LUAN PORFIRIO E SILVA

RASTREAMENTO FACIAL E REFINAMENTO DE PONTOS FIDUCIAIS 3D BASEADO NA REGIÃO DO NARIZ EM AMBIENTES NÃO CONTROLADOS

CURITIBA

2017

LUAN PORFIRIO E SILVA

RASTREAMENTO FACIAL E REFINAMENTO DE PONTOS FIDUCIAIS 3D BASEADO NA REGIÃO DO NARIZ EM AMBIENTES NÃO CONTROLADOS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.
Orientadora: Profa. Dra. Olga R. P. Bellon
Coorientador: Prof. Dr. Luciano Silva

CURITIBA

2017

SI586r Silva, Luan Porfirio e
Rastreamento facial e refinamento de pontos fiduciais 3D baseado na região do nariz em ambientes não controlados / Luan Porfirio e Silva. – Curitiba, 2017.
78 f. : il. color. ; 30 cm.

Dissertação - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática, 2017.

Orientadora: Olga R. P. Bellon.
Coorientador: Luciano Silva.

1. Rastreamento da face. 2. Alinhamento de face 3D. 3. Refinamento de pontos fiduciais. I. Universidade Federal do Paraná. II. Bellon, Olga R. P. III. Silva, Luciano. IV. Título.

CDD: 006.693

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor CIÊNCIAS EXATAS
Programa de Pós-Graduação INFORMÁTICA

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **LUAN PORFIRIO E SILVA** intitulada: **RASTREAMENTO FACIAL E REFINAMENTO DE PONTOS FIDUCIAIS 3D BASEADO NA REGIÃO DO NARIZ EM AMBIENTES NÃO CONTROLADO**, após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 26 de Maio de 2017.

OLGA REGINA PEREIRA BELLON
Presidente da Banca Examinadora (UFPR)

RODRIGO MINETTO
Avaliador Externo (UTFPR)

ROBERTO PEREIRA
Avaliador Interno (UFPR)

LUCIANO SILVA
Co-orientador - Avaliador Interno (UFPR)

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
Setor de Ciências Exatas
Coordenação de Pós-Graduação em Informática

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

O rastreamento da face utiliza a informação temporal para inferir a posição da face em cada *frame*. Uma de suas aplicações é em cenários não controlados, onde os métodos de detecção da face falham. As abordagens atuais presentes na literatura são baseadas em pontos fiduciais, porém encontram dificuldades quando aplicadas em ambientes não controlados, devido este cenário não ser trivial. Para lidar com esta dificuldade, o presente trabalho propõe uma abordagem baseada no método de rastreamento visual de Nam e Han, inicializando-o com a região do nariz detectada no *frame* de vídeo com melhor qualidade da face. A região do nariz, ao invés da face inteira, foi escolhida devido sua menor probabilidade de estar oclusa, ser invariante a expressões faciais e visível em grande variações de pose. Foram realizados experimentos na base de dados 300 *Videos in the Wild* e *Point and Shoot Challenge*, em comparação ao rastreamento de face. Já o alinhamento de faces em ambientes sem restrições consiste em localizar pontos fiduciais com precisão, auxiliando em tarefas como reconstrução 3d, reconhecimento e análise de expressões faciais. Em situações onde existem variações de poses, a aparência da face difere da face frontal, dificultando a tarefa de alinhamento. Para contornar esta situação, este trabalho propõe refinar a localização dos pontos fiduciais com regressão em cascata e regressão com vetores de suporte (SVR), partindo da pose estimada com a região do nariz. Experimentos foram realizados na base de dados 3D *Face Alignment in the Wild*, demonstrando resultados amplamente superiores ao alinhamento baseado na pose e comparáveis ao estado-da-arte.

Palavras-chave: rastreamento da face; alinhamento de face 3d; refinamento de pontos fiduciais.

ABSTRACT

Face tracking uses temporal information to infer the position of the face in each frame. One of its applications is in-the-wild environments where face detection methods fail to perform robustly. Current approaches presented in the literature are based on facial landmarks. Therefore, they have limitations when applied in in-the-wild environments as estimating the landmarks in such scenarios is not trivial. To address this issue, this work propose an approach based on a Nam and Han's visual tracking method, initializing it with the nose of the best quality face in the video sequence. The nose region, rather than the entire face was chosen due to it being unlikely to be occluded, mostly invariant to facial expressions and visible in a long range of head poses. We performed experiments on the 300 Videos in the Wild and Point and Shoot Challenge datasets. Face alignment in unrestricted environments, however, consists of accurately locating landmarks, assisting in tasks such as 3D reconstruction, recognition and facial expression analysis. In situations where there are variations of poses, the face appearance differs from the frontal face, making the alignment task difficult. To overcome this issue, this work proposes to refine the landmarks by using cascade regression and support vector regression (SVR), starting from the pose estimated with the nose region. Experiments were performed on the 3D Face Alignment in the Wild dataset, showing results broadly superior to pose-based alignment and comparable to the state-of-the-art.

Keywords: face tracking; 3d face alignment; landmark refinement.

CAPÍTULO 1

INTRODUÇÃO

Segundo Jain *et al.* [28], o reconhecimento facial é um dos problemas amplamente estudados no campo de visão computacional. Com esta finalidade, as abordagens existentes assumem que o primeiro estágio é a detecção da face. Em consideração a vídeos, a detecção da face pode ser associada à informação temporal, de forma que a localização da região de interesse nos *frames* subsequentes sejam estimados através do rastreamento.

Embora as principais abordagens de rastreamento da face utilizem pontos fiduciais [44,59,65], tais métodos encontram dificuldades em lidar com ambientes sem restrições, os quais apresentam oclusões parciais, variações de escala, problemas de iluminação, resolução e diferentes orientações (pose).

O Rastreamento visual genérico é uma alternativa sem pontos fiduciais que tem sido aplicado com sucesso em estimar a localização de diferentes regiões, incluindo faces [25, 36,38,55]. Desta forma, foi proposto localizar faces em vídeos não controlados através do rastreamento visual estado-da-arte de Nam e Han [38], utilizando como região de interesse apenas o nariz, possibilitando aumentar a confiabilidade do resultado.

O nariz é um componente facial que já foi comprovado ser eficiente para a biometria em Jain *et al.* [9], Zavan [67] e Zehngut *et al.* [68]. Ao contrário dos olhos e orelhas, o nariz é um elemento facial visível em *frames* de perfil, e, diferentemente da boca, ele é praticamente invariante a deformações provocadas por expressões faciais. Outro fator relevante é que tal região não se prejudica por oclusão mediante acessórios ou pelos faciais.

Dado que a informação extraída pelo primeiro *frame* reflete diretamente no desempenho do rastreamento, foi proposto combiná-lo com a escolha automática de *frame* de melhor qualidade [46,67] para inicialização do rastreamento, evitando que o resultado obtido seja prejudicado em casos onde o primeiro *frame* contenha dificuldades como baixa iluminação, contraste, foco ou orientação distante do frontal.

Já desafio do alinhamento facial pode ser formulado como, a partir de uma imagem facial, localizar com precisão pontos pré-definidos em regiões discriminantes da face, tais como olhos, nariz, boca, sobrancelha e contorno, traçando, desta forma, a geometria da face. Por meio do alinhamento facial é possível extrair informações que auxiliam no reconhecimento facial, reconstrução 3d e análise de expressões faciais.

Embora o alinhamento facial tenha alcançado resultados com grande precisão nas principais bases de dados em alinhamento 2d [42, 43], estas por sua vez contém faces com orientação próxima ao frontal, sem a existência de poses extremas (i.e. faces em perfil), evento este que apresenta maior dificuldade.

Por meio do alinhamento 3d é possível inferir os pontos fiduciais de faces 2d e quantificá-los com maior consistência, inclusive em poses extremas [29]. Desta forma, o presente trabalho propõe realizar o alinhamento 3d em ambientes sem restrições, utilizando como informação precedente classificação de orientação da cabeça proposta em Zavan *et al.* [16]. Para aumentar a precisão em localizar os pontos faciais, a informação de orientação foi combinada com regressão 2d em cascata de Xiong e De la Torre [60] e regressão com vetores de suporte (SVR).

Este trabalho está estruturado da seguinte forma: No capítulo 2 são introduzidos os trabalhos relacionados ao rastreamento e alinhamento; O capítulo 3 relata o método de rastreamento desenvolvido e experimentos realizados; No capítulo 4 é descrito em detalhes o alinhamento 3d e apresenta os experimentos realizados; No capítulo 5 são expostas as conclusões em função dos resultados alcançados e sugestões para trabalhos futuros.

CAPÍTULO 2

TRABALHOS RELACIONADOS

Neste capítulo são descritas as principais abordagens relacionadas aos temas desenvolvidos neste trabalho: o rastreamento visual genérico que, através da informação temporal, consiste em localizar a região de interesse em vídeos constituídos de cenários não controlados; o alinhamento 2d e 3d, que representa encontrar com precisão pontos fiduciais em faces de ambientes não controlados.

2.1 Rastreamento

Os algoritmos de rastreamento visual são representados em duas categorias, generativo e discriminativo. O primeiro utiliza modelos generativos para encontrar candidatos prováveis nos *frames* consecutivos, e a maioria das abordagens são baseadas na análise dos componentes principais (PCA) [41] ou representações esparsas [5, 30, 37, 71, 76]. O segundo é um método de aprendizado de máquina com intuito de treinar classificadores binários para distinguir o alvo e a região de fundo, e são baseados principalmente em Haar [4, 20, 21, 23, 47, 54, 70] e filtros de correlação [7, 14, 24, 26, 69], porém apresentam baixa performance em ambientes não controlados,e.g. variação de iluminação, deformação e oclusão parcial.

Recentemente, abordagens discriminativas baseadas em redes neurais convolucionais (CNN) têm ganho atenção com resultados estado-da-arte em diversas pesquisas na visão computacional, tais como classificação de imagens [33], reconhecimento de objetos [17], detecção e segmentação [18].

Neste contexto, Wang *et al.* [56] realiza o rastreamento de objetos utilizando uma CNN originalmente pré-treinada em distinguir a região de interesse, tendo como informação de saída um mapa pixel-a-pixel indicando a probabilidade de cada pixel pertencer ao alvo nos *frames* analisados. Hong *et al.* [25] utiliza uma CNN pré-treinada em classificação de

objetos, cuja saída é avaliada por um descritor de características através de máquina de vetores de suporte (SVM).

Os métodos baseados em CNN relacionados acima exploram somente informações obtidas na última camada da CNN. De acordo com Ma *et al.* [36] tal informação é insuficiente para capturar detalhes mais precisos como a mudança de aparência do alvo no decorrer do vídeo. Em contraste, a extração de informações de múltiplas camadas convolucionais têm se mostrado mais favorável à obtenção de bons resultados em rastreamento visual [36,55].

Nam e Han [38] propõe a MDNet (*Multi-Domain Network*), uma CNN que treina um conjunto de vídeos de objetos diversos com respectivas anotações, alcançando resultados estado-da-arte nos desafios *Visual Object Tracking* [32] e *Object Tracking Benchmark* [58].

A CNN do rastreamento MDNet [38] consiste em duas partes: as camadas compartilhadas, representadas por três camadas convolucionais e duas camadas totalmente conectadas; e uma camada totalmente conectada adicional denominada camada de domínio específico que possui K ramificações, onde K é representado pela quantidade de vídeos de treinamento, de forma que cada K ramificação faz uma classificação binária de região de interesse e fundo.

Esta divisão permite que as camadas compartilhadas obtenham uma representação genérica de todos os vídeos, de forma que a informação na última camada (camada de domínio específico) não é utilizada na etapa de testes. Desta forma, durante a etapa de testes, o método MDNet [38] extrai regiões em torno da área de interesse manualmente anotada no *frame* inicial do vídeo e realiza um treinamento online, possibilitando que o rastreamento seja adaptável ao vídeo de teste, independente da região de interesse.

Uma das aplicações do rastreamento é a obtenção da região da face em vídeos, porém em ambientes não controlados não é uma tarefa trivial, devido a grande variabilidade de características existentes.

Outro fator que prejudica o rastreamento de face é a informação adquirida no *frame* inicial, o qual delimita a região de interesse a ser estimada nos *frames* seguintes, que pode conter baixa iluminação, oclusão ou orientação distante do frontal. Neste contexto, Zavan [67] detecta a região da face através da Faster-RCNN de Ren *et al.* [40] e quantifica a

qualidade desta com base em parâmetros definidos por Abaza *et al.* [1]. Adicionalmente, utiliza a região do nariz para classificar a pose através de máquina de vetores de suporte (SVM), possibilitando estimar o melhor *frame* para iniciar o rastreamento, o que proporcionou aumento de precisão em 5% em relação ao rastreamento iniciado no primeiro *frame*.

2.2 Alinhamento de faces

Existem diferentes configurações de pontos fiduciais faciais adotadas para o alinhamento 2d. Sun *et al.* [48] utiliza cinco pontos esparsos, nestes casos os pontos fiduciais têm como finalidade rotacionar a imagem através de transformação afim, porém apenas cinco pontos são insuficientes para tarefas como reconstrução 3d e análise de expressões faciais.

Em Sagonas *et al.* [43] é proposto uma base de dados com imagens 2d e 68 pontos fiduciais manualmente anotados, tal como demonstrado na imagem 2.1, possibilitando extrair mais informações da face. Estas bases de dados consistem de imagens em ambiente sem restrições, apresentando variações de expressões faciais espontâneas, diferentes condições de iluminação e resolução da imagem, permitindo comparar o desempenho dos métodos de alinhamento facial existentes em imagens extraídas do cenário real.

Figura 2.1: Exemplo da base de dados 300W [43] com os 68 pontos fiduciais 2d fornecidos. Os pontos foram conectados para facilitar a visualização da geometria da face.

Shen *et al.* [45] disponibiliza uma base de dados com vídeos e o mesmo padrão de 68 pontos fiduciais anotados, através do qual Xiao *et al.* [59] e Yang *et al.* [65] utilizam informação temporal para localizar os pontos fiduciais.

O desempenho dos métodos de alinhamento facial são comparados, principalmente, através do cálculo de erro de precisão da distância entre os pontos fiduciais estimados pelo alinhamento e a localização dos pontos manualmente anotados, normalizados pela distância intra-ocular, de forma que, quanto mais próximo de zero seja o resultado, mais preciso é o alinhamento.

Embora o alinhamento facial tenha alcançado resultados com grande precisão, tal como demonstrado por Yang *et al* [65], Xiao *et al.* [59] e Zhu *et al.* [78] nas principais bases de dados em alinhamento 2d [43, 45], estas bases contém imagens cujo sujeito apresenta a orientação da cabeça próxima ao frontal, sem a existência de imagens com faces em poses extremas (ex.: faces em perfil), tal como exemplificado na figura 2.2.

Imagens com grandes variações de orientação apresentam maior dificuldade para o alinhamento, uma vez que a geometria da face apresentada em imagens frontais difere das imagens com poses extremas. Desta forma, o alinhamento facial em grandes variações de orientação é um desafio ainda não solucionado.

Figura 2.2: Exemplos da base de dados 300W [43], sem grandes variações de pose do sujeito.

A base de dados AFLW [31](*Annotated Facial Landmarks in the Wild*) contém imagens faciais em diferentes ângulos, porém, com apenas 21 pontos anotados e em regiões visíveis da face, ou seja, a depender do ângulo do sujeito são demarcados os pontos fiduciais. Tal configuração é suscetível à falsa ideia de bom resultado de alinhamento, uma vez que somente pontos correspondentes à parte visível do rosto é computada, tal como exemplificado por Zhu *et al.* [78] na imagem:

Figura 2.3: Alinhamento incorreto na base AFLW [31]. Em cada sujeito, a primeira imagem demonstra os pontos fiduciais existentes na base de dados, na segunda imagem de cada sujeito são demonstrados 68 pontos fiduciais detectados. Fonte: Zhu *et al.* [78]

O alinhamento 3d tem ganho atenção recentemente [8, 16, 19, 34, 75], principalmente a partir da base de dados *3d Face Alignment in the Wild* (3DFAW), [29], que consiste em imagens 2d e seus respectivos 66 pontos fiduciais em três dimensões.

A base 3DFAW [29] contém imagens com grande variabilidade de características, tais como expressões faciais, iluminação, resolução da imagem e pose, aumentando o grau de dificuldade na tarefa de alinhamento.

Para localizar os pontos fiduciais na face, as abordagens de alinhamento são classificadas por duas principais vertentes, generativos e discriminativos:

Classificadores generativos aprendem um modelo de probabilidade conjunta, $p(x, y)$, onde x são os valores de entrada e y são as classes. Desta forma, métodos generativos tratam o alinhamento como um problema de otimização em encontrar parâmetros ideais correspondentes à aparência e geometria da face, sendo subdivididos em duas principais categorias, representações holísticas e modelos deformáveis por partes.

O método de modelos de aparência ativos (*Active Appearance Models - AAM*) proposto por Cootes *et al.* [11] é o trabalho seminal na abordagem holística, o qual gera modelos estatísticos lineares da geometria e aparência da face, seguindo as principais etapas: a geometria da face é normalizada através de transformação de similaridade e análise dos componentes principais (PCA); o resultado do alinhamento é obtido nas imagens de teste através da busca pelos parâmetros gerados no treinamento que melhor se ajustem ao modelo de aparência sintetizado no treinamento.

Em [2,50,51] são propostas extensões do método de alinhamento AAM [11], aplicando-o em ambientes não controlados [51], utilizando diferentes representações de imagem [2,50] ou diferentes estratégias de ajustes [50, 51] do modelo com a face.

As abordagens de alinhamento holísticas dependem de dois fatores: o poder de representação do modelo treinado e a dificuldade em otimizar tal modelo no cenário de teste. Uma vez que os modelos são gerados para toda a região da face, as abordagens holísticas não são ideais em cenários sem restrição, tais como em imagens com oclusões parciais na face, dada a grande variabilidade neste cenário. Desta forma, podem ser empregados os modelos generativos baseados em partes da face.

Os modelos por partes são métodos generativos baseados em partes faciais, podendo ser a partir de aparência individual para cada parte facial, tal como *Active Shape Models* [12], o qual combina modelos de partes faciais com modelo de distribuição de pontos (PDM), ou métodos baseados em modelos generativos para todas as partes faciais simultaneamente, tal como exemplo o trabalho proposto por Tzimiropoulos e Pantic [52], o qual, usando PCA, combina um modelo linear estatístico das partes faciais concatenadas com a forma da face, superando resultados baseados em AAM em cenários não controlados.

Já os métodos de alinhamento baseados em classificadores discriminativos têm como princípio básico o mapeamento dos valores de entrada x para as classes y em $p(y|x)$ diretamente, aprendendo um conjunto funções discriminativas, possibilitando ser empregados tanto o aprendizado local, de forma que cada ponto facial é inferido independentemente, quanto um modelo global, o qual estima a localização de todos os pontos faciais simultaneamente.

Nas abordagens de aprendizado local têm-se como principais algoritmos de alinhamento as seguintes classes: Modelos de restrição local (*Constrained Local Models*), que emprega detecções locais para cada ponto facial, regularizado posteriormente por um modelo global [3, 13]; restrição de regressão local, o qual aplica regressão local independente em cada ponto facial em conjunto com grafos para restringir o espaço de busca [53].

Já no segundo caso, onde as abordagens de alinhamento cujos pontos faciais são estimados simultaneamente (toda a geometria da face), os algoritmos são baseados em:

florestas de regressão baseado em voto [10,15,64], redes neurais convolucionais [48,73,74] e regressão em cascata [44,60,61], sendo esta última umas das abordagens mais utilizadas no alinhamento facial, devido a obtenção de resultados estado-da-arte e baixo custo de processamento.

Na regressão em cascata é adotada uma geometria genérica da face como entrada, normalmente a face média extraída do conjunto de treinamento. A cada estágio do algoritmo de cascata é realizada a regressão dos pontos fiduciais, refinando a localização dos pontos a cada iteração, tal como exemplo o método de descida supervisionado SDM [60], proposto por Xiong e De la Torre.

O método de aprendizado SDM [60] visa a aprender séries de regressões lineares em cascata, aproximando o valor resultante (coordenadas dos pontos faciais) à localização desejada na imagem, para tanto, extrai regiões ao redor de cada ponto fiducial e aplica o descritor de características SIFT (*Scale Invariant Feature Transform*) [35]. Dada uma imagem $\mathbf{d} \in \mathbb{R}^{m \times 1}$ com m pixels e x_* pontos faciais manualmente anotados, o treinamento consiste em minimizar a seguinte formulação sobre Δx:

$$f(x_o + \Delta x) = \|h(d(x_0 + \Delta x)) - \theta_*\|_2^2 \tag{2.1}$$

onde x_o são os pontos faciais da face média, extraídos a partir conjunto de treinamento, $h(d(x))$ é a função de extração de características (SIFT), e $\theta_* = h(d(x))$ representa os valores SIFT extraídos nas regiões anotadas.

Diferentes variações de regressão em cascata foram propostas a partir do SDM [60]: Xiong e De La Torre [61] estendem o método SDM, dividindo o espaço de busca em regiões de direções de gradiente similares; Em [62,77] são utilizadas várias geometrias faciais ao invés de somente a face média para inicialização, aumentando a precisão do alinhamento em diferentes poses.

Yang *et al.* [63] descreve a importância da utilização da informação da pose do sujeito para o alinhamento, através da qual é possível utilizar um modelo de inicialização específico para cada pose, aumentando a precisão em localizar pontos fiduciais. No entanto, Yang *et al.* [63] não utiliza bases de dados com imagens em poses faciais extremas.

Embora métodos de alinhamento tenham alcançado significativos resultados em existentes bases de dados de imagens 2d, conforme descrito anteriormente, tais bases de dados por sua vez não apresentam imagens com grandes variações à pose do sujeito [43] ou não possuem uma avaliação criteriosa em tais condições [31].

Para contornar tais limitações, a utilização da informação de profundidade da face na tarefa de alinhamento facial em imagens 2d [8, 16, 19, 34, 75] permite inferir a localização de pontos fiduciais em poses distantes do frontal, possibilitando avaliar com maior confiabilidade o desempenho dos métodos de alinhamento em imagens com poses extremas, tal como descrito por Jeni et al. [29].

Em [8, 75] o alinhamento 3d é subdividido em duas etapas através de regressão com redes neurais convolucionais. Gou *et al.* [19] aproxima a localização dos pontos fiduciais 2d mediante regressão da geometria da face, enquanto Li *et al.* [34] propõem um algoritmo de força bruta para classificação em 2d, e ambos recuperam a informação de profundidade (3d) utilizando modelos deformáveis 3d.

O método proposto por Zavan *et al.* [16] é uma abordagem de alinhamento de pontos fiduciais faciais 3d alternativa, que utiliza como principal informação a pose extraída a partir da região do nariz, desprezando as características faciais específicas da face durante o alinhamento. Sua formulação consiste nas seguintes etapas, tal como demonstrado na figura 2.4:

1. Detecção do nariz através de um método estado-da-arte [40];

2. Classificação da orientação do sujeito a partir da região do nariz através de redes neurais convolucionais (CNN);

3. Encaixe de uma face neutra 3d com a informação de pose, ajustando-se conforme escala e translação pela região detectada do nariz.

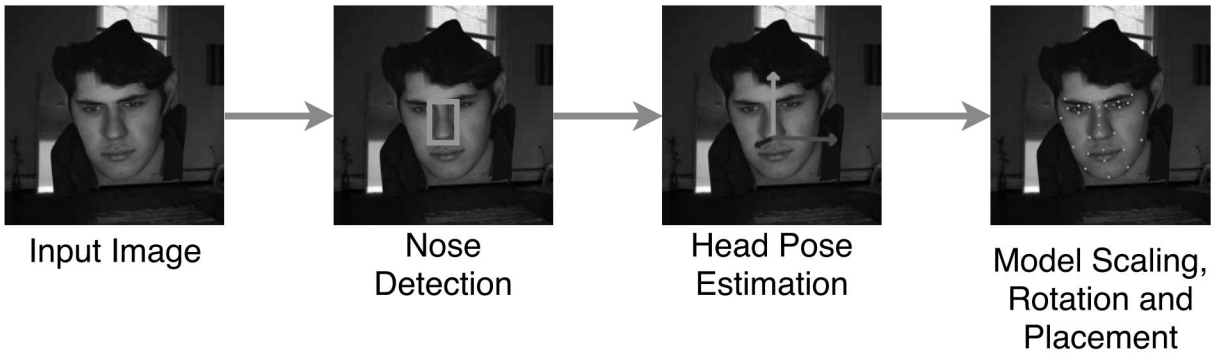


Figura 2.4: Principais etapas do alinhamento de pontos fiduciais 3d. Fonte: Zavan *et al.* [16]

São classificadas variações de orientação da face do sujeito em torno de dois eixos de movimentação: O eixo vertical (*yaw*) e eixo lateral (*pitch*). O resultado final do alinhamento proposto por Zavan *et al.* [16] é, para todas as imagens, o mesmo modelo de face com expressão neutra, rotacionado e ajustado conforme a pose estimada, possibilitando ser expandido futuramente através da utilização de um refinamento dos pontos fiduciais.

CAPÍTULO 3

RASTREAMENTO DO NARIZ EM AMBIENTES NÃO CONTROLADOS

O presente trabalho visa a contribuir para o rastreamento em ambientes sem restrições através da utilização da região do nariz como alvo para aumentar a confiabilidade do resultado, uma vez que o nariz é, em relação à face e outros componentes, menos suscetível à variações de expressões faciais e oclusão. Para tanto, foi realizado a combinação da escolha automática de melhor *frame* de por Zavan [67] com o rastreamento visual estado-da-arte MDNet de Nam e Han [38].

A análise de qualidade de *frames* de Zavan [67] possui as seguintes etapas: A região da face é inicialmente detectada através do método estado-da-arte Faster-RCNN de Ren *et al.* [40] e a qualidade da mesma é estimada através da média geométrica aferida pelo contraste, brilho, foco, nitidez e iluminação. Dentro da face é detectado o nariz por intermédio da Faster-RCNN [40], possibilitando inferir a pose com classificador SVM (máquinas de vetores de suporte).

Os *frames* cuja a região da face ou nariz não foram detectados são desconsiderados pela avaliação de qualidade e classificação de pose. O método de rastreamento do nariz é representado no algoritmo 1 e diagrama da figura 3.1, possibilitando identificar a integração da análise de qualidade [67] com o rastreamento [38].

Após escolha de melhor *frame* de inicialização, o vídeo é desmembrado em duas partes e o rastreamento é realizado, para ambos os casos, a partir do *frame* de melhor qualidade. O resultado é posteriormente reordenado para a sequência de *frames* original.

Figura 3.1: Diagrama de qualidade da imagem e rastreamento do nariz. Linhas vermelhas e azuis são as áreas detectadas e resultado do rastreamento, respectivamente.

Algorithm 1 Rastreamento do nariz. A função *rastreia* representa o método de rastreamento visual MDNet [38]. A função *melhorframe* representa a escolha de melhor *frame* [67].

função RASTREAMENTONARIZ(*frames*)
  $CarregaArquiteturaMDNet()$
  $CarregaModelo()$
  $melhorNariz, nariz \leftarrow melhorframe(frames)$
  $InicializaRastreamento(frames[melhorNariz])$
  $resultado[melhorNariz] \leftarrow nariz$
  para i ← melhorNariz-1 **até** 0 **faça**
    $resultado[i] \leftarrow rastreia(frames[i], resultado[i+1])$
  fim para
  para i ← melhorNariz+1 **até** tam(*frames*) **faça**
    $resultado[i] \leftarrow rastreia(frames[i], resultado[i-1])$
  fim para
  devolve *resultado*
fim função

3.1 Bases de dados

Para avaliação de desempenho do rastreamento é necessário utilizar vídeos com anotação da respectiva região de interesse. Kristan *et al.* [32] e Wu *et al.* [58] disponibilizam bases de dados com diferentes tipos de objetos para avaliação de rastreamento, sendo em sua minoria faces. Wolf *et al.* [57] utiliza vídeos retirados da internet contendo faces em ambientes sem restrições, no entanto, não apresentam grandes variações de translação e

poses, facilitando a tarefa de rastreamento.

A base de dados 300 *Videos in the Wild* (300VW) de Shen *et al.* [45] dispõe de vídeos de faces em ambientes não controlados, categorizados em três níveis de dificuldade, além de 68 pontos fiduciais, possibilitando extrair a região do nariz e da face para avaliação.

Beveridge *et al.* [6] propõem a base de dados *Point and Shoot Challenge* (PaSC), a qual consiste de imagens e vídeos sem restrições com alto grau de dificuldade, no entanto, não contém anotação da região da face e nariz para avaliação.

Desta forma, foram selecionadas para avaliação de desempenho do rastreamento do nariz em relação ao rastreamento de face as bases de dados com maior variação de características, 300VW [45] e PaSC [6], sendo que nesta última foram anotados manualmente a região da face e nariz em 100 vídeos escolhidos aleatoriamente.

3.1.1 300 Videos in the Wild - 300VW

A base de dados 300VW [45] foi desenvolvida com intuito de avaliar o desempenho de algoritmos de rastreamento da face, dispondo de 114 vídeos com aproximadamente um minuto de duração cada, totalizando 218.597 *frames*.

Não foi possível comparar os resultados obtidos no rastreamento do nariz com abordagens baseadas em pontos fiduciais descritas 300VW [45] devido a indisponibilidade de tais métodos para realizar os testes e pela diferente métrica de avaliação relatada em [45], que desconsideraram alguns *frames* do conjunto de testes, como exemplo a exclusão de faces em perfil.

A base 300VW [45] está subdividida em 50 vídeos para treinamento e 64 vídeos para testes, sendo este agrupado por nível crescente de dificuldade, a seguir:

a) Categoria um: 31 vídeos em boa qualidade, apresentando poucas variações de pose, expressões faciais e oclusões parciais;

b) Categoria dois: 19 vídeos de dificuldade intermediária, apresentando variações de iluminação, expressões faciais, pose e poucas oclusões;

c) Categoria três: 14 vídeos em ambientes completamente sem restrições, apresentando maior incidência de oclusão, mudança de iluminação, grande variação de pose e expressões

faciais.

Os 50 vídeos do conjunto de treinamento apresentam características variadas conforme as três categorias apresentadas no conjunto de testes.

3.1.1.1 Experimentos 300VW

Para comparação com o desempenho do rastreamento do nariz, o rastreamento da face empregado é o rastreamento visual estado-da-arte MDNet [38], utilizando o treinamento original disponibilizado pelo autor. Para todos os vídeos de testes (64 vídeos) o método de rastreamento da face foi inicializado a partir da região da face manualmente anotada no primeiro *frame* de cada vídeo.

Visando maior acurácia do rastreamento do nariz, foi utilizado a região do nariz presente nos 50 vídeos do conjunto de treinamento disponíveis na base de dados 300VW [45] durante o treino.

No conjunto de testes foram realizados dois experimentos com o rastreamento do nariz:

Inicialmente foi utilizada a região do nariz detectada automaticamente para iniciar o rastreamento, partindo do melhor *frame* de vídeo, denominado nos experimentos como *Nariz detectado*.

Em segundo momento, o rastreamento do nariz foi realizado mediante a região do nariz anotada manualmente após escolha de melhor *frame*, possibilitando comparar de forma justa com o rastreamento da face. Desta forma foi evitado rastrear a região incorreta em casos onde a detecção do nariz falha após avaliação de qualidade do melhor *frame*, tal como demonstrado na imagem 3.2. Dentre os 64 vídeos de teste da base 300VW [45], 5 apresentaram detecção incorreta. O resultado do rastreamento do nariz a partir da anotação manual é descrito nos experimentos como *Nariz anotado*.

Figura 3.2: Detecções incorretas do nariz (em verde) durante a etapa de escolha de melhor *frame* de inicialização ocasionam no rastreamento da região errada (região esperada em azul).

O desempenho do rastreamento visual é avaliado *frame* a *frame*, utilizando duas métricas, o coeficiente de interseção [27], denominado também como taxa de sucesso [32, 58] e a precisão [32], que identifica a taxa de acerto em relação à distância da região estimada e a respectiva região anotada (*ground-truth*). (Algoritmo 2)

Algorithm 2 Métricas de avaliação de precisão e coeficiente de interseção. A região estimada é representada por pred e o *ground-truth* é representado por gt

função COEFICIENTE($pred, gt$)
  $intersecao \leftarrow calculaIntersecao(pred, gt)$
  $iArea \leftarrow intersecao.largura * intersecao.altura$
  $pArea \leftarrow pred.largura * pred.altura$
  $gArea \leftarrow gt.largura * gt.altura$
  devolve $min(iArea/pArea, iArea/gArea)$
fim função
função PRECISAO($pred, gt$)
  devolve $l2norm(centro(pred), centro(gt))$
fim função

A seguir são demonstrados os resultados obtidos no conjunto de testes da base 300VW [45] para todos os 64 vídeos existentes neste conjunto de vídeos. Em segundo momento, são demonstrados resultados agrupados conforme a categoria de dificuldade.

Em consideração ao quantitativo total do conjunto de vídeos de testes (64 vídeos), os resultados obtidos demonstram que o rastreamento do nariz apresenta grande precisão em estimar a localização da região do nariz, conforme figura 3.3, alcançando precisão de traslação em 90,61% quando iniciado a partir da detecção automática do nariz no melhor *frame* e 97,67% de precisão quando o rastreamento é iniciado a partir do anotação manual da região do nariz, também iniciado o rastreamento no melhor *frame*.

Já o rastreamento da face alcançou a precisão de 96.68%. Para todos os casos foi levado em consideração o acerto com o limiar de 20 pixels de distância, conforme adotado pela avaliação de rastreamento visual em [32].

Em uma avaliação de precisão mais restrita, reduzindo o erro para o limiar de 10 pixels de distância, o rastreamento do nariz alcança a taxa de acerto de 82.30% e 92.09%, iniciando-o pela detecção automática e anotação manual, respectivamente. Nesta margem de erro o rastreamento da região da face obtém acerto de 76.20%, comprovando melhor desempenho do rastreamento do nariz em relação à precisão em localizar a região esperada.

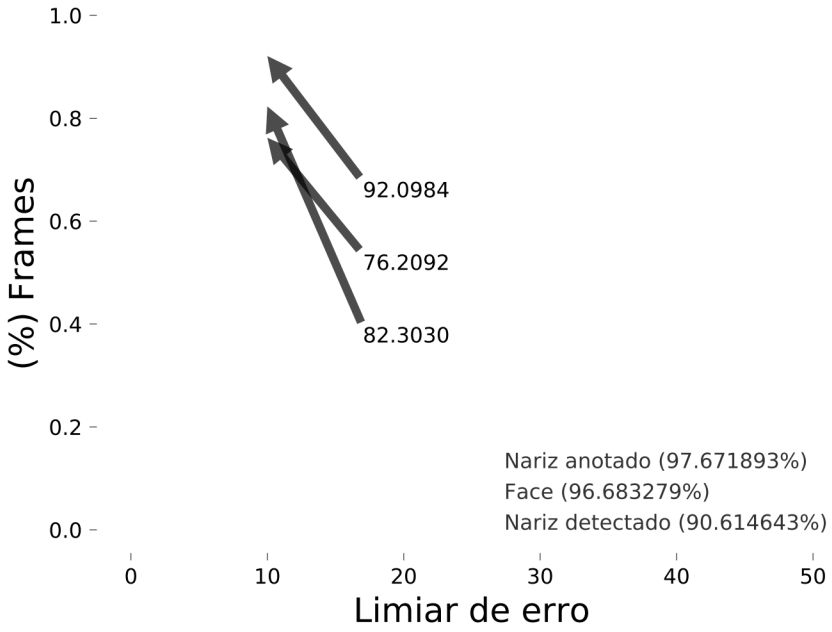


Figura 3.3: Comparação da métrica de precisão para todos os vídeos de teste da base 300VW [45]. Entre parênteses porcentagem de *frames* cujo limiar de acerto é de 20 pixels de distância. As setas demarcam o limiar de 10 pixels.

Embora demonstrado que o rastreamento pelo nariz seja eficiente, tal região não obtém o resultado desejado quando comparado à face na avaliação do coeficiente de interseção, cálculo este que indica que a sobreposição da região estimada em relação ao *ground-truth*, para todos os *frames* avaliados no conjunto de testes da base 300VW [45], tanto na avaliação conjunta de 64 vídeos (figura 3.4), quanto na avaliação de cada categoria separada por nível de dificuldade (figuras 3.8, 3.12(b) e 3.16(b)).

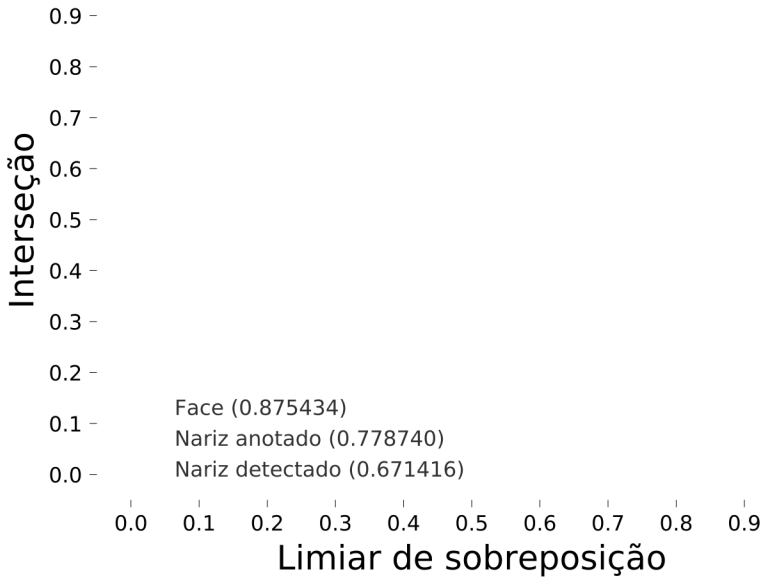


Figura 3.4: Comparação do coeficiente de interseção para todos os vídeos de teste da base 300VW [45]. A área sob a curva está disposta entre parênteses.

A análise visual nos resultados obtidos indica que tal evento é ocasionado devido os seguintes fatores: O resultado do rastreamento do nariz é ligeiramente maior do que sua região esperada (*ground-truth*), devido a dificuldade em separar com precisão a região do nariz da face (figura 3.5). Já no rastreamento da face, a região de fundo a ser destacada do alvo de rastreamento apresenta notável diferença visual, o que favorece o controle de escala em se manter dentro da região de interesse.

Figura 3.5: Exemplos de resultados do rastreamento do nariz com área maior que a região esperada. Rastreamento a partir da detecção em azul, rastreamento a partir da anotação manual em verde, *ground-truth* em vermelho.

Outro aspecto negativo é a variação de pose do indivíduo em mais de 90 graus, fazendo com que o nariz esteja totalmente ocluso. Desta forma, o rastreamento do nariz não consegue prever a localização correta, ocasionando na perda do rastreamento. Mesmo que esta oclusão também atrapalhe no rastreamento da face, este apresenta maior probabilidade de acerto, uma vez que sua predição engloba a região da cabeça (computando como resultado positivo). A figura 3.6 demonstra tal evento, no qual o rastreamento da região do nariz foi comprometido.

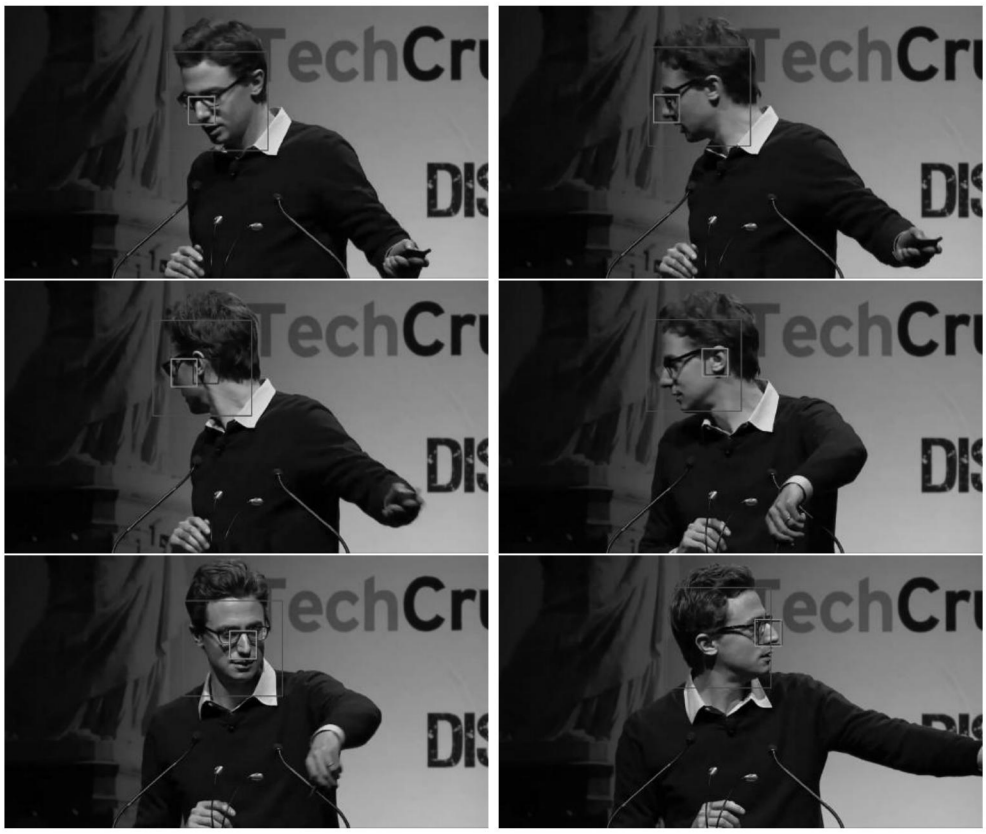


Figura 3.6: Exemplos de resultados de rastreamentos: nariz detectado em azul, nariz anotado em verde, rastreamento da face em vermelho.

Os resultados a seguir destacam o desempenho independente de cada uma das três categorias existentes na bases de dados 300VW [45]:

Na sequência de 31 vídeos da categoria um, o rastreamento do nariz iniciado pela região manualmente anotada e rastreamento da face apresentaram resultados semelhantes de precisão (figura 3.7), devido este subconjunto de vídeos apresentar boa qualidade e pouca variação de pose, tornado um conjunto menos desafiador para a realização do rastreamento por ambas abordagens.

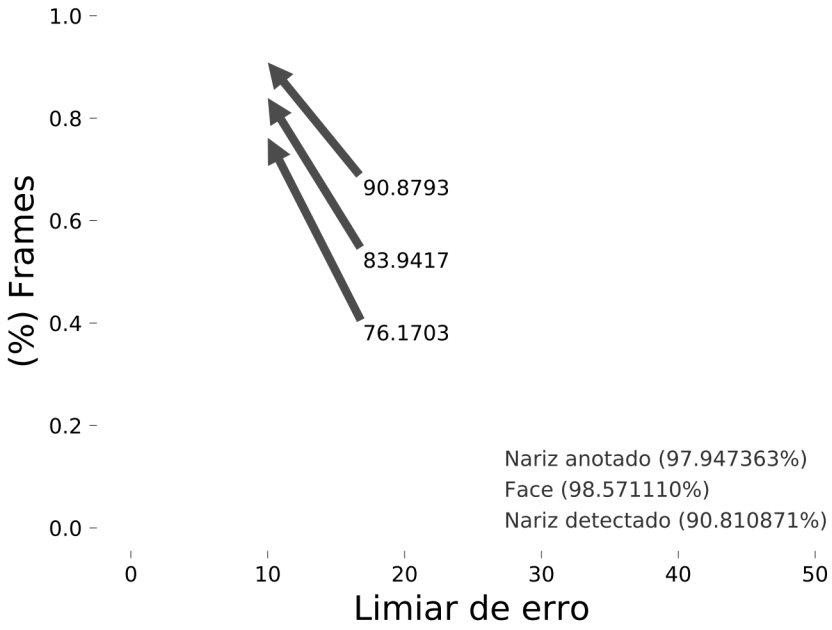


Figura 3.7: Gráfico de precisão para os vídeos de teste da categoria um. Entre parênteses porcentagem de *frames* cujo limiar de acerto é de 20 pixels de distância. As setas demarcam o limiar de 10 pixels.

Considerando os dados computados nos vídeos da categoria um, o nariz obtém uma precisão maior em relação à face se levada em consideração uma margem de erro minimizada (10 pixels de distância), de forma que a face apresenta 76.17% de acerto, o nariz com inicialização manualmente anotada alcança 90.87% e o rastreamento iniciado pela detecção automática do nariz atinge a taxa de 83.94% de precisão.

Em relação ao coeficiente de interseção, a face obtém valor superior aos testes de rastreamento do nariz, devido a maior complexidade deste em restringir com precisão o tamanho da região de interesse a ser localizada, conforme destacado o gráfico de interseção na figura 3.8. As figuras 3.9, 3.10, 3.11 demonstram alguns exemplos resultantes dos métodos avaliados nos vídeos da categoria um.

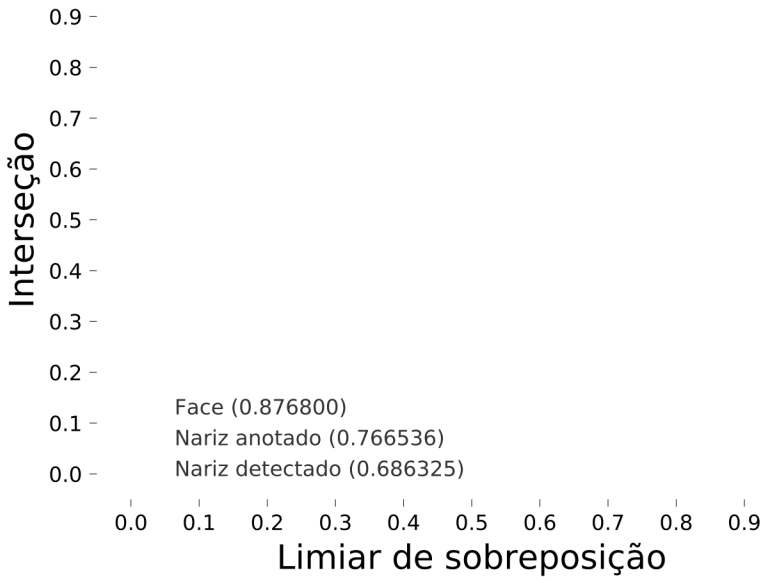


Figura 3.8: Coeficiente de interseção para os vídeos de teste da categoria um. A área sob a curva está disposta entre parênteses.

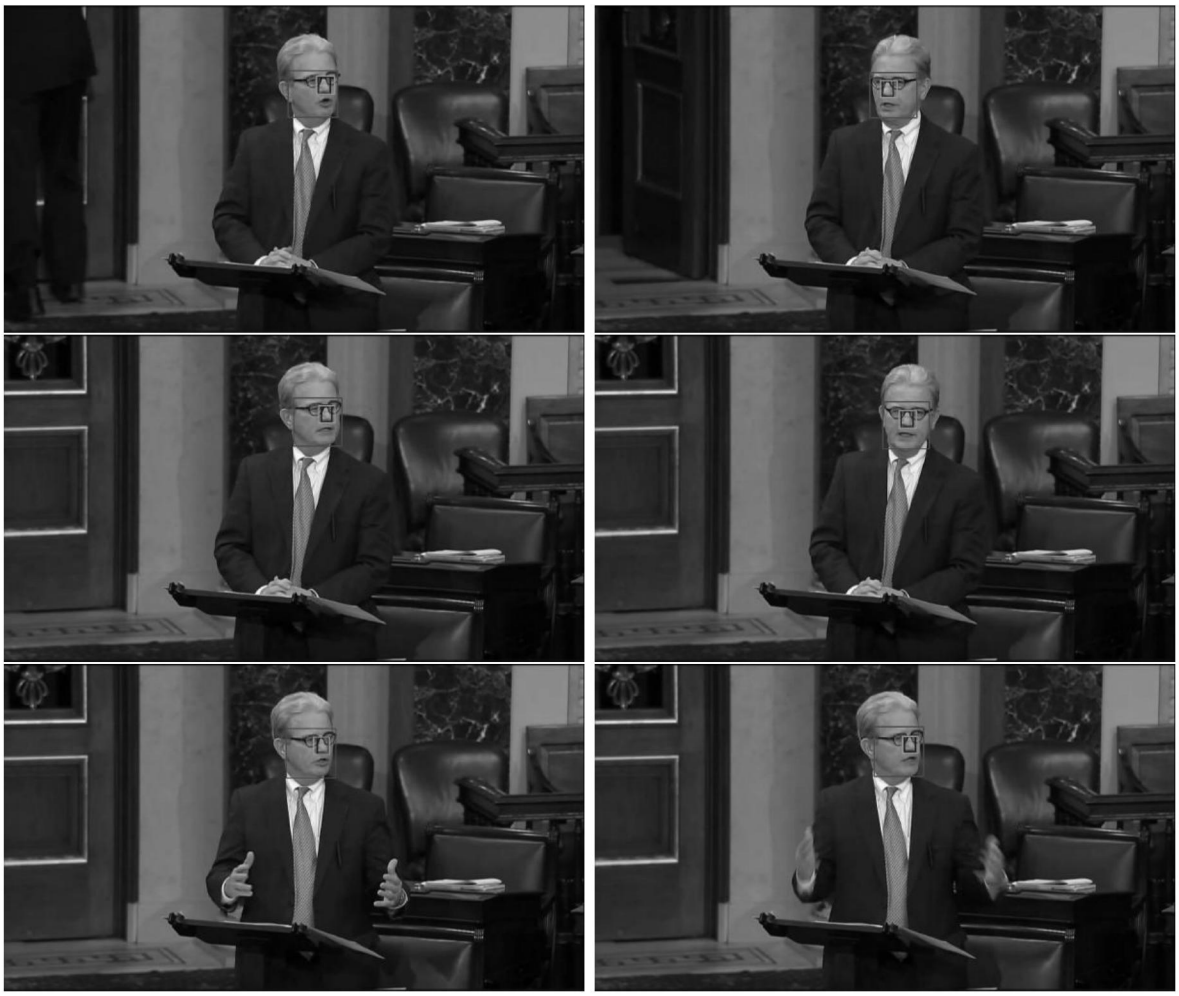


Figura 3.9: Exemplo de resultados obtidos na categoria um. Os vídeos nesta categoria apresentam boa qualidade e poucas variações de pose. Rastreamentos de nariz detectado em azul, nariz anotado em verde, face em vermelho.

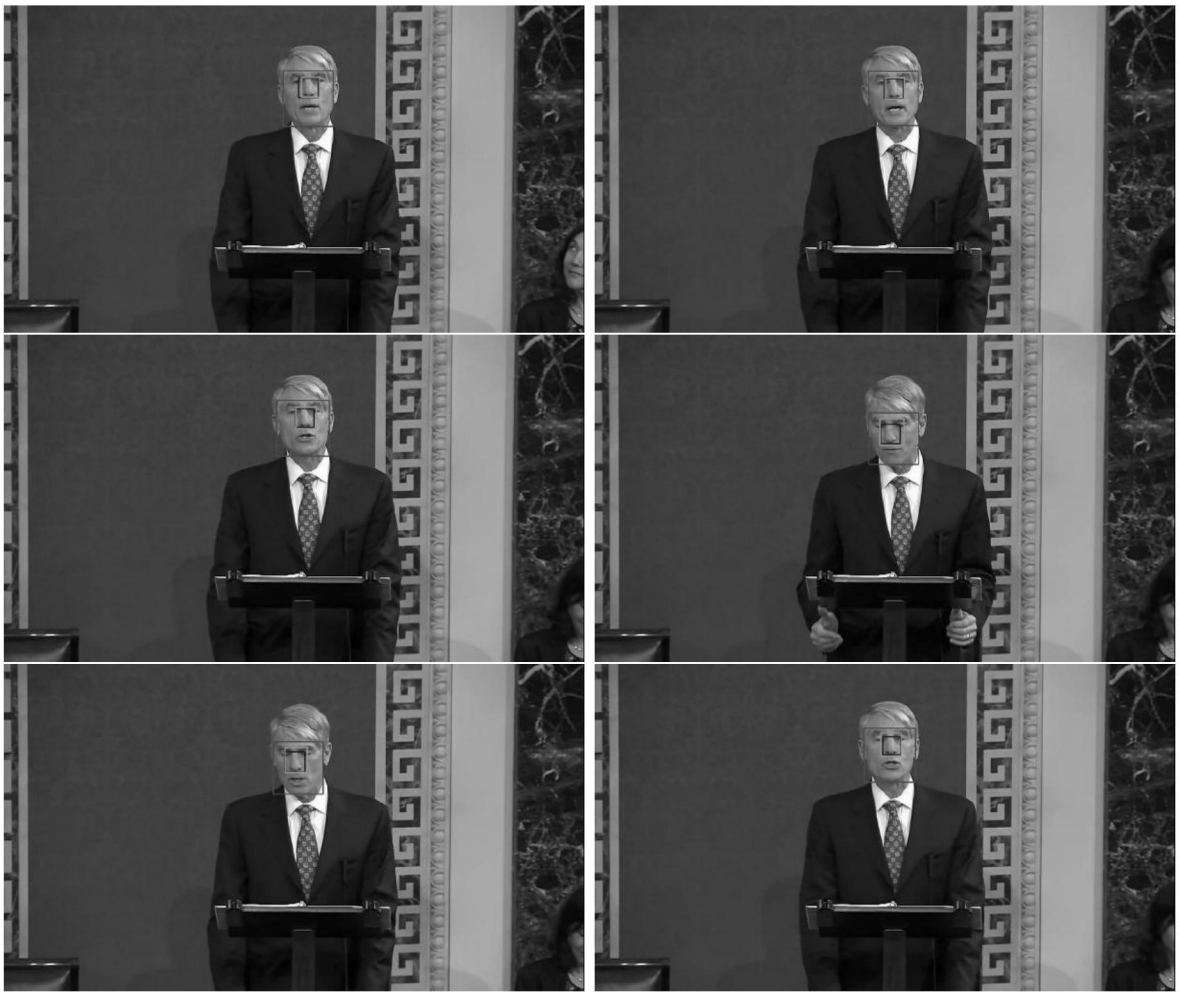
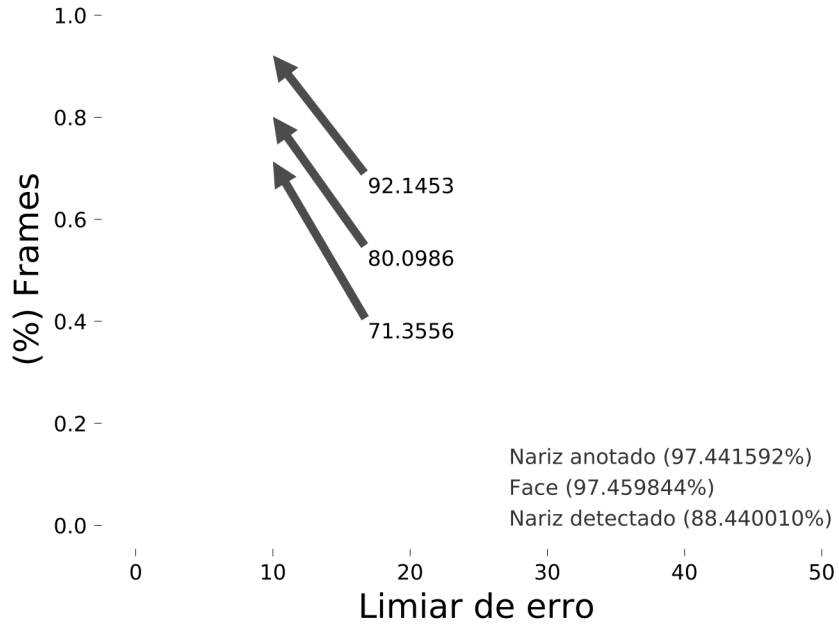


Figura 3.10: Exemplo de resultados obtidos na categoria um. Os vídeos nesta categoria apresentam boa qualidade e poucas variações de pose. Rastreamentos de nariz detectado em azul, nariz anotado em verde, face em vermelho.

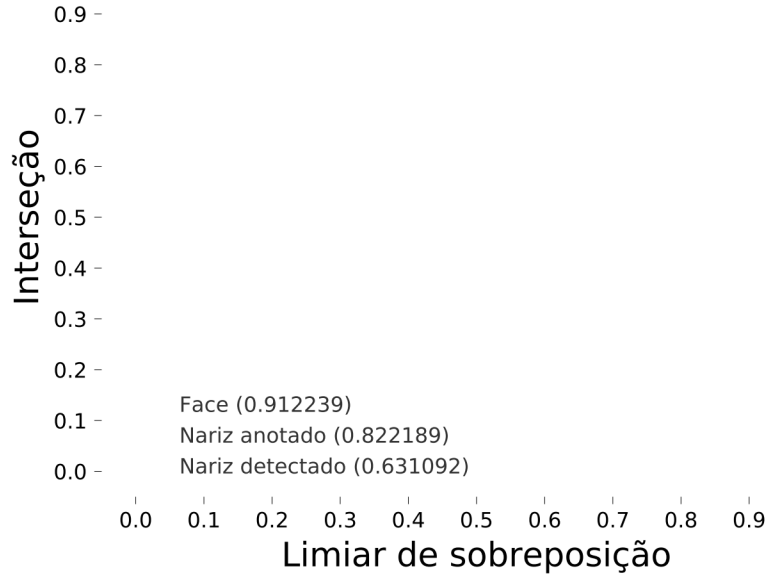
Figura 3.11: Exemplo de resultados obtidos na categoria um. Os vídeos nesta categoria apresentam boa qualidade e poucas variações de pose. Rastreamentos de nariz detectado em azul, nariz anotado em verde, face em vermelho.

O conjunto de vídeos da categoria dois dispõe de 19 vídeos, onde existem algumas variações como pose e iluminação, tornando-a mais desafiadora em relação ao conjunto anterior. Os resultados de precisão indicam confiabilidade idêntica no rastreamento do nariz (partindo da anotação manual) (97.441%) em relação à face (97.459%) (imagem 3.12(a)).

Ainda na categoria dois, se considerando o acerto restrito de 10 pixels de distância, o melhor desempenho é do rastreamento do nariz iniciado pelo anotação manual do melhor *frame*, com precisão em 92.14%, contra o rastreamento da face, o qual atinge a precisão de 80.09%. O rastreamento do nariz a partir da detecção automática alcança 71.35% de precisão.



(a) Gráfico de precisão. Entre parênteses porcentagem de *frames* cujo limiar de acerto é de 20 pixels de distância. As setas demarcam o limiar de 10 pixels



(b) Coeficiente de interseção. A área sob a curva está disposta entre parênteses

Figura 3.12: Comparação das métricas de precisão e coeficiente de interseção para os vídeos de teste da categoria dois.

O resultado do rastreamento do nariz a partir da detecção automática para o coeficiente de interseção foi amplamente inferior aos demais para a categoria dois. Tal evento ocorreu devido detecções erradas da região correspondente ao nariz no *frame* de inicialização (ex: imagem 3.2). Exemplos do resultado obtido nesta categoria são demonstrados nas figuras 3.13, 3.14, 3.15.

Figura 3.13: Exemplo de resultados obtidos na categoria dois. Rastreamentos de nariz detectado em azul, nariz anotado em verde, face em vermelho.

Figura 3.14: Exemplo de resultados obtidos na categoria dois. Rastreamentos de nariz detectado em azul, nariz anotado em verde, face em vermelho.

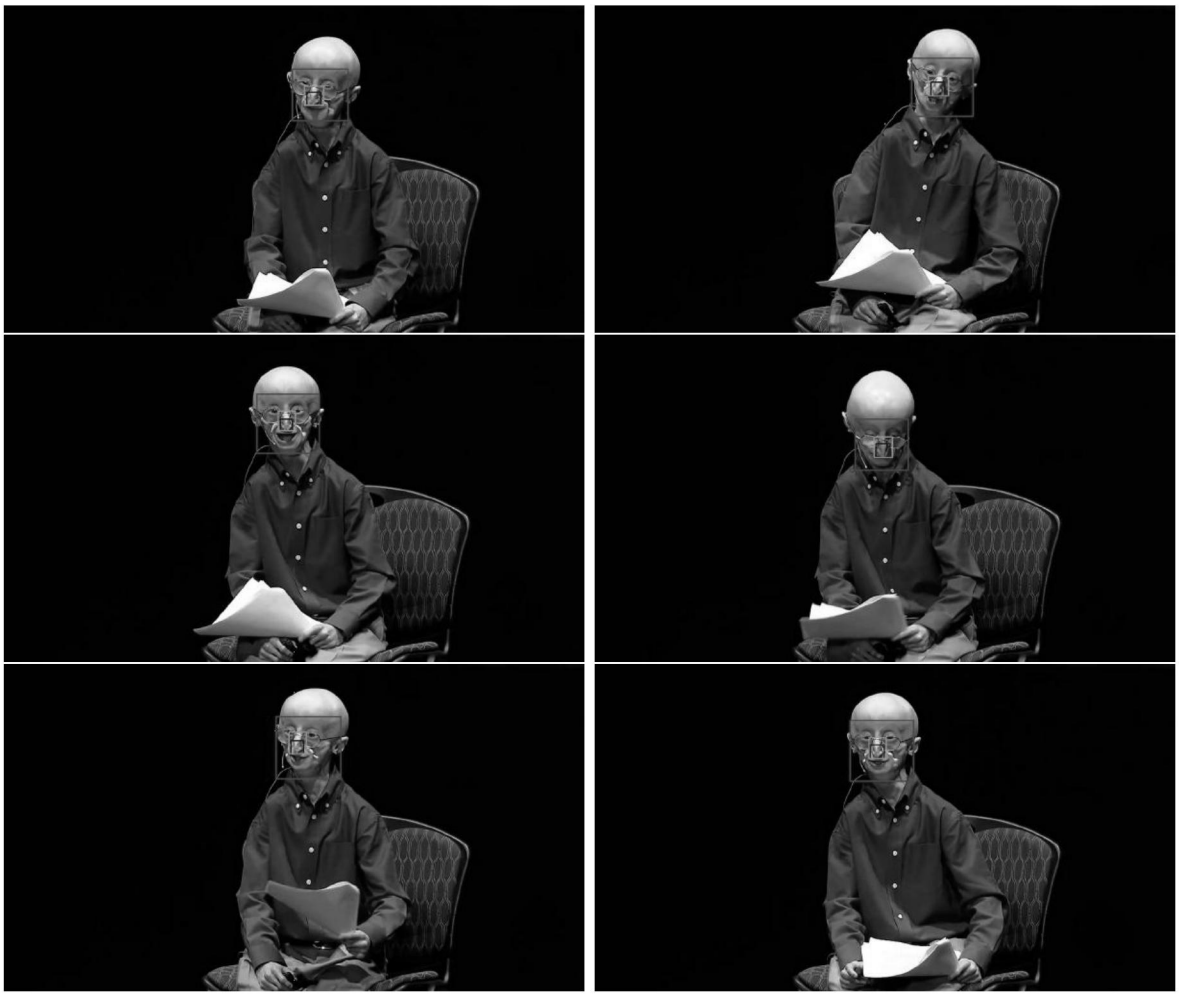
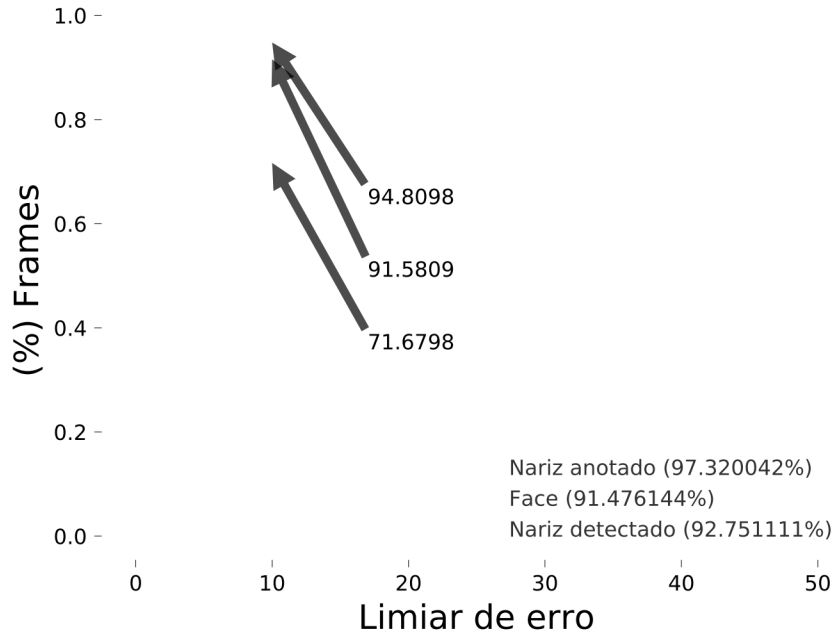


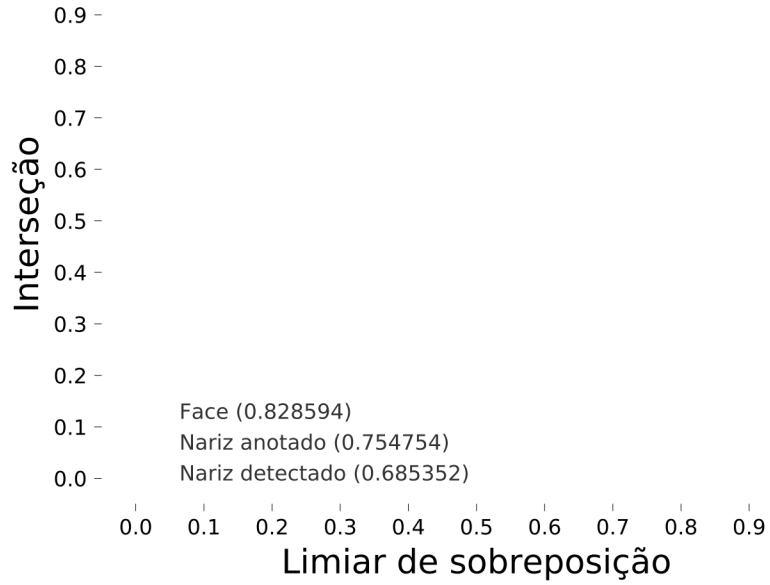
Figura 3.15: Exemplo de resultados obtidos na categoria dois. Rastreamentos de nariz detectado em azul, nariz anotado em verde, face em vermelho.

A categoria três do conjunto de testes apresenta vídeos em cenários sem restrições, possuindo com maior incidência variações de iluminação, pose, oclusão, expressões faciais, tornando este o subconjunto mais desafiador.

Os resultados de precisão alcançados nesta categoria indicam que a abordagem proposta de rastreamento pela região do nariz é superior em relação ao rastreamento da face neste cenário, atingindo a precisão em 92.75% nos vídeos onde o rastreamento do nariz é iniciado pela detecção automática e precisão de 97,32% para o rastreamento do nariz iniciando com a região manualmente anotada, ambos os casos superiores ao rastreamento da face, o qual alcança 91.47% de precisão. A figura 3.16 demonstra a precisão e coeficiente de interseção computados para os vídeos da categoria três da base de dados 300VW [45].



(a) Gráfico de precisão. Entre parênteses porcentagem de *frames* cujo limiar de acerto é de 20 pixels de distância. As setas demarcam o limiar de 10 pixels



(b) Coeficiente de interseção. A área sob a curva está disposta entre parênteses

Figura 3.16: Comparação das métricas de precisão e coeficiente de interseção para os vídeos de teste da categoria dois.

Em consideração à uma margem de acerto em 10 pixels de precisão, o rastreamento do nariz alcança uma taxa de 91.58%, sendo notavelmente superior à face (71.67%), mesmo que utilizando a região detectada automaticamente para inicialização do rastreamento. O rastreamento do nariz a partir da região manualmente anotada alcança a taxa de 94.80% de precisão.

As figuras 3.17, 3.18, 3.19 demonstram *frames* de vídeos com resultados obtidos no subconjunto de vídeos pertencentes à categoria três, permitindo a análise visual dos resultados em vídeos sem restrições, como oclusão parcial e variações de pose.

Figura 3.17: Exemplo onde a região do nariz está oclusa. Rastreamentos de nariz detectado em azul, nariz anotado em verde, face em vermelho.

Figura 3.18: Exemplo com variação de pose e inconsistência do rastreamento da face. Rastreamentos de nariz detectado em azul, nariz anotado em verde, face em vermelho.

Figura 3.19: Exemplo com variação de pose. Rastreamentos de nariz detectado em azul, nariz anotado em verde, face em vermelho.

Com base nos resultados apresentados é possível identificar que o rastreamento do nariz é eficiente para localizar o sujeito em vídeos que apresentam cenários sem restrições, apresentando confiabilidade comparável ao rastreamento facial.

No que se refere ao coeficiente de interseção, o rastreamento da face é superior ao método desenvolvido em todos os cenários, principalmente devido a região da face ser mais discriminante em relação ao fundo, de forma que, embora preciso, o rastreamento do nariz não é superior ao rastreamento da face.

3.1.2 Point and Shoot Challenge - PaSC

A base de dados PaSC [6] contém 9376 imagens e 2802 vídeos, desenvolvida para avaliar métodos de reconhecimento facial. A base PaSC [6] apresenta diferentes resoluções de imagem, além de variações de pose, expressões faciais, escala e variação de iluminação, tornando-a mais desafiadora também para o rastreamento.

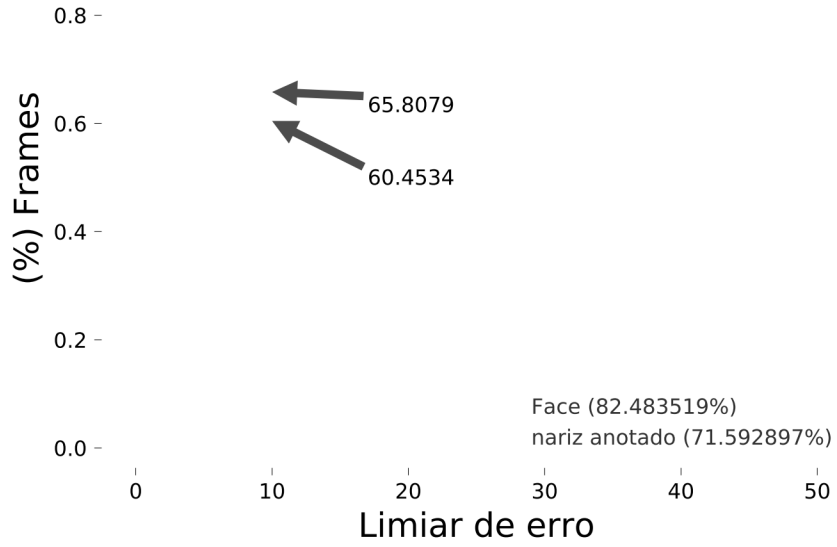
Devido a base PaSC [6] não dispor de anotação manual da região da face para avaliar o desempenho do rastreamento, foram selecionados de forma aleatória 100 vídeos e anotados manualmente, para cada *frame*, a região do nariz e da face do sujeito principal, totalizando 14.866 *frames* anotados.

3.1.2.1 Experimentos - PaSC

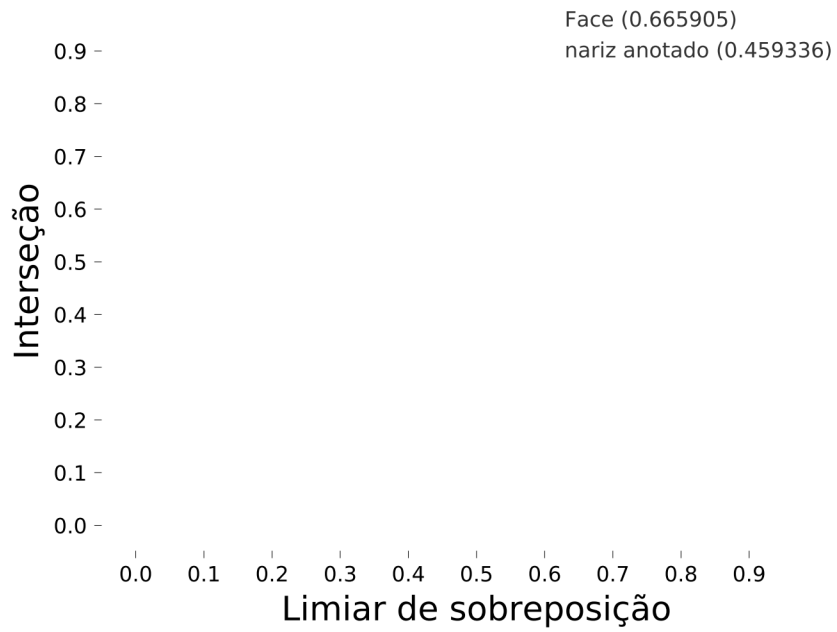
A avaliação nos 100 vídeos da base PaSC [6] foi realizada pelo método de rastreamento do nariz combinado com *frame* de melhor qualidade e região do nariz manualmente anotada, utilizando o treinamento realizado com o subconjunto de treino da base 300VW [45]. Os resultados obtidos foram comparados ao rastreamento da região da face, realizada com o método de rastreamento visual estado-da-arte [38] a partir do primeiro *frame* com a face manualmente anotada.

O rastreamento da face na base de dados PaSC [6] apresentou resultado amplamente superior ao nariz, devido a baixa resolução dos *frames* e, principalmente, em situações onde existem grandes variações de escala, de forma a reduzir drasticamente a região do nariz, prejudicando seu rastreamento.

Embora a precisão aferida para o rastreamento do nariz seja superior ao rastreamento da face, levando-se em consideração o limiar de 10 pixels como margem de erro (precisão de 65,30% para o nariz e 60,45% para a face), tal proporção de acerto não é mantida no decorrer da avaliação de precisão dos *frames*. No que se refere ao limiar de 20 pixels de distância, o rastreamento da região da face é superior ao nariz, atingindo a precisão de 82,48% e 71,59%, respectivamente, conforme demonstrado na imagem 3.20(a).



(a) Gráfico de precisão. Entre parênteses porcentagem de *frames* cujo limiar de acerto é de 20 pixels de distância. As setas demarcam o limiar de 10 pixels



(b) Coeficiente de interseção. A área sob a curva está disposta entre parênteses

Figura 3.20: Comparação das métricas de precisão e coeficiente de interseção para 100 vídeos da base de dados PaSC [6].

Na figura 3.20(b) é demonstrado o resultado obtido pelo rastreamento do nariz e face mediante a avaliação do coeficiente de interseção, de forma que o desempenho de rastreamento do nariz é inferior à face, fator este ocasionado principalmente por variações existentes de escala do sujeito presente no vídeo e baixa resolução de imagem, demonstrando maior consistência do rastreamento da face neste cenário.

Nas imagens a seguir são demonstrados *frames* de vídeos da base de dados PaSC [6] com resultados obtidos pelo rastreamento da face e da região do nariz, sendo possível identificar as situações onde o rastreamento do nariz falha, tal como exemplo a imagem 3.21, no qual o rastreamento do nariz perde precisão em escala. Este fenômeno foi ocasionado principalmente pela variação de tamanho do alvo e baixa resolução do vídeo, comprometendo o resultado final do rastreamento do nariz, enquanto o rastreamento da face apresenta resultados mais coerentes com o esperado.

A imagem 3.22 demonstra falha do rastreamento do nariz (em azul) em vídeo com baixa resolução e variação de escala. Já o rastreamento da face (em vermelho), consegue localizar o sujeito em movimento, mesmo que perdendo precisão em estimar a escala da face.

Já em situações onde o vídeo de teste apresenta *frames* com boa resolução, o resultado do rastreamento do nariz atinge resultados visivelmente melhores em relação à face, que por sua vez apresentou dificuldade em englobar a região de interesse com precisão, tal como demonstrado nas imagens 3.23 e 3.24.

Com base nos resultados obtidos na base de dados PaSC [6], é possível identificar que, nas situações onde o sujeito presente no vídeo apresenta variações de escala e o vídeo apresenta baixa qualidade, o rastreamento do nariz é comprometido devido a dificuldade em extrair com precisão a região do nariz em relação a face, sendo então preferível a utilização de rastreamento da face.

Figura 3.21: Exemplo de *frames* de vídeo com baixa qualidade e variação de escala. Rastreamento do nariz em azul e da face em vermelho.

Figura 3.22: Exemplo de vídeo com baixa qualidade e variação de escala. Rastreamento do nariz em azul e da face em vermelho.

Figura 3.23: Exemplo de vídeo de melhor qualidade apresentando variação de escala e pose. Rastreamento do nariz em azul e da face em vermelho.

Figura 3.24: Exemplo de vídeo de melhor qualidade apresentando variação de escala e pose. Rastreamento do nariz em azul e da face em vermelho.

CAPÍTULO 4

ALINHAMENTO 3D EM AMBIENTES NÃO CONTROLADOS

Visto que o alinhamento 3d possibilita inferir a posição dos pontos fiduciais em faces com maior precisão para imagens de faces em poses extremas [29], e que a informação da pose pode auxiliar como etapa inicial no alinhamento [16,63], o presente trabalho tem como objetivo contribuir para o alinhamento de imagens de faces 2d em poses extremas, propondo uma abordagem alternativa de alinhamento de pontos fiduciais 3d em imagens de faces 2d a partir da informação da pose adquirida pela região do nariz.

Figura 4.1: Etapas do alinhamento desenvolvido: refinamento 2d dos pontos fiduciais previamente estimados pela pose e obtenção da informação 3d com base no 2d detectado.

O alinhamento de pontos fiduciais 3d em imagens faciais 2d desenvolvido contém duas principais etapas: o refinamento 2d a partir de pontos fiduciais previamente posicionados, conforme classificação de orientação do nariz através de Zavan *et al.* [16]; regressão do eixo z, que consiste em estimar a profundidade de cada um dos pontos 2d, conforme exemplificado na imagem 4.1.

4.1 Refinamento de pontos faciais 2D (x,y)

Para o refinamento de pontos 2d foi utilizado a regressão em cascata de Xiong e De la Torre [60], substituindo o uso da face média frontal para a face neutra rotacionada conforme a pose estimada pelo nariz, aumentando a precisão do alinhamento 2d.

Durante o treinamento do refinamento, a região da face é aumentada em 8% em relação à anotação inicial disponível, em seguida a face é recortada e redimensionada para o valor 250x250 pixels, normalizando desta forma, variações de escala e translação existentes nas imagens de treinamento. Para cada imagem de face do conjunto de treinamento, a face neutra dos pontos fiduciais 3d é encaixada mediante translação, escala e rotação, através da informação de pose 3d. Desta forma, o algoritmo de regressão em cascata aprende a minimizar o erro entre a face neutra rotacionada e os pontos manualmente anotados (*ground-truth*), simulando a regressão a ser aplicada durante o teste, dado que neste foi utilizado uma inicialização baseado na pose, evitando a face média frontal para todos as imagens, conforme descrito em [60].

Na figura 4.2 é possível visualizar a diferença entre utilizar pontos faciais da face média frontal e a face neutra encaixada mediante rotação 3d na face do sujeito, no qual o segundo é mais próximo do resultado desejado e converge com maior precisão em imagens onde a pose do sujeito apresenta grandes variações. O treinamento do refinamento é então realizado seguindo a formulação descrita na equação 2.1.

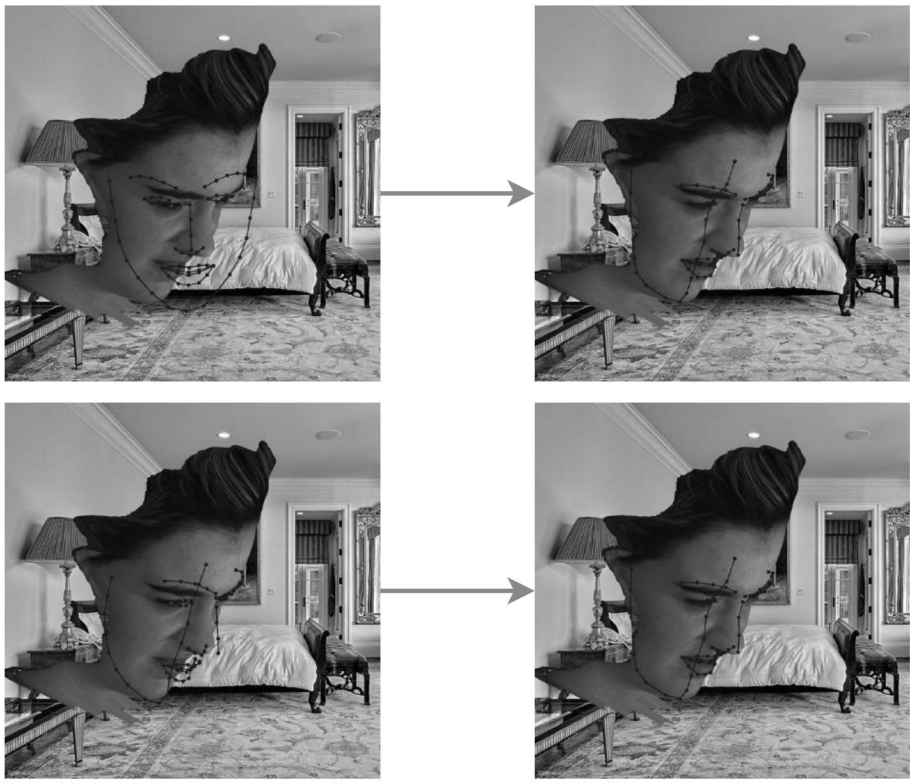


Figura 4.2: Diferença entre os pontos fiduciais da face média frontal e face neutra encaixada por rotação, na primeira e segunda linha. As imagens da direita representam os pontos manualmente anotados.

4.2 Regressão eixo Z

O alinhamento do eixo z, correspondente à profundidade da face, foi obtido através da aplicação de regressão com vetores de suporte, estimando de forma independente cada ponto facial n para i imagens de treinamento. Para tanto, foi necessário normalizar os valores das coordenadas x, y e z previamente, tornando-os invariantes à escala e translação. Dada as coordenadas de pontos 2d (x, y) estimados, o valor de seu respectivo z é dado por:

$$\hat{z}_{ni} = f(\hat{x}_{ni}, \hat{y}_{ni}), \tag{4.1}$$

onde $\hat{x}, \hat{y}$ e $\hat{z}$ representam os valores normalizados de x, y e z, respectivamente, conforme aplicado inicialmente por Zhao *et al.* [75]:

$$\hat{x}_{ni} = \frac{x_i - \overline{x_i}}{(\sigma(x_i) + \sigma(y_i))/2}, \tag{4.2}$$

$$\hat{y}_{ni} = \frac{y_i - \overline{y_i}}{(\sigma(x_i) + \sigma(y_i))/2}, \tag{4.3}$$

$$\hat{z}_{ni} = \frac{z_i - \overline{z_i}}{(\sigma(x_i) + \sigma(y_i))/2}. \tag{4.4}$$

Os valores $\sigma(x_i)$ e $\sigma(y_i)$ consistem no desvio padrão, enquanto $\overline{x_i}, \overline{y_i}$ e $\overline{z_i}$ representam a média dos valores de x_i, y_i e z_i.

Uma vez normalizados, os valores das coordenadas 2d (x, y) de cada ponto fiducial são associados ao respectivo z e são realizados n treinamentos de regressão com vetores de suporte (SVR), correspondendo a um treinamento independente para cada ponto fiducial. Na etapa de testes, para cada ponto fiducial, a predição dos valores de z são realizadas simultaneamente para todas as imagens.

O método de alinhamento foi desenvolvido na linguagem de programação Python, com o auxílio da biblioteca Scikit-learn [39]. No algoritmo 3 é representado o pseudo-código do método de alinhamento de pontos fiduciais 3d em imagens de face 2d:

Algorithm 3 Estimando pontos fiduciais 3d em imagens faciais 2d. As funções *Estima_2d, refinamento_SDM e SVR* representam o alinhamento inicial pela pose ([16]), o refinamento 2d (4.1) e o alinhamento para a profundidade (4.2).

```
função REFINAMENTO_3D(imagens)
    Carrega_modelos_SVR()
    para i ← 1 até tamanho(imagens) faça
        lms ← Estima_2d(imagens{i})
        xy[i] ← refinamento_SDM(lms, imagens{i})
    fim para
    para n ← 1 até tamanho(lms) faça
        z[n] ← SVR(xy, modelo_SVR[n])
    fim para
    lms_3d ← {xy, z}
    devolve lms_3d
fim função
```

4.3 Bases de dados

A base de dados *3D Face Alignment in the Wild* (3DFAW) [29] foi desenvolvida com intuito de avaliar o desempenho de algoritmos de alinhamento facial 3d a partir de imagens 2d, e é composta por 23606 imagens 2d em cenários controlados e ambientes sem restrição, apresentando grande variação de orientação, diferentes condições de iluminação, resolução da imagem e expressões faciais, sendo então utilizada para avaliação do alinhamento 3d.

Tulyakov e Sebe [49] geram uma base de dados com 68 pontos fiduciais 3d a partir da sintetização de faces 3d em diferentes vistas extraídas da base controlada BU-4DFE [66]. No entanto, esta base não foi utilizada para avaliação devido a semelhança com um subconjunto de imagens existentes na base 3DFAW [29].

4.3.1 Base de dados 3DFAW

A base 3DFAW [29] consiste na combinação de quatro bases de dados de imagens de faces com características diferentes: BU-4DFE [66], BP4D-Spontaneous [72], MultiPIE [22] e quadros de vídeos extraídos da internet, e seus respectivos pontos fiduciais, conforme exemplificado nas figuras 4.3, 4.4 e 4.5 (pontos da face conectados em linhas azuis para melhor visualização).

As bases BU-4DFE [66] e BP4D-Spontaneous [72] são bases de dados com imagens de faces 3d que foram renderizadas em sete poses diferentes a cada 15 graus, apresentando variações de expressões faciais. Foram utilizadas 9.555 imagens da base BU-4DFE e 6.510 imagens da base BP4D-Spontaneous faciais, respectivamente. A base de dados MultiPIE [22] é uma base controlada e com grande variações de pose, nesta base foram coletadas 7000 imagens cujos pontos faciais não estavam oclusos. As imagens retiradas de vídeos na internet são completamente sem restrições, totalizando 541 imagens faciais nesta categoria.

Figura 4.3: Exemplos de imagens da base BU-4DFE [72] com variações de poses e expressões faciais.

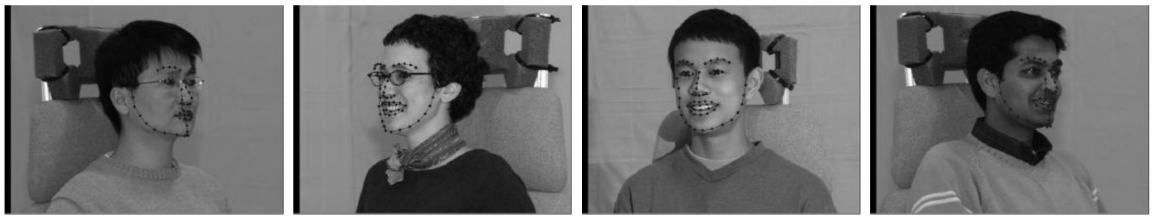


Figura 4.4: Exemplos de imagens faciais da base controlada MultiPIE [22] com diferentes poses.

Figura 4.5: Exemplos de imagens faciais de quadros de vídeos retirados na internet, apresentando diferentes condições de iluminação, baixa qualidade, oclusão parcial e expressões faciais.

A base 3DFAW [29] está subdividida em três conjuntos: 13.969 imagens para treino, 4.725 imagens para validação e 4.912 imagens para para teste, dispondo de 66 pontos faciais 3d e o retângulo correspondente à região da face para todas as imagens, inclusive no conjunto de teste, uma vez que esta base tem por finalidade avaliar somente a consistência do alinhamento 3d, desconsiderando possíveis perturbações originadas pela detecção automática da face.

A avaliação na base de dados 3DFAW [29] é realizada em duas métricas de erro, de forma que, quanto menor o valor resultante, maior precisão do método avaliado:

Ground Truth Error (GTE): Medida amplamente utilizada também na avaliação do alinhamento 2d. Afere a distância euclideana entre o resultado obtido no alinhamento e

o valor da anotação (de todos os pontos), normalizado pela distância intra-ocular:

$$E(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\frac{\|X_k - Y_k\|_2}{d_i} \tag{4.5}$$

onde N é o número de pontos faciais, X são os pontos faciais 3d estimados, Y são as anotações 3d correspondentes (*ground-truth*), k são os pontos faciais e d_i é a distância intra-ocular 3d para a i-ésima imagem.

Cross View Ground Truth Consistency Error (CVGTCE): Avalia a consistência dos pontos 3d estimados em diferentes vistas para o mesmo sujeito, como definido a seguir:

$$E_{vc}(X,Y,P) = \frac{1}{N}\sum_{k=1}^{N}\frac{\|(s\mathbf{R}X_k + t) - Y_k\|_2}{d_i}, \tag{4.6}$$

onde $P = \{\text{s},\mathbf{R},\text{t}\}$ são parâmetros de transformação rígida (escala, rotação e translação), obtidos conforme:

$$P = \{s,R,t\} = \underset{s,R,t}{\operatorname{argmin}}\sum_{k=1}^{N}\|Y_k - (sRX_k + t)\|_2^2 \tag{4.7}$$

4.3.1.1 Experimentos - 3DFAW

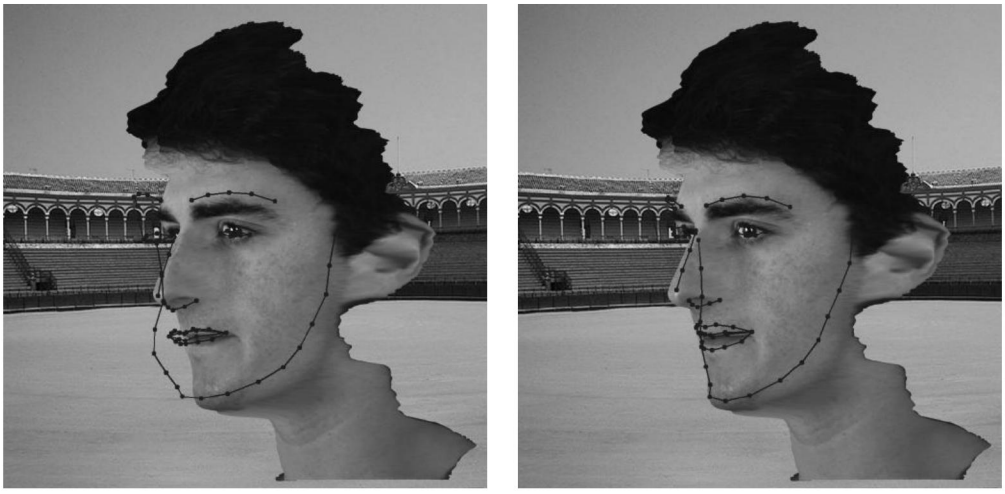
O método de alinhamento foi avaliado em dois momentos: no subconjunto de validação, possibilitando compará-lo com alinhamento tomado como base de Zavan *et al.* [16], o alinhamento 2d sem informação de pose de Xiong e de La Torre [60] e o estado-da-arte de Bulat e Tzimiropoulos [8]; posteriormente, avaliado no conjunto de testes da base de dados 3DFAW [29], confrontando com os principais métodos de alinhamento 3d [8,16,19,34,75]. Em ambos os casos, no treinamento da estimativa de pose as imagens foram classificadas de acordo com a orientação do sujeito, variando entre -60 e 60 graus, a cada 7.5 graus, em torno dos eixos vertical e lateral, conforme relatado em Zavan *et al.* [16].

Para avaliação no conjunto de validação (4725 imagens) foram utilizadas na etapa de treino as 13.969 imagens disponíveis do conjunto de treinamento. No treinamento do refinamento de pontos 2d, realizado com regressão em cascata, foram adicionadas dez amostras de cada imagem de face, mediante variação uniforme de translação (-5%, 5%),

rotação (-$\pi/4$, $\pi/4$) e escala (-10%, 10%) dos pontos fiduciais 2d providos na base de dados sobre a respectiva imagem, totalizando 139.690 imagens faciais no conjunto de treinamento do refinamento de pontos fiduciais 2d.

A inserção destas variações de translação no conjunto de treinamento permite que o algoritmo de regressão em cascata seja mais robusto à diferentes inicializações durante a etapa de testes, uma vez que a estimativa inicial dos pontos fiduciais gerados através da utilização do método utilizado [16] é impreciso no alinhamento local (pontos fiduciais independentes).

A imagem a seguir apresenta resultados obtidos no conjunto de validação com e sem adição de dados no treinamento, onde é possível identificar visualmente que o acréscimo de dados no treinamento do refinamento 2d em 4.6(b) consegue minimizar o erro de ajuste local dos pontos fiduciais 2d com maior precisão em relação ao refinamento treinado com as 13.969 imagens originais 4.6(a).



(a) 13.969 imagens de treinamento (b) 139.690 imagens de treinamento

Figura 4.6: Resultado do refinamento 2d obtido no conjunto de validação da base de dados 3DFAW [29] com e sem adição de imagens no treinamento.

Na tabela a seguir são relacionados os valores aferidos pela distância intra-ocular (GTE) referentes ao alinhamento de pontos fiduciais 2d e 3d, alcançados através do treinamento com 13.969 imagens e 139.690 imagens confirmando que, no caso onde o treinamento do refinamento 2d é realizado com acréscimo de imagens, a precisão do alinhamento é superior em 15%.

Tabela 4.1: Resultado obtido no conjunto de validação da base de dados 3DFAW [29] para os eixos 2d (x,y) e 3d (x,y,z), para treinamentos com 13.690 imagens e 139.690 imagens.

Eixos	Resultados - GTE (%)	
	Refinamento 2d com 13.969 imagens	Refinamento 2d com 139690 imagens
XY (2d)	4.274	3.526
XYZ (3d)	6.676	5.613

Já no treinamento do alinhamento em profundidade (eixo z), correspondente ao 3d, realizado através de regressão com de vetores de suporte (SVR), foram utilizadas para o treinamento somente os pontos fiduciais correspondentes às 13.969 imagens de treinamento. O acréscimo de dados para o treinamento do SVR impactou na variação de apenas 1% de precisão, sendo preferível então não utilizar 139.690 amostras durante o treino para alinhamento em profundidade, conforme descrito na tabela a seguir:

Tabela 4.2: Resultado da regressão com SVR para o eixo z em imagens do conjunto de validação da base de dados 3DFAW [29], para 13.969 e 139.690 amostras no treinamento.

Eixos	Resultados - GTE (%)	
	Regressão SVR com 13.969 amostras	Regressão SVR com 139.690 amostras
Z (1d)	3.724	3.710
XYZ (3d)	5.613	5.598

Na primeira linha da tabela 4.2 são relacionados os resultados obtidos para o eixo z, utilizando 13.969 amostras e 139.690 amostras no treinamento de regressão, sendo possível identificar que o resultado é praticamente inalterado para os dois casos. Desta forma é então preferível utilizar o treinamento com 13.969 amostras para estimar a profundidade dos pontos fiduciais.

A segunda linha relata os valores referentes ao alinhamento 3d (XYZ) nos treinamentos acima relacionados, confirmando que o impacto gerado pela adição de pontos fiduciais no treinamento do alinhamento para a profundidade não proporciona ganho considerável no resultado acumulado dos três eixos (XYZ).

Dada a disponibilidade dos pontos fiduciais 3d das faces do conjunto de validação da base de dados 3DFAW [29], foi realizado inicialmente a avaliação do desempenho individual de cada eixo (x, y e z), em seguida a avaliação em 2d (XY) e 3d (X,Y,Z), comparado os resultados obtidos com: o método de alinhamento baseado na pose utilizado

como base [16], o alinhamento 2d a partir da face média [60] (não utiliza a informação da pose) e o estado-da-arte [8], observando a métrica 4.5, conforme destacado na tabela a seguir:

Tabela 4.3: Resultados obtidos no conjunto de validação da base de dados 3DFAW [29] para os eixos 2d (x,y), 3d (x,y,z) e eixos x, y e z independentes. Em azul e vermelho estão destacados 1º e 2º melhores resultados, respectivamente. O método [60] não realiza alinhamento 3d.

	Resultados - GTE (%)				
Eixos	XY	XYZ	X	Y	Z
Bulat e Tzimiropoulos [8]	3.626	4.940	2.12	2.48	2.77
Zavan *et al.* [16]	7.787	10.442	4.97	4.94	5.75
Xiong e De La Torre [60]	4.736	-	3.37	3.36	-
Refinamento	3.526	5.613	2.19	2.28	3.72

Na segunda coluna da tabela 4.3 são relacionados os resultados para o alinhamento 2d (XY), sendo possível identificar que o método de alinhamento desenvolvido supera o estado-da-arte em 2.75% nesta categoria e é 25% mais preciso do que o alinhamento pela face média de [60], enfatizando o ganho em utilizar a informação precedente de pose para o alinhamento. Em relação ao método tomado como base [16], a precisão do alinhamento é superior em 54%.

Na terceira coluna da tabela 4.3 é descrito o resultado do alinhamento 3d (XYZ), no qual o desempenho do alinhamento 3d é superior ao método base [16] em mais de 46%, confirmando a consistência do resultado também no conjunto de pontos fiduciais 3d. Em consideração ao estado-da-arte [8], o resultado obtido é inferior devido menor precisão alcançada ao estimar a profundidade dos pontos fiduciais (correspondentes à linha Z), tal como destacado o desempenho individual do eixo z na última linha da tabela. O método de alinhamento [60] gera apenas pontos fiduciais 2d, não sendo possível o comparativo nesta categoria.

As três últimas linhas da tabela 4.3 correspondem ao erro de translação do ponto fiducial estimado em relação à localização esperada (*ground-truth)* para cada um dos eixos x,y e z, independentes, sendo possível identificar a proximidade dos valores obtidos entre o refinamento e o método estado-da-arte [8].

Na figura 4.7 é monstrado o comparativo de desempenho do refinamento e o alinhamento utilizado como base [16], enfatizando a relevância de utilizar o refinamento para alinhar os pontos fiduciais 2d.

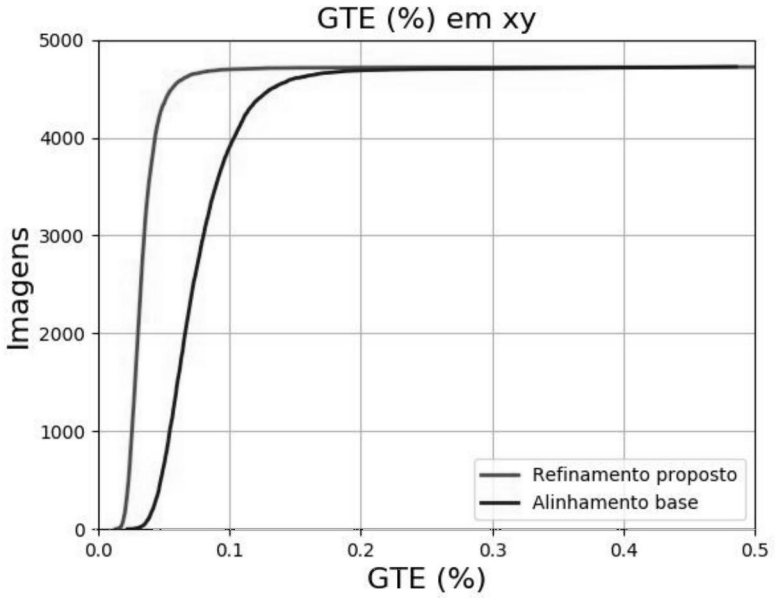


Figura 4.7: Gráfico de desempenho do alinhamento em XY para o refinamento (em vermelho) e o alinhamento base [16] (em azul).

A figura 4.8 relata o comparativo de desempenho entre o método de refinamento desenvolvido e o alinhamento utilizado como base [16] para os resultados do alinhamento de pontos fiduciais 3d, confirmando resultados mais precisos obtidos por meio do método de refinamento.

Na imagem 4.9 são monstradas imagens de faces do conjunto de validação da base 3DFAW [29], possibilitando visualizar o resultado obtido pelo alinhamento 3d em comparação ao método base [16] e ao resultado esperado (*ground-truth* fornecido na base de dados) em faces com diferentes variações de pose e expressões faciais.

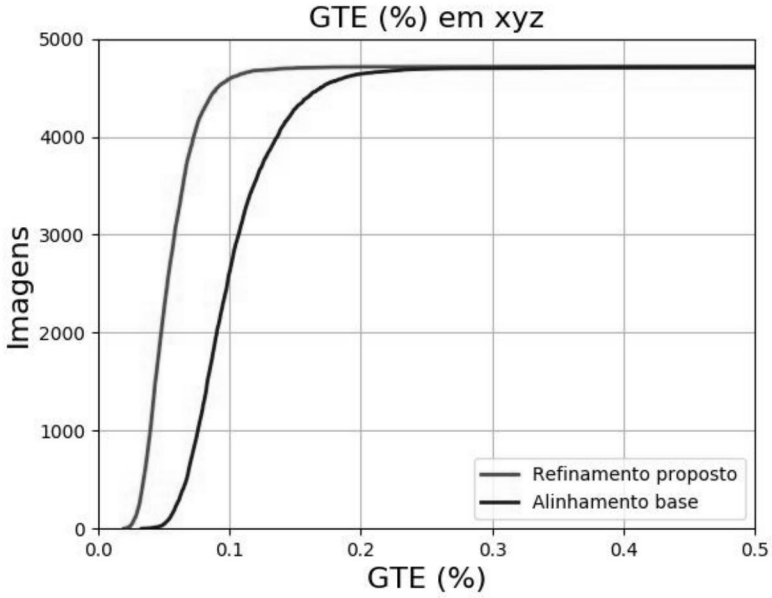


Figura 4.8: Gráfico de desempenho do alinhamento em XYZ para o refinamento (em vermelho) e o alinhamento base [16] (em azul).

Figura 4.9: Exemplo de resultados obtidos no conjunto de validação da base de dados 3DFAW [29]. Método base [16] na primeira coluna, alinhamento 3d e resultado esperado na terceira coluna.

Em segundo momento, o método de alinhamento de pontos fiduciais 3d foi avaliado no conjunto de testes da base de dados 3DFAW [29]. Para tanto, o treinamento consistiu nas imagens dos conjuntos de treino e validação, totalizando 18.694 imagens. No treinamento do refinamento de pontos 2d, realizado com regressão em cascata, foram adicionadas oito amostras de cada imagem de face mediante variações de translação(-5%, 5%), rotação ($-\pi/4$, $\pi/4$) e escala (-10%, 10%), consistindo em 149.552 imagens para treino do refinamento 2d.

Uma vez que, para o subconjunto de testes não é disponibilizado os pontos fiduciais 3d, apenas a imagem 2d e a região da face, o comparativo de desempenho (GTE e CVGTCE) deve ser realizado em uma plataforma online disponibilizada pelos autores Jeni *et al.* [29]. Desta forma foi possível aferir apenas o desempenho da combinação de x,y e z, correspondentes aos pontos 3d.

Na tabela abaixo são relacionados os resultados estados-da-arte [8, 16, 19, 34, 75] no conjunto de testes da base 3DFAW [29], em conjunto ao resultado obtido pelo método de alinhamento 3d, destacado na tabela como refinamento, comparados através das métricas GTE (equação 4.5) e CVGTCE (equação 4.6), em ordem crescente de desempenho:

Tabela 4.4: Resultados obtidos no subconjunto de testes da base de dados 3DFAW [29] para GTE e CVGTCE, em ordem crescente.

Métodos	Resultados	
	% CVGTCE	% GTE
Bulat and Tzimiropoulos [8]	3.4767	4.5623
Zhao *et al.* [75]	3.9700	5.8835
Refinamento	**4.035**	**6.317**
Li *et al.* [34]	4.891	7.589
Gou *et al.* [19]	4.9488	6.2071
Zavan *et al.* [16]	5.9093	10.8001

Conforme relacionado na tabela acima, o método de refinamento aumenta a precisão do alinhamento 3d em 31% na métrica de avaliação CVGTCE, e um ganho de 41% se considerado o GTE, ambos em relação ao método base [16] que fora utilizado como inicialização.

Em relação ao resultado descrito pelo método estado-da-arte [8], o alinhamento 3d desenvolvido possui erro superior devido dois fatores: menor precisão em estimar a profundidade (eixo z), conforme citado anteriormente em experimentos realizados no conjunto de validação; erros obtidos na etapa inicial em detectar a região do nariz ou estimar a pose para algumas imagens do conjunto de teste, haja visto que em situações onde a pose estimada é completamente incoerente à pose da imagem, o modelo de inicialização de pontos fiduciais da face média encaixado na face não consegue convergir para o esperado na etapa de refinamento, gerando um resultado de alinhamento 3d inconsistente com a face.

Na figura a seguir são monstrados resultados obtidos em imagens do conjunto de testes da base de dados 3DFAW [29], apresentando poses extremas e diferentes expressões faciais. Na primeira coluna são relacionados os resultados do método base [16] e na segunda coluna são apresentados resultados obtidos com o alinhamento 3d, sendo possível identificar visualmente maior precisão do método alinhamento 3d desenvolvido em detrimento do método base.

Figura 4.10: Exemplo de resultados do método base e refinamento, na primeira e segunda coluna, respectivamente, em poses extremas da base 3DFAW [29].

Na figura 4.11 são monstrados resultados do alinhamento 3d obtidos em imagens com faces em poses extremas e diferentes expressões faciais do conjunto de testes da base 3DFAW [29], e suas respectivas projeções 3d rotacionadas em diferentes vistas, possibilitando visualizar o modelo da face 3d gerada pelo alinhamento de pontos fiduciais.

Figura 4.11: Resultado do alinhamento obtido no conjunto de testes da base 3DFAW [29]. Para cada imagem de amostra na primeira coluna, têm-se o respectivo resultado do alinhamento 3d rotacionado em diferentes vistas, na segunda e terceira coluna. Pontos fiduciais conectados por linhas para melhor visualização.

Durante os experimentos com o conjunto de testes da base 3DFAW [29] ocorreram as seguintes falhas em imagens de faces na etapa de alinhamento inicial [16]: detecção errada da região do nariz, fazendo com que as etapas seguintes em estimar a pose e encaixar o modelo da face média neutra sobre a face ocorram em região distante do local esperado, tal como exemplificado na imagem 4.12(a); estimando pose completamente incoerente com a imagem facial de teste, conforme figuras 4.12(b) e 4.12(c).

Nestes casos onde ocorreram má inicialização os pontos fiduciais não proporcionam o

encaixe global sobre a face, de forma que o refinamento 2d da etapa seguinte não consiga convergir para o resultado esperado, impossibilitando o alinhamento.

(a) Detecção do nariz em região errada (b) Pose estimada incoerente (c) Pose estimada incoerente

Figura 4.12: Exemplo de imagens que apresentaram detecção do nariz e estimativa de ângulo incoerentes com a imagem de teste.

Em algumas situações do alinhamento 3d em imagens do conjunto de testes da base 3DFAW [29], o refinamento 2d utilizado atinge o mínimo local durante a regressão e não convergiu para o resultado esperado, principalmente no refinamento dos pontos fiduciais referentes ao contorno da face, conforme relatado na figura 4.13.

Figura 4.13: Falhas no refinamento, principalmente na região do contorno da face.

CAPÍTULO 5

CONCLUSÃO

Este trabalho abrangeu duas etapas intermediárias utilizadas na análise biométrica da face, o rastreamento facial em vídeos e o alinhamento 3d de face em imagens 2d, ambos os casos consistindo de imagens em ambientes sem restrições.

Foi proposto integrar a escolha de melhor *frame* ao rastreamento da região do nariz, confrontando com o rastreamento da região da face através do método estado-da-arte de Nam e Han [38]. Experimentos foram realizados nas bases de dados 300VW [45] e PaSC [6], compostas de vídeos sem restrições, apresentando variação de orientação, expressões faciais, oclusões parciais, variações de iluminação, escala e resolução do vídeo, dificultando a tarefa de rastreamento.

Foi demonstrado na base de dados 300VW [45] que o rastreamento do nariz alcança taxa de precisão em translação de 97,67%, similar ao rastreamento da região da face (96,68% de precisão). Em relação ao cenário mais desafiador desta base, o rastreamento do nariz supera amplamente a face (97,32% e 91.47%) em translação. Contudo, não é trivial delimitar o tamanho da região do nariz com precisão durante o rastreamento, dado a semelhança dos pixels da região de interesse com a face, fazendo com que o rastreamento do nariz perca precisão em escala.

Testes realizados em 100 vídeos manualmente anotados da base de dados PaSC [6] demonstram a dificuldade encontrada pelo rastreamento do nariz em situações em que existem variações de escala, devido a redução da região de interesse a ser rastreada. Desta forma, o rastreamento da face é mais robusto em todas as avaliações realizadas referentes ao coeficiente de interseção, sendo este então preferível para realizar a tarefa de rastreamento do sujeito em vídeos em relação ao rastreamento de nariz.

Em segundo momento, foi elaborado um método de alinhamento 3d com a finalidade de encontrar pontos fiduciais em imagens sem restrição. Para tal, foi exposto uma abordagem

que consiste em refinar com precisão a localização dos pontos fiduciais da face estimados conforme a classificação de orientação obtida pela região do nariz.

Foram realizados experimentos nos subconjuntos de validação e teste da base de dados sem restrição 3DFAW [29]. No primeiro caso, o alinhamento desenvolvido superou o método utilizado como alinhamento inicial de Zavan *et al.* [16] em mais de 54%, o alinhamento sem auxílio da pose de Xiong e De la Torre [60] em 25% e o estado-da-arte de Bulat e Tzimiropoulos [8] em 2,75%, no que se refere ao alinhamento 2d (XY).

Em relação ao 3d (XYZ), o método de alinhamento supera o método base [16] no subconjunto de validação em 46%, porém demonstra resultado inferior ao estado-da-arte [8], devido a menor precisão em estimar a profundidade dos pontos fiduciais (eixo Z).

O experimento realizado no subconjunto de testes da base 3DFAW [29] avaliou o método apresentado em comparação aos trabalhos mais relevantes no alinhamento facial 3d, alcançando resultados competitivos com o estado-da-arte [8]. Em relação ao método utilizado como base [16] foi constatado maior precisão em 41% na métrica de avaliação GTE e 31% em relação à avaliação aferida pelo CVGTCE.

Em alguns casos onde o nariz foi detectado incorretamente, ou a pose do sujeito foi estimada de forma incoerente com o esperado, ocasionaram falhas no alinhamento, não convergindo para o resultado esperado.

Em trabalhos futuros o refinamento de pontos fiduciais 3d pode ser estendido com a informação temporal, auxiliando no rastreamento de faces em ambientes sem restrições.

BIBLIOGRAFIA

[1] Ayman Abaza, Mary Ann Harrison, Thirimachos Bourlai, e Arun Ross. Design and evaluation of photometric image quality measures for effective face recognition. *IET Biometrics*, 2014. 15

[2] Epameinondas Antonakos, Joan Alabort-i Medina, Georgios Tzimiropoulos, e Stefanos Zafeiriou. Hog active appearance models. *International Conference on Image Processing (ICIP)*. IEEE, 2014. 18

[3] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, e Maja Pantic. Robust discriminative response map fitting with constrained local models. *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013. 18

[4] Boris Babenko, Ming-Hsuan Yang, e Serge Belongie. Visual tracking with online multiple instance learning. *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009. 13

[5] Chenglong Bao, Yi Wu, Haibin Ling, e Hui Ji. Real time robust l1 tracker using accelerated proximal gradient approach. *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. 13

[6] J Ross Beveridge, P Jonathon Phillips, David S Bolme, Bruce A Draper, Geof H Givens, Yui Man Lui, Mohammad Nayeem Teli, Hao Zhang, W Todd Scruggs, Kevin W Bowyer, et al. The challenge of face recognition from digital point-and-shoot cameras. *IEEE BTAS*, 2013. , 24, 43, 44, 45, 67

[7] David S Bolme, J Ross Beveridge, Bruce A Draper, e Yui Man Lui. Visual object tracking using adaptive correlation filters. *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010. 13

[8] Adrian Bulat e Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. *European Conference on Computer Vision (ECCV)*. Springer, 2016. 17, 20, 56, 59, 62, 63, 68

[9] Kyong I Chang, Kevin W Bowyer, e Patrick J Flynn. Multiple nose region matching for 3d face recognition under varying facial expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2006. 11

[10] Tim F Cootes, Mircea C Ionita, Claudia Lindner, e Patrick Sauer. Robust and accurate shape model fitting using random forest regression voting. *European Conference on Computer Vision (ECCV)*. Springer, 2012. 19

[11] Timothy F. Cootes, Gareth J. Edwards, e Christopher J. Taylor. Active appearance models. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2001. 17, 18

[12] Timothy F Cootes, Christopher J Taylor, David H Cooper, e Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 1995. 18

[13] David Cristinacce e Timothy F Cootes. Feature detection and tracking with constrained local models. *Proceedings of the British Machine Vision Conference (BMVC)*, 2006. 18

[14] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, e Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 13

[15] Matthias Dantone, Juergen Gall, Gabriele Fanelli, e Luc Van Gool. Real-time facial feature detection using conditional regression forests. *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. 19

[16] Flávio H de Bittencourt Zavan, Antônio CP Nascimento, Luan P e Silva, Olga RP Bellon, e Luciano Silva. 3d face alignment in the wild: A landmark-free, nose-based

approach. *European Conference on Computer Vision (ECCV)*. Springer, 2016. , 12, 17, 20, 21, 50, 51, 53, 56, 57, 59, 60, 61, 62, 63, 65, 68

[17] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, e Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *International Conference on Machine Learning (ICML)*, 2014. 13

[18] Ross Girshick, Jeff Donahue, Trevor Darrell, e Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 13

[19] Chao Gou, Yue Wu, Fei-Yue Wang, e Qiang Ji. Shape augmented regression for 3d face alignment. *European Conference on Computer Vision (ECCV)*. Springer, 2016. 17, 20, 56, 62, 63

[20] Helmut Grabner, Michael Grabner, e Horst Bischof. Real-time tracking via on-line boosting. *British Machine Vision Conference (BMVC)*, 2006. 13

[21] Helmut Grabner, Christian Leistner, e Horst Bischof. Semi-supervised on-line boosting for robust tracking. *European conference on computer vision (ECCV)*. Springer, 2008. 13

[22] R. Gross, I. Matthews, J. Cohn, T. Kanade, e S. Baker. Multi-pie. *International Conference on Automatic Face Gesture Recognition (FG)*, 2008. , 54, 55

[23] Sam Hare, Amir Saffari, e Philip HS Torr. Struck: Structured output tracking with kernels. *International Conference on Computer Vision (ICCV)*. IEEE, 2011. 13

[24] João F Henriques, Rui Caseiro, Pedro Martins, e Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015. 13

[25] Seunghoon Hong, Tackgeun You, Suha Kwak, e Bohyung Han. Online tracking by learning discriminative saliency map with convolutional neural network. *International Conference on Machine Learning (ICML)*, 2015. 11, 13

[26] Zhibin Hong, Zhe Chen, Chaohui Wang, Xue Mei, Danil Prokhorov, e Dacheng Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 13

[27] Adam Hoover, Gillian Jean-Baptiste, Xiaoyi Jiang, Patrick J Flynn, Horst Bunke, Dmitry B Goldgof, Kevin Bowyer, David W Eggert, Andrew Fitzgibbon, e Robert B Fisher. An experimental comparison of range image segmentation algorithms. *IEEE ansactions on Pattern Analysis and Machine Intelligence (TPAMI)*. 26

[28] Anil Jain, Patrick Flynn, e Arun A Ross. *Handbook of biometrics*. Springer Science & Business Media, 2007. 11

[29] László A Jeni, Sergey Tulyakov, Lijun Yin, Nicu Sebe, e Jeffrey F Cohn. The first 3d face alignment in the wild (3dfaw) challenge. *European Conference on Computer Vision (ECCV)*. Springer, 2016. , 12, 17, 20, 50, 54, 55, 56, 57, 58, 59, 60, 62, 63, 64, 65, 66, 68

[30] Xu Jia, Huchuan Lu, e Ming-Hsuan Yang. Visual tracking via adaptive structural local sparse appearance model. *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. 13

[31] Martin Köstinger, Paul Wohlhart, Peter M Roth, e Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. *International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011. , 16, 17, 20

[32] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, e Roman Pflugfelder. The visual object tracking vot2015 challenge results. *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2015. 14, 23, 26, 27

[33] Alex Krizhevsky, Ilya Sutskever, e Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 2012. 13

[34] Mengtian Li, Laszlo Jeni, e Deva Ramanan. Brute-force facial landmark analysis with a 140,000-way classifier. *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 17, 20, 56, 62, 63

[35] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 2004. 19

[36] Chao Ma, Jia-Bin Huang, Xiaokang Yang, e Ming-Hsuan Yang. Hierarchical convolutional features for visual tracking. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 11, 14

[37] Xue Mei e Haibin Ling. Robust visual tracking using l1 minimization. *International Conference on Computer Vision (ICCV)*. IEEE, 2009. 13

[38] Hyeonseob Nam e Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016. 11, 14, 22, 23, 25, 43, 67

[39] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 2011. 53

[40] Shaoqing Ren, Kaiming He, Ross Girshick, e Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NIPS)*, 2015. 14, 20, 22

[41] David A Ross, Jongwoo Lim, Ruei-Sung Lin, e Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision (IJCV)*, 2008. 13

[42] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, e Maja Pantic. 300 faces in-the-wild challenge. *Image and Vision Computing*, 2016. 12

[43] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, e Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. *Proceedings of the International Conference on Computer Vision Workshops*. IEEE, 2013. , 12, 15, 16, 20

[44] Enrique Sánchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, e Michel Valstar. Cascaded continuous regression for real-time incremental face tracking. *European Conference on Computer Vision (ECCV)*. Springer, 2016. 11, 19

[45] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, e Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. *International Conference on Computer Vision Workshop (ICCV Workshops)*. IEEE, 2015. , 16, 24, 25, 26, 27, 28, 30, 38, 43, 67

[46] L.P. Silva, F.H.B. Zavan, L. Silva, e O.R.P. Bellon. Follow that nose: tracking faces based on the nose region and image quality feedback. *Conference on Graphics, Patterns and Images (SIBGRAPI)*. 2016. 11

[47] Severin Stalder, Helmut Grabner, e Luc Van Gool. Beyond semi-supervised tracking: Tracking should be as simple as detection, but not simpler than recognition. *Computer Vision Workshops (ICCV Workshops)*. IEEE, 2009. 13

[48] Yi Sun, Xiaogang Wang, e Xiaoou Tang. Deep convolutional network cascade for facial point detection. *Proceedings of the conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013. 15, 19

[49] Sergey Tulyakov e Nicu Sebe. Regressing a 3d face shape from a single image. *Proceedings of the IEEE International Conference on Computer Vision*, páginas 3748–3755, 2015. 54

[50] Georgios Tzimiropoulos, Joan Alabort-i Medina, Stefanos Zafeiriou, e Maja Pantic. Generic active appearance models revisited. *Asian Conference on Computer Vision (ACCV)*. Springer, 2012. 18

[51] Georgios Tzimiropoulos e Maja Pantic. Optimization problems for fast aam fitting in-the-wild. *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE, 2013. 18

[52] Georgios Tzimiropoulos e Maja Pantic. Gauss-newton deformable part models for face alignment in-the-wild. *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014. 18

[53] Michel Valstar, Brais Martinez, Xavier Binefa, e Maja Pantic. Facial point detection using boosted regression and graph models. *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010. 18

[54] Paul Viola e Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, 2004. 13

[55] Lijun Wang, Wanli Ouyang, Xiaogang Wang, e Huchuan Lu. Visual tracking with fully convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 11, 14

[56] Naiyan Wang, Siyi Li, Abhinav Gupta, e Dit-Yan Yeung. Transferring rich feature hierarchies for robust visual tracking. *arXiv preprint arXiv:1501.04587*, 2015. 13

[57] Lior Wolf, Tal Hassner, e Itay Maoz. Face recognition in unconstrained videos with matched background similarity. *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011. 23

[58] Yi Wu, Jongwoo Lim, e Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015. 14, 23, 26

[59] Shengtao Xiao, Shuicheng Yan, e Ashraf A Kassim. Facial landmark detection via progressive initialization. *Proceedings of the International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2015. 11, 16

[60] Xuehan Xiong e Fernando De la Torre. Supervised descent method and its applications to face alignment. *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013. , 12, 19, 51, 56, 59, 68

[61] Xuehan Xiong e Fernando De la Torre. Global supervised descent method. *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. 19

[62] Junjie Yan, Zhen Lei, Dong Yi, e Stan Li. Learn to combine multiple hypotheses for accurate face alignment. *Proceedings of the International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2013. 19

[63] Heng Yang, Wenxuan Mou, Yichi Zhang, Ioannis Patras, Hatice Gunes, e Peter Robinson. Face alignment assisted by head pose estimation. *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA, 2015. 19, 50

[64] Heng Yang e Ioannis Patras. Sieving regression forest votes for facial feature detection in the wild. *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE, 2013. 19

[65] Jing Yang, Jiankang Deng, Kaihua Zhang, e Qingshan Liu. Facial shape tracking via spatio-temporal cascade shape regression. *Proceedings of the International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2015. 11, 16

[66] L. Yin, X. Chen, Y. Sun, T. Worm, e M. Reale. A high-resolution 3d dynamic facial expression database. *International Conference on Automatic Face Gesture Recognition (FG)*, 2008. 54

[67] F. H. B. Zavan. Nose pose estimation in the wild and its applications on nose tracking and 3d face alignment. *Dissertação de Mestrado, Universidade Federal do Paraná - UFPR*, 2016. 11, 14, 22, 23

[68] Niv Zehngut, Felix Juefei-Xu, Rishabh Bardia, Dipan K Pal, Chandrasekhar Bhagavatula, e Marios Savvides. Investigating the feasibility of image-based nose biometrics. *International Conference on Image Processing (ICIP)*. IEEE, 2015. 11

[69] Kaihua Zhang, Lei Zhang, Qingshan Liu, David Zhang, e Ming-Hsuan Yang. Fast visual tracking via dense spatio-temporal context learning. *European Conference on Computer Vision (ECCV)*. Springer, 2014. 13

[70] Kaihua Zhang, Lei Zhang, e Ming-Hsuan Yang. Real-time compressive tracking. *European Conference on Computer Vision (ECCV)*. Springer, 2012. 13

[71] Tianzhu Zhang, Bernard Ghanem, Si Liu, e Narendra Ahuja. Robust visual tracking via multi-task sparse learning. *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. 13

[72] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, e Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 2014. , 54, 55

[73] Zhanpeng Zhang, Ping Luo, Chen Change Loy, e Xiaoou Tang. Facial landmark detection by deep multi-task learning. *European Conference on Computer Vision (ECCV)*. Springer, 2014. 19

[74] Zhanpeng Zhang, Ping Luo, Chen Change Loy, e Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. 19

[75] Ruiqi Zhao, Yan Wang, C Fabian Benitez-Quiroz, Yaojie Liu, e Aleix M Martinez. Fast and precise face alignment and 3d shape reconstruction from a single 2d image.

European Conference on Computer Vision (ECCV). Springer, 2016. 17, 20, 52, 56, 62, 63

[76] Wei Zhong, Huchuan Lu, e Ming-Hsuan Yang. Robust object tracking via sparsity-based collaborative model. *Computer vision and pattern recognition (CVPR)*. IEEE, 2012. 13

[77] Shizhan Zhu, Cheng Li, Chen Change Loy, e Xiaoou Tang. Face alignment by coarse-to-fine shape searching. *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. 19

[78] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, e Stan Z Li. Face alignment across large poses: A 3d solution. *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. , 16, 17