UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

RAFAEL GOMES MONTEIRO

DETECÇÃO DA DIREÇÃO DO OLHAR VIA WEBCAM

Dissertação de Mestrado submetida ao Corpo Docente do Departamento de Ciência da Computação do Instituto de Matemática, e Núcleo de Computação Eletrônica da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Mestre em Informática.

Orientador : Prof. Dr. Josefino Cabral Melo Lima
Co-Orientador: Prof. Dr. Antonio Carlos Gay Thomé

Rio de Janeiro
2012

Ficha Catalográfica

G633 Gomes Rafael Monteiro
Detecção da direção do olhar via webcam / Rafael Gomes Monteiro. - Rio de Janeiro: UFRJ, 2012.
106 f.: il.

Orientador: Josefino Cabral Melo Lima
Co-orientador: Antônio Carlos Gay Thomé
Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro, Instituto de Matemática, Instituto Tércio Pacitti, Programa de Pós-Graduação em Informática, 2012.

1. Processamento de Imagens. 2. Sistemas de Rastreamento. 3. Webcam – Teses. I. Lima, Josefino Cabral Melo (Orient.). II. Thomé, Antônio Carlos Gay (Co-orient.). III. Universidade Federal do Rio de Janeiro, Instituto de Matemática, Instituto Tércio Pacitti. IV. Título

Rafael Gomes Monteiro

Detecção da direção do olhar via webcam

Dissertação de Mestrado submetida ao Corpo Docente do Departamento de Ciência da Computação do Instituto de Matemática, e Núcleo de Computação Eletrônica da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Mestre em Informática.

Aprovado em: Rio de Janeiro, 14 de fevereiro de 2012

_____

Prof. Dr. Josefino Cabral Melo Lima (Orientador)

_____

Prof. Dr. Antonio Carlos Gay Thomé (Co-orientador)

_____

Profa. Dra. Simone Diniz Junqueira Barbosa

_____

Prof. Dr. Ricardo Farias

_____

Prof. Dr. Adriano Joaquim de Oliveira Cruz

_____

Prof. Dr. Paulo Roma Cavalcanti

Rio de Janeiro

2012

À Angélica, por me dar um sentido.

AGRADECIMENTOS

À minha família, pelas bases que me deram e por todo apoio, amor e principalmente compreensão pelo fato de eu ter estado extremamente ausente nos últimos três anos.

Ao professor Thomé, por ter acreditado em mim desde o momento em que este trabalho era apenas uma ideia, e por ter me orientado compartilhando informações, conhecimento e sabedoria.

Aos voluntários que se dispuseram a participar dos experimentos, possibilitando a criação da base de imagens utilizada neste trabalho.

Aos amigos e colegas que contribuíram direta ou indiretamente no desenvolvimento deste trabalho.

À Angélica, pela paciência e por estar sempre ao meu lado durante o desenvolvimento deste trabalho, me dando força, motivação e contribuindo diretamente nos nossos brainstormings.

RESUMO

GOMES, Rafael Monteiro. **Detecção da direção do olhar via webcam.** 2012. 106 f. Dissertação (Mestrado em Informática) – Instituto de Matemática, Instituto Tércio Pacitti, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2012.

Dentre as formas de Interação Humano-Computador descritas na literatura existem os sistemas de rastreamento do olhar, que consistem em estimar o ponto na tela para onde o usuário está focando a visão. Neste trabalho foi realizado um estudo sobre os sistemas de rastreamento do olhar existentes na literatura e foi desenvolvido um protótipo funcional para servir de base para pesquisas futuras. As principais contribuições deste trabalho estão no fato de se fazer uso de uma *webcam* comum para capturar as imagens, e o uso de iluminação ambiente, visto que a maioria dos sistemas utiliza câmeras de alta resolução e iluminação infravermelha, o que aumenta o custo. O sistema apresentou bons resultados, conseguindo alcançar cerca de 5.6^{o} de precisão nos experimentos realizados.

Palavras-chave: Detecção do olhar, processamento de imagens, webcam.

Gaze detection via webcam

ABSTRACT

GOMES, Rafael Monteiro. **Detecção da direção do olhar via webcam.** 2012. 106 f. Dissertação (Mestrado em Informática) – Instituto de Matemática, Instituto Tércio Pacitti, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2012.

Amongst the several existing ways of human-computer interaction described in the literature there are the eye tracking systems, which estimate the point on the screen where the user is looking at. In this dissertation we present a study on the gaze tracking systems available in the literature and a working prototype was developed to serve as a basis for future research. The major contributions of this study are in the usage of a common webcam to capture images and the use of ambient light, as most systems use high-resolution cameras and infrared lighting, which increases their costs. The system showed good results, reaching to around 5.6^{o} of accuracy in the experiments.

Palavras-chave: Detecção do olhar, processamento de imagens, webcam.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AAM	*Active Appearance Models*
EM	*Expectation - Maximization*
EOG	Eletro-oculografia
IHC	Interação Humano-Computador
HMD	*Head-Mounted Display*
RANSAC	*Random Sample Consensus*
RNA	Rede Neural Artificial
SVM	*Support Vector Machines*

SUMÁRIO

1 INTRODUÇÃO

A Interação Humano-Computador (IHC) é uma área que estuda a interação entre as pessoas e os computadores, envolvendo conceitos de áreas como Ciência da Computação, Psicologia e Linguística, dentre outras (FILHO, 2003). De maneira geral, essa interação deve ser simples, amigável e intuitiva, de forma a atingir a maior quantidade de pessoas. Porém, antes de tudo é fundamental que tal interação seja no mínimo possível, o que remete ao conceito de acessibilidade. Uma interface gráfica muito bem construída não será necessariamente acessível a um deficiente visual, por exemplo.

Acessibilidade é um assunto bastante importante, sendo inclusive tema da lei 10.098 de 19 de dezembro de 2000, que "estabelece normas gerais e critérios básicos para a promoção da acessibilidade das pessoas portadoras de deficiência ou com mobilidade reduzida, e dá outras providências"(BRASIL, 2000). Acessibilidade em IHC consiste em garantir que as pessoas possam acessar o computador, independente das suas limitações. Isso pode ser feito de várias formas, dependendo do tipo da limitação.

Para portadores de deficiência visual, por exemplo, existem programas que

atuam como leitores de tela, como é o caso do sistema DOSVOX (BORGES, 2002), desenvolvido pelo Núcleo de Computação Eletrônica da Universidade Federal do Rio de Janeiro. Esse sistema se comunica com o usuário através de síntese de voz, que consiste em uma técnica para transformar texto em voz, de maneira que o computador lê para o usuário o que está escrito na tela.

Já para o caso de pessoas que possuam alguma deficiência motora que as impeçam de utilizar os dispositivos de entrada padrão, como teclado e mouse, existem sistemas que capturam a entrada de dados do usuário a partir de outras fontes de informação. Isso pode ser feito, por exemplo, através do reconhecimento de fala, detecção de piscadas de olhos, rastreamento do movimento do olhar, etc (EDWARDS, 1995).

O rastreamento do olhar consiste em tentar estimar o ponto na tela para onde o usuário está olhando com base na captura de imagens dos seus olhos. Existem vários tipos de sistemas que realizam essa tarefa (HANSEN E JI, 2010). Em alguns sistemas o usuário utiliza um capacete que possui câmeras direcionadas para os seus olhos. Geralmente há um sensor que estima a posição da cabeça, enquanto a câmera fornece imagens em alta resolução dos olhos, para estimar a direção do olhar. Combinando essas informações e a posição da tela do computador é possível estimar o ponto para onde o usuário está olhando. Apesar de serem bastante precisos, há o problema desses sistemas serem invasivos, devido ao hardware que deve ser acoplado no usuário.

Existem também sistemas não-invasivos, onde tipicamente há uma (ou mais de uma) câmera de vídeo posicionada à frente do usuário, apontada na sua direção, capturando imagens que são processadas por um *software* que tenta estimar o olhar do usuário. Apesar de geralmente serem menos precisos, possuem um custo mais baixo, sendo mais acessíveis à maioria das pessoas.

1.1 Objetivos

O principal objetivo neste trabalho foi desenvolver um sistema de detecção da direção do olhar com base em imagens capturadas a partir de uma câmera de vídeo. Complementando o foco principal da pesquisa, estabeleceu-se como objetivo específico que o sistema possuísse baixo custo, de forma que o mesmo fosse capaz de operar com uma *webcam* comum, tornando-o mais próximo de ser acessível para o público em geral.

O sistema foi desenvolvido em módulos, que atuam de forma sequencial e que correspondem às fases principais necessárias para que ele atinja o objetivo:

1. Localização do olho: módulo responsável por processar a imagem e determinar a posição dos olhos do usuário na imagem. Essa informação é passada para os módulos seguintes, para processamento posterior.

2. Detecção do ponto fixo: módulo responsável por localizar e rastrear um ponto fixo na face do usuário. Isso é necessário para que a direção do olhar seja estimada, com base no deslocamento entre os olhos com relação ao ponto fixo.

3. Estimação da direção do olhar: módulo que, com base nas informações obtidas através dos outros módulos, estima a direção do olhar do usuário, utilizando um processo de calibração.

1.2 Motivação

A principal motivação para o desenvolvimento deste trabalho está relacionada às aplicações do produto gerado, pois espera-se que ele contribua para aumentar a

acessibilidade de pessoas portadoras de algum tipo de deficiência motora que as impeçam de acessar o computador da forma tradicional, através do teclado e do mouse. Essa tecnologia também pode ser utilizada pelo público em geral como um substituto do mouse, ou complemento ao uso do mesmo. Outra aplicação interessante seria a utilização da tecnologia em caixas eletrônicos de bancos, principalmente para aumentar a segurança, pois o cliente poderia digitar sua senha apenas olhando para os dígitos, o que dificulta que esta seja descoberta por pessoas que estejam próximas a ele.

Além disso, espera-se que o sistema contribua para a consolidação desta área de conhecimento no PPGI e para o desenvolvimento de pesquisas futuras. A existência de um sistema funcional possibilitará que novas pesquisas, mais focadas e direcionadas, possam buscar melhores resultados em cada módulo.

1.3 Resumo dos resultados

O protótipo desenvolvido apresentou bons resultados, apesar de utilizar imagens capturadas por uma *webcam*. O sistema obteve em média cerca de 5.6^o de precisão[1], o que não é considerada uma precisão excelente se comparada ao estado da arte, onde sistemas comerciais conseguem obter cerca de 0.5^o de precisão (vide seção 2.3). Apesar disso, ele permite distinguir certas áreas na tela para onde o usuário está olhando, o que possibilitaria o controle de interfaces gráficas projetadas de forma a compensar o erro gerado pelo sistema. Por exemplo: num monitor de 24" *widescreen* com resolução de 1920x1080 píxeis a uma distância de 60cm do usuário, o sistema conseguiria distinguir, com bastante precisão, botões com tamanho de cerca de 210x210 píxeis. Isso permitiria montar uma interface gráfica contendo cerca de 45 botões, dispostos numa grade 9x5.

[1] Vide seção 4.3.1 para maiores informações sobre a mensuração da acurácia em graus

1.4 Organização da dissertação

O capítulo 2 fala sobre a área de rastreamento do olhar, apresentando um breve histórico, a organização da área e trabalhos relacionados. O capítulo 3 apresenta a abordagem proposta, especificando a divisão do sistema e detalhando as técnicas utilizadas em cada módulo, apresentando alguns dos resultados obtidos. O capítulo 4 apresenta de forma detalhada os resultados do trabalho nos testes realizados, avaliando o desempenho de cada módulo e a precisão geral do sistema. O capítulo 5 expõe as conclusões e sugestões para trabalhos futuros.

2 RASTREAMENTO DO OLHAR

Rastreamento do olhar consiste em determinar e acompanhar ao longo do tempo os pontos no espaço para onde uma pessoa está olhando. Duchowski (2003, apud HANSEN E JI, 2010) define duas grandes áreas de aplicação desse tipo de tecnologia: diagnóstico e interação.

Na área de diagnóstico, o objetivo é obter informações do olhar para fins de análise do comportamento do usuário. Um exemplo prático é a análise do olhar de um usuário enquanto o mesmo observa uma página da Internet. Deseja-se saber se os elementos visuais da página estão bem organizados, para quais *links* o usuário olha, etc. Na área de interação o objetivo é que o sistema reaja ao olhar do usuário, de forma que haja uma interação entre o usuário e o computador através do movimento dos olhos. Uma pessoa que não possua mobilidade dos membros, por exemplo, poderia controlar o cursor do mouse apenas olhando para a tela.

2.1 Histórico

O interesse científico em rastrear o olhar de uma pessoa remonta ao século XIX, tendo como marco inicial os estudos de Louis Émile Javal sobre a movimentação dos olhos de uma pessoa enquanto esta lê um texto (JAVAL, 1879). Javal observou que, ao ler um texto, os olhos não se movimentam de forma contínua e constante através de cada linha de texto, mas realizam saltos entre fragmentos de texto e se fixam nesses fragmentos por um curto período de tempo. A análise feita por Javal foi feita a partir de observações a olho nu, sem o uso de tecnologias para rastrear o movimento dos olhos.

Em 1893, M. Lamare construiu um aparato onde um sensor sonoro foi colocado sobre a pálpebra superior do olho de um indivíduo. Cada movimento do olho emitia um som, que era capturado pelo sensor e convertido na informação de que houve movimento dos olhos num dado instante de tempo (LAMARE, 1893). Apesar de ser um método invasivo, produzia uma informação mais precisa do que a informação obtida por meio de observações a olho nu.

Um grande avanço nesta área ocorreu quando E. Schott construiu um dispositivo para capturar os movimentos dos olhos através de eletrodos que mediam a diferença de potencial dos músculos da face, próximos aos olhos (SCHOTT, 1922). Essa técnica é conhecida como eletro-oculografia (EOG), e um exemplo disso é ilustrado na figura 2.1.

Com o advento do computador, os sistemas de rastreamento do olhar mais recentes utilizam soluções computacionais para atingir seus objetivos. As imagens são capturadas através de câmeras de vídeo enquanto o usuário olha para a tela do computador. A partir das imagens capturadas, o sistema tenta estimar o ponto na tela para onde o usuário está olhando.

Figura 2.1: Exemplo de sistema de eletro-oculografia (DUCHOWSKI, 2007)

2.2 Trabalhos relacionados

Segundo HANSEN E JI (2010), que recentemente publicaram um levantamento feito sobre as pesquisas que vêm sendo realizadas na área, duas subáreas concentram o esforço e interesse dos pesquisadores: a localização do olho na imagem e a estimativa da direção do olhar. Esta seção aborda as duas subáreas, explicando seus objetivos e apresentando alguns trabalhos existentes na literatura.

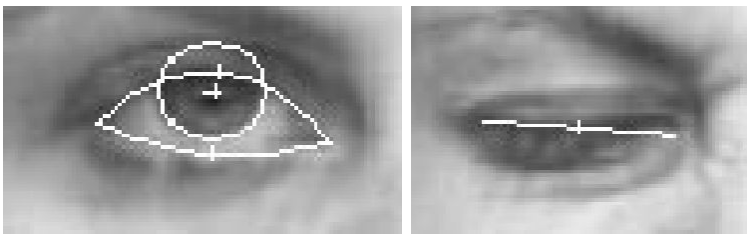
2.2.1 Localização do olho na imagem

A localização do olho na imagem visa encontrar, com precisão, a posição dos olhos em cada quadro obtido pelo dispositivo de captura. Isso é feito através do uso de variadas técnicas de processamento de imagens e reconhecimento de padrões. Os aspectos principais são a detecção da existência ou não dos olhos em uma imagem, a determinação precisa da posição desses e o seu rastreamento através dos quadros

sucessivos de vídeo. Vários são os desafios encontrados, como a oclusão do olho pelas pálpebras, a distinção entre o olho aberto e fechado, variação em tamanho, refletividade, posição da cabeça, dentre outros fatores.

Existem várias abordagens para a localização do olho, como as que utilizam modelos baseados na forma, através do uso de formas fixas ou deformáveis (IVINS E PORRILL, 1998; TIAN; KANADE E COHN, 2000; HANSEN E PECE, 2005). Esses modelos são construídos com base em descritores (*features*), informações sobre bordas ou utilizando reposta de filtros. Também há abordagens que utilizam modelos baseados na aparência, através de casamento de padrões (*template matching*), utilizando medidas de similaridade (HUANG E WECHSLER, 1999; HUANG E MARIANI, 2000; PENTLAND; MOGHADDAM E STARNER, 1994). Existem também os métodos híbridos, que combinam técnicas utilizadas nas duas outras abordagens (HANSEN et al., 2003; ISHIKAWA et al., 2004; MATSUMOTO E ZELINSKY, 2000).

TIAN; KANADE E COHN (2000) utilizaram um modelo de formas fixas baseados na detecção de dois estados do olho: aberto e fechado. A figura 2.2 ilustra o modelo utilizado por eles. O casamento do modelo é baseado no uso de detectores de bordas e de cantos, logo é necessário o uso de imagens com alto contraste.



(a) (b)

Figura 2.2: Modelo de dois estados de TIAN; KANADE E COHN (2000). (a) Modelo do olho aberto e (b) modelo do olho fechado

ZHU; FUJIMURA E JI (2002) desenvolveram um sistema baseado na aparência utilizando iluminação infravermelha. Eles construíram um iluminador que possui lâmpadas de infravermelho próximas à câmera e distantes da mesma, conforme ilustra a figura 2.3. Quando a iluminação próxima a câmera é ativada, a maioria da luz é refletida de volta para a câmera, o que produz um efeito de pupila clara. Quando a iluminação mais distante é ativada, a pupila fica mais escura. Com base nisso, os quadros de vídeo são capturados de forma que a iluminação é alternadamente trocada em cada frame.

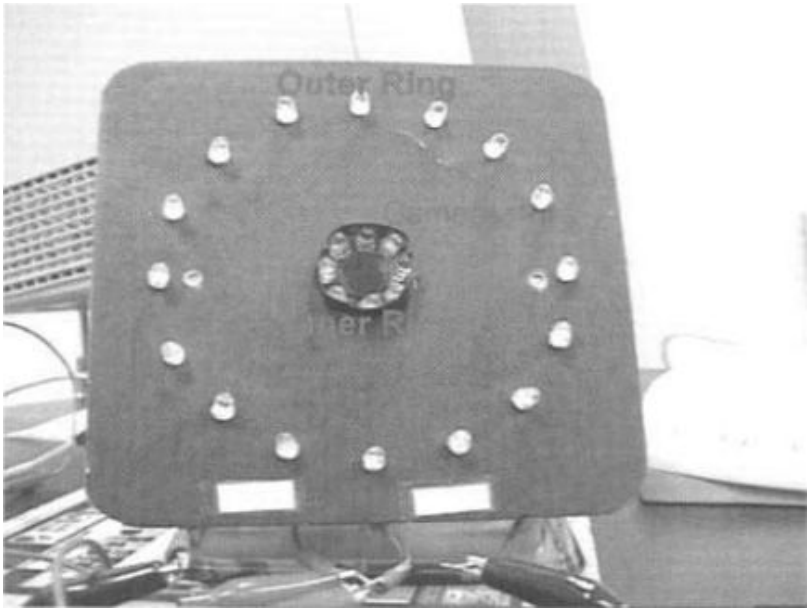


Figura 2.3: Iluminador infravermelho desenvolvido por ZHU; FUJIMURA E JI (2002)

Primeiro é feita uma pré-seleção dos componentes conexos através da análise da diferença entre dois quadros de vídeo, sendo um capturado com a iluminação infravermelha próxima à câmera e outro com a iluminação distante, como ilustra a figura 2.4. Máquinas de vetores de suporte (*Support Vector Machines - SVM*) são usadas para refinar a localização e o rastreamento é feito através do algoritmo *Mean-Shift* (COMANICIU; RAMESH E MEER, 2000).

(a) (b) (c)

Figura 2.4: Exemplo de captura com iluminação infravermelha alternada: (a) e (b) são quadros subsequentes e (c) diferença entre os quadros

Sistemas que utilizam iluminação infravermelha são bastante comuns devido a sua alta reflexividade no globo ocular, o que facilita a tarefa de localização dos olhos. A figura 2.5 ilustra a diferença das imagens do olho capturadas com a iluminação próxima e distante da câmera, gerando o efeito de pupila clara e pupila escura. Porém, esse tipo de sistema gera um custo adicional, pois requer o uso de lâmpadas especiais. Também é necessário o uso de câmeras específicas que consigam captar a iluminação infravermelha ou adaptação de câmeras comuns para esse objetivo, o que pode ser feito através da remoção de filtros existentes em suas lentes.

(a) (b)

Figura 2.5: Exemplo de captura com iluminação (a) próxima à câmera e (b) distante da câmera (HANSEN E JI, 2010)

DROEGE; GEIER E PAULUS (2007) construíram um sistema bastante robusto e de baixo custo utilizando iluminação infravermelha. A detecção da íris foi baseada numa adaptação da Transformada de Hough, utilizando a média entre as

coordenadas do centro das duas íris como base. O reflexo da iluminação infravermelha na íris é utilizado para medir o deslocamento do olhar, através da diferença das suas coordenadas e das coordenadas do centro das íris. O sistema apresentou bons resultados, conseguindo distinguir pontos numa grade de 5x7 blocos na tela.

Posteriormente, DROEGE; SCHMIDT E PAULUS (2008) fizeram um estudo sobre algoritmos para detectar a posição da íris em imagens com baixa resolução. Os algoritmos pesquisados atuam de forma semelhante, detectando as bordas da pupila para posteriormente refinar a detecção e tentar estimar o seu centro. As mesmas técnicas podem ser utilizadas para detectar os centros das íris, quando a resolução da imagem é insuficiente para distinguir a fronteira entre a íris e a pupila. A conclusão do estudo é que não há um algoritmo ótimo, mas sim algoritmos melhores para situações específicas.

VALENTI E GEVERS (2008) desenvolveram uma técnica chamada de curvatura isolux, que é robusta a mudanças de iluminação ambiente, sem a necessidade do uso de iluminação infravermelha. O sistema obteve excelentes resultados utilizando uma *webcam*. Porém, é necessário capturar as imagens de forma que a íris apareça com um formato circular simétrico, ou o desempenho do sistema é prejudicado.

CRISAFULLI; IANNIZZOTTO E LA ROSA (2009) desenvolveram uma forma competitiva de rastrear o olhar através do uso de várias técnicas que, quando combinadas, produziriam um resultado mais apurado. Para isso eles utilizaram cinco técnicas distintas para detectar as coordenadas da íris e em seguida escolhem o melhor resultado através de uma análise entre as distâncias das coordenadas retornadas entre cada técnica. Os autores realizaram uma redução da área de busca da íris através da definição de sub-regiões geradas a partir a região obtida pelo detector de faces.

2.2.2 Estimativa da direção do olhar

A estimativa da direção do olhar visa tentar determinar o ponto para onde o usuário está olhando na tela. Geralmente isso é feito através de um processo de calibração, onde o usuário olha para alguns pontos pré-definidos, que são armazenados pelo sistema e interpolados para gerar a informação da direção do olhar nos quadros sucessivos de vídeo. Os aspectos principais são a estimativa da linha de visão do usuário e a estimativa da posição do globo.

A estimativa da linha de visão pode ser feita com apenas uma câmera através de um processo de interpolação, com base em informações obtidas através de um processo de calibração (BROLLY E MULLIGAN, 2004; EBISAWA E SATOH, 1993). O objetivo é estimar um ponto para onde o usuário está olhando em algum objeto de interesse (ex.: a tela do computador).

A estimativa da posição do globo pode ser feita através do uso de mais de uma câmera, que tentam estimar a posição da cabeça e a posição do globo ocular, combinando essas informações de forma a tentar obter um modelo em três dimensões da posição do olho, permitindo estimar a linha de visão do usuário (OHNO E MUKAWA, 2004; VILLANUEVA; CABEZA E PORTA, 2006). Essa linha de visão pode ser combinada com informações sobre o ambiente, de forma a estimar o objeto para onde o usuário está olhando.

Nos últimos anos o número de pesquisas relacionadas à estimativa da direção do olhar vem aumentando consideravelmente. Segundo VILLANUEVA et al. (2008), modelos matemáticos e geométricos apresentam benefícios potenciais por fornecerem conhecimento a priori sobre o comportamento de aspectos do sistema, como movimentação da cabeça, acurácia, áreas de erro, etc.

JI E ZHU (2002) utilizaram redes neurais para estimar a direção do olhar do usuário, combinando a informação do centro da pupila com a reflexão gerada pela iluminação infravermelha. O sistema desenvolvido não exigia calibração por usuário, podendo aproveitar uma mesma calibração para outros usuários. O rastreamento foi feito utilizando iluminação infravermelha, de forma que as pupilas pudessem ser facilmente detectadas.

COUTINHO E MORIMOTO (2006) também utilizaram iluminação infravermelha para localizar os olhos e propõem um procedimento para estimação do olhar que visa compensar a diferença angular entre os eixos dos olhos e da visão. Isso foi feito utilizando uma câmera e cinco fontes de luz, gerando um padrão de reflexão na íris que é posteriormente utilizado para estimar a direção do olhar. O sistema possui um desempenho de 30 quadros por segundo com alta acurácia. Uma visão do *hardware* utilizado é exibida na figura 2.6. Nota-se uma semelhança com o sistema desenvolvido por ZHU; FUJIMURA E JI (2002), exibido na figura 2.3.

Figura 2.6: Sistema desenvolvido por COUTINHO E MORIMOTO (2006)

KUNKA E KOSTEK (2009) desenvolveram um sistema não intrusivo de rastreamento do olhar sem a necessidade do uso de iluminação infravermelha. Após a detecção da face, os olhos são detectados utilizando a Transformada Circular de Hough (HOUGH, 1959; DUDA E HART, 1972). Os cantos dos olhos são detectados como pontos fixos, para que seja possível medir o movimento dos olhos e estimar a direção do olhar. A detecção dos cantos dos olhos é feita através de uma varredura vertical dos píxeis próximos aos olhos de forma a buscar áreas de maior variância que, supostamente, seriam os cantos dos olhos.

TUNHUA et al. (2010) construíram um sistema também não intrusivo para rastreamento do olhar que faz o rastreamento da face entre quadros de vídeo através do algoritmo *Camshift* (WANG et al., 2009), a localização dos olhos através de um detector treinado pelo algoritmo *Adaboost* (VIOLA E JONES, 2001a) e seu rastreamento através do algoritmo *Optical Flow* de Lucas-Kanade (SCHREIBER, 2008). Como ponto fixo, foi utilizado o nariz. O sistema gera melhores resultados quando a cabeça permanece imóvel durante o seu uso.

CUONG E HOANG (2010) desenvolveram um sistema utilizando uma *web-cam* comum que, assim como o sistema desenvolvido por KUNKA E KOSTEK (2009), também faz a detecção dos cantos dos olhos para serem utilizados como pontos fixos, utilizando o detector de cantos definido por HARRIS E STEPHENS (1988). A detecção da íris é feita utilizando uma técnica similar à Transformada de Hough. O sistema obteve bons resultados, mesmo trabalhando com imagens em baixa resolução (320x240 píxeis).

Existem diversas maneiras de combinar as abordagens utilizadas nos trabalhos, principalmente com relação ao *hardware*, seja utilizando uma câmera e uma fonte de luz, uma câmera e várias fontes de luz, várias câmeras e várias fontes de luz, e assim por diante. Cada abordagem possui as suas vantagens e desvantagens, que

correspondem basicamente à precisão obtida e o custo gerado. De forma geral, o uso de mais de uma câmera, de câmeras de alta resolução ou de um ambiente com iluminação controlada produz maior precisão, porém gera um custo maior (HANSEN E JI, 2010). É comum o uso de capacetes com câmeras acopladas, também conhecidos como *head-mounted displays* (HMD), onde câmeras são posicionadas bem próximas aos olhos, conseguindo capturar imagens em altíssima resolução (BOENING et al., 2006). Esse tipo de sistema costuma gerar ótimos resultados, apesar de ser considerado invasivo.

2.2.3 Direções futuras de pesquisa

A área apresenta vários problemas em aberto, o que permite que muitas pesquisas possam ser realizadas. HANSEN E JI (2010) definiram algumas direções para pesquisas futuras na área, como:

1. Limitar o uso de luz infravermelha, devido aos custos de mais um dispositivo de iluminação;

2. Evitar o uso de capacetes com câmeras acopladas, devido a ser um método considerado invasivo;

3. Oferecer configuração mais flexível, no sentido de não exigir um ambiente altamente controlado, tornando os sistemas mais baratos e acessíveis;

4. Diminuir a necessidade de calibração do equipamento, pois o processo pode ser cansativo para o usuário;

5. Diminuir os custos utilizando *hardware* mais simples, como câmeras mais baratas, iluminação ambiente ao invés de luz infravermelha, que é bastante comum;

6. Aumentar o grau de tolerância a óculos, lentes de contato, variações nas condições do ambiente;

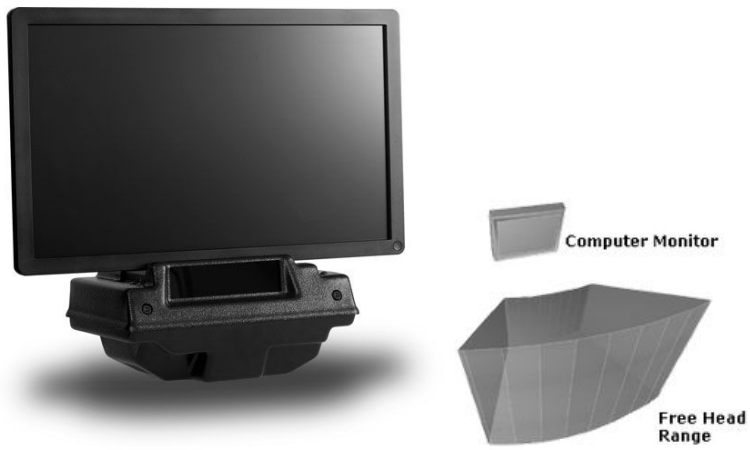
7. Realizar estudos sobre a análise dos movimentos oculares, no sentido da interpretação dos estados cognitivos e afetivos relacionados ao comportamento do olhar do usuário.

2.3 Sistemas comerciais de rastreamento do olhar

Além dos trabalhos existentes na literatura, existem também os sistemas de rastreamento do olhar comerciais, disponíveis no mercado. Os preços variam muito, podendo chegar a R$ 180.000,00. Esta seção apresenta alguns dos sistemas comerciais existentes e cita suas principais características.

O sistema EyeFollower 2.0, da empresa Interactive Minds, é um sistema bastante robusto, permitindo que o usuário movimente livremente a cabeça e possuindo uma acurácia menor que 0.4 graus. A figura 2.7 ilustra o sistema e também mostra graficamente o espaço permitido para a movimentação da cabeça do usuário. O sistema é não intrusivo, pois nenhum artifício precisa ser colocado no usuário, que precisa apenas estar à frente do monitor e da câmera. A iluminação utilizada é infravermelha, presente na maioria dos sistemas comerciais de rastreamento do olhar por gerar imagens mais nítidas do que as obtidas com iluminação natural.

O EyeLink 1000, da empresa SR Research, é um sistema bastante similar ao EyeFollower, que também permite a movimentação da cabeça do usuário e possui uma acurácia de 0.5 graus. Tolera movimentos rápidos da cabeça, mas exige a fixação de uma marcação manual no usuário, conforme ilustrado na figura 2.8.



(a) (b)

Figura 2.7: EyeFollower 2.0, da empresa Interactive Minds: (a) visão do sistema e (b) espaço de movimentação da cabeça

(a) (b)

Figura 2.8: EyeLink 1000, da empresa SR Research: (a) visão do sistema e (b) detalhe do marcador artificial

O T60XL, da Tobii, possui precisão entre 0.4 e 0.6 graus, dependendo da variação da iluminação do ambiente e da distância do usuário para a câmera. O sistema também tolera movimentos da cabeça e, além da posição do olhar, retorna outras informações, como o diâmetro da pupila, que pode ser utilizada para estudos sobre a variação do tamanho ao longo do tempo. A taxa de captura é de 60 quadros por segundo (60Hz). O sistema também conta com uma *toolbox* para o Matlab. A

figura 2.9 ilustra o sistema. O preço estimado no Brasil é de R$ 180.000,00[1].

Figura 2.9: T60XL, da empresa Tobii

O easyGaze da DESIGNinteractive possui precisão informada no site de menos de 1 grau. Captura as imagens utilizando iluminação infravermelha e pode alcançar a taxa de 55 quadros por segundo. O sistema vem no formato de uma base que deve ser posicionada abaixo do monitor do usuário ou de qualquer outro objeto de interesse, conforme ilustra a figura 2.10. Seu custo é de US$ 8.300,00[2].

Figura 2.10: easyGaze, da empresa DESIGNinteractive

Similar ao eyeGaze, o TM3 da empresa eyeTech também vem no formato de uma base a ser posicionada abaixo do monitor, como é ilustrado na figura 2.11. Possui precisão informada no site de 0.5 graus, também alcança 55 quadros por segundo e utiliza iluminação infravermelha e tolera movimentos de cabeça.

[1]Segundo consulta feita ao representante de vendas no Brasil em novembro de 2011. O preço pode variar de acordo com a cotação do Dólar.

[2]Segundo informações no site http://designinteractive.net

Figura 2.11: TM3, da empresa eyeTech

O RED da SMI também vem no formato de base. A precisão informada no site é de menos de 0.4 graus e o sistema consegue capturar imagens a uma taxa superior a 166 quadros por segundo. A figura 2.12 ilustra o sistema.

Figura 2.12: RED, da empresa SMI

De forma geral, pode-se observar que a maioria dos sistemas comerciais utiliza iluminação infravermelha, possui precisão de cerca de 0.5^o e tolera movimentos da cabeça. Em alguns sistemas observou-se a existência de mais de uma câmera, sendo uma fixa, com visão ampla, e outra com zoom, focada apenas nos olhos. A primeira câmera realiza a detecção da face e, com base na sua posição, a segunda câmera é direcionada mecanicamente para os olhos do usuário. Esses e outros fatores aumentam a precisão desses sistemas, porém encarecem seu custo.

3 MODELO PROPOSTO

A proposta? deste trabalho foi estudar, avaliar e comparar os sistemas de rastreamento do olhar disponíveis na literatura e desenvolver um protótipo funcional, que sirva de base para pesquisas futuras. O sistema foi desenvolvido de forma a considerar alguns dos problemas em aberto da área, citados na seção 2.2.3. As principais características são duas: usar uma *webcam* comum para a captura das imagens e usar iluminação ambiente, sem a necessidade do uso de luz infravermelha, comum em grande parte dos trabalhos existentes na literatura (HANSEN E JI, 2010).

3.1 Análise das alternativas

Foi feito um estudo sobre a área de forma a compreender como são construídos os sistemas de rastreamento do olhar. Várias são as abordagens utilizadas na literatura, conforme citado na seção 2. Optou-se por utilizar técnicas que necessitem de pouco processamento, produzindo um resultado rápido para que o protótipo possa ser utilizado sem muitas exigências de *hardware*.

Na localização do olho, foram utilizadas técnicas baseadas em modelos de

formas fixas, pois utilizam poucas variáveis livres para casar os modelos com as formas existentes na imagem. O uso de poucas variáveis livres reduz o processamento, visto que requer menos cálculos a serem realizados para encontrar os valores ideais dessas variáveis, pois há um conjunto menor de combinações possíveis.

Já quanto à estimativa do olhar, foi utilizado um modelo neural simples para transformar os dados de calibração numa função de mapeamento entre o deslocamento dos olhos e as coordenadas de tela. A execução de uma rede neural pode ser feita basicamente através de um produto entre matrizes, o que pode ser rapidamente computado.

3.2 Configuração do sistema

O sistema foi desenvolvido em módulos, conforme ilustra a figura 3.1. O desenvolvimento em módulos é vantajoso porque o desempenho do sistema pode ser medido separadamente e para que pesquisas mais focadas possam ser realizadas em cada módulo. Quanto ao *hardware*, a câmera utilizada foi uma *webcam* Logitech QuickCam Pro 9000, com sensor de 2 megapíxeis, capaz de capturar imagens na resolução de 1600x1200 píxeis.

O Módulo 1 é responsável por localizar os olhos do usuário, mais especificamente as coordenadas do centro das íris. Ele localiza os dois olhos e retorna o ponto médio entre os centros das duas íris, denotado por (O_X, O_Y). Internamente, esse módulo localiza a face do usuário, passando-a para o Módulo 2, que faz a detecção do ponto fixo, cujas coordenadas são representadas por (F_X, F_Y). Um ponto fixo é um ponto que permanece estacionário entre os quadros de vídeo e é necessário para que se possa calcular o deslocamento dos olhos. Conforme os mesmos se movimentam, o ponto fixo permanece na mesma posição, então pode-se calcular o deslocamento

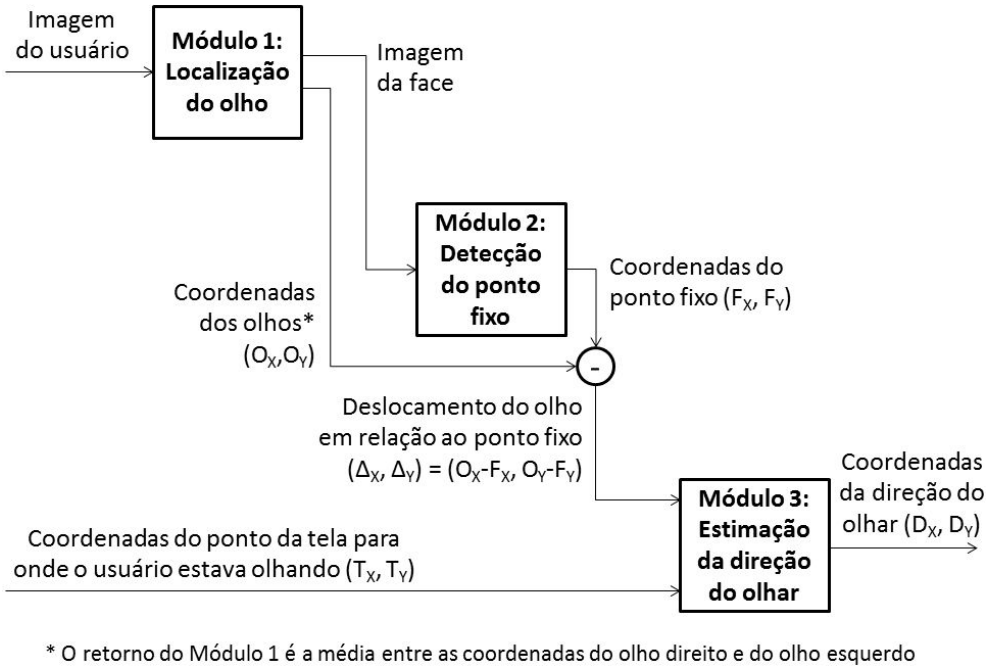


Figura 3.1: Visão geral dos módulos da abordagem proposta

com base na diferença entre as coordenadas dos olhos e as coordenadas do ponto fixo, da seguinte forma: $(\Delta_X, \Delta_Y) = (O_X - F_X, O_Y - F_Y)$. O Módulo 3 utiliza essa informação de deslocamento para estimar a direção do olhar. Para isso, ele utiliza também as coordenadas do ponto na tela para onde o usuário estava olhando no momento da captura daquele quadro, denotados por (T_X, T_Y). O objetivo desse módulo é fazer um mapeamento entre o deslocamento dos olhos e o ponto na tela.

3.3 Metodologia de desenvolvimento

O processo de rastreamento do olhar geralmente é dividido em duas fases (HANSEN E JI, 2010), denominadas calibração e teste. Em ambas as fases, o usuário deve olhar para alguns pontos previamente definidos na tela e o sistema irá

capturar algumas imagens do usuário enquanto o mesmo olha para cada ponto. Os pontos são exibidos em sequência através de um programa de captura. Uma das formas de se fazer a captura é solicitar que o usuário pressione uma tecla enquanto está olhando para o ponto atualmente exibido, para que o sistema faça a captura e passe para o próximo ponto. Isso é feito para que o usuário possa piscar o olho livremente entre as capturas das imagens em cada ponto, mas permaneça com o olho aberto durante as capturas.

Na fase de calibração, são exibidos nove pontos, dispostos numa grade com 3x3 pontos. Esses pontos serão utilizados para que o sistema aprenda a fazer o mapeamento entre o deslocamento dos olhos e as coordenadas de tela. Em seguida, são capturados 25 pontos, dispostos numa grade com 5x5 pontos, para a fase de teste. Nessa fase o sistema irá tentar mapear o olhar do usuário, capturado nos 25 pontos, para as coordenadas da tela, com base nas informações obtidas na calibração. Essas quantidades de pontos foram definidas com base nos trabalhos relacionados, que utilizam grades de pontos com quantidades similares. Os pontos utilizados para calibração e teste são ilustrados na figura 3.2.

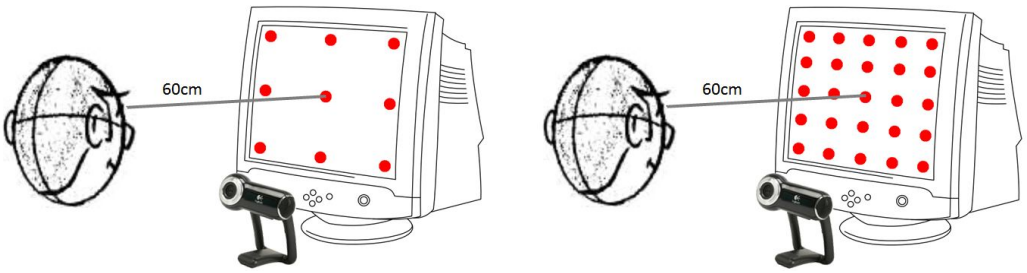
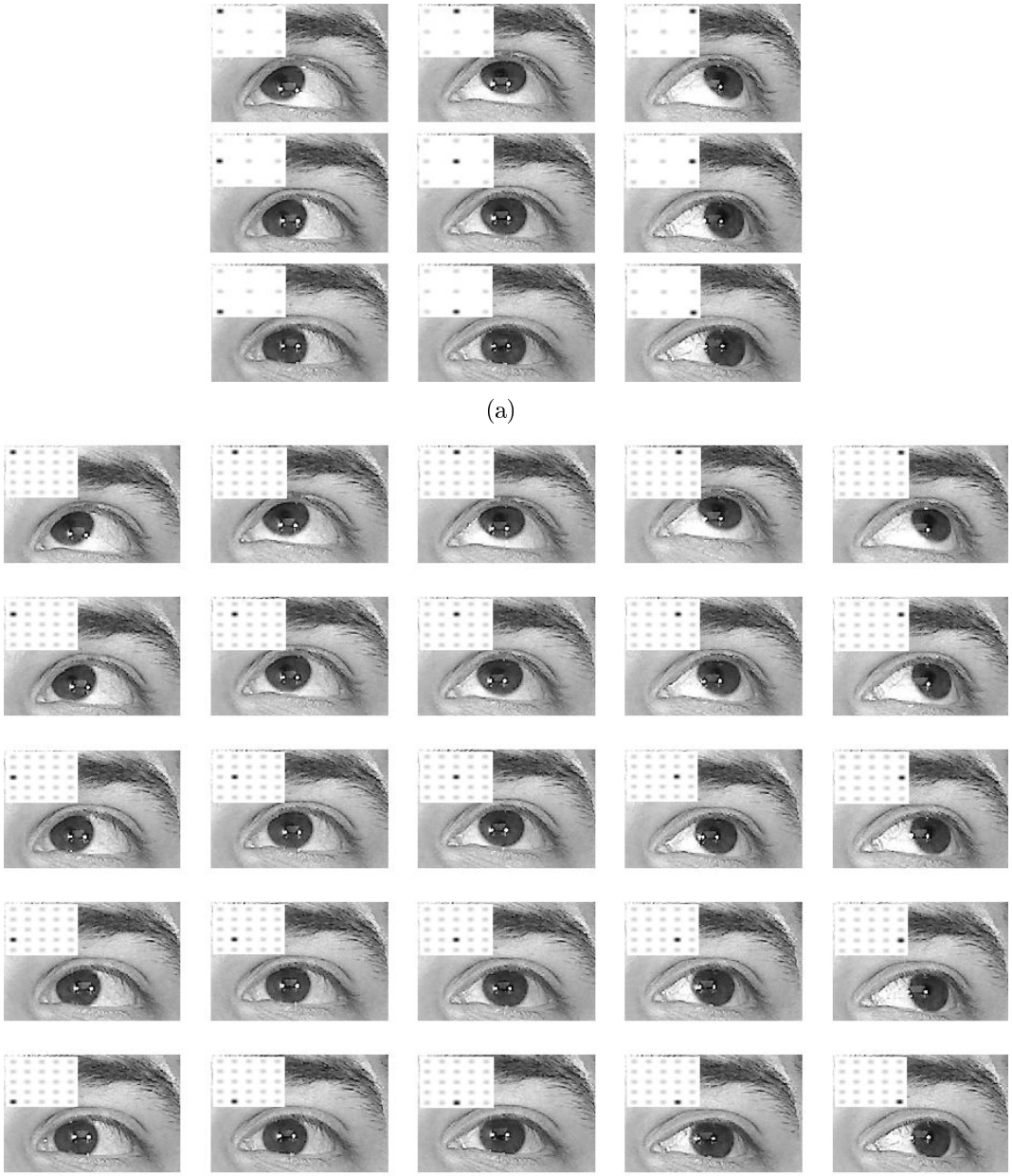


Figura 3.2: Pontos utilizados nas fases de (a) calibração e (b) teste

A figura 3.3 ilustra exemplos de imagens dos olhos de um usuário olhando para cada um dos pontos das fases de calibração e de teste.



(a)

(b)

Figura 3.3: Exemplos de imagens dos olhos olhando para cada ponto (a) da calibração e (b) do teste. O ponto para onde o usuário estava olhando está destacado em cada figura.

Os experimentos foram realizados a partir de uma base de dados de imagens criada num ambiente controlado. Foram capturadas imagens de cinco usuários distintos. Em cada captura, o usuário permanecia a uma distância fixa de 60cm de um monitor com tela de 24". Para cada ponto de tela, tanto na calibração quanto no teste, foram capturados cinco quadros de vídeo ao invés de apenas um. Isso foi feito porque permitiria medir a estabilidade da estimação do olhar em quadros sucessivos onde o usuário estaria olhando fixamente para o mesmo ponto.

Em cada quadro, o sistema capturava a imagem do usuário através da *webcam*, e armazenava essa imagem juntamente com a coordenada do ponto que estava sendo exibido na tela para o usuário. Essas informações foram utilizadas na fase de calibração, para definir a função de mapeamento, e na fase de teste, para avaliar os resultados do mapeamento. Uma restrição do sistema é que o usuário não poderia movimentar a cabeça durante a captura das imagens.

3.4 Construção do Módulo 1: Localização do olho na imagem

O módulo de localização do olho busca extrair as coordenadas do centro da íris na imagem. Isso é feito através de uma série de passos, que serão explicados em detalhes nas subseções seguintes.

3.4.1 Detecção e delimitação da face

O primeiro passo para realizar o rastreamento do olhar é a detecção das coordenadas dos olhos na imagem. A figura 3.4 ilustra um exemplo de imagem capturada pela *webcam*.

Figura 3.4: Exemplo de captura da imagem do usuário

Para reduzir a área de busca e diminuir o processamento necessário para a localização dos olhos, primeiro é feita a localização da face do usuário. Existem vários trabalhos na literatura abordando o assunto. MOUTINHO (2005) fez um estudo sobre a área e desenvolveu um sistema de localização de faces, obtendo bons resultados nos testes realizados. Um método bastante utilizado para localização de faces é o método proposto por VIOLA E JONES (2001b), considerado praticamente o método padrão para realizar tal procedimento (JENSEN, 2008).

O método realiza uma varredura na imagem através de uma janela de detecção, que é passada por toda a imagem para tentar detectar uma face. Essa janela é passada com tamanhos variáveis, visando detectar faces com diversos tamanhos. Para cada tamanho da janela, são computados descritores (*features*) baseados nos descritores de HAAR (1910), ilustrados na figura 3.5.

O valor de um descritor é calculado através da diferença entre a soma dos

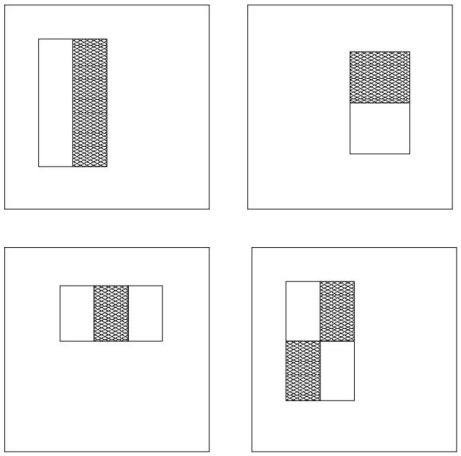


Figura 3.5: Exemplos de descritores utilizados no detector de faces de VIOLA E JONES (2001b)

píxeis abaixo da região escura e a soma dos píxeis abaixo da região clara. A operação de soma dos píxeis em uma região foi simplificada pelos autores através da criação de uma nova imagem, chamada de imagem integral. Seja i a imagem original e I a imagem integral, o valor do píxel $I(x, y)$ equivale à soma dos píxeis acima e à esquerda do mesmo, inclusive, na imagem i, conforme a equação 3.1. Isso é ilustrado na figura 3.6.

$$I(x, y) = \sum_{\substack{x' \leq x \\ y' \leq y}} i(x', y') \tag{3.1}$$

Essa imagem pode ser gerada em uma única varredura através de computações simples envolvendo somas e subtrações, varrendo a imagem linha a linha e calculando o valor do píxel $I(x, y)$ através da equação 3.2.

$$I(x, y) = i(x, y) + I(x - 1, y) + I(x, y - 1) - I(x - 1, y - 1) \tag{3.2}$$

Após a criação dessa imagem, o cálculo da soma dos píxeis num dado retângulo da imagem (necessário para a computação dos descritores) pode ser feito com

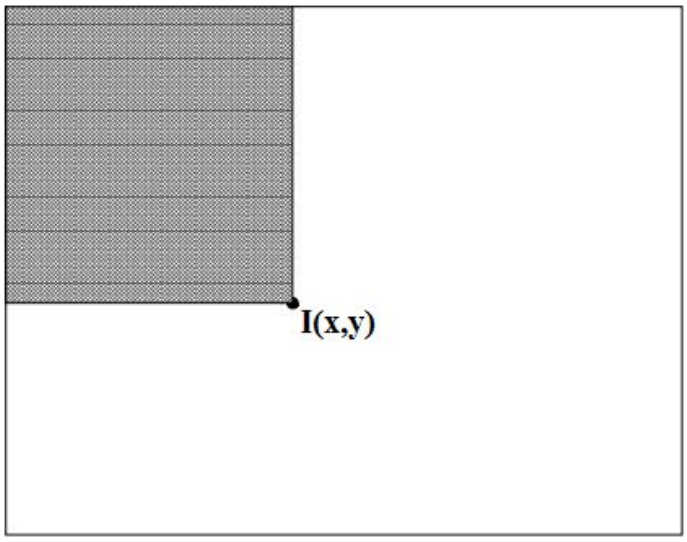


Figura 3.6: Ilustração da imagem integral, onde cada píxel $I(x, y)$ corresponde à soma dos píxeis acima e à esquerda do mesmo na imagem original

complexidade constante, através de quatro referências à imagem. Para calcular a soma dos píxeis da área D da figura 3.7 basta acessar os píxeis da imagem integral correspondente nos pontos 1, 2, 3 e 4 e calcular $P_4 - P_2 - P_3 + P_1$.

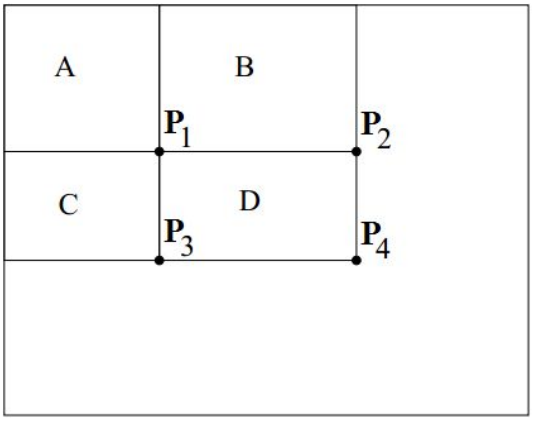


Figura 3.7: Exemplo de uso da imagem integral para o cálculo de área

Ao invés de treinar um classificador mais complexo, o método utiliza uma cascata de classificadores simples (como *perceptrons*, por exemplo), de forma a redu-

zir o processamento da localização. Cada classificador é treinado de forma a possuir uma taxa de aceitação mais alta do que a taxa de rejeição, com base em um dos descritores da janela de detecção. Dessa forma, quando um classificador aceita um descritor, o próximo descritor é passado para o classificador seguinte na cascata, e assim por diante. O processo é ilustrado na figura 3.8. Classificadores mais discriminativos são colocados no início da cascata, de forma que imagens que claramente não sejam faces sejam descartadas num estágio mais inicial do processo. Quando uma região é aceita por todos os classificadores, esta é classificada como sendo uma face. O treinamento é realizado utilizando uma versão modificada do algoritmo de classificação *Adaboost* (FREUND E SCHAPIRE, 1995).

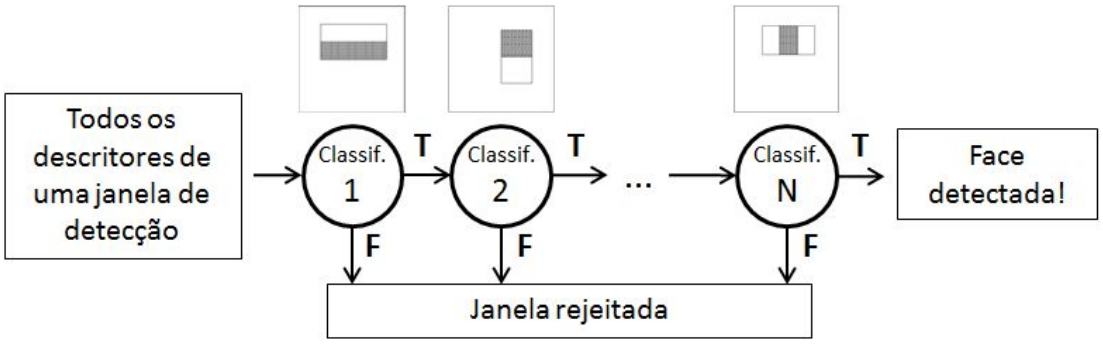


Figura 3.8: Visão geral do processo de localização de faces para uma janela de detecção

A figura 3.9 ilustra um exemplo de detecção da face ilustrada na figura 3.4, utilizando o método de VIOLA E JONES (2001b). Caso o algoritmo detecte mais de uma face, o protótipo considera a maior delas, pois o sistema foi projetado para ser utilizado por apenas um usuário, posicionado à frente da câmera. Logo, sua face deverá se sobressair a eventuais faces que venham a aparecer na imagem (ex.: pessoas ao fundo da imagem).

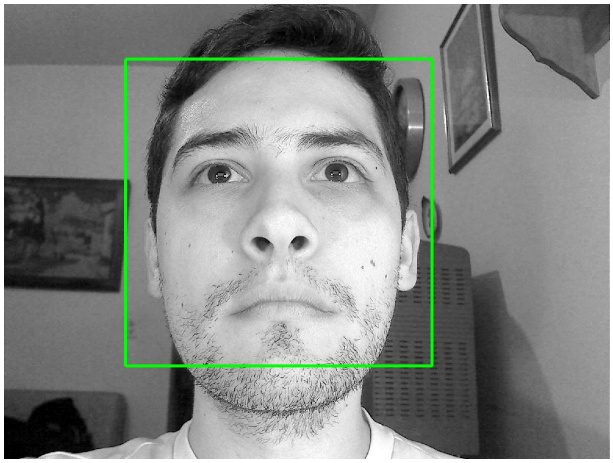


Figura 3.9: Resultado do detector de faces de VIOLA E JONES (2001b)

3.4.2 Redução da área de busca

Com base na imagem da face, faz-se necessário realizar a localização dos componentes de interesse (no caso, os dois olhos). Para reduzir o processamento realizado na imagem, pode-se tentar delimitar a área de busca desses componentes. A região retornada pelo detector de faces geralmente possui a mesma distribuição espacial interna dos componentes da face (olhos, boca, nariz, etc). Isso se deve ao fato do treinamento ter usado imagens frontais recortadas de forma similar entre si, conforme ilustra a figura 3.10.

Figura 3.10: Exemplos de faces usadas no treinamento por VIOLA E JONES (2001b)

Com base nisso, sub-regiões foram definidas de forma a reduzir ainda mais a área de busca. A figura 3.11 ilustra a divisão dessas regiões. Os valores dos percentuais foram obtidos de forma empírica a partir de observações feitas durante os experimentos. Uma abordagem similar foi realizada por CRISAFULLI; IANNIZZOTTO E LA ROSA (2009), que definiram os parâmetros da área de busca de forma empírica.

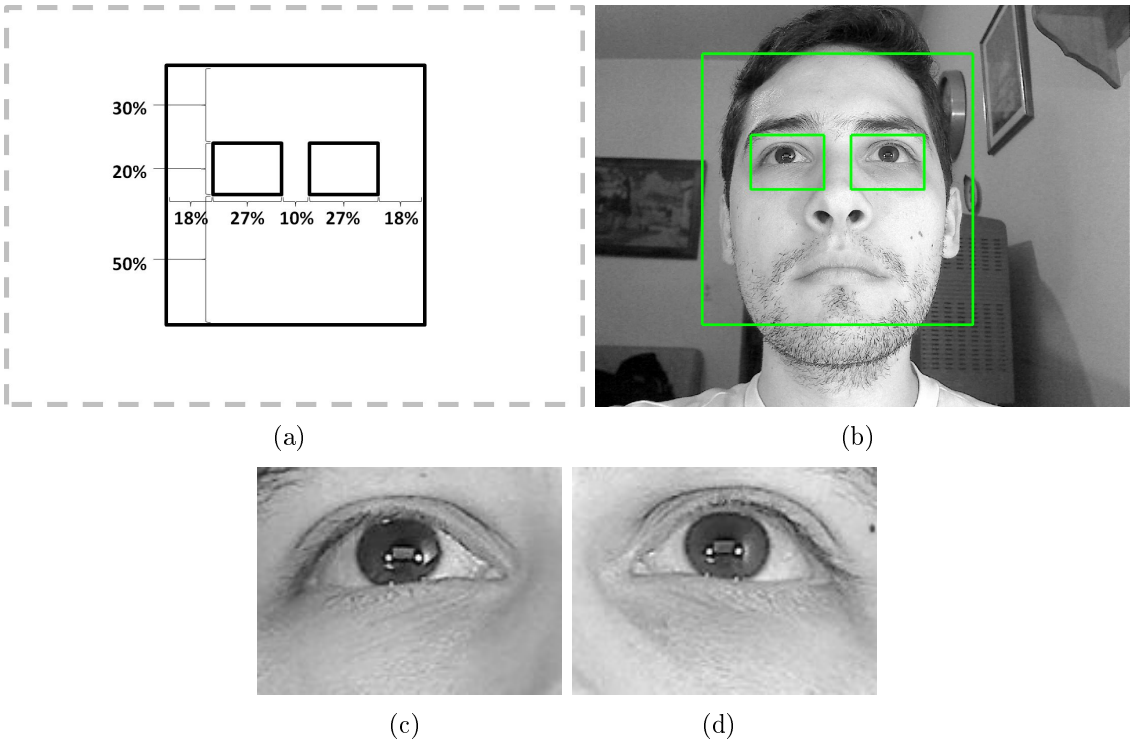


Figura 3.11: (a) Sub-regiões que contém os olhos e (b) exemplo numa imagem real, olhos (c) direito e (d) esquerdo recortados

3.4.3 Localização da íris

A localização da íris na imagem consiste em localizar com precisão as coordenadas da íris, para que seja possível realizar o rastreamento do olhar. O formato do olho humano sofre muita variação dependendo do ângulo a partir do qual a imagem

foi capturada, como ilustra a figura 3.12.

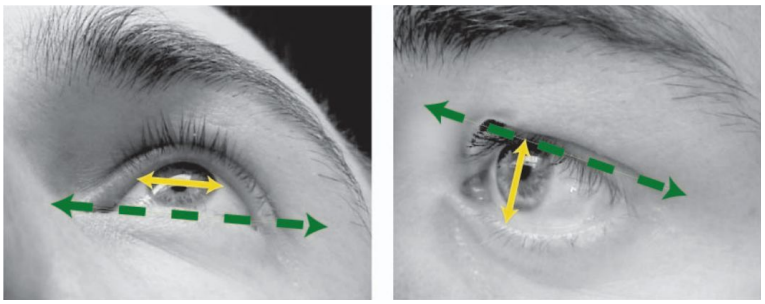


Figura 3.12: Variação do olho de acordo com ângulo de captura (HANSEN E JI, 2010)

Vários são os fatores que podem dificultar a localização da íris, como variação na iluminação, tamanho, oclusão pelas pálpebras, refletividade, etc (HANSEN E JI, 2010). O módulo 2 tenta determinar com precisão a localização da íris na imagem. Para isso, é feito um pré-processamento, visando eliminar o ruído na imagem e, em seguida, a íris é detectada através de técnicas de processamento de imagens. As subseções seguintes irão explicar com detalhes os passos utilizados na abordagem proposta.

3.4.3.1 Ajuste de luminosidade

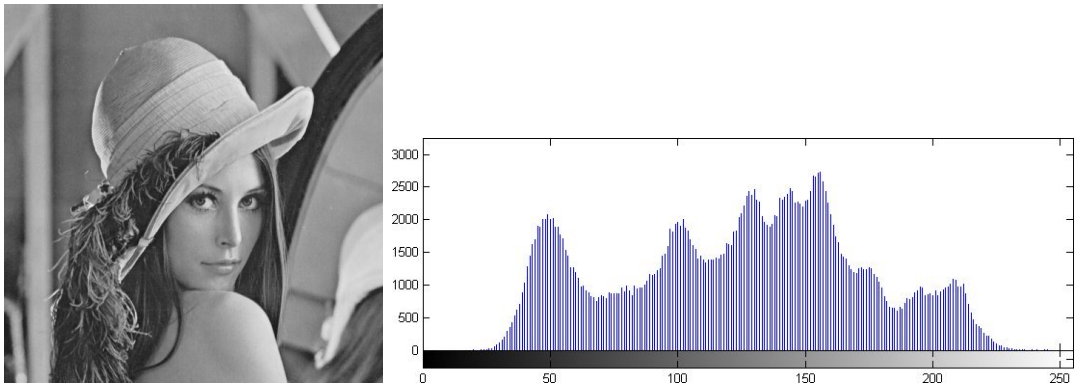
Um dos problemas encontrados no processamento da imagem do olho é a variação na iluminação. O ideal é que as imagens possuam iluminação uniforme, porém isso dificilmente acontece. Ajustes de luminosidade são feitos modificando o histograma da imagem. Seja $X = [X_{i,j}]$ uma imagem em tons de cinza com dimensões $M \times N$, onde $1 \leq i \leq M$, $1 \leq j \leq N$ e $X_{i,j}$ representa o valor da luminância do píxel nas coordenadas (i, j), que pode assumir valores discretos na faixa de $[0, 255]$ (onde o menor valor é o preto e o maior valor, o branco). O

histograma dessa imagem pode ser definido através da função

$$h(X_k) = n^k, \tag{3.3}$$

onde n^k representa o número de píxeis que possuem o nível de cinza X_k para $k = [0, 255]$.

A figura 3.13 mostra uma imagem e seu histograma correspondente. Imagens mais escuras tendem a ter uma concentração maior de píxeis à esquerda do histograma, enquanto que imagens mais claras tendem a ter uma concentração maior de píxeis à direita do histograma.



(a) (b)

Figura 3.13: (a) Exemplo de imagem e (b) histograma correspondente

Uma das operações que podem ser feitas com o histograma é chamada de Expansão de Histograma, que consiste em tentar normalizar a faixa de valores para que ela ocupem toda a extensão do histograma. A figura 3.14 ilustra a aplicação dessa técnica numa imagem. Pode-se notar que a técnica distribuiu melhor os valores dos píxeis e a imagem ficou mais nítida.

Para fins de normalizar a imagem a ser processada nesse módulo, essa técnica foi aplicada, com o objetivo de tornar a imagem da íris o mais escura possível,

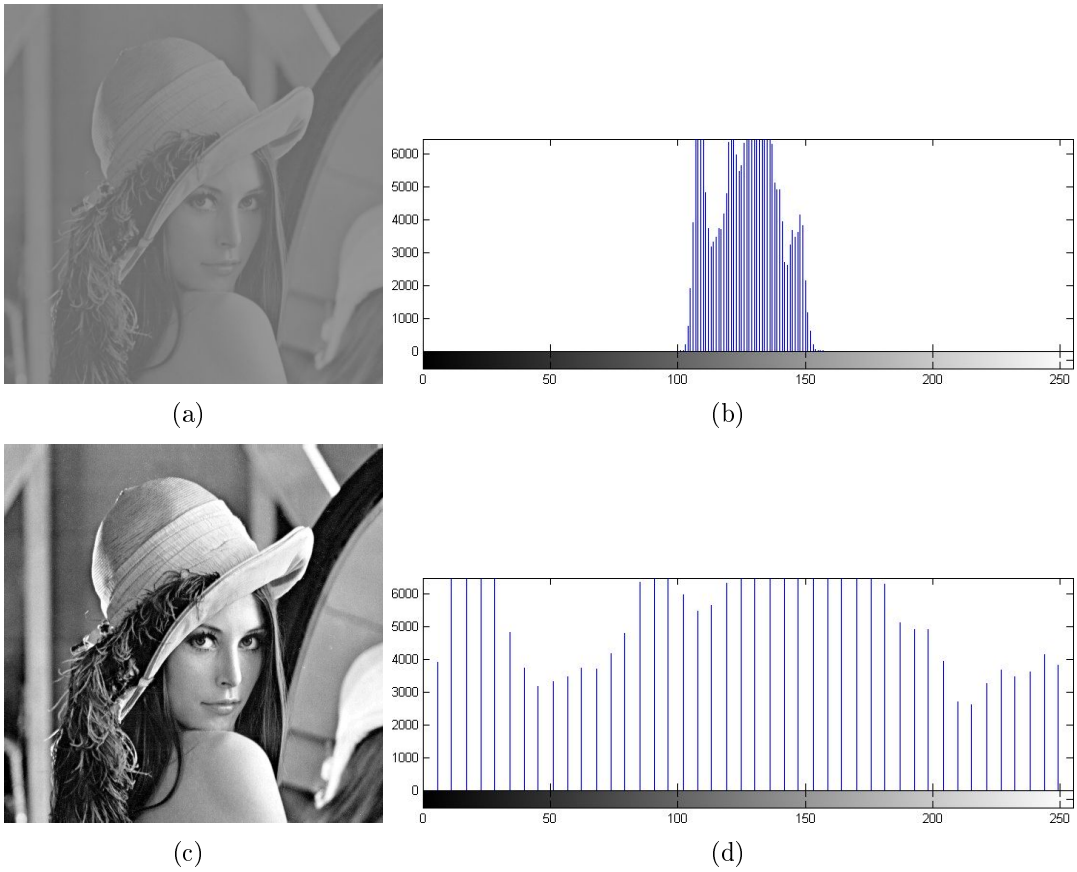


Figura 3.14: (a) Exemplo de imagem e (b) histograma correspondente, (c) resultado da expansão de histograma e (d) histograma expandido

destacando-a do restante da imagem, facilitando a sua localização. A figura 3.15 ilustra a aplicação dessa técnica em uma imagem capturada durante os experimentos.

3.4.3.2 Eliminação de ruído

A presença de ruído na imagem pode atrapalhar a localização de elementos na mesma. Ruído pode ser gerado por diversos fatores, sendo os principais o meio de aquisição e o meio de transmissão da imagem. Várias são as técnicas utilizadas

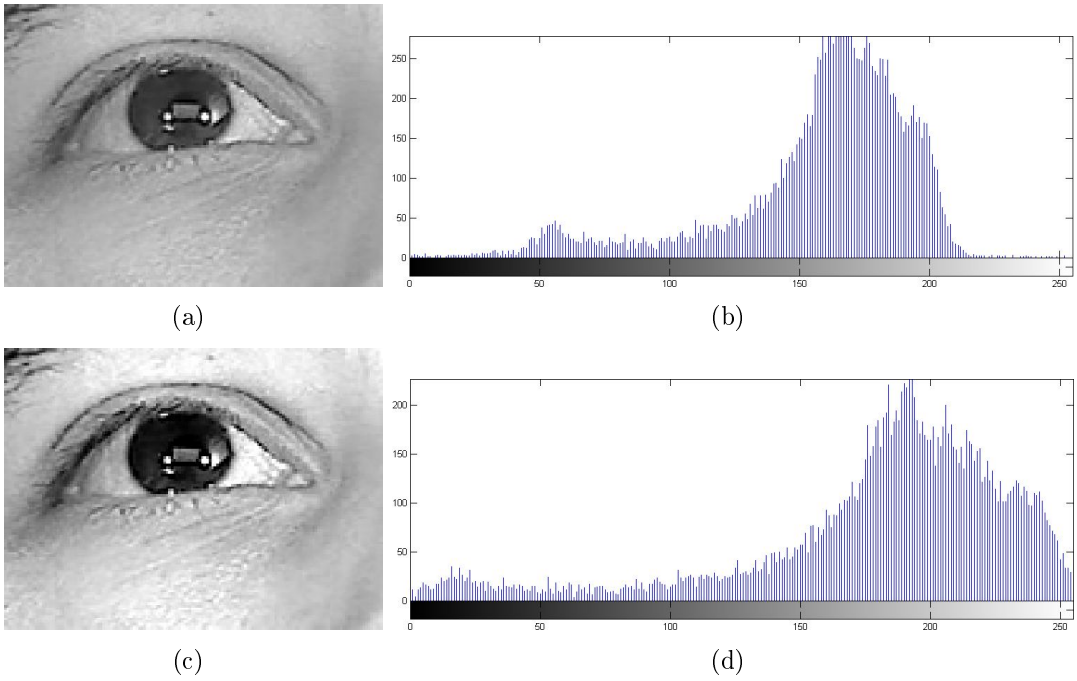


Figura 3.15: (a) Exemplo de imagem do olho e (b) histograma correspondente, (c) resultado da expansão de histograma e (d) histograma expandido

para a remoção de ruído na imagem, sendo as mais comuns o filtro de média e o filtro de gaussiana (GONZALEZ E WOODS, 2006).

O filtro de média consiste em substituir cada píxel da imagem original pela média dos píxeis vizinhos. Matematicamente isso pode ser definido como sendo a convolução da imagem original com uma máscara $N \times N$ onde o valor de cada elemento é igual a $1/N^2$. A máscara mais comumente utilizada é a máscara 3x3, definida na equação 3.4 (GONZALEZ E WOODS, 2006).

$$M = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \frac{1}{9} \tag{3.4}$$

O tamanho da máscara possui impacto direto na quantidade de suavização gerada na imagem. Uma máscara muito grande irá gerar o efeito de um borrão (forte em-

baçamento) na imagem. A convolução C da máscara M numa imagem I (denotada por $M * I$) é realizada através da equação 3.5.

$$C(x,y) = \frac{1}{LC} \sum_{l=-\infty}^{\infty} \sum_{c=-\infty}^{\infty} I(l,c)M(x-l, y-c) \tag{3.5}$$

A figura 3.16 ilustra a aplicação do filtro de média na imagem da figura 3.15(c). Pode-se observar a redução de ruído na imagem, principalmente na área da pele,

(a) (b)

Figura 3.16: (a) Imagem do olho e (b) resultado da aplicação do filtro de média

abaixo do olho. Porém, esse filtro acaba suavizando também as bordas da imagem, gerando perda de informação, pois as bordas são úteis na localização de objetos numa imagem. Um filtro similar ao filtro de média é o filtro de gaussiana. A diferença principal entre esse filtro e o filtro de média é que a máscara utilizada não possui distribuição uniforme, mas é gerada através do cálculo de uma função gaussiana, como na equação 3.6.

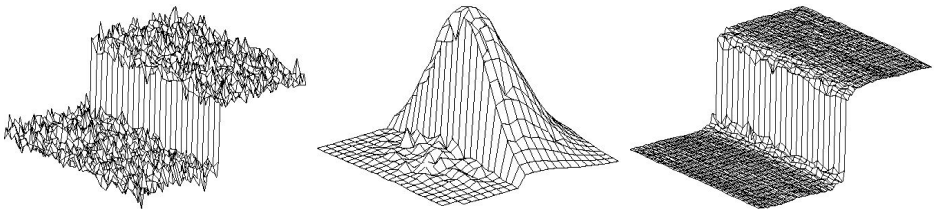
$$G(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.6}$$

A aplicação do filtro de gaussiana na imagem da figura 3.15(c) é ilustrada na figura 3.17. Houve menos redução de ruído na imagem do que no filtro de média, porém as bordas foram mais preservadas. Ainda assim, houve uma suavização das bordas, pois o filtro suaviza a imagem como um todo, não distinguindo áreas contínuas de áreas de bordas.

(a) (b)

Figura 3.17: (a) Imagem do olho e (b) resultado da aplicação do filtro de gaussiana

O desejável seria utilizar um filtro que reduzisse o ruído na imagem e preservasse as bordas. Existe um filtro que realiza essa tarefa, chamado de filtro bilateral (TOMASI E MANDUCHI, 1998). Esse filtro cria uma máscara adaptável na imagem, considerando tanto a diferença entre o valor de luminosidade do píxel central e os valores dos píxeis vizinhos e a distância entre o píxel central e os píxeis vizinhos. A figura 3.18 ilustra a aplicação do filtro bilateral numa imagem contendo uma borda com ruído. A figura 3.18(b) representa uma máscara criada pelo filtro bilateral para



(a) (b) (c)

Figura 3.18: (a) Imagem de borda com ruído, (b) exemplo de máscara criada pelo filtro bilateral para o píxel central e (c) resultado da aplicação do filtro bilateral na imagem inteira (HAMARNEH, 2002)

o píxel central da figura 3.18(a), que se situa na parte superior da representação em 3D da imagem (seria um píxel claro, de luminosidade alta). Pode-se notar que a máscara irá considerar com mais peso os píxeis da região direita da imagem, pois são mais similares ao píxel central do que os píxeis do lado esquerdo. A máscara

possui uma distribuição gaussiana, similar a do filtro de gaussiana, apresentado anteriormente, o que gera uma boa redução de ruído. Para cada píxel é criada uma máscara, que é aplicada à imagem, centrada naquele píxel, gerando o píxel resultante na imagem final. O resultado é ilustrado na figura 3.18(c). A aplicação do filtro bilateral na imagem do olho utilizada anteriormente (figura 3.15(c)) é ilustrada na figura 3.19.

(a) (b)

Figura 3.19: (a) Imagem do olho e (b) resultado da aplicação do filtro bilateral

O filtro conseguiu reduzir o ruído da imagem preservando as bordas, o que será útil na etapa posterior do trabalho, onde será feita a detecção das bordas com o objetivo de localizar a íris na imagem.

3.4.3.3 Detecção da íris

A próxima etapa do processo é a detecção das coordenadas da íris, mais precisamente do centro da mesma na imagem. Como dito na seção 3.1, neste trabalho optou-se por utilizar um modelo de formas fixas, devido à computação mais rápida dos resultados. Optou-se por um modelo circular da íris, por sua simplicidade e eficácia pois, como as imagens do usuário são capturadas frontalmente, a íris mantém o seu formato aproximadamente circular.

O primeiro passo para esse processo é a detecção das bordas da imagem. Existem várias abordagens para a detecção de bordas existentes na literatura. Dentre as mais comuns estão o filtro de SOBEL E FELDMAN (1968) e o filtro de CANNY (1986). As técnicas para detecção de bordas existentes na literatura são relativamente antigas, mas produzem bons resultados e são amplamente utilizadas até hoje. O levantamento mais recente sobre a área de detecção de bordas foi publicado por ZIOU E TABBONE em 1998. Isso também pode ser observado nos diversos livros existentes sobre processamento de imagens (JAIN, 1989; PEDRINI E SCHWARTZ, 2008; CONGI; AZEVEDO E LETA, 2008; GONZALEZ E WOODS, 2006).

O filtro de SOBEL E FELDMAN (1968) é um operador diferencial discreto baseado em uma aproximação do gradiente da imagem. Duas máscaras de tamanho $3x3$ píxeis são aplicadas na imagem através de convolução. O resultado é combinado através da equação em 3.7.

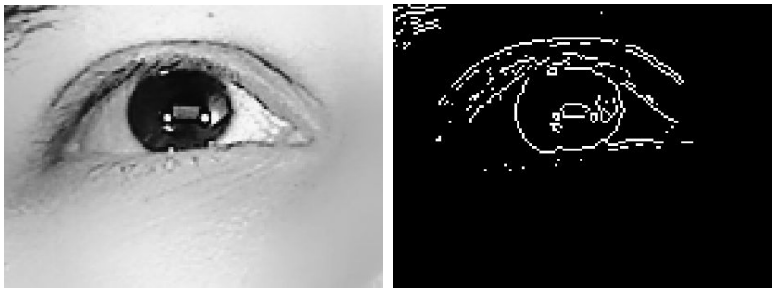
$$G = \sqrt{G_x^2 + G_y^2}, \tag{3.7}$$

onde G_x e G_y resultam, respectivamente, da aplicação da primeira e da segunda máscara à imagem, conforme equações em 3.8.

$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * I, G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * I, \tag{3.8}$$

onde I representa a imagem a ser processada. Após isso, geralmente é realizada uma binarização da imagem, de forma a gerar uma imagem em preto e branco contendo apenas as bordas, em branco. A figura 3.20 mostra o resultado da aplicação do filtro de Sobel na imagem do olho da figura 3.19(b).

O filtro de CANNY (1986) é um método que busca bordas próximas ao valor máximo do gradiente, obtendo bom desempenho e de uso bastante difundido. Ele utiliza um filtro baseado na convolução da imagem com um filtro gaussiano, como o filtro citado na seção 3.4.3.2, o que suaviza a imagem antes da detecção de bordas ser



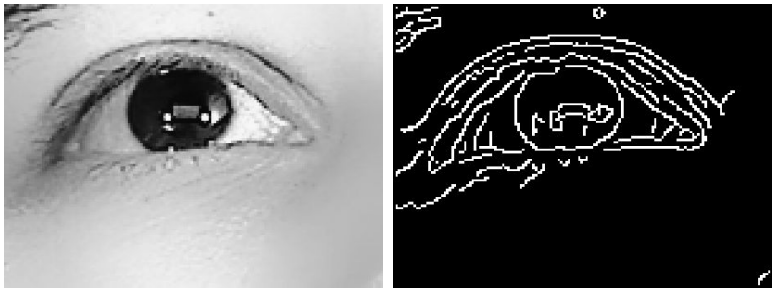
(a) (b)

Figura 3.20: (a) Imagem do olho e (b) resultado da aplicação do filtro de Sobel

realizada. Em seguida, ele aplica a equação 3.7, a mesma utilizada no filtro de Sobel, para encontrar as bordas da imagem. Depois, é computado o ângulo de direção de cada ponto da imagem de bordas. Isso pode ser feito através da equação 3.9. Por fim, algumas métricas são aplicadas na imagem resultante de forma a classificar cada píxel como borda ou não borda.

$$\Theta = arctan\left(\frac{G_y}{G_x}\right) \tag{3.9}$$

A figura 3.21 mostra o resultado da aplicação do filtro de Canny na imagem do olho da figura 3.19(b).



(a) (b)

Figura 3.21: (a) Imagem do olho e (b) resultado da aplicação do filtro de Canny

Comparando a figura 3.20(b) com a figura 3.21(b), pode-se notar que a última conseguiu mais êxito em detectar as bordas da imagem, principalmente na região da íris. Nos experimentos realizados, optou-se por utilizar o filtro de Canny, pois apresentou melhores resultados em obter as bordas de interesse.

A partir da imagem contendo as bordas, será feita a detecção da íris na imagem, que se assemelha a um círculo. A detecção de formas geométricas simples em imagens pode ser feita através da Transformada de Hough (HOUGH, 1959; DUDA E HART, 1972). A transformada original foi desenvolvida para detectar segmentos de reta, mas pode ser utilizada para detectar quaisquer formas que podem ser definidas por uma equação analítica, como um círculo, que é o caso da íris.

No caso mais simples, de detecção de linhas retas, as mesmas são representadas por coordenadas polares, da seguinte forma:

$$\rho = xcos(\theta) + ysin(\theta) \tag{3.10}$$

A figura 3.22 ilustra essa representação. A detecção das linhas se baseia na transformação da imagem para um espaço de parâmetros, denominado espaço de Hough, inicializado como uma matriz de zeros. Após a extração das bordas de uma imagem através de um método qualquer de detecção de bordas, é feita uma varredura na imagem e, para cada píxel (x, y) que esteja sobre uma borda, varia-se θ no intervalo de $[0, \pi)$ para calcular o valor de ρ, de acordo com a equação 3.10. Em cada iteração, o ponto (ρ, θ) no espaço de Hough é incrementado em uma unidade.

A detecção dos segmentos de reta consiste basicamente em detectar os picos existentes no espaço de Hough, que correspondem a retas na imagem. Para cada ponto (ρ, θ) no espaço de Hough, é feita uma varredura nos píxeis da imagem original que estão contidos nessa reta, de forma a encontrar as coordenadas dos segmentos de reta existentes. Isso é feito porque um pico pode ter sido gerado por várias retas

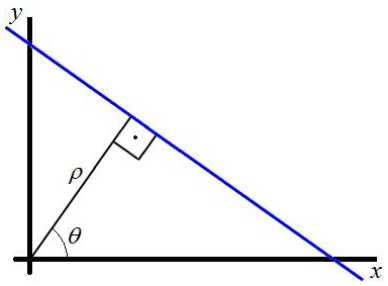


Figura 3.22: Representação de uma reta em coordenadas polares

colineares. Por fim, faz-se uma análise dos segmentos detectados, de forma a unir segmentos que estejam muito próximos e eliminar segmentos muito pequenos, de acordo com parâmetros definidos conforme cada caso.

Com relação à detecção dos picos, o método padrão consiste em encontrar o maior pico e, em seguida, com base numa vizinhança de supressão, os pontos vizinhos ao pico são marcados com zero, de forma a evitar a detecção de picos muito próximos. Em seguida, o próximo pico é detectado e o processo se repete. Normalmente define-se uma quantidade máxima de picos e um limiar que define um limite inferior para que um ponto seja considerado um pico. O procedimento é encerrado quando a quantidade máxima de picos é atingida ou quando é detectado um pico que possui valor inferior ao limiar definido. Para uma escolha ideal desses parâmetros é necessário um conhecimento prévio do domínio onde o método será aplicado. A escolha incorreta pode fazer com que algumas bordas sejam ignoradas, gerando falsos negativos, ou pode considerar ruído como se fosse uma borda, gerando falsos positivos. A figura 3.23 ilustra a aplicação da Transformada de Hough numa imagem contendo segmentos de reta.

Para a detecção de formas circulares, o método descrito é adaptado para detectar círculos com um determinado raio r. É feita uma varredura na imagem de bordas e, para cada píxel $P(x, y)$, são analisados os píxeis que estejam a uma distância r deste píxel, numa varredura circular, centrada em $P(x, y)$. Para cada

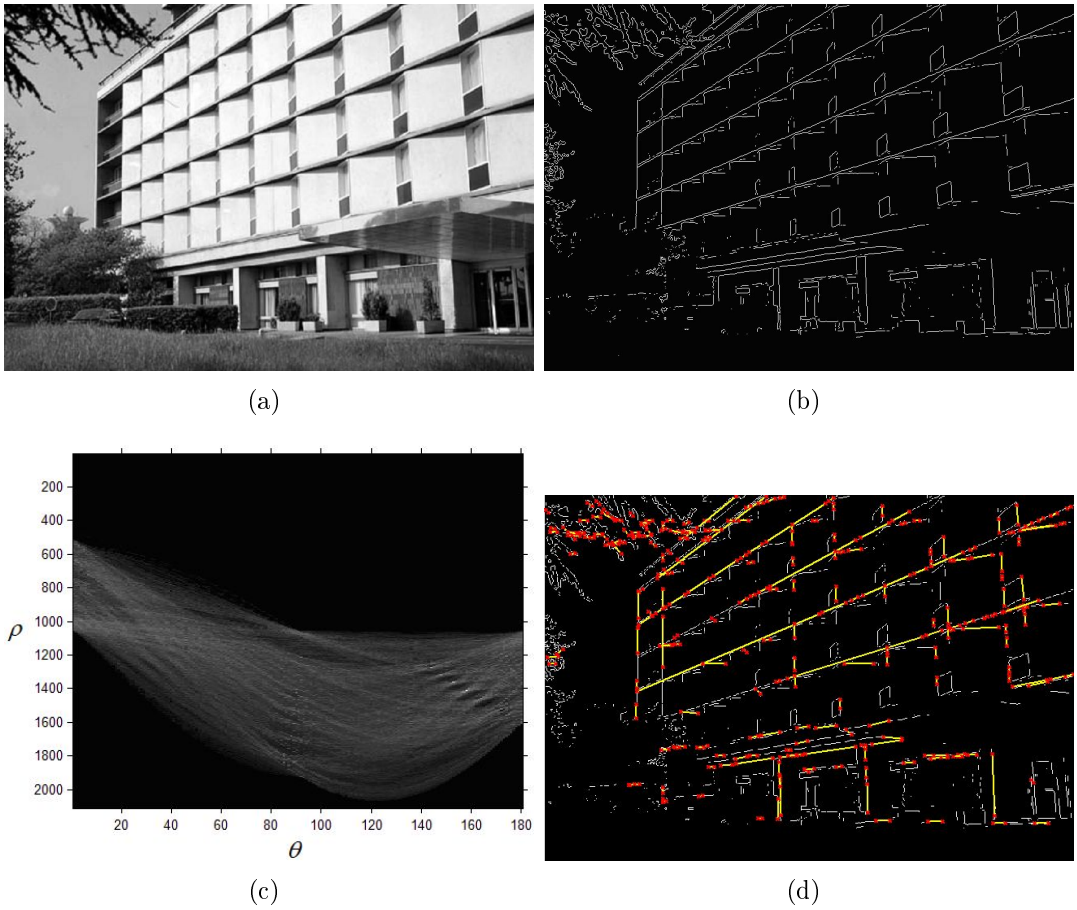
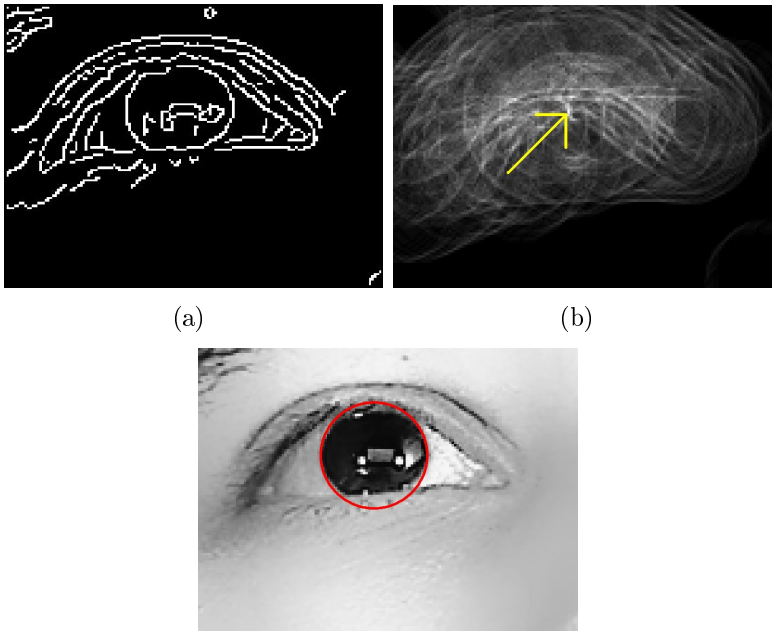


Figura 3.23: (a) Imagem original, (b) bordas detectadas pelo filtro de Sobel, (c) espaço de Hough e (d) bordas detectadas pela Transformada de Hough

um desses píxeis ao redor que represente uma borda, é feito o incremento de uma unidade no espaço de Hough nas coordenadas (x, y). Isso gera um espaço onde os picos representam pontos que correspondam ao centro de uma circunferência de raio r. Um processo similar foi realizado por CUONG E HOANG (2010) em seu sistema de rastreamento do olhar. Apesar de os mesmos não terem utilizado diretamente a Transformada de Hough, o algoritmo desenvolvido por eles realiza basicamente a mesma tarefa.

A figura 3.24 mostra o espaço de Hough gerado pela aplicação da Transfor-

mada de Hough na imagem de bordas do olho, obtida pelo filtro de Canny. Foi utilizado um raio $r = 21$ píxeis. O resultado da detecção é exibido na figura 3.24(c) através de um círculo de raio r centrado no píxel correspondente ao pico do espaço de Hough, indicado pela seta na figura 3.24(b).



(a) (b)

(c)

Figura 3.24: (a) Imagem das bordas do olho e (b) espaço de Hough gerado com raio $r = 21$ píxeis e (c) resultado da detecção na imagem original

Um problema encontrado nessa abordagem é que o raio do círculo que se deseja localizar deve ser conhecido a priori. Isso é um problema porque o tamanho da íris pode variar levemente de pessoa para pessoa, além de poder variar devido a outros fatores como a distância da câmera, distorções na lente, etc. Porém, há uma relação entre o tamanho da íris e o tamanho da face.

Como a face já foi detectada anteriormente, foi definida uma faixa de raios possíveis em torno de um raio médio, definido através de medidas extraídas a partir dos experimentos realizados. Um processo similar foi descrito por KUNKA E KOS-

TEK (2009) em seu sistema de rastreamento de olhar, onde foi definida uma faixa de valores de raios para tentar localizar o melhor raio.

O raio médio de uma íris foi definido como sendo 2% da largura da face, o que se mostrou um valor razoável nos experimentos realizados. A partir do raio médio, a faixa de raios possíveis é definida como sendo o intervalo de 80% a 120% do raio médio, com um salto de 5% entre os valores, gerando nove variações do raio médio. Esses valores foram definidos através de experimentos.

Para cada variação de raio, foi aplicada a transformada de Hough para detectar os círculos, escolhendo-se o círculo correspondente ao maior pico no espaço de Hough. Em seguida, foi necessário criar uma métrica para pontuar os círculos detectados de forma a escolher qual seria o melhor círculo, ou seja, qual círculo melhor representa a íris do usuário. Uma característica comum da íris é que ela é geralmente mais escura do que o seu redor (fundo do olho e pálpebras). Com base nessa característica, foi criada uma máscara, ilustrada na figura 3.25.

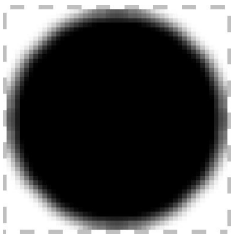


Figura 3.25: Máscara utilizada para a validação dos círculos detectados pela Transformada de Hough

Essa máscara é criada para cada variação de raio, e o raio da máscara é igual ao raio da variação. A máscara é aplicada na região da imagem correspondente à candidata à íris detectada, para cada uma das nove detecções. A aplicação da máscara é feita através do cálculo da diferença absoluta entre os píxeis da máscara e os píxeis da região candidata à íris, extraída da imagem original, contendo o olho.

É calculada a média dessa diferença absoluta e, quanto menor a média, melhor é o casamento entre a máscara e a região candidata. A região com a menor média de diferenças é escolhida como sendo a íris detectada. A figura 3.26 ilustra a aplicação desta forma de validação na imagem do olho usada nos exemplos anteriores. São exibidas as nove regiões candidatas e a pontuação atribuída de acordo com a máscara referente àquela região. No exemplo, o melhor raio seria $r = 21$, correspondente à íris da primeira imagem na segunda linha, com pontuação igual a 0.20408, sendo a menor de todas.

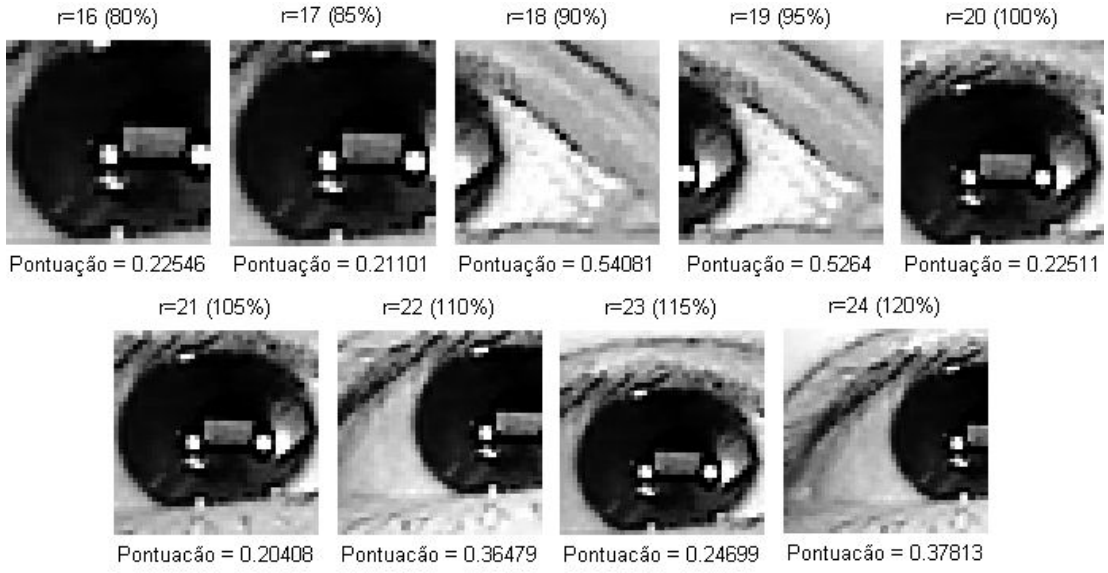


Figura 3.26: Regiões candidatas à íris e pontuação atribuída a cada região

3.4.4 Refinamento e estimação das coordenadas

Em certas situações, as coordenadas da íris não são detectadas com muita precisão. Isso ocorre principalmente quando o usuário está olhando para as regiões periféricas da tela. Nesses casos, há uma certa distorção na imagem devido ao deslocamento do globo ocular que faz com que o formato da íris seja elíptico ao

invés de circular, como ilustra a figura 3.27.

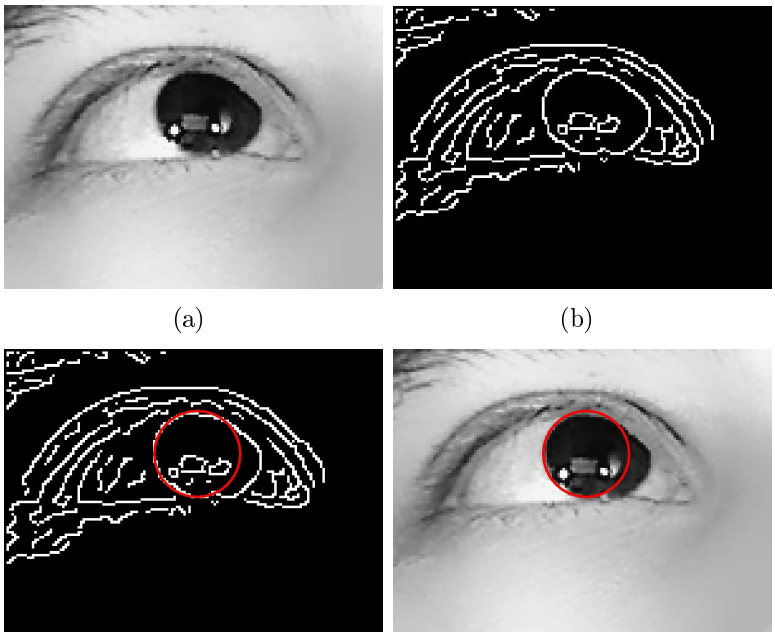


Figura 3.27: Resultado da detecção quando o formato da íris é elíptico: (a) Imagem original, (b) bordas, (c) resultado da detecção sobre as bordas e (d) sobre a imagem original

Pode-se notar na figura 3.27(c) que o melhor casamento de uma elipse na imagem de bordas foi detectado na região esquerda da íris, abrangendo a maioria dos píxeis brancos, pertencentes à borda da mesma. Logo, fez-se necessário um estudo com o objetivo de melhorar a precisão da localização inicial da íris. Apesar da falta de precisão, a localização inicial do olho apresentou bons resultados. Geralmente, o ponto central detectado pelo melhor círculo, conforme a validação realizada na seção anterior, se localiza dentro da área da íris. Como a íris tende a ser mais escura do que a região ao seu redor, como foi explicado anteriormente, foi criada uma variação da máscara utilizada para a validação das regiões candidatas, com o objetivo de refinar a localização da íris. Essa máscara está ilustrada na figura 3.28.

Figura 3.28: Máscara utilizada para o refinamento da detecção da íris

As cores da imagem foram normalizadas para que a mesma pudesse ser visualizada, mas a área cinza seriam píxeis com valor igual a zero, e as áreas em preto e branco representam píxeis cujo valor são menores ou maiores que zero, respectivamente, de forma que a soma dos píxeis dessas duas áreas seja igual a -1 e 1, e o somatório dos píxeis da máscara seja igual a zero. Isso é feito para normalizar a máscara, de forma que seja dado o mesmo peso para a íris (área de baixa intensidade) e para a região externa (área de alta intensidade). A máscara é recortada na parte superior e inferior, pois geralmente nessas áreas há muita oclusão causada pelas pálpebras. Essa máscara atua como um filtro de convolução, e é aplicada na imagem do olho com o objetivo de obter um melhor casamento no centro da íris, onde a resposta do filtro deve ser a maior possível.

Novamente há o problema do raio da íris ser desconhecido. Porém, há a informação do melhor raio obtido pela validação das regiões candidatas, feita anteriormente. Tomando como base esse raio, é definida uma nova faixa de raios ao redor deste, com tamanhos variando de 80% a 120% do melhor raio, com um salto de 10% entre os valores dos raios, gerando cinco variações do melhor raio. Para cada raio, é criada a máscara da figura 3.28 e é feita a convolução dessa máscara com a imagem do olho. Porém, agora a convolução é feita com uma região ao redor do centro da íris, detectado previamente. Dessa forma, o pico da convolução representa o melhor casamento da máscara de raio r com a região da íris, podendo refinar a detecção visto que pode haver um deslocamento da máscara no interior da imagem. Esse processo é feito para cada um dos cinco raios definidos na faixa e o raio que

obtiver a melhor resposta (maior valor do pico da convolução) é escolhido como o melhor candidato. As coordenadas desse raio representam o centro da íris.

O refinamento da detecção da íris na figura 3.27 é ilustrado na figura 3.29. Na primeira linha da imagem estão as cinco regiões candidatas, recortadas de acordo com os raios da faixa de raios previamente definidas, centradas na localização inicial do olho, realizada anteriormente. Na segunda linha estão os resultados das convoluções das máscaras com cada imagem, e as setas indicam as coordenadas do pico da convolução. Na terceira linha é apresentado o resultado da detecção, onde cada círculo está centrado na coordenada do pico da convolução. No caso, foi selecionada a imagem da quarta coluna, que possui maior pontuação, igual a 0.46048.

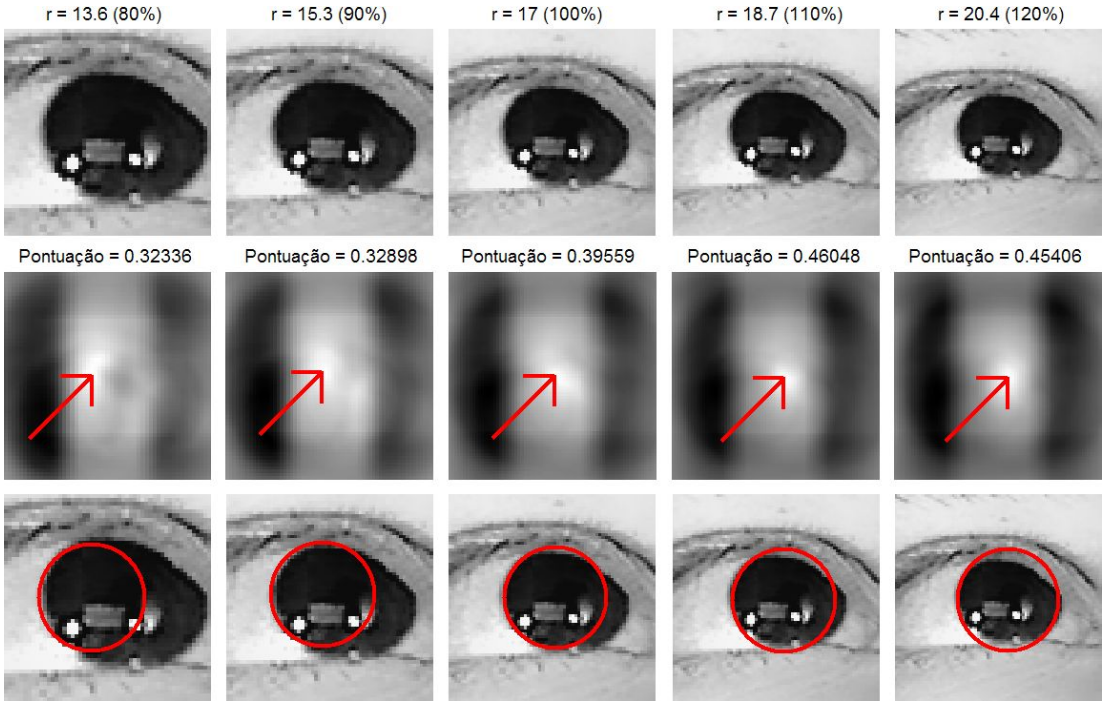


Figura 3.29: Refinamento da detecção da íris da figura 3.27.

3.5 Construção do Módulo 2: Detecção do ponto fixo

A principal informação necessária para a estimação da direção do olhar é o movimento dos olhos do usuário. Porém, apesar de uma das restrições do sistema ser que o usuário mantenha a cabeça imóvel, pode ocorrer algum movimento involuntário. As coordenadas do olho, detectadas no Módulo 1, informam a posição do centro da íris na imagem. Para medir o deslocamento, é necessário ter um ponto de referência, fixo na face do usuário. A partir daí é feito o cálculo da diferença entre as coordenadas dos olhos e desse ponto fixo, o que permite mensurar o deslocamento dos olhos com relação a esse ponto.

3.5.1 Escolha e localização

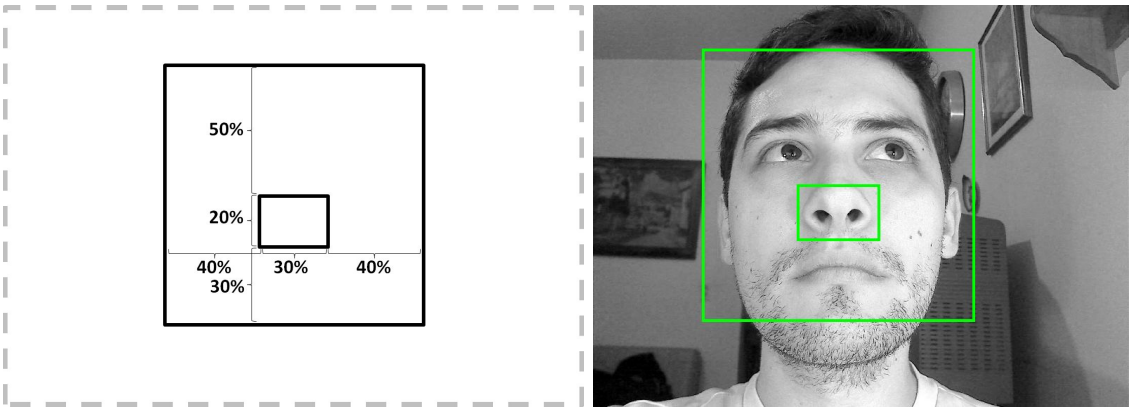
O primeiro passo para a detecção do ponto fixo é escolher qual ponto fixo será utilizado. Um ponto fixo é uma região da face que será detectada no primeiro quadro de vídeo e rastreada nos demais quadros. Um ponto fixo ideal deve ser:

1. Universal: todos os usuários devem possuir o ponto fixo na face.
2. Imutável: deve permanecer visível em todos os quadros de vídeo, sem sofrer alterações, para que possa ser rastreado.
3. Estático: sua posição deve permanecer fixa com relação aos demais componentes da face, principalmente aos olhos.

Foram feitos experimentos com alguns pontos fixos. O primeiro ponto fixo escolhido foi o nariz do usuário, por atender os requisitos anteriores. É universal porque está presente em todas as faces. É imutável pois possui uma estrutura

relativamente fixa na face do usuário, e sua posição se altera muito pouco quando a cabeça faz algum movimento. É razoavelmente estático, pois a posição é fixa com relação aos olhos. TUNHUA et al. (2010) utilizaram o nariz como ponto fixo em seu sistema de rastreamento do olhar obtendo bons resultados.

A detecção do nariz foi feita de forma similar à redução da área de busca, descrita na seção 3.4.2. Estabeleceu-se uma região fixa, ilustrada na figura 3.30.



(a) (b)

Figura 3.30: (a) Sub-região que contem o nariz e (b) exemplo numa imagem real

Em alguns casos essa abordagem não funciona muito bem, pois a posição do nariz pode variar, principalmente devido ao ângulo da câmera. A figura 3.31 ilustra uma situação onde o nariz não foi muito bem enquadrado pela região. Pode-se notar que a região detectada enquadra também um pedaço do lábio superior do usuário.

Para contornar esse problema, foi feita uma análise da área ao redor da região padrão definida na máscara de forma a enquadrar melhor o nariz do usuário. Geralmente o nariz encontra-se horizontalmente no centro da região retornada pelo detector de faces. Logo, apenas o deslocamento vertical foi ajustado nessa análise. Como a região detectada será rastreada posteriormente, ela deve possuir uma boa diversidade de informações que permitam que ela seja rastreada. Um bom indicador

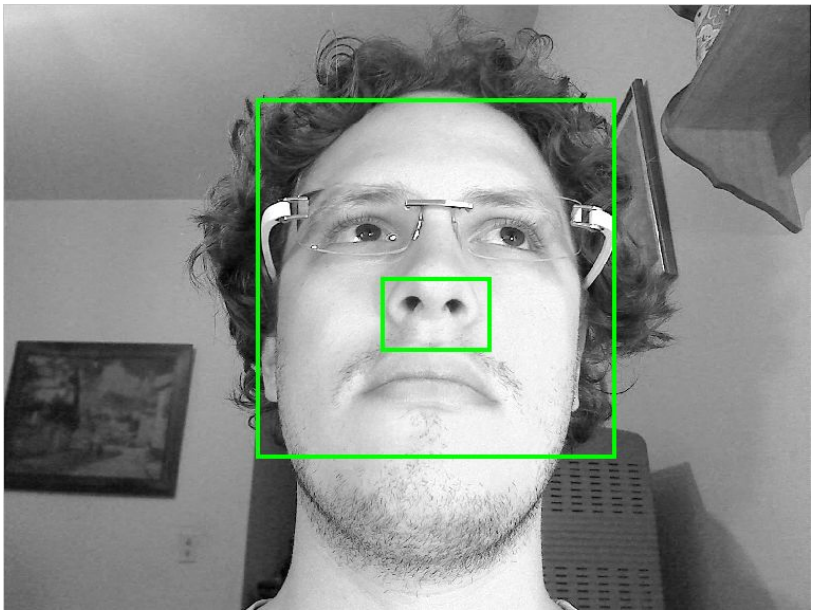


Figura 3.31: Exemplo de enquadramento incorreto do nariz

disso é o desvio padrão. Quanto maior for o desvio padrão de uma região, maior a variação dos píxeis contidos na mesma. Uma região com baixo desvio padrão tende a gerar resultados ruins quando a mesma é rastreada (ex.: um trecho da face contendo apenas a pele, de forma plana).

Foi definida uma extensão das margens superior e inferior com base nas coordenadas da região da máscara, para que seja feita uma busca por uma melhor região. Essa extensão abrange uma área de 25% da altura da região fixa, tanto para cima quanto para baixo, conforme ilustra a figura 3.32. As linhas na horizontal delimitam os limites superior e inferior da região de busca. Em seguida, essa região é varrida de cima para baixo, considerando todas as sub-regiões possíveis com as mesmas dimensões da região fixa original, mas variando a altura, a partir do topo, com saltos de um píxel. Para cada região, é calculado o desvio padrão dos píxeis nela contidos, como ilustra o gráfico da figura 3.33. É selecionada a região que contém o maior desvio padrão, indicada pela seta na imagem, cujo eixo $x = 15$, o que significa que a

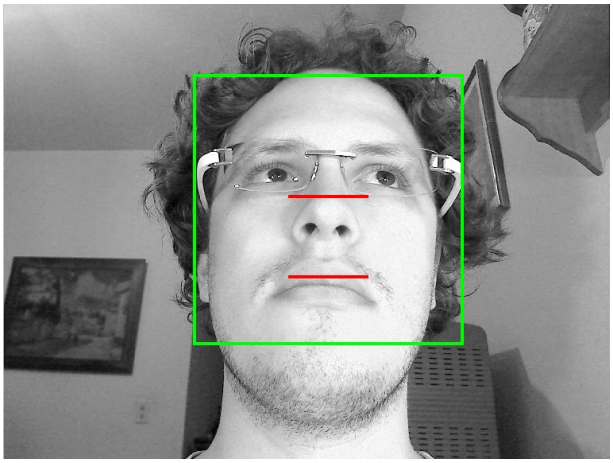


Figura 3.32: Limites superior e inferior para o ajuste da detecção da região do nariz

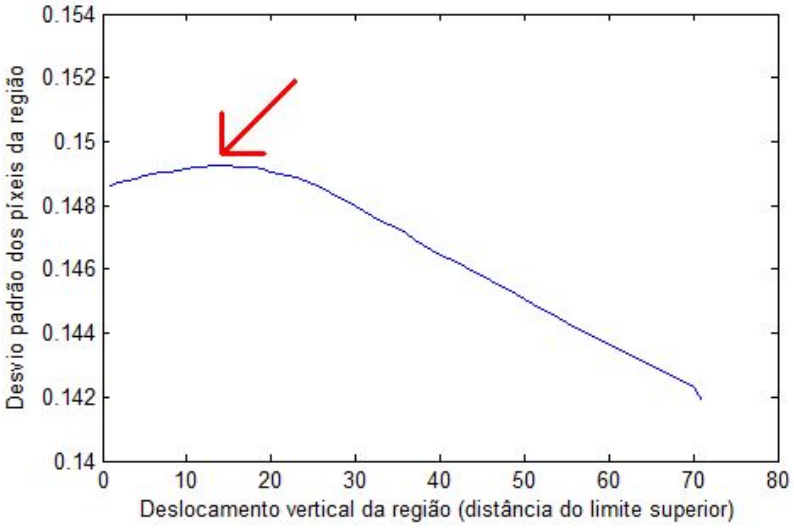


Figura 3.33: Desvio padrão das regiões entre as linhas superior e inferior da figura 3.32

região selecionada será a região cujo topo está 15 píxeis abaixo da linha superior que delimita a área de busca. A figura 3.34 ilustra a detecção resultante da aplicação desse processo na imagem da figura 3.31.

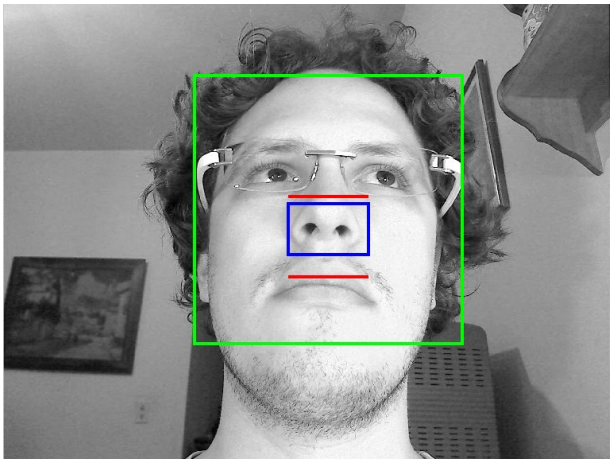


Figura 3.34: Enquadramento ajustado do nariz

Para fins de comparação, outras regiões foram utilizadas nos experimentos. Uma região que apresentou bons resultados foi a região acima dos olhos, compreendendo as sobrancelhas, como ilustra a figura 3.35. O problema desse ponto é que o usuário não poderá mover as sobrancelhas durante o uso do sistema. Além disso, muitas pessoas não possuem o centro das sobrancelhas bem definido, o que pode implicar num ponto fixo com poucas informações para ser rastreado.

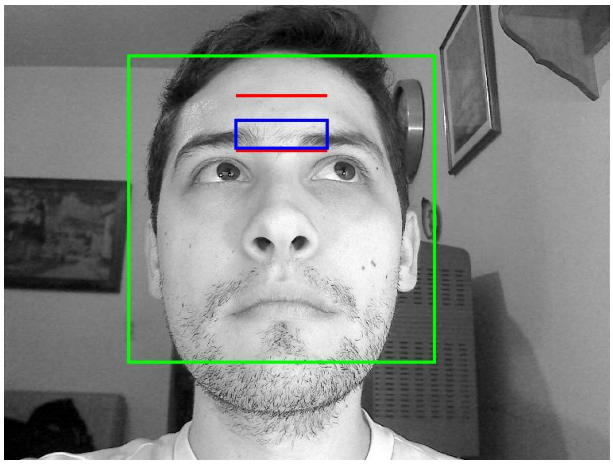
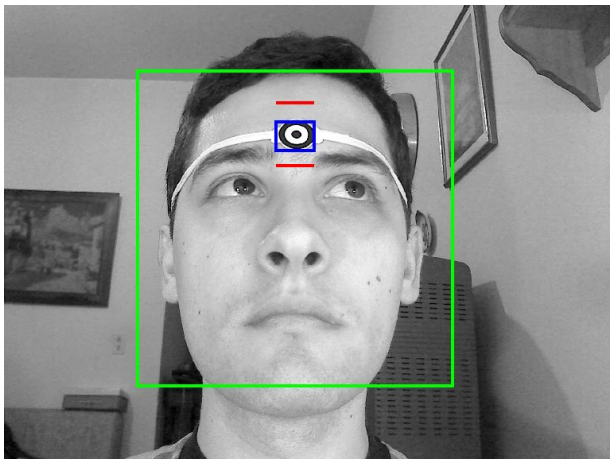


Figura 3.35: Uso das sobrancelhas como ponto fixo

Também foram realizados experimentos posicionando uma marcação artificial na face do usuário. Isso foi feito apenas para fins comparativos, pois esperava-se desenvolver o protótipo sem a necessidade de artifícios externos. Foi utilizada uma marcação no formato de um alvo, com círculos concêntricos alternando entre o preto e o branco. Dessa forma, o ponto seria bem destacado do fundo da imagem e mais facilmente rastreável. Um exemplo do uso desse marcador é exibido na figura 3.36.

(a)



(b)

Figura 3.36: (a) Uso de um marcador artificial como ponto fixo e (b) sua detecção

3.5.2 Rastreamento

No primeiro quadro de vídeo, o ponto fixo é detectado da forma descrita na seção anterior. Nos quadros subsequentes, ele é rastreado, com base na imagem extraída no primeiro quadro. Como a cabeça do usuário está numa posição relativamente fixa, não há a necessidade de localizar o ponto fixo na imagem inteira. Para reduzir o processamento, foi definida uma área de busca um pouco maior que a área original onde o ponto fixo foi detectado. Então, é feita uma busca nessa região de forma a tentar encontrar a imagem do ponto fixo, salva anteriormente. A figura 3.37 ilustra a área de busca utilizada para a localização do nariz na imagem. Áreas similares foram definidas para a busca dos outros pontos fixos (sobrancelhas e marcação artificial).

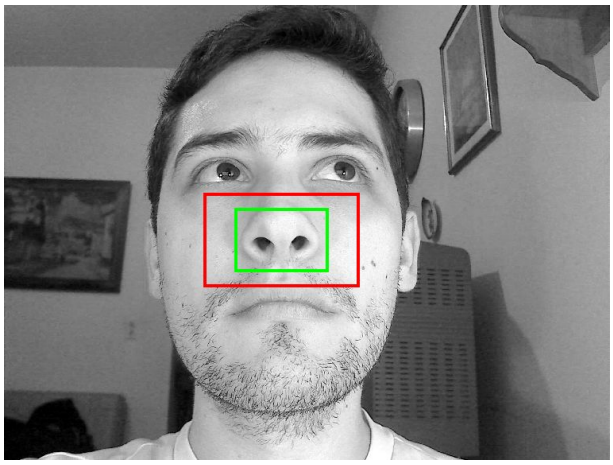


Figura 3.37: Área de busca utilizada para o rastreamento do nariz (em vermelho)

Uma abordagem muito utilizada para fazer o rastreamento da imagem é a técnica de *Phase Correlation* (ZITOVA E FLUSSER, 2003). Essa técnica é similar à convolução, demonstrada na seção 3.4.3.2, porém atua no domínio da frequência e é menos sensível a ruído e a variações na iluminação. Ela computa o espectro de potência cruzada da imagem de busca e da imagem de referência através da

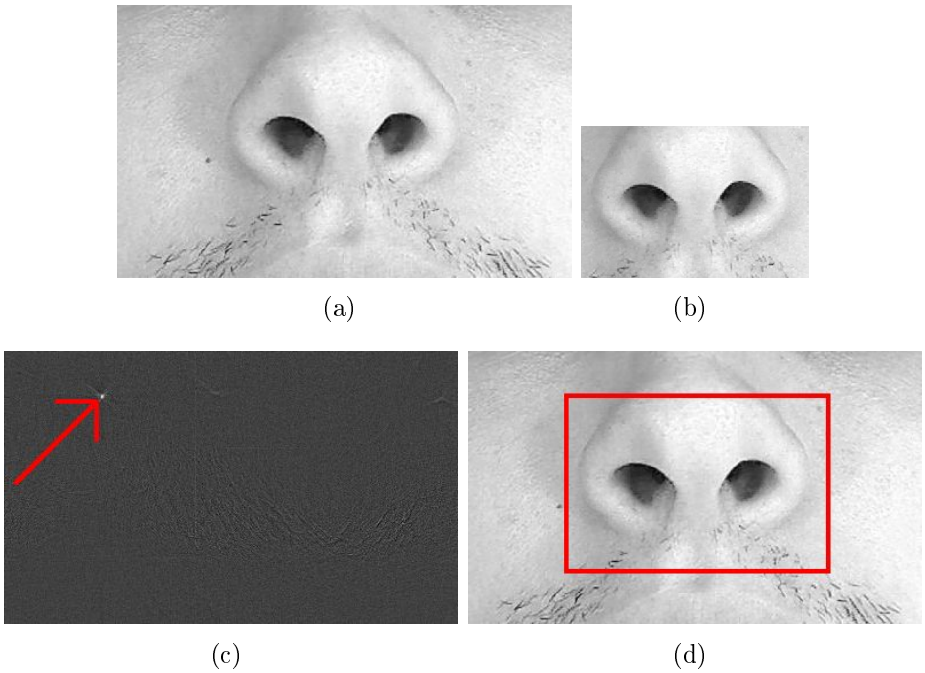
equação 3.11, retornando uma função cujo pico representa a localização espacial das coordenadas superior esquerda da imagem de referência na imagem de busca.

$$c = F^{-1}\left\{\frac{F\{G_a\}F\{G_b^*\}}{|F\{G_a\}F\{G_b^*\}|}\right\}, \tag{3.11}$$

onde G_a e G_b são a imagem de busca e a imagem de referência, respectivamente. As coordenadas da localização da imagem de referência na imagem de busca são feitas encontrando-se o pico do resultado, em c, através da equação 3.12.

$$(\Delta x, \Delta y) = \arg\max_{(x,y)}\{c\} \tag{3.12}$$

As figuras 3.38 e 3.39 ilustram, respectivamente, o resultado do rastreamento do nariz do usuário e da marcação artificial.



(a) (b)

(c) (d)

Figura 3.38: (a) Imagem da área de busca, (b) imagem do ponto fixo a ser buscado, (c) função retornada pelo *Phase Correlation* e (d) resultado do rastreamento

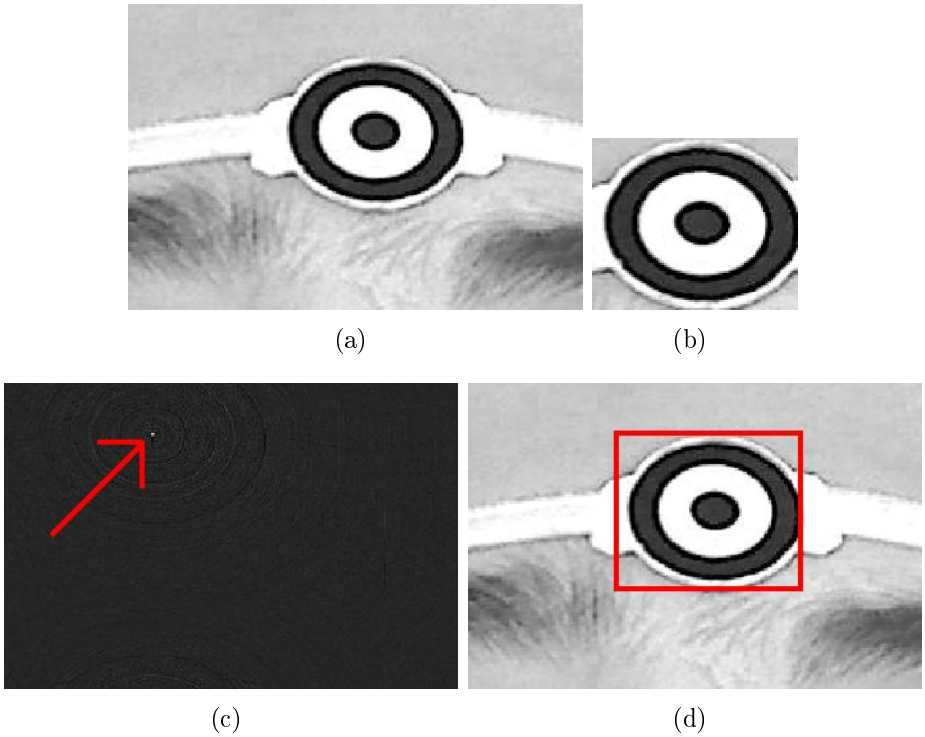


Figura 3.39: (a) Imagem da área de busca, (b) imagem do ponto fixo a ser buscado, (c) função retornada pelo *Phase Correlation* e (d) resultado do rastreamento

3.6 Construção do Módulo 3: Estimação da direção do olhar

A estimação da direção do olhar consiste em determinar o ponto na tela para onde o usuário está olhando. Ela recebe informações sobre as coordenadas dos olhos do usuário e do ponto fixo e calcula a diferença entre elas, gerando a informação sobre o deslocamento dos olhos com relação ao ponto fixo, através da equação

$$(\Delta x, \Delta y) = (O_x - F_x, O_y - F_y), \tag{3.13}$$

onde (O_x, O_y) representam as coordenadas dos olhos e (F_x, F_y) representam as coordenadas do ponto fixo. No caso das coordenadas dos olhos, como são detectados os dois olhos, considera-se a média entre as coordenadas de cada olho. O processo começa com uma fase de calibração para, em seguida, ser feita a estimação do olhar.

3.6.1 Processo de calibração

Para que o sistema possa estimar o ponto na tela para onde o usuário está olhando, é necessário realizar um processo de calibração. O processo consiste em exibir uma série de pontos para o usuário e pedir que o mesmo olhe para esses pontos, e o sistema captura a imagem do usuário enquanto este olha para cada um dos pontos. Neste trabalho foi utilizada uma grade com 3x3 pontos, exibida anteriormente na figura 3.2(a).

Foi desenvolvido um programa para capturar as imagens do usuário. Para cada ponto, o programa captura cinco imagens através da *webcam*, para se ter uma noção da estabilidade do sistema quando o usuário olha fixamente para o mesmo ponto. O sistema também armazena as coordenadas do ponto para onde o usuário está olhando, representado por (T_x, T_y). O próximo passo é definir uma função que mapeie as coordenadas do deslocamento do olho $(\Delta x, \Delta y)$ para as coordenadas de tela (T_x, T_y) para onde o usuário está olhando. Para isso, foi treinada uma Rede Neural Artificial (RNA).

RNAs são modelos matemáticos inspirados na organização do cérebro, funcionando através da interconexão de neurônios artificiais chamados de *Perceptrons*. Seu uso é feito em situações onde existam dados de entrada e dados de saída, de forma que seja necessário construir uma função que mapeie as entradas nas saídas desejadas. Isso é feito através de um algoritmo de treinamento, que ajusta os pesos da rede para que a mesma desempenhe o papel da função desejada (BRAGA; CARVALHO E LUDEMIR, 1998; HAYKIN, 1999). O uso de RNAs no processo de calibração de sistemas de rastreamento do olhar é bastante eficaz, como demonstrado por JI E ZHU (2002), que obtiveram bons resultados usando essa técnica.

Com base nas informações de treinamento, compostas por nove pontos de ca-

libração e cinco imagens capturadas por ponto, o conjunto de treinamento da RNA consistiu de 45 pontos. A figura 3.40 ilustra um exemplo de dados de treinamento. Os valores do primeiro gráfico representam a diferença em píxeis entre as coordenadas dos olhos e o ponto fixo, ou seja, as coordenadas do deslocamento do olho. Os valores do segundo gráfico representam as coordenadas de tela, normalizadas para o intervalo entre zero e um. As coordenadas do deslocamento também foram normalizadas posteriormente, para o treinamento da RNA. A tabela 3.1 apresenta alguns dos dados da figura 3.40.

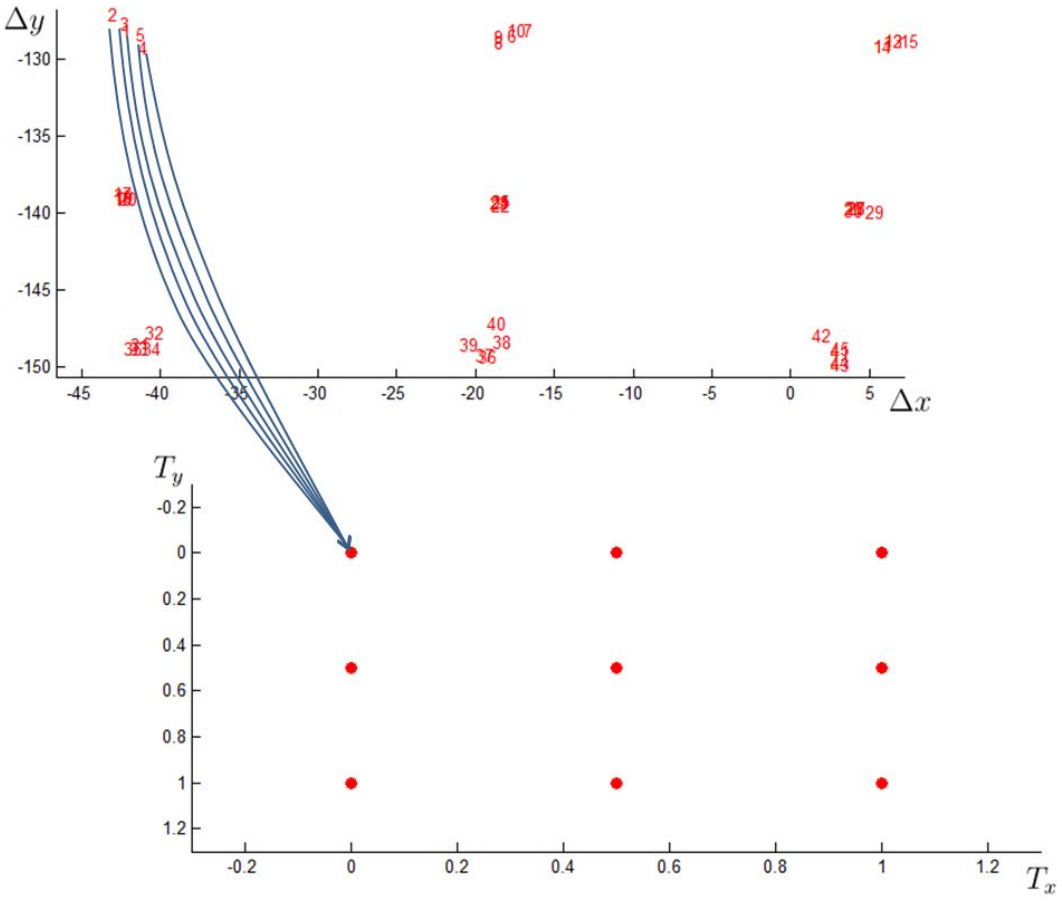


Figura 3.40: Ilustração dos dados de treinamento da RNA, que deverá mapear as entradas (coordenadas do deslocamento do olho $(\Delta x, \Delta y)$) para as saídas (coordenadas de tela (T_x, T_y))

Tabela 3.1: Exemplo de conjunto de treinamento para a RNA, ilustrado na figura 3.40

Amostra	Entradas: $(\Delta x, \Delta y)$	Saídas: (T_x, T_y)
1	(-12.300, -132.300)	(0.00, 0.00)
2	(-12.440, -132.290)	(0.00, 0.00)
3	(-12.300, -132.800)	(0.00, 0.00)
4	(-12.300, -132.300)	(0.00, 0.00)
5	(-11.800, -132.290)	(0.00, 0.00)
6	(5.840, -132.720)	(0.50, 0.00)
7	(6.520, -132.720)	(0.50, 0.00)
8	(6.160, -131.970)	(0.50, 0.00)
9	(7.245, -131.970)	(0.50, 0.00)
10	(5.820, -132.740)	(0.50, 0.00)
11	(26.650, -133.005)	(1.00, 0.00)
12	(27.880, -132.560)	(1.00, 0.00)
...	...	...
45	(25.495, -147.735)	(1.00, 1.00)

Para cada usuário, foi construída uma RNA para mapear as entradas nas saídas, gerando assim a função que irá fazer a estimativa da direção do olhar daquele usuário. Devido à simplicidade dos dados, optou-se por um modelo simples com apenas uma camada de *Perceptrons* com função de transferência linear, ilustrado na figura 3.41.

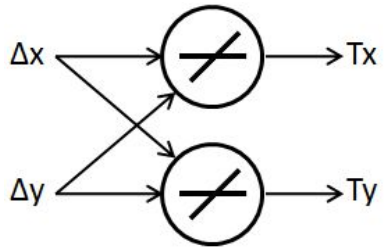


Figura 3.41: Arquitetura da RNA

Também foram realizados testes utilizando uma rede com função de trans-

ferência não linear no estilo de curvatura-S (como a tangente hiperbólica). Porém, observou-se que a curva acabava casando com os dados de treinamento mas gerando distorções que faziam com que a mesma não generalizasse os dados de teste. Como em cada dimensão havia a variação de apenas três pontos no treinamento (devido à grade 3x3 utilizada), a função não tinha informações sobre os pontos intermediários e poderia gerar interpolações incorretas dos mesmos. Isso é ilustrado na figura 3.42.

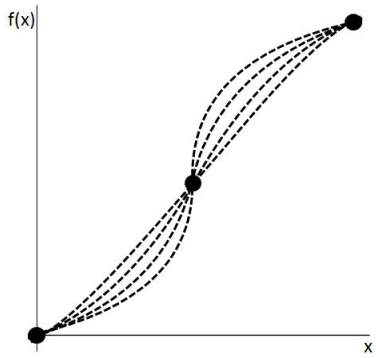


Figura 3.42: Exemplos de possibilidades de má interpolação causados pelo uso de uma RNA com função de transferência não linear

3.6.2 Estimação

Após a calibração, a estimação do olhar pode ser feita passando as coordenadas do deslocamento do olho $(\Delta x, \Delta y)$ dos dados de teste para a RNA treinada, que irá retornar as coordenadas estimadas da direção do olhar (D_x, D_y). Para testar a estimação, foi utilizada uma grade com 5x5 pontos, exibida anteriormente na figura 3.2(b). Para cada ponto, também foram capturadas cinco imagens, num total de 125 capturas. A figura 3.43 ilustra um exemplo de dados de teste.

Após a apresentação dos dados de teste à rede, as saídas obtidas representam a estimativa da direção do olhar do usuário. Um exemplo de resultado é exibido na figura 3.44. Os círculos representam os 25 pontos de teste e as cruzes representam

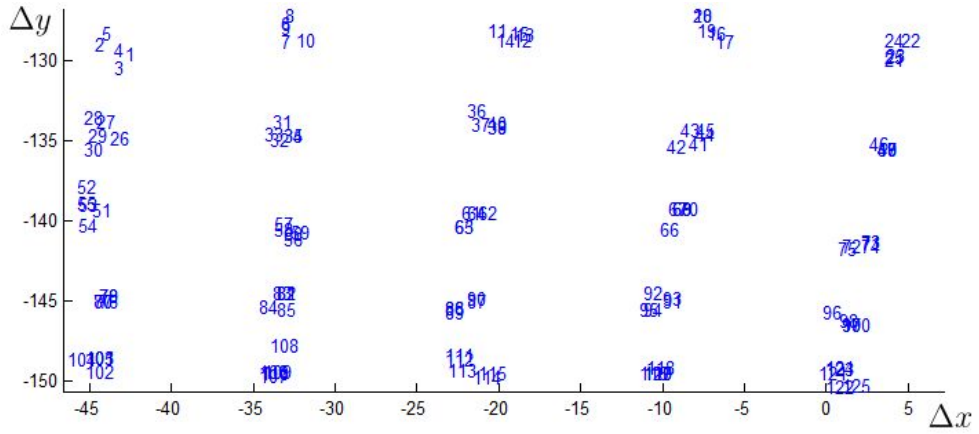


Figura 3.43: Ilustração dos dados de teste da RNA

a estimativa do ponto para onde o usuário estava olhando, e estão ligadas para o ponto original exibido na tela no momento da captura.

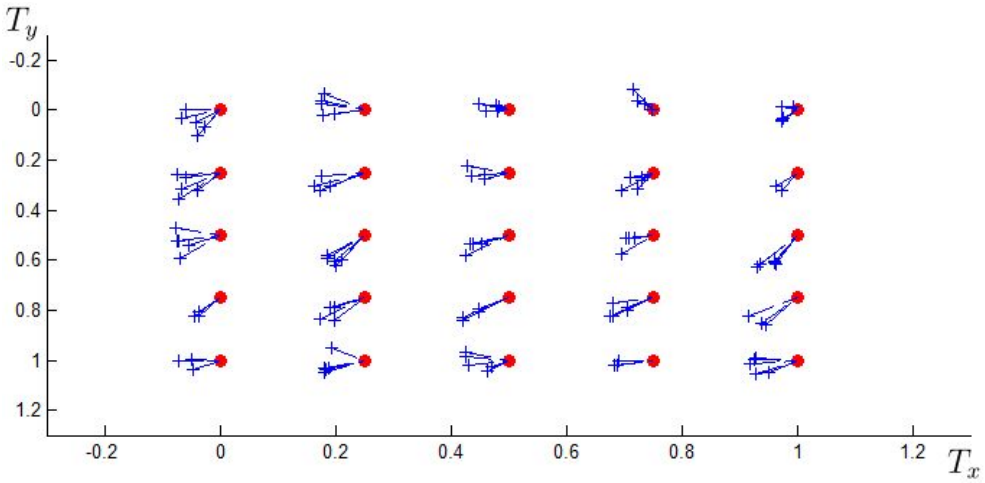


Figura 3.44: Resultado da estimação do olhar do usuário

Apenas para ilustrar o que discutido na seção 3.6.1 sobre o uso de saídas lineares na RNA, a figura 3.45 ilustra a estimativa da direção do olhar para mesmos dados utilizados no exemplo anterior, mas utilizando funções de transferência não lineares para treinar a RNA. Observa-se um comportamento similar ao discutido

sobre a figura 3.42, onde há erros de interpolação dos pontos que a rede desconhecia.

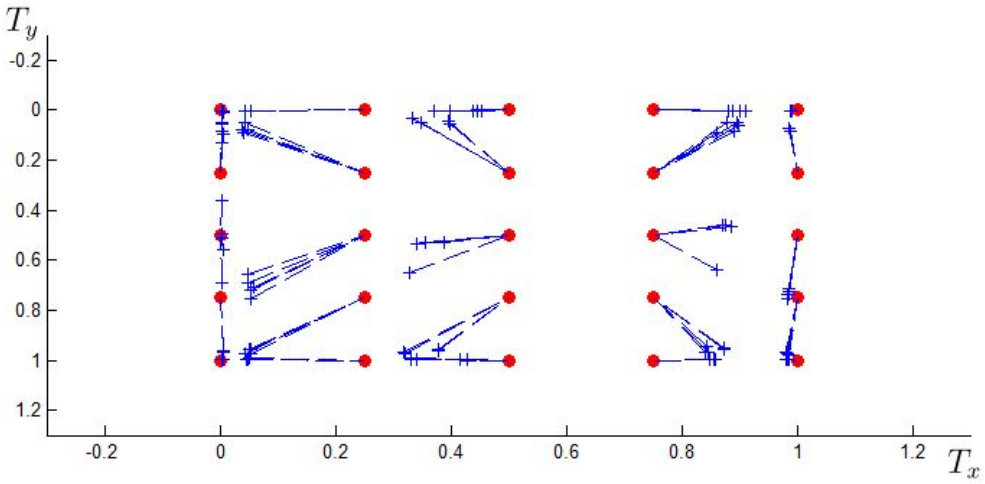


Figura 3.45: Resultado da estimação do olhar do usuário utilizando uma RNA com função de transferência não linear

Durante os experimentos, foi levantada a hipótese de se treinar apenas uma RNA para todos os usuários, ao invés de treinar uma RNA para cada um. Porém, a distribuição espacial dos dados de entrada é extremamente variável, havendo muitas interseções entre os dados de diferentes usuários e de diferentes pontos fixos. A figura 3.46 ilustra a distribuição espacial dos dados de entrada de dois usuários e três pontos fixos utilizados. Devido a esse problema, o treinamento de uma única rede para todos os usuários ficaria inviável.

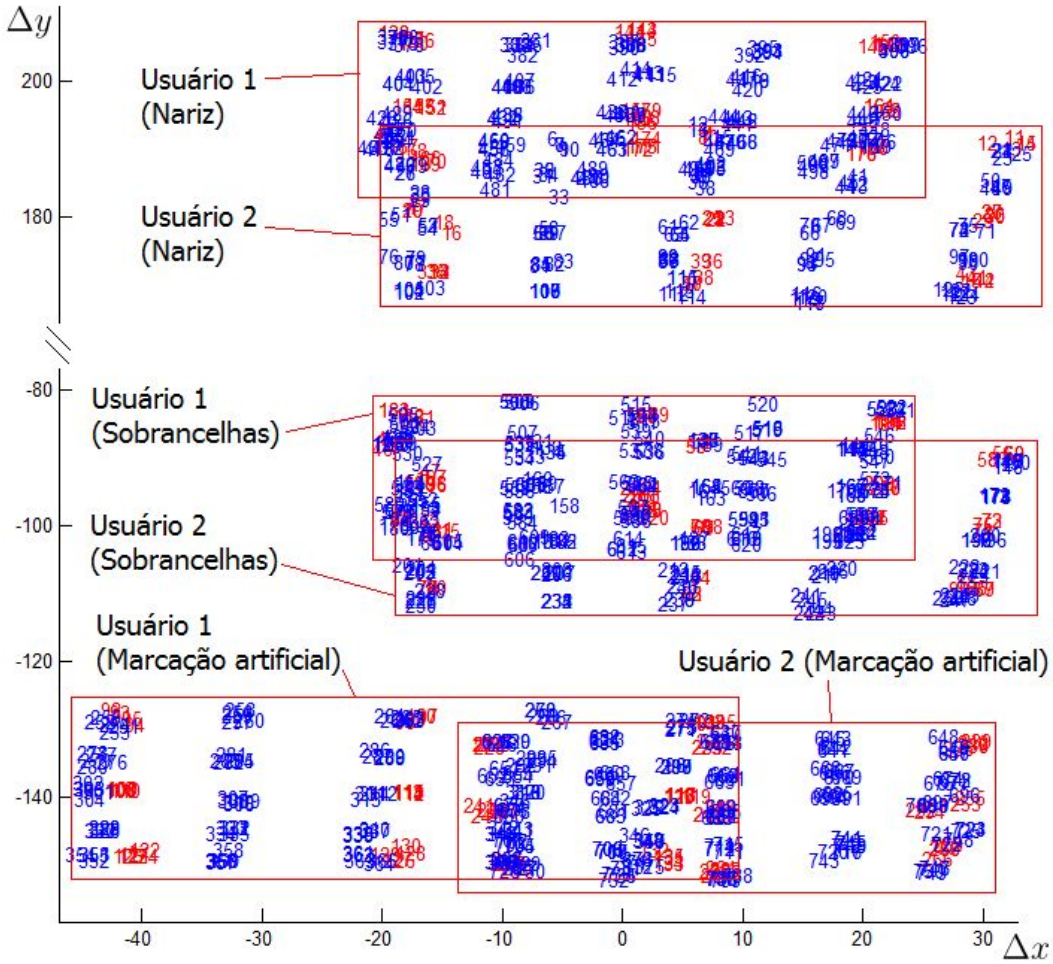


Figura 3.46: Distribuição espacial (em píxeis) dos dados de entrada de dois usuários e três pontos fixos

4 TESTES, RESULTADOS E ANÁLISES

Os testes foram realizados através do uso de bases de dados de imagens. Existem várias bases de imagens de faces na literatura, o que permitiria avaliar o Módulo 1, que faz a localização do olho na imagem, pois sua avaliação necessitaria apenas de imagens de faces contendo a marcação do centro da íris, comum na maioria das bases. Porém, para a avaliação dos Módulos 2 e 3, seria necessário utilizar uma base de dados criada especificamente para avaliar sistemas de rastreamento do olhar. Essas bases contêm, além das informações típicas de uma base de faces (coordenadas da face e dos seus componentes), a informação sobre o ponto na tela para onde o usuário estava olhando no momento da captura.

Segundo estudo realizado, existem duas bases de imagens sobre rastreamento do olhar na literatura: *"A Comprehensive Head Pose And Gaze Database"* (WEIDENBACHER et al., 2007) e *"A Natural Head Pose and Eye Gaze Dataset"* (ASTERIADIS et al., 2009). Porém, elas não puderam ser utilizadas neste trabalho. A primeira encontrava-se incompleta, possuindo apenas imagens com variação horizontal do olhar, sem o movimento vertical do olho. Na segunda os usuários podiam movimentar a cabeça durante a captura da imagem, e uma restrição neste trabalho é que a cabeça permaneça imóvel durante a captura (vide seção 3.3). Além disso,

nesta última base as imagens foram capturadas em baixa resolução (640x480 píxeis).

Optou-se então por gerar uma base de dados para avaliar o desempenho do protótipo. A base foi criada a partir de um programa de captura, desenvolvido especificamente para este projeto. Em cada captura, foram capturadas imagens do usuário olhando para nove pontos distintos na tela na fase de calibração e 25 pontos na fase de teste, totalizando 34 pontos. Para cada ponto, foram capturados cinco quadros de vídeo, totalizando 170 quadros por captura.

Foram capturadas imagens de cinco usuários, sendo 4 homens e 1 mulher, caucasianos, com idade entre 27 e 30 anos, sem uso de óculos. Para cada usuário, foram feitas duas capturas: uma utilizando o marcador artificial, para medir sua acurácia, e outra sem o marcador, para medir a acurácia do uso do nariz e das sobrancelhas como pontos fixos, totalizando 10 capturas, num total de 1700 imagens na base de dados (340 imagens por usuário). As imagens foram capturadas na resolução de 1600x1200 píxeis. Cada imagem da base contém informações sobre as coordenadas do ponto na tela para onde o usuário estava olhando no instante da captura, além de informações sobre as coordenadas dos centros das duas íris, que foram marcadas manualmente.

Para evitar que os usuários movimentassem a cabeça durante as capturas, foi improvisado um suporte para a cabeça. A câmera foi fixada no suporte, para que sua posição em relação ao usuário se mantivesse fixa. A iluminação foi feita através de duas luminárias, posicionadas de cada lado do monitor, de forma a manter a face do usuário bem iluminada. Uma visão geral do ambiente é ilustrada na figura 4.1, juntamente com a visão de um usuário utilizando o sistema.

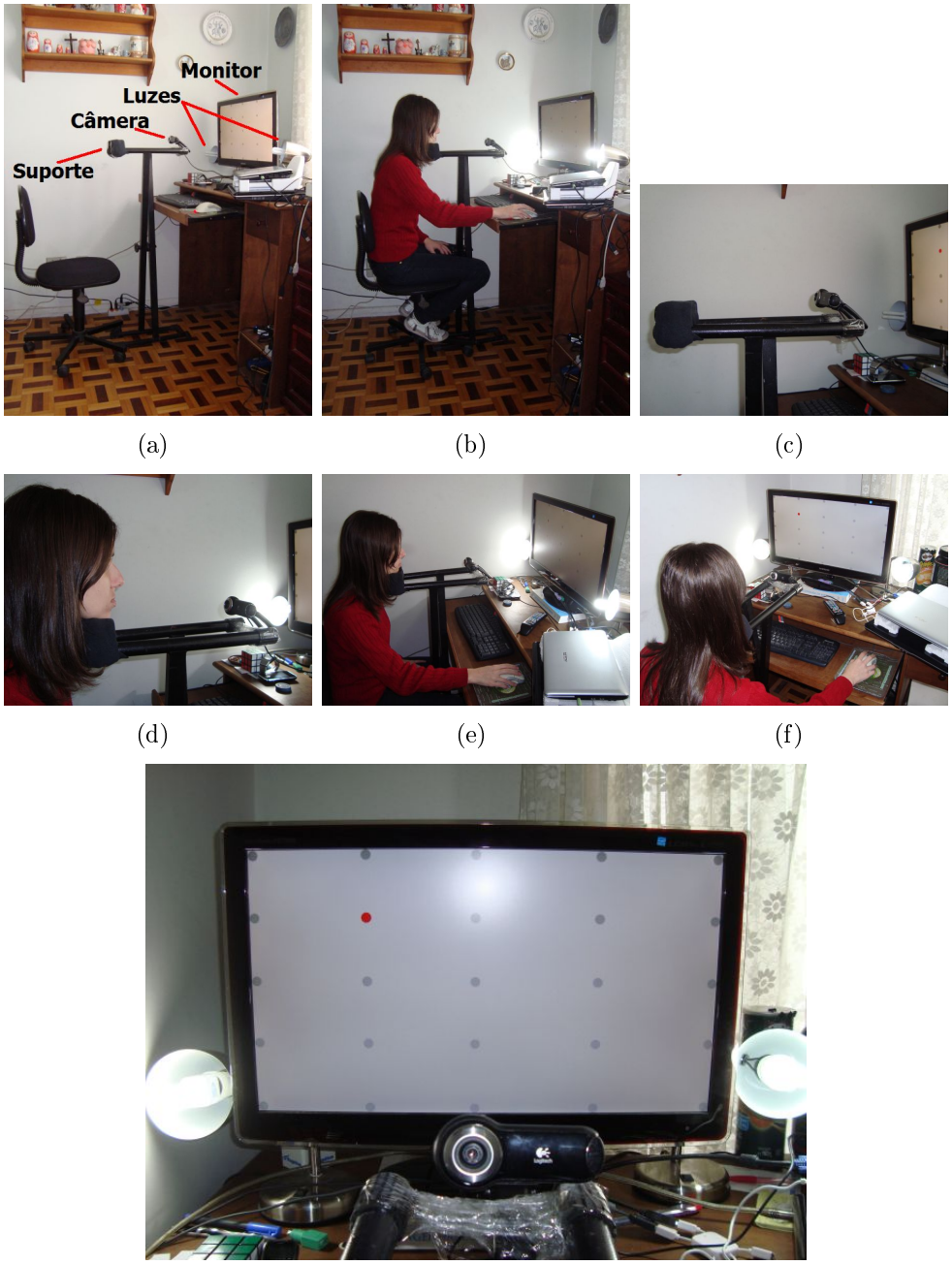


Figura 4.1: Ambiente de captura: (a) visão geral, (b) usuário no ambiente, (c-d) detalhe do suporte, (e-f) outros ângulos de visão e (g) ponto de vista do usuário

4.1 Localização do olho na imagem

A localização do olho na imagem consiste das etapas descritas na seção 3.4. Quanto aos resultados, foram mensurados os acertos referentes à detecção da face e redução da área de busca, assim como a precisão da localização da íris na imagem.

4.1.1 Detecção da face e redução da área de busca

O detector de faces utilizado (VIOLA E JONES, 2001b) retorna uma região contendo as coordenadas do retângulo que contém a face. Em seguida, é feita a redução da área de busca, através da divisão da região da face em sub-regiões, conforme descrito na seção 3.4.

4.1.1.1 Resultados

Para medir o acerto desse procedimento, foi feita uma varredura na base de imagens e, para cada imagem, a face foi detectada e o acerto foi quantificado através dos seguintes indicadores:

1. Acertos da face: percentual de situações onde a face detectada continha no seu interior as coordenadas dos olhos, previamente marcadas.

2. Acertos do olho esquerdo: percentual de situações onde a região do olho esquerdo continha as coordenadas do olho esquerdo.

3. Acertos do olho direito: percentual de situações onde a região do olho direito continha as coordenadas do olho direito.

4. Acertos de ambos os olhos: percentual de situações onde as duas regiões obtiveram um acerto, simultaneamente.

Os resultados são apresentados na tabela 4.1.

Tabela 4.1: Resultados da detecção da face e redução da área de busca

Usuário	Quantidade de imagens	Acertos da face	Acertos do olho esquerdo	Acertos do olho direito	Acertos de ambos os olhos
1	340	100.00%	100.00%	100.00%	100.00%
2	340	100.00%	100.00%	100.00%	100.00%
3	340	100.00%	100.00%	100.00%	100.00%
4	340	100.00%	100.00%	100.00%	100.00%
5	340	100.00%	100.00%	100.00%	100.00%
Total	1700	100.00%	100.00%	100.00%	100.00%

4.1.1.2 Análise

A redução da área de busca conseguiu englobar o olho em todas as imagens da base de dados, mostrando ser uma forma robusta de isolar as áreas de interesse para diminuir o processamento a ser realizado.

4.1.2 Localização da íris

O Módulo 1 retorna as coordenadas dos centros das duas íris. Cada imagem da base possui as coordenadas reais das duas íris, marcadas manualmente. Com base nessas informações, pode-se medir o acerto da localização da íris na imagem.

4.1.2.1 Resultados

Para medir a acurácia da localização da íris, foi utilizada a métrica definida por JESORSKY; KIRCHBERG E FRISCHHOLZ (2001). Ela é calculada da seguinte forma:

1. As íris são detectadas, gerando dois pares de coordenadas $(OEsq_x, OEsq_y)$ e $(ODir_x, ODir_y)$.
2. Em seguida, calcula-se a distância euclidiana entre cada par de coordenadas e os pares correspondentes à marcação real (denominados $(MEsq_x, MEsq_y)$ e $(MDir_x, MDir_y)$), marcados manualmente, gerando duas distâncias.
3. A maior distância é dividida pela distância entre os dois olhos, utilizando a marcação manual. Essa distância é chamada de distância relativa dos olhos.

O processo é sumarizado na equação 4.1. Quanto menor o valor, menor o erro gerado pelo Módulo 1. Um valor igual a zero corresponderia a um acerto absoluto. A interpretação de outros valores é simples: um valor igual a 0.5 corresponde a um erro igual à distância entre os dois olhos, pois a maior distância de erro equivale à metade da distância entre os dois olhos; um erro menor ou igual a 0.25, por exemplo, corresponde a uma acurácia de aproximadamente a largura do olho, e assim por diante. Essa métrica é bastante robusta porque é invariante ao tamanho da face.

$$\begin{aligned} DistEsq &= \sqrt{(OEsq_x - MEsq_x)^2 + (OEsq_y - MEsq_y)^2} \\ DistDir &= \sqrt{(ODir_x - MDir_x)^2 + (ODir_y - MDir_y)^2} \\ DistOlhos &= \sqrt{(MDir_x - MEsq_x)^2 + (MDir_y - MEsq_y)^2} \\ erro &= \frac{max(DistEsq, DistDir)}{DistOlhos} \end{aligned} \tag{4.1}$$

O Módulo 1 também faz um refinamento da localização da íris, conforme explicado na seção 3.4.4. Para fins de comparação, foi feita uma análise dos resultados da localização da íris com e sem o refinamento, de forma a validar a sua eficácia.

Os resultados são apresentados na tabela 4.2. Foram consideradas apenas as imagens que obtiveram acerto em ambos os olhos na detecção da face e redução da área de busca (última coluna da tabela 4.1).

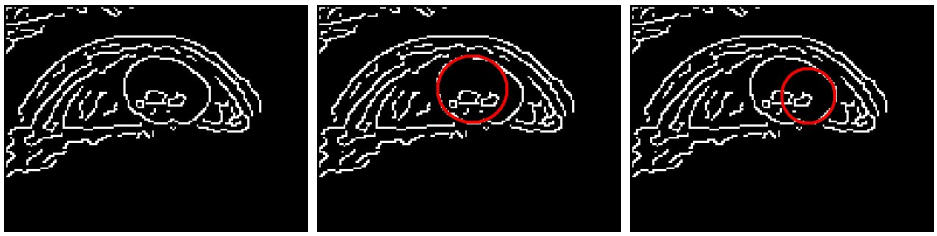
Tabela 4.2: Resultados da localização da íris. A linha "Total" corresponde à média de cada coluna, exceto na coluna "Quantidade de imagens", onde corresponde à soma.

Usuário	Quantidade de imagens	Erro sem o refinamento	Erro com o refinamento	Diferença absoluta
1	340	0.0153	0.0150	0.0003
2	340	0.0137	0.0203	-0.0066
3	340	0.0163	0.0162	0.0001
4	340	0.0167	0.0196	-0.0029
5	340	0.0116	0.0178	-0.0062
Total	1700	0.0147	0.0178	-0.0031

4.1.2.2 Análise

Considerando a métrica de erro definida na equação 4.1, o refinamento da detecção aparentemente não apresentou uma diferença significativa no sentido de melhorar a precisão da localização da íris. Na maioria dos casos, a precisão até ficou pior, como nos casos dos usuários 2, 4 e 5. Porém, posteriormente observou-se que as coordenadas retornadas pelo refinamento são mais estáveis do que as coordenadas retornadas pela Transformada de Hough, utilizada pelo Módulo 1 para localizar a íris. Isso ocorre porque é feita uma convolução com uma máscara suavizada, o que faz com que ela se ajuste de forma mais precisa na íris, enquanto que a Transformada

de Hough procura fazer um casamento exato com as bordas da imagem. Em certos casos, isso poderia gerar ambiguidade, conforme ilustra a figura 4.2.



(a) (b) (c)

Figura 4.2: Confusão que poderia ser gerada pela Transformada de Hough: (a) Imagem de bordas e (b)(c) possibilidades de detecção da íris

A ambiguidade pode fazer com que as possibilidades de detecção possam ser escolhidas alternadamente, causando instabilidade na localização da íris. A estabilidade proporcionada pelo refinamento faz com que a estimação da direção do olhar possua uma acurácia maior, como será visto na seção 4.3.

4.2 Detecção do ponto fixo

O Módulo 2 faz a detecção de um ponto fixo na face do usuário no primeiro quadro de vídeo e realiza o rastreamento desse ponto nos quadros subsequentes. A exibição de resultados desse módulo será feita graficamente, visto que a base de imagens não possui informação sobre a posição do ponto fixo na imagem. Isso seria inviável de ser realizado porque a seleção desse ponto pelo algoritmo é diferente de uma seleção manual, pois a seleção do algoritmo visa localizar a área que possui maior desvio padrão dos píxeis nela contidos, e uma seleção manual seria feita de forma empírica.

4.2.1 Resultados

A seguir são apresentados alguns resultados gráficos da detecção e do rastreamento dos pontos fixos. Conforme descrito na seção 3.5, foram utilizados três pontos fixos: nariz, sobrancelhas e marcação artificial. A figura 4.3 ilustra os resultados de cada uma das 10 capturas realizadas. Como cada captura é composta por 170 quadros de vídeo, para a apresentação dos resultados foi exibido apenas um dos quadros e o resultado do rastreamento de cada quadro na mesma imagem, para fins de ilustração. O retângulo tracejado indica a seleção do ponto fixo no primeiro quadro de vídeo e os demais 169 retângulos contínuos indicam o seu rastreamento nos demais quadros de vídeo.

4.2.2 Análise

Pode-se notar uma estabilidade muito boa no rastreamento da marcação artificial, pois é um ponto fixo bastante distinto na imagem, sendo de fácil localização. O nariz apresentou uma boa estabilidade, porém sua localização varia em relação a posição dos olhos quando o usuário realiza algum movimento com a cabeça, o que prejudica a estimação da direção do olhar (vide seção 4.3.2). As sobrancelhas tiveram uma instabilidade muito alta no rastreamento, principalmente nas figuras 4.3(c)-(e) pois são elementos inconstantes na imagem devido principalmente à sua movimentação pelos usuários. Portanto, não é recomendável o seu uso como ponto fixo em um sistema de rastreamento do olhar.

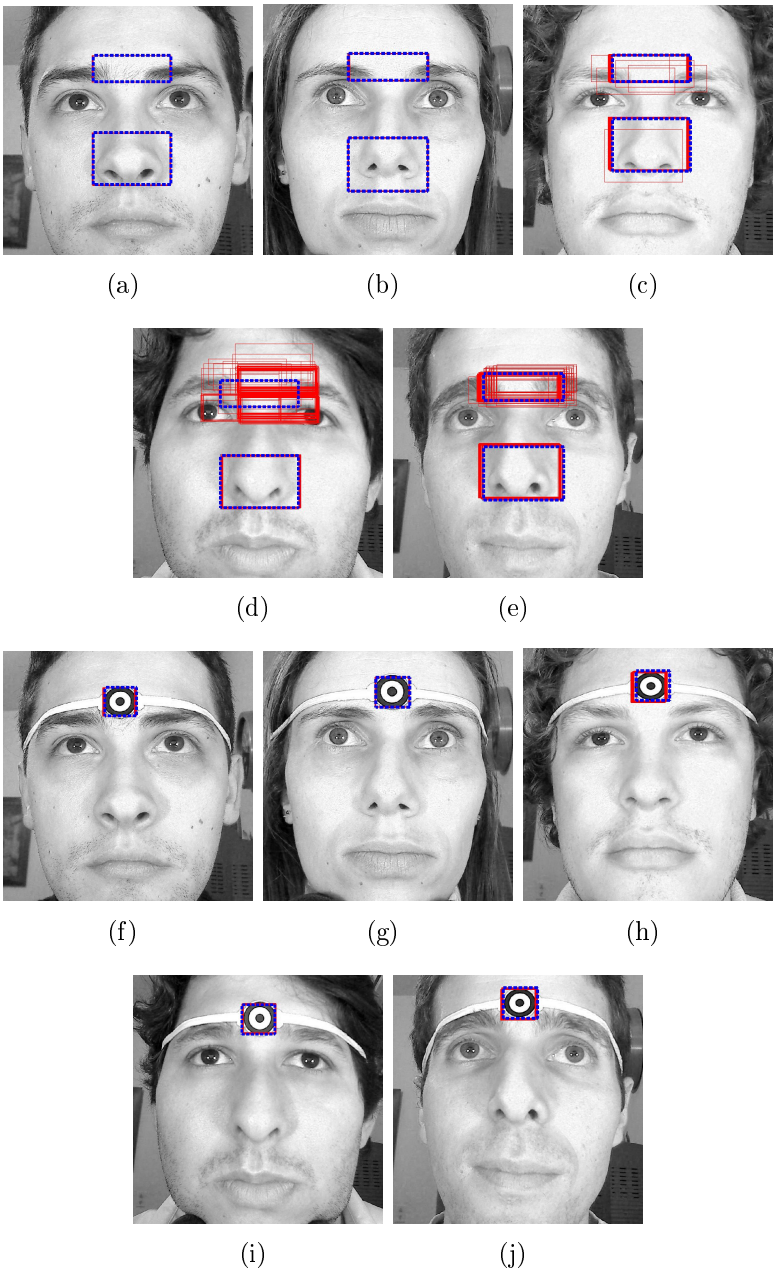


Figura 4.3: Resultados da detecção e rastreamento do ponto fixo para (a-e) nariz, sobrancelha e (f-j) marcação artificial. O retângulo tracejado indica a detecção do ponto fixo no primeiro quadro de vídeo e os demais retângulos indicam o seu rastreamento nos quadros seguintes

4.3 Estimação da direção do olhar

A estimação da direção da olhar foi feita utilizando RNAs, conforme explicado na seção 3.6. Para cada usuário, foi treinada uma RNA utilizando os dados de calibração, compostos por nove pontos, sendo cinco capturas por ponto. Logo, o conjunto de treinamento continha 45 amostras. Após o treinamento, foi passado para a rede o conjunto de teste, composto por 25 pontos, totalizando 125 amostras. Foi feita uma análise da estimação utilizando os três pontos fixos estudados, de forma a medir o seu impacto no desempenho deste módulo. Também foi feito um comparativo do impacto do refinamento da localização da íris na estimação da direção do olhar.

4.3.1 Resultados

Para cada usuário, foi feita a estimação da direção do olhar para os 25 pontos de tela utilizados na fase de teste (125 pontos no total, pois foram capturados cinco quadros de vídeo em cada ponto de tela). Para cada ponto de teste, foi calculada a distância euclidiana do ponto estimado para o ponto desejado e foi tirada uma média dessas distâncias. Esse valor indica o erro médio da estimação e foi calculado para cada usuário. A medida de acurácia mais utilizada na literatura é o acerto do sistema em graus (HANSEN E JI, 2010), neste trabalho denominado de cone de erro. A figura 4.4 ilustra o cone de erro. O raio do cone equivale ao erro médio da estimação.

O cálculo do ângulo do cone de erro pode ser feito através da equação 4.2.

$$\Theta = 2 \times arctan\left(\frac{raio}{distancia}\right), \tag{4.2}$$

onde *raio* equivale ao erro médio da estimação e *distancia* equivale à distância do

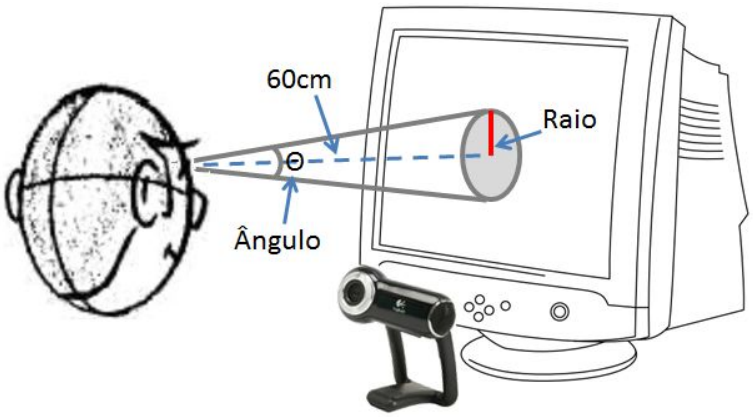


Figura 4.4: Ilustração do cone de erro

usuário para a tela (neste trabalho isso foi fixado em 60cm). A tabela 4.3 exibe os resultados da estimação do olhar em graus para cada usuário da base de dados, para cada um dos três pontos fixos. Quanto menor o valor do grau, mais precisa é a estimação, pois o cone de erro é mais fechado em uma região menor.

Tabela 4.3: Resultados da estimação do olhar utilizando cada um dos três pontos fixos. A coluna "Melhor resultado" indica a situação onde ocorreu a melhor precisão através do par X/Y, onde X pode ser N=Nariz, MA=Marcação artificial e S=Sobrancelhas, e Y pode ser S=Sem o refinamento e C=Com o refinamento.

Usuário	Sem o refinamento			Com o refinamento			Melhor resultado
	Nariz	Sobrancelhas	Marcação artificial	Nariz	Sobrancelhas	Marcação artificial	
1	7.33^o	6.30^o	7.54^o	6.69^o	5.63^o	6.19^o	S/C
2	9.70^o	8.92^o	6.80^o	8.36^o	6.88^o	4.64^o	MA/C
3	14.06^o	13.36^o	7.55^o	13.20^o	11.79^o	4.84^o	MA/C
4	5.53^o	30.57^o	5.48^o	3.62^o	30.20^o	6.37^o	N/C
5	15.25^o	37.66^o	7.85^o	15.04^o	37.31^o	5.90^o	MA/C
Média	10.37^o	19.36^o	7.05^o	9.38^o	18.36^o	5.59^o	MA/C

O gráfico na figura 4.5 ilustra uma visão da última linha da tabela, sumarizando os resultados gerais do protótipo.

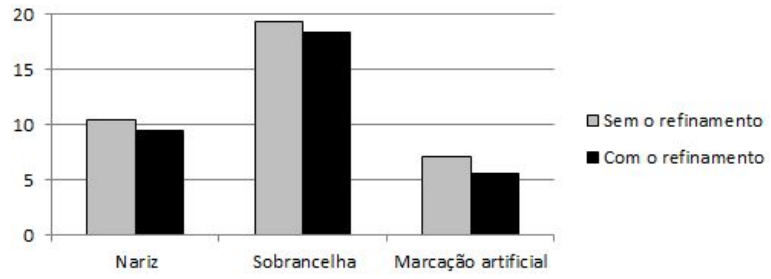


Figura 4.5: Gráfico representando a última linha da tabela 4.3. Quanto menor o valor, maior a precisão.

4.3.2 Análise

Pode-se notar que, na grande maioria dos casos, os melhores resultados foram obtidos com o refinamento da localização da íris. As figuras 4.6 e 4.7 ilustram bem isso. Nelas pode-se visualizar graficamente os resultados obtidos. Cada retângulo representa uma captura, contendo os 25 pontos de teste. Para cada ponto, estão representados os pontos referentes à estimativa da direção do olhar nos cinco quadros de vídeo capturados enquanto o usuário olhava para aquele ponto. Nota-se um melhor agrupamento dos pontos nas capturas realizadas com o refinamento, como foi argumentado na seção 4.1.2.2.

A marcação artificial foi quem gerou os melhores resultados. Os piores resultados foram obtidos nos usuários 4 e 5, com o uso da sobrancelha como ponto fixo. Isso se deve ao fato desses usuários não possuírem sobrancelhas muito acentuadas, o que mostra que o uso dessa característica como ponto fixo é desaconselhável. O uso do nariz como ponto fixo apresentou bons resultados, principalmente para o usuário 4, onde gerou o melhor resultado. Porém, este ponto fixo gerou os piores resultados para o usuário 3. Isso se deve pelo fato do mesmo ter movimentado o nariz durante a captura das imagens. De forma geral, a marcação artificial parece ser o ponto fixo mais estável para ser utilizado.

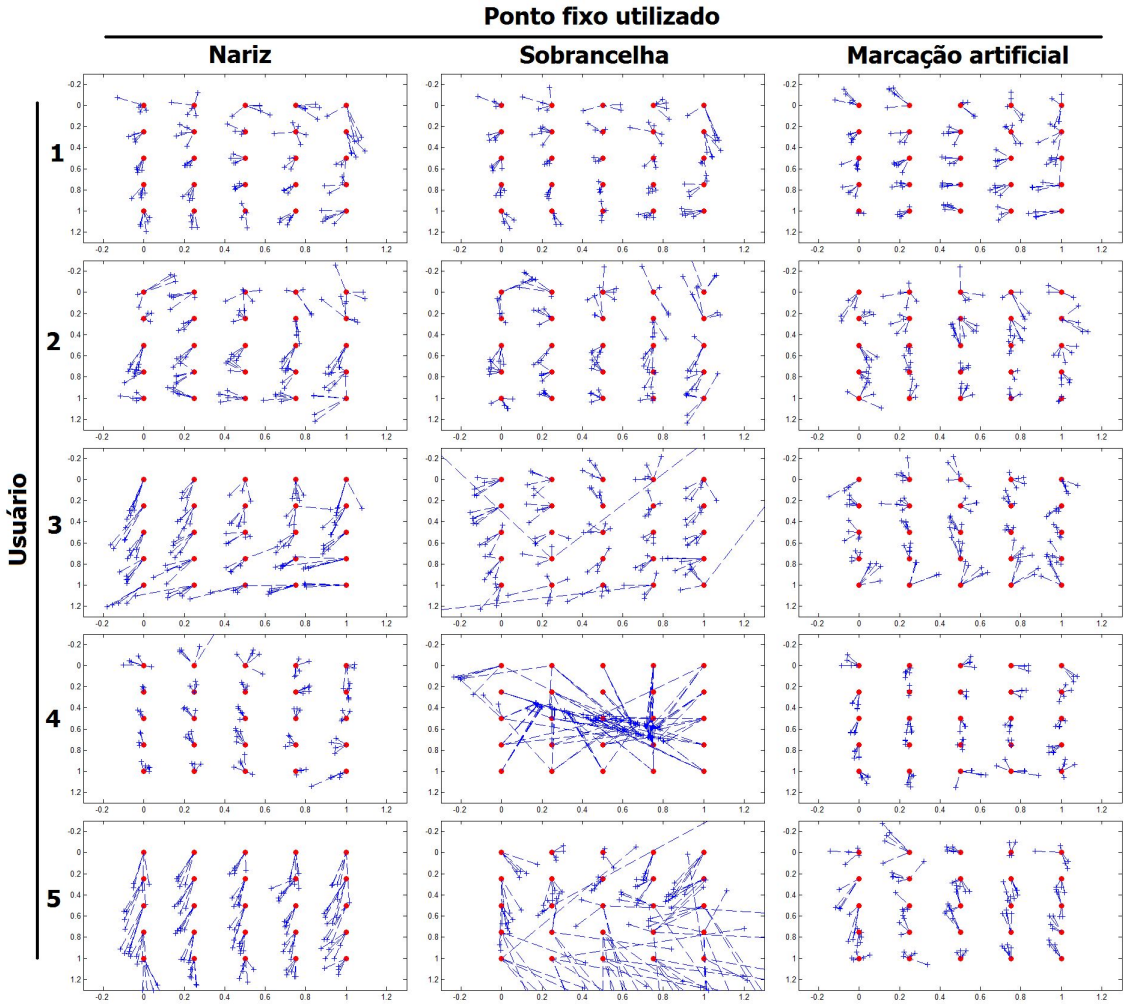


Figura 4.6: Visualização gráfica dos resultados da estimativa da direção olhar sem o refinamento.

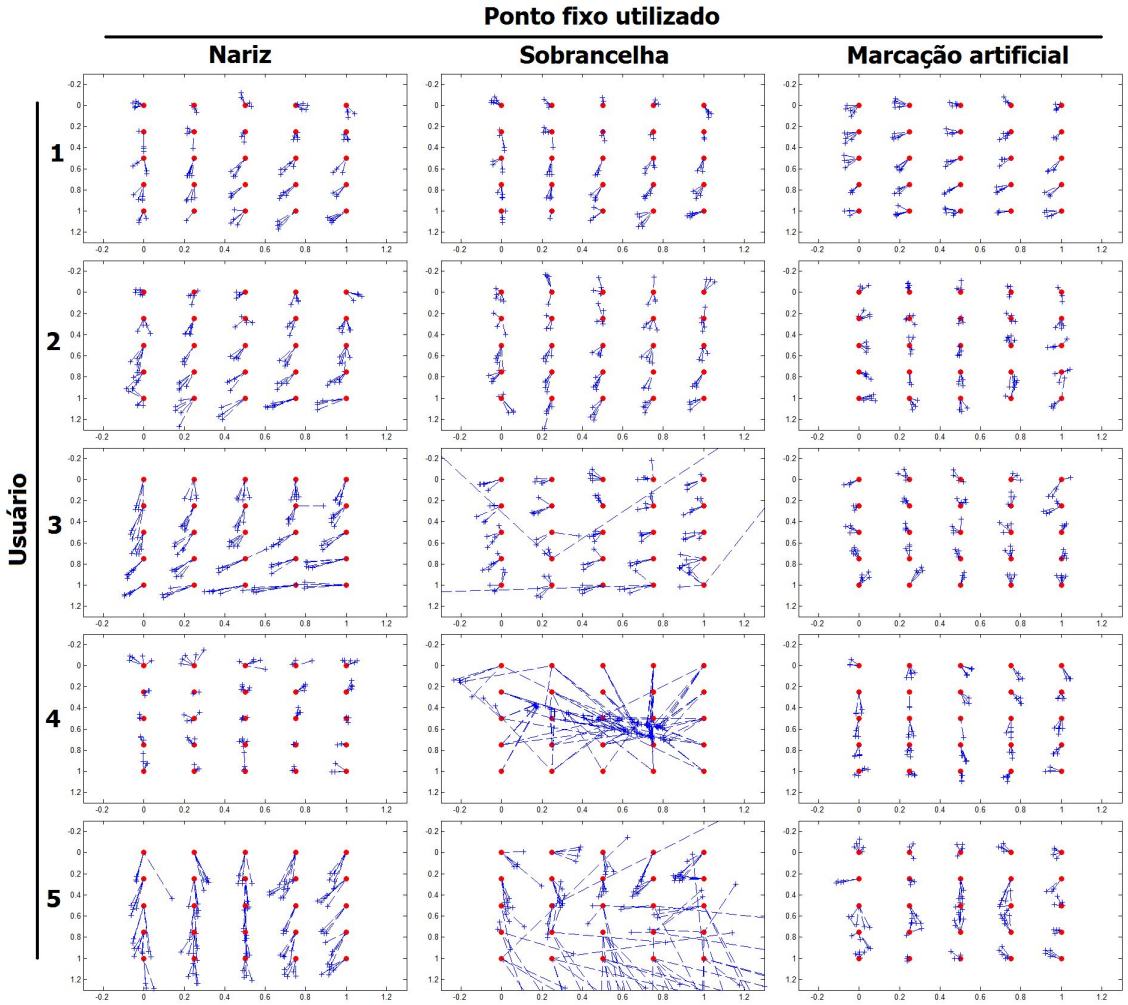


Figura 4.7: Visualização gráfica dos resultados da estimativa da direção olhar com o refinamento.

5 CONCLUSÕES E SUGESTÕES DE TRABALHOS FUTUROS

Neste trabalho foi feito um estudo sobre rastreamento do olhar e desenvolvido o protótipo de um sistema de detecção da direção do olhar. O protótipo foi desenvolvido de forma modular, para melhor dividir as etapas típicas existentes nesse tipo de sistema.

Com relação ao Módulo 1 (Localização do olho na imagem), verificou-se que a localização da face para delimitar a área de busca é uma boa forma de reduzir o processamento a ser feito na imagem, pois este é aplicado somente às regiões de interesse (no caso, as regiões dos olhos). Utilizar a Transformada Circular de Hough para encontrar círculos na imagem de bordas é uma boa forma de localizar a íris. Apesar da mesma poder assumir um formato elíptico em algumas situações, a transformada consegue aproximar com bastante precisão a sua localização na imagem. Verificou-se que um refinamento da localização trouxe bons resultados para a estimação da direção do olhar. Apesar do refinamento não ter melhorado diretamente a precisão da localização, este apresentou uma forma mais estável de definir a posição da íris na imagem.

No Módulo 2 (Detecção do ponto fixo) foram estudadas algumas formas de detectar e rastrear um ponto fixo na imagem, para que a estimativa da direção do olhar possa ser feita com base no deslocamento dos olhos em relação a esse ponto fixo. Foram feitos experimentos utilizando as sobrancelhas, o nariz e uma marcação artificial. A marcação artificial apresentou os melhores resultados, apesar de ser um método um tanto invasivo, pois um objeto deve ser colocado no usuário antes das capturas. Foi aplicada a técnica de *Phase Correlation* para rastrear os pontos fixos, o que demonstrou bons resultados.

O Módulo 3 (Estimação da direção do olhar) utilizou RNAs para realizar a estimação da direção do olhar, utilizando as informações da posição dos olhos com e sem o refinamento da localização e utilizando os três pontos fixos abordados no estudo. Verificou-se que os melhores resultados são obtidos utilizando a marcação artificial e fazendo a localização da íris com o refinamento. Também foi verificado que um modelo neural simples com uma única camada é suficiente para gerar bons resultados.

A principal contribuição deste trabalho foi ter conseguido obter bons resultados utilizando uma *webcam* comum ao invés de utilizar uma câmera mais robusta. Além disso, não foi utilizada iluminação infravermelha, presente na maioria dos trabalhos pelo fato de gerar imagens com um maior contraste da íris com o meio exterior, facilitando sua localização e rastreamento. Ao invés disso, foi utilizada iluminação natural.

O acerto final do sistema (cerca de 5.6^o de precisão) pode ser considerado satisfatório para uma primeira versão do protótipo. Os sistemas comerciais possuem precisão média de cerca de 0.5^o, conforme visto na seção 2.3. Mas esses sistemas utilizam câmeras com altíssima resolução e iluminação infravermelha, o que resulta em imagens de maior qualidade. No levantamento da área realizado por HANSEN E

JI (2010), foi constatado que sistemas que utilizam *webcam* possuem precisão média entre 2^o-4^o (considerando o estado da arte na literatura). Isso é ilustrado na figura 5.1, que sumariza a precisão média dos sistemas levantados pelos autores.

Cameras	Lights	Gaze Info	Head pose	Calibration	Accuracy (deg)	References	Comments
1	0	PoR	—	—	$2-4$	[47], [46], [157]	web-camera
1	0	LoG/LoS	—	Fully	$1-2$	[151], [144], [145]	
1	0	LoG	≈	—	<1	[79]	*[a]
1	1	PoR	—	—	1-2	[103], [156], [70]	*[b]
1	2	PoR	✓	Fully	$1-3$	[105], [100], [43]	
1+1	1	PoR	✓	Fully	3	[112]	Mirrors
1(+1)	4	PoR	✓	—	$<1-2.5$	[164], [20]	
2	0	PoR	✓	—	1	[109]	*[c]
2+1	1	LoG	✓	—	0.7-1	[135]	pan/tilt
2+2	2	PoR	✓	Fully	0.6	[8]	Mirrors
2	2(3)	PoR	✓	Fully	$<1-2$	[128], [127]	*[d]
3	2	PoR	✓	Fully	—	[139][11]	
1	1	PoR	—	—	0.5-1.5	[6], [133], [136], [160]	*[e]

Fig. 9. Comparison of gaze estimation methods with respective prerequisites and reported accuracies (e.g., based on different data and scenarios). The "cameras" column shows the number of cameras necessary for the methods. An additional "+1" means that an extra pan and tilt camera is used. If this is given in parenthesis, the pan and tilt is used in the implementation, but not necessary by the method. The column "Lights" indicates the number of light sources needed and with an additional set of parenthesis to indicate if extra lights have been used in the implementations. "Gaze info" describes the type of gaze information being inferred by the method (PoR), optical (LoG), or visual axes (LoS). When LoG/LoS is used, it is implicitly assumed that an additional 3D scene model is needed to get the point of regard. The column "Head pose" shows if the methods are head pose invariant (√), if approximate solutions are proposed (≈), or an external head pose unit is needed (—). The "Calibration" column indicates if explicit calibration of scene geometry and cameras are needed prior to use. [a]Additional markers, iris radius, parallel with screen. [b]Polynomial approximation. [c]3D face model. [d]Experiments have been conducted with three glints, but two ought to be sufficient. [e]Appearance-based.

Figura 5.1: Precisão dos trabalhos relacionados levantados por HANSEN E JI (2010). Destaque para a precisão dos sistemas que utilizam *webcam* (retângulo).

Esses sistemas utilizam técnicas diferentes das utilizadas neste trabalho, cujo foco foi apostar no uso de técnicas mais simples, tentando reduzir o processamento. O sistema da referência [46] (HANSEN et al., 2003) na figura 5.1 utiliza modelos de formas ativas (*Active Appearance Models* - *AAM*), que combina informações de forma com aparência em um modelo único do olho. O modelo é posteriormente casado com a imagem através da variação de parâmetros, utilizando uma regra de aprendizagem HANSEN E JI (2010). O sistema da referência [47] (HANSEN E PECE, 2005) combinou as técnicas de maximização de expectativa (*Expectation* -

Maximization - EM) e o método RANSAC (*RANdom SAmple Consensus*) para realizar o casamento da íris com a imagem. Isso é feito utilizando um modelo elíptico, combinado com informações sobre os píxeis vizinhos às bordas. Os autores do sistema da referência [157] (WILLIAMS; BLAKE E CIPOLLA, 2006) propõem um modelo de regressão de processo Gaussiano esparso e semissupervisionado que aprendem um mapeamento utilizando apenas dados parcialmente marcados, apresentando bons resultados.

Como propostas de trabalhos futuros, ficam as seguintes sugestões:

1. Estudar técnicas que possibilitem tolerar movimentos de cabeça. A eliminação dessa restrição facilitaria o uso do sistema pelos usuários. Isso poderia ser feito através da construção de um modelo tridimensional da cabeça do usuário, combinando essa informação com a posição dos olhos. Uma sugestão seria utilizar o *Kinect*, da Microsoft.

2. Fazer um estudo mais aprofundado da localização e rastreamento do ponto fixo, principalmente utilizando outros tipos marcação artificial, variando seu tamanho e posicionamento e evitando que o mesmo sofra variações durante a captura das imagens. Sugere-se experimentar o uso dos cantos dos olhos como marcação natural. Apesar do seu formato variar de pessoa para pessoa, parecem ser atributos bastante estáveis.

3. Estudar outras técnicas de localização da íris, como a localização através do uso de classificadores (como RNAs, por exemplo), casamento de padrões e com o uso de modelos de formas deformáveis.

4. Fazer uma análise temporal do rastreamento. Isso poderia ser feito através do estudo do movimento dos olhos enquanto os mesmos seguem um ponto que se move na tela, suavizando o rastreamento através de filtros temporais, e analisar o comportamento do usuário enquanto o mesmo assiste a um vídeo.

REFERÊNCIAS

ASTERIADIS, S. et al. A natural head pose and eye gaze dataset. In: ICMI-MLMI 2009. INTERNATIONAL CONFERENCE ON MULTIMODAL INTERFACES, 2009, Cambridge, Mass.. WOKSHOP ON MACHINE LEARNING FOR MULTIMODAL INTERFACES. Cambridge, Mass. **Proceedings...** New York: ACM/IEEE, 2009.

BOENING, G. et al. Mobile eye tracking as a basis for real-time control of a gaze driven head-mounted video camera. In: ETRA '06. EYE TRACKING RESEARCH & APPLICATIONS, 2006, San Diego CA. **Proceedings …** New York: ACM, 2006. p. 56-56.

BORGES, A. **O que é o DOSVOX**. Disponível em: <http://intervox.nce.ufrj.br/dosvox/intro.htm>. Acesso em: Nov. 2009.

BRAGA, A. P. ; CARVALHO, A. P. L. F. ; LUDEMIR, T. B. **Fundamentos de redes neurais artificiais.** 2. ed. Rio de Janeiro: LTC, 1998.

BRASIL. Lei n. 10.098, de 19 de dezembro de 2000. **Diário Oficial [da] República Federativa do Brasil**, Brasíia:, DF, 20 dez. 2000.Estabelece normas gerais e critérios básicos para a promoção da acessibilidade das pessoas portadoras de deficiência ou com mobilidade reduzida, e dá outras providências.

BROLLY, X. L. C. ; MULLIGAN, J. B. Implicit calibration of a remote gaze tracker. In: CVPRW'04. CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOP, 2004, Washington, DC, **Proceedings …** Washington, DC: IEEE, 2004. v. 8. p. 134.

CANNY, J. A computational approach to edge detection. **IEEE Transaction on Pattern Analysis and Machine Intelligence**, New York, v. 8, n. 6, p. 679-698, Nov. 1986.

COMANICIU, D. ; RAMESH, V. ; MEER, P. Real-time tracking of non-rigid objects using Mean-Shift. In: CVPR 2000. CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2000. Hilton Head, SC. **Proceedings …** New York: IEEE, 2000. V. 2. p. 2142-2151.

CONGI, A. ; AZEVEDO, E. ; LETA, F. R. **Computação gráfica**, teoria e prática. Rio de Janeiro: Campus, 2008.

COUTINHO, F. L. ; MORIMOTO, C. H. Free head motion eye gaze tracking using a single camera and multiple light sources. SIBGRAPI´06. BRAZILIAN SYMPOSIUM

ON COMPUTER GRAPHICS AND IMAGE PROCESSING, 19., 2006, Manaus. **Proceedings …** Los Alamitos, CA: IEEE, 2006. p. 171-178.

CRISAFULLI, G. ; IANNIZZOTTO, G. ; LA ROSA, F. Two competitive solutions to the problem of remote eye-tracking. In: HSI '09. CONFERENCE ON HUMAN SYSTEM INTERACTIONS, 2., 2009. Catania, Italy. **Proceedings …** HSI '09. 2ND CONFERENCE ON. Anais... Catania, Italy: IEEE, 2009. p. 356-362.

CUONG, N. H. ; HOANG, H. T. Eye-gaze detection with a single WebCAM based on geometry features extraction. In: ICARCV. INTERNATIONAL CONFERENCE ON CONTROL AUTOMATION ROBOTICS VISION, 11., 2010, Singapore. **Proceedings …** Singapore: IEEE, 2010. p. 2507-2512.

DROEGE, D. ; GEIER, T. ; PAULUS, D. Improved low cost gaze tracker. In: COGAIN. ANNUAL CONFERENCE ON COMMUNICATION BY GAZE INTERACTION, 3., 2007, Leicester, UK. **Proceedings …** Leicester, UK: COGAIN Association, 2007. p. 37-40.

DROEGE, D. ; SCHMIDT, C. ; PAULUS, D. A Comparison of pupil center estimation algorithms. In: COGAIN. ANNUAL CONFERENCE ON COMMUNICATION BY GAZE INTERACTION, 4., 2008, Prague. **Proceedings …** Prague: COGAIN Association, 2008. p. 23-26.

DUCHOWSKI, A. **Eye tracking methodology**: theory and practice. London: Springer-Verlag, 2007.

DUDA, R. O. ; HART, P. E. Use of the Hough transformation to detect lines and curves in pictures. **Communication of ACM**, New York, v. 15, n. 1, p. 11-15, Jan. 1972.

EBISAWA, Y. ; SATOH, S. Effectiveness of pupil area detection technique using two light sources and image difference method. In: ANNUAL INTERNATIONAL CONFERENCE OF THE IEEE ENGENEERING IN MEDICINE AND BIOLOGY SOCIETY, 15., 1993. San Diego, CA, **Proceedings …** San Diego, CA: IEEE, 1993. p. 1268-1269, 1993.

EDWARDS, A. D. **Extra-ordinary human-computer interaction**: interfaces for users with disabilities. Cambridge: Cambridge University Press, 1995.

FILHO, A. M. D. S. Percepção humana na interação humano-computador. **Revista Espaço Acadêmico**, Maringá, v. 3, n. 25, jun 2003.

FREUND, Y. ; SCHAPIRE, R. A decision-theoretic generalization of on-line learning and an application to on-line learning and an appliction to boosting. In: EUROPIAN CONFERENCE ON COMPUTATIONAL LEARNING THEORY, 2., 1995, Barcelona. **Proceedings...** London: Springer-Verlag, 1995.

GONZALEZ, R. C. ; WOODS, R. E. **Digital image processing**. 3. ed. Upper Saddle River: Prentice-Hall, 2006.

HAAR, A. Zur theorie der orthogonalen funktionensysteme**. Mathematische Annalen**, Berlin, v. 69, n. 3, p. 331-371, 1910. 10.1007/BF01456326.

HAMARNEH, G. Time Varying Shape, Spatio-Temporal Models. Deformable Spatio-Temporal Shape Models. In FISHER, R. (Ed.) **CVonline**: On-Line Compendium of Computer Vision 2002. Disponível em: http://www.dai.ed.ac.uk/CVonline/.

HANSEN, D. et al. Eye typing using markov and active appearance models. In: IEEE WORKSHOP ON APPLICATIONS ON COMPUTER VISION, 6., 2002, Washington, DC. **Proceedings …** Washington, DC: IEEE, 2003. p. 132-136.

HANSEN, D. ; JI, Q. In the eye of the beholder: a survey of models for eyes and gaze. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, New York, v. 32, n. 3, p. 478-500, Mar. 2010.

HANSEN, D. ; PECE, A. E. C. Eye tracking in the wild. **Computer Vision and Image Understanding**, San Diego, CA, v. 98, n. 1, p. 182-210, Apr. 2005.

HARRIS, C. ; STEPHENS, M. A Combined corner and edge detection. In: ALVEY VISION CONFERENCE, 4., 1988, Manchester, UK. **Proceedings...** Manchester, UK: Organising Committee AVC, 1988. p. 147-151.

HAYKIN, S. **Neural networks**: a comprehensive foundation. Upper Saddle River: Prentice-Hall, 1999.

HOUGH, P. V. C. MachineX analysis of bubble chamber pictures. In: INTERNATIONAL CONFERENCE ON HIGH ENERGY ACCELERATORS AND INSTRUMENTATION. 1959, Geneva. **Proceedings...** Geneva: European Organization for Nuclear Research, 1959.

HUANG, J. ; WECHSLER, H. Eye detection using optimal wavelet packets and radial basis functions (rbfs). **International Journal of Pattern recognition and Artificial Intelligence**, Singapore, v. 13, n. 7, Nov. 1999.

HUANG, W. ; MARIANI, R. Face detection and precise eyes location. In: ICPR'00. INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, 15., 2000, Barcelona. **Proceedings …** New York: IEEE, 2000. V. 4.

ISHIKAWA, T. et al. Passive driver gaze tracking with active appearance models. In: WORLD CONGRESS ON INTELLIGENT TRANSPORTATION SYSTEMS, 11., 2004, Nagoya, Japan. **Proceedings …** Ontario: ITS/STI, 2004.

IVINS, J. P. ; PORRILL, J. A deformable model of the human iris for measuring small 3-dimensional eye movements. **Machine Vision and Applications**, New York, v. 11, n. 1, p. 42-51, 1998.

JAIN, A. K. **Fundamentals of digital image processing**. Upper Saddle River: Prentice-Hall, 1989.

JAVAL, E. Essai sur la physiologie de la lecture. **Annales D'Oculistique**, Paris: v. 80, p. 135- 149, 1879.

JENSEN, O. H. **Implementing the Viola-Jones face detection algorithm**. 2008. Dissertação (Master Informatics and Mathematical) – Technical University of Denmark, Lyngby, 2008.

JESORSKY, O. ; KIRCHBERG, K. J. ; FRISCHHOLZ, R. Robust face detection using the hausdorff distance. In: AVBPA '01. INTERNATIONAL CONFERENCE ON AUDIO- AND VIDEO-BASED BIOMETRIC PERSON AUTHENTICATION, 3., 2001. Halmstad, Swenden. **Proceedings...** London: Springer-Verlag, 2001. p. 90-95.

JI, Q. ; ZHU, Z. Eye and gaze tracking for interactive graphic display. In: INTERNATIONAL SYMPOSIUM SMART GRAPHICS, 2., 2002, Hawthorne, NY. **Proceedings...** New York: ACM, 2002. p.79-85.

KUNKA, B. ; KOSTEK, B. Non-intrusive infrared-free eye tracking method. In: SIGNAL PROCESSING ALGORITHMS, ARCHITECTURES, ARRANGEMENTS, AND APPLICATIONS CONFERENCE PROCEEDINGS. 2009. Poznan. **Proceedings …** New York: IEEE, 2009. p. 105-109.

LAMARE, M. Des mouvements des yeux pendants la lecture. **Comptes Rendus de la Société Française D'Ophthalmologie**, Paris, p. 35-64, 1893.

MATSUMOTO, Y. ; ZELINSKY, A. An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In: INTERNATIONAL CONFERENCE ON AUTOMATIC FACE AND GESTURE RECOGNITION, 4., 2000. Grenoble, FR. **Proceedings …** New York: IEEE, 2000. p. 499-504.

MOUTINHO, A. M. **Identificação de padrões faciais usando redes neurais artificiais**. 2005. Dissertação (Mestrado em Informática) – Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005.

OHNO, T. ; MUKAWA, N. A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. EYE TRACKING RESEARCH & APPLICATIONS SYMPOSIUM, 2004, San Antonio, TX. **Proceedings …** New York: ACM, 2004. p. 115-122.

PEDRINI, H. ; SCHWARTZ, W. R. **Análise de Imagens digitais**. Rio de Janeiro: Thompson, 2008.

PENTLAND, A. ; MOGHADDAM, B. ; STARNER, T. View-based and modular eigenspaces for face recognition. CVPR'94. INTERNATIONAL CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 1994, Seattle, WA. **Proceedings …** New York: IEEE, 1994. P. 84-91.

SCHOTT, E. Über die registrierung des nystagmus und anderer augenbewegungen vermittels des saitengalvanometers. **Deutsches Archiv für klinisches Medizin**, Munchen, v. 140, p. 79-90, 1922.

SCHREIBER, D. Generalizing the Lucas-Kanade algorithm for histogram-based tracking. **Pattern Recognition Letters**, Amsterdam, v. 29, n. 7, p. 852-861, May 2008.

SOBEL, I. ; FELDMAN, G. **A 3x3 Isotropic Gradient Operator for Image Processing**. 1968. Presented at a talk at the Stanford Artificial Project. Never published.

TIAN, Y. ; KANADE, T. ; COHN, J. F. Dual-state parametric eye tracking. In: IEEE INTERNATIONAL CONFERENCE ON AUTOMATIC FACE AND GESTURE RECOGNITION, 4., 2000. Grenoble, FR. **Proceedings …** New York: IEEE, 2000.

TOMASI, C. ; MANDUCHI, R. Bilateral filtering for gray and color images. In: ICCV. INTERNATIONAL CONFERENCE ON COMPUTER VISION, 6., 1998, Bombay. **Proceedings …** New York: IEEE, 1998. p. 839-846.

TUNHUA, W. et al. Real-time non-intrusive eye tracking for human-computer interaction. In: CCISE. INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND EDUCATION, 5., 2010, Singapore. **Proceedings …** New York: IEEE, 2010. p. 1092-1096.

VALENTI, R. ; GEVERS, T. Accurate eye center location and tracking using isophote curvature. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2008, Anchorage, Alaska. **Proceedings …** New York: IEEE, 2008. p. 1-8.

VILLANUEVA, A. ; CABEZA, R. ; PORTA, S. Eye tracking: pupil orientation geometrical modeling. **Image and Vision Computing**, Guildford, Eng., v. 24, n. 8, p. 663-679, Jul. 2006.

VILLANUEVA, A. et al. **D5.6 Report on New Approaches to Eye Tracking. Summary of new algorithms**. Frederiksberg : Communication by Gaze Interaction (COGAIN), 2008. (IST-2003-511598: Deliverable 5.6).

VIOLA, P. ; JONES, M. Robust real-time object detection. INTERNATIONAL WORKSHOP ON STATISTICAL AND COMPUTATIONAL THEORIES OF VISION

– MODELING, LEARNING, COMPUTING, AND SAMPLING INTERNATIONAL JOURNAL OF COMPUTER VISION, 2., 2001, Vancouver. **Proceedings …** Vancouver: IEEE, 2001.

VIOLA, P. ; JONES, M. Rapid object detection using a boosted cascade of simple features. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2001, Kauwai, HI. **Proceedings...** New York: IEEE, 2001. v. 1. p. 511-518.

WANG, Z. et al. CamShift guided particle filter for visual tracking. **Pattern Recognition. Letters**, Amsterdam, v. 30, n. 4, p. 407-413, Mar. 2009.

WEIDENBACHER, U. et al. A Comprehensive Head Pose And Gaze Database. In: INTERNATIONAL CONFERENCE ON INTELLIGENT ENVIRONMENTS, 3., 2007, Ulm, Germany. **Proceedings …** Ulm Germany: IET/GICC, 2007. p. 455-458.

WILLIAMS, O. ; BLAKE, A. ; CIPOLLA, R. Sparse and semi-supervised visual apping with the S3GP. In: CVPR '06. 2006 IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2006, New York. **Proceedings …** New York: IEEE, 2006, v. 1, p. 230-237.

ZHU, Z. ; FUJIMURA, K. ; JI, Q. Real-time eye detection and tracking under various light conditions. In: SYMPOSIUM ON EYE TRACKING RESEARCH AND APPLICATIONS, 2002, New Orleans. **Proceedings …** New York: ACM, 2002.

ZIOU, D. ; TABBONE, S. Edge detection techniques - an overview. **International Journal of Pattern Recognition and Image Analysis**, Moscow, v. 8, p. 537-559, 1998.

ZITOVA, B. ; FLUSSER, J. Image registration methods: a survey. **Image and Vision Computing**, Guildford, Eng., v. 21, p. 977-1000, 2003.