UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

FÁBIO CISNE RIBEIRO

TÉCNICAS PARA AVALIAÇÃO E IDENTIFICAÇÃO DE POSTURAS HUMANAS EM TEMPO REAL POR VISÃO COMPUTACIONAL MONOCULAR

Fortaleza-CE, Brasil
Abril de 2009

Universidade Federal do Ceará
Centro de Tecnologia
Departamento de Engenharia de Teleinformática
Programa de Pós Graduação em Engenharia de Teleinformática

DISSERTAÇÃO DE MESTRADO

FÁBIO CISNE RIBEIRO

TÉCNICAS PARA AVALIAÇÃO E IDENTIFICAÇÃO DE POSTURAS HUMANAS EM TEMPO REAL POR VISÃO COMPUTACIONAL MONOCULAR

ORIENTADOR:
PROF. DR. JOSÉ MARQUES SOARES

CO-ORIENTADOR:
PROF. DR. PAULO CÉSAR CORTEZ

Dissertação de Mestrado apresentada à Coordenação do Curso de Pós-Graduação em Engenharia de Teleinformática da Universidade Federal do Ceará como parte dos requisitos para obtenção do grau de Mestre em Engenharia de Teleinformática.

Fortaleza-CE, Brasil
Abril de 2009

Fábio Cisne Ribeiro

Técnicas para avaliação e identificação de posturas humanas em tempo real por visão computacional monocular.

Esta dissertação foi julgada adequada na defesa para a obtenção do título de Mestre em Engenharia de Teleinformática e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Teleinformática da Universidade Federal do Ceará.

ORIENTADOR: _____
PROF. DR. JOSÉ MARQUES SOARES

CO-ORIENTADOR: _____
PROF. DR. PAULO CÉSAR CORTEZ

Banca Examinadora:

_____
PROF. DR. JOSÉ MARQUES SOARES

_____
PROF. DR. PAULO CÉSAR CORTEZ

_____
PROF. DR. GIOVANNI CORDEIRO BARROSO

_____
PROF. DR. GUILHERME DE ALENCAR BARRETO

_____
PROF. DR. ANTÔNIO MACÁRIO CARTAXO DE MELO

Fortaleza-CE, abril de 2009

AGRADECIMENTOS

Aos meus amados familiares; meu pai, Anibal; minha mãe, Gláucia; meu irmão, Leonardo; e minha irmã, Manuelle; pelo amor, apoio, paciência e compreensão que sempre me foram fornecidos, ao longo de minha vida.

Ao meu amigo e orientador Dr. Jose Marques Soares pela ajuda, incentivo e apoio que sempre me fornece; e, principalmente, pela grande amizade conquistada.

Aos amigos e a todas as pessoas que me ajudaram nesta jornada.

A Deus, pela graça da vida.

Ao suporte financeiro recebido pelas agências de fomento CNPq e FUNCAP desde o Projeto COGEST.

E a Samsung pela infra-estrutura disponibilizada.

RESUMO

O presente trabalho disserta sobre as contribuições desenvolvidas para melhoria da qualidade e da eficiência de um sistema de visão computacional monocular voltado para a identificação, em tempo real, de posturas humanas, sem a utilização de hardware especializado ou de marcadores sobre o corpo. O sistema extrai de uma seqüência de imagens informações articulares que definem a postura humana. A técnica empregada executa o ajuste de um modelo humanóide tridimensional sobre a imagem do ser humano a cada imagem da seqüência. Para isso, é necessário processar a imagem capturada pela câmera, separando as informações de cor da pele e de fundo e, em seguida, minimizar a diferença entre essa imagem e as imagens sintéticas, geradas pela projeção do modelo tridimensional no plano imagem, de acordo com um mecanismo de predição. As primeiras contribuições deste trabalho para o sistema descrito consistem na concepção de um conjunto de métricas de similaridade que são usadas pelo algoritmo de minimização. Tais métricas utilizam informações de contornos, superfície de não recobrimento ou computam a diferença de *pixels*. Outras contribuições ao sistema, conjugadas às métricas de similaridade, incluem o uso de banco de dados em memória para auxílio à predição, o uso de restrições biomecânicas dinâmicas para impedir posturas ergonomicamente inválidas, a classificação independente das diferentes regiões de pele para identificação dos membros, entre outras adaptações e modificações. A fim de validar o trabalho, foram realizadas experimentações usando-se diversas combinações. Os resultados são comparados e apresentados.

Palavras-chaves: Identificação de posturas humanas, visão computacional, métricas de similaridade entre imagens, restrições biomecânicas.

ABSTRACT

This work presents contributions developed to improve the quality and efficiency of a monocular computer vision system aimed to identify human pose in real-time without the use of specialized hardware or markers over the body. The system extracts join information from a sequence of images that defines the human pose. The method used fits a humanoid three-dimensional model over human image for each image of the sequence. For this, it is necessary to segment the image captured by a camera, separating skin color information from background, and then minimizing the difference between segmented and synthetic images, generated by the projection of the three-dimensional model in an image according to a prediction method. The first contributions of this work to the system are the design of a set of similarity metrics that are used by the minimization algorithm. These metrics use contour information, non-overlapping surface or images pixel difference. Other contributions, combined to the metric of similarity include using memory database to aid prediction, biomechanical dynamic restrictions to prevent poses ergonomically invalid, different skin regions classification to differentiating body parts, among other adaptations and modifications. In order to validate this work, experiments were performed using different combinations. The results are compared and presented.

Key-words: Human pose estimation, computer vision, image similarity biomechanical restrictions.

SUMÁRIO

Lista de Figuras

Lista de Tabelas

Capítulo 1

1 Introdução

As interfaces homem-máquina evoluíram extraordinariamente desde a invenção dos primeiros computadores. Os cartões perfurados, os teclados, os *mouse*s e as telas sensíveis ao toque compõem um pouco dessa história que continua em franca evolução.

Uma das tecnologias inovadoras de interação com a máquina, na qual se concentram diversos trabalhos de pesquisa nos dias atuais, é a **visão computacional**, que engloba um conjunto de métodos e técnicas empregadas para processar e extrair informações de imagens digitais. O presente trabalho disserta sobre os resultados de pesquisas relacionadas à **identificação de posturas humanas** por visão computacional monocular, em seqüência de vídeo, em tempo real.

1.1 Identificação de gestos por visão computacional

Os gestos constituem um meio de comunicação natural do homem na vida cotidiana e reforçam outros veículos, como a fala. Podem ser usados como veículo único de comunicação, como no caso da linguagem de sinais utilizada por portadores de problemas auditivos.

Um gesto pode ser descrito como uma seqüência de posturas da mão, do braço, do corpo e da face. Neste último caso, usa-se com mais freqüência o termo “expressão” no lugar de gestos. Destaca-se aqui que a comunicação gestual humana é extremamente complexa e envolve todos esses elementos em conjunto. Os movimentos das partes do corpo humano não possuem um significado intrínseco. É a interpretação que associa ao movimento um sentido, a partir de um conhecimento *a priori* da semântica associada ao gesto.

Máquinas também são concebidas para serem pilotadas por gestos, através de interfaces mecânicas ou eletrônicas, como mostra nas Figuras 1-1 e 1-2. Sistemas computacionais vêm sendo pesquisados e desenvolvidos de maneira a identificar posturas humanas e codificá-las em parâmetros que permitam o seu uso como instrumento de manipulação de máquinas ou de objetos virtuais [1–9]. Uma das maneiras de aplicação dos parâmetros identificados é a animação de atores virtuais, objetos humanóides articulados, de maneira a restituir os gestos praticados pelo ser humano [1,10].

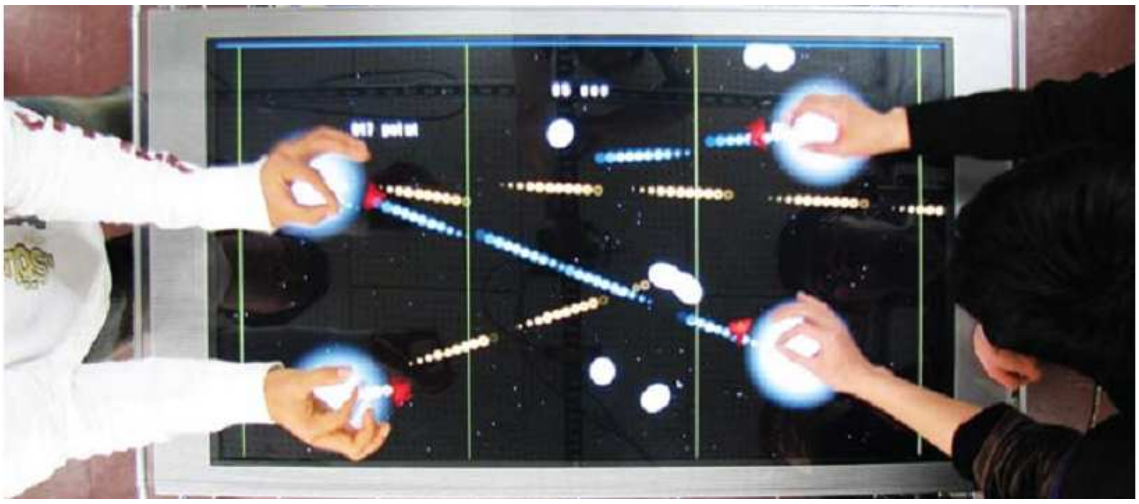

Figura 1-1: jogo eletrônico utilizando como interface homem-máquina os gestos da mão [11].

Figura 1-2: robô controlado pelo humano através de gestos da mão [12].

Uma das maneiras de capturar tais parâmetros é a instrumentação do corpo do ator com sensores eletrônicos, colocados em pontos-chave do corpo humano, como nas extremidades (pés e mãos) ou proximidades das articulações. A posição e a orientação de partes do corpo não instrumentadas com sensores podem ser calculadas por algoritmos apropriados, muitas vezes baseados em cinemática inversa. Em geral, tais sensores apresentam resultados precisos e em tempo real,

mas tais recursos são muito caros e, em geral, incômodos para o ator humano, como mostrado na Figura 1-3 [13].

Figura 1-3: captura de movimentos usando *hardware* especializado [13].

A visão computacional pode fornecer um mecanismo mais barato e menos incômodo ao ser humano para a identificação de posturas. Para isso, uma seqüência de imagens de vídeo, capturadas ou não em tempo real, pode ser processada e analisada. Em geral, tais métodos são menos incômodos ao ser humano, mas, em contrapartida, são menos precisos, principalmente devido à complexidade de serem extraídas informações tridimensionais de imagens que são bidimensionais. Além disso, o processamento de imagens para uma aquisição mais precisa é custosa e, em geral, só se alcança uma aquisição em tempo real impondo-se uma série de limitações ao cenário de aquisição (iluminação, cores de fundo, partes do corpo humano, tipo de movimento, etc.) [14].

Muitas técnicas são exploradas para identificação de posturas humanas em seqüências de imagens de vídeo. Algumas técnicas usam suporte de marcadores coloridos ou luminosos distribuídos pelo corpo, como mostrado na Figura 1-4 [15], o que restringe bastante o cenário de aquisição. Outras são baseadas em características de regiões ou manchas nas imagens (conhecidas como técnicas 2D) [8,16,3], fazendo identificações rústicas da postura. Técnicas bastante exploradas atualmente se apóiam no uso de modelos geométricos do corpo inteiro ou de parte a ser identificada. Estas últimas procuram estabelecer a correspondência entre

características das imagens sintéticas construídas a partir dos modelos e das imagens de vídeo processadas. A escolha e a complexidade da técnica usada, em geral, dependem da aplicação. Entretanto, se o objetivo for a restituição da postura humana em objetos humanóides virtuais [1], a técnica baseada em modelos torna-se mais apropriada, visto que, como o modelo e o ator podem possuir estruturas equivalentes, a partir dos parâmetros utilizados para posicionar o modelo de maneira adequada ao estabelecimento da correspondência, pode-se extraí-los e codificá-los para posterior animação de atores virtuais [17].

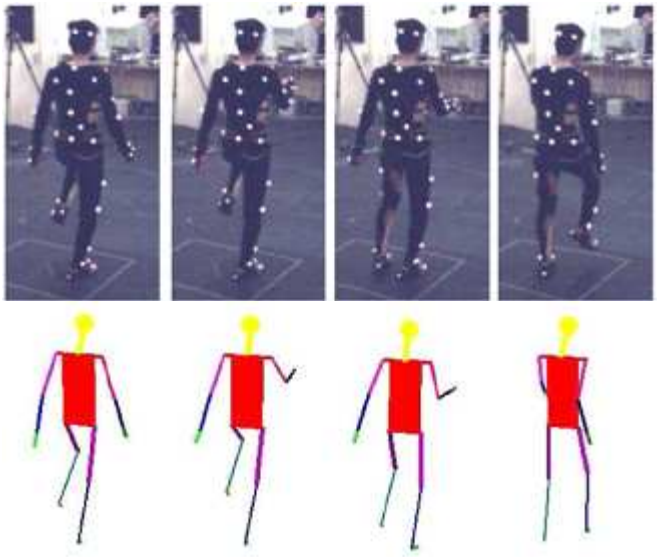

Figura 1-4: uso de marcadores sobre o corpo para identificação da postura [15].

Além disso, conhecimentos gerais da biomecânica do corpo humano podem ser usados para auxiliar os algoritmos na predição das posturas possíveis durante o processamento das imagens [18].

Para exemplificar tal técnica, pode-se conceber um modelo geométrico composto de segmentos conectados por articulações com dimensões e posições semelhantes ao do ator humano (pessoa cujos gestos estão sendo analisados), como exemplificado na Figura 1-5 [19]. O estado do modelo pode ser descrito pelos parâmetros de posição e os ângulos das articulações.

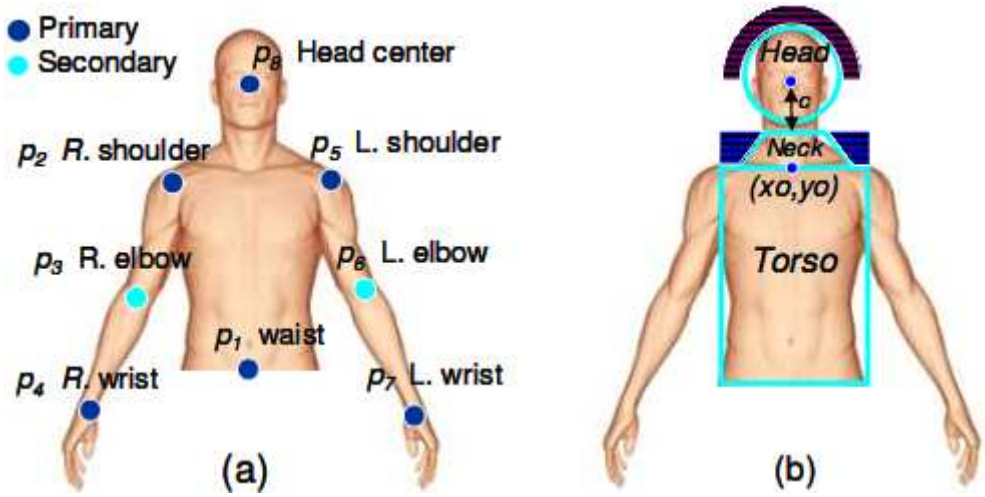

Figura 1-5: modelo geométrico composto de segmentos conectados por articulações com dimensões e posições semelhantes ao do corpo humano [19].

O procedimento de reconhecimento do gesto consiste em pesquisar, para cada imagem, o estado do modelo que apresenta a melhor correspondência segundo um critério de similaridade qualquer. Busca-se a melhor correspondência da imagem do modelo articulado com uma imagem capturada, como nos exemplos das Figuras 1-6 e 1-7, minimizando a diferença entre as características extraídas de cada imagem [17,20].

Figura 1-6: modelo 3D correspondente à postura humana [20].

Figura 1-7: modelo 3D correspondente à postura humana [17].

Como métrica de similaridade entre as características da imagem capturada pela câmera e a imagem de síntese, criada pela projeção do modelo, podem ser utilizadas diversas técnicas como, por exemplo, a distância entre os contornos [21], o fluxo óptico [7], e a taxa de não-recobrimento [1]. Para extrair tais características, as imagens devem ser processadas por algoritmos adequados [14, 22–24].

Algumas abordagens utilizam um conjunto de imagens tomadas de diversas câmeras simultaneamente, como é mostrado nos exemplos das Figuras 1-8, 1-9 e 1-10. Tais métodos permitem diminuir o problema da ausência de informações de profundidade, visto que câmeras podem ser dispostas em diversos ângulos, de maneira a melhorar a precisão na identificação de posturas. Técnicas estereoscópicas, no entanto, aumentam o custo computacional, pela necessidade de processamento simultâneo de mais de uma imagem, além de exigir cenários mais complexos para disposição e configuração de equipamentos [25,26].

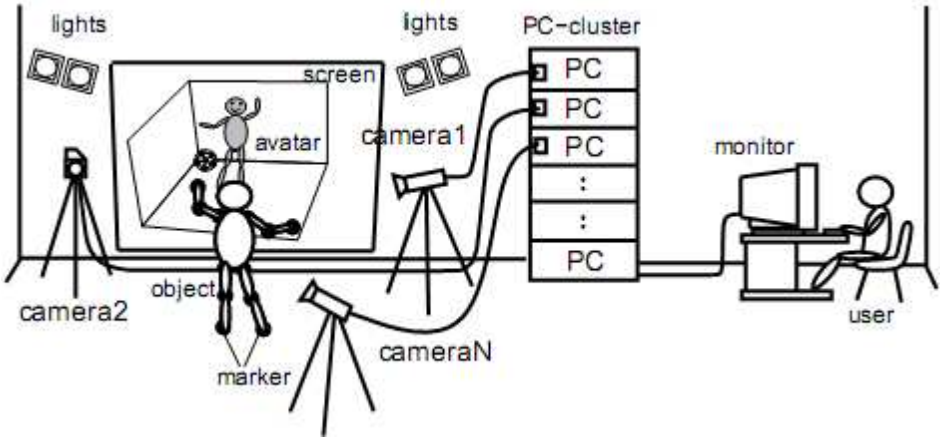

Figura 1-8: aquisição de gestos através de várias câmeras simultaneamente, usando vários computadores em *cluster* [26].

Figura 1-9: aquisição de gestos através de várias câmeras de alta resolução simultaneamente [27].

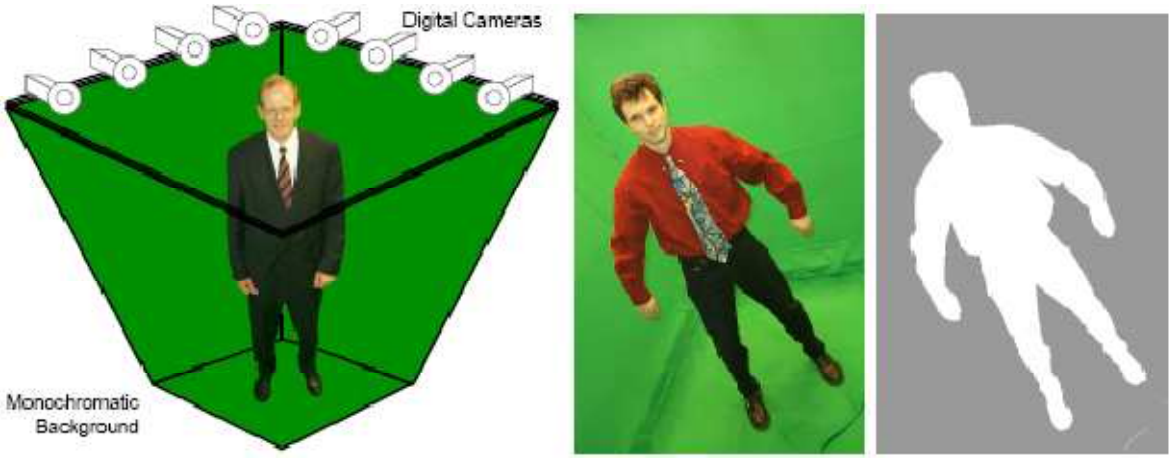

Figura 1-10: aquisição de gestos através de dezesseis câmeras simultaneamente e um ambiente com cor de fundo controlada [3].

Técnicas monoculares requerem, em geral, equipamentos mais simples e configuração menos complexa, apresentando, no entanto, fortes restrições relativas à ausência de informação de profundidade e oclusão (superposição de partes do corpo).

1.2 O projeto COGEST

O presente trabalho é uma continuação do projeto COGEST (Comunicação Gestual à Distância com Humanóides Virtuais), que estendeu o trabalho de Soares [1] e foi desenvolvido de janeiro de 2005 a julho de 2007, financiado pelo CNPq com recursos do CT-INFO (processo número 506226/2004-2), sob a coordenação do Professor Dr. Giovanni Cordeiro Barroso.

Devido à natureza multidisciplinar do projeto COGEST, foram desenvolvidos trabalhos ligados a algoritmos de otimização multidimensionais [28], algoritmos de segmentação em classes de cor para separação da pele humana [29] e uma infra-estrutura para colaboração à distância [10]. Tal sistema foi desenvolvido com o objetivo de capturar informações sobre as posturas do indivíduo e de reconstruí-las em ambientes virtuais colaborativos. No entanto, o projeto culminou numa infra-estrutura para produção de diversos trabalhos de pesquisa, possibilitando experimentos nas várias áreas do conhecimento do referido projeto.

1.3 Aquisição de gestos por visão computacional monocular sem marcadores

A função do sistema de aquisição de gestos proposto por Soares [1] é extrair de uma seqüência de imagens os parâmetros que permitam atribuir a um modelo humanóide a mesma postura do ator. Estes parâmetros são estimados a partir do tratamento das imagens capturadas do ator por uma única câmera de vídeo (*webcam*). A técnica empregada consiste em ajustar um modelo humanóide tridimensional sobre o ator em cada imagem da seqüência de vídeo.

Na Figura 1-11 é mostrado um diagrama com o funcionamento do sistema e, em seguida, as etapas são especificadas pela legenda.

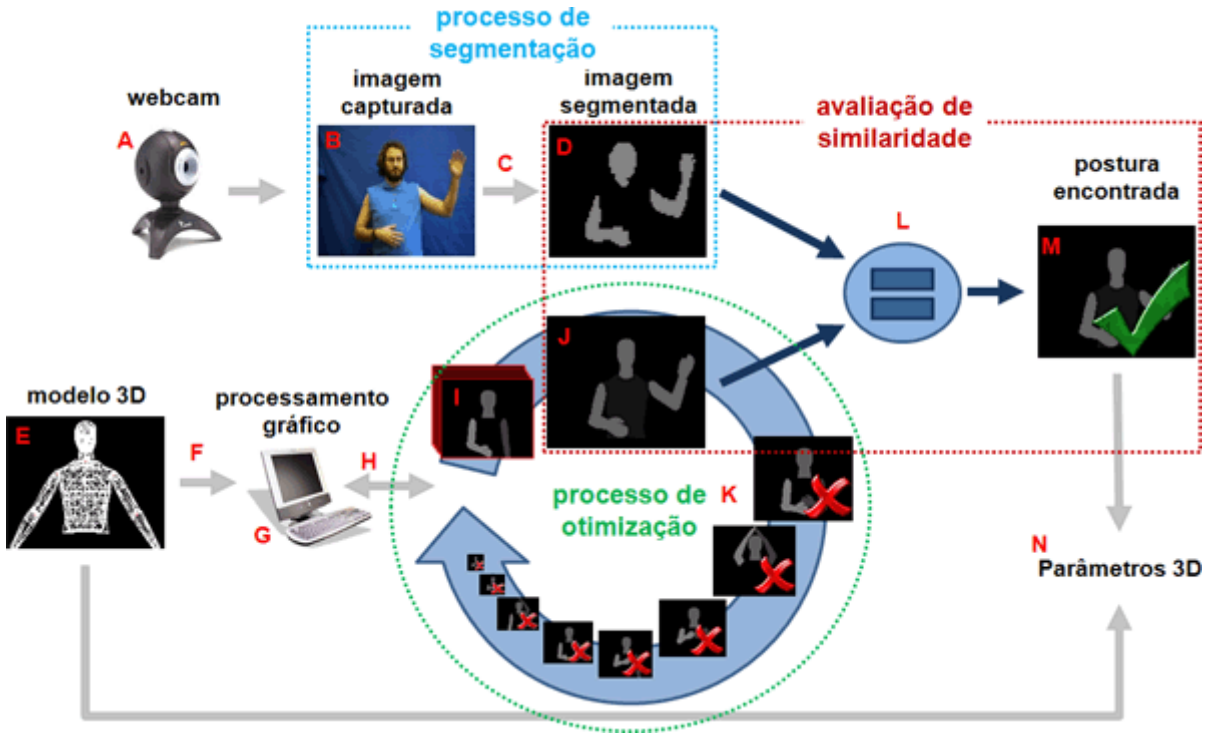

Figura 1-11: aquisição de imagens baseada em modelos tridimensionais do corpo humano por visão monocular.

A. *Webcam* (dispositivo de captura de imagem);

B. Imagem adquirida pelo dispositivo de captura (A);

C. Processo de segmentação, em que se identificam na imagem (B) as regiões de pele;

D. Imagem segmentada em (C);

E. Modelo tridimensional com dimensões semelhantes as do ator da imagem (B);

F. Envio do modelo 3D (E) para a placa gráfica;

G. Processamento gráfico realizado na placa gráfica do computador, onde são geradas as imagens no plano a partir da projeção do modelo 3D (E) nas posturas determinadas pelo processo de otimização (K);

H. Transferência da imagem da placa gráfica para a memória do computador;

I. Imagens sintéticas do modelo 3D geradas em (G) que ilustram o conjunto de dados para avaliação do processo de otimização;

J. Imagem sintética a ser comparada com a imagem segmentada (D) pela métrica de avaliação de similaridade (L);

K. Processo de otimização, que avalia as imagens (I) através da métrica de similaridade (L);

L. Métrica que avalia a similaridade entre as imagens (D) e (J) e define qual imagem sintética contém a melhor correspondência da postura do ator (B);

M. Imagem do processo de otimização (J) identificada como semelhante à imagem processada do ator (D) pela métrica de avaliação de similaridade (L);

N. Obtenção dos parâmetros do modelo 3D que foram utilizados para gerar a imagem (M), que podem ser armazenados ou transmitidos para um sistema de restituição dos gestos.

Um sistema de aquisição de imagens baseado em modelos tridimensionais do corpo humano pode ser estruturado nas seguintes etapas:

Construção da Representação Geométrica (E) – Nesta etapa, o modelo é construído com segmentos dispostos hierarquicamente e ajustado às dimensões do ator humano, de maneira a construir imagens que permitam medidas aproximativas com relação à imagem real. Para representação das articulações, alguns padrões internacionais foram estabelecidos. A representação geométrica e a codificação da

animação de humanóides virtuais são especificadas pelo padrão ISO/IEC MPEG-4 [30,31] tanto para o corpo como para a face humana. A utilização de padrões para a construção dos modelos usados no sistema de aquisição de gestos facilita a aplicação posterior dos parâmetros identificados em modelos virtuais usados em ferramentas ou players que adotam o padrão MPEG-4.

O modelo de humanóide usado no projeto COGEST e no presente trabalho é composto por um conjunto de articulações organizadas de forma hierárquica que representam a parte superior do corpo humano seguindo as especificações H-Anim MPEG-4 [32]. O modelo possui 8 articulações e 23 graus de liberdade, como apresentado na Figura 1-12. Para animar partes do corpo, são efetuadas rotações nos ângulos dos eixos articulares que definem os graus de liberdade do movimento. A articulação *HumanoidRoot* é a raiz de toda a hierarquia e divide o corpo em duas partes: superior e inferior. Os movimentos articulares são apresentados de maneira detalhada no APÊNDICE A.

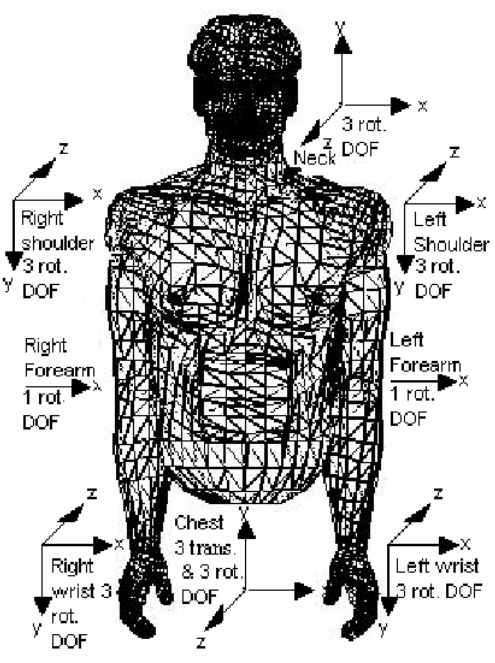

Figura 1-12: modelo 3D da parte superior do corpo humano com 8 articulações e 23 graus de liberdade.

Aquisição e processamento da Imagem de vídeo (A à D) – Nesta etapa, a imagem (ou imagens, no caso de sistemas baseados em múltiplas câmeras) é capturada e processada. A forma de processamento da imagem depende da métrica

de comparação a ser usada. Em geral, procura-se separar informação sobre a posição humana a partir de informação de cores da imagem, destacando as partes do corpo humano. A prática mais comum para esse tipo de sistema é a segmentação da imagem pela cor da pele, como é mostrado na Figura 1-13. A cor da pele humana possui propriedades de crominância particulares, variando pouco entre indivíduos de raças diferentes [33,34]. A segmentação entre regiões de pele e de fundo pode ser obtida através de técnicas simples de limiarização, em que cada *pixel* é associado a uma classe em função de seu valor numérico. Para isso, utilizam-se com freqüência os sistemas colorimétricos HSV e YCrCb, que separam informações de crominância e luminância em diferentes canais. O maior problema desta fase da segmentação é a presença de ruídos na imagem devido a variações de iluminação ou a presença de objetos de cor semelhante a da pele nos cenários de aquisição, o que representa uma limitação para o sistema. Alguma melhora pode ser obtida através de técnicas de extração do fundo fixo de cena ou de reconhecimento de regiões de fundo com treinamentos realizados com redes neurais.

No caso de uso de contornos, estes podem ser obtidos com facilidade a partir do uso de filtros sobre as imagens para destacar as bordas do modelo e da imagem do ator, neste último caso, após a segmentação, evita-se capturar objetos de fundo de cena.

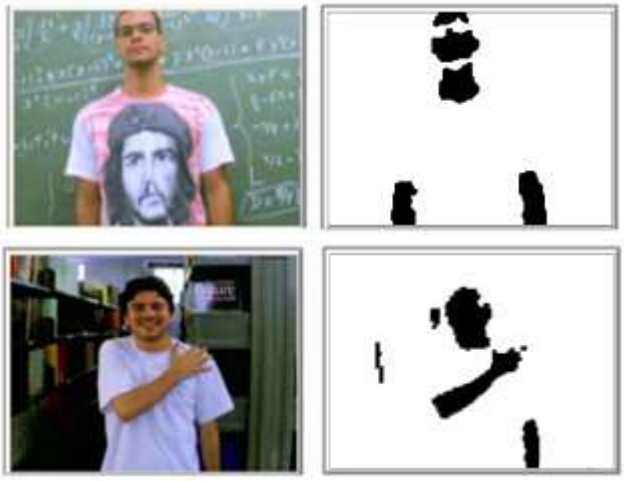

Figura 1-13: imagens segmentadas em cor de pele e fundo.

Estimação da pose e projeção do modelo (G, H, I e J) – Para cada imagem extraída e processada do vídeo, a pose correspondente deve ser identificada. Em

técnicas baseadas no uso de modelos, projeta-se o modelo no plano imagem de acordo com um mecanismo de predição escolhido, ou baseado em informações registradas de posições anteriores, buscando-se a correspondência entre as imagens de vídeo processada e a imagem sintetizada.

Para cada postura candidata do modelo 3D, é calculada a função de custo de acordo com a métrica de similaridade. O presente trabalho é focado, de maneira especial, na implementação de técnicas e de métricas que possam expressar, de forma qualitativa, a similaridade entre a imagem gerada com a projeção do modelo e a imagem de vídeo capturada pela câmera e segmentada. Um exemplo desse processamento é ilustrado na Figura 1-14, em que é usada uma métrica de similaridade chamada de minimização da superfície de não-recobrimento.

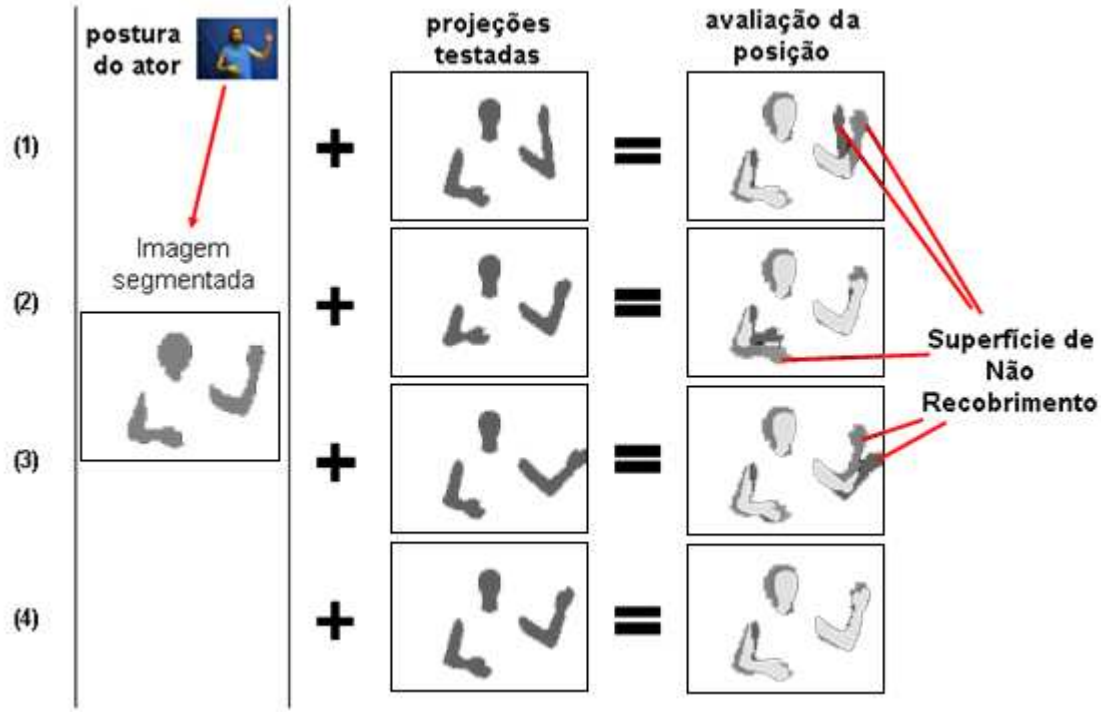

Figura 1-14: superposição das imagens segmentada e projetada para análise de similaridade.

Minimização da diferença e métrica de similaridade (L) – Para encontrar a posição aproximada, utilizam-se algoritmos de otimização iterativos que procuram minimizar a diferença entre as duas imagens. Algumas heurísticas podem ser utilizadas para a busca da postura aproximada. Exemplos de algoritmos de otimização multidimensional usados com este objetivo são o algoritmo de *Powel* [35]

e o *Downhill Simplex* [35–37], sendo este último adotado no presente trabalho. As iterações realizam perturbações nos parâmetros do modelo (translação do corpo e rotação nos graus de liberdade das articulações), partindo de informações de posturas anteriores, devendo respeitar limitações biomecânicas estáticas ou dinâmicas. As restrições estáticas definem os valores mínimos e máximos configurados de maneira fixa para cada grau de liberdade. As restrições dinâmicas estabelecem os valores mínimos e máximos em função da posição de outros graus de liberdade ou outras articulações. Por exemplo, o movimento de rotação do braço pode ser limitado diferentemente em função da flexão aplicada no cotovelo, visto que o braço poderia, para algumas posições do cotovelo, atravessar o tronco.

Como critério de parada para o algoritmo iterativo de otimização, são usualmente adotados valores de tolerância mínimos entre iterações sucessivas, ajustados em função do cenário da captura, do tamanho da imagem utilizada, do tipo de métrica utilizada, entre outros fatores. Para aquisição em tempo real, é necessário também limitar o número de iterações.

É importante observar que tais métodos heurísticos não garantem um resultado preciso e que, devido à natureza complexa da identificação de postura de objetos com diversos graus de liberdade, é grande a incidência de mínimos locais. Alguns mecanismos convencionais devem ser inseridos no algoritmo para reduzir a incidência de mínimos locais. No caso do *Downhill Simplex*, por exemplo, sugere-se a re-iniciação do processo de otimização uma ou duas vezes, visando perturbar o resultado encontrado como pretensa solução, de maneira a escapar de possíveis mínimos locais. Uma discussão sobre a incidência de mínimos locais no ambiente utilizado pelo projeto COGEST é feita em [28].

Extração, codificação e aplicação dos parâmetros identificados (N) – Uma vez que o algoritmo tenha localizado a postura aproximada através da técnica descrita, os parâmetros usados para definir a postura do modelo geométrico podem ser extraídos e codificados. Neste caso, é importante recorrer aos padrões, como o MPEG-4 BAP (*Body Animation Parameters*) [31], visando a restituição em modelos humanóides compatíveis. Os parâmetros podem ser gravados em arquivos para restituição off-line ou transmitidos através de algum canal de comunicação ou rede para animar atores virtuais em ambientes sintéticos.

1.4 Ambiente virtual colaborativo

Uma aplicação simples, originada no trabalho de Soares [1], é um ambiente virtual colaborativo que valoriza a comunicação gestual [38]. Este ambiente permite o compartilhamento à distância de aplicações de interface bidimensionais imersas em um mundo virtual tridimensional, como apresentado na Figura 1-15. Ações dos usuários sobre o objeto compartilhado são representadas por objetos humanóides no ambiente virtual (o detalhamento desta infra-estrutura pode ser encontrado em [10]).

Figura 1-15: interface do ambiente virtual colaborativo.

Nesse ambiente virtual, as animações geradas pelo sistema de aquisição de gestos por visão monocular podem ser usadas para animar os humanóides virtuais, seja através da re-execução dos parâmetros BAP armazenados em disco, ou ainda

em tempo real, neste caso sendo os parâmetros identificados enviados através da rede imediatamente após a aquisição.

1.5 Objetivos do trabalho

O objetivo geral deste trabalho é o aprimoramento das técnicas e métricas do projeto COGEST para identificação de posturas humanas em tempo real, visando a qualidade da aquisição, isto é, a diminuição dos erros entre a postura real e a postura identificada pelo sistema.

Como objetivos específicos, são tratados os seguintes pontos:

- Melhoria do desempenho do sistema, visando possibilitar um maior número de iterações ao longo do processo de otimização e, conseqüentemente, melhor convergência em direção à postura do ator em tempo real. Nesse sentido, estudos foram realizados relativamente ao processamento gráfico do modelo e à transferência de dados entre a placa gráfica e a memória do computador, para comparação de imagens, uma operação de custo elevado;
- Diminuição do espaço de busca e melhoria nos aspectos relacionados à predição da postura, usando restrições biomecânicas dinâmicas e tratando aspectos relativos a ambigüidades registradas nas métricas de similaridade;
- Melhoria da maneira como são feitas as perturbações dos graus de liberdade que inicializam o processo de otimização, visando uma convergência mais rápida;
- Memorização de posturas através do uso de técnicas de banco de dados;
- Estudo da eficiência, usando variações nas métricas de similaridade;
- Combinação de técnicas e de métricas para identificação das melhores soluções para a identificação de postura em tempo real.

Dessa maneira, registram-se contribuições em todas as fases do sistema de aquisição de gestos do projeto COGEST.

1.6 Trabalhos Relacionados

Diversos trabalhos de pesquisa vêm sendo desenvolvidos no contexto da análise de imagens para a extração de características e de posturas humanas. Alguns desses trabalhos são brevemente discutidos a seguir.

Bergh, Koller-Meier & Gool [16] extraem a silhueta de seis câmeras processada por seis computadores, separam as características das imagens, usando *wavelet Haar*, e criam um conjunto de 50 posturas únicas obtidas por um treinamento de 6000 amostras.

Gupta, Mittal & Davis [39], através da visão estereoscópica, extraem a silhueta e, utilizando a cinemática dos membros, identificam a postura mesmo, quando há oclusões na imagem.

Anam-Dong & Seongbuk-gu [40] desenvolveram um método para interpretação visual de gestos que pode ser utilizado na interação natural robô-homem (HRI). Usando duas câmeras (visão do robô), o sistema aprende e extrai os gestos de todo o corpo de um ator, incluindo sua movimentação espacial, através do Modelo Oculto de Markov (HMM).

Um sistema rápido de detecção de poses, usando múltiplas câmeras em um ambiente 3D flexível, foi desenvolvido por Mittal, Zhao & Davis [2]. Estes autores utilizam um trabalho de decomposição de silhuetas para procurar a melhor posição, eles utilizam uma função de probabilidade que integra as informações das visões. A oclusão é tratada de forma que a função utiliza as informações somente das partes que são visíveis em cada câmera.

Watanabe & Yachida [9] propõem um método baseado em um formato de imagem próprio, criado a partir de duas câmeras, permitindo a identificação de posturas mais complexas, como a oclusão de membros ou o deslocamento em profundidade de partes do corpo.

O trabalho de Zou *et al.* [17] necessita de uma calibragem no início do processo onde são localizadas manualmente as articulações do ator, de forma a gerar um modelo humano esquelético. A partir disto, o ator é localizado nas imagens pelo reconhecimento de forma do modelo esquelético, e as posturas são calculadas através de cinemática inversa.

Poppe [8] estima poses a partir das silhuetas de uma seqüência de imagens, localizando as extremidades dos membros e utilizando a cinemática inversa para calcular a postura completa, tendo como restrições a necessidade de que o ator esteja de frente e por completo na visão de uma única câmera.

J. Zhao, Li & Keong [41] extraem pontos, com ou sem marcadores de imagens 2D monoculares, usando filtros espaciais, predição linear e inter-correlações para recuperar movimentos humanos. Um esqueleto 3D é adaptado com as características angulares. Uma função de energia representa o resíduo da diferença entre os pontos do modelo e dos extraídos da imagem.

Agarwal & Triggs [7] realizam a recuperação da postura usando regressão direta, através de uma grade de densidade do gradiente do histograma, seguida por uma matriz para aprender posições básicas de postura humana. O sistema é treinado com um banco de imagens em posturas pré-definidas, tendo como vantagem trabalhar na presença de fundos naturais, complexos, sem a necessidade de segmentações.

Weik & Liedtke [3] desenvolveram um sistema totalmente automático para extração de movimentos 3D. Necessita da calibração do fundo monocromático do ambiente para extração da silhueta das seis câmeras. Usando um esqueleto dinâmico, são estimadas posições arbitrárias e, através de informações de volume capturadas, um algoritmo ICP (*Iterative Closest Point*) detecta a postura de forma hierárquica pela cinemática do corpo humano.

Utilizando um modelo humano de três níveis Zhang *et al.* [4], localizam o corpo humano em visões arbitrárias. A saída é um mapa proposto com probabilidades de configurações do corpo. Utilizando uma mistura de visão monocular com SMC (*Sequential Monte Carlo*), torna-se possível tratar casos de oclusões e mudanças no ângulo de visão.

Shakhnarovich, Viola & Darrel [5] utilizam um método de busca através de uma *hashtable* modificada, de forma a encontrar soluções nas vizinhanças. É treinada com um conjunto de aproximadamente 150 mil posições obtidas pelo programa comercial Poser® com 18 graus de liberdade, incluindo a rotação do corpo. Utiliza a silhueta como pré-processamento da imagem.

Vilariño & Rekeczky [42] mostram que interferências na imagem, nos objetos ou na iluminação da cena produzem situações difíceis de modelar como, por exemplo, contornos descontínuos e ocultos, necessitando conectar tais segmentos e eliminar aqueles correspondentes a ruído.

Zhiwei *et al.* [43] comentam que a eficácia do reconhecimento da imagem reside na habilidade do aprendizado adaptativo, propondo a utilização de métodos gerais de aprendizado, os quais podem simplificar a estrutura do sistema de reconhecimento, simulando o cérebro humano, que pensa mais eficazmente.

Pela contextualização do problema tratado neste trabalho, esses trabalhos correlatos foram analisados, buscando-se inspiração para as soluções e melhorias propostas para o sistema de aquisição do projeto COGEST.

1.7 Organização da dissertação

Esta dissertação está dividida em seis capítulos. No segundo capítulo, são apresentadas as técnicas básicas desenvolvidas visando melhorar a qualidade e o desempenho do sistema de aquisição do projeto COGEST. No terceiro capítulo, são tratadas as métricas de similaridade, para comparar as imagens, segmentada e do modelo. No quarto capítulo, são apresentados e discutidos os resultados, comparando-se as métricas e as técnicas propostas neste trabalho, através da análise do resultado de diversas simulações, a fim de determinar as vantagens e as desvantagens de cada métrica e técnica desenvolvida. A conclusão é feita no quinto capítulo, apresentando-se sugestões para a continuidade e melhoria das técnicas desenvolvidas.

Capítulo 2

2 Técnicas de Apoio ao Sistema de Identificação de Posturas Humanas por Visão Computacional

No contexto do projeto COGEST, algumas contribuições foram desenvolvidas, com o objetivo de tratar as limitações e expandir as funcionalidades do protótipo de Soares [1], algumas do ponto de vista da interface, outras relacionadas à qualidade da análise das imagens e identificação das posturas humanas.

As seções a seguir apresentam as contribuições desta dissertação no contexto do projeto COGEST.

2.1 Plataforma de desenvolvimento

Obedecendo as características do projeto COGEST, os desenvolvimentos realizados para esta dissertação utilizam *software* livre ou código aberto e multiplataforma (programas que podem ser executados em qualquer sistema operacional). A linguagem de programação é a C++, utilizando o compilador g++. Abaixo, estão listadas as coleções de subprogramas utilizados no desenvolvimento do programa, conhecido na ciência da computação como biblioteca, e suas respectivas funções:

- ***C++***: linguagem de programação;
- ***Gtk+/Glade***: interface gráfica e *framework* de acesso ao sistema operacional;
- ***Intel® OpenCV*** [44]: processamento gráfico de imagem, manipulação de vídeos e captura de imagens de dispositivos de captura como a *webcam*;

- ***OpenGL***: interface de programação para aplicações gráfica 2D e 3D, padronizando a comunicação entre as aplicações e os diversos tipos de placas gráficas;
- ***OpenGL/Cg Shading Language***: linguagem de programação utilizada para o desenvolvimento de algoritmos que são executados dentro da placa gráfica, aproveitando os recursos de hardware da mesma.

2.2 Características do sistema de aquisição de gestos

Como apresentado na seção 1.3, para cada imagem de vídeo cuja postura se deseja identificar, no processo de otimização, diversas comparações são realizadas entre a imagem de vídeo segmentada e imagens de síntese geradas com o modelo 3D em posturas candidatas, a fim de se identificar a que melhor se aproxima.

Uma importante restrição do sistema é que, na primeira imagem de vídeo, o ator e o modelo devem estar posicionados de maneira equivalente. Isto se deve ao fato de que a técnica utiliza parâmetros da posição anterior do modelo para o processamento da postura subseqüente.

Algumas restrições e dificuldades inerentes ao método de aquisição utilizado neste trabalho advêm do fato de se usar uma única câmera, visto que informações tridimensionais são obtidas a partir de imagens bidimensionais. A avaliação de cada uma das poses do modelo tridimensional tem elevado custo de processamento, pois é necessária a sua projeção no plano e a transferência da imagem gerada da placa de vídeo para a memória principal. Isto é feito para que possam ser efetuadas as avaliações de correspondências entre as imagens segmentada e projetada.

Utilizando a técnica original em computadores modernos, podem ser realizadas até 150 avaliações por postura, permitindo que o sistema seja executado em tempo real com desempenho em torno dos quinze quadros de câmera por segundo (fps). O problema da transferência de dados é discutido com maiores detalhes na seção 2.5.

As contribuições do presente trabalho ao de Soares [1] visam dois objetivos: melhorar a qualidade da aquisição (identificar a postura com maior precisão) e

reduzir o tempo de cada avaliação, o que, conseqüentemente, possibilita a realização de mais avaliações na tentativa de melhorar a qualidade da aquisição.

Como o sistema de aquisição de gestos utiliza câmeras de vídeo que capturam imagens e as apresentam invertidas na tela, quando nos reportarmos ao braço esquerdo do modelo, estaremos nos referenciando ao braço que esta à direita na imagem (como se fosse a imagem refletida por um espelho).

2.3 Métricas para avaliação de postura

A principal contribuição deste trabalho é o estudo e a implementação de diferentes métricas de avaliação de similaridade entre a imagem segmentada e a imagem projetada. Tal assunto é extenso, sendo, portanto, o Capitulo 3 inteiramente dedicado à apresentação das métricas.

2.4 Ajuste das dimensões do modelo tridimensional

Um dos pré-requisitos naturais para a boa qualidade na aquisição de gestos com a técnica empregada é que as dimensões do modelo tridimensional correspondam, ao máximo, às do indivíduo (chamado aqui de ator). O protótipo anterior não oferecia uma interface amigável para este ajuste. Um editor tridimensional (3D) interativo foi desenvolvido para este fim e sua interface é apresentada na Figura 2-1. Nesta nova funcionalidade do sistema, um modelo pode ser carregado e suas dimensões ajustadas sobre uma imagem capturada.

Esta interface possibilita a manipulação da escala, rotação e translação de todos os membros do modelo tridimensional. Existe uma total liberdade de manipulação das articulações, como é mostrado na Figura 2-2. Esta manipulação pode ser realizada com o ponteiro do mouse, clicando e arrastando sobre o membro, ou através das barras deslizantes e dos botões que são encontrados à esquerda da figura.

Além da parte que sobrepõe o modelo à imagem do ator, a interface possui uma segunda visão do modelo tridimensional, exibida na parte inferior direita da janela. Esta segunda imagem mostra o modelo na mesma configuração de ajuste,

porém utilizando um ponto de vista independente, o que permite visualizar aspectos de profundidade durante o ajuste das dimensões do modelo.

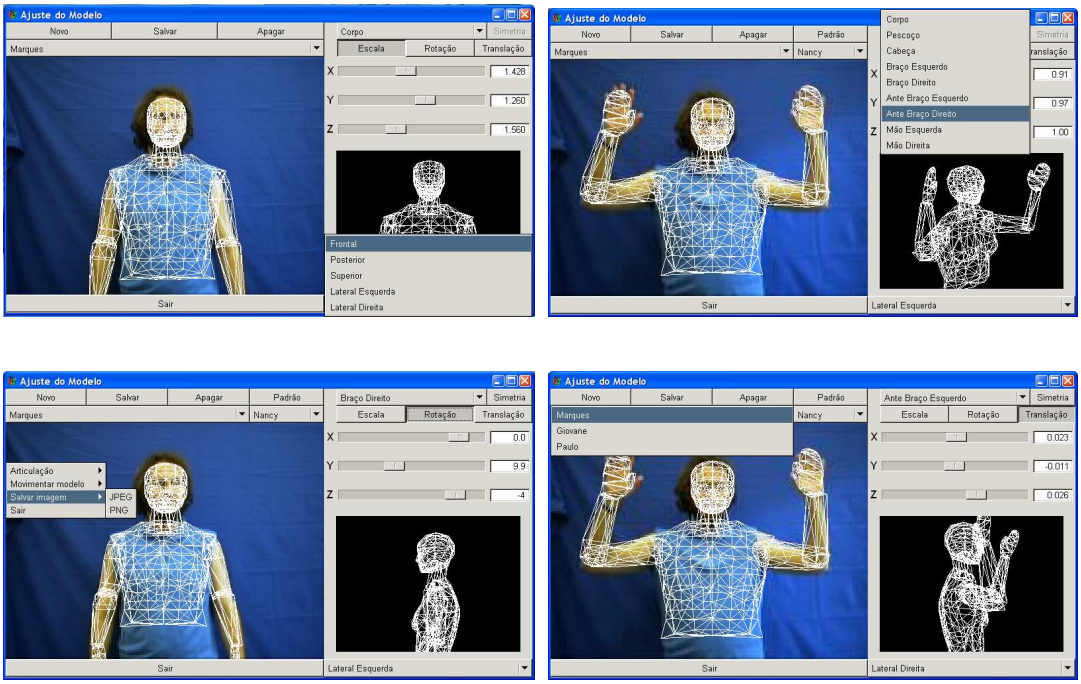

Figura 2-1: interface gráfica para o ajustamento do modelo tridimensional às dimensões do ator.

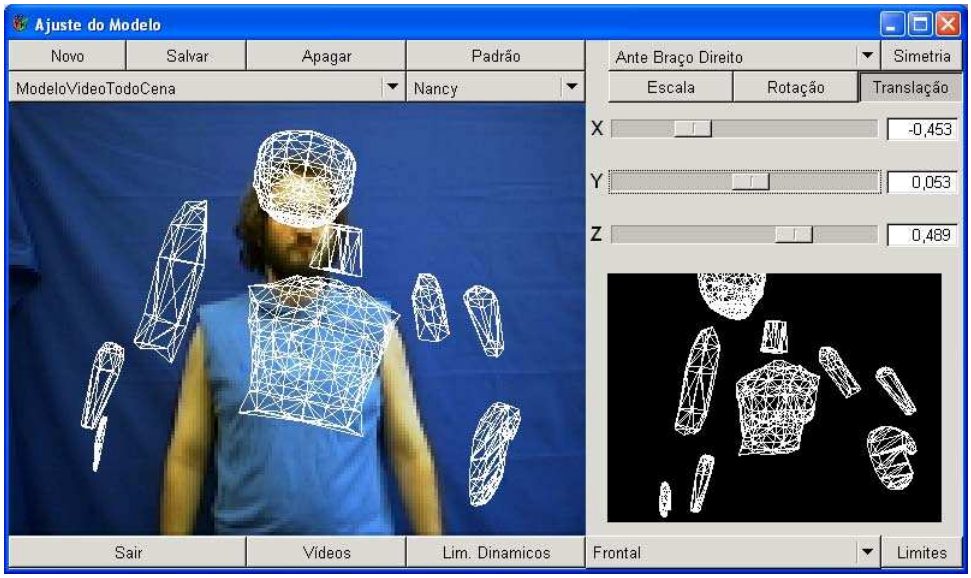

Figura 2-2: liberdade de manipulação das articulações.

Para realizar o dimensionamento do modelo, deve ser ajustada a posição e o tamanho de cada membro, começando pelo de nível hierarquicamente superior, seguindo-se sempre na direção dos membros hierarquicamente inferiores.

Na Figura 2-3 são mostrados, para efeito de demonstração, os passos de como realizar um ajuste do modelo, descritos abaixo:

1. Ajusta-se a posição e o tamanho do tronco do modelo;
2. Ajusta-se a posição e o tamanho da cabeça;
3. Ajusta-se a posição, o tamanho, e o ângulo dos braços;
4. Ajusta-se a posição, o tamanho, e o ângulo dos antebraços;
5. Ajusta-se a posição, o tamanho, e o ângulo das mãos.

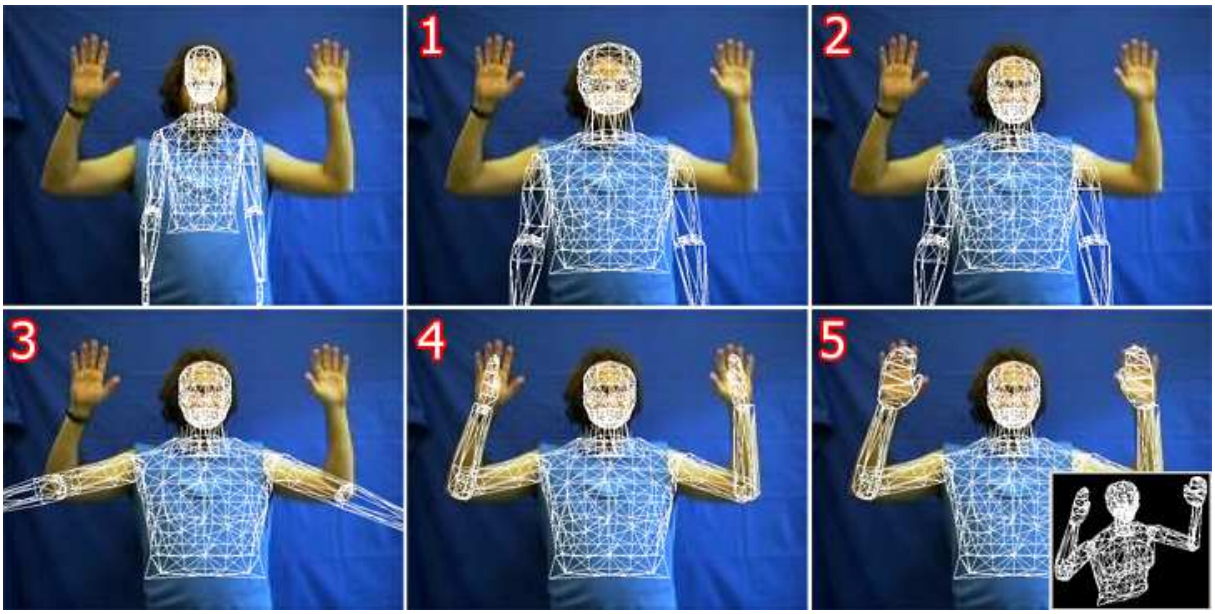

Figura 2-3: demonstração do ajuste do modelo.

A seqüência mostrada na imagem inicia com o modelo sem ajuste, finalizando-se, na imagem 5, com a postura ajustada. A imagem da segunda visão exibe a mesma postura em outra perspectiva.

2.4.1 Análise de posturas

Através desta interface, é possível analisar e ajustar os parâmetros das restrições biomecânicas que são explicadas em detalhes na seção 2.6. As posturas podem ser visualizadas, e as restrições biomecânicas estáticas que definem o limite máximo e mínimo de cada articulação alteradas, como mostrado nas Figuras 2-4 e 2-5.

O exemplo da Figura 2-4 mostra o modelo em três posturas. A diferença entre as posturas está no eixo Z, correspondente à abdução do braço esquerdo. Nas duas primeiras imagens, observa-se que o ângulo do braço foi alterado, mas a última imagem não apresenta diferença para a segunda, devido à ação das restrições biomecânicas que impediram esta postura, por considerá-la, neste caso, ergonomicamente desfavorável.

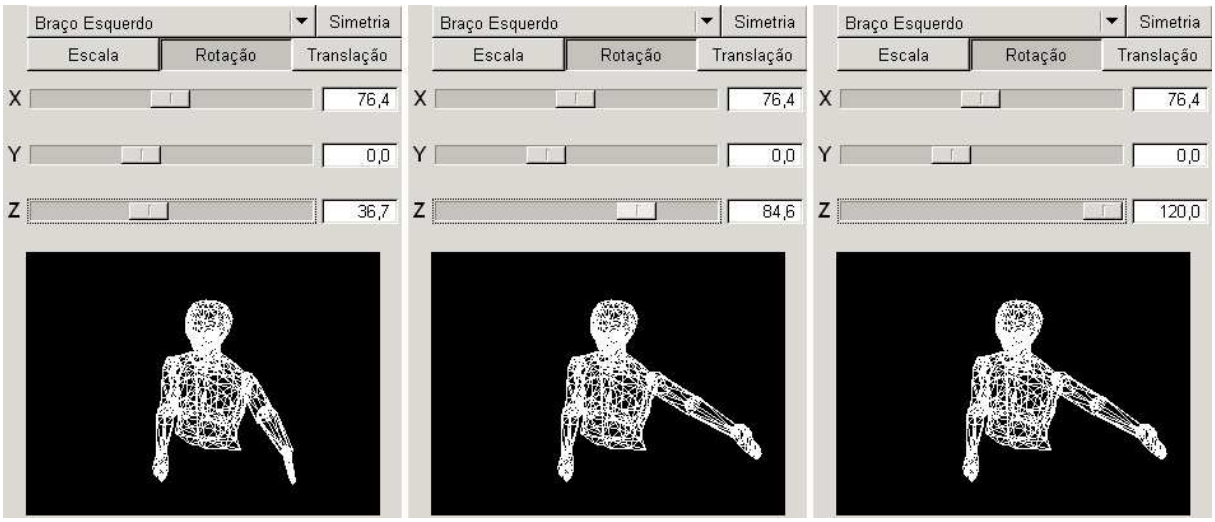

Figura 2-4: postura ergonomicamente desfavorável impedida pela restrição dinâmica da abdução do braço esquerdo.

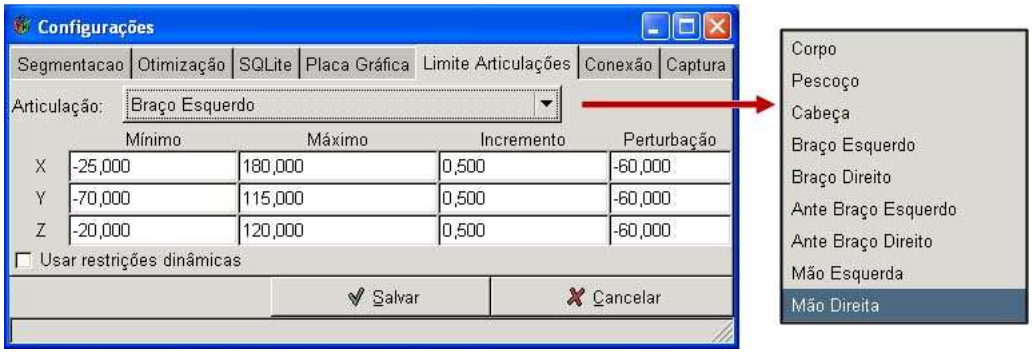

Figura 2-5: definição dos limites estáticos das articulações.

2.4.2 Criação de vídeos sintéticos

Integrada ao ajuste do modelo, existe uma ferramenta de criação de vídeos sintéticos, como apresentado na Figura 2-6. Essa ferramenta foi utilizada para criar os vídeos que são usados para avaliar o resultado das técnicas. Os vídeos gerados para a avaliação das técnicas são discutidos na seção 4.1.

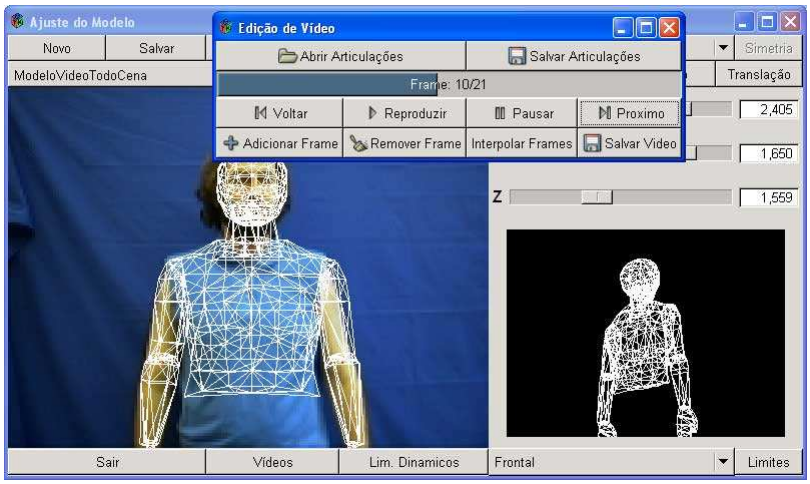

Figura 2-6: geração de vídeos sintéticos usando o ajuste do modelo.

2.5 Estudo das placas gráficas para melhoria do desempenho

A avaliação da postura realiza a transferência da imagem projetada da placa gráfica (*Graphics Processing Unit* - GPU) para a memória principal do computador, pois a imagem projetada é sintetizada dentro da placa gráfica, a partir das informações do modelo tridimensional. No protótipo de Soares [1], essa transferência representa, em média, 50% do tempo gasto para avaliar cada imagem de vídeo, tornando-se o principal limitador de desempenho do sistema.

2.5.1 Alteração do formato de armazenamento da imagem

As imagens utilizadas pelas métricas de similaridade necessitam de, no máximo, 16 cores para representar as diferentes partes dos membros, podendo, portanto, ser resumidas em imagens de 4 bits.

Contudo, a imagem padrão do OpenGL, que é utilizada pelo protótipo original, usa 24 bits de informação para cada *pixel*, 8 bits para cada canal de cor no padrão RGB.

Foram realizados testes na tentativa de transferir somente um canal de cor do RGB. Contudo, isso não reduz o tempo de transferência, pois, ainda que apenas uma cor seja armazenada na memória, são transferidos os três canais, o que representa uma restrição da placa de vídeo.

Experimentações mostraram que a única possibilidade de melhoria na transferência entre a placa gráfica e a memória é através da utilização de uma extensão do OpenGL que permite a criação de *framebuffers* auxiliares (regiões de memória dentro da GPU). Os *framebuffers* aceitam o padrão GL_R3_G3_B2, um espaço de cor que utiliza 3 bits para o canal vermelho e verde, e 2 bits para o canal azul, somando 8 bits para os três canais de cor, reduzindo três vezes a quantidade de bits por *pixel*. Esta extensão não é padrão das placas gráficas, mas é encontrada em quase todas as GPUs da NVidia®.

Mesmo com uma redução de três vezes no tamanho da imagem, o tempo de transferência foi reduzido no máximo em duas vezes e, em média 25%, o que representa um ganho de 10% no tempo total do sistema de aquisição. Os resultados do emprego dessa técnica são discutidos na seção 4.6.1.

2.5.2 Programação da placa gráfica

Existem várias linguagens de programação implementadas na placa de vídeo. Uma experimentação foi realizada utilizando a linguagem *Cg (C for Graphics)*, desenvolvida pela NVidia®, através de uma extensão da *OpenGL Shading Language* [45]. No entanto, todas essas linguagens se concentram em modificar a funcionalidade fixa durante o processamento de cada vértice e de cada *pixel* que passam através dos estágios da *pipeline* gráfica.

Foi desenvolvido um *shader* (programa que roda na GPU) para ler toda a textura (imagem armazenada na GPU), *pixel* a *pixel*, contando-os e armazenando o resultado em outra textura. Apenas esta segunda textura, de tamanho reduzido, ocupando menos de 1% do tamanho da textura original, é transferida para a memória do computador. Porém, o custo de acesso à textura dentro do *shader* se mostrou proibitivo. Por esse motivo, essa técnica foi descartada.

Algum tempo após esse estudo, foi lançada uma nova série de placas gráficas que suportam uma nova linguagem de programação com menos restrições, chamada *CUDA (Compute Unified Device Architecture)* [46]. Estas GPUs têm uma estrutura e hardware totalmente diferentes. Foram realizados testes mostrando ser possível calcular o histograma dentro da GPU mais rápido do que no processador. Por não representar o foco principal dessa dissertação, cujo trabalho foi desdobrado

em outras técnicas, a exploração da tecnologia CUDA é deixada para trabalhos futuros.

2.6 Restrições biomecânicas estáticas e dinâmicas das articulações dos membros

Para diminuir o espaço de busca no processo de otimização, busca-se evitar a avaliação de posturas irreais para o modelo tridimensional. Nesse sentido, é necessário que a estimação de poses leve em consideração as restrições biomecânicas das articulações do corpo humano. No trabalho de Soares [1], eram verificadas apenas as restrições estáticas de cada grau de liberdade (GL), isto é, os limites mínimos e máximos de flexão, abdução e extensão de cada articulação, que eram estipulados sem levar em conta o estado dos demais graus de liberdade. O APÊNDICE A (Movimentos articulares do humanóide virtual) apresenta um estudo sobre os graus de liberdade do modelo adotado e como são realizados os movimentos dos membros.

As restrições biomecânicas estáticas (RBE) definem somente um limite inferior e superior para os GL das articulações, mas não leva em consideração a relação de limitação entre eles. A posição e a orientação de um eixo podem influenciar na amplitude de outro, definindo uma relação entre os GL das articulações envolvidas no movimento. Modelar essa característica anatômica é uma tarefa extremamente complexa [18].

Uma restrição para um grau de liberdade é considerada dinâmica quando seus limites possuem dependência com relação a valores de outros graus de liberdade, o que permite maior redução na ocorrência e na análise de posturas ergonomicamente desconfortáveis, tais como as exemplificadas na Figura 2-7. Tais restrições diminuem o espaço de busca e auxiliam na tarefa de predição de posturas.

Para especificar restrições entre os GL, é necessário definir uma hierarquia entre os mesmos, de forma que seja possível delimitar o movimento dos GL hierarquicamente inferiores de acordo com o estado dos de nível superior.

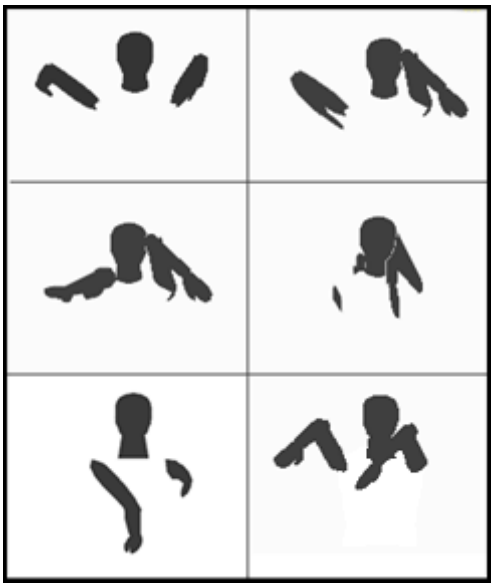

Figura 2-7: posturas ergonomicamente impossíveis realizadas pelo modelo.

A partir da análise do modelo e da observação dos resultados de testes realizados no sistema, convencionou-se que, para a articulação do ombro, a flexão é o GL hierarquicamente superior. Devido à rotação ser o GL que causa maior problema na otimização com relação a mínimos locais, este foi definido como o GL hierarquicamente inferior; ou seja, a rotação dependerá do valor de ambos os outros GL: flexão e abdução. A abdução, por sua vez, fica em um nível hierárquico intermediário, dependendo apenas da flexão.

Resumidamente a rotação do ombro depende tanto da abdução quanto da flexão do ombro, e a abdução depende apenas da flexão do ombro.

2.6.1 Restrições biomecânicas dinâmicas

Foram realizados estudos preliminares que estabeleceram relações dinâmicas entre os GL dos braços (membros utilizados na aquisição de posturas). As relações foram exploradas, inicialmente, a partir de posturas de modelos vivos (bolsistas do laboratório) e modelados por Redes de Petri Coloridas, utilizando formulas complexas que definiam os limites dinâmicos. Entretanto, ocorreram três fatos que levaram ao seu abandono.

Foram feitas algumas alterações no modelo, sendo necessário refazer as restrições biomecânicas dinâmicas (RBD), e como estas eram feitas com fórmulas, sua manutenção se tornou trabalhosa.

O uso de fórmulas, principalmente utilizando a função raiz quadrada, torna o cálculo computacionalmente custoso. O uso de tabelas facilita a manutenção e o aprimoramento das restrições, além de ter um melhor desempenho.

É importante observar que o modelo computacional do corpo utilizado no sistema de aquisição é simplificado, não contendo toda a complexidade inerente ao corpo humano. Isso limita a consideração das restrições biomecânicas estudadas em anatomia humana para o sistema proposto. A utilização de pessoas na análise das posturas não se adequou perfeitamente ao modelo.

Por isso, as restrições são definidas a partir da observação da movimentação do próprio modelo 3D simplificado. Assim, o estudo das RBD é realizado usando a interface de ajuste do modelo exibida na Figura 2-1.

Observa-se que as restrições variam de maneira considerável em intervalos de vinte graus e, assim, definem-se os limites inferiores e superiores da abdução e da rotação, com relação à flexão, para cada intervalo. Na Tabela 2-1 são apresentados tais limites para o ombro direito.

Na Tabela 2-2 são apresentados os limites inferiores e superiores da rotação com relação à flexão para o ombro direito em cada intervalo.

Os limites do ombro esquerdo foram produzidos por simetria do ombro direito. Contudo, é possível definir limites diferenciados em casos especiais como os comentados anteriormente.

A aplicação destas RBD apresentadas impede que o sistema de aquisição de gestos, durante o processo de otimização, realize o cálculo da função custo para posturas ergonomicamente impossíveis, como aquelas apresentadas na Figura 2-7, diminuindo o espaço de busca e, por conseguinte, o custo computacional do processo de otimização.

Observa-se, ainda que, para aplicações que requeiram menor grau de precisão das posturas identificadas, pode-se aumentar o conjunto de restrições, diminuindo o espaço de busca.

Tabela 2-1: relação da interdependência da flexão com a abdução e a rotação para o ombro direito (valores em graus).

Flexão	**Abdução**		**Rotação**	
	Inferior	**Superior**	**Inferior**	**Superior**
160 à 180	-10	10	-70	-30
140 à 160	-20	15	-70	-30
120 à 140	-30	20	-90	-10
100 à 120	-40	20	-100	5
80 à 100	-50	20	-110	5
60 à 80	-60	20	-115	5
40 à 60	-70	15	-115	10
20 à 40	-80	10	-115	30
0 à 20	-90	0	-90	30
-20 à 0	-110	0	-80	50
-40 à -20	-120	10	-110	70

Tabela 2-2: relação da interdependência da abdução com a rotação para o ombro direito (valores em graus).

Abdução	**Rotação**	
	Inferior	**Superior**
0 à 20	-30	20
-20 à 0	-50	40
-40 à -20	-70	50
-60 à -40	-90	70
-80 à -60	-100	70
-100 à -80	-90	70
-120 à -100	-80	70
-140 à -120	-70	70

Uma pesquisa bastante promissora, a qual fica para trabalhos futuros, é o estudo dos aspectos ergonômicos da postura humana. Uma vez que as restrições biomecânicas dinâmicas foram mapeadas, as posturas são classificadas em

diversos níveis de dificuldade, priorizando as posturas ergonomicamente mais fáceis. Este recurso, além de auxiliar no processo de estimação da posição do modelo 3D, pode ser utilizado como critério adicional de classificação na avaliação da postura durante o processo de otimização, podendo ser útil para eliminar ambigüidades (vide seção 2.11).

2.7 Penalização de posturas irreais

O processo de identificação de posturas deste trabalho utiliza o algoritmo de otimização iterativo multidimensional *DownHill Simplex* (DHS) de Nelder-Mead, implementado a partir do trabalho de Press et al. [35]. Este algoritmo é apresentado detalhadamente no APÊNDICE B.

O *DownHill Simplex* foi modificado para comportar as restrições impostas pelo modelo, conforme visto na seção anterior (restrições dinâmicas e estáticas).

Durante a execução do algoritmo de otimização, quando uma postura é impedida por uma restrição biomecânica, redefinem-se os valores dos graus de liberdade para o limite que foi ultrapassado (mínimo ou máximo). O DHS, devido a sua heurística de busca, durante um processo de otimização, por diversas vezes ultrapassa tais limites, principalmente quando a postura desejada se encontra próxima a algum limite. Isso faz com que o processo de otimização avalie posturas semelhantes diversas vezes.

Observa-se, entretanto, que tais limites são raramente alcançados em movimentos humanos reais, pois os gestos naturais realizados por pessoas, ao se comunicarem, geralmente, não se aproximam dos limites físicos.

Para evitar a avaliação da mesma postura diversas vezes, e impedir uma convergência forçada pelas restrições biomecânicas, penaliza-se de forma aleatória o resultado da função de custo. Além disso, como a verificação da similaridade entre as imagens segmentada e de síntese, criada pela projeção do modelo numa configuração definida pelo algoritmo de otimização, é um processo muito caro para este sistema, simula-se um valor alto como resultado da função de custo, o que penaliza a postura irreal sem a necessidade de análise das imagens.

Este valor atribuído ao custo da postura é um número aleatório entre 0,9 e 1,0. Como os valores obtidos das avaliações encontram-se entre 0,0 e 1,0, a penalização utiliza a faixa dos valores 10% piores. O uso de números aleatórios inibe a convergência do DHS, evitando tais mínimos locais.

Em especial, quando a postura do ator está próxima aos limites definidos para as restrições biomecânicas, o processo de iniciação do algoritmo do DHS realiza avaliações que ultrapassam os limites das restrições. Quando as perturbações do *simplex* definem valores fora dos limites para os graus de liberdade, estes recebem o próprio valor limite. Este fato incorre na avaliação de diversas posturas iguais, o que acarreta uma convergência do DHS para um mínimo local. A aplicação da penalização impediu que este tipo de convergência fosse induzido pelos limites biomecânicos.

2.8 Iniciação do *simplex* com parâmetros aleatórios

Como o *Downhill Simplex* (DHS) é determinístico, para a mesma seqüência, considerando os mesmos valores de entrada, sempre serão obtidos os mesmos parâmetros de resultado no processo de identificação da postura.

Quando um mínimo local ocorre para alguma imagem da seqüência, o algoritmo é levado a propagar este erro às próximas imagens da seqüência. Isso ocorre devido aos parâmetros de entrada para cada processo de otimização, usando o *Downhill Simplex*, serem aqueles encontrados como resultado da análise da imagem anterior. Essa maneira de iniciar o algoritmo de otimização é assumida porque, em um sistema de identificação em tempo real, presume-se que a nova postura seja próxima à anterior. Assim, caso os parâmetros identificados na otimização anterior não apresentem uma boa correspondência, a probabilidade de incorrer em mal resultado na otimização seguinte é elevada.

Esta iniciação direcionada é necessária, pois, devido à multiplicidade de posturas possíveis, é impossível avaliar todas em tempo real e, caso a iniciação realize uma grande abertura no espaço de busca, a convergência do DHS é prejudicada.

2.8.1 Iniciação do *Downhill Simplex* usada por Soares [1]

A iniciação dos parâmetros utilizada no trabalho de Soares [1] também é determinística. Assim, de forma a diminuir a propagação de identificações provocadas por uma iniciação imprópria, devido a uma má correspondência do processo anterior, desenvolveu-se uma técnica de perturbação aleatória nos parâmetros de iniciação do *simplex*.

O processo de identificação da postura é realizado em três etapas, correspondendo a três execuções do algoritmo de otimização. Tal técnica, proposta por Soares [1], visa minimizar a ocorrência de mínimos locais, pois caso o DHS venha a convergir para um mínimo local, a iniciação das outras etapas podem fazê-lo fugir do mínimo local.

As etapas da identificação de postura possuem diferentes limites máximos para o número de iterações em cada uma das três etapas de otimizações: 20 na primeira, 30 na segunda e 100 na última. Esta configuração é definida de forma que a primeira etapa faça uma busca local, na tentativa de achar a postura, caso esta esteja próxima da de referência. A segunda tenta fugir de algum possível mínimo local encontrado na etapa anterior. Por esse motivo, estas duas etapas têm um número menor de iterações que a terceira, que objetiva realizar a convergência para uma postura mais precisa.

A iniciação é realizada perturbando os parâmetros encontrados no processo de identificação prévio. O *simplex* é representado por uma matriz com "*n+1*" linhas e "*n*" colunas, onde "*n*" representa o número de graus de liberdade do modelo. Cada linha representa um vértice do *simplex* [47]. O primeiro vértice do *simplex* possui a configuração da postura prévia, localizada no processamento da imagem anterior ou a configuração de repouso, caso seja a primeira imagem. Os demais vértices, representados a partir da segunda linha, formam uma matriz quadrada, cujos elementos da diagonal sofrem uma perturbação no processo de iniciação, que é chamada aqui de "iniciação diagonal". Na Tabela 2-3 é mostrado um exemplo da iniciação diagonal de um *simplex* de dimensão 5 em que:

- cada elemento p_i representa um parâmetro recuperado do processamento da imagem anterior;

- cada elemento x_i representa um valor prefixado de perturbação para cada grau de liberdade das articulações do modelo. Estes valores foram estimados através da análise do comportamento do sistema de aquisição, considerando um limite médio de variação para cada grau de liberdade entre duas imagens, considerando movimentos moderados (sem a execução de gestos abruptos). A perturbação é, em princípio, executada no sentido positivo. Caso o limite biomecânico seja alcançado, o sentido da perturbação é invertido.

Tabela 2-3: *simplex* da iniciação diagonal.

Parâmetros do *Simplex*				
p_1	p_2	p_3	p_4	p_5
$p_1 \pm x_1$	p_2	p_3	p_4	p_5
p_1	$p_2 \pm x_2$	p_3	p_4	p_5
p_1	p_2	$p_3 \pm x_3$	p_4	p_5
p_1	p_2	p_3	$p_4 \pm x_4$	p_5
p_1	p_2	p_3	p_4	$p_5 \pm x_5$

2.8.2 Iniciação aleatória (IAL)

Para melhor tratar a incidência de mínimos locais que ocorrem em função da propagação de erros herdados do processamento de imagens anteriores, e aumentar o espaço de busca na tentativa de fugir de mínimos locais, introduziu-se uma iniciação aleatória ao processo de iniciação do DHS.

Inicialmente, verificou-se o comportamento do sistema iniciando todas as três etapas de otimização com iniciação aleatória. Essa medida, entretanto, mostrou-se ineficaz devido à possibilidade de um grande afastamento da postura de referência, comprometendo a convergência para a posição ideal.

Em seguida, utilizando a iniciação aleatória nas duas primeiras etapas, a convergência da postura foi recuperada. Verificou-se, entretanto, o aumento da

ocorrência de ambigüidades (ver seção 2.11), pois a primeira etapa, responsável em convergir para posturas próximas, foi prejudicada pela abertura do espaço de busca da IAL, que pode levar o DHS a convergir para uma postura ambígua.

Por fim, utilizou-se a iniciação aleatória somente na segunda etapa, que é a etapa para expandir o espaço de busca na tentativa de fugir dos mínimos locais.

Para melhorar a distribuição no espaço de busca, evitando que todos os vértices do *simplex* fiquem distantes da referência, a intensidade da perturbação é realizada em níveis crescentes. Na Tabela 2-4 é mostrado um exemplo da iniciação aleatória de um *simplex* de dimensão cinco, onde:

- p_i é o conjunto dos parâmetros de referência;
- x_i é um valor predefinido de perturbação, pertencente a cada grau de liberdade das articulações do modelo, o mesmo utilizado na iniciação diagonal;
- a_i é a intensidade da perturbação realizada, calculada conforme a Equação ((2-1), que é igual a um valor aleatório entre 0 (zero) e 1,2 multiplicado pelo índice normalizado do *simplex*.

$$a_i = rand\,[0,1] * \frac{i}{\max(i)} * 1{,}2 \qquad (2\text{-}1)$$

Tabela 2-4: *simplex* da iniciação aleatória.

Parâmetros do *Simplex*				
p_1	p_2	p_3	p_4	p_5
$p_1 + (x_1 \times a_1)$	$p_2 + (x_2 \times a_1)$	$p_1 + (x_3 \times a_1)$	$p_1 + (x_4 \times a_1)$	$p_1 + (x_5 \times a_1)$
$p_1 + (x_1 \times a_2)$	$p_2 + (x_2 \times a_2)$	$p_1 + (x_3 \times a_2)$	$p_1 + (x_4 \times a_2)$	$p_1 + (x_5 \times a_2)$
$p_1 + (x_1 \times a_3)$	$p_2 + (x_2 \times a_3)$	$p_1 + (x_3 \times a_3)$	$p_1 + (x_4 \times a_3)$	$p_1 + (x_5 \times a_3)$
$p_1 + (x_1 \times a_4)$	$p_2 + (x_2 \times a_4)$	$p_1 + (x_3 \times a_4)$	$p_1 + (x_4 \times a_4)$	$p_1 + (x_5 \times a_4)$
$p_1 + (x_1 \times a_5)$	$p_2 + (x_2 \times a_5)$	$p_1 + (x_3 \times a_5)$	$p_1 + (x_4 \times a_5)$	$p_1 + (x_5 \times a_5)$

A aleatoriedade inserida na iniciação permitiu ao DHS fugir de mínimos locais que são bastante recorrentes na movimentação de um ser humano, o que seria

impossível em uma abordagem determinística, isso sem prejudicar a convergência e a limitação proposital do espaço de busca no início da otimização.

2.8.3 Outras formas de iniciação do *simplex*

Outras variações de iniciação foram testadas. Contudo, visando o tempo real, não se pode permitir que o algoritmo passe de 150 iterações.

Reduzindo as execuções do processo de otimização para uma etapa (com 150 iterações) ou duas etapas (com 50 e 100 iterações), observou-se maior incidência de mínimos locais, prejudicando a qualidade da identificação de posturas.

Por outro lado, aumentando o número de etapas da otimização com um número menor de iterações, em torno de 40 e 25 iterações, observou-se menor incidência de mínimos locais. Esta abordagem resulta em uma aquisição de qualidade inferior, visto que não se chega a uma convergência mais acurada em nenhuma das etapas. Nestas configurações, foram testadas iniciações com perturbações incrementais dos parâmetros do *simplex*, visando à melhoria da convergência. Entretanto, nenhum dos referidos testes trouxe bons resultados, optando-se pela configuração original de três etapas.

2.9 Classificação das regiões de pele da imagem capturada para identificação dos membros correspondentes

A segmentação da imagem do ator identifica apenas as regiões de pele, não sendo possível diferenciar na imagem: os braços e a cabeça do ator, pois todos pertencem à mesma classe, “cor referente a região de pele”, o que limita a avaliação da similaridade das imagens.

Como na segmentação não há distinção entre os braços, as métricas podem avaliar como um bom resultado, por exemplo, uma imagem em que um braço ocupe a região do outro.

Por isso, a identificação independente dos membros do ator permite a diminuição da ambigüidade (ver seção 2.11), uma vez que, após a segmentação pela cor de pele, podem ser identificadas e coloridas diferentemente as regiões

correspondentes à cabeça e aos braços do ator. Torna-se possível penalizar a avaliação de posturas em que um membro está ocupando a área de outro. Verifica-se também a redução do número de iterações no processo de otimização, obtendo-se a convergência mais rapidamente.

Para realizar este processo, primeiramente, é extraído o contorno da imagem segmentada e, a partir das formas geométricas encontradas, são identificados os membros. Na Figura 2-8 é mostrado como o processo é realizado, e o detalhamento de seu funcionamento é descrito a seguir.

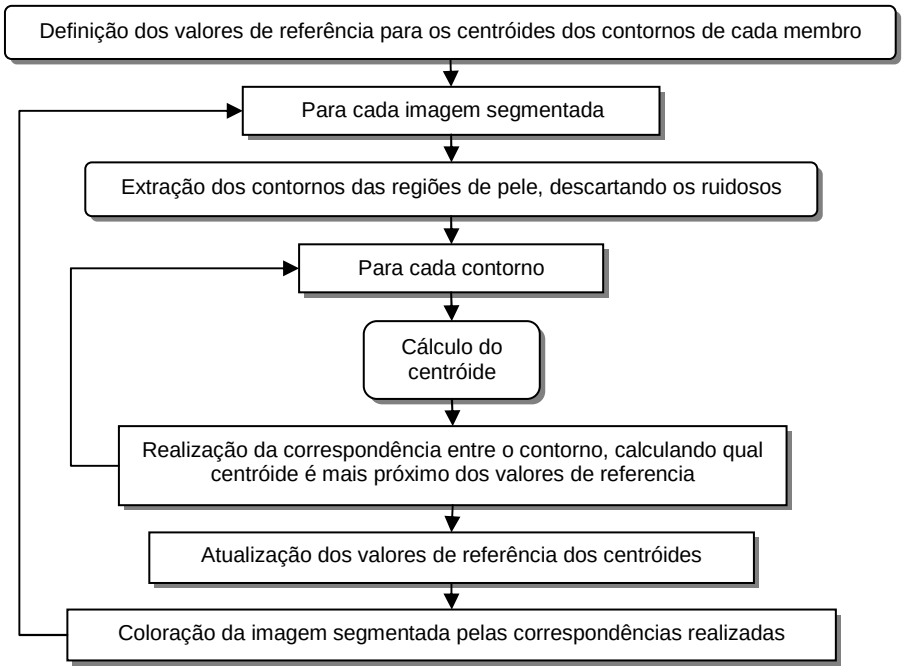

Figura 2-8: processo de classificação das regiões de pele da imagem segmentada através da coloração independente dos membros correspondentes.

Os pontos de referências padrões são definidos de acordo com os centróides do contorno da postura inicial do modelo calibrada no ajuste do modelo pelas dimensões físicas do ator.

A cada imagem segmentada é calculado o contorno desta, utilizando a função de extração de contornos disponível na biblioteca OpenCV [44], baseado no trabalho de Suzuki *et al.* [48]. São descartados os pequenos contornos, com menos de 40 pontos, considerados ruídos. Este número foi encontrado empiricamente observando vários contornos calculados pelo sistema. Os três maiores contornos são separados, e são calculados seus centróides.

A cabeça é identificada pela sua localização, pois a variação de seu centróide é pequena. A cada nova imagem segmentada, as regiões referentes aos braços são localizadas pela proximidade de seus respectivos centróides, relativamente aos centróides de referência. Para melhorar a precisão da referência para a próxima imagem de vídeo a ser processada, os centróides calculados para a imagem corrente, tornam-se os centróides de referência para a imagem seguinte.

E, ao final, é colorida a imagem segmentada com diferentes cores para cada membro, de acordo com a correspondência realizada.

Na Figura 2-9 é mostrado um exemplo do resultado de uma classificação. Os pontos amarelos são os valores de referência dos centróides. Os pontos em azul, verde e vermelho representam o centróide dos contornos das áreas, após serem identificadas, respectivamente, como braço direito, braço esquerdo e cabeça.

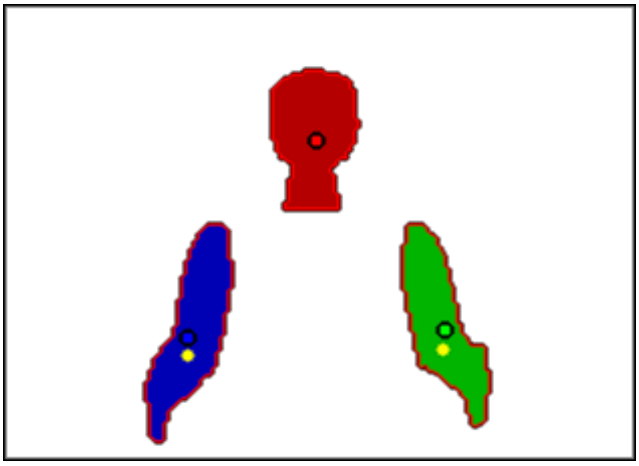

Figura 2-9: classificação das regiões de pele da imagem segmentada através da coloração, independente dos membros correspondentes.

2.10 Predição de posturas através de Banco de Dados

Para evitar o reprocessamento de posturas idênticas, é possível armazenar dados e utilizar ferramentas de busca, o que evita um novo processo de otimização completo.

Tal solução, porém, traz restrições e possíveis problemas. A primeira restrição é que a ferramenta utilizada para armazenamento e busca das posturas seja mais eficiente temporalmente do que o processo de identificação atual, realizado através

de avaliações de similaridade entre imagens. Outra restrição vem da necessidade de treinamento do sistema para o ambiente em questão, visto que as posturas devem ser previamente armazenadas.

O principal problema dessa abordagem é a possibilidade de armazenamento de posturas mal identificadas, devido a convergências de mínimos locais. Caso isso ocorra, o sistema irá propagar o erro em imagens semelhantes, não tentando melhorar o resultado da postura.

Para implementação desta técnica, foi realizada uma pesquisa sobre métodos de busca, tendo como foco principal o desempenho no acesso aos dados armazenados, que se demonstrou ser uma forte restrição em alguns casos, como encontrado em vários bancos de dados comerciais.

As melhores soluções são aquelas que permitem o armazenamento em memória dos dados sobre as posturas, pois qualquer acesso a disco tornaria inviável a solução. Neste caso, existem soluções manuais como matrizes em memória. Foi estudada a viabilidade de uso de tabelas de dispersão (*HashTables*), reconhecidamente eficiente em termos de desempenho na busca. Entretanto, a definição da função *hash* para indexação das chaves, principalmente devido à multidimensionalidade do sistema, torna a inclusão de registro demorada. Experimentou-se, ainda, uma implementação de *hashmap*, desenvolvida pelo Google®, chamada de “google-sparsehash”, que é mais eficiente do que a *HashTable* padrão. Entretanto, esta apresenta o mesmo problema de indexação multidimensional. Os fatos relatados levaram ao descarte do uso de tabelas de dispersão.

A outra abordagem verificada foi a utilização de banco de dados em memória. Alguns sistemas gerenciadores de bancos de dados (SGBD) comerciais possuem suporte a tabelas em memória, e existem também bancos de dados de código fonte aberto, como o MySQL e o SQLite, que suportam este tipo de armazenamento. Entretanto, os SGBD, no estilo cliente-servidor, como o MySQL, são mais lentos, devido ao *overhead* de comunicação.

Por outro lado, o SQLite, que é desenvolvido visando desempenho, mostrou-se mais favorável para o propósito deste trabalho. O código fonte do SQLite foi

compilado com o programa de aquisição de gestos para um ganho maior de desempenho.

Devido à infinidade de posturas que se pode avaliar, bem como à possibilidade de armazenamento de resultados ambíguos ou errados, decidiu-se utilizar essa abordagem não como uma métrica direta para encontrar as posturas correspondentes no banco de dados, mas como um suporte para a predição de posturas. Assim, ao invés de obter diretamente a postura correspondente, utiliza-se o resultado da pesquisa no processo de iniciação do algoritmo *Downhill Simplex*, realizando-se uma predição da postura, com o intuito de facilitar a pesquisa e com a possibilidade de fuga de mínimos locais.

2.10.1 Predição

Para pesquisar a postura no banco de dados, dispõe-se da imagem segmentada, mas como pesquisar uma imagem inteira é computacionalmente custoso, divide-se tal imagem em uma matriz 10x10 regiões, totalizando 100 valores, em que cada célula da matriz representa o número de *pixels* identificados como pele na região correspondente da imagem segmentada. Isso realiza uma aproximação dos dados, visto que não se sabe exatamente quais os *pixels* da região foram segmentados como pele, mas apenas a quantidade de *pixels*.

Como esta aproximação é realizada em regiões localizadas e próximas, isso acaba fazendo com que posturas parecidas sejam armazenadas em conjunto, facilitando a predição de posturas e a fuga de mínimos locais.

As posturas pesquisadas são usadas na iniciação do processo de otimização, aumentando-se o número de vértices do *simplex* usado no algoritmo do *DownHill Simplex*. Esta abordagem permite melhorar o processo de otimização, que, dessa maneira, já é iniciado com posturas próximas à desejada.

Quando pelo menos uma postura é encontrada no banco de dados, é executado um processo de otimização especial, diferente do mostrado na seção 2.8.1. Neste processo, são realizadas apenas as duas primeiras etapas de otimização, melhorando o desempenho do sistema.

As posturas encontradas no banco de dados são usadas apenas na iniciação da primeira etapa. Além disso, como foi removida a terceira etapa, aumentou-se o número de iterações da segunda etapa de 30 para 50.

O desafio no uso desta técnica é não permitir que posturas erradas sejam armazenadas definitivamente no banco de dados. Para evitar esta situação, é usado um conjunto de regras que verifica se os resultados das avaliações podem ser considerados satisfatórios.

A solução foi permitir que apenas os menores valores absolutos da função de custo, resultantes das avaliações, fossem classificados como bons resultados, permitindo que apenas estes sejam armazenados no banco de dados.

Para identificar os menores valores das avaliações, os resultados são normalizados entre 0,0 e 1,0. Para isso, primeiramente, registram-se o maior e o menor valor obtido na função de custo entre as primeiras 500 avaliações realizadas, visto que não se conhece *a priori* o que vem a ser um bom resultado para essa função. Durante esse período de aprendizado, nenhum valor é armazenado no banco de dados. A cada avaliação, esses valores, mínimo e máximo, são reajustados, caso necessário.

Finalmente, para avaliar se uma postura encontrada pelo sistema é suficientemente confiável para ser armazenada, verifica-se se o seu valor na função de custo apresenta uma variação de, no máximo, 10% do limite inferior. Este percentual foi definido através de testes experimentais.

Obviamente, se o sistema de aquisição for iniciado com posturas incorretas, os valores limites permitirão o armazenamento de valores inconsistentes. Entretanto, lembra-se aqui ser uma restrição deste sistema o posicionamento correto do ator em relação ao modelo no primeiro quadro do vídeo.

Na Figura 2-10 é mostrado o resumo dos passos do processo de predição de posturas.

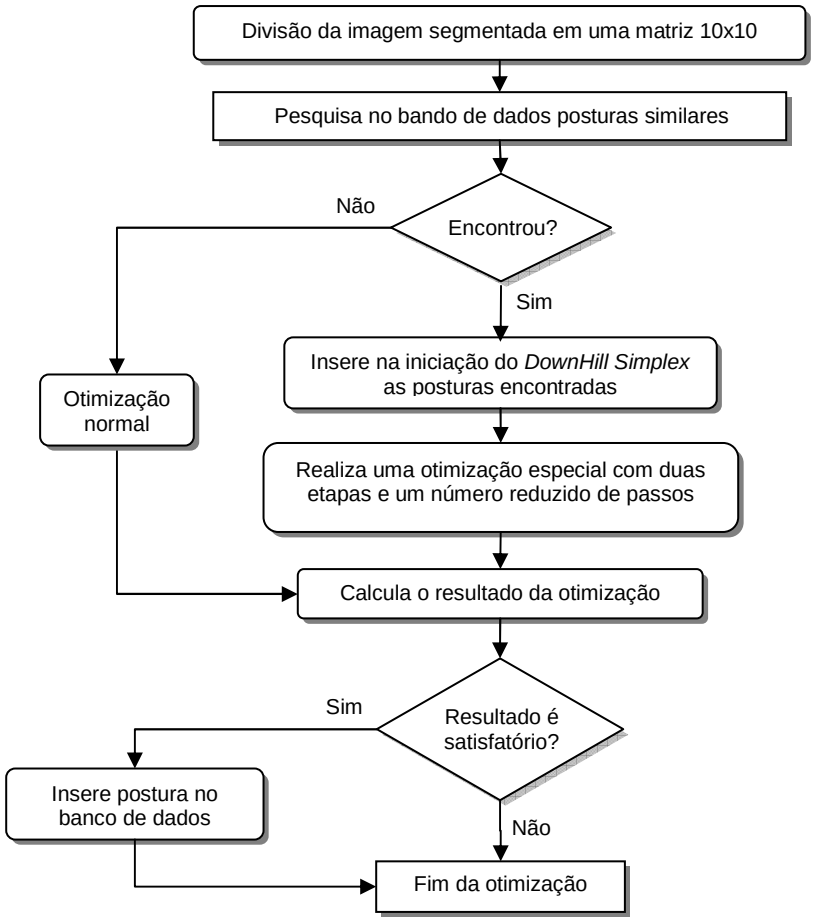

Figura 2-10: processo de predição de posturas através do banco de dados.

2.10.2 Armazenamento dos dados e otimizações

Todos os comandos executados no BD utilizam o padrão SQL, sigla em inglês de Linguagem de Consulta Estruturada, desenvolvido originalmente no início dos anos 70 nos laboratórios da IBM. SQL é um padrão de comunicação com o BD, assim como o OpenGL é um padrão de comunicação com as placas gráficas.

Para aumentar o desempenho, os comandos SQL são compilados uma única vez no início do programa, assim, nenhum dos comandos SQL das consultas e das inserções de dados necessitam ser interpretados, pois os mesmos já estão prontos para execução.

Numa primeira abordagem, os dados eram armazenados no BD em 100 campos, um para cada valor da matriz. Contudo, a busca seguindo o padrão SQL

era realizada no formato, “c1=v1 AND c2=v2 AND c3=v3 AND... c100=v100”, onde “c1” a “c100” representam os campos, e “v1” a “v100” os valores procurados.

Para aumentar a eficiência na busca, foi desenvolvida outra forma de armazenamento. Os campos armazenam valores inteiros de 32bits, mas os dados armazenados são histogramas da imagem fracionada, que contém no máximo 12x16 *pixels*. Assim, o histograma pode chegar a um valor máximo de 192, necessitando apenas de 8 bits para armazenamento. Assim, ao invés de armazenar os valores em 100 campos inteiros, todos os valores são armazenados em um campo binário de 100 bytes. Com isso, a busca é realizada com o uso de um único campo no formato “c=`XXXX`”, onde XXXX representam os 100 bytes resultantes da composição dos valores de cada campo.

Para comparar o desempenho das técnicas usando cem e apenas um campo, foi calculado o tempo necessário para realizar 500 pesquisas no BD. Além da média das 500 pesquisas, também foram armazenados o tempo da pesquisa mais rápida e da mais lenta, para se observar a variação de tempo, como é mostrado na Tabela 2-5, concluindo-se que o uso do campo no formato binário reduziu o tempo médio de busca no BD em 14 vezes.

Tabela 2-5: tempo de 500 pesquisas no banco de dados usando as duas técnicas.

	Campo Inteiro	**Campo Binário**
Média	505 ms	35 ms
Menor	472 ms	31 ms
Maior	621 ms	47 ms

Como o banco de dados encontra-se em memória, para evitar o estouro da mesma, foi definida uma quantidade de dados que o dito banco pode suportar. Esta quantidade é configurável pelo usuário, que pode modificá-la de acordo com a memória disponível no computador utilizado. Utilizou-se um limite de 100.000 registros, que nos testes realizados ocupou 15.753KB de memória. O máximo de registros utilizados não ultrapassou 30.000 registros.

2.11 Ambigüidades

As técnicas de identificação de posturas traduzidas em informações tridimensionais, usando como base imagens bidimensionais, estão sujeitas ao problema de ambigüidade. Isso se deve à ausência de informação de profundidade.

As métricas utilizadas tentam evitá-las, mas não há como garantir resultados não ambíguos. Algumas situações de ambigüidade são apresentadas a seguir.

2.11.1 Posturas diferentes, mas visualmente parecidas

Na Figura 2-11 são mostradas duas imagens de sobreposição da projeção do modelo 3D em duas configurações diferentes sobre a mesma imagem segmentada. Observa-se aqui um exemplo de ambigüidade em que as imagens, quase idênticas aos olhos humanos, são geradas a partir de parâmetros articulatórios bem diferentes. Somente através de uma inspeção minuciosa é possível diferenciá-las.

A imagem da esquerda foi gerada com o modelo na postura inicial, em que todos os graus de liberdade recebem o valor zero. O valor da função de custo, usando a métrica da superfície de não-recobrimento, é igual a 0,33906; a imagem da direita, por outro lado, tem um custo menor 0,15681 mas com parâmetros articulatórios bem diferentes, como mostra a Tabela 2-6.

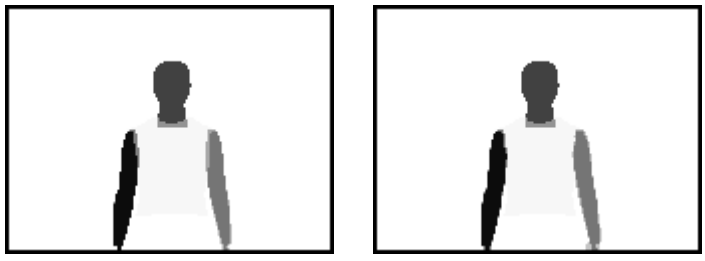

Figura 2-11: ambigüidade inerente ao método.

Note que isso pode ocorrer devido à diferença anatômica entre o ser humano e o modelo 3D rígido e é um problema de difícil solução.

Nesses casos, o algoritmo de otimização prioriza a postura com menor valor da avaliação da função de custo, não sendo possível identificar com precisão a posição correta.

Tabela 2-6: ângulos dos graus de liberdade do modelo em ambigüidade.

Ombro esquerdo			Cotovelo esquerdo	Ombro direito			Cotovelo direito	Resultado da avaliação
X	Y	Z		X	Y	Z		
0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,33906
21,0	35,0	2,0	-1,0	12,0	-17,0	0,0	-3,0	0,15681

2.11.2 Braços invertidos

Outra situação em que é recorrente a ocorrência de ambigüidade é mostrada na Figura 2-12, em que a imagem (a) corresponde à imagem segmentada, (b) e (c) ao modelo em diferentes posturas, e (d) e (e) à imagem de superposição gerada, respectivamente, pela soma das imagens (b) e (c) com a imagem segmentada (a).

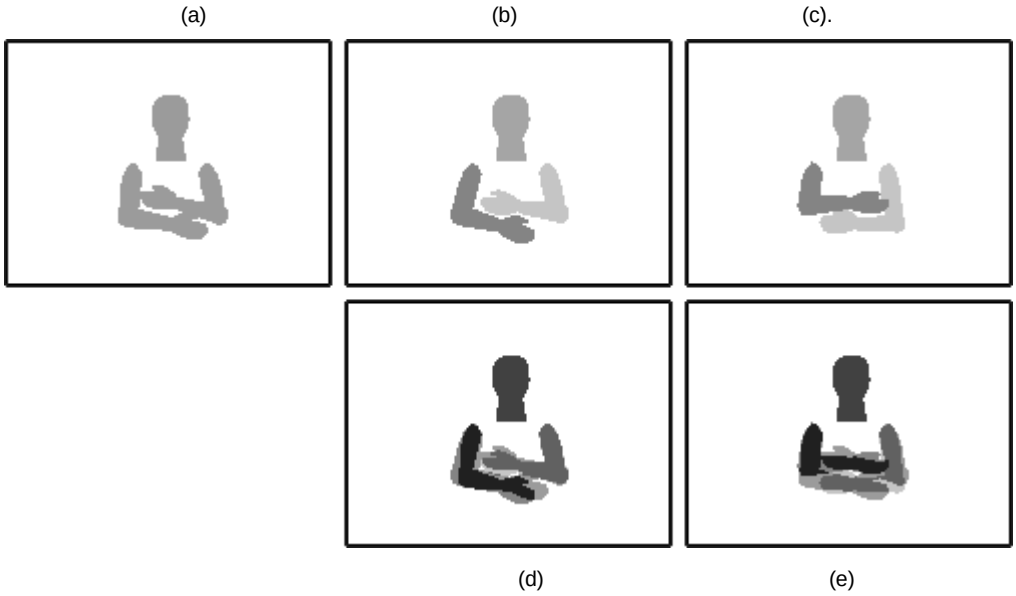

Figura 2-12: outro exemplo de ambigüidade inerente ao método.

A ambigüidade observada na função de custo na métrica da superfície de não recobrimento ocorre porque as imagens de superposição (d) e (e) contêm quase a mesma quantidade de *pixels* não recobertos para ambos os modelos, mesmo estando o modelo com os braços em posições invertidas nas duas imagens.

2.11.3 Partes oclusas ou encobertas

Nas técnicas utilizadas, as regiões de pele superpostas ou oclusas são de difícil identificação, podendo implicar em ambigüidades por permitirem que tais membros assumam qualquer postura sem haver como diferenciá-las.

Outros fatores que incidem em um problema de ambigüidade semelhante e sem solução ocorrem quando um membro está encoberto por outra parte do corpo. Por exemplo, o braço atrás do tronco, ou quando uma parte do corpo está fora da região capturada pela câmera.

Fora do campo de visualização do sistema, um membro pode assumir qualquer postura e não será possível ao sistema diferenciar entre diferentes configurações de parâmetros articulatórios candidatos.

A seqüência de imagem à esquerda da Figura 2-13 mostra um exemplo de ambigüidade quando parte do braço se encontra fora da área capturada. A parte com fundo preto é a imagem real da visão do sistema, e a parte com fundo cinza foi inserida especialmente para mostrar a postura que não está sendo visualizada pelo sistema.

As imagens à direita da Figura 2-13 mostram outro exemplo. Como o sistema não pode visualizar o final do membro, grande parte da percepção de profundidade é perdida, incorrendo-se em um erro e em que o braço é posicionado em direção à câmera. Em ambos os casos, não há como verificar o erro pelo resultado da função de custo.

2.11.4 Rotação dos braços

O problema que reflete a principal ocorrência de ambigüidade no sistema aqui apresentado é a rotação do braço.

Como exemplo, tomam-se as seqüências de vídeo em que as mãos não são visualizadas na posição de inicialização. Na Figura 2-14, simula-se uma situação em que a parte com fundo preto corresponde à área capturada pela câmera, e a parte com fundo cinza corresponde à parte da cena que não está sendo visualizada pelo sistema (por não ter sido capturada pela câmera).

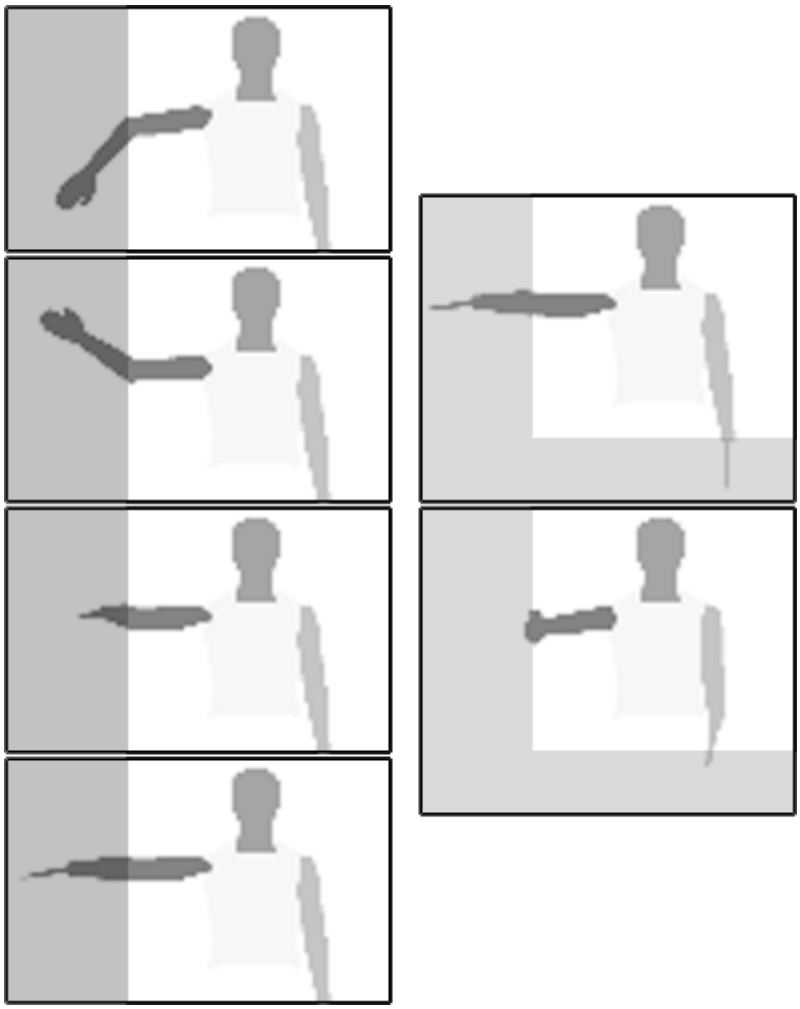

Figura 2-13: ambigüidade causada pela não visualização parcial ou completa de algum membro.

Na imagem da esquerda o modelo se encontra em sua postura inicial, com todos os graus de liberdade em zero, e na imagem da direita o modelo sofreu uma rotação de 90 graus nos dois braços.

Observa-se que as imagens consideradas pelo sistema (as partes com fundo preto) são quase iguais, mesmo tendo sido obtidas a partir de parâmetros que contêm grandes diferenças de valores.

Todos os problemas de ambigüidade aqui apresentados constituem uma limitação da abordagem de identificação de posturas em sistemas de visão computacional monoculares.

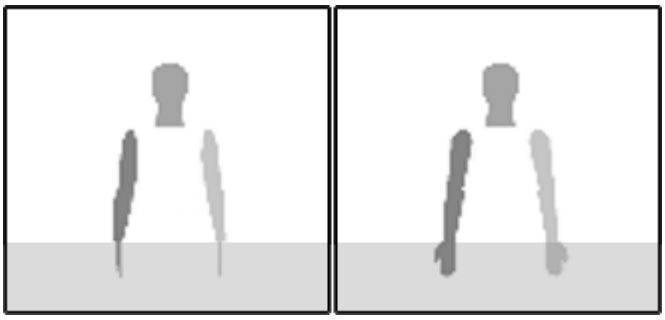

Figura 2-14: ambigüidade causada pela rotação do braço.

2.12 Quadro resumo com as técnicas desenvolvidas

As técnicas desenvolvidas e descritas neste capítulo são sintetizadas na Tabela 2-7, indicando resumidamente os resultados alcançados.

Tabela 2-7: resumo com as técnicas desenvolvidas e os resultados alcançados.

Técnica	Objetivo	Resultados alcançados
Estudo das placas gráficas para melhoria do desempenho.	Redução do tempo de transferência de dados (imagem projetada) da placa gráfica para a memória principal do computador.	Usando-se *framebuffers* auxiliares no padrão GL_R3_G3_B2, o tempo de transferência foi reduzido em aproximadamente 25% em média.
Restrições biomecânicas dinâmicas das articulações dos membros.	Evitar a avaliação de posturas irreais para o modelo tridimensional.	Melhora da qualidade da aquisição e diminuição no espaço de busca, pois impede-se a avaliação de muitas posturas ergonomicamente desfavoráveis.
Penalização de posturas irreais.	Evitar a avaliação da similaridade para configurações de posturas irreais no *Downhill Simplex*, impedindo a convergência forçada pelas restrições biomecânicas.	Melhora do desempenho, mas diminuição na qualidade da aquisição.
Iniciação do *simplex* com parâmetros aleatórios.	Reduzir a convergência para mínimos locais.	Redução da incidência de mínimos locais, sem, entretanto, prejudicar o desempenho do algoritmo de otimização.
Classificação das regiões de pele da imagem capturada para identificação dos membros correspondentes.	Identificar independentemente os membros, cabeça e braços, do ator de forma a reduzir a incidência de ambigüidades na avaliação da similaridade.	Não se demonstrou uma técnica robusta, apresentando bons resultados apenas em situações particulares.
Predição de posturas através do uso de um banco de dados.	Evitar o reprocessamento de posturas idênticas.	Demonstrou-se eficiente, melhorando o desempenho do sistema e reduzindo erros de identificação. Contudo há a necessidade de realizar um treinamento prévio.

No próximo capítulo, são abordadas, de maneira particular, as métricas de similaridade usadas para comparar a imagem de vídeo segmentada com as imagens sintetizadas a partir da projeção do modelo.

Capítulo 3

3 Métricas de similaridade para avaliação da postura

As métricas de similaridade são responsáveis por mensurar a semelhança entre a imagem capturada e as imagens sintetizadas pela projeção do modelo tridimensional ao longo do processo de otimização.

Para cada postura estimada para o modelo 3D, uma taxa de erro é calculada pela métrica. Essa taxa deve ser minimizada a fim de se encontrar a posição mais aproximada do modelo.

Neste capítulo são apresentadas as diferentes métricas de similaridade utilizadas no sistema de aquisição de gestos desenvolvido no presente trabalho.

3.1 Distância entre os centróides dos contornos

Esta técnica realiza o cálculo da similaridade pela comparação dos centróides dos polígonos gerados pelos pontos vetorizados dos contornos das imagens. Estes formam o contorno das regiões da imagem correspondentes à cabeça, ao braço esquerdo e ao braço direito, como é apresentado na Figura 3-1. Posteriormente, o algoritmo analisa os centróides para os contornos das imagens segmentadas, obtidos pelo processamento da imagem capturada pela câmera, e de síntese, obtida pela projeção do modelo tridimensional no plano, realizando, em seguida, a correspondência entre os valores encontrados e os contornos correspondentes.

Usando esta técnica, o algoritmo pode encontrar diversas posturas ambíguas para os membros. Para reduzir este problema, os perímetros dos contornos também são considerados no cálculo da correspondência, permitindo tratar a ambigüidade em alguns casos, sendo insuficiente em outros. Alguns problemas decorrem da oclusão (superposição de regiões de pele), bem como de inconsistências na

segmentação provocadas por ruídos, fazendo com que um mesmo membro seja dividido em mais de um contorno.

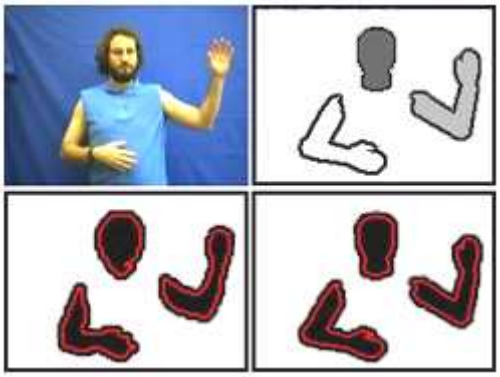

Figura 3-1: extração dos contornos dos braços e da cabeça.

3.1.1 Extração dos contornos

Algumas técnicas utilizadas para cálculo de distância entre contorno de imagens levam em consideração a existência de pontos correspondentes, par a par, e permitem quantificar de forma precisa as diferenças entre dois contornos.

Dois métodos bastante utilizados são os seguintes:

– Distância média dos pontos que representam o contorno, que é definida por

$$\overline{d} = \frac{1}{N}\sum_{i=1}^{N} d_i \qquad (3\text{-}1)$$

em que "*N*" é o número de pontos, e "d_i", a distância entre os dois pontos correspondentes. Esta média mostra de forma global a distância entre os contornos.

– Índice de mérito de Pratt [49], que é definido por

$$F_{Prat} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1+\alpha\, d_i^2} \qquad (3\text{-}2)$$

em que "*N*" é o número de pontos, "d_i" é a distância entre dois pontos correspondentes, e "α" é um parâmetro associado ao tamanho do contorno. Este índice corresponde a uma medida relativa do comportamento global das distâncias entre contornos, que varia no intervalo [0,1], em que 1 equivale à comparação de dois contornos iguais.

A biblioteca OpenCV oferece três métodos que permitem calcular a distância entre contornos correspondentes. Entretanto, como não existe correspondência precisa entre o número de pontos do contorno da imagem segmentada e o da projeção do modelo, nenhum dos dois métodos apresentados é aplicável. Por isso, outras técnicas foram utilizadas.

Para verificar a similaridade entre contornos, foram exploradas as propriedades que podem ser calculadas através dos pontos que formam os mesmos. O primeiro método proposto foi o cálculo do centro de massa dos pontos que representam o contorno, que pode ser executado com baixo custo computacional. Entretanto, para contornos ruidosos, o que é freqüente em imagens segmentadas, o método não apresenta resultados satisfatórios. Por exemplo, o centro de massa de quatro pontos formando uma região retangular coincide exatamente com o seu centro gravitacional. Caso o contorno encontrado possua cinco pontos, sendo dois destes muito próximos, ainda que seja imperceptível visualmente, o cálculo do centro de massa ficará deslocado para o vértice dos pontos vizinhos, como mostra o exemplo na Figura 3-2.

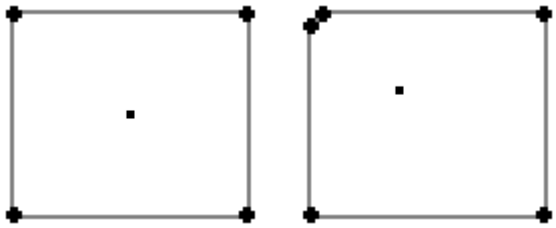

Figura 3-2: centro de massa dos pontos para dois objetos aparentemente similares.

O cálculo do centróide permite solucionar este problema, apesar de possuir um custo computacional maior do que o método anterior. Através dele, pode-se calcular a similaridade entre contornos utilizando o centro gravitacional que não depende do número de pontos, levando em conta a estrutura geométrica do objeto. Na Figura 3-3 é mostrado o mesmo exemplo da Figura 3-2, usando o centróide. O cálculo do centróide foi realizado conforme as Equações (3-3), em que “*N*” é o número de pontos, “x_i” e “y_i” são as coordenadas de cada ponto respectivo do contorno.

$$A(P) = \frac{1}{2}\sum_{i=1}^{N} x_i y_{i+1} - y_i x_{i+1}$$

$$x = \frac{\sum_{i=1}^{N} (x_{i+1} + x_i) \times (x_i y_{i+1} - y_i x_{i+1})}{3A(P)} \qquad (3\text{-}3)$$

$$y = \frac{\sum_{i=1}^{N} (y_{i+1} + y_i) \times (x_i y_{i+1} - y_i x_{i+1})}{3A(P)}$$

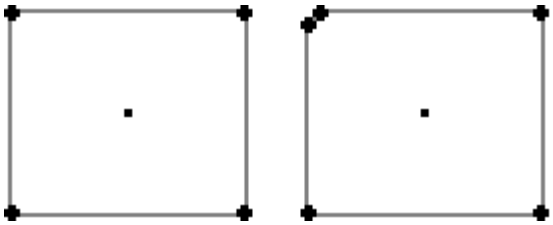

Figura 3-3: centróide dos pontos para dois objetos aparentemente similares.

3.1.2 Função de avaliação da similaridade entre contornos

Esta técnica realiza o cálculo da similaridade pela comparação dos centróides dos polígonos gerados pelos pontos vetorizados relativos aos contornos das regiões de cor de pele da imagem segmentada e de síntese. Estes formam o contorno das regiões da imagem correspondentes à cabeça, ao braço esquerdo e ao braço direito.

Os contornos são extraídos através de métodos implementados pelo OpenCV [44], baseados no trabalho de Suzuki *et al.* [48]. Em seguida, são selecionados os maiores polígonos. Estes representam as maiores áreas de informação na imagem que geralmente estão relacionadas aos membros do ator. Esta seleção, automaticamente, realiza um filtro de ruído na imagem segmentada, pois os ruídos presentes na imagem segmentada, geralmente são áreas pequenas, descartadas nesta seleção.

Posteriormente, o algoritmo analisa os centróides da imagem segmentada e do modelo, realizando a correspondência entre os centróides de cada contorno referente aos membros respectivos. Entretanto, na tentativa de encontrar os pares de membros correspondentes, o algoritmo pode incorrer em ambigüidades. Um exemplo é a ocorrência de oclusões, quando há a superposição de duas ou mais

regiões de pele. Outras decorrem de inconsistências na segmentação, fazendo com que um mesmo membro seja dividido em mais de um contorno.

Como a técnica não diferencia o formato dos polígonos, utilizou-se, para cada parte do corpo, como parâmetro adicional, o perímetro em conjunto com o centróide. O algoritmo executado por essa métrica de similaridade é representado na Figura 3-4.

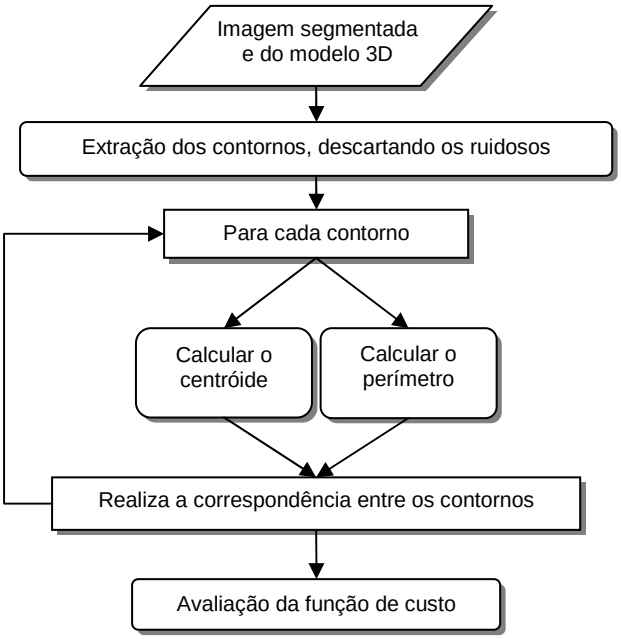

Figura 3-4: processo de aquisição utilizando a métrica da distância dos centróides dos contornos.

A função que calcula a similaridade por esta técnica é apresentada na Equação (3-4) e corresponde ao somatório resultante da soma do módulo das distâncias dos contornos e do módulo da diferença dos perímetros, multiplicado por uma constante "*k*", em que "*N*" é o número de contornos, "dc_i" e "dp_i" são, respectivamente, a diferença dos centróides e a diferença dos perímetros.

$$f_C = \sum_{i=1}^{N} \left(\left| dc_i \right| + k \left| dp_i \right| \right) \tag{3-4}$$

A constante *k* é um valor encontrado de forma empírica, e foi inserida para equilibrar a diferença dos perímetros em relação à distância entre os centróides. A função prioriza os perímetros, pois estes possuem valores absolutos e gradiente

maiores do que os centróides, contudo, os testes mostraram que o processo de aquisição tem um resultado melhor, as se minimizar a diferença entre os centróides. Através desta análise, foi inserida esta constante com valores inferiores à unidade, sendo definido o valor numérico da constante *k* igual a 0,6.

3.1.3 Limitações

A métrica da distância entre os contornos possui um elevado grau de ambigüidade para posturas bastante diferentes, impossibilitando o uso desta. Na Figura 3-5 é mostrado um exemplo, em que os braços de ambas as imagens possuem centróides e perímetros semelhantes, o que as faria semelhantes pela métrica aqui apresentada.

Esta limitação levou a uma adaptação da técnica que permite reduzir de maneira significativa esse tipo de ambigüidade, como apresentado na próxima subseção.

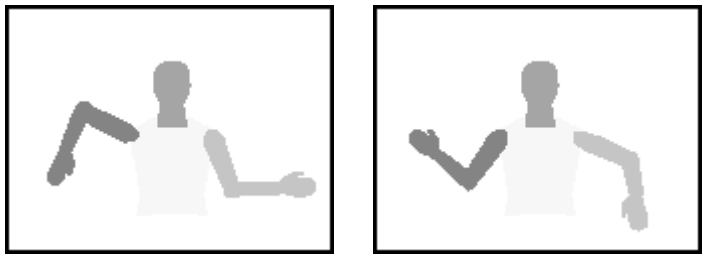

Figura 3-5: problema da métrica pela distância entre os contornos.

3.2 Distância entre os centróides dos contornos com divisão da imagem

Para reduzir os problemas de ambigüidade da técnica anterior, as imagens foram fracionadas em uma grade, subdividindo assim as curvas correspondentes aos contornos em várias sub-regiões, possibilitando uma maior precisão em relação à posição dos centróides, como é mostrado na Figura 3-6. Nesta abordagem, a função de custo é resultante da comparação das frações correspondentes na grade de cada uma das imagens (segmentada e projetada), aplicando sobre estas um

conjunto de regras para verificar a diferença de cada fração da imagem segmentada com a fração correspondente na imagem do modelo.

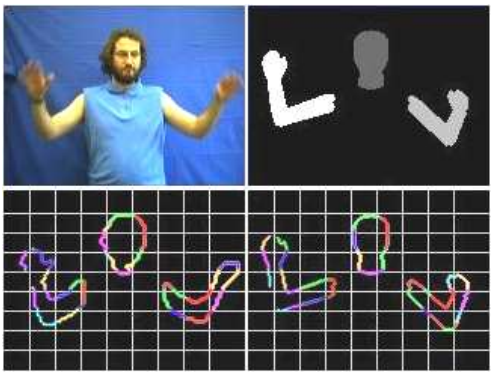

Figura 3-6: imagens (a) capturada da câmera, (b) contorno da imagem segmentada dividida em grade, (c) projeção do modelo 3D e (d) contorno do modelo dividido em grade.

3.2.1 Função de avaliação

A métrica utilizada possui uma abordagem diferente da anterior em relação à avaliação da função de custo. Nesta, a avaliação é realizada comparando as frações correspondentes na grade, aplicando sobre estas um conjunto de regras para verificar a diferença entre cada *n-ésima* fração da imagem segmentada com a mesma fração na imagem do modelo. Na Figura 3-7 é apresentado o fluxograma do processo utilizado para a verificação de similaridade.

Cada fração é avaliada atribuindo-se um valor de acordo com um conjunto de regras. Após avaliar todas as frações, é calculada a média, resultando no custo da avaliação da similaridade.

As regras para avaliação de cada fração são apresentadas a seguir:

- Caso as frações correspondentes não possuam contorno, são ignoradas e não são contabilizadas no calculo da média;
- Caso o número de contornos entre as frações correspondentes sejam iguais, é calculado o centróide de ambas. O valor atribuído a esta fração é igual ao valor da distância entre os centróides, dividido pela largura da fração. Isso normaliza o valor entre zero e um;

- Caso os números de contornos entre as frações correspondentes sejam diferentes, é atribuído o valor 1 (um), o que representa a penalização máxima para a fração.

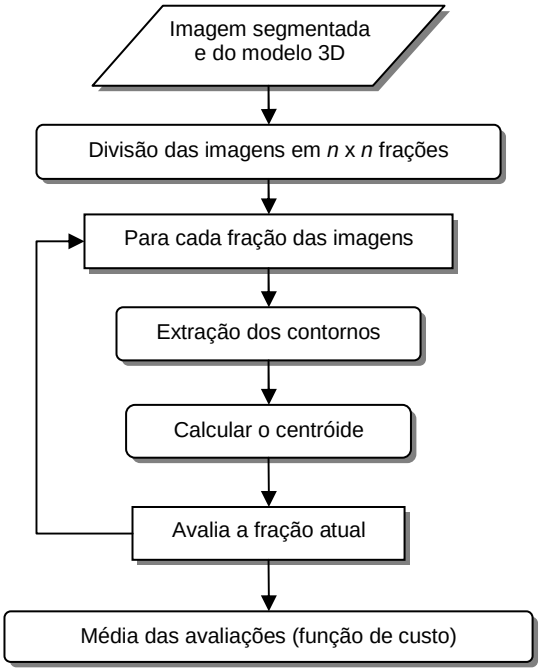

Figura 3-7: processo de aquisição utilizando a métrica da distância dos centróides dos contornos com divisão da imagem.

3.2.2 Limitações

O fracionamento da imagem e a extração dos contornos para cada uma das frações elevam o custo computacional, reduzindo bastante o desempenho desta métrica em comparação à apresentada anteriormente.

3.3 Superfície de não recobrimento

Esta métrica analisa o histograma da imagem de superposição, resultante da soma da imagem segmentada com a imagem de projeção do modelo 3D, como exemplificado na Figura 3-8. São utilizadas cores diferentes na imagem segmentada e no modelo, a fim de evitar coincidências de cores de regiões diferentes após a soma. Cada região colorida na imagem resultante indica o recobrimento ou o não recobrimento de uma região de pele. A taxa de não recobrimento contabiliza o

número de *pixels* de regiões de pele não coincidentes entre as imagens, segmentada e projetada.

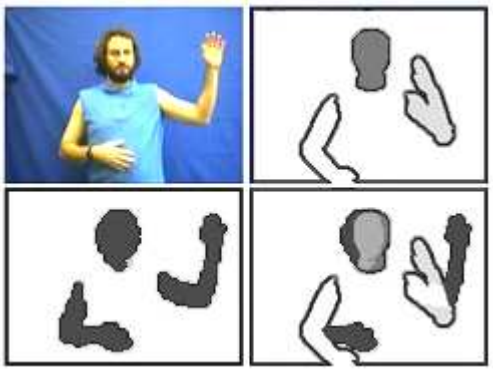

Figura 3-8: imagens (a) capturada da câmera, (b) segmentada, (c) projeção do modelo 3D e (d) superposição das imagens segmentada e do modelo.

3.3.1 Função de avaliação

O histograma da imagem de superposição é calculado para obter-se o número de *pixels* pertencentes a cada cor, e esta cor representa um tipo de informação da avaliação, por exemplo: braço recoberto, braço não recoberto, pele não recoberta.

Para cada postura dada ao modelo 3D, é calculada uma "taxa de não recobrimento", que mede o número de *pixels* não coincidentes na imagem de superposição (Figura 3-9). Esta função é expressa na Equação (3-5):

$$F(q) = \sum_{c=1}^{m} \left(\frac{\left| A_c \cup B_c(q) \right| - \left| A_c \cap B_c(q) \right|}{\left| A_c \cup B_c(q) \right|} \right) \qquad (3\text{-}5)$$

onde: "*q*" é o vetor de parâmetros da articulação, que descreve a postura candidata, A_c é o conjunto de *pixels* pertencentes à *c-ésima* classe de cores na imagem segmentada, $B_c(q)$ é o conjunto de *pixels* portadores da *c-ésima* classe de cores na projeção do modelo, *m* é o número de classes de cores utilizadas, e |*X*| designa o número de *pixels* de um conjunto *X*.

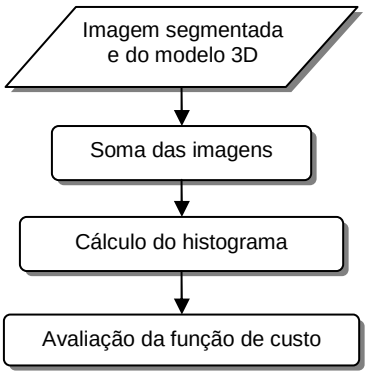

Figura 3-9: processo de aquisição utilizando a métrica da taxa de não recobrimento.

3.3.2 Limitações

Não há como a métrica diferenciar se existe uma grande área descasada ou é devido à soma de pequenas áreas descasadas. Na Figura 3-10 é mostrado um exemplo, onde na figura da esquerda existem regiões descasadas, decorrentes das pequenas áreas. Neste caso, a soma dos *pixels* não recobertos soma 533 *pixels*. Na figura da direita existem duas grandes áreas descobertas, somando 450 *pixels*. Observa-se, visualmente, que a postura da esquerda está muito mais próxima da postura desejada do que a da direita, mas, mesmo assim, esta tem uma quantidade de *pixels* descoberta maior.

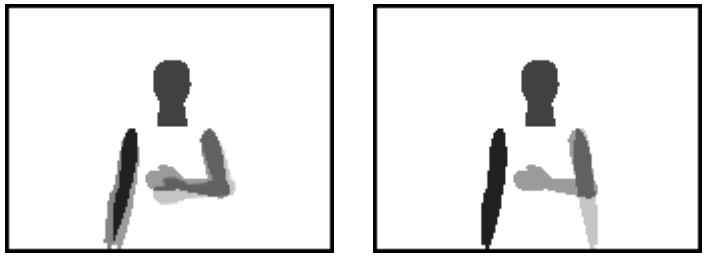

Figura 3-10: situações onde a métrica de não recobrimento falha em sua análise.

3.4 Superfície de não recobrimento com divisão da imagem

De forma semelhante à realizada com os contornos, aplicando a divisão da imagem para a superfície de não recobrimento, esta insere informações com relação à posição, como é mostrado na Figura 3-11.

A métrica da superfície de não recobrimento, sozinha, tem a qualidade de ser possível mensurar de forma precisa a diferença global entre as imagens devido sua análise ser em *pixels*. Contudo, esta métrica não tem informação sobre a posição de onde se encontra o descasamento, e se este descasamento é bem localizado, como é mostrado na Figura 3-10.

A divisão da imagem possibilita a identificação das regiões onde se encontra o descasamento e torna possível identificar áreas isoladas, onde não houve recobrimento, para poder penalizar a avaliação destas.

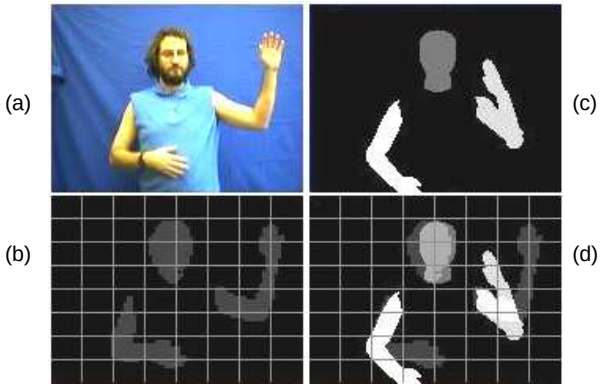

Figura 3-11: imagens (a) capturada da câmera, (b) segmentada com divisão da imagem, (c) projeção do modelo 3D e (d) superposição das imagens segmentada e do modelo.

3.4.1 Função de avaliação

Nesta métrica, a avaliação é realizada comparando as frações correspondentes na grade, aplicando sobre estas um conjunto de regras para verificar a diferença de cada fração da imagem segmentada com a mesma fração na imagem do modelo (Figura 3-12).

Para cada fração é calculada a taxa de não recobrimento, a mesma utilizada na métrica não fracionada da seção 3.3.1, mostrada na Equação (3-5). Caso não exista nenhum *pixel*, a fração é ignorada.

Ao final, é realizada a média das avaliações de cada fração não ignorada, resultando no custo da avaliação das imagens.

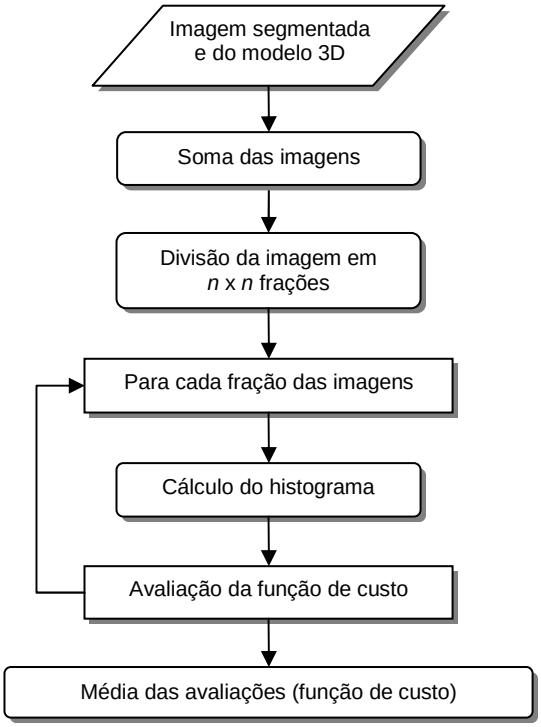

Figura 3-12: processo de aquisição utilizando a métrica da taxa de não recobrimento fracionada.

3.4.2 Função de avaliação quadrática

Na tentativa de penalizar mais as frações isoladas não recobertas, para evitar grandes áreas não recobertas, foi desenvolvida uma variação da função de avaliação, calculando a média quadrática ao invés da média aritmética.

Esta média quadrática penaliza caso existam frações com um alto nível de descasamento, tendendo a aumentar a média, caso as frações não estejam com valores próximos.

3.4.3 Limitações

Devido à penalização de frações onde não há recobrimento, há a possibilidade de penalizar regiões pequenas que não foram recobertas nas extremidades das frações.

Ao fracionar a imagem em sessenta e quatro partes, com a configuração do modelo utilizada, os braços são divididos em sua extremidade. Neste caso, uma

linha de *pixels* ficou em uma fração isolada, quando estas não são recobertas, a penalização efetuada nestas frações pela função de custo é muito elevada, e, em algumas vezes, não representando a real diferença.

Na Figura 3-13 é mostrado que fracionando a imagem em 8x8, os braços são separados. No caso do braço esquerdo, ficou uma linha de um *pixel* de espessura em outra fração. Isso ocasionou vários erros de avaliação, pois estas frações eram penalizadas por só terem informação da imagem segmentada, mesmo o modelo estando ao lado da mesma.

Assim, o fracionamento foi realizado em 7x7 partes, deixando-o como mostra a imagem na Figura 3-14. Isto não impede que tais erros aconteçam, porém os testes mostraram que não apresentavam mais erros visíveis.

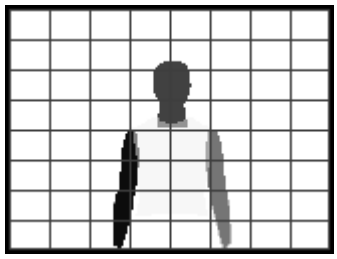

Figura 3-13: divisão do modelo em oito partes.

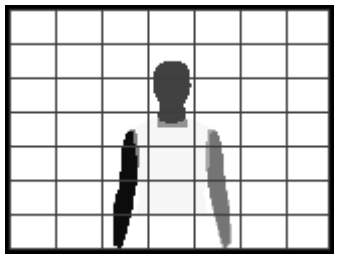

Figura 3-14: divisão do modelo em sete partes.

Este problema é recorrente a qualquer técnica de fracionamento usada, inclusive na métrica mostrada na seção 3.2.

3.5 Diferença em *pixel* entre as imagens

Como o interesse é saber o quanto a imagem segmentada está diferente da imagem projetada, pode-se utilizar a diferença entre as mesmas, de forma a definir sua similaridade. Sendo então realizada a diferença entre a imagem segmentada e a

imagem de projeção do modelo. A imagem resultante, Figura 3-15(d), é uma imagem na qual os *pixels* desta representam as áreas de descasamentos entre as imagens. Assim, conta-se o número de pixels diferentes de zero desta imagem, dividindo pela contagem do número de pixels da imagem segmentada.

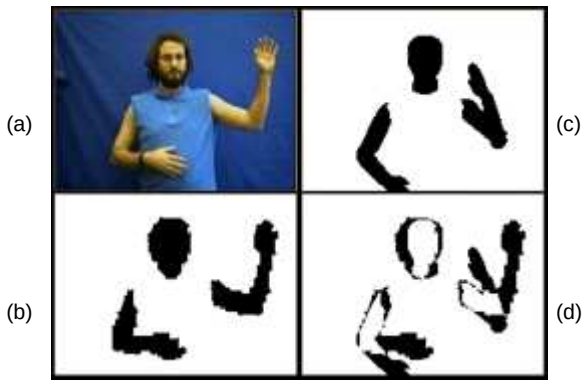

Figura 3-15: imagens (a) capturada da câmera, (b) segmentada, (c) projeção do modelo 3D e (d) diferença entre as imagens segmentada e do modelo.

3.5.1 Função de avaliação

A avaliação é o resultado da contagem do número de *pixels* diferentes de zero da imagem gerada pela diferença entre a imagem segmentada e a projeção do modelo, proporcional à quantidade de *pixels* da imagem segmentada. É mostrada na Equação (3-6) como o calculo é realizado, e na Figura 3-16, o processo desta métrica.

$$F(q) = \frac{|A(q) - S|}{|S|} \tag{3-6}$$

em que "*q*" é o vetor de parâmetros da articulação que descreve a postura candidata, "*A(q)*" é o conjunto de *pixels* portadores da projeção do modelo, "*S*" é o conjunto de *pixels* pertencentes à imagem segmentada, e |*X*| designa o número de *pixels* de um conjunto *X*.

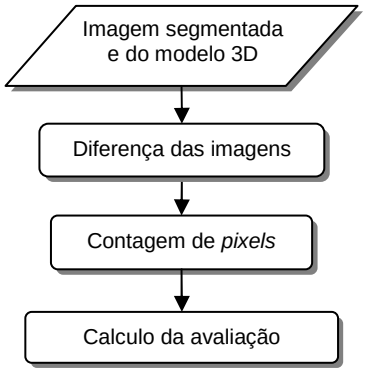

Figura 3-16: processo de aquisição utilizando a métrica da diferença em *pixel*.

3.5.2 Função de avaliação com sub-amostragem

Foi adicionado um nível de processamento especial, na tentativa de penalizar grandes áreas descasadas. Para obter as grandes áreas da imagem, foi utilizada a operação morfológica de erosão, de forma a deixar na imagem somente as áreas de grande descasamento.

A utilização da operação de morfologia através do OpenCV tornou-se muito lenta, impossibilitando seu uso em tempo real, diminuindo a velocidade abaixo dos cinco quadros por segundo, pois a implementação da erosão do OpenCV é complexa e realizada de forma interativa.

A estratégia utilizada para substituir a morfologia foi reduzir o tamanho da imagem para a proporção desejada da máscara da erosão. Os testes demonstraram que reduzir cinco vezes o tamanho da imagem trouxe bons resultados, sendo estes satisfatórios, assim como a morfologia. A redução da imagem tem um custo computacional baixo.

Assim, após a imagem ser analisada no primeiro passo, seu tamanho é reduzido em cinco vezes e, novamente, é realizada a contagem dos *pixels*, formando o segundo passo. A Figura 3-17 mostra tal imagem como resultado do processo na Figura 3-15(d). É mostrado na Equação (3-7) como o calculo é realizado, e na Figura 3-18 é exibido o processo desta métrica.

Figura 3-17: imagem reduzida representando o segundo passo do cálculo da diferença em *pixel* entre as imagens: segmentada e do modelo.

$$F(q) = \frac{|A(q) - S|}{|S|} * \left(1 + \frac{|B(q) - S|}{|S|}\right) \tag{3-7}$$

em que "*q"* é o vetor de parâmetros da articulação que descreve a postura candidata, "*A(q)"* é o conjunto de *pixels* portadores da projeção do modelo, *"B(q)"* é o conjunto de *pixels* portadores da projeção do modelo reduzida, "*S"* é o conjunto de *pixels* pertencentes à imagem segmentada, e |*X*| designa o número de *pixels* de um conjunto *X*.

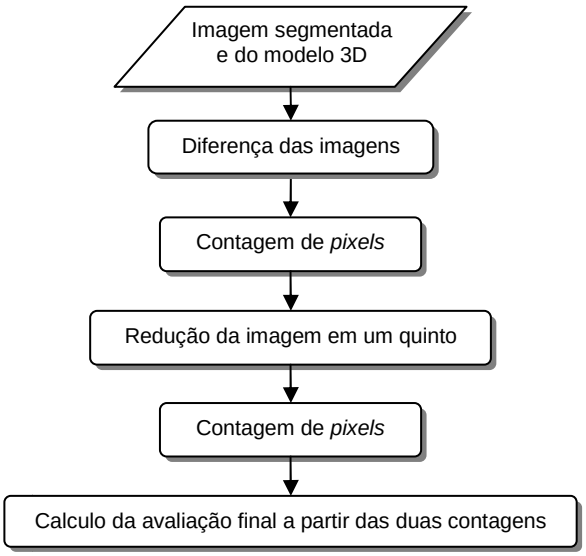

Figura 3-18: processo de aquisição utilizando a métrica da diferença em *pixel* com sub-amostragem.

3.5.3 Limitações

Como a contagem de *pixels* realizada não leva em consideração a cor dos membros, não há diferença entre os braços, e, assim, não há como a métrica diferenciar se os membros estão invertidos, como foi mostrado um exemplo desta ambigüidade na Figura 2-12 da seção 2.11.

Por este motivo, não é possível utilizar, em conjunto com esta métrica, a técnica da classificação das regiões de pele pela identificação dos membros correspondentes da seção 2.9.

3.6 Quadro resumo com as métricas desenvolvidas

As métricas desenvolvidas e descritas neste capítulo são sintetizadas na Tabela 3-1, indicando resumidamente os resultados alcançados.

Tabela 3-1: resumo com as métricas desenvolvidas e os resultados alcançados.

Métrica	Objetivo	Resultados alcançados
(1) Distância entre os centróides dos contornos	Analisar a similaridade pela distância dos contornos dos membros.	Métrica que apresentou os piores resultados, visto que permite um elevado grau de ambigüidades.
(2) Distância entre os centróides dos contornos com divisão da imagem	Analisar a similaridade pela distância dos contornos dos membros, considerando sua posição espacial.	Melhoria da qualidade em relação à métrica (1), embora com resultados ainda insatisfatórios.
(3) Superfície de não recobrimento	Verificar as regiões de descasamento pela diferença entre as imagens.	Métrica com boa relação qualidade e desempenho, conseguindo-se obter resultados satisfatórios com um desempenho intermediário.
(4) Superfície de não recobrimento com divisão da imagem	Verificar as regiões de descasamento pela diferença entre as imagens, considerando sua posição espacial.	A métrica apresentou os melhores resultados em relação à qualidade, mas com prejuízo do desempenho, chegando ao dobro do tempo de processamento.
(5) Superfície de não recobrimento com divisão da imagem usando a função quadrática de avaliação	Verificar as regiões de descasamento pela diferença entre as imagens, considerando sua posição espacial, e penalizar mais as regiões espacialmente descasadas.	A função quadrática penaliza excessivamente regiões descasadas, obtendo melhores resultados do que a métrica original (3), mas resultados piores do que a métrica (4), embora com o mesmo desempenho.
(6) Diferença em *pixel* entre as imagens	Avaliar a quantidade de *pixels* resultante da diferença entre as imagens.	A métrica apresenta o melhor desempenho alcançado entre todas as outras, permitindo uma avaliação de similatidade simples e rápida, obtendo resultados próximos aos da métrica (3).
(7) Diferença em *pixel* entre as imagens com sub-amostragem	Avaliar a quantidade de *pixels* resultante da diferença entre as imagens e, em seguida, avaliar as áreas com grande diferença através de uma sub-amostragem.	A sub-amostragem obteve resultados de qualidade próximos à melhor métrica (4), superiores aos obtidos em (6), mas com uma leve degradação de desempenho.

O próximo capítulo mostra, detalhadamente, os resultados alcançados por este trabalho, comparando-os entre si, explicitando a forma como tais dados foram obtidos.

Capítulo 4

4 Experimentação e análise de dados

A partir das modificações, novas técnicas e métricas propostas para o sistema de aquisição de gestos humanos, um conjunto de experimentações foram realizadas e serão apresentadas nas próximas subseções.

Para fins de melhor precisão na análise comparativa das técnicas, foram utilizados apenas 8 graus de liberdade, correspondendo aos ombros e cotovelos.

4.1 Base de vídeos sintéticos para testes

A utilização de vídeos reais ou ao vivo compromete a comparação das métricas, pois não há como saber quais são os parâmetros articulatórios exatos da postura de um ser humano. E também como o modelo tridimensional (3D) utilizado para identificar as posturas tem uma estrutura simplificada com relação ao corpo humano, os parâmetros articulatórios humanos não podem ser utilizados diretamente ao modelo 3D.

Para comparar os resultados dos algoritmos, foram geradas quatro seqüências de vídeo com o uso do modelo 3D usado na síntese das posturas de avaliação. O uso de vídeos sintéticos permite o cálculo do erro de forma mais precisa, já que se sabe, *a priori*, o conjunto de parâmetros utilizados para definir a postura do modelo em cada quadro.

Assim, para cada técnica, os resultados podem ser obtidos pelo cálculo do erro, que é a diferença dos graus de liberdade entre os parâmetros encontrados no modelo e os armazenados durante a geração do vídeo.

Foi averiguado no artigo [50], que ambigüidades são causadas em algumas situações pelo fato de os membros não estarem por completo dentro da visão do

sistema. Quando algum membro não está sendo visualizado, o sistema permite qualquer postura neste membro, não havendo como distingui-las.

Então foram realizados mais quatro vídeos sintéticos, com as mesmas seqüências de movimentos dos anteriores, alterando a posição do modelo de forma que o mesmo fique todo dentro da visão do sistema, totalizando oito vídeos, dois tipos de vídeos com quatro movimentos cada.

Para que o modelo ficasse sempre todo dentro da visão do sistema, seria necessário que sua posição ficasse muito afastada da visão, deixando o modelo pequeno e sem qualidade, por ser representado por poucos *pixels*.

Neste caso, não foi possível nem analisar os resultados da aquisição, pois antes disso o processo de segmentação foi prejudicado, perdendo informações devido aos algoritmos de redução de ruídos confundirem partes do modelo, como se os mesmos fossem ruídos causados pela baixa quantidade de *pixels*.

Foi definida uma posição intermediária, de forma que em algumas posturas somente parte das mãos fica fora da visão do sistema.

Os vídeos realizam movimentos predefinidos no modelo, no intuito de testar características pertinentes a um sistema de visão computacional, como será explicado posteriormente. Todos os vídeos foram gerados em três etapas. Eles iniciam com a postura em descanso, todas as articulações com valores zeros, como é mostrado na Figura 4-1. A primeira etapa realiza o movimento somente no braço esquerdo; em seguida um movimento semelhante é feito com o braço direito. Na última etapa, o movimento é feito simultaneamente em ambos os braços, terminando com o modelo voltando à postura de descanso.

A primeira seqüência foi definida para testar a percepção de profundidade do sistema, pelo movimento de ambos os braços do modelo, que é realizado para frente e de forma perpendicular ao corpo. Amostras do primeiro vídeo são mostradas na Figura 4-2. O mesmo movimento é realizado para o segundo tipo de vídeo, ou seja, o quinto vídeo, com a visão mais afastada, como é mostrado na Figura 4-3.

Este movimento analisa como o sistema se comporta, devido a sua visualização ser uma imagem bidimensional, e o modelo possuir três dimensões. Os vídeos contêm 146 quadros.

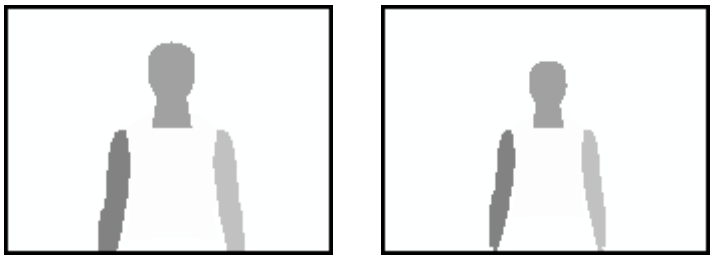

Figura 4-1: postura de descanso dos dois tipos de vídeo, todos os vídeos iniciam e terminam na postura de descanso.

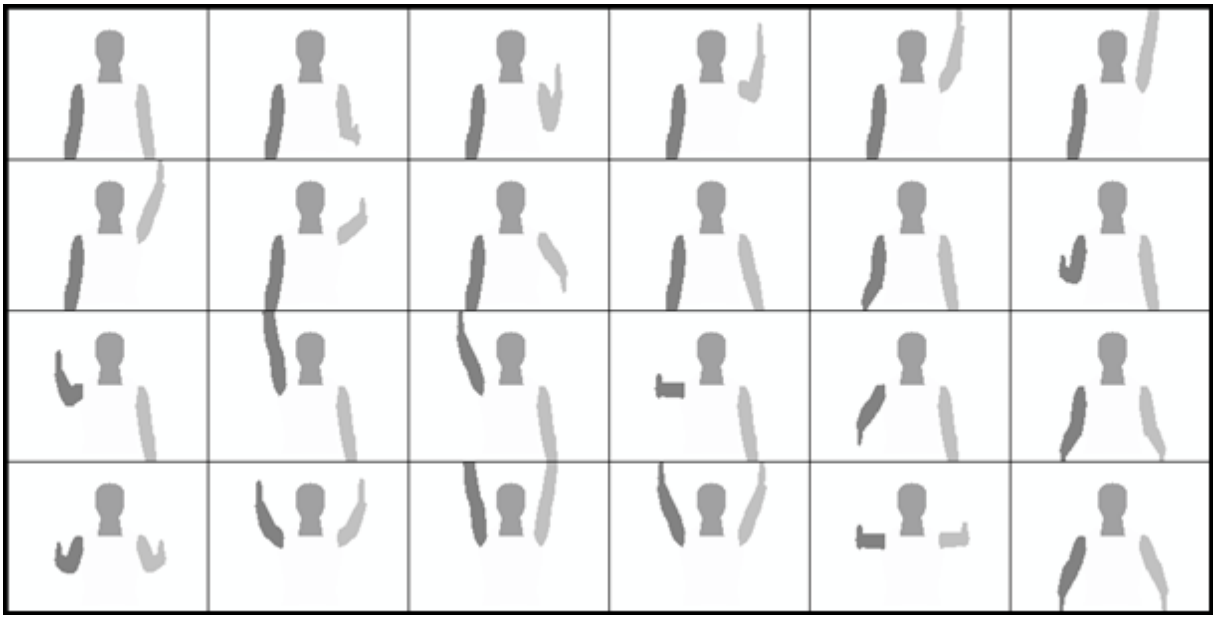

Figura 4-2: quadros do primeiro vídeo.

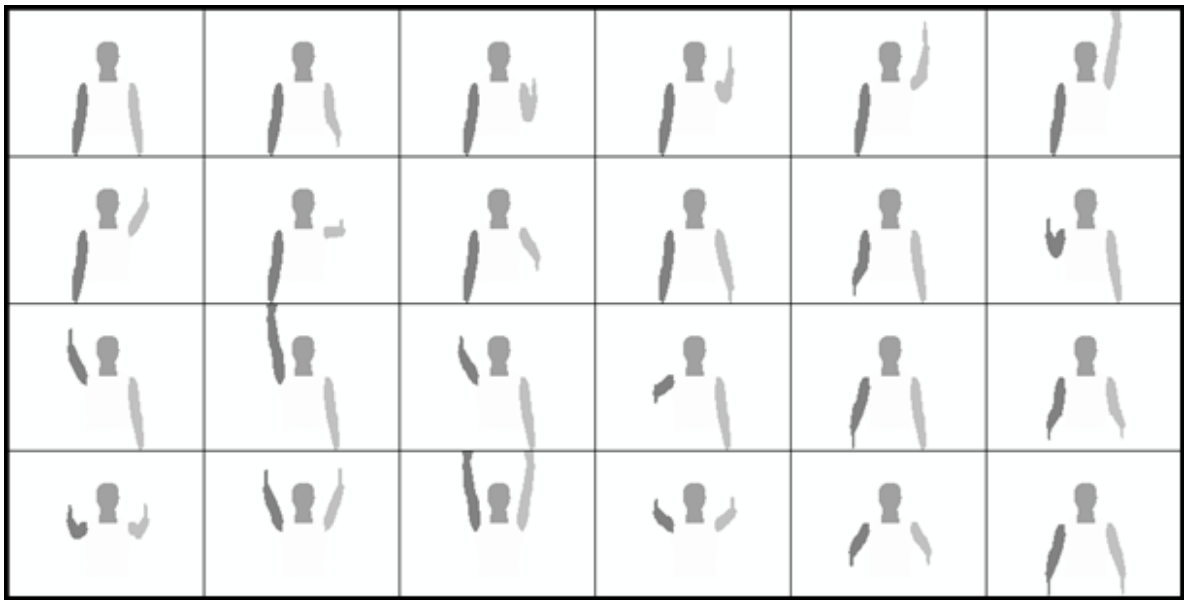

Figura 4-3: quadros do quinto vídeo.

A segunda seqüência realiza a flexão dos cotovelos, movimentando os antebraços de forma paralela ao tronco, esta seqüência contém 84 quadros. São mostrados respectivamente nas Figuras 4-4 e 4-5 o segundo e o sexto vídeos.

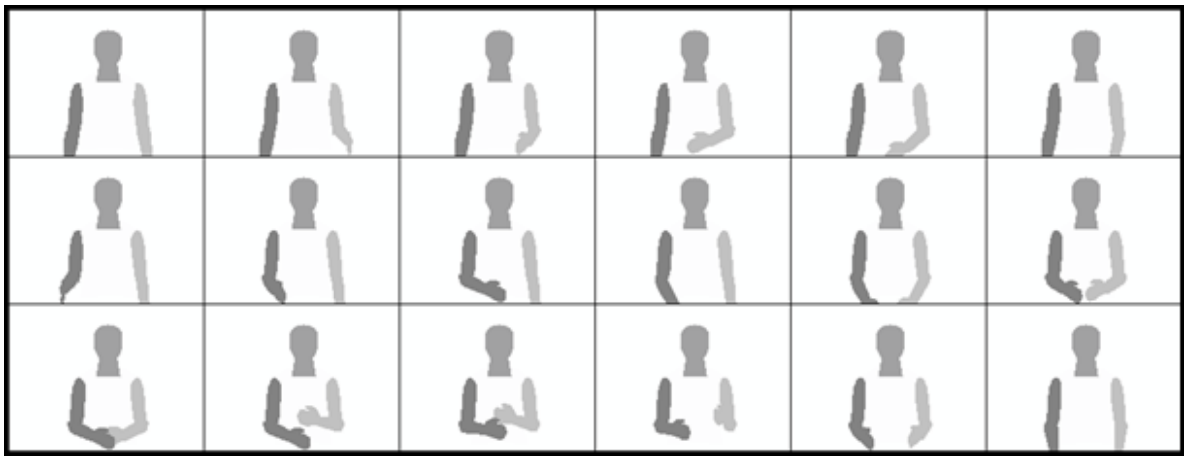

Figura 4-4: quadros do segundo vídeo.

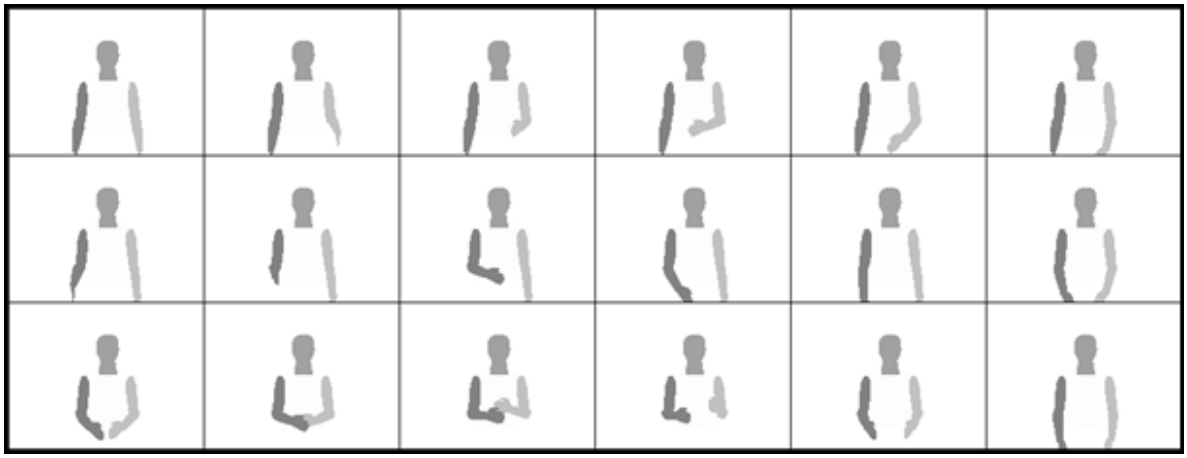

Figura 4-5: quadros do sexto vídeo.

Na terceira seqüência, é feito o movimento de abdução dos braços, contendo 66 quadros. No terceiro vídeo, (Figura 4-6) partes dos braços não são visualizados durante o movimento, por isso é utilizado o sétimo vídeo (Figura 4-7).

Como no terceiro vídeo o modelo está próximo da visão, durante todo o vídeo as mãos não estão sendo visualizadas por completo.

A criação de vídeos mais afastados se faz necessário devido a problemas que podem ser causados pela não visualização completa dos membros explicados nas seções 2.11.3 e 2.11.4.

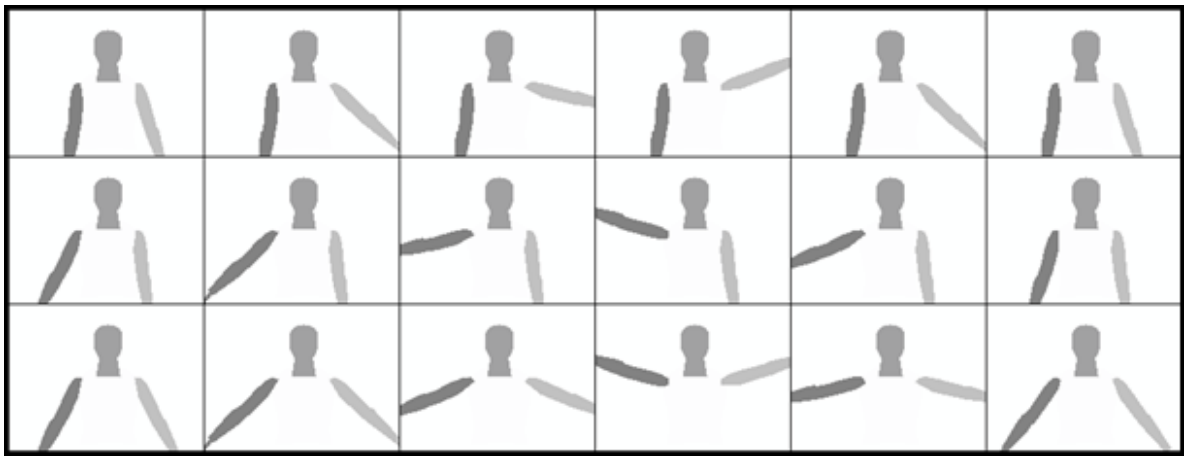

Figura 4-6: quadros do terceiro vídeo.

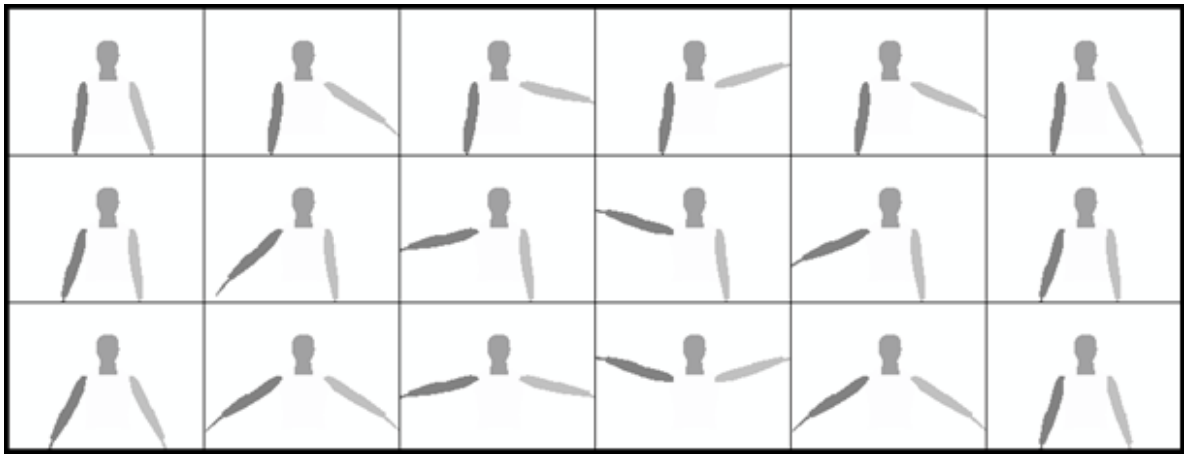

Figura 4-7: quadros do sétimo vídeo.

Os movimentos mais complexos em um sistema de aquisição de gestos estão presentes na quarta seqüência, contendo 142 quadros, que inclui a oclusão dos membros. O quarto e oitavo vídeos são mostrados, respectivamente, nas Figuras 4-8 e 4-9.

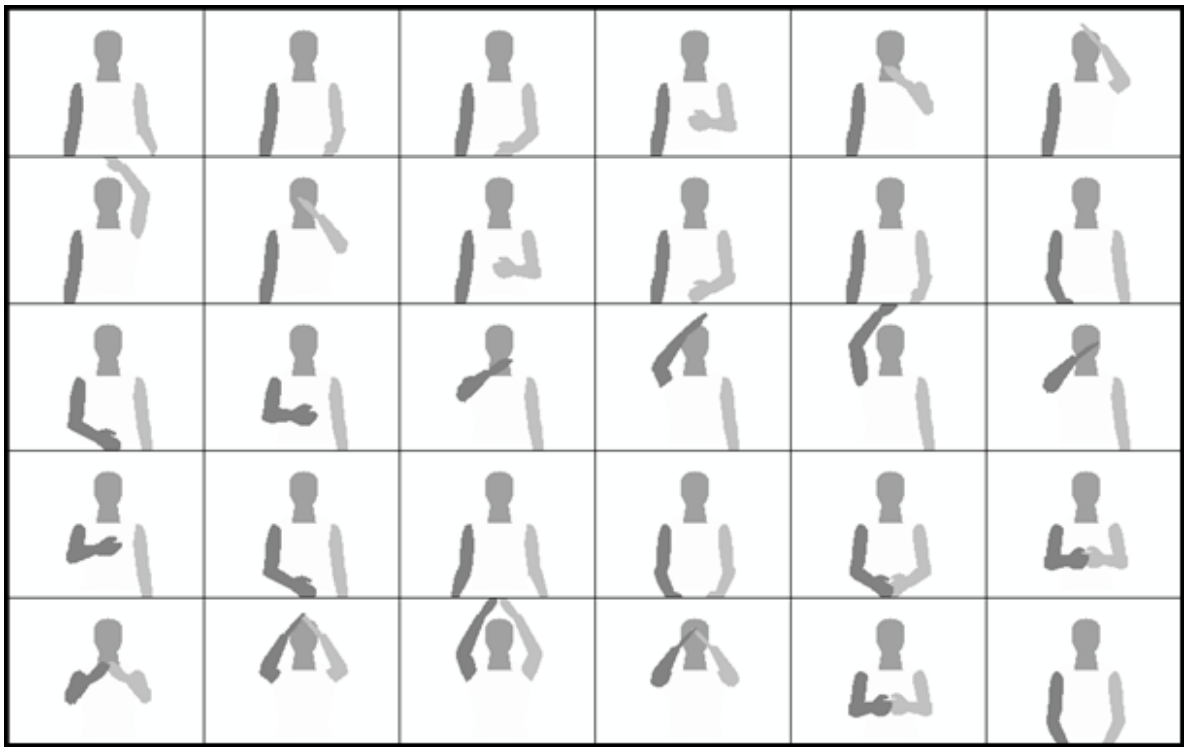

Figura 4-8: quadros do quarto vídeo.

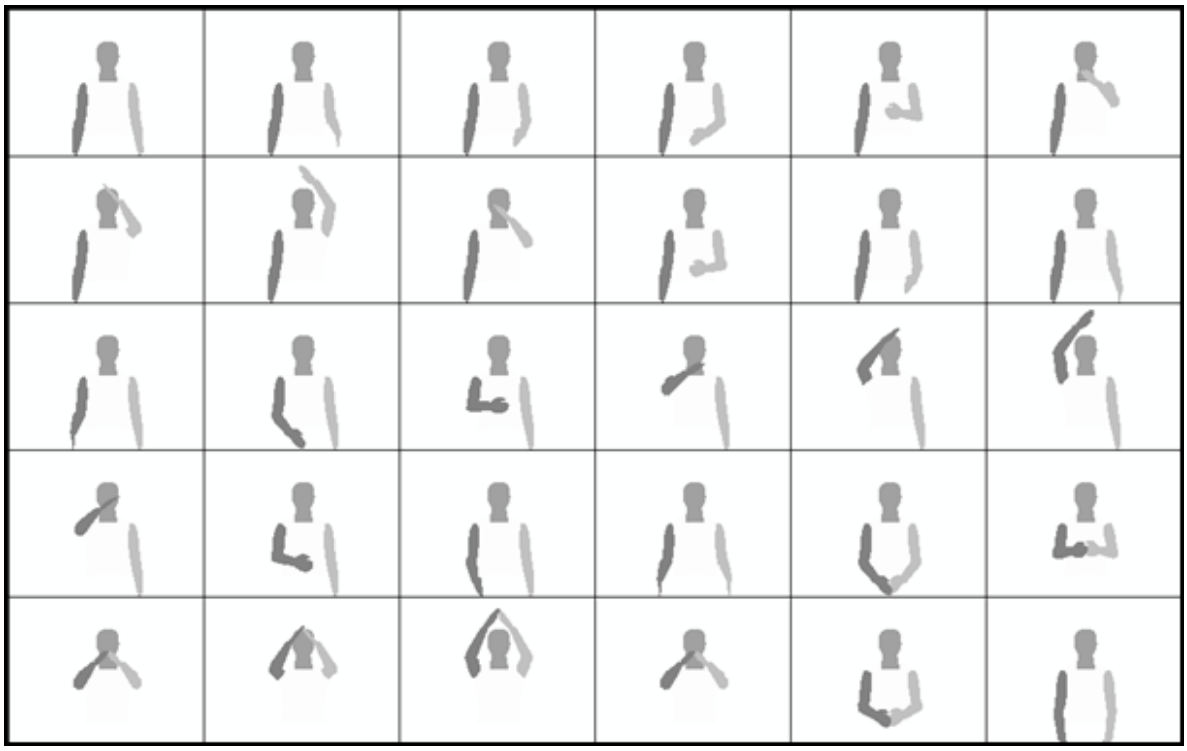

Figura 4-9: quadros do oitavo vídeo.

4.2 Conjunto de configurações do sistema

Devido à existência de diversas técnicas desenvolvidas, a análise dos resultados foi realizada de forma que pudesse ser analisado o conjunto de todas as configurações possíveis.

Totalizam 1.024 diferentes possíveis configurações do sistema, pois existem 5 métricas, 8 vídeos e 5 técnicas que podem ou não ser utilizadas. Todas as opções somariam ($5 \times 8 \times 2^5$) = 1.280, contudo algumas técnicas não podem ser utilizadas em conjunto, quais sejam: as métricas baseadas na diferença em *pixel* entre as imagens, e a técnica da classificação das regiões de pele da imagem, capturada para identificação dos membros correspondentes.

As métricas “Distância entre os centróides dos contornos” e “Distância entre os centróides dos contornos com divisão da imagem” não obtiveram resultado satisfatório, ambas foram ignoradas e não entram na comparação das métricas.

As cinco métricas comparadas são as seguintes:

- Superfície de não recobrimento;
- Superfície de não recobrimento com divisão da imagem;
- Superfície de não recobrimento com divisão da imagem usando a função quadrática de avaliação;
- Diferença em *pixel* entre as imagens;
- Diferença em *pixel* entre as imagens com sub-amostragem.

As cinco técnicas que podem ou não ser utilizadas são listadas a seguir:

- Restrições biomecânicas estáticas e dinâmicas das articulações dos membros;
- Penalização de posturas irreais;
- Iniciação do *simplex* com parâmetros aleatórios;
- Classificação das regiões de pele da imagem capturada para identificação dos membros correspondentes;
- Predição de posturas através de Banco de Dados.

4.3 Cálculo do erro e resultado de desempenho

Ao final de cada processo de otimização, uma postura é encontrada, esta representa a postura identificada pelo sistema como similar à do quadro do vídeo, e, assim, deseja-se que a diferença entre estas posturas seja nula. Cada postura possui oito graus de liberdade, representando as articulações dos membros.

A cada quadro do vídeo é calculado o erro quadrático médio do quadro, que corresponde à raiz quadrada da diferença ao quadrado de cada grau de liberdade da postura encontrada, e seu respectivo valor, utilizado para geração do vídeo. Os valores de referência são aqueles usados para a produção da seqüência de vídeo.

Como mostrado na seção 2.11, que trata sobre as ambigüidades nas análises das posturas, é melhor um erro estar disperso, proporcionalmente, em todos os graus de liberdade do que concentrado em apenas alguns. Por isso, calcula-se o erro quadrático médio dos graus de liberdade, pois este valoriza a dispersão dos valores, resultando em erros maiores do que a média aritmética.

Dessa maneira, a cada quadro do vídeo, é calculado o erro quadrático médio dos graus de liberdades entre as duas posturas: aquela encontrada pelo sistema e a que gerou a postura de referência na seqüência do vídeo. É mostrado na Equação (4-1) como o cálculo é realizado. Utiliza-se a sigla em inglês RMSE de *root mean square error*.

$$RMSE(pr, pe) = \sqrt{\sum_{i=1}^{n} \frac{(pr_i - pe_i)^2}{n}} \qquad (4\text{-}1)$$

Na Equação (4-1), "pr" e "pe" representam as posturas a serem comparadas, "pr_i" é o grau de liberdade de índice "i" da postura de referência, "pe_i" é o grau de liberdade de índice "i" da postura encontrada, "n" é a quantidade de graus de liberdade das posturas que, neste caso, são oito.

No final do vídeo, é calculada a média de todos os erros quadráticos médios de cada quadro do vídeo. Na Equação (4-2) é mostrado como esse cálculo é

realizado. Usa-se a sigla em inglês MRMSE de *mean of root mean square error*, para simplificar.

$$MRMSE = \frac{\sum_{j=0}^{m} RMSE_j}{m} \tag{4-2}$$

Na Equação (4-2), $RMSE_j$ representa a raiz quadrada do erro quadrático médio do quadro do vídeo de índice "j", e *"m"* é o número de quadros do vídeo.

Assim como na comparação entre os graus de liberdade, é importante considerar um erro elevado para um ou poucos quadros, mesmo que a média entre todos os quadros da seqüência seja pequena, o que indica a ocorrência de mínimos locais, sendo provável sua propagação para outros quadros.

Por esse motivo, além da média do RMSE, calcula-se o desvio padrão. Usa-se a sigla em inglês SDRMSE de *standard deviation of root mean square error*.

$$SDRMSE = \sqrt{\frac{\sum_{j=0}^{m}\left(RMSE_j\right)^2}{m} - MRMSE^2} \tag{4-3}$$

Na Equação (4-3), $RMSE_j$ corresponde ao erro quadrático médio do quadro do vídeo de índice "j", *"m"* é o número de quadros do vídeo, e $MRMSE$, à média do erro quadrático médio, calculado na Equação (4-2). É mostrado na Figura 4-10 como estes erros são calculados.

Como todos os dados utilizados no cálculo são em graus, o cálculo do MRMSE e do SDRMSE mantém a unidade de ângulo em graus.

Para cálculo do resultado de desempenho, é medido o tempo total que o sistema necessitou para o processo, calculando-se o número de quadros por segundo. Utiliza-se a sigla em inglês FPS (*frames per second*) para referência a essa medida, correspondendo à informação que indica se o sistema pode ser executado em tempo real. Neste trabalho, busca-se, idealmente, um processamento em torno de 25 fps ou mais, visto que as *webcams* utilizadas e as comumente encontradas no mercado em geral não conseguem capturar mais que 30 fps. Entretanto, para algumas aplicações, uma taxa mínima de 8 fps pode ser aceita.

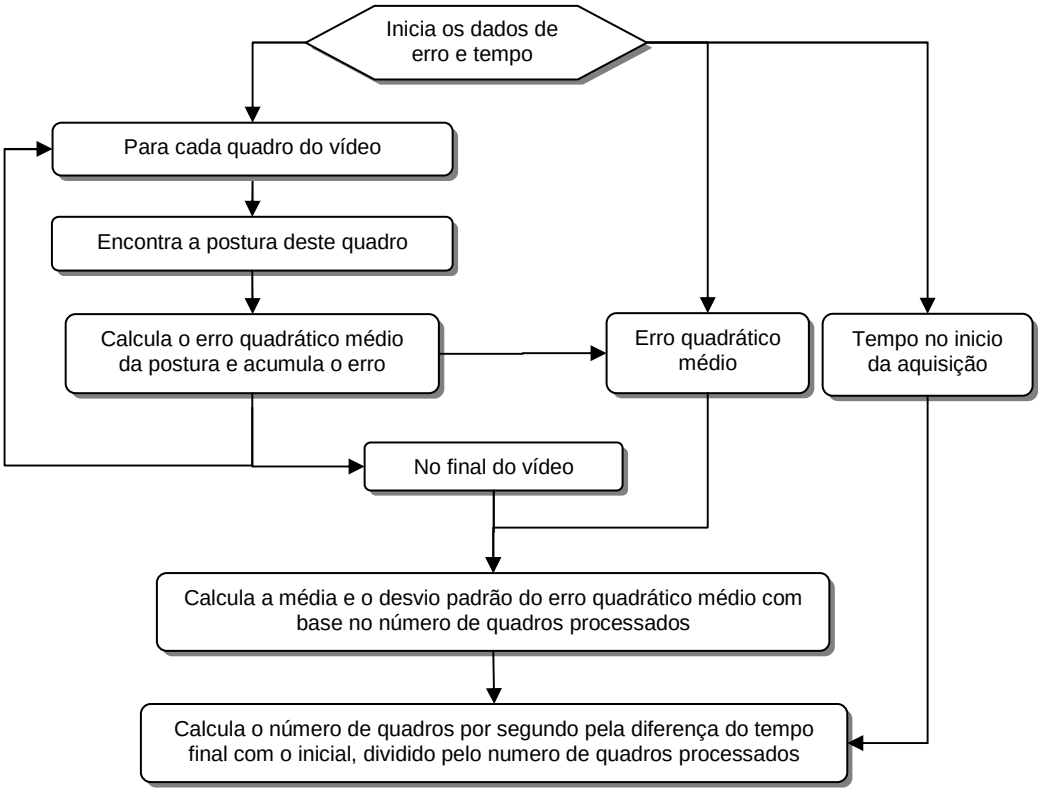

Figura 4-10: processo de cálculo dos erros.

4.4 Coleta dos dados dos resultados

Para comparação qualitativa dos dados, o programa deve registrar os resultados em arquivo, para efetuar o cálculo do erro, mas, tal procedimento, reduz consideravelmente o desempenho do sistema. Assim, foram realizadas execuções distintas para captura dos parâmetros de erro e dos parâmetros de desempenho.

Conforme descrito na seção 2.8, o algoritmo de otimização é baseado no *Downhill Simplex* e, como este algoritmo é determinístico, as execuções que não possuem iniciação aleatória, sempre são iguais, independente do momento em que o sistema foi executado.

Uma amostra de vinte configurações diferentes foi escolhida, sendo comparadas três execuções distintas, confirmando que a implementação desenvolvida é determinística. Esta análise foi realizada diversas vezes durante o desenvolvimento do trabalho, pois erros no algoritmo podem ocasionar uma alteração em sua execução, e esta análise comprova a corretude do algoritmo.

Portanto, para verificar o resultado das métricas que não utilizam iniciação aleatória, os dados foram coletados somente uma vez. Por outro lado, para as execuções que usam a iniciação do *simplex*, usando dados aleatórios, o sistema foi executado cinco vezes, sendo descartados o maior e menor dado coletado, realizando a média dos três dados intermediários.

4.5 Configuração do equipamento

Os resultados foram gerados através de um computador com as seguintes configurações de *hardware* e *software*:

- Processador Intel Core2 Duo E6550 @ 2.33 GHz;
- Placa mãe Intel D945GCNL – FSB 1066;
- Memória Kingston 2GB DDR2 667MHz Dual Channel;
- Placa gráfica Zogis NVidia GeForce 8500 GT – GPU 450MHz BS 400MHz;
- Microsoft Windows XP SP3;
- Driver NVidia GeForce Release 181.20.

4.6 Explicação dos gráficos e comparação com o sistema base

Nas subseções seguintes serão mostrados vários gráficos com os dados das experimentações. Aqui é apresentada a maneira como interpretar os dados através de um gráfico fictício, mostrado na Figura 4-11.

Na Figura 4-11, cada par de linhas da mesma cor, uma contínua e uma tracejada, comparam duas técnicas em relação ao mesmo dado estatístico, que pode ser o MRMSE, o SDRMSE ou o FPS. Para facilitar a visualização, além das cores, símbolos representam os dados: o quadrado representa o MRMSE, o triângulo representa o SDRMSE, e o losango representa o FPS. Cada linha vertical representa uma das oito seqüências de vídeo apresentadas na seção 4.1.

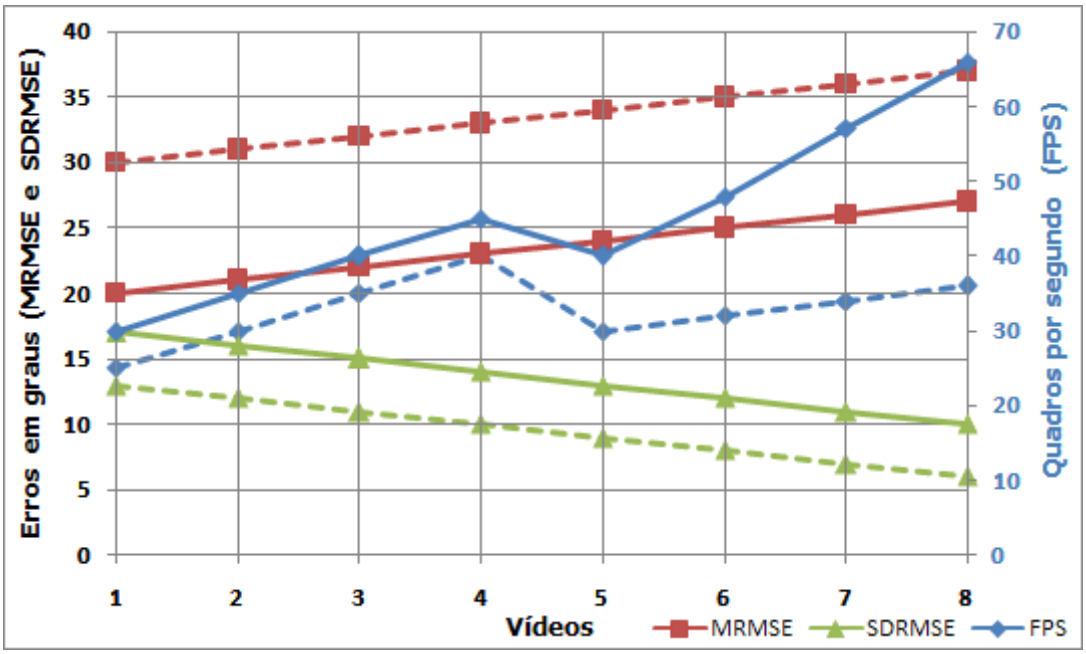

Figura 4-11: gráfico fictício comparando a técnica Y (linha tracejada) com a técnica X (linha contínua).

Ainda, com relação ao gráfico da Figura 4-11, a escala ao lado esquerdo representa o erro em graus para as medidas MRMSE e SDRMSE, que, no gráfico exemplificado, variam entre 0 e 40º. Ao lado direito, a escala representa o número de quadros por segundo obtidos para cada seqüência, variando, no exemplo, de 0 a 70 fps. Dessa maneira, ao analisar o MRMSE e o SDRMSE, deve-se referenciar à escala da esquerda, ao passo que, no caso do FPS, deve-se olhar para a escala da direita.

Por exemplo, analisando a Figura 4-11, pode-se ver que a técnica Y é pior do que a X em todos os MRMSE, porém é melhor em todos os SDRMSE. Por exemplo, no sétimo vídeo, o MRMSE da técnica X é de 26º, enquanto o da técnica Y é 36º, representando um aumento de 10 graus, ou 28% do erro. No SDRMSE houve uma redução de 11º da técnica X para 7º da técnica Y, ou seja, houve uma melhora no valor do desvio padrão do erro quadrático médio em 57%.

Ainda como parte da análise, observa-se que do quarto para o quinto vídeo, justamente quando há a troca dos tipos de vídeo em relação à distância da câmera, houve uma redução do desempenho (FPS), invertendo a tendência de subida, e que, depois desta redução, os FPS da técnica X aumentaram proporcionalmente mais do que a técnica anterior Y.

Ao final de cada gráfico, será exibido um resumo dos dados do respectivo gráfico, em valores numéricos e percentuais, como se exemplifica abaixo, ainda com os dados fictícios da Figura 5-11.

- A média da média do erro quadrático médio (MRMSE) passou de 23,5 para 33,5, indicando uma piora de 30%, sendo representada, por isso, com o símbolo negativo (-30%);
- A média do desvio padrão do erro quadrático médio (SDRMSE) passou de 13,5 para 9,5, reduzindo em 42%. Para enfatizar a ocorrência de uma melhora, usando o símbolo ”+” (+42%);
- A média do número de quadros por segundo (FPS) passou de 32,8 para 45,1, melhorando em +38%.

Seguindo este padrão, os valores serão sempre apresentados como no resumo abaixo.

Resumo:

- MRMSE 23,5 ► 33,5 = – 30%;
- SDRMSE 13,5 ► 9,5 = + 42%;
- FPS 32,8 ► 45,1 = + 38%.

4.6.1 Sistema base de comparação

As próximas subseções realizam comparações das técnicas e das métricas. Estas comparações geralmente tomam como parâmetro de comparação o sistema na configuração do protótipo de Soares [1] chamado aqui de “sistema base”.

O sistema base é a primeira configuração do sistema de aquisição, que representa a métrica da superfície de não recobrimento, **sem a utilização de nenhuma das técnicas apresentadas** neste trabalho.

O gráfico mostrado na Figura 4-12 é o resultado do sistema base de comparação.

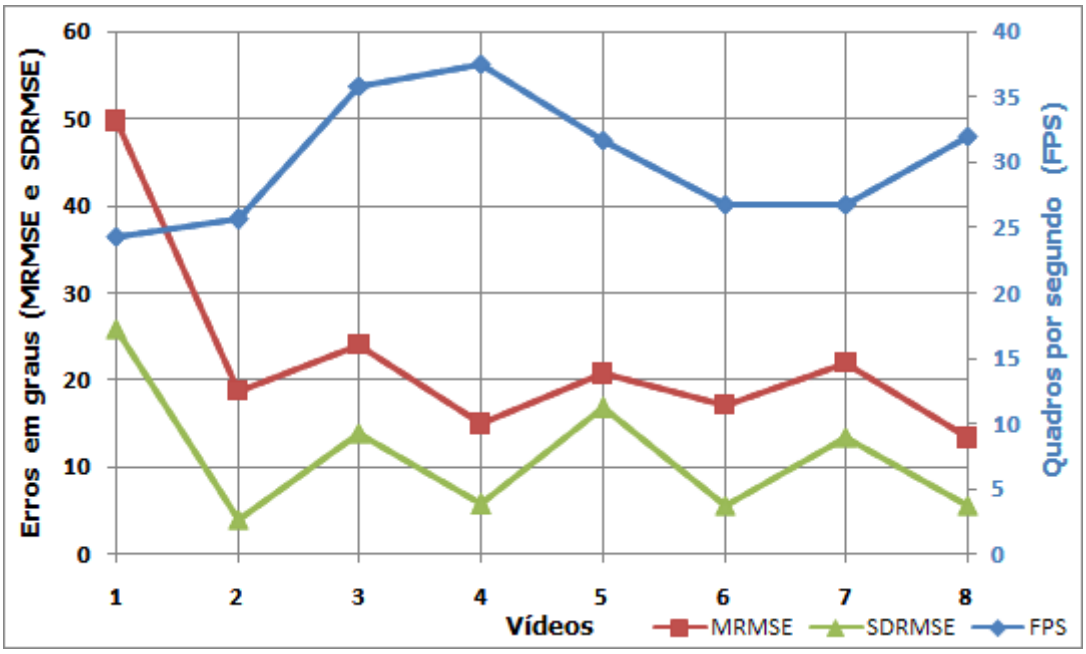

Figura 4-12: gráfico do resultado do sistema base que é a métrica da superfície de não recobrimento sem utilizar as técnicas desenvolvidas.

Observa-se, no vídeo 1, o elevado erro, tanto no MRMSE, quanto no SDRMSE, demonstrando que em vários quadros deste vídeo houve a incidência de mínimos locais. Este vídeo, em particular, mostra resultados muito ruins, indicando que, nos quadros onde ocorreram os mínimos locais, a postura encontrada é muito diferente daquela que originou a imagem.

Nos demais vídeos (de 2 a 8), os erros se localizam em torno dos 20 graus, com um desvio padrão médio de 10 graus entre quadros. Embora ainda elevados, tais erros são considerados razoáveis, em alguns casos.

Por exemplo, os erros encontrados no quarto vídeo são admissíveis: MRMSE de 16 graus e SDRMSE de 6 graus. Supondo que os dados sigam uma distribuição normal, em que 95% dos valores se encontram a uma distância duas vezes do desvio padrão, ter-se-á uma grande probabilidade de o maior RMSE ser algo em torno dos 28 graus. Erros desta magnitude, estando bem distribuídos entre os graus de liberdade são visíveis, mas ocorrendo em poucos quadros de uma comunicação que use a restituição de gestos identificados, seriam pouco perceptíveis visualmente.

4.7 Resultado do estudo das placas gráficas para melhoria do desempenho

O sistema base transfere 24 bits por *pixel* entre a memória principal e a placa gráfica, usando 8 bits para cada canal RGB. Conseguiu-se transferir apenas 8 bits para os três canais, usando um formato interno especial de imagem.

Os resultados coletados têm como objetivo comparar os formatos internos utilizados na aplicação, usando imagens de tamanhos diferentes, em diversas placas gráficas.

Na Tabela 4-1 é mostrado o resultado de desempenho em imagens transferidas entre a placa gráfica e a memória principal do computador, apresentando-se, para cada placa, o número de imagens transferidas por segundo, usando a configuração de 24 bits e de 8 bits, além do ganho percentual quando se passa da primeira para a segunda configuração. São comparadas diversas placas gráficas da série GeForce da NVidia®, utilizando diferentes tamanhos de imagem: 160x120, 320x240, 128x128 e 512x512 (largura X altura em *pixels*).

Tabela 4-1: desempenho das placas gráficas em imagens transferidas por segundo.

		GeForce					
		5200	**5700 LE**	**6200**	**7300LE**	**7600GT**	**8500GT**
512x512	**24bits**	144	144	464	311	418	1.092
	8bits	257	268	668	542	597	1.322
	%	**78%**	**86%**	**44%**	**74%**	**43%**	**21%**
128x128	**24bits**	1.525	1.601	2.973	1.997	2.950	10.648
	8bits	2.138	2.209	3.422	2.675	3.273	13.171
	%	**40%**	**38%**	**15%**	**34%**	**11%**	**24%**
320x240	**24bits**	512	562	2.770	1.843	2.807	3.292
	8bits	1.068	1.123	3.560	2.601	3.541	3.940
	%	**109%**	**100%**	**29%**	**41%**	**26%**	**20%**
160x120	**24bits**	2.134	2.138	8.889	5.853	9.355	9.910
	8bits	4.004	4.004	11.642	9.235	11.375	12.060
	%	**88%**	**87%**	**31%**	**58%**	**22%**	**22%**

Reduzindo de GL_RGB8 que ocupa 8 bits para cada canal, ou seja 24 bits ao todo, para GL_R3_G3_B2, que são 8 bits ao todo, não se verificou proporcionalmente, o aumento do número de imagens transferidas por segundo. O ganho é maior quando as imagens são maiores. Entretanto, houve uma melhoria significativa (de 22%) desta transferência nas placas gráficas utilizadas atualmente.

4.8 Resultado das restrições biomecânicas dinâmicas das articulações dos membros

As restrições impediram que posturas impossíveis ou ergonomicamente desconfortáveis fossem utilizadas, diminuindo-se, assim, o espaço de busca e a possibilidade de mínimos locais e ambigüidades.

Na Figura 4-13 é mostrado o gráfico que compara os resultados do sistema base com e sem o uso das restrições biomecânicas dinâmicas. A explicação de como interpretar este gráfico encontra-se na seção 4.6.

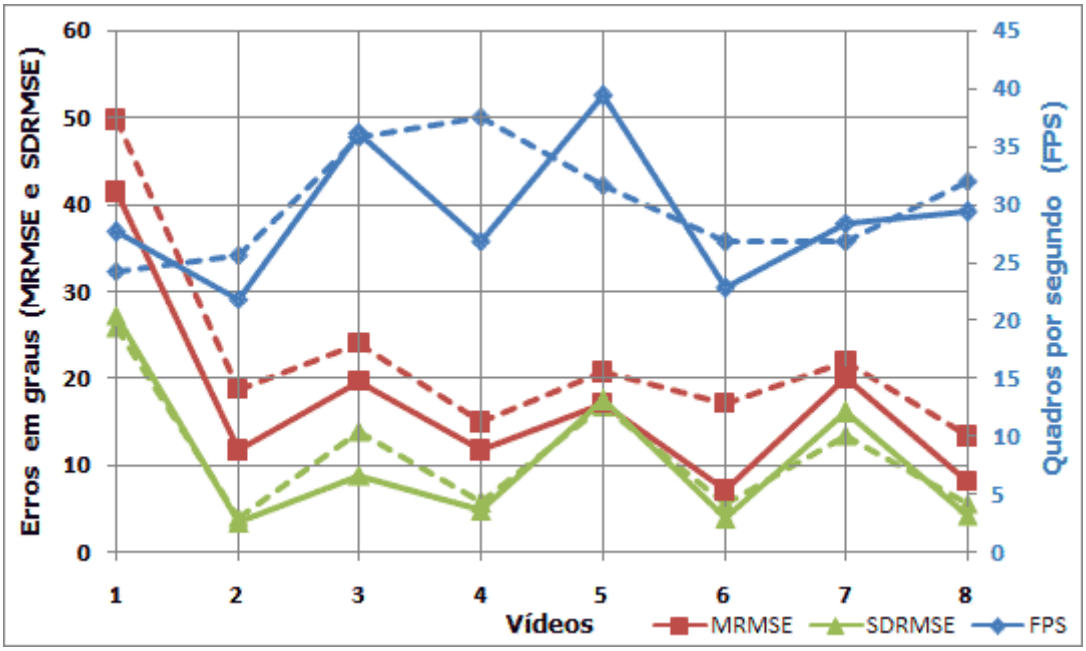

Figura 4-13: gráfico comparativo usando as restrições biomecânicas dinâmicas das articulações dos membros (linha continua) com o sistema base (linha tracejada).

Resumo: (vide seção 4.6)

- MRMSE 22,6 ► 17,1 = + 24%;
- SDRMSE 11,4 ► 10,8 = + 5%;
- FPS 30,1 ► 29,0 = – 3%.

Os resultados mostram uma significante melhoria no MRMSE e uma pequena melhoria no SDRMSE. Registrou-se, entretanto, a perda de qualidade no SDRMSE no primeiro e no sétimo vídeo.

As restrições adicionaram um calculo a mais no processo de avaliação de similaridade, causando uma degradação no desempenho. Mesmo assim, a melhoria da qualidade da identificação permite a convergência mais rápida do sistema, deixando-o mais eficiente em alguns casos. Por exemplo, nos vídeos 1, 3, 5 e 7, a eficiência da convergência superou o custo do cálculo adicionado ao processo, elevando a taxa de FPS.

Esta técnica, sozinha, permitiu uma redução de 24% no MRMSE, valor bastante representativo, demonstrando um bom ganho de qualidade da aquisição do sistema, sem prejudicar o SDRMSE.

4.9 Resultado da penalização de posturas irreais

A penalização das posturas que atingem o limite das restrições biomecânicas eleva o desempenho do sistema, pois tais posturas não são avaliadas.

Na Figura 4-14 é mostrado o gráfico comparativo dos resultados do sistema base, com e sem o uso da penalização de posturas irreais. Recorda-se aqui que a forma de interpretação deste gráfico é apresentada na seção 4.6.

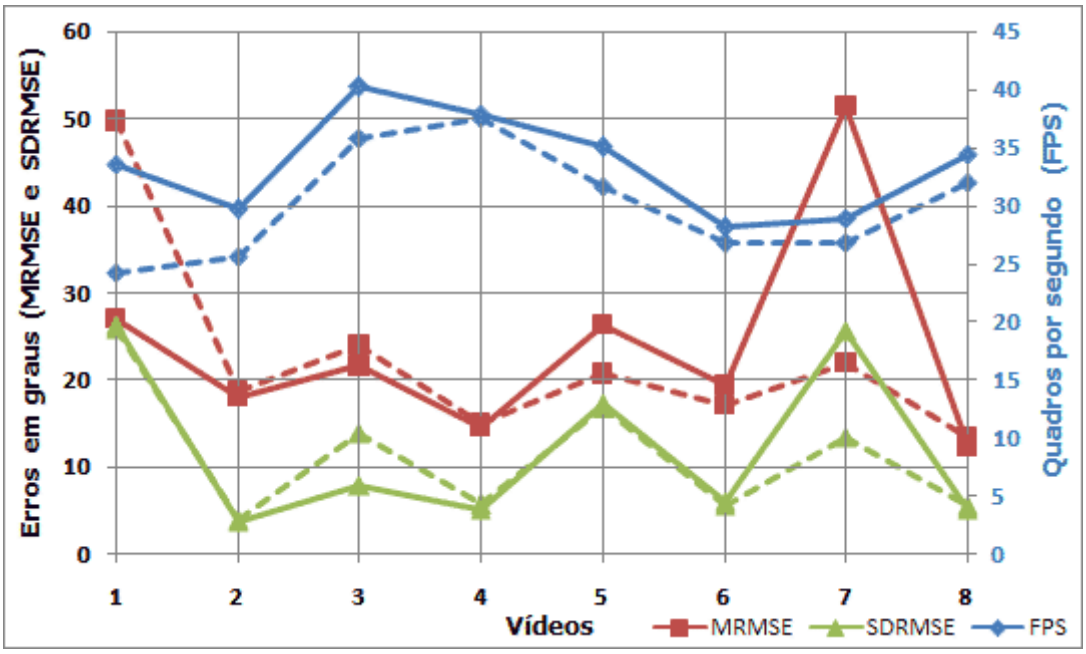

Figura 4-14: gráfico comparativo usando a penalização de posturas irreais (linha continua) com o sistema base (linha tracejada).

Resumo: (vide seção 4.6)

- MRMSE 22,6 ► 23,9 = – 6%;
- SDRMSE 11,4 ► 12,2 = – 7%;
- FPS 30,1 ► 33,5 = + 11%.

Houve um ganho de desempenho pelo descarte das posturas irreais, mas registrou-se uma degradação na qualidade da aquisição. Quando posturas reais se aproximam de seus limites, esta técnica acaba provocando a convergência do sistema para um mínimo local.

Observa-se uma inversão no gráfico. Nos vídeos 1 a 4, em que o modelo está mais próximo da câmera, houve melhoria tanto no MRMSE, quanto no SDRMSE. Já nos vídeos 5 a 8, com o modelo afastado da câmera, os resultados foram piores.

O resultado geral foi pior, pois o ganho obtido nos quatro primeiros vídeos foi menor do que a degradação dos resultados nos outros. Os resultados extremos foram registrados para os vídeos 1 e 7. A análise indica que o vídeo 7 piorou mais do que o ganho no vídeo 1. O MRMSE do vídeo 1 foi de 50 para 27 graus, enquanto o do vídeo 7 foi de 22 para 51. Registra-se também que não houve melhora no

SDRMSE no primeiro vídeo, enquanto no sétimo houve uma degradação de 10 graus.

Por isso, o resumo foi analisado separadamente para os vídeos de 1 a 4 e de 5 a 8.

Resumo do vídeo 1 ao 4, os vídeos mais próximos da visão:

- MRMSE 26,9 ► 20,3 = + 24%;
- SDRMSE 12,4 ► 10,8 = + 13%;
- FPS 30,8 ► 35,4 = + 15%.

Resumo do vídeo 5 ao 8, os vídeos mais afastados da visão:

- MRMSE 18,3 ► 27,4 = – 50%;
- SDRMSE 10,4 ► 13,5 = – 30%;
- FPS 29,3 ► 31,6 = + 8%.

Se considerarmos somente os quatro primeiros vídeos, esta técnica apresenta um ganho de qualidade ao sistema. Um ganho em erro de 24%, em conjunto com um ganho de 15% no desempenho, é um resultado expressivo, mas um aumento de 50% no erro para outros tipos de vídeo impossibilita seu uso no sistema em qualquer situação.

Conclui-se, assim, que esta técnica é dependente do posicionamento da câmera em relação ao ator.

4.10 Resultado da iniciação do *simplex* com parâmetros aleatórios

A adição de uma etapa de aleatoriedade na iniciação do *simplex* no algoritmo de otimização do sistema diminui a incidência de mínimos locais pela expansão do espaço de busca, causando uma enorme redução no MRMSE e, principalmente, no SDRMSE.

Na Figura 4-15 é mostrado o gráfico comparativo dos resultados do sistema base, com e sem o uso da iniciação aleatória.

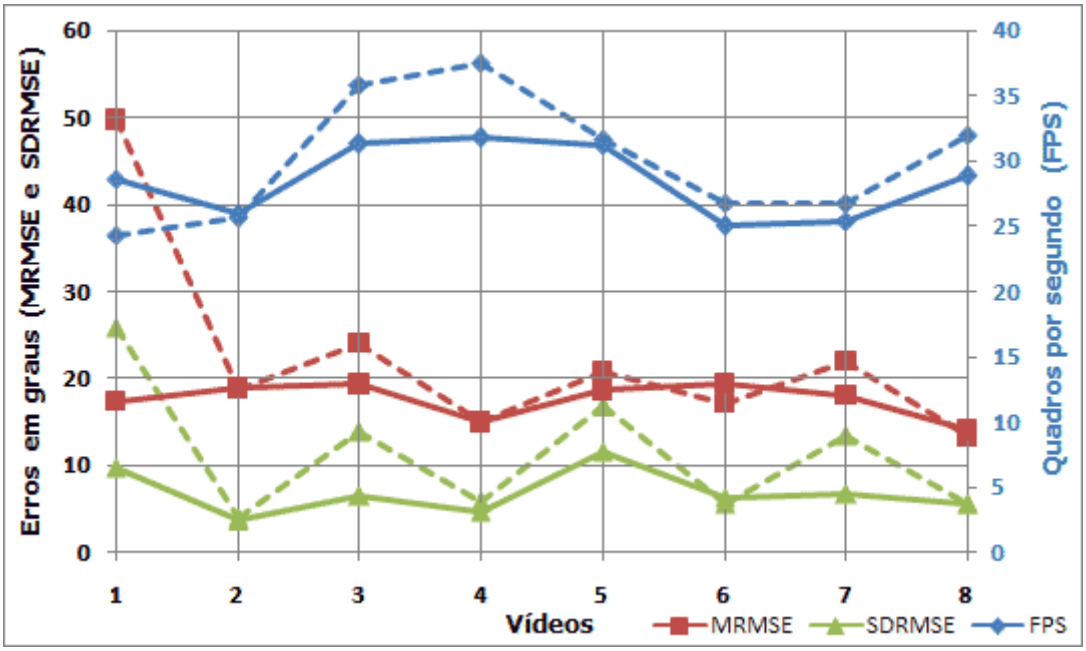

Figura 4-15: gráfico comparativo usando a iniciação do *simplex* com parâmetros aleatórios (linha continua) com o sistema base (linha tracejada).

Resumo:

- MRMSE 22,6 ► 17,6 = + 22%;
- SDRMSE 11,4 ► 6,9 = + 40%;
- FPS 30,1 ► 28,5 = – 5%.

Somente no vídeo 6 foi registrada uma piora no resultado. Uma melhora de 40% no SDRMSE, reduzindo seu valor para 6,9 graus, mostra uma grande estabilidade entre os quadros, de forma que mesmo quando ocorrem mínimos locais em quadros intermediários, estes não se propagam.

Por outro lado, registra-se uma leve redução no desempenho, pois a aleatoriedade pode retardar a convergência do DHS para uma etapa posterior.

Sua grande vantagem, que pode ser observada pelo gráfico, é a estabilidade. Os vídeos onde os erros eram mais elevados obtiveram resultados próximos aos que não apresentaram erros muito elevados, suavizando as linhas de erros.

Em geral, esta técnica tende a melhorar os resultados. Quando isto não ocorre, dificilmente ela piora os erros em níveis elevados, de maneira que prejudique significativamente a aquisição das posturas. Um exemplo é mostrado pelo vídeo 6, cuja piora registrada foi de apenas 2 graus no MRMSE e 1 grau no SDRMSE.

Isso torna o uso da técnica de iniciação aleatória preferível às demais técnicas de iniciação do algoritmo DHS.

4.11 Resultado da classificação das regiões de pele da imagem capturada para identificação dos membros correspondentes

Foi possível identificar os membros quando não há oclusão na imagem. Em todos os vídeos, o algoritmo identificou os braços de forma correta, como exemplificado na Figura 4-16. Além disso, mesmo após uma oclusão, a classificação foi bem sucedida devido à realimentação dos centróides de controle.

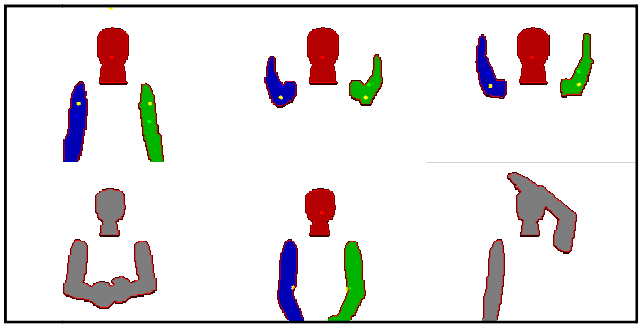

Figura 4-16: classificação das regiões de pele da imagem capturada para identificação dos membros correspondentes.

Existem casos, entretanto, onde este algoritmo falha. Realizando movimentos similares ao mostrado na seção 2.11.2, o sistema de adaptação pode, após uma oclusão dos braços, classificá-los de forma invertida.

Na Figura 4-17 é mostrado o gráfico comparativo dos resultados do sistema base, com e sem o uso da classificação das regiões de pele.

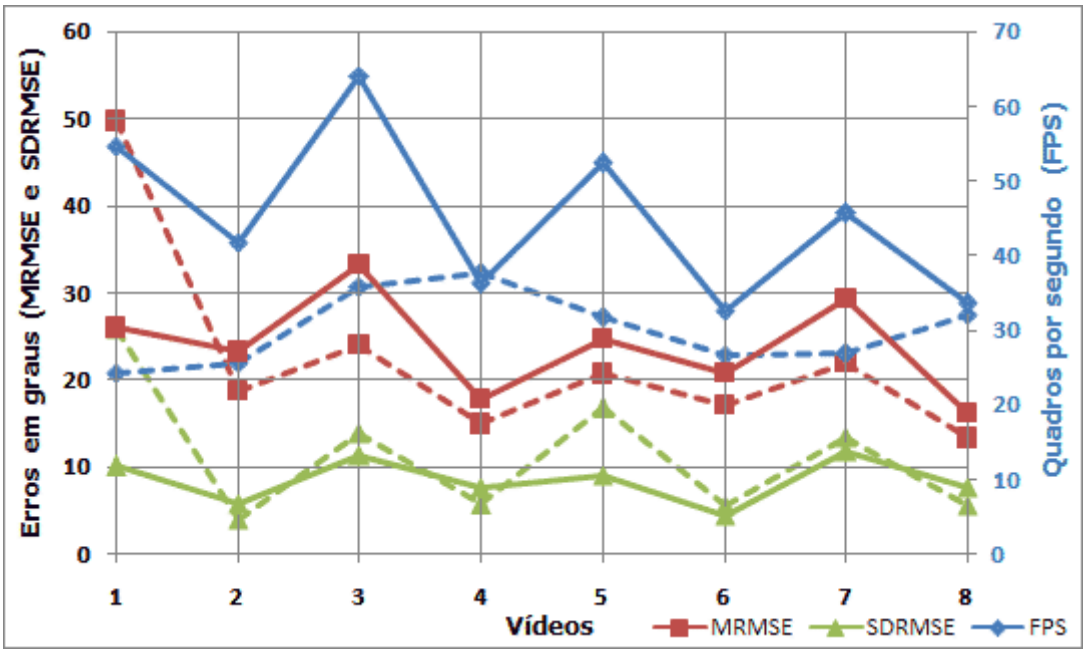

Figura 4-17: gráfico comparativo usando a classificação das regiões de pele (linha continua) com o sistema base (linha tracejada).

Resumo:

- MRMSE 22,6 ► 23,9 = – 6%;
- SDRMSE 11,4 ► 8,5 = + 25%;
- FPS 30,1 ► 45,1 = + 50%.

Os dados indicam que, mesmo com a piora do MRMSE, houve um melhora significativa do SDRMSE, mostrando que a técnica evitou que alguns quadros do vídeo tivessem grandes erros. É melhor obter erros maiores e distribuídos do que grandes erros localizados, e esta técnica melhora este aspecto.

Obteve-se um grande ganho em desempenho, causado pela diminuição dos passos para convergência do algoritmo de otimização.

Seu uso ainda pode ser aprimorado, realizando uma otimização mais direcionada. Foram realizadas experimentações executando-se passos de otimização separados para os braços, mas não foram obtidos resultados satisfatórios.

Tem-se em perspectiva a identificação ocasional de uma má adaptação de um dos membros e processamento de apenas essa região do corpo para um melhor ajuste do modelo.

4.12 Resultado da predição de posturas através de banco de dados

O uso de um banco de dados previamente treinado melhora muito todas as características do sistema.

Na Figura 4-18, é mostrado o gráfico comparativo dos resultados do sistema base, com e sem o uso da predição de posturas através de banco de dados.

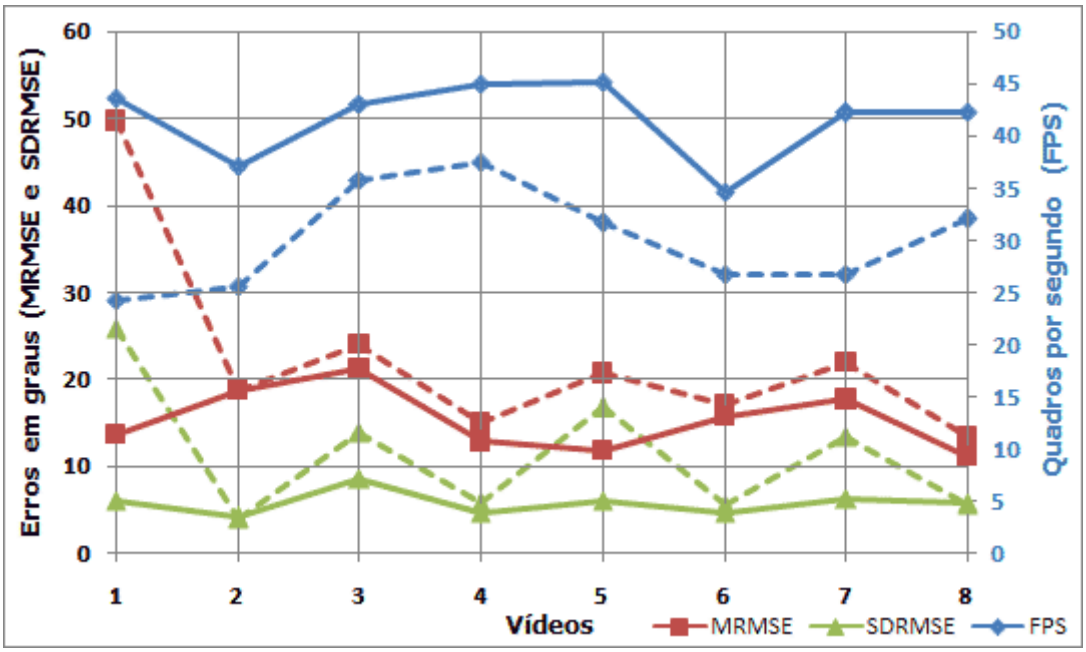

Figura 4-18: gráfico comparativo usando a predição de posturas através de banco de dados (linha continua) com o sistema base (linha tracejada).

Resumo:

- MRMSE 22,6 ► 15,4 = + 32%;
- SDRMSE 11,4 ► 5,8 = + 49%;
- FPS 30,1 ► 41,6 = + 39%.

Observa-se que o uso do banco de dados para predição de posturas melhorou expressivamente o funcionamento do sistema de aquisição em todos os aspectos. Entretanto, é importante lembrar que esta abordagem exige um treinamento prévio no ambiente de aquisição.

4.13 Resultado das técnicas aplicadas em conjunto

4.13.1 Todas as técnicas simultaneamente

Usando todas as técnicas em conjunto, podemos verificar como estas se comportam combinadas, o que não resulta, obrigatoriamente, em melhora para o sistema.

Na Figura 4-19, é mostrado o gráfico com os resultados da métrica da superfície de não recobrimento, utilizando todas as técnicas desenvolvidas em comparação ao sistema base, quais sejam:

- Restrições biomecânicas estáticas e dinâmicas das articulações dos membros;
- Penalização de posturas irreais;
- Iniciação do *simplex* com parâmetros aleatórios;
- Classificação das regiões de pele da imagem capturada para identificação dos membros correspondentes;
- Predição de posturas através de Banco de Dados.

Registrou-se piora em quatro ocasiões: o SDRMSE dos vídeos 2 e 8, e o MRMSE dos vídeos 2 e 3, com diferenças em torno de apenas 1 grau. Em casos como os vídeos 1 e 5, o ganho foi significativo para todos os dados.

O número de FPS atingido nesta configuração representa o dobro do máximo usável em um sistema em tempo real. Isto pode ser utilizado de outras formas na tentativa de reduzir ainda mais os erros encontrados, como, por exemplo, aumentar o número de passos da otimização ou o tamanho do quadro capturado pela câmera.

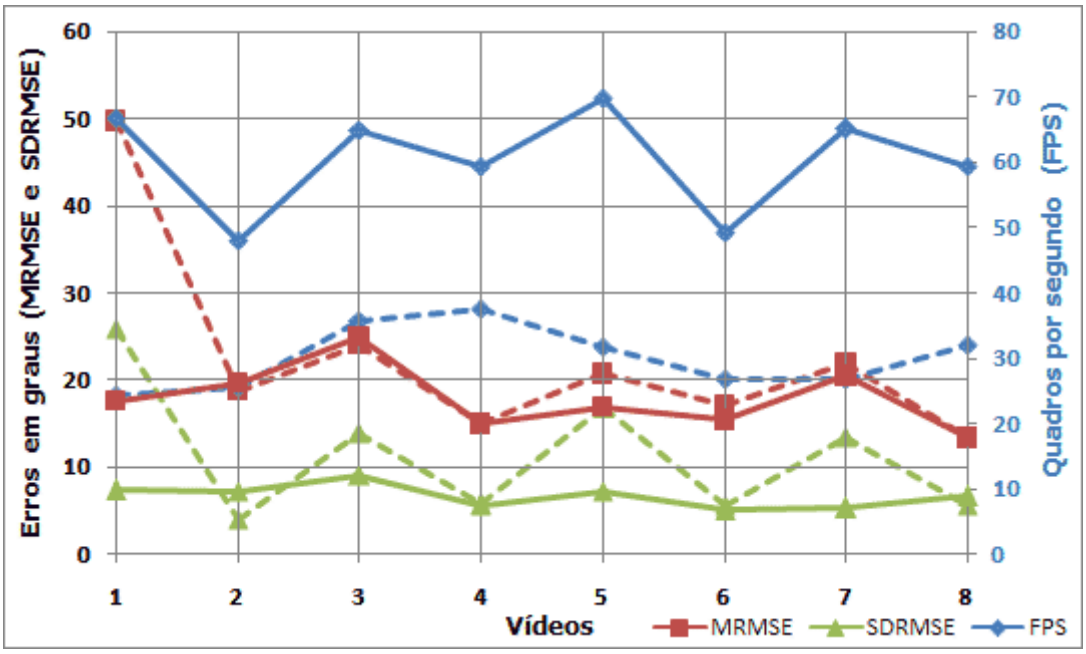

Figura 4-19: gráfico comparando a métrica da superfície de não recobrimento com todas as técnicas desenvolvidas em conjunto (linha continua) e o sistema base que usa a métrica da superfície de não recobrimento sem as técnicas desenvolvidas (linha tracejada).

Resumo:

- MRMSE 22,6 ► 17,9 = + 21%;
- SDRMSE 11,4 ► 6,7 = + 41%;
- FPS 30,1 ► 60,3 = + 101%.

4.13.2 Melhor configuração

O uso de todas as técnicas, em conjunto, não representa a melhor configuração. A combinação de algumas técnicas que apresentam bons resultados isoladamente não garante melhoria do processo de aquisição, quando combinadas.

O conjunto de técnicas que obteve o melhor resultado foi aquele em que se exclui a técnica de classificação das regiões de pele, ou seja, utilizando: restrições biomecânicas, penalização de posturas irreais, iniciação aleatória e banco de dados. Na Figura 4-20, é mostrado um gráfico com os resultados do melhor conjunto, comparando-o ao sistema base.

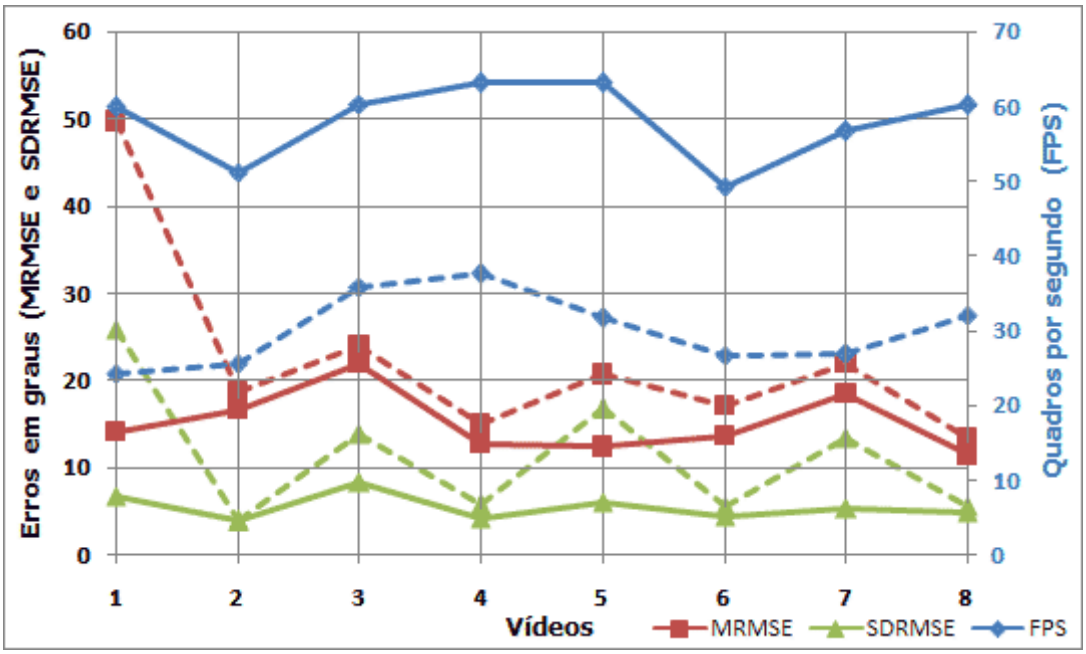

Figura 4-20: gráfico comparando o melhor conjunto de técnicas usando a métrica da superfície de não recobrimento (linha continua) e o sistema base (linha tracejada).

Resumo:

- MRMSE 22,6 ► 15,2 = + 33%;
- SDRMSE 11,4 ► 5,5 = + 51%;
- FPS 30,1 ► 58,0 = + 93%.

Esta configuração apresenta melhora em todos os dados. Há casos em que as técnicas isoladas obtiveram melhores resultados do que o apresentado por este conjunto, como se observa na Figura 4-13 em que os MRMSE dos vídeos: 2, 3, 6 e 8 são menores. Contudo, esta configuração obteve uma média melhor, tanto no MRMSE quanto no SDRMSE.

A comparação deste resultado com a técnica de predição de posturas através do uso do banco de dados apresenta detalhes interessantes que serão abordados na próxima subseção.

A comparação com a configuração da seção anterior, que utiliza todas as técnicas em conjunto, comprova que o uso da totalidade das técnicas não trouxe o melhor resultado. A configuração apresentada nesta subseção apresenta melhores resultados, mas com uma leve degradação no desempenho.

4.13.3 Comparação da melhor configuração com a predição de posturas através de banco de dados

Foi observada uma semelhança entre os gráficos da predição de posturas através de banco de dados (Figura 4-18) com o gráfico da melhor configuração (Figura 4-20). Por isso, estes são comparados nesta subseção através do gráfico mostrado na Figura 4-21.

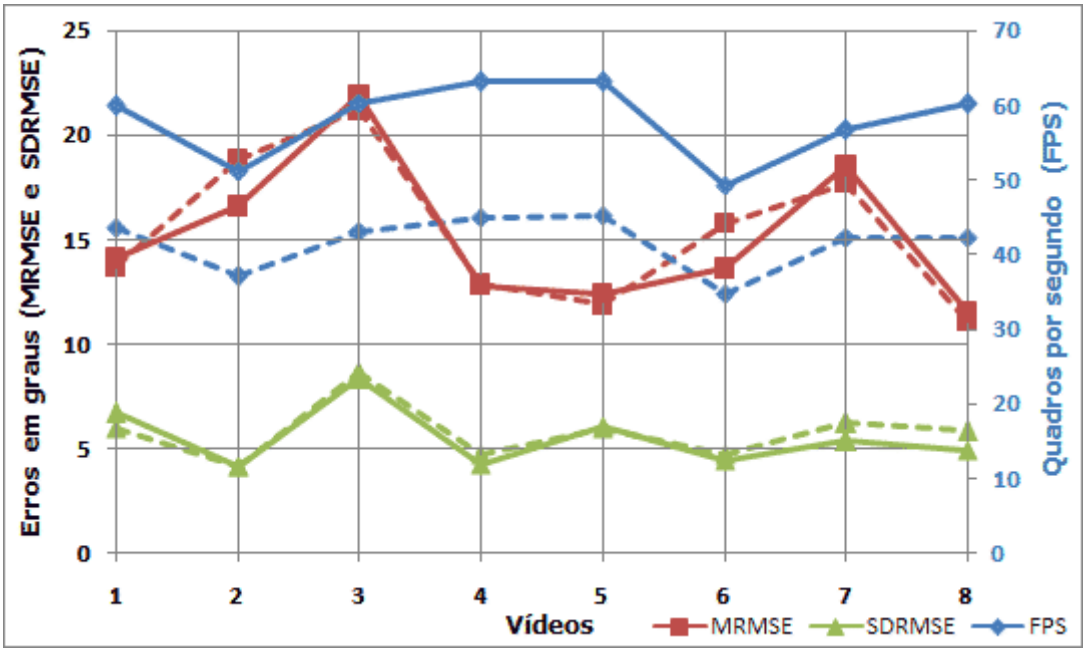

Figura 4-21: gráfico comparando a técnica da predição de posturas através de banco de dados (linha continua) e o melhor conjunto de técnicas (linha tracejada) ambos usando a métrica da superfície de não recobrimento.

Resumo:

- MRMSE 15,4 ► 15,2 = + 1%;
- SDRMSE 5,8 ► 5,5 = + 4%;
- FPS 41,6 ► 58,0 = + 39%.

Observa-se uma grande semelhança nas linhas que representam os erros, demonstrando que o bom resultado de qualidade da configuração mostrada na subseção anterior é bastante influenciado pela técnica da predição de posturas através de banco de dados.

As outras técnicas, entretanto, possibilitaram uma melhoria no desempenho em 39%, e ainda uma leve diminuição do erro.

4.14 Resultados da métrica da superfície de não recobrimento com divisão da imagem

4.14.1 Comparação com o sistema base sem utilizar as técnicas desenvolvidas

O sistema base, que usa a métrica da superfície de não recobrimento sem as técnicas anteriormente apresentadas, e a métrica que utiliza a divisão da imagem, nas mesmas condições, mostrou uma melhoria na qualidade da aquisição, mas ao custo de um desempenho muito inferior.

Na Figura 4-22 é mostrado um gráfico com os resultados desta métrica, em comparação ao sistema base.

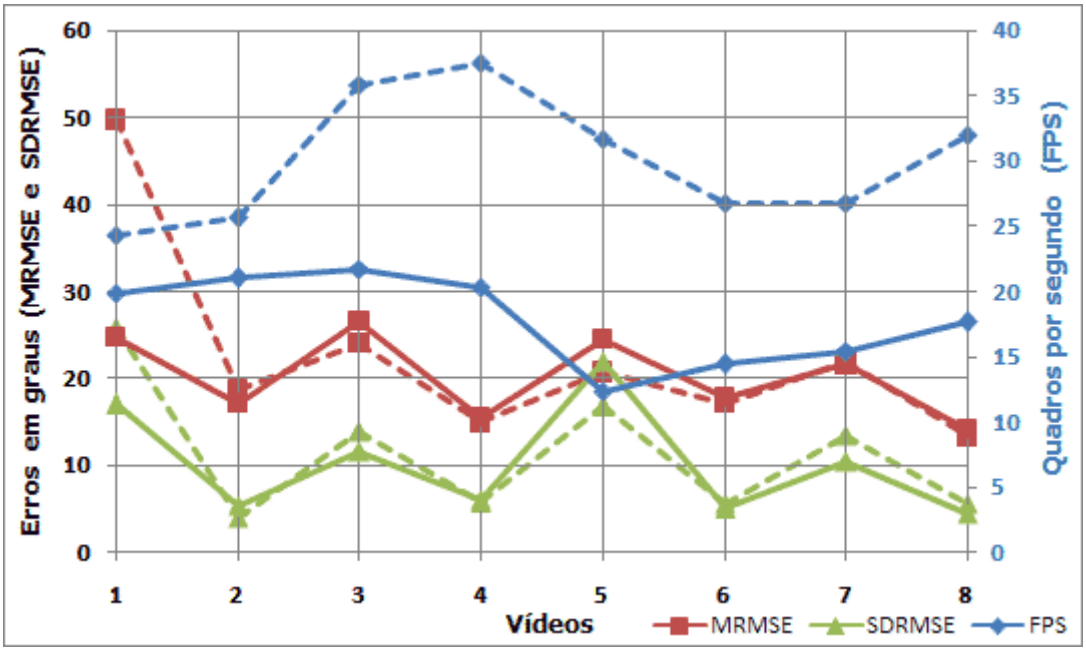

Figura 4-22: gráfico comparando a métrica da superfície de não recobrimento com divisão da imagem (linha continua) e o sistema base que é a métrica da superfície de não recobrimento (linha tracejada).

Resumo:

- MRMSE 22,6 ► 20,3 = + 10%;
- SDRMSE 11,4 ► 10,3 = + 10%;
- FPS 30,1 ► 17,8 = – 41%.

Esta métrica tem a vantagem de evitar que grandes áreas fiquem descasadas, por causa da análise das divisões da imagem, mas o procedimento de dividir a imagem e analisar cada fração separadamente acarreta em um peso computacional muito elevado. Percebe-se que o uso desta métrica evita resultados com grandes erros e, principalmente, sua propagação.

A perda de desempenho é muito elevada para ser utilizada como métrica principal, mas pode ser utilizada em situações especiais em que a qualidade da aquisição prevaleça em relação ao tempo.

4.14.2 Comparação com a melhor configuração da métrica de superfície de não recobrimento

Na Figura 4-23 é mostrado um gráfico comparando a melhor configuração obtida usando a métrica da superfície de não recobrimento com divisão da imagem, e a melhor configuração da métrica da superfície de não recobrimento mostrada na seção 4.13.2.

A melhor configuração desta métrica foi o conjunto, retirando apenas a técnica de iniciação aleatória, ou seja, utilizando: restrições biomecânicas, penalização de posturas irreais, classificação das regiões de pele e banco de dados.

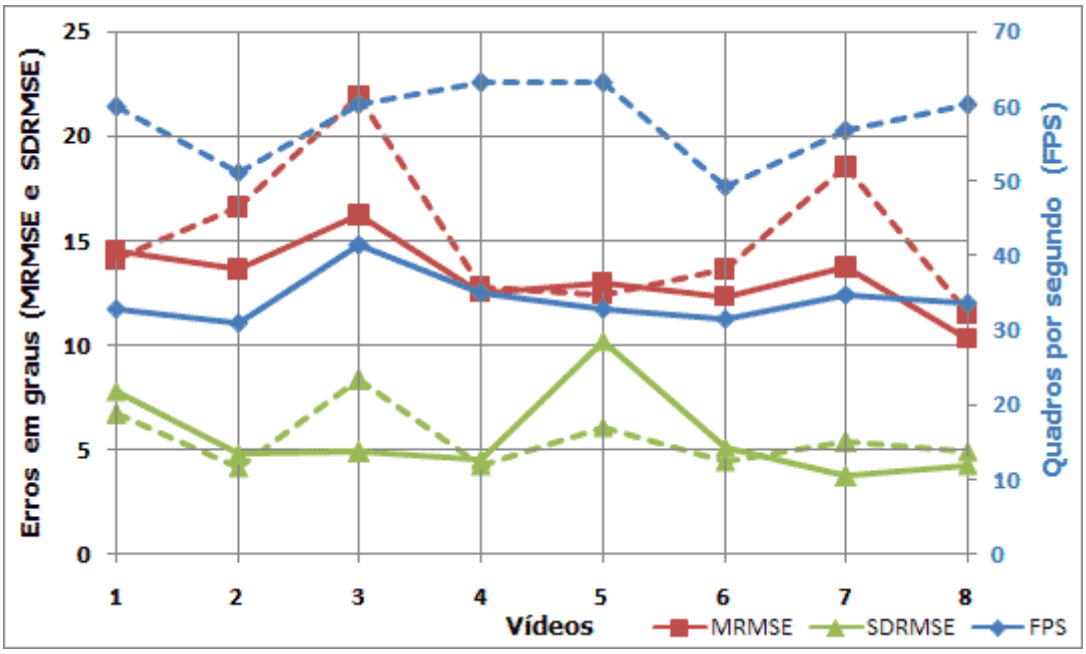

Figura 4-23: gráfico comparando o melhor conjunto de técnicas usando a métrica da superfície de não recobrimento com divisão da imagem (linha continua) e a melhor configuração da métrica da superfície de não recobrimento (linha tracejada).

Resumo:

- MRMSE 15,2 ► 13,3 = + 13%;
- SDRMSE 5,5 ► 5,7 = – 2%;
- FPS 58,0 ► 34,2 = – 41%.

Apesar da técnica de iniciação aleatória não ter, neste caso, melhorado o resultado, o que não é um fato comum, seu uso não prejudicou muito o resultado desta configuração, aumentando o MRMSE de 13,3 para 13,8, diminuindo o ganho de + 13% para + 9%.

Esta métrica tem a característica de evitar que posturas ambíguas sejam identificadas, em um nível muito melhor do que a métrica do sistema base, justamente por causa da divisão da imagem. Entretanto, a divisão proporciona outro tipo de ambigüidade, como mostrado na seção 3.4.3.

Este fato foi registrado no vídeo 5, em que o SDRMSE é quase o dobro. Isso se deve ao fato de alguns quadros do vídeo terem obtido erros elevados. Por outro

lado, o pequeno aumento do MRMSE mostra que tal ambigüidade não se propagou em demasia.

Uma análise detalhada dos dados do vídeo 5 mostra que, dos 144 quadros, houve seis incidências de mínimos locais, quatro destas com valores de RMSE na faixa dos 30º que não se propagaram, ficando apenas em seus respectivos quadros isolados, e dois casos onde obtiveram-se erros de 55º, sendo tal erro propagado para mais três quadros.

É válido salientar que o processamento de quatro quadros a uma taxa de 20 fps dura apenas 0,2 segundos, ou seja, a ambigüidade é visível, mas o tempo é bastante pequeno e, por isso, não se considera uma propagação demasiada.

Mesmo utilizando a melhor configuração em ambas as métricas, aquela com a divisão da imagem conseguiu obter melhor qualidade final no MRMSE. O erro MRMSE de 13,3º é o menor registrado no sistema de aquisição. Mas, conforme dito anteriormente, a perda de desempenho não justificaria sua utilização.

Numericamente, foram obtidos erros MRMSE menores de 18,9º, como será mostrado na seção 4.18.1, mas estes pioraram os dados de SDRMSE e FPS de forma que não compensa as diferenças.

4.15 Resultados da métrica da superfície de não recobrimento com divisão da imagem usando a função quadrática de avaliação

4.15.1 Comparação da métrica da superfície de não recobrimento com divisão da imagem, usando a função quadrática de avaliação, com o sistema base sem utilizar as técnicas desenvolvidas

Esta métrica penaliza ainda mais as posturas que possuem frações não recobertas. Isto se dá devido ao cálculo do erro quadrático médio das frações na função de custo.

Na Figura 4-24 é mostrado um gráfico comparando os resultados desta métrica e o sistema base.

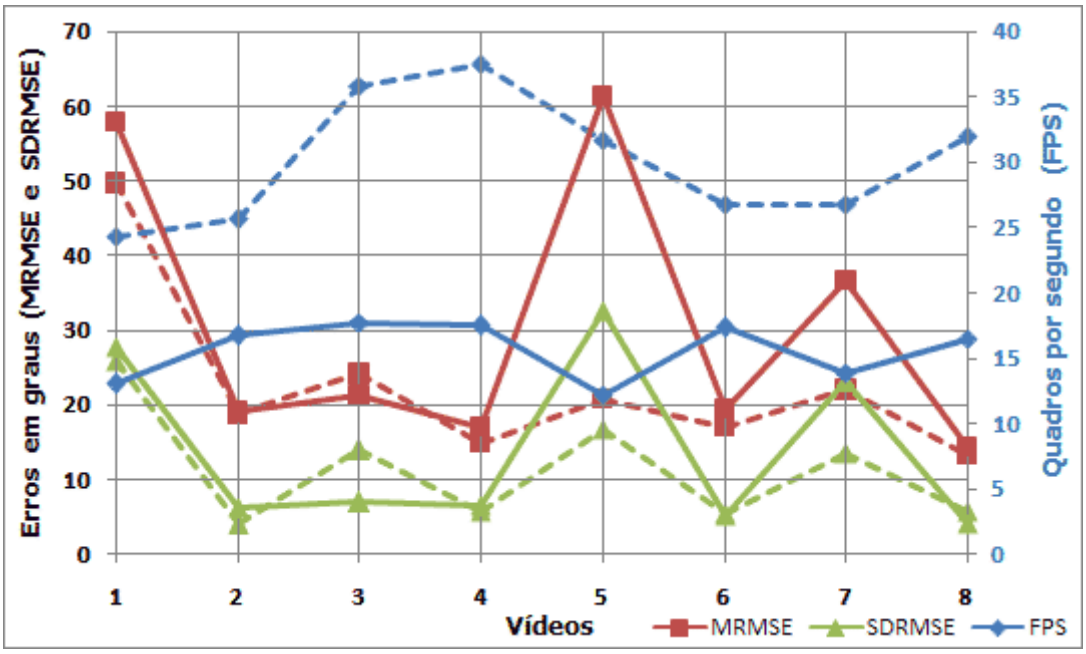

Figura 4-24: gráfico comparando a métrica da superfície de não recobrimento com divisão da imagem usando a função quadrática de avaliação (linha continua) e o sistema base (linha tracejada).

Resumo:

- MRMSE 22,6 ► 30,9 = – 37%;
- SDRMSE 11,4 ► 14,1 = – 24%;
- FPS 30,1 ► 15,6 = – 48%.

O incremento na penalização se tornou excessivo. Em posturas próximas à desejada, existe a possibilidade da existência de frações não recobertas. Outras métricas levariam o algoritmo de otimização a buscar outras posturas próximas, mas a penalização desta métrica pode prejudicar a convergência.

Esta métrica piorou até mesmo o primeiro vídeo, que já possuía um erro bastante elevado, situação em que todas as demais métricas e técnicas desenvolvidas apresentaram melhora de resultado. O quinto vídeo, que é a mesma seqüência de movimentos do primeiro vídeo, diferente apenas na proximidade da câmera, obteve uma piora elevadíssima em todos os dados.

4.15.2 Comparação com a melhor configuração da métrica de superfície de não recobrimento

A utilização das técnicas desenvolvidas podem melhorar os maus resultados obtidos por esta métrica isolada. Principalmente, o uso da predição através do banco de dados, pois este pode iniciar o algoritmo com posturas semelhantes à desejada, de forma a evitar que os erros comentados na subseção anterior venham a prejudicar em demasia a convergência do algoritmo.

Na Figura 4-25 é mostrado um gráfico comparando a melhor configuração obtida na métrica da superfície de não recobrimento com divisão da imagem, usando a função quadrática de avaliação e a melhor configuração da métrica da superfície de não recobrimento mostrada na seção 4.13.2.

A melhor configuração desta métrica foi o conjunto total, excetuando a técnica de classificação das regiões de pele, ou seja, utilizando: restrições biomecânicas, penalização de posturas irreais, iniciação aleatória e banco de dados.

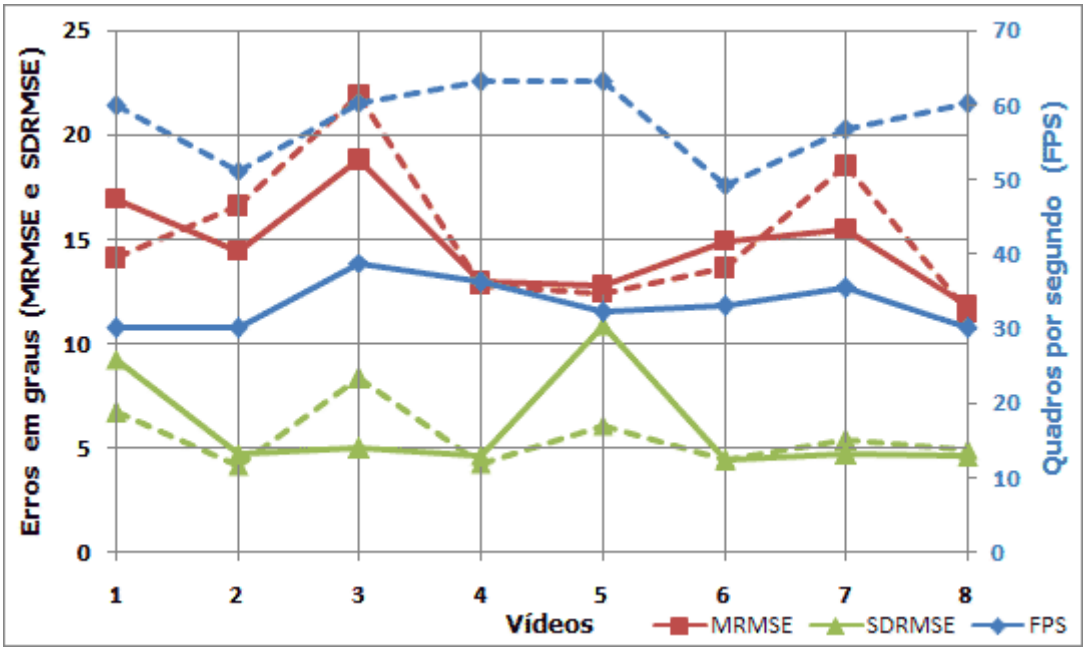

Figura 4-25: gráfico comparando a melhor configuração da métrica da superfície de não recobrimento com divisão da imagem usando a função quadrática de avaliação (linha continua) e a melhor configuração da métrica da superfície de não recobrimento (linha tracejada).

Resumo:

- MRMSE 15,2 ► 14,8 = + 3%;
- SDRMSE 5,5 ► 6,0 = – 9%;
- FPS 58,0 ► 33,2 = – 43%.

As técnicas desenvolvidas em conjunto com esta métrica melhoram muito seu resultado em comparação com a métrica isolada, obtendo erros comparáveis às outras métricas.

Mesmo assim, esta métrica não superou o resultado da métrica em que foi baseada, que é a superfície de não recobrimento com divisão da imagem sem a função quadrática de avaliação.

Assim, conclui-se que esta métrica pode ser ignorada e utiliza-se apenas a superfície de não recobrimento com divisão da imagem.

4.16 Resultados da métrica da diferença em *pixel* entre as imagens

4.16.1 Comparação da métrica da diferença em *pixel* entre as imagens com o sistema base sem utilizar as técnicas desenvolvidas

A grande vantagem desta métrica é sua simplicidade, apesar de sua boa funcionalidade. Métricas como a superfície de não recobrimento com divisão da imagem representam um algoritmo complexo, com bastantes cálculos. A métrica apresentada nessa seção, entretanto, utiliza uma idéia simples que mostra resultados satisfatórios.

Na Figura 4-26, é mostrado um gráfico, comparando os resultados desta métrica e o sistema base.

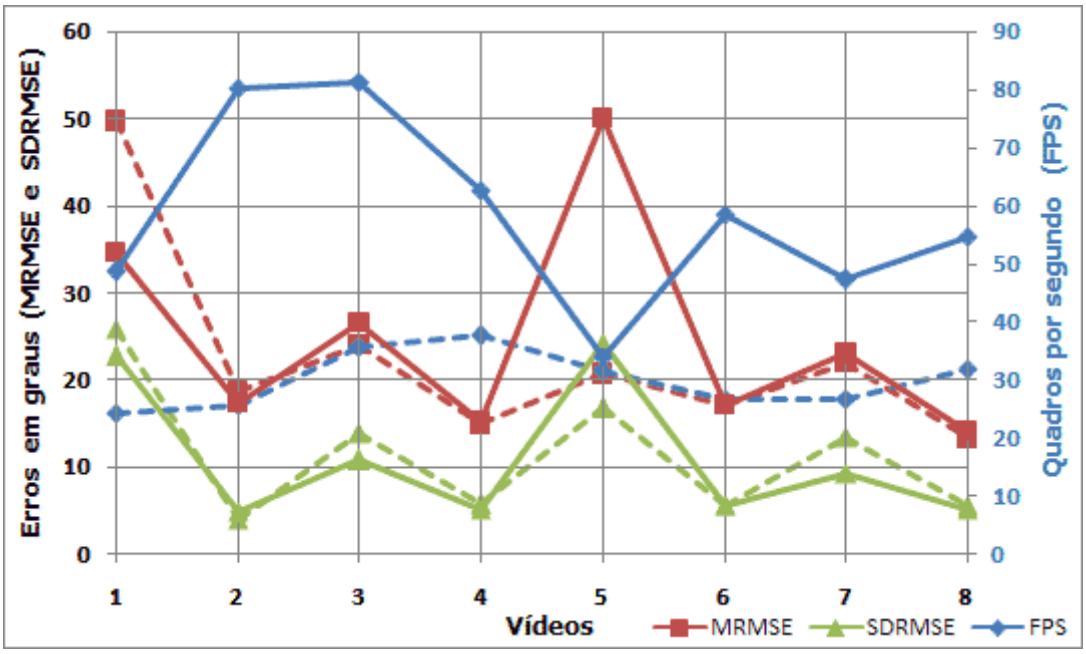

Figura 4-26: gráfico comparando a métrica da diferença em *pixel* entre as imagens (linha continua) e o sistema base (linha tracejada).

Resumo:

- MRMSE 22,6 ► 24,8 = – 10%;
- SDRMSE 11,4 ► 11,0 = + 3%;
- FPS 30,1 ► 58,4 = + 94%.

O resultado geral do MRMSE desta métrica foi inferior, devido ao do vídeo 5, que apresentou um erro muito elevado. Os demais resultados foram similares. Com esta técnica, o sistema atingiu valores de desempenhos surpreendentes, como 80 fps nos vídeos 2 e 3.

Sua principal diferença encontra-se no desempenho obtido que foi, em média, quase o dobro da métrica do sistema base. Com este ganho de desempenho, é possível duplicar o numero de iterações do DHS, visando melhorar a qualidade da identificação.

4.16.2 Comparação com a melhor configuração da métrica de superfície de não recobrimento

A melhor configuração desta métrica foi o conjunto utilizando: restrições biomecânicas, penalização de posturas irreais e iniciação aleatória, sem incluir o uso do banco de dados e a classificação das regiões de pele, sendo esta última incompatível com a métrica da diferença em *pixel* entre as imagens.

Na Figura 4-27, é mostrado um gráfico comparando a melhor configuração obtida usando a métrica da diferença em *pixel* entre as imagens e a melhor configuração da métrica da superfície de não recobrimento mostrada na seção 4.13.2.

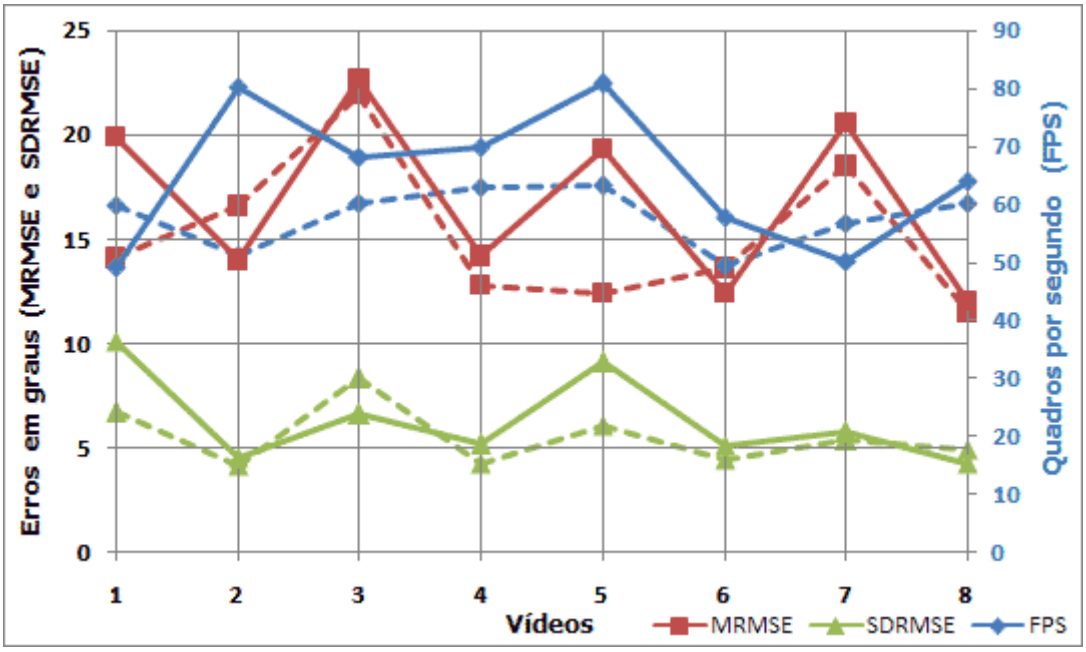

Figura 4-27: gráfico comparando a melhor configuração da métrica da diferença em *pixel* entre as imagens (linha continua) e a melhor configuração da métrica da superfície de não recobrimento (linha tracejada).

Resumo:

- MRMSE 15,2 ► 16,9 = – 11%;
- SDRMSE 5,5 ► 6,3 = – 14%;
- FPS 58,0 ► 65,0 = + 12%.

É oportuno salientar que os bons resultados desta métrica não se devem ao uso do banco de dados, que representa um diferencial em desempenho e qualidade da melhor configuração da superfície de não recobrimento. Em outras palavras, esta métrica consegue ter resultados próximos à comparada, sem necessitar de treinamento prévio para alimentar um banco de dados, obtendo-se, ainda assim, um desempenho superior.

O uso de banco de dados, entretanto, piora a qualidade da aquisição na métrica da diferença em *pixel* entre as imagens, pois a contagem de pixels é realizada sobre a imagem completa, sem identificar partes do corpo, o que permite grande ambigüidade na comparação de posturas.

4.17 Resultados da métrica da diferença em *pixel* entre as imagens com sub-amostragem

4.17.1 Comparação da métrica da diferença em *pixel* entre as imagens com sub-amostragem com o sistema base, sem utilizar as técnicas desenvolvidas

A sub-amostragem foi proposta para penalizar as posturas com grandes áreas descasadas, o que deveria melhorar erros grandes e reduzir a ambigüidade.

Na Figura 4-28, é mostrado um gráfico comparando os resultados desta métrica e o sistema base.

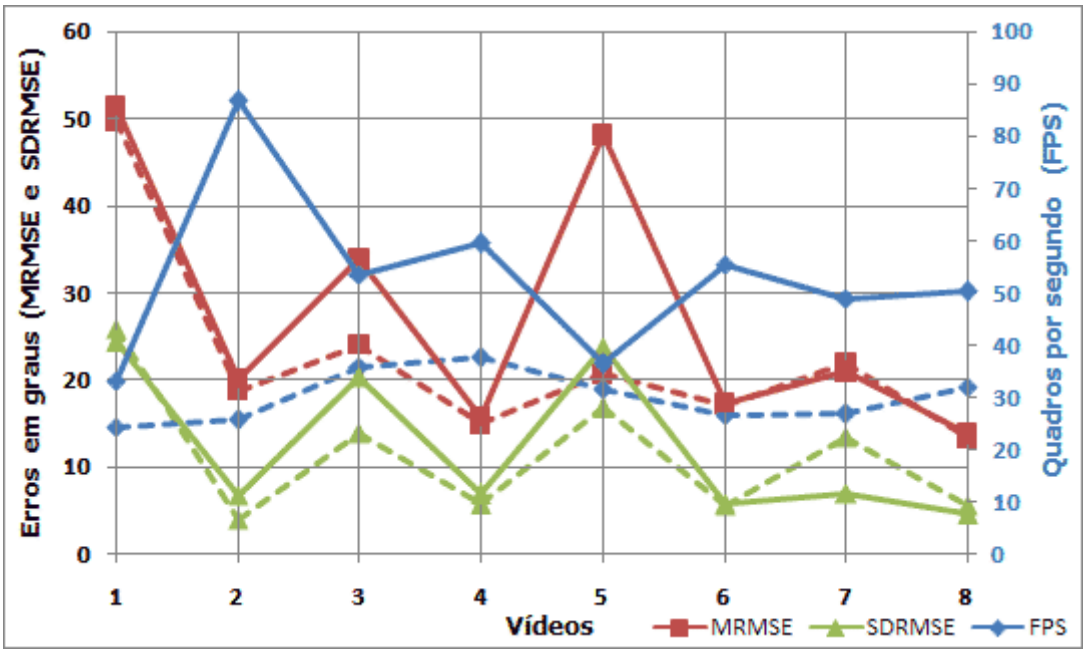

Figura 4-28: gráfico comparando a métrica da diferença em *pixel* entre as imagens com sub-amostragem (linha continua) e o sistema base (linha tracejada).

Resumo:

- MRMSE 22,6 ► 27,7 = – 22%;
- SDRMSE 11,4 ► 12,4 = – 9%;
- FPS 30,1 ► 53,0 = + 76%.

Observa-se que esta métrica sozinha não conseguiu cumprir com seu propósito, pois não obteve melhores resultados, em comparação à métrica sem o uso da sub-amostragem mostrada na seção 4.16.1. Além disso, não conseguiu evitar os mínimos locais registrados no vídeo 7.

4.17.2 Comparação com a melhor configuração da métrica de superfície de não recobrimento

A melhor configuração desta métrica foi o conjunto utilizando: restrições biomecânicas e iniciação aleatória, ou seja, sem incluir as técnicas do banco de dados, a penalização de posturas irreais e a classificação das regiões de pele, sendo a última incompatível com esta métrica da diferença em *pixel* entre as imagens.

Na Figura 4-29, é mostrado um gráfico comparando a melhor configuração obtida usando a métrica da diferença em *pixel* entre as imagens com sub-amostragem e a melhor configuração da métrica da superfície de não recobrimento mostrada na seção 4.13.2.

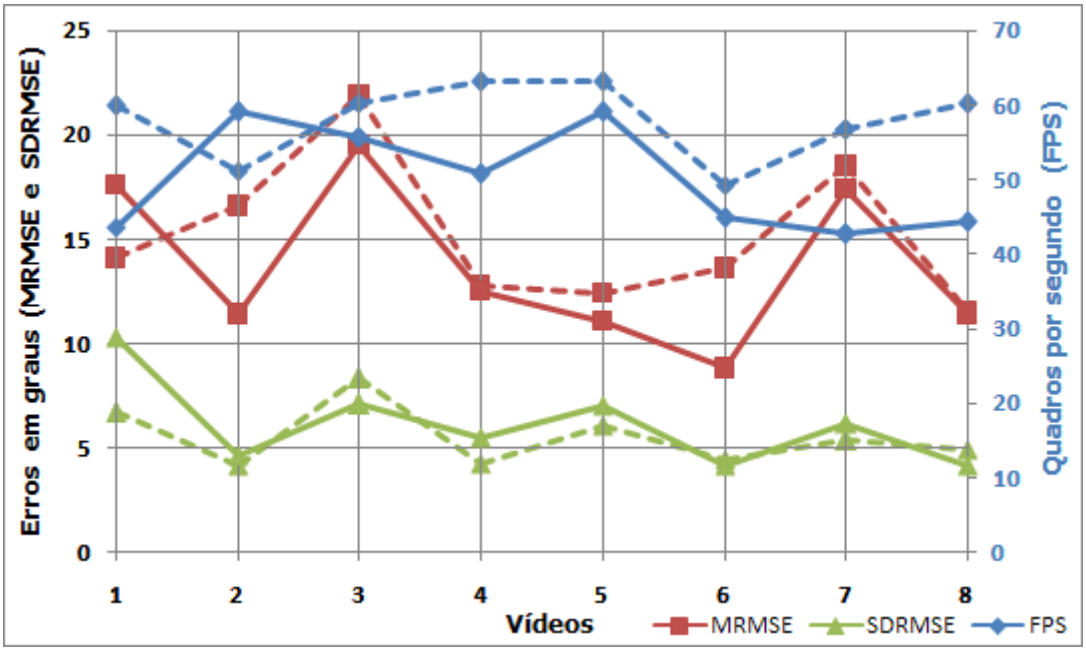

Figura 4-29: gráfico comparando a melhor configuração da métrica da diferença em *pixel* entre as imagens com sub-amostragem (linha continua) e a melhor configuração da métrica da superfície de não recobrimento (linha tracejada).

Resumo:

- MRMSE 15,2 ► 13,7 = + 10%;
- SDRMSE 5,5 ► 6,1 = – 11%;
- FPS 58,0 ► 50,0 = – 14%.

Apesar desta métrica sozinha não ter obtido bons resultados, a agregação das técnicas desenvolvidas melhoram bastante seu resultado, obtendo o segundo melhor MRMSE de todos os resultados, perdendo apenas para a métrica da superfície de não recobrimento com divisão da imagem em sua melhor configuração (seção 4.14.2). Por isso, a próxima subseção compara estas duas métricas.

4.17.3 Comparação com a métrica da superfície de não recobrimento com divisão da imagem, ambas em suas melhores configurações

Como foi visto anteriormente, estas foram as duas melhores métricas com relação ao MRMSE. Assim, sua comparação é importante para analisar detalhadamente os demais dados obtidos.

A Figura 4-30 mostra um gráfico comparando a melhor configuração obtida usando a métrica da diferença em *pixel* entre as imagens com sub-amostragem e a melhor configuração da métrica da superfície de não recobrimento com divisão da imagem (seção 4.14.2).

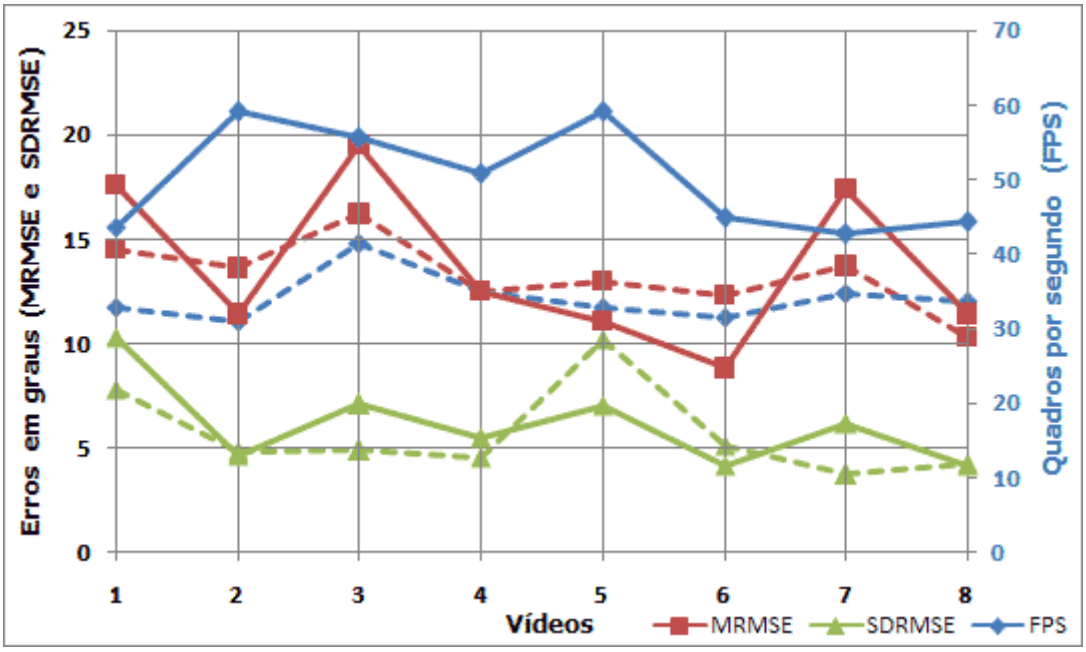

Figura 4-30: gráfico comparando a melhor configuração da métrica diferença em *pixel* entre as imagens com sub-amostragem (linha continua) e a melhor configuração da métrica da superfície de não recobrimento (linha tracejada).

Resumo:

- MRMSE 13,3 ► 13,7 = – 3%;
- SDRMSE 5,7 ► 6,1 = – 8%;
- FPS 34,2 ► 50,0 = + 46%.

A nova métrica, em sua melhor configuração, obteve ótimos resultados, quase tão bons quando a melhor configuração em qualidade já obtida, que é a métrica da superfície de não recobrimento com divisão da imagem da seção 4.14.2. Mas se considerarmos que esta nova métrica tem um desempenho muito superior, podemos considerá-la como melhor no geral, pois um ganho de 46% em 50 fps possibilita alterar parâmetros do sistema para reduzir os erros.

Contudo, a métrica comparada da seção 4.14.2 é mais estável, e pode ser observado no gráfico da Figura 4-30 que as linhas tracejadas não variam tanto quando as continuas.

4.17.4 Discussão dos resultados das métricas baseadas na diferença em *pixels* entre as imagens

Ambas as métricas da diferença em *pixel*, com e sem a sub-amostragem, não utilizam a predição de posturas através do uso de um banco de dados em suas melhores configurações. Mesmo assim, obtiveram resultados próximos das outras métricas, como mostra a última comparação realizada (seção anterior), em que os resultados foram tão bons quanto a configuração com menor erro do sistema, e ainda obtendo um ótimo desempenho.

Estas novas métricas foram os últimos desenvolvimentos deste trabalho, ficando os aprimoramentos para trabalhos futuros. A métrica não funcionou em conjunto com o banco de dados, mas pode-se ainda explorar formas de fazer sua integração, buscando melhores resultados, assim como os observados na seção 4.12, com ganhos na faixa dos 30%.

O interesse desta subseção é mostrar as perspectivas de resultado desta métrica, que se estima com grande potencial para tornar-se a mais eficiente em qualidade e tempo entre todas as demais.

Outro fator que dá crédito à métrica da diferença em *pixel* entre as imagens é a forma de penalização das posturas, visto que a sub-amostragem não permite a ocorrência das ambigüidades observadas nas métricas com divisão da imagem como mostrado na seção 3.5.3.

Atualmente, os processos de sub-amostragem, e de contagem de *pixels* antes e depois da sub-amostragem são realizados separadamente. Entretanto, é possível unificá-los em um único processo, o que aumentará o desempenho desta métrica.

4.18 Melhores resultados

Para verificar a importância que as técnicas e métricas representaram nos resultados, analisamos as melhores configurações.

4.18.1 Comparação de qualidade global de métricas e técnicas

Para efeito de comparação de qualidade das métricas e técnicas desenvolvidas neste trabalho, usamos a média dos resultados obtidos dos 8 vídeos para cada configuração, resultando em 128. Os dados foram ordenados pelo menor MRMSE e, para análise estatística, foram separados os 40 primeiros.

Das 40 configurações, 39 utilizam a técnica do banco de dados representando 98% dos melhores resultados. As restrições biomecânicas estão presentes em 22 destas, representando 55% das configurações, a iniciação aleatória ficou equilibrada em 50%, e a penalização de posturas irreais e a classificação das regiões de pele ficaram com 45% e 37%, respectivamente, demonstrando que estas técnicas não foram fator predominante na melhoria da qualidade.

Na Figura 4-31 são mostrados os gráficos com tais percentuais.

O resultado das três métricas baseadas na superfície de não recobrimento representa quase integralmente, 95% do resultado das melhores configurações. As duas métricas baseadas na superfície de não recobrimento com divisão da imagem representam 78% das melhores configurações e demonstram sua predominância para obtenção dos menores erros.

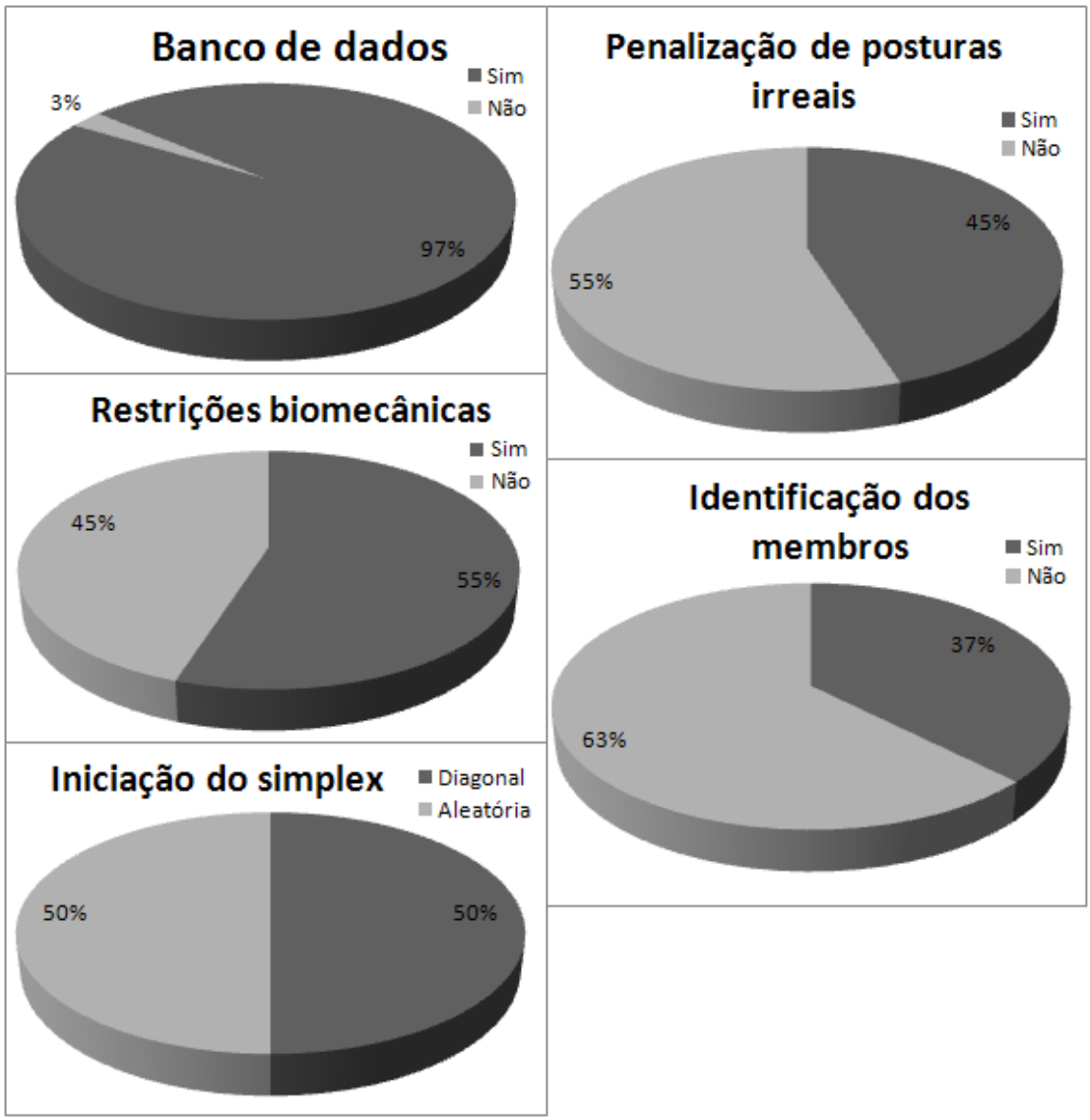

Figura 4-31: gráficos do percentual representando o uso das técnicas que compõem o conjunto das 40 configurações com menores MRMSE.

A métrica da diferença em *pixels* entre as imagens não apresentou nenhum resultado na lista. E a métrica da diferença em *pixels* entre as imagens com sub-amostragem está presente em apenas 2 das 40 configurações, representando somente 5%.

Na Figura 4-32 são mostrados tais percentuais graficamente.

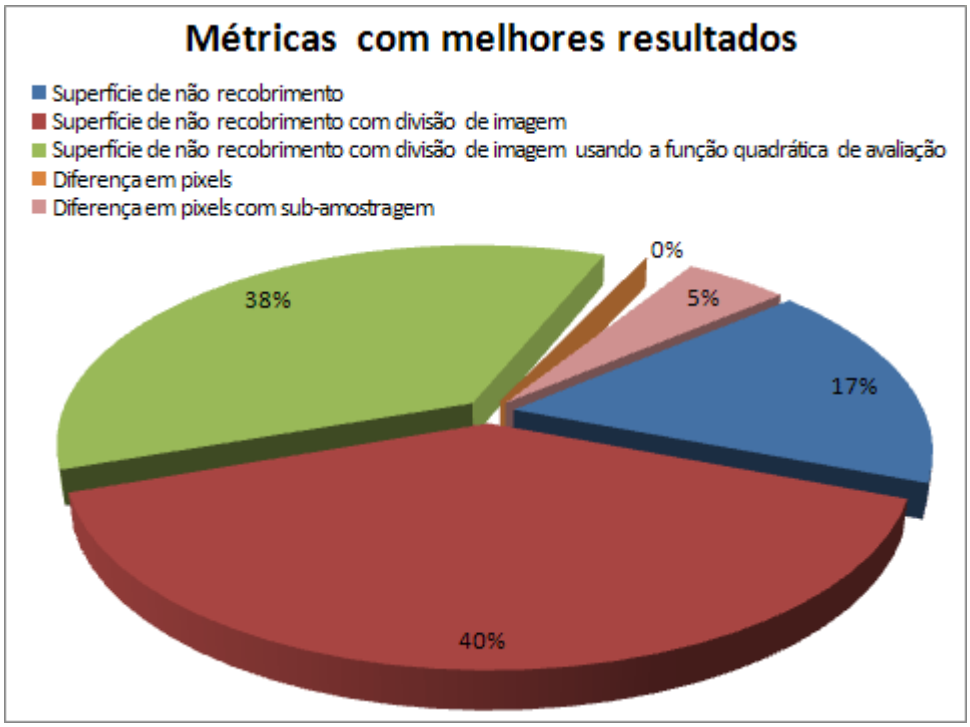

Figura 4-32: gráficos do percentual representando o uso das métricas de avaliação de similaridade que compõem o conjunto das 40 configurações com menor MRMSE.

Os seis melhores resultados utilizam a métrica da superfície de não recobrimento com divisão da imagem, juntamente com a técnica das restrições biomecânicas dinâmicas e banco de dados.

Na 9ª posição, encontra-se a métrica da diferença em pixels entre as imagens com sub-amostragem, e é a única configuração nesta lista que não utiliza o banco de dados.

Dos dezesseis melhores resultados, treze utilizam a métrica da superfície de não recobrimento com divisão da imagem. Isso demonstra a importância desta com relação à obtenção de qualidade na aquisição.

A análise destas 40 melhores configurações é relevante para mostrar quais métricas e técnicas contribuem para a melhoria do erro.

Por outro lado, todas estas configurações obtiveram bons resultados com relação à qualidade da aquisição, demonstrado pela faixa de erros obtidos, com MRMSE entre 12,9 e 15,6 graus, e SDRMSE entre valores de 5,4 a 7,7 graus. Onde qualquer destes resultados é considerado um ótimo resultado.

O desempenho em FPS varia entre 18,9 e 58,0 fps. Apesar desta variação ser muito grande, o menor valor, 19 fps, já pode ser considerado tempo real. Um desempenho de 58 fps, entretanto, vai além da expectativa, possibilitando uma nova gama de técnicas, e aperfeiçoamento das atuais, na tentativa de melhorar a qualidade.

4.18.2 Comparação de desempenho

Além da comparação dos melhores resultados em qualidade, separamos também para análise estatística as 40 configurações com melhor desempenho, ou seja, as com maior FPS. Esta lista contém valores de FPS entre 45 e 69 fps. Por outro lado, nem todas estas configurações obtiveram bons resultados de qualidade, registrando-se valores de MRMSE entre 13,7 e 27,7, e SDRMSE entre 5,5 e 14,2.

A penalização de posturas irreais se mostra presente em 65% das configurações com maior desempenho. As técnicas de restrições biomecânicas e iniciação aleatória não influenciaram muito o desempenho. Adicionalmente, a classificação das regiões de pele não ajuda o desempenho do sistema, estando presente em apenas 17% destas.

A técnica do banco de dados está presente na metade das configurações, com um percentual de 52%. Mas é valido ressaltar que, como será visto a seguir, grande parte dos resultados com maior desempenho fazem parte das métricas baseadas na diferença em *pixel* entre as imagens, e como esta não obteve bons resultados em conjunto com o banco de dados, faz o percentual da técnica de banco de dados menos representativa no conjunto das configurações com melhor desempenho. A técnica do banco de dados, entretanto, melhora bastante o desempenho do sistema quando observada nas métricas baseadas na superfície de não recobrimento.

Na Figura 4-33 são mostrados os gráficos com tais percentuais.

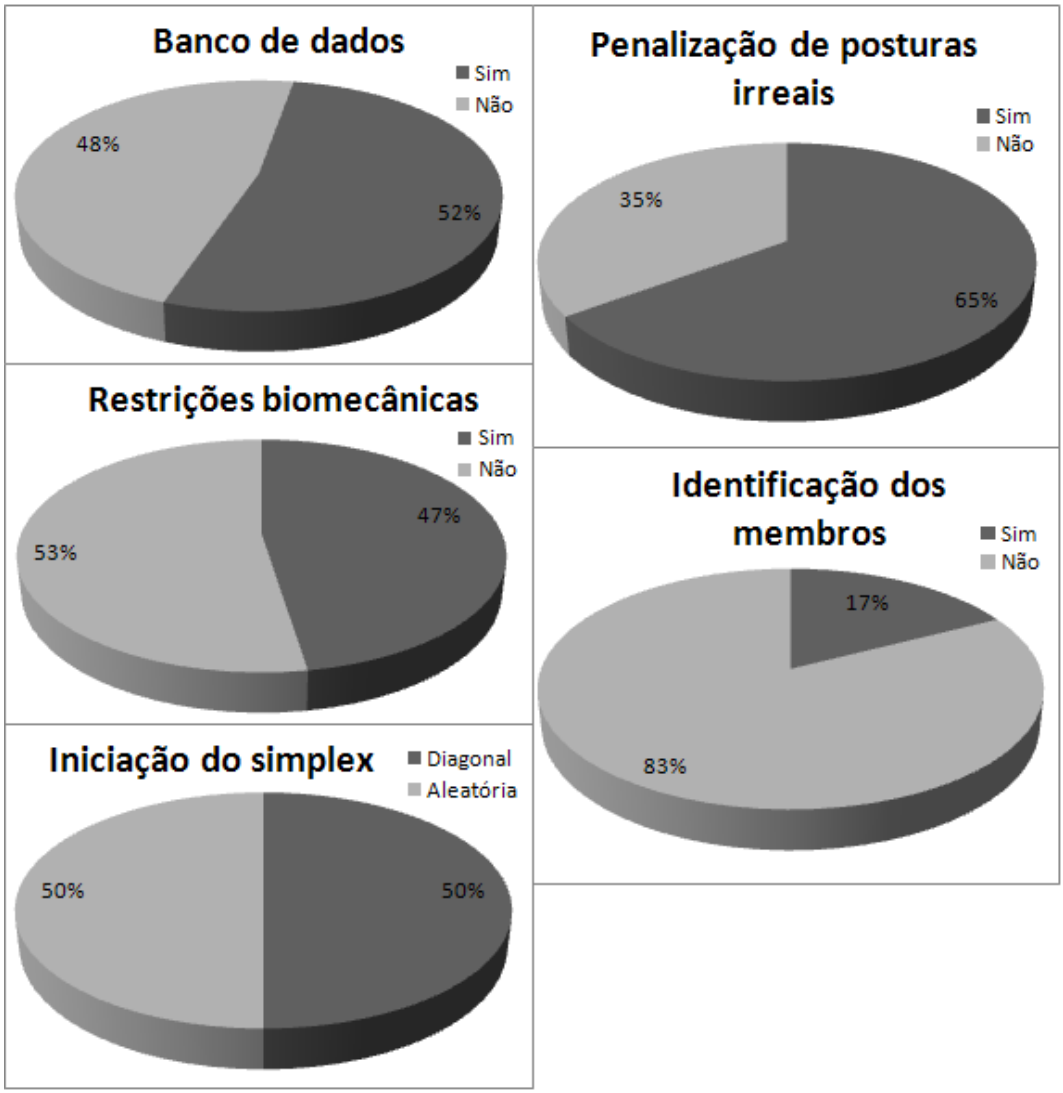

Figura 4-33: gráficos do percentual representativo das técnicas no conjunto das 40 configurações com melhor desempenho.

O resultado das métricas neste conjunto de configurações está concentrado nas duas métricas baseadas na diferença em *pixel* entre as imagens, representando 73% das configurações com maior desempenho.

As duas métricas baseadas na superfície de não recobrimento com divisão da imagem, que apresentaram os melhores resultados em qualidade, não estão presentes em nenhuma das configurações de desempenho.

Na Figura 4-34 é mostrado tais percentuais graficamente.

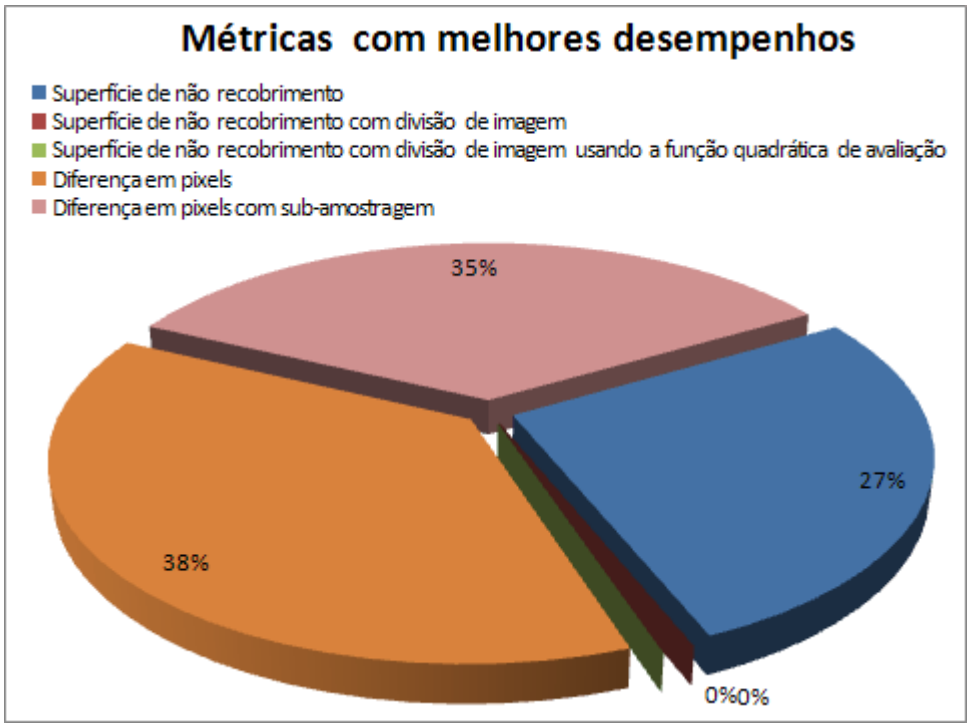

Figura 4-34: gráficos do percentual representativo das métricas de avaliação de similaridade no conjunto das 40 configurações com melhor desempenho.

Não existe um padrão nos primeiros resultados de desempenho, como mostrado para os resultados de qualidade na subseção anterior.

Observa-se, entretanto, uma distribuição equilibrada das três métricas que não obtiveram valores nulos em percentuais na figura acima. Isso mostra que as duas métricas baseadas na diferença em *pixel* entre as imagens, juntas, representam os melhores desempenhos. Por outro lado, isoladas, elas têm valores numéricos próximo aos da métrica da superfície de não recobrimento.

Capítulo 5

5 Conclusão

Este trabalho disserta sobre os resultados de pesquisas relacionadas à identificação de posturas humanas por visão computacional monocular em seqüência de vídeo, em tempo real.

O sistema de aquisição de gestos proposto visa extrair de uma seqüência de imagens os parâmetros que permitam atribuir a um modelo humanóide a mesma postura do ator na imagem. O processo consiste em ajustar um modelo humanóide tridimensional sobre o ator em cada imagem da seqüência de vídeo.

Foram desenvolvidas e avaliadas diversas técnicas, procurando contornar as limitações conhecidas no sistema descrito em [1,50,51], buscando a melhoria na qualidade da aquisição, reduzindo o erro em graus dos ângulos das articulações da postura. E buscou-se também a redução do tempo de processamento, aumentando o desempenho do sistema.

A combinação das técnicas pode levar o sistema a uma maior robustez quando adequada ao cenário. Observa-se que as métricas e técnicas desenvolvidas e experimentadas apresentam comportamentos diferentes em configurações ambientais distintas. Dessa maneira, o usuário pode escolher a configuração mais adequada, de acordo com as variáveis do cenário de aquisição, que podem envolver, dentre outras, a velocidade e o tipo dos movimentos praticados pelo ator, o número de posturas que devem ser identificadas por segundo, a possibilidade de registro prévio de posições comuns e a precisão desejada.

Discutem-se, a seguir, alguns dos resultados obtidos com o emprego das técnicas e das métricas de similaridade desenvolvidas neste trabalho.

A métrica da superfície de não recobrimento com divisão da imagem, em conjunto com as técnicas desenvolvidas, obteve uma melhoria considerável no

resultado do sistema de aquisição de gestos, quando comparado ao trabalho de Soares [1], diminuindo em torno de 67% o erro em graus nas posturas identificadas.

Quando o ambiente de aquisição permite a realização de treinamento prévio, a técnica de predição de posturas utilizando um banco de dados demonstra-se particularmente aplicável e eficiente, melhorando bastante o desempenho do sistema, reduzindo erros de identificação e, mesmo quando estes ocorrem devido à incidência de mínimos locais, evitando que estes erros se propagem para os próximos quadros do vídeo.

Para uma aquisição sem treinamento, a métrica da diferença em *pixel* entre as imagens com sub-amostragem revela-se a mais recomendável, visto que ela se destaca de maneira incomparável dentre aquelas que não utilizam o banco de dados para predição. A grande vantagem desta métrica é o seu alto desempenho (taxa em fps), associado a níveis de erro próximos às demais métricas desenvolvidas.

A diferença em *pixel* com sub-amostragem apresenta, em perspectiva, grande potencial na melhoria da qualidade da identificação, caso se consiga utilizá-la em conjunto com a predição utilizando o banco de dados, bem como o aumento do desempenho através da otimização de seu processo.

A métrica da superfície de não recobrimento com divisão da imagem, embora apresente resultados de melhor qualidade na identificação de posturas, possui um desempenho comprometedor. Entretanto, o aumento em capacidade de memória e de desempenho que vêm sendo observados continuamente nos computadores pessoais, pode trazer novas perspectivas de uso para esta técnica.

Dentre as demais técnicas exploradas, em ordem de importância, em termos de contribuição com a melhoria dos resultados, temos: (i) uso de banco de dados, (ii) iniciação aleatória dos valores dos vértices do *simplex*, (iii) uso de restrições biomecânicas e (iv) penalização de posturas irreais. Observa-se que a técnica da penalização de posturas irreais melhora o desempenho, prejudicando, entretanto, a qualidade da aquisição.

A iniciação aleatória dos valores dos vértices do *simplex* mostrou-se bastante importante para reduzir a incidência de mínimos locais, sem, entretanto, prejudicar o desempenho do sistema e a convergência do algoritmo.

As restrições biomecânicas dinâmicas, por sua vez, contribuem com a diminuição do espaço de busca, evitando a avaliação e a convergência para posturas ergonomicamente desfavoráveis.

A técnica da classificação das regiões de pele da imagem para identificação dos membros apresentou bons resultados apenas em situações particulares, não se demonstrando uma técnica robusta. Entretanto, sugere-se que esta técnica seja melhor explorada em conjunto com a métrica de superfície de não recobrimento devido ao seu grande potencial para a eliminação de ambigüidades.

Além do aprimoramento das técnicas e métricas implementadas, cujas limitações e perspectivas já foram mencionadas, pode-se colocar como meta para trabalhos futuros a pesquisa de recursos que permitam a identificação de atributos do ambiente, de maneira a definir, automaticamente, a configuração a ser usada para o cenário de aquisição.

Publicações

As seguintes publicações registram algumas das contribuições realizadas ao longo deste trabalho:

- VI Encontro de Pesquisa e Pós-Graduação – EMPPG do Cefet-CE, realizado em Fortaleza – CE nos dias 23 a 25 de agosto de 2006 [52];
- II Workshop de Visão Computacional – WVC2006 realizado em São Carlos – SP nos dias 16 a 18 de outubro de 2006 [50];
- XVII Congresso Brasileiro de Automática – CBA2008, realizado em Juiz de Fora – MG, nos dias 14 a 17 de setembro de 2008 [51].

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Soares, José Marques. "Contribution à la communication gestuelle dans les environnements virtuels collaboratifs." *PhD Thesis of Institut National de Télécommunications.* Evry, France, 2004. 133p.

[2] Mittal, Anurag, Liang Zhao, and Larry S. Davis. "Human Body Pose Estimation Using Silhouette Shape Analysis." *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03).* 2003.

[3] Weik, Sebastian, and C. E. Liedtke. "Hierarchical 3D Pose Estimation for Articulated Human Body Models from a Sequence of Volume Data." *Robot Vision In Lecture Notes in Computer Science.* Vol. 1998/2001. University of Hanover, Germany: Springer Berlin / Heidelberg, Monday, January 01, 2001. pg. 27-34.

[4] Zhang, Jiayong, Jiebo Luo, Robert Collins, and Yanxi Liu. "Body Localization in Still Images Using Hierarchical Models and Hybrid Search." *Computer Vision and Pattern Recognition Conference.* Vol. 2. EUA, 2006. pages 1536–1543.

[5] Shakhnarovich, Gregory, Paul Viola, and Trevor Darrel. "Fast Pose Estimation with Parameter-Sensitive Hashing." *Ninth IEEE International Conference on Computer Vision (ICCV 2003).* Vol. 2. 2003.

[6] Micilotta, Antonio S., Eng-Jon Ong, and Richard Bowden. "Real-time Upper Body Detection and 3D Pose Estimation in Monoscopic Images." *Centre for Vision, Speech and Signal Processing, University of Surrey.* Surrey, United Kingdom, 2003.

[7] —. "A Local Basis Representation for Estimating Human Pose from Cluttered Images." *Proceedings of the 7th Asian Conference on Computer Vision, 2006.* France, 2006.

[8] Poppe, Ronald. "Real-time pose estimation from monocular image sequence using silhouettes." University of Twente, April 2004.

[9] Watanabe, T., and M. Yachida. "Real Time Gesture Recognition Using Eigenspace from Multi Input Image Sequences." *Systems and Computers in Japan.* Vols. 30, Issue 13. nov. 1999. pp. 61-72.

[10] Soares, José Marques, Francisco José Marques Anselmo, Carlos Maurício Jaborandy Mattos, Patrick Anderson Moreira Macelino, Giovanni Cordeiro Barroso, e Paulo César Cortez. “Uma Infra-estrutura para a Colaboração à Distância com Suporte à Comunicação Gestual.” *XXXIII Seminário Integrado de Software e Hardware – SEMISH 2006.* Campo Grande, MS, 2006. CD-ROM.

[11] Sato, T., H. Mamiya, H. Koike, and K and Fukuchi. "An augmented tabletop video game with pinching gesture recognition." *In ACM SIGGRAPH ASIA 2008 Artgallery: Emerging Technologies (Singapore, December 10 - 13, 2008). SIGGRAPH Asia '08.* ACM, New York, NY, 2008. 38-38.

[12] Loper, M. M., N. P. Koenig, S. H. Chernova, C. V. Jones, and O. C. Jenkins. "Mobile human-robot teaming with environmental tolerance." *In Proceedings of the 4th ACM/IEEE international Conference on Human Robot interaction (La*

Jolla, California, USA, March 09 - 13, 2009). HRI '09. ACM. New York, NY, 2009. 157-164.

[13] Moustakas, Konstantinos, Georgios Nikolakis, Dimitrios Tzovaras, Sebastien Carbini, Olivier Bernier, and Jean Emmanuel Viallet. "3D content-based search using sketches." *Personal and Ubiquitous Computing - Computer Science.* Springer London, 20 April 2007.

[14] Moeslund, T., and E. Granun. "A Survey of Computer Vision-Based Human Motion Capture." *Computer Vision and Image Understanding.* Vol. 81. 2001. p. 231-268.

[15] Kohli, Pushmeet, Jonathan Rihan, Matthieu Bray, and Philip H. S. Torr. "Simultaneous Segmentation and Pose Estimation of Humans Using Dynamic Graph Cuts." *International Journal of Computer Vision.* Vols. Volume 79, Issue 3. 2008-09-08. 285-298.

[16] Bergh, Michael Van den, Esther Koller-Meier, and Luc Van Gool. "Real-Time Body Pose Recognition Using 2D or 3D Haarlets." *International Journal of Computer Vision.* 21 February 2009.

[17] Zou, Beiji, Shu Chen, Cao Shi, and Umugwaneza Marie Providence. "Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking." *Pattern Recognition.* Vols. 42, Issue 7. January 2009. Pages 1559-1571.

[18] Maciel, A. “Modelagem de Articulações para Humanos Virtuais Baseada em Anatomia.” *Universidade Federal do Rio Grande do Sul.* Porto Alegre, RS, agosto de 2001.

[19] Zhu, Youding, B. Dariush, and K. Fujimura. "Controlled human pose estimation from depth image streams." *Computer Vision and Pattern Recognition Workshops, 2008.* CVPRW 08. IEEE Computer Society Conference on, 23-28 June 2008. Page(s):1 - 8.

[20] Qiang, Chen, Zheng EnLiang, and Liu YunCai. "Pose estimation based on human detection and segmentation." *Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University.* Shanghai 200240, China, 2008-10-23.

[21] Ouhaddi, Hocine, and Patrick Horain. "3D Hand Gesture Tracking by Model Registration." *Proc. IWSNHC3DI’99.* 1999. 70-73.

[22] Gavrila, D. M. "The Visual Analysis of Human Movement: A Survey." *Computer Vision and Image Understanding. Daimler-Benz Research.* Vols. 73, n. 1. Wilhelm Runge St. 11, 89081 Ulm, Germany, 1999. p. 82-98.

[23] Laxton, B. "Monocular Human Pose Estimation." *University of California.* San Diego, 2007.

[24] Gall, J., R. Bodo, and Hans-Peter. "Clustered Stochastic Optimization for Object Recognition and Pose Estimation." *Seidel Max-Planck-Institute for Computer Science.* Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, 2007.

[25] Gavrila, D. M. "The analysis of human motion and its application for visual surveillance." *International Workshop on Visual Surveillance,.* Fort Collins, USA, 1999. p. 3-5.

[26] Yonemoto, S., A. Matsumoto, D. Arita, and R. Taniguchi. "A Real-Time Motion Capture System with Multiple Camera Fusion." *ICIAP '99: Proceedings of the 10th International Conference on Image Analysis and Processing.* Society, IEEE

Computer, Washington, DC, USA, 1999.

[27] Gross, Ralph, and Jianbo Shi. "The CMU Motion of Body (MoBo) Database." *Robotics Institute, Carnegie Mellon University.* Pittsburgh, Pennsylvania, June 2001.

[28] Rodrigues, Halisson Rodrigo Fernandes. “Um novo algoritmo de otimização derivado dos métodos downhill simplex e simulated annealing para identificação de posturas humanas em tempo real.” *Dissertação de Mestrado no Departamento de Engenharia de Teleinformática, Universidade Federal do Ceará.* Fortaleza, Maio de 2008.

[29] Silva, Wylkson Pinheiro, Tibério Menezes Oliveira, José Marques Soares, e Paulo César Cortez. “Segmentação da Pele Humana em Imagens de Vídeo utilizando Wavelet e Redes Neurais.” *Anais do II Workshop de Visão Computacional.* São Carlos - SP, 2006.

[30] —. "The MPEG Handbook." *Elsevier 2nd edition.* 2004.

[31] Preda, M., and F. Prêteux. "MPEG-4 Human Virtual Body Animation." *In WALSH, A. E., BOURGES-SÉVENIER, M., MPEG-4 Jump-Start.* Upper Saddle River: Prentice Hall, 2002. 460p.

[32] *Web3D Consortium, Information technology – Computer graphics and image processing – Humanoid animation (H-Anim), ISO/IEC FCD 19774:200x. Disponível em: http://h-anim.org (Data de consulta: 28/04/2004).*

[33] Phung, S. L., Chai, D., and A. Bouzerdoum. "Adaptive skin segmentation in color images." *In ICASSP’03, editor, IEEE International Conference on Acoustics, Speech, and Signal Processing.* Vol. 3. April 2003. pages 353–356.

[34] Wong, K., K. Lam, and W. Siu. "An efficient color compensation scheme for skin color segmentation." *In International Symposium on Circuits and Systems - ISCAS’03.* Vol. 2. 2003. pages 676–679.

[35] Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Verrerling. "Numerical Recipes in C, the Art of Scientific Computing." *Cambridge University Press.* 1988. p. 735.

[36] Nelder, J. A., and R. Mead. "A simplex method for function minimisation." *Computer Journal.* Vol. 7. 1965. 308-313.

[37] Mckinnon, K. I. M. "Convergence of the Nelder-Mead simplex Method to a Non-Stationary Point." *SIAM Journal on Optimization.* 1996.

[38] Soares, J. Marques, P. Horain, e A. Bideau. “Sharing and immersing applications in a 3D virtual world.” *Laval Virtual 5th virtual reality international conference (VRIC 2003).* Laval, France, 13-18 May 2003. pp. 27-31.

[39] Gupta, Abhinav, Anurag Mittal, and Larry S. Davis. "Constraint Integration for Efficient Multiview Pose Estimation with Self-Occlusions." *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Vols. 30, Issue 3. Univ. of Maryland, College Park, March 2008. On page(s): 493-506.

[40] Anam-dong, and Seongbuk-gu. "Automatic Gesture Recognition for Intelligent Human-Robot Interaction." *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition.* Southampton, UK, April 10-12, 2006.

[41] Zhao, Jianhui, Ling Li, and Kwoh Chee Keong. "Motion Recovery Based On Feature Extraction From 2D Images." *Computer Vision and Graphics - Computational Imaging and Vision.* Vol. 32. Springer Netherlands, Saturday, March 11, 2006. pg 1075-1081.

[42] Vilariño, D. L., and Cs. Rekeczky. "Analogic CNN Algorithms Implementing Pixel-Level Snakes: Experiments on the ACE4K Chip Operating within the ACE-BOX Computational Infrastructure." *MTA SZTAKI Technical Report, DNS-7-2003.* Budapest, 2003.

[43] Zhiwei, Jiang, Yi Wensheng, Zhao Xiaoming, and Yao Min. "Applications of Generalized Learning in Image Recognition." *2005 First International Conference on Neural Interface and Control Proceedings.* Wuhan, China, May 2005. pg. 26-28.

[44] *Intel Corporation. Open Source Computer Vision Library – OpenCV.* Disponível em: http://www.intel.com/research/mrl/research/opencv (Data de consulta: 10/06/2008).

[45] Rost, Randi J. "OpenGL Shading Language." *Addison Wesley.* 2004.

[46] NVidia. "CUDA Zone." *Disponível em: http://www.nvidia.com/cuda (Data de consulta: 05/10/2008).*

[47] Press, W. H., S. A. Teukolsky, W. T. Vetterling, and Flannery B. P. *Numerical Recipes in C.* Second Edition. London: Cambridge University Press, 1992.

[48] Suzuki, S., and K. Abe. "Topological Structural Analysis of Digital Binary." *CVGIP.* Vols. 30, n.1. 1985. pp. 32-46.

[49] Ferreira, Carlos. "Avaliação Quantitativa de um Método Automático de Extracção de Contornos Em Tomogramas Pulmonares." *VIII Jornadas de Classificação e Análise de Dados.* Porto, 2001.

[50] Ribeiro, Fábio Cisne, Adail Nunes da Silva, José Marques Soares, Giovanni Cordeiro Barroso, e Paulo César Cortez. “Métricas de Avaliação de Similaridade para Identificar Posturas Humanas em Imagens de Vídeo.” *Anais do II Workshop de Visão Computacional.* São Carlos - SP, 2006.

[51] Ribeiro, Fábio Cisne, Halisson R. Rodrigues, José Marques Soares, e Paulo César Cortez. “Otimização Multidimensional com Restrições Biomecânicas Dinâmicas para Identificação de Posturas Humanas por Visão Computacional em Tempo Real.” *Anais do XVII Congresso Brasileiro de Automática.* Juiz de Fora - MG, 2008.

[52] Ribeiro, Fábio Cisne, Adail Nunes da Silva, José Marques Soares, Giovanni Cordeiro Barroso, e Paulo César Cortez. “Métricas de avaliação de similaridade baseadas em superfície de não-recobrimento e distância entre contornos para identificação de posturas humanas.” *Anais do VI Encontro de Pesquisa e Pós Graduação do Cefet Ce.* Fortaleza - CE, 2006.

[53] Spendley, W., G. R. Hext, et F. R. Himsworth. «Sequencial Application of Simplex Designs in Optimization and Evolutionary Operation.» *Technometrics.* 1962. 441-461.

[54] Walters, F. H. et al. «Sequencial simplex optimization.» *CRC Press LLC.* Corporate Blvd., N. W., Boca Raton, Florida 333431, 2000.

APÊNDICE A – Movimentos articulares do humanóide virtual

Os movimentos articulares podem ser observados através de planos imaginários e em eixos perpendiculares ao movimento. Por convenção, tais movimentos são definidos com relação à posição anatômica (ou posição de repouso). Nesta posição, o corpo do humanóide virtual é referenciado de acordo com três planos mutuamente ortogonais, conforme ilustrado na Figura A-1.

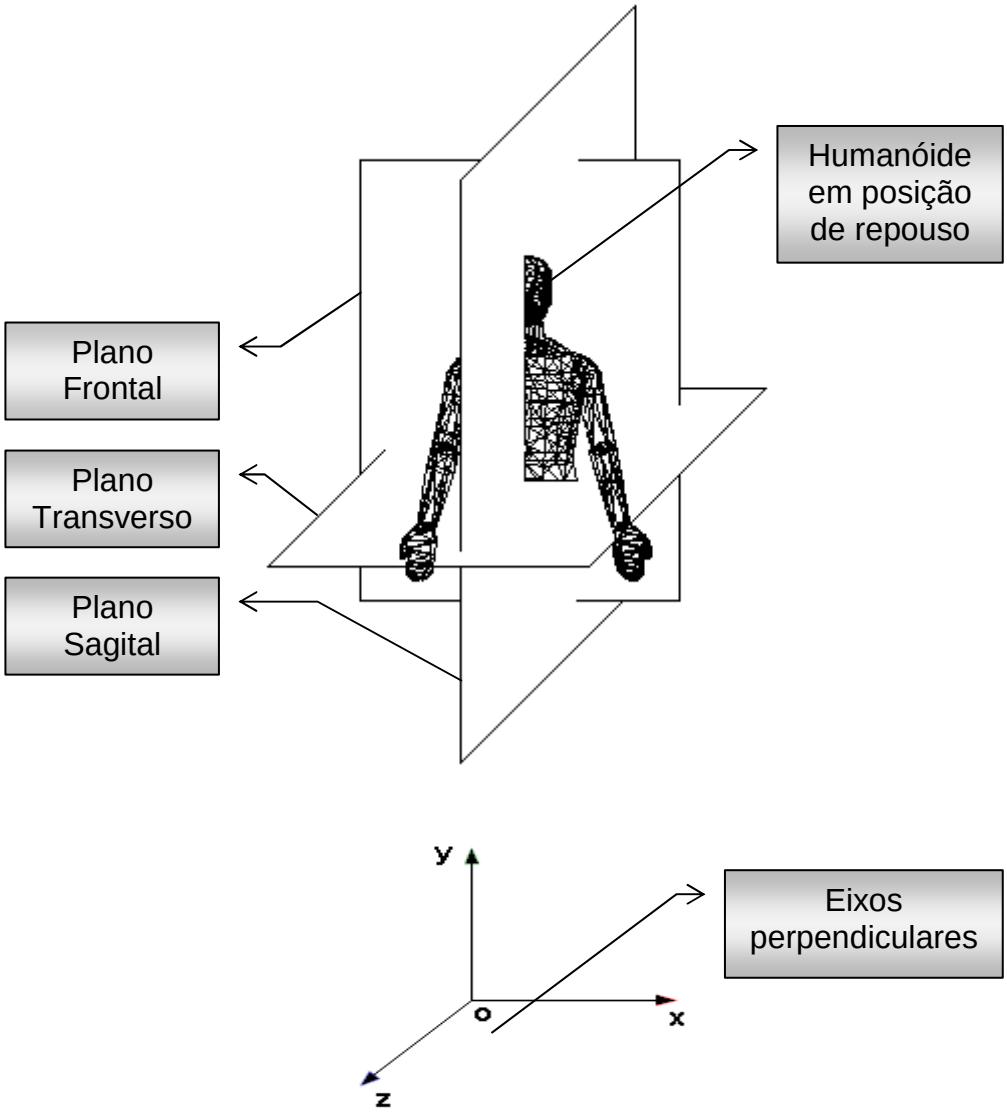

Figura A-1: eixos e planos ortogonais usados como referência para os movimentos articulares do humanoide virtual.

Para os exemplos mostrados a seguir são considerados como referenciais os planos: frontal, horizontal, sagital. O movimento é feito com o braço direito, através da articulação do ombro, de um humanóide que está de frente para o leitor. Na Figura A-1 o plano ***yz*** coincide com o plano sagital, o plano ***xy*** coincide com o plano frontal, e o plano ***xz*** coincide com o plano transverso.

O **plano Sagital** divide o corpo simetricamente em partes direita e esquerda. As ações articulares ocorrem em torno de um eixo horizontal ou transversal ***x*** e incluem os movimentos de flexão e extensão. No exemplo da Figura A-2 o movimento pode ser observado usando como referência o plano sagital e o eixo ***x*** (ver Figura A-1). Seguindo a seqüência 1→2→3→4→5→6 tem-se uma flexão; já a seqüência inversa, ou seja, 6→5→4→3→2→1 tem-se uma extensão.

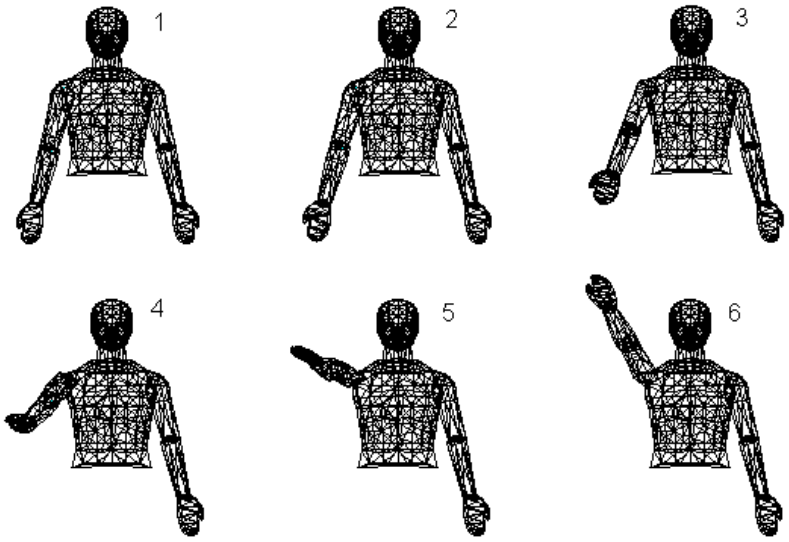

Figura A-2: movimento do braço: ascendente (flexão) e descendente (extensão).

O **plano Coronal** ou **Frontal** divide o corpo em partes anterior (ventral) e posterior (dorsal) e incluem os movimentos de abdução e adução. As ações articulares ocorrem em torno de um eixo ântero-posterior ***z***. No exemplo na Figura A-3, o movimento pode ser observado usando como referência o plano frontal e o eixo ***z*** (ver Figura A-1). Seguindo a seqüência 1→2→3→4→5→6, tem-se uma abdução; já a seqüência inversa, ou seja, 6→5→4→3→2→1 tem-se uma adução.

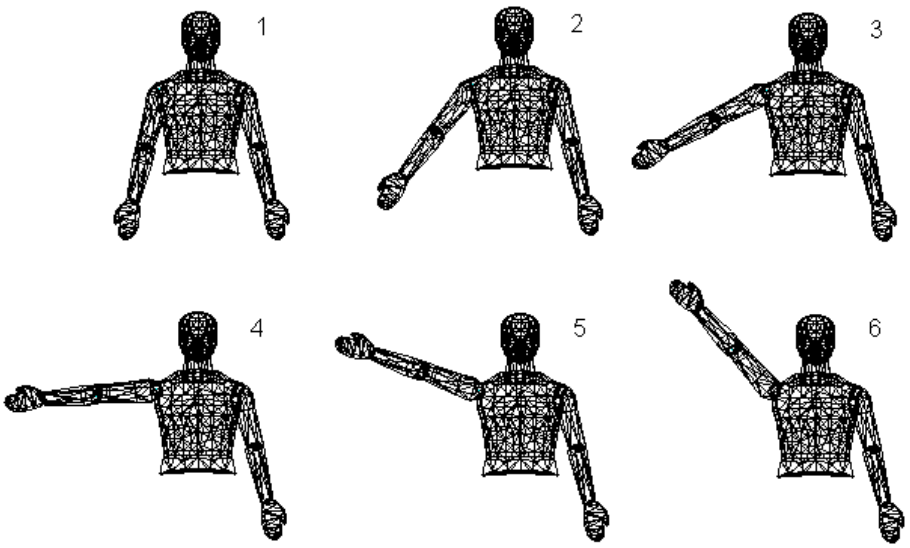

Figura A-3: movimento do braço: ascendente (abdução) e descendente (adução).

O **plano Transversal** ou **Horizontal** divide o corpo em partes superior (cranial) e inferior (caudal). As ações articulares ocorrem em torno de um eixo longitudinal ou vertical ***y***. No exemplo da Figura A-4, o movimento pode ser observado usando como referência o plano horizontal e o eixo ***y*** (ver Figura A-1). Seguindo a seqüência 1→2→3→4→5→6, tem-se uma rotação interna; já a seqüência inversa, ou seja, 6→5→4→3→2→1 tem-se uma rotação externa.

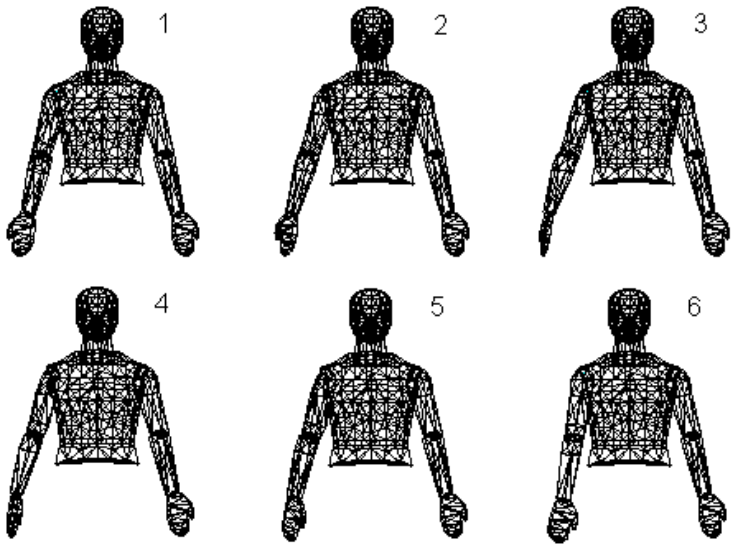

Figura A-4: movimento de rotação sobre a articulação do ombro direito.

Observando a rotação interna no início do movimento, Figura A-4(1), a palma da mão está voltada para frente, de forma a ficar visível ao leitor. No fim do

movimento, Figura A-4(6), a palma da mão está voltada para trás, escondida da visão do leitor.

Por se tratar de um modelo simplificado, determinados movimentos de algumas articulações não são executados, como os movimentos de rotação do antebraço: pronação e supinação. Mas é possível executar a flexão e extensão do antebraço, como pode ser visto na Figura A-5.

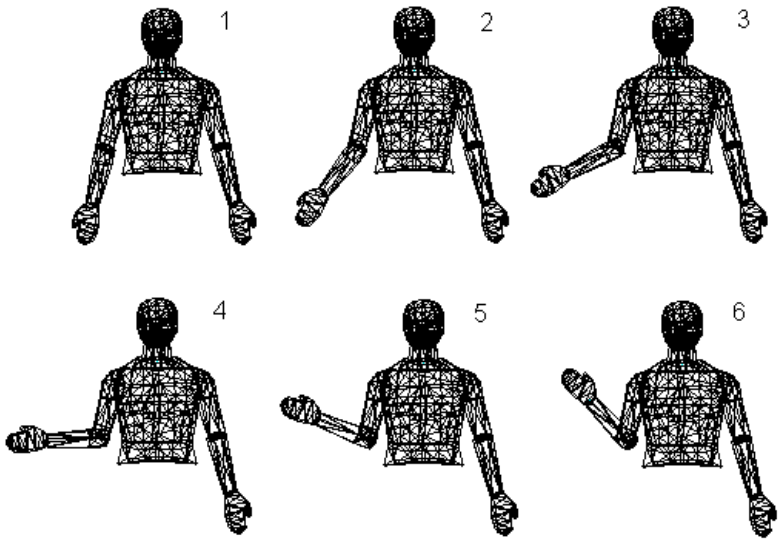

Figura A-5: movimento de flexão e extensão do antebraço.

Como os eixos sempre acompanham os movimentos da articulação, é fácil perceber a mudança no referencial, principalmente, quando se trata de uma composição de movimentos, por exemplo, uma flexão somada a uma abdução. Tomando como exemplo a articulação do ombro com os eixos de modo que sua origem ***o*** fique sobre o centro da articulação do ombro, conforme ilustra a Figura A-6.

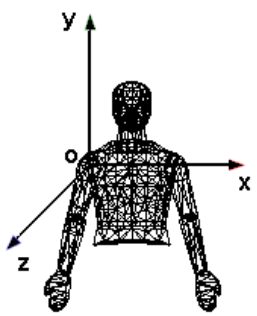

Figura A-6: eixo com o centro na articulação do ombro.

Imaginando o braço direito na mesma direção e sentido contrário ao indicado pela componente ***y***, ou seja, apontando para baixo, como mostrado na Figura A-6. Um movimento de ascensão do braço (flexão) de 90°, realizado em torno do eixo ***x***, tomando como referencial o plano ***xy***, muda as direções dos eixos ***y*** e ***z***, fazendo ***y*** se situar na mesma direção, e sentido contrário a ***z***, e ***z*** assume a posição de ***y***. Ao se executar um novo movimento, como a abdução, a partir desta posição, será levada em consideração a nova configuração dos eixos articulares.

APÊNDICE B –*Downhill Simplex* de *Nelder & Mead*

O algoritmo de otimização de Nelder-Mead [36], chamado de *Downhill Simplex*, é bastante difundido e utilizado na resolução de problemas de estimação de parâmetros. Ele pertence à classe dos métodos de busca direta que usam comparação de valores da função custo e não necessitam calcular nenhuma derivada. O algoritmo usa o conceito de *simplex*, cuja natureza geométrica introduz um nível de abstração intuitivo, facilitando a compreensão.

B.1 Definição de *simplex*

O *simplex* pode ser definido como uma figura geométrica, cujo número de vértices é igual ao número de dimensões mais um. Em outras palavras, se *n* é o número de dimensões do espaço, então, o *simplex* é definido como tendo *n+1* pontos. Por exemplo, em duas dimensões o *simplex* é representado por um triângulo no plano bidimensional; em três dimensões, um tetraedro no espaço tridimensional, etc. Um *simplex* de quatro ou mais dimensões deve ter cinco ou mais vértices. Neste caso, não é fácil mostrar um exemplo visual, devido ao seu elevado número de dimensões. Entretanto, suas propriedades são análogas às propriedades dos *simplexes* que podem ser visualizados.

Os *simplexes* com duas ou três dimensões são muito usados na literatura como ferramenta didática para explicar o método e para mostrar sua convergência. Os *simplexes* de zero e uma dimensão apresentam pouca relevância didática para este trabalho, portanto não fazem parte das explicações.

Cada vértice x_i dos *simplexes* na Figura B-1 representa um conjunto de parâmetros de uma condição experimental, de um estado do sistema ou solução do problema. É através de seus vértices que se determina a direção e a extensão do movimento de transformação do *simplex*, definindo assim os passos de busca do algoritmo.

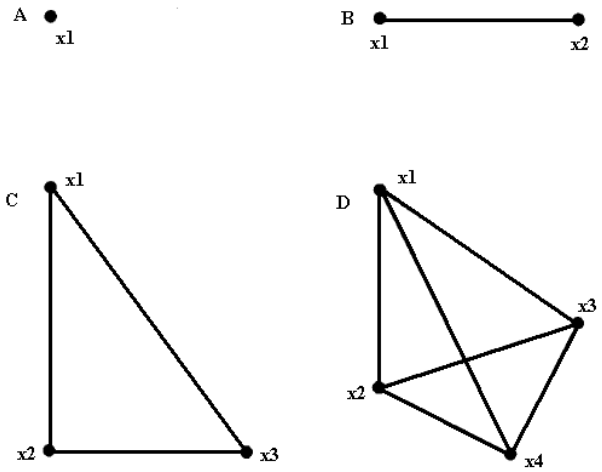

Figura B-1: *simplex* em dimensão zero (A), em uma dimensão (B), em duas dimensões (C), em três dimensões (D).

B.2 Funcionamento do algoritmo

Durante cada iteração, o algoritmo tenta melhorar os valores dos parâmetros modificando o *simplex*, usando como ponto de partida o vértice com o pior resultado. Isto é realizado por meio de transformações geométricas, como: reflexão, expansão e contração. O *simplex* move-se através do espaço de parâmetros, buscando o mínimo, e, em um determinado momento, pode contrair a si mesmo até atingir a aproximação desejada. O algoritmo *simplex* original contava apenas com o movimento de reflexão no qual retira um vértice (usualmente o pior, considerando o problema) e o projeta em direção ao centróide dos vértices restantes, criando assim um novo vértice no lado oposto. Este novo vértice, assim como os demais, guarda uma configuração de parâmetros que gera uma das soluções do problema, como é mostrado na Figura B-2. A principal contribuição de Nelder-Mead em relação ao algoritmo *simplex* original de Spendlry *et al*. [53] é a modificação que permite ao *simplex* expandir-se em direções favoráveis e se contrair em direções desfavoráveis. Por causa desta característica, alguns autores usam o termo “*simplex* de tamanho variável” em contraste com o seu predecessor o “*simplex* de tamanho fixo” [54]. A reflexão é o movimento chave do método. Observa-se que as distâncias de *W* a *P* e de *P* a *R* devem ser as mesmas.

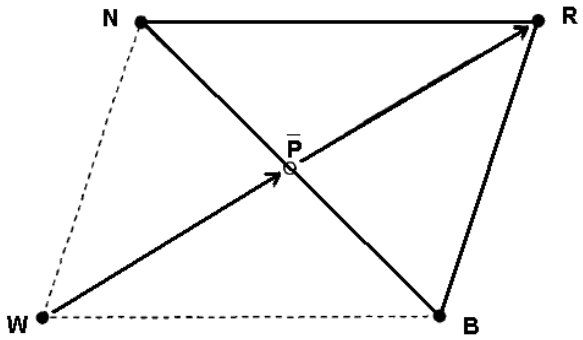

Figura B-2: geração do vértice R pela projeção de W em direção a P.

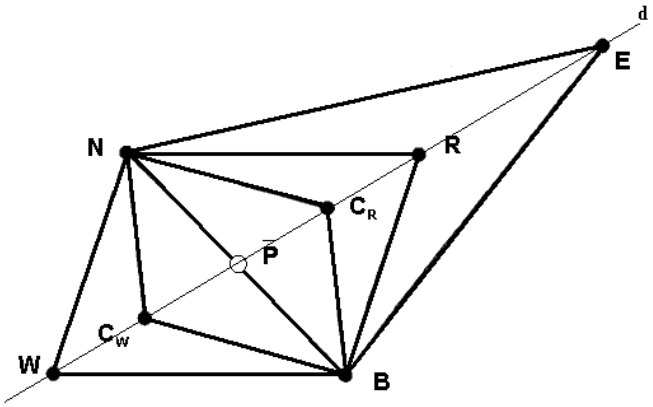

Figura B-3: os possíveis movimentos do *simplex* W-N-B.

Os movimentos do *simplex*, segundo Walters *et al*. [54], para um *simplex* no $\Re^2$ (espaço bidimensional no domínio dos reais) de tamanho variável formado pelos vértices *W*, *N* e *B* podem ser vistos na Figura B-3. As linhas são usadas apenas para visualizar o *simplex* e suas transformações; *d* é a reta que passa por *P* e *W*, indicando a direção da projeção do ponto *W* em relação a *P*. Neste caso, *W* é o pior vértice, ou seja, o que detém uma configuração de parâmetros com o pior resultado; *N* é o segundo pior e *B* é o melhor vértice; *R* é a reflexão de *W*, em relação a *P*; *E* é a expansão de *R*; C_R é a contração de *R*; C_W é a contração de *W*; *P* é o centróide formado por todos os vértices com exceção de *W*. Calcula-se *P* usando-se a média das coordenadas de todos os pontos com exceção de *W*. No exemplo da Figura B-3, *P* é o ponto médio de *N* e *B* e é usado como referência para os movimentos possíveis a partir de *W* ou de *R*.

Outro exemplo, seguindo o mesmo raciocínio usado no exemplo anterior, é mostrado na Figura B-4. Este exemplo mostra como o algoritmo realiza os passos seguindo a mesma idéia do exemplo anterior.

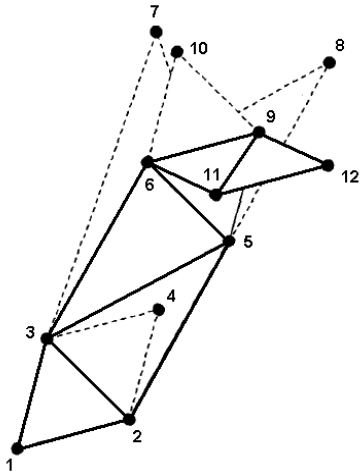

Figura B-4: exemplo de passos do algoritmo.

Para este *simplex*, Figura B-4, formado pelos vértices *1*, *2* e *3*, descrevem-se os passos do algoritmo, incluindo a numeração nos vértices para ajudar na percepção da seqüência de passos executada. Inicialmente, os seus vértices são classificados de acordo com o resultado da avaliação de seus parâmetros. O algoritmo identifica o pior, o intermediário e o melhor vértice. Após a classificação dos vértices, tenta-se a reflexão do pior ponto *W*, obtendo-se *R*. Se *R* for melhor do que *B* tem-se uma indicação clara de se estar indo em uma direção favorável. Por isto, faz sentido experimentar uma projeção mais distante, resultando em um movimento chamado de expansão que é representado pela letra *E*.

Nelder-Mead sugerem realizar o movimento de expansão com duas vezes à distância de *P* a *R*. No exemplo mostrado na Figura B-4, se admite conhecida a classificação dos vértices do *simplex*. Assim, após a reflexão de *1* (*W*), obtém-se *4* (*R*), como *4* é melhor do que *3* (*B*), então, tenta-se a expansão *5* (*E*). Agora, têm-se duas situações:

1. se a expansão *E* for melhor ou igual a *B*, usa-se o *simplex BNE*;
2. se a expansão *E* for pior que *B*, usa-se o *simplex BNR*.

No primeiro caso, obtém-se o *simplex 2-3-5,* resultante desta transformação. No segundo caso, produz-se o *simplex 2-3-4*. O próximo movimento é a reflexão de *2*, obtendo *6,* seguida por uma expansão até *7*. Como a expansão falha, o *simplex* resultante é *3-5-6*. Seguindo este raciocínio, o movimento seguinte é uma reflexão de *3*, que não apresenta bom resultado, vértice *8*, mas não é tão ruim quanto *3*.

Assim, executa-se uma contração C_R, obtendo-se o *simplex* BNC_R ou, pela numeração *5-6-9*. O próximo movimento é a reflexão de *5*, resultando no ponto *10*. Como *10* é pior do que *5*, executa-se uma contração $C_{W,}$ obtendo-se o *simplex* BNC_W ou, pela numeração *6-9-11*. Por fim, tem-se uma reflexão do vértice *6*, obtendo o *simplex 9-11-12*.

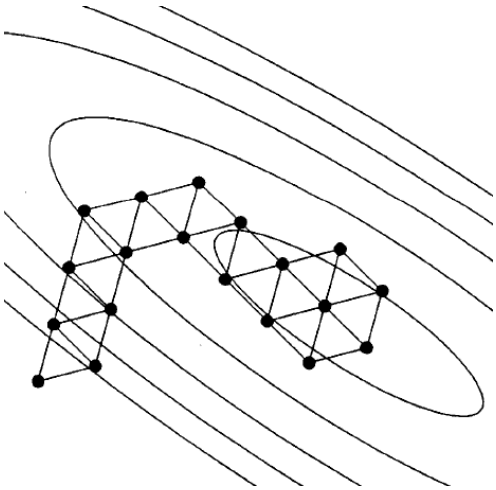

Figura B-5: o "*simplex* de tamanho fixo" executa a convergência usando somente o movimento de reflexão.

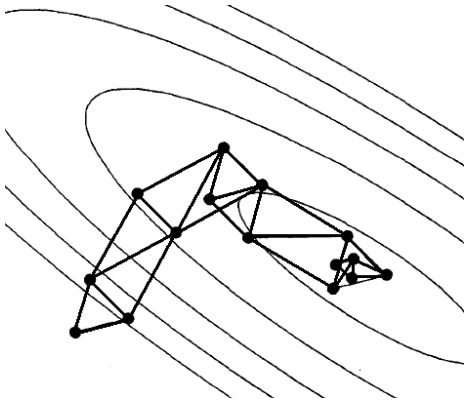

Figura B-6: o "*simplex* de tamanho variável" executa a convergência usando passos de reflexão, expansão e contração.

As execuções dos dois algoritmos, o *simplex* de tamanho fixo e o de tamanho variável, aplicados a um mesmo problema com a mesma inicialização estão ilustrados nas Figura B-5 e 7-6.

A implementação sugerida por Press *et al*. [35] segue os mesmos princípios estabelecidos por Walters *et al.* [54], porém com uma ligeira diferença no movimento

de contração. Em resumo, o funcionamento do algoritmo e as diferenças de cada implementação são descritas a seguir.

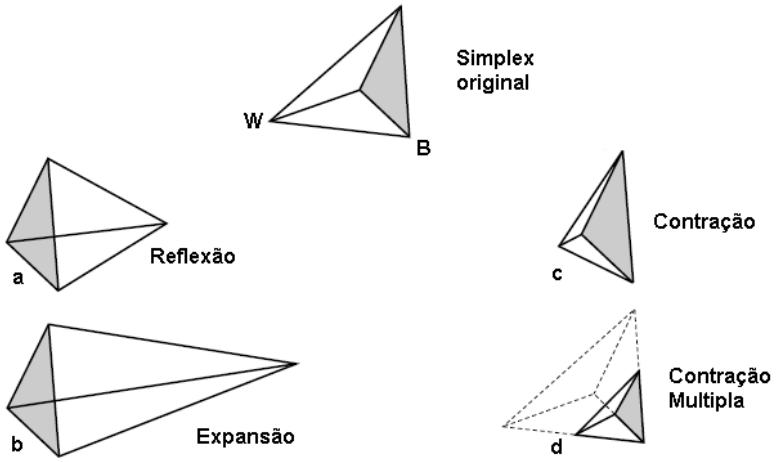

Figura B-7: possíveis movimentos do *simplex* segundo Press *et al*. [35]

No *simplex* original da Figura B-7 tem-se: *B* e *W* são o melhor e o pior vértice, respectivamente; em "*a"* tem-se a reflexão de *W*; em "*b"*, uma reflexão de *W*, seguida de uma expansão; em "*c"* tem-se uma contração de *W*; em "*d"* tem-se uma contração de todos os demais vértices em direção ao vértice *B*.

O algoritmo começa com um *simplex* inicial e, depois de classificar os vértices, tenta descartar o pior deles. Para isto, lança mão dos movimentos ou transformações do *simplex,* como os ilustrados na Figura B-7, e emprega as seguintes regras:

1. Reflete-se o *simplex* em direção à face oposta do pior vértice;

2. Se a reflexão cair em um vértice o qual é pior do que o melhor, mas é melhor do que o segundo pior, então usa-se o *simplex* formado por estes vértices: melhor, segundo pior e o vértice refletido;

3. Se a reflexão cair em um vértice o qual é melhor do que o melhor vértice do *simplex*, então se tenta uma expansão naquela direção;

4. Se a reflexão cair em um vértice que é pior do que o segundo pior vértice, então o *simplex* contrai-se em uma dimensão, movendo o pior ponto na direção do centróide;

5. Se esta contração ainda não melhora o segundo pior vértice, então se faz uma contração múltipla, movendo todos os vértices, com exceção do melhor, na direção do melhor vértice.

Estes passos são repetidos até um critério de parada ser satisfeito.

A diferença do algoritmo ora descrito [35], do apresentado anteriormente [54] estão nos passos *4* e *5*, no qual, em Walters *et al*. [54], o passo *5* não é realizado, e o *4* é executado da seguinte maneira: se a reflexão cair em um vértice que é pior do que o segundo pior vértice, então têm-se duas situações. Na primeira situação, se a reflexão for melhor do que o pior vértice, faz-se uma contração do vértice refletido e se usa o *simplex* formado pelos vértices melhor, segundo pior e a contração do vértice refletido. Na segunda situação, se a reflexão for pior do que o pior vértice, faz-se uma contração do pior vértice e se usa o *simplex* formado pelos vértices melhor, segundo pior e a contração do pior vértice.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo