pela (REF) nas três principais bases de dados de referência.	30
4.1	Performance das três contramedidas utilizando o Protocolo de Avaliação Intra Teste	38
4.2	Performance das três contramedidas aplicando o Protocolo de Avaliação Inter Teste	38

Lista de Acrônimos e Notação

<i>AUC</i>	<i>Area Under the Curve</i>
<i>DoG</i>	<i>Difference of Gaussians</i>
<i>EER</i>	<i>Equal Error Rate</i>
<i>FAR</i>	<i>False Acceptance Rate</i>
<i>FRR</i>	<i>False Rejection Rate</i>
<i>GLCM</i>	<i>Gray Level Co-occurrence Matrix</i>
<i>HMM</i>	Hidden Markov Model
<i>HTER</i>	<i>Half Total Error Rate</i>
<i>LBP</i>	<i>Local Binary Patterns</i>
<i>LBP – TOP</i>	<i>Local Binary Patterns from Three Orthogonal Planes</i>
<i>MLP</i>	<i>Multi Layer Perceptron</i>
<i>PLS</i>	<i>Partial Least Square</i>
<i>ROC</i>	<i>Receiver Operating Characteristics</i>
<i>SVM</i>	<i>Support Vector Machines</i>

Sumário

1	Introdução	17
1.1	Contextualização e Motivação	17
1.1.1	Ataque de <i>Replay</i>	19
1.1.2	Ataque na Referência Biométrica	20
1.1.3	Ataque <i>man-in-the-middle</i>	20
1.1.4	Ataque de Spoofing	21
1.2	Objetivos	21
1.3	Organização do Trabalho	21
2	Revisão da Literatura	23
2.1	Bases de Dados de Referência	23
2.1.1	NUAA	24
2.1.2	Replay Attack	25
2.1.3	CASIA FASD	26
2.2	Spoofing em Reconhecimento de Face	27
2.2.1	Presença de Vitalidade	28
2.2.2	Características da Cena	29
2.2.3	Discrepância Relativa à Qualidade da Imagem	30
2.3	Considerações Finais	32
3	Metodologia de Avaliação	35
3.1	Medidas de Desempenho	35
3.2	Protocolo de Avaliação Intra Base de Dados	36
3.3	Protocolo de avaliação Inter-base de dados	36
4	Resultados Parciais e Plano de Trabalho	37
4.1	Protocolo de Avaliação Intra Base de Dados	37
4.2	Protocolo de Avaliação Inter Base de Dados	38
4.3	Plano de Trabalho	39
5	Conclusões Parciais	41

Introdução

1.1 Contextualização e Motivação

Em uma sociedade moderna, o processo de autenticação é uma tarefa importante para proteger dados e recursos sejam ele físicos ou digitais. Consistindo da confirmação de uma identidade requerida, o processo de autenticação é o primeiro e o mais crítico na cadeia de segurança restringindo acesso a usuários não autorizados.

Para a tarefa de confirmação de uma identidade, utilizam-se elementos que devem corresponder unívocamente ao identificador associado a um determinado usuário. Estes elementos são chamados de fatores de autenticação. Centralizados no usuário que está requerendo a identidade, estes fatores podem ser utilizados isoladamente ou combinados a fim de reforçar a segurança. Os fatores de autenticação são classificados em aquilo que o usuário:

- **Sabe:** Por exemplo, uma senha ou uma frase de segurança;
- **Possui:** Por exemplo, um *token* de segurança, uma chave de cadeado ou um cartão;
- **É:** Por exemplo, uma característica física ou comportamental.

Cada um destes fatores apresentados possui um conjunto de vantagens e desvantagens. O uso mais comum de senhas, é para o acesso lógico a sistemas computacionais (computadores, e-mail, banco, cartão de crédito e muitos outros). A senha possui a vantagem de ser naturalmente imutável ao longo do tempo, ou seja, caso a mesma não seja mudada, ela continuará tendo o mesmo valor ao longo do tempo. Senhas contudo podem ser tão complexas quanto se queira, ficando a critério de seu detentor criar uma senha que ao mesmo tempo seja segura (de difícil adivinhação para um eventual ataque) e de fácil memorização. Estes critérios, claramente antagônicos, são o principal ponto de ataque a sistemas computacionais baseados em autenticação com senhas. Como um exemplo de vulnerabilidade no uso de senhas, em fevereiro de 2012 78 contas de e-mail de membros do governo da Síria foram invadidas divulgando informações confidenciais¹. Destas 78 contas de e-mail, 33 a senha era '12345' ou '123456' incluindo a senha do próprio presidente.

¹<http://www.dailymail.co.uk/news/article-2100111/New-York-spin-doctor-coached-Syrian-dictator-Assad-swing-sympathies-US-public.html>

Tokens geralmente são associados a um segundo fator de autenticação. Como exemplo, um cartão de crédito é um *token* que acompanhado de um segundo fator, como a senha do cartão, reforça a segurança de transações financeiras. Há bancos que disponibilizam para seus clientes tokens que geram números aleatórios a cada 30 segundos com a finalidade de reforçar o acesso à serviços de *internet banking*. Na Figura 1.1) pode-se observar um token em formato de chaveiro. Estes *tokens* são objetos físicos que podem ser facilmente perdidos, e uma vez perdidos a autenticação fica comprometida.



Figura 1.1: Token em formato de chaveiro

Biometria é a ciência de reconhecer a identidade de uma pessoa baseada em seus atributos físicos e/ou comportamentais, tais como a face, as impressões digitais, veias da mão, voz e a íris (?). O uso da biometria como fator de autenticação possui algumas vantagens. Naturalmente, não é possível esquecer uma característica biométrica e dificilmente esta característica desaparece repentinamente (talvez em casos de acidentes graves). Características biométricas são intrínsecas a pessoa que as possui e portanto é intransferível. Como desvantagem, a biometria pode variar ao longo do tempo. Como exemplo, a voz humana sofre variações quando estamos doentes ou em situações de *stress*; nossos traços faciais infelizmente mudam a medida que envelhecemos. Vale ressaltar que métodos de autenticação baseados em biometria são probabilísticos, ou seja, pode ser que o sistema de autenticação rejeite uma entrada autêntica devido à uma série de fatores externos.

As características humanas para serem utilizadas em uma método de autenticação biométrica devem satisfazer alguns requisitos, dentre eles destacam-se:

- Universalidade (toda pessoa deve possuí-la);
- Unicidade (deve permitir distinguir as pessoas);
- Estabilidade (não deve se alterar demasiadamente ao longo do tempo);
- Coletabilidade (deve poder ser medida quantitativamente);
- Desempenho (deve possibilitar um reconhecimento preciso, em tempo hábil);

- Aceitabilidade (deve ser aceitos facilmente por seus usuários);
- Circunvenção (deve dificultar a possibilidade de fraudes).

A Tabela 1.1 apresenta um comparativo realizado por (?) entre as características biométricas mais utilizadas. É possível observar que nenhuma das biometrias apresentadas consegue atender todos estes requisitos com excelência e a escolha de qual utilizar deve levar em conta a natureza e as exigências de cada aplicação (?).

Tabela 1.1: Comparaçao das características biométricas mais utilizadas

Característica	Universalidade	Unicidade	Estabilidade	Coletabilidade	Desempenho	Aceitabilidade	Circunvenção
Face	Alta	Baixa	Média	Alta	Baixa	Alta	Baixa
Impressão Digital	Média	Alta	Alta	Média	Alta	Média	Média
Geometria das mãos	Média	Média	Média	Alta	Média	Média	Média
Veias da mão/dedo	Média	Média	Média	Média	Média	Média	Alta
Íris	Alta	Alta	Alta	Média	Alta	Baixa	Alta
Assinatura	Baixa	Baixa	Baixa	Alta	Baixa	Alta	Baixa
Voz	Média	Baixa	Baixa	Média	Baixa	Alta	Baixa

Sistemas de autenticação biométrica podem ser grosseiramente representados segundo o fluxograma da Figura 1.2.

Primeiramente o trato biométrico é capturado via algum tipo de **sensor**. Após esta captura, o trato biométrico capturado é **processado** a fim de extrair as características biométricas e geração da referência biométrica. Quando se está efetuando o cadastro de uma referência biométrica, estas características são **armazenadas** em uma base de dados para acessos futuros. Quando se está efetuando a **autenticação**, estas características biométricas serão utilizadas no processo de comparação com alguma identidade requerida no banco de dados. Conforme pode ser observado na mesma Figura 1.2, ataques podem ser efetuados em qualquer ponto da arquitetura (?). As próximas subseções irão discorrer sobre cada um dos possíveis ataques e soluções para mitigá-los.

1.1.1 Ataque de *Replay*

O ataque de replay consiste da utilização de dados previamente submetidos da identidade alvo para o sistema de autenticação a fim de obter o acesso não autorizado. Estes dados podem ser obtidos interceptando (*sniffing*) o canal de dados entre o sensor e a unidade de processamento de dados biométricos durante uma autenticação bem sucedida da identidade alvo. Para deter ataques dessa natureza, o sistema de autenticação biométrica deve assegurar

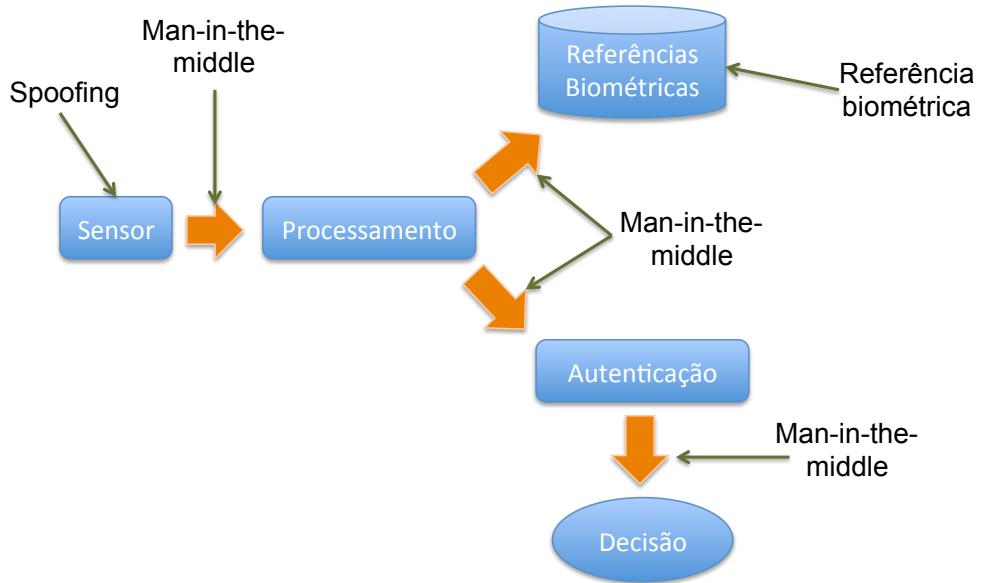


Figura 1.2: Fluxograma básico de um sistema de autenticação biométrica e os seus possíveis pontos de ataque.

que o dado fornecido não foi injetado artificialmente. Uma forma de mitigar a incidência deste tipo de ataque é fazer uso da característica probabilística da própria biometria. É praticamente impossível dois processos de captura independentes e em intervalos de tempo distintos gerar exatamente o mesmo dado biométrico. Se isto ocorrer é provável que este dado foi interceptado e está sendo injetado no sistema de autenticação (?).

1.1.2 Ataque na Referência Biométrica

O ataque nas referências biométricas consiste em atacar o seu local de armazenamento. Com este tipo de ataque pode-se: adicionar uma biometria falsa no sistema de armazenamento, copiar as referências biométricas armazenadas, remover alguma referência biométrica ou modificar as referências biométricas existentes (?). Dentre estas possibilidades a mais perigosa é a cópia de referências biométricas, pois as mesmas podem ser usadas, através de engenharia reversa, para gerar biometrias falsas. (?) demonstrou que é possível gerar impressões digitais falsas através do processo de engenharia reversa com referências biométricas baseadas em minúcias. Com estas impressões digitais fabricadas, foi possível violar um sistema de autenticação baseado em impressões digitais.

1.1.3 Ataque *man-in-the-middle*

O ataque *man-in-the-middle* ou homem do meio é uma forma de ataque em que os dados trocados entre os componentes do sistema de biometria são, de alguma forma, interceptados e alterados pelo atacante. Neste tipo de ataque o atacante pode: interceptar dados sensor, interceptar dados enviados para armazenamento e interceptar dados de decisão do sistema de

biometria. Mecanismos como encriptação dos dados antes de serem transmitidos e/ou prover canais de comunicação seguro podem mitigar este tipo de ataque.

1.1.4 Ataque de Spoofing

O ataque de *spoofing* em sistemas de autenticação biométrica, é um tipo de ataque em que o atacante forja o trato biométrico alvo apresentando uma biometria falsa ao sensor, burlando o sistema de autenticação. Em sistemas de autenticação baseados em biometria há duas motivações para se forjar um trato biométrico. A **primeira** motivação é o atacante obter o trato biométrico de outra pessoa a fim de tomar sua identidade para obter privilégios. Em sistemas de autenticação baseados na voz, o atacante pode gravar a voz da identidade alvo e usar esta gravação como entrada em sistema de autenticação baseado na biometria da voz. Em sistemas de autenticação baseados em impressões digitais o atacante pode obter alguma impressão latente da identidade alvo e gerar um dedo artificial contento a impressão digital roubada. A saber em (?), (?) e (?) são trabalhos relacionados a ataques de *spoofing* em sistemas de autenticação baseados em impressões digitais, em (?), (?) e (?) são trabalhos relacionados a ataques de *spoofing* baseados na biometria da iris e em (?) e (?) são trabalhos relacionados a ataques de *spoofing* em sistemas de autenticação baseados na biometria de locutor. A **segunda** motivação é um atacante gerar um trato biométrico totalmente artificial (sem se basear em uma biometria real), a fim de enganar o sistema de cadastro e autenticação biométricos. Com isso, o atacante pode compartilhar esta biometria falsa com outros atacantes.

Melhores práticas de segurança orientam utilizar mecanismos como: criptografia de dados e criação de canais seguros para mitigar ataques na maioria dos casos citados anteriormente (?). No caso dos ataques de *spoofing*, o sensor de biometria (ponto alvo deste tipo de ataque) é o único ponto do fluxograma da Figura 1.2 em que nenhum dos mecanismos são efetivos para mitigá-los, tornando-se assim o ponto mais frágil a ataques. Por esta razão, ataques dessa natureza serão o ponto central desta dissertação.

1.2 Objetivos

Este trabalho tem como objetivo desenvolver uma metodologia para avaliar contramedidas para detecção de ataques de *spoofing* em sistemas de autenticação de face.

1.3 Organização do Trabalho

Este trabalho está organizado na forma que segue. No Capítulo 2 são apresentados os conceitos base para esta pesquisa e uma breve revisão da literatura. Na Seção 2.1 são apresentadas as principais bases de dados de referência. Na Seção 2.2 são apresentadas algumas contramedidas publicadas na literatura.

Já no Capítulo 3 é apresentado a metodologia de avaliação do projeto de mestrado em questão.

O Capítulo 4 apresenta os resultados parciais obtidos com a aplicação da metodologia apresentada no Capítulo 3.

Por fim, no Capítulo 5 são apresentadas as conclusões parciais obtidas no projeto de mestrado em questão.

Capítulo 2

Revisão da Literatura

Por sua natureza não intrusiva, autenticação utilizando a biometria da face é uma das áreas mais ativas e desafiadoras no campo da biometria. Apesar do significativo progresso da tecnologia de reconhecimento facial nas últimas décadas em uma série de tópicos como o envelhecimento dos indivíduos e reconhecimento em cenários de iluminação complexa ainda são desafios de pesquisa na área. Avanços na área foram amplamente relatados em (?) e em (?). No entanto, a tarefa de verificar se o rosto apresentado a uma câmera é realmente um rosto de uma pessoa real, e não uma tentativa de forjar uma identidade (*spoof*) tem sido quase sempre esquecido. A Figura 2.1 apresenta dois fluxos de execução. O primeiro (a) apresenta o processo de aquisição da biometria da face em um acesso real. Já o segundo fluxo apresenta o processo de aquisição da biometria da face em uma tentativa de ataque de *spoofing*.

Recentemente, a mídia tem documentado algumas situações de ataques de *spoofing* em sistemas de reconhecimento de face em produção no mercado. Usando fotografias simples, um grupo de pesquisa da Universidade de Hanói mostrou que, com relativa facilidade, é possível burlar os sistemas de autenticação face em produção nos laptops Lenovo, Asus e Toshiba (?). Desde o lançamento da versão *Ice Cream Sandwich*, o sistema operacional Android vêm com um sistema de autenticação de face embarcado com a finalidade de desbloquear o uso do celular. Desde então, tem sido amplamente demonstrado em toda a WEB como é possível burlar esta barreira de autenticação. Como resposta, uma contramedida baseada no piscar de olhos, foi introduzida na versão mais recente do sistema operacional Android.

Este Capítulo está organizado da seguinte maneira: Na Seção 2.1 serão apresentadas as principais bases de dados de referência para o estudo de ataques de *spoofing* em sistemas de reconhecimento facial. A Seção 2.2 apresenta uma sucinta revisão da literatura apresentando algumas estratégias para lidar com o problema. Por fim na Seção 2.3 apresenta as considerações finais do Capítulo.

2.1 Bases de Dados de Referência

Com o surgimento de pesquisas relacionadas a ataques de *spoofing*, alguma bases de dados utilizadas para validar métodos propostos foram surgindo. Nesta Seção algumas dessas bases de dados serão descritas.

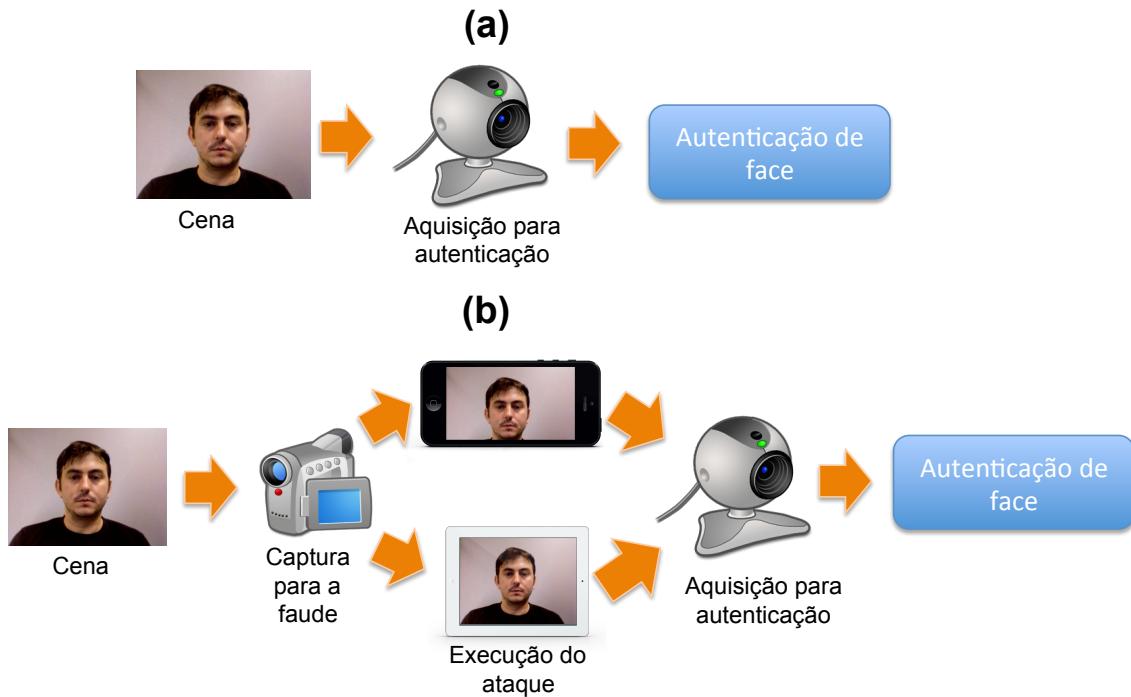


Figura 2.1: (a) Fluxo da informação em um acesso real (b) Fluxo da informação em um ataque de spoofing

2.1.1 NUAA

Construída para estudar o cenário de ataques utilizando fotografias impressas em papel, a base de dados NUAA¹ consiste de capturas de fotografias de pessoas em frente à câmera de um *notebook* e de ataques feitos à mesma câmera com fotos impressas de alta qualidade destas mesmas pessoas. A base de dados possui gravações de 15 pessoas distintas divididas em 3 seções espaçadas em duas semanas e cada seção possui quatro gravações por pessoa com condições de iluminação distintas. Na Figura 2.2 são apresentados exemplos dessa base de dados.



Figura 2.2: Ataques efetuados com fotografias na base de dados NUAA

¹<http://parnec.nuaa.edu.cn/xtan/data/NuaaImposterdb.html>

2.1.2 Replay Attack

A base de dados Replay Attack (?) consiste de gravações de vídeos de curta duração ($\sim 10s$) tanto de simulações de acessos reais e quanto simulações de ataques em 50 identidades diferentes, utilizando um computador portátil. Ela contém 1.200 vídeos sendo 200 contemplando simulações de acessos reais e 1000 contemplando simulações de ataques em três cenários diferentes com duas diferentes condições de iluminação e suporte. Os cenários de ataque incluem condições em que o atacante:

1. **print**: Exibe uma cópia da face da identidade alvo impressa em alta resolução com papel fotográfico de tamanho A4;
2. **mobile**: Exibe fotografias e vídeos obtidas com um telefone celular modelo iPhone 3GS;
3. **highdef**: Exibe fotografias e vídeos em alta resolução (1024×768) com um iPad.

As condições de iluminação incluem:

1. **controlled**: O fundo da cena é uniforme e iluminada com uma luz fluorescente;
2. **adverse**: O fundo da cena não é uniforme e é iluminada pela luz natural.

As condições de suporte incluem:

1. **hand-based**: O atacante segura a mídia de ataque com as mãos;
2. **fixed**: O atacante fixa a mídia de ataque em um suporte fixo impossibilitando qualquer tipo de movimento durante a tentativa de ataque.

A Figura 2.3 apresenta alguns exemplos de acessos reais e ataques em diferentes cenários. Na linha superior, pode-se observar amostras no cenário controlado e na linha inferior pode-se observar amostras no cenário com iluminação não controlada. Colunas da esquerda para a direita mostram exemplos de acesso real, ataques efetuados com fotografias impressas em papel, celular e *tablet* respectivamente. Na Tabela 2.1.2 são apresentadas as quantidades de vídeos para cada cenário descrito.

A base de dados Replay Attack fornece um protocolo para avaliar objetivamente uma dada contramedida. Tal protocolo define três conjuntos mutuamente exclusivos e são eles o conjunto de treino (train set), calibração (devel set) e o conjunto de teste (test set). O conjunto de treino deve ser utilizado para treinar uma contramedida, o conjunto de calibração é usado ajustar hiper-parâmetros de uma contramedida e para definir uma valor de limiar de detecção de ataques para ser utilizado no conjunto de teste que deve ser utilizado apenas para reportar resultados. Como medida de desempenho, o protocolo recomenda o uso da medida (*HTER*) que é definida como:

$$HTER = \frac{FAR(\tau, D) + FRR(\tau, D)}{2}, \quad (2.1)$$

onde τ é o limiar de detecção de ataques, D é um subconjunto da base de dados, *FAR* é a taxa de falsas aceitações e *FRR* é a taxa de falsas rejeições. Neste protocolo recomenda-se para o valor de τ o valor de *EER* obtido no conjunto de calibração (devel set).



Figura 2.3: Algumas capturas da base de dados Replay Attack (cortesia de (?)).

Tabela 2.1: Número de vídeos em cada subconjunto da base de dados. Células sinalizadas com o operador "+", indicam a quantidade de vídeos com suporte manual e com suporte fixo respectivamente.

Type	Train	Devel.	Test	Total
Real-access	60	60	80	200
Print-attack	30+30	30+30	40+40	100+100
Mobile-attack	60+60	60+60	80+80	200+200
Highdef-attack	60+60	60+60	80+80	200+200
Total	360	360	480	1200

2.1.3 CASIA FASD

Composta por 50 identidades, a base de dados CASIA FASD possui simulações de acessos reais e uma variedade de simulações de ataques. Tal variedade é obtida através três tipos de ataque em três tipos de resolução (baixa, normal e alta). Os tipos de ataques são: ataques impressos em papel em que o atacante deforma o papel a fim de gerar um ataque mais efetivo (*warped*), máscaras de papel em que o atacante a veste na execução do ataque (*cut*) e os ataques utilizando vídeos sendo exibidos utilizando um iPad. Pode ser observado na Figura 2.4 exemplos dessa base de dados. No total, o banco de dados consiste de 600 vídeos subdivididos em subconjuntos mutuamente exclusivos para treinamento e teste; 240 e 360, respectivamente.

O objetivo desta base de dados é investigar a efetividade de diferentes tipos de ataques com as suas respectivas resoluções. Para isso, a base de dados possuí um protocolo de avaliação composto de sete cenários. Para avaliar o impacto da resolução da imagem nos ataques, três cenários são descritos utilizando todos os tipos de ataque e são eles testes com: (1) baixa resolução, (2) resolução normal e (3) alta resolução. Para avaliar o impacto do tipo de ataque mais três cenários são descritos utilizando ataques de todas as resoluções e são eles testes com (4) ataques com fotografias impressas em papel, (5) ataques de utilizando máscaras de papel e (6) ataques com vídeo. O sétimo cenário consiste da avaliação utilizando toda a base de dados.

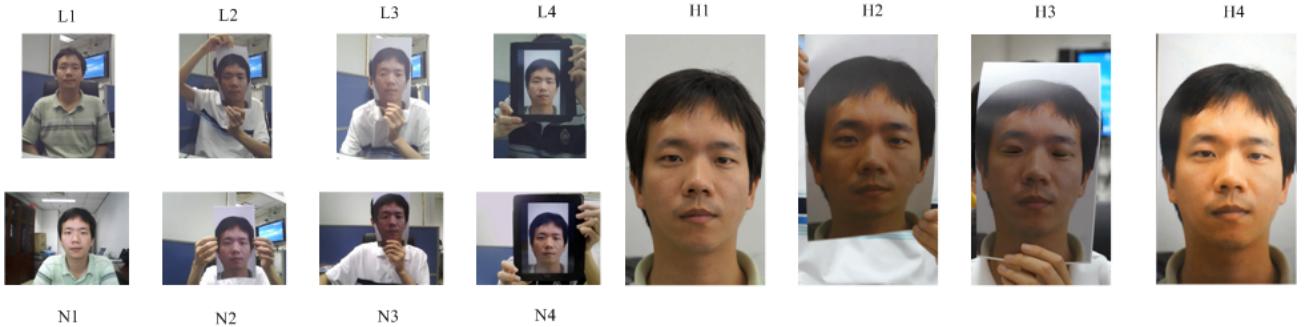


Figura 2.4: Exemplos de acessos reais e ataques da base de dados CASIA FASD (cortesia de (?))

Como métrica para reportar os resultados, recomenda-se o cálculo do EER no subconjunto de teste.

2.2 Spoofing em Reconhecimento de Face

Um sistema de autenticação baseado na biometria de face pode ser forjada de diversas maneiras (?) e são elas a apresentação para a câmera de:

- Fotos com a face do usuário alvo;
- Vídeos com a face do usuário alvo;
- Máscaras construídas a partir da face do usuário alvo;
- Maquiagem na tentativa de imitar a identidade do usuário alvo;
- Cirurgia plástica na tentativa de imitar a identidade do usuário alvo.

Embora seja possível falsificar um sistema de autenticação de face utilizando maquiagem, cirurgia plástica ou máscaras; fotografias e vídeos são provavelmente as ameaças mais comuns. Além disso, devido a crescente popularidade das redes sociais na WEB, (facebook, youtube, flickr, instagram e outros) uma grande quantidade de conteúdo multimídia, especialmente vídeos e fotografias, estão disponíveis e estes dados podem ser utilizados facilmente para atacar um sistema de autenticação de faces. Para mitigar os sucessos dos ataques dessa natureza, contramedidas eficazes devem ser pesquisadas e desenvolvidas.

Contramedidas para ataques de *spoofing* em reconhecimento de face podem ser classificados quanto à dependência da colaboração do usuário. Métodos que são ditos colaborativos, partem do princípio que a pessoa que está efetuando a autenticação deve favorecer o mesmo, executando alguma atividade do tipo desafio-resposta. Em (?) e (?) o usuário é orientado a falar um texto gerado automaticamente e os movimentos labiais são correlacionados com reconhecimento de fala a fim de gerar uma checagem forte acerca da presença de um usuário em frente à câmera.

Métodos que não são colaborativos, operam com imagens ou vídeos capturados por câmeras convencionais sem exigir uma interação com o usuário que está efetuando a autenticação. Uma

vantagem clara nas abordagens desse tipo é que a usabilidade de sistemas de autenticação de face não é onerada, já que o usuário não toma ciência de que uma checagem de sua presença em frente a câmera está sendo efetuada. Dada a vantagem descrita, métodos dessa natureza serão explorados neste trabalho.

Estratégias não colaborativas podem ser classificados em estratégias que exploram:

- Presença de vitalidade (*liveness detection*);
- Características da cena;
- Discrepância relativa a qualidade da imagem;

As próximas sub-seções apresentam cada uma das estratégias e os trabalhos relacionados a elas.

2.2.1 Presença de Vitalidade

Presença de vitalidade ou *liveness detection* consiste na seleção de características faciais que apenas pessoas vivas conseguem reproduzir.

O piscar de olhos é uma tarefa involuntária que os seres humanos executam constantemente. Um ser humano comum pisca de forma involuntária em média uma vez a cada 2 ou 6 segundos para manter os olhos limpos e umedecidos. Este intervalo pode variar drasticamente em situações de *stress* e/ou de alta concentração aumentando este intervalo para mais de 20 segundos. Contudo, não importa a situação de *stress* em que se está submetido; em algum momento este movimento irá ocorrer e não há estabelecido um limite máximo estabelecido em que um ser humano consegue suportar sem piscar os olhos. Apoiado nesta hipótese, (?) desenvolveu uma contramedida baseada no piscar dos olhos com o objetivo de bloquear ataques efetuados com fotografia. O sistema desenvolvido modela a piscadela utilizando cadeias escondidas de Markov (HMM) mapeando os estados de olho aberto para olho fechado e olho aberto novamente. Experimentos foram conduzidos utilizando uma base de dados criada pelos autores e livremente disponível para download² mostraram uma acurácia de 95,7% contra uma taxa de falsos positivos abaixo de 0,1%.

Apoiado na hipótese de que faces vivas apresentam padrões de movimento em certas regiões da face altamente descorrelacionados se comparados ataques, (?) desenvolveu uma heurística baseada em fluxo ótico para explorar tal característica. Como referência para a heurística foram selecionados a região do centro da face e das orelhas com pode ser observado na Figura 2.5.

A estratégia da contramedida pode ser sumarizada como segue:

1. Detectar a região da face;
2. Definir se região facial está se movendo mais horizontalmente ou mais verticalmente observando as velocidades do movimento;
3. Delimitar a região do centro da face e das orelhas (Figura 2.5);

²http://www.cs.zju.edu.cn/gpan/database/db_blink.html

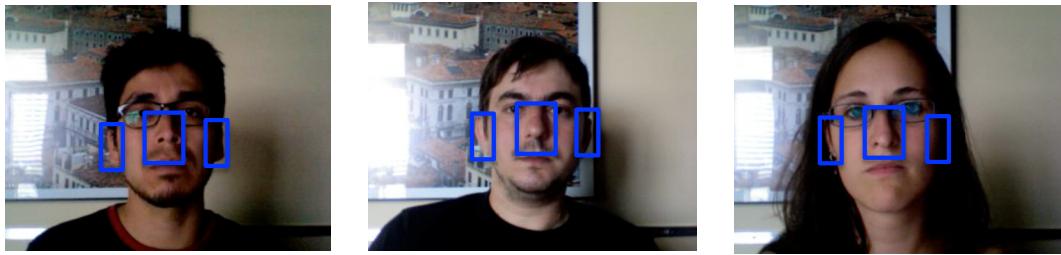


Figura 2.5: Seleção de partes faciais para a execução da heurística apresentada em (?)

4. Se o movimento for mais horizontal, computar a razão das velocidades dos movimentos das orelhas e do centro da face das componentes horizontais; caso contrário utilizar as componentes verticais;
5. Será considerado ataque se a razão das velocidades de movimento das orelhas e do centro da face for maior que um determinado limiar α .

A performance foi avaliada utilizando uma base de dados construída sobre a base de dados de face XM2VTS³. Os acessos reais foram gerados utilizando o subconjunto *Head Rotation Shot* desta base de dados e os ataques foram gerados com fotografias impressas em papel das mesmas imagens utilizadas para os acessos reais e regravados utilizando uma câmera de computador. Com esta base de dados criada um $EER = 0,5\%$ foi obtido. Esta base dados não foi disponibilizada publicamente pelos autores de modo que qualquer tentativa de reprodução dos resultados fica impossibilitada.

2.2.2 Características da Cena

Contramedidas que buscam características da cena buscam combinar a relação das características faciais com as características de onde a face está inserida.

A contramedida proposta por (?) mede a correlação do movimento da região facial em relação ao seu fundo. Como medida de movimento é utilizada uma simples diferença das intensidades dos pixels em quadros sucessivos. O movimento acumulado entre esta diferença (M_D), para um determinado *RoI* e seu respectivo fundo, é calculado usando a seguinte equação:

$$M_D = \frac{1}{S_D} \sum_{(x,y) \in D} |I_t(D) - I_{t-1}(D)|, \quad (2.2)$$

em que D é o *RoI*, S_D é a área do *RoI* e I_t é a intensidade de um pixel na imagem t .

Para introduzir o coeficiente de movimento em um classificador, 5 medidas são computadas em uma janela de n segundos. As medidas são as seguintes: o mínimo de M_D , o máximo, a média, o desvio padrão e a proporção R composta entre a soma de todos os componentes não-DC e DC (Direct Current) tomadas como base na transformada de Fourier do sinal gerado

³<http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/>

por M_D (ver Equação 2.3). Estas medidas são a entrada para uma rede Neural do tipo MLP (Multi-Layer Perceptron) a fim de detectar ataques.

$$R = \frac{\sum_{i=1}^N |FFT_i|}{|FFT_0|}. \quad (2.3)$$

Configurada com uma camada intermediária com 5 neurônios e considerando janelas de tempo com 20 quadros, esta contramedida foi avaliada utilizando o subconjunto de ataques de fotografia da base de dados Replay Attack (?) e apresentou $HTER = 9\%$.

2.2.3 Discrepância Relativa à Qualidade da Imagem

Contramedidas baseada na discrepância relativa à qualidade da imagem apoia-se na hipótese que o processo de amostragem e quantização de uma mídia de ataque (fotografias, vídeos e etc.) geram padrões de imagem degradados em relação a captura de pessoas reais.

Pela razão de possuir propriedades reflexivas distintas, mídias de ataque apresentam padrões distintos de faces reais. Apoiada nesta hipótese, (?) explora características de textura utilizando LBP analizando quadros individuais. A Figura 2.6 exibe o diagrama de blocos da contramedida proposta.

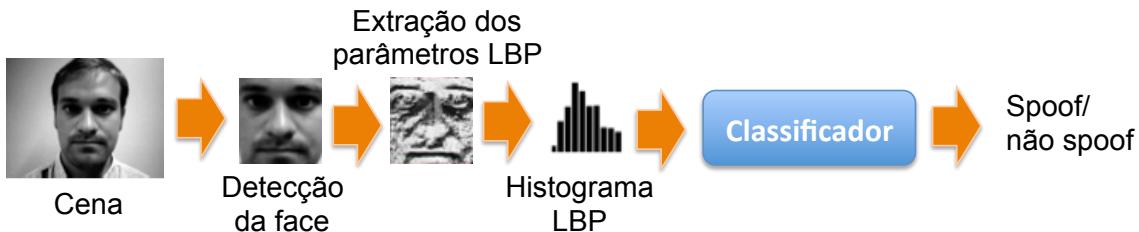


Figura 2.6: Fluxo dos dados da contramedida baseada em LBP

Neste trabalho, as faces são segmentadas e geometricamente normalizadas para 64×64 pixels. Em seguida os parâmetros LBP configurado seguindo a configuração $LBP_{8,1}^{u2}$ são extraídos e histogramados. Estes histogramas são a entrada do classificador que detecta ataques.

A Tabela 2.2 apresenta a performance do algoritmo em termos de HTER em três bases de dados de referência; a base de dados Replay Attack, a base de dados CASIA FASD e a base de dados NUAA utilizando SVM e LDA como classificadores.

Tabela 2.2: Performance em termos de $HTER(\%)$ da contramedida proposta pela (REF) nas três principais bases de dados de referência.

	Replay Attack		NUAA		CASIA-FASD	
	Conj. dev	Conj. test	Conj. dev	Conj. test	Conj. dev	Conj. test
$LBP_{8,1}^{u2} + LDA$	19,60	17,17	0,06	18,32	17,08	21,01
$LBP_{8,1}^{u2} + SVM$	14,84	15,16	0,11	19,03	16,00	18,17

Pode-se observar uma performance satisfatória nas três bases de dados entre $\sim 15\%$ e $\sim 20\%$. Contudo uma análise da performance nos conjuntos de desenvolvimento e teste na base de dados NUAA sugere uma baixa capacidade de generalização da contramedida.

Ainda analizando texturas, (?) propôs uma contramedida utilizando a dinâmica de uma textura ao longo do tempo utilizando o descritor $LBP - TOP$. Complementar ao descritor LBP , o descritor $LBP - TOP$ além de observar as componentes espaciais (direção X e Y), ele observa padrões de textura orientados no tempo (direção X e T e direção Y e T) como pode-se observar na Figura 2.7.

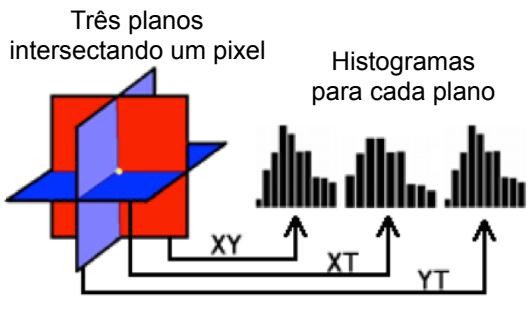


Figura 2.7: Representação da extração de parâmetros utilizando $LBP - TOP$

Neste trabalho, as faces são segmentadas e geometricamente normalizadas para 64×64 pixels. Em seguida os parâmetros $LBP - TOP$ seguindo a configuração $LBP - TOP_{8,8,8,1,1,1}^{u2}$ são extraídos e histogramados. Estes histogramas são a entrada de um classificador do tipo SVM que detecta ataques. Avaliado utilizando a base de dados Replay Attack, esta contramedida apresentou um $HTER = 7.60\%$ superando o trabalho apresentado por (?) em $\sim 50\%$ nesta base de dados.

Apoiados na hipótese que imagens/vídeos utilizadas para ataques concentram mais informação em uma banda específica de frequência, (?) apresenta uma contramedida explorando tal característica utilizando filtros de diferença de gaussianas (DoG).

Como pode ser observado no diagrama de blocos da Figura 2.8, quatro sequências de filtros DoG são aplicados na imagem. Cada filtro possuí uma máscara de 3×3 e as configurações das variâncias de cada filtro são:

- $\sigma_1 = 0,5$ e $\sigma_2 = 1$;
- $\sigma_1 = 1$ e $\sigma_2 = 1,5$;
- $\sigma_1 = 1,5$ e $\sigma_2 = 2$;
- $\sigma_1 = 1$ e $\sigma_2 = 2$.

Após a filtragem as imagens são reescaladas para 128×128 e estes dados são a entrada de um classificador SVM. Avaliada utilizando a base de dados CASIA FASD, esta contramedida apresentou um EER de 17%.



Figura 2.8: Fluxo dos dados da contramedida baseada em filtros DoG

Apoiado na hipótese de que as dimensões de uma face utilizando uma mídia para ataque são menores do que uma face real e que as variações de movimento facial em um ataque também são menores, (?) propôs uma contramedia analizando o espectro Fourier. A expectativa com esta análise é que as imagens utilizadas para ataque contém menos componentes de alta frequência do que imagens de acessos reais. Avaliado utilizando uma base de dados construída pelos próprios autores e não disponibilizada publicamente, obteve-se uma acurácia de 100% na detecção dos ataques.

Para detectar padrões de ruídos em ataques de spoofing, (?). desenvolveu uma contramedida analizando vídeos combinando diversos elementos. Primeiramente os quadros capturados são filtrados aplicando na sequência um filtro Gaussiano e um filtro da Mediana respectivamente. Estas imagens filtradas são subtraídas da imagem original obtendo o chamado ruído residual da imagem. Este ruído residual é analizado no domínio da frequência através da transformada de Fourier 2D. Todos os quadros de um vídeo capturado são combinados utilizando a técnica chamada Rítmico Visual (?) gerando uma imagem única caracterizando toda uma aquisição. Com esta etapa de pré-processamento concluída uma descrição utilizando Matriz de Co-ocorrência (GLCM) com 4 orientações são computadas. Uma matriz de co-ocorrência descreve a frequência de ocorrência de níveis de cinza entre pares de pixels. Através dessa matriz 12 medidas são extraídas para ser a entrada do classificador que detectará os ataques. Os classificadores avaliados foram o PLS e o SVM.

Com uma base de dados combinando o subconjunto de ataques utilizando fotografias da base de dados Replay Attack e uma base de dados criado pelos autores uma performance $\sim 100\%$ em termos de AUC foi obtida.

2.3 Considerações Finais

Neste capítulo foram apresentadas as principais bases de dados de referência para o estudo de ataques de *spoofing* para autenticação de faces e uma breve revisão das contramedidas apresentadas na literatura. É possível observar que as contramedidas apresentadas são avaliadas utilizando métricas distintas e muitas vezes em bases de dados privadas impossibilitando uma

comparação honesta das mesmas.

Capítulo 3

Metodologia de Avaliação

Neste capítulo é apresentada a metodologia de avaliação do projeto em questão.

As contramedidas serão avaliadas seguindo dois critérios. O primeiro critério avaliará a performance de cada contramedida, em termos de taxa de detecção de ataques, utilizando cada uma das bases de dados de referência. Este critério de avaliação é chamado de Protocolo de Avaliação Intra Base de Dados. O segundo critério também avaliará a performance de cada contramedida, em termos de taxa de detecção de ataques, mas simulando um cenário mais realístico. Tal cenário procurará avaliar a capacidade de generalização da detecção dos ataques. Este critério de avaliação é chamado de Protocolo de Avaliação Inter Base de Dados.

Este capítulo está organizado da seguinte maneira: A Seção 3.1 descreve as medidas de desempenho que serão utilizadas neste projeto. Já Seção 3.2 apresenta o Protocolo de Avaliação Intra Base de Dados. Por fim a Seção 3.2 apresenta o Protocolo de Avaliação Inter Base de Dados.

3.1 Medidas de Desempenho

Serão utilizadas as seguintes métricas de desempenho para avaliar as contramedidas:

- *HTER*: Média entre as Taxas de Falsas Aceitações (FAR) e Falsas Rejeições (FRR) para um dado limiar τ como pode-se observar na equação:

$$HTER = \frac{FAR(\tau, D) + FRR(\tau, D)}{2}, \quad (3.1)$$

onde τ é o limiar de detecção, D é a base de dados, *FAR* é a taxa de falsas aceitações na base de dados D e *FRR* é a taxa de falsas rejeições na base de dados D ;

- *FAR100*: Valor esperado para a Taxa de Falsa Rejeição (FRR) quando a Taxa de Falsa Aceitação (FAR) é igual a 1/100. Esta medida é útil para avaliar contramedidas com requisitos de segurança mais rigorosos;
- *FAR1000*: Semelhante a medida FAR100, onde obtém-se o valor esperado para a Taxa de Falsa Rejeição (FRR) quando a Taxa de Falsa Aceitação (FAR) é igual a 1/1000. Sistemas com requisitos ainda mais rigorosos adotam tal medida.

3.2 Protocolo de Avaliação Intra Base de Dados

Este protocolo avaliará a performance das contramedidas em termos de detecção de ataques. As contramedidas serão treinadas, calibradas e os resultados serão reportados em cada uma das bases de dados de referência utilizando as métricas descritas na seção 3.1. O valor de τ na Equação 3.1 será obtido estimando o valor da pontuação no ponto de EER no subconjunto de desenvolvimento. Os resultados serão reportados na subconjunto de teste.

3.3 Protocolo de avaliação Inter-base de dados

Com o objetivo de mitigar qualquer tipo de viés no processo de aferição de performance em classificadores, comumente bases de dados são divididas em conjuntos mutuamente exclusivos de treinamento e teste. Mesmo com este cuidado, viéses relativos à construção da base de dados podem acontecer, prejudicando a avaliação de uma contramedida. O Protocolo de Avaliação Inter Base de Dados tem o objetivo de mitigar este problema. Neste protocolo cada uma das contramedidas serão treinadas e calibradas utilizando uma base de dados. Resultados serão reportados utilizando outras base de dados.

O valor de τ na Equação 3.1 será obtido estimando o valor da pontuação no ponto de EER no subconjunto de desenvolvimento em uma base de dados. Os resultados serão reportados no subconjunto de teste de outras bases de dados.

Capítulo 4

Resultados Parciais e Plano de Trabalho

Neste capítulo são apresentados os primeiros experimentos realizados seguindo a metodologia proposta no capítulo 3. A metodologia foi aplicada em três contramedidas sendo duas explorando a características relativas a qualidade da imagem e uma relativa às características da cena. Os códigos fontes destas contramedidas estão livremente disponíveis na WEB e são elas:

- Contramedida que explora a relação de movimento da face em relação a cena proposta por (?)¹;
- Contramedida que explora texturas utilizando LBP proposta por (?)²;
- Contramedida que explora a dinâmica de uma textura ao longo do tempo utilizando *LBP-TOP*(?)³.

As contramedidas propostas por (?) e (MINHA REF) admitem apenas vídeos como entrada. Por esta razão a metodologia foi aplicada apenas nas bases de dados Replay Attack e CASIA FASD que possuem apenas vídeos. Neste experimento, a configuração de cada contramedida foi selecionada seguindo orientações apresentadas em cada trabalho de referência. Mais detalhes sobre as configurações de cada contramedida podem ser obtidos na Capítulo 2.

Em especial, a contramedida proposta por (?) utiliza redes neurais do tipo MLP para classificação. Dada a natureza aleatória na inicialização dos pesos sinápticos, experimentos com esta contramedida foram executados cinco vezes. A média dos resultados é reportada como resultado final.

4.1 Protocolo de Avaliação Intra Base de Dados

A Tabela 4.1 exibe a performance obtida com cada contramedida utilizando o Protocolo de Avaliação Intra Base de Dados.

Em termos de HTER é possível observar um desempenho satisfatório de todas as contramedidas testadas em ambas as bases de dados. A análise da performance de cada contramedida

¹<http://pypi.python.org/pypi/antispoofing.motion/1.0.2>

²<http://pypi.python.org/pypi/antispoofing.lbp/1.1.0>

³<http://pypi.python.org/pypi/antispoofing.lbptop/>

Tabela 4.1: Performance das três contramedidas utilizando o Protocolo de Avaliação Intra Teste

Contramedida	Database	HTER(%) dev test		FAR100 (%)	FAR1000 (%)
Correlação de Movimento	Replay Attack	11.66	11.79		
	CASIA FASD	24.91	31.36		
$LBPTOP_{8,8,8,1,1,1}^{u2}$	Replay	8.17	8.51	28.77	53.72
	CASIA FASD	21.77	22.27	86.15	98.42
$LBP_{8,1}^{u2}$	Replay	14.41	15.45	46.10	73.38
	CASIA FASD	23.00	22.54	86.48	98.84

no conjunto de calibração e no conjunto de teste sugere uma boa capacidade de generalização em ambas as bases de dados já que suas performances são semelhantes. Este comportamento pode ser também observado através das curvas ROC na Figura 4.1. As curvas azuis e vermelhas (linha pontilhada e linha sólida) são as performances obtidas no conjunto de desenvolvimento e teste. É possível observar que ambas as curvas estão quase sobrepostas corroborando com o resultado obtido na Tabela 4.1. Porém, analisando a performance das contramedidas com restrições de segurança mais rigorosas os resultados não são satisfatórios. Analisando a taxa de Falsa Rejeições quando a Taxa de Falsas Aceitações está em 0,1% (FAR 1000) temos um FRR acima de 98% para todas as contramedidas inviabilizando o seu uso para aplicações que necessitam altos requisitos de segurança.

A comparação da performance de cada contramedida em diferentes bases de dados sugere que a base de dados CASIA FASD possui ataques mais difíceis de serem detectados que os da base de dados Replay Attack.

4.2 Protocolo de Avaliação Inter Base de Dados

A Tabela 4.2 exibe a performance obtida com cada contramedida utilizando o Protocolo de Avaliação Inter Base de Dados.

Tabela 4.2: Performance das três contramedidas aplicando o Protocolo de Avaliação Inter Teste

Contra-medida	Conj. de treino e calibração	Conj. de teste	HTER(%) dev test		FAR100(%)	FAR1000(%)
Correlação de Movimento	Replay Attack	CASIA FASD	11.66	61.78		
	CASIA FASD	Replay Attack	24.91	54.56		
$LBPTOP_{8,8,8,1,1,1}^{u2}$	Replay Attack	CASIA FASD	8.17	51.05	100	100
	CASIA FASD	Replay Attack	21.77	61.11	99.52	99.99
$LBP_{8,1}^{u2}$	Replay Attack	CASIA FASD	14.41	48.06	100	100
	CASIA FASD	Replay Attack	23.00	57.64	98.65	99.90

Em termos de HTER é possível observar um desempenho distante do obtido no experimento anterior sinalizando um forte enviesamento no processo de treinamento. Tal performance sugere que as contramedidas publicadas não possui um poder de generalização tão bom quanto reportado. Este comportamento pode ser também observado através das curvas ROC na Figura 4.1. As curvas azuis e verdes (linha pontilhada e linha tracejada) são as performances obtidas no conjunto de calibração de uma base de dados e no conjunto de teste de outra base de dados. É possível observar que as curvas estão bem distantes, ou seja, não é possível ter uma performance satisfatória para qualquer valor de limiar τ . Corroborando com estes resultados, uma análise das Taxas de Falsas Rejeições quando as Taxas de Falsas Aceitações estão em 0,1% (FAR 1000) temos um FRR acima de 98% para todas as contramedidas impossibilitando o seu uso para aplicações que necessitam altos requisitos de segurança.

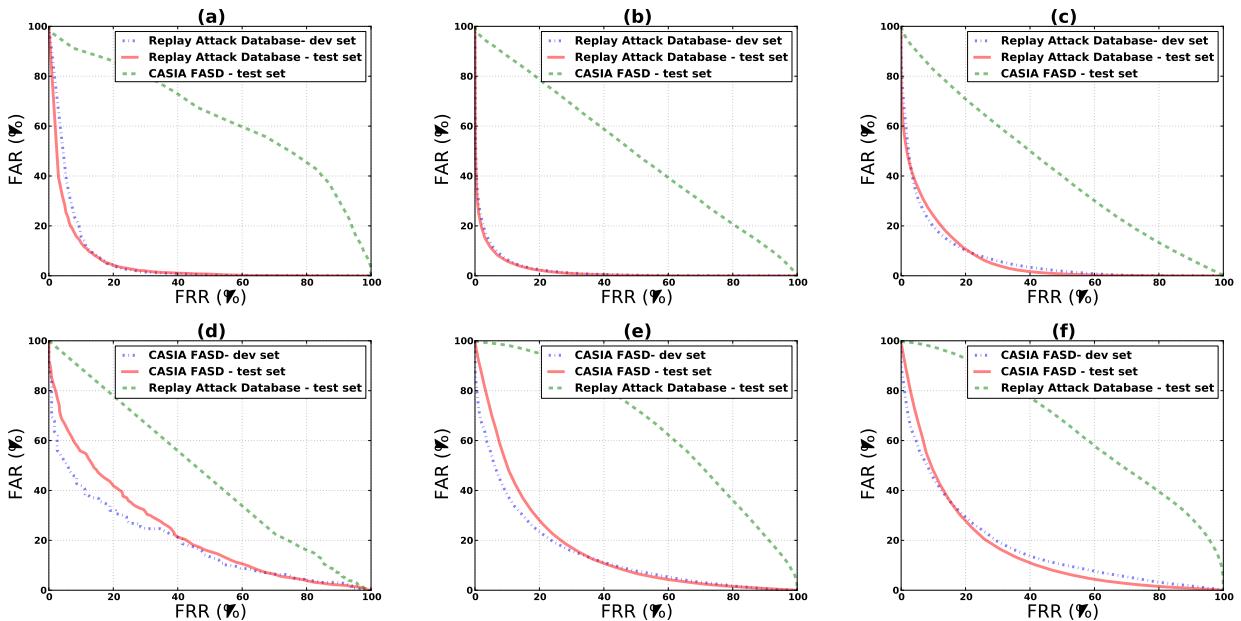


Figura 4.1: Curvas ROC aplicando os dois protocolos de avaliação. (a) Correlação de movimento treinada e calibrada com a Base de Dados Replay Attack Database (b) Contramedida *LBP – TOP* treinada e calibrada com a Base de Dados Replay Attack (c) Contramedida *LBP* treinada e calibrada com a Base de Dados Replay Attack (d) Correlação de movimento treinada e calibrada com a Base de Dados CASIA-FASD (e) Contramedida *LBP – TOP* treinada e calibrada com a Base de Dados CASIA-FASD (f) Contramedida *LBP* treinada e calibrada com a Base de Dados CASIA-FASD.

4.3 Plano de Trabalho

Discutir

Capítulo **5**

Conclusões Parciais

O estudo de detecção de ataques de *spoofing* em sistemas de autenticação de face é bastante recente. Diversas contramedidas vem sendo publicadas nos últimos meses, porém cada trabalho apresenta avaliações utilizando métricas diferentes em bases de dados muitas vezes privadas impossibilitando a reprodução de resultados e eventual comparação.

A principal contribuição deste projeto de mestrado é prover uma metodologia de avaliação de contramedidas contra ataques de *spoofing* que possibilite uma comparação justa entre as mesmas. Para tal, a metodologia define dois protocolos de avaliação com a finalidade de medir a:

- Performance, em termos de detecção de ataques, em bases de dados de referência;
- Performance, em termos de detecção de ataques, em um cenário similar a um cenário real;

Como avaliação preliminar, a metodologia foi aplicada a três contramedidas da literatura. Os resultados obtidos indicam que as contramedidas apesar de apresentarem resultados satisfatórios em testes nas bases de dados de referência, sua aplicabilidade em um cenário real ainda é impossibilitada.