

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

**SEGMENTAÇÃO DE PESSOAS EM IMAGENS  
ESTÁTICAS BASEADA EM ESQUELETO**

JULIO CEZAR SILVEIRA JACQUES JUNIOR

Tese apresentada como requisito à obtenção do grau de Doutor em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Prof<sup>a</sup>. Dr<sup>a</sup>. Soraia Raupp Musse

Co-orientador: Prof. Dr. Cláudio Rosito Jung

**Porto Alegre  
2012**



J19s

Jacques Junior, Julio Cezar Silveira

Segmentação de pessoas em imagens estáticas baseada em esqueleto / Julio Cezar Silveira Jacques Junior. – Porto Alegre, 2012.

114 f.

Tese (Doutorado) – Fac. de Informática, PUCRS.

Orientador: Prof. Dr. Soraia Raupp Musse.

Co-orientador: Prof. Dr. Cláudio Rosito Jung.

1. Informática. 2. Processamento de Imagens. 3. Semântica.  
4. Esqueleto. I. Musse, Soraia Raupp. II. Jung, Cláudio Rosito.  
III. Título.

CDD 006.61

**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

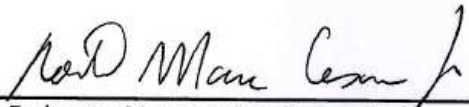
### TERMO DE APRESENTAÇÃO DE TESE DE DOUTORADO

Tese intitulada "Segmentação de Pessoas em Imagens Estáticas Baseada em Esqueleto", apresentada por Julio Cezar Silveira Jacques Junior, como parte dos requisitos para obtenção do grau de Doutor em Ciência da Computação, Sistemas Interativos e de Visualização, aprovada em 20/04/2012 pela Comissão Examinadora:

  
\_\_\_\_\_  
Profa. Dra. Soráia Raupp Musse - PPGCC/PUCRS  
Orientadora

  
\_\_\_\_\_  
Prof. Dr. João Batista Souza de Oliveira - PPGCC/PUCRS

  
\_\_\_\_\_  
Prof. Dr. Jacob Scharcanski - UFRGS

  
\_\_\_\_\_  
Prof. Dr. Roberto Marcondes Cesar Junior - USP

Homologada em 08/06/2012, conforme Ata No. 012, pela Comissão Coordenadora.

  
\_\_\_\_\_  
Prof. Dr. Paulo Henrique Lemelle Fernandes  
Coordenador.

**PUCRS**

**Campus Central**  
Av. Ipiranga, 6681 - P. 32 - sala 507 - CEP: 90619-900  
Fone: (51) 3320-3611 - Fax (51) 3320-3621  
E-mail: [ppgcc@pucrs.br](mailto:ppgcc@pucrs.br)  
[www.pucrs.br/facin/pos](http://www.pucrs.br/facin/pos)



## AGRADECIMENTOS

Agradeço a meus pais, Julio (*in memorian*) e Lucia, que sempre com muito amor, dedicação, simplicidade e honestidade me educaram e ensinaram o quanto é importante dar valor às pequenas coisas, aos pequenos gestos e, principalmente às pessoas.

Às minhas irmãs, Letícia e Luciana, que sempre estiveram presente, me incentivando e me inspirando. Aos meus cunhados, André e Douglas, pela força e principalmente pelo amor dedicado por eles à essas duas mulheres tão especiais. Às minhas queridas sobrinhas, Isabelle, Gabriela e Julia, que encham de alegria nossas vidas. Talvez elas nem saibam o quando foram importantes.

A minha esposa, amiga e companheira, Cristiane, por estar sempre ao meu lado.

Agradeço também a todos os familiares e amigos que nunca me deixaram sem o apoio necessário para que eu pudesse chegar ao fim desse trabalho.

A minha orientadora, professora Soraia Raupp Musse e ao professor Cláudio Rosito Jung (co-orientação), pela confiança, incentivo e amizade conquistados. Pela atenção com que sempre me receberam e pela dedicação, orientação segura e esforços, que tornaram possível a execução desse trabalho.

A todo o pessoal do Centro de Pesquisa em Computação Aplicada da PUCRS e do laboratório VHLab, que me ajudaram direta ou indiretamente.

Aos professores e funcionários da PUCRS, que de alguma maneira, também contribuíram para o meu desenvolvimento.

A HP Brasil, pelo apoio financeiro.





# SEGMENTAÇÃO DE PESSOAS EM IMAGENS ESTÁTICAS BASEADA EM ESQUELETO

## RESUMO

A segmentação (automática ou semi-automática) de pessoas em imagens estáticas é uma tarefa bastante desafiadora, principalmente devido a diversos fatores do mundo real, como por exemplo, fatores relacionados à iluminação da cena onde a imagem foi capturada, sombras, ruídos na imagem, oclusão, alta similaridade do objeto de interesse com o fundo da cena e a falta de informação inerente de profundidade quando uma cena é capturada em uma imagem 2D. Nessa tese é apresentado um modelo para segmentação de pessoas em imagens baseado em esqueleto. Os dados de entrada para o modelo proposto, associados ao modelo de esqueleto, podem ser obtidos de forma automática (utilizando um algoritmo para estimativa de pose 2D de pessoas em imagens, por exemplo) ou manual (através de interação com usuário), dependendo da aplicação em questão. O modelo de esqueleto guia a segmentação da pessoa na imagem levando em consideração informações de cor, luminosidade, restrições de ângulos e parâmetros antropométricos. De uma forma geral, a idéia principal da abordagem proposta é construir um grafo ao redor do modelo de esqueleto, para uma determinada imagem de entrada, e buscar o melhor caminho nesse grafo que satisfaça uma determinada condição (por exemplo, aquela que maximiza certo critério de energia), gerando assim o contorno da pessoa na imagem.

Também está sendo proposta nessa tese uma abordagem para avaliar quantitativamente os resultados experimentais obtidos, a partir de informações fornecidas através de interação com usuário. Os resultados experimentais demonstram que o modelo proposto gera resultados satisfatórios para imagens não triviais, contendo pessoas com aparências e poses variadas (podendo haver membros parcialmente ocultos), em diversos ambientes complexos (e não controlados), com diferentes iluminações e qualidade de imagem, entre outros fatores. Os resultados obtidos com a utilização do modelo proposto também foram comparados com os obtidos por um trabalho considerado estado-da-arte e os experimentos indicam que o nosso modelo gera resultados mais coerentes para o contorno da pessoa, enquanto que os contornos obtidos pelo trabalho em questão apresentam formas mais suaves.

O modelo de segmentação proposto é capaz de gerar um contorno fechado (para cada pessoa na imagem) contendo informação semântica, ou seja, cada ponto do contorno resultante está associado a uma determinada parte do corpo, que pode ser utilizada para diversos fins (por exemplo, construção de humanos virtuais com características extraídas da imagem, métodos para estimativa de roupas em imagens, estimativa da forma humana sobre as roupas, entre outros).

**Palavras-chave:** segmentação de pessoas em imagens, análise de imagens, informação semântica.



# SKELETON-BASED HUMAN SEGMENTATION IN STILL IMAGES

## ABSTRACT

The segmentation of people (automatic or semi-automatic) in still images is a very challenging task, mainly due to several factors in the real world, such as those related to the lighting of the scene where the image was captured, shadows, image noise, occlusions, high similarity of the object of interest with the background of the scene and the lack of information inherent in depth when a scene is captured into a 2D image. In this work we present a skeleton-based model for human segmentation in still images. The input data of the model, related to the skeleton model, can be obtained automatically (using an algorithm for 2D pose estimation of people in images, for example) or manually (through user interaction), depending on the particular application. The skeleton model is used to guide the segmentation by taking into account color information, brightness, angle constraints and anthropometric parameters. In a general way, the main idea of the proposed approach is to build a graph around the skeleton model, for a given input image, and find out the best path in this graph that satisfies a certain condition (e.g., the one that maximizes a certain energy criterion), thus generating the contour of the person in the picture.

It is also being proposed in this work an approach to measure quantitatively the experimental results, from information provided through user interaction. The experimental results demonstrate that the proposed model generates satisfactory results for non-trivial images containing people with varied appearances and poses (containing self-occlusions), in various complex environments (and uncontrolled), with different lighting conditions and image quality. The results obtained using the proposed model was also compared with those obtained by a work considered state of the art. Our experiments indicate that the proposed model adapts better to the contours, while the human body shape priors in the confronted work enforce a smoother contour.

The proposed segmentation model generates a closed contour (for each person in the image) with semantic information included, (e.g., each contour point is associated with a particular body part), which can be used for various purposes (for example, construction of virtual humans with features extracted from the image, methods for clothes estimation in images, estimation of the human shape under the clothes, etc.).

**Keywords:** person segmentation in still images, image analysis, semantic information.



## LISTA DE FIGURAS

Figura 1.1	Ilustração do método proposto por Hasler e colaboradores [1]. Imagem da esquerda: silhueta da pessoa, segmentada manualmente. Imagem central e direita: resultado da estimativa de forma 3D. . . . .	30
Figura 1.2	Ilustração do método proposto por Zhou e sua equipe [2]. Imagem da esquerda: imagem original. Imagem central: modelo 3D sobreposto na imagem original. Imagem à direita: resultado após a manipulação da imagem. . . . .	30
Figura 1.3	Ilustração do método proposto por Jacques Junior e colaboradores [3]. Imagem da esquerda: face detectada de maneira automática. Imagem central: segmentação de <i>pixels</i> com tons de pele e região do tronco. Imagem à direita: estimativa da pose da parte superior do corpo. . . . .	31
Figura 2.1	Ilustração do método proposto por Hornung e sua equipe [4]. Imagem original é exibida à esquerda. Resultado obtido é exibido à direita. . . . .	35
Figura 2.2	Ilustração do método proposto por Freifeld e equipe [5]. Imagem da esquerda: representação 2D da pessoa contendo informação semântica. Imagem central e direita: resultado da estimativa de pose e segmentação. . . . .	36
Figura 2.3	Ilustração do método proposto por Hu [6] e sua equipe. Imagem da esquerda: imagem de entrada com o tronco detectado (em vermelho) e limites da região do fundo estimada (em azul). Imagem à direita: resultado da segmentação. . . . .	37
Figura 2.4	Ilustração do método proposto por Mori e sua equipe [7]. Imagem da esquerda: imagem de entrada. Ao centro, esqueleto estimado. À direita, resultado da segmentação associada à pessoa. . . . .	38
Figura 2.5	(a) Imagem de entrada. (b-c) resultado detector de bordas Canny, com duas escalas distintas. (d) Mapa de probabilidade de contornos, extraído usando abordagem proposta por Martin e equipe [8]. (e) Resultado do algoritmo <i>Normalized Cuts</i> . (f) Mapa de <i>superpixels</i> gerado para essa imagem. . . . .	38
Figura 2.6	<i>Templates</i> de forma caracterizados em uma representação de árvore, propostos no trabalho de Lin e sua equipe [9]. . . . .	39
Figura 2.7	Ilustração do resultado do trabalho de Lin e sua equipe [9]. Sendo (a), o conjunto inicial de hipóteses detectadas; (b) resultado inicial da segmentação; (c) resultado final da detecção; (d) resultado final da segmentação. . . . .	39
Figura 2.8	Ilustração do método proposto por Su e colaboradores [10]. As imagens representam, da esquerda para a direita: imagem sem <i>flash</i> ; imagem com <i>flash</i> ; resultado da segmentação na imagem sem <i>flash</i> ; e resultado da segmentação na imagem com <i>flash</i> . . . . .	40

Figura 2.9	Ilustração do resultado do trabalho de Guan e sua equipe [11]. À esquerda, imagem de entrada. Ao centro, objeto segmentado sobreposto na imagem. À direita, estimativa 3D da forma e pose da pessoa. . . . .	41
Figura 2.10	Ilustração do resultado do trabalho de Hu e sua equipe [12]. . . . .	41
Figura 2.11	Ilustração do resultado do trabalho de Ren e colaboradores [13]. As imagens representam, respectivamente (da esquerda para a direita): imagem de entrada; mapa de bordas; pose estimada; e segmentação resultante. . . . .	42
Figura 3.1	Visão geral do modelo proposto. . . . .	47
Figura 3.2	Modelo de esqueleto adotado para guiar a segmentação. . . . .	47
Figura 3.3	(a) Ilustração dos três grafos principais ( $A$ , $B$ e $C$ ). Em vermelho, grafo $A$ gerado para parte superior do corpo. Em verde e magenta, grafos $B$ e $C$ , gerados para a parte inferior do corpo (lado direito e esquerdo, respectivamente). (b) Contorno resultante sobreposto na imagem de entrada (convertida para escala de cinza), com esqueleto de entrada ilustrado em ciano.	50
Figura 3.4	(a) Imagem de entrada (e respectivo “osso”, em verde). (b) Magnitude do gradiente da imagem (a) (após conversão para escala de cinza). (c) Nós dos grafos (asteriscos) e caminhos gerados (linhas em amarelo). (d) Zoom na imagem (c) (região representada pelo retângulo azul). (e) Contorno obtido pelo caminho que maximiza o valor de energia. . . . .	50
Figura 3.5	Ilustração do grafo gerado para o contorno externo do braço direito (ilustrado na Figura 3.4). . . . .	51
Figura 3.6	Ilustração da conexão entre dois grafos adjacentes (canela e pé direito, por exemplo). Em (a), os ângulos formados na conexão entre as duas partes. Em (b), regiões associadas a cada parte do corpo podem se interceptar ou deixar “buracos”, dependendo do ângulo formado no ponto de conexão. . . .	54
Figura 3.7	Detalhes da conexão entre dois grafos adjacentes (canela e pé direito, por exemplo). (a) Há intersecção quando o ângulo $\alpha < 180^\circ$ , assim, alguns níveis devem ser removidos, ilustrados à esquerda em magenta, então os dois grafos são conectados por um “nível de conexão”, ilustrado à direita por uma linha em magenta. (b) Há “buracos” quando $\alpha > 180^\circ$ (ilustrado à esquerda), assim, alguns níveis devem ser criados (ilustrado à direita). . . . .	54
Figura 3.8	Ilustração da modelagem do grafo da mão direita (setor circular). Em (a) é exibido o grafo em forma retangular, antes da conexão entre os dois lados (esquerdo e direito, em relação ao “osso”), feita pela inserção do setor circular. Em (b) é ilustrado o setor circular criado, assim como a diminuição do “osso” dessa extremidade em questão (distância $d/2$ ). Em (c) é ilustrado o grafo resultante. . . . .	55

Figura 3.9	Ilustração da modelagem do ombro. Em (a) é ilustrado a diminuição do “osso” do ombro direito, no lado do pescoço. Em (b) é exibido o grafo gerado para o ombro direito, já conectado ao grafo do braço associado. . . . .	56
Figura 3.10	Ilustração da modelagem da cabeça. (a) Estrutura hexagonal do grafo da cabeça e sua conexão com os “ossos” dos ombros (em magenta). (b) Ponto de intersecção entre o grafo da cabeça e o grafo do ombro direito. (c) Grafo da cabeça conectada aos ombros. (d) Grafo da cabeça resultante, normalizado pela largura usada em todo grafo $A$ . . . . .	57
Figura 3.11	Ilustração da modelagem do tronco. Em (a), dois grafos criados para o tronco. Em (b), ilustração do segmento de reta gerado para a criação do grafo do tronco, para o lado direito (que vai do ponto médio do ombro ao início da perna). Em (c), conexão entre o tronco $\times$ coxa. . . . .	57
Figura 3.12	Ilustração da modelagem da coxa (parte interna). Em (a), detalhe da conexão entre tronco $\times$ coxa, antes do tratamento especial. Em (b), tratamento especial para a parte interna do grafo da coxa direita (redução na parte superior) e ilustração do “comprimento do quadril” $l_q$ . . . . .	58
Figura 3.13	Ilustração da conexão entre níveis adjacentes de um grafo. . . . .	59
Figura 3.14	Ilustração dos vetores $t_k$ e $u$ , usados no cálculo do valor de energia de cada aresta do grafo (com restrição antropométrica). . . . .	61
Figura 3.15	Em (a), intersecção ou buracos criados na região de conexão entre a canela e o pé direito. Em (b), pontos salientados (em verde) onde o vetor $u$ está associado às duas partes do corpo. Em (c), ilustração do vetor resultante $u$ , associado às duas partes em questão. . . . .	61
Figura 3.16	(a) Imagem de entrada e seu respectivo “osso” (em verde). (b) Região estimada para o contorno (linhas contínuas). (c) Imagem binária composta pela linha do “osso” apenas. (d) Transformada da Distância $TD_i$ da imagem exibida em (c). (e) mapa resultante $R_i$ de valores de distâncias, usado no cálculo da energia do contorno para essa parte do corpo. . . . .	62
Figura 3.17	Ilustração da distância antropométrica em uma região de conexão entre duas partes do corpo. (a) <i>Pixel</i> em destaque, considerado em uma região de conexão. (b) Zoom na imagem (a) e ilustração do fator de distância $a$ usado no cálculo de $R_{ij}$ . (c) Mapa de distâncias antropométricas resultante para duas partes do corpo (canela e pé direito). . . . .	63
Figura 3.18	Ilustração da influência do mapa de distâncias antropométricas. Em (a), imagem em escala de cinza da canela e pé direito. Em (b), magnitude do gradiente da imagem (a). Em (c), magnitude do gradiente da imagem (a), multiplicada, ponto a ponto, pelo mapa de distâncias antropométricas associado (ilustrado na Figura 3.17(c)). . . . .	64

Figura 3.19	Ilustração das alternativas de caminhos, do nível 1 ao nível 2, passando pelo vértice $S_{2,2}$ . . . . .	65
Figura 3.20	Ilustração de grafos sobrepostos. Nesse caso, as arestas associadas ao braço esquerdo não estão conectadas diretamente ao braço oposto devido à estrutura do grafo. . . . .	66
Figura 3.21	Ilustração dos caminhos gerados para cada grafo principal. (a) Três grafos principais ( $A$ , $B$ e $C$ ) gerados para uma determinada pessoa em uma imagem. (b) Caminho gerado para o grafo da parte superior do corpo (grafo $A$ ). (c) Caminho gerado para o grafo da parte inferior direita do corpo (grafo $B$ ). (d) Caminho gerado para o grafo da parte inferior esquerda do corpo (grafo $C$ ). . . . .	66
Figura 3.22	Ilustração da conexão entre braços×tronco e parte interna das pernas. (a) Caminhos gerados pelos três grafos principais sobrepostos na imagem de entrada (elipses indicam regiões do contorno que devem ser conectadas, para originar um contorno fechado). (b) Caminhos gerados pelos três grafos principais, conectados uns aos outros, após tratamento especial (resultando em um contorno fechado). . . . .	67
Figura 3.23	Ilustração dos pontos de conexão entre os contornos dos braços×tronco. . . . .	67
Figura 3.24	Ilustração dos pontos de conexão entre os contornos da parte interna das pernas. . . . .	68
Figura 3.25	Ilustração da segmentação por cores. Em (a), é ilustrado a região de aprendizado (retângulo verde) e busca (retângulo preto) da cor de uma determinada parte do corpo (braço direito). Em (b), <i>pixels</i> da imagem com baixa e alta similaridade em relação à cor aprendida (em vermelho e azul, respectivamente) são ilustrados. . . . .	69
Figura 3.26	Diferentes formas de setar as larguras das regiões de aprendizado (retângulos verdes) e busca (retângulos pretos) das cores predominantes. (a) Larguras estimadas diretamente a partir das distâncias entre os dois pontos que formam cada “osso”. (b) Zoom na imagem ilustrada em (a). (c) Larguras estimadas a partir dos valores esperados para aquela parte do corpo $w_i$ . . . . .	70
Figura 3.27	Ilustração da segmentação inicial usada no estágio de aprendizado do modelo de cor. (a) “Osso”, região de aprendizado e busca de uma determinada parte do corpo. (b) Imagem de entrada. (c) Segmentação inicial, usando [14]. (d) Contornos da segmentação inicial. (e) Região de aprendizado inicial. (f) Região de aprendizado final. . . . .	72



Figura 3.28	(a) Parte do corpo em análise. (b) Conversão da imagem (a) para escala de cinza. (c) Distâncias de <i>Mahalanobis</i> computadas para essa parte do corpo. (d) Magnitude do gradiente gerado a partir da imagem (b). (e) Magnitude do gradiente gerado a partir da imagem (c). (f) Magnitude do gradiente gerado pela média dos gradientes normalizados usados para gerar as imagens (d) e (e). . . . .	76
Figura 3.29	(a) Parte do corpo em análise. (b) Conversão da imagem (a) para escala de cinza. (c) Distâncias de <i>Mahalanobis</i> computadas para essa parte do corpo (valores escuros representam distâncias pequenas e valores claros o contrário). (d) Magnitude do gradiente gerado a partir da imagem (b). (e) Magnitude do gradiente gerado a partir da imagem (c). (f) Magnitude do gradiente gerado pela média dos gradientes normalizados usados para gerar as imagens (d) e (e). . . . .	77
Figura 3.30	(a) Região de aprendizado da cor predominante (retângulo verde). (b) Duas cores predominantes detectadas. (c) Mapa de distâncias de <i>Mahalanobis</i> para a cor 1 (associada a região verde na imagem (b)). (d) Mapa de distâncias de <i>Mahalanobis</i> para a cor 2 (associada a região vermelha na imagem (b)). (e) Mapa de distâncias de <i>Mahalanobis</i> final, usado para calcular o gradiente $\nabla D_i$ . . . . .	77
Figura 3.31	Ilustração das distâncias de <i>Mahalanobis</i> nas conexões entre duas partes adjacentes. Em (a), imagem RGB de entrada (usada para aprendizado e busca dos modelos de cor - regiões de busca são ilustradas por retângulos pretos). Em (b), três mapas distâncias de <i>Mahalanobis</i> para três partes do corpo (braço, antebraço e mão direita). Em (c), mapa distâncias de <i>Mahalanobis</i> gerado pela combinação dos três mapas exibidos em (b), usando o critério que retém o valor mínimo para cada ponto. . . . .	78
Figura 4.1	Ilustração do <i>ground truth</i> gerado para avaliação quantitativa e suas características. (a) Ilustração do contorno fechado. (b) Ilustração da semântica associada a cada parte do corpo. (c) Partes ocultas (parcialmente ou totalmente) são estimadas pelo usuário. (d) O contorno do cabelo não é considerado, ou seja, o usuário gera o <i>ground truth</i> esperado da cabeça. (e) Roupas folgadas são consideradas objetos que obstruem partes do corpo. . . . .	81
Figura 4.2	Resultado experimental obtido (em azul) e <i>ground truth</i> (em verde). . . . .	83
Figura 4.3	Resultado experimental obtido com e sem restrição de ângulos no cálculo da energia, respectivamente. . . . .	85

Figura 4.4	Resultado experimental obtido com e sem utilização de <i>PCA</i> na segmentação que utiliza informação de cor. (a) Distâncias de <i>Mahalanobis</i> concatenadas gerando um único mapa para todo o corpo (sem <i>PCA</i> ). (b) Resultado do modelo proposto usando a versão v6. (c) Distâncias de <i>Mahalanobis</i> concatenadas gerando um único mapa para todo o corpo (com <i>PCA</i> ). (d) Resultado do modelo proposto usando a versão v7. . . . .	85
Figura 4.5	Ilustração do número de reduções das dimensões do modelo de cor, de determinadas partes do corpo, com a utilização de <i>PCA</i> . “Ossos” ilustrados em amarelo (contínuo) possuem dimensão original (3D); em azul (tracejado) foram reduzidos para 2D; e em vermelho (pontilhado) foram reduzidos para 1D. . . . .	86
Figura 4.6	Resultado experimental obtido com e sem restrições de distâncias antropométricas, respectivamente. (a) Usando a versão v2. (b) Usando a versão v4. . . . .	86
Figura 4.7	Resultados experimentais obtidos utilizando a versão v1 do modelo proposto, considerados muito bons através de inspeção visual. . . . .	87
Figura 4.8	Resultados experimentais obtidos utilizando a versão v1 do modelo proposto, considerados aceitáveis através de inspeção visual. . . . .	88
Figura 4.9	Limitação do modelo: partes do corpo que não estão aproximadamente no mesmo plano da imagem podem produzir resultados indesejados. . . . .	89
Figura 4.10	Esqueletos informados por 24 usuários, sobrepostos nas imagens de entrada. . . . .	91
Figura 4.11	Contornos obtidos a partir dos dados de entrada informados por 24 usuários, sobrepostos nas imagens usadas (resultados obtidos exibidos em verde e <i>ground truth</i> em azul). . . . .	92
Figura 4.12	Tempo (em segundos) que os usuários levaram para informar os dados de entrada: curva de aprendizagem, assumindo-se que as imagens foram exibidas para os usuários sempre na mesma ordem (a → b → c). . . . .	93
Figura 4.13	Estudo de caso usando estimativa de pose automática ( <i>Kinect</i> ) × manual (usuário). Em (a) e (d), imagens de entrada, capturadas pela câmera do <i>Kinect</i> . Em (b) e (e), resultados (em vermelho), usando os dados de entrada capturados pelo <i>Kinect</i> ( <i>ground truth</i> exibido em azul). Em (c) e (f), resultados (em vermelho), usando os dados de entrada informados pelo usuário ( <i>ground truth</i> exibido em azul). . . . .	95
Figura 4.14	Problema associado à estimativa automática da pose (resultado em vermelho e <i>ground truth</i> em azul). Em (a), “osso” associado à uma determinada parte do corpo fora da região desejada (antebraço direito e mão direita). Em (b), <i>zoom</i> na imagem (a). . . . .	96

Figura 4.15	(a) Resultado gerado pelo modelo proposto. Modelo de esqueleto informado pelo usuário ilustrado em ciano. (b) Resultados obtidos por Freifeld e sua equipe [5]. (c) Dados de entrada utilizados no trabalho de Freifeld e sua equipe [5] . . . . .	97
Figura 4.16	(a) Resultado gerado pelo modelo proposto. (b) Resultado obtidos por Freifeld e sua equipe [5]. (c) Resultado obtidos usando <i>Grab-Cut</i> [15] com inicialização manual, utilizado como comparativo no trabalho de Freifeld e sua equipe [5]. . . . .	98
Figura 4.17	(a) Imagem de entrada. (b) Resultado gerado pelo modelo proposto. (c) Resultado obtido usando <i>Graph Cuts</i> [16] com inicialização manual. . . . .	98
Figura D.1	Imagens originais (porém reduzidas), usadas para ilustrar os resultados obtidos nesse trabalho. . . . .	109
Figura D.2	Imagens originais (porém reduzidas), usadas para ilustrar os resultados obtidos nesse trabalho. . . . .	110



## LISTA DE TABELAS

Tabela 3.1	Partes do corpo relacionadas ao modelo de esqueleto usado. Primeira coluna: índice de cada parte; segunda coluna: parte do corpo; terceira coluna: pontos que formam cada parte; quarta coluna: parâmetro usado no cálculo da estimativa da largura de cada parte. . . . .	48
Tabela 4.1	Características usadas para avaliar a energia em cada versão avaliada. . . . .	81
Tabela 4.2	Ilustração dos erros computados (em <i>pixels</i> e normalizados) para uma única imagem. Para esse exemplo, o erro médio foi 2.4442 (em <i>pixels</i> ) e 0.0053 (normalizado). O erro máximo foi 3.5881 (em <i>pixels</i> ) e 0.0079 (normalizado). Desconsiderando as partes do corpo, tratando o contorno como uma única curva, o erro global foi 2.0969 (em <i>pixels</i> ) e 0.0046 (normalizado). . . . .	84
Tabela 4.3	Erro avaliado para as diferentes versões do modelo proposto, usando uma base de dados com 277 imagens. . . . .	84
Tabela 4.4	Erro avaliado para as imagens exibidas na Figura 4.7, usando a versão v1 do modelo proposto. . . . .	88
Tabela 4.5	Erro avaliado para as imagens exibidas na Figura 4.8, usando a versão v1 do modelo proposto. . . . .	89
Tabela 4.6	Tempo de processamento (em segundos) avaliado para 3 imagens da base de dados usada, para todas as versões do modelo proposto. . . . .	90
Tabela 4.7	Altura média estimada (em <i>pixels</i> ) por 24 usuários para as 3 pessoas contidas nas 3 imagens exibidas na Figura 4.10 e altura armazenada no <i>ground truth</i> . . . . .	91
Tabela 4.8	Erro avaliado para 3 imagens (ilustradas na Figura 4.10), a partir dos dados de entrada fornecidos por 24 usuários. . . . .	92
Tabela 4.9	Tempo médio (em segundos) que os usuários levaram para informar os dados de entrada para cada imagem e média global. . . . .	93
Tabela 4.10	Erro avaliado para 8 imagens, para dados de entrada adquiridos de forma automática ( <i>Kinect</i> ) e manual (informados pelo usuário). . . . .	94
Tabela C.1	Parâmetros usados no modelo proposto. . . . .	107



## LISTA DE ABREVIATURAS

GMM – *Gaussian Mixture Models* (Modelo de Mistura de Gaussianas)

IQP – *Integer Quadratic Programming*

MCMC – *Markov Chain Monte Carlo* (Cadeias de Markov Monte Carlo)

MGM – *Multivariate Gaussian Model* (Modelo Gaussiano Multivariado)

PCA – *Principal Component Analysis* (Análise de Componentes Principais)

SCAPE – *Shape Completion and Animation of People*





# SUMÁRIO

1. Introdução	27
1.1 Motivação . . . . .	29
1.2 Objetivos . . . . .	31
2. Segmentação de imagens	33
2.1 Conceitos básicos . . . . .	33
2.2 Trabalhos Relacionados . . . . .	34
2.3 Contexto desse trabalho no estado-da-arte . . . . .	42
3. Modelo Proposto	45
3.1 Modelo de esqueleto: dados de entrada . . . . .	46
3.2 Geração do grafo . . . . .	49
3.2.1 Definição do Grafo . . . . .	51
3.2.2 Conectando os grafos e casos especiais . . . . .	53
3.2.3 Mapa de Energias convencional . . . . .	58
3.2.4 Mapa de Energias - com restrições antropométricas . . . . .	60
3.2.5 Programação Dinâmica . . . . .	64
3.3 Inserindo informação de cor no modelo . . . . .	68
3.3.1 Aprendendo o modelo de cor . . . . .	69
3.3.2 Aprendendo o modelo de cor com a utilização de <i>PCA - Principal Component Analysis</i> . . . . .	72
3.3.3 Confrontando o modelo de cor . . . . .	74
3.3.4 Mapa de Energias com informação de cor . . . . .	75
4. Resultados Experimentais	79
4.1 Criação de <i>ground truth</i> para análise quantitativa . . . . .	79
4.2 Características usadas no modelo proposto . . . . .	81
4.3 Sensibilidade do modelo . . . . .	90
4.4 Esqueleto de entrada adquirido de forma manual × automática . . . . .	93
4.5 Comparação qualitativa com estado-da-arte . . . . .	95

5. Considerações Finais e Trabalhos Futuros	99
A. Apêndice - Trabalhos publicados	103
A.1 Artigos completos publicados em periódicos . . . . .	103
A.2 Artigos completos publicados em anais de congressos . . . . .	103
A.3 Artigo aceito para publicação . . . . .	104
B. Apêndice - Prêmios recebidos	105
C. Apêndice - Lista de parâmetros e valores padrão	107
D. Anexo - Base de dados	109
Referências Bibliográficas	111

## 1. Introdução

Devido a um crescimento tecnológico bastante acelerado, existem atualmente diversos sistemas baseados em técnicas de processamento de imagens digitais ou visão computacional que visam detectar, identificar, rastrear, monitorar e compreender o comportamento dos mais diversos tipos de objetos, com a utilização de uma câmera de vídeo (ou múltiplas câmeras) e um computador (ou múltiplos computadores). Tal avanço tecnológico possibilitou fácil acesso a diversos meios de aquisição de imagens digitais, como por exemplo, câmeras de vídeo ou câmeras fotográficas com baixo custo, assim como equipamentos de visualização (monitores, celulares, projetores, impressoras, etc) e processamento (computadores, celulares, câmeras digitais, etc) bastante acessíveis. Com isso, o interesse na manipulação de imagens digitais (edição, análise, processamento, etc) tem sido objeto de pesquisa e atraído grande atenção da comunidade científica nas últimas décadas. Conforme Gonzales e Woods [17], o interesse em métodos de processamento de imagens digitais decorre de duas áreas principais de aplicação: melhoria de informação visual para a interpretação humana e o processamento de dados de cenas para percepção automática através de máquinas (também relacionado hoje em dia com o termo “visão computacional”). O fácil acesso à tecnologia também pode ser considerado hoje em dia um grande motivador para aumentar o interesse nessa área de pesquisa.

A tecnologia relacionada ao processamento de imagens digitais e visão computacional pode ser aplicada nas mais variadas áreas do conhecimento (Física, Biologia, Matemática, Medicina, Astronomia, Geologia, Psicologia, etc). Problemas típicos relacionados à visão computacional que normalmente utilizam técnicas de processamento de imagens são: detecção e reconhecimento automático de indivíduos/objetos, automação industrial (desenvolvimento e inspeção de objetos), aplicações militares, aplicações biométricas (reconhecimento automático de impressões digitais), monitoramento automático de tráfego, pessoas ou multidões, aplicações na medicina, interpretação de imagens aéreas ou capturadas via satélite, dentre várias outras.

O processamento de imagens digitais abrange ampla escala de *hardwares*, *softwares* e fundamentos teóricos. Etapas fundamentais relacionadas ao processamento de imagens vão desde a aquisição da imagem digital propriamente dita, passando por outras etapas, como por exemplo, pré-processamento, segmentação, representação e descrição dos dados, reconhecimento e interpretação dos resultados [17]. Obviamente, todas essas etapas não precisam necessariamente estar envolvidas em todas as aplicações de processamento de imagens ou visão computacional, estando isso diretamente relacionado com sua aplicação.

Conforme relatado por Gonzalez e Wood [17], geralmente o primeiro passo em análise de imagens é a segmentação da imagem. A segmentação subdivide uma imagem em suas partes ou objetos constituintes. O nível até o qual essa subdivisão deve ser realizada depende do problema sendo resolvido. Ou seja, a segmentação deve parar quando os objetos de interesse na aplicação tiverem sido isolados. Em geral, a segmentação autônoma é uma das tarefas mais difíceis em processamento

de imagens. Esse passo determina o eventual sucesso ou fracasso na análise. De fato, a segmentação efetiva pode aumentar a probabilidade de sucesso nas etapas posteriores do processamento (se houverem).

Os algoritmos de segmentação de imagens são geralmente baseados em uma das seguintes propriedades básicas: descontinuidade e similaridade [17]. Na primeira categoria, a abordagem é particionar a imagem baseado em mudanças bruscas de valores, que podem indicar, por exemplo, a intensidade de nível de cinza em um ponto da imagem. As principais abordagens da segunda categoria baseiam-se em limiarização, crescimento de regiões, e divisão e fusão de regiões. Ainda que grandes avanços tenham sido feitos nessa área, conforme relatado no trabalho de McGuinness [18], não há uma técnica padrão para selecionar um determinado algoritmo a ser usado em uma determinada aplicação. Essa deficiência está associada com a ambiguidade inerente em se determinar o propósito e escopo da segmentação.

Em trabalhos recentes encontrados na literatura existe um grande interesse em se determinar o contorno, ou forma (2D ou 3D) de objetos conhecidos, com aplicação em diversas áreas. Algumas abordagens utilizam novas propriedades para segmentar objetos, além da descontinuidade e similaridade de suas partes, como por exemplo, forma dos objetos, simetria, perspectiva da câmera, entre outras [19]. Uma área que tem despertado grande interesse da comunidade científica é a estimativa automática ou semi-automática da pose e/ou forma (2D ou 3D) [20, 21], que está intimamente relacionada com a segmentação da pessoa em uma imagem ou sequência de imagens. Guan e colaboradores [11] relatam que enquanto a estimativa de pose 3D de seres humanos a partir de câmeras monoculares não calibradas tem recebido grande atenção nos últimos anos, há pouquíssima pesquisa na área de segmentação de pessoas (extração automática da forma do corpo) em imagens. A natureza articulada e deformável do corpo humano faz com que a estimativa de sua forma seja uma tarefa extremamente desafiadora, podendo ser utilizada em diversas aplicações que variam desde computação gráfica (criação de personagens virtuais a partir de imagens, por exemplo) à vigilância automática baseada em visão computacional. Entretanto, como relatado no trabalho de Zhou e sua equipe [2], apenas algumas técnicas para estimativa detalhada da forma humana, a partir de imagens, são encontradas na literatura nos tempos atuais.

A segmentação automática de pessoas em imagens, com a utilização de técnicas de processamento de imagens e visão computacional, ainda é um problema em aberto, devido à influências de inúmeros fatores do mundo real, como por exemplo, iluminação da cena, ruídos na imagem, oclusão de membros da pessoa (parcial ou total), muita similaridade com o fundo da imagem ou perda de informação de profundidade relacionado quando uma cena é capturada em uma imagem bidimensional [3] assim como a outros fatores associados à dinâmica do ser humano (grande variabilidade de poses, formas distintas do corpo, roupas, etc). Entretanto, alguns trabalhos encontrados recentemente na literatura demonstram que essa área de pesquisa tem se tornado foco de atenção nos últimos anos, podendo ser aplicado em diversas áreas [2, 5, 11], como por exemplo, estimativa de pose e forma da pessoa (2D ou 3D), edição/manipulação da imagem, entre outras. De uma forma geral, a grande maioria dos trabalhos que utilizam algum método para estimativa de pose ou forma

de um indivíduo, ou manipulação da imagem de uma pessoa em uma imagem digital, utiliza alguma técnica de segmentação, seja ela automática, semi-automática, ou manual.

## 1.1 Motivação

A segmentação de pessoas em sequências de imagens (vídeo), ou trabalhos que estejam envolvidos com isso, são temas que têm atraído grande atenção da comunidade científica nos últimos anos [22, 23], podendo ser utilizada em diversas áreas, como por exemplo, na detecção, reconhecimento e monitoramento de pessoas, grupos de pessoas, entre outras. Uma vantagem na utilização de sequências de imagens é a capacidade de se poder estimar o movimento dos objetos na cena, a partir de informação temporal, que pode facilitar sua segmentação se o objeto estiver em movimento. Outra vantagem, em se tratando de múltiplas imagens de um mesmo objeto/indivíduo capturadas de diferentes locais, é que pode-se tentar resolver problemas de oclusão, como por exemplo, segmentar partes de uma pessoa/objeto que estão mais visíveis em uma imagem do que em outra. Porém, em se tratando de imagens estáticas, onde não há informação de profundidade, tempo e tão pouco de movimento, há uma grande dificuldade em se segmentar objetos complexos (pessoas, por exemplo) de forma automática ou semi-automática. Alguns trabalhos, relacionados a seguir, utilizam técnicas semi-automáticas ou manuais para realizar tal tarefa, podendo gerar informação de entrada para outros fins, como por exemplo, edição/manipulação de imagens, estimativa 3D de pose ou forma das pessoas, entre outros.

Recentemente, Hasler e colaboradores [1] propuseram um método para estimativa 3D da forma do corpo de uma pessoa em uma única imagem, múltiplas imagens ou até mesmo em pinturas digitalizadas. Nesse trabalho, os autores utilizam como dados de entrada a silhueta da pessoa, extraída da imagem, e marcações informadas através de usuário (posição das mãos e pés). A partir de um modelo 3D inicial, o método é capaz de computar correspondências entre sua forma e a silhueta extraída da imagem. Os parâmetros da pose e forma do corpo são determinados de forma a coincidir com a silhueta observada. Os autores relatam que devido a condições de iluminação e qualidade das imagens utilizadas, a maioria dos seus resultados foram gerados a partir de uma segmentação manual das pessoas nas imagens. Entretanto, em algumas imagens cuja a segmentação é considerada simples, foi utilizado um método semi-automático de segmentação de imagens (*Grab-Cut*, [15]). Uma característica do modelo proposto [1] é que, ao final da estimativa 3D da forma do corpo da pessoa, o método é capaz de estimar informações biométricas relacionadas à mesma, como por exemplo, peso e altura da pessoa, utilizando um simples regressor linear treinado a partir de uma base de dados. A Figura 1.1 exibe o resultado desse método para um exemplo apresentado.

Zhou e sua equipe [2] propõem um método para edição realista da forma do corpo de pessoas em imagens, a partir de uma segmentação manual. Um fator motivador desse trabalho é que ferramentas usadas para edição da forma do corpo de pessoas em imagens são normalmente tarefas realizadas em baixo nível, e usualmente limitadas à modificações locais, como por exemplo, remoções de imperfeições na pele, pequenas rugas, etc. Nesse trabalho [2], é assumido que uma modificação

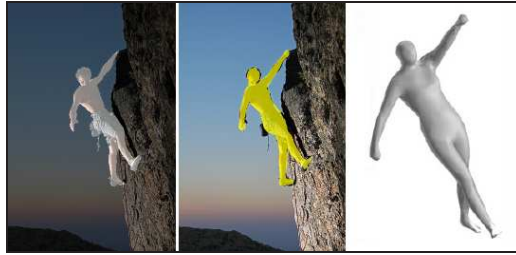


Figura 1.1: Ilustração do método proposto por Hasler e colaboradores [1]. Imagem da esquerda: silhueta da pessoa, segmentada manualmente. Imagem central e direita: resultado da estimativa de forma 3D.

mais radical da forma do corpo humano, como por exemplo, engordar, emagrecer ou até mesmo tornar a pessoa com uma aparência mais forte, requer uma edição que seja consistente globalmente, demandando habilidades que até mesmo para pessoas especialistas possa se tornar uma tarefa bastante tediosa e demorada. Uma possível solução, mencionada pelos autores, seria atuar em um nível mais elevado, operando sobre as partes do corpo (por exemplo, braços, pernas, tronco, etc), modificando-as em relação à sua parte no esqueleto, o que também é uma tarefa bastante desafiadora, devido às seguintes razões: (i) uma deformação realista requer efeitos de deformações espacialmente variadas para cada parte do indivíduo (não se tratando de uma simples escala ao longo do eixo do esqueleto de cada parte do corpo); e (ii) não é muito clara a forma que se deve fazer modificação em partes individuais, de forma que estas estejam coerentes globalmente.

Para alcançar tal objetivo, em um primeiro momento, é utilizado um modelo 3D de uma pessoa (3D *morphable model*), o qual é sobreposto na imagem de entrada, de forma que haja um encaixe perfeito entre o modelo 3D e a pessoa na imagem. Esse primeiro passo é feito com intervenção manual do usuário. O usuário pode, nessa etapa, esculpir o modelo 3D, reajustando alguns parâmetros, de forma que o mesmo se encaixe ao corpo da pessoa na imagem, como por exemplo, largura ou altura. Por fim, a forma da pessoa na imagem pode ser deformada, de modo que seus contornos sejam o mais semelhantes possível com as modificações feitas no modelo 3D. A Figura 1.2 ilustra parte desse processo.



Figura 1.2: Ilustração do método proposto por Zhou e sua equipe [2]. Imagem da esquerda: imagem original. Imagem central: modelo 3D sobreposto na imagem original. Imagem à direita: resultado após a manipulação da imagem.

Os trabalhos apresentados na Seção 1.1, assim como os trabalhos exibidos na Seção 2.2, serviram como motivação para o desenvolvimento do modelo de segmentação de pessoas em imagens estáticas proposto nessa tese, devido ao interesse que esse tema tem recebido nos últimos anos e pelo fato

de ser uma tarefa bastante desafiadora.

## 1.2 Objetivos

Esse trabalho teve como objetivo geral inicial investigar métodos que pudessem auxiliar no processo de segmentação (automática ou semi-automática) de pessoas em imagens estáticas. Atingido tal objetivo, está sendo proposta nessa tese uma solução viável para a resolução do problema (segmentação de pessoas em imagens), que leve em consideração características antropométricas da forma humana (baseada em um modelo de esqueleto), além das características oriundas da imagem sendo analisada (cor, bordas, luminosidade, etc).

No decorrer do curso de doutorado foi proposta uma técnica para segmentação e estimativa automática da pose de pessoas em imagens estáticas, publicado em uma conferência da área [3]. Nesse trabalho [3], a segmentação da pessoa é realizada sem intervenção manual, inicializada a partir de um detector de faces automático [24], onde o objetivo inicial é encontrar cores predominantes em regiões específicas, estimadas a partir de parâmetros antropométricos. O resultado final desse trabalho é um método para estimativa de poses de pessoas em imagens estáticas (basicamente da parte superior do corpo - tronco e membros superiores). A Figura 1.3 ilustra parte desse processo.

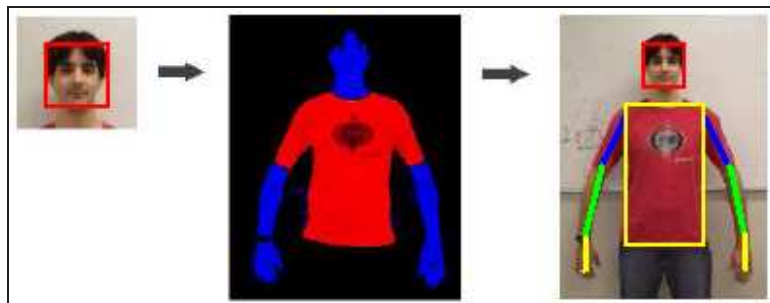


Figura 1.3: Ilustração do método proposto por Jacques Junior e colaboradores [3]. Imagem da esquerda: face detectada de maneira automática. Imagem central: segmentação de *pixels* com tons de pele e região do tronco. Imagem à direita: estimativa da pose da parte superior do corpo.

Entretanto, devido a inúmeros fatores, que fazem com que não seja trivial a resolução desse problema de forma automática (tanto como o de segmentação como o de estimativa de poses), pretendeu-se também investigar vantagens/desvantagens de métodos que permitam intervenção com o usuário, assim como estender o trabalho proposto em [3] para segmentar o corpo todo (ao invés de somente a parte superior do corpo). O resultado final desse processo investigativo resultou no modelo proposto nessa tese, para segmentação de pessoas em imagens estáticas baseada em esqueleto, descrito em detalhes no Capítulo 3. No Capítulo 2 são apresentados alguns conceitos fundamentais sobre processamento de imagens, assim como trabalhos relacionados à segmentação de pessoas em imagens, considerados estado-da-arte. Resultados experimentais e estudos de caso (análise quantitativa, sensibilidade do modelo proposto, aquisição automática dos dados de entrada e comparação com estado-da-arte) são apresentados no Capítulo 4. Por fim, considerações finais e sugestões para trabalhos futuros são apresentadas no Capítulo 5. Trabalhos publicados e prêmios

recebidos durante o doutorado são listados nos Apêndices A e B, respectivamente. No Apêndice C é apresentada uma lista detalhada dos principais parâmetros usados no modelo proposto, assim como os valores padrão adotados. Imagens usadas neste trabalho para ilustrar algum resultado do modelo proposto são ilustradas no seu formato original (em cores e sem cortes, porém redimensionadas) no Apêndice D.



## 2. Segmentação de imagens

Este capítulo visa introduzir alguns conceitos fundamentais sobre segmentação de imagens. O objetivo é apresentar algumas técnicas conhecidas de segmentação de imagens que podem, ou poderiam ser utilizadas para segmentação de pessoas em imagens estáticas. Nesse capítulo também são apresentados alguns trabalhos considerados estado-da-arte no que tange segmentação de pessoas em imagens estáticas.

### 2.1 Conceitos básicos

Conforme mencionado anteriormente, os algoritmos de segmentação de imagens são geralmente baseados em uma das seguintes propriedades básicas: descontinuidade e similaridade [17]. Na categoria de algoritmos baseados em descontinuidades, a abordagem é particionar a imagem baseado em mudanças bruscas de valores, que podem indicar, por exemplo, a intensidade de nível de cinza em um ponto da imagem. De uma forma geral, os três tipos de descontinuidades mais utilizados nesse processo de segmentação são: detecção de pontos, linhas e bordas. Na prática, a maneira mais comum de procurar por descontinuidades é através de uma varredura na imagem com a utilização de uma determinada máscara (ou filtro), que irá salientar alguma característica desejada na imagem (como pontos, linhas ou contornos, por exemplo).

As principais abordagens da segunda categoria (similaridade) baseiam-se em limiarização, crescimento de regiões, e divisão e fusão de regiões. A técnica de limiarização é uma das mais importantes abordagens para a segmentação de imagens. Suponha que seja construído o histograma de níveis de cinza de uma determinada imagem  $f(x,y)$ , composta por objetos iluminados sobre um fundo escuro, de maneira que os *pixels* do objeto e os do fundo tenham seus níveis de cinza agrupados em dois grupos dominantes. Uma maneira óbvia de extrair os objetos do fundo é através da seleção de um limiar  $T$  que separa os dois grupos. Então, cada *pixel*  $(x,y)$  tal que  $f(x,y) > T$  é denominado um ponto do objeto; caso contrário, o ponto é denominado parte do fundo. Entretanto, na grande maioria das aplicações envolvendo processamento de imagens (em condições não controladas), dificilmente o histograma terá somente dois grupos dominantes, podendo haver diversos picos, associados aos diversos objetos contidos na imagem, fazendo com que não seja considerada uma tarefa trivial a escolha de múltiplos valores de limiar para isolar, de forma efetiva, um objeto de interesse, principalmente no que tange a escolha de valores de limiar de forma automática.

A segmentação orientada a regiões baseia-se nos conceitos de crescimento, divisão e fusão de regiões. O crescimento de regiões é um procedimento que agrupa *pixels* ou sub-regiões em regiões maiores. A mais simples dessas abordagens é a agregação de *pixels*, que começa com um conjunto de pontos denominados “semente” e, a partir deles, a região cresce, de maneira que *pixels* que possuem propriedades similares (como níveis de cinza, textura ou cor) são anexados às essas “sementes”. Dois problemas imediatos, relacionado à essa abordagem, são a seleção de “sementes” que representam

adequadamente as regiões de interesse, bem como a seleção de propriedades apropriadas para a inclusão de pontos nas várias regiões durante o processo de crescimento. A seleção de um ou mais pontos iniciais pode frequentemente se basear na natureza do problema, assim como, se as “sementes” são adquiridas de maneira automática, semi-automática ou manual. Outros fatores que devem ser levados em consideração são os critérios de similaridade usados, que estão intimamente relacionados com o tipo de dado em questão (imagem monocromática, colorida, térmica, etc) e o estabelecimento de uma condição de parada (o crescimento de uma região deveria parar quando uma determinada condição for satisfeita). O procedimento utilizado em abordagens baseadas em crescimento de regiões parte de um conjunto de “sementes”. Uma alternativa à essa abordagem seria subdividir a imagem de entrada em um conjunto de regiões arbitrárias e disjuntas, e então realizar a divisão e/ou fusão das regiões na tentativa de satisfazer alguma condição pré-estabelecida.

O conceito de segmentação de uma imagem em descontinuidades ou em similaridade de valores pode ser aplicado tanto em imagens estáticas como em imagens dinâmicas (que variam com o tempo). Nesse último caso, porém, o movimento pode frequentemente ser usado como uma pista poderosa para melhorar a performance dos algoritmos de segmentação. Em aplicações de imageamento, o movimento é originado a partir de um deslocamento relativo entre o sistema de coordenadas do sensor e a cena sendo observada, como em aplicações de robótica, navegação autônoma e análise dinâmica de cenas (podendo ser considerado tanto no domínio espacial quanto no domínio das frequências).

## 2.2 Trabalhos Relacionados

Nesta seção são apresentados alguns trabalhos, considerados estado-da-arte no que tange a segmentação automática ou semi-automática de pessoas em imagens estáticas. Alguns trabalhos não associados diretamente à segmentação de pessoas também são relatados, por serem considerados importantes ou porque poderiam ser utilizados na construção de um novo modelo de segmentação de pessoas em imagens estáticas.

Alguns métodos propostos para segmentação automática de pessoas em imagens inicializam seus modelos a partir de alguma pré-determinada informação, como por exemplo, região da face [3] (usando um detector automático de faces), região ou pose estimada da pessoa (detecção automática de pessoas em imagens, ou estimativa de pose), como em [5], ou a partir da região do torso [6] (parte superior da pessoa), por exemplo, entre outras formas. Por outro lado, alguns trabalhos se propõem a detectar e segmentar as pessoas de forma simultânea, como em [9], ou [7], por exemplo. Além disso, métodos semi-automáticos podem ser uma alternativa para a resolução do problema, como em [4, 11, 15], por exemplo.

Hornung e sua equipe [4] apresentam um método para animar personagens em imagens (fotos ou pinturas digitalizadas), com a utilização de movimentos capturados do mundo real (*motion capture*). Dada a imagem de uma pessoa, ou personagem similar a um ser humano, o método faz uma estimativa do modelo de câmera perspectiva usado na composição dessa imagem, assim

como da pose 3D do personagem contido na mesma, e transfere o movimento de um esqueleto 3D para o personagem na imagem, gerando uma impressão de movimento realístico. Nesse trabalho é considerado um modelo genérico de esqueleto 3D de um personagem virtual, e o usuário informa pontos de correspondência entre esse modelo 3D e o personagem na imagem 2D. A extração do contorno do personagem é feita de maneira semi-automática, na qual um conjunto de *templates* de formas (*shape templates*), organizados hierarquicamente, é encaixado às partes do corpo desse personagem. Dado uma imagem de entrada, um determinado *template* de forma é selecionado automaticamente, explorando-se a pose que melhor se encaixa ao personagem, associado também ao modelo de câmera perspectiva estimado. Então esse *template* é deformado com a utilização de um algoritmo que preserva algumas características desejadas, como por exemplo, forma e proporção (denominado *As Rigid As Possible*, ou “tão rígido quando possível”), objetivando adequar o *template* inicial ao modelo de câmera estimado. Posteriormente, um algoritmo de segmentação baseado em informações de contornos (*snakes 2D – Active Shape Models*) é usado para aprimorar o encaixe entre o *template* ao contorno do personagem na imagem. Por fim, regiões do contorno não condizentes com o contorno esperado, são ajustadas manualmente com auxílio do usuário. Dessa forma o personagem é segmentado, e animado com dados de movimentos capturados. A Figura 2.1 ilustra o resultado desse trabalho. Regiões ocultas do personagem, como do fundo de cena, são reconstruídas com uma técnica de síntese de texturas.

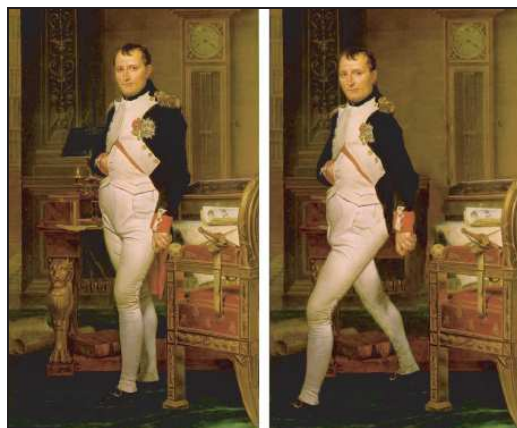


Figura 2.1: Ilustração do método proposto por Hornung e sua equipe [4]. Imagem original é exibida à esquerda. Resultado obtido é exibido à direita.

No trabalho de Freifeld e equipe [5] é proposto um modelo 2D da silhueta humana que pode ser utilizado para segmentação automática de pessoas em imagens. O modelo é construído a partir de uma base de dados de aprendizado (SCAPE – *Shape Completion and Animation of People*, [25]), que representa detalhadamente formas e poses do corpo humano, de maneira natural, assim como essas variam em uma população. Uma característica do modelo é que não se trata de um contorno simples de uma forma, visto que inclui informação semântica, ou seja, as partes do corpo são representadas no contorno por cores (ou índices) diferentes, como ilustrado na Figura 2.2 (à esquerda), o que torna possível que duas partes do corpo fiquem sobrepostas (por exemplo, o braço na frente do tronco), mantendo uma conectividade coerente do contorno. A inicialização do método é feita

automaticamente com a utilização de um detector automático de pessoas e pose [26]. O modelo de deformação do contorno é composto por três partes: variação de forma, mudanças do ponto de vista (de câmera) e rotação das partes. O resultado é um modelo 2D articulado e parametrizável. A pose e a forma estimadas são refinadas com a utilização de uma função de custo que segmenta a cena em objeto e fundo (*foreground* e *background*, respectivamente) baseada em um modelo de segmentação semi-automático (*Grab-Cut*, [15]). Um resultado desse trabalho é ilustrado com auxílio da Figura 2.2.

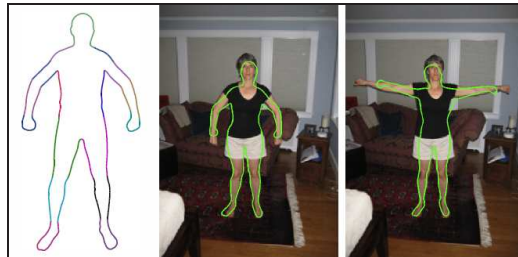


Figura 2.2: Ilustração do método proposto por Freifeld e equipe [5]. Imagem da esquerda: representação 2D da pessoa contendo informação semântica. Imagem central e direita: resultado da estimativa de pose e segmentação.

Hu [6] e sua equipe propõem um método para segmentação automática de roupas de pessoas em imagens estáticas sem qualquer modelo pré-definido de roupa. Tal abordagem não é usada especificamente para a segmentação da pessoa propriamente dita, porém pode ser usada para auxiliar nesse processo. Nessa abordagem, as roupas são extraídas usando um modelo bastante conhecido para segmentação de imagens, *Graph Cuts* [16], onde as “sementes” do *foreground* (objeto a ser segmentado) e do *background* (fundo da cena) são obtidas de forma automática. As “sementes” do *foreground* são obtidas com a utilização de um detector de tronco, baseado na segmentação de cores dominantes. As “sementes” do *background* são estimadas com base na Triangularização de *Delaunay* [27]. Após obter as “sementes” do *foreground* e *background*, a distribuição de cores de ambos são modeladas com a utilização de Misturas de Gaussianas (GMM - *Gaussian Mixture Models*).

Os autores utilizam nesse trabalho [6] um modelo probabilístico para segmentar *pixels* com tons de pele, criado a partir dos *pixels* da região da face. Relatam que remover *pixels* com tons de pele do objeto e associá-los ao fundo da cena gera resultados melhores, pois esses podem influenciar na distribuição de valores tanto do fundo da cena como do objeto, gerando um resultado de segmentação não muito acurado. Partem da hipótese que os tons de pele de um indivíduo são similares aos tons de pele de sua face. Dessa forma, utilizam um algoritmo (*k-means* [28]) para segmentar a região da face e assim, segmentam os *pixels* de tons de pele, assumindo que esses *pixels* são usualmente dominantes na região da face. A Figura 2.3 ilustra um resultado desse trabalho.

No trabalho de Mori e sua equipe [7] é proposta uma abordagem onde o reconhecimento é guiado pela segmentação. Os autores consideram problemático tentar detectar, de forma automática e individual, as partes do corpo de uma pessoa em uma imagem. Ilustram esse problema com a situação onde é aplicado um *zoom* em uma imagem, na região do braço de uma pessoa, por exemplo,



Figura 2.3: Ilustração do método proposto por Hu [6] e sua equipe. Imagem da esquerda: imagem de entrada com o tronco detectado (em vermelho) e limites da região do fundo estimada (em azul). Imagem à direita: resultado da segmentação.

e essa região pode se assemelhar à imagem de um gramado, ou o tronco de uma árvore. Porém, em um contexto global (com uma mão, um ombro, um torso, etc), as partes podem fazer mais sentido, ou seja, muitas características de baixo nível agregam informações apenas quando consideradas dentro de seu contexto. Nesse trabalho, o modelo parte de um conjunto de características de baixo nível, com informações independentes de contexto, que usualmente representam partes salientes, possuindo informações suficientes em si mesmas para criar uma configuração parcial (por exemplo, “se isso é um cotovelo e aquilo é um torso, então aquele deve ser o braço”). Dessa forma, existe um problema combinatorial para se determinar quais partes devem ser postas juntas para originar uma configuração parcial. Os autores utilizam algumas restrições globais, como por exemplo, escalas relativas, localização e cores, para remover combinações impossíveis. O restante da configuração é realizado através de uma busca pelas partes restantes.

Uma característica de baixo nível utilizada nesse trabalho [7] é o resultado de um método para estimativa de contornos proposto por Martin e colaboradores [8], que combina informação de brilho e de textura para remover contornos aglomerados. Outra característica usada é adquirida com a utilização de um algoritmo de segmentação denominado *Normalized Cuts* [29], com o objetivo de agrupar em regiões *pixels* semelhantes (relatam que muitas partes salientes do corpo “saltam aos olhos”, em regiões individuais). Também é usada uma forma de segmentação que gera como resultado uma imagem com pequenos segmentos, denominados *superpixels* [30], pois tem se mostrado uma abordagem que retém visualmente todas as estruturas de uma imagem, além de reduzir drasticamente a etapa de análises (de  $400k$  de *pixels* para 200 *superpixels*, por exemplo). Outras características de baixo nível como iluminação (*shading*) e foco também são usadas. Os autores utilizam uma base de dados de treinamento (definida de forma empírica por especialistas) para criar um descritor de iluminação (*shading*) para membros do corpo. Assumem que os membros podem ser associados a cilindros (onde algumas características de iluminação podem ser salientes) provendo alguma noção de 3D. Em relação ao foco da imagem, partem da hipótese que o fundo da cena normalmente perde informação de textura ou foco (característica encontrada nas imagens de jogadores de *baseball* usadas). Dessa forma, os autores criam um modelo baseado em regras, que utiliza uma busca exaustiva para detectar membros e torso de uma pessoa, com a utilização de descritores locais e restrições globais. Uma consideração importante dos autores é sobre o desafio em se estabelecer um valor para a pose final estimada (como um percentual de acerto, ou *score*,

por exemplo). A Figura 2.4 ilustra o resultado desse trabalho. O método de estimativa de contornos usado nesse trabalho é ilustrado na Figura 2.5(d). As imagens ilustradas na Figura 2.5(e) e Figura 2.5(f) ilustram resultados dos algoritmos de segmentação usados para extrair características de baixo nível da imagem, *Normalized Cuts* e *superpixels*, respectivamente.

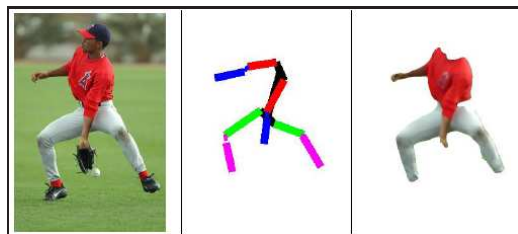


Figura 2.4: Ilustração do método proposto por Mori e sua equipe [7]. Imagem da esquerda: imagem de entrada. Ao centro, esqueleto estimado. À direita, resultado da segmentação associada à pessoa.

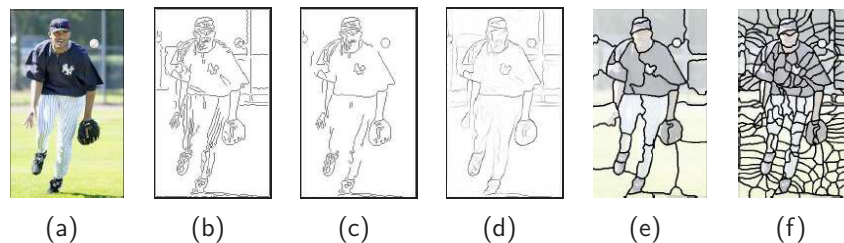


Figura 2.5: (a) Imagem de entrada. (b-c) resultado detector de bordas Canny, com duas escalas distintas. (d) Mapa de probabilidade de contornos, extraído usando abordagem proposta por Martin e equipe [8]. (e) Resultado do algoritmo *Normalized Cuts*. (f) Mapa de *superpixels* gerado para essa imagem.

Lin e sua equipe [9] apresentam um modelo hierárquico baseado em *template-matching* para detecção e segmentação de pessoas em imagens. *Template-matching* é uma técnica bastante conhecida em processamento de sinais, porém é bastante sensível à mudanças de escala e rotação, também como é considerada computacionalmente cara. Os autores salientam que a detecção de uma pessoa é um problema fundamental em análise de imagens, pois pode prover inicialização para técnicas de segmentação, sistemas de rastreamento e identificação de indivíduos. Também classificam as abordagens para detecção de pessoas em imagens em duas categorias: baseadas em forma (*shape-based*) e baseadas em objeto de *foreground* (*blob-based*). As formas, em técnicas baseadas em forma, podem ser modeladas como segmentos de curvas locais, ou diretamente com um modelo hierárquico global de forma, ou então representadas por descritores globais ou locais. Abordagens baseadas em forma possuem a vantagem de não necessitar de técnicas de subtração de fundo (*background subtraction*), porém têm a necessidade de “varrer” toda a imagem, para encontrar o melhor *matching*, podendo gerar diversos alarmes falsos. Por outro lado, abordagens baseadas em objeto de *foreground* (*blob-based*) são computacionalmente mais eficientes, porém, seus resultados dependem de técnicas de subtração de fundo.



A abordagem proposta em [9] utiliza detectores de partes locais e globais, utilizando *template-matching*, através da decomposição de modelos globais de forma para a construção de uma estrutura em forma de árvore de *templates* de forma. Características de baixo nível, como bordas, são usadas para fazer o *matching* entre uma determinada região da imagem e um determinado *template*, gerando um conjunto de hipóteses de pessoas detectadas. A segmentação e estimativa da pose são obtidas de forma automática com a utilização de síntese de partes detectadas, com a utilização de um modelo *Bayesiano*. A Figura 2.6 ilustra uma árvore de *templates* de forma usado neste trabalho. A Figura 2.7 ilustra resultados desse trabalho.

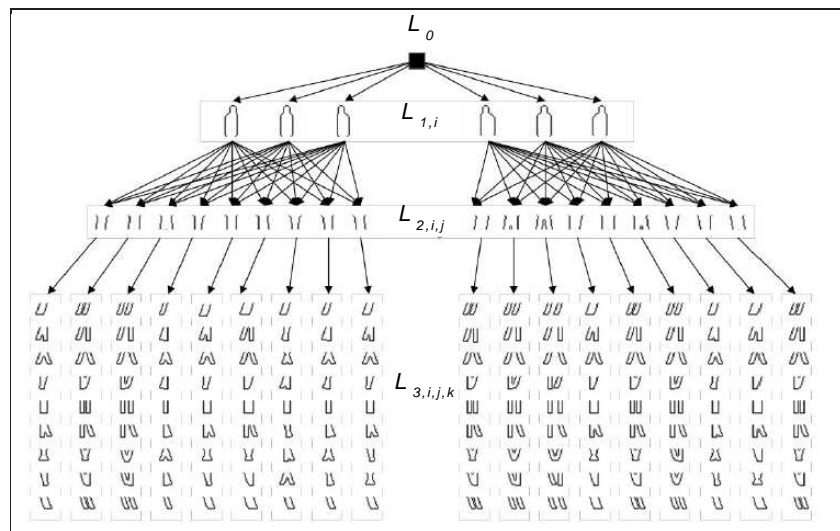


Figura 2.6: *Templates* de forma caracterizados em uma representação de árvore, propostos no trabalho de Lin e sua equipe [9].

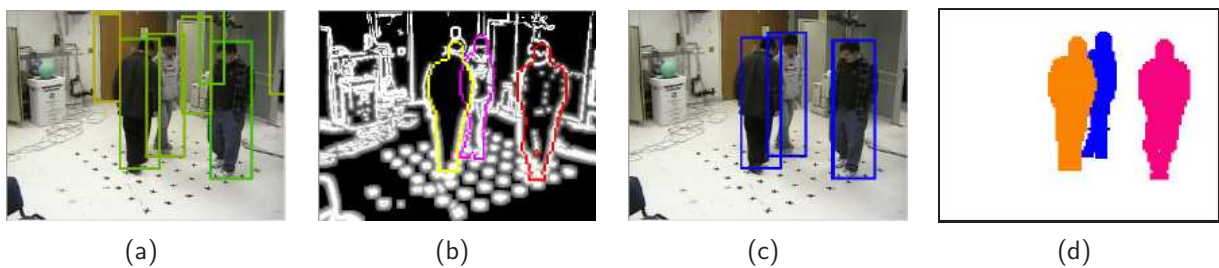


Figura 2.7: Ilustração do resultado do trabalho de Lin e sua equipe [9]. Sendo (a), o conjunto inicial de hipóteses detectadas; (b) resultado inicial da segmentação; (c) resultado final da detecção; (d) resultado final da segmentação.

Similarmente, Gravila [31] propõem um modelo hierárquico usando *template-matching*, representado em uma estrutura de árvore, combinado com uma abordagem *Bayesiana*. Entretanto, os objetos são descritos a partir de um conjunto de treinamento, baseados em forma ou exemplos, que cubram um determinado conjunto de aparências devido à transformações geométricas (rotação e escala, por exemplo) e variação intra-classe (diferentes pedestres, poses, etc). Como critério de

similaridade entre exemplos, nesse trabalho utiliza-se a distância de *Chamfer*, baseada na orientação das bordas, extraídas da imagem.

No trabalho de Su e colaboradores [10], é proposta uma técnica que utiliza um par de imagens, com e sem *flash*, de um mesmo objeto para segmentá-lo, a qual os autores chamam de *Flash-cut*. Essa técnica baseia-se na hipótese de que apenas o objeto de interesse é significativamente influenciado pelo *flash* e que as mudanças geradas no fundo da cena (*background*) são menos significativas, podendo ser segmentadas facilmente (se o *background* estiver distante). A técnica suporta uma variação pequena de movimento do *background* assim como do *foreground* (objeto em questão). Uma desvantagem dessa abordagem é que não pode ser utilizada em imagens genéricas, encontradas na *web*, devido à restrição do par de imagens sobre uma mesma cena assim como da necessidade do uso do *flash*. A Figura 2.8 ilustra parte desse processo.



Figura 2.8: Ilustração do método proposto por Su e colaboradores [10]. As imagens representam, da esquerda para a direita: imagem sem *flash*; imagem com *flash*; resultado da segmentação na imagem sem *flash*; e resultado da segmentação na imagem com *flash*.

Recentemente, Guan e sua equipe [11] propuseram uma abordagem semi-automática para estimar a forma do corpo e a pose de pessoas em imagens (ou pinturas digitalizadas). Nessa abordagem, são computados parâmetros de forma e pose de um modelo 3D de um corpo humano. É utilizado um modelo 3D de forma aprendido a partir de uma base de dados de treinamento denominada *SCAPE* [25], que incorpora grande variação de formas de pessoas como em poses. A partir de informações adquiridas através do usuário (altura estimada da pessoa na imagem e outros pontos de controle) é feita uma estimativa inicial de um modelo articulado 3D da pose e forma da pessoa na imagem. A partir dessa estimativa inicial, são gerados mapas de regiões contidas dentro do objeto, fora do objeto e ao longo do contorno do objeto, usados para segmentar a imagem com o algoritmo de segmentação *Grab-Cut* [15]. Os autores também utilizam um modelo linear de forma do corpo humano (com baixa dimensionalidade) no qual variações devido à altura da pessoa são concentradas ao longo de uma única dimensão, tornando possível a estimativa da forma do corpo com restrições de alto nível. Os autores também formulam o problema de estimativa de forma a partir da iluminação (sombreamento) contida na imagem (*shading – shape from shading*). Dessa forma, é estimada a pose, forma do corpo e iluminação da cena, que produzem um corpo sintetizado que se encaixa de maneira adequada às evidências encontradas na imagem de entrada. O resultado dessa abordagem é um modelo de corpo que pode ser medido, animado, editado, para uma grande variedade de aplicações. A Figura 2.9 ilustra o resultado desse trabalho.

Nos trabalhos de Hu e sua equipe [12] e Ren e colaboradores [13] são propostas abordagens para estimativa de pose de pessoas em imagens estáticas que também podem ser usados para inicializar



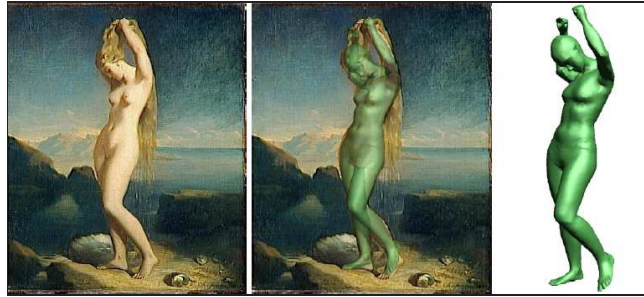


Figura 2.9: Ilustração do resultado do trabalho de Guan e sua equipe [11]. À esquerda, imagem de entrada. Ao centro, objeto segmentado sobreposto na imagem. À direita, estimativa 3D da forma e pose da pessoa.

um modelo de segmentação ou até mesmo para segmentar as imagens diretamente a partir de seus resultados obtidos (uma vez que ambos utilizam algum critério, baseado em segmentação de imagens, para realizar tal estimativa, como por exemplo, segmentação da região do tronco e cor de pele [12] ou baseada em contornos [13]).

No trabalho de Hu e sua equipe [12] é proposta uma abordagem para estimativa de pose da parte superior do corpo de pessoas em imagens estáticas, a partir de três informações observadas em um estágio inicial: região da face, *pixels* em tons de pele e região do tronco. Então as juntas (ou articulações), que ligam as partes do corpo dessa pessoa, são inicializadas de acordo com as observações feitas e restrições com base em heurísticas definidas. O método de MCMC (*Markov chain Monte Carlo*), baseado em exemplos, é utilizado para determinar a estimativa final da pose. A Figura 2.10 ilustra o resultado desse trabalho.

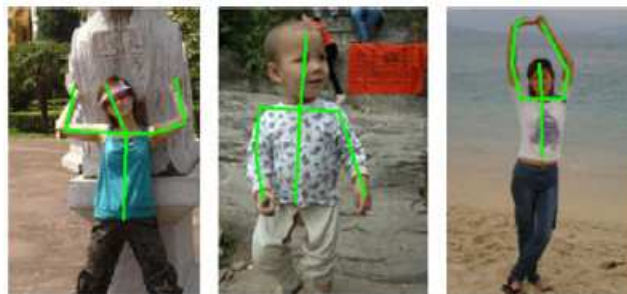


Figura 2.10: Ilustração do resultado do trabalho de Hu e sua equipe [12].

Ren e colaboradores [13] propõem uma abordagem que incorpora restrições entre pares de partes do corpo, como por exemplo, escala, posição relativa, simetria da roupa e contorno suave na conexão entre partes, para estimar pose de pessoas em imagens de forma automática. Possíveis candidatos à parte do corpo são adquiridos através de uma abordagem *bottom-up*, usando características como paralelismo e restrições impostas entre pares de partes do corpo. De forma a originar uma estimativa de pose final, a partir do agrupamento das partes detectadas, é utilizada uma abordagem denominada *Integer Quadratic Programming – IQP*, a qual é relatada pelos autores por poder agregar mais informações do que programação dinâmica [32] (tipicamente aplicada à problemas de otimização), por exemplo. Nesse trabalho são utilizadas 15 imagens, segmentadas por especialista, usadas para

o treinamento de um detector de baixo nível de partes do corpo, assim como para o aprendizado de determinadas restrições entre partes (exemplos de restrições são: conexão entre as partes superiores das pernas devem estabelecer um determinado critério, assim como a posição relativa entre braços e pernas, por exemplo). A Figura 2.11 ilustra o resultado desse trabalho.

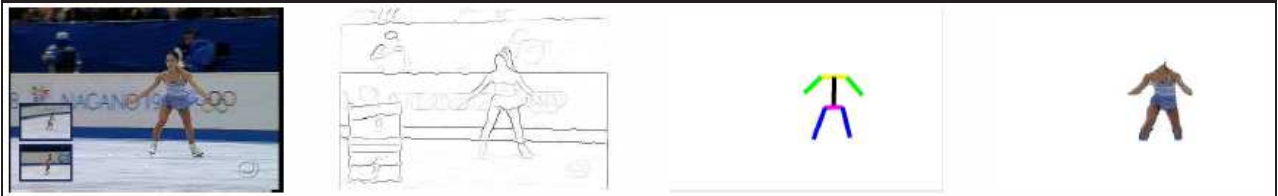


Figura 2.11: Ilustração do resultado do trabalho de Ren e colaboradores [13]. As imagens representam, respectivamente (da esquerda para a direita): imagem de entrada; mapa de bordas; pose estimada; e segmentação resultante.

### 2.3 Contexto desse trabalho no estado-da-arte

O modelo proposto nesta tese, descrito em detalhes no próximo capítulo, apresenta uma abordagem para segmentação de pessoas em imagens estáticas baseada em esqueleto. O modelo baseado em esqueleto guia a segmentação da pessoa na imagem levando em consideração informações de cor, contornos, restrições de ângulos e parâmetros antropométricos. De uma forma geral, a idéia principal da abordagem proposta é construir um grafo ao redor do modelo de esqueleto, para uma determinada imagem de entrada, e buscar o melhor caminho nesse grafo que satisfaça uma determinada condição (por exemplo, que maximize certo critério de energia), gerando assim o contorno (ou silhueta) da pessoa na imagem.

Na abordagem proposta não são usados modelos 3D complexos da forma humana, como em [1, 2, 11] nem base de dados para aprendizado de formas, aparências e/ou poses, como em [5, 7, 9, 31]. Uma característica importante, que deve ser salientada do modelo proposto, é que o resultado dessa abordagem gera um contorno fechado (onde o ponto inicial é igual ao ponto final) com informação semântica embutida, ou seja, cada ponto do contorno resultante está associado à uma determinada parte do corpo (similar ao trabalho de Freifeld e sua equipe [5]).

Resultados experimentais, exibidos em detalhes no Capítulo 4, demonstram que o modelo proposto gera resultados satisfatórios para imagens não triviais, contendo pessoas com aparências e poses variadas, em diversos ambientes complexos (e não controlados), com diferentes iluminações e qualidade de imagem, podendo haver membros parcialmente ocultos, entre outros fatores.

Uma breve comparação de resultados gerados a partir do modelo proposto com resultados ilustrados no trabalho de Freifeld e sua equipe [5], considerado estado-da-arte, demonstram que o nosso modelo gera resultados mais coerentes para o contorno da pessoa, enquanto que os contornos ilustrados em [5] apresentam formas mais suaves.

Possíveis aplicações que poderiam se beneficiar de tal informação semântica, gerada a partir da utilização da abordagem proposta, são modelos para construção de humanos virtuais (avatar) com

características extraídas da imagem (como geometria ou textura) [33], métodos para estimativa de roupas em imagens [6] ou estimativa da forma humana sobre as roupas [34, 35], entre outros. O modelo proposto nessa tese é descrito em detalhes a seguir.



### 3. Modelo Proposto

Segmentar pessoas em imagens estáticas com a utilização de técnicas de visão computacional é uma tarefa bastante desafiadora, assim como a de se obter informações semânticas das pessoas contidas nessas imagens. Isso se deve à grande variabilidade de aparências e poses que essas podem assumir, e fatores relacionados à imagem, como ruído, iluminação, entre outros [3]. Segmentar partes de uma pessoa (em imagens estáticas), como braços, tronco, membros, entre outras, também não é considerada uma tarefa trivial. Segmentar uma imagem em segmentos quaisquer, sem informação semântica, pode ser diferente de segmentar um objeto específico de interesse, pois nesse último precisa-se saber qual parte da imagem é de interesse, no caso, o que é o braço, o que é a perna, a cabeça, etc, e não somente “segmento 1”, “segmento 2”, “segmento 3”, etc. Métodos de processamento de vídeo, por exemplo, podem utilizar informações que não estão disponíveis em imagens estáticas, como as de tempo e movimento, e assim segmentar/detectar pessoas combinando essas características usando diferentes técnicas [36].

A segmentação de uma imagem em segmentos quaisquer pode ser usada como dado de entrada para diversas aplicações (relacionados à figura humana), por exemplo, em técnicas que fazem estimativa de pose, forma humana, edição/manipulação de imagens, entre outras [1, 3, 4]. Por outro lado, a segmentação pode ser o resultado final de uma técnica que utilize alguma informação a priori sobre o objeto a ser segmentado, como a forma do objeto, localização, base de treinamento (aprendizado de informações), entre outras, podendo ser aplicado em segmentação de cor de pele [37], roupas [6], objetos, etc. Dessa forma, um aspecto que deve ser levado em consideração quando se deseja segmentar um objeto de interesse é a quantidade e qualidade de informações envolvidas para resolução do problema.

Conforme relatado por Hornung e sua equipe [4], a aquisição da postura 2D de um ser humano de forma interativa, que poderia ser utilizada para inicializar um método de segmentação, por exemplo, tem algumas vantagens quando comparada a métodos automáticos, pois a intervenção manual normalmente leva alguns minutos e gera resultados superiores em poses onde há alguma ambiguidade, se comparado a técnicas de estimativa de pose automáticas. Por outro lado, existem métodos automáticos que poderiam [3, 7, 12, 13, 26] ser usados para estimar poses de pessoas em imagens estáticas sem qualquer intervenção humana, porém ainda não obtendo resultados tão bons quanto os obtidos de forma interativa, devido a diversos fatores, como grande variação de poses e aparências que as pessoas podem assumir, complexidade da cena capturada na imagem em análise, entre outros.

O modelo de esqueleto utilizado nessa tese para guiar a segmentação pode ser obtido manualmente (através do usuário) ou de forma automática, com a utilização de um algoritmo para estimativa automática de poses 2D em imagens (como [26], por exemplo). Sob o ponto de vista semi-automático, a ideia inicial da segmentação baseada em esqueleto é que o usuário informe alguns dados pertencentes à pessoa na imagem (por exemplo, altura, posição da cabeça, mãos,

pés, etc – associadas ao modelo de esqueleto), e a partir daí a segmentação é feita de forma automática. Sob o ponto de vista automático, a ideia é que esses dados sejam adquiridos sem intervenção com o usuário, fazendo que o modelo proposto seja capaz de extrair o contorno de uma pessoa em uma imagem de forma totalmente automatizada. O modelo proposto nesta tese, descrito em detalhes neste capítulo, é definido inicialmente a partir de informações adquiridas de maneira semi-automática, ou seja, os dados de entrada relacionados ao modelo de esqueleto são informados através de interação com usuário. O objetivo dessa decisão é simplificar o sistema de entrada de dados, tornando-o também mais preciso. Dessa forma, os resultados gerados pelo modelo proposto não devem (ou não deveriam) ser influenciados, por exemplo, por um dado de entrada adquirido de forma imprecisa que por ventura possa ocorrer ao utilizar um algoritmo automático para estimativa de poses 2D de pessoas em imagens. Um estudo de caso, detalhando as vantagens/desvantagens dessas duas abordagens (semi-automática  $\times$  automática), assim como resultados experimentais do modelo proposto em uma configuração totalmente automática é apresentado na Seção 4.4.

Desconsiderando-se a variabilidade de poses e aparências que uma pessoa possa assumir, e fatores relacionados a qualidade da imagem em análise, como iluminação, ruído, entre outros, deve ser também considerado o tipo de dado em estudo, ou seja, número de pessoas na imagem, a distância que as pessoas estavam da câmera quando fotografadas (relacionados ao tamanho que a pessoa vai assumir na imagem - proporção da pessoa  $\times$  resolução da imagem), distância que as pessoas estavam umas das outras (podendo gerar oclusão ou ambiguidade), ambiente da cena (relacionado ao fundo da imagem - homogêneo, texturizado, controlado), se a(s) pessoa(s) está(ão) com todo o corpo aparente na imagem, entre outros fatores.

Nas próximas seções são descritas em detalhes as etapas que compõem o modelo de segmentação baseado em esqueleto. A Figura 3.1 exibe uma visão geral do modelo proposto. Mais especificamente, na Seção 3.1 são descritos os dados de entrada para a geração do contorno (esqueleto de entrada e altura estimada da pessoa na imagem). A definição do grafo criado ao redor do modelo esqueleto, assim como a forma na qual suas conexões especiais são definidas e a maneira como a energia de seus caminhos é medida são descritos na Seção 3.2. Na Seção 3.2.5 é detalhado o método que encontra o melhor caminho desse grafo, originando o contorno da pessoa na imagem. Na Seção 3.3 é apresentada uma abordagem que utiliza informações de cores predominantes da pessoa na imagem, para auxiliar no cálculo que irá gerar seu contorno. Estudos de caso e resultados experimentais são apresentados no Capítulo 4. Por fim, considerações finais e sugestões para trabalhos futuros são apresentadas no Capítulo 5.

### **3.1 Modelo de esqueleto: dados de entrada**

O modelo de segmentação proposto, baseado em esqueleto, utiliza uma série de parâmetros estimados a partir de dados antropométricos, como por exemplo, largura média de um braço, perna, tronco, etc. Tais parâmetros determinam, por exemplo, regiões de aprendizado e busca de cores predominantes, restrições de ângulos e distâncias, assim como regiões de busca dos contornos, e

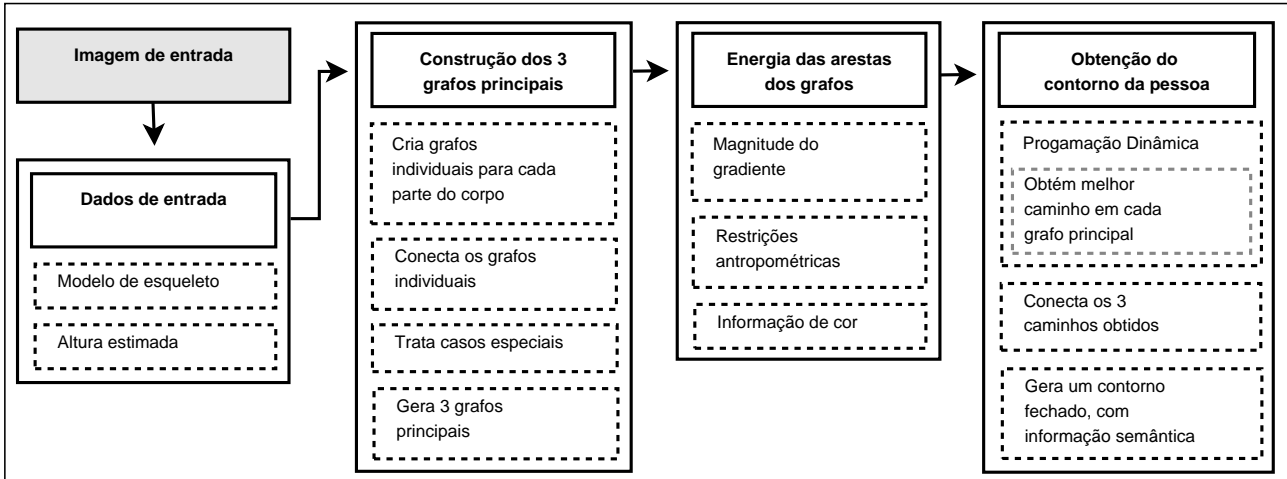


Figura 3.1: Visão geral do modelo proposto.

são definidos a partir de um modelo de esqueleto, ilustrado com auxílio da Figura 3.2. O modelo de esqueleto é composto por 19 pontos de controle que definem 16 segmentos de reta, ou “ossos”. Todos os 19 pontos definidos no modelo de esqueleto (ilustrados na Figura 3.2) devem ser informados por um usuário/especialista como dados de entrada (no caso semi-automático) ou adquiridos de forma automática por um algoritmo de estimativa de poses 2D de pessoas em imagens, para cada imagem analisada (a ser segmentada). Os pontos devem ser associados às suas respectivas partes do corpo, ou seja, devem ser adquiridos em uma determinada ordem para que o modelo possa relacionar cada ponto de entrada à sua respectiva parte do corpo.

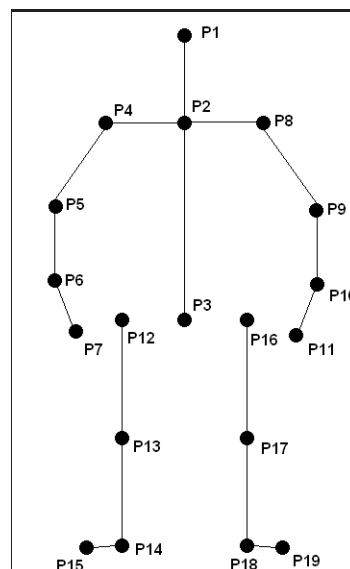


Figura 3.2: Modelo de esqueleto adotado para guiar a segmentação.

Todos os “ossos” do modelo de esqueleto têm suas larguras estimadas, parametrizadas em função da altura  $H$  de uma pessoa de altura média. Mais precisamente, para uma determinada parte do corpo com índice  $i$ , referenciado na Tabela 3.1, a correspondente largura  $w_i$  é dada por

$$w_i = H f_{w_i}, \quad (3.1)$$

onde os fatores de proporcionalidade  $f_{w_i}$  são derivado a partir de valores antropométricos [38]. A Tabela 3.1 exibe todos os “ossos” do modelo de esqueleto, assim como os valores correspondentes de  $f_{w_i}$ .

Tabela 3.1: Partes do corpo relacionadas ao modelo de esqueleto usado. Primeira coluna: índice de cada parte; segunda coluna: parte do corpo; terceira coluna: pontos que formam cada parte; quarta coluna: parâmetro usado no cálculo da estimativa da largura de cada parte.

$i$	Parte do corpo	Pontos	$f_{w_i}$
0	Cabeça	(P2 - P1)	0.0883
1	Tronco	(P2 - P3)	0.1751
2	Braço esquerdo	(P5 - P4)	0.0608
3	Antebraço esquerdo	(P6 - P5)	0.0492
4	Mão esquerda	(P7 - P6)	0.0593
5	Braço direito	(P9 - P8)	0.0608
6	Antebraço direito	(P10 - P9)	0.0492
7	Mão direita	(P11 - P10)	0.0593
8	Coxa esquerda	(P13 - P12)	0.0912
9	Canela esquerda	(P14 - P13)	0.0608
10	Pé esquerdo	(P15 - P14)	0.0564
11	Coxa direita	(P17 - P16)	0.0912
12	Canela direita	(P18 - P17)	0.0608
13	Pé direito	(P19 - P18)	0.0564
14	Ombro esquerdo	(P4 - P2)	0.0608
15	Ombro direito	(P8 - P2)	0.0608

Como mencionado no parágrafo anterior, outro dado de entrada importante a ser considerado no modelo proposto é a altura da pessoa na foto. Na abordagem proposta, há duas formas de se estimar a altura de uma pessoa em uma imagem (em coordenadas de imagem), através de intervenção manual:

- Quando a pessoa está em pé na imagem, com todo o corpo visível (desconsiderando-se oclusão de membros superiores, como braços, por exemplo), o usuário simplesmente informa um ponto no topo da cabeça e outro logo abaixo dos pés da pessoa, obtendo-se a altura  $H$  diretamente.
- Por outro lado, se a pessoa estiver sentada, ou com parte do corpo oculta (principalmente os membros inferiores), sua altura  $H$  é estimada a partir do tamanho de sua face combinada



com valores antropométricos, como descrito a seguir. Mais precisamente, o usuário informa um ponto no topo da cabeça da pessoa e outro na altura do queixo da mesma para se obter o comprimento da face  $h_f$ . Dessa forma, a altura dessa pessoa é então estimada utilizando-se a Equação 3.2.

$$H = \frac{h_f}{0.125}, \quad (3.2)$$

onde 0.125 é um peso (ou fração) derivado de valores antropométricos [38].

Uma alternativa ao usuário informar o tamanho da face ( $h_f$ ) seria capturar esse valor com a utilização de um algoritmo de detecção de faces. Entretanto, quando a pessoa está de corpo inteiro na imagem, a estimativa direta da altura  $H$  tem se mostrado mais convincente (gerando melhores resultados) que a altura estimada pelo tamanho de sua face (conforme Equação 3.2), devido a diversos fatores, como perspectiva da câmera, pessoas com tamanhos diferentes, entre outros. Dessa forma, essa abordagem alternativa tornar-se-ia atrativa apenas em situações onde a altura não pudesse ser adquirida diretamente, necessitando talvez de uma confirmação do usuário, ou utilizando alguma métrica para detectar a ocorrência dessa situação automaticamente. Conforme mencionado anteriormente, na Seção 4.4 é apresentado um estudo de caso onde os dados de entrada são adquiridos sem intervenção do usuário e a segmentação é realizada de forma automática.

Os dados obtidos nessa primeira etapa, como altura estimada, pontos de controle (esqueleto de entrada) e estimativas de larguras, são utilizados em diversas etapas do modelo de segmentação baseado em esqueleto proposto, descritos em detalhes nas próximas seções.

### 3.2 Geração do grafo

Conforme mencionado anteriormente, a ideia por trás do modelo proposto é criar um grafo ao redor do esqueleto de cada pessoa na imagem e encontrar o melhor caminho nesse grafo, baseado em um determinado critério (por exemplo, o melhor caminho é aquele que maximiza um determinado valor de energia), que irá compor o contorno ou silhueta da pessoa. Basicamente, são criados três grafos principais,  $A$ ,  $B$  e  $C$  (um para a parte superior do corpo, grafo  $A$ , e outros dois para a parte inferior do corpo, grafos  $B$  e  $C$ ). Para cada grafo é gerado um caminho (o melhor caminho), que representa o contorno da pessoa para aquele determinado subgrupo de partes do corpo. No final do processo, os três caminhos obtidos são conectados, originando um único contorno que irá representar a silhueta da pessoa na imagem. A Figura 3.3(a) ilustra os três grafos gerados para uma determinada imagem (mais especificamente os níveis e limites de cada grafo), assim como a Figura 3.3(b) ilustra o contorno resultante, representado pela conexão dos melhores caminhos, obtidos para cada grafo principal.

A Figura 3.4 ilustra a obtenção do contorno para uma única parte do corpo. De uma forma geral, para cada lado (esquerdo e direito) de um determinado “osso” é criado um grafo, como ilustrado nas Figuras 3.4(c-d) e o objetivo é criar dois caminhos (um para cada lado/grafos), considerados

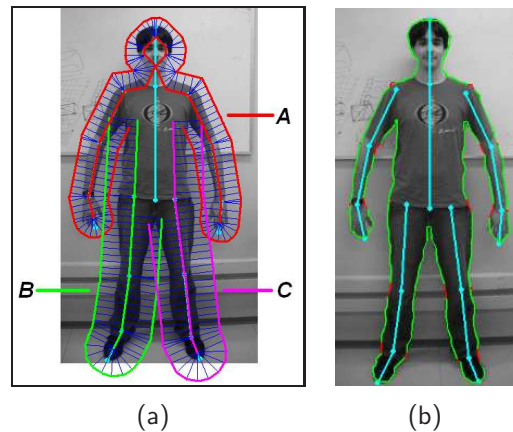


Figura 3.3: (a) Ilustração dos três grafos principais ( $A$ ,  $B$  e  $C$ ). Em vermelho, grafo  $A$  gerado para parte superior do corpo. Em verde e magenta, grafos  $B$  e  $C$ , gerados para a parte inferior do corpo (lado direito e esquerdo, respectivamente). (b) Contorno resultante sobreposto na imagem de entrada (convertida para escala de cinza), com esqueleto de entrada ilustrado em ciano.

os melhores, como ilustrado na Figura 3.4(e) (em vermelho), que irão compor o contorno de uma determinada parte do corpo.

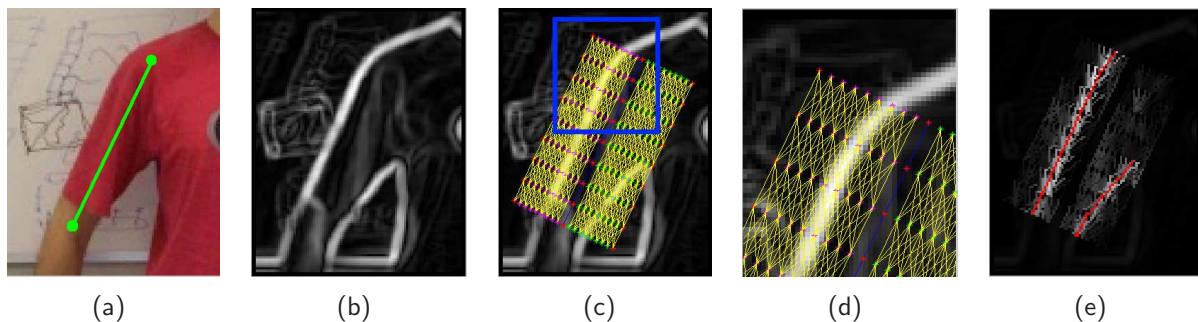


Figura 3.4: (a) Imagem de entrada (e respectivo “osso”, em verde). (b) Magnitude do gradiente da imagem (a) (após conversão para escala de cinza). (c) Nodos dos grafos (asteriscos) e caminhos gerados (linhas em amarelo). (d) Zoom na imagem (c) (região representada pelo retângulo azul). (e) Contorno obtido pelo caminho que maximiza o valor de energia.

A hipótese assumida nessa abordagem é que cada parte aproximadamente retangular do corpo (como braço, antebraço, perna, tronco, etc), definida a partir de dois pontos de controle (como descrito na Seção 3.1) é limitada por um contorno, que deve ter inclinação similar à do seu respectivo “osso”, gerado pela conexão entre os dois pontos de controle que o formam, assim como deve respeitar algumas restrições de distância (restrições baseadas em antropometria). Entretanto, nem sempre o contorno de uma determinada parte do corpo vai ser uma linha paralela à seu respectivo “osso”, como por exemplo, para o contorno da cabeça, mãos, pés, ou até mesmo por causa das ondulações causadas por roupas, oclusões, etc, fazendo com que a resolução desse problema não seja considerada trivial.

A busca pelo melhor caminho é feita com a utilização de uma técnica chamada *programação*

*dinâmica* [32], tipicamente aplicada em problemas de otimização. Conforme [32], *programação dinâmica*, assim como o método de *dividir para conquistar*, são formas de solucionar problemas combinando suas soluções em subproblemas. A forma na qual os grafos são criados, caminhamentos possíveis, energia calculada, restrições antropométricas e detalhes da programação dinâmica são descritos a seguir.

### 3.2.1 Definição do Grafo

Considere  $G_i = (S, E)$  um grafo gerado para cada parte do corpo  $i$ , constituído por um conjunto finito  $S$  de vértices (nodos, ou pontos) e um conjunto  $E$  de arestas (ou linhas), tal que  $E \subseteq [S]^2$ , ou seja, os elementos de  $E$  são pares de elementos do conjunto  $S$ . Os vértices formam uma estrutura tipo *grid* (grade ou rede) e são posicionados ao longo de uma região onde seja esperado que o contorno da pessoa esteja (a Figura 3.4(d) exibe o grafo para o contorno externo do braço direito de uma pessoa). Os vértices formam níveis ao longo do *grid*, onde cada nível é ortogonal ao segmento de reta que conecta seus dois pontos de controle ( $P_1$  e  $P_2$ ), associados a uma determinada parte do corpo (“osso”). A extensão do grafo, assim como o número de vértices em cada nível, é baseada em valores estimados por antropometria, descritos na Tabela 3.1, que provê a largura estimada para cada parte do corpo. Os vértices são nomeados  $S_{m,n}$ , onde  $m = 1, \dots, M$  denota o nível de cada vértice e  $n = 1, \dots, N$  é a posição de cada vértice em cada nível, de forma que menores valores de  $n$  estão relacionados à vértices mais próximos do seu respectivo “osso”. Cada vértice em um determinado nível é conectado com os  $k$  vértices mais próximos do nível seguinte, com exceção dos níveis das bordas, onde o número de ligações é menor. A utilização de valores muito grandes para  $k$  ( $k > 3$ , por exemplo) faz com que o contorno de uma determinada parte do corpo possa assumir mudanças de direções muito bruscas, dependendo das bordas encontradas na imagem, podendo gerar efeitos desagradáveis (efeito *zig-zag*). Assim, definiu-se experimentalmente  $k = 3$ . Um exemplo de um grafo, com seus vértices e arestas é ilustrado com auxílio da Figura 3.5.

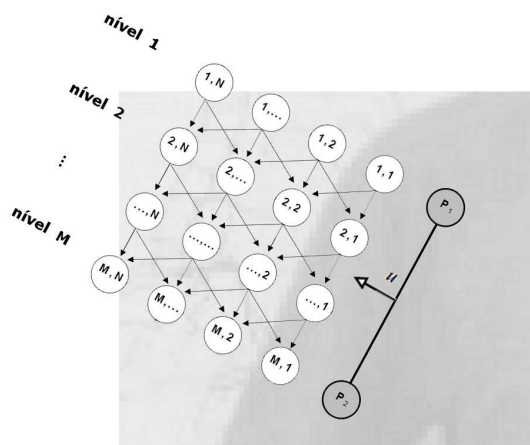


Figura 3.5: Ilustração do grafo gerado para o contorno externo do braço direito (ilustrado na Figura 3.4).

O número de níveis do grafo  $G_i$ , gerado para uma determinada parte do corpo  $i$ , e representado por  $M$ , (conforme ilustrado na Figura 3.5), é calculado diretamente através da distância (em *pixels*) entre os dois pontos de controle ( $\mathbf{P}_1$  e  $\mathbf{P}_2$ ), multiplicada por um determinado fator  $s_4$  (setado experimentalmente para  $s_4 = 0.1$ ). Essa discretização é feita com o objetivo de acelerar o cálculo do contorno (pois não é criado um vértice para cada *pixel* da imagem, por exemplo) sem desprezar a curvatura que um determinado contorno pode assumir. Assim, o número de níveis para o grafo  $G_i$  é dado pela Equação 3.3.

$$M = \text{round}(s_4 \|\mathbf{P}_1 - \mathbf{P}_2\|), \quad (3.3)$$

onde  $\text{round}(\cdot)$  representa o valor arredondado de um número.

A localização do vértice  $S_{1,N}$  (vértice do primeiro nível mais distante o ponto  $\mathbf{P}_1$ , ilustrado na Figura 3.5) está associada a uma determinada distância do seu respectivo “osso”, estimada por antropometria, na qual é esperado que o contorno de uma determinada parte do corpo esteja (em coordenadas de imagem). Para manter certa coerência global entre os 3 grafos principais ( $A$ ,  $B$  e  $C$ ), essa distância  $d$  é fixada como a largura ( $w_i$ ) máxima estimada para o subconjunto de partes do corpo que compõem cada grafo principal, definida na Equação 3.4 para o grafo  $A$  (parte superior do corpo) e na Equação 3.5 para os grafos  $B$  e  $C$  (parte inferior do corpo), conforme indicado na Tabela 3.1.

$$d_A = w_2, \quad (3.4)$$

$$d_B = d_C = w_8 \quad (3.5)$$

Assim, podemos definir o vértice  $S_{1,N}$  (ilustrado na Figura 3.5), como descrito na Equação 3.6.

$$S_{1,N} = (\mathbf{P}_{1x} - d \sin(\beta), \mathbf{P}_{1y} + d \cos(\beta)), \quad (3.6)$$

onde  $\beta$  é o ângulo formado pelo vetor criado a partir dos pontos  $\mathbf{P}_1$  e  $\mathbf{P}_2$ . Os demais vértices desse mesmo nível podem ser criados com a diminuição do valor de  $d$ , de acordo com algum critério estabelecido (número de subdivisões  $N$ , por exemplo). O número de vértices em cada nível (ou número de subdivisões  $N$ ) também sofre uma discretização em função da largura ( $w_i$ ) máxima estimada para o subconjunto de partes do corpo que o compõem, com o objetivo de acelerar o cálculo do contorno sem desprezar a curvatura que o mesmo pode assumir. Assim, o número de vértices em cada nível do grafo  $A$  (parte superior do corpo) é dado pela Equação 3.7. Seguindo o mesmo critério, o número de vértices em cada nível dos grafos  $B$  e  $C$  (parte inferior do corpo) é dado pela Equação 3.8

$$N_A = w_2 s_6, \quad (3.7)$$

$$N_B = N_C = w_8 s_6, \quad (3.8)$$

onde  $s_6$  é um fator de proporcionalidade (setado experimentalmente para  $s_6 = 0.33$ ). Uma justificativa para o valor de  $s_6$  ser maior que o valor de  $s_4$  (relacionados ao valor de  $N$  e  $M$ , respectivamente),

baseia-se na hipótese de que o contorno deve ser uma linha com orientação similar à do seu respectivo “osso”, necessitando de menos subdivisões para representar contornos com muitas curvas nessa direção de crescimento (relacionado ao valor de  $s_4$  e consequentemente  $M$ ). Os vértices dos outros níveis podem ser definidos de maneira similar, através de multiplicação vetorial (de vetores criados a partir dos pontos  $P_1$  e  $P_2$ ,  $M$  e  $d$ ), por exemplo.

### 3.2.2 Conectando os grafos e casos especiais

A definição do grafo descrita na seção anterior está focada em uma única parte do corpo. Os 3 grafos principais ( $A$ ,  $B$  e  $C$ ), usados na abordagem proposta, são formados por diversas partes do corpo. Assim, os grafos gerados para cada parte do corpo devem ser conectados uns aos outros, de uma maneira coerente, para formar o grafo da parte superior (grafo  $A$ ) e os dois grafos da parte inferior (grafos  $B$  e  $C$ ).

Quando uma determinada parte do corpo é conectada à outra, as regiões delimitadas por seus respectivos grafos podem se interceptar ou deixar “buracos”, dependendo do ângulo  $\alpha$ , formado no ponto de conexão. A Figura 3.6 exemplifica essa situação, para a conexão entre a canela e o pé direito de uma pessoa em uma imagem. De uma forma geral, há intersecção quando  $\alpha < 180^\circ$ , assim, alguns níveis devem ser removidos, e “buracos” são originados quando  $\alpha > 180^\circ$ , então alguns níveis devem ser criados.

A Figura 3.7 detalha esse procedimento, de criação/remoção de níveis, para a conexão entre os grafos da canela e pé direito exibidos na Figura 3.6. No contorno externo da conexão, ilustrado na Figura 3.7(a) os grafos se interceptam, dessa forma alguns níveis devem ser removidos (ilustrados na Figura 3.7(a), à esquerda, por linhas tracejadas em magenta) e então os dois grafos são conectados por um “nível de conexão” (ilustrado na Figura 3.7(a), à direita, por uma linha tracejada em magenta). O “nível de conexão” é um segmento de reta que vai do ponto que conecta os dois “ossos” (nesse caso, o ponto  $P_{14}$  do modelo de esqueleto, associado ao tornozelo direito) até o ponto criado pela intersecção entre os dois segmentos de retas que delimitam aquele lado do grafo (ilustrado por um sinal de + em azul). No contorno da parte interna há um “buraco” não preenchido, e alguns níveis devem ser criados para formar uma conexão suave entre os dois grafos desse lado do “osso” (ilustrado na Figura 3.7(b)). Nesse caso também é gerado um “nível de conexão”, que irá conectar os dois grafos, porém, dependendo do ângulo  $\alpha$  formado nesse ponto, os níveis de cada grafo podem ser expandidos até encontrar o “nível de conexão”.

Esse procedimento de conexão de dois grafos adjacentes, exemplificado para a conexão entre a canela e o pé direito, também é realizado entre as seguintes partes do corpo: para o grafo da parte superior do corpo, entre ombro×braço, braço×antebraço, antebraço×mão, assim como para os grafos da parte inferior do corpo, entre tronco×coxa, coxa×canela e canela×pé. Também deve ser mencionado que quando  $\alpha = 180^\circ$  os grafos são simplesmente concatenados. Como o número de nodos em cada nível é o mesmo, para cada grafo principal ( $A$ ,  $B$  e  $C$ ), a conectividade entre os grafos individuais é mantida ao longo da direção do “osso”.

A seguir são detalhados os casos especiais, como por exemplo, a modelagem das mãos, pés,

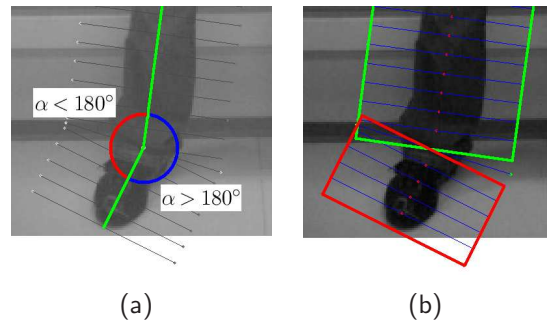


Figura 3.6: Ilustração da conexão entre dois grafos adjacentes (canela e pé direito, por exemplo). Em (a), os ângulos formados na conexão entre as duas partes. Em (b), regiões associadas a cada parte do corpo podem se interceptar ou deixar “buracos”, dependendo do ângulo formado no ponto de conexão.

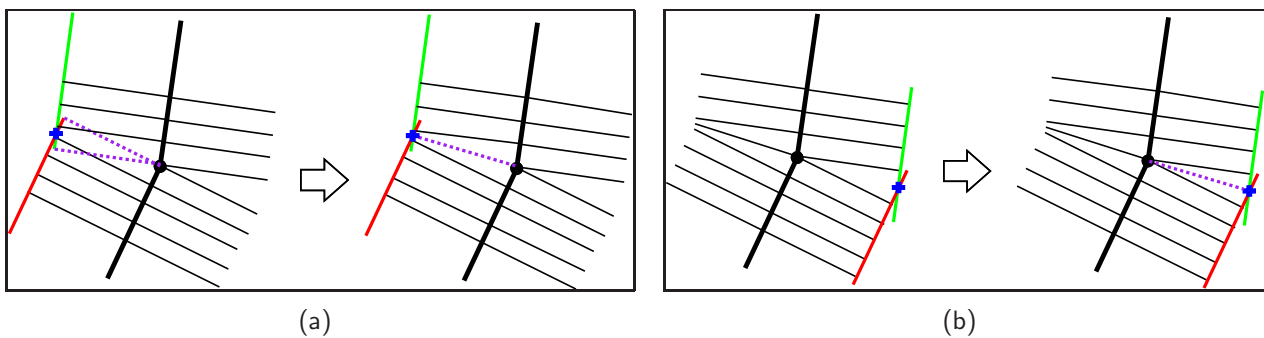


Figura 3.7: Detalhes da conexão entre dois grafos adjacentes (canela e pé direito, por exemplo). (a) Há intersecção quando o ângulo  $\alpha < 180^\circ$ , assim, alguns níveis devem ser removidos, ilustrados à esquerda em magenta, então os dois grafos são conectados por um “nível de conexão”, ilustrado à direita por uma linha em magenta. (b) Há “buracos” quando  $\alpha > 180^\circ$  (ilustrado à esquerda), assim, alguns níveis devem ser criados (ilustrado à direita).

cabeça, ombros, tronco e como essas partes são conectadas umas as outras de maneira coerente, originando os três grafos principais usados nesse trabalho.

### Modelagem das mãos e pés

Mesmo que a maioria das partes do corpo gere um grafo em uma região retangular, conforme descrito anteriormente, algumas partes específicas apresentam formas diferentes, como a cabeça, mãos e pés. Na abordagem proposta, as mãos e pés são modelados por um setor circular onde os níveis são definidos por segmentos de reta radiais, discretizados por um ângulo de  $22.5^\circ$ , setado experimentalmente.

Inicialmente, ambas as partes (mãos e pés) tem seu respectivo “osso” diminuído em suas extremidades por uma distância  $d/2$ , ou seja, a metade da largura usada na criação do seu respectivo grafo (ilustrado na Figura 3.8(b)). O objetivo dessa redução é fazer com que a região esperada para a ocorrência do contorno esteja dentro da região definida pelo seu grafo, expresso pelo setor



circular, e não na borda dessa mesma região (assumindo-se que o ponto extremo do esqueleto das mãos e pés está associado ao ponto extremo dos mesmos membros na imagem). Posteriormente, os níveis criados pelo setor circular (para cada mão e pé) são conectados ao seu respectivo grafo (criado anteriormente pela região retangular, ilustrado na Figura 3.8(a) para a mão direita de uma determinada pessoa), de forma que níveis sobrepostos sejam removidos (ilustrado na Figura 3.8(c)). Esse procedimento liga os dois lados/grafos de cada braço e perna (em relação aos seus respectivos “ossos” - lado esquerdo e direito).

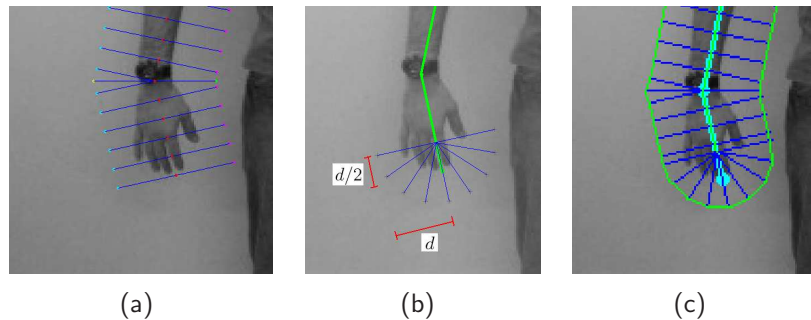


Figura 3.8: Ilustração da modelagem do grafo da mão direita (setor circular). Em (a) é exibido o grafo em forma retangular, antes da conexão entre os dois lados (esquerdo e direito, em relação ao “osso”), feita pela inserção do setor circular. Em (b) é ilustrado o setor circular criado, assim como a diminuição do “osso” dessa extremidade em questão (distância  $d/2$ ). Em (c) é ilustrado o grafo resultante.

### Modelagem dos ombros

A modelagem dos ombros é feita da seguinte forma. Inicialmente, os “ossos” dos ombros são diminuídos por um determinado fator (setado experimentalmente para a metade da largura do pescoço,  $w_p/2$ , onde  $w_p = 0.0333H$ , baseado em [38]), no lado do pescoço, conforme ilustrado na Figura 3.9(a). A geração do seu grafo inicial é similar à uma parte do corpo normal (braço por exemplo), porém no caso do ombro o grafo é criado apenas para um lado do “osso” (lado superior), conforme ilustrado na Figura 3.9(b), usando-se a mesma largura dos grafos dos braços (objetivando manter uma coerência global no grafo da parte superior do corpo). Posteriormente, os grafos criados para cada ombro são conectados aos seus respectivos braços, de maneira similar à descrita no início da Seção 3.2.2, porém apenas na parte externa em relação ao “osso” (uma vez que os grafos dos ombros são criados apenas para um único lado do “osso”).

### Modelagem da cabeça

A cabeça é modelada por uma forma hexagonal. Basicamente, o “osso” do esqueleto associado à cabeça é diminuído na sua parte superior por um determinado fator (setado experimentalmente para  $w_p/2$ ) e na sua parte inferior por  $w_p$ , conforme ilustrado na Figura 3.10(a). A largura do hexágono também é setada experimentalmente para  $w_p$ . O objetivo desse procedimento é criar

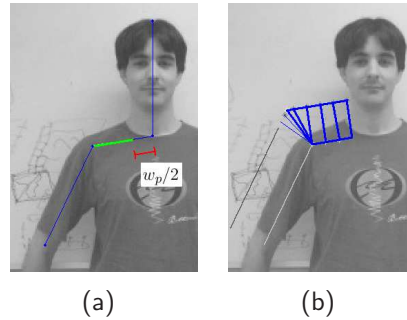


Figura 3.9: Ilustração da modelagem do ombro. Em (a) é ilustrado a diminuição do “osso” do ombro direito, no lado do pescoço. Em (b) é exibido o grafo gerado para o ombro direito, já conectado ao grafo do braço associado.

uma forma simplificada para o grafo da cabeça, fazendo com que os níveis desse grafo englobem a região esperada para a ocorrência do seu contorno. O ponto inferior desse hexágono é conectado aos “ossos” dos ombros, ilustrado por uma linha em magenta na Figura 3.10(a).

A partir dessa configuração inicial, o grafo da cabeça (mais especificamente associado à parte inferior hexágono) é criado com a mesma largura dos grafos dos braços, objetivando manter uma coerência global no grafo da parte superior do corpo (grafo  $A$  - essa coerência é mantida para todo o grafo da cabeça). O ponto de intersecção entre as fronteiras do grafo do ombro e o grafo da parte inferior do hexágono da cabeça são adquiridos (ilustrado na Figura 3.10(b)), para ambos os lados do corpo (esquerdo e direito). Esses pontos de intersecção irão unir o grafo criado para a cabeça aos dois ombros da pessoa na imagem, conforme ilustrado na Figura 3.10(c). Os demais níveis do hexágono são criados e conectados uns aos outros como uma parte do corpo qualquer. Por fim, os níveis do hexágono (que não são conectados aos pontos de intersecção com os ombros) são normalizados pela largura do braço (também usada para criar o grafo da cabeça, como mencionado anteriormente), de forma que o grafo resultante tenha uma forma mais “arredondada”, conforme ilustrado na Figura 3.10(d).

Os procedimentos apresentados até aqui descrevem como os diversos grafos, associados às várias partes do corpo, são conectados uns aos outros, assim como definem como o grafo da parte superior do corpo é criado (grafo  $A$  - ilustrado em vermelho na Figura 3.3(a)). Os procedimentos usados para criar os grafos para a parte inferior do corpo (lado esquerdo e direito) são descritos a seguir.

### Modelagem do tronco

O tronco é modelado por dois diferentes grafos (um para cada lado do corpo - esquerdo/direito), ilustrados com auxílio da Figura 3.11(a). Basicamente, são criados dois segmentos de reta, que conectam cada fêmur (pontos  $P_{12}$  e  $P_{16}$ , associados ao modelo de esqueleto) aos seus respectivos ombros, no ponto médio de cada “osso” do ombro, conforme ilustrado na Figura 3.11(b) para o lado direito do corpo de uma determinada pessoa. Esse segmento de reta é então diminuído na sua parte superior, tanto para o grafo criado para o lado direito como para o grafo criado para o lado esquerdo



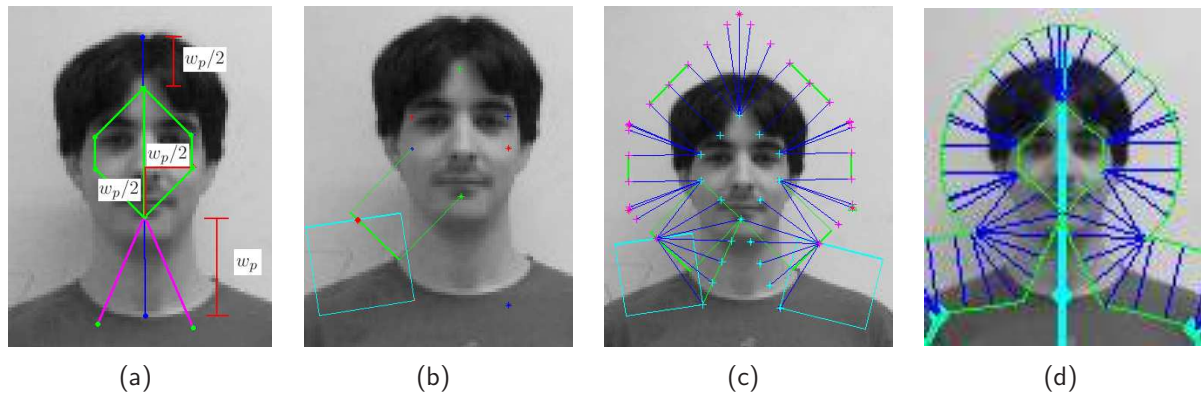


Figura 3.10: Ilustração da modelagem da cabeça. (a) Estrutura hexagonal do grafo da cabeça e sua conexão com os “ossos” dos ombros (em magenta). (b) Ponto de intersecção entre o grafo da cabeça e o grafo do ombro direito. (c) Grafo da cabeça conectada aos ombros. (d) Grafo da cabeça resultante, normalizado pela largura usada em todo grafo  $A$ .

do corpo, por um determinado fator  $l_p$  (onde,  $l_p = 0.0980H$ ), relacionado ao comprimento do peito de uma pessoa de tamanho mediano (baseado em valores antropométricos [38]), objetivando desconsiderar a região localizada entre o ombro e a região da axila. O segmento de reta que irá compor o grafo do tronco para o lado direito de uma determinada pessoa é ilustrado em verde, na Figura 3.11(b). É importante salientar que a largura do grafo do tronco é a mesma largura usada para todo grafo da parte inferior do corpo (grafos  $B$  e  $C$ ), objetivando manter certa coerência global.

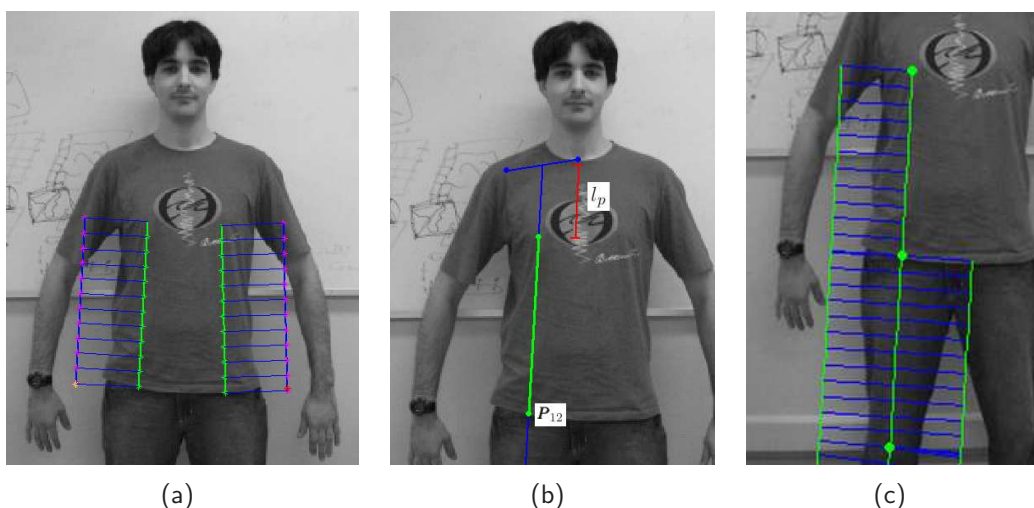


Figura 3.11: Ilustração da modelagem do tronco. Em (a), dois grafos criados para o tronco. Em (b), ilustração do segmento de reta gerado para a criação do grafo do tronco, para o lado direito (que vai do ponto médio do ombro ao início da perna). Em (c), conexão entre o tronco  $\times$  coxa.

Após criar o grafo para o lado direito do tronco, o mesmo pode ser conectado ao grafo da perna direita, similarmente à conexão entre a canela e o pé (descrito anteriormente, no início da

Seção 3.2.2), porém nesse caso o grafo do tronco é conectado apenas ao lado externo da perna, ilustrado na Figura 3.11(c). Esse procedimento é repetido para o lado esquerdo do corpo.

### Modelagem da coxa (parte interna)

Como pode ser visto na Figura 3.12(a), o grafo criado para a coxa direita (similarmente para a coxa esquerda) de uma determinada pessoa, inicia no ponto definido pelo início de seu respectivo “osso” (ponto  $P_{12}$ , associado ao modelo de esqueleto) até sua conexão com a canela (ponto  $P_{13}$ ). Entretanto, na parte interna da coxa, pode-se verificar (visualmente, na Figura 3.12(a)) que seu grafo (na parte superior) cobre uma determinada região onde geralmente o contorno esperado daquela parte não a atinge (associada à região do quadril). Dessa forma, o grafo da coxa (lado esquerdo e direito do corpo) é diminuído na sua parte interna superior por um determinado fator  $l_q$  (onde  $l_q = 0.0492H$ ), associado ao comprimento do quadril de uma pessoa de altura mediana (derivado a partir de valores antropométricos [38]). O resultado desse procedimento, para o lado direito do corpo de uma determinada pessoa, é ilustrado na Figura 3.12(b).

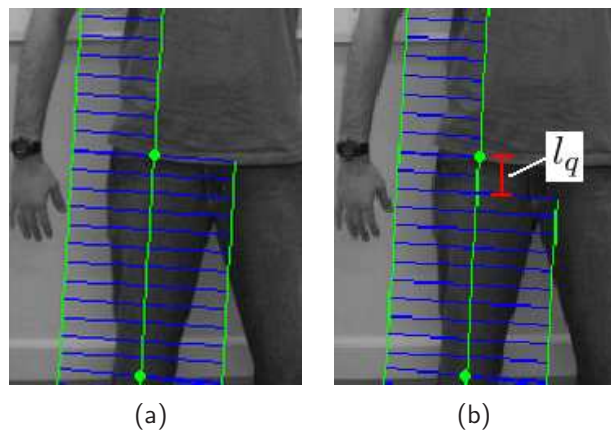


Figura 3.12: Ilustração da modelagem da coxa (parte interna). Em (a), detalhe da conexão entre tronco  $\times$  coxa, antes do tratamento especial. Em (b), tratamento especial para a parte interna do grafo da coxa direita (redução na parte superior) e ilustração do “comprimento do quadril”  $l_q$ .

Os procedimentos utilizados para a criação dos três grafos principais (grafos  $A$ ,  $B$  e  $C$ ) foram descritos até aqui. Nas próximas Seções (3.2.3 e 3.2.4) são descritos dois métodos utilizados para calcular a energia das arestas desses grafos, usadas na obtenção do contorno da pessoa na imagem.

### 3.2.3 Mapa de Energias convencional

Definido o grafo  $G_i = (S, E)$  para uma determinada parte do corpo  $i$ , onde  $E$  é o conjunto de arestas desse grafo (pares de elementos do conjunto  $S$ ), assim como definidos os três grafos principais (grafos  $A$ ,  $B$  e  $C$ ), compostos por subconjuntos de partes do corpo, é preciso definir como são calculadas as energias de suas arestas. A energia de cada aresta irá representar o custo

associado a um determinado caminho do grafo, que liga dois níveis adjacentes, podendo ser calculada por diferentes formas.

Uma medida de energia amplamente utilizada em processamento de imagens para detecção de contornos é o valor da magnitude do gradiente dessa imagem. O gradiente  $\nabla I = (I_x, I_y)^T$  de uma imagem em escala de cinza (considere que a imagem de entrada, no espaço de cor RGB, é transformada para escala de cinza) pode ser computado através da convolução dessa imagem com um filtro de Sobel [17], por exemplo. A magnitude (ou módulo)  $\|\nabla I\|$  do gradiente de uma imagem é dada pela Equação 3.9.

$$\|\nabla I\| = \sqrt{I_x^2 + I_y^2} \quad (3.9)$$

A energia de cada aresta pode ser medida diretamente pelos valores de  $\|\nabla I\|$  ao longo dos *pixels* da aresta (somando os valores, por exemplo). Entretanto, as arestas podem assumir tamanhos diferentes, conforme ilustrado na Figura 3.13, onde são exibidos (de forma salientada) quatro vértices conectados por três arestas. Dessa forma, a energia  $w(e_k)$  de cada aresta  $e_k$  é obtida pelo valor médio de suas energias ao longo da aresta, conforme descrito na Equação 3.10.

$$w(e_k) = \frac{1}{q_k} \sum_{j=1}^{q_k} \|\nabla I(x_j, y_j)\|, \quad (3.10)$$

onde  $q_k$  é o número de *pixels* ao longo de cada aresta e  $(x_j, y_j)$  representa a posição de cada *pixel* em coordenadas de imagem.

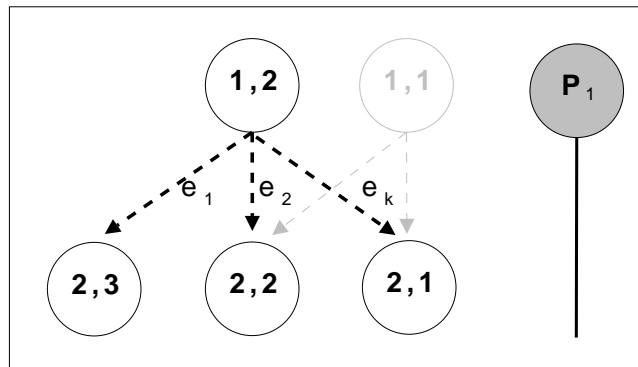


Figura 3.13: Ilustração da conexão entre níveis adjacentes de um grafo.

O cálculo de energia para cada aresta, apresentado nessa seção, leva em consideração a mudança de intensidade nos níveis de cinza de uma determinada imagem (usando a magnitude do gradiente), ou seja, onde pode haver a ocorrência de uma borda (ou contorno). Entretanto, esse cálculo não leva em consideração a inclinação desse contorno e sua distância em relação a algum critério determinado (como por exemplo, distâncias baseadas em antropometria). Dessa forma, qualquer borda saliente na imagem em análise poderia compor o contorno de uma determinada parte do corpo, mesmo que sua orientação seja completamente diferente da orientação esperada para essa determinada parte do corpo, assim como, não estabelecendo nenhuma relação de distância em função de seu respectivo

“osso”. Assim o contorno de uma determinada parte do corpo poderia estar sobre o seu “osso” gerador, ou muito distante dele, o que não é esperado em uma situação real. Na Seção 3.2.4 é apresentada uma forma alternativa para o cálculo de energia de cada aresta, onde são levados em consideração outros fatores, além da descontinuidade de valores da imagem em escala de cinza.

### 3.2.4 Mapa de Energias - com restrições antropométricas

O cálculo de energia para cada aresta do grafo pode envolver diversos fatores. Nesse trabalho, parte-se da hipótese que o contorno de uma determinada parte do corpo deve ter no mínimo duas relações fundamentais com seus respectivos “ossos”:

- Similaridade entre ângulos: o contorno deve estabelecer uma relação com o seu respectivo “osso”, de forma que tenham inclinação similar (salvo a cabeça, mãos e pés);
- Distância antropométrica: o contorno deve estar próximo à uma determinada região, estimada através de valores antropométricos, para cada parte do corpo;

#### Similaridade entre ângulos

Objetivando encontrar contornos com orientação similar a do seu respectivo “osso”, é modificada a Equação 3.10, do cálculo de energia de cada aresta, de forma que seja levado em consideração também a orientação de cada aresta, a direção do vetor gradiente (computado da imagem), e a orientação do “osso” associado à cada parte do corpo. Assim, o cálculo de energia  $w(e_k)$  de cada aresta  $e_k$  é definido através da Equação 3.11.

$$w(e_k) = \frac{1}{q_k} \sum_{j=1}^{q_k} |\mathbf{t}_k \cdot \nabla I(x_j, y_j)| |\mathbf{u} \cdot \mathbf{t}_k|, \quad (3.11)$$

onde  $\mathbf{t}_k$  é o vetor unitário normal ao segmento de reta que representa a aresta  $e_k$  e  $\mathbf{u}$  é o vetor unitário normal ao formado pelos pontos que compõem cada “osso” ( $\mathbf{P}_1$  e  $\mathbf{P}_2$ ). O primeiro termo da Equação 3.11 faz com que bordas da imagem alinhadas com a orientação da aresta do grafo tenham prioridade. O segundo termo prioriza arestas que tenham orientação similar à do seu respectivo “osso”. A Figura 3.14 ilustra os vetores  $\mathbf{t}_k$  e  $\mathbf{u}$ .

É importante salientar que o vetor  $\mathbf{u}$  (usado no cálculo da energia) pode estar associado a duas partes do corpo. Isso ocorre quando a energia é calculada para um determinado ponto que está localizado em uma região de conexão entre duas partes do corpo. Nesse caso,  $\mathbf{u}$  é um vetor unitário obtido pela soma entre os dois vetores (normalizada) associados a cada “osso” (ou parte do corpo). A Figura 3.15 ilustra esse procedimento para a conexão entre a canela e o pé direito de uma determinada pessoa em uma imagem.

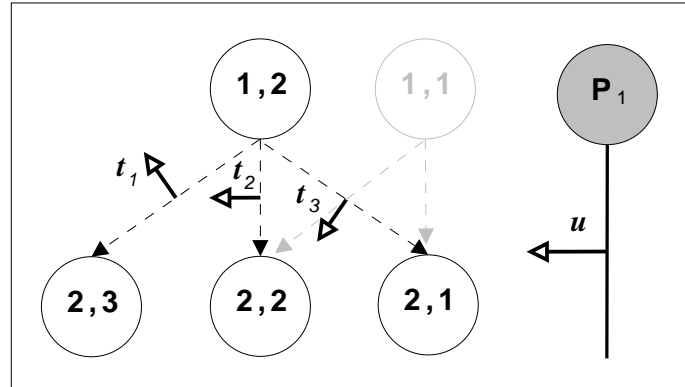


Figura 3.14: Ilustração dos vetores  $t_k$  e  $u$ , usados no cálculo do valor de energia de cada aresta do grafo (com restrição antropométrica).

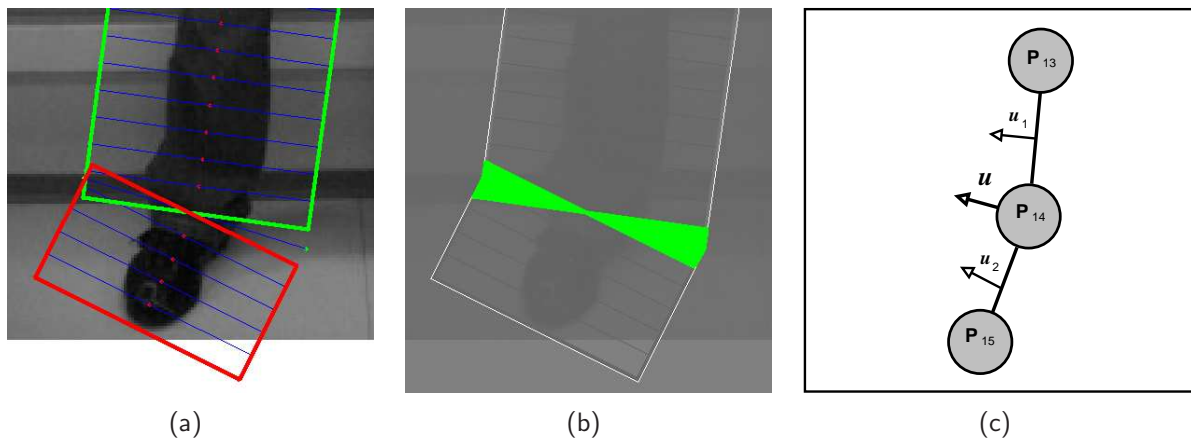


Figura 3.15: Em (a), intersecção ou buracos criados na região de conexão entre a canela e o pé direito. Em (b), pontos salientados (em verde) onde o vetor  $u$  está associado às duas partes do corpo. Em (c), ilustração do vetor resultante  $u$ , associado às duas partes em questão.

### Distância antropométrica

Objetivando estabelecer a segunda relação fundamental (listada no início da Seção 3.2.4), no cálculo de energia das arestas do grafo, é assumido que vértices muito próximos do seu respectivo “osso” podem não estar associados ao contorno de uma determinada parte do corpo, assim como vértices muito distantes do “osso” podem estar associados a contornos de outras estruturas (como o fundo da cena, por exemplo), que não o da parte do corpo em análise. Objetivando tratar esse tipo de problema, é criada uma restrição, baseada em valores antropométricos, para auxiliar no cálculo da estimativa do contorno de uma determinada parte do corpo, ilustrada com auxílio da Figura 3.16, para a canela direita de uma determinada pessoa em uma imagem.

A Figura 3.16(a) exibe uma determinada parte do corpo e seu respectivo “osso”, em verde. A Figura 3.16(b) exibe duas linhas paralelas ao “osso” ilustrado na Figura 3.16(a). Essas duas linhas paralelas (posição esperada do contorno) estão a uma distância  $d_2$  do “osso”, onde  $d_2$  é a metade da largura estimada para àquela determinada parte do corpo, ou seja,  $d_2 = w_i/2$ , onde  $w_i$  é a largura estimada por antropometria, conforme descrito na Tabela 3.1. A Figura 3.16(c) exibe uma imagem

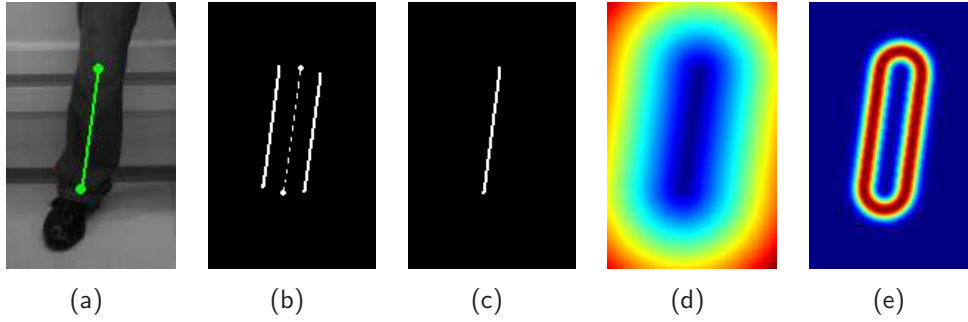


Figura 3.16: (a) Imagem de entrada e seu respectivo “osso” (em verde). (b) Região estimada para o contorno (linhas contínuas). (c) Imagem binária composta pela linha do “osso” apenas. (d) Transformada da Distância  $TD_i$  da imagem exibida em (c). (e) mapa resultante  $R_i$  de valores de distâncias, usado no cálculo da energia do contorno para essa parte do corpo.

binária, contendo apenas o segmento de reta associado ao “osso” dessa determinada parte do corpo. A Figura 3.16(d), exibe uma matriz  $TD_i$ , resultado da Transformada da Distância aplicada para a imagem ilustrada na Figura 3.16(c), associada a uma determinada parte do corpo  $i$ . A transformada da distância é uma operação normalmente realizada sobre uma imagem binária, produzindo uma imagem em escala de cinza cujas intensidades estão associadas às distâncias de um determinado ponto em relação ao ponto válido mais próximo da imagem binária (no nosso caso, em relação ao ponto mais próximo contido no segmento de reta associado ao “osso” - na Figura 3.16(d), pontos azuis representam distâncias pequenas e pontos vermelhos distâncias maiores).

O mapa final  $R_i$  (normalizado), utilizado como restrição antropométrica para o cálculo da energia do contorno para essa determinada parte do corpo, ilustrado na Figura 3.16(e), é descrito por uma função Gaussiana (o fator de escala da Gaussiana é dado por  $w_i/4$ ), onde o valor de cada *pixel*  $(x, y)$ , contido em  $R_i$ , é dado pela Equação 3.12. O procedimento descrito para essa determinada parte do corpo, ilustrado com auxílio da Figura 3.16, é repetido para cada parte do corpo, de forma que cada região possua suas próprias restrições de distâncias, baseadas em valores antropométricos.

$$R_i(x, y) = e^{-\frac{(TD_i(x, y) - w_i/2)^2}{(w_i/4)^2}} \quad (3.12)$$

Na Figura 3.16(e) pode-se perceber que pontos com nível máximo (valor igual a 1, em vermelho) são os locais mais prováveis de se encontrar o contorno dessa determinada parte do corpo, enquanto que pontos com valores mínimos (valor igual a 0, zero, em azul) indicam locais com probabilidade nula de existência de um contorno. Dessa forma, a energia de uma determinada aresta do grafo pode sofrer alguma “penalidade” por não estar coerente à essa medida de distância e ter seu valor decaído, fazendo com que pontos que satisfazem essa condição sejam favorecidos. Esse procedimento é realizado diretamente pela multiplicação, ponto a ponto, do *pixel* em análise por  $R_i$ , alterando consideravelmente seu valor. Assim, o peso final  $w(e_k)$  de cada aresta  $e_k$ , para uma determinada parte do corpo  $i$ , é dado pela Equação 3.13.



$$w(e_k) = \frac{1}{q_k} \sum_{j=1}^{q_k} |\mathbf{t}_k \cdot \nabla I(x_j, y_j)| |\mathbf{u} \cdot \mathbf{t}_k| R_i(x_j, y_j). \quad (3.13)$$

É importante salientar que o grafo é influenciado por partes do corpo adjacentes, nas regiões de conexão. Nessas regiões de conexão, como por exemplo, na conexão entre a canela e o pé direito (ilustrada na Figura 3.15(a-b)), a distância antropométrica é computada como uma média ponderada ( $R_{i'}$ ) entre os mapas de distâncias relacionados às duas partes do corpo em questão ( $i$  e  $p$ ). Essa ponderação é proporcional à distância que um determinado *pixel* está em relação a cada parte do corpo sendo analisada, definida na Equação 3.14. Dessa forma as distâncias antropométricas nas regiões de conexão apresentarão conexões suaves. A Figura 3.17(c) ilustra as distâncias antropométricas computadas para a conexão entre a canela e o pé direito.

$$R_{i'}(x, y) = aR_i(x, y) + (1 - a)R_p(x, y), \quad (3.14)$$

onde  $a$  é o fator de proximidade ( $a \in [0..1]$ , sendo “proximidade máxima” = 1) que o *pixel* ( $x, y$ ) está em relação à parte do corpo  $i$ ,  $R_i$  é a distância computada para o *pixel* ( $x, y$ ) usando os valores associados à parte do corpo  $i$  (conforme Equação 3.12) e  $R_p$  é a distância computada para o mesmo *pixel* usando os valores associados à parte do corpo  $p$  (conforme Equação 3.12). A Figura 3.17(a-b) ilustra o percentual de proximidade  $a$  de um determinado *pixel* ( $x, y$ ) em relação à parte do corpo  $i$  (canela).

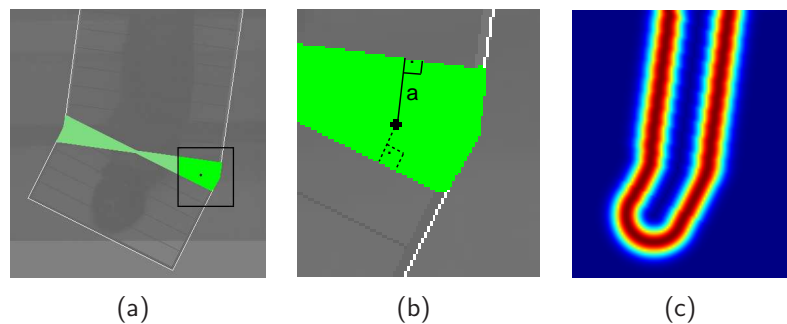


Figura 3.17: Ilustração da distância antropométrica em uma região de conexão entre duas partes do corpo. (a) *Pixel* em destaque, considerado em uma região de conexão. (b) Zoom na imagem (a) e ilustração do fator de distância  $a$  usado no cálculo de  $R_{i'}$ . (c) Mapa de distâncias antropométricas resultante para duas partes do corpo (canela e pé direito).

A caráter ilustrativo, a Figura 3.18 exhibe o mapa de magnitudes de uma determinada parte do corpo, sem ( $\|\nabla I\|$ ) e com ( $\|\nabla I\| R_i(x, y)$ ) a influência do mapa de distâncias antropométricas, respectivamente.

De uma forma geral, os procedimentos descritos até aqui descrevem basicamente como os grafos são criados para cada parte do corpo e como são conectados uns aos outros de forma a originar os três grafos principais (grafos  $A$ ,  $B$  e  $C$ ) usados nesse trabalho. Também foi descrita a forma na qual a energia de cada aresta do grafo é calculada, com a utilização de restrições de ângulos e

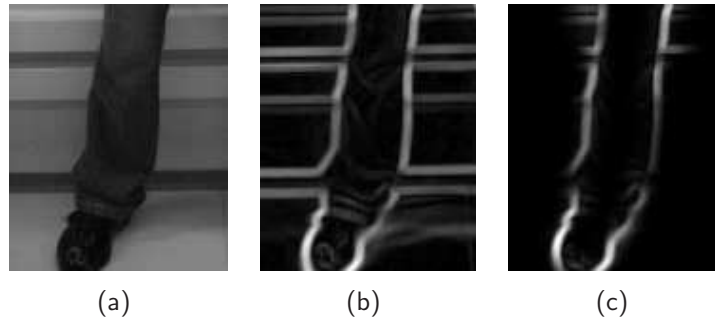


Figura 3.18: Ilustração da influência do mapa de distâncias antropométricas. Em (a), imagem em escala de cinza da canela e pé direito. Em (b), magnitude do gradiente da imagem (a). Em (c), magnitude do gradiente da imagem (a), multiplicada, ponto a ponto, pelo mapa de distâncias antropométricas associado (ilustrado na Figura 3.17(c)).

distâncias antropométricas, que são a base para a obtenção do contorno de uma pessoa em uma imagem, a partir da utilização do modelo proposto. A seguir é detalhado o método usado para a obtenção do contorno, com a utilização de programação dinâmica.

### 3.2.5 Programação Dinâmica

Conforme definido em [32], *programação dinâmica* (ou *dynamic programming*), assim como o método de *dividir para conquistar*, são formas de solucionar problemas combinando suas soluções em subproblemas. Entretanto, o método *dividir para conquistar* particiona o problema em subproblemas independentes, resolve os subproblemas de forma recursiva e combina suas soluções para resolver o problema original. Por outro lado, *programação dinâmica* é utilizada quando os subproblemas não são independentes, ou seja, quando os subproblemas compartilham subproblemas. Um algoritmo de *programação dinâmica* resolve cada subproblema apenas uma vez e salva sua solução em uma “tabela”, evitando o recálculo toda vez que o subproblema em questão estiver envolvido.

*Programação dinâmica* é tipicamente aplicada a problemas de otimização, onde tais problemas podem assumir diversas soluções. Nesses casos, cada solução está associada a um valor, e o objetivo final é encontrar a solução com valor ótimo (mínimo ou máximo). Usualmente, define-se “uma” solução ótima ao invés de “a” solução ótima, uma vez que pode haver diversas formas de se encontrar o valor ótimo.

Como descrito no início da Seção 3.2, o modelo de segmentação de pessoas proposto nessa tese baseia-se na definição de três grafos principais ( $A$ ,  $B$  e  $C$ ), onde o objetivo é encontrar o melhor caminho para cada grafo (do nível 1 ao nível  $M$ ), que maximiza o valor de energia de suas arestas. O algoritmo de *Dijkstra* [32] é utilizado para encontrar o caminho de menor custo em grafos direcionados e ponderados, onde todas as arestas possuem pesos não negativos. No modelo proposto, esse algoritmo sofreu uma pequena alteração, de modo que o critério utilizado seja o oposto, ou seja, o melhor caminho do grafo é aquele que maximiza o valor de energia de suas arestas (caminho de maior custo). Salienta-se que o grafo é acíclico, de modo que os cálculos do



custo máximo e mínimo são semelhantes.

Considere o vértice  $S_{2,2}$  do grafo  $A$ , descrito na Seção 3.2.2, e ilustrado na Figura 3.19. Os três caminhos possíveis para se chegar a  $S_{2,2}$  são:  $E(S_{1,3}, S_{2,2})$ ,  $E(S_{1,2}, S_{2,2})$  e  $E(S_{1,1}, S_{2,2})$ , referenciados também pelas arestas  $e_1$ ,  $e_2$  e  $e_3$ , cada qual com seu respectivo valor de energia (assumindo-se que o número de caminhos possíveis  $k$ , foi setado com base em experimentos para  $k = 3$ , conforme descrito na Seção 3.2.1).

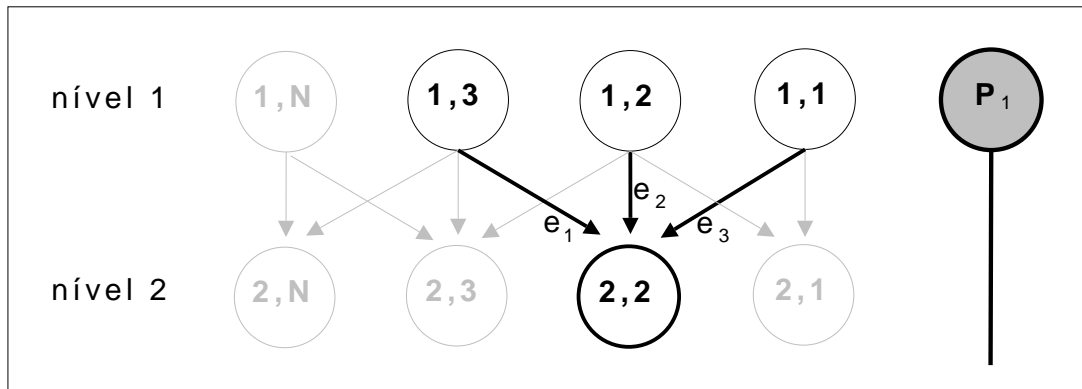


Figura 3.19: Ilustração das alternativas de caminhos, do nível 1 ao nível 2, passando pelo vértice  $S_{2,2}$ .

Dessa forma, o melhor caminho para se chegar ao nível 2 (nível do vértice  $S_{2,2}$ ), vindo do nível 1, e passando por  $S_{2,2}$ , é através da aresta  $e_k$  que possui maior valor de energia. Esse procedimento é realizado para todos os vértices do nível 2. Então, para cada vértice  $n$  do nível 2 é armazenado, em uma tabela, o melhor caminho para se chegar até ele e o valor de energia de sua aresta  $e_k$ . A verificação do melhor caminho para se chegar aos vértices dos próximos níveis segue a mesma lógica, porém, levando em consideração os valores de energias acumuladas para chegar aos níveis anteriores. Dessa forma, a obtenção do melhor caminho não é baseada apenas em uma decisão local, pois também leva em consideração as energias acumuladas. Ao chegar ao último nível, o maior valor de energia acumulado representará o melhor caminho desse grafo, que liga o nível 1 ao nível  $M$ . Assim, o caminho que resultou esse valor acumulado pode ser resgatado e compor o contorno de uma determinada parte do corpo.

É importante salientar que o caminho obtido para cada grafo principal ( $A$ ,  $B$  e  $C$ ) é computado a partir do primeiro nível até o último nível do grafo, fazendo com que os braços sejam computados um após o outro (assim como as pernas são computadas separadamente, uma vez que estão em grafos distintos,  $B$  e  $C$ , respectivamente). Dessa forma, mesmo que os braços (ou pernas) estejam cruzados (visualmente sobrepostos, conforme ilustrado com auxílio da Figura 3.20), as arestas associadas ao braço (ou perna) esquerdo não estão conectadas diretamente ao braço (ou perna) oposto devido à estrutura do grafo, organizada em níveis.

A Figura 3.21(b-d) ilustra o melhor caminho obtido para cada grafo principal ilustrado na Figura 3.21(a).

Os três caminhos obtidos, gerados para cada grafo principal, (ilustrados na Figura 3.21) são independentes, ou seja, não geram um contorno fechado para o corpo da pessoa. Para gerar um

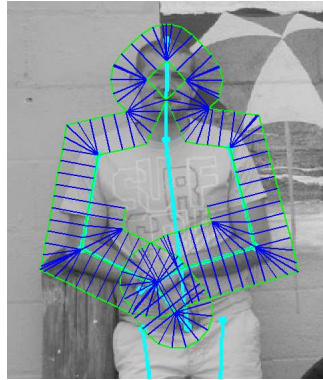


Figura 3.20: Ilustração de grafos sobrepostos. Nesse caso, as arestas associadas ao braço esquerdo não estão conectadas diretamente ao braço oposto devido à estrutura do grafo.

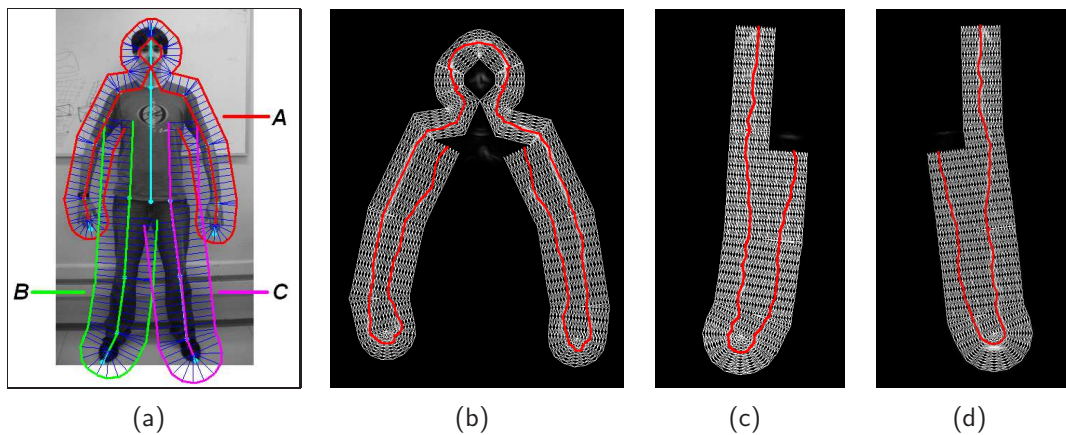


Figura 3.21: Ilustração dos caminhos gerados para cada grafo principal. (a) Três grafos principais (*A*, *B* e *C*) gerados para uma determinada pessoa em uma imagem. (b) Caminho gerado para o grafo da parte superior do corpo (grafo *A*). (c) Caminho gerado para o grafo da parte inferior direita do corpo (grafo *B*). (d) Caminho gerado para o grafo da parte inferior esquerda do corpo (grafo *C*).

contorno fechado esses caminhos devem ser conectados uns aos outros. A Figura 3.22(a) ilustra os três caminhos obtidos para uma determinada pessoa em uma imagem, sobrepostos na imagem de entrada. O objetivo agora é conectar esses três caminhos, de forma que haja um ponto de conexão que ligue o contorno do braço direito ao início do tronco; outro ponto de conexão que ligue o contorno do braço esquerdo de maneira similar ao braço direito e por último, um ponto de conexão que ligue os contornos na parte interna das pernas, conforme ilustrado na Figura 3.22(b), originando um contorno fechado. A forma na qual é realizada esse procedimento é descrita a seguir.

#### Conectando os contornos dos braços ao tronco

Os contornos dos braços são conectados ao tronco de forma independente, ou seja, o procedimento descrito a seguir é realizado para o braço esquerdo e para o braço direito. Para cada braço, é verificado qual ponto do seu contorno interno está mais próximo do início do contorno do tronco

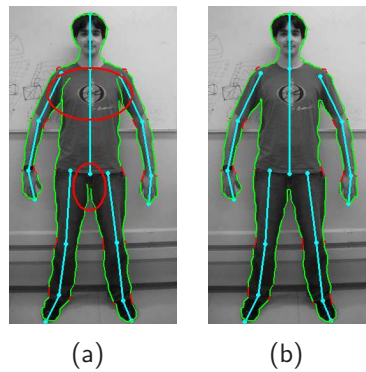


Figura 3.22: Ilustração da conexão entre braços×tronco e parte interna das pernas. (a) Caminhos gerados pelos três grafos principais sobrepostos na imagem de entrada (elipses indicam regiões do contorno que devem ser conectadas, para originar um contorno fechado). (b) Caminhos gerados pelos três grafos principais, conectados uns aos outros, após tratamento especial (resultando em um contorno fechado).

associado (ilustrados na Figura 3.23 por um sinal de “+” em amarelo e branco, respectivamente). Os pontos dos contornos dos braços a partir dos pontos de conexão são desconsiderados. Assim, o contorno de cada braço é conectado ao contorno do tronco por esses pontos de conexão, conforme ilustrado na Figura 3.22(b).

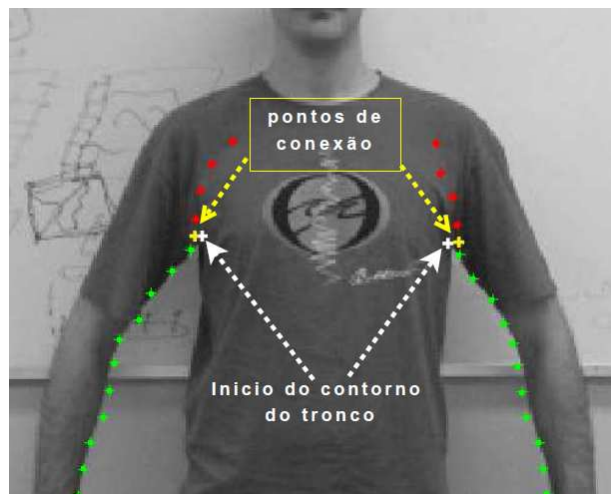


Figura 3.23: Ilustração dos pontos de conexão entre os contornos dos braços×tronco.

#### Conectando os contornos da parte interna das pernas

Para conectar o contorno interno da perna direita ao contorno interno da perna esquerda, originando assim um contorno fechado da pessoa na imagem, é efetuado o seguinte procedimento. Inicialmente calcula-se a distância Euclidiana de todos os pontos do contorno interno da coxa esquerda em relação a todos os pontos do contorno interno da coxa direita. O ponto que obtiver a menor distância é considerado ponto de conexão. Se houver mais de um ponto de conexão (ou seja,

mais de um ponto com a mesma distância mínima, ilustrados em vermelho na Figura 3.24), então o ponto mais próximo de  $P_{13}$  (associado ao modelo de esqueleto) é considerado o ponto de conexão (salientado por uma seta na Figura 3.24). A Figura 3.22(b) ilustra o contorno resultante para uma imagem exemplo, após a conexão feita na parte interna das coxas.

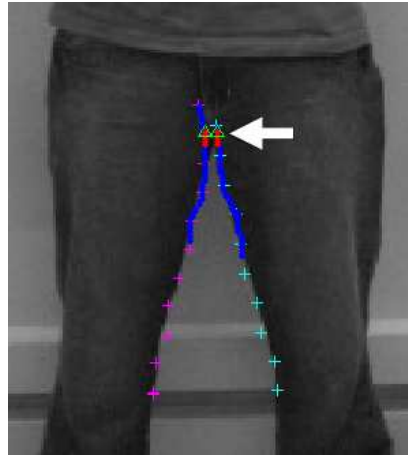


Figura 3.24: Ilustração dos pontos de conexão entre os contornos da parte interna das pernas.

Os procedimentos descritos até aqui resultam em um contorno fechado para uma pessoa em uma imagem, gerados a partir do modelo proposto nessa tese. Uma característica importante do contorno gerado é que possui informação semântica associada, ou seja, cada ponto do contorno está associado a uma determinada parte do corpo. Tal informação semântica torna possível, por exemplo, que duas partes do corpo fiquem sobrepostas (como os braços na frente do tronco), mantendo ainda uma conectividade coerente do contorno (uma vez que se sabe quais partes do grafo estão associadas a quais partes do corpo e suas respectivas regiões de adjacência). Essa informação semântica pode ser utilizada para diversos fins e é ilustrada nos resultados experimentais por um contorno vermelho, que separa duas partes adjacentes do corpo (podendo ser observada na Figura 3.22(b)). Na Seção 3.3 é apresentada uma abordagem alternativa para o cálculo de energia das arestas, onde é considerada também informação de cor (cor RGB do *pixel*, por exemplo), contida na grande maioria das imagens digitais atuais.

### 3.3 Inserindo informação de cor no modelo

O método apresentado na Seção 3.2 descreve, de uma forma geral, a obtenção do contorno de uma pessoa em uma imagem como sendo um caminho em um determinado grafo (mais especificamente três caminhos - um para cada grafo principal ( $A$ ,  $B$  e  $C$ )). Esses caminhos foram obtidos por um algoritmo que avalia o valor de energia das arestas desses grafos, onde uma determinada condição deve ser satisfeita (aquela que maximiza o valor de suas energias). Entretanto, os procedimentos descritos na Seção 3.2 não levam em consideração informações de cor para calcular o valor de energia das arestas desses grafos (cor RGB do *pixel*, por exemplo, contida na grande maioria das imagens digitais atuais).

Nessa seção é apresentada uma variação do modelo de segmentação por cores, proposto por Jacques Junior e sua equipe [3], de maneira que a energia das arestas desses grafos (descritos na Seção 3.2) também inclua tal característica (informação de cor). Resumidamente, a ideia dessa abordagem é criar, em um primeiro estágio, um modelo de cor para algumas partes do corpo (etapa de aprendizado da cor) e em um momento posterior avaliar a similaridade de cada *pixel* da imagem (em uma determinada região de busca) em relação ao seu modelo de cor associado (etapa de confronto). O resultado desse processo irá criar um mapa de distâncias, onde pode-se determinar quais regiões da imagem possuem cores similares às encontradas no corpo (ou roupas) da pessoa na imagem. A Figura 3.25 ilustra parte desse processo, onde as regiões de aprendizado e busca de uma determinada cor são ilustradas na Figura 3.25(a) e *pixels* da imagem com cores similares ao modelo de cor daquela determinada parte do corpo (possuindo distâncias pequenas) são ilustrados na Figura 3.25(b), em tons de azul.

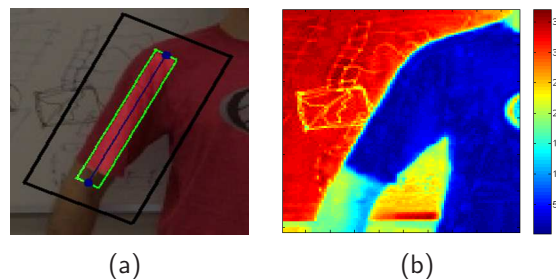


Figura 3.25: Ilustração da segmentação por cores. Em (a), é ilustrado a região de aprendizado (retângulo verde) e busca (retângulo preto) da cor de uma determinada parte do corpo (braço direito). Em (b), *pixels* da imagem com baixa e alta similaridade em relação à cor aprendida (em vermelho e azul, respectivamente) são ilustrados.

Diferentemente do modelo proposto em [3], onde é criado um modelo de cor para *pixels* com tons de pele e outro para *pixels* associados à região do tronco da pessoa na imagem, agora é criado um modelo de cor para cada parte do corpo da pessoa na imagem (com exceção dos ombros), como descrito em detalhes nas Seções 3.3.1 e 3.3.2.

Deve ser salientado que uma análise sobre qual espaço de cor melhor se adapta à resolução desse problema (segmentação de pessoas em imagens estáticas) não é objetivo principal dessa tese (apesar de que isso possa ser feito em estudos futuros), sendo adotado o espaço de cor RGB pelo fato de ser amplamente utilizado em processamento de imagens digitais. A seguir são descritos os métodos para aprendizagem (Seções 3.3.1 e 3.3.2) e busca das cores predominantes (Seção 3.3.3). A forma na qual a informação de cor é utilizada para auxiliar no cálculo de energia das arestas dos grafos é descrita na Seção 3.3.4.

### 3.3.1 Aprendendo o modelo de cor

O objetivo dessa etapa é criar um modelo de cor para cada parte do corpo (dados dois pontos de controle que formam o “osso” de cada parte, associados ao modelo de esqueleto) que posteriormente

será usado para avaliar a cor de cada *pixel* na imagem (em uma determinada região de busca), utilizando uma determinada medida de distância. O objetivo desse procedimento é associar partes do corpo em análise à *pixels* com distâncias pequenas (em relação à seus respectivos modelos de cor), gerando assim informações relevantes, utilizadas no modelo de segmentação proposto nessa tese. Basicamente 14 partes do corpo são utilizadas nesse processo, dados seus pontos de controle (são elas: cabeça, tronco, braço direito e esquerdo, antebraço direito e esquerdo, mão direita e esquerda, coxa direita e esquerda, canela direita e esquerda, pé direito e esquerdo). As únicas duas partes do corpo associadas ao modelo de esqueleto, descrito na Seção 3.1, que não passam por esse processo de aprendizado e busca de cores predominantes (ao menos diretamente) são os ombros (direito e esquerdo). Isso ocorre, porque a região de busca do tronco quando combinada com a região de busca dos braços (conforme descrito na Seção 3.3.4) é suficiente para englobar as regiões dos ombros, fazendo com que essa parte do corpo seja inserida de forma indireta nessa abordagem.

Inicialmente é definida uma região de aprendizado  $Tr_i$  (ilustrada na Figura 3.25(a) por um retângulo verde) ao redor de cada parte do corpo  $i$ . Essa região será usada para o aprendizado da cor (ou cores) predominante de cada parte do corpo. A região de aprendizado é um retângulo, com eixo central igual à da parte do corpo correspondente, que idealmente deveria conter apenas *pixels* relacionados àquela parte do corpo. Na abordagem proposta, o comprimento desse retângulo é igual à distância entre os dois pontos que definem cada parte do corpo (“osso”), e sua largura é uma fração  $s_1$  (setada experimentalmente para 0.4) da largura esperada para aquela parte do corpo  $w_i$ , conforme definido na Tabela 3.1. É importante salientar que a largura  $w_i$  de cada parte do corpo é usada ao invés da distância 2D entre os dois pontos que formam cada “osso” (que poderia ser calculada diretamente em coordenadas de imagem). De fato, usar diretamente a distância dos pontos 2D de cada parte do corpo que não está no plano da imagem (por exemplo, braços paralelos ao chão) pode levar a resultados incoerentes, devido à questões de perspectiva, como ilustrado com auxílio da Figura 3.26.

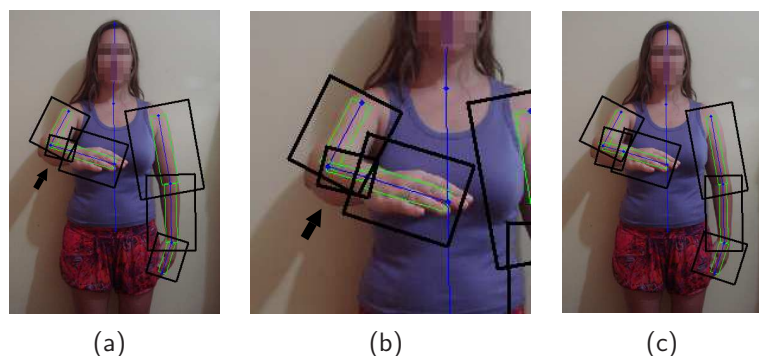


Figura 3.26: Diferentes formas de setar as larguras das regiões de aprendizado (retângulos verdes) e busca (retângulos pretos) das cores predominantes. (a) Larguras estimadas diretamente a partir das distâncias entre os dois pontos que formam cada “osso”. (b) Zoom na imagem ilustrada em (a). (c) Larguras estimadas a partir dos valores esperados para aquela parte do corpo  $w_i$ .

Na Figura 3.26 são ilustradas regiões de aprendizado (retângulos verdes) e busca (retângulos



pretos) das cores predominantes dos membros superiores de uma determinada pessoa, definidas a partir de duas formas. Na Figura 3.26(a-b) foram usadas diretamente as distâncias 2D dos pontos de cada parte do corpo (ou seja, as larguras das regiões de aprendizado e busca são uma fração das distâncias entre os pontos que formam cada “osso”), enquanto que na Figura 3.26(c) as larguras das regiões de aprendizado e busca são definidas com base nas larguras esperadas para aquelas partes do corpo (com base em  $w_i$ ). Pode-se observar, na Figura 3.26(a), que regiões simétricas (como braço esquerdo e direito) podem possuir larguras diferentes. Também é ilustrado na Figura 3.26(a-b), que a região de busca do antebraço direito (salientada por uma seta) ficou muito pequena, fazendo com que essa determinada parte do corpo não seja corretamente avaliada, pois parte do braço ficou fora da região de busca. Conforme será detalhado na Seção 3.3.3, as larguras das regiões de busca e aprendizado das partes do corpo são definidas de forma similar, porém com proporções diferentes em função de  $w_i$ . Dessa forma, não se teria como solucionar o problema de estimar a largura de uma parte do corpo diretamente a partir da distância entre os dois pontos que a formam, ilustrado com auxílio da Figura 3.26(a-b), com a utilização de um único critério, uma vez que as distâncias entre os dois pontos que originam cada parte do corpo dificilmente serão iguais. Por exemplo, ao aumentar a largura do antebraço direito, usando uma fração da distância 2D entre os dois pontos que o formam, poder-se-ia solucionar parte do problema, porém a largura do antebraço esquerdo seria aumentada nas mesmas proporções (fazendo com que ficasse ainda mais exagerada). Já na Figura 3.26(c), onde as larguras das regiões foram estimadas a partir da largura esperada para aquela parte do corpo ( $w_i$ ), pode-se observar que as regiões são visualmente mais coerentes (assumindo valores iguais para partes simétricas do corpo), suavizando o problema gerado pela perspectiva da câmera.

Para obter as cores predominantes de cada parte do corpo, é inicialmente utilizado um algoritmo não supervisionado para segmentação de imagens [14], com o objetivo de se obter as principais regiões contidas em  $Tr_i$  (similarmente à abordagem descrita em [3]), como ilustrado na Figura 3.27(c-e). Na maioria dos casos, a maior dessas regiões está relacionada com a cor predominante. Entretanto, há alguns casos (camisetas com escritas, ilustrações, sombras, etc) onde a maior região segmentada pode não estar relacionada com a cor predominante. Objetivando tratar esse tipo de problema, as  $N_r$  maiores regiões segmentadas contidas em  $Tr_i$ , com valor de área maior que um determinado valor de limiar  $T_a$  são mantidas (valores setados através de experimentos:  $N_r = 3$  e  $T_a = 0.1 \#Tr_i$ , onde  $\#Tr_i$  é a área de  $Tr_i$ ). Na Figura 3.27(f), pode-se observar que algumas pequenas regiões segmentadas (visíveis na Figura 3.27(e), por exemplo) foram eliminadas pelo limiar de área.

Para uma determinada parte do corpo, consideram-se as  $N_j \leq N_r$  maiores regiões segmentadas que satisfazem o critério de área mínima (diferentemente de [3], onde é criado um modelo de cor apenas para a maior região). A distribuição de cores dentro de cada região segmentada é representada por um Modelo Gaussiano Multivariado (MGM - *Multivariate Gaussian Model*), que requer o cálculo de um vetor média ( $\mu_{ij}$ ) e uma matriz de covariância ( $C_{ij}$ ,  $3 \times 3$ ), onde  $1 \leq j \leq N_j$  está relacionado a cada diferente modelo, criado para uma determinada região segmentada, associado

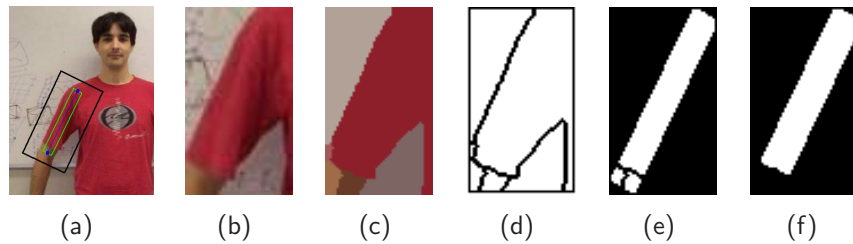


Figura 3.27: Ilustração da segmentação inicial usada no estágio de aprendizado do modelo de cor. (a) “Osso”, região de aprendizado e busca de uma determinada parte do corpo. (b) Imagem de entrada. (c) Segmentação inicial, usando [14]. (d) Contornos da segmentação inicial. (e) Região de aprendizado inicial. (f) Região de aprendizado final.

à uma específica parte do corpo  $i$ . Esses parâmetros  $\mu_{ij}$  e  $C_{ij}$  são obtidos com a utilização de um algoritmo que remove a influência de possíveis *outliers* [39].

Uma vantagem de se utilizar um modelo Gaussiano simples é que pode-se verificar facilmente quando uma determinada amostra se adequa ao modelo, usando a distância de *Mahalanobis*, o que se torna mais complicado ao utilizar modelos mais complexos (como GMM – *Gaussian Mixture Models*, por exemplo).

### 3.3.2 Aprendendo o modelo de cor com a utilização de *PCA - Principal Component Analysis*

Na Seção 3.3.1 foi apresentada uma proposta para criação de um modelo de cor associado à uma determinada parte da imagem (ou parte do corpo). Porém, em algumas situações o modelo de cor criado pode não representar de forma adequada a cor predominante em uma determinada região, mais especificamente quando essa região possui valores muito homogêneos em todas suas camadas de cores (RGB), por exemplo, para regiões de cor preta, representada por  $R = 0$ ,  $G = 0$  e  $B = 0$ , ou branca, representada por  $R = 255$ ,  $G = 255$  e  $B = 255$  (no espaço de cor RGB), onde a variação dos 3 canais de cor é muito baixa ou as vezes chegando a zero. Na prática, isso acaba ocorrendo quando os canais de cor (ou dois deles) são bastante correlacionados. Nesses casos, o determinante da matriz de covariância gerada pode ser próximo de zero ou igual a zero, indicando que essa matriz não pode ser invertida, ou sua inversa pode ser instável numericamente. Um problema associado à impossibilidade da matriz ser invertida é que esse procedimento é necessário no cálculo da distância de *Mahalanobis*, conforme definido na Equação 3.18, usada para verificar a coerência entre um determinado *pixel* e o modelo de cor estimado.

A análise dos componentes principais (*Principal Component Analysis – PCA*) é utilizada para reduzir a dimensão do modelo de cor gerado (para 2D ou 1D), através de uma transformação linear que combina a informação dos três canais de cores. Essa transformação geralmente é efetuada quando os canais de cor (ou dois deles) são bastante correlacionados.

Na análise dos componentes principais (*PCA*), são calculados os autovalores e autovetores da matriz de covariância criada. O problema mencionado acima, sobre o determinante da matriz de covariância ser quase nulo, normalmente está relacionado com alguns autovalores muito pequenos,



quando comparados com os demais. A matriz de covariância representa um elipsóide, e os tamanhos dos semi-eixos (autovalores) representam a variância ao longo da direção correspondente (autovetor). Se um dos autovalores é muito pequeno, então o elipsóide fica achatado (e o determinante próximo de zero).

Considere  $\lambda_k$  e  $\mathbf{v}_k$  os autovalores e autovetores, respectivamente, da matriz de covariância  $\mathbf{C}_{ij}$ , onde  $\lambda_1 \geq \lambda_2 \geq \lambda_3$ . A proporção do autovalor  $\lambda_k$  sobre os demais é representada por  $r_k$ , definida na Equação 3.15.

$$r_k = \frac{\lambda_k}{\sum_{t=1}^3 \lambda_t}, \quad (3.15)$$

para  $k = 1, 2, 3$ . Assim, a direção que possuir uma representatividade muito pequena, apontada por  $\mathbf{v}_k$ , é desconsiderada, ou seja, com valor de  $r_k$  menor que um determinado valor de limiar  $T_p$  (setado experimentalmente para  $T_p = 0.001$ ). Mais especificamente, uma dimensão  $k$  do modelo de cor é eliminada se

$$r_k < T_p, \quad (3.16)$$

para  $1 \leq k \leq 3$ . Dessa forma, a dimensão do modelo de cor criado pode permanecer 3D, caso a representatividade associada a cada autovalor seja maior que o valor mínimo definido  $T_p$ , ou então pode-se criar um modelo 2D ou até 1D.

Quando uma determinada dimensão é removida, o novo modelo de cor (2D ou 1D) é criado da seguinte forma:

- **Redução para 2D:** apenas uma dimensão é removida do modelo original ( $\mathbf{C}_{ij} 3 \times 3$ ), associada ao menor valor de  $r_k$  que não satisfaz a condição do valor mínimo  $T_p$ . Um novo vetor 2D é criado ( $\mathbf{F}$ ) com os mesmos *pixels* utilizados para a criação do modelo original (modelo 3D), porém com apenas os componentes que irão compor o novo modelo de cor (associados aos autovetores  $\mathbf{v}_k$ , computados na análise dos componentes principais). Consideram-se mesmo *pixels* aqueles que não foram removidos por serem considerados *outliers*, através da utilização do algoritmo proposto em [39]. Assim, cada *pixel* com cor  $\mathbf{c} = (R, G, B)^T$  desse novo vetor, terá seu valor transformado como descrito na Equação 3.17.

$$\mathbf{c}' = (F_1, F_2, F_3) = \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \mathbf{v}_3^T \end{bmatrix} \mathbf{c} \quad (3.17)$$

A criação do novo modelo de cor 2D requer um novo cálculo de um vetor média  $\boldsymbol{\mu}_{ij}$  e uma nova matriz de covariância  $\mathbf{C}_{ij}$  (agora com dimensão  $2 \times 2$ ), gerados a partir do novo vetor de características  $\mathbf{F} = (F_1, F_2)$ , de maneira similar à descrita anteriormente, utilizando o algoritmo proposto em [39].

- **Redução para 1D:** duas dimensões são removidas do modelo original ( $3 \times 3$ ), associada aos dois menores valores de  $r_k$  que não satisfazem a condição do valor mínimo  $T_p$ . De maneira similar à redução para o 2D, um novo vetor 1D é criado ( $\mathbf{F}$ ) com os mesmos *pixels* utilizados para a criação do modelo original, porém com apenas os componentes que irão compor o novo modelo de cor. Assim, cada *pixel* com cor  $\mathbf{c} = (R, G, B)^T$  desse novo vetor, terá seu valor transformado como descrito na Equação 3.17.

A criação do novo modelo de cor 1D requer o cálculo de um valor de média ( $\mu_{ij}$ ) e um valor de desvio padrão ( $\sigma_{ij}$ ), gerados a partir do novo vetor 1D de características  $\mathbf{F} = (F_1)$ .

Deve-se salientar que, tanto para a redução para o 2D como para 1D, devem ser armazenados também seus respectivos autovetores  $\mathbf{v}$  para que a mesma transformação feita na etapa da criação do modelo de cor também possa ser feita na etapa de confronto (como descrito na Seção 3.3.3).

### 3.3.3 Confrontando o modelo de cor

Para localizar *pixels* relacionados a uma determinada parte do corpo  $i$  (após a criação de um modelo de cor para essa parte), uma região de busca  $Te_i$  é definida. Diferentemente da região de aprendizado  $Tr_i$ , a região de busca deve ser grande o suficiente para englobar todos os *pixels* relacionados a cada parte do corpo.

De uma forma geral, a região de busca  $Te_i$  é uma versão aumentada da região de aprendizado  $Tr_i$ . O comprimento da região de busca é igual ao comprimento da região de aprendizado  $Tr_i$ , multiplicado por um fator  $s_2$  (setado experimentalmente para  $s_2 = 1.15$ ). A largura de  $Te_i$  é baseada em valores estimados para cada parte do corpo (descritos na Tabela 3.1) multiplicados também por outro determinado fator  $s_3$  (setado experimentalmente para  $s_3 = 2$ , para poder suportar pessoas com diferentes tamanhos, ou com larguras acima da média – esse fator pode ser aumentado para englobar partes do corpo de pessoas muito obesas ou que estão vestindo roupas muito largas, por exemplo). A Figura 3.25(a) ilustra os pontos de controle (que compõem o “osso”) para uma determinada parte do corpo (pontos azuis), a região de aprendizado  $Tr_i$  usada para o aprendizado do modelo de cor (retângulo verde) dessa determinada parte do corpo e sua respectiva região de busca  $Te_i$  (retângulo preto).

Para calcular a coerência entre o modelo de cor de uma determinada parte do corpo  $i$  (e modelo de cor  $j$ ) e a cor de um *pixel* da região de busca  $Te_i$ , é utilizada a distância quadrada de *Mahalanobis*. Para cada *pixel* com cor  $\mathbf{c}$ , a distância quadrada de *Mahalanobis*  $D_{ij}^2(\mathbf{c})$  é obtida através da Equação 3.18 (nos casos 3D e 2D).

$$D_{ij}^2(\mathbf{c}) = (\mathbf{c} - \boldsymbol{\mu}_{ij})^T \mathbf{C}_{ij}^{-1} (\mathbf{c} - \boldsymbol{\mu}_{ij}), \quad 1 \leq j \leq N_i \quad (3.18)$$

Para que um *pixel* da região de busca (no espaço de cor RGB) possa ser confrontado com o modelo de cor 2D, sua cor  $\mathbf{c} = (R, G, B)^T$  deve ser transformada para 2D ( $\mathbf{c}' = (F_1, F_2)^T$ ). Essa transformação é feita de maneira similar à utilizada na criação do modelo de cor, ou seja, com a

utilização do autovetor  $v$ , computado na análise dos componentes principais (*PCA*), como definido na Equação 3.17.

Nos casos onde o modelo de cor é reduzido para 1D, sua cor  $c = (R, G, B)^T$  deve ser transformada para 1D ( $c' = (F_1)$ ). Essa transformação é feita de maneira similar à utilizada na redução para o 2D, ou seja, com a utilização do autovetor  $v_1$ , computado na análise dos componentes principais (*PCA*), como definido na Equação 3.17. A distância de *Mahalanobis* para o caso 1D é definida na Equação 3.19.

$$D_{ij}^2(c') = \frac{(c' - \mu_{ij})^2}{\sigma_{ij}^2}, \quad (3.19)$$

onde  $\mu_{ij}$  e  $\sigma_{ij}$  são a média e desvio padrão, respectivamente, para uma determinada parte do corpo  $i$  (e modelo de cor  $j$ ), do modelo de cor 1D.

O confronto entre o modelo de cor criado para uma determinada parte do corpo  $i$  e uma região de busca  $Te_i$  gera um mapa de distâncias  $D_{ij}^2$ . A Figura 3.25(b) ilustra um mapa de distâncias de *Mahalanobis*, gerado a partir do confronto entre uma região de busca e um modelo de cor criado a partir da região definida na Figura 3.25(a) (também ilustrada na Figura 3.27(f)). Pode-se verificar, na Figura 3.25(b), que *pixels* com distâncias muito pequenas (em azul) possuem cor semelhante à da região usada para criação do modelo de cor para essa determinada parte do corpo (como ilustrado com auxílio da Figura 3.25(a)). Também é possível verificar na Figura 3.25(b) que o mapa de distâncias de *Mahalanobis* gerados para essa determinada parte do corpo possui uma característica importante: as regiões associadas ao modelo de cor (exibidas em tons de azul) possuem alto contraste em relação às outras partes da imagem. Essas regiões de descontinuidades geralmente oferecem informações relevantes para a obtenção do contorno de uma pessoa em uma imagem, como descrito na Seção 3.3.4.

### 3.3.4 Mapa de Energias com informação de cor

Os mapas de distâncias de *Mahalanobis*  $D_{ij}$  gerados na etapa de confronto dos modelos de cor com as regiões de busca, para cada parte do corpo (descritos na Seção 3.3.3) também podem ser utilizados para auxiliar no cálculo do valor de energia das arestas dos grafos (descrito na Seção 3.2). Conforme descrito na Seção 3.2.3, o gradiente da imagem em escala de cinza  $\nabla I$  foi utilizado como atributo para avaliar o valor de energia das arestas, assim como na Seção 3.2.4 outros fatores foram utilizados, como por exemplo, restrições de ângulos e distâncias antropométricas.

A ideia agora é criar um único mapa distâncias de *Mahalanobis*  $D_i$  para cada parte do corpo, e a partir desse mapa, computar o seu gradiente  $\nabla D_i$  e combiná-lo com o gradiente gerado a partir da imagem em escala de cinza  $\nabla I$ . Esse procedimento irá compor uma nova matriz de gradientes  $\nabla I'$ , onde serão inseridas informações de luminosidade e cores predominantes do objeto de desejo, servindo como dado de entrada para o modelo descrito na Seção 3.2. A Equação 3.20 define o valor do gradiente de um *pixel*  $(x, y)$  para essa nova matriz de gradientes.

$$\nabla I'(x, y) = \frac{1}{2}(\nabla I(x, y) + \nabla D_i(x, y)), \quad (3.20)$$

onde as magnitudes do gradiente  $\|\nabla I\|$  e  $\|\nabla D_i\|$  estão normalizados no intervalo  $[0,1]$ .

A Figura 3.28 ilustra parte desse processo. Considere a imagem exibida na Figura 3.28(a) como sendo a parte do corpo sendo analisada (no caso, o braço direito) e sua respectiva imagem em escala de cinza na Figura 3.28(b). As distâncias de *Mahalanobis* obtidas para essa parte do corpo, com a utilização do modelo descrito na Seção 3.3.3, são ilustradas na Figura 3.28(c) (onde valores escuros representam distâncias pequenas e valores claros o contrário). As Figuras 3.28(d) e (e) representam a magnitude do gradiente (normalizado) das imagens ilustradas nas Figuras 3.28(b) e (c), respectivamente. A Figura 3.28(f) ilustra a magnitude do gradiente ( $\|\nabla I'\|$ ) obtido pela média (definida na Equação 3.20) dos gradientes usados para gerar as Figuras 3.28(d) e (e). Essa nova matriz de gradientes  $\nabla I'$ , composta pela combinação entre a imagem em escala de cinza e as distâncias de *Mahalanobis* de cada parte do corpo, será usada para auxiliar no cálculo de energia das arestas (definido na Equação 3.13). Assim, o modelo de segmentação baseado em esqueleto também estará considerando informação de cor, associada àquela determinada parte do corpo, ao invés de considerar apenas valores em níveis de cinza da imagem de entrada.

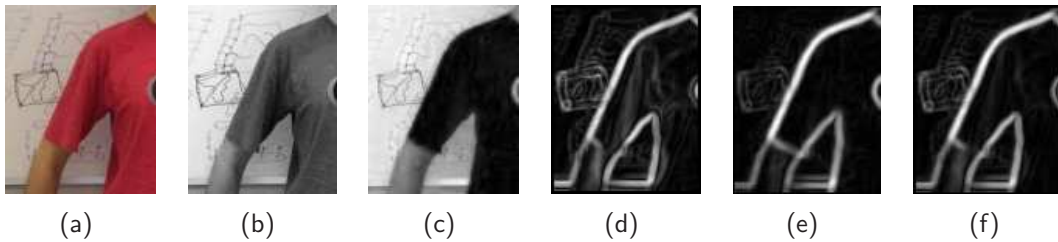


Figura 3.28: (a) Parte do corpo em análise. (b) Conversão da imagem (a) para escala de cinza. (c) Distâncias de *Mahalanobis* computadas para essa parte do corpo. (d) Magnitude do gradiente gerado a partir da imagem (b). (e) Magnitude do gradiente gerado a partir da imagem (c). (f) Magnitude do gradiente gerado pela média dos gradientes normalizados usados para gerar as imagens (d) e (e).

Ao transformar uma imagem colorida para tons de cinza, perde-se muita informação. Assim, regiões coloridas, visualmente muito diferentes umas das outras, podem parecer muito parecidas após a transformação para a escala de cinza (por exemplo, conforme ilustrado na Figura 3.29(a-b)). A inserção da informação de cor no cálculo de energia das arestas visa atenuar esse problema. A Figura 3.29 ilustra outra situação onde a combinação de informação de luminosidade e cor pode ser proveitosa (ilustração do tronco de uma determinada pessoa).

Conforme descrito na Seção 3.3.1, uma determinada parte do corpo pode ter associada até  $N_r$  modelos de cor (onde  $N_r = 3$ ), fazendo com que até  $N_r$  mapas de distâncias de *Mahalanobis* sejam criados para uma determinada parte do corpo. Nos casos onde o número de cores predominantes de uma determinada parte do corpo for maior que um, vai haver mais de uma distância associada a cada *pixel*  $(x, y)$  (cada qual associada ao seu respectivo modelo de cor). Nesse caso, para compor o

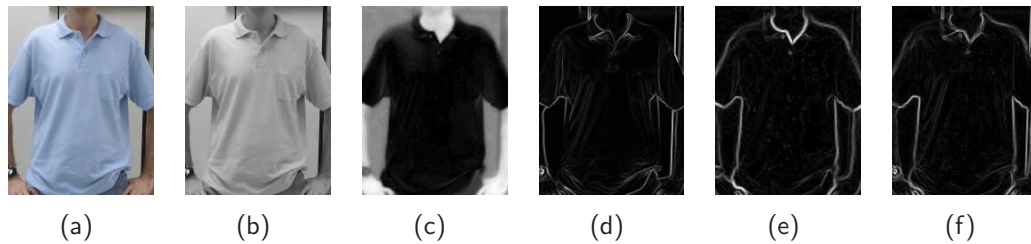


Figura 3.29: (a) Parte do corpo em análise. (b) Conversão da imagem (a) para escala de cinza. (c) Distâncias de *Mahalanobis* computadas para essa parte do corpo (valores escuros representam distâncias pequenas e valores claros o contrário). (d) Magnitude do gradiente gerado a partir da imagem (b). (e) Magnitude do gradiente gerado a partir da imagem (c). (f) Magnitude do gradiente gerado pela média dos gradientes normalizados usados para gerar as imagens (d) e (e).

mapa de distâncias de *Mahalanobis* final  $D_i$  de cada parte do corpo, considera-se a menor distância associada a cada *pixel*  $(x, y)$  como a distância resultante para aquele ponto. A Figura 3.30 ilustra esse processo, onde para uma determinada parte do corpo de uma pessoa em uma imagem foram criados dois modelos de cor (associados à duas cores predominantes, ilustrados na Figura 3.30(a-b)), os quais foram usados para gerados dois mapas de distâncias de *Mahalanobis* (um para cada modelo de cor, ilustrados na Figura 3.30(c-d)). O mapa de distâncias de *Mahalanobis* resultante para essa parte do corpo é ilustrado na Figura 3.30(e), utilizando-se do critério da menor distância.

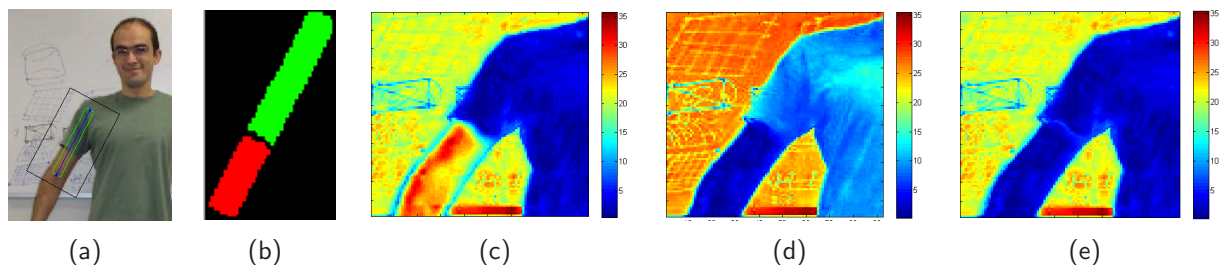


Figura 3.30: (a) Região de aprendizado da cor predominante (retângulo verde). (b) Duas cores predominantes detectadas. (c) Mapa de distâncias de *Mahalanobis* para a cor 1 (associada a região verde na imagem (b)). (d) Mapa de distâncias de *Mahalanobis* para a cor 2 (associada a região vermelha na imagem (b)). (e) Mapa de distâncias de *Mahalanobis* final, usado para calcular o gradiente  $\nabla D_i$ .

Em regiões de conexão entre duas partes do corpo os mapas de distâncias de *Mahalanobis* se interceptam, uma vez que são regiões retangulares expandidas em relação à seus respectivos “ossos” geradores, conforme descrito na Seção 3.3.3, ou podem até deixar “buracos” (dependendo do ângulo formado no ponto de conexão entre os dois “ossos”). No primeiro caso, quando um *pixel*  $(x, y)$  da imagem está associado a mais de um mapa de distâncias de *Mahalanobis*, é utilizada abordagem similar à descrita no parágrafo anterior, ou seja, o valor de distância associado a esse *pixel* será a menor distância em relação aos mapas de distâncias o qual pertence. No segundo caso não é diferente, porém é necessário computar a distância de *Mahalanobis* associada aos  $n$  modelos de cor

para aquele ponto. A Figura 3.31 ilustra esse processo, onde um mapa de distâncias *Mahalanobis* é gerado pela combinação de três mapas (distâncias de *Mahalanobis* para as regiões do braço, antebraço e mão direita), usando o critério que retém o valor mínimo de cada ponto. Deve ser salientado que a Figura 3.31 é usada apenas para ilustrar como um *pixel* associado à duas partes do corpo recebe apenas um valor de distância de *Mahalanobis*, uma vez que as regiões de busca de cada parte do corpo (que geram os mapas de distâncias de *Mahalanobis* para as mesmas) são definidas em função de seus respectivos “ossos” e dados antropométricos (descritos na Tabela 3.1 e ilustradas na Figura 3.31(a)), que definem principalmente suas inclinações e dimensões (o que não está sendo preservado na Figura 3.31(b-c) - as imagens exibidas na Figura 3.31(b-c) ilustram regiões que englobam a real região de busca de cada parte do corpo).

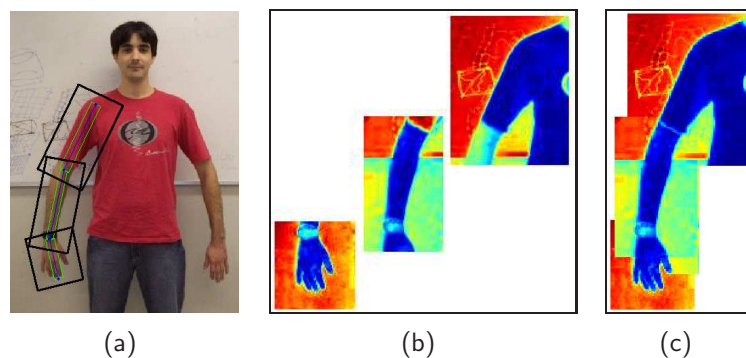


Figura 3.31: Ilustração das distâncias de *Mahalanobis* nas conexões entre duas partes adjacentes. Em (a), imagem RGB de entrada (usada para aprendizado e busca dos modelos de cor - regiões de busca são ilustradas por retângulos pretos). Em (b), três mapas distâncias de *Mahalanobis* para três partes do corpo (braço, antebraço e mão direita). Em (c), mapa distâncias de *Mahalanobis* gerado pela combinação dos três mapas exibidos em (b), usando o critério que retém o valor mínimo para cada ponto.

No próximo capítulo são apresentados resultados experimentais gerados a partir da utilização do modelo proposto nessa tese, assim como são apresentados estudos de caso envolvendo aquisição automática do esqueleto, sensibilidade do modelo em relação à diferentes usuários (diferentes dados de entrada para uma mesma imagem e tempo médio de captura dos dados de forma empírica) e comparação com um modelo considerado estado-da-arte.



## 4. Resultados Experimentais

Nesse capítulo são apresentados resultados experimentais do modelo baseado em esqueleto para segmentação de pessoas em imagens. Os resultados são avaliados quantitativamente (com exceção do estudo de caso apresentado na Seção 4.5), com a utilização de uma base de dados contendo 277 imagens (adquiridas desde bases de dados públicas [40–42], imagens encontradas na *web*, a imagens adquiridas em nosso laboratório com a utilização de uma câmera fotográfica convencional). Cada imagem utilizada nos experimentos (das 277 imagens da base de dados) teve seu *ground truth* (tabela verdade) demarcado manualmente por um usuário (mais especificamente, 3 indivíduos se dividiram para efetuar essa tarefa), com o objetivo de poder avaliar quantitativamente os resultados do modelo proposto. A criação do *ground truth* é descrita em detalhes a seguir. De uma forma geral, nesse capítulo são apresentados cinco experimentos, onde foram analisados os seguintes aspectos:

- Criação de *ground truth* para análise quantitativa;
- Características usadas no modelo proposto (tipos de energia avaliada);
- Sensibilidade do modelo (experimento com usuários - tempo médio computado para entrada de dados e variação dos dados de entrada);
- Esqueleto de entrada adquirido de forma manual (usuário) × automática (*kinect*);
- Comparação qualitativa com estado-da-arte.

### 4.1 Criação de *ground truth* para análise quantitativa

Objetivando avaliar quantitativamente os resultados deste trabalho, também está sendo proposta uma metodologia para medir o erro entre o contorno gerado para uma determinada pessoa em uma imagem e seu contorno esperado (estimado manualmente). O termo *ground truth* pode ser utilizado para diferentes finalidades, dependendo da aplicação em uso, porém sua característica principal é informar o que é dito como verdade (ou considerado como certo/verdadeiro). Por exemplo, em algoritmos de subtração de fundo (*background subtraction*) pode-se criar um *ground truth* onde para cada quadro da sequência de vídeo seja criado um mapa que informe onde está localizado o(s) objeto(s) em movimento na cena (*foreground*). Assim, o pesquisador pode medir o quão seu modelo segmentou corretamente o fundo dos objetos em movimento.

No caso de segmentação de pessoas em imagens estáticas, mais especificamente no caso deste trabalho, onde também há informação semântica envolvida, a criação do *ground truth* não é tão trivial, devido à diversos fatores:

- Grande complexidade e variação de poses e aparência que as pessoas podem assumir;

- Grande variação de roupas e acessórios que as pessoas podem usar;
- Grande variação de cortes/tipos de cabelos (curtos, compridos, encaracolados, arrepiados, etc);
- Oclusão parcial ou total de membros;
- Conforme List e sua equipe [43], inspeção manual é dependente de usuário, ou seja, duas pessoas analisando a mesma imagem podem (e provavelmente irão) produzir diferentes dados de *ground truth*;
- A criação do *ground truth* é uma tarefa extremamente tediosa e demorada.

Dessa forma, o *ground truth* criado nesse trabalho possui algumas simplificações que devem ser esclarecidas, de maneira que o leitor possa compreender o que está sendo avaliado e de que maneira.

### **Simplificações adotadas:**

1. De uma forma geral, o *ground truth* é um contorno fechado (ponto inicial = ponto final) que representa a silhueta esperada para uma pessoa em uma imagem;
2. Cada ponto desse contorno é associado à uma determinada parte do corpo (assim, pode-se saber quais pontos do contorno compõem o braço esquerdo, por exemplo);
3. Membros parcialmente ou totalmente ocultos são estimados pelo usuário;
4. O contorno da cabeça é feito de maneira que o cabelo não seja levado em consideração, ou seja, como se a pessoa fosse careca ou estivesse com a cabeça raspada;
5. Roupas justas servem para guiar o contorno da pessoa, enquanto que roupas muito folgadas são relacionadas a objetos que estão ocultando uma determinada parte do corpo.

A Figura 4.1 ilustra cada uma das simplificações enumeradas acima. A Figura 4.1(a) ilustra o *ground truth* gerado, de forma que não haja falhas ou buracos nessa curva, ou seja, o contorno é fechado (o ponto inicial dessa curva coincide com o ponto final da mesma). Figura 4.1(b) exibe a informação semântica embutida em cada dado de *ground truth* (ilustrada por diferentes cores), que pode ser útil em uma análise local dos resultados. Na Figura 4.1(c) é exibida uma imagem na qual o usuário estimou suas partes ocultas, assim como ilustra o contorno da cabeça sem considerar saliências do cabelo. Na Figura 4.1(d) é exibido novamente o detalhe do contorno da cabeça estimada pelo usuário, onde o cabelo é desconsiderado. Roupas muito salientes ou largas são consideradas objetos que obstruem partes do corpo, como ilustrado na Figura 4.1(e). Também pode ser observado na Figura 4.1(e) que uma pequena porção do tronco dessa pessoa foi estimado pelo usuário (devido ao braço esquerdo da mesma estar obstruindo-o). De uma forma geral, o *ground truth* gerado servirá para analisar quantitativamente os resultados desse trabalho. Essa análise será feita de diferentes formas, como descrito a seguir.



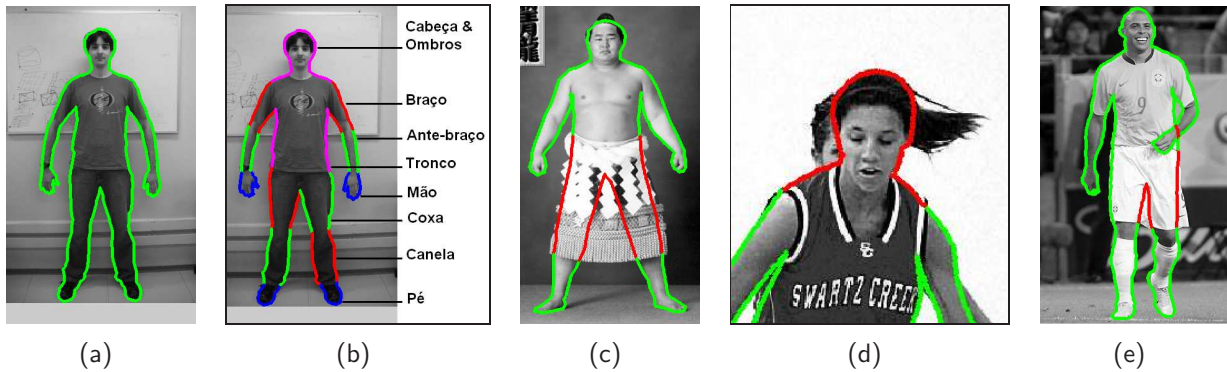


Figura 4.1: Ilustração do *ground truth* gerado para avaliação quantitativa e suas características. (a) Ilustração do contorno fechado. (b) Ilustração da semântica associada a cada parte do corpo. (c) Partes ocultas (parcialmente ou totalmente) são estimadas pelo usuário. (d) O contorno do cabelo não é considerado, ou seja, o usuário gera o *ground truth* esperado da cabeça. (e) Roupas folgadas são consideradas objetos que obstruem partes do corpo.

## 4.2 Características usadas no modelo proposto

Nessa seção são avaliadas as características usadas no modelo baseado em esqueleto. Basicamente são analisadas 7 diferentes formas (ou versões, descritas com auxílio da Tabela 4.1) de se calcular a energia das arestas dos grafos. A energia calculada é utilizada na aquisição do contorno de cada pessoa, desde a forma mais simples (apenas utilizando a magnitude do gradiente de uma imagem em escala de cinzas, detalhada na Seção 3.2.3) até a sua forma mais completa (utilizando informações de luminosidade/cor, distâncias antropométricas e restrições de ângulos, conforme descrito na Seção 3.3.4). O objetivo desse experimento é avaliar o impacto causado por cada característica usada. Para isso, a taxa de acerto de cada versão utilizada é avaliada como descrito a seguir.

Tabela 4.1: Características usadas para avaliar a energia em cada versão avaliada.

Característica	v1	v2	v3	v4	v5	v6	v7
Gradiente da luminosidade	×	×	×	×	×	×	×
Cor (sem <i>PCA</i> )						×	
Cor (com <i>PCA</i> )	×						×
Distâncias antropométricas	×	×	×				
Restrição de ângulos	×	×		×			

Para avaliar o erro entre o *ground truth* e o resultado gerado com a utilização do modelo proposto, foi utilizada uma versão modificada da distância de *Hausdorff*, proposta por Hossain e sua equipe [44]. A distância de *Hausdorff* (definida na Eq. 4.1) é uma métrica muito utilizada em várias aplicações envolvendo processamento de imagens e visão computacional, incluindo detecção de objetos em movimento, casamento de padrões, rastreamento e reconhecimento de objetos, entre outras.

$$H(A, B) = \max(h(A, B), h(B, A)), \quad (4.1)$$

onde

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|. \quad (4.2)$$

Intuitivamente, o significado da distância de *Hausdorff* para a validação dos resultados obtidos pela técnica proposta é descrito a seguir. Considere dois conjuntos de dados,  $A = \{a_1, \dots, a_m\}$  e  $B = \{b_1, \dots, b_n\}$ , por exemplo, onde o conjunto  $A$  representa os pontos estimados para o contorno do braço direito de uma determinada pessoa e o conjunto  $B$  representa os pontos dessa mesma parte do corpo, porém informados pelo usuário (*ground truth*). Para cada ponto do conjunto  $A$  é calculado a distância *Euclidiana* em relação a todos os pontos do conjunto  $B$ . A distância mínima é armazenada em um vetor de “mínimos”. Esse procedimento é repetido para o conjunto  $B$ , em relação ao conjunto  $A$ . A distância de *Hausdorff* é a distância máxima obtida, armazenada nesses vetores de mínimos. Conforme relatado por Hossain e sua equipe [44], o valor máximo obtido é muito sensível a ruídos, propondo em seu trabalho a utilização do valor médio (média dos valores mínimos, definida na Eq. 4.3) ao invés do valor máximo. Essa métrica, denominada em nosso trabalho por distância de *Hausdorff* modificada, é utilizada para avaliar os resultados experimentais.

$$h'(A, B) = \frac{1}{m} \sum_{a \in A} \min_{b \in B} \|a - b\| \quad (4.3)$$

Os resultados experimentais são avaliados quantitativamente de três diferentes formas:

1. Erro médio: calcula-se a distância de *Hausdorff* modificada para cada parte do corpo e o erro armazenado para cada imagem é a média das distâncias computadas. Permite avaliar a taxa de acerto de cada imagem considerando a semântica associada às partes do corpo.
2. Erro máximo (ou erro representativo): calcula-se a distância de *Hausdorff* modificada para cada parte do corpo e o erro armazenado para cada imagem é a distância máxima computada. Permite avaliar a taxa de acerto de cada imagem, considerando a semântica associada às partes do corpo, e salientando imperfeições associadas à alguma determinada parte do corpo.
3. Erro global: desconsidera-se a semântica associada às partes do corpo, e a distância de *Hausdorff* modificada é aplicada ao contorno completo. Permite avaliar a taxa de acerto de cada imagem de uma forma global.

O erro médio, assim como o erro máximo, é computado para 14 partes do corpo (e não para 16 partes, conforme definido no modelo de esqueleto apresentado na Tabela 3.1), uma vez que a cabeça e os ombros são avaliados como uma única parte (ilustrado na Figura 4.1(b) e exemplificado com auxílio da Tabela 4.2).

Deve-se salientar que os erros avaliados nesse experimento consideram todos os pontos do contorno gerados pelo método proposto, ou seja, incluindo os pontos ocultos estimados na etapa de

criação do *ground truth*. Também é importante mencionar que os erros avaliados foram normalizados em função da altura estimada para cada pessoa na imagem. Essa normalização é feita pelo fato das imagens possuírem diferentes resoluções, o que poderia afetar a análise dos resultados (por exemplo, em uma imagem com alta resolução, um  $erro = 5 \text{ pixels}$  de distância pode ser considerado pequeno, enquanto que em uma imagem em baixa resolução esse mesmo valor possa ser considerado moderado ou grande). Além disso, pessoas contidas em imagens com mesma resolução podem apresentar tamanhos diferentes (uma pessoa em relação à outra ou devido à distância que estavam da câmera quando fotografadas), fazendo com que alguma normalização seja desejável.

A Figura 4.2 ilustra um resultado gerado pelo modelo proposto (em azul) e o *ground truth* gerado pelo usuário (em verde), sobrepostos na imagem de entrada (convertida para escala de cinza). A Tabela 4.2 exibe os erros avaliados (distância de *Hausdorff* modificada - em *pixels* e normalizados em função da altura) para a imagem exibida na Figura 4.2 (usando a versão v1 do modelo proposto), onde é possível identificar qual parte do corpo foi melhor segmentada (nesse caso, a canela esquerda, com  $erro = 0.0031$ ), assim como a parte que obteve menor taxa de acerto (nesse caso, a canela direita, com  $erro = 0.0079$ ).

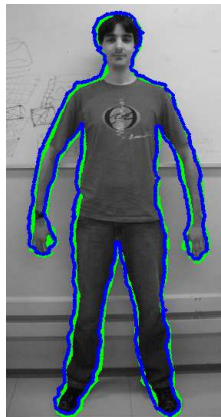


Figura 4.2: Resultado experimental obtido (em azul) e *ground truth* (em verde).

A Tabela 4.3 exibe os erros (normalizados em função da altura, para as 7 versões utilizadas para computar a energia das arestas dos grafos - v1 a v7) avaliados para as 277 imagens da base de dados, usando como métrica a distância de *Hausdorff* modificada, conforme mencionado anteriormente.

Com base nas Tabelas 4.3 e 4.1, podemos concluir que a versão do modelo proposto que possuiu maior taxa de acerto (menor erro médio e menor erro global) e conseqüentemente a que gerou melhores resultados foi a versão mais completa (v1). Entretanto, também pode-se observar que a informação de cor influenciou pouco nos resultados (na forma como foi utilizada), devido a baixa diferença nos erros obtidos entre as versões v1 e v2. A restrição de ângulos também teve pouca influência em uma análise global dos resultados, como pode-se observar na baixa diferença dos erros obtidos nas versões v2 e v3 (com e sem restrições de ângulos + distâncias antropométricas), assim como entre as versões v4 e v5 (com e sem restrições de ângulos + magnitude do gradiente). No geral, as três primeiras versões possuíram erros relativamente próximos, em uma análise global,

Tabela 4.2: Ilustração dos erros computados (em *pixels* e normalizados) para uma única imagem. Para esse exemplo, o erro médio foi 2.4442 (em *pixels*) e 0.0053 (normalizado). O erro máximo foi 3.5881 (em *pixels*) e 0.0079 (normalizado). Desconsiderando as partes do corpo, tratando o contorno como uma única curva, o erro global foi 2.0969 (em *pixels*) e 0.0046 (normalizado).

Parte do corpo	Erro em <i>pixels</i>	Erro normalizado
Cabeça & Ombros	1.7658	0.0039
Tronco	2.7072	0.0059
Braço direito	3.3372	0.0073
Antebraço direito	2.4797	0.0054
Mão direita	2.0953	0.0046
Braço esquerdo	1.4770	0.0032
Antebraço esquerdo	2.5983	0.0057
Mão esquerda	2.0119	0.0044
Coxa direita	3.2027	0.0070
Canela direita	3.5881	0.0079
Pé direito	3.5214	0.0077
Coxa esquerda	1.9560	0.0043
Canela esquerda	1.4050	0.0031
Pe esquerdo	2.0738	0.0045

Tabela 4.3: Erro avaliado para as diferentes versões do modelo proposto, usando uma base de dados com 277 imagens.

Versão	v1	v2	v3	v4	v5	v6	v7
Erro médio							
Média	0.0082	0.0083	0.0084	0.0110	0.0110	0.0165	0.0165
Desvio padrão	0.0023	0.0024	0.0024	0.0033	0.0031	0.0063	0.0063
Erro máximo							
Média	0.0174	0.0173	0.0173	0.0249	0.0248	0.0363	0.0362
Desvio padrão	0.0084	0.0081	0.0079	0.0093	0.0092	0.0130	0.0126
Erro global							
Média	0.0063	0.0064	0.0065	0.0084	0.0085	0.0127	0.0127
Desvio padrão	0.0019	0.0020	0.0020	0.0026	0.0025	0.0130	0.0050

fazendo com que as diferenças entre elas possam ser percebidas apenas em condições específicas (localmente, em algumas imagens onde uma ou outra determinada característica possa fazer a diferença). A Figura 4.3, por exemplo, ilustra dois resultados experimentais obtidos ao utilizar o modelo proposto nas suas versões v2 e v3 (com e sem a restrição de ângulos, respectivamente). Pode-se observar que a imagem ilustrada na Figura 4.3(a) possui uma diferença sutil em relação a Figura 4.3(b) (mais especificamente na região interna das pernas e nos pés, onde o contorno é mais “suave”).

A utilização da análise dos componentes principais (*PCA*) teve pouca influência nessa análise global dos resultados, porém obteve um pequeno ganho observado na versão v6, em relação a v7 (Tabela 4.3). Em algumas imagens onde a informação de cor é muito pouca, percebe-se certo

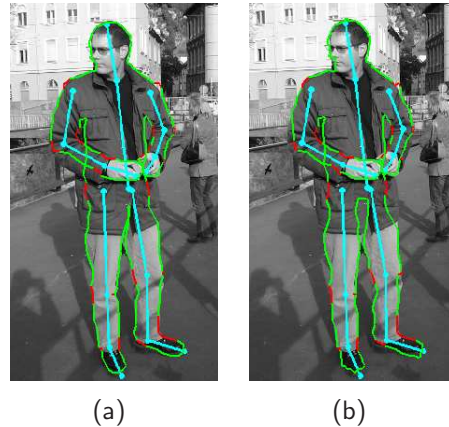


Figura 4.3: Resultado experimental obtido com e sem restrição de ângulos no cálculo da energia, respectivamente.

ganho na utilização do *PCA*, como ilustrada na Figura 4.4 (aumentando o contraste no mapa de distâncias de *Mahalanobis*, usado no cálculo de energia das arestas, e conseqüentemente no resultado final). Outra vantagem da utilização do *PCA* é na redução da dimensionalidade do modelo de cor (podendo chegar a 1D ao invés de 3D, conforme descrito no modelo proposto), ilustrada com auxílio da Figura 4.5.

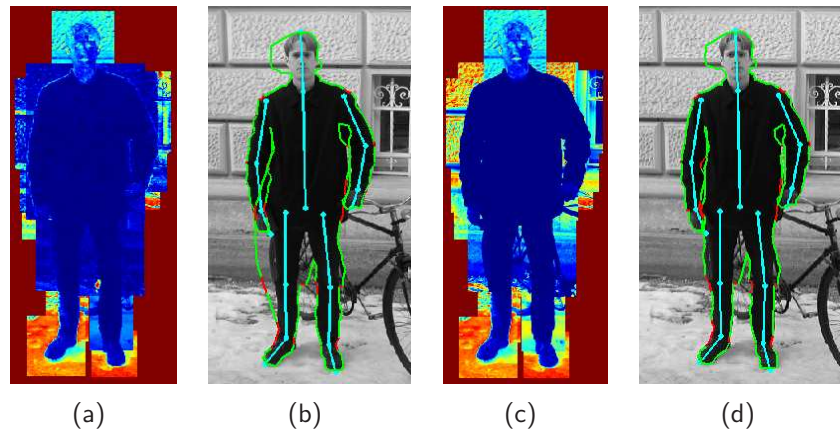


Figura 4.4: Resultado experimental obtido com e sem utilização de *PCA* na segmentação que utiliza informação de cor. (a) Distâncias de *Mahalanobis* concatenadas gerando um único mapa para todo o corpo (sem *PCA*). (b) Resultado do modelo proposto usando a versão v6. (c) Distâncias de *Mahalanobis* concatenadas gerando um único mapa para todo o corpo (com *PCA*). (d) Resultado do modelo proposto usando a versão v7.

A característica usada que mais influenciou positivamente nos resultados foram as distâncias antropométricas, como pode-se observar na grande diferença entre os erros obtidos para as versões v2 e v4 (Tabela 4.3). A Figura 4.6 ilustra dois resultados experimentais gerados a partir do modelo proposto utilizando as versões v2 e v4 (com e sem restrições de distâncias antropométricas, respectivamente).

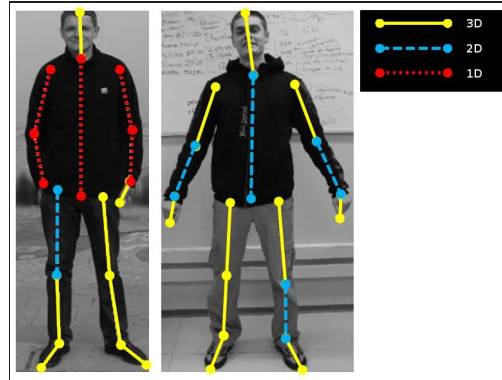


Figura 4.5: Ilustração do número de reduções das dimensões do modelo de cor, de determinadas partes do corpo, com a utilização de *PCA*. “Ossos” ilustrados em amarelo (contínuo) possuem dimensão original (3D); em azul (tracejado) foram reduzidos para 2D; e em vermelho (pontilhado) foram reduzidos para 1D.

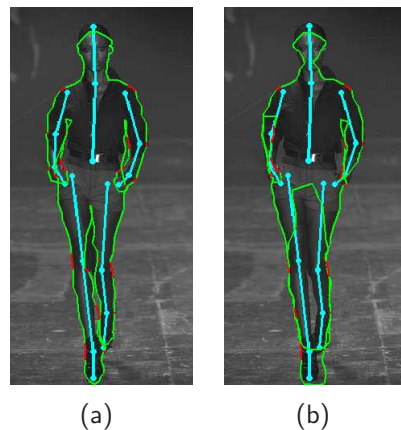


Figura 4.6: Resultado experimental obtido com e sem restrições de distâncias antropométricas, respectivamente. (a) Usando a versão v2. (b) Usando a versão v4.

A Figura 4.7 ilustra resultados experimentais obtidos com a utilização do modelo proposto (versão v1), considerados muito bons através de inspeção visual. Em contrapartida, a Figura 4.8 ilustra resultados experimentais considerados aceitáveis, ou seja, os contornos resultantes nesses exemplos não superaram expectativas desejadas (através de inspeção visual). As imperfeições ilustradas na Figura 4.8 podem ser decorrentes de diversos fatores: baixo contraste entre a pessoa e o fundo, camuflagem, oclusão parcial de partes do corpo, baixa resolução da imagem, problemas de iluminação, entre outros fatores. É importante salientar que as imagens exibidas nas Figuras 4.7 e 4.8 foram recortadas das imagens originais (após a segmentação), por critério de simplicidade (muitas imagens com diferentes fundos e tamanhos) e também foram transformadas para escala de cinza para melhor visualização dos resultados. O critério utilizado para recortar essas imagens foi usar a região que englobasse todos os *pixels* da pessoa segmentada. Entretanto, com o objetivo de ilustrar a complexidade (ou simplicidade) das imagens como um todo (fundo da cena, número de pessoas, entre outros fatores), as imagens originais também são exibidas (de forma reduzida) no Anexo D



(com auxílio das Figuras D.1 e D.2). As imagens ilustradas nas Figuras 4.7 e 4.8 exibem o esqueleto de entrada informado manualmente (em ciano) e o contorno resultante (em verde e vermelho). As partes em vermelho do contorno simbolizam pontos do contorno que estão em uma região que divide duas partes do corpo adjacentes, fazendo possível que a informação semântica embutida nos resultados também possa ser visualizada.



Figura 4.7: Resultados experimentais obtidos utilizando a versão v1 do modelo proposto, considerados muito bons através de inspeção visual.

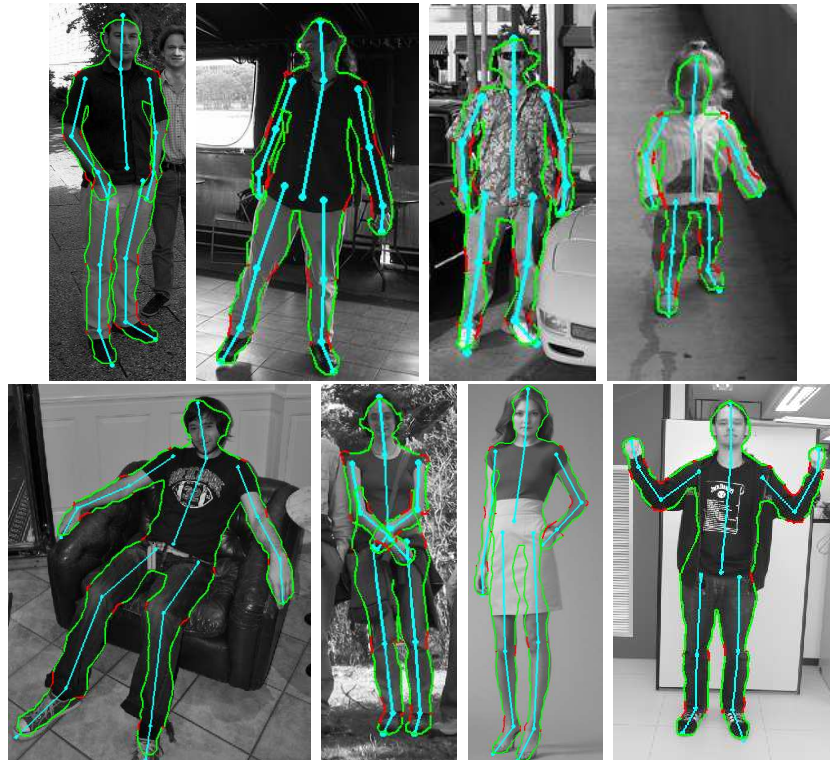


Figura 4.8: Resultados experimentais obtidos utilizando a versão v1 do modelo proposto, considerados aceitáveis através de inspeção visual.

A Tabela 4.4 exibe os erros avaliados para as imagens ilustradas na Figura 4.7. A Tabela 4.4 serve para ilustrar valores de erros associados à resultados considerados muito bons através de inspeção visual, uma vez que os valores apresentados até aqui (erros normalizados em função da altura) dificultam a interpretação de qual valor de erro deveria ser considerado baixo/médio/alto. Por outro lado, a Tabela 4.5 exibe os erros computados para as imagens ilustradas na Figura 4.8, associados à resultados considerados aceitáveis (através de inspeção visual).

Tabela 4.4: Erro avaliado para as imagens exibidas na Figura 4.7, usando a versão v1 do modelo proposto.

	Erro médio	Erro máximo	Erro global
Média	0.0069	0.0137	0.0053
Desvio padrão	0.0018	0.0051	0.0014

Outro aspecto que deve ser mencionado em relação ao modelo proposto, além de avaliar e ilustrar seus resultados, é informar quais seriam suas limitações, ou seja, em que situações o modelo não deve funcionar corretamente ou pode apresentar resultados incoerentes. Uma limitação do modelo proposto é tratar poses onde o movimento dos membros (das pessoas contidas nas imagens) não estão aproximadamente no mesmo plano da imagem (o que afeta as estimativas antropométricas na imagem projetada), ilustrada com auxílio da Figura 4.9. Por exemplo, considere a pessoa contida na imagem mais à esquerda, ilustrada na Figura 4.9, com o braço direito apontando para frente.



Nesse exemplo, os pontos do esqueleto associado ao braço esquerdo (braço + antebraço + mão) estariam praticamente sobrepostos (no caso 2D). Dessa forma, a informação 2D associada a esse determinado membro é praticamente nula. A utilização de um algoritmo de estimativa 3D de pose de pessoas em imagens poderia ser utilizado para identificar esse tipo de situação e então aplicar um tratamento especial adequado. Outros fatores podem fazer com que os resultados gerados sejam indesejáveis, como mencionados anteriormente: grande complexidade da pose, oclusão parcial ou total de membros, problemas associados a fatores de iluminação, entre outros.

Tabela 4.5: Erro avaliado para as imagens exibidas na Figura 4.8, usando a versão v1 do modelo proposto.

	Erro médio	Erro máximo	Erro global
Média	0.0099	0.0290	0.0092
Desvio padrão	0.0021	0.0132	0.0030



Figura 4.9: Limitação do modelo: partes do corpo que não estão aproximadamente no mesmo plano da imagem podem produzir resultados indesejados.

Além disso, o custo computacional associado ao modelo proposto pode se tornar relevante quando o mesmo estiver associado a uma aplicação comercial, onde o processamento em tempo-real normalmente é desejável. Como o tempo de processamento de cada imagem no modelo proposto está relacionado ao tamanho, ou altura (em *pixels*) da pessoa nessa imagem, a avaliação apresentada na Tabela 4.6 foi realizada levando-se em consideração esse fator. Basicamente, a Tabela 4.6 exibe o tempo de processamento gasto para cada versão do modelo proposto, avaliado para 3 imagens, contendo uma única pessoa em cada, com alturas distintas. As imagens foram selecionadas com base nos seguintes critérios: (i) pessoa com menor altura da base de dados usada (igual a 150 *pixels*); (ii) pessoa com altura igual a média das alturas da base de dados usada (igual a 463 *pixels*); e (iii) pessoa com maior altura da base de dados usada (igual a 1239 *pixels*). O protótipo do modelo proposto foi implementado em MATLAB sem nenhuma otimização, ou seja, acredita-se que os valores apresentados na Tabela 4.6 possam melhorar significativamente se o modelo for implementado de maneira otimizada em uma linguagem compilada como C/C++, por exemplo, ao invés de uma linguagem interpretada como o MATLAB. O *hardware* utilizado foi uma máquina *Workstation HP xw8600*, com processador *Intel Xeon, Core2 Quad, 2.83GHz*, com 3Gb de memória.

Com base nos valores apresentados na Tabela 4.6 podemos concluir que o modelo sofre uma queda de performance significativa quando a informação de cor é inserida, como pode ser observado na diferença entre os tempos apresentados para as versões v1, v6 e v7 em relação às demais (v2, v3, v4 e v5) que não utilizam tal característica. Acredita-se que isso esteja relacionado com a forma na qual esse módulo foi implementado, sem otimização e utilizando laços para percorrer varios pontos da imagem (o que normalmente aumenta o tempo de processamento em aplicações rodadas no MATLAB). Também pode ser observado, com auxílio da Tabela 4.6, o melhor e pior tempo de execução para cada versão do modelo, assim como o tempo de processamento em função da altura de cada pessoa. No melhor caso o modelo segmentou uma imagem utilizando aproximadamente 2 segundos, enquanto que no pior caso a aplicação necessitou em torno de 100 segundos para ser executada.

Tabela 4.6: Tempo de processamento (em segundos) avaliado para 3 imagens da base de dados usada, para todas as versões do modelo proposto.

	v1	v2	v3	v4	v5	v6	v7
Altura mínima (150)	10.28	2.34	2.33	2.22	2.34	11.01	10.46
Altura média (463)	43.47	14.18	14.54	14.08	14.5	45.43	45.59
Altura máxima (1239)	100.89	19.91	21	19.58	20.82	101.33	101.95

### 4.3 Sensibilidade do modelo

Nessa seção é apresentado um estudo de caso para avaliar a sensibilidade do modelo proposto em relação aos dados de entrada (modelo de esqueleto). Nesse experimento, 24 usuários (pessoas de diversas áreas do conhecimento e não especificamente de processamento de imagens) foram instruídos a clicar em 3 imagens (contendo uma única pessoa em cada imagem) para inserir os dados de entrada (estimativa da altura e pontos do esqueleto), visando avaliar se o modelo proposto é sensível a pequenas modificações em relação aos dados de entrada. O objetivo secundário desse experimento é analisar o tempo que uma pessoa leva para informar os dados de entrada.

Cada usuário que realizou o experimento recebeu as seguintes instruções:

- Para cada imagem, você deve estimar a altura da pessoa (clikando em um ponto no topo da cabeça e outro abaixo dos pés, onde considera a base da pessoa na imagem);
- Após estimar a altura, você deve informar os pontos do esqueleto (associados ao modelo de esqueleto), em uma ordem pré-determinada;
- Primeira restrição: os pontos informados devem estar centralizados nas suas respectivas partes do corpo. Por exemplo, ao informar o ponto do joelho, você deve clicar no centro do joelho e não na sua fronteira;

- Segunda restrição: nas partes do corpo associadas à extremidades (mãos, cabeça e pés), você deve informar como ponto extremo, o ponto da imagem mais extremo em relação àquela parte, ou seja, clicar no final da mão, no final do pé, no início da cabeça, ao invés de clicar no centro da mão ou pé ou cabeça, por exemplo.

A Figura 4.10 ilustra as 3 imagens usadas no experimento, com os esqueletos informados pelos 24 usuários sobrepostos na imagem de entrada, onde é possível observar a variação nos dados de entrada, relacionadas aos pontos do esqueleto. A Tabela 4.7 exhibe a média de altura estimada pelos 24 usuários (em *pixels*) para cada pessoa em cada imagem (exibidas na Figura 4.10), mostrando também uma certa variação. Na última linha da Tabela 4.7 são exibidas as alturas armazenadas no *ground truth* para as mesmas imagens. É importante salientar que as imagens exibidas na Figura 4.10 foram redimensionadas, ou seja, podem possuir resoluções diferentes (assim, as alturas exibidas na Tabela 4.7 não representam a mesma distância, ou proporção, em *pixels* exibidas na Figura 4.10).

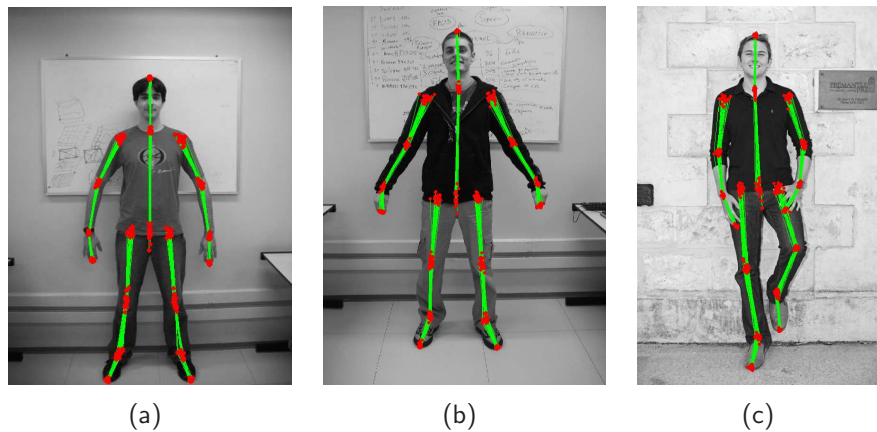


Figura 4.10: Esqueletos informados por 24 usuários, sobrepostos nas imagens de entrada.

Tabela 4.7: Altura média estimada (em *pixels*) por 24 usuários para as 3 pessoas contidas nas 3 imagens exibidas na Figura 4.10 e altura armazenada no *ground truth*.

	imagem (a)	imagem (b)	imagem (c)
Altura Média	464.3462	643.4615	653.4615
Desvio padrão	12.3643	15.5775	16.8623
<i>Ground truth</i>	457	636	633

A Tabela 4.8 exhibe os erros avaliados nesse experimento (erro médio, erro máximo e erro global) de maneira individual (para cada imagem) e global (para as 3 imagens agrupadas, (a), (b) e (c), ilustradas na Figura 4.10), usando a versão v1 do modelo proposto. Com base na Tabela 4.8, podemos concluir que os resultados foram coerentes em relação a diferentes dados de entrada, uma vez que o desvio padrão foi relativamente baixo para cada imagem, assim como para todas as imagens, em uma análise global. A Figura 4.11 ilustra os diversos contornos obtidos para cada imagem usada nesse experimento, sobrepostos na imagem de entrada.

Tabela 4.8: Erro avaliado para 3 imagens (ilustradas na Figura 4.10), a partir dos dados de entrada fornecidos por 24 usuários.

	imagem (a)	imagem (b)	imagem (c)	imagens (a,b,c)
Erro médio				
Média	0.0057	0.0061	0.0054	0.0057
Desvio padrão	0.0005	0.0008	0.0005	0.0007
Erro máximo				
Média	0.0091	0.0109	0.0099	0.0099
Desvio padrão	0.0014	0.0035	0.0014	0.0024
Erro global				
Média	0.0046	0.0042	0.0042	0.0043
Desvio padrão	0.0002	0.0004	0.0003	0.0004

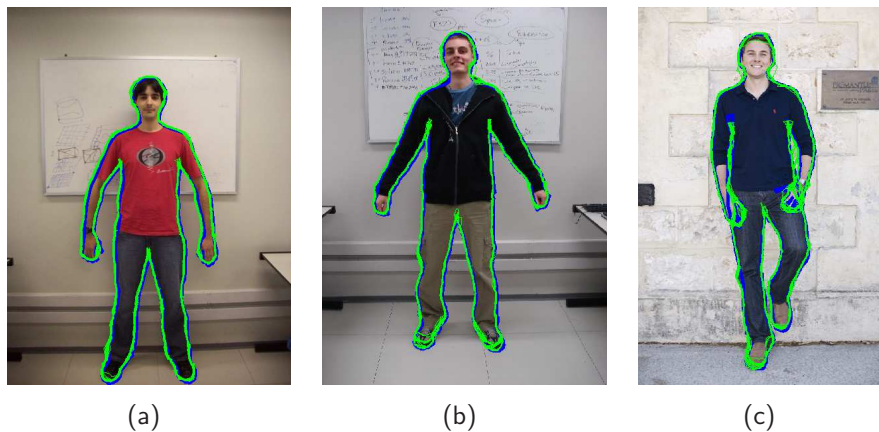


Figura 4.11: Contornos obtidos a partir dos dados de entrada informados por 24 usuários, sobrepostos nas imagens usadas (resultados obtidos exibidos em verde e *ground truth* em azul).

A Tabela 4.9 exibe os tempos que os usuários levaram para informar os dados de cada imagem. É importante salientar que as imagens foram exibidas aos usuários sempre na mesma ordem, ou seja, primeiro cada usuário informou os dados de entrada para a imagem (a), depois para a imagem (b) e posteriormente para a imagem (c). É possível observar no gráfico exibido na Figura 4.12 (assim como na Tabela 4.9) que houve uma curva de aprendizagem, ou seja, os usuários foram informando os dados de entrada de maneira mais ágil (ou mais rápida) à medida que as imagens eram exibidas. Embora apenas 3 imagens tenham sido usadas, espera-se que essa curva tenda a atingir um limite de tempo mínimo. De qualquer forma, o experimento apresentado na Seção 4.3 não visa demonstrar que um usuário pode informar os dados de entrada em tempo ótimo, e sim ilustrar como é simples e prática a entrada de dados via usuário, sem que seja necessário demasiada interação e perda de tempo.

Na próxima seção é apresentado um estudo de caso onde o esqueleto de entrada é obtido de forma automática, fazendo com que o modelo de segmentação de pessoas baseado em esqueleto proposto nessa tese possa ser executado sem qualquer intervenção com o usuário.

Tabela 4.9: Tempo médio (em segundos) que os usuários levaram para informar os dados de entrada para cada imagem e média global.

	imagem (a)	imagem (b)	imagem (c)	imagens (a,b,c)
Tempo médio	113.4642	87.4088	76.2418	92.3716
Desvio padrão	29.2118	29.8061	27.0879	32.4056

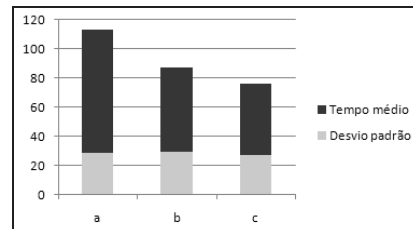


Figura 4.12: Tempo (em segundos) que os usuários levaram para informar os dados de entrada: curva de aprendizagem, assumindo-se que as imagens foram exibidas para os usuários sempre na mesma ordem (a → b → c).

#### 4.4 Esqueleto de entrada adquirido de forma manual × automática

Nessa seção é apresentado um experimento onde os dados de entrada (altura e pontos associados ao modelo de esqueleto) são capturados de forma automática, fazendo com que o modelo proposto seja executado sem qualquer intervenção manual. Nesse estudo de caso foram utilizadas 8 imagens, capturadas com um dispositivo especial (*Kinect*), bastante popular nos dias de hoje, no qual os pontos do esqueleto podem ser adquiridos facilmente através de um software disponível na internet <sup>1</sup>.

É importante salientar que o *Kinect* foi utilizado apenas na captura dos pontos associados ao modelo de esqueleto de cada pessoa em um único quadro de um vídeo (ou seja, de uma única imagem), os quais foram pós-processados de maneira que a altura de cada pessoa na imagem pudesse ser estimada. As informações disponíveis por este dispositivo, como por exemplo, mapa de profundidades (capturadas por um sensor infravermelho) assim como informação de movimento (uma vez que o dispositivo captura vídeos) não estão sendo consideradas. O estudo de caso apresentado nessa seção, utilizando o *Kinect* para captura dos dados de entrada de forma automática, objetiva apenas demonstrar que o modelo proposto pode ser executado de forma automatizada se os dados de entrada forem obtidos de maneira similar. Deve-se salientar que a utilização do *Kinect* inviabiliza qualquer experimento com imagens que não foram capturadas pelo mesmo, como por exemplo, imagens encontradas na *web*, adquiridas em nosso laboratório por uma câmera fotográfica convencional ou de base de dados públicas [40–42], conforme mencionado no início desse capítulo. Entretanto, existem modelos para obtenção de pose articulada de pessoas em imagens propostos na literatura [3, 7, 12, 13, 26] que poderiam ser usados para obtenção dos dados de entrada do modelo proposto de forma automática, porém não foram usados por não oferecerem código fonte (ou programa executável) disponível para testes ou por não estimarem a pose do corpo todo (como em [3],

<sup>1</sup>*Kinect* para Windows: <http://www.microsoft.com/en-us/kinectforwindows/>

por exemplo). Ferrari e sua equipe [40] disponibilizam na internet um software para estimativa 2D de poses de pessoas em imagens <sup>2</sup>, entretanto tal abordagem gera um modelo de esqueleto bastante diferente do utilizado nessa tese (principalmente nos pontos de conexão entre duas partes adjacentes - articulações), o qual deveria passar por algumas adaptações ou etapas de pós-processamento para que pudesse ser usado de forma adequada. De qualquer forma, o objetivo desse estudo de caso é ilustrar o funcionamento do modelo proposto nessa tese em uma configuração totalmente automática, e não salientar qual algoritmo para estimativa de pose 2D de pessoas em imagens melhor se adapta à resolução desse problema.

A Figura 4.13 ilustra 2 imagens usadas nesse experimento, capturadas pela câmera do *Kinect*. As Figura 4.13(b) e (e) exibem (em verde) o esqueleto detectado automaticamente pelo dispositivo. As Figura 4.13(c) e (f) exibem (em verde) o esqueleto informado pelo usuário. É possível perceber que algumas partes do esqueleto detectado de forma automática nem sempre estão centralizadas na sua respectiva parte do corpo em relação à imagem de entrada (como por exemplo, na Figura 4.13(b), o antebraço direito, ou Figura 4.13(e), a canela esquerda). Esse problema, associado ao esqueleto detectado automaticamente, faz com que o contorno daquela determinada parte do corpo possa não ser segmentada de maneira precisa, uma vez que diversos fatores são levados em consideração (como distâncias antropométricas e regiões de busca, por exemplo). Em ambos os casos (automático e manual) a altura de cada pessoa foi obtida diretamente pela distância entre o ponto associado ao topo da cabeça e um ponto abaixo à região dos pés (no caso automático, foi utilizado o ponto médio entre os pontos  $P_{16}$  e  $P_{19}$ , associados ao modelo de esqueleto, como base da pessoa para estimativa da altura).

A Tabela 4.10 exhibe os erros avaliados para as 8 imagens utilizadas nesse experimento (em comparação com o *ground truth* criado para as mesmas), com dados de entrada adquiridos de forma automática (usando o *Kinect*) e manual (informados pelo usuário).

Tabela 4.10: Erro avaliado para 8 imagens, para dados de entrada adquiridos de forma automática (*Kinect*) e manual (informados pelo usuário).

	Automático ( <i>Kinect</i> )	Manual (usuário)
Erro médio		
Média	0.0110	0.0049
Desvio padrão	0.0024	0.0015
Erro máximo		
Média	0.0281	0.0111
Desvio padrão	0.0090	0.0032
Erro global		
Média	0.0079	0.0041
Desvio padrão	0.0023	0.0014

Como pode ser observado na Tabela 4.10 (e com auxílio da Figura 4.13), os resultados gerados com esqueleto obtido de forma automática não superaram os resultados obtidos com o esqueleto

<sup>2</sup>[http://www.vision.ee.ethz.ch/~calvin/articulated\\_human\\_pose\\_estimation\\_code/](http://www.vision.ee.ethz.ch/~calvin/articulated_human_pose_estimation_code/)



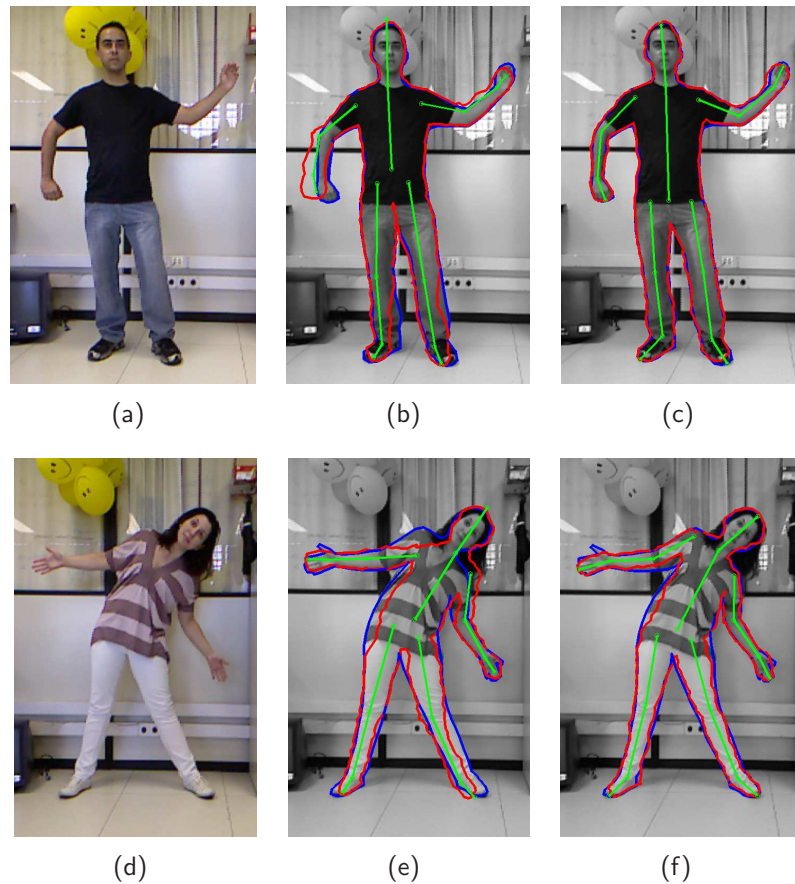


Figura 4.13: Estudo de caso usando estimativa de pose automática (*Kinect*)  $\times$  manual (usuário). Em (a) e (d), imagens de entrada, capturadas pela câmera do *Kinect*. Em (b) e (e), resultados (em vermelho), usando os dados de entrada capturados pelo *Kinect* (*ground truth* exibido em azul). Em (c) e (f), resultados (em vermelho), usando os dados de entrada informados pelo usuário (*ground truth* exibido em azul).

informado pelo usuário. Isso justifica-se pelo fato da pose, estimada de forma automática, nem sempre estar centralizada no corpo da pessoa na imagem, podendo às vezes nem mesmo estar inserida na região da pessoa, como ilustrado na Figura 4.14. Por outro lado, esse experimento demonstra que é possível tornar o modelo proposto totalmente automático, uma vez que os dados de entrada também sejam adquiridos de maneira automática, produzindo ainda resultados satisfatórios.

#### 4.5 Comparação qualitativa com estado-da-arte

Nessa seção é apresentada uma breve comparação qualitativa entre alguns resultados gerados pelo modelo proposto nessa tese e resultados gerados no trabalho de Freifeld e sua equipe [5], considerado estado-da-arte. É importante salientar que os dados de entrada utilizados nesse estudo de caso, além de serem adquiridos de maneiras diferentes (manualmente no modelo proposto e de forma automática em [5]), também diferem entre si (diferentes modelos de esqueleto).

A Figura 4.15(a) ilustra 3 imagens, nas quais as pessoas tiveram seus dados de entrada (altura

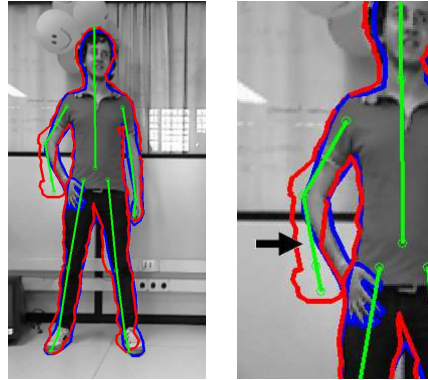


Figura 4.14: Problema associado à estimativa automática da pose (resultado em vermelho e *ground truth* em azul). Em (a), “osso” associado à uma determinada parte do corpo fora da região desejada (antebraço direito e mão direita). Em (b), *zoom* na imagem (a).

e pontos associados ao modelo de esqueleto) informados através de interação com usuário, para que pudessem ser processados pela abordagem proposta. O modelo de esqueleto informado pelo usuário, para cada imagem ilustrada na Figura 4.15(a), é exibido em ciano. Por outro lado, os resultados gerados pelo trabalho de Freifeld e sua equipe [5] (fornecidos pelos autores e ilustrados na Figura 4.15(b)) foram obtidos com a utilização de um detector automático de pessoas e pose [26]. A Figura 4.15(c) ilustra os dados de entrada (gerados de maneira automática usando [26]) utilizados no trabalho de Freifeld e sua equipe [5].

Qualitativamente, os resultados experimentais exibidos na Figura 4.15(a-b) demonstram que o modelo apresentado nesta tese gera resultados mais coerentes para o contorno da pessoa, enquanto que os contornos gerados pelo trabalho de Freifeld e sua equipe [5] apresentam formas mais suaves. Deve ser salientado que nesse estudo de caso está sendo desconsiderada a forma na qual os dados de entrada foram obtidos e suas diferenças.

Freifeld e sua equipe [5] também compararam seus resultados com os obtidos usando *Grab-Cut* [15], conforme ilustrado com auxílio da Figura 4.16(b-c). É importante mencionar que a informação semântica associada ao contorno, assim como sua própria conectividade, estão sendo desconsideradas nesse tipo de análise, uma vez que o objeto segmentado (*blob* ou objeto de *foreground*) pelo *Grab-Cut* não oferece esse tipo de informação.

Objetivando ilustrar o problema mencionado no parágrafo anterior, considere por exemplo, uma pessoa em uma imagem, com os braços cruzados na frente do tronco. O método *Grab-Cut*, assim como o algoritmo *Graph Cuts* [16], iria gerar um único *blob* (objeto segmentado ou objeto de *foreground*), onde os braços cruzados e o tronco iriam estar unidos por um conjunto de *pixels*, conforme ilustrado na Figura 4.17(c).

A seguir, no Capítulo 5, são apresentadas considerações finais sobre o modelo de segmentação de pessoas em imagens estáticas baseado em esqueleto, proposto nessa tese, assim como sugestões de aperfeiçoamentos e trabalhos futuros.



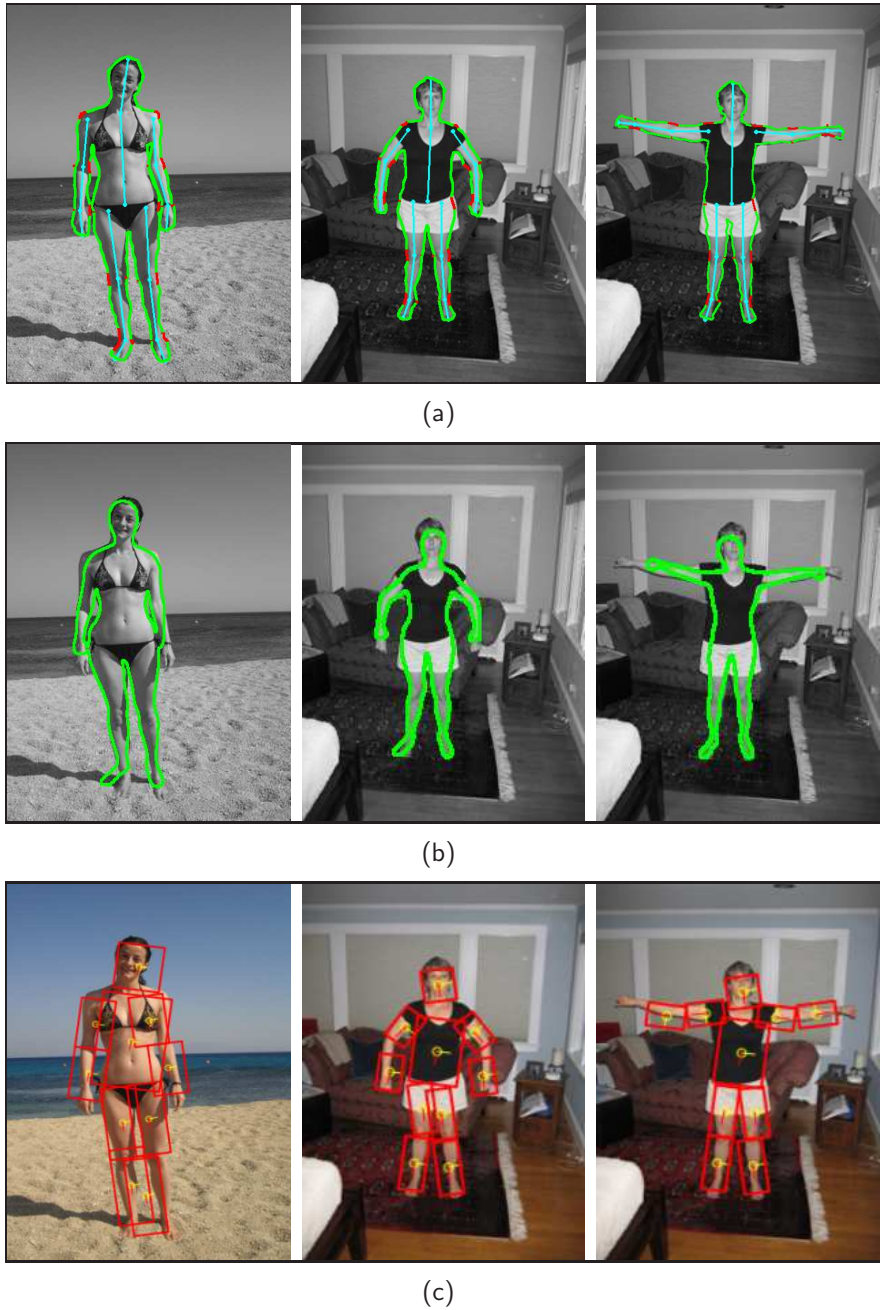


Figura 4.15: (a) Resultado gerado pelo modelo proposto. Modelo de esqueleto informado pelo usuário ilustrado em ciano. (b) Resultados obtidos por Freifeld e sua equipe [5]. (c) Dados de entrada utilizados no trabalho de Freifeld e sua equipe [5]



Figura 4.16: (a) Resultado gerado pelo modelo proposto. (b) Resultado obtidos por Freifeld e sua equipe [5]. (c) Resultado obtidos usando *Grab-Cut* [15] com inicialização manual, utilizado como comparativo no trabalho de Freifeld e sua equipe [5].

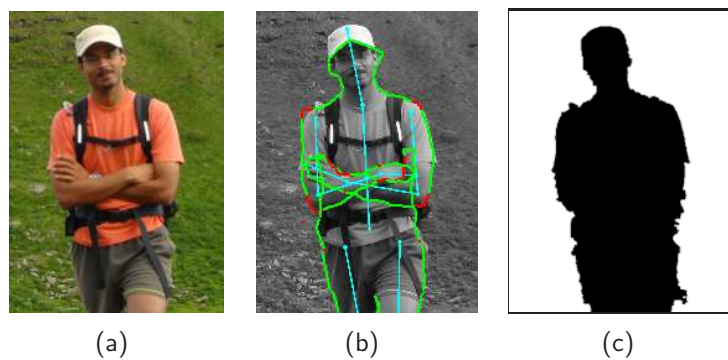


Figura 4.17: (a) Imagem de entrada. (b) Resultado gerado pelo modelo proposto. (c) Resultado obtido usando *Graph Cuts* [16] com inicialização manual.

## 5. Considerações Finais e Trabalhos Futuros

Segmentar pessoas em imagens estáticas, com a utilização de técnicas de visão computacional, é uma tarefa bastante desafiadora, assim como a de se obter informações semânticas das pessoas contidas nessas imagens, devido a diversos fatores do mundo real, como por exemplo, grande variabilidade de aparências e poses que essas podem assumir, assim como fatores relacionados à iluminação da cena onde a imagem foi capturada, sombras, ruídos na imagem, oclusão, alta similaridade do objeto de interesse com o fundo da cena e a falta de informação inerente de profundidade quando uma cena é capturada em uma imagem 2D [3].

No decorrer do curso de doutorado foi proposta uma técnica para segmentação e estimativa automática da pose 2D de pessoas em imagens estáticas, publicada em uma conferência da área [3]. Em [3], a segmentação da pessoa é realizada sem intervenção manual, inicializada a partir de um detector de faces automático [24], onde o objetivo inicial é encontrar cores predominantes em regiões específicas, estimadas a partir de parâmetros antropométricos. O resultado final desse trabalho [3] é um método para estimativa de poses 2D de pessoas em imagens estáticas (basicamente da parte superior do corpo - tronco e membros superiores). Entretanto, conforme relatado por Hornung e sua equipe [4], a aquisição da postura 2D de um ser humano de forma interativa tem algumas vantagens quando comparada à métodos automáticos, pois a intervenção manual normalmente leva alguns minutos e gera resultados superiores em poses onde há alguma ambiguidade, se comparado à técnicas de estimativa de pose automáticas. Dessa forma, devido a inúmeros fatores que fazem com que não seja trivial a resolução desse problema (tanto o de estimativa de poses como o de segmentação automática), pretendeu-se também investigar vantagens/desvantagens de métodos que permitam intervenção com o usuário, assim como estender o trabalho proposto em [3] para segmentar o corpo todo (ao invés de somente a parte superior do corpo). O resultado final desse processo investigativo resultou no modelo proposto nessa tese, para segmentação de pessoas em imagens estáticas baseada em esqueleto.

Na abordagem proposta não são usados modelos 3D complexos da forma humana, como em [1, 2, 11] nem base de dados para aprendizado de formas, aparências e/ou poses, como em [5, 7, 9, 31]. O modelo de esqueleto guia a segmentação da pessoa na imagem, levando em consideração informações de cor, luminosidade, restrições de ângulos e parâmetros antropométricos. De uma forma geral, a idéia principal da abordagem proposta é construir um grafo ao redor do modelo de esqueleto, para uma determinada imagem de entrada, e buscar o melhor caminho nesse grafo que satisfaça uma determinada condição (por exemplo, aquela que maximiza certo critério de energia), gerando assim o contorno da pessoa na imagem.

Uma característica importante, que deve ser salientada do modelo proposto, é que o resultado dessa abordagem gera um contorno fechado (onde o ponto inicial é igual ao ponto final) com informação semântica embutida, ou seja, cada ponto do contorno resultante está associado a uma determinada parte do corpo (similar ao trabalho de Freifeld e sua equipe [5], considerado estado-

da-arte). Tal informação semântica torna possível, por exemplo, que duas partes do corpo fiquem sobrepostas (como os braços na frente do tronco, ou pernas cruzadas), mantendo ainda uma conectividade coerente do contorno (uma vez que se sabe quais partes do grafo estão associadas a quais partes do corpo e suas respectivas regiões de adjacência). Tal característica pode ser utilizada para diversos fins, como por exemplo a construção de humanos virtuais baseada imagem (como geometria ou textura [33]), métodos para estimativa de roupas em imagens [6], estimativa da forma humana sobre as roupas [34, 35], entre outros.

Em processamento de imagens é bastante comum a utilização de base de dados gerada por especialista para avaliar experimentos. Para avaliar as características usadas no modelo proposto, assim como outros aspectos referentes ao mesmo, foi proposta uma abordagem para analisar quantitativamente os resultados experimentais obtidos, a partir de informações adquiridas manualmente, descrita em detalhes na Seção 4.1. A metodologia proposta permite avaliar, de maneira local ou global, o erro entre o contorno gerado para uma determinada pessoa em uma imagem e seu contorno esperado (estimado manualmente). As simplificações adotadas, assim como desafios enfrentados, são discutidas na Seção 4.1 e podem servir como ponto de partida para trabalhos futuros.

As características usadas no modelo de segmentação proposto, para avaliar a energia do contorno, foram avaliadas no estudo de caso apresentado na Seção 4.2. Com base no experimento apresentado na Seção 4.2, pôde-se verificar que algumas características usadas tiveram pouca influência nos resultados, assim como, também pôde-se observar quais delas tiveram maior impacto nos resultados. Com base nas métricas de erro empregadas, a abordagem utilizada que apresentou melhores resultados foi aquela que levava em conta todas as características mencionadas (informações de cor, luminosidade, restrições de ângulos e parâmetros antropométricos).

O modelo de segmentação proposto também foi avaliado em relação à sensibilidade em função dos dados de entrada. Na Seção 4.3 foi apresentado um estudo de caso onde 24 usuários (pessoas de diversas áreas do conhecimento e não especificamente de processamento de imagens) foram instruídos à clicar em 3 imagens (contendo uma única pessoa em cada imagem) para inserir os dados de entrada (altura e pontos do esqueleto, associados ao modelo de esqueleto). Com base nesse experimento, pôde-se concluir que, apesar dos usuários informarem os dados de entrada de maneira variada, os resultados mantiveram-se satisfatórios, fazendo com que pequenas variações em relação aos dados de entrada não acarretassem alterações muito impactantes nos contornos obtidos. O experimento conduzido no estudo de caso apresentado na Seção 4.3 também pode demonstrar o quão simples pode ser a entrada de dados via usuário, uma vez que apenas alguns cliques são necessários, fazendo com que a média de tempo gasto para cada imagem seja menor que 2 minutos.

O modelo proposto nessa tese, descrito em detalhes no Capítulo 3, utiliza dados de entrada (associados ao modelo de esqueleto) que podem ser obtidos de forma automática (utilizando um algoritmo para estimativa de pose 2D de pessoas em imagens [7, 12, 13, 26], por exemplo) ou manual (informados por um usuário), dependendo da aplicação em questão. O estudo de caso apresentado na Seção 4.4 indica que os resultados de segmentação obtidos com os dados de entrada inseridos de forma manual geram resultados mais coerentes do que os obtidos com os dados de entrada

adquiridos de forma automática (uma vez que existem diversos desafios a serem superados em se tratando de métodos automáticos para estimativa de pose 2D de pessoas em imagens). Entretanto, o experimento apresentado na Seção 4.4 demonstra que os resultados de segmentação obtidos com os dados de entrada adquiridos de forma automática também podem ser considerados satisfatórios, mesmo que não superem os obtidos a partir de dados adquiridos através de intervenção com usuário.

Os resultados obtidos com a utilização do modelo proposto também foram comparados (qualitativamente) com os obtidos por um trabalho considerado estado-da-arte [5], no estudo de caso apresentado na Seção 4.5. Os experimentos indicam que o modelo proposto nessa tese gera resultados mais coerentes para o contorno da pessoa, enquanto que os contornos obtidos pelo trabalho em questão [5] apresentam formas mais suaves. De uma forma geral, os experimentos realizados demonstram que o modelo proposto nessa tese gera resultados satisfatórios para imagens não triviais, contendo pessoas com aparências e poses variadas (podendo haver membros parcialmente ocultos), em diversos ambientes complexos (e não controlados), com diferentes iluminações e qualidade de imagem, entre outros fatores.

Uma limitação do modelo proposto é tratar poses onde o movimento dos membros (das pessoas contidas nas imagens) não está aproximadamente no mesmo plano da imagem (o que afeta as estimativas antropométricas na imagem projetada). Outros fatores podem fazer com que os resultados gerados sejam indesejáveis, como por exemplo, grande complexidade da pose, oclusão parcial ou total de membros, problemas associados a fatores de iluminação, entre outros. Dessa forma, pretende-se investigar, em trabalhos futuros, alternativas para minimizar os problemas relacionados aos fatores mencionados, assim como tratar poses mais complexas, principalmente àquelas onde os membros das pessoas não estão aproximadamente no mesmo plano da imagem. A estimativa de pose 3D de uma pessoa em uma imagem estática também é um trabalho bastante desafiador. De certa forma, outra sugestão para trabalho futuro seria utilizar o contorno obtido através da utilização do modelo proposto, combinado com os valores de energia associados a cada ponto do contorno, para construir um modelo de estimativa de pose 3D de pessoas em imagens estáticas, assumindo-se que regiões ocultas do contorno deveriam apresentar valores de energia baixo em relação à pontos não ocultos, fazendo com que essa informação gerada pelo modelo proposto (energia associada ao contorno) se torne relevante, podendo ser utilizada também para outras finalidades que não a de segmentação de pessoas em imagens estáticas.



## A. Apêndice - Trabalhos publicados

Durante o período do curso de doutorado, foram realizados trabalhos em áreas relacionadas à visão computacional porém não exclusivamente sobre segmentação de pessoas em imagens estáticas. Abaixo segue a lista de trabalhos publicados e submetidos.

### A.1 Artigos completos publicados em periódicos

- ★ Jacques Junior, J. C. S.; Musse, Soraia Raupp; Jung, Cláudio Rosito. ***Crowd analysis using computer vision: a survey***. IEEE Signal Processing Magazine (SPM), 2010.
- Silveira, César L. B.; Cavalheiro, Gerson Geraldo H.; Jung, Cláudio R.; Jacques Junior, J. C. S.; Musse, Soraia R. ***An improved background subtraction algorithm and concurrent implementations***. Parallel Processing Letters (PPL), v. 20, p. 71, 2010.

### A.2 Artigos completos publicados em anais de congressos

- Braun, H.; Souto Junior, H.; Jacques Junior, J. C. S.; Dihl, L. L.; Braun, Adriana; Musse, Soraia Raupp; Jung, Claudio Rosito; Thielo, M. R.; Keshet, R. ***Making Them Alive***. X Simpósio Brasileiro de Games e Entretenimento Digital (SBGAMES), 2011, Salvador - Bahia.
- ★ Jacques Junior, J. C. S.; Dihl, L. L.; Jung, Cláudio Rosito; Thielo, M. R.; Keshet, R.; Musse, Soraia Raupp. ***Human upper body identification from images***. IEEE International Conference on Image Processing (ICIP), 2010
- Queiroz, Rossana; Cohen, M.; Juliano L Moreira; Braun, Adriana; Jacques Junior, J. C. S.; Musse, S. R. ***Generating Facial Ground Truth with Synthetic Faces***. XXIII Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens (SIBGRAPI), 2010, Gramado/RS.
- Braun, H. ; Hocevar, R.; Queiroz, R.; Cohen, M.; Moreira, J. L.; Jacques Junior, J. C. S.; Braun, Adriana; Musse, Soraia Raupp; Samadani, R. ***VhCVE: A Collaborative Virtual Environment Including Facial Animation and Computer Vision***. VIII Simpósio Brasileiro de Games e Entretenimento Digital (SBGAMES), 2009, Rio de Janeiro. p. 207-213.
- Jacques Junior, J. C. S.; Moreira, J. L.; Braun, Adriana; Musse, Soraia Raupp; Said, Amir. ***A Template-Matching Based Method to Perform Iris Detection in Real-Time Using Synthetic Templates***. Proceedings of the 2009 11th IEEE International Symposium on Multimedia (ISM), 2009, San Diego, p. 142-147.

- Paravisi, M.; Werhli, Adriano; **Jacques Junior, J. C. S.**; Rodrigues, R.; Bicho, A.; Jung, Cláudio Rosito; Musse, Soraia Raupp . ***Continuum Crowds with Local Control***. Proceedings of Computer Graphics International (CGI), 2008, Istambul. v. 1. p. 108-115.

### **A.3 Artigo aceito para publicação**

- **Jacques Junior, J. C. S.**; Cláudio Rosito; Musse, Soraia Raupp. ***Skeleton-based human segmentation in still images***. Submitted to IEEE International Conference on Image Processing (ICIP), 2012. Orlando, Florida, U.S.A.



## B. Apêndice - Prêmios recebidos

Abaixo segue uma lista de prêmios recebidos durante o período do curso de doutorado:

- 2011: 2º melhor artigo do X Simpósio Brasileiro de Games e Entretenimento Digital (SBGAMES), Salvador - Bahia.
- ★ 2011: IBM Ph.D. Fellowship Awards<sup>1</sup>.  
*(O IBM Ph.D. Fellowship é direcionado aos doutorandos que estudam, em suas teses, questões que contribuam para a solução de problemas de interesse da IBM e que representem contribuições científicas significativas para as áreas de Ciência da Computação, Engenharia Elétrica e de Materiais, Ciências Físicas, Matemática, e Ciência de Serviços. Os prêmios são concedidos anualmente para doutorandos que se destacam por sua atuação acadêmico-científica em todo o mundo, através de um processo global de seleção altamente competitivo).*
- 2008: Melhor Filme na categoria Técnica da Mostra de Vídeos do XXI *Brazilian Symposium on Computer Graphics and Image Processing* (SIBGRAPI 2008).

---

<sup>1</sup>Mais informações em: <https://www.ibm.com/developerworks/university/phdfellowship/>



## C. Apêndice - Lista de parâmetros e valores padrão

A Tabela C.1 exibe uma lista detalhada dos principais parâmetros usados no modelo proposto, assim como os valores padrão adotados e a seção onde são definidos no texto. É importante salientar que, dependendo da versão do modelo proposto usada (detalhada com auxílio da Tabela 4.1), alguns parâmetros podem não ser usados.

Tabela C.1: Parâmetros usados no modelo proposto.

Informação complementar	Variável adotada	Valor padrão
Número de ligações entre um nodo do grafo com o nível adjacente (Seção 3.2.1).	$k$	3
Fator usado para setar o número inicial de níveis do grafo de uma determinada parte do corpo (Seção 3.2.1).	$s_4$	0.1
Fator usado para setar o número de nodos em um nível do grafo, para uma determinada parte do corpo (Seção 3.2.1).	$s_6$	0.33
Ângulo usado para criar o setor circular dos grafos das mãos e pés (Seção 3.2.2).		$22.5^\circ$
Fator de escala da Gaussiana, usada na restrição de distâncias antropométricas (Seção 3.2.4).		$w_i/4$
Fator usado para definir a largura da região de aprendizado da cor predominante (Seção 3.3.1).	$s_1$	0.4
Número máximo de modelos de cor criados para cada parte do corpo (Seção 3.3.4).	$N_r$	3
Fator usado para desconsiderar um modelo de cor em função de sua área (Seção 3.3.1).		0.1
Limiar usado para eliminar a dimensão do modelo de cor (PCA, Seção 3.3.2).	$T_p$	0.001
Fator usado para setar o comprimento da região de busca de uma cor (Seção 3.3.3).	$s_2$	1.15
Fator usado para setar a largura da região de busca de uma cor (Seção 3.3.3).	$s_3$	2



## D. Anexo - Base de dados

Nas Figuras D.1 e D.2 são ilustradas algumas imagens usadas neste trabalho (mais especificamente aquelas que foram usadas para ilustrar algum resultado), no seu formato original (em cores e sem cortes, porém redimensionadas). As imagens foram adquiridas de bases de dados públicas [40–42], encontradas na *web*, ou capturadas através de câmeras fotográficas convencionais em nosso laboratório.



Figura D.1: Imagens originais (porém reduzidas), usadas para ilustrar os resultados obtidos nesse trabalho.

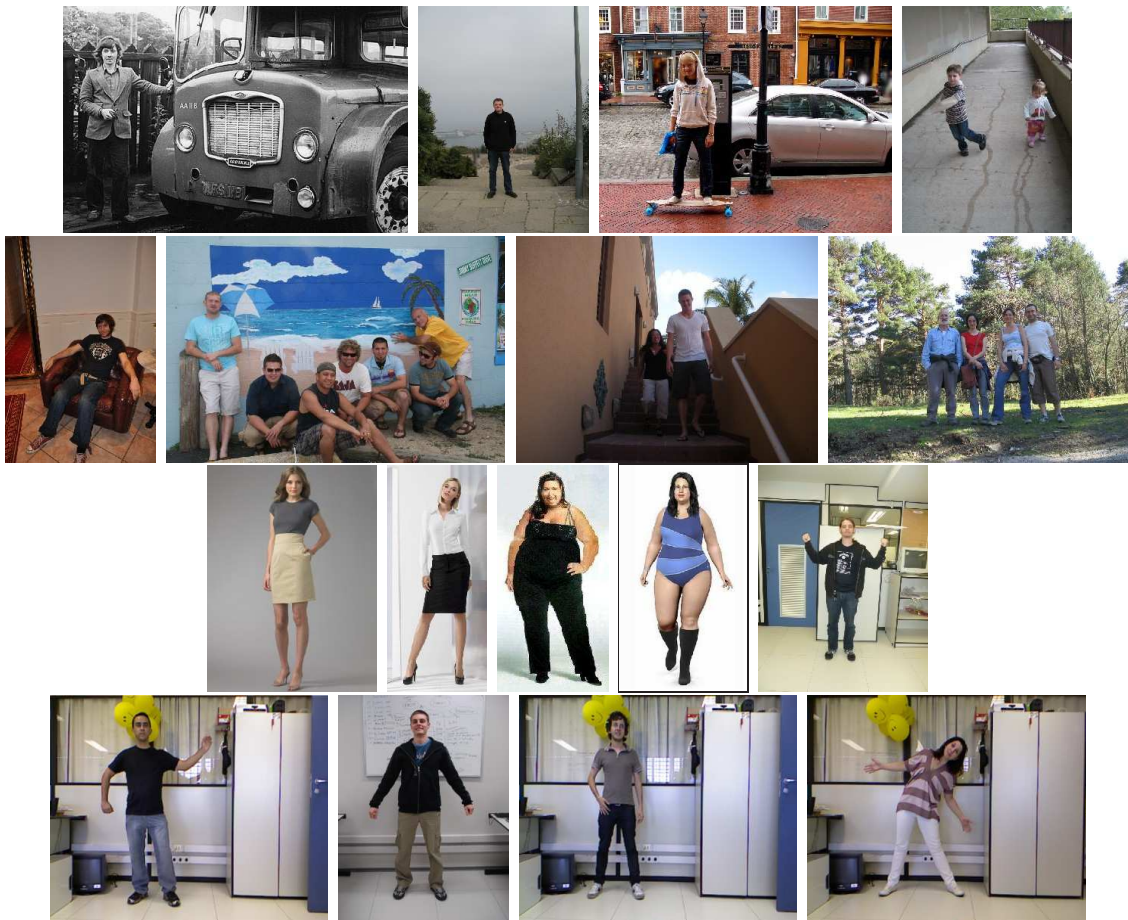


Figura D.2: Imagens originais (porém reduzidas), usadas para ilustrar os resultados obtidos nesse trabalho.



## Referências Bibliográficas

- [1] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans peter Seidel, "Multilinear pose and body shape estimation of dressed subjects from image sets," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [2] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han, "Parametric reshaping of human bodies in images," in *ACM SIGGRAPH papers*, New York, USA, 2010, pp. 126:1–126:10.
- [3] J. C. S. Jacques Jr., L. Dihl, C. R. Jung, M. R. Thielo, R. Keshet, and S. R. Musse, "Human upper body identification from images," in *International Conference on Image Processing. IEEE Computer Society Conference on*, Hong Kong, China, 2010, pp. 1717 – 1720.
- [4] Alexander Hornung, Ellen Dekkers, and Leif Kobbelt, "Character animation from 2d pictures and 3d motion data," *ACM Trans. Graph.*, vol. 26, January 2007.
- [5] Oren Freifeld, Alexander Weiss, Silvia Zuffi, and Michael J.Black, "Contour people: A parameterized model of 2d articulated human shape," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [6] Zhilan Hu, Hong Yan, and Xinggang Lin, "Clothing segmentation using foreground and background estimation based on the constrained delaunay triangulation," *Pattern Recognition*, vol. 41, pp. 1581–1592, May 2008.
- [7] Greg Mori, Xiaofeng Ren, Alexei A. Efros, and Jitendra Malik, "Recovering human body configurations: Combining segmentation and recognition," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, pp. 326–333, 2004.
- [8] David R. Martin, Charless C. Fowlkes, and Jitendra Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 530–549, 2004.
- [9] Zhe Lin, Larry S. Davis, David Doermann, and Daniel DeMenthon, "Hierarchical part-template matching for human detection and segmentation," *Computer Vision, IEEE International Conference on*, vol. 0, pp. 1–8, 2007.
- [10] Jian Sun, Jian Sun, Sing Bing Kang, Zong-Ben Xu, Xiaoou Tang, and Heung-Yeung Shum, "Flash cut: Foreground extraction with flash and no-flash image pairs," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 1–8, 2007.
- [11] Peng Guan, Alexander Weiss, Alexandru O. Balan, and Michael J. Black, "Estimating human shape and pose from a single image," in *International Conference on Computer Vision, IEEE Computer Society Conference on*, 2009, pp. 1381 –1388.
- [12] Zhilan Hu, Guijin Wang, Xinggang Lin, and Hong Yan, "Recovery of upper body poses in static images based on joints detection," *Pattern Recognition Letters*, vol. 30, pp. 503–512, April 2009.

- [13] Xiaofeng Ren, Alexander C. Berg, and Jitendra Malik, "Recovering human body configurations using pairwise constraints between parts," *Computer Vision, IEEE International Conference on*, vol. 1, pp. 824–831, 2005.
- [14] C. R. Jung, "Unsupervised multiscale segmentation of color images," *Pattern Recognition Letters*, vol. 28, no. 4, pp. 523–533, March 2007.
- [15] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, "Grabcut: interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, pp. 309–314, August 2004.
- [16] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," *Computer Vision . in Proceedings of the Eighth IEEE International Conference on*, vol. 1, pp. 105–112 vol.1, 2001.
- [17] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing (3rd Edition)*, Prentice Hall, August 2007.
- [18] Kevin McGuinness, *Image Segmentation, Evaluation, and Applications*, PhD in electronic engineering, School of Electronic Engineering – Dublin City University (DCU), 2009.
- [19] Tammy Riklin-Raviv, Nir Sochen, and Nahum Kiryati, "On symmetry, perspectivity, and level-set-based segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 1458–1471, 2009.
- [20] M. Dimitrijevic, V. Lepetit, and P. Fua, "Human body pose detection using bayesian spatio-temporal templates," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 127–139, 2006.
- [21] Ronald Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.
- [22] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 1, pp. 44–58, 2006.
- [23] Rafael Muñoz-Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, and A. Carmona-Poyato, "Depth silhouettes for gesture recognition," *Pattern Recognition Letters*, vol. 29, no. 3, pp. 319 – 329, 2008.
- [24] P. A. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [25] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis, "Scape: shape completion and animation of people," *ACM Trans. Graph.*, vol. 24, pp. 408–416, July 2005.
- [26] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pp. 1014–1021, 2009.
- [27] Leonidas Guibas and Jorge Stolfi, "Primitives for the manipulation of general subdivisions and the computation of voronoi," *ACM Trans. Graph.*, vol. 4, pp. 74–123, April 1985.
- [28] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100–108, 1979.



- [29] Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi, "Contour and Texture Analysis for Image Segmentation," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 7–27, June 2001.
- [30] Xiaofeng Ren and Jitendra Malik, "Learning a classification model for segmentation," *Computer Vision, IEEE International Conference on*, vol. 1, pp. 10–17, 2003.
- [31] Dariu M. Gavrilă, "A bayesian, exemplar-based approach to hierarchical shape matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, pp. 1408–1421, August 2007.
- [32] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to Algorithms, Second Edition*, McGraw-Hill Science/Engineering/Math, July 2001.
- [33] H. Braun, H. Souto Junior, J. C. S. Jacques Junior, L. L. Dihl, Adriana Braun, Soraia Raupp Musse, Claudio Rosito Jung, M. R. Thielo, and R. Keshet, "Making them alive," in *Simpósio Brasileiro de Games e Entretenimento Digital (SBGAMES)*, Salvador - Bahia, 2011, pp. 1 – 10.
- [34] Alexandru O. Bălan and Michael J. Black, "The naked truth: Estimating body shape under clothing," in *Proceedings of the 10th European Conference on Computer Vision: Part II*, Berlin, Heidelberg, 2008, ECCV '08, pp. 15–29, Springer-Verlag.
- [35] Peng Guan, Oren Freifeld, and Michael J. Black, "A 2d human body model dressed in eigen clothing," in *Proceedings of the 11th European conference on Computer vision: Part I*, Berlin, Heidelberg, 2010, ECCV'10, pp. 285–298, Springer-Verlag.
- [36] Hanzi Wang and David Suter, "Tracking and segmenting people with occlusions by a sample consensus based method," in *IEEE International Conference on Image Processing*, 2005, vol. 2, pp. 410–413.
- [37] Changhyung Lee, Morgan T. Schramm, Mireille Boutin, and Jan P. Allebach, "An algorithm for automatic skin smoothing in digital portraits," in *Proceedings of the 16th IEEE international conference on Image processing*, Piscataway, NJ, USA, 2009, pp. 3113–3116.
- [38] Alvin R. Tilley, *The measure of man and woman - Human factors in design*, John Wiley & Sons, inc, 2002.
- [39] D. Pena and F. Prieto, "Multivariate outlier detection and robust covariance matrix estimation," *Technometrics*, vol. 43, pp. 286–310, 2001.
- [40] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Computer Vision and Pattern Recognition, IEEE Conference on*, june 2008, pp. 1 –8.
- [41] Lubomir Bourdev and Jitendra Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Computer Vision, 12th IEEE International Conference on*, 2009, pp. 1365–1372.
- [42] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, june 2005, vol. 1, pp. 886 –893 vol. 1.

- [43] T. List, J. Bins, J. Vazquez, and R. B. Fisher, "Performance evaluating the evaluator," in *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance.*, 2005, pp. 129–136.
- [44] M. Hossain, M. Dewan, Kiok Ahn, and Oksam Chae, "A linear time algorithm of computing hausdorff distance for content-based image analysis," *Circuits, Systems, and Signal Processing*, vol. 31, pp. 389–399, 2012.