

Tiago de Freitas Pereira

ESTUDO COMPARATIVO DE TÉCNICAS DE DETECÇÃO DE ATAQUES DIRETOS À SISTEMAS DE
AUTENTICAÇÃO DE FACE

Campinas
2012

Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Tiago de Freitas Pereira

ESTUDO COMPARATIVO DE TÉCNICAS DE DETECÇÃO DE ATAQUES DIRETOS À SISTEMAS DE
AUTENTICAÇÃO DE FACE

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

Resumo

Palavras-chave:

Abstract

Abstract

Key-words:

Sumário

| | | |
|----------|---|-----------|
| 1 | Introdução | 11 |
| 1.1 | Contextualização e motivação | 11 |
| 1.1.1 | Ataque de <i>replay</i> | 13 |
| 1.1.2 | Ataque na referência biométrica | 14 |
| 1.1.3 | Ataque <i>man-in-the-middle</i> | 14 |
| 1.1.4 | Ataque de spoofing | 15 |
| 1.2 | Objetivos, Hipóteses e Resultados esperados | 15 |
| 1.3 | Organização do trabalho | 15 |
| 2 | Revisão da literatura | 17 |
| 2.1 | Bases de dados de referência | 18 |
| 2.1.1 | NUAA | 18 |
| 2.1.2 | Replay Attack | 19 |
| 2.1.3 | CASIA FASD | 20 |
| 2.2 | Spoofing em reconhecimento de face | 21 |
| 2.2.1 | Presença de vitalidade | 22 |
| 2.2.2 | Características da cena | 23 |
| 2.2.3 | Discrepância relativa à qualidade da imagem | 24 |
| 2.3 | Considerações Finais | 26 |
| 3 | Metodologia de avaliação | 27 |
| 3.1 | Medidas de desempenho | 27 |
| 3.2 | Protocolo de Avaliação Intra Base de Dados | 28 |
| 3.3 | Protocolo de avaliação Inter-base de dados | 28 |
| 4 | Resultados e conclusões parciais | 29 |
| 4.1 | Resultados preliminares | 29 |
| 4.1.1 | Protocolo de Avaliação Intra Base de Dados | 29 |
| 4.1.2 | Protocolo de Avaliação Inter Base de Dados | 30 |
| 4.2 | Conclusões parciais | 30 |

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 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 atacante) e de fácil memorização. Estes critérios, claramente antagônicos, é 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 com a senha do cartão reforça a segurança do mesmo. Há bancos que disponibilizam para seus clientes tokens que geram um número aleatório 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 por parte do usuário.



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 drasticamente ao longo do tempo. Como exemplo, a voz humana pode variar repentinamente quando estamos doentes; nossos traços faciais infelizmente envelhecem ao longo do tempo. 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 ??.

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 figura 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

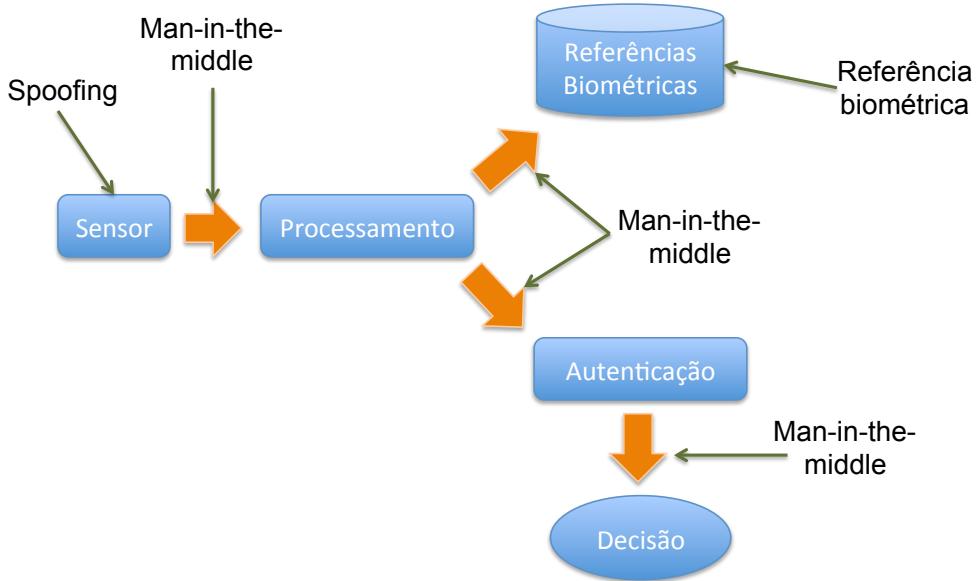


Figura 1.2: Fluxograma básico de um sistema de autenticação biométrica e os seus possíveis pontos de ataque.

que o dado fornecido não foi injetado artificialmente. Uma forma de se defender 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 de referências biométricas consiste em atacar o local de armazenamento das referências biométricas. 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. Um outro ponto de destaque neste tipo de ataque é que uma vez violada uma referência biométrica, a mesma perde a característica da unicidade.

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. Em sistemas de autenticação baseados na voz, o atacante pode gravar a voz da identidade alvo e usar esta gravação como entrada no sistema de autenticação. 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 da voz humana. 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 para a 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 e este será o ponto central desta dissertação.

1.2 Objetivos, Hipóteses e Resultados esperados

Avaliação das contramedidas e generalização.

1.3 Organização do trabalho

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

Já no capítulo é apresentado a metodologia de avaliação do 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 iluminação exterior complexo ainda são desafios de pesquisa na área. Avanços na área foram amplamente relatados 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 os fluxos da informação no caso de um acesso real e em no caso de um ataque de spoofing.

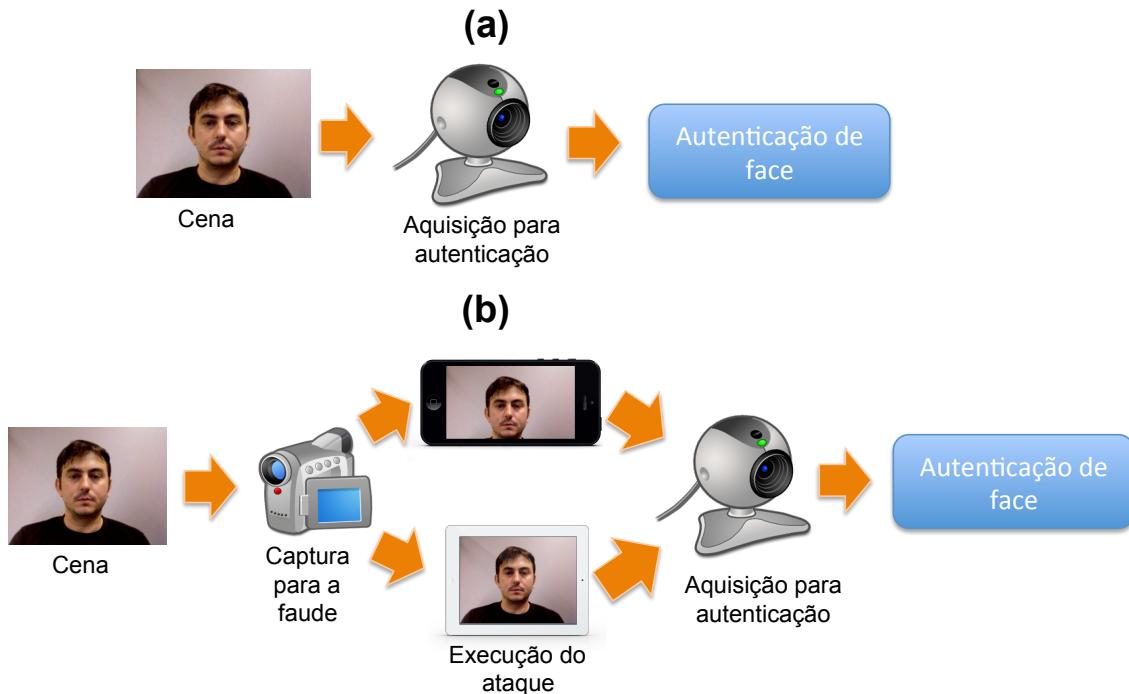


Figura 2.1: (a) Fluxo da informação em um acesso real (b) Fluxo da informação em um ataque de spoofing

Recentemente, a mídia tem documentado algumas situações de ataques em sistemas de

reconhecimento de face em produção. 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 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.

Deficiências contra ataques de spoofing não é exclusividade da biometria facial. Podem ser observados em (? , ?, ?) que sistemas biométricos baseados em impressões digitais sofrem de fraqueza semelhante. A mesma deficiência em sistemas de reconhecimento de íris foram diagnosticados em (? , ?). Finalmente, em (? , ?) ataques spoofing para reconhecimento de locutor são abordados.

Este capítulo está organizado da seguinte maneira: Na Seção 2.1 serão apresentados 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 é apresentado 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.

2.1.1 NUAA

Construída para o cenário de ataques utilizando fotografias impressas em papel, a base de dados NUAA¹ consiste de capturas de vídeo 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 ataques 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 papel com alta resolução e impressa em papel fotográfico de tamanho A4;
2. **mobile**: Exibe fotografias e vídeos obtidas com um telefone celular modelo iPhone 3GS e exibidas com o mesmo dispositivo;
3. **highdef**: Exibe fotografias 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.

Fig. ref fig: Banco de Dados mostrar 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 com fotografias impressas em papel, celular e ataques com *tablet* respectivamente.

A base de dados Replay Attack fornece um protocolo para avaliar objetivamente uma dada contramedida. Tal protocolo define três conjuntos não sobrepostos e são eles o conjunto de treino (train set), calibração (devel set) e o conjunto de teste (test set). O conjunto de treino (train set) deve ser utilizado para treinar uma contramedida, o conjunto de calibração é usado para ajustar hiper-parâmetros de uma contramedida e para definir um 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: Some frames of real access and spoofing attempts (courtesy of (REF IVANA)).

Tabela 2.1: Número de vídeos e 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. A variedade de ataques 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. Exemplos da base de dados pode ser visto na Fig ???. No total, o banco de dados consiste de 600 vídeos e os assuntos são divididos em subconjuntos para treinamento e teste (240 e 360, respectivamente).

O objetivo desta base de dados é investigar a efetividade de diferentes tipos e 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.

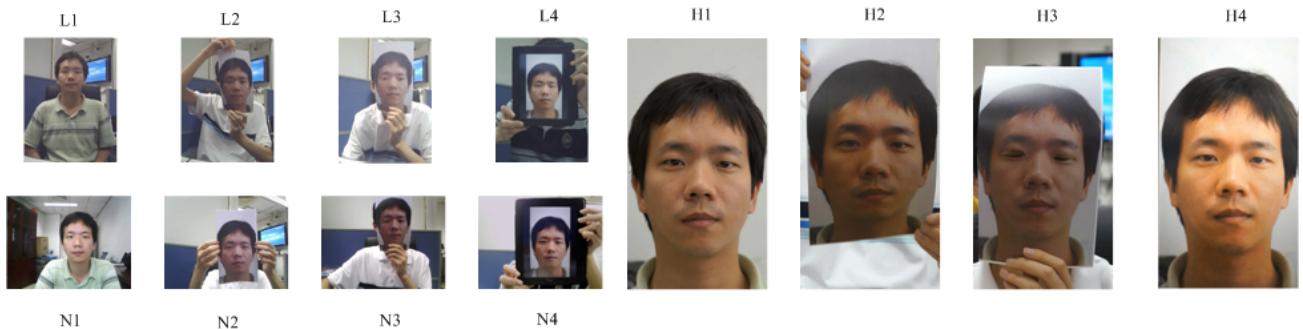


Figura 2.4: Exemplos de acessos reais e ataques da base de dados CASIA FASD (cortesia de ZHANG)

Para reportar os resultados recomenda-se o uso do EER obtido na base de dados 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 para 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 deve 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 verticalmente (Figura 2.5);

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

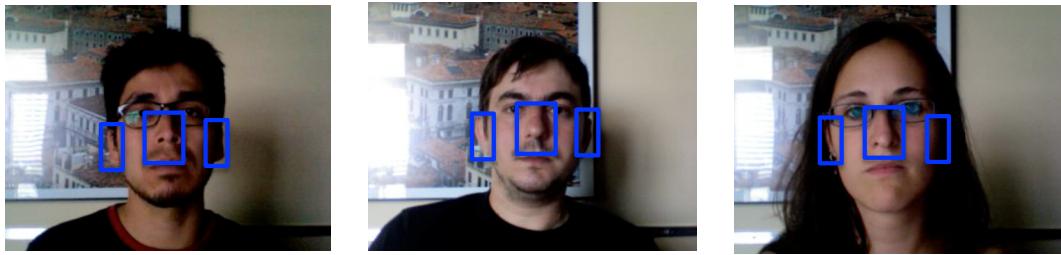


Figura 2.5: Seleção de partes faciais

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 XM2VTS (REF). 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 de 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 o N pontos transformada de Fourier do sinal gerado por M_D (ver Equação ref eq: motionR). Estas medidas computadas servem de 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 (IVANA REF) 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 de 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 apresenta padrões distintos de faces reais. Apoiada nesta hipótese (?)³ explora características de textura utilizando LBP analizando quadros individuais. A Figura 2.6 exibe o diagrama de blocos da contramedida proposta.

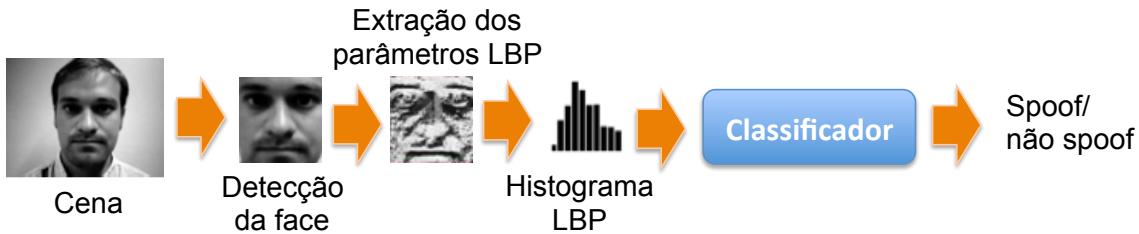


Figura 2.6: Fluxo dos dados da contramedida baseada em LBP

Neste trabalho, as faces são segmentadas e geométricamente 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 |

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

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, (MINHAS REF) 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.

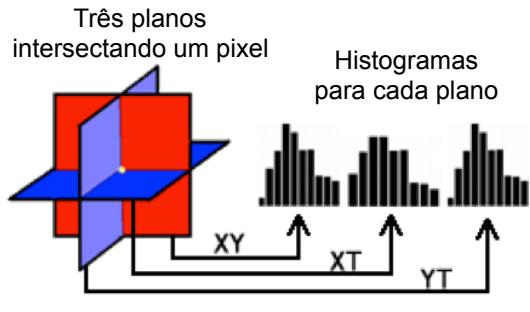


Figura 2.7: Representação da extração de parâmetros utilizando $LBP - TOP$

Neste trabalho, as faces são segmentadas e geométricamente 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 em (IVANA) em $\sim 50\%$.

Apoiados na hipótese que imagens/vídeos utilizadas para ataques concentram mais informação em uma banda específica de frequência, (?) apresenta uma contramedida separando bandas específicas de frequência 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 possui 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. A performance foi avaliada utilizando a base de dados CASIA é um EER de 17% foi alcançado.

Apoiado na hipótese de que as dimensões de uma face utilizando uma mídia para ataque em são menores do que uma face real e as variações de movimento facial em um ataque são menores



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

do que em uma face real, (REF FOURIER) propôs uma contramedia analizando o espectro Fourier. A expectativa com esta análise é que as imagens utilizadas para ataque contém menos componentes de alta frequênciade 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%

Para detectar padrões de ruído em ataques de spoofing, Pinto et al. desenvolveu uma contramedida analizando vídeos combinando diversos elementos. Primeiramente os quadros capturados são filtrados utilizando um filtro Gaussiano e uma filtro na Mediana respectivamente. Estas imagens filtradas são subtraídas na imagem original obtendo o ruído residual da imagem. Este ruído residual é analizado no domínio da frequênciade através da transformada de Fourier 2D. Todos os quadros de um vídeo capturado são combinados utilizando a técnica chamada Rítmo Visual (REF 25 DO PAPER DO HELIO) gerando uma imagem única caracterizando toda 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 frequênciade ocorrência de níveis de cinza entre pares de pixels. Através dessa matriz 12 medidas são extraídas e são elas: ESCREVER. Estas medidas são 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 avaliará a performance de cada contramedida, em termos de taxa de detecção de ataques, em um cenário mais realístico avaliando assim a capacidade de generalização das mesmas. 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, 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.

3.3 Protocolo de avaliação Inter-base de dados

Mesmo utilizando conjuntos distintos de treinamento, calibração e teste com o objetivo de mitigar qualquer tipo de enviesamento certos viéses no processo de construção da base de dados podem acontecer. Este protocolo tem o objetivo de mitigar este tipo de viés e fornecer uma métrica de avaliação justa das contramedidas apresentadas na literatura.

Neste protocolo cada uma das contramedidas será treinada e calibrada utilizando uma base de dados. Resultados serão reportados utilizando outra base de dados.

Capítulo 4

Resultados e Conclusões Parciais

MLP TREINADA 5x.

Neste capítulo são apresentados os primeiros experimentos realizados seguindo a metodologia proposta no capítulo 3. A metodologia foi aplicada a 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 (?);
- Contramedida que explora texturas utilizando LBP proposta por (?);
- Contramedida que explora a dinâmica de uma textura ao longo do tempo (MINHA REF).

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.

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 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 tracejada 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% (FAR1000) temos uma

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

| Countermeasure | Database | HTER(%) dev test | | FAR100 (%) | FAR1000 (%) |
|-----------------------------|---------------|---------------------|-------|------------|-------------|
| Correlation | 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 |

performance acima de 98% para todas as contramedidas iviabilizando o seu uso para aplicações que necessitam altos requisitos de segurança.

A performance obtida em cada base 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: $HTER(\%)$ of each countermeasure applying the intra-test ($D_1 = D_2$) and the inter-test ($D_1 \neq D_2$) protocol.

| Contramedida | Conj. de treino e calibração | Conj. de teste | HTER(%) dev test | | FAR100 (%) | FAR1000 (%) |
|-----------------------------|---------------------------------|-------------------|---------------------|-------|------------|-------------|
| Correlation | 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 possuem 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 tracejada e linha pontilhada) 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 da performance das taxas de Falsas Rejeições quando as Taxas de Falsas Aceitações estão em 0,1% (FAR1000) temos uma performance acima de 98% para todas as contramedidas impossibilitando o seu uso para aplicações que necessitam altos requisitos de segurança.

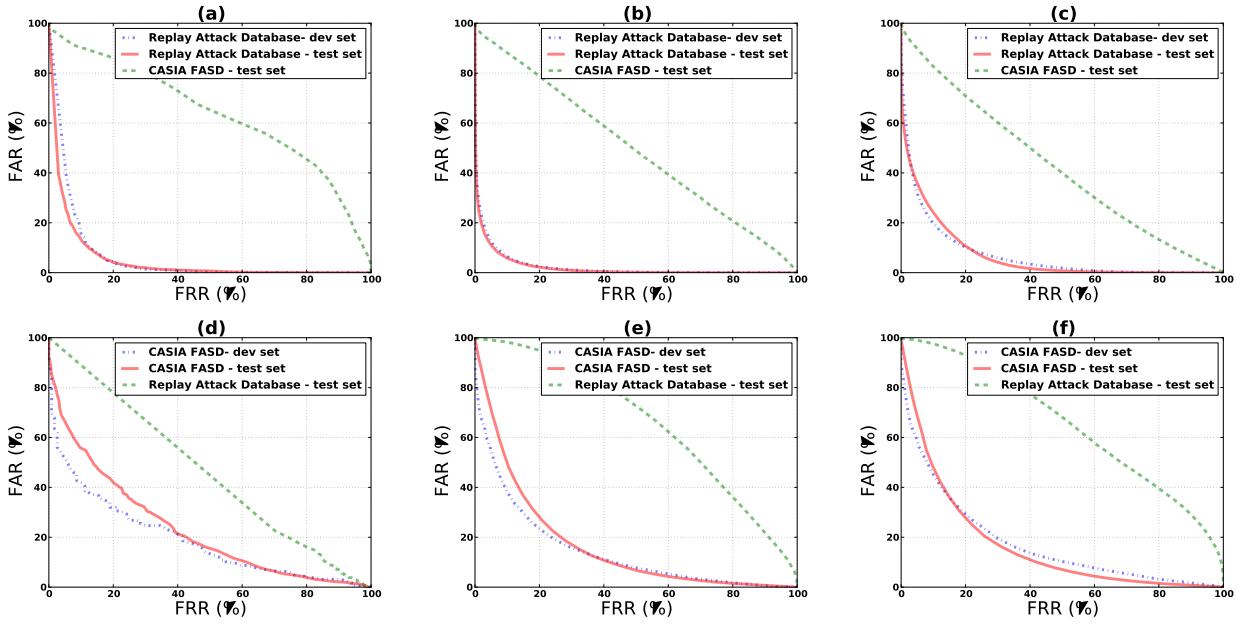


Figura 4.1: ROC curves of each countermeasure using the intra-test and the inter-test protocol. (a) Correlation with frame differences countermeasure trained and tuned with the Replay Attack Database (b) $LBP - TOP$ countermeasure trained and tuned with the Replay Attack Database (c) LBP countermeasure trained and tuned with the Replay Attack Database (d) Correlation with frame differences countermeasure trained and tuned with the CASIA-FASD (e) $LBP - TOP$ countermeasure trained and tuned with the CASIA-FASD (f) LBP countermeasure trained and tuned with the CASIA-FASD.

4.3 Conclusões parciais

O estudo de detecção de ataques de *spoofing* em sistemas de autenticação de face é bastante recente. BLA BLA

Diversas contramedidas vem sendo publicadas nos últimos meses, porém cada trabalho apresentam avaliações utilizando métricas diferentes e bases de dados muitas vezes privadas impossibilitando a reprodução de resultados e eventual comparação.

O objetivo dessa dissertação de mestrado é prover uma metodologia de avaliação de contramedidas de ataques de spoofing explorando dois aspectos. O primeiro deles é a aferição da performance de contramedidas em termos de detecção de ataques em bases de dados públicas. O segundo aspecto consiste da aferição da performance das contramedidas em termos de detecção de ataques em um cenário mais realístico avaliando a capacidade de generalização de cada contramedida. Para isso dois protocolos foram propostos, respectivamente o Protocolo de

Avaliação Intra Base de Dados e o Protocolo de Avaliação Inter Base de Dados.

Aplicada a três contramedidas o Protocolo de Avaliação Intra Base de Dados evidenciou uma boa capacidade de generalização de todas porém a performance das mesmas com requisitos de segurança mais rigorosos não foi satisfatória. O Protocolo de Avaliação Inter Base de Dados ajudou a evidenciar que todas as contramedidas avaliadas possuem um forte enviesamento por conta da base dados sugerindo sobretrainimento inviabilizando a aplicabilidade de tais contramedidas em um cenário real.

4.4 Trabalhos futuros e cronograma

Aplicação de pelo menos uma contramedida relativa a liveness. Mais 3 contramedidas.