PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DETECÇÃO DE ESQUELETOS 2D BASEADA EM IMAGENS

HUMBERTO DUMONT SOUTO JÚNIOR

Dissertação de Mestrado apresentada como requisito para obtenção do título de Mestre em Ciência da Computação pelo Programa de Pós-graduação da Faculdade de Informática. Área de concentração: Ciência da Computação.

Orientadora: Profa. Dra. Soraia Raupp Musse

Porto Alegre, Brasil
2011

Dados Internacionais de Catalogação na Publicação (CIP)

S728d Souto Júnior, Humberto Dumont
Detecção de esqueletos 2D baseada em imagens / Humberto Dumont Souto Júnior. – Porto Alegre, 2011.
73 p.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientadora: Profa. Dra. Soraia Raupp Musse.

1. Informática. 2. Redes Neurais (Computação). 3. Visão por Computador. 4. Percepção Visual (Computação). 5. Processamento de Imagens. 6. Esqueleto. I. Musse, Soraia Raupp. II. Título.

CDD 006.37

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada **"Detecção de Esqueletos 2D Baseada em Imagens"**, apresentada por Humberto Dumont Souto Júnior, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Sistemas Interativos e de Visualização, aprovada em 23/03/2011 pela Comissão Examinadora:

Profa. Dra. Soraia Raupp Musse– Orientadora — PPPGCC/PUCRS

Profa. Dra. Milene Selbach Silveira - — PPGCC/PUCRS

Profa. Dra. Silvia Silva da Costa Botelho – — FURG

Homologada em 30/08/11, conforme Ata No. 17 pela Comissão Coordenadora.

Prof. Dr. Fernando Luís Dotti
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320–3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

“Não é fácil ter paciência diante dos que têm excesso de paciência.”

Carlos Drummond de Andrade

AGRADECIMENTOS

A minha orientadora Soraia, pela dedicação durante as orientações.

A minha noiva Luciele pelo apoio e por compreender minha ausência em momentos que ela precisou.

A meus pais Humberto e Laura e meus avós Eliseu, Leda e Maria, pelo apoio nos momentos difíceis.

Ao amigo Ulisses pela grande ajuda e paciência.

Aos grandes amigos que fiz nestes 2 anos de mestrado, junto ao Laboratório de pesquisa VHLAB (Adriana, Henry, Rafael, Rossana, Juliano, Leandro, Fernando, Luiz, Anderson, Júlio, ...).

Aos colegas de mestrado, os quais se tornaram grandes amigos.

Este trabalho foi desenvolvido em colaboração com a HP Brazil P&D.

DETECÇÃO DE ESQUELETOS 2D BASEADA EM IMAGENS

RESUMO

Existem diversas situações em que a detecção de postura de uma pessoa em fotografias desempenha um papel importante, tais como: sistemas de segurança, medição de desempenho atlético, aplicações em sistemas de realidade virtual, dentre outras. Além disso, saber a posição dos membros de uma pessoa em uma fotografia pode revelar inclusive muito sobre o contexto que a fotografia foi tirada. Podendo revelar, por exemplo, se a cena trata-se de uma festa e se a pessoa está feliz. Esta dissertação apresenta um conjunto de métodos que, quando utilizados juntamente, realizam a tarefa de estimar a postura de uma pessoa a partir de uma única imagem, sem que haja intervenção externa de um usuário e sem que se tenha qualquer conhecimento prévio sobre a fotografia (tais como posição do fotógrafo em relação à pessoa, textura de roupa ou se há membros oclusos).

Palavras-chave: Detecção; Esqueleto; Visão computacional; Rede Neural.

DETECTION OF 2D SKELETONS BASED ON IMAGES

ABSTRACT

There are several situations in which posture detection of humans in pictures has an important role, such as: security systems, measuring athletic performance, virtual reality applications, among others. Besides that, knowing the members locations of a person in a picture may reveal information about the picture context. One example of context can be if the picture is illustrating a party with a happy person. This work presents a set of methods to provide automatic posture detection from a single image, without any manual intervention or any previous knowledge about the picture (for example, photographer's position, clothes texture or even ocluded members).

Keywords: Skeleton detection; Computer vision; Neural network.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

MLP	*Multi Layer Perceptron*
MRF	*Markov Random Field*
MCMC	*Markov Chain Monte Carlo*
3D	Três dimensões
2D	Duas Dimensões
MSE	Mean Squared Error
VHLAB	*Virtual Humans Simulation Laboratory*

SUMÁRIO

1. INTRODUÇÃO

O processo de estimar automaticamente a postura de pessoas em fotografias não é uma atividade simples de ser realizada, devido ao grande número de graus de liberdade presentes no corpo humano e das possíveis oclusões inerentes a observação monocular [26]. Contudo, esta tarefa pode ser de suma importância para diversas áreas, pois permite que seja identificada a atividade que está sendo desempenhada por uma pessoa [18]. Dessa forma, é possível citar algumas potenciais aplicações diretas para o processo de detecção de postura:

- sistemas de segurança;
- busca em banco de dados de imagens por fotografias em que pessoas estão realizando certa atividade;
- controle de pedestres em aplicações de tráfego;
- medição de desempenho atlético;
- aplicações em sistemas de realidade virtual.

Atualmente, uma aplicação comercial que faz uso direto da detecção automática de postura é o Kinect[1], o qual trata de um dispositivo que traz uma série de sensores (tais como câmera RGB, infravermelho e microfone) para o videogame Xbox 360, produzido pela empresa Microsoft. O aparelho, quando posicionado em frente à televisão, é capaz de realizar a detecção do movimento do corpo inteiro das pessoas em sua frente, além do reconhecimento de rosto e voz, dispensando por completo a utilização de controles e *joysticks* [22]. A Figura 1.1 mostra uma imagem do jogo “Harry Potter e os Talismãs da Morte” sendo jogado utilizando-se o dispositivo Kinect.

Nas próximas seções serão discutidos o problema, a motivação, os objetivos e a estrutura desta dissertação.

1.1 Problema

Neste trabalho definiu-se “postura global” como sendo as informações de alto nível sobre a posição do corpo de uma pessoa. Sendo ainda, a postura em si e a definição das posições das principais articulações do corpo.

Tento em vista o conceito introduzido no parágrafo anterior, o problema sob investigação desta pesquisa é a seguir caracterizado: dada uma fotografia que apareça uma pessoa, a

[1] Mais informações disponíveis em: http://www.xbox.com/pt-br/kinect

Figura 1.1 – Imagem do jogo “Harry Potter e os Talismãs da Morte” sendo jogado utilizando-se o dispositivo Kinect acoplado ao videogame Xbox 360 [22], produzindo pela empresa Microsoft.

qual esteja com a face ligeiramente voltada para frente, automaticamente encontrar uma postura global (quais membros estão aparentes do ponto de vista em que a fotografia foi tirada) e estimar um esqueleto 2D para esta pessoa (coordenada das principais articulações do corpo).

Para que tal tipo de processamento seja possível, é necessário que o fundo da imagem (*background*) seja removido para que o *foreground* possa ser utilizado. Esta tarefa geralmente não é simples de ser realizada, embora existam muitos trabalhos de pesquisa recentes tais como [15], [17] e [30].

A questão de pesquisa deste trabalho é realizar a tarefa de tendo-se como entrada simplesmente uma fotografia na qual apareça uma pessoa, encontrar a postura desta sem que haja qualquer intervenção externa tais como cliques do usuário, por exemplo.

1.2 Motivação

Existem diversas situações em que a detecção de postura desempenha um papel importante, como as citadas anteriormente: sistemas de segurança, medição de desempenho atlético, aplicações em sistemas de realidade virtual, dentre outras. Contudo, a principal motivação na escolha do tema para este trabalho foi o desafio de realizar a tarefa de estimar a postura para uma pessoa a partir de uma única imagem (e não um vídeo, ocasião em que se tem uma sequencia de imagens e, consequentemente, informação temporal), sem que haja intervenção do usuário e sem que se tenha qualquer conhecimento prévio sobre a fotografia (tais como posição do fotógrafo em relação à pessoa, textura de roupa ou se há

membros oclusos).

Como já foi dito anteriormente, saber a posição dos membros de uma pessoa em uma fotografia pode revelar muito sobre a imagem. Esta informação pode revelar desde a atividade que está sendo desempenhada, assim como é utilizado no Kinect [22], a até mesmo se a cena trata-se de uma festa e se a pessoa está feliz. Portanto, o nível de precisão das poses obtidas pode comprometer os resultados de processos posteriores, tornando este um grande desafio em aberto.

1.3 Objetivos

De forma mais específica, são objetivos desse trabalho:

- Realizar um estudo sobre as técnicas utilizadas na literatura para a detecção de poses;
- Definir uma estrutura para o modelo de detecção de posturas;
- Desenvolver o modelo de detecção de posturas;
- Avaliar os resultados obtidos.

1.4 Estrutura da Dissertação

O Capítulo 2 apresenta uma fundamentação teórica sobre alguns conceitos necessários para o entendimento deste trabalho. Já no Capítulo 3, são apresentados diversos trabalhos científicos relacionados ao tema desta dissertação, os quais serviram como referencial teórico para o desenvolvimento do modelo que será apresentado. O Capítulo 4 faz uma introdução aos métodos utilizados para realizar a segmentação da imagem e após apresenta o modelo proposto para a detecção de postura, sendo este dividido em três etapas.

Os resultados obtidos nos testes do modelo são expostos no Capítulo 5. Já no Capítulo 6 são expostas algumas considerações sobre o trabalho realizado, destacadas as principais contribuições e sugeridos alguns trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são introduzidos, resumidamente, alguns conceitos básicos que se farão necessários ao entendimento tanto deste trabalho como de alguns outros presentes no capítulo de revisão bibliográfica.

2.1 *Template Matching*

Template Matching consiste em um método bastante utilizado em visão computacional, contudo, a fim de definir do que se trata este método, primeiramente deve-se esclarecer o significado das palavras *template* e *matching* [3].

A palavra *template* é definida por Brunelli como sendo algo desenvolvido para servir de modelo para alguma coisa. Brunelli também define *matching* como sendo uma comparação de similaridade com relação a forma, analisando-se a semelhança ou a diferença. Um *template* pode adicionalmente exibir alguma variabilidade, ou seja, nem todas as instâncias precisam ser exatamente iguais. Um exemplo importante disto é o fato um objeto poder ser observado de diferentes pontos de vista. Mudanças de iluminação, sensores de imagens ou configuração dos sensores podem também criar variações significantes. Ainda outra forma de variabilidade deriva da diferença físicas de uma pessoa para outra, que acabam causando variações nos padrões da imagem correspondente: considerando as muitas variações de poses, todas elas partilham uma estrutura básica (o esqueleto), mas também exibem diferenças, por exemplo, nem todas pessoas possuem o braço com a mesma proporção em relação a sua altura. Outra fonte importante de variabilidade consiste na evolução temporal de um único objeto, um interessante exemplo é a boca durante a fala [3].

Dessa forma, o método *Template Matching*, em visão computacional, nada mais é do que o ato de comparar uma imagem com um modelo, ou conjunto deles, a fim de que se possa chegar a conclusão de que esta imagem se parece ou não com esse modelo.

2.2 Antropometria

Antropometria significa, literalmente, a medição de pessoas [27]. Contudo, ela surgiu para ser utilizada com um sentido mais específico, realizar um estudo comparativo dos tamanhos e proporções do corpo humano. A antropometria já estava presente desde a época da Renascença, quando as medidas do corpo e suas proporções eram utilizadas por pintores e escultores a fim de reproduzir uma imagem humana em tamanho não natural, mas com proporções reais. No entanto, a antropometria não serve somente para estimar as proporções do corpo humano, mas também para dizer quais as posições possíveis que

um corpo pode se deformar, restringindo movimentos e ângulos.

Muitos dos dados de antropometria utilizados nos dias de hoje são derivados de dados medidos manualmente em um conjunto de população há aproximadamente 50 anos atrás [27]. Contudo, melhorias na alimentação e nos cuidados com a saúde nos últimos anos resultaram em um aumento nas dimensões médias da população, tornando assim, estes dados de 50 anos obsoletos. Desta forma, pesquisas conduzidas pelos exércitos britânico e americano tem se esforçado para remediar esta situação fazendo uso de um escâner 3D para obter, com exatidão e rapidez, dados do corpo humano em três dimensões [27].

Assim sendo, o método antropométrico continua sendo utilizado nos dias de hoje na indústria (para determinar o posicionamento de botões e instrumentos em um painel de carro ou avião, por exemplo), para solucionar problemas de visão computacional e também na criação de humanos virtuais [27].

O livro *The Measure of Man and Woman: Human Factors in Design* [28] de Alvin R. Tilley & Henry Dreyfuss Associates, por ser bastante completo, foi utilizado como base para todos os cálculos antropométricos realizados neste trabalho. Um resumo das médias deste livro é apresentado na Tabela 2.1, na qual w_{hi} representa o comprimento do osso de índice i em relação a altura da pessoa e w_{wi} a largura do osso i em relação ao próprio comprimento, juntamente com a Figura 2.1.

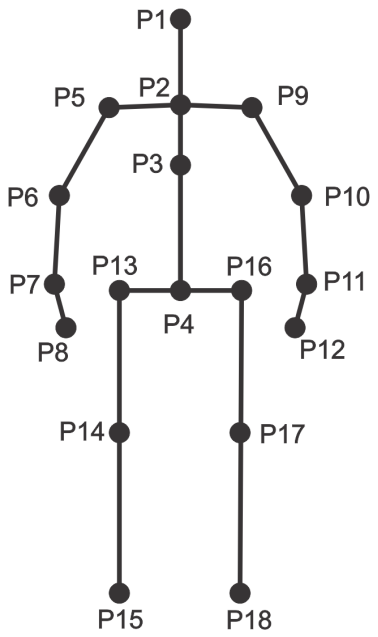


Figura 2.1 – Resumo dos ossos e articulações do livro *The Measure of Man and Woman: Human Factors in Design* [28]. As medidas antropométricas resumidas para cada ponto desta Figura são informadas na Tabela 2.1

Tabela 2.1 – Resumo das medidas antropométricas para o corpo humano, sendo w_{hi} o comprimento do osso de índice i em relação a altura da pessoa e w_{wi} a largura do osso i em relação ao próprio comprimento. Primeira coluna: índice da parte do corpo; segunda coluna: Parte do corpo (osso); terceira coluna: as duas articulações que formam cada osso; quarta coluna: pesos utilizados para calcular o comprimento de cada osso; quinta coluna: pesos utilizados para calcular a largura de cada osso.

i	**Osso**	**Articulação**	w_{hi}	w_{wi}
0	Cabeça	(P1 - P2)	*0.20*	*0.0883*
1	Peito	(P2 - P3)	*0.098*	*0.1751*
2	Abdomem	(P3 - P4)	*0.172*	*0.1751*
3	Clavícula Direita	(P2 - P5)	*0.102*	*não usada*
4	Braço Direito	(P5 - P6)	*0.159*	*0.0608*
5	Antebraço Direito	(P6 - P7)	*0.146*	*0.0492*
6	Mão Direita	(P7 - P8)	*0.108*	*0.0593*
7	Clavícula Esquerda	(P2 - P9)	*0.102*	*não usada*
8	Braço Esquerdo	(P9 - P10)	*0.159*	*0.0608*
9	Antebraço Esquerdo	(P10 - P11)	*0.146*	*0.0492*
10	Mão Esquerda	(P11 - P12)	*0.108*	*0.0593*
11	Quadril Direito	(P4 - P13)	*0.050*	*não usada*
12	Coxa Direita	(P13 - P14)	*0.241*	*0.0912*
13	Panturrilha Direito	(P14 - P15)	*0.240*	*0.0608*
15	Quadril Esquerdo	(P4 - P16)	*0.050*	*não usada*
16	Coxa Esquerda	(P16 - P17)	*0.241*	*0.0912*
17	Panturrilha Esquerda	(P17 - P18)	*0.240*	*0.0608*

2.3 Redes Neurais Artificiais

Redes Neurais Artificiais, ou simplesmente redes neurais, tem sido amplamente usadas para solucionar diversos problemas de reconhecimento de padrões, bem como de aproximação de funções complexas [11]. Estas estruturas têm sido utilizadas com sucesso em aplicações de domínios extremamente diversificados, desde controle robótico, até visão computacional, passando por aproximações de curvas complexas, predição de condições do mercado financeiro, diagnóstico médico, entre outras [12].

Dentre os diversos tipos de redes neurais, pode-se dizer que a rede neural do tipo perceptron de múltiplas camadas (MLP - do inglês *Multi Layer Perceptron*) é a mais popular [10].

A Figura 2.2 apresenta a topologia genérica de uma rede neural do tipo MLP. Os elementos $i_1, ..., i_n$ denotam os n *perceptrons* da camada de entrada, que por definição não realizam processamento sobre os dados, sendo então classificados como transparentes.

Os m elementos da camada intermediária, ou escondida, são apresentados como $h_1, ..., h_m$. Esta camada, diferentemente da anterior, apresenta processamento sobre os dados de entrada assim como a terceira camada, ou camada de saída, de onde os resultados serão lidos. Os elementos $o_1, ..., o_p$ denotam os p neurônios desta última camada.

Para cada neurônio *ativo* existe um conjunto de pesos associados, estes têm função

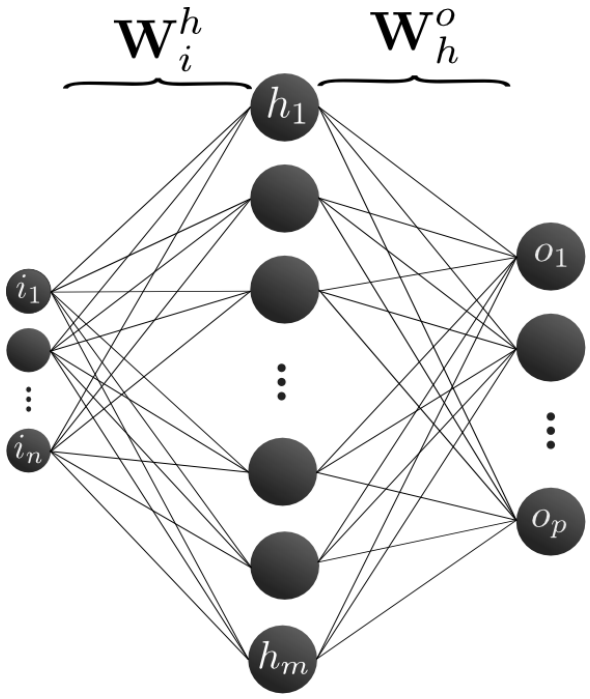


Figura 2.2 – Topologia Genérica de uma Rede Neural do tipo MLP [10].

de ponderar as saídas da camada anterior. Os pesos, ou *sinapses*, são os coeficientes que efetivamente aprenderão os exemplos apresentados à rede. W_i^h e W_h^o, por sua vez representam as matrizes de pesos da camada de entrada pra camada oculta e da camada oculta para a camada de saída, respectivamente.

A saída de cada neurônio da camada escondida obedece a um formalismo matemático expresso pela Equação 2.1, sendo que net_h e h_j denotam: a função de ativação e um neurônio genérico da camada escondida, respectivamente e b_j representa o *bias* (peso especial que normalmente é aplicado nas camadas de uma rede neural para aumentar os graus de liberdade, permitindo uma melhor adaptação, por parte da rede neural, ao conhecimento à ela fornecido) do neurônio h_j.

$$h_j = net_h(b_j + \sum_{k=1}^{n} w_j^k \cdot i_k) \tag{2.1}$$

Já a saída da rede neural, obtida dos neurônios da camada o, obedece à Equação 2.2, na qual net_o denota a função de ativação escolhida para a camada de saída, o_j representa um neurônio genérico nesta camada e b_j representa o viés deste neurônio.

$$o_j = net_o(b_j + \sum_{k=1}^{m} w_j^k \cdot h_k) \tag{2.2}$$

As funções de ativação dos neurônios podem apresentar diversos comportamentos,

dentre as mais utilizadas temos as funções: linear e sigmóides (logística e tangente hiperbólica). Estas funções apresentam importante papel no que se refere ao tipo de dados a serem classificados ou preditos. Com funções lineares pode-se alcançar maiores magnitudes nas saídas, já com funções não lineares como as sigmóides tem-se curvas mais acentuadas [12].

Em geral, assim como neste trabalho, as redes MLP são treinadas através da aplicação de *retropropagação do erro de aprendizado*. Mais informações sobre este tema podem ser obtidas em várias literaturas, dentre elas pode-se citar [11, 12].

O capítulo a seguir apresenta um estudo sobre diversos trabalhos relacionas a detecção de posturas.

3. TRABALHOS RELACIONADOS

Neste capítulo será apresentada uma revisão bibliográfica dos trabalhos referentes à detecção de posturas de pessoas em imagens. A maior parte dos trabalhos atuais relacionados a este tema utilizam imagens de vídeo, facilitando a tarefa de estimar as posições dos membros do corpo, e até mesmo a de remoção do *background*, pois a ocorrência temporal ajuda no processo de segmentação. Embora o foco deste trabalho seja estimar posturas de pessoas em fotografias, ou seja, um único *frame* e não uma sequência deles, optou-se por também relatar alguns dos métodos que focam em vídeos pois os resultados obtidos são similares e, portanto, passíveis de comparação com modelo que será apresentado no Capítulo 4 desta dissertação.

A seguir serão apresentados alguns dos trabalhos analisados durante o processo de revisão bibliográfica. Para fins de melhor organização, os trabalhos foram divididos em dois grupos, sendo eles:

- Detecção de posturas baseada em vídeo;
- Detecção de posturas baseada em fotografia.

3.1 Detecção de Posturas Baseada em Vídeo

Em McIntosh et al. [21] é proposta uma técnica para detecção de articulações que se baseia na combinação de um modelo existente de estimativa de articulações com um método robusto de localização de deformações físicas. Mais especificamente, a técnica desenvolvida utiliza um método chamado *Cardboard People* [16], o qual representa os membros de uma pessoa por partes planas (exatamente como se fosse uma “pessoa de papelão”, Figura 3.1(a)), sendo realizados alguns aperfeiçoamentos em cima deste modelo tendo em vista uma melhor precisão de encaixe no corpo da pessoa da imagem (Figura 3.1(b)). Além disso, a técnica *Cardboard People* foi originalmente desenvolvida para ser utilizada em imagens estáticas, então, foi realizada também uma adaptação desta para trabalhar com imagens dinâmicas. O resultado obtido foi uma técnica robusta capaz de encontrar as posições das articulações e membros do corpo de pessoas em vídeos.

Outra técnica sobre detecção de posturas baseada em vídeo é abordada em Dimitrijevic et al. [6]. Neste trabalho é apresentada uma abordagem que realiza um treinamento a partir de um banco de dados de modelos de silhuetas humanas em uma pose de caminhada. Esses modelos consistem em pequenas sequências de silhuetas 2D obtidas a partir de dados de captura de movimento, permitindo assim incorporar informações de movimento dentro destes modelos e ainda ajudando a diferenciar movimento de pessoas de objetos

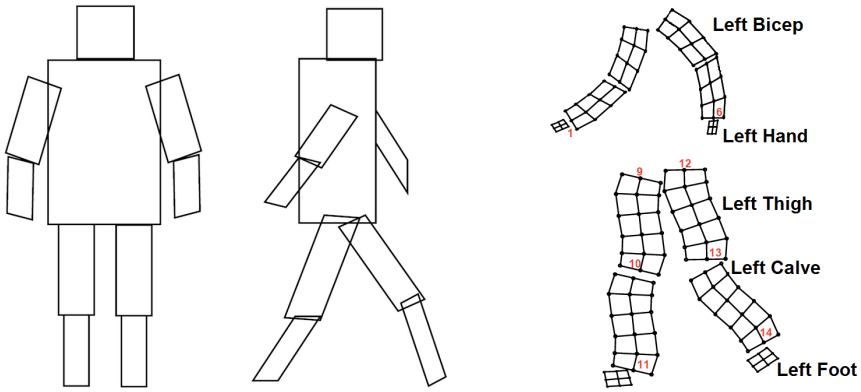


Figura 3.1 – (a) Modelo proposto em [16] para representar posturas, sendo os membros de uma pessoa representados por partes planas. (b) Alteração proposta por McIntosh et al. [21] em cima do modelo de [16] visando uma um melhor encaixe no corpo da pessoa da imagem.

cujos contornos se assemelham aos dos seres humanos. Durante a fase de treinamento é feita a utilização de técnicas estatísticas de aprendizagem para estimar e armazenar a relevância das diferentes partes da silhueta para, posteriormente, ser empregada na tarefa de reconhecimento. Em tempo de execução, estas informações são utilizadas para converter a distância de *Chamfer* [25, 8] na estimativa de probabilidade mais significante. Outra característica deste trabalho é que os modelos podem lidar com seis diferentes pontos de vista de câmera, com exceção das vistas frontal e traseira, bem como em diferentes escalas.

Para demonstrar a eficácia desta técnica, são utilizados vídeos de pessoas andando em frente a *backgrounds* não homogêneos, em ambientes internos e externos e, além disso, adquiridos com uma câmera em movimento, o que torna difícil a utilização de técnicas tais como subtração de fundo. Um exemplo dos resultados obtidos é mostrado na Figura 3.2.

Como anteriormente descrito, a detecção de posturas em vídeos pode utilizar de múltiplas fontes de câmeras bem como do tempo. Na próxima seção são apresentados trabalhos baseados em fotografia, os quais não dispõem deste tipo de informações.

3.2 Detecção de Posturas Baseada em Fotografia

Vários trabalhos sobre detecção de esqueleto em fotografias são encontrados na literatura, atualmente. Pode-se citar como exemplo Mori et al. [24], o qual trata o problema de, em uma imagem, ao se tentar detectar individualmente uma parte do corpo humano, perceber-se que estas partes confundem-se frequentemente com objetos do fundo. Por exemplo, um braço pode facilmente se parecer com um tronco de árvore ou até mesmo com um pedaço de grama. Isto acontece porque somente o contexto global da imagem

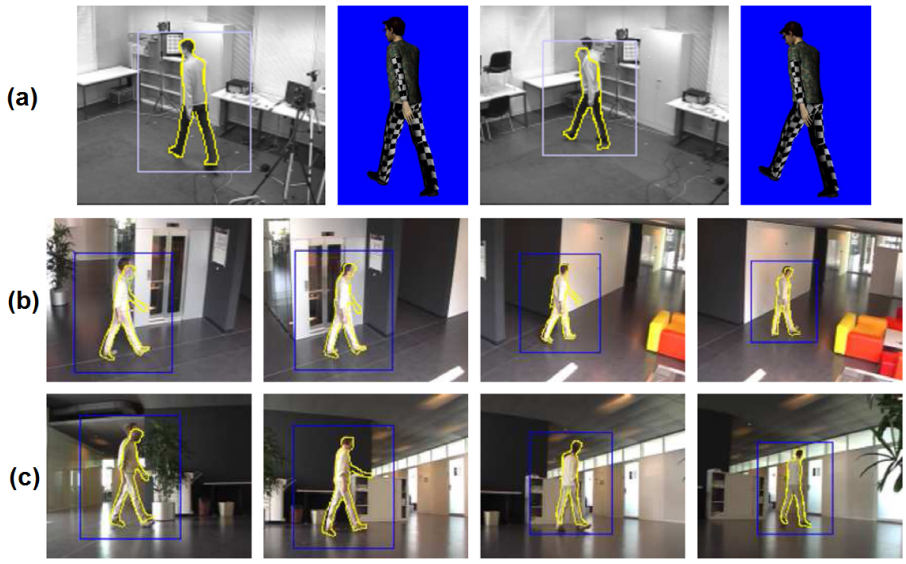


Figura 3.2 – Resultados do modelo de [6]: (a) recuperação correta da postura demonstrada no modelo 3D; (b) postura detectada corretamente mesmo com a câmera posicionada muito acima do que as câmeras utilizadas para criar os *templates*; (c) silhueta detectada corretamente com a câmera colocada a apenas 30cm acima do chão.

está sendo levado em consideração. Desta forma, propõe-se a utilização de um método que realize uma busca em "baixo nível" na fotografia, ou seja, levando em consideração informações que estão sempre presentes independentemente do contexto. Partindo-se de "saliências", que podem ser um cotovelo, uma mão ou um rosto, e fazendo-se perguntas baseadas no contexto, ("se isto é um cotovelo e aquilo um torso, então esta linha deve ser um braço") estas saliências vão se interligando. Além disso, são impostas restrições globais sobre as configurações do corpo humano a fim de prevenir que posições impossíveis de se obter com o corpo humano sejam encontradas. As configurações parciais restantes são então estendidas para as figuras humanas completas a fim de buscar os membros faltantes e, para auxiliar a segmentação das saliências, é utilizado o detector de bordas Canny [4] e também realiza-se uma sobre-segmentação da imagem em um grande número de pequenas regiões, as quais foram chamadas de "*superpixels*".

A tarefa de encontrar as partes do corpo é realizada em cima dos *superpixels*. Estes são analisados levando-se em consideração o contorno, a forma, o sombreamento e o foco. Após encontrados os segmentos referentes ao corpo, parte-se para a tarefa de combinar partes de modo a, por exemplo, juntar duas partes de uma mesma perna. Isto é realizado com o auxílio de uma função sigmoide, a qual foi treinada previamente com um banco de imagens de partes do corpo, combinando-as linearmente.

Feito isto, a próxima etapa trata de unir as diferentes partes para construção do esqueleto. Ela é realizada gerando-se uma lista com todas as possibilidades de configurações

possíveis para um corpo humano completo. Após, para cada torço, é utilizado um método simples de montagem de configurações o qual independentemente seleciona o melhor membro para ser conectado as articulações do respectivo torço (quadris e ombros). Três exemplos de resultados obtidos podem ser vistos na Figura 3.3.

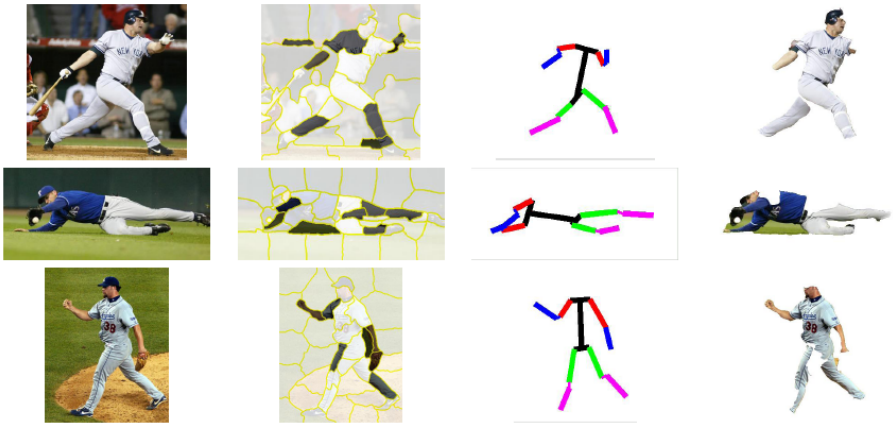


Figura 3.3 – Resultados obtidos com a técnica de Mori et al. [24].

Em Mori et al. [23] o problema considerado é utilizar uma única imagem que contenha a figura de uma pessoa, localizar as posições de suas articulações, e usar esta informação para estimar a configuração e a postura do corpo em 3D. A abordagem básica é armazenar um número de exemplares de imagens 2D de corpos humanos em diferentes configurações e pontos de vista da câmera. Em cada uma dessas imagens armazenadas, a localização das articulações (cotovelos, joelhos, etc.) foram manualmente marcadas e rotuladas para serem futuramente utilizadas. É então utilizada a técnica de *Template Matching* [3] em conjunto com um modelo cinemático de deformação, para se encontrar o modelo que melhor corresponda à postura da pessoa na fotografia.

Outro trabalho sobre estimativa de esqueleto humano em fotografias que apresenta bons resultados é o de Hu et al. [13], o qual propõe um método para recuperação de poses somente para a parte superior do corpo humano. Neste trabalho, três observações são primeiramente detectadas para se inicializar as articulações: face, pele e torso. A face é localizada baseando-se no detector de AdaBoost [29] (Figura 3.4(a)). A pele, que é importante para a detecção dos membros, é segmentada usando-se o método de Markov Random Field (MRF) [20] (Figura 3.4(b)). O torso (Figura 3.4(c)) é a parte chave que conecta a maioria das outras partes do corpo e é relativamente estável em relação a detecção da face, então, foi desenvolvido um detector de torso que utiliza um modelo de deformação o qual representa o torso, Figura 3.5(a). Modificando-se os parâmetros deste modelo: largura w, altura h, inclinação θ e posição do pescoço na imagem (x_0, y_0); este modelo é capaz de estimar diferentes torsos. Desta forma, é construído um vetor com hipóteses de torsos e em seguida estas são avaliadas por uma distribuição probabilística para encontrar qual

melhor se ajusta na imagem. Então, um esqueleto 2D provido de 11 articulações é utilizado para descrever a pose superior do corpo. Resultados obtidos neste trabalho são mostrados na Figura 3.6.

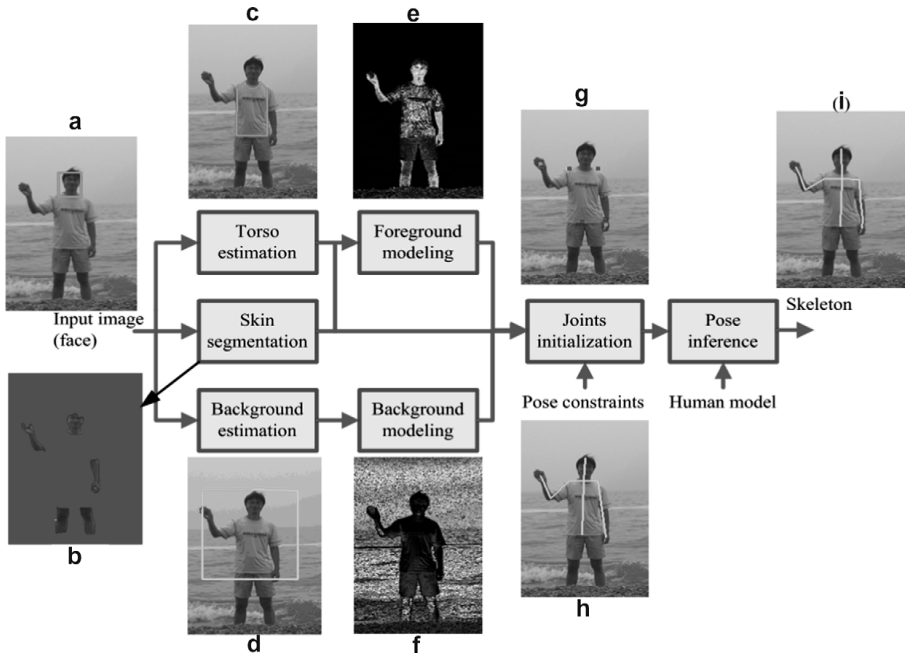


Figura 3.4 – Diagrama de blocos do sistema de detecção de posturas de partes superior do corpo do modelo proposto por Hu et al. [13].

Tendo-se algumas premissas, (ombros, quadris e pescoço são diretamente localizados pelo torso que foi detectado) e assumindo-se que a mão está visível, mãos candidatas são determinadas a partir das regiões de pele detectadas. Os cotovelos são inicializados utilizando-se heurística e a distribuição de cores do primeiro plano (Figura 3.4(e)) e do plano de fundo (Figura 3.4(f)). Na maioria dos casos, a inicialização das juntas não é precisa, mas próxima a ela. Deste modo, a pose final é inferida por meio da utilização do método de *Markov Chain Monte Carlo* (MCMC), no qual a distribuição proposta é baseada no *Random-walk sampler* [9] que é facilmente exemplificado e avaliado.

O trabalho de Lee & Cohen [18] propõe o uso de uma abordagem adaptativa, a qual consiste em um modelo humano que é usado para sintetizar as regiões da imagem que correspondem a formas humanas (dada uma hipótese) e, portanto, separando a pessoa do fundo. Esta abordagem é útil para avaliar as hipóteses de postura, porém surge o problema de como criar estas hipóteses. Desta forma, é proposto resolver este problema através da construção de um modelo gerador de imagem usando o modelo de MCMC para procurar o espaço de soluções (hipóteses).

O modelo corporal humano utilizado para ser encaixado no corpo da pessoa na fotografia é uma união de três partes: a estrutura cinemática, a forma e as roupas do corpo. O

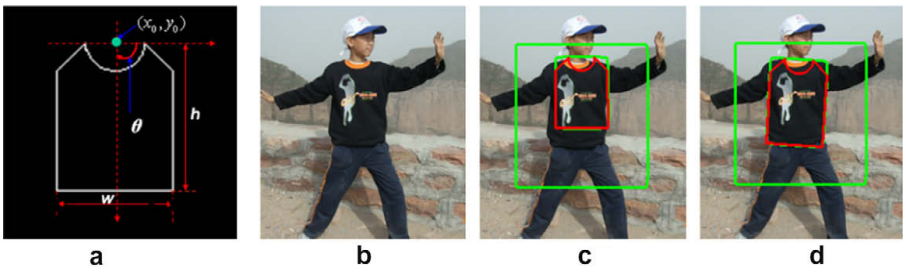


Figura 3.5 – Modelo de deformação de torso e o processo de estimativa do mesmo, proposto por Hu et al. [13]. (a) Modelo de deformação do torso; (b) imagem de entrada; (c) inicialização do torso; (d) torso estimado.

Figura 3.6 – Resultados obtidos por [13] com a utilização de seu método.

componente cinemático é representado por uma estrutura articulada que possui 31 graus de liberdade (Figura 3.7(a)). Já o modelo de formas prováveis (Figura 3.7(b)) é deformado para poder se ajustar dentro da imagem conforme necessário visando melhor representar a postura do corpo, mais especificamente, é utilizado um modelo probabilístico de formas que codificam a variabilidade da forma humana. Assim, cada componente do corpo (torço, membros, etc.) é aproximadamente representado por um cilindro cônico em 3D, sendo que cada cilindro tem três parâmetros: comprimento, altura do topo e altura base (a proporção do cilindro permanece constante mesmo com a variação das medidas). Já o modelo de roupas (Figura 3.7(c)) descreve o tipo de roupas que a pessoa está usando e permite estimar quando regiões de pele estão visíveis. Para isto, este modelo apresenta três parâmetros: o comprimento das mangas, comprimento inferior da roupa e o comprimento das meias.

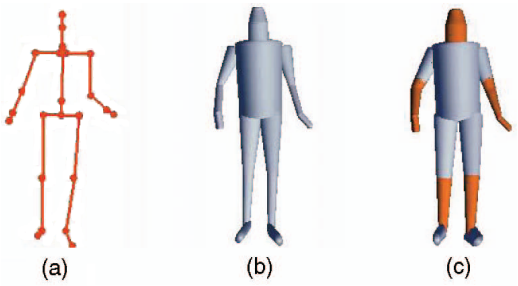


Figura 3.7 – Modelo humano de [18]: (a) modelo cinemático caracterizado pela posição das articulações; (b) forma aproximada do corpo humano representada por um conjunto de cilindros cônicos; (c) roupas da pessoa.

A partir do modelo humano proposto, dado um exemplo de estado que represente uma pose candidata, uma imagem humana pode ser sintetizada e comparada com a verdadeira imagem. Esta abordagem baseada em modelos é atraente uma vez que tenta "separar os dados" do ponto de vista da geração da imagem, resolvendo assim simultaneamente o problema da segmentação. Um exemplo do resultado obtido neste trabalho é mostrado na Figura 3.8.

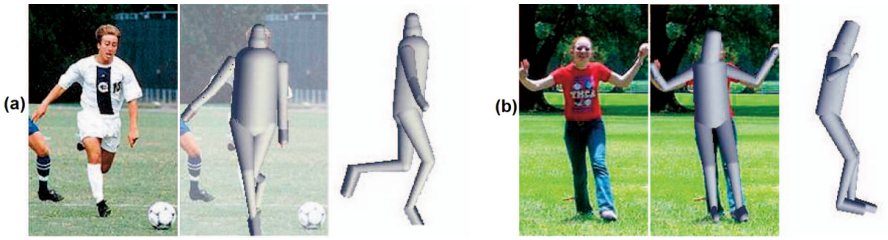
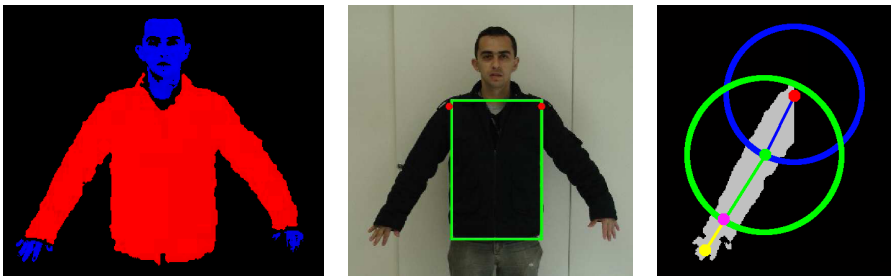


Figura 3.8 – Resultado obtido por [18]: para imagens (a) e (b), imagem original, postura estimada e vista lateral.

Jacques Jr. et al. desenvolveram um modelo [14] que tem por objetivo estimar a postura da parte superior do corpo de pessoas em imagens estáticas utilizando um modelo 2D combinado com dados antropométricos. Para o funcionamento deste modelo é necessário que a pessoa na fotografia esteja em uma pose aproximadamente frontal. Desta forma, o detector de faces de Viola & Jones [29] é utilizado para encontrar as coordenadas x_f e y_f de centro e o raio R da face da pessoa. A face é uma parte muito importante neste modelo, uma vez que ela é utilizada para definir regiões de busca para as partes do corpo que se pretende encontrar. Assim sendo, um método baseado em cores dominantes é utilizado para realizar a segmentação da pessoa na fotografia. O resultado da segmentação consiste em uma imagem RGB de 8 bits sendo os canais R (vermelho) e B (azul) representados respectivamente pelas regiões da parte superior e pele do corpo. Um exemplo de segmentação realizada, utilizando-se este método, é mostrado na Figura 3.9(a).



(a) (b) (c)

Figura 3.9 – (a) Imagem segmentada. (b) Torso estimado (pontos vermelhos são os locais dos ombros estimados por parâmetros antropométricos). (c) Localização do cotovelo direito (ponto verde), pulso direito (ponto magenta) e mão direita (ponto amarelo).

Para estimar as partes superiores do corpo primeiramente é estimado um *bounding box* do torso, uma região de interesse (considerando largura, altura e posição central) na qual espera-se que o torso esteja localizado. Um exemplo do *bounding box*, referente a Figura 3.9(a), é mostrado na Figura 3.9(b). É esperado que os cantos superiores do *bounding box* sejam as posições dos ombros da pessoa, contudo, devido a imprecisão do método de segmentação, em alguns casos esses valores são estimados de forma equivocada. Desta forma, visando minimizar os erros de classificação, são utilizados parâmetros antropométricos, baseados em [7], para refinar a posição dos ombros.

Tendo em vista a detecção dos membros, inicialmente são eliminados os pixels que estão dentro do *bounding box* do torso. Assim, os pixels restantes são esperados que façam parte dos membros, contanto que estes não se encontrem em frente ou atrás do torso da pessoa. Em uma segunda etapa, tenta-se estimar a localização dos cotovelos e pulso. Para isso, é assumido que os pixels relacionados com o braço ou são classificados como torso (podendo ser as mangas da camisa) ou são classificados como pele. Para tentar isolar os braços, inicialmente exclui-se todos os pixels relacionados ao torso fora de uma região circular com raio R_e centrado na posição do ombro como ilustrado na Figura 3.9(c) (circulo azul). O valor de R_e foi definido por parâmetros antropométricos [7] e equivale a $R_e = 2,9R$, sendo R o raio da face. Espera-se que o segmento compreendido dentro do círculo de raio R_e corresponda ao braço. Neste segmento são aplicadas as seguintes operações morfológicas: *filling* e *thinning*. A extremidade da linha resultante mais distante do ombro é associada como cotovelo, conforme ilustrado no ponto verde da Figura 3.9(c). A partir deste momento, é possível estimar se a pessoa está vestindo uma blusa de mangas curtas ou longas, baseado na área dos membros restantes (antebraços).

Para encontrar os pulsos, é utilizada uma abordagem similar a empregada para encontrar os cotovelos, porém, define-se um raio $R_f = 3,15R$, também por antropometria, centrado no posição do cotovelo, o qual foi encontrado anteriormente. Após aplicar-se os

mesmos métodos morfológicos, o pulso é encontrado (ponto magenta da Figura 3.9(c)). Finalmente, para se obter a mão remove-se os pixels de pele e torso que estão dentro do círculo do cotovelo, encontrado anteriormente. Após, pega-se o maior componente de pixels de pele conectados (*blob*) que se encontra próximo ao pulso e calcula-se o centroide desta região. Este centroide é tido como a posição da mão (ponto amarelo da Figura 3.9(c)).

3.3 Contexto deste trabalho no estado-da-arte

Este trabalho utiliza uma abordagem diferente das apresentadas nas seções anteriores deste capítulo. O modelo que será apresentado provê uma união de métodos, em forma de um *pipeline*, nos quais não há intervenção do usuário nem busca exaustiva em banco de dados. Além disso, este é capaz de encontrar posturas coerentes mesmo quando na imagem as informações sobre a posturas são difíceis de serem percebidas.

A seção a seguir apresenta o modelo proposto nesta dissertação para detecção de posturas 2D baseada em fotografias.

4. MODELO PROPOSTO

Neste capítulo será apresentado o modelo desenvolvido para realizar a detecção de esqueletos 2D em imagens, partindo-se da segmentação (Seção 4.1) e após introduzindo métodos (Seção 4.2) que visam encontrar a postura global e as articulações e ossos referentes ao esqueleto da pessoa presente na fotografia.

4.1 Segmentação

O processo de segmentação deve ser capaz de automaticamente reconhecer as partes superior, inferior e pele do corpo da pessoa na imagem. O resultado dessa segmentação é denominado *blob* ou *foreground*, o qual desconsidera informações sobre o fundo (*background*) da imagem, mantendo somente a parte referente ao corpo da pessoa (*blob*). Para isso, foi escolhido o método baseado em cores dominantes desenvolvido por Jacques Jr. et al. [14], o qual foi apresentado no Capítulo 3. A ideia é, assumindo-se que a pessoa na fotografia se encontra em uma pose aproximadamente frontal, encontrar o centro e o raio da face utilizando o detector de face de Viola & Jones [29] e então estimar, baseado em dados antropométricos [28], as partes superior, inferior e de pele do corpo.

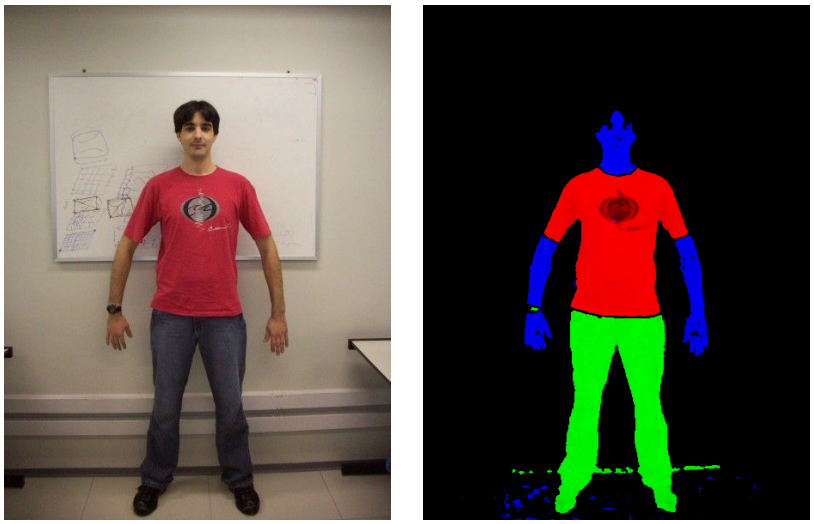
O resultado do processo de segmentação consiste em uma imagem RGB de 8 bits sendo que os canais R (vermelho), G (verde) e B (azul) representam respectivamente as regiões da parte superior, parte inferior e pele do corpo. Um exemplo de segmentação realizada, utilizando-se este método, é mostrado na Figura 4.1. Uma limitação deste método ainda é a dificuldade em trabalhar com partes heterogêneas do corpo, como as roupas ou grandes diferenças de iluminação na fotografia.

Contudo, como o método apresentado nem sempre gera segmentações ideais, neste trabalho também foram utilizadas imagens segmentadas manualmente, pois a qualidade dos resultados é mais controlada. Para isso, foi utilizado o software *Interactive Segmentation Tool*[1]. Este software consiste em uma método semiautomático que o usuário seleciona com o botão esquerdo do mouse, definindo o *foreground*, e com o botão direito o *background*, conforme mostrado na Figura 4.2. A saída do software consiste em uma imagem binária, sendo o *foreground* representado pela cor branca e o *background* pela cor preta.

Dessa forma, para realizar uma segmentação manual cujo resultado seja similar ao obtido pelo método automático [14], procedeu-se da seguinte maneira:

1. segmentou-se separadamente as partes superior (Figura 4.3(a)), inferior (Figura 4.3(b)) e de pele do corpo (Figura 4.3(c)) utilizando-se o software *Interactive Segmentation Tool*;

[1] *Interactive Segmentation Tool* - Disponível em http://kspace.cdvp.dcu.ie/public/interactive-segmentation/



(a) (b)

Figura 4.1 – Exemplo de segmentação obtida com o método proposto por Jacques Jr. et al. [14]: (a) Fotografia original; (b) Segmentação obtida, sendo os canais R (vermelho), G (verde) e B (azul) representam respectivamente as regiões da parte superior, parte inferior e pele do corpo.

2. combinou-se as imagens resultantes do primeiro passo em uma imagem RGB de 8 bits de modo que as Figuras 4.3(a), 4.3(b) e 4.3(c) constituíssem respectivamente os canais R (vermelho), G (verde) e B (azul) desta imagem (Figura 4.3(d)).

4.2 Estimativa do Esqueleto

O processo de detecção do esqueleto 2D é totalmente realizado tendo-se por base a imagem adquirida com os métodos de segmentação apresentados na Seção 4.1. Assim sendo, o modelo para detecção de poses proposto neste trabalho é composto por quatro métodos distintos. São eles:

- Método das Projeções;
- Método das Redes Neurais;
- Método de Identificação;
- Método do Ajuste.

O modelo proposto neste trabalho foi desenvolvido com o intuito de formar um *pipeline*, ou seja, uma linha de processamento em que cada um dos métodos anteriormente citados compreende uma etapa do processo. Dessa forma, as informações adquiridas durante a

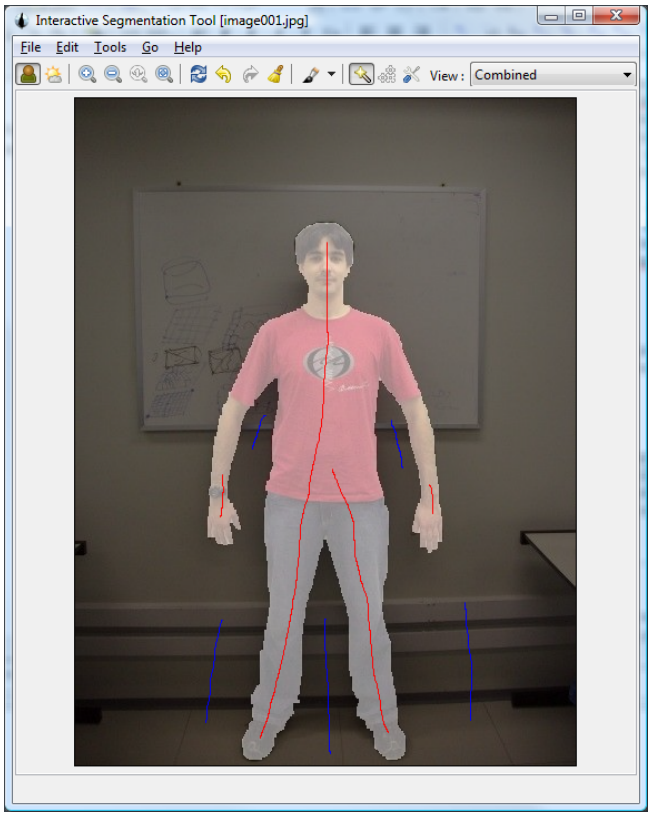
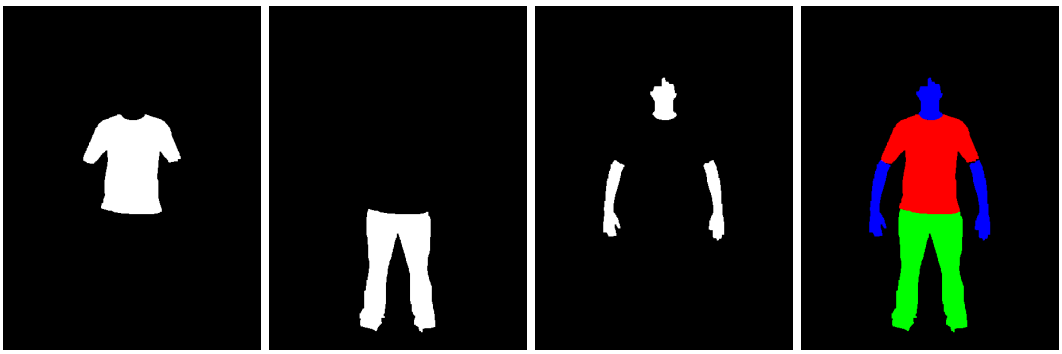


Figura 4.2 – Software *Interactive Segmentation Tool*. Linhas vermelhas são as marcações do *foreground* e as azuis o *background* feitas pelo usuário.



(a) (b) (c) (d)

Figura 4.3 – Passos realizados para segmentar manualmente uma imagem utilizando-se o software *Interactive Segmentation Tool*: (a) Segmentação da parte superior do corpo; (b) Segmentação da parte inferior do corpo; (c) Segmentação das partes de pele do corpo; (d) Imagens a, b e c combinadas em uma imagem RGB, representando respectivamente os canais R (vermelho), G (verde) e B (azul).

execução de cada método vão sendo processadas e repassadas a todas as etapas posteriores do processo. A Figura 4.4 demonstra o fluxograma do modelo proposto para a

detecção de posturas.

Nas próximas seções cada um dos métodos desenvolvidos serão detalhadamente descritos.

4.2.1 Método das Projeções

A finalidade deste método é projetar a imagem segmentada nos eixos vertical e horizontal visando obter informações sobre a disposição das partes do corpo da pessoa na fotografia. As informações adquiridas por este método não revelarão diretamente informações sobre a postura da pessoa, mas servirão de base para o Método das Redes Neurais e Método de Identificação.

O funcionamento deste método consiste em primeiramente dividir a imagem em *parte superior* e *parte inferior* aproximadamente na altura do quadril da pessoa, conforme é ilustrado na Figura 4.5(a) pela linha em cor magenta. Esta divisão é realizada estimando-se inicialmente a altura da roupa inferior da pessoa e, para isto, é utilizada a biblioteca *cvBlob* [19] (uma extensão para a biblioteca *OpenCV* [2]) que encontra *blobs* em uma imagem. Dessa forma, pega-se o maior *blob* encontrado no canal verde da imagem como sendo o *blob* correspondente à roupa inferior da pessoa e calcula-se a altura H_r deste conforme a Equação 4.1, lembrando que as coordenadas verticais da imagem crescem de cima para baixo.

$$H_r = Max_y - Min_y \tag{4.1}$$

Conhecendo H_r, a altura da linha do quadril L_q da pessoa é calculada conforme a Equação 4.2 para localizar-se 5% (estimado a partir de testes) abaixo do ponto mais alto do *blob* referente à roupa inferior da pessoa (Min_y). A Figura 4.5(b) mostra uma ampliação da área tracejada da Figura 4.5(a).

$$L_c = Min_y + (H_r \times 0,05) \tag{4.2}$$

Posteriormente, são realizadas, para cada um dos três canais da imagem, as projeções vertical (Figura 4.6(a)) e horizontal da contagem dos pixels, conforme Figura 4.6(a). Após, é realizado um ajuste de curvas utilizando-se a função *polyfit* do software MATLAB a fim de se obter as equações características das curvas, e assim possibilitando que os pontos de máximos e mínimos de cada curva sejam encontrados (Figura 4.6(b)). O grau dos polinômios varia de acordo com o tipo de curva em que o ajuste será realizado, por exemplo, para a roupa da parte superior do corpo o polinômio possui quinze graus e para as outras curvas, pele e roupa da parte inferior do corpo, cinco graus. Este valores foram obtidos mediante a realização testes com diferentes imagens, quando foi possível perceber que para representar a curva da roupa da parte superior do corpo é necessário um polinômio com o

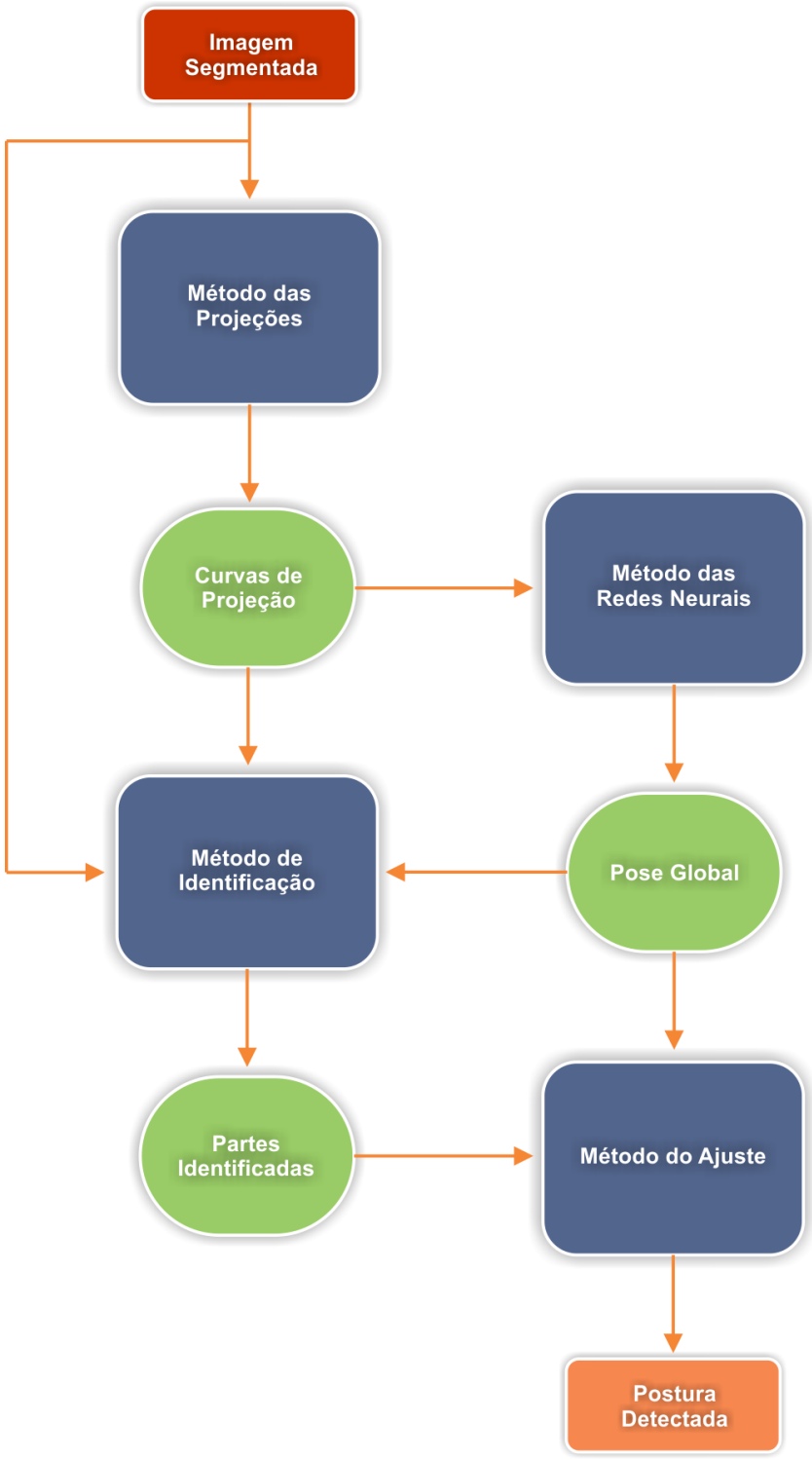
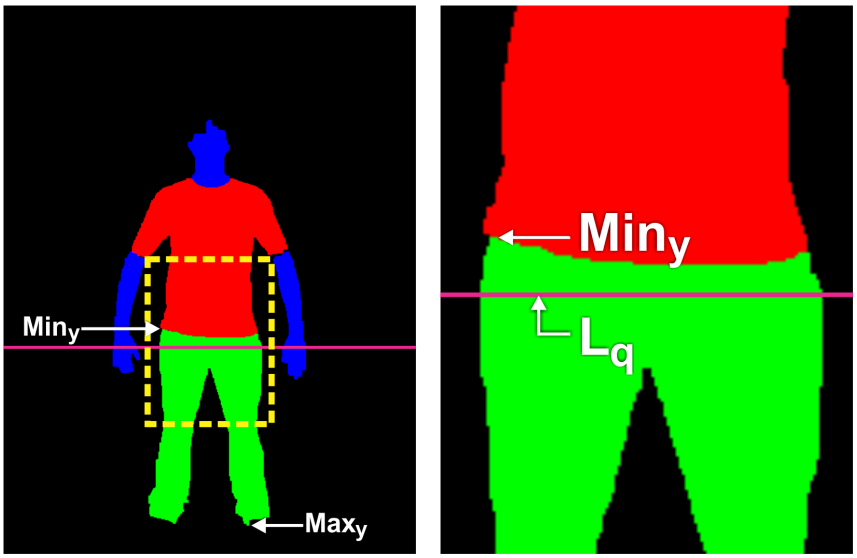


Figura 4.4 – Modelo proposto para realizar a detecção de posturas 2D a partir de uma imagem, o qual foi desenvolvido em forma de um *pipeline*. Neste modelo, as informações adquiridas durante o avanço das etapas vão sendo processadas e são repassadas às etapas posteriores do processo.



(a) (b)

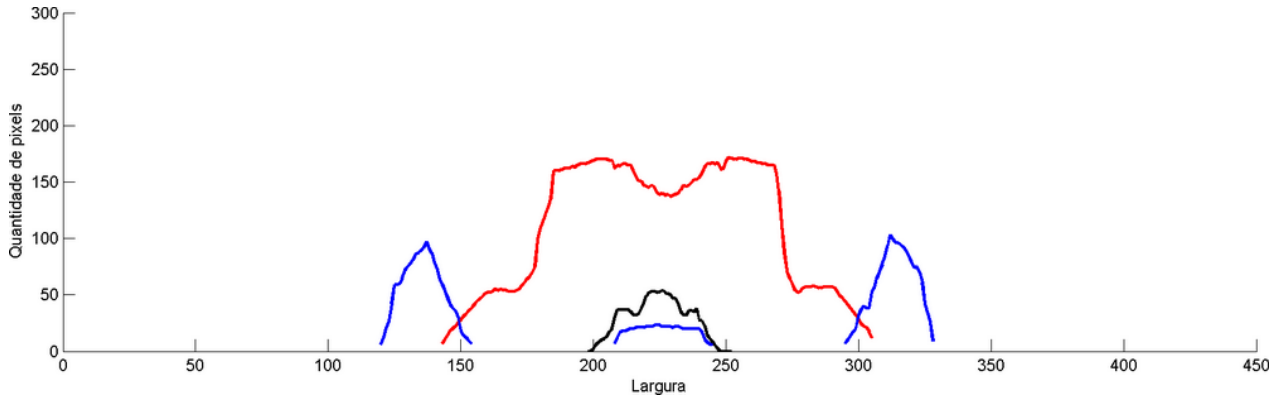
Figura 4.5 – Divisão da imagem em *parte superior* e *parte inferior* na altura do quadril da pessoa: (a) Imagem com a linha do quadril demostrada ao longo do eixo horizontal em cor magenta. (b) Ampliação da área tracejada da Figura 4.5(a).

grau três vezes maior do que para representar as curvas das outras partes.

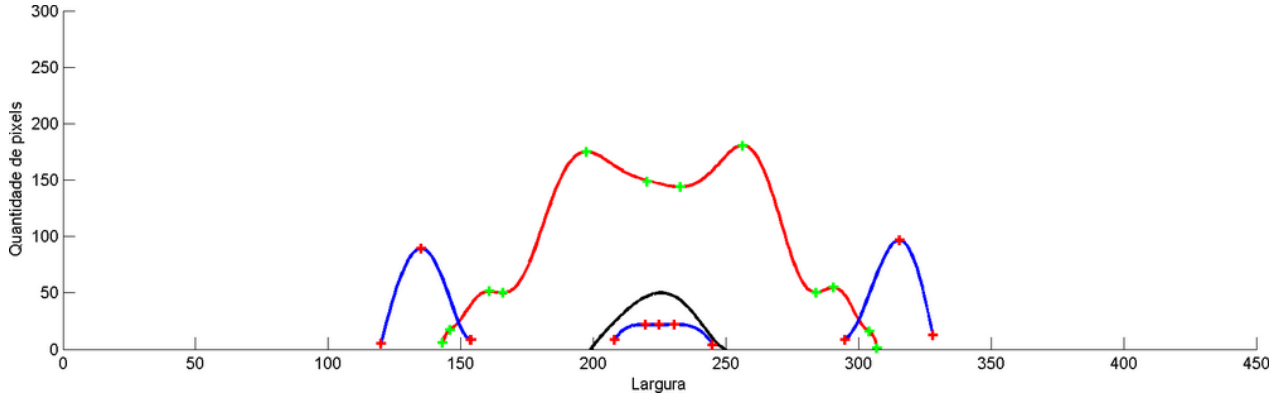
Para encontrar os pontos de inflexão das curvas, ou seja, pontos de máximos e mínimos, é utilizado o critério da primeira derivada. Dessa forma, a derivada de primeira ordem é encontrada e, em seguida, suas raízes (as quais são as abscissas dos pontos de inflexão). Depois disso, as raízes são aplicadas ao polinômio característico da curva, obtendo-se as ordenadas dos pontos de máximos e mínimos.

Visando determinar, pada cada par abscissa-ordenada encontrado, se este trata-se de um ponto de máximo ou mínimo da curva, é utilizado o critério da segunda derivada. Dessa forma, obtêm-se a derivada de segunda ordem do polinômio e aplica-se à ela o valor da abscissa de cada ponto de inflexão. Se o valor resultante for maior que zero então se trata de um ponto de mínimo, caso contrário um ponto de máximo. A Figura 4.7 mostra todas as projeções e pontos de inflexão referentes ao *blob* mostrado na Figura 4.1(a).

Apesar de as informações adquiridas até então serem de grande importância, por si só elas não fornecem dados definitivos sobre a postura da pessoa na imagem. Desta forma, os métodos das redes neurais e de identificação utilizam estas informações e tentam transformá-las em dados relevantes, os quais começam a levantar informações sobre a postura final que será encontrada para a pessoa da fotografia.



(a)



(b)

Figura 4.6 – Método das Projeções: realiza a contagem dos pixels para cada canal do *blob*. As curvas preta, vermelha e azul correspondem respectivamente à projeção da cabeça da pessoa e às projeções dos canais vermelho e azul da parte superior da imagem: (a) sem ajuste de curvas e; (b) após realizado o ajuste de curvas e identificados os pontos de inflexão de cada curva.

4.2.2 Método das Redes Neurais

Uma das características mais marcantes das redes neurais é a sua notável capacidade de lidar com padrões de complexidade significativa e que dependem de um número considerável de variáveis apresentando bons resultados, tanto para tarefas de predição quanto para tarefas de regressão [31].

Portanto, devido a qualidade dos resultados alcançados, as redes neurais artificias são amplamente utilizadas nas tarefas de predição de classes, bem como de regressão de séries de dados. Além de apresentarem bons resultados para exemplos previamente apresentados, as redes neurais têm bons resultados para exemplos que não foram apresentados, mesmo que eles extrapolem os exemplos apresentados, como provado por Yang & Kavli et al. [31]. Por todas essas características esse método de aprendizado de máquina se apresenta como uma boa alternativa a ser aplicada neste trabalho.

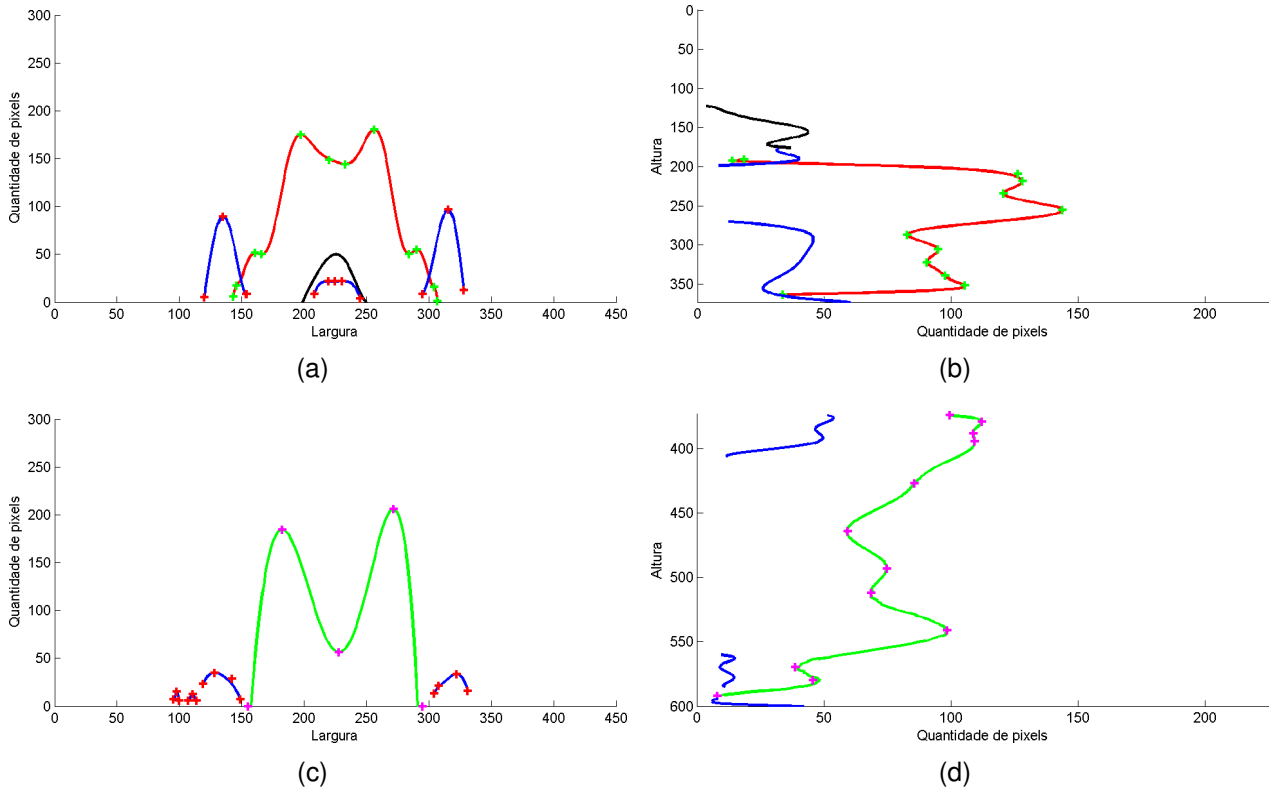


Figura 4.7 – Projeções da imagem da Figura 4.1. As curvas preta, vermelha, verde e azul correspondem respectivamente à projeção da cabeça da pessoa e às projeções dos canais vermelho, verde e azul da imagem: (a) Projeção vertical - parte superior; (b) Projeção horizontal - parte superior; (c) Projeção vertical - parte inferior; (d) Projeção horizontal - parte inferior.

Assim sendo, este modelo, baseado em redes neurais artificiais, foi desenvolvido a fim de se tentar estimar a *Pose Global 2D* da pessoa na fotografia, ou seja, as informações de alto nível sobre a posição do corpo de uma pessoa. Mais precisamente, defini-se neste trabalho que a pose global 2D é descrita como informações globais de postura, tais como os membros que estão aparentes e os tipos de roupas que estão sendo utilizados.

A base deste método consiste em redes neurais que utilizam como entrada os somatórios das contagens vertical de pixels de cada canal de cor da imagem segmentada, ou seja, as curvas de projeção vertical das partes superior e inferior geradas pelo método das projeções (sem o *fitting*, mostrado na Seção 4.2.1). Como saída, estas redes neurais produzem informações que caracterizam a *Pose Global* da pessoa.

Para produzir informações, uma rede neural necessita primeiramente passar por uma etapa de *treinamento*. Essa deve ser realizada apresentando-se à rede exemplos de entradas que a rede irá receber e a respectiva saída que deverá produzir. Quanto maior for este banco de dados utilizado para o treinamento da rede neural, melhor será seu aprendizado e, consequentemente, melhores serão os resultados produzidos posteriormente por esta

rede neural.

Dessa forma, foram reunidas 164 imagens retiradas dos bancos de dados de imagens H3D Dataset [1] e INRIA Person Dataset [5], além de algumas que foram capturadas especialmente para este trabalho. Estas imagens foram então segmentadas manualmente com o auxílio do software *Interactive Segmentation Tool*, descrito na Seção 4.1, tendo-se por objetivo evitar possíveis erros que viessem a ocorrer durante um processo de segmentação automática.

Este novo banco de dados de imagens foi então dividido aleatoriamente em duas partes: 85% para constituirem o *banco de dados de treinamento* (conhecido também como *banco de conhecimento* ou *banco de aprendizado*) das redes neurais (totalizando 139 imagens); e 15% para constituirem o conjunto de teste da RNA (25 imagens).

Visando aumentar ainda mais o banco de dados de treino, cada uma das 139 imagens selecionadas para serem utilizadas na etapa de aprendizagem das redes neurais foi escalada dez vezes. O que resultou em um banco de dados de treinamento composto por 1390 imagens.

Posteriormente, todas 1390 imagens foram "espelhadas" (invertidas horizontalmente). Este processo possibilitou que o banco de conhecimento ficasse mais homogêneo, contendo, por exemplo, o mesmo número de exemplares de pessoas com o braço direito oculto e de exemplares com o braço esquerdo oculto. Além disso, este processo dobrou o número de imagens do banco de aprendizado, totalizando 2780 imagens.

Para que o treinamento da rede neural pudesse ser realizado, todas as imagens do banco de conhecimento foram manualmente classificadas segundo suas respectivas *Poses Globais*, as quais são representadas pelos seguintes critérios:

1. tipo de camisa (sem camisa, manga cavada, manga curta ou manga comprida);
2. braço direito aparente (sim ou não);
3. braço esquerdo aparente (sim ou não);
4. mão direita aparente (sim ou não);
5. mão esquerda aparente (sim ou não);
6. tipo de calças (sunga de banho, calção/saia curta, bermuda/saia longa/calça capri ou calça comprida);
7. perna direita aparente (sim ou não);
8. perna esquerda aparente (sim ou não).

Cada um destes critérios é tratado como um problema individual, ou seja, cada critério possui sua própria rede neural. Isto porque a separação das características a serem preditas facilita o aprendizado da rede neural, aumentando a acurácia dos resultados encontrados. Além disso, para cada critério, tem-se o número de neurônios na camada de saída correspondente a quantidade possível de dados de saída. Por exemplo, para o critério "tipo de camiseta" têm-se quatro neurônios na camada de saída e para o critério "braço direito aparente" têm-se dois neurônios na camada de saída.

O número de neurônios presente na camada de saída é a única diferença entre as configurações de todas oito redes neurais (uma para cada critério da pose global). Para todas redes foram utilizadas as seguintes configurações:

- Rede Neural do tipo MLP (*Multi Layer Perceptron*);
- Algoritmo de aprendizado *Levenberg-Marquardt*;
- *Mean Squared Error* (MSE) [12];
- Função de ativação tangente hiperbólica para os neurônios da camada escondida.
- Função de ativação linear para os neurônios da camada de saída;

Foi escolhido o algoritmo de aprendizado de *Levenberg-Marquardt* pois este tende a ser um algoritmo de rápida convergência, exigindo poucas iterações para a obtenção de bons resultados [12]. Já a métrica de regressão MSE, que significa Erro Médio Quadrático (do inglês *Mean Square Error*), foi escolhida por ser bastante utilizada na literatura e, mediante testes, mostrou-se adequada para ser utilizada neste trabalho. Para os neurônios da camada de saída foi utilizada uma função de ativação linear pois sua imagem suporta melhor os valores esperados para o intervalo saída.

Como não existe uma equação para calcular o número de neurônios da camada escondida, foram realizados diversos testes de desempenho, chegando-se ao número de dez neurônios. Com esta quantidade de neurônios na camada escondida foram obtidos os melhores resultados, taxa de aprendizagem e tempo de aprendizagem (resultados estes que serão apresentados no Capítulo 5). O número de dez neurônios para a camada escondida mostrou-se a melhor opção para todas as oito redes. A Figura 4.8 mostra um fluxograma ilustrando os passos realizados para proceder-se com o treinamento das redes neurais, conforme explicado nesta seção.

Importante ainda salientar que a classificação feita nas imagens do banco de treinamento da rede neural só precisa ser feita uma vez. Quando o banco de dados está treinado, as detecções podem ser realizadas.

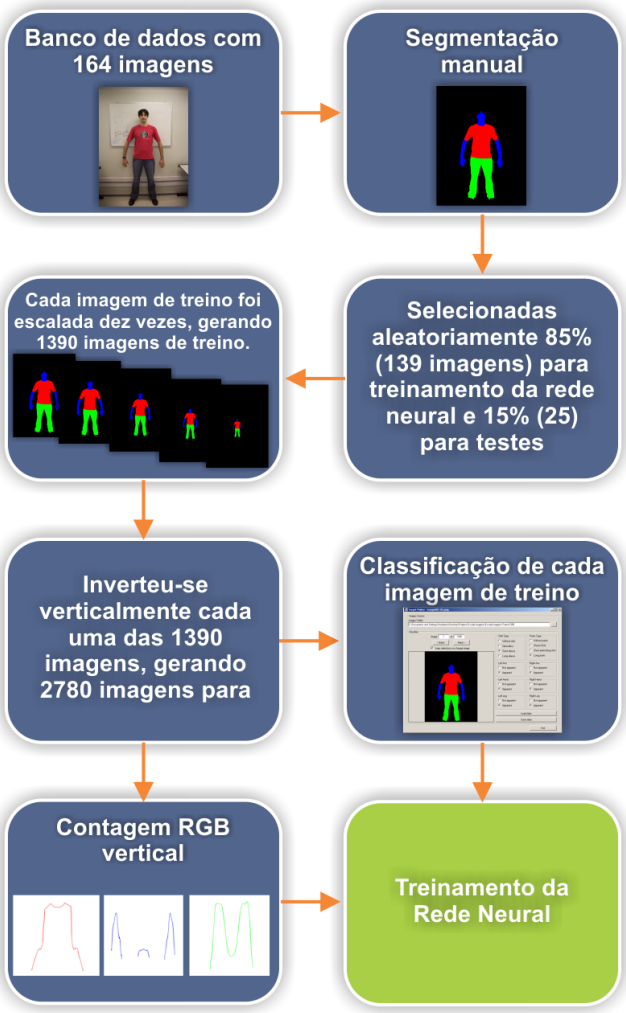


Figura 4.8 – Fluxograma explicando o modelo proposto, o qual faz uso de redes neurais artificiais.

4.2.3 Método de Identificação

Este método tem o objetivo de tentar identificar as partes do corpo presentes na imagem. Utilizando as informações adquiridas durante o processamento dos métodos das projeções e das redes neurais, a imagem resultante do processo de segmentação e de regras definidas baseadas na semântica do corpo humano e antropometria, este método gera como saída uma imagem semelhante à imagem do processo de segmentação, porém com as partes do corpo identificadas.

A identificação das partes do corpo na imagem de saída é realizada por meio de cores. Conforme as partes do corpo da pessoa vão sendo identificadas, estas são pintadas na nova imagem com cores distintas da seguinte forma:

- Branco - Mãos e/ou pele dos braços (quando a pessoa não estiver usando mangas compridas);

- Amarelo - Perna direita;
- Laranja - Pele da perna direita (quando a pessoa não estiver usando calças compridas);
- Magenta - Perna esquerda;
- Roxo - Pele da perna direita (quando a pessoa não estiver usando calças compridas).

Este método é totalmente procedural e utiliza-se de regras criadas a partir de observações realizadas principalmente nas curvas de projeções, e fazendo uso dos pontos de máximos, mínimos e áreas providas pelo método das projeções. Os dados sobre a postura global da pessoa, os quais são resultantes do método das redes neurais, são utilizados para “saber o que estamos procurando na imagem”, ou seja, não se tentará identificar um braço direito quando a pose global informar que este braço não está aparente.

Grande parte das regras deste método, criadas para a identificação das partes do corpo, utilizam a área das curvas de projeção. Por exemplo, o primeiro passo realizado neste método é invalidar as curvas de pele que possuam área inferior a uma certa porcentagem da área da curva da cabeça da pessoa (esta porcentagem varia de acordo com o tipo de roupa informada pela pose global, ou seja, manga comprida, curta, cavada ou sem camisa). Este procedimento evita que curvas decorrentes de erros de segmentação possam atrapalhar o resto do processo.

Basicamente, este método tenta encontrar curvas que possam representar as informações da pose global. Para isto, são utilizadas heurísticas tais como: se a pose global informa “manga comprida” e “ambas mãos aparentes”, então tenta-se encontrar as mãos procurando-se duas curvas de pele com áreas parecidas, primeiramente na parte superior, posteriormente na parte inferior e, ainda, uma curva na parte superior e outra na parte inferior do corpo. Se não forem encontradas curvas que caracterizem duas mãos separadas, conclui-se que as mãos devem estar juntas, dessa forma, procura-se uma curva que contenha área plausível de representar essas duas mão.

Identificar as pernas é uma tarefa relativamente menos complicada do que estimar a posição dos braços e mãos. Isto porque as pernas, por servirem de apoio para o corpo e por possuírem grau de liberdade mais limitado do que os braços, dificilmente não se encontrarão uma ao lado da outra. Dessa forma, visando encontrar a linha que separa as pernas direita e esquerda, foram construídas regras que preveem três situações diferentes quando a pose global informa que ambas pernas estão aparentes, são elas:

1. As pernas estão juntas;
2. As pernas estão abertas;

3. As pernas foram separadas nas curvas de projeção;

Assim sendo, as situações 1 e 2 são tratadas da mesma maneira, pois haverá somente uma curva de roupa na projeção vertical inferior que pode realmente caracterizar as pernas (não necessariamente existe somente uma curva na projeção vertical inferior, porém somente uma tem reais condições de ser uma curva que caracteriza as pernas). Dessa forma, calcula-se o ponto médio entre o ponto inicial e o final desta curva e pega-se o ponto de inflexão (máximo ou mínimo) com menor distância euclidiana do ponto médio. É então traçada uma linha que passa por este ponto e pelo ponto central da cabeça, a qual foi denominada de *Linha de Orientação do Torso*. Assim sendo, a parte da curva que está compreendida do ponto inicial até a *Linha de Orientação do Torso* representa a perna direita da pessoa, já a parte compreendida entre a *Linha de Orientação do Torso* e o ponto final da curva, representa a perna esquerda da pessoa. A Figura 4.9 ilustra a técnica utilizada.

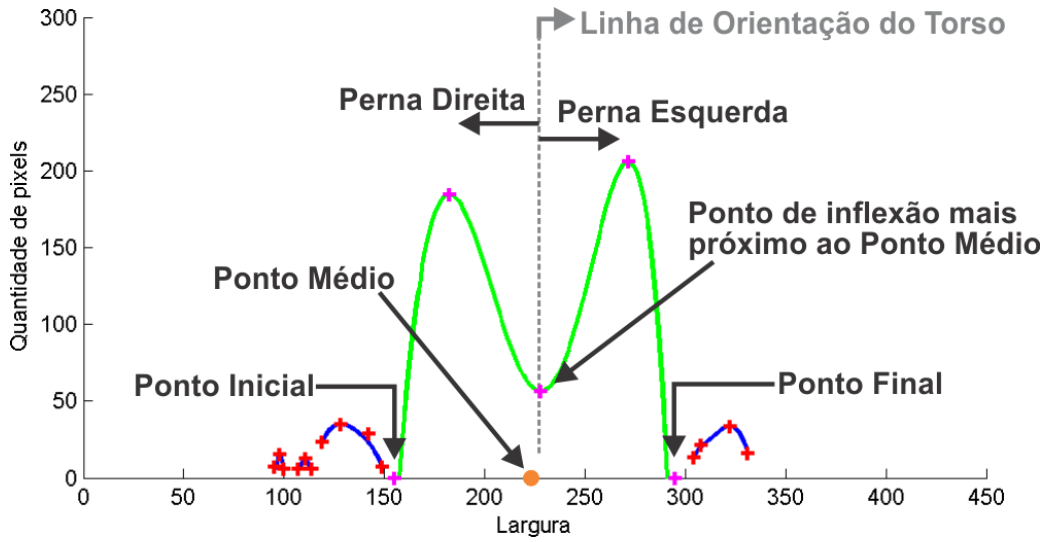


Figura 4.9 – Ilustração da técnica utilizada para encontrar as áreas pertencentes às pernas direita e esquerda quando há somente uma curva de roupa na projeção vertical inferior que pode realmente caracterizar as pernas.

Para a situação 3, quando são encontradas duas curvas para representar as pernas, simplesmente considera-se a curva mais a esquerda do gráfico como sendo referente a perna direita da pessoa e a curva mais a direta do gráfico como sendo referente a perda esquerda da pessoa. Neste caso, para encontrar a *Linha de Orientação do Torso* calcula-se o ponto médio entre o ponto final da curva mais a esquerda e o ponto inicial da curva mais a direita, então, se traça a *Linha de Orientação do Torso* passando por este ponto e pelo ponto central da cabeça. A Figura 4.10 ilustra o procedimento para a situação 3.

Este método consiste basicamente em regras construídas a partir de heurística, baseadas nas informações provindas dos métodos das Projeções e das Redes Neurais, mas

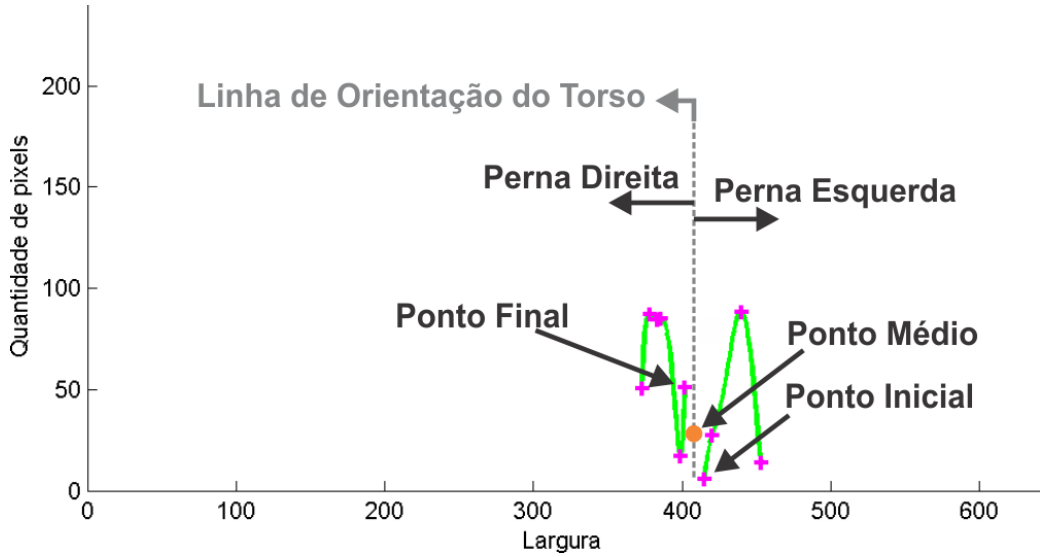


Figura 4.10 – Ilustração da técnica utilizada para encontrar as áreas pertencentes às pernas direita e esquerda quando são encontradas duas curvas de roupa, na projeção vertical inferior, para representar as pernas.

também utiliza a imagem segmentada para prover o resultado das partes do corpo devidamente identificadas. Um exemplo do resultado obtido com a aplicação do *Método de Identificação* é apresentado na Figura 4.11.

(a) (b)

Figura 4.11 – *Método de Identificação*: (a) Imagem segmentada manualmente; (b) Imagem resultante do *Método de Identificação*, a qual foi gerada automaticamente a partir de regras que utilizam como entrada as informações provenientes dos métodos das *Projeções* e das *Redes Neurais*.

4.2.4 Método do Ajuste

O funcionamento deste método consiste em, a partir das informações obtidas pelos métodos anteriores, gerar por antropometria [28] um possível esqueleto, provido de ossos e articulações, para o dado *blob* da pessoa na fotografia. Cada osso deste esqueleto é então rotacionado em torno de sua articulação de mais alta hierarquia e, para cada ângulo de rotação, é medida a probabilidade deste ser o ângulo mais adequado. Essa probabilidade é calculada de modo distinto para cada osso do corpo avaliando-se, para cada ângulo, os pixels que envolvem o osso e a posição deste em relação ao *blob*.

Para traçar o esqueleto a partir de antropometria fez-se necessário estimar a altura da pessoa. Esta altura é estimada de duas formas diferentes: por antropometria em relação ao raio da face (seguindo as proporções apresentadas na Tabela 2.1); e por estimativa, a partir de dados adquiridos principalmente pelo método de identificação (apresentado na Seção 4.2.3).

A estimativa da altura da pessoa é realizada a partir dos seguintes dados conhecidos:

- P_{cf} - ponto central da face (encontrado anteriormente por meio da utilização do detector de faces de Viola & Jones [29]);
- θ - ângulo de orientação de $\vec{T}$ (*Linha de Orientação do Torso* fornecida pelo método de identificação);
- P_{lp} - último ponto da imagem que foi identificado como perna (direita ou esquerda) pelo método de identificação;
- d - distância entre P_{cf} e P_{lp};
- α - ângulo de orientação da linha $\vec{E}$ que passa pelos pontos P_{cf} e P_{lp};
- β - ângulo localizado entre $\vec{T}$ e $\vec{E}$, o qual é dado por $\beta = \theta - \alpha$.

Assim sendo, a altura estimada da pessoa h_p é calculada utilizando-se a Equação 4.3. Uma ilustração dos cálculos realizados para encontrar h_p é apresentada na Figura 4.12.

$$h_p = d \times cos(\beta) \tag{4.3}$$

Uma vez calculadas as alturas, por antropometria baseada no raio da face e h_p, deve-se então escolher qual é a mais adequada para ser utilizada. A altura encontrada para h_p pode ser incoerente por diversas razões, tais como: falhas na segmentação; foto com perspectiva; posturas não eretas (tais como uma pessoa sentada ou agachada); entre outras. Assim sendo, quando detectada uma dessas situações, opta-se por utilizar a altura baseada no raio da face. No caso de não ser detectada nenhuma incoerência, utiliza-se a altura com o maior valor encontrado.

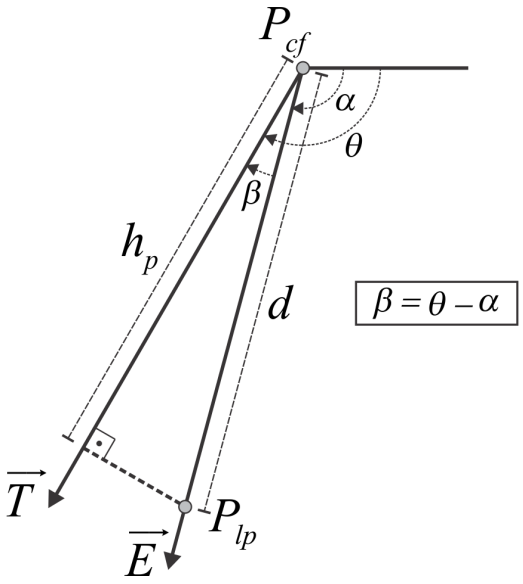


Figura 4.12 – Método utilizado para estimar a altura da pessoa na fotografia. Esta altura é utilizada como base para os cálculos dos tamanhos dos ossos por antropometria.

Tendo-se a altura da pessoa, são então calculados os tamanhos dos ossos da pessoa. Assim sendo, dá-se início ao processo que tenta encontrar a melhor posição de cada osso no *blob* da pessoa na imagem resultante do método de identificação.

Os ossos do pescoço, tronco e quadris não necessitam de técnicas especiais para serem posicionados corretamente. Este fato se dá porque o ângulo do tronco já foi estimado anteriormente no método de identificação, assim sendo, atribui-se ao pescoço o mesmo ângulo do abdomem da pessoa, já os quadris são posicionados perpendicularmente ao tronco. Além disso, como as mãos possuem grau de rotação limitado, para posicioná-las também não se faz necessária a criação de uma técnica específica, por isso, neste trabalho, as mão são posicionadas utilizando-se o ângulo obtido para o antebraço.

Entretanto, para o posicionamento dos outros ossos foi necessário criar métodos específicos. Dessa forma, o método do ajuste é composto por quatro módulos, são eles:

1. Ajusta Clavícula;
2. Ajusta Braço;
3. Ajusta Antebraço;
4. Ajusta Perna.

A seguir cada um dos módulos acima citados serão explicados.

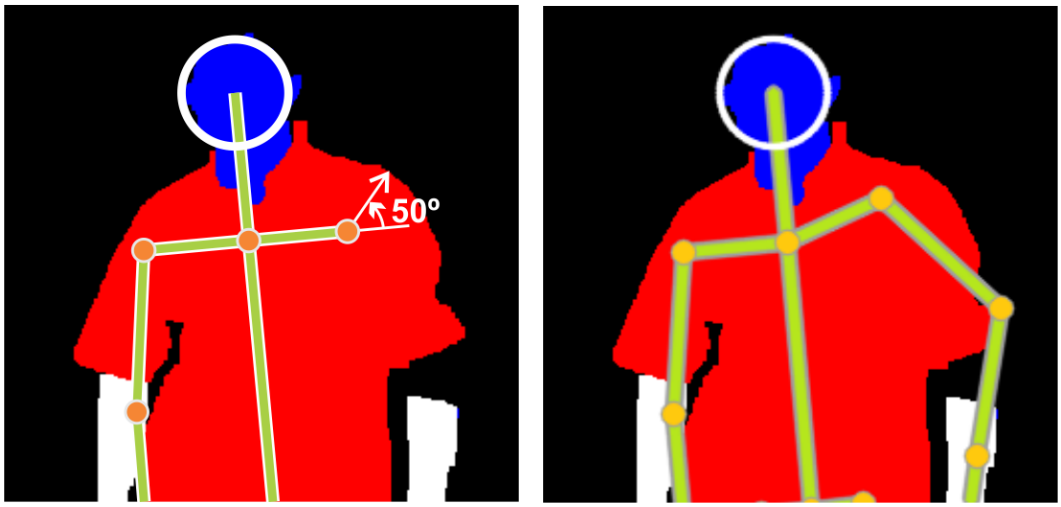
Ajusta Clavícula

Este módulo tem por objetivo posicionar as clavículas direita e esquerda da pessoa.

Primeiramente as clavículas são posicionadas a 90º em relação ao pescoço e então verifica-se se pelo menos 30% dos pontos que envolvem o ponto final destes ossos se encontram em uma região fora do *blob* (cor preta). Caso isso se verifique, procura-se perpendicularmente ao osso em questão um ponto dentro do *blob* da pessoa (considerando que a pessoa pode estar com o braço levantado e, por isso, o ponto final da clavícula estar fora do *blob*). Por antropometria o osso da clavícula pode possuir uma inclinação de 60º a 90º em relação ao pescoço, então, caso o *blob* seja encontrado acima do ponto final do osso, inclina-se a clavícula proporcionalmente a distância de seu ponto final até o *blob*, de modo que a clavícula fique localizada dentro deste.

Se nenhum *blob* for encontrado acima do ponto final da clavícula, provavelmente a pessoa está em perspectiva na fotografia. Assim sendo, o osso tem seu tamanho reduzido de 10% em 10% até que 70% dos pixels que envolvem o ponto final da clavícula pertençam ao *blob* da pessoa.

No caso de pelo menos 70% dos pixels que envolvem o ponto final da clavícula estarem dentro do *blob* da pessoa, procura-se identificar se esta pessoa está com os ombros inclinados para cima. Para isso, procura-se a um ângulo de 50º em relação ao osso, a partir do ponto final da clavícula, o final do *blob*. Assim sendo, o ângulo da clavícula é alterado proporcionalmente à distância encontrada, sempre respeitando as restrições antropométricas. A Figura 4.13(a) ilustra a busca pela borda do *blob* para calcular-se a inclinação da clavícula.



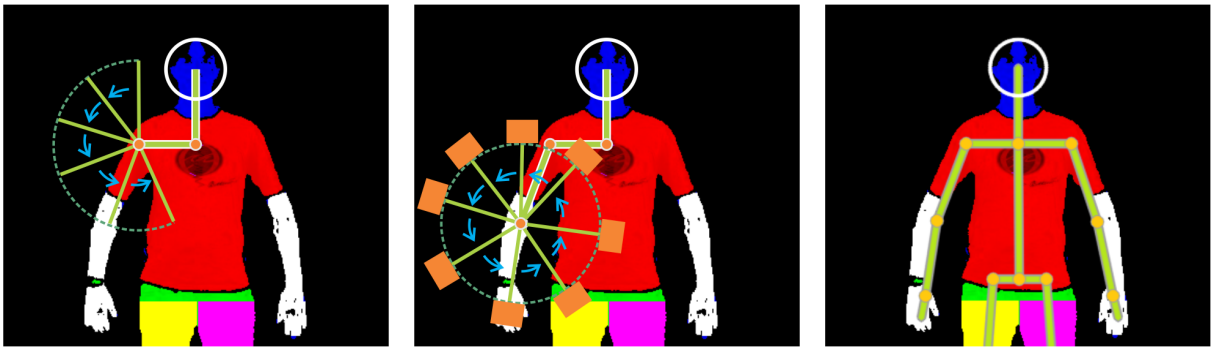
(a) (b)

Figura 4.13 – Método do ajuste para a clavícula: (a) busca pela borda do *blob* para calcular-se a inclinação da clavícula; (b) Clavícula ajustada.

Ajusta Braço

Para ajustar o braço a estratégia é diferente da utilizada para a clavícula. Segundo as restrições antropométricas, o braço pode girar livremente em 210 graus. Assim sendo, uma vez que a posição da clavícula é conhecida, este módulo posiciona o braço voltado para cima, 90 graus em relação à clavícula, e então rotaciona o braço até o ângulo de -120 graus, sentido horário para o braço direito e anti-horário para o braço esquerdo (sempre do ponto de vista da pessoa na fotografia), calculando para todo o comprimento do braço a distância perpendicular do osso até o primeiro pixel de cor preta. A Figura 4.14(a) ilustra esta ação.

A média dos valores encontrados para as distâncias até a borda externa do braço vão sendo armazenados em um vetor. Após testar todo intervalo de 210 graus (de 90 à -120 graus), considera-se o ângulo mais adequado para representar o braço aquele que produziu uma média de distâncias mais próximas à medida esperada (conforme Tabela 2.1) para a largura do braço.



(a) (b) (c)

Figura 4.14 – *Método do Ajuste* para os braços: (a) Ajuste do braço: rotaciona-se o braço em um intervalo de 210 graus procurando-se a melhor posição; (b) Ajuste do antebraço: a busca pela melhor posição estende-se por 360 graus. Além disso, os retângulos em laranja (os quais possuem comprimento e largura dados por antropometria) ilustram a projeção da mão na extremidade do antebraço para verificar a possibilidade de, naquela posição, estar localizada a mão; (c) Braços posicionados automaticamente.

Ajusta Antebraço

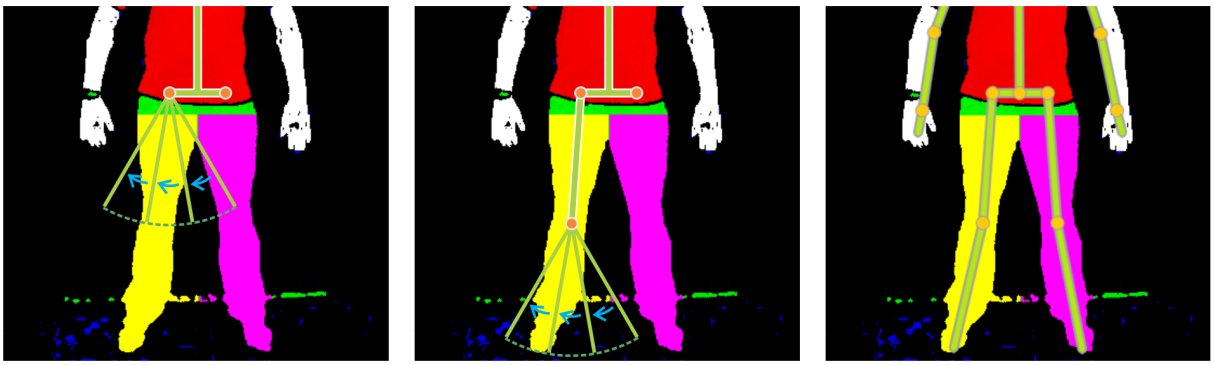
O procedimento para posicionar o antebraço é equivalente ao utilizado no ajuste do braço, contudo, a busca pela posição é estendida a 360 graus. Além disso, se a postura global informar que a mão está aparente, antes de se procurar pela média das distâncias mais próxima da largura esperada para o antebraço, projeta-se a mão na extremidade do antebraço e verifica-se a possibilidade de, naquela posição, estar localizada a mão. A

Figura 4.14(b) ilustra a busca pela mão através da utilização de retângulos em laranja, os quais possuem comprimento e largura estimados por antropometria.

Ajusta Perna

A técnica criada para posicionar as pernas também consiste em rotacionar o osso em busca da melhor posição, contudo, as pernas direita e esquerda já foram devidamente identificadas e separadas uma da outra pelo método de identificação. Dessa forma, sendo θ o angulo de orientação do torso (calculado no método de identificação), rotaciona-se o osso da perna partindo-se do ângulo $\theta - 30°$ e parando em $\theta + 30°$. Para cada grau que a perna é rotacionada calcula-se, para ambos os lados do osso, a distância perpendicular deste até o primeiro pixel de cor preta, ou seja, fora do *blob* da perna da pessoa, então, calcula-se as medianas destas distâncias. Para encontrar o ângulo que melhor ajusta o osso no centro da perna, verifica-se qual ângulo entre $\theta - 30°$ e $\theta + 30°$ produziu medianas parecidas em ambos os lados da perna (excluindo-se os casos em que as medianas resultaram o valor zero, pois significa que o osso estava fora do *blob* da perna).

O processo de ajuste da perna é exatamente o mesmo para a parte coxa e para a panturrilha. Ilustrações do processo são apresentadas nas Figuras 4.15(a) e 4.15(b).



(a) (b) (c)

Figura 4.15 – *Método do Ajuste* para as pernas: (a) Ajuste da coxa: rotaciona-se o osso em um intervalo de 60 graus procurando-se a melhor posição; (b) Ajuste da panturrilha: a busca pela melhor posição estende-se pelos mesmos 60 graus; (c) Pernas posicionadas automaticamente.

Quando este modelo detecta que a perna está mais comprida do que deveria, seja por problemas de perspectiva ou erro ao estimar a altura da pessoa, ele reduz o tamanho dos ossos da perna (coxa e panturrilha) em 25% e refaz todo o processo tentando ajustar a perna. No Capítulo 5 serão apresentados exemplos de quando esta situação ocorre.

No próximo capítulo são apresentados alguns resultados obtidos com a aplicação do modelo apresentado nesta seção.

5. RESULTADOS

Neste capítulo são apresentados os resultados obtidos com o modelo proposto para realizar a detecção de esqueleto 2D em imagens, o qual é executado em forma de um *pipeline* composto por quatro métodos distintos, e sem intervenção do usuário. Para melhor visualização do *pipeline* proposto, este capítulo apresentará lado a lado os resultados obtidos em cada um dos métodos que compõem o modelo. A Seção 5.1 apresenta alguns resultados obtidos a partir de imagens segmentadas manualmente, já a Seção 5.2 apresenta resultados gerados utilizando-se imagens segmentadas automaticamente.

5.1 Resultados com Imagens Segmentadas Manualmente

A seguir serão expostos resultados obtidos a partir da execução do *pipeline* proposto no Capítulo 4 para imagens segmentadas manualmente.

A Figura 5.1 mostra um caso que todos os passos do *pipeline* funcionaram como esperado. O método das redes neurais conseguiu encontrar uma postura global a qual confere as características da foto, o método de identificação conseguiu identificar as curvas da mão e das pernas e o método do ajuste encontrou corretamente a posição dos ossos do corpo.

Um caso em que o método das redes neurais não acertou por completo a postura global da pessoa é mostrado na Figura 5.2. Para esta imagem, apesar do erro na detecção da postura global, o método de identificação conseguiu identificar com êxito os membros do corpo. Já o método do ajuste tentou erradamente, devido a detecção da pose global, encontrar a mão esquerda da pessoa, que na verdade não estava aparente.

Na Figura 5.3 o método das redes neurais falhou em dois dos oito critério (mãos direita e esquerda aparentes), contudo, isto não impediu que o método de identificação obtivesse sucesso ao identificar as partes do corpo. O método do ajuste identificou a necessidade de curvar a clavícula esquerda em 10 graus (conforme explicado na Seção 4.2.4 deste trabalho). Além disso, o método do ajuste foi afetado pelo erro na postura global e colocou mãos no esqueleto estimado.

A Figura 5.4 é um exemplo da importância de se calcular corretamente o ângulo de orientação do torso, caso contrário não seria possível separar as duas pernas e o torso ficaria fora do *blob* da pessoa. Contudo, para esta fotografia o fato da pessoa estar em perspectiva e ainda com a cabeça inclinada para a frente de seu corpo implicou no deslocamento de todo o esqueleto e, dessa forma, impossibilitando que o método do ajuste obtivesse sucesso na tarefa de posicionar os braços corretamente, mesmo tendo inclinado em 20 graus a clavícula esquerda.

O erro ao estimar a pose global da pessoa da Figura 5.5 (manga comprida ao invés

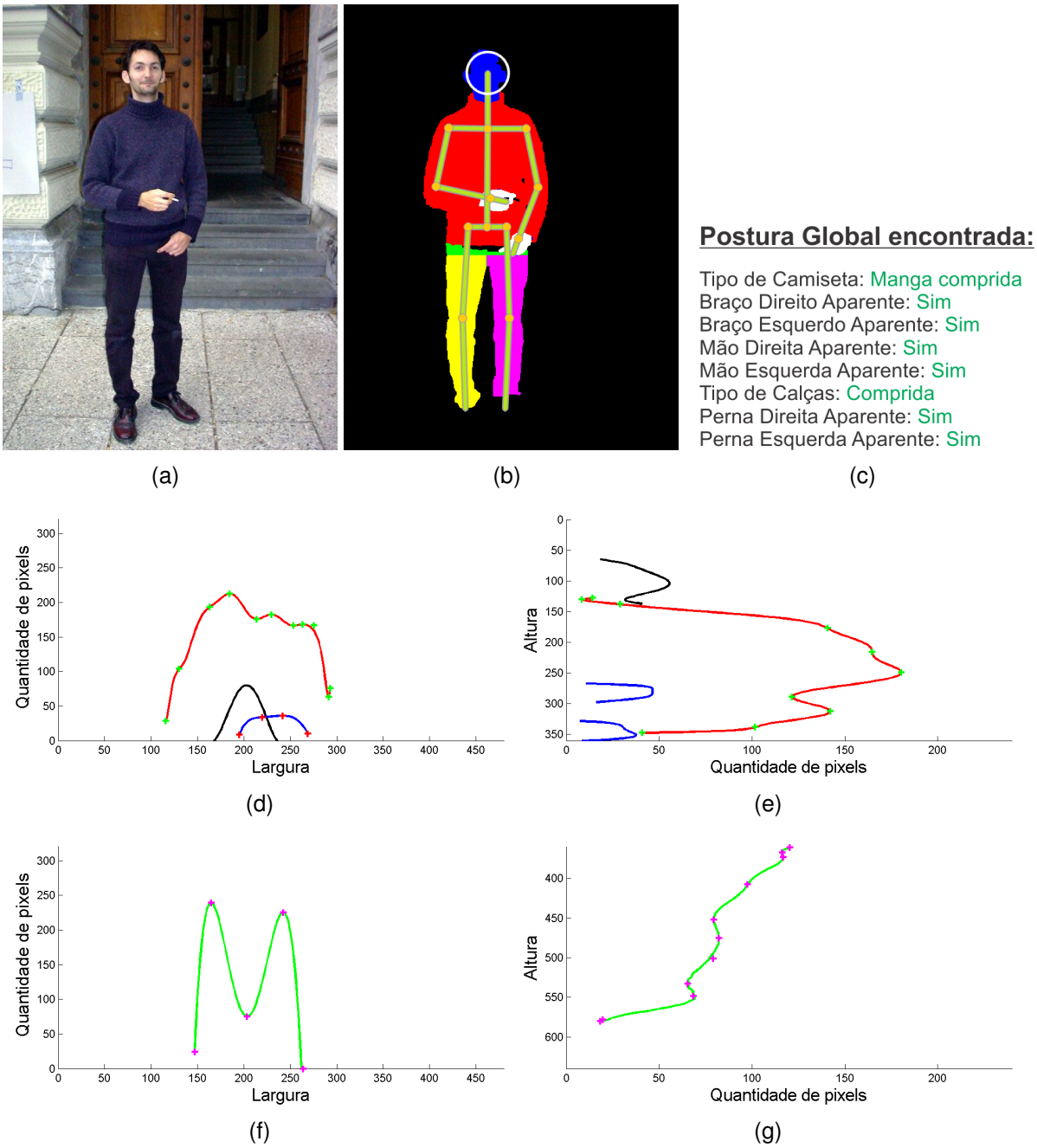


Figura 5.1 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste; (c) Pose global gerada pelo método redes neurais, o qual conseguiu encontrar uma postura que confere com as características da foto; (d) Projeção vertical da parte superior do corpo; (e) Projeção horizontal da parte superior do corpo (f) Projeção vertical da parte inferior do corpo; (g)Projeção horizontal da parte inferior do corpo.

de curta) implicou na falha do método de ajuste ao tentar posicionar o braço esquerdo e também na procura indevida por mãos que, na verdade, não estão aparentes.

As Figuras 5.6 e 5.7 são mais dois casos em que todos os métodos do modelo funcio-

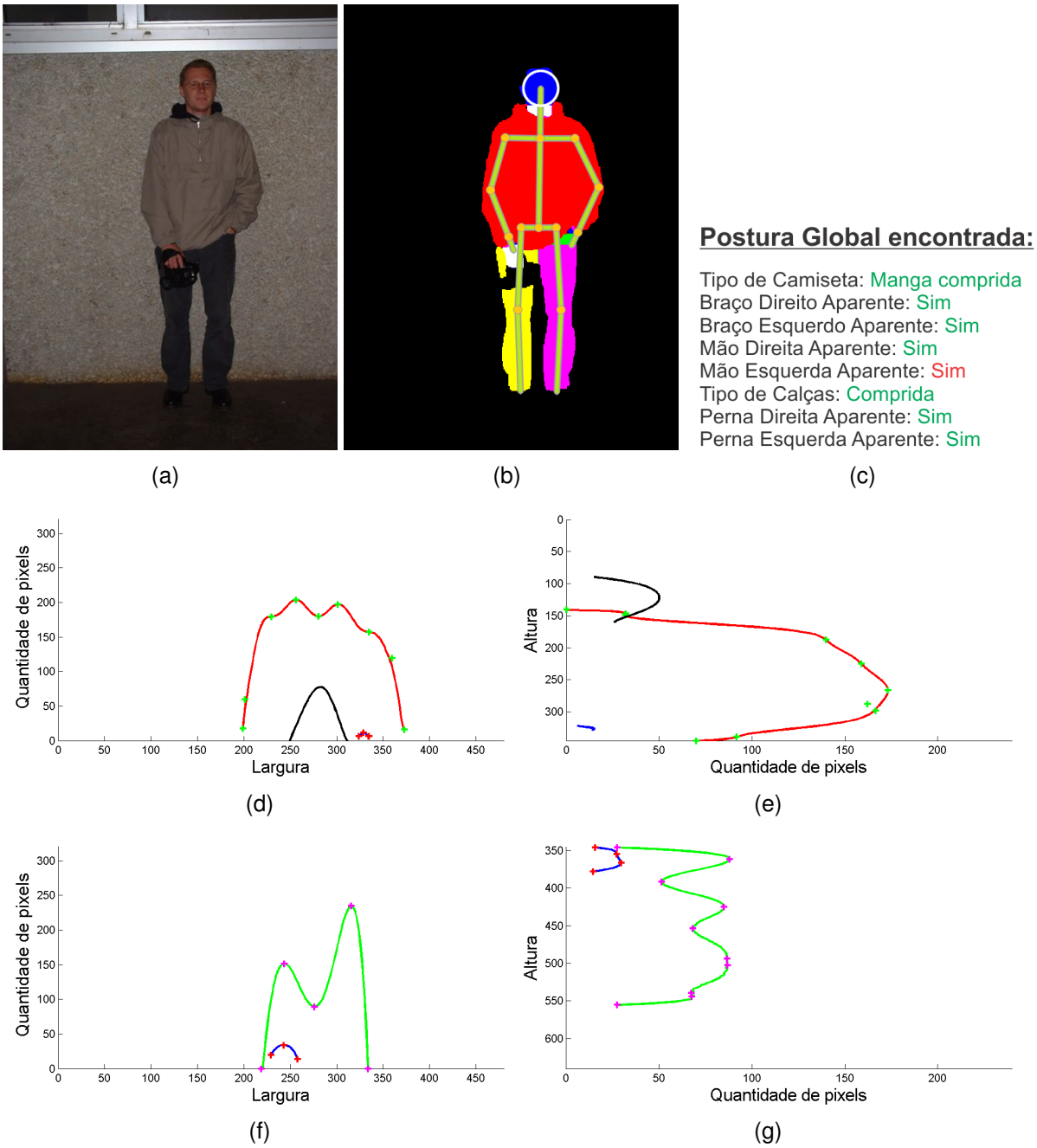


Figura 5.2 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste. Um erro na detecção da pose global implicou na tentativa de se encontrar a mão esquerda, que na verdade não estava aparente; (c) Pose global gerada pelo método das redes neurais. Em vermelho é mostrado a falha no processo de estimativa da pose global da pessoa; (d) Projeção vertical da parte superior do corpo; (e) Projeção horizontal da parte superior do corpo (f) Projeção vertical da parte inferior do corpo; (g)Projeção horizontal da parte inferior do corpo.

naram conforme esperado.

Conforme se pôde perceber durante a apresentação dos resultados nesta seção, o mé-

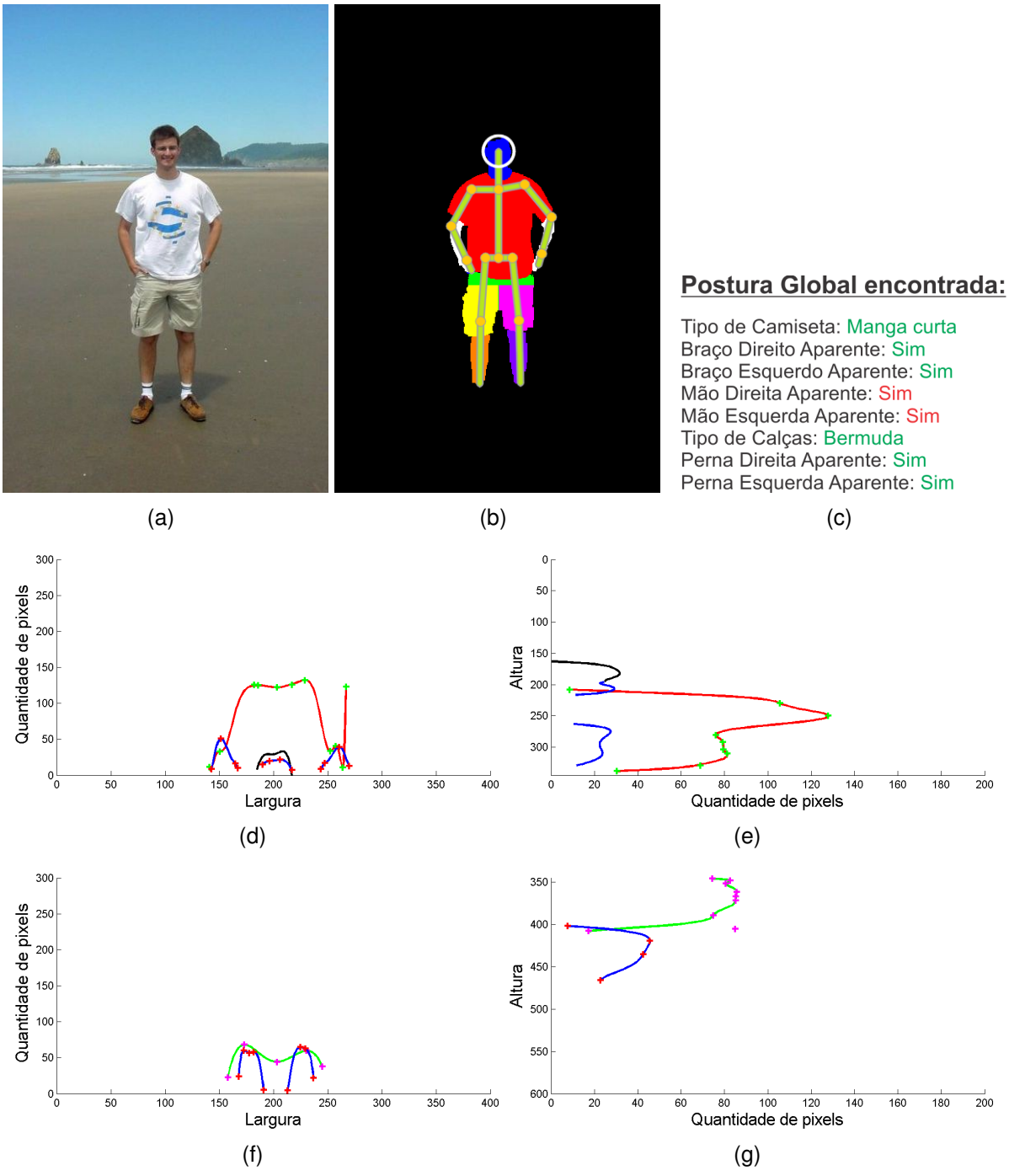


Figura 5.3 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste. A Clavícula esquerda foi inclinada em 10 graus para melhor se ajustar ao *blob* da pessoa. Além disso, o método do ajuste foi afetado pela postura global parcialmente incorreta, e, dessa forma, colocando indevidamente mãos no esqueleto; (c) Pose global gerada pelo método redes neurais. As classificações tidas como incorretas estão em vermelho; (d) Projeção vertical superior; (e) Projeção horizontal superior; (f) Projeção vertical inferior; (g)Projeção horizontal inferior.

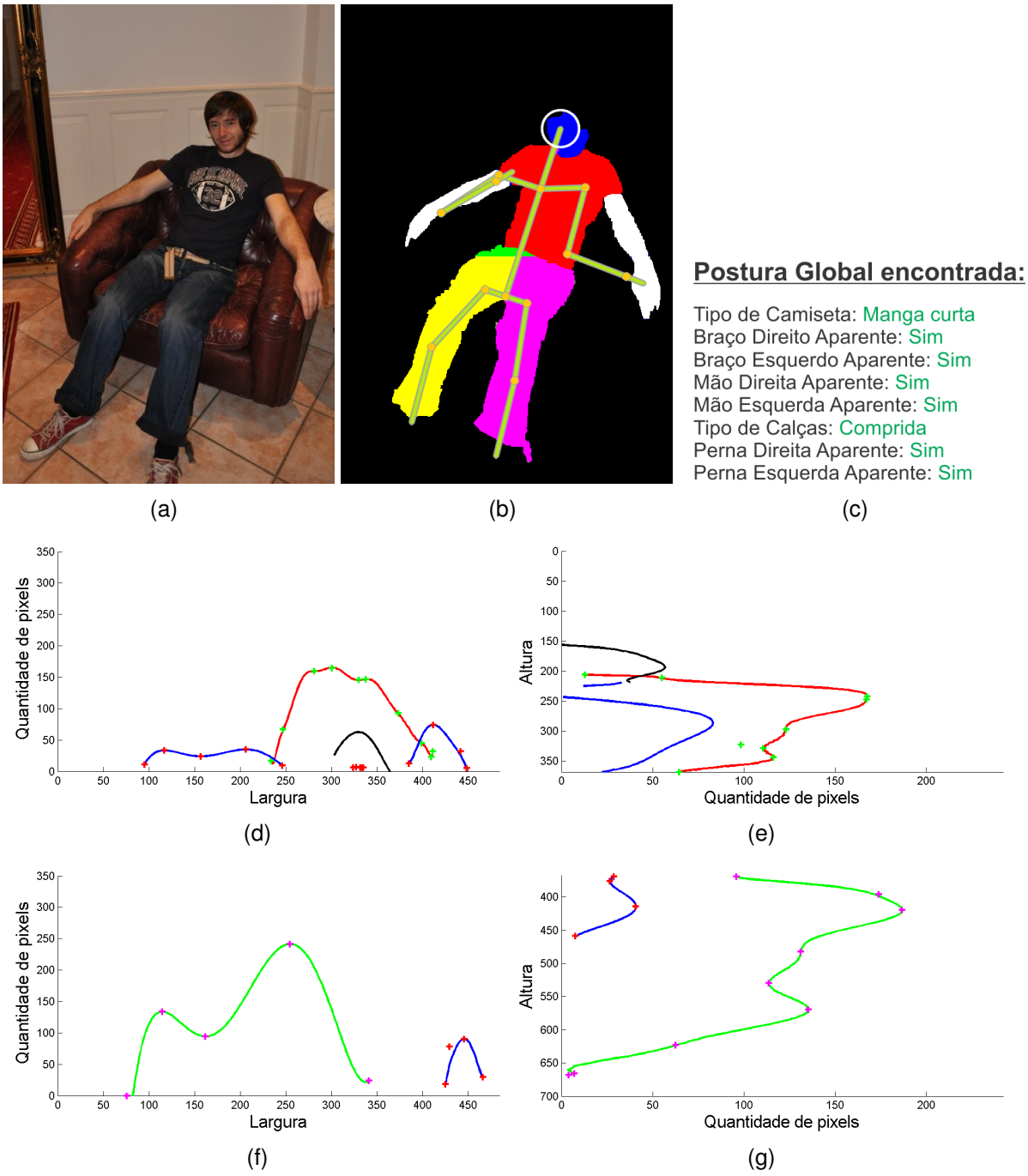


Figura 5.4 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste. O fato da pessoa estar em perspectiva e ainda com a cabeça inclinada para a frente de seu corpo implicou no deslocamento de todo o esqueleto e, dessa forma, impossibilitando que o método do ajuste obtivesse sucesso na tarefa de posicionar os braços corretamente; (c) Pose global gerada pelo método redes neurais , o qual conseguiu encontrar uma postura coerente com as características da foto; (d) Projeção vertical superior; (e) Projeção horizontal superior; (f) Projeção vertical inferior; (g)Projeção horizontal inferior.

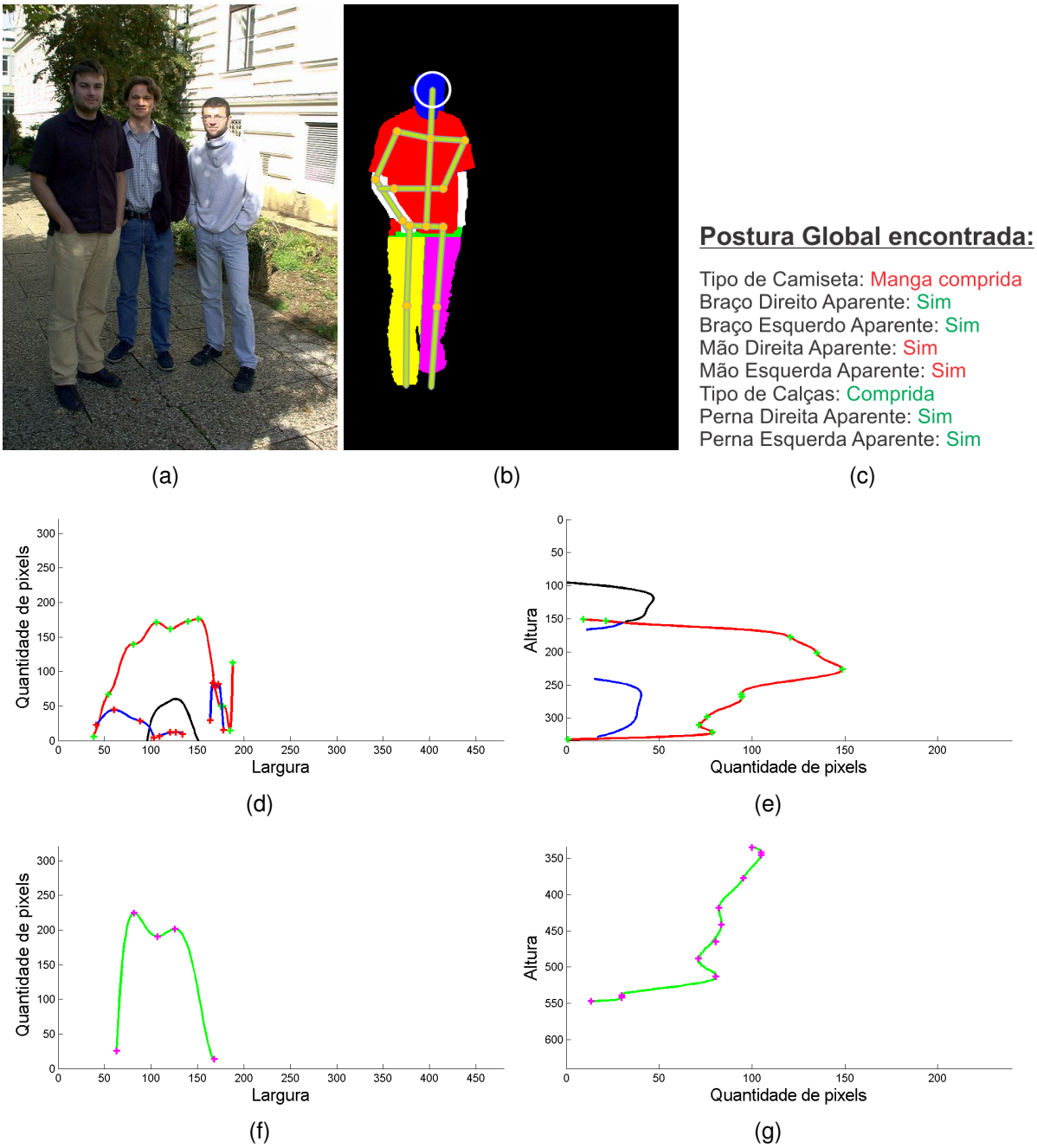


Figura 5.5 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste. O erro ao estimar a pose global da pessoa implicou na falha do método de ajuste ao tentar posicionar o braço esquerdo e também na procura indevida por mãos que não estão aparentes; (c) Pose global gerada pelo método redes neurais. Os critérios classificados incorretamente estão marcados de vermelho; (d) Projeção vertical superior; (e) Projeção horizontal superior; (f) Projeção vertical inferior; (g)Projeção horizontal inferior.

todo das redes neurais, o qual é responsável por encontrar a pose global da pessoa na imagem, apresenta muitas vezes classificações que não condizem com a “real pose global

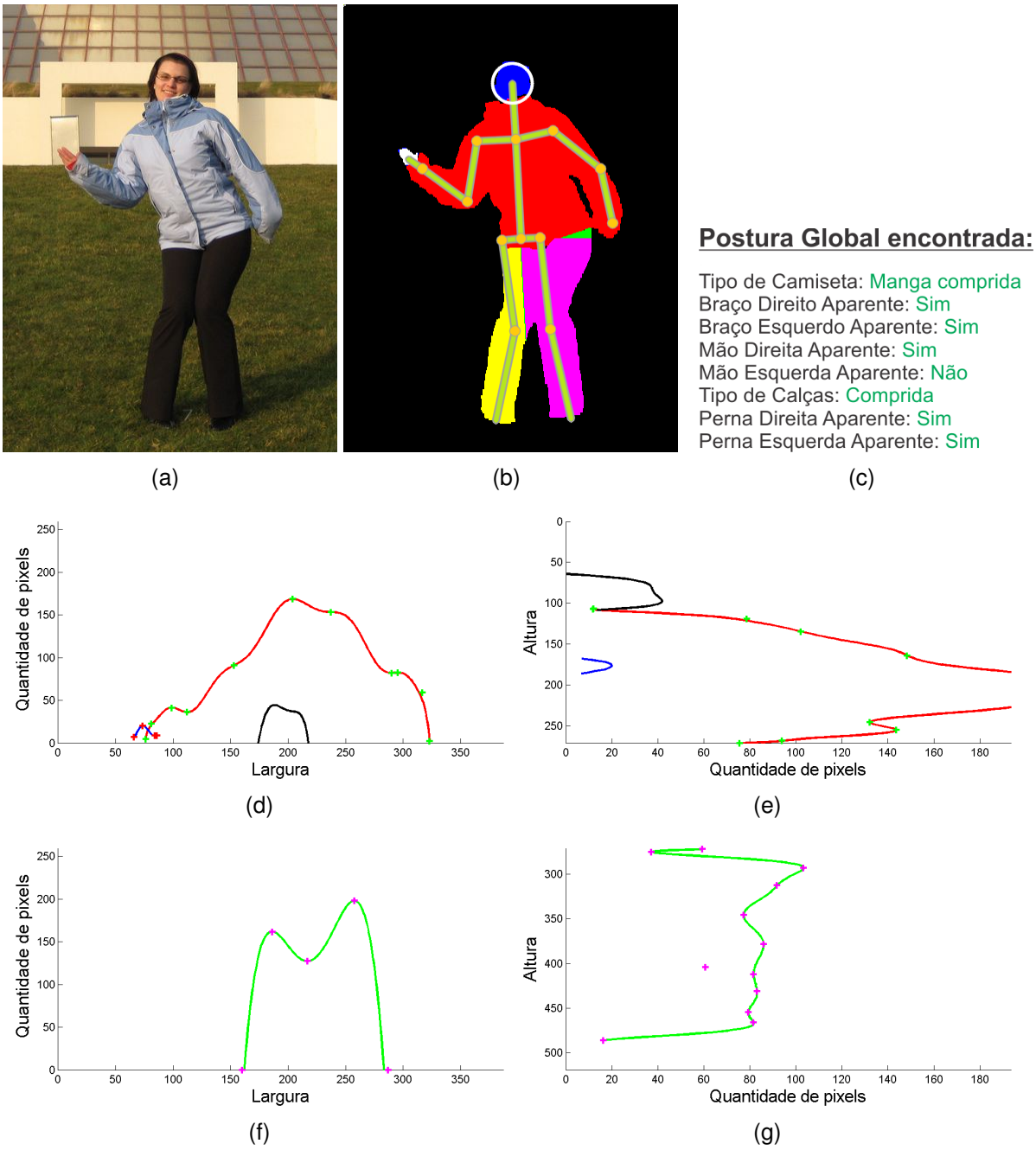


Figura 5.6 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste; (c) Pose global gerada pelo método redes neurais, o qual conseguiu encontrar uma postura que confere com as características da foto; (d) Projeção vertical da parte superior do corpo; (e) Projeção horizontal da parte superior do corpo (f) Projeção vertical da parte inferior do corpo; (g)Projeção horizontal da parte inferior do corpo.

da pessoa". Dessa forma, como a pose global é uma característica capaz de ser avaliada como correta ou não, foi montada a Tabela 5.1 visando melhor apresentar o potencial do método das redes neurais. Esta tabela mostra a porcentagem de erros encontrados

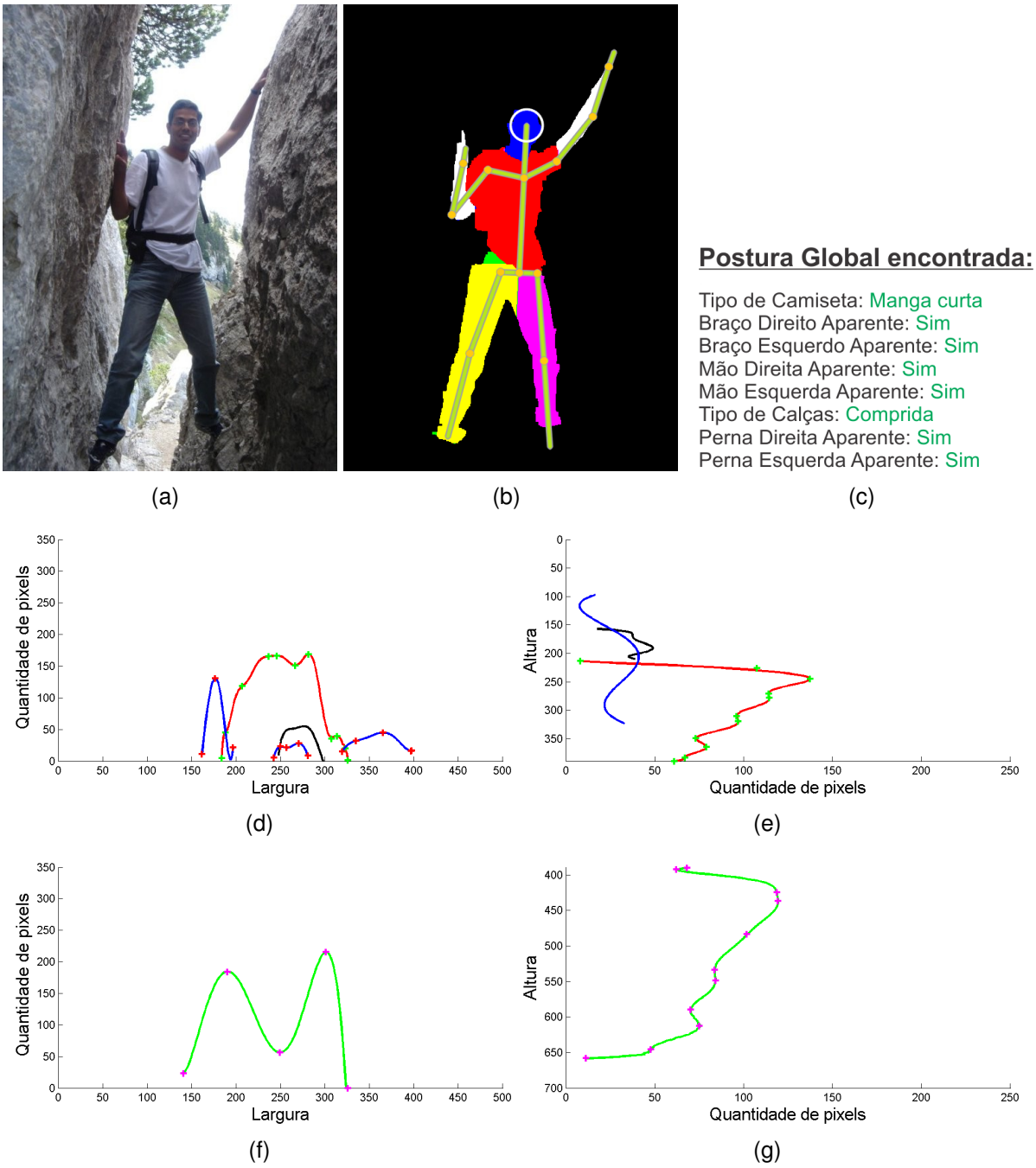


Figura 5.7 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste; (c) Pose global gerada pelo método redes neurais, o qual conseguiu encontrar uma postura que confere com as características da foto; (d) Projeção vertical da parte superior do corpo; (e) Projeção horizontal da parte superior do corpo (f) Projeção vertical da parte inferior do corpo; (g)Projeção horizontal da parte inferior do corpo.

para cada critério que representa a pose global, considerando uma amostra de 25 imagens manualmente segmentadas.

Tabela 5.1 – Porcentagem de erros encontrados para cada critério que representa a pose global, considerando uma amostra de 25 imagens manualmente segmentadas.

	Tipo de camiseta	Braço direito aparente	Braço esquerdo aparente	Mão direita aparente	Mão esquerda aparente	Tipo de calças	Perna direita aparente	Perna esquerda aparente
% de erro	4%	12%	4%	36%	24%	0%	0%	4%

5.2 Resultados com Imagens Segmentadas Automaticamente

Nesta seção serão apresentados alguns resultados obtidos a partir da execução do modelo proposto neste trabalho utilizando-se imagens segmentadas automaticamente. Neste caso, os resultados não são tão bons quanto os apresentados na Seção 5.1 pois o processo de segmentação automático ainda é um problema em aberto e, o modelo utilizado neste trabalho para realizar a segmentação automática [14] ainda encontra-se em desenvolvimento. Dessa forma, há muitas falhas no processo de segmentação que “confundem” o método das projeções e, consequentemente, todos os outros métodos apresentados neste trabalho. Além disso, o banco de dados utilizado para o treinamento do método das redes neurais não possui nenhuma imagem segmentada automaticamente, não estando preparado para lidar com os erros que por ventura resultam do processo de segmentação automática.

Assim sendo, a Figura 5.8 mostra um bom resultado obtido pelos métodos do *pipeline* proposto neste trabalho utilizando-se como entrada uma imagem segmentada automaticamente. A única falha ocorrida para esta imagem foi no método das redes neurais, que errou ao classificar a mão direita da pessoa como estado aparente.

Na Figura 5.9 o resultado obtido também foi coerente. Novamente, o único erro obtido foi no método das redes neurais, classificando, incorretamente, a mão esquerda como aparente. Provavelmente este erro ocorreu por haver erros de segmentação de pele, como pode-se perceber ao observar o canal azul da imagem, o qual faz referência a pele da pessoa na fotografia.

A Figura 5.10 apresenta outro resultado no qual a postura global foi classificada incorretamente, mas, desta vez o critério classificado de forma incorreta foi o tipo de camiseta. Mais uma vez acredita-se que este erro ocorreu devido às falhas na segmentação da pele, que classificou como pixels de pele alguns pontos do *background* da imagem. No entanto, neste caso o erro não influenciou o resultado final.

Para as Figuras 5.11, 5.12 e 5.13 os métodos desenvolvidos para detecção de pose apresentaram resultados esperados, detectando com sucesso as poses das pessoas na fotografia. Contudo, na Figura 5.11 o método de identificação obteve sucesso ao tentar identificar as mãos da pessoa, no entanto, como as mãos foram identificadas corretamente na postura global, o método do ajuste posicionou as mãos na extremidade dos antebraços.

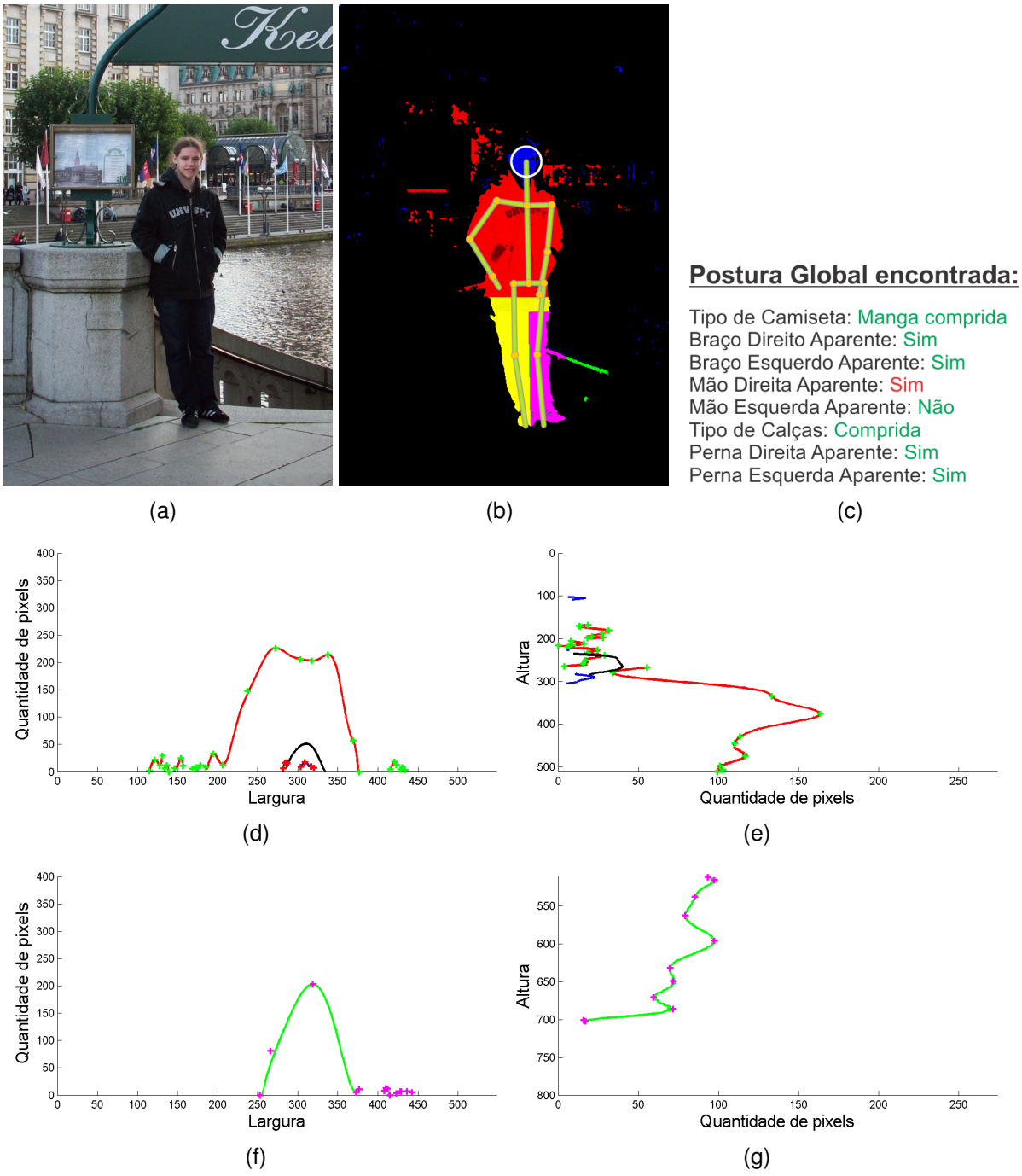


Figura 5.8 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste; (c) Pose global gerada pelo método redes neurais. Somente um dos oito critérios da pose global foi classificado de forma incorreta, o qual está destacado em vermelho; (d) Projeção vertical superior; (e) Projeção horizontal superior; (f) Projeção vertical inferior; (g)Projeção horizontal inferior.

Já a Figura 5.14 apresenta um caso no qual embora a postura global tenha falhado em somente um critério (braço direito aparente) a detecção da postura foi prejudicada pela falha

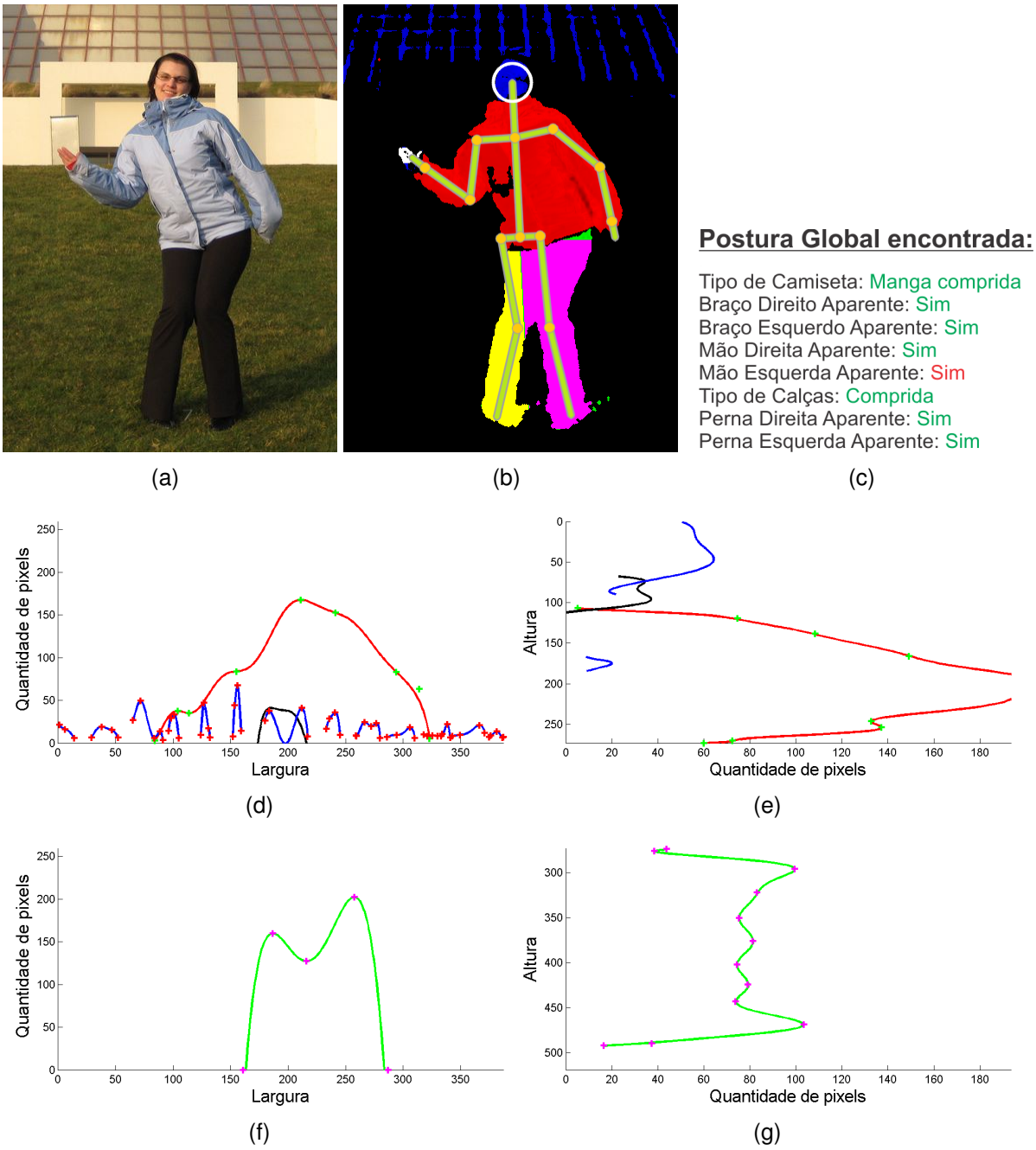


Figura 5.9 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste; (c) Pose global gerada pelo método redes neurais. Somente uma classificação incorreta, a qual está destacada em vermelho; (d) Projeção vertical superior; (e) Projeção horizontal superior; (f) Projeção vertical inferior; (g)Projeção horizontal inferior.

do método de segmentação automática. No caso, foram segmentadas equivocadamente várias pequenas regiões como sendo de roupa da parte inferior do corpo. Além disso, estas regiões não conseguiram ser eliminadas pelo método de identificação, que acabou detectando pontos de perna muito abaixo do que deveria. Isto acarretou em uma estimativa

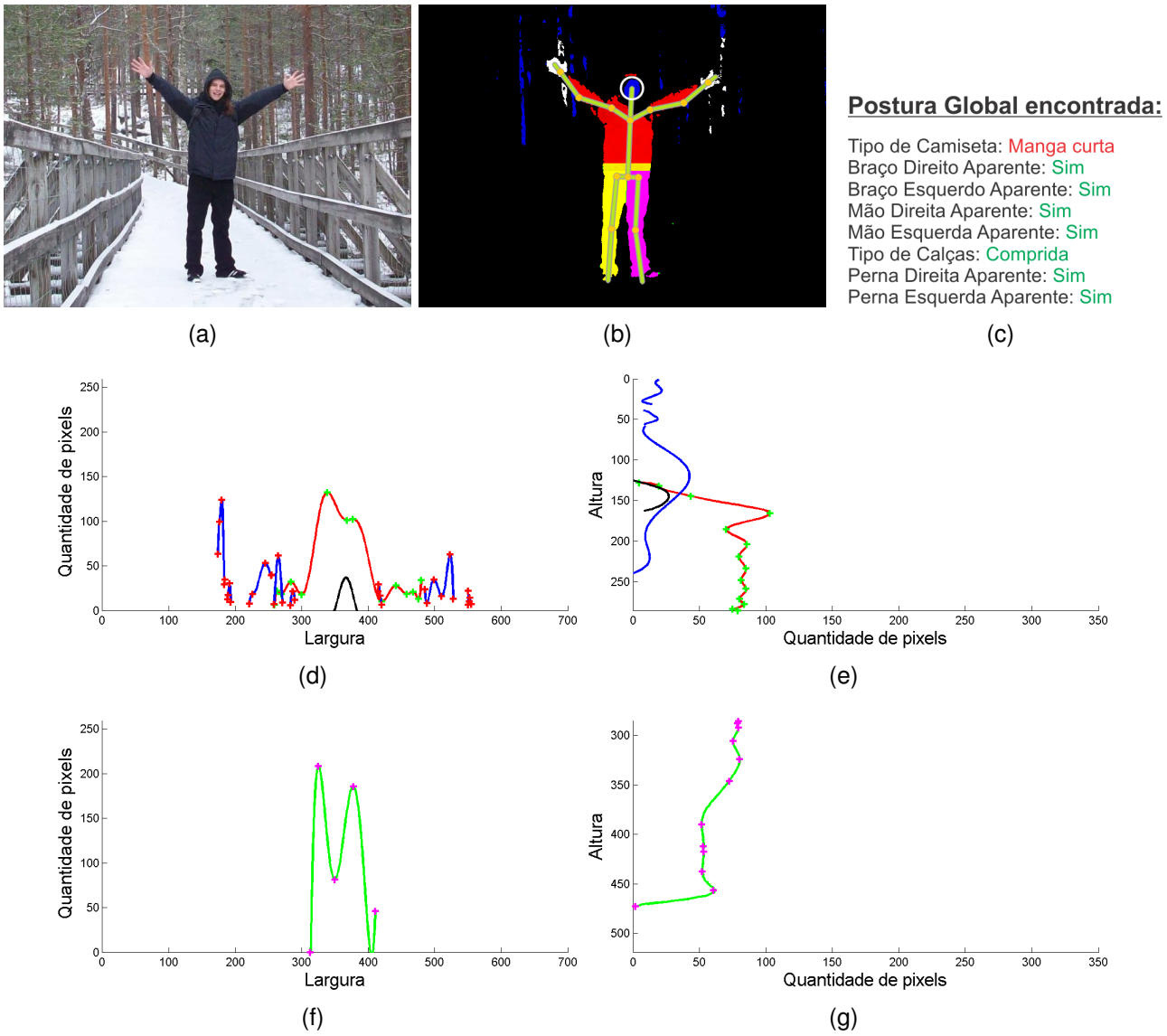


Figura 5.10 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste; (c) Pose global gerada pelo método redes neurais; (d) Projeção vertical superior; (e) Projeção horizontal superior; (f) Projeção vertical inferior; (g)Projeção horizontal inferior.

incorreta para a altura da pessoa e, consequentemente, em uma super-estimativa do tamanho dos ossos. Dessa forma, embora as clavículas tenham sido corretamente posicionadas, o restante dos ossos não obtiveram o mesmo êxito.

Da mesma forma que foi apresentada na Seção 5.1 uma tabela contendo as taxas de erro do método das redes neurais para imagens segmentadas à mão, é apresentada a Tabela 5.2. Esta tabela mostra a porcentagem de erros encontrados para cada critério que representa a pose global, considerando uma amostra de 26 imagens segmentadas automaticamente. No entanto, não é possível realizar uma comparação quantitativa entre

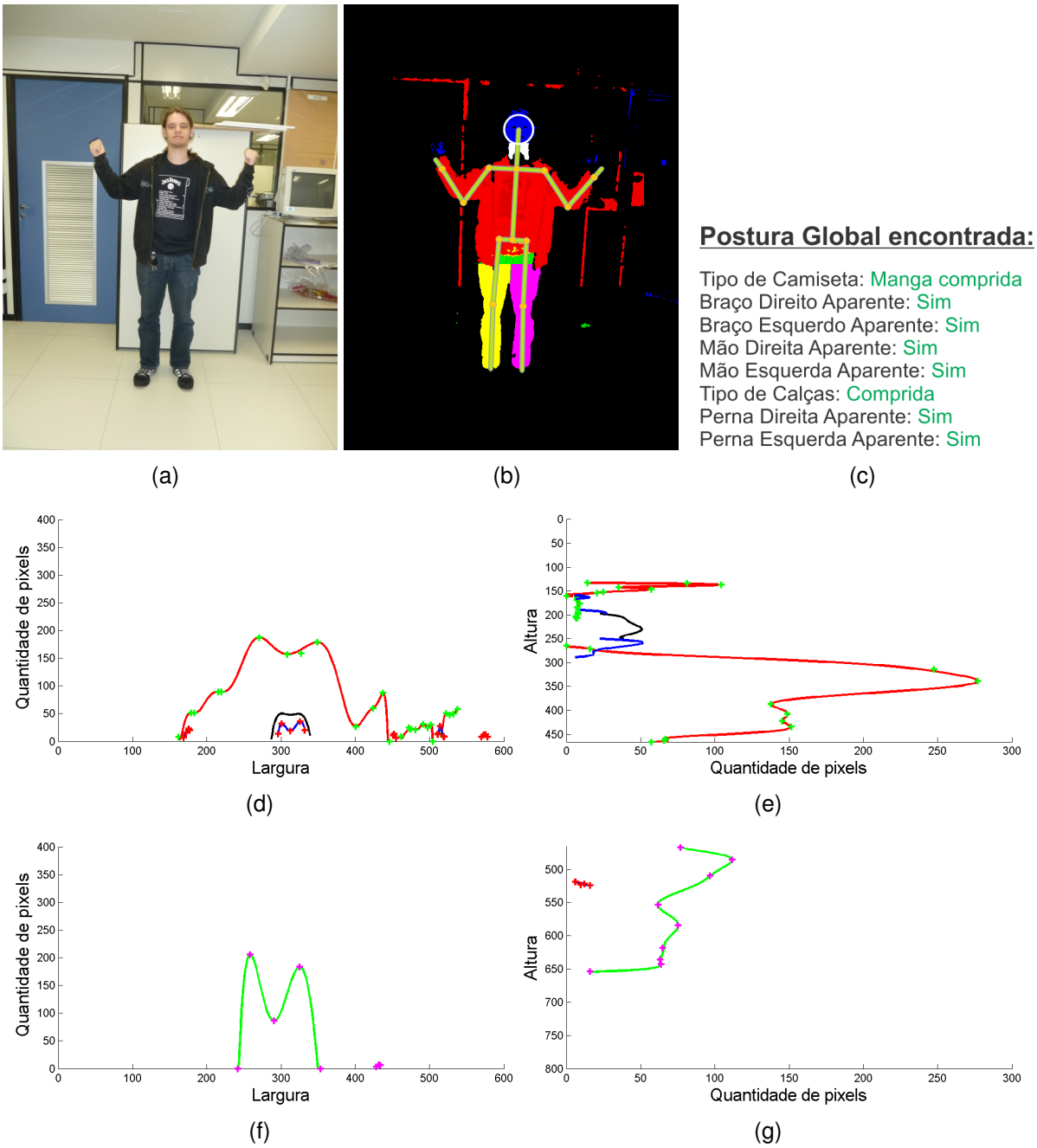


Figura 5.11 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste; (c) Pose global gerada pelo método redes neurais; (d) Projeção vertical superior; (e) Projeção horizontal superior; (f) Projeção vertical inferior; (g)Projeção horizontal inferior.

as duas tabelas pois as imagens utilizadas não foram as mesmas.

No próximo capítulo são realizadas as considerações finais, bem como uma avaliação dos métodos apresentados neste capítulo e algumas sugestões de trabalhos futuros.

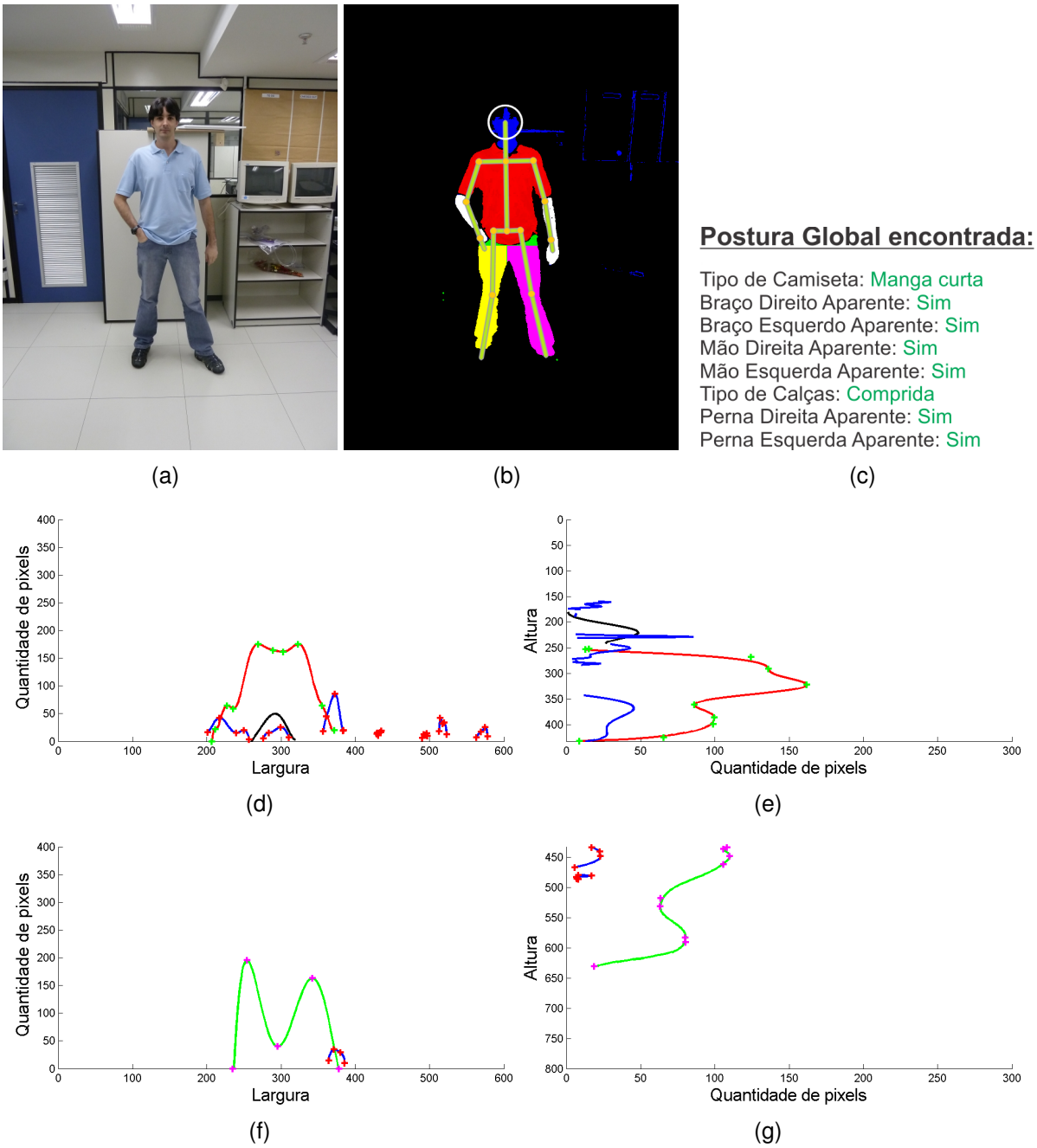


Figura 5.12 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste; (c) Pose global gerada pelo método redes neurais; (d) Projeção vertical superior; (e) Projeção horizontal superior; (f) Projeção vertical inferior; (g)Projeção horizontal inferior.

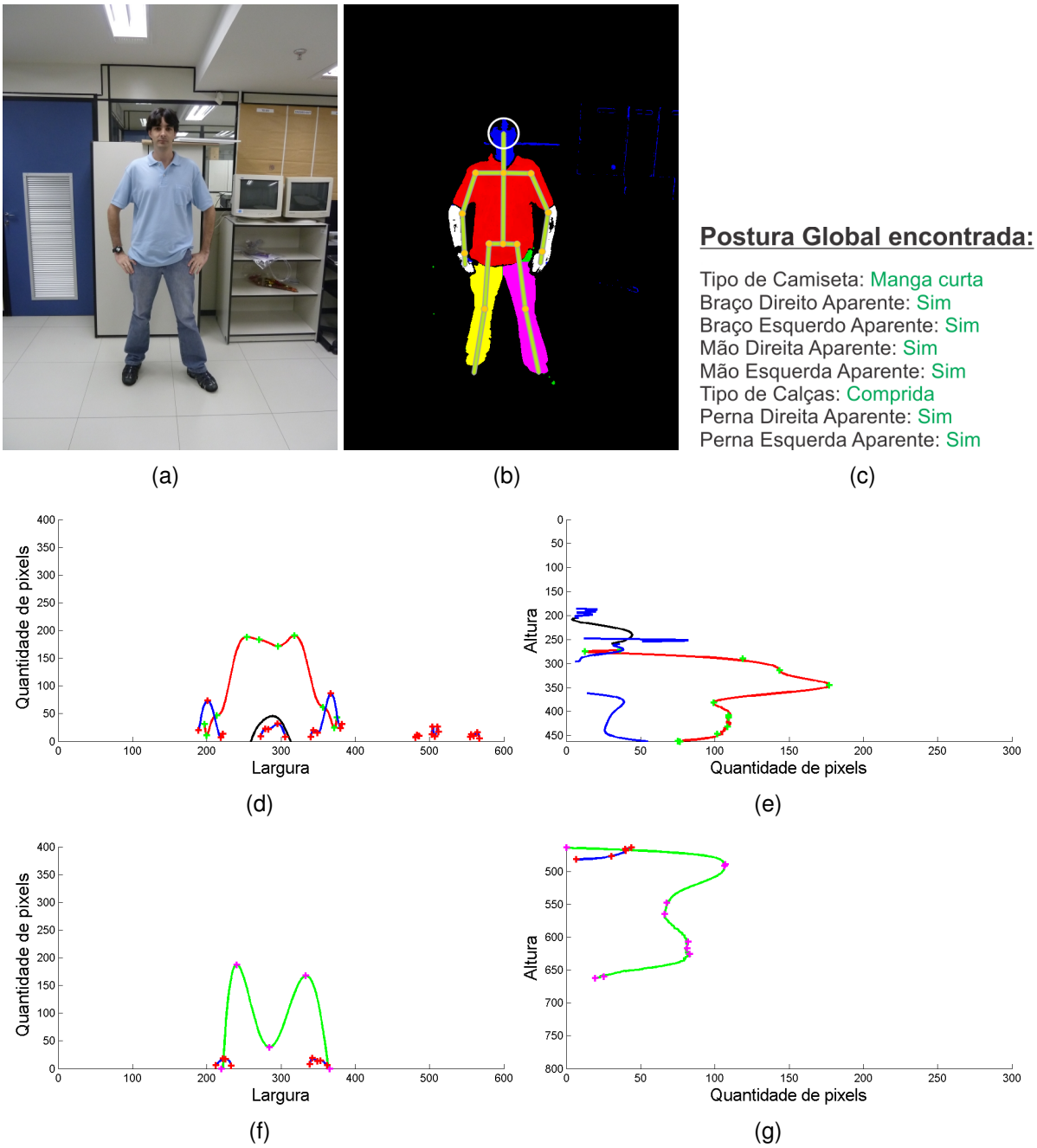


Figura 5.13 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste; (c) Pose global gerada pelo método redes neurais; (d) Projeção vertical superior; (e) Projeção horizontal superior; (f) Projeção vertical inferior; (g)Projeção horizontal inferior.

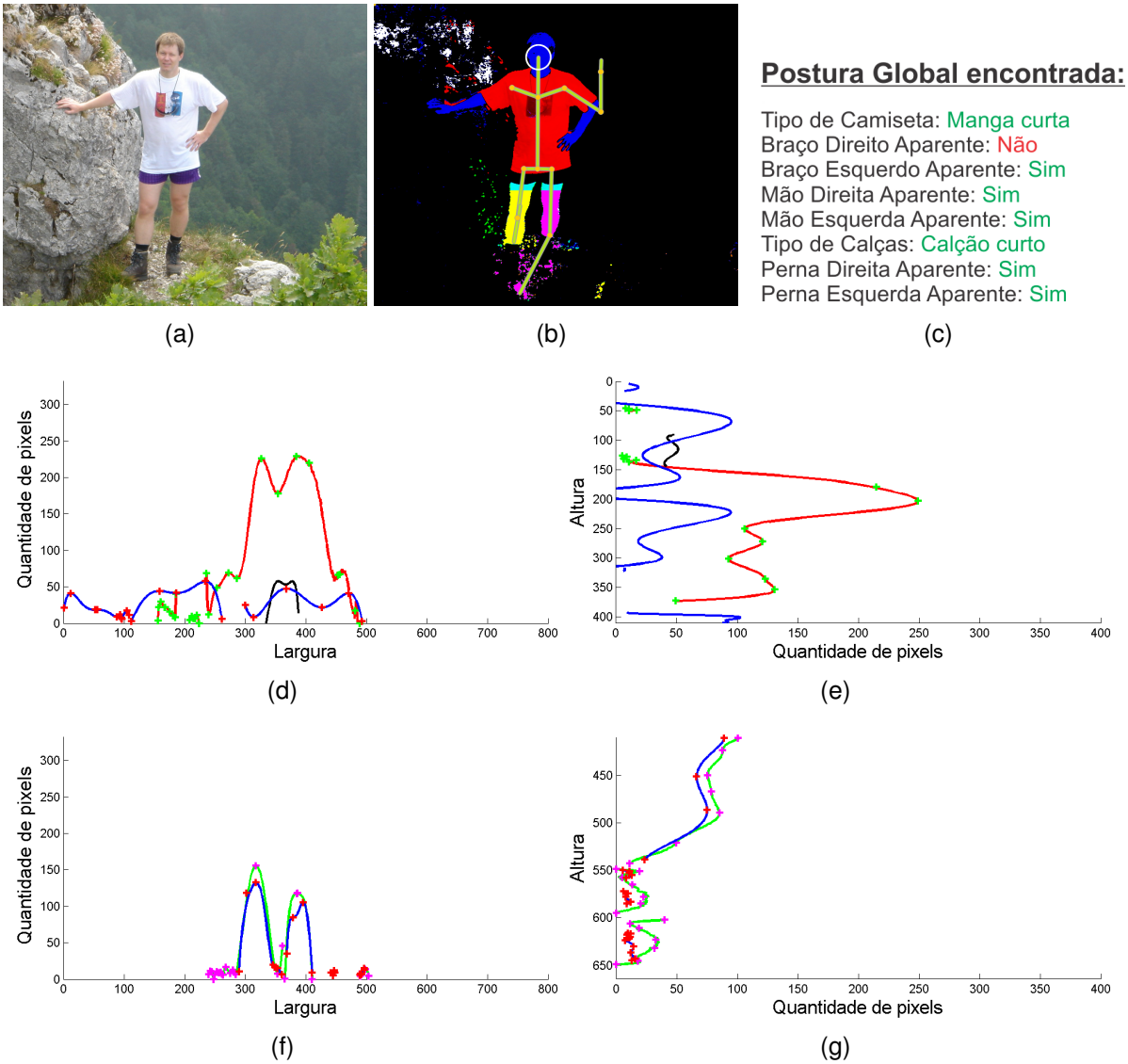


Figura 5.14 – Resultados: (a) Imagem original; (b) Resultados encontrados pelos métodos de identificação e do ajuste; (c) Pose global gerada pelo método redes neurais; (d) Projeção vertical superior; (e) Projeção horizontal superior; (f) Projeção vertical inferior; (g)Projeção horizontal inferior.

Tabela 5.2 – Porcentagem de erros encontrados para cada critério treinado, considerando uma amostra de 26 imagens.

	Tipo de camiseta	Braço direito aparente	Braço esquerdo aparente	Mão direita aparente	Mão esquerda aparente	Tipo de calças	Perna direita aparente	Perna esquerda aparente
% de erro	30,7%	3,85%	0%	7,69%	30,77%	7,69%	3,85%	0%

6. CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentado um modelo para detecção de esqueletos 2D a partir de imagens, desenvolvido no âmbito de pesquisa do laboratório de pesquisa VHLAB, e foi apresentada uma abordagem baseada em um *pipeline* de modo que cada etapa deste *pipeline* repasse as informações adquiridas para as etapas posteriores, com o intuito de de acumular conhecimento.

A contribuição principal deste trabalho é prover um modelo no qual não há intervenção do usuário, o treinamento de postura foi realizado somente uma vez (não existe busca exaustiva em banco de dados) e que é capaz de encontrar posturas coerentes mesmo quando na imagem estas informações são difíceis de serem vistas, que é o caso de quando as mãos não estão aparentes e os braços estão juntos ao corpo (com a mesma cor do tronco).

Realizar uma avaliação para um modelo de detecção de posturas baseada em fotografias não é uma tarefa fácil de ser realizada uma vez que não se dispõe de um *Ground Truth* (base de dados com as informações corretas sobre os dados que queremos avaliar, no caso, a pose 2D da pessoa na fotografia), que é o caso deste trabalho. Neste caso, deve ser realizada um avaliação visual e, dessa forma, os métodos apresentados neste trabalho mostraram-se capazes de, seguindo o *pipeline* proposto, encontrar um esqueleto 2D de uma pessoa para a maioria das imagens testadas. Além disso, dependendo da aplicação, mesmo as poses com pouca precisão podem ser úteis. Contudo, um dos objetivos deste trabalho é retornar as informações sobre a pose encontrada ao método de segmentação automático. Assim, esta segmentação pode ser melhorada, o que consequentemente melhoraria o funcionamento do modelo apresentado neste trabalho.

6.1 Trabalhos Futuros

Há várias possibilidades de melhorias para o modelo desenvolvido. Uma delas poderia ser, como citado anteriormente, repassar as informações sobre a postura encontrada (tais como posição dos ossos e articulações) para o método de segmentação automática. Desta forma, a segmentação seria facilitada, e com uma segmentação de melhor precisão os esqueletos encontrados seriam mais adequados.

Outra melhoria possível seria aumentar o número imagens do banco de dados de treinamento do método das redes neurais, inclusive incluindo imagens segmentadas automaticamente. Além disso, outras configurações para a rede podem ser exploradas para tentar obter melhor precisão nos resultados deste método.

Seria desejável realizar melhorias no método apresentado no sentido de diminuir o nú-

mero de heurísticas utilizadas (aumentando a confiabilidade do modelo) e, também, diminuir o custo computacional (objetivando-se utilizar o método em sistemas de tempo real).

Além disso, poderia-se criar um *Ground Truth* para tornar possível a realização de uma avaliação quantitativa do método desenvolvido.

Uma proposta futura seria disponibilizar o modelo para testes juntamente com a bases trabalhadas e arquivos contendo as informações das poses detectadas. Assim, outras pessoas poderiam utilizar os resultados obtidos com o modelo apresentado neste trabalho para comparação de seus próprios métodos.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Bourdev, L.; Malik, J. “Poselets: Body part detectors trained using 3d human pose annotations”. In *International Conference on Computer Vision*, sep 2009.

[2] Bradski, G. “The OpenCV Library”. *Dr. Dobb’s Journal of Software Tools*, 2000.

[3] Brunelli, R. “Template matching techniques in computer vision: Theory and practice”. Wiley, May 2009.

[4] Canny, J. “A computational approach to edge detection”. *IEEE Transactions Pattern Analysis and Machine Intelligence 8*, November 1986, 679–698.

[5] Dalal, N.; Triggs, B. “Inria person dataset”. Capturado em: http://pascal.inrialpes.fr/data/human/, Março 2009.

[6] Dimitrijevic, M.; Lepetit, V.; Fua, P. “Human body pose detection using bayesian spatio-temporal templates”. *Comput. Vis. Image Underst. 104*, 2, 2006, 127–139.

[7] Drillis, R.; Contini, R. *Body Segment Parameters*. School of Engineering Science, New York University, 1966.

[8] Gavrila, D. M.; Giebel, J.; Munder, S. “Vision-based pedestrian detection: the protector system”. In *Intelligent Vehicles Symposium, 2004 IEEE*, 2004, pp. 13–18.

[9] Gilks, W. R. “Markov chain monte carlo in practice”. Chapman & Hall/CRC, December 1995.

[10] Guimarães Jr, D. S.; Corrêa, U. B.; Carro, L.; Reis, R. “Estimador de potência e atraso em standard-cells utilizando redes neurais artificiais”. *IBERCHIP Workshop 16*, 2010.

[11] Haykin, S. “Neural networks: A comprehensive foundation”, 2nd ed. Prentice Hall, Upper Saddle River, NJ, USA, 1998.

[12] Haykin, S. “Neural networks and learning machines”, 3 ed. Prentice Hall, November 2008, 936p.

[13] Hu, Z.; Wang, G.; Lin, X.; Yan, H. “Recovery of upper body poses in static images based on joints detection”. *Pattern Recognition Letters 30*, 5, 2009, 503–512.

[14] Jacques, J. C.; Dihl, L.; Jung, C.; Thielo, M.; Keshet, R.; Musse, S. “Human upper body identification from images”. In *ICIP*, 2010, IEEE, pp. 1717–1720.

[15] Jacques, J. C.; Jung, C.; Musse, S. “A background subtraction model adapted to illumination changes”. In *IEEE International Conference on Image Processing*, 2006, pp. 1817–1820.

[16] Ju, S. X.; Black, M. J.; Yacoob, Y. “Cardboard people: A parameterized model of articulated image motion”. *Automatic Face and Gesture Recognition, IEEE International Conference on 0*, 1996, 38.

[17] Lee, D.-Y.; Ahn, J.-K.; Kim, C.-S. “Fast background subtraction algorithm using two-level sampling and silhouette detection”. In *Proceedings of the 16th IEEE international conference on Image processing*, Piscataway, NJ, USA, 2009, ICIP’09, IEEE Press, pp. 3141–3144.

[18] Lee, M. W.; Cohen, I. “A model-based approach for estimating human 3d poses in static images”. *IEEE Trans. Pattern Anal. Mach. Intell. 28*, 6, 2006, 905–916. Student Member-Lee, Mun Wai and Member-Cohen, Isaac.

[19] Liñán, C. C. “cvBlob - Blob library for OpenCV”. Disponível em: http://cvblob.googlecode.com, 2010 Janeiro.

[20] Maglogiannis, I.; Vouyioukas, D.; Aggelopoulos, C. “Face detection and recognition of natural human emotion using markov random fields”. *Personal Ubiquitous Comput. 13*, 1, 2009, 95–101.

[21] McIntosh, C.; Hamarneh, G.; Mori, G. “Human limb delineation and joint position recovery using localized boundary models”. In *WMVC ’07: Proceedings of the IEEE Workshop on Motion and Video Computing*, Washington, DC, USA, 2007, IEEE Computer Society, p. 31.

[22] Microsoft. “Kinect para xbox 360”. Disponível em: http://www.xbox.com/pt-br/kinect, Janeiro 2011.

[23] Mori, G. “Recovering 3d human body configurations using shape contexts”. *IEEE Transactions Pattern Analysis and Machine Intelligence 28*, 7, 2006, 1052–1062. Senior Member-Malik, Jitendra.

[24] Mori, G.; Ren, X.; Efros, A. A.; Malik, J. “Recovering human body configurations: combining segmentation and recognition”. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2004, vol. 2, pp. II–326–II–333 Vol.2.

[25] Olson, C. F.; Huttenlocher, D. P. “Automatic target recognition by matching oriented edge pixels”. *IEEE Transactions on Image Processing 6*, 1997, 103–113.

[26] Rogez, G.; Orrite-Uru nuela, C.; Martínez-del Rincón, J. “A spatio-temporal 2d-models framework for human pose recovery in monocular sequences”. *Pattern Recogn. 41*, 9, 2008, 2926–2944.

[27] Thalmann, N. M.; Thalmann, D., Eds. “Handbook of virtual humans”, 1 ed. John Wiley, Chichester, 2004, 468p.

[28] Tilley, A. R.; Associates, H. D. “The measure of man and woman: Human factors in design”. Wiley, New York, 2001, 104p.

[29] Viola, P.; Jones, M. J. “Robust real-time face detection”. *International Journal of Computer Vision 57*, 2, 2004, 137–154.

[30] Vrubel, A.; Bellon, O. R. P.; Silva, L. “Planar background elimination in range images: a practical approach”. In *Proceedings of the 16th IEEE international conference on Image processing*, Piscataway, NJ, USA, 2009, ICIP’09, IEEE Press, pp. 3161–3164.

[31] Yang, L.; Kavli, T.; Carlin, M.; Clausen, S.; de Groot, P. F. M. “An evaluation of confidence bound estimation methods for neural networks”. In *Advances in Computational Intelligence and Learning: Methods and Applications*, 2000, Kluwer Academic Publishers, pp. 71–84.