Relatório Técnico do INE

Levantamento e comparação entre abordagens de estado da arte para detecção de vivacidade em biometria facial

Alexandre Gonçalves Silva

Elder Rizzon Santos

Fabio Rafael Segundo

Gabriel Arthur Gerber Andrade

Luiz Henrique Susin

Thiago Ângelo Gelaim

Thiago Fonseca Lemos

Relatório Técnico INE 002/2023

UFSC

Universidade Federal de Santa Catarina
Departamento de Informática e Estatística

SUMÁRIO

Resumo

Mecanismos para verificação de identidade mediante biometria constituem uma importante categoria de abordagens para a segurança da informação. A detecção de vivacidade adiciona uma verificação nesse processo de forma a garantir que a biometria apresentada seja legítima e não uma cópia ou artefato artificial. Assim sendo, neste relatório apresenta-se um levantamento do estado da arte em abordagens para detecção de vivacidade no contexto de biometria facial. Além disso, também são apresentados os principais conjuntos de dados publicamente disponíveis para treinamento ou avaliação de métodos. Por último, é apresentada uma análise comparativa quanto ao desempenho das principais abordagens.

Lista de Abreviaturas e Siglas

APCER - *Attack Presentation Classification Error Rate*

APNRR - *Attack Presentation Non-Response Rate*

BPCER - *Bona-fide Presentation Classification Error Rate*

BPNRR - *Bona Fide Presentation Non-Response Rate*

DFT - Transformada Discreta de Fourier

EER - *Equal Error Rate*

FFT - Transformada Rápida de Fourier

HOG - *Histogram of Oriented Gradients*

PAD - *Presentation Attack Detection*

PAI - *Presentation Attack Instruments*

PS-PD - *PAD Subsystem Processing Duration*

SVM - *Support Vector Machine*

1. INTRODUÇÃO

Mecanismos para verificação de identidade mediante biometria constituem uma importante categoria de abordagens para a segurança da informação (BHATTACHARYYA et al, 2009). Tendo em vista que a biometria utiliza de características distintivas, tais como formato do rosto, som da voz e padrão das veias nas mãos para reconhecer uma pessoa, estas abordagens são utilizadas para adicionar camadas de segurança ou para facilitar a autenticação de usuários.

Uma fragilidade do uso automatizado de biometria refere-se à duplicação ou construção artificial das características biométricas e a consequente utilização delas para passar pela verificação. As abordagens para lidar com essa fragilidade, independente da natureza da biometria, constituem diferentes formas para detecção de vivacidade. Em outras palavras, a detecção de vivacidade verifica se a entrada obtida pelo sensor realmente é de uma pessoa e não é uma cópia, um molde ou qualquer outro artefato que simule a biometria (RAHEEM et al, 2019).

Considerando-se a forma de aquisição da biometria, o presente relatório limita-se ao escopo da biometria facial dado um contexto mais amplo de pesquisa aplicada ao desenvolvimento de uma plataforma de pagamentos mediante o rosto dos clientes. Tendo em vista esse escopo e contexto, além de considerar trabalhos acadêmicos é imprescindível analisar também as normas vigentes referentes à segurança desse tipo de aplicação.

A principal normativa referente à detecção de vivacidade é a ISO/IEC 30170, a qual será descrita nas subseções a seguir tendo em vista que ela delimita um escopo para a detecção de vivacidade e também estabelece requisitos de conformidade para aplicações e para avaliação das abordagens.

Após a apresentação da normativa, no Capítulo 2 apresenta-se uma taxonomia para classificação das abordagens de detecção e, em seguida, são apresentadas as principais abordagens para cada categoria. No Capítulo 3 são apresentados os principais conjuntos de dados públicos disponíveis para avaliar as abordagens. Após o levantamento, apresenta-se uma análise comparativa entre principais comparações entre abordagens (Capítulo 4).

1.1. ISO/IEC 30107

A ISO/IEC 30107 partes 1, 2, 3 e 4 tratam da caracterização e avaliação de métodos para detecção de ataques de apresentação (*Presentation Attack Detection* - PAD). As normativas abordam o assunto considerando sistemas baseados em biometria em geral, não apenas os que utilizam a face como informação biométrica. O escopo das normativas também está restrito à ataques no sensor, ou seja, ataques 'apresentados' ao mecanismo que captura as informações biométricas (parte destacada na Figura 1).

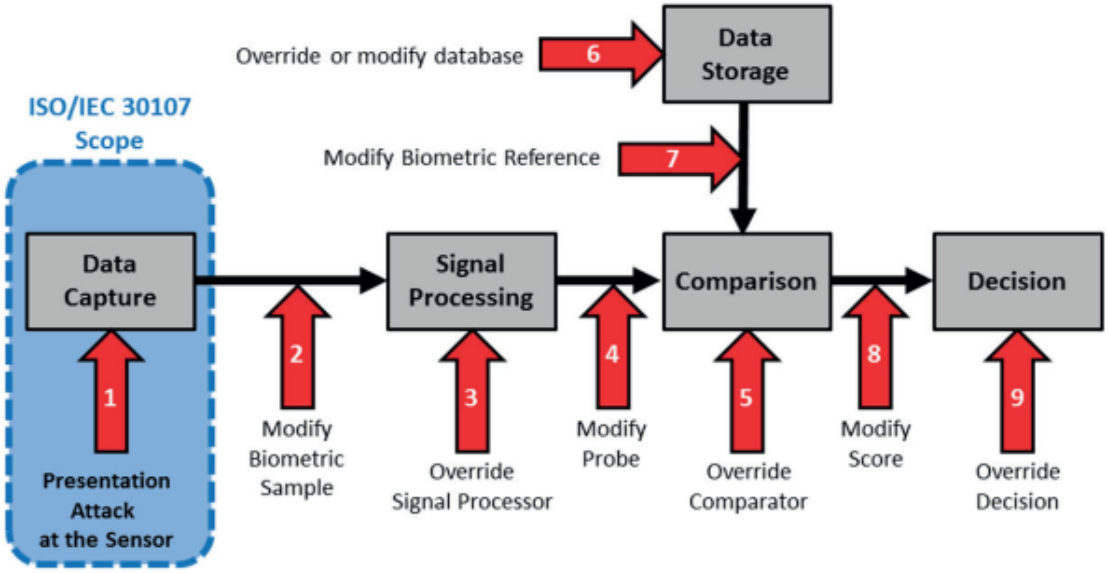

Figura 1 - Escopo da normativa com relação a possíveis pontos de ataque em sistemas biométricos. Fonte: ISO/IEC 30107.

A primeira parte da normativa apresenta um conjunto de possibilidades e formas de ataques e defesas/detecções, dentre elas encontra-se a detecção de vivacidade (liveness detection) a qual é o escopo do presente projeto/relatório e é objeto de estudo principal da parte 3 da ISO, a qual normatiza um método para avaliação de PAD com ênfase em vivacidade.

Nas seções 6.2.1 e 6.2.2 da parte 1 da ISO, são apresentados os conceitos de vivacidade ativa e passiva. As abordagens ativas utilizam alguma forma de desafio/resposta, sendo o desafio oculto ou visível/interativo. Exemplos dessa abordagem são exibir um flash e verificar se a resposta condiz com o esperado ou solicitar que o usuário sorria ou pisque. Em geral, essas abordagens não são consideradas seguras uma vez que o atacante tem clareza do

método de detecção de vivacidade sendo utilizado e pode criar artefatos específicos. Já as abordagens passivas buscam detectar a vivacidade sem qualquer estímulo, utilizando-se apenas os dados obtidos do sensor.

Independente da abordagem para detecção de vivacidade, a normativa define um framework genérico referente à detecção de ataques de apresentação. O framework é utilizado, seção 7 da parte 3, para especificar diferentes possibilidades de avaliação da detecção, cada uma com critérios e níveis de aceitação distintos. O primeiro nível considera apenas o subsistema de detecção, o segundo nível considera o subsistema de captura mais o de detecção e o terceiro nível considera o sistema completo.

De acordo com essa divisão, o escopo atual do presente levantamento está no desenvolvimento do subsistema de detecção de ataque de apresentação. Quanto à localização e forma dos subsistemas, a normativa prevê e possibilita que os mesmos estejam em diferentes locais, não necessariamente todos embutidos em um único dispositivo. Dessa forma, já é previsto, por exemplo, que o subsistema de detecção seja executado em um servidor separado do dispositivo de captura.

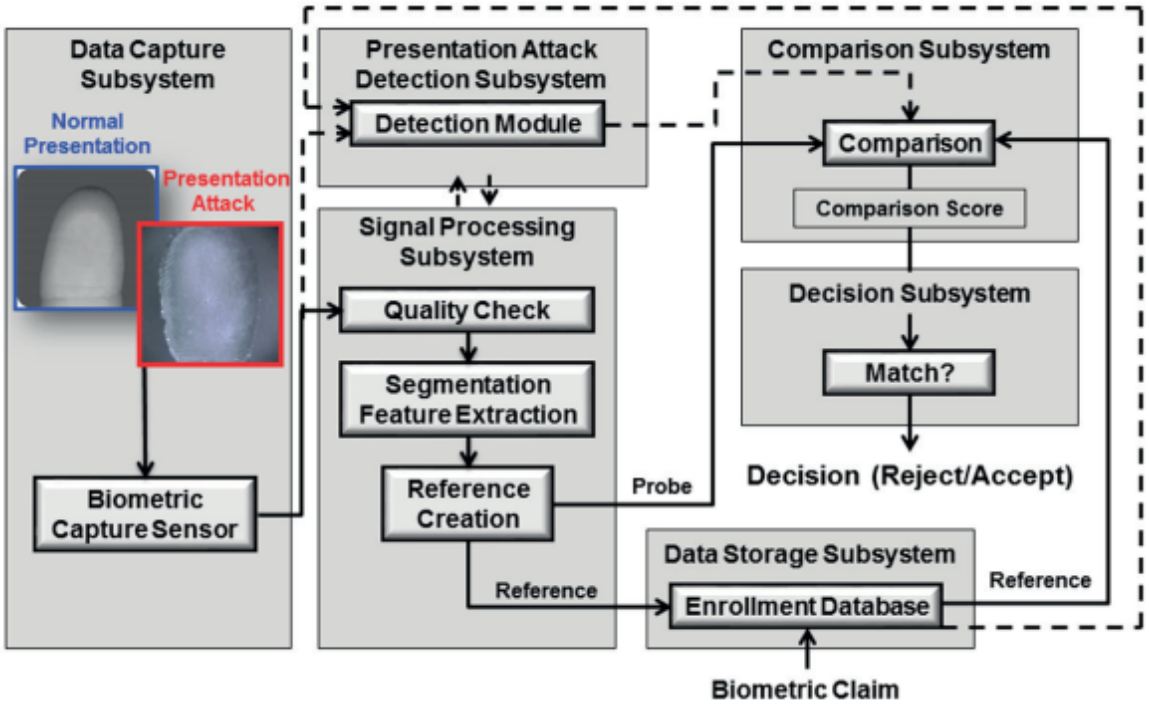

Figura 2 - Framework genérico de um sistema de detecção de ataques de apresentação. Fonte: ISO/IEC 30107.

1.1.1 Método de avaliação do subsistema de PAD

O cerne da avaliação do subsistema de PAD está na medição da habilidade do sistema em classificar corretamente ataques de apresentação e apresentações legítimas (bona fide presentations). Em contextos de uso onde o método de detecção não utiliza um hardware específico e está mais dependente de um algoritmo (como é o caso do presente projeto), a avaliação do sistema pode ser realizada com amostras advindas de um conjunto de dados (parte 3, seção 7 da ISO).

A normativa também estabelece que a avaliação da detecção pode ser realizada em três níveis de especificidade:

i. avaliações genéricas e amplas de qualquer dispositivo para uma aplicação desconhecida;

ii. avaliações específicas para uma aplicação nas quai o conjunto de tipos de ataques é previamente estabelecido analisar apenas o contexto de uso da aplicação;

iii. avaliações específicas para um produto, utilizadas para averiguar uma declaração de desempenho referente à uma categoria de ataques específica.

Nas seções 8, 9 e 10 da terceira parte da normativa, são estabelecidos os critérios e formas de construção dos artefatos utilizados durante a avaliação para testar o mecanismo de detecção (*Presentation Attack Instruments*). No Anexo A da mesma normativa são apresentadas, de maneira geral, formas para obtenção das informações necessárias para elaborar os instrumentos e também algumas sugestões para sua construção. Salienta-se que, quanto ao processo de construção dos instrumentos, a normativa apresenta recomendações, sendo o mais relevante os consequentes tipos de ataques a serem considerados devido às definições genéricas dos tipos de instrumentos.

Assim sendo, a partir do estabelecimento de um conjunto base de instrumentos para os ataques, são apresentados alguns tipos de ataques a serem considerados (entretanto não trata-se de uma lista exaustiva) os quais são divididos entre ataques estáticos e dinâmicos.

Os ataques estáticos adotam artefatos que não simulam aspectos comportamentais associados à característica biométrica (ex.: movimentos, piscadas, entre outros), na normativa descrevem-se duas formas para esse tipo de ataque:

1. Ataque com impressão 2D: apresentação de uma impressão em papel, transparência ou outro meio ao sensor.
2. Ataque com objeto 3D: apresentação ao sensor de um artefato construído, por exemplo, com gesso, impressão 3D ou o uso de uma máscara.

Já os ataques dinâmicos simulam comportamentos associados à característica biométrica e, na normativa, são exemplificados com as seguintes possibilidades:

1. Ataques com vídeos de dispositivos móveis, tablets ou laptops: o principal cenário desse tipo de ataque é a obtenção de um pequeno conjunto de imagens ou vídeos advindas de redes sociais e a reprodução do mesmo para simular vivacidade.
2. Ataques de replay: esse tipo de ataque difere-se do anterior apenas quanto ao conteúdo sendo reproduzido. Nesse caso, considera-se a reprodução de uma característica biométrica obtida pelo próprio sensor do mecanismo de detecção. Ainda considera-se que a reprodução pode ser modificada.

Finalmente, as três últimas seções da normativa abordam detalhes do processo de avaliação em si. A seção 11 aborda aspectos a serem considerados quando faz-se necessário avaliar algum processo específico do produto, por exemplo, processo de cadastramento de usuários, verificação e mecanismos ‘offline’ de PAD - os quais por alguma questão de projeto não apresentam seu resultado imediatamente após a apresentação. A normativa não estabelece exatamente quando essas recomendações devem ser adotadas. Nossa interpretação sugere que seja o caso quando a avaliação tratar de um produto específico. Aspectos referentes à organização do teste e definições de metodologia, equipe e equipamentos são apresentados na seção 12 sendo, em nossa opinião, recomendações dirigidas às instituições que realizam as avaliações.

As métricas adotadas para avaliar o desempenho do mecanismo de detecção são descritas na seção 13 da normativa. O processo de avaliação é organizado visando a obtenção de dados e, em uma primeira etapa, o relatório de avaliação deve apresentar as seguintes informações, referentes ao método adotado:

- quantidade e descrição dos instrumentos de ataque utilizados;
- quantidade de sujeitos utilizados nos testes
- quantidade de artefatos criados em função de cada sujeito e de cada material utilizado
- quantidade de fontes das quais cada artefato foi produzido

- quantidade de materiais testados
- descrição da informação disponibilizada pela saída do mecanismo de detecção
- ordenação das apresentações com e sem ataques e se alguns sujeitos foram reutilizados
- ordenação das apresentações ao mecanismo habilitado e desabilitado e se alguns sujeitos foram reutilizados

A partir da execução dos testes, são calculadas as seguintes métricas avaliativas referentes à habilidade do subsistema de PAD para classificar corretamente ataques de apresentação, sendo que as mais importantes são as seguintes:

- APCER (*attack presentation classification error rate*): taxa de erro da classificação de ataques de apresentação - proporção de ataques utilizando o mesmo tipo de instrumento incorretamente classificadas como apresentações legítimas
- BPCER (*bona fide presentation classification error rate*): taxa de erro da classificação de apresentações legítimas - proporção de apresentações legítimas incorretamente classificadas como ataques em um cenário específico

Além destas taxas principais, seguindo-se a descrição do fabricante do mecanismo, o avaliador também deve estabelecer critérios para determinar quando o mecanismo não apresenta uma resposta, tanto para ataques quanto para apresentações legítimas. Dessa forma, a avaliação do subsistema de detecção inclui mais duas medidas referente à quantidade de ataques e apresentações legítimas que o sistema não ofereceu resposta:

- APNRR (*attack presentation non-response rate*): taxa de falta de resposta à ataques
- BPNRR (*bona fide presentation non-response rate*): taxa de falta de resposta às apresentações legítimas

A última medida calculada na avaliação de subsistemas de PAD é o tempo de duração do processamento do mecanismo (PS-PD - PAD Subsystem Processing Duration), o qual deve ser relatado para os ataques e apresentações legítimas separadamente. No sumário apresentado na seção 13.2.5 essa última medida é opcional e as quatro anteriores são obrigatórias.

A norma não estabelece valores mínimos ou máximos para conformidade. Ao consultar a metodologia de testes de uma empresa certificadora (iBeta[1]), nota-se que a mesma

[1] www.ibeta.com

disponibiliza dois níveis de testes e as respectivas taxas para cada nível. No primeiro nível, os avaliadores que constroem os artefatos não tem experiência com esse tipo de avaliação e, para cada tipo de instrumento, possuem um limite de 8h para a confecção dos mesmos. Nesse caso, os aparelhos e máquinas disponíveis para construção são os que podem ser encontrados em escritórios ou casas, com um limite de $30,00 para a construção.

Já no segundo nível, alguns avaliadores possuem experiência, o limite de tempo é de até 4 dias por tipo de instrumento, os equipamentos de construção são mais avançados (impressora 3D, máscaras de resinas e látex) e o limite financeiro é de $300,00. A taxa de erro de apresentações legítimas (BPCER) e a taxa de rejeições a usuários legítimos (FNIR na ISO, a qual não está considerada na presente análise por estar fora do escopo do subsistema de detecção de PAD) não pode ultrapassar 15% para ambos níveis.

2. ABORDAGENS PARA DETECÇÃO DE VIVACIDADE

Em Ramachandra et al (2017), os autores apresentam uma taxonomia referente aos métodos para detecção de ataques de apresentação em sistemas de reconhecimento facial. Esta seção resume a classificação proposta pelo referido estudo, e considera também duas abordagens adicionais mais recentes (métodos baseados em aprendizado de máquina e métodos baseados no funcionamento do corpo humano, respectivamente) que não estão contempladas na taxonomia.

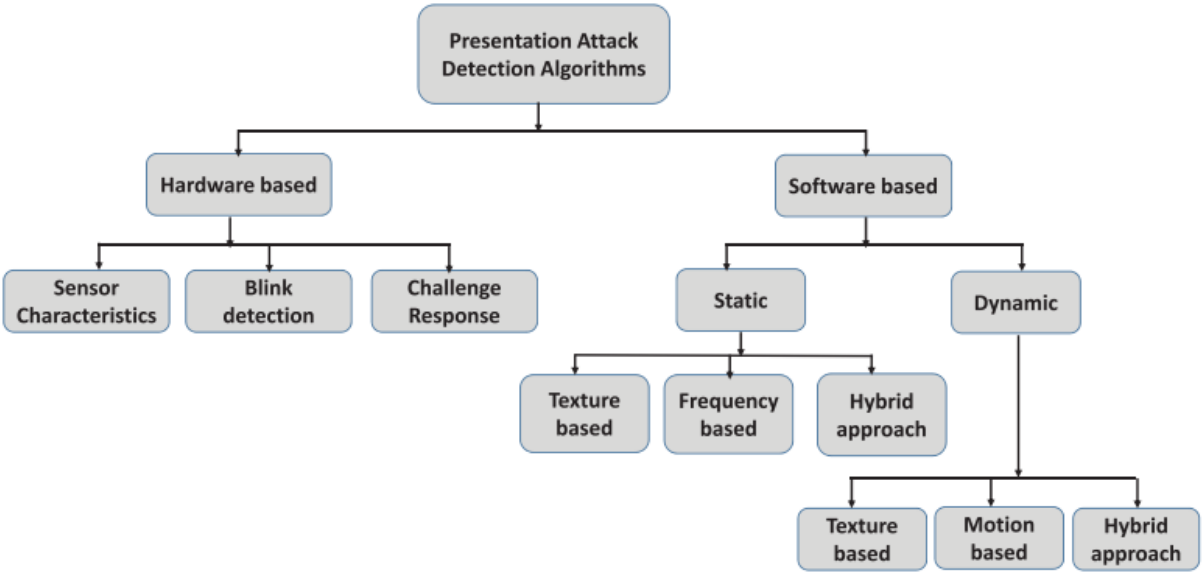

Figura 3 - Taxonomia de métodos para detecção de vivacidade. Fonte: Ramachandra et al (2017)

A taxonomia distingue, primeiramente, os métodos que utilizam um componente de *hardware* adicional que trabalha em conjunto com o sensor de reconhecimento facial, dos que utilizam apenas algoritmos para determinar se a amostra capturada do rosto origina de um ataque de apresentação ou de uma apresentação bona-fide (também chamada de apresentação real ou viva). O presente estudo possui como foco métodos baseados em *software* e, portanto, nenhuma das seções a seguir contém métodos baseados em *hardware*.

Em seguida, o estudo diferencia os métodos entre estáticos e dinâmicos. Métodos estáticos são projetados para trabalhar sobre uma única imagem sem a necessidade de qualquer informação temporal. Ainda assim, métodos estáticos podem ser aplicados com

vídeos, caso em que cada quadro do vídeo é analisado independentemente e uma decisão final do vídeo (sequência de quadros) é tomada com base em uma decisão majoritária. Por outro lado, métodos dinâmicos exploram a informação temporal de um vídeo apresentado para o sensor de reconhecimento facial. Normalmente, a informação temporal é obtida pelo movimento relativo entre cada quadro do vídeo. Por essa razão, métodos dinâmicos costumam requerer mais tempo e mais esforço computacional do que os métodos estáticos.

A presente pesquisa foca em métodos estáticos tendo em vista o interesse em soluções de baixo custo computacional para detectar a vivacidade de imagens ou de vídeos bem curtos (0.5 segundos). Ainda assim, os métodos dinâmicos (baseados em movimento) permanecem neste levantamento teórico, mas focamos em categorizar os métodos quanto ao terceiro nível hierárquico da taxonomia: métodos baseados em textura, métodos baseados na qualidade da imagem, métodos baseados em frequência, métodos baseados em movimentação e métodos híbridos. A Tabela 1 apresenta um sumário de vantagens e desvantagens das abordagens baseadas em software e foi elaborada com base no trabalho de Ramachandra e Busch, 2017.

Abordagem	**Custo**	**Ataque c/ imagem**	**Ataque c/ vídeo**	**Generalização**	**Medida de vivacidade**	**Eficácia depende de**
Textura	Baixo	Efeito		Ruim		Resolução da imagem
Frequência	Baixo		Efeito	Ruim		Dispositivo
Híbrido	Alto	Efeito	Efeito	Razoável		
Movimento	Alto	Efeito		Ruim	Sim	

Tabela 1 - Vantagens e desvantagens das abordagens baseadas em software.

Percebe-se que os métodos baseados em textura oferecem um baixo custo computacional contra ataques com fotos, enquanto métodos baseados em frequência são efetivos contra ataques com vídeos. As outras duas abordagens requerem um maior custo computacional. Métodos híbridos são efetivos contra ambos tipos de ataques considerados pelo estudo em questão e, mais importante, são capazes de detectar ataques desconhecidos durante a fase de treinamento do método (o que denomina-se “generalização”). Por fim, métodos baseados em movimento utilizam de um processamento maior para oferecer uma medida de vivacidade (como, por exemplo, o piscar dos olhos).

Nas subseções seguintes apresentam-se as principais implementações acerca das abordagens de software presentes na taxonomia e adicionando-se abordagens mais recentes conforme descrito anteriormente.

2.1 Métodos baseados em textura

O objetivo principal desse tipo de método é identificar diferenças em texturas de faces genuínas e impostoras. São exemplos de diferenças detectáveis: pigmentação em impressões, deformação de sombras e reflexão da luz na face. É uma abordagem bem-sucedida na detecção de ataques de apresentação; discrimina, com eficiência características de artefatos de ataque, como presença de pigmentos devido a defeitos de impressão, deformação de sombra devido a um ataque de apresentação, e reflexão especular devido a um ataque com máscara, por exemplo.

Os métodos tradicionais para detecção e caracterização de texturas utilizam, principalmente, os seguintes descritores: LBP, LPQ, HOG, GLCM, WLD e LDP (mais informações sobre os descritores estão presentes no Glossário deste relatório).

Métodos baseados em textura podem ser estáticos ou dinâmicos. Métodos dinâmicos exploram informações temporais do vídeo reproduzido que é apresentado para o sensor de reconhecimento facial. Essas informações buscam detectar o movimento relativo ao longo dos quadros do vídeo. Esta é a razão pela qual essa variante dos métodos requer mais tempo e mais recursos computacionais. Por outro lado, de acordo com Freitas Pereira et al (2014) métodos dinâmicos usualmente possuem uma alta eficácia em detectar ataques de impressão.

Métodos dinâmicos baseados em textura, em particular, utilizam descritores de textura-movimento para explorar a mudança da textura ao longo do vídeo capturado. Esses descritores extraem características de três planos ortogonais para combinar informações espaciais e temporais. A extração dos planos ortogonais é uma simplificação necessária da inviável alternativa de extrair informações de múltiplos quadros do vídeo. A pesquisa apresentada em Freitas Pereira et al (2014) foi uma das primeiras implementações desta abordagem e baseou-se em Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) e demonstrou um desempenho razoável na base de ataques disponibilizada pela Idiap[2]. Além do descritor LBP-TOP, os descritores WLD-TOP e LDP-TOP também apresentam eficácia em descrever tanto padrões da aparência quanto padrões dos movimentos horizontais e verticais para a detecção de ataques de apresentação (MEI et al, 2015; PHAN et al, 2016).

[2] https://www.idiap.ch/en

Em suma, segundo Ramachandra et al (2017), métodos baseados em textura são uma técnica eficiente (baixo custo computacional) contra ataques de fotos impressas mas carecem de generalidade e dependem da resolução da imagem.

2.1.1 Principais implementações de métodos baseados em textura

Benlamoudi et al (2015) apresentam um método estático baseado em textura que, a partir de cada frame, registra (detecta) a face e normaliza as intensidades da imagem em níveis de cinza para uma dimensão de 128x128. O LPQ é aplicado por bloco e seus histogramas são concatenados para formar parte do descritor. Este processo se repete para tamanhos cada vez menores de blocos, em um sistema multi-nível. O descritor final consiste no histograma médio dos 250 níveis adotados. A classificação binária (real ou fake) é feita por SVM (Support Vector Machine). Para o conjunto de dados CASIA (600 vídeos), o EER (Equal Error Rate) foi de 11,39%.

Boulkenafet et al (2016) também propõem um método estático baseado em textura o qual utiliza o espaço de cores YCbCr e aplica LBP em cada componente. Os três histogramas gerados são conectados para formar o descritor. A classificação binária (real ou fake) é feita por SVM. O método foi avaliado com teste cruzada de três bases (CASIA, MSU e Replay-Attack), e apresentou um HTER médio entre 20,4 e 46,0 para os experimentos realizados. o EER foi de 3,30%, se posicionando entre os 10 melhores resultados para variadas bases de imagens.

Já Freitas Pereira et al (2014) descrevem uma abordagem que faz uso do descritor LBP-TOP para realizar uma detecção dinâmica baseada em textura. Esse descritor considera apenas os três planos ortogonais que intersectam o centro do pixel a ser avaliado em três direções XY, XT e YT, onde T é o eixo temporal (a sequência de quadros). A Figura 4 ilustra o processamento dos eixos.

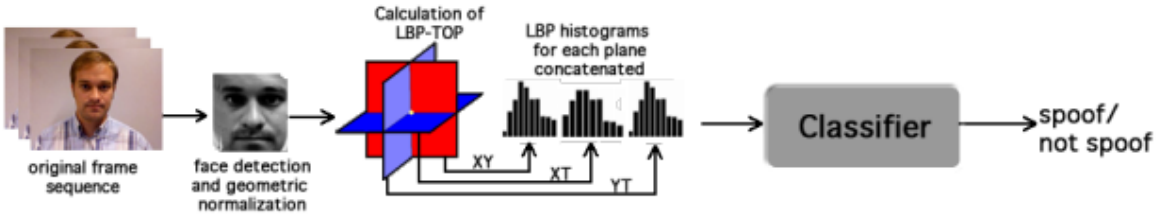

Figura 4 - Uso do descritor LBP-TOP para detecção de vivacidade. Fonte: Freitas Pereira et al (2014)

A direção XY é a mesma utilizada por métodos estáticos baseados em LBP. Com o uso destes três planos ortogonais, três diferentes histogramas são gerados e, então, concatenados. Com esta abordagem, o tamanho do histograma de LBP-TOP é de apenas 3×2^p. Além da simplificação computacional, LBP-TOP tem a vantagem de gerar histogramas independentes para cada plano de intersecção, em espaço e tempo, os quais podem ser tratados tanto em combinação quanto individualmente.

Os experimentos que os autores realizaram com o conjunto de dados da Idiap mostraram um aumento da eficácia com relação a dois métodos estáticos baseados em LBP: um proposto pela Idiap, e outro proposto pela universidade de Oulu. A análise mostrou que LBP-TOP melhora quando utilizada com múltiplas resoluções de Rt; configuração esta que permitiu o método alcançar um HTER de 7.60% com um classificador SVM. Adicionalmente, a comparação quantitativa de um estudo de 2020 indica que LBP-TOP é superior a outros dois métodos dinâmicos de textura (WDL-TOP e LDP-TOP) relatados na literatura.

Jia et al (2020) analisam o uso do LBP-TOP com relação a outros dois descritores de textura e o sumário da análise pode ser observado na Tabela 2 a seguir.

Resultados da avaliação em % no conjunto de dados Oulu-NPU DB com o classificador Softmax						
Método	ACER Impressão	ACER Display	BPCER20	BPCER10	ACER	Rank
LBP-TOP	44.17	34.58	61.67	51.67	47.08	21
WLD-TOP	55.00	35.83	84.17	80.33	55.00	30
LDP-TOP	45.83	41.67	75.00	64.17	50.83	27

Tabela 2 - Comparação do desempenho de descritores de textura. Fonte: Jia et al (2020).

Comparando-se os métodos descritos na presente seção, a abordagem proposta por Boulkenafet et. al 2016 destaca-se pela eficiência e acurácia obtida, além disso, este trabalho tem uma expressiva importância evidenciada pela alta quantidade de citações ao referido trabalho.

2.1.2 Implementação de métodos baseados em qualidade da imagem

Este método é similar à detecção baseada em textura, porém utiliza-se métricas de distorção de imagem: reflexão, embaçamento (blurriness), distorção cromática, e distorção na

diversidade de cores. Uma das implementações mais relevantes encontradas durante o levantamento foi a proposta por Liu (2014), na qual utiliza-se como entrada apenas duas imagens ou frames (ou seja, é um método estático) e adota o desfoque como o principal descritor para a detecção.

A primeira etapa do método extrai da entrada três características (foco do Laplacian modificado - LAPM, histograma de potência, e GLOH - Gradient Location and Orientation Histogram). Ainda quanto à entrada, o método captura uma imagem com foco no nariz e outra com foco no fundo. Imagens provenientes de uma pessoa real possuem uma diferença no desfoque entre estas duas regiões. Quando a captura foca no fundo, a região do nariz fica desfocada. Quando a captura foca no nariz, o fundo fica desfocado. A diferença entre as imagens deve-se às regiões estarem em diferentes campos de profundidade.

Por outro lado, ataques (provenientes de fotos e de tablets) produzem imagens similares apesar do ajuste de foco. O autor conclui que não apenas a diferença de foco é maior quando comparado com o fundo, como a região do fundo é mais fácil de ser extraída do que a região da orelha (utilizada pelo trabalho antecessor). O método utiliza uma medida comum para efeitos de foco: Sum-Modified-Laplacian (SML). A SML mede a distância entre quaisquer objetos e o plano de foco. Quanto mais próximo do plano focalizado, maior o valor SML. Ou seja, os pontos que se encontram no campo com a profundidade desejada para o foco possuem o maior valor.

Os experimentos apresentaram bons resultados no conjunto de dados em comparação com o método anterior, que compara a diferença de foco entre o nariz e a orelha com uma combinação híbrida de descritores de qualidade e de frequência (LAPM, histogramas de potência, e GLOH). Um lado contra da abordagem, segundo o autor, é que a qualidade da imagem pode afetar significativamente o desempenho do método.

2.2 Métodos baseados em frequência

Os métodos baseados em análise de frequência fazem uso de sequências de imagens onde há variação, mesmo que pequena, de pose e expressões faciais. Assume-se, portanto, que em imagens falsas essa variação não existirá ou será mais sutil e, por isso, estes métodos são eficientes contra ataques de apresentação do tipo impresso e exibição de vídeo.

A maneira mais difundida para realizar esse tipo de detecção é utilizar o valor de energia da imagem no domínio da frequência, que pode ser obtido por meio de manipulações matemáticas oriundas da transformada de Fourier (LI et al, 2004).

De maneira inicial, é necessário que cada uma das imagens das sequências adquiridas seja transformada em escala de cinza de 8 bits. Isso é fundamental uma vez que a escala de cinza descreve a quantidade de luz, sem as componentes de cor. Alguns autores também aplicam o processo de binarização, responsável por deixar a imagem apenas em preto e branco por meio de uma aproximação. Se os pixels tiverem uma coloração mais próxima de branco do que de preto, tornam-se puramente brancos, e vice-versa. Resultados preliminares desenvolvidos mostraram que este processo pode atrapalhar ao invés de auxiliar na análise de Liveness, mas pode ser útil em processos de filtragem de determinadas frequências.

Com a imagem já em escala de cinza, é necessário que a imagem seja transformada em um vetor bidimensional contendo a quantidade de luz de cada pixel. Esta matriz pode então ser manipulada matematicamente.

As imagens supracitadas ainda se encontram no domínio espacial, ou seja, são projeções bidimensionais onde cada ponto no espaço representa um pixel, contendo uma informação de quantidade de luz. Para que possam ser trabalhadas no domínio da frequência, deve-se utilizar a transformada rápida de Fourier (FFT), criada pelo estatístico John Turkey. A Figura 5 ilustra um exemplo de apresentação legítima e um ataque bem como as respectivas representações após o processamento. Esse algoritmo é fundamental para encontrar a transformada discreta de Fourier (DFT), que permitirá plotar o espectro de Fourier. É importante, também, mover a componente de frequência zero pro centro do espectro.

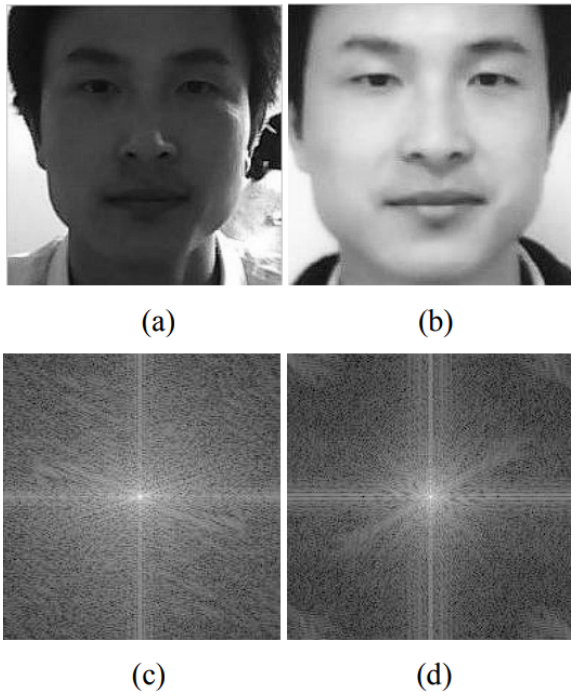

Figura 5 - Diferenças entre uma imagem real e uma imagem falsa e suas respectivas representações no domínio da frequência: (a) Imagem Real; (b) Imagem Falsa; (c) Espectro de Fourier em magnitude logarítmica de (a); (d) Espectro de Fourier de (b). Fonte: Parveen et al (2016)

Dado o espectro de Fourier, é possível calcular a quantidade de energia total da imagem utilizando a equação abaixo:

$$x = \iint |F(u,v)|dudv$$

Múltiplos autores apontam que o valor de energia de uma imagem real tende a ser maior que o de uma imagem falsa. Contudo, para dizer que uma sequência de imagens possui energia superior a outra é necessário o uso de um descritor de frequência (FD), que corresponde ao desvio padrão do valor de energia de cada imagem da sequência. Este descritor é representado pela equação abaixo, onde n é o índice da imagem e x corresponde ao valor de energia da imagem.

$$FD = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - x_m)^2\right)^{\frac{1}{2}}$$

Existem múltiplos fluxogramas de avaliação de sequências de imagens, descritas por autores distintos. Todos os avaliados também fazem uso de limiares de FD para descrever se uma sequência é real ou se corresponde a um ataque. Considerando apenas os valores de

energia e de FD, já é possível, segundo os autores dos estudos analisados, avaliar se a sequência se trata de uma face real ou forjada, conforme demonstrado pela Figura 6:

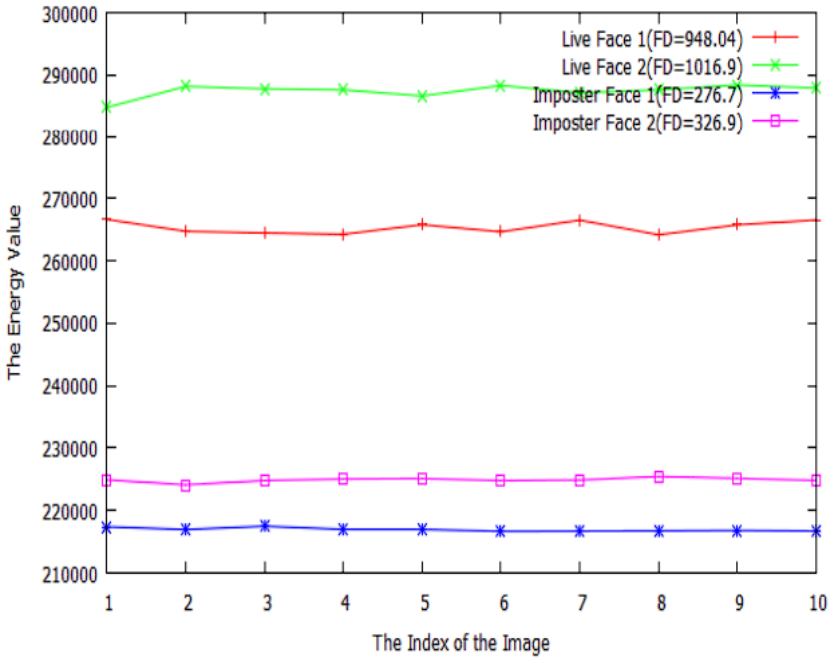

Figura 6 - Quatro sequências de imagens diferentes com seus determinados valores de FD e de energia. Fonte: Parveen et al, 2016.

Apesar de o gráfico da Figura 2 explicitar as diferenças no valor de energia entre uma imagem real e uma falsa ser da ordem de 10^4, é difícil generalizar o método para diferentes dispositivos de captura. O método também apenas tende a ser efetivo apenas contra dois tipos de ataques diferentes.

Contudo, por se tratar de uma análise matemática, os resultados são obtidos rapidamente e são independentes de modelo de treino, apenas necessitam de ajuste de limiares a partir do dispositivo de entrada.

Este método pode ser uma excelente alternativa para criar uma filtragem inicial, que descarte sequência de imagens com baixa energia (inferiores a um limiar de FD) que podem significar um ataque. Caso seja superior a um limiar, é possível aplicar outro método de maior confiança, ainda que implique em maior custo computacional.

2.3 Métodos baseados em movimentação

O principal diferencial dessa categoria é o uso de informações temporais referentes à biometria apresentada. No contexto da biometria facial, essa categoria visa identificar piscar de olhos, rotação da cabeça, movimento de músculos faciais, entre outras características.

As técnicas de análise de textura apresentam bons resultados para avaliação da vivacidade e as técnicas baseadas na análise de frames consecutivos estão evoluindo e apresentando bons resultados, principalmente na detecção de *deep fakes* – vídeos falsos com alto grau de realismo (ROUAST et al, 2018).

Aquelas técnicas baseadas na avaliação de mais de um frame consecutivo no tempo tentam detectar mudanças que podem ser voluntárias ou involuntárias. Aquelas voluntárias são dependentes de estímulos e não são muito desejadas quando se deseja um comportamento mais natural, não estimulado da pessoa submetida.

A verificação do batimento cardíaco pode ser detectada pela mudança de coloração da pele em frames consecutivos e já vem sendo utilizado para PAD e *deep fakes*. A coloração da pele muda pela oxigenação do sangue e a incidência da luz no tecido com sangue mais ou menos oxigenado. Câmeras web convencionais podem ser usadas para extrair essa informação desde que em uma distância adequada e iluminação dentro de uma faixa apropriada.

Em Hernandez-Ortega et al (2022), os autores apresentam uma análise da detecção de batimentos cardíacos especificamente em *deep fakes*. Para tanto, é necessário se ter uma porção da pele do rosto disponível para avaliação. Normalmente a testa e as bochechas são mais utilizadas por permitirem melhor avaliação.

As técnicas normalmente incluem a extração de informações advindas da frequência de características das imagens da pele e seus picos comparando com imagens de frames consecutivos.

2.4 Métodos baseados em características aprendidas (*data-driven*)

Esta categoria não está presente no levantamento de Ramachandra e Busch, 2017 porém, têm se destacado como estado da arte em abordagens para classificação de imagens (KHAIRNAR et al, 2023). Desta maneira, sob a perspectiva da aprendizagem de máquina, a

detecção de vivacidade pode ser modelada como um problema de classificação, ou seja, dada uma imagem classificá-la em uma apresentação real ou um ataque de apresentação.

Considerando-se as abordagens apresentadas por Ali et al (2021) e Khairnar et al (2023), as implementações com melhores desempenho são as que utilizam-se de redes neurais convolucionais. Em Cai et al (2022) os autores propõem o uso da arquitetura convolucional ResNet (HE et al, 2016) na detecção de vivacidade em biometria facial. Comparando-se com as principais abordagens para detecção de vivacidade, esse tipo de abordagem substitui o LBP por redes neurais que extraem *Meta Patterns* (uma representação implícita dos descritores). Portanto, trata-se de um método híbrido que combina a imagem RGB e os *Meta Patterns*.

Os autores constataram, por meio do uso de bases diferentes para treinamento e teste, que o método também é capaz de detectar ataques desconhecidos durante o treinamento, conforme ilustrado na Figura 7.

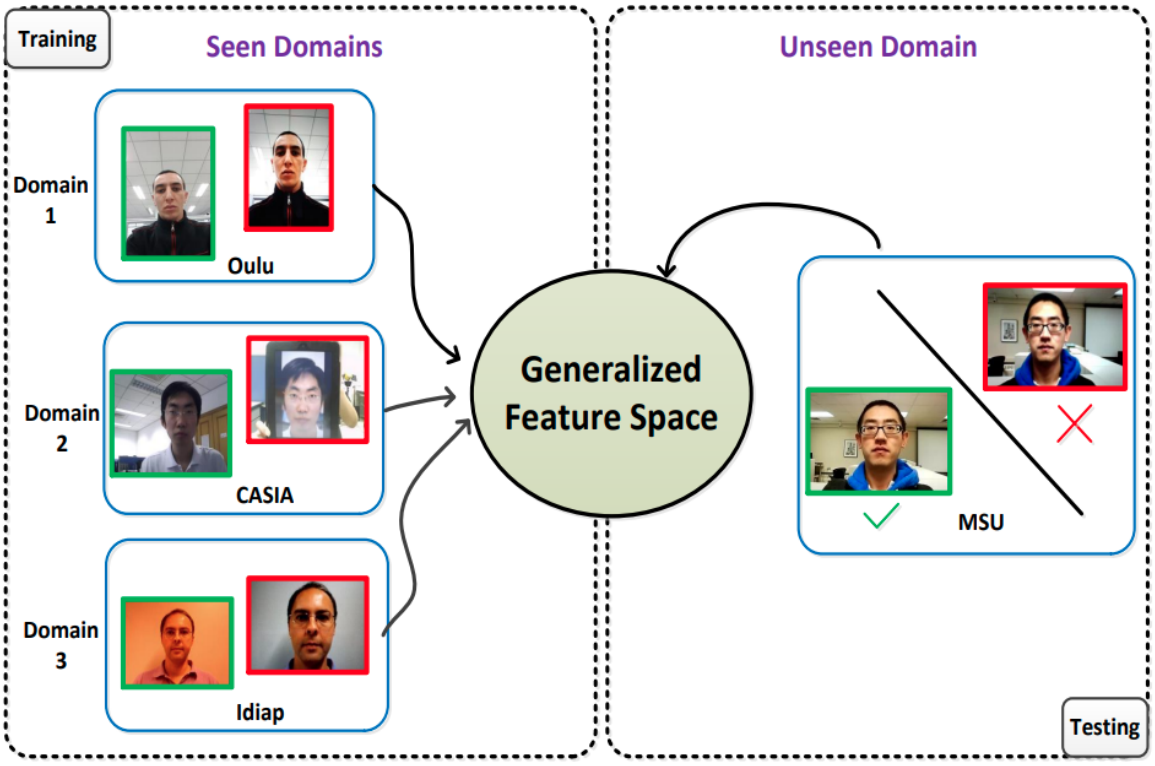

Figura 7 - Validação cruzada utilizada no treino e teste da ResNet. Fonte: He et al (2016).

O modelo proposto consiste em uma rede neural capaz de extrair meta padrões mediante o uso da rede ResNet-50 (RM) em dois fluxos de imagens, um usando RGB e o outro usando os meta padrões, para fazer uma classificação binária. Além disso, utiliza um

módulo de fusão hierárquica (HFM) nas camadas do seu modelo. A ligação entre as entradas, a rede neural e a detecção é sintetizada na Figura 8:

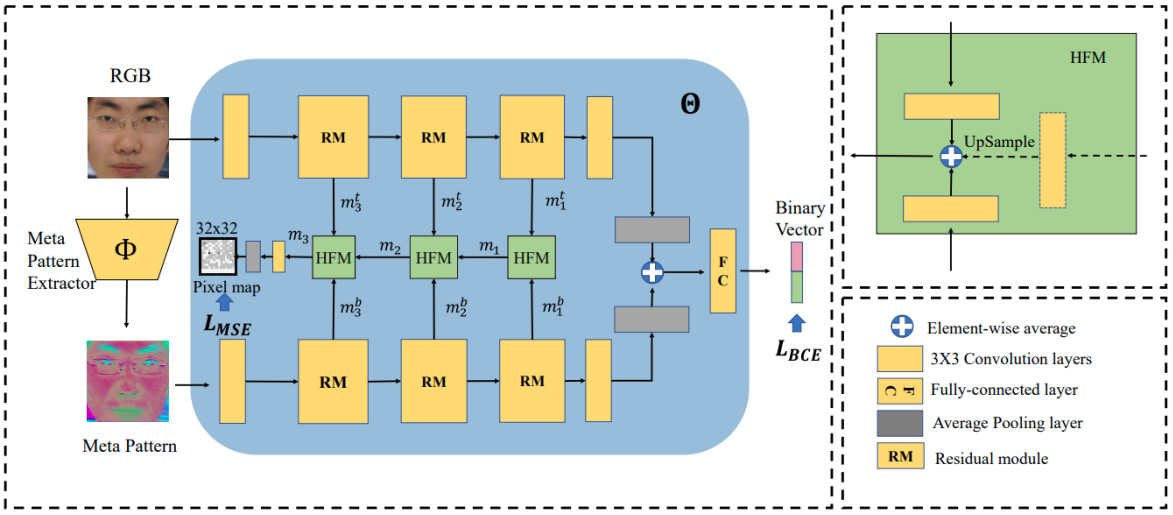

Figura 8 - Modelo para detecção de vivacidade com a ResNet. Fonte: He et al (2016).

A avaliação do modelo foi realizada utilizando-se de validação cruzada, ou seja, o modelo foi treinado utilizando-se de alguns conjuntos de dados públicos e testado (avaliado) com conjuntos de dados não utilizados durante o treinamento. Esta abordagem avaliativa é mais difícil do que a tradicional em que os conjuntos de treino e teste são obtidos a partir de um único conjunto de dados. Desta maneira é possível avaliar o desempenho considerando-se sensores e materiais (artefatos) de ataque não vistos durante o treinamento. A Tabela 3 apresenta os dois protocolos utilizados para a realização desta avaliação (os conjuntos de dados utilizados são descritos no Capítulo 3.

Protocolo	Treinamento	Teste
MICO	MSU-MFSD (M), IDIAP REPLAY-ATTACK (I)	CASIA-FASD (C), OULU-NPU (O)
MICY	MSU-MFSD (M), IDIAP REPLAY-ATTACK (I)	CASIA-FASD (C), ROSE-YOUTU (Y)

Tabela 3 - Protocolos de avaliação propostos por He et al (2016).

A abordagem proposta por He et. al 2016 caracteriza um método que pode ser aplicado com diferentes arquiteturas de redes neurais, com ou sem a utilização de meta padrões. Ademais, a abordagem de validação cruzada também é comumente utilizada em

competições da área e também na avaliação de soluções comerciais por entidades certificadoras. No Capítulo 4 são apresentados os resultados da utilização da ResNet, VGG e Senet para a detecção de vivacidade.

2.5 Métodos híbridos

A literatura também conta com métodos que combinam técnicas descritas nas seções anteriores. Em Määttäj et al (2012), apresenta-se um método que estende a abordagem de detecção de falsificação usando análise de microtextura baseada em LBP e introduz dois recursos complementares de baixo nível à descrição da face: Gabor wavelets e Histogram of Oriented Gradients (HOG). A proposta adota dois recursos de textura, LBP e wavelets de Gabor, para descrever não apenas as microtexturas, mas também mais informações macroscópicas.

Além disso, a descrição da forma local é introduzida usando histograma de gradientes orientados. Cada descritor de baixo nível produz sua própria representação de face na qual o mapa de kernel homogêneo é aplicado para transformar os dados em uma representação linear compacta. Cada vetor em seu próprio espaço de recursos transformado é então alimentado a um classificador SVM linear e a fusão de nível de pontuação das saídas SVM individuais determina se há uma pessoa viva ou uma imagem falsa na frente da câmera.

Já Kim et al (2015) descrevem um método baseado em imagem única para discriminar máscaras de papel 2-D das faces vivas. A proposta trabalha com informações de frequência usando espectro de potência para diferenciar forma e com textura usando padrão binário local para diferenciar detalhamento. Para a análise de frequência, é empregado um método baseado em espectro de potência que explora não apenas as informações de baixa frequência, mas também as informações que residem nas regiões de alta frequência. Para análise de textura, é empregado um método de descrição baseado em Padrão Binário Local (LBP) cujo desempenho é invariante com relação a escalas-de-cinza e rotações.

O método proposto combina o valor de decisão do classificador SVM treinado com base no espectro de potência com o classificador SVM treinado com base em LBP. Esses dois valores de decisão, produzidos por diferentes metodologias de extração de recursos, são então concatenados como um vetor de recursos 2D e usados para treinar um método baseado em fusão.

Wen et al (2015) apresentam um método baseado em análise de distorção de imagem (IDA). Quatro características diferentes (reflexo especular, desfoque, momento cromático e diversidade de cores) são extraídas para formar o vetor de características IDA. Um classificador SVM é utilizado para a determinação final da classe entre apresentação real ou ataque. Primeiro, as amostras falsas (spoof) são divididas em K grupos de acordo com o tipo de ataque. Depois, um conjunto de treinamento específico é construído combinando todas as amostras genuínas e um único grupo de amostras falsas, resultando em K conjuntos de treinamento. Esses conjuntos são então utilizados para treinar k modelos SVM que irão compor o classificador final do método.

3. CONJUNTOS DE DADOS

Previamente à alta disponibilidade de métodos baseados em aprendizado de máquina, os conjuntos de dados de apresentações legítimas e ataques eram utilizados principalmente para a avaliação e ranqueamento das abordagens. Além disso, entidades certificadoras também mantêm seus próprios conjuntos avaliativos utilizados durante o processo de certificação para abordagens comerciais.

Apesar da maioria dos levantamentos e análises referentes a conjuntos de dados biométricos estarem concentrados em reconhecimento facial, alguns trabalhos apresentam panoramas específicos para a detecção de vivacidade. Yu et al (2021) levanta o estado da arte de deep-learning para ataques de apresentação e relata 36 bases de dados. Os autores descrevem, para cada base, o número de indivíduos, modalidades, sensores especializados, tipos de arquivos (imagens, vídeos), configurações do ambiente, e tipos de ataques.

Costa-Pazo et al (2021) analisa o problema de generalização dos métodos de detecção de ataques de apresentação em sistemas de reconhecimento facial. Introduz uma nova versão da agregação de bases denominada GRAD-GPAD proposta pelos mesmos autores em 2019. A agregação conta, assim, com 13 bases RGB. Também categoriza e rotula cada base com relação a grupos demográficos: sexo, idade e cor de pele. Por sua vez, Zhang et al (2020) apresenta um novo banco de dados, CelebA-Spoof, e relata 13 bases de dados utilizadas por trabalhos relacionados. Os autores incluem o relato do número de sensores, condições de iluminação e de ambientes utilizados durante a coleta.

Nenhum dos trabalhos citados anteriormente teve como foco levantar as bases de dados disponíveis e assim, infelizmente, cada relato está incompleto de alguma maneira. Entretanto, cada trabalho possui alguma contribuição ímpar para o levantamento das bases disponíveis, seja nas bases avaliadas ou nos parâmetros utilizados para descrevê-los.

Assim sendo, com o objetivo de testar as principais abordagens elencadas no Capítulo 2, foi realizado um levantamento de conjuntos de dados considerando-se os seguintes requisitos:

- O sensor
    - é uma câmera monocular que captura o espectro visível (VIS) de cores (RGB),
    - é de um telefone celular.

- Os ataques
    - são de fotos impressão e/ou de vídeo, e
    - utilizam instrumentos de baixo custo e/ou que não seriam suspeitos no ambiente controlado em que reconhecimento será implantado (ou seja, há bases cujos ataques - com, por exemplo, máscaras, manequins, laptops, LCDs - estão fora do escopo atual desta pesquisa).
- A coleta foi feita
    - no interior de um ambiente (*indoors*),
    - sobre condições variadas de iluminação (como fonte natural/artificial, intensidade forte/normal/fraca, direção frontal/lateral/traseira),
    - com um grupo representativo de pessoas (como raça, gênero, idade),
    - com variabilidade de instrumentos de ataque (como tipo celulares/tablets, qualidade, marca) e do manuseio do mesmo (como ângulo, deformação).

Foram identificadas três bases cujos dados estão abertos ao público (apesar de protegidos legalmente para o uso comercial), e que são compatíveis com o escopo da pesquisa. O conjunto CelebA-Spoof[3] é, segundo Yu et al 2022, a maior base de ataques de apresentação capturada com uma simples câmera comercial. Esse conjunto compreende imagens de apresentações sendo que quatro tipos de ataques são reproduzidos com imagens digitais de 10.177 celebridades retiradas de redes sociais.

Por outro lado, os conjuntos MSU-MFSD[4] e OULU-NPU[5], apresentam cenários com aparelhos móveis. O MSU-MFSD compreende vídeos curtos de aproximadamente 12 segundos; dois tipos de ataques são reproduzidos com 55 voluntários. Diferente do CelebA-Spoof, são reproduzidos ataques em vídeo (por meio de um iPhone 5S). O OULU-NPU também compreende vídeos curtos (5 segundos) de ataques de impressão e de vídeo. Diferente do MSU-MFSD, a sua coleta é mais diversificada. Por exemplo, OULU-NPU estuda a possibilidade que o sensor utilizado pelo método de detecção seja um dentre seis diferentes aparelhos móveis: Samsung Galaxy S6 edge, HTC Desire EYE, MEIZU X5, Sony XPERIA C5 Ultra Dual, OPPO N3.

[3] https://github.com/ZhangYuanhan-AI/CelebA-Spoof
[4] https://sites.google.com/site/huhanhomepage/datasetcode
[5] https://sites.google.com/site/oulunpudatabase/

Na Tabela 4 apresenta-se um sumário dos conjuntos de dados e as respectivas quantidades de indivíduos utilizados para registrar as apresentações legítimas e os ataques, bem como o tipo das imagens ou vídeos e os tipos de ataques disponíveis.

Nome do base	**Desenvolvido por**	**Nº de indivíduos**	**Modalidade**	**Ataques**
CelebA-Spoof	CUHK NJTU SenseTime	10177	RGB	impressão reprodução imagem máscara de papel
MSU-MFSD	Case MSU	55	RGB	impressão reprodução de vídeo
OULU-NPU	OULU NPU	55	RGB	impressão reprodução de vídeo

Tabela 4 - Sumário quantitativo dos conjuntos de dados considerados nesta pesquisa.

As Tabelas 5, 6 e 7 consideram os principais tipos de ataques disponíveis (impressão: de fotos inteiras, recortes; reprodução de vídeos: inteira, parcial; uso de máscaras de papel), um detalhamento acerca da sua origem e qual (quais) conjuntos disponibilizam o respectivo ataque.

Ataques com impressão			
foto impressa	inkjet	a3-glossy	OULU-NPU
	laserjet	a3	MSU-MFSD
	desconhecida	a4	CelebA-Spoof
	poster		CelebA-Spoof
	photo		
recorte	rosto		CelebA-Spoof
	parte superior do corpo		
	região		

Tabela 5 - Ataques de impressão com relação aos conjuntos de dados.

Ataques com vídeos			
Reprodução de vídeo	mobile	mobile-capture	MSU-MFSD
		camera-capture	
	laptop	mobile-capture	OULU-NPU
Reprodução de imagem	mobile	mobile-capture	CelebA-Spoof
		camera-capture	
		laptop-capture	
		tablet-capture	
	laptop	mobile-capture	
		camera-capture	
		laptop-capture	
		tablet-capture	
	tablet	mobile-capture	
		camera-capture	
		laptop-capture	
		tablet-capture	

Tabela 6 - Ataques com vídeo com relação aos conjuntos de dados.

Ataques com máscara	
papel	CelebA-Spoof

Tabela 7 - Ataques com máscaras com relação aos conjuntos de dados.

Por último, apresenta-se uma organização dos ataques mediante variações nos critérios de ambiente de captura (se a apresentação foi realizada em um ambiente aberto ou fechado), tipo de iluminação, tipo de sensor, ângulo e distorções (quando o instrumento de ataque permite, por exemplo uma folha de papel que pode ser dobrada). Tais variações e os respectivos conjuntos de dados são resumidos nas Tabelas 8, 9, 10, 11 e 12 respectivamente.

Environment	
Indoors	CelebA-Spoof, MSU-MFSD, OULU-NPU
Outdoors	CelebA-Spoof

Tabela 8 - Ambiente de captura com relação aos conjuntos de dados.

Illumination	
1 variation or not-reported	MSU-MFSD
3 variations	OULU-NPU
4 variations	CelebA-Spoof

Tabela 9 - Variações de iluminação com relação aos conjuntos de dados.

Input sensors (PAD)	
mobile	CelebA-Spoof, OULU-NPU, MSU-MFSD
camera	CelebA-Spoof, MSU-MFSD
tablet	CelebA-Spoof
laptop	

Tabela 10 - Tipos de sensores com relação aos conjuntos de dados.

Angle	
1 variation or not-reported	MSU-MFSD, OULU-NPU
4 variations	CelebA-Spoof

Tabela 11 - Variações de ângulo das apresentações com relação aos conjuntos de dados.

Shape (when PAI is malleable)	
1 variation or not-reported	MSU-MFSD, OULU-NPU
4 variations	CelebA-Spoof

Tabela 12 - Variações de distorções com relação aos conjuntos de dados.

O levantamento dos conjuntos de dados fundamentou a implementação de abordagens baseadas em dados e a avaliação de todas as abordagens implementadas durante a presente pesquisa. Os conjuntos foram selecionados especialmente considerando-se os ataques mais comuns e que podem ser realizados com poucos equipamentos e baixo custo. Existem conjuntos de dados com ataques e sensores mais elaborados (máscaras de silicone e de alta fidelidade, por exemplo), entretanto tais ataques serão considerados em etapas futuras do desenvolvimento.

4. COMPARAÇÃO ENTRE ABORDAGENS

O principal resultado deste levantamento foi fundamentar a decisão inicial acerca de quais abordagens são mais promissoras para a detecção de vivacidade. Assim sendo, foram realizadas duas comparações, a primeira, com o intuito de ser um estudo preliminar em menor escala e a segunda utilizando-se mais recursos computacionais para considerar um conjunto de dados maior.

Ambas comparações foram realizadas mediante o desenvolvimento de modelos analíticos e modelos baseados em dados. Esta diferenciação impacta no tempo de execução (inferência) do mecanismo pronto e também no custo para o treinamento do modelo. No caso das abordagens analíticas, não é necessária uma etapa de treinamento e a sua execução tem baixo custo computacional. Já as abordagens baseadas em dados necessitam treinamento e o custo computacional depende da arquitetura utilizada.

Com relação aos modelos analíticos, foram implementados e testados modelos para detecção de ataques com base na frequência do sinal capturado, textura da imagem e detecção de batimentos cardíacos em vídeos. Já os modelos baseados em dados utilizam-se de técnicas de aprendizagem de máquina, especialmente abordagens supervisionadas. Nesse grupo, foram implementados e testados modelos de redes neurais tradicionais e convolucionais. No contexto da primeira comparação, foi implementada uma rede neural tradicional, sem recursos convolucionais. A Tabela 13 apresenta o desempenho dos modelos implementados considerando-se o treinamento e avaliação com apenas um conjunto de dados.

Abordagem	Erro em Ataques	Erro em Legítimas
Textura	1,16%	30,28%
Rede neural	2,50%	20,00%
Freqüencia	26,40%	20,90%
Textura Dinâmica	13%	47%

Tabela 13 - Desempenho de métodos de detecção de vivacidade com apenas um conjunto de dados.

Mesmo considerando-se um escopo reduzido de conjunto de dados, observa-se o melhor desempenho das abordagens baseadas em dados com relação às analíticas. Entretanto,

implementações de modelos para análise de textura de imagens apresentam um desempenho relevante considerando-se que são mais rápidas e econômicas do que as redes neurais.

Após a comparação preliminar, foram mantidas as abordagens analíticas com melhor desempenho e adicionaram-se abordagens baseadas em redes neurais convolucionais, pré-treinadas e sem treino inicial. Além disso, foi possível também ampliar os conjuntos de dados para treinamento e avaliação. Os resultados obtidos nesta etapa comparativa estão sintetizados na Tabela 14.

Abordagem	**Conjunto de dados**	**Erro em Ataques**	**Erro em Legítimas**
Textura	MSU + UVAD	1%	30%
Textura	MSU + UVAD + OULU	7%	21%
Textura	MSU + UVAD + OULU + RECOD	2,5%	56%
Frequência + *Visual Rythm*	MSU	19%	14%
Textura	MSU + RECOD + dados privados	1%	30%
Resnet (Convolucional)	MSU + RECOD + dados privados	2%	20%
VGG (Convolucional)	MSU + RECOD + dados privados	1%	2%
Frequência	MSU + ERCOD + dados privados	19%	14%

Tabela 14 - Síntese dos resultados comparativos utilizando-se de 2 a 3 conjuntos de dados.

A partir dessa análise, nota-se o desempenho superior das abordagens baseadas em dados e da abordagem baseada em textura. Assim sendo, estas abordagens foram priorizadas com a execução de mais épocas de treinamento. Além disso, com essas abordagens, também realizou-se uma etapa de verificação com uma simulação para garantir que não houvesse overfitting com os dados, o que foi diagnosticado com a VGG, por exemplo.

Para esta última análise comparativa, consolidou-se um conjunto de dados com um total de 18.000 imagens, sendo metade de ataques e metade de apresentações reais fornecidas por uma empresa atuante na área. Utilizando-se esses dados e os modelos com melhor desempenho nas análises anteriores foram obtidos os seguintes resultados (Tabela 15):

Abordagem	Erro em Ataques	Erro em Legítimas
LBP	2,1%	6,7%
MobileNet	1,8%	4,6%
Resnet 50	0,8%	4,3%
Resnet 101 (4 épocas)	3,4%	2,9%
Resnet 101 (3 épocas)	2,1%	3,2%
Resnet 50	1,6%	2,3%
CNN	5,0%	1,6%
Senet	1,2%	0,2%

Tabela 15 - Resultados finais obtidos com ajustes em todos os modelos.

Os resultados dessa etapa indicaram que o balanceamento do conjunto de dados, bem como a análise do viés, aprimoraram significativamente o desempenho de todos os modelos, inclusive do método analítico (textura LBP), o qual passamos a integrar com uma rede neural na etapa final do método.

Os resultados obtidos já alcançam e, para algumas categorias, ultrapassam o estado da arte em métodos acadêmicos porém não superam as melhores abordagens comerciais as quais utilizam conjuntos de dados ainda maiores.

Considerando-se a Tabela 15, é possível observar uma melhora significativa em quase todos os modelos. Tal resultado é corriqueiro em modelos de dados, uma vez que quanto maior a quantidade de exemplares disponíveis para treinamento, maior será a capacidade de generalização do modelo. Outro aspecto de destaque na Tabela 15 refere-se à quantidade de épocas utilizadas nos treinamentos. Usualmente esse tipo de abordagem beneficia-se de mais tempo de treinamento, ou seja, mais épocas. Entretanto, percebeu-se que ao ultrapassar 10 épocas os resultados não melhoraram significativamente e também aumentou o sobre ajuste (overfitting) dos modelos ao conjunto de treinamento.

Referências bibliográficas

ALI, S. F.; KHAN, M. A.; ASLAM, A. S. Fingerprint matching, spoof and liveness detection: classification and literature review. **Frontiers of Computer Science**, v. 15, n. 1, 29 set. 2020.

BENLAMOUDI, A. et al. **Face spoofing detection using Multi-Level Local Phase Quantization (ML-LPQ) Face spoofing detection using Multi-Level Local Phase Quantization (ML-LPQ)**. International Conference on Automatic control, Telecommunications and Signals (ICATS15). **Anais**...2015. Acesso em: 5 dez. 2023

BHATTACHARYYA, D. et al. Biometric authentication: A review. **International Journal of u-and e-Service, Science and Technology**, v. 2, n. 3, p. 13–28, 2009.

BOULKENAFET, Z.; KOMULAINEN, J.; HADID, A. **Face Spoofing Detection Using Colour Texture Analysis**. IEEE Transactions on Information Forensics and Security. **Anais**...ago. 2016. Acesso em: 1 fev. 2021

CAI, R. et al. **Learning Meta Pattern for Face Anti-Spoofing**. IEEE Transactions on Information Forensics and Security. **Anais**...Institute of Electrical and Electronics Engineers, 1 jan. 2022. Acesso em: 6 dez. 2023

COSTA‐PAZO, A. et al. Face presentation attack detection. A comprehensive evaluation of the generalisation problem. **IET Biometrics**, v. 10, n. 4, p. 408–429, 28 jun. 2021.

FREITAS PEREIRA, T. DE et al. Face liveness detection using dynamic texture. **EURASIP Journal on Image and Video Processing**, v. 2014, n. 1, 7 jan. 2014.

HE, K. et al. **Deep Residual Learning for Image Recognition**. arXiv (Cornell University). **Anais**... In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. Cornell University, 10 dez. 2015.

HERNANDEZ-ORTEGA, J. et al. DeepFakes Detection Based on Heart Rate Estimation: Single- and Multi-frame. In: **Handbook of Digital Face Manipulation and Detection**. [s.l.] Springer, 2022. p. 255–273.

ISO/IEC 30107 Biometric presentation attack detection. [s.l.] ISO/IEC, 2016.

JIA, S. et al. Face presentation attack detection in mobile scenarios: A comprehensive evaluation. **Image and Vision Computing**, v. 93, p. 103826, nov. 2019.

KHAIRNAR, S. et al. Face Liveness Detection Using Artificial Intelligence Techniques: A Systematic Literature Review and Future Directions. **Big Data and Cognitive Computing**, v. 7, n. 1, p. 37, 1 mar. 2023.

KIM, G. et al. **Face liveness detection based on texture and frequency analyses**. 5th IAPR International Conference on Biometrics (ICB). **Anais**...1 mar. 2012. Acesso em: 6 dez. 2023

LI, J. et al. **Live face detection based on the analysis of Fourier spectra**. (N. K. Ratha, Ed.)Proceedings of the SPIE. **Anais**...25 ago. 2004. Disponível em: <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5404/1/Live-face-detection-based-on-the-analysis-of-Fourier-spectra/10.1117/12.541955.full>

LIU, Y. **Face liveness detection by focusing on frontal faces and image backgrounds**. International Conference on Wavelet Analysis and Pattern Recognition. **Anais**...1 jul. 2014. Acesso em: 19 set. 2023

MÄÄTTÄJ.; HADID, A.; PIETIKÄINENM. Face spoofing detection from single images using texture and local shape analysis. **IET Biometrics**, v. 1, n. 1, p. 3, 2012.

MEI, L. et al. WLD-TOP Based Algorithm against Face Spoofing Attacks. **Lecture Notes in Computer Science**, v. 9428, p. 135–142, 1 jan. 2015.

PARVEEN, S. et al. Face Liveness Detection Using Dynamic Local Ternary Pattern (DLTP). **Computers**, v. 5, n. 2, p. 10, 24 maio 2016.

PHAN, Q.-T. et al. **FACE spoofing detection using LDP-TOP**. 2016 IEEE International Conference on Image Processing (ICIP). **Anais**...1 set. 2016. Acesso em: 5 dez. 2023

RAHEEM, E. A.; AHMAD, S. M. S.; ADNAN, W. A. W. Insight on face liveness detection: A systematic literature review. **International Journal of Electrical and Computer Engineering (IJECE)**, v. 9, n. 6, p. 5165–5175, 1 dez. 2019.

RAMACHANDRA, R.; BUSCH, C. Presentation Attack Detection Methods for Face Recognition Systems. **ACM Computing Surveys**, v. 50, n. 1, p. 1–37, 20 mar. 2017.

ROUAST, P. V. et al. Remote heart rate measurement using low-cost RGB face video: a technical literature review. **Frontiers of Computer Science**, v. 12, n. 5, p. 858–872, 18 set. 2018.

WEN, D.; HAN, H.; JAIN, A. K. Face Spoof Detection With Image Distortion Analysis. **IEEE Transactions on Information Forensics and Security**, v. 10, n. 4, p. 746–761, abr. 2015.

YU, Z. et al. Deep Learning for Face Anti-Spoofing: A Survey. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 45, p. 1–22, 2022.

ZHANG, Y. et al. **CelebA-Spoof: Large-Scale Face Anti-spoofing Dataset with Rich Annotations**. Computer Vision – ECCV 2020. **Anais**...2020.

Glossário

GLCM (Gray Level Co-occurrence Matrices): Calcula a frequência com que pares de pixels com um valor e deslocamento específicos ocorrem na imagem. Método com poder de representação global discriminante na detecção de ataques.

HOG (Histogram of Oriented Gradients): Métodos baseados em histograma de gradientes orientados capturam as estruturas de borda ou gradiente da imagem facial para distinguir artefatos reais de ataques. São usados para extrair características relativas ao contraste, luminosidade e formas.

LBP (Local Binary Pattern) : Descreve um padrão binário para uma vizinhança circular, a partir de limiarização dada para o valor do pixel central. Métodos baseados em LBP mostram excelente desempenho em detecção de ataques com fotos impressas e com máscaras.

LBP-TOP (Local Binary Pattern from Three Orthogonal Planes): Adaptação de LBP para capturar informações temporais de um vídeo. Define três planos ortogonais que intersectam o pixel avaliado, e aplica LBP três vezes. O descritor, assim, combina tanto informações espaciais quanto temporais. A resolução é ajustável.

LDP (Local Derivative Pattern): Codifica padrões direcionais com base nas variações locais da derivada. Consegue capturar mais informações que LBP, o qual captura padrões locais de primeira ordem. O LDP, por outro lado, permite trabalhar com outras ordens de padrões de textura.

LDP-TOP (Local Derivative Pattern from Three Orthogonal Planes): Adaptação de LDP para capturar informações temporais de um vídeo. Define três planos ortogonais que intersectam o pixel avaliado, e aplica LDP três vezes. O descritor combina informações espaciais e temporais com diferentes direções dos sutis movimentos de cabeça. A resolução é ajustável.

LPQ (Local Phase Quantization): Tem estrutura semelhante ao LBP, mas codifica informação de fase extraída da STFT (Short-Term Fourier Transform) para cada patch da imagem. Explora propriedades da convolução em frequência que contribuem para a invariância ao borramento.

WLD (Weber Local Descriptor): Descritor de textura inspirado na Lei de Weber-Fechner, acerca da relação existente entre a magnitude física de um estímulo e a percepção de sua

intensidade. A Lei de Werber diz que o tamanho de uma diferença perceptível possui uma proporção constante com o estímulo original. Assim, por exemplo, em um ambiente com muito barulho, uma pessoa precisa gritar para ser escutada enquanto que um sussurro da mesma pessoa é suficiente em uma sala silenciosa. Produz um histograma de gradiente e orientação.

WLD-TOP (Weber Local Descriptor from Three Orthogonal Planes): Adaptação de WLD para capturar informações temporais de um vídeo. Define três planos ortogonais que intersectam o pixel avaliado, e aplica WLD três vezes. O descritor, assim, combina tanto informações espaciais quanto temporais. A resolução é ajustável.