UNIVERSIDADE DO VALE DO RIO DOS SINOS
CIÊNCIAS EXATAS E TECNOLÓGICAS.
PROGRAMA INTERDISCIPLINAR DE PÓS-GRADUAÇÃO EM
COMPUTAÇÃO APLICADA

Um método para detecção da orientação de pessoas em seqüências de vídeo

por

GUILHERME IZIDORO LAZZARI

Dissertação submetida a avaliação como requisito parcial para a obtenção do grau de Mestre em Computação Aplicada

Orientador: Prof Dr. Claudio Rosito Jung

São Leopoldo, Janeiro de 2008

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Lazzari, Guilherme Izidoro

Um método para detecção da orientação de pessoas em seqüências de vídeo
/ por Guilherme Izidoro Lazzari — São Leopoldo: Ciências Exatas e Tecnológicas da Unisinos, 2008.

60 f.: il.

Dissertação (mestrado) — Universidade do Vale do Rio dos Sinos. Ciências Exatas e Tecnológicas. Programa Interdisciplinar de Pós-Graduação em Computação Aplicada, São Leopoldo, BR–RS, 2008. Orientador: Jung, Claudio Rosito.

1. Foco de atenção. 2. Detecção de orientação de pessoas. 3. Estimativa de cabeça e postura. 4. Rastreamento de pessoas. 5. Processamento de imagens. I. Jung, Claudio Rosito. II. Título.

Dedico esta dissertação especialmente à minha mãe Marli,
meu pai Nei,
meu irmão Vinícius,
e minha noiva Aline.

Agradecimentos

Gostaria inicialmente de agradecer ao meu pai Nei, minha mãe Marli e meu irmão Vini pelo suporte e motivação para realizar esse trabalho. Aos meus sogros Joemir e Inelve e ao meu cunhado Dudu. À minha tia Erli, e primos Ricardo e Renato.

Um agradecimento mais que especial à minha noiva Aline, pelo amor e companheirismo demonstrados nesses dois anos de mestrado. Mas principalmente por me fazer acreditar no meu sonho de ser mestre. Muito obrigado meu amor.

Ao meu orientador Cláudio Jung, pela total compreensão dos problemas que enfrentei durante o mestrado, mas principalmente pelo apoio e motivação.

Queria agradecer ainda aos meus amigos de longa data Francesco Monttemagiore, Otavio Gaspareto, Guilherme Holsbach, Juliano Bergamini, Georgenes Zapalaglio e Leandro Motta. Ao pessoal da Hewlett-Packard pelas constantes conversas sobre meu trabalho. Em especial, ao meu amigo Eduardo Basso, que sempre me ajudou quando precisei.

Obrigado aos que participaram das filmagens e à todos que de alguma forma ou outra acabaram ajudando no meu trabalho.

Por fim, agradeço a Deus por ter me dado saúde e forças para seguir em frente e nunca desistir.

Resumo

O crescente uso de câmeras de circuito fechado de televisão permite que a atividade humana fique sob constante monitoramento. Adquirir informação sobre essas pessoas em uma seqüência de vídeo é de fundamental importância para diversas áreas como segurança, publicidade entre outras. Entre os fatores a serem analisados, a direção de observação da pessoa é a que fornece o melhor indicativo de sua intenção ou objetivo. Apesar de existirem trabalhos com o intuito de detectar a orientação, muitos sistemas usam apenas imagens de alta qualidade e captura com a pessoa próxima da câmera, usando traços da face com olhos e boca. Em imagens de mais baixa resolução, a face é representada por apenas alguns pixels, e a informação geométrica é comprometida.

Esta dissertação apresenta um novo método para determinar a orientação global de uma pessoa em uma seqüência de vídeo. O método consiste em uma abordagem estatística baseada na correlação cruzada de padrões de pele da face de um quadro atual com padrões previamente aprendidos em uma etapa anterior de treinamento. O sistema proposto combina a informação de movimentação à orientação fornecida pela face, gerando assim uma orientação global mais robusta. O sistema discretiza a informação de orientação em oito diferentes orientações baseadas na rosa dos ventos. A fim de demonstrar a validade do método, é apresentado ainda um protótipo do método proposto sendo executado sobre uma seqüência de vídeo, com pessoas em posições variadas gravadas exclusivamente para este trabalho. O método foi avaliado quantitativamente utilizando matrizes de confusão.

Palavras-chave: Foco de atenção, Detecção de orientação de pessoas, Estimativa de cabeça e postura, Rastreamento de pessoas, Processamento de imagens.

TITLE: "A METHOD FOR DETECTING PEOPLE ORIENTATION IN VIDEO SEQUENCES"

Abstract

The increasing use of CCTV systems set the monitoring of human activity to higher levels. Such monitoring has several applications, ranging from surveillance to publicity, among others. One important aspect to analise in such video sequences is the focus of attention, which is closely related to the orientation of the person in the scene. Despite the existence of some works to detect the focus of attention (orientation), most of them are focused on higher resolution images, and explore geometrical features of the face. However, this task is more complex for lower resolution images, in which the face is captured with only a few pixels, and geometrical features are not salient.

This work presents a new method to determine the orientation of people monitored by a video camera. The proposed approach is based on cross-correlations between orientation templates of the face using skin color detection, combined with an estimate of orientation based on displacement vectors obtained with tracking algorithms. The possible orientation vectors are discretized into eight directions. A prototype of the proposed model was tested in an indoor video sequence, and results were quantitatively evaluated using confusion matrices.

Keywords: Gaze Detection, Head and Pose Estimation, People Tracking, Image Processing.

Sumário

Lista de Abreviaturas

FPS	Quadros por Segundo (*Frames per Second*)
SSD	Soma das Diferenças Quadráticas (*Sum of Squared Differences*)
WVFOA	Foco de Atenção Visual sem Curso Definido (*Wandering Visual Focus of Attention*)
MCMC	Monte Carlo via Cadeias de Markov (*Markov Chain Monte Carlo*)

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O olhar humano exerce um papel fundamental na comunicação, por exemplo, fornecendo dicas do interesse e intenção das pessoas. Com o uso cada vez maior de circuitos fechados de TV, onde a atividade humana está sob constante observação, determinar para onde uma pessoa está olhando tornou-se um dos principais indicativos para um observador entender o que está ocorrendo na cena. Aplicações que fazem uso deste tipo de tecnologia são várias, como sistemas de vigilância, eventos esportivos ou até mesmo por empresas de publicidade que pretendem monitorar o comportamento de pessoas em frente a um *outdoor*. Recentemente, sistemas de monitoramento para o reconhecimento da atividade humana tornaram-se área de muito interesse para pesquisas em visão computacional, sendo classificados como – sistemas inteligentes de monitoramento por vídeo. Além disso, analisar o olhar humano a partir de sistemas deste tipo é um grande desafio, devido à perspectiva de visão da câmera e tipo de imagem obtida [Buccolieri et al., 2005].

Conforme [Voit et al., 2007], o uso de equipamentos instalados em pessoas que participam da cena é proibitivo em cenários da vida real. Por esse motivo, um dos grandes desafios em sistemas que pretendem analisar o comportamento humano é como lidar com as características do ambiente em questão. Podemos notar uma grande variedade de sistemas com o propósito de controlar grandes ambientes externos com muita população (ex. estádios de futebol e estacionamentos), outros construídos para ambientes internos com poucas pessoas (ex. salas de aula e laboratórios) e ainda sistemas que visam monitorar apenas uma pessoa frente a um objeto (ex. anúncios publicitários). Em todos esses cenários, notamos diferentes configurações de posicionamento e qualidade da câmera, que influenciam na imagem final capturada. Ainda, alguns casos são controlados, sem variação na luz ambiente, o que não é verdade para ambientes externos, por exemplo.

Entretanto, a maior parte dos sistemas mencionados acima visa apenas detectar e rastrear a pessoa na cena, ou seja, o *tracking*. Isso significa que determinar o olhar de uma pessoa em imagens de segurança é um

problema desafiador que vinha recebendo pouca ou nenhuma atenção nos últimos anos [Robertson and Reid, 2006a], mas mesmo assim podemos ver alguns trabalhos preliminares sobre este problema específico como [Robertson et al., 2005]. Ainda, o problema de como detectar para onde uma pessoa está olhando em imagens de baixa resolução e onde a câmera não é frontal tem sido tema de pesquisas recentes. Esse fato mostra que algumas técnicas tradicionais de processamento de imagens para detecção do olhar de uma pessoa não são aplicáveis para câmeras de vigilância, o que impulsiona cada vez mais os pesquisadores a investigarem novos métodos para realizar esta tarefa.

1.1 Motivação

Analisar o olhar de uma pessoa significa desenvolver um sistema capaz de analisar o movimento de seus olhos. Em visão computacional esta tarefa é possível assumindo a disponibilidade de imagens de alta resolução. Entretanto, este não é o tipo de imagem utilizada quando se pretende monitorar pessoas em grandes ambientes. Exemplos são as câmeras de vigilância encontradas em *shopping centers*, eventos esportivos ou até mesmo por agências de publicidade querendo avaliar se as pessoas estão prestando atenção em um determinado anúncio ou cartaz exposto ao público. Além disso, do ponto de vista psicológico, o olhar está diretamente ligado a orientação da cabeça [Langton et al., 2000]. Seres humanos usam três fatores para determinar o olhar de outra pessoa: primeiro a direção dos olhos, segundo a orientação da cabeça e terceiro a orientação dos ombros. Os olhos recebem prioridade, mas a direção geral do olhar é estimada baseada na relação dos três fatores. Assim, quando o olhar não pode ser medido a partir dos olhos, ele ainda pode ser estimado pela posição da cabeça. Isto é o que pode ser feito quando se está tratando imagens de baixa resolução, onde os olhos não são suficientemente visíveis [Ba, 2007]. Além disso, quando uma pessoa foca sua atenção em um determinado ponto da cena, ela tende a alinhar seu corpo, cabeça e olhos por questões de conforto.

Dentre os trabalhos existentes para detecção do olhar humano, a maioria trabalha com a extração de traços do rosto como olhos e boca, para obter uma idéia de onde a pessoa está olhando como [Zhu and Yang, 2002, Viola and Jones, 2001, Lienhart and Maydt, 2002]. Como já mencionado, as imagens obtidas pela câmera precisam ser de alta qualidade para que isso seja possível, pois as características da face precisam estar bem definidas. Além disso, muitos dos métodos existentes assumem uma câmera frontal e não contemplam grandes variações de movimentos da cabeça. Outro fator interessante é que nesse tipo de imagem a cabeça ocupa praticamente toda a imagem, ou seja, não possui interferências de objetos

do fundo da cena. Esses métodos são normalmente usados em sistemas de interação homem-computador, com comprovada funcionalidade em robótica, por exemplo [Chella et al., 2005] e [Sakaue et al., 2006]. Porém, esse não é o tipo de imagem que será tratada nesse trabalho e, portanto, um método precisa ser proposto para a detecção do foco do olhar humano.

(a) (b)

Figura 1.1 – (a) Detecção facial utilizando características da face para imagens de perto e câmera frontal [Lienhart and Maydt, 2002]. (b) Imagem capturada com câmera de vigilância, onde os traços da face não são visíveis devido à imagem de longa distância e com baixa qualidade

Na Figura 1.1(b), vemos uma pessoa caminhando na cena, mas qual sua intenção ou objetivo? Perguntas como estas podem ser respondidas através da análise do movimento desta pessoa, e conseqüentemente analisando um possível comportamento. Ainda, é preciso ter a mesma resposta se a pessoa estiver parada na cena, observando algum objeto, por exemplo. Imagine que uma empresa de publicidade posicionou a câmera no topo de um *outdoor* e necessita verificar se pedestres estão prestando atenção neste anúncio. Em casos como este, onde a pessoa está parada, faz-se uso da postura da cabeça. Como mencionado anteriormente, a orientação da cabeça fornece uma excelente indicativa do foco de atenção desta pessoa. Desta forma, determinar a orientação da cabeça é de fundamental importância neste tipo de aplicação.

Entre fatores que contribuem para uma estimativa do foco de atenção de uma pessoa, destacam-se os seguintes:

- Deslocamento na cena: através do *tracking*, pode ser feita uma análise quadro a quadro da cena e analisar o quanto uma pessoa moveu-se na cena. Desta forma, obtém-se um vetor deslocamento que serve como indicativo da orientação da pessoa, ou seja, onde ela estava e para onde está indo. Na verdade, alguns trabalhos como [Morellas et al., 2003] e [Grimson et al., 1998] determinam a orientação apenas pelo vetor deslocamento.

- Posição da cabeça: um outro fator que merece bastante atenção é a posição/orientação da cabeça. Estudos do comportamento

humano [Zhao et al., 2002, Hu et al., 2005, Smith et al., 2006a, Odobez and Ba, 2007] comprovaram que a posição ou postura da cabeça é um indicativo de onde a pessoa está olhando, ou ainda, qual seu foco de atenção. Em imagens de baixa qualidade, em que o tamanho da cabeça obtida gira em torno de 20 a 40 *pixels* [Robertson and Reid, 2006a] e traços do rosto não são facilmente visíveis, a postura da cabeça é um excelente indicativo para se determinar a orientação da pessoa.

- Postura do corpo: existem autores [Buccolieri et al., 2005, Spagnolo et al., 2003, Rosales and Sclaroff, 2000] que analisam o corpo inteiro da pessoa a fim de se obter uma postura que caracterize uma ação ou comportamento. Como citado anteriormente por [Ba, 2007], a postura dos ombros, por exemplo, é um indicativo da orientação da pessoa.

A idéia de criar um método que possa combinar pelo menos dois dos itens acima parece ser promissor. O trabalho de [Robertson and Reid, 2006a] é um bom exemplo disso, no qual o autor considera a direção do movimento e a postura da cabeça. A Seção 2.2 apresenta mais detalhadamente este e outros trabalhos relacionados, que de alguma forma comprovam o interesse de pesquisas recentes com esse tipo de abordagem. O fato de poder contribuir nessa área, investigando e fornecendo novos métodos ou até mesmo técnicas para definir a orientação de uma pessoa em vídeo, certamente é um incentivo para este trabalho. Assim, espera-se que os resultados obtidos nesse trabalho sirvam de entrada para futuras pesquisas relacionadas com este tema.

1.2 Objetivos e Contribuição

Um dos principais objetivos deste trabalho é estudar e desenvolver métodos para detectar a direção de orientação de pessoas em seqüências de vídeo. Como será discutido na Seção 2.1.4, algoritmos consagrados de detecção facial não são totalmente aplicáveis ao tipo de imagens em estudo neste trabalho, portanto, novos métodos precisam ser propostos. Junta-se a isso o fato de que pessoas estão cada vez mais sob constante monitoramento através de câmeras de vigilância, além do crescente número de artigos publicados sobre esse assunto como [Ba and Odobez, 2006, Xie and Lin, 2007, Voit et al., 2007, Pflugfelder and Bischof, 2007]. Esse fato certamente é um dos motivadores para esta dissertação. Ainda, faz parte deste estudo uma revisão dos principais trabalhos sobre detecção de orientação, técnicas existentes e problemas que precisam ser enfrentados.

Como principal resultado deste estudo, esta dissertação apresenta um método para detecção da orientação de pessoas em seqüências de vídeo, abordando

o fato das imagens capturadas serem semelhantes às imagens de câmeras de vigilância tradicionalmente usadas em monitoramento de estabelecimentos comerciais. Sistemas tradicionais usam abordagens analíticas, onde o olhar humano é determinado pelas características da face da pessoa. O grande problema é que estes métodos requerem imagens de alta resolução e câmeras frontais, ou seja, existe um *trade-off* entre a exatidão e resolução da imagem em abordagens analíticas. Porém, esse tipo de imagem é muito diferente das que serão vistas nesse trabalho, já que em quase 100% dos casos torna-se impossível achar características no rosto das pessoas usando câmeras de vigilância.

A solução proposta emprega algumas técnicas conhecidas e trabalhos recentes promissores de processamento de imagens e visão computacional como auxílio nas etapas de detecção e rastreamento de pessoas na cena, por exemplo [Jacques et al., 2005] e [Cezar Silveira Jacques et al., 2006]. Pretendeu-se tratar a questão de pessoas paradas na cena, onde a estimativa baseada na movimentação torna-se inválida e as únicas informações presentes são posturas do corpo e cabeça. Muitos métodos estimam a direção da orientação baseando-se apenas na informação provida pelo *tracking*, que tendem a apresentar sérios problemas quando quando uma pessoa pára na cena e gira seu corpo/cabeça sem se mover.

Outro objetivo deste trabalho é a implementação de um protótipo em Matlab [IET, 2005] para a validação do método proposto. Através de seqüências de vídeo gravadas com diferentes pessoas, pretende-se ainda mostrar a aplicabilidade, problemas encontrados e motivar os trabalhos futuros. A avaliação, por sua vez, será composta principalmente por métricas quantitativas que permitam medir o grau de acerto nas classificações de orientação das pessoas na cena.

1.3 Metodologia

A seguinte metodologia foi aplicada para o desenvolvimento desta dissertação:

- Revisão bibliográfica relativa a soluções para detecção de orientação de pessoas em seqüências de vídeo, focando principalmente em trabalhos que não fazem uso de câmera frontal e onde as imagens normalmente são de baixa ou média resolução. Por exemplo: câmeras de vigilância;
- Definição dos cenários alvo e requisitos para a construção de um sistema de detecção de orientação que respeitasse os aspectos do item acima;
- Escolha das técnicas a serem utilizados em cada etapa do desenvolvimento deste trabalho;
- Definição e formulação matemática do método proposto: descrição da arquitetura e de seus principais componentes;

- Implementação de um protótipo em Matlab a ser executado em um ambiente de simulação;
- Execução e avaliação dos experimentos para validação do modelo proposto;
- Análise da contribuição do método e proposta de possíveis áreas de atuação para trabalhos futuros.

1.4 Organização da Dissertação

O restante deste documento encontra-se organizado como segue. O Capítulo 2 apresenta uma revisão bibliográfica da área de processamento de imagens e visão computacional voltadas para a detecção de orientação de pessoas em seqüências de vídeo, discutindo ainda os trabalhos relacionados mais relevantes para esta dissertação. O Capítulo 3 descreve o método proposto, tratando desde os cenários alvo e técnicas usadas, até a formalização de cada etapa do método. No Capítulo 4, são apresentados aspectos de implementação do método proposto, métricas usadas para avaliação e os resultados encontrados. Este capítulo inclui também exemplos resultantes de um protótipo construído para mostrar o funcionamento do método. Por fim, o Capítulo 5 finaliza a dissertação com considerações finais e uma proposta de trabalhos futuros.

Capítulo 2

Aspectos Gerais sobre Detecção de Orientação

Este capítulo visa introduzir ao leitor os aspectos gerais sobre detecção de orientação de pessoas em seqüências de vídeo, apresentando os principais fundamentos e técnicas de processamentos de imagens e visão computacional relacionadas com esse assunto. Desta forma, espera-se formar uma base teórica e conceitual para o entendimento deste trabalho. Primeiramente, será apresentada uma terminologia comum para a área. Na seqüência, serão discutidas as principais etapas de uma arquitetura genérica de detecção de orientação, focando nas utilizadas nesta dissertação. Por fim, são apresentados trabalhos relacionados com o tema e que formam a base do desenvolvimento do presente trabalho.

2.1 Fundamentos e Terminologias

2.1.1 Arquitetura genérica

Conforme [Hu et al., 2004], o *framework* de processamento de cenas dinâmicas inclui os seguintes estágios: modelagem do ambiente, detecção do movimento, classificação de objetos em movimento, rastreamento, entendimento e descrição de comportamentos, identificação humana e fusão de dados de múltiplas câmeras.

Modelos de ambientes podem ser classificados em 2D – no plano da imagem, ou em 3D – em coordenadas do mundo real. Quanto à câmera, pode ser dinâmica ou estática. Na etapa de detecção ocorre a segmentação e classificação das regiões correspondentes às pessoas em relação ao resto da imagem. Nessa etapa eliminam-se objetos que não são de interesse, como carros e prédios, por exemplo. Os processos seguintes de *tracking* e entendimento do comportamento são totalmente dependentes disso. Essa etapa de classificação pode não ser necessária em algumas aplicações onde se sabe que os objetos em movimento tratam-se somente de pessoas.

Em uma visão de alto nível, a arquitetura genérica pode ser resumida nas seguintes etapas, descritas na Figura 2.1.

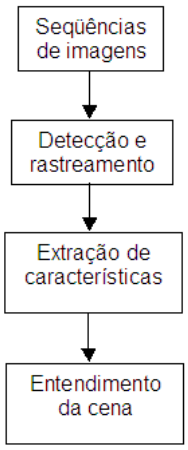


Figura 2.1 – Arquitetura genérica de visão computacional para processamento de cenas

- **Seqüência de imagens**: Etapa inicial do sistema, onde é feita a captura de vídeo;
- **Detecção e rastreamento**: Faz parte dessa etapa a detecção e acompanhamento dos objetos na cena, explicada mais detalhadamente abaixo;
- **Extração de características**: Nessa etapa os objetos são processados a fim de se eliminar os que não são de interesse da aplicação, extraindo-se características apenas dos objetos alvo. Faz-se uso de técnicas de refinamento para, por exemplo, calcular a altura, posição da cabeça de uma pessoa na cena e calcular a cor de pele.
- **Entendimento da cena**: Uma vez tendo esses dados extraídos, cabe a cada aplicação determinar o significado dos dados obtidos na etapa anterior. A análise da saída dessa etapa é que vai confirmar se o sistema está funcionando corretamente ou não. No contexto desse trabalho, faz parte dessa etapa a detecção de orientação propriamente dita.

O rastreamento de objetos em uma seqüência de vídeo, ou *tracking*, consiste em determinar a posição de cada objeto de interesse na cena ao longo do tempo. Há uma grande variedade de algoritmos de acompanhamento de objetos e pessoas existentes na literatura como [Wren et al., 1997, Haritaoglu et al., 2000, Pai et al., 2003, Adam et al., 2006, Ramanan et al., 2007], e o desempenho de cada técnica depende bastante do tipo de câmera (posição, estática ou em movimento, etc.) e as características dos objetos a serem acompanhados. Um bom *survey*

para acompanhamento de objetos pode ser encontrado em [Yilmaz et al., 2006]. Em particular, quando câmeras estáticas são empregadas (como no caso deste trabalho), algoritmos de subtração (ou remoção) de fundo são normalmente utilizados para detectar objetos em movimento.

Na subtração de fundo, basicamente se obtém um modelo matemático do fundo da cena, e subtrai-se cada quadro do vídeo desse modelo. Os *pixels* que possuem uma diferença significativa com relação ao fundo da cena são associados a objetos em movimento (*foreground*) e os demais são classificados como fundo da imagem (*background*). As abordagens para subtração de fundo diferenciam-se umas das outras pelo tipo de modelo usado para aproximar o fundo e pelas estratégias de atualização do modelo (para adaptação a mudanças de iluminação, por exemplo). Além disso, há o problema gerado pela falsa identificação de sombras como objetos do *foreground*.

Há diversas abordagens para o problema da remoção de fundo, tais como [Haritaoglu et al., 2000, Wang et al., 2005b, Tian et al., 2005, Jacques et al., 2005, Cezar Silveira Jacques et al., 2006], entre várias outras. Neste trabalho, adotou-se o modelo adaptativo de *background* para seqüências de vídeo monocromáticas proposto em [Cezar Silveira Jacques et al., 2006], que inclui detecção de sombras e adaptação a variações suaves de luminosidade. Ainda entre os principais fatores que motivaram a escolha desse método estão o fato apresentar uma boa relação entre acurácia e tempo de execução, além de disponibilidade do código. A seção 3.2 descreve o método mais detalhadamente, mas um resultado pode ser visto na Figura 2.2.

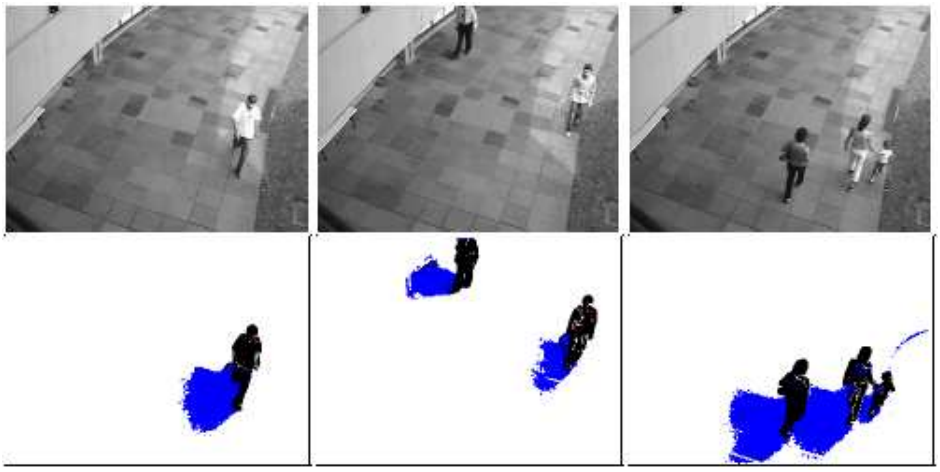


Figura 2.2 – Detecção e rastreamento

Acima, quadros extraídos de uma seqüência de vídeo. Abaixo, objetos detectados como *foreground* (preto) e *pixels* de sombra (azul)

Salienta-se ainda que existem diversas outras abordagens para os problemas de remoção de fundo e acompanhamento de objetos, como os baseados em diferença

temporal [Yang et al., 2004] [Wang et al., 2005a] e fluxo ótico [Lipton et al., 1998]. Entretanto, a detecção e o acompanhamento dos objetos não são os focos principais deste trabalho, mas ferramentas utilizadas para a detecção da orientação das pessoas em movimento capturadas por câmeras estáticas.

2.1.2 Detecção de orientação

Ao falarmos em orientação de uma pessoa, podemos muitas vezes notar o termo foco de atenção visual. Segundo [Stiefelhagen, 2002], obter o conhecimento sobre o foco de atenção de uma pessoa é o maior passo para um melhor entendimento do que usuários fazem, o que e com quem interagem ou até a que se referem. O estudo do foco de atenção usualmente requer a determinação da direção de observação, ou seja, o olhar da pessoa. Como dito, a posição da cabeça é um bom indicador do foco de atenção da pessoa [Ba, 2007].

2.1.3 Estimativa de postura da cabeça

A detecção e rastreamento da cabeça são componentes essenciais em aplicações de vídeo relacionadas com o entendimento do comportamento humano. São comumente usados como um primeiro passo, antes da aplicação de algoritmos para tarefas de alto nível, como reconhecimento de face e expressões faciais ou estimativa da direção de observação [Ba and Odobez, 2004].

A modelagem da cabeça não é uma tarefa trivial, uma vez que sua aparência pode variar devido a diversos fatores:

- Variação na escala: o tamanho da cabeça pode variar conforme a pessoa move-se na cena, ficando mais perto ou mais distante da câmera;
- Orientação da cabeça: a pessoa realiza movimentos de cabeça, olhando para os lados ou simplesmente caminhando na cena, o que causa mudança de orientação da cabeça em relação à câmera;
- Condições de iluminação: dependendo das fontes de iluminação, a aparência da cabeça está sujeita a variações uma vez que nem todas as partes da cabeça refletem a luz igualmente;
- Características físicas: pessoas diferem umas das outras fisicamente, como presença de barba, cabelo, bigode, tom de pele, entre outros fatores;

2.1.4 Técnicas de detecção de faces baseadas na cor de pele

Entre os métodos de detecção de faces baseados em características da face, o uso de cor da pele[1] ganhou muita popularidade nos últimos anos. A cor permite rápido processamento e é um meio bastante robusto às variações geométricas do padrão da face, além de permitirem a detecção em imagens de menor resolução, onde características espaciais da face podem não estar visíveis.

Ao construir um sistema que usa a cor da pele como característica a ser encontrada, os pesquisadores usualmente enfrentam três problemas que são: o espaço de cores a ser usado, como modelar a distribuição de cor da pele e de que forma será o processamento do resultado da segmentação da cor [Vezhnevets et al., 2003]. Apesar da cor da pele de uma pessoa variar em um intervalo bastante grande, de tons claros a tons escuros, estudos mostram que as diferenças entre os valores dos *pixels* de pele recaem largamente nas intensidades e não na cor [Ba, 2007]. Diversos espaços de cores tem sido usados para a modelagem dos *pixels* de pele, assim como métodos com o mesmo propósito.

Os itens a seguir descrevem com detalhes os espaços de cores mais utilizados:

- **RGB**: *Red, Green, Blue*, foi originado a partir de aplicações para monitores CRT, quando era conveniente descrever cores como uma combinação de três raios coloridos (vermelho, verde e azul). É um dos espaços de cores mais usados para armazenamento de imagens. Entretanto, devido à alta correlação entre os canais, significante percentual de não uniformidade, mistura de crominância e luminância fazem do RGB uma escolha não muito favorável para algoritmos de reconhecimento baseados na cor [Vezhnevets et al., 2003]. Apesar disso, esse modelo pode ser visto em [Brand and Mason, 2000] e [Jones and Rehg, 1999].

- **RGB normalizado**: É uma representação simples e facilmente obtida a partir da normalização do RGB. A soma dos três componentes normalizados é igual a 1, onde o terceiro componente pode ser omitido pois não traz nenhuma informação significativa, reduzindo assim a dimensão do espaço de cor. Usado em trabalhos como [Soriano et al., 2000] e [Brown et al., 2001], é uma representação interessante para superfícies sem brilho, pois enquanto ignorando a luz ambiente, este espaço de cores é invariante à mudanças da orientação relativamente à fonte de luz [Skarbek and Koschan, 1994]. A equação abaixo define a normalização do RGB.

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}, \quad b = \frac{B}{R+G+B}. \tag{2.1}$$

[1] Conforme [Ba, 2007], o termo "cor da pele" não é uma propriedade física de um objeto, pelo contrário, é um fenômeno perceptivo e, portanto, é subjetivo ao conceito humano.

- **HSV**: *Hue, Saturation, Value*, é caracterizada por ser uma transformação não-linear do sistema de cores RGB. H define a cor dominante (como vervelho, verde, violeta e amarelo) de uma área, S mede a policromia de uma área em proporção ao seu brilho e V é relacionado à luminância da cor. Por ser intuitivo, é bastante popular em trabalhos que tratam a segmentação pela cor da pele [Zarit et al., 1999, Sigal et al., 2000]. A conversão entre os espaços de cor RBG e HSV é definida pela equação abaixo.

$$\begin{aligned} H &= \arccos \frac{\frac{1}{2}((R-G)+(R-B))}{\sqrt{((R-G)^2+(R-B)(G-B))}}. \\ S &= 1-3\frac{\min(R,G,B)}{R+G+B}. \\ V &= \frac{1}{3}(R+G+B). \end{aligned} \tag{2.2}$$

- **YCrCb**: É um sinal RGB não linear codificado, comumente usado por estúdios de televisões européias e para compressão de imagens. A cor é representada por *luma* (luminância para um RGB não linear), construído como uma soma ponderada dos valores RGB, e dois diferentes valores C_r e C_b que são formados subtraindo-se *luma* dos componentes vermelho e azul do RGB. Como pode ser visto na equação a seguir, a simplicidade da transformação e a separação dos componentes de luminância e crominância tornam o YCrCb muito atrativo para a modelagem de cor de pele [Phung et al., 2002, Hsu et al., 2002, Chai and Bouzerdoum, 2000].

$$\begin{aligned} Y &= 0.299R+0.587G+0.114B, \\ C_r &= R-Y, \\ C_b &= B-Y. \end{aligned} \tag{2.3}$$

A seguir estão descritos descritos os três principais tipos de métodos usados para se modelar a cor da pele humana. O objetivo aqui é introduzir uma métrica que diferencie *pixels* de pele para não pele [Vezhnevets et al., 2003].

- **Explícitos**: Através de um conjunto de regras, um classificador baseado em algum espaço de cor define explicitamente os limites do que pode ser pele. Este método de modelagem tem atraído muitos pesquisadores devido à sua simplicidade e excelentes resultados, como pode ser visto pela equação usada por [Kovac et al., 2003]. A principal dificuldade do método é definir um bom espaço de cor e boas regras de decisão

empiricamente. Segundo [Kovac et al., 2003], um *pixel* com coordenadas (R, G, B) é classificado como pele se:

$$\begin{aligned} &R > 95 \text{ e } G > 40 \text{ e } B > 20 \text{ e} \\ &\max\{R,G,B\} - \min\{R,G,B\} > 15 \text{ e } |R-G| > 15 \text{ e} \\ &R > G \text{ e } R > B \end{aligned} \tag{2.4}$$

- **Não paramétricos**: A idéia de métodos não paramétricos de modelagem de pele é estimar uma distribuição de cor de pele a partir de dados de treinamento sem derivar um modelo específico de cor de pele. Como resultado, normalmente se tem um mapa de probabilidades de pele, onde é definido um valor que caracteriza a probabilidade de cada ponto para um espaço de cor. Os métodos não paramétricos mais conhecidos são os baseados em histogramas [Brand and Mason, 2000, Gomez, 2002].

- **Paramétricos**: Modelos paramétricos definem que a distribuição de cor de pele pode ser modelada, por exemplo, por uma função densidade de probabilidade baseada em uma Gaussiana elíptica [Vezhnevets et al., 2003, Hsu et al., 2002, Menser and Wien, 2000], definida como:

$$p(\boldsymbol{c}|\text{skin}) = \frac{1}{(2\pi)^{3/2}|\boldsymbol{\Sigma}_s|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{c}-\boldsymbol{\mu}_s)^T\boldsymbol{\Sigma}_s^{-1}(\boldsymbol{c}-\boldsymbol{\mu}_s)\right), \tag{2.5}$$

  onde $\boldsymbol{c}$ é um vetor de cor, $\boldsymbol{\mu}_s$ é a média da distribuição que representa a classe de *pixels* da pele, e $\boldsymbol{\Sigma}_s$ é a matriz de covariância, estimados a partir dos dados de treinamento. [Menser and Wien, 2000] faz uso direto do valor de $p(\boldsymbol{c}|\text{skin})$ para definir a probabilidade de $\boldsymbol{c}$ ser pele. Já [Terrillon et al., 2000] busca o mesmo propósito, mas calculando a distância Mahalanobis do vetor de cores $\boldsymbol{c}$ para o vetor de médias $\boldsymbol{\mu}_s$, dada uma matriz de covariância $\boldsymbol{\Sigma}_s$.

Um desafio que os pesquisadores enfrentam é como escolher o melhor espaço de cores para detecção de pele, ou melhor, descobrir se existe um espaço ideal para a classificação de pele. Modelos explícitos são rápidos, eficientes e simples, assim são usados por muitos pesquisadores exatamente por esses motivos, mas exigem uma boa definição empírica dos limites do que pode ser pele ou não. O trabalho de [Kovac et al., 2003] é um bom exemplo de uso desse modelo. Modelos não paramétricos como os histogramas e modelos paramétricos como os baseados em distribuições Gaussianas multivariadas mostraram-se bem adaptados para *pixels* de pele, como pode ser visto no trabalho de [Phung et al., 2005], onde é apresentado um estudo comparativo de várias técnicas de segmentação baseados na cor da pele. Conforme [Ba, 2007], existem duas desvantagens principais relacionadas com o uso

da cor da pele. Primeiro, é que a cor da pele pode ser vista em outras partes ou objetos da imagem. Segundo, é que modelos baseados na cor da pele são sensíveis a condições de iluminação. A Seção 3 apresenta como o método proposto neste trabalho tenta contornar esse problema.

2.2 Trabalhos relacionados

Esta seção foca em trabalhos voltados à detecção de orientação ou detecção do foco de atenção de pessoas em seqüências de vídeo.

No trabalho de [Smith et al., 2006a], os autores visam determinar o foco de atenção de pessoas que caminham por uma área sem um destino ou rumo definido, para um cartaz na vitrine de uma loja. Os autores chamam esse tipo de foco de atenção de WVFOA (*Wandering Visual Focus of Attention*) e citam esse trabalho como o pioneiro a estudar esse assunto. Basicamente, o WVFOA de uma pessoa visível é definido como estando em um de dois estados possíveis: focado (ele ou ela está olhando para o anúncio) ou não focado (ele ou ela não está olhando para o anúncio). Como pode ser visto na Figura 2.3, pessoas passando focam suas atenções no anúncio de diferentes localizações e com uma variedade de diferentes posturas de cabeça. Para inferir o WVFOA de cada pessoa em cada passo da cena, os autores usam a localização da cabeça e estimativas de postura fornecidas por um filtro MCMC, o qual monitora e mantém identidade de pessoas no tempo, mesmo sobre oclusão. Métodos tradicionais de detecção de faces como visto em [Jones and Viola, 2003] não funcionariam para essa aplicação, uma vez que não mantém identidade das pessoas sobre os quadros do vídeo.

Figura 2.3 – Foco de atenção visual

A Figura 2.4 é uma aplicação desenvolvida para coletar informações estatísticas do número de observadores, tempo gasto olhando o anúncio e número total de pessoas expostas ao anúncio [Smith et al., 2006a].

Como visto nos trabalhos de [Stiefelhagen, 2004, Kruger et al., 2000, Rae and Ritter, 1998, Zhao et al., 2002], algumas abordagens fazem uso de redes neurais para se detectar a orientação horizontal e vertical da cabeça a partir de imagens da face. Abordagens baseadas em redes neurais em geral não requerem um

Figura 2.4 – Configuração experimental para determinação do foco de atenção

rastreamento de características detalhadas da face, pois a região da face inteira pode ser usada para estimar a postura da cabeça, permitindo assim seu uso também em imagens de baixa resolução. Em [Stiefelhagen et al., 2000], os autores definem um modelo baseado em redes neurais para estimar os ângulos de posição da cabeça a partir de imagens da face. Nesta abordagem, imagens pré-processadas da face são usadas como entrada para uma rede neural, a qual é treinada a fim de estimar os ângulos de orientação da cabeça para uma imagem de entrada. A Figura 2.5 mostra um exemplo de aplicação desenvolvida no início das pesquisas para esta dissertação, cujo objetivo era detectar posturas utilizando redes neurais.

Figura 2.5 – Aplicação de redes neurais para detecção de postura.

No trabalho de [Voit et al., 2006], é apresentado um sistema para estimativa da postura de cabeça humana usando múltiplas câmeras. Para cada câmera é aplicada uma rede neural, sendo que a saída é agregada usando-se um filtro Bayesiano. O sistema é avaliado em um sala controlada (Figura 2.6) onde são gravados vídeos de baixa resolução, inclusive com uma iluminação ruim, onde o tamanho da cabeça capturada varia entre 20 e 25 *pixels*.

[Brumitt et al., 2000] foca na relação entre indivíduos através do olhar de uma pessoa com outra durante uma ocasião social. Alguns estudos mais recentes mostram forte evidência de que pessoas naturalmente olham para os objetos com os quais interagem [Maglio et al., 2000, Smith et al., 2006b, Räihäa and Duchowski, 2006]. O trabalho de [Gee and Cipolla, 1994] talvez seja um dos primeiros trabalhos importantes direcionados ao estudo do foco de atenção em uma imagem. Nesse trabalho, o autor usa medidas de estruturas da face (olhos, nariz e boca), onde é extraído um vetor normal que indica para onde a pessoa está olhando. Outro exemplo de trabalho nesse sentido é o de [Yao et al., 2001]. O trabalho

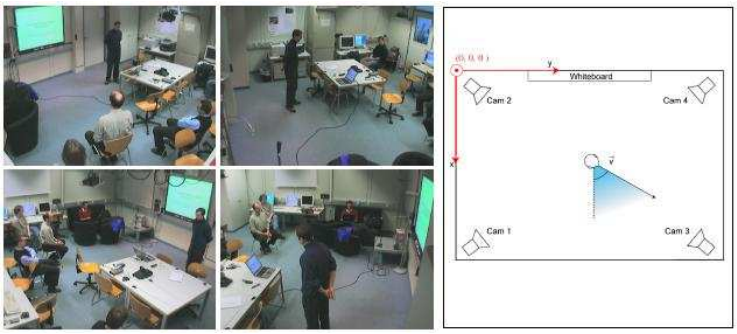


Figura 2.6 – Configuração de uma sala monitorada.

de [Wang and Ji, 2004] usa uma combinação de SVMs (*Support Vector Machines*) para detectar faces com diferentes visões. Essas faces são divididas em sete visões, onde cada uma delas modela uma típica postura sobre uma cena complexa.

A conhecida biblioteca de processamento de imagens OpenCV [Bradski et al., 2005] também implementa um algoritmo de detecção facial, no qual usa uma abordagem estatística para a detecção da face, originalmente desenvolvida por Viola e Jones em [Viola and Jones, 2001], posteriormente analisada e estendida por Lienhart em [Lienhart and Maydt, 2002]. Entretanto, seu uso não é indicado no contexto deste trabalho, pois apresenta bons resultados somente em imagens capturadas com câmera frontal, ou seja, diferentes das usadas nessa dissertação.

O trabalho de [Robertson and Reid, 2006b] foi um dos principais motivadores para esta dissertação, pois visa determinar o foco de atenção em imagens de baixa qualidade. O método proposto pelos autores também é visto em outros trabalhos de mesma autoria como [Robertson and Reid, 2006a] e [Robertson and Reid, 2005]. A proposta dos autores é estimar para onde uma pessoa está olhando em imagens onde a cabeça possui um tamanho entre 20 e 40 *pixels*, ou seja, tipo de imagens normalmente geradas por câmeras de vigilância. São detectados os *pixels* da pele na face para detectar a orientação da cabeça, que é discretizada em 8 diferentes orientações relativas à câmera. Um método rápido de amostragem retorna a distribuição de probabilidade sobre posições de cabeça previamente observadas. A posição do corpo relativa à câmera é aproximada usando a velocidade do corpo, obtida por um *tracking* baseado em cor. Os autores fazem então uma combinação da direção com a informação de postura da cabeça, determinando-se o foco de atenção da pessoa. A Figura 2.7 mostra a aplicabilidade do sistema descrito. Porém, este trabalho não considera a situação da pessoa parada na cena.

[Ba and Odobez, 2004] considera o rastreamento da cabeça e a estimativa de postura como dois problemas acoplados. O autor define uma configuração probabilística onde modelos de cabeça são aprendidos e incorporados em um filtro

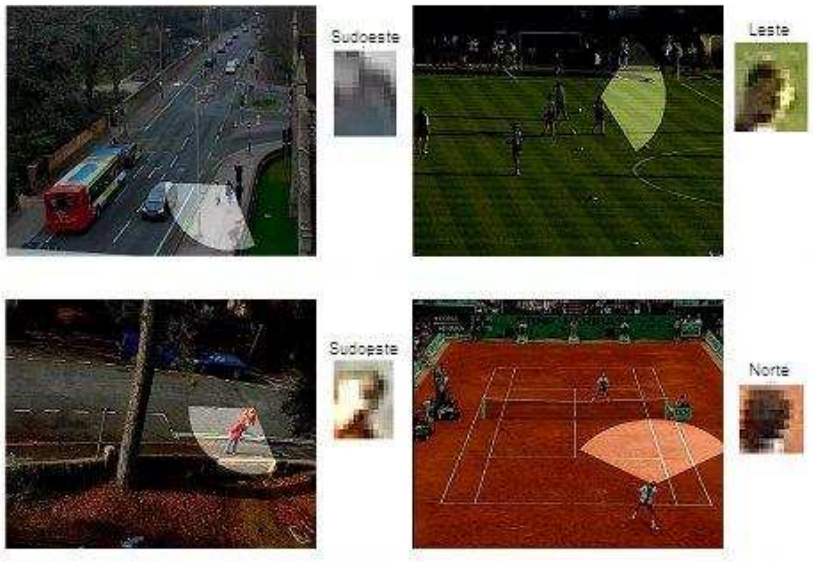


Figura 2.7 – Detecção da postura da cabeça em diferentes cenários.

de partículas de estados misturados. A configuração espacial da cabeça (posição e escala) e sua postura são representados em um modelo combinado de espaço de estados. A distribuição posterior combinada do estado dada a seqüência das imagens é estimada em cada instante e propagada para o próximo instante de tempo usando a dinâmica de estados. Nesse trabalho, o autor trata apenas ângulos em um intervalo de -90 a +90 graus, discretizados em passos de 22,5 graus. A Figura 2.8 mostra um resultado da detecção da orientação da cabeça.

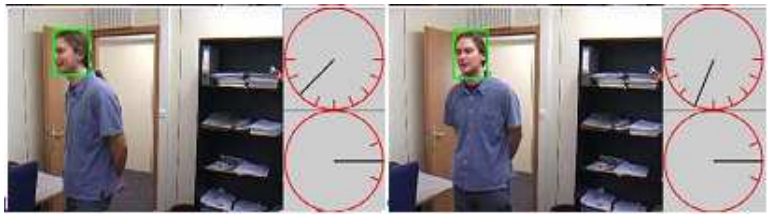


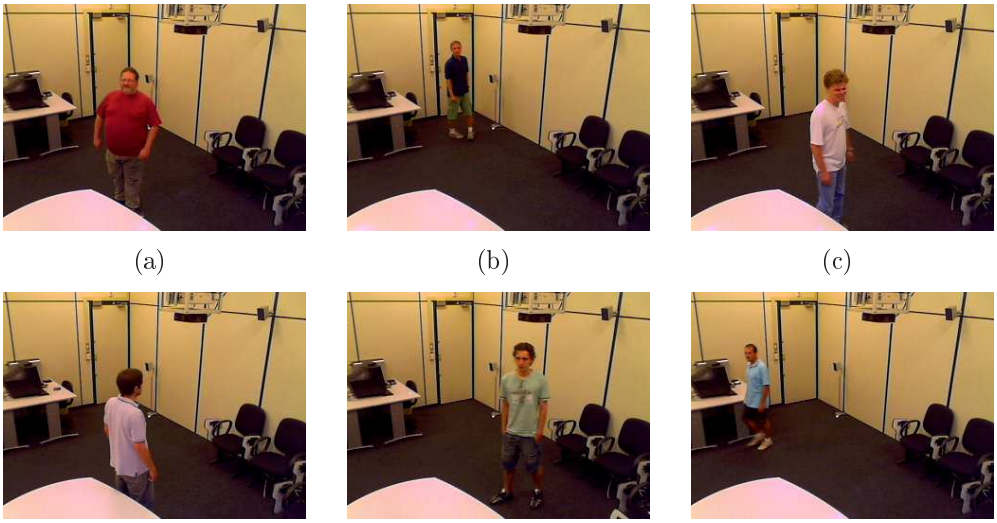
Figura 2.8 – Exemplo de método probabilístico de detecção facial.

Capítulo 3

Método Proposto

3.1 Arquitetura escolhida para o método

O objetivo deste trabalho é detectar a orientação de pessoas capturadas por uma câmera de vídeo estática, posicionada em uma quina superior de uma sala fechada. Tal câmera fornece imagens onde o tamanho da cabeça varia entre 20 a 40 *pixels*, de modo que as características faciais não são proeminentes quando as pessoas estão longe da câmera. Exemplos de imagens capturadas pela câmera do protótipo são ilustradas na Figura 3.1.



(a) (b) (c)

(d) (e) (f)

Figura 3.1 – (a), (c) e (d) Pessoas em posições diferentes e próximas à câmera, onde as pernas estão escondidas por uma mesa. (b) Pessoa está no ponto máximo de distância da câmera. (e) Pessoa um pouco mais afastada da câmera, sendo possível ver o corpo inteiro. (f) Pessoa está em um nível intermediário entre o fundo e a frente da sala.

Com base no tipo de imagens fornecidas pela câmera, optou-se por utilizar a estratégia a seguir. A Figura 3.2 mostra a arquitetura geral do sistema. Pela

figura é possível ver que o sistema é bastante modular, facilitando que diferentes métodos sejam usados além dos propostos nesse trabalho. Permite-se assim uma continuidade das pesquisas e experimentos com o método proposto, sendo possível, por exemplo, substituir o módulo atual de detecção da face por algum outro.

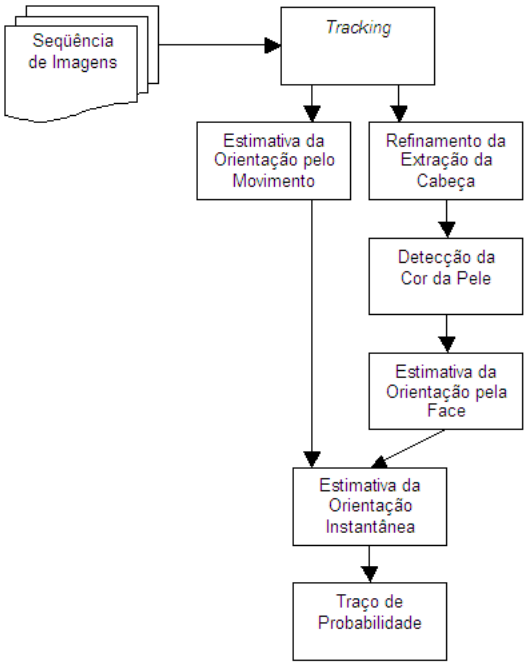


Figura 3.2 – Arquitetura do método proposto.

A idéia central do método proposto é estimar a orientação da pessoa com base em sua movimentação na cena (assumindo que as pessoas normalmente olham para onde estão caminhando) e com base na posição de sua cabeça/face. Essas duas estimativas são então combinadas em uma única medida de orientação.

Para realizar o acompanhamento das pessoas, é aplicado inicialmente um algoritmo de remoção do fundo da cena, que separa discrimina entre *pixels* do fundo da imagem e *pixels* em movimento. O acompanhamento (*tracking*) dos objetos em movimento é realizado através do casamento de máscaras, utilizando a cabeça estimada da pessoa como o identificador da pessoa.

Para determinar a orientação utilizando informação apenas da cabeça, a estimativa desta (obtida inicialmente no processo de *tracking*) é refinada, e a detecção de *pixels* da face com base na informação da cor da pele é obtida. A região da face/cabeça é então centrada e reescalonada para um vetor de tamanho padrão (10×10 *pixels*), garantindo assim que o tamanho da face seja o mesmo tanto para pessoas perto quanto para as que estão mais longe da câmera. Um conjunto de 8 vetores caracterizando faces em 8 diferentes orientações é obtido em um período de treinamento, e comparado com o vetor extraído da pessoa a ser analisada no período de execução através da correlação cruzada.

Finalmente, uma análise de coerência temporal é aplicada, e a informação de movimento é combinada com a orientação da face, fornecendo assim a orientação global da pessoa. Cada uma dessas etapas é descrita em detalhes a seguir.

3.2 Remoção de Fundo

Uma abordagem tradicional para aplicações que usam câmera estática é a subtração (ou remoção) do fundo, que consiste em obter um modelo matemático do fundo estático, e compará-lo com cada novo quadro da seqüência de vídeo. Entretanto, mudanças de iluminação como sombras podem afetar a detecção gerando falsos *pixels* de *foreground*, o que significa que o modelo do fundo da cena precisa ser adaptado de acordo. Em função disso, o modelo escolhido para detecção e rastreamento da pessoa na cena é o proposto em [Cezar Silveira Jacques et al., 2006]. Este modelo implementa um modelo adaptativo de fundo, trata as sombras com bastante eficiência e possui uma boa relação entre desempenho e tempo de execução. Além disso, o código foi fornecido pelo autor exclusivamente para o desenvolvimento desta dissertação. O método escolhido é explicado mais detalhadamente a seguir.

3.2.1 Obtenção do modelo de *background*

Seja $I_t(x, y)$ a intensidade do *pixel* (x, y) em um quadro t. O primeiro passo é denominado período de treinamento, cujo objetivo é extrair um modelo do fundo estático a partir de uma série de quadros. Nesta etapa, consideramos $\{I_1(x, y), ..., I_N(x, y)\}$ os N quadros da seqüência de imagens usados no período de treinamento, a partir dos quais se calculam a mediana $\lambda(x, y)$ e o desvio padrão $\sigma(x, y)$ de cada *pixel* (x, y).

Assumindo que o ruído é uniforme e identicamente distribuído para todos os *pixels*, o histograma de $\sigma(x, y)$ deveria fornecer um um pico saliente na posição do efetivo desvio padrão do ruído da imagem. Entretanto, se houver objetos em movimento durante a fase de treinamento, alguns *pixels* podem apresentar estimativas errôneas do desvio padrão do ruído. Assim, picos secundários no histograma causados por objetos em movimento podem aparecer. De fato, a tendência é que, em tais *pixels* não-estacionários, a estimativa do ruído seja exagerada (pois a diferença de intensidade causada por objetos em movimento tende a aumentar o desvio padrão). Assim, a estimativa σ_m do ruído global da imagem é obtida através da posição do primeiro pico do histograma de $\sigma(x, y)$.

No período de teste, um *pixel* (x, y) é classificado como pertencente ao *foreground* se as diferenças de luminosidade em uma vizinhança de (x, y) forem suficientemente distintas do modelo de fundo $\lambda(x, y)$. Mais especificamente, a

diferença $D_t(x, y)$ entre o quadro em análise $I_t(x, y)$ e o modelo de fundo é dada por

$$D_t(x, y) = |I_t(x, y) - \lambda(x, y)| * A, \quad A = \frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}, \tag{3.1}$$

onde $*$ denota a convolução. Assim, $D_t(x, y)$ é uma média ponderada das diferenças em uma vizinhança 3×3 de cada *pixel* (x, y). Um *pixel* (x, y) é então classificado como *foreground* quando

$$D_t(x, y) > k\sigma_m, \tag{3.2}$$

onde k é um parâmetro ajustável (foi utilizado o valor $k = 6$, conforme sugerido pelos autores).

3.2.2 Eliminação de sombras

A fim de eliminar as sombras existentes na cena, que podem influenciar no resultado do *foreground* detectado, [Cezar Silveira Jacques et al., 2006] assumem que a intensidade de *pixels* de sombra é diretamente proporcional à luz incidente e, conseqüentemente, *pixels* sombreados são versões escaladas (escuras) dos *pixels* correspondentes no modelo de *background*. De forma análoga, o mesmo é assumido para regiões iluminadas.

Seja $R(x, y)$ uma região de 3×3 centralizada em cada *pixel* de *foreground* (x, y). Este *pixel* é classificado como sombreado se:

$$\mathrm{std}_R \left(\frac{I_t(x, y)}{\lambda(x, y)} \right) < L_{\mathrm{std}} \quad \text{e } L_{\mathrm{low}} \leq \left(\frac{I_t(x, y)}{\lambda(x, y)} \right) < 1, \tag{3.3}$$

onde $\mathrm{std}_R \left(\frac{I_t(x,y)}{\lambda(x,y)} \right)$ é o desvio padrão das quantidades $I_t(x, y)/\lambda(x, y)$ sobre a região R, e $L_{\mathrm{std}}, L_{\mathrm{low}}$ são os limiares. Ou seja, L_{std} controla o máximo desvio padrão dentro da vizinhança sendo analisada, e L_{low} previne a classificação incorreta de objetos escuros com intensidades de *pixels* muito baixas como *pixels* sombreados. De forma similar, temos o cálculo para considerar regiões iluminadas:

$$\mathrm{std}_R \left(\frac{I_t(x, y)}{\lambda(x, y)} \right) < L_{\mathrm{std}} \quad \text{and } 1 \leq \left(\frac{I_t(x, y)}{\lambda(x, y)} \right) < L_{\mathrm{high}}, \tag{3.4}$$

onde L_{high} previne a falsa classificação de objetos muito brilhantes como *pixels* iluminados. Os parâmetros sugeridos pelo artigo de [Cezar Silveira Jacques et al., 2006] foram analisados e mostraram bons resultados para o tipo de imagem utilizada nesse trabalho, portanto, foram mantidos como $L_{\mathrm{std}} = 0.05$, $L_{\mathrm{low}} = 0.5$ e $L_{\mathrm{high}} = 1.3$.

3.2.3 Processo de adaptação do *background*

A adaptação do fundo é feita nos *pixels* que são classificados como sombra ou iluminados para um período suficientemente grande de tempo. Seja $C_t(x, y)$ uma máscara binária que retorna 1 se o *pixel* (x, y) é classificado como sombra (ou iluminado) em um quadro t, e 0 caso contrário. Ainda, seja M o tempo de adaptação, ou seja, o número de quadros usados na etapa de adaptação do fundo da cena. Assim, para cada quadro t, o termo $A_t(x, y)$ na equação (3.5) representa o número de vezes que um *pixel* (x, y) foi detectado como sombra ou claridade nos M quadros anteriores.

$$A_t(x, y) = \sum_{i=t-M+1}^{t} C_i(x, y), \tag{3.5}$$

O modelo do fundo é então recomputado no *pixel* (x, y) se:

$$A_t(x, y) \geq pM, \tag{3.6}$$

onde, $0 \leq p \leq 1$ representa a fração mínima do período de adaptação necessário para a adaptação do fundo. Nesse trabalho foi usado um $p = 0.8$ e $M = 50$ quadros da seqüência de vídeo, seguindo a sugestão de [Cezar Silveira Jacques et al., 2006]. Um exemplo de subtração de fundo com este método em um dos quadro do vídeo usado no protótipo desta dissertação pode ser visto na Figura 3.2.3. Após a obtenção do *foreground* e eliminação das sombras, um processo de refinamento (seqüência de abertura e fechamento morfológicos) ainda é aplicado para eliminar pontos de *foreground* isolados e preencher pequenos buracos que podem surgir nos objetos de *foreground* decorrentes da detecção. A Seção 4.2 apresenta uma análise completa dos resultados de todo o processo, desde a etapa de captura até a estimativa da orientação da pessoa.

3.3 Acompanhamento das pessoas

Após a obtenção dos objetos de *foreground* obtidos pela remoção de sombra, é necessário acompanhá-los ao longo do tempo. Conforme visto no capítulo anterior, há diversas abordagens para este problema. Uma das soluções envolve o casamento de máscaras (*template matching*) entre quadros adjacentes da seqüência de vídeo, como adotado nos trabalhos de [Porikli et al., 2006, Adam et al., 2006, Musse et al., 2007]. Dada a máscara identificada no quadro inicial (quando o objeto aparece na cena pela primeira vez), o objetivo desses algoritmos é encontrar a máscara mais similar no quadro adjacente.

Neste trabalho, visa-se extrair uma região retangular em torno da cabeça de

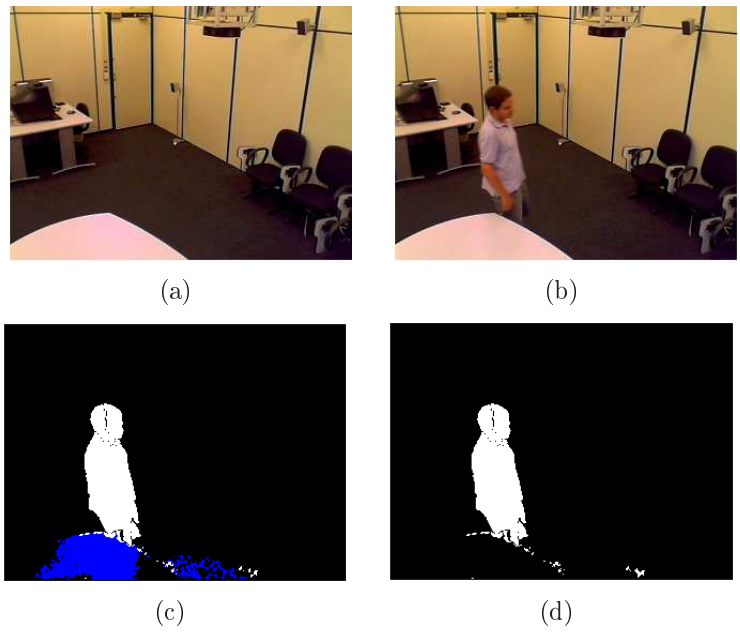


Figura 3.3 – (a) Imagem do fundo da cena para treinamento do *background*. (b) Imagem com uma pessoa caminhando na sala em um quadro t da aplicação. (c) Objeto de *foreground* detectado pelo modelo, com *pixels* de sombra marcados em azul. (d) Imagem com a eliminação das sombras, ainda sem o refinamento para descartar pontos que não fazem parte do objeto.

cada pessoa como sendo a máscara que será acompanhada. Primeiramente, cada componente conexo de objetos do *foreground* (chamados de *blob*) é extraído, e apenas *blob* com área maior do que um limiar T_{area} são mantidos como candidados a pessoas. Conforme a posição da câmera e observação de uma grande seqüência de imagens capturadas com a câmera protótipo, este valor de limiar foi configurado como sendo $T_{\text{area}} = 800$ *pixels*.

Além disso, assume-se que as pessoas surgem apenas nas extremidades da imagem. À medida que a pessoa entra progressivamente na cena, seu *blob* aumenta progressivamente, até que estabilize em área e "descole" das bordas da imagem. Assim, um objeto só é capturado se estiver a uma distância mínima das bordas da imagem.

Dado um conjunto conexo de *pixels* do *foreground* que satisfaça as condições acima, calcula-se o *bounding-box* ao redor do objeto, além de seu ponto mínimo em y e médio em x[1]. Assumindo que a cabeça da pessoa esteja sempre direcionada para a parte superior da imagem e esteja aproximadamente centralizada no seu *blob*, sua posição é inicialmente estimada através do ponto superior-central do *bounding-box*.

Uma vez determinada a máscara que representa o objeto de interesse (no caso, a cabeça da pessoa), foi utilizada uma métrica de casamento baseada na Soma

[1]Neste trabalho, x representa a linha, y a coluna, e o ponto $(0, 0)$ representa a origem no canto superior esquerdo da imagem.

das Diferenças Quadráticas, ou SSD - *Sum of Squared Differences*, similarmente à abordagem proposta em [Musse et al., 2007]. Assim, a posição da máscara no quadro atual é utilizada para definir uma região de busca nas redondezas dessa posição no quadro adjacente, e a máscara é então deslocada para a posição que produzir o menor erro no quadro adjacente[2]. Salienta-se que as métricas de casamento propostas em [Porikli et al., 2006, Adam et al., 2006] são mais robustas com relação a oclusões e variações de postura, mas o casamento por SSD se mostrou suficiente para o tipo de vídeo abordado neste trabalho.

3.4 Refinamento da extração da cabeça

A extração da cabeça é um passo fundamental para o correto funcionamento do método proposto. A seção anterior descreveu uma abordagem simples para obter uma máscara retangular que aproxime o *bounding box* da cabeça, que é utilizada para realizar o acompanhamento em quadros consecutivos. Tal aproximação é eficiente para efeitos de acompanhamento da pessoa, mas pode incluir parte do pescoço e tronco da pessoa, ou até representar apenas uma porção da cabeça.

Para refinar a estimativa da cabeça, foi implementado um algoritmo baseado na análise da silhueta da parte superior do objeto. Em um primeiro momento, a área preferencial da ocorrência da cabeça da pessoa é estimada como tendo 15% da altura do objeto. O *blob* da pessoa é varrido de cima para baixo nessa região de interesse, e são obtidos dois vetores $\boldsymbol{X}^{\min}$ e $\boldsymbol{X}^{\max}$, contendo respectivamente a posição x de início e fim do *blob* em cada linha da região de interesse. Assumindo que a maioria valores nesses vetores de fato correspondem à cabeça (e apenas alguns ao ruído ou tronco), calcula-se a largura estimada da face através de $S = \text{med}(X^{\max}) - \text{med}(X^{min}) + 1$, onde med representa a mediana do vetor. A mediana foi escolhida ao invés da média por apresentar mais invariância quando valores relativos ao tronco estiverem presentes em $\boldsymbol{X}^{\min}$ e $\boldsymbol{X}^{\max}$.

Para refinar ainda mais a estimativa da cabeça, os vetores $\boldsymbol{X}^{\min}$ e $\boldsymbol{X}^{\max}$ são restringidos a uma região de tamanho $S \times S$, assumindo que a altura da cabeça é similar a sua largura. Então, um filtro da mediana (usando uma vizinhança de 3 *pixels*) é usado para suavizar o contorno, $\bar{X}_i^{\min} = \text{med}(X_{i-1}^{\min}, X_i^{\min}, X_{i+1}^{\min})$ e $\bar{X}_i^{\max} = \text{med}(X_{i-1}^{\max}, X_i^{\max}, X_{i+1}^{\max})$. Os limites esquerdo e direito da cabeça são calculados respectivamente por $x^{\min} = \text{med}(X^{\min})$ e $x^{\max} = \text{med}(X^{\max})$, e a estimativa final do tamanho da cabeça é dado por $S' = x^{\max} - x^{\min} + 1$.

Considerando que o limite superior da cabeça coincide com o limite superior do objeto da pessoa ($y^{\min}$), e os limites laterais da cabeça são dados por $x^{\max}$ e $x^{\min}$,

[2] A região de busca é ainda restringida a *pixels* classificados como *foreground*, para evitar que a máscara ache um mínimo em alguma região do *background*.

o cálculo da *bounding-box* da cabeça é concluída por $y^{\max} = y^{\min} + S'$.

Figura 3.4 – Exemplos do *bounding-box* da pessoa detectada em diferentes quadros da seqüência de vídeo. A linha amarela mostra os valores dos vetores $\boldsymbol{X}^{\min}$ e $\boldsymbol{X}^{\max}$ para a área do busto. A linha vermelha ilustra os valores médios entre os dois vetores, e o quadro rosa é o resultado do refinamento da região da cabeça.

A Figura 3.4 mostra alguns exemplos do recorte de cabeça implementado para esta dissertação. É possível ver que mesmo em situações onde o *bounding-box* é bem maior que a largura do corpo, como quando a pessoa está de braços abertos, ou até mesmo quando o objeto de *foreground* não é totalmente perfeito, este simples algoritmo apresenta uma excelente extração da cabeça. A Figura 3.5 é um exemplo desse segundo caso, no entanto a área da boca e queixo foram perdidas.

(a) (b)

Figura 3.5 – (a) Imagem mostrando uma deformação na detecção da pessoa, mas que não alterou a detecção da cabeça devido ao refinamento que foi implementado na imagem (b).

3.5 Estimativa da orientação

Como resultado do acompanhamento das pessoas, tem-se acesso às trajetórias de cada pessoa capturada. Além disso, o processo de refinamento da estimativa da cabeça permite um acesso direto aos *pixels* que compõem a cabeça, de modo que se pode verificar quais deles pertencem à face. Na abordagem proposta para o cálculo da orientação das pessoas, o vetor velocidade proveniente do acompanhamento indica a direção na qual a pessoa se movimenta (e, provavelmente, está olhando). Além disso, uma estimativa da orientação da cabeça também é calculada com base na detecção de *pixels* da face, conforme descrito a seguir.

3.5.1 Estimativa de orientação pela movimentação

A primeira estimativa de orientação é dada pela movimentação da pessoa na cena, assumindo que uma pessoa está olhando na direção em que caminha. Nessa estimativa, é utilizado o vetor velocidade instantânea $\boldsymbol{m}$, calculado pela diferença entre a posição fornecida pelo algoritmo de *tracking* em dois quadros consecutivos.

Claramente, há problemas em estimar a orientação da pessoa pelo vetor velocidade quando ela estiver parada. O algoritmo de *tracking* deve fornecer pequenos deslocamentos (devido ao ruído e pequenos movimentos da pessoa), que não se refletem necessariamente em mudanças de orientação da pessoa. A magnitude de $\boldsymbol{m}$ pode ser utilizado para discriminar movimentos efetivos da pessoa em detrimento de pequenas oscilações que ocorrem em pessoas paradas. A idéia é que, se $\|\boldsymbol{m}\|$ é grande, então deve haver uma maior confiança na direção do movimento fornecida por $\boldsymbol{m}$.

Embora se possa determinar a orientação de movimento diretamente pelo ângulo de $\boldsymbol{m}$, optou-se neste trabalho discretizar as possíveis orientações em um conjunto de 8 vetores base unitários $\boldsymbol{m}_i^b$, nas direções Norte, Nordeste, Leste, Sudeste, Sul, Sudoeste, Oeste e Noroeste, onde i denota a orientação. Dado o vetor velocidade instantânea $\boldsymbol{m}$ de uma pessoa sendo acompanhada, sua distância (angular) com relação à orientação fornecida por cada vetor base é:

$$a_i = \cos^{-1}\left(\frac{< \boldsymbol{m}_i^b, \boldsymbol{m} >}{\|\boldsymbol{m}\|}\right), \tag{3.7}$$

onde $< \cdot, \cdot >$ denota o produto interno, e $\|\cdot\|$ é a norma Euclideana de vetores.

Conforme mencionado acima, a magnitude do vetor de movimento também deve ser levada em consideração, pois indica o quão significativa é a movimentação da pessoa. Assim, o poder de discriminação de a_i deve ser amplificado se $\|\boldsymbol{m}\|$ é grande, e atenuado se $\|\boldsymbol{m}\|$ é pequeno. A métrica proposta de similaridade é dada por

$$s_i^m = \exp\left(-\frac{1}{2}\left(\frac{a_i\|\boldsymbol{m}\|}{\varsigma_m}\right)^2\right), \tag{3.8}$$

onde ς_m é um limiar que controla o deslocamento (em módulo) que aumenta a separabilidade entre as orientações. Se $\|\boldsymbol{m}\| >> \varsigma_m$, então a orientação base $\boldsymbol{m}_i^b$ que minimiza a distância angular a_i terá um peso grande com relação às demais orientações. Por outro lado, se $\|\boldsymbol{m}\| << \varsigma_m$, os valores s_i^m serão aproximadamente similares, indicando que a seleção da orientação pelo critério de movimentação não é muito confiável. Com base nas imagens utilizadas neste trabalho, fixou-se experimentalmente o valor de ς_m em 0.5.

Para gerar uma métrica de similaridade o_i^m para a orientação baseada no movimento que possa ser interpretada como uma probabilidade de pertinência a

cada uma das classes, define-se

$$o_i^m = \frac{s_i^m}{\sum_{j=1}^{8} s_j^m}. \tag{3.9}$$

3.5.2 Estimativa de orientação pela face

A posição da face com relação à câmera também pode ser utilizada para estimar a orientação da cabeça. De fato, a quantidade de *pixels* da pele detectados na região da face pode ser explorada para estimar se a pessoa está olhando frontalmente para a câmera (quantidade grande de *pixels* de pele), lateralmente (menor quantidade) ou de costas para a câmera. Para tal estimativa, é necessário detectar *pixels* com cor da pele.

3.5.2.1 Modelagem da cor da pele

Como visto na Seção 2.1.4, o uso da cor da pele para a detecção de face em imagens coloridas é muito útil, pois fornece um meio fácil de localização da face sem levar em consideração textura e propriedades geométricas [N. A. Abdul Rahim, 2006]. Ainda, em imagens de média resolução onde os traços do rosto nem sempre são visíveis, o uso da cor de pele torna-se fundamental. Existe uma grande abundância de técnicas e abordagens que tratam a cor da pele, os quais foram muito bem revisados em [Hjelmas and Low, 2001], [Yang et al., 2002] e [Vezhnevets et al., 2003].

Neste trabalho, optou-se por um usar um método que explicitamente define a distribuição de cor pele através de um conjunto de regras. As principais vantagens que motivaram a escolha desta abordagem foram a simplicidade, rápida implementação e por apresentarem bons resultados. Entre os métodos existentes, o escolhido foi o modelo de [N. A. Abdul Rahim, 2006], que utiliza informação adicional ao padrão *RGB*, vinda dos sub-espaços de cores *HSV* e *YCbCr*, descritos na Seção 2.1.4.

Conforme os autores, no espaço *RGB* a região de cor de pele não é bem distinta em todos os 3 canais, como pode ser vista pela simples observação de um histograma. No espaço *HSV*, o canal *H (Hue)* mostra uma significante discriminação de regiões de pele. Os canais *Cb* e *Cr* possuem crominância similar para regiões de pele, e variando os valores de intensidade de *Y* (luminância), não há alteração na distribuição de cor de pele nestes sub-espaços. Assim, a luminância meramente caracteriza o brilho de um valor de crominância. Conforme [Phung et al., 2002], estes valores apresentam boa representação para todas as raças humanas.

Alguns testes foram feitos em faces de diferentes tonalidades para validar a aplicabilidade do modelo escolhido para esta dissertação. Alguns resultados

destes experimentos podem ser vistos na Figura 3.6, sobre um conjunto de imagens encontradas na *internet*. No trabalho original, os autores ainda utilizam operações morfológicas para preencher os buracos na face detectada, o que não se mostrou necessário para a detecção de orientação proposta neste trabalho. O algoritmo de detecção de *pixels* da pele foi avaliado quadros de vídeo da câmera protótipo usados nessa dissertação, conforme ilustrado na Figura 3.7, indicando a validade do modelo escolhido para detecção de pele.

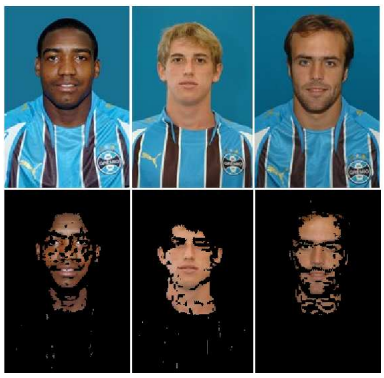


Figura 3.6 – Detecção de pele em diferentes tonalidades.

Para o espaço RGB, foram utilizadas as mesmas regras de [Kovac et al., 2003] da Equação (2.4) para a cor de pele sob luz do dia uniforme. Para luz do dia lateral ou sob a luz de um *flash* a equação utilizada é

$$\begin{aligned} &R > 220 \text{ e } G > 210 \text{ e } B > 170 \text{ e} \\ &|R - G| = 15 \text{ e } R > B \text{ e } G > B. \end{aligned} \tag{3.10}$$

Assim, a **Regra A** na detecção de *pixels* da pele é definida como Equação(2.4) $\cup$ Equação(3.10), ou seja, basta que uma delas ocorra.

Já para o espaço $CbCr$, foram construídos os seguintes planos (**Regra B**) baseados na sua distribuição no espaço 2D.

$$\begin{aligned} Cr &\geq 1.5862Cb + 20 \text{ e} \\ Cr &\leq 0.3448Cb + 76.2069 \text{ e} \\ Cr &\leq -4.5652Cb + 234.5652 \text{ e} \\ Cr &\geq -1.15Cb + 301.75 \text{ e} \\ Cr &\geq -2.2857Cb + 432.85 \end{aligned} \tag{3.11}$$

No espaço *HSV*, foram definidos dois limites de corte como sendo a **Regra C**:

$$H < 25 \text{ ou } H > 230 \tag{3.12}$$

Dadas essas regras, um *pixel* é classificado como pele se satisfazer **Regra A** $\cap$ **Regra B** $\cap$ **Regra C**, ou seja, se satisfizer simultaneamente as três regras [N. A. Abdul Rahim, 2006].

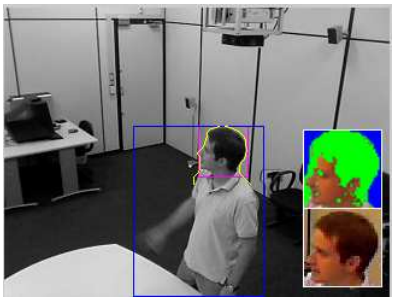


Figura 3.7 – Detecção de pele em um vídeo da câmera protótipo

3.5.2.2 Orientações da face

Para cada amostra da face, é construído um vetor que é a serialização da imagem, cujos elementos podem assumir valores no conjunto $\{1, 0, -1\}$, representando pele (1), fundo (0) e não-pele (-1). A idéia básica é comparar o vetor de amostras de uma imagem de teste com outras previamente classificadas em diferentes orientações, medindo a similaridade entre elas. O método proposto pode ser dividido aqui em duas etapas: aprendizagem e classificação.

Na etapa de aprendizagem, são gerados vetores protótipos $\boldsymbol{\tau}_i$, para $i = 1, .., 8$, que representam cada uma das 8 possíveis orientação da cabeça: Sul (S), Sudoeste (SO), Oeste (O), Noroeste (NO), Norte (N), Nordeste (NE), Leste (L) e Sudeste (SE). A Figura 3.8 mostra essas orientações, assim como exemplos de padrões de pele obtidos para cada orientação da face. Cada imagem j da base de treinamento é classificada manualmente em uma das 8 classes i, e a máscara da cabeça é percorrida (de cima para baixo, esquerda para direita), gerando um vetor $\boldsymbol{\tau}_i^j$ cuja posição pode ser 1 (pele), 0 (fundo) ou -1 (não-pele). O vetor protótipo $\boldsymbol{\tau}_i$ da classe i é então obtido como sendo a média (na variável j) de todos os vetores $\boldsymbol{\tau}_i^j$ que foram manualmente assinalados à classe i no período de treinamento.

Na etapa de teste, um vetor de faces $\boldsymbol{\nu}_f$ é extraído da mesma maneira que no período de treinamento, e sua coerência com cada vetor protótipo $\boldsymbol{\tau}_i$ é avaliada. Embora diferentes métricas de similaridade possam ser utilizadas, a correlação cruzada mostrou bons resultados para discriminar diferentes orientações. Assim, coerência inicial ϑ_i entre o vetor de teste $\boldsymbol{\nu}_f$ e o protótipo de cada classe $\boldsymbol{\tau}_i$ é dada

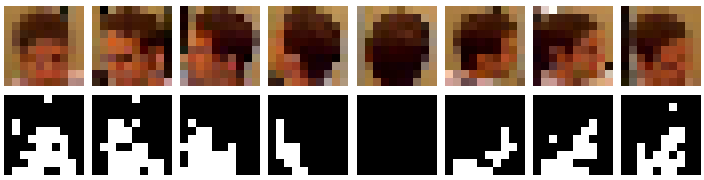


Figura 3.8 – Orientações possíveis
S, SO, O, NO, N, NE, L e SE, respectivamente.

por

$$\vartheta_i = \frac{\boldsymbol{\nu}_f^T \boldsymbol{\tau}_i}{\|\boldsymbol{\nu}_f\| \|\boldsymbol{\tau}_i\|}. \tag{3.13}$$

Por fim, utiliza-se uma função *soft-max* [Andrew, 1999] para potencializar as separabilidade entre as classes, através de um coeficiente de temperatura t. Quanto menor o valor de t, maior a potencialização das diferenças (o valor $t = 0.1$ foi utilizado neste trabalho, determinado empiricamente). A equação que calcula o vetor de probabilidades pela face o_i^f utilizando *soft-max* é definido como

$$o_i^f = \frac{\exp\left(-\left(1 - \frac{\vartheta_i}{t}\right)\right)}{\sum_{j=1}^{8} \exp\left(-\left(1 - \frac{\vartheta_j}{t}\right)\right)}. \tag{3.14}$$

3.5.3 Estimativa de orientação instantânea

A estimativa instantânea da orientação da atenção o_i é dada pela combinação das probabilidades das orientações pela movimentação o_i^m e pela face o_i^f. A combinação dessas probabilidade é dada pela multiplicação marginal dos valores de cada vetor, conforme a Equação (3.15), normalizando os produtos de modo que a soma dos valores de pertinência seja unitário. Assim, a orientação instantânea conjunta o_i pode ser interpretada como a probabilidade de uma pessoa possuir orientação na direção do vetor $\boldsymbol{m}_i^b$, usando tanto a informação de movimento quanto orientação da face.

Salienta-se que, se todos os valores de o_i^m forem parecidos, o_i tende a ser igual a o_i^f. Se todos os valores de o_i^f forem parecidos, o_i tende a ser igual a o_i^m. Quando a pessoa está parada (ou com pouco movimento), as orientações com base na movimentação são aproximadamente equiprováveis, ou seja, os valores de o_i^m são parecidos. Assim, a orientação da atenção será dada pela orientação da face.

$$o_i = \frac{o_i^m o_i^f}{\sum_{j=1}^{8} o_j^m o_j^f} \tag{3.15}$$

3.5.3.1 Traço das probabilidades

Até aqui foi descrita a detecção da direção da atenção da pessoa com base no movimento e na orientação da cabeça em um dado instante. Deste processo resulta um vetor $\boldsymbol{o}(t) = ((\boldsymbol{o}_1(t), \boldsymbol{o}_2(t)), ..., \boldsymbol{o}_k(t)), k = 8$ que representa a probabilidade de cada orientação em cada instante de tempo t.

Considerando que o foco de atenção tende a persistir temporalmente, e com o objetivo de evitar variações bruscas na detecção do foco de atenção, optou-se por estimar a direção da atenção da pessoa a partir do histórico recente das detecções do sistema. Deste processo resulta um vetor de probabilidades $\bar{\boldsymbol{o}}(t)$ que representa a estimativa do foco de atenção a partir de sucessivas detecções de movimentação e orientação da cabeça, fornecendo assim uma aproximação mais robusta da atenção da pessoa. Neste trabalho, o vetor $\bar{\boldsymbol{o}}$ é referido como traço das probabilidades do foco de atenção.

Antes do sistema processar qualquer imagem, o traço das probabilidades é inicializado como uma distribuição equiprovável ($\bar{o}_i(0) = \frac{1}{8}$), representando o não conhecimento do foco de atenção. Ao passo que o sistema processa quadros do vídeo e calcula as probabilidades instantâneas $\boldsymbol{o}(t)$, o traço das probabilidades é atualizado conforme

$$\bar{o}_i(t) = \frac{o_i(t)(\bar{o}_i(t-1))^\alpha}{\sum_{j=1}^{8} o_j(t)(\bar{o}_j(t-1))^\alpha}, \tag{3.16}$$

onde $\alpha > 0$ é um coeficiente de memória. Se $\alpha \approx 0$, a observação atual carrega um peso maior com relação à coerência temporal. Se α igual a 1, considera-se igualmente todas as observações passadas, incluindo a atual. Neste trabalho, o valor de memória α foi definido empiricamente como 0.5.

Com base no traço das probabilidades do foco de atenção, a direção da atenção instantânea $\boldsymbol{\omega}(t)$ é expressa em função dos vetores unitários de orientação $\boldsymbol{m}_i^b$ através de

$$\boldsymbol{\omega}(t) = \boldsymbol{m}_j^b, \text{ onde } j = \underset{i}{\operatorname{argmax}}\, \bar{o}_i \tag{3.17}$$

Capítulo 4

Resultados Experimentais

Esta seção apresenta a validação do método proposto, assim como os experimentos realizados nessa dissertação.

4.1 Ambiente de execução

Para a avaliação do método proposto nesta dissertação, foi implementado um protótipo em Matlab R2006a. Todos os experimentos foram executados em Notebook HP Intel Celeron M 1.73 GHz, 512 MB RAM. Os vídeos usados pela aplicação foram gravados em uma sala com luz ambiente controlada na Universidade do Vale do Rio dos Sinos utilizando uma *webcam* Logitech. 10 vídeos coloridos de 10 diferentes pessoas foram obtidos a 14 FPS com a resolução de 320×240 *pixels*, totalizando mais de 10.000 imagens entre posições válidas e imagens utilizadas para o treinamento do plano de fundo da cena.

Tabela 4.1 – Tabela de amostras usadas na aprendizagem das posições

Posição	Número de Imagens
Norte	150
Nordeste	182
Leste	143
Sudeste	176
Sul	155
Sudoeste	199
Oeste	123
Noroeste	116

Conforme descrito na Seção 3.5.2.2, o método necessita de um período de aprendizagem (dos vetores de orientação com base em *pixels* cor de pele detectados na face) antes de executar propriamente. Na etapa de aprendizagem foi utilizado um subconjunto de imagens do total de imagens adquiridas, conforme sumarizado na Tabela 4.1. Salienta-se que cada vídeo tem entre aproximadamente de 50 a 200

imagens de cada posição, e então o número de amostras para cada orientação é levemente diferente.

4.1.1 Análise Visual

Para facilitar a visualização das orientações detectadas, são apresentados 5 conjuntos de vetores, um para cada tipo de orientação analisada pelo método proposto. Os vetores são representados com base nos pontos cardeais e colaterais (ver Figura 4.1), sendo possível representar as 8 posições descritas na seção 3.5.2.2. A magnitude de cada uma das oito orientações é dada pelo respectivo valor de probabilidade, conforme o seguinte padrão de cores:

- **amarelo**: o_i^m, estimativa de orientação pela movimentação;
- **verde**: o_i^f, estimativa de orientação pela face;
- **ciano**: o_i, estimativa instantânea da orientação;
- **azul**: $\bar{o}_i$, traço de probabilidades;

Além disso, o vetor vermelho isolado representa a orientação final instantânea $\boldsymbol{\omega}$.

Figura 4.1 – Vetor descritivo das posições

4.2 Resultados

A Tabela 4.2 mostra a probabilidade de cada uma das 8 posições em um quadro (ilustrado na Figura 4.2(a)) da seqüência de imagens, para cada um dos tipos de posição analisada. O valor em destaque na tabela é a orientação final $\boldsymbol{\omega}$, estimada pelo método proposto. Como é possível verificar, a saída do método indicou corretamente a posição Leste. O quadro no vídeo representa um momento em que a pessoa está caminhando na cena, sendo possível ver que pela análise da movimentação, a posição Nordeste como segunda maior probabilidade. Nesse exemplo, a informação da face recebe menor peso do que a informação de movimentação, mas mesmo assim a tabela mostra que orientação Leste também

foi estimada pelo método proposto. A quarta coluna na tabela mostra o valor de orientação instantânea, destacado pela função *soft-max*, ressaltando a posição Leste como preferida.

Tabela 4.2 – Tabela de probabilidades para cada orientação em um instante i

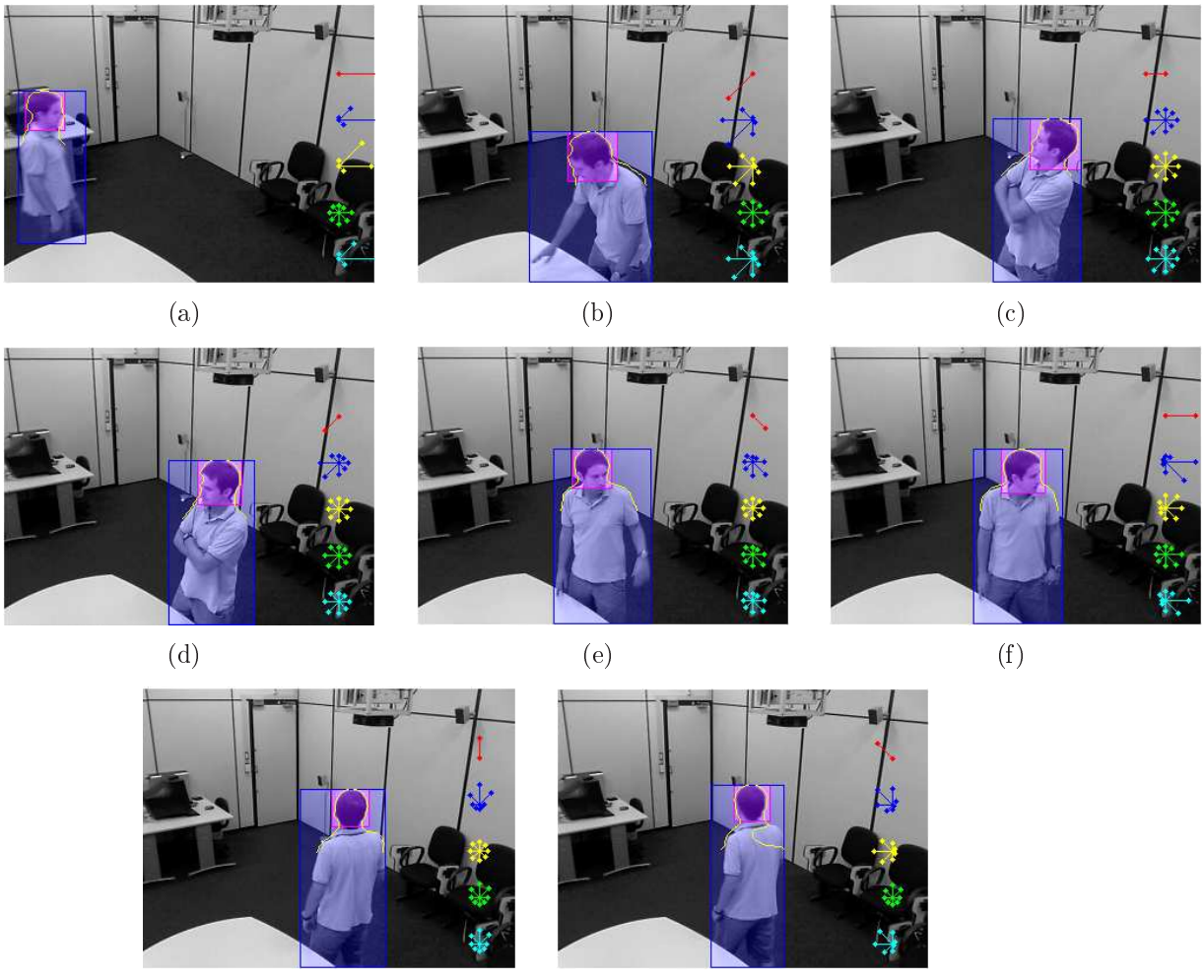
	o_i^m	o_i^f	o_i	$\bar{o}_i$
L	0.4187	0.1676	0.5516	**0.6731**
SE	0.0862	0.1559	0.1056	0.0903
S	0.0036	0.1424	0.0040	0.0019
SO	0.0000	0.1263	0.0000	0.0000
O	0.0000	0.1146	0.0000	0.0000
NO	0.0031	0.1138	0.0028	0.0005
N	0.0800	0.0928	0.0583	0.0257
NE	0.4084	0.0865	0.2777	0.2084

A Figura 4.2 ilustra o resultado da detecção da orientação para uma seqüência de um mesmo vídeo através do método proposto nesta dissertação. Na Figura 4.2(a) percebe-se que a pessoa está entrando na cena, tendo como orientação a posição Leste. Aqui é possível notar que o vetor da face recebe pouco peso no cálculo, uma vez que o movimento já fornece uma boa estimativa do objetivo ou foco de atenção dessa pessoa. Nas Figuras 4.2(b)-(h), a pessoa está com pouca movimentação na cena, realizando basicamente movimentos de tronco.

Na Figura 4.3 podemos ver a identificação da orientação em diferentes quadros do vídeo, para diferentes pessoas. Na Figura 4.3(a), a pessoa está caminhando na cena para o Leste, sendo corretamente identificada pelo sistema. Na Figura 4.3(b), como pode ser visto pelo vetor amarelo de movimentação, a pessoa está parada na cena. Nesse caso o sistema considerou a face com maior peso para identificar corretamente a orientação Sudoeste. Já as Figuras 4.3(c),(d) e (h) mostram pessoas na direção Noroeste, Sul e Oeste respectivamente, corretamente identificadas. A análise do vetor amarelo na Figura 4.3(e) indica que a pessoa está caminhando na cena. Nesse caso o vetor resultante pesou para a direção do movimento. Na Figura 4.3(f), a pessoa está de costas para a câmera, mas a cabeça está levemente direcionada para a posição Noroeste, identificada pela detecção da face por cor da pele. Já o exemplo ilustrado na Figura 4.3(g) pode ser considerado um erro. Pelo menos no exato instante de captura do quadro em questão, o rosto da pessoa está levemente inclinado para a esquerda gerando uma sombra sobre a própria pele, sendo um ponto problemático para o identificador de cor da pele.

4.3 Métricas de Análise

Os resultados da classificação foram avaliados quantitativamente através da análise da matriz de confusão. Os valores exatos de orientação (*ground truth*) foram

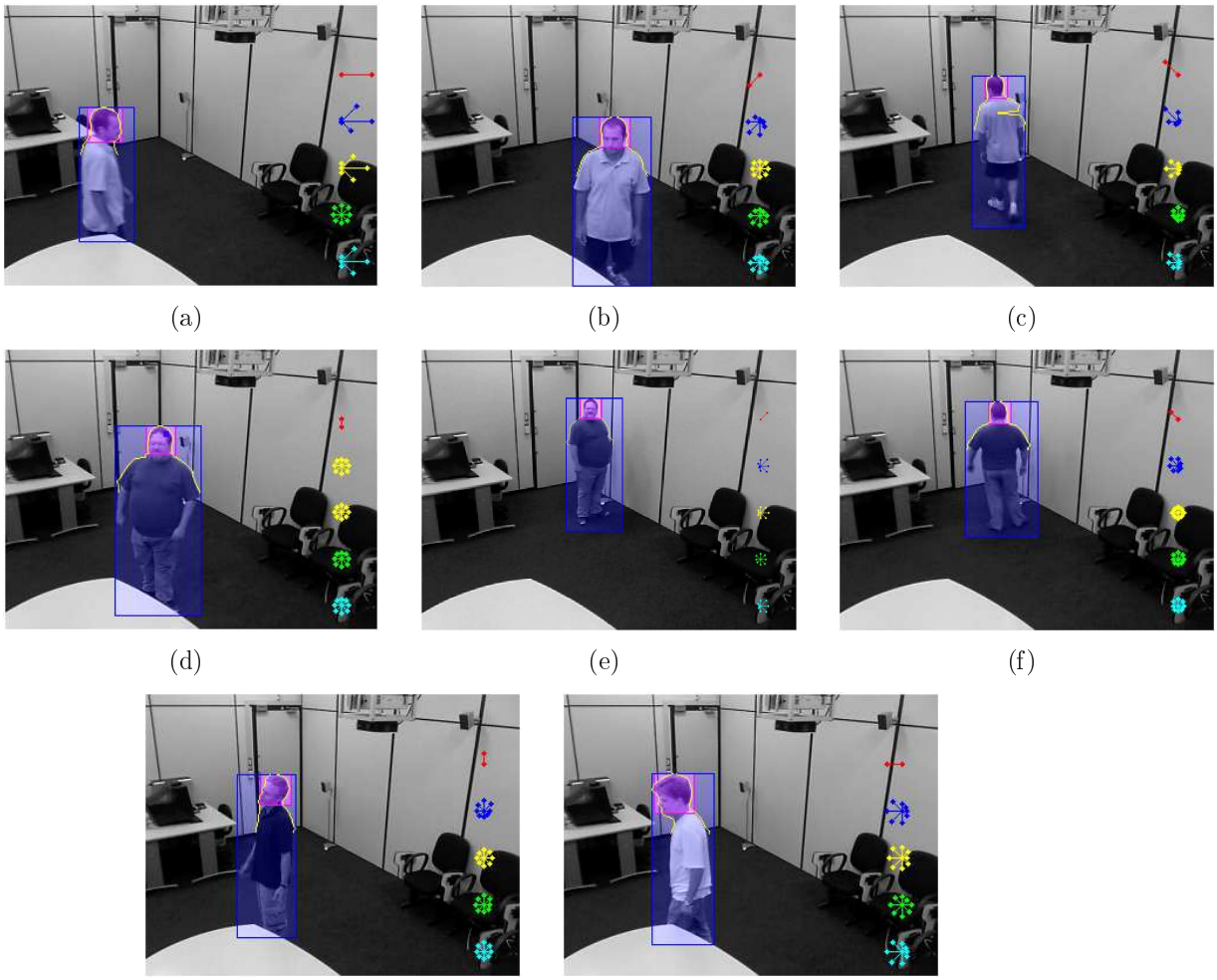


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figura 4.2 – Resultados para uma seqüência de imagens



(a) (b) (c)

(d) (e) (f)

(g) (h)

Figura 4.3 – Outros resultados de detecção de orientação

obtidos manualmente, e então pode haver confusão entre duas classes adjacentes (por exemplo, S e SE). A Tabela 4.3 mostra nas linhas a orientação que deveria ter ocorrido e nas colunas a informação retornada pelo método proposto (em porcentagem de classificação em cada uma das classes). Percebe-se que em todas as orientações, a maioria das amostras ($> 70\%$) foi classificada corretamente ou em uma orientação adjacente. Considera-se que o pior resultado ocorreu nas amostras da classe SO, onde aproximadamente 24% dos valores foi erroneamente classificado como pertencente à classe L.

Tabela 4.3 – Matriz de confusão(%) para as imagens de validação

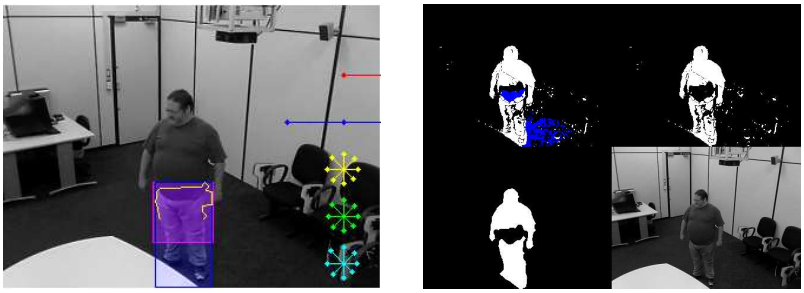
-	L	SE	S	SO	O	NO	N	NE
L	95,35	4,65	0	0	0	0	0	0
SE	21,05	60,52	14,47	0	1,32	2,64	0	0
S	23,63	5,45	69,09	1,83	0	0	0	0
SO	24,12	2,51	10,55	42,21	20,10	0,51	0	0
O	4,06	0	1,62	17,07	65,04	7,31	1,65	3,25
NO	0	4,31	6,03	11,20	23,27	45,69	9,50	0
N	0	0	0	0	0	26,53	71,43	2,04
NE	0	0	0	0	0	0	100	0

4.3.1 Problemas encontrados

Esta seção tem por objetivo levantar os problemas encontrados, para que soluções sejam propostas e estudadas futuramente.

Após a captura e análise de diversos vídeos com diferentes pessoas, observou-se que para alguns quadros do mesmo vídeo o objeto detectado de *tracking* não foi satisfatório. Como foi descrito neste trabalho, o sistema possui um algoritmo para refinamento da cabeça baseado no *bounding-box* do objeto detectado, logo, este precisa funcionar corretamente. Como pode ser visto na Figura 4.4, em uma determinada imagem da seqüência de imagens o *blob* que representa o objeto foi partido ao meio, ocasionando uma falsa informação do tamanho dos limites da pessoa. Este tipo de problema não está relacionado ao método proposto nesta dissertação, mas sim na camada mais inferior da arquitetura, que é responsável pela detecção e rastreamento da pessoa na cena de vídeo. As imagens da Figura 4.2 mostram que quando a informação do objeto é correta, o refinamento da região da cabeça também funciona de maneira correta.

Outros problemas vistos não foram tratados nesse trabalho, como a questão de pessoas carecas e com cores de pele fora do padrão, por exemplo. Para estes tipos de pessoa, o padrão gerado pelo detector de pele é praticamente igual ao padrão gerado para uma pessoa de frente (no caso de carecas), ou a detecção de pele pode falhar (pessoas com cor de pele muito fora do padrão). Apesar de muitos autores terem



(a) (b)

Figura 4.4 – Problema com os limites do objeto detectado

motivado o uso de detecção da cor da pele, muitos deles não tratam situações como pessoas de costas para a câmera, por exemplo. Por outro lado, pessoas de óculos e/ou com barba também são potenciais fontes de problemas, pois influenciam na detecção de *pixels* da pele. Entretanto, alguns resultados preliminares, como os ilustrados nas Figuras 4.3(d)-(e), indicam que o método proposto pode ter resultados satisfatórios nessas situações.

Outro problema identificado é ilustrado na Figura 4.3(g), onde o algoritmo de estimativa pela face não identificou o padrão correto devido à sombra gerada pela inclinação do rosto. Questões como essas, assim como o problema citado no parágrafo anterior certamente merecem atenção futuramente na continuação das pesquisas após esta dissertação.

Capítulo 5

Conclusões

Esta dissertação apresentou um novo método para a determinação da orientação global de uma pessoa em uma seqüência de vídeo. ' Foi apresentada uma abordagem estatística baseada na correlação cruzada de padrões de pele da face de um quadro atual com padrões previamente aprendidos em uma etapa anterior de treinamento. Através da junção da informação de orientação fornecida pela face e pelo movimento das pessoas, o método proposto mostrou resultados satisfatórios. Com a arquitetura escolhida para validar este trabalho, pode-se concluir que o método proposto é viável, porém ficou claro que alguns casos de falhas na obtenção da pessoa na cena podem ser bastante prejudiciais para o funcionamento do sistema.

A análise da matriz de confusão gerada mostrou que a maior parte dos erros de classificação ocorrem entre orientações adjacentes, como por exemplo, na troca de uma posição Oeste para uma Sudoeste. Este fato não é um problema grave, pois pode-se pensar em uma posição intermediária para esses casos como sendo a média entre as duas posições com maior probabilidade. É importante dizer que o sistema apresentado utilizada apenas 8 posições possíveis de acordo com os pontos cardeais e colaterais, separadas assim por um ângulo de 45 graus.

Um fator importante que precisa ser investigado e que está listado na seção a seguir é a questão de pessoas com barba, carecas ou que fazem uso de acessórios como óculos, por exemplo. Um problema que inicialmente havia causado preocupação para esta dissertação era como detectar tons de pele escuros, mas em teoria o modelo de detecção de cor de pele deve funcionar nestes casos também, como indicado na Figura 3.6. Entretanto, nenhuma pessoa com tom de pele mais escura foi avaliada nos testes de orientação, sendo necessária uma validação mais profunda.

Por fim, resta dizer que o protótipo implementado em Matlab foi de fundamental importância para uma análise visual das orientações. Para pesquisas futuras, têm-se então uma aplicação modular, sendo facilmente possível a utilização de outras técnicas e métodos para melhorias futuras.

5.1 Trabalhos Futuros

Esta seção visa motivar tópicos de pesquisa para trabalhos futuros relacionadas à esta dissertação.

- Análise de outros métodos para detecção de cor de pele, alguns deles já apresentados como trabalhos relacionados nesta dissertação;
- Implementação do detector dos padrões de pele com redes neurais;
- Estudo de outras técnicas para detectar a orientação da face que não sejam baseados na cor da pele, buscando solucionar problemas de pessoa inteiramente carecas.
- Estudo mais aprofundado sobre o desempenho do método proposto aplicado a pessoas com bigodes, barba ou outros aparatos como óculos escuros que podem influenciar na obtenção dos padrões por cor da pele;
- Avaliar o sistema em ambientes externos ou iluminação variável;
- Avaliar a possibilidade de detecção de múltiplas pessoas na cena;
- Tornar o sistema tempo-real, através de implementação em uma linguagem compilada.

Pretende-se enviar um trabalho relacionado com esta dissertação à conferência ICPR - *International Conference on Pattern Recognition*.

Bibliografia

[Adam et al., 2006] Adam, A., Rivlin, E., and Shimshoni, I. (2006). Robust fragments-based tracking using the integral histogram. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 798–805, Washington, DC, USA. IEEE Computer Society.

[Andrew, 1999] Andrew, A. M. (1999). *REINFORCEMENT LEARNING: AN INTRODUCTION by Richard S. Sutton and Andrew G. Barto, Adaptive Computation and Machine Learning series, MIT Press (Bradford Book), Cambridge, Mass., 1998, xviii + 322 pp, ISBN 0-262-19398-1.*, volume 17. Cambridge University Press, New York, NY, USA.

[Ba and Odobez, 2004] Ba, S. and Odobez, J. (2004). A probabilistic framework for joint head tracking and pose estimation. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 4:264–267 Vol.4.

[Ba, 2007] Ba, S. O. (2007). *Joint head tracking and pose estimation for visual focus of attention recognition.* PhD thesis, EPFL, Lausanne.

[Ba and Odobez, 2006] Ba, S. O. and Odobez, J.-M. (2006). Head pose tracking and focus of attention recognition algorithms in meeting rooms. In Stiefelhagen, R. and Garofolo, J. S., editors, *CLEAR*, volume 4122 of *Lecture Notes in Computer Science*, pages 345–357. Springer.

[Bradski et al., 2005] Bradski, G., Kaehler, A., and Pisarevsky, V. (2005). Learning-based computer vision with intel's open source computer vision library. *Intel Technology Journal*, 9:119–130.

[Brand and Mason, 2000] Brand, J. and Mason, J. (2000). A comparative assessment of three approaches to pixel-level human skin-detection. *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, 1:1056–1059 vol.1.

[Brown et al., 2001] Brown, D. A., Craw, I., and Lewthwaite, J. (2001). A som based approach to skin detection with application in real time systems. In Cootes, T. F. and Taylor, C. J., editors, *BMVC*. British Machine Vision Association.

[Brumitt et al., 2000] Brumitt, B., Krumm, J., Meyers, B., and Shafer, S. (2000). Let there be light: Comparing interfaces for homes of the future. In *IEEE Personal Communications*.

[Buccolieri et al., 2005] Buccolieri, F., Distante, C., and Leone, A. (2005). Human posture recognition using active contours and radial basis function neural network. *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*, pages 213–218.

[Cezar Silveira Jacques et al., 2006] Cezar Silveira Jacques, J., Rosito Jung, C., and Musse, S. (2006). A background subtraction model adapted to illumination changes. *Image Processing, 2006 IEEE International Conference on*, pages 1817–1820.

[Chai and Bouzerdoum, 2000] Chai, D. and Bouzerdoum, A. (2000). A bayesian approach to skin color classification in ycbcr color space. *TENCON 2000. Proceedings*, 2:421–424 vol.2.

[Chella et al., 2005] Chella, A., Dindo, H., and Infantino, I. (2005). A system for simultaneous people tracking and posture recognition in the context of human-computer interaction. *Computer as a Tool, 2005. EUROCON 2005.The International Conference on*, 2:991–994.

[Gee and Cipolla, 1994] Gee, A. and Cipolla, R. (1994). Estimating gaze from a single view of a face. *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, 1:758–760 vol.1.

[Gomez, 2002] Gomez, G. (2002). On selecting colour components for skin detection. *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 2:961–964 vol.2.

[Grimson et al., 1998] Grimson, W. E. L., Stauffer, C., Romano, R., and Lee, L. (1998). Using adaptive tracking to classify and monitor activities in a site. In *CVPR '98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 22, Washington, DC, USA. IEEE Computer Society.

[Haritaoglu et al., 2000] Haritaoglu, I., Harwood, D., and Davis, L. S. (2000). W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830.

[Hjelmas and Low, 2001] Hjelmas, E. and Low, B. (2001). Face detection: a survey. *Comput. Vision Image Understanding*, 83:236–274.

[Hsu et al., 2002] Hsu, R.-L., Abdel-Mottaleb, M., and Jain, A. (2002). Face detection in color images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):696–706.

[Hu et al., 2005] Hu, N., Huang, W., and Ranganath, S. (2005). Attentive behavior detection by non-linear head pose embedding and mapping. *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, pages 1–4.

[Hu et al., 2004] Hu, W., Tan, T., Wang, L., and Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(3):334–352.

[IET, 2005] IET (2005). Introducing matlab7 and simulink 6. *The IEE Review*, 51(3):11–11.

[Jacques et al., 2005] Jacques, J., Jung, C., and Musse, S. (2005). Background subtraction and shadow detection in grayscale video sequences. *Computer Graphics and Image Processing, 2005. SIBGRAPI 2005. 18th Brazilian Symposium on*, pages 189–196.

[Jones and Rehg, 1999] Jones, M. and Rehg, J. (1999). Statistical color models with application to skin detection. *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, 1:–280 Vol. 1.

[Jones and Viola, 2003] Jones, M. and Viola, P. (2003). Fast multi-view face detection. Technical report, Mitsubishi Electric Research Labs. Technical Report TR2003-96.

[Kovac et al., 2003] Kovac, J., Peer, P., and Solina, F. (2003). Human skin color clustering for face detection. *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, 2:144–148 vol.2.

[Kruger et al., 2000] Kruger, V., Bruns, S., and Sommer, G. (2000). Efficient head pose estimation with gabor wavelet networks. *British Machine Vision Conference*, pages 12–14.

[Langton et al., 2000] Langton, S. R., Watt, R. J., and Bruce, I. (2000). Do the eyes have it? cues to the direction of social attention. *Trends Cogn Sci*, 4(2):50–59.

[Lienhart and Maydt, 2002] Lienhart, R. and Maydt, J. (2002). An extended set of haar-like features for rapid object detection. *Image Processing. 2002. Proceedings. 2002 International Conference on*, 1:I–900–I–903 vol.1.

[Lipton et al., 1998] Lipton, A. J., Fujiyoshi, H., and Patil, R. S. (1998). Moving target classification and tracking from real-time video. In *WACV '98: Proceedings of the 4th IEEE Workshop on Applications of Computer Vision (WACV'98)*, page 8, Washington, DC, USA. IEEE Computer Society.

[Maglio et al., 2000] Maglio, P. P., Matlock, T., Campbell, C. S., Zhai, S., and Smith, B. A. (2000). Gaze and speech in attentive user interfaces. In *ICMI '00: Proceedings of the Third International Conference on Advances in Multimodal Interfaces*, pages 1–7, London, UK. Springer-Verlag.

[Menser and Wien, 2000] Menser, B. and Wien, M. (2000). Segmentation and tracking of facial regions in color image sequences. In Ngan, K. N., Sikora, T., and Sun, M.-T., editors, *Visual Communications and Image Processing 2000*, volume 4067, pages 731–740. SPIE.

[Morellas et al., 2003] Morellas, V., Pavlidis, I., and Tsiamyrtzis, P. (2003). Deter: detection of events for threat evaluation and recognition. *Mach. Vision Appl.*, 15(1):29–45.

[Musse et al., 2007] Musse, S. R., Jung, C. R., Julio C. S. Jacques, J., and Braun, A. (2007). Using computer vision to simulate the motion of virtual agents: Research articles. *Comput. Animat. Virtual Worlds*, 18(2):83–93.

[N. A. Abdul Rahim, 2006] N. A. Abdul Rahim, C. W. Kit, J. S. (2006). Rgb-h-cbcr skin colour model for human face detection. *MMU International Symposium on Information and Communications Technologies (M2USIC 2006)*.

[Odobez and Ba, 2007] Odobez, J.-M. and Ba, S. (2007). A cognitive and unsupervised map adaptation approach to the recognition of the focus of attention from head pose. *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1379–1382.

[Pai et al., 2003] Pai, C.-J., Tyan, H.-R., Liang, Y.-M., Liao, H.-Y. M., and Chen, S.-W. (2003). Pedestrian detection and tracking at crossroads. *Image Processing,*

2003. ICIP 2003. Proceedings. 2003 International Conference on, 2:II–101–4 vol.3.

[Pflugfelder and Bischof, 2007] Pflugfelder, R. and Bischof, H. (2007). People tracking across two distant self-calibrated cameras. *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 393–398.

[Phung et al., 2005] Phung, S., Bouzerdoum, A., S., and Chai, D., S. (2005). Skin segmentation using color pixel classification: analysis and comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(1):148–154.

[Phung et al., 2002] Phung, S. L., Bouzerdoum, A., and Chai, D. (2002). A novel skin color model in ycbcr color space and its application to human face detection. *Image Processing. 2002. Proceedings. 2002 International Conference on*, 1:I–289–I–292 vol.1.

[Porikli et al., 2006] Porikli, F., Tuzel, O., and Meer, P. (2006). Covariance tracking using model update based on lie algebra. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1:728–735.

[Rae and Ritter, 1998] Rae, R. and Ritter, H. (1998). Recognition of human head orientation based on artificial neural networks. *Neural Networks, IEEE Transactions on*, 9(2):257–265.

[Ramanan et al., 2007] Ramanan, D., Forsyth, D., and Zisserman, A. (2007). Tracking people by learning their appearance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(1):65–81.

[Räihäa and Duchowski, 2006] Räihäa, K.-J. and Duchowski, A. T., editors (2006). *ETRA '06: Proceedings of the 2006 symposium on Eye tracking research and applications*, New York, NY, USA. ACM Press.

[Robertson and Reid, 2005] Robertson, N. and Reid, I. (2005). Behaviour understanding in video: a combined method. *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, 1:808–815 Vol. 1.

[Robertson and Reid, 2006a] Robertson, N. and Reid, I. (2006a). Estimating gaze direction from low-resolution faces in video. In *Proc. 9th IEEE Eur. Conf. on Computer Vision, Graz, Austria*, volume 3952/2006, pages 402–415.

[Robertson and Reid, 2006b] Robertson, N. and Reid, I. (2006b). A general method for human activity recognition in video. *Comput. Vis. Image Underst.*, 104(2):232–248.

[Robertson et al., 2005] Robertson, N. M., Reid, I. D., and Brady, J. M. (2005). What are you looking at? gaze estimation in medium-scale images. *HAREM05, 16th British Machine Vision Conference*, 1.

[Rosales and Sclaroff, 2000] Rosales, R. and Sclaroff, S. (2000). Learning and synthesizing human body motion and posture. *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 506–511.

[Sakaue et al., 2006] Sakaue, F., Kobayashi, M., Migita, T., and Shakunaga, T. (2006). A real-life test of face recognition system for dialogue interface robot in ubiquitous environments. *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 3:1155–1160.

[Sigal et al., 2000] Sigal, L., Sclaroff, S., and Athitsos, V. (2000). Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 2:152–159 vol.2.

[Skarbek and Koschan, 1994] Skarbek, W. and Koschan, A. (1994). Colour image segmentation – a survey. technical report 32. Technical report, Technical University of Berlin, Department of Computer Science.

[Smith et al., 2006a] Smith, K., Ba, S. O., Gatica-Perez, D., and Odobez, J.-M. (2006a). Tracking the multi person wandering visual focus of attention. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 265–272, New York, NY, USA. ACM.

[Smith et al., 2006b] Smith, T. J., Whitwell, M., and Lee, J. (2006b). Eye movements and pupil dilation during event perception. In *ETRA '06: Proceedings of the 2006 symposium on Eye tracking research & applications*, pages 48–48, New York, NY, USA. ACM.

[Soriano et al., 2000] Soriano, M., Martinkauppi, B., Huovinen, S., and Laaksonen, M. (2000). Skin detection in video under changing illumination conditions. *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, 1:839–842 vol.1.

[Spagnolo et al., 2003] Spagnolo, P., Leo, M., Leone, A., Attolico, G., and Distante, A. (2003). Posture estimation in visual surveillance of archaeological sites. *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.*, pages 277–283.

[Stiefelhagen, 2002] Stiefelhagen, R. (2002). Tracking focus of attention in meetings. *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pages 273–280.

[Stiefelhagen, 2004] Stiefelhagen, R. (2004). Estimating Head Pose with Neural Networks. In *POINTING 2004, Visual Observation of Deictic Gestures, In association with ICPR 04, Cambridge, UK*.

[Stiefelhagen et al., 2000] Stiefelhagen, R., Yang, J., and Waibel, A. (2000). Simultaneous tracking of head poses in a panoramic view. *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, 3:722–725 vol.3.

[Terrillon et al., 2000] Terrillon, J.-C., Shirazi, M., Fukamachi, H., and Akamatsu, S. (2000). Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 54–61.

[Tian et al., 2005] Tian, Y.-L., Lu, M., and Hampapur, A. (2005). Robust and efficient foreground analysis for real-time video surveillance. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 1182–1187 vol. 1.

[Vezhnevets et al., 2003] Vezhnevets, V., Sazonov, V., and Andreeva, A. (2003). A survey on pixel-based skin color detection techniques. In *Proceedings of the International Conference on Computer Graphics (GRAPHICON03)*, pages 85–92.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 1:I–511–I–518 vol.1.

[Voit et al., 2006] Voit, M., Nickel, K., and Stiefelhagen, R. (2006). A bayesian approach for multi-view head pose estimation. *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, pages 31–34.

[Voit et al., 2007] Voit, M., Nickel, K., and Stiefelhagen, R. (2007). Neural network-based head pose estimation and multi-view fusion. In *Multimodal Technologies for Perception of Humans*, pages 291–298. Springer.

[Wang et al., 2005a] Wang, G., Wong, T.-T., and Heng, P.-A. (2005a). Real-time surveillance video display with salience. In *VSSN '05: Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, pages 37–44, New York, NY, USA. ACM.

[Wang and Ji, 2004] Wang, P. and Ji, Q. (2004). Multi-view face detection under complex scene based on combined svms. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 4:179–182 Vol.4.

[Wang et al., 2005b] Wang, Y., Tan, T., Loe, K.-F., and Wu, J.-K. (2005b). A probabilistic approach for foreground and shadow segmentation in monocular image sequences. *Pattern Recognition*, 38(11):1937–1946.

[Wren et al., 1997] Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A. (1997). Pfinder: real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):780–785.

[Xie and Lin, 2007] Xie, J. and Lin, X. (2007). Gaze direction estimation based on natural head movements. *Image and Graphics, 2007. ICIG 2007. Fourth International Conference on*, pages 672–677.

[Yang et al., 2002] Yang, M.-H., Kriegman, D., and Ahuja, N. (2002). Detecting faces in images: a survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(1):34–58.

[Yang et al., 2004] Yang, T., Li, S. Z., Pan, Q., and Li, J. (2004). Real-time and accurate segmentation of moving objects in dynamic scene. In *VSSN '04: Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks*, pages 136–143, New York, NY, USA. ACM.

[Yao et al., 2001] Yao, P., Evans, G., and Calway, A. (2001). Face tracking and pose estimation using affine motion parameters. In Austvoll, I., editor, *Proceedings of the 12th Scandinavian Conference on Image Analysis*, pages 531–536. Norwegian Society for Image Processing and Pattern Recognition.

[Yilmaz et al., 2006] Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13.

[Zarit et al., 1999] Zarit, B., Super, B., and Quek, F. (1999). Comparison of five color models in skin pixel classification. *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 1999. Proceedings. International Workshop on*, pages 58–63.

[Zhao et al., 2002] Zhao, L., Pingali, G., and Carlbom, I. (2002). Real-time head orientation estimation using neural networks. *Image Processing. 2002. Proceedings. 2002 International Conference on*, 1:I–297–I–300 vol.1.

[Zhu and Yang, 2002] Zhu, J. and Yang, J. (2002). Subpixel eye gaze tracking. *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 124–129.