UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Everton Simões da Motta

# Desenvolvimento de um método para a captura de movimentos humanos usando uma câmera RGB-D

São José do Rio Preto
2016

Everton Simões da Motta

# Desenvolvimento de um método para a captura de movimentos humanos usando uma câmera RGB-D

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Financiadora: CAPES - DS.

Orientador: Dr. Antonio Carlos Sementille

São José do Rio Preto
2016

Motta, Everton Simões da.
Desenvolvimento de um método para a captura de movimentos humanos usando uma câmera RGB-D / Everton Simões da Motta. -- São José do Rio Preto, 2016
68 f. : il., tabs.

Orientador: Antonio Carlos Sementille
Dissertação (mestrado) – Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Letras e Ciências Exatas

1. Computação - Matemática. 2. Processamento de imagens - Técnicas digitais. 3. Detectores óticos. 4. Movimento. 5. Algoritmos de computador. I. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. II. Título.

CDU – 518.72:76

Ficha catalográfica elaborada pela Biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

Everton Simões da Motta

# Desenvolvimento de um método para a captura de movimentos humanos usando uma câmera RGB-D

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Financiadora: CAPES - DS.

Comissão Examinadora

Dr. Antonio Carlos Sementille
UNESP - Bauru
Orientador

Dr. Valter Vieira de Camargo
UFSCar - São Carlos

Dr. João Fernando Marar
UNESP - Bauru

São José do Rio Preto
22 de dezembro de 2016

*Dedico este trabalho a Deus, meu filho Gabriel, minha esposa Flávia, meus pais e aos meus irmãos.*

*In memoriam: Às minhas avós Benedita e Leonor, que me proporcionaram bons momentos. Ao meu filho Miguel, que me foi tirado o privilégio de vê-lo crescer e ensiná-lo as coisas simples da vida.*

# Agradecimentos

Agradeço a Deus por me dar força em momentos difíceis, à minha amada esposa Flávia que está sempre ao meu lado e de quem tenho muito orgulho, ao meu filho Gabriel por me oferecer tantos momentos felizes, aos meus pais Iracema e Ademar que demonstraram amor e me ajudaram de muitas formas.

A minha irmã Elga, Suelen e Adriana (e seu esposo Emerson) que proporcionam conversas agradáveis sempre tentando me motivar, mostrando como querem meu bem. As minhas irmãs Andrea e Eliane, que apesar da distância também torcem muito por mim.

Ao meu sobrinho Bruno pelas conversas nas madrugadas, ao meu irmão Ademar Júnior que desde pequeno tenho ele como exemplo. A minha tia Ana e toda sua família por me receberem tão bem em Bauru.

A minha sogra Elisabete pela ajuda e carinho e a todos os outros familiares que depositam tanta confiança em mim, como: Fábio, Fernando, Cibele, Adriely, Danilo, etc.

Aos meus amigos Rodrigo e Francine que já fazem parte de nossa família. Aos meus padrinhos Elder e Geilza e minha comadre Aninha.

Aos meu colegas do Laboratório SACI: Marcelo, Rafael, Tiago, Calos Cubas, Danilo Salioni, em especial ao meu amigo Ivan e sua família.

Aos professores Aparecido Nilceu Marana, João Paulo Papa, João Fernando Marar, Ildeberto Aparecido Rodello e não posso esquecer do Valter Vieira de Camargo.

Por fim ao meu orientador Dr. Antonio Carlor Sementille pela paciência e sabedoria. Além de ser um excelente orientador, mostrou ser uma pessoa incrível em todos os aspectos.

*"Não espere por grandes líderes;*
*faça você mesmo, pessoa a pessoa.*
*Seja leal às ações pequenas porque*
*é nelas que está a sua força.*
*(Madre Teresa de Calcutá)*

# Resumo

Sistemas de captura de movimentos humanos vêm sendo cada dia mais estudados, tanto pela área de Visão Computacional, quanto por grandes empresas do setor de entretenimento. São sistemas capazes de rastrear a posição e orientação das articulações do corpo e sua trajetória no espaço durante um intervalo de tempo. São utilizados em diversas aplicações, tais como em jogos digitais, animação de personagens virtuais para cinema e televisão, reconhecimento gestual, medicina de reabilitação, e outras. O surgimento de novos dispositivos de baixo custo e boa resolução que fornecem informações de profundidade, tem motivado novas pesquisas para a utilização dos mesmos. No entanto, sistemas que se baseiam somente em informações de profundidade (geralmente sistemas em tempo real) não apresentam uma alta acurácia no rastreamento do movimento. Considerando este contexto, o presente trabalho teve como objetivo principal, o desenvolvimento de um método de captura de movimentos humanos utilizando único sensor RGB-D, combinando informações de textura da imagem e de profundidade, que são associados a um esqueleto virtual, conseguindo-se uma maior acurácia em comparação com os métodos baseados apenas em profundidade. Tal método não visa aplicações de tempo real, mas sim, uma maior acurácia em comparação com os métodos baseados apenas em profundidade.

**Palavras-chaves:** Captura de movimentos, câmeras RGB-D, sistemas ópticos de captura, esqueletos virtuais.

# Abstract

Human motion capture systems are being increasingly studied, in the area of computer vision and also by major entertainment industries. These systems are able to track the position and orientation of joints of the body and its trajectory in space over a period of time. They are used in various applications such as digital games, animation of virtual characters for film and television, gesture recognition, medical rehabilitation, etc. The emergence of new low-cost and good resolution devices that provide depth information has prompted new research. However, systems that are based only on depth information (usually real-time systems) do not present a high accuracy in movement tracking. Considering this context, this thesis project presents the development of a method for capturing human movements using only one RGB-D sensor, combining captured texture from the image and depth information in order to obtain a higher accuracy, which are associated with a virtual skeleton. This method is not intended for real-time applications, but those that require greater accuracy, such as the animation of virtual characters in video and medical rehabilitation.

**Key-words**: Motion capture, RGB-D cameras, optical capture systems, virtual skeletons.

# Lista de figuras


Figura 1 – Exoesqueleto Gypsy da empresa Meta Motion . . . . . 19
Figura 2 – Rastreador Eletromagnético “Flock of Birds” . . . . . 20
Figura 3 – Sistema de Posicionamento Ultassonico Hx19 da empresa Hexamite . . 21
Figura 4 – Sistema de captura de movimentos óptico da empresa PhaseSpace . . 22
Figura 5 – Sistema inercial IGS-190 da empresa Meta Motion . . . . . 23
Figura 6 – Fragmentos dos arquivos ASF e AMC. . . . . 25
Figura 7 – Fragmento de arquivo BVA . . . . . 27
Figura 8 – Exemplo de um esqueleto humanoide com marcadores e representação hierárquica de arquivo BVH. . . . . 28
Figura 9 – Princípio do método de cálculo de triangulação de imagens para a obtenção do mapa de profundidade. . . . . 30
Figura 10 – Como são geradas regiões de sombras no mapa de profundidade utilizando luz estruturada . . . . . 31
Figura 11 – Comportamento dos feixes de luz infravermelha projetadas em superfícies de diferentes materiais; . . . . . 32
Figura 12 – Visão geral do funcionamento do sensor ToF . . . . . 33
Figura 13 – Linha do tempo com sensores de luz estruturada (azul) e sensores ToF (preto). . . . . 34
Figura 14 – Especificações de hardware dos modelos do sensor Kinect . . . . . 35
Figura 15 – Classificação dos métodos de captura de movimentos humanos . . . . . 36
Figura 16 – Divisão de métodos MoCap com único sensor de profundidade . . . . . 37
Figura 17 – *Pipeline* do método de MoCap usado no Kinect . . . . . 39
Figura 18 – Diferença entre abordagem que utiliza deformações daquela que utiliza combinação de partes . . . . . 40
Figura 19 – Pipeline do método de Yang e Ramanan (2013) . . . . . 41
Figura 20 – Divisão das Partes do corpo (YANG; RAMANAN, 2013) . . . . . 42
Figura 21 – Agrupamento dos métodos de MoCap de acordo com o tipo de informação utilizada. . . . . 43
Figura 22 – Informações utilizadas pelo método desenvolvido. . . . . 44
Figura 23 – Pipeline do Método Desenvolvido . . . . . 45
Figura 24 – Esqueleto suportado pelo Kinect v2 . . . . . 46
Figura 25 – Imagem capturada no processo de calibração . . . . . 48
Figura 26 – Fluxograma de captura de dados usando o Kinect v2. . . . . 49
Figura 27 – Estrutura de dados que armazena as informações capturadas pelo dispositivo kinect . . . . . 50
Figura 28 – Fluxograma da extração de fundo da imagem RGB . . . . . 51

Figura 29 – Etapas do método de Yang e Ramanan (2013) utilizadas no método desenvolvido. . . . . 52
Figura 30 – Matriz que armazena dados das partes do corpo identificadas na imagem RGB. . . . . 52
Figura 31 – Fluxograma da Detecção de partes do corpo . . . . . 53
Figura 32 – Dados das articulações inferidas para o esqueleto virtual. . . . . 54
Figura 33 – Informações para encontrar o eixo z das articulações . . . . . 55
Figura 34 – Fluxograma da extração de fundo da imagem RGB . . . . . 55
Figura 35 – Ambiente Experimental . . . . . 56
Figura 36 – Tripé com suporte especial para o Kinect v2. . . . . 57
Figura 37 – Erros apresentados pelo Kinect . . . . . 58
Figura 38 – Interface da ferramenta criada para o posicionamento manual das articulações sobre a imagem do ator. . . . . 60
Figura 39 – Média da distância das articulações por *frame* . . . . . 62
Figura 40 – Média das articulações por *frame* . . . . . 62
Figura 41 – Avaliação visual das articulações. . . . . 63

# Lista de tabelas

Tabela 1 – Especificações das informações obtidas por meio do Kinect v2. . . . . 57
Tabela 2 – Resultados da aplicação do método com relação as imagens com e sem extração de fundo. . . . . 59
Tabela 3 – Sequência de marcação das articulações para o Ground Truth . . . . . 61

# Lista de abreviaturas e siglas

| | |
|---|---|
| 2D | Bidimensional |
| 3D | Tridimensional |
| AMC | *Acclaim Motion Capture data* |
| ASF | *Acclaim Skeleton File* |
| BVA | *BioVision Animation* |
| BVH | *BioVision Hierarchy* |
| FPS | *frames per second* |
| MoCap | *Motion Capture* |
| RGB | *Red, Green and Blue* |
| RGB-D | *Red, Green, Blue and Depth* |
| SACI | Sistemas Adaptativos e Computação Inteligente |
| SVM | *Support Vector Machines* |
| ToF | *Time of Flight* |
| V2 | Versão 2 |

# Sumário


**1 INTRODUÇÃO . . . . . 15**
**1.1 Motivação . . . . . 15**
**1.2 Objetivos . . . . . 16**
**1.3 Organização da Dissertação . . . . . 16**

**2 CAPTURA DE MOVIMENTOS . . . . . 18**
**2.1 Tipos de Sistemas de Captura . . . . . 18**
2.1.1 Mecânicos . . . . . 18
2.1.2 Magnéticos . . . . . 20
2.1.3 Acústicos . . . . . 20
2.1.4 Ópticos . . . . . 21
2.1.5 Inerciais . . . . . 23
**2.2 Classificação Baseada em Sensores e Fontes . . . . . 24**
**2.3 Formatos de Arquivos . . . . . 24**
2.3.1 ASF/AMC . . . . . 25
2.3.2 BVA . . . . . 27
2.3.3 BVH . . . . . 27

**3 MÉTODOS E SENSORES RGB-D USADOS NA CAPTURA DE MOVIMENTOS . . . . . 29**
**3.1 Tecnologias Usadas em Dispositivos de Profundidade . . . . . 29**
3.1.1 Imageamento Estéreo . . . . . 29
3.1.2 Luz Estruturada . . . . . 30
3.1.3 ToF - *Time of Flight* . . . . . 32
**3.2 Dispositivos de Profundidade . . . . . 34**

**4 MÉTODOS DE CAPTURA DE MOVIMENTOS E DETECÇÃO DE POSE . . . . . 36**
**4.1 Classificação de Métodos de Captura de Movimentos Humanos . . 36**
**4.2 Trabalhos Correlatos . . . . . 37**

**5 MÉTODO PARA A CAPTURA DE MOVIMENTOS HUMANOS USANDO UMA CÂMERA RGB-D . . . . . 43**
**5.1 O Método de Captura de Movimentos Desenvolvido . . . . . 44**
**5.2 Implementação do Método . . . . . 47**
5.2.1 Materiais . . . . . 47
5.2.2 Processo de Calibração . . . . . 48

5.2.3 Captura dos dados . . . . . 48
5.2.4 Extração da Imagem de Fundo . . . . . 50
5.2.5 Detecção das partes do corpo do ator . . . . . 51
5.2.6 Geração do Esqueleto Virtual Tridimensional . . . . . 53

**6 TESTES E ANÁLISE DE RESULTADOS . . . . . 56**
**6.1 Configuração do Ambiente Experimental . . . . . 56**
**6.2 Captura e Validação das Informações . . . . . 57**
**6.3 Situações de Erro na Captura de Movimentos com o Kinect . . . . 58**
**6.4 Imagem de Entrada Com ou Sem Fundo . . . . . 58**
**6.5 Método Desenvolvido e o Usado pelo Kinect . . . . . 59**
6.5.1 Ground Truths . . . . . 60
6.5.2 Avaliação Qualitativa . . . . . 61
6.5.3 Avaliação Visual . . . . . 63

**7 CONCLUSÕES . . . . . 64**

**REFERÊNCIAS . . . . . 66**

# 1 INTRODUÇÃO

Nas últimas décadas, a investigação sobre captura de movimentos humanos no campo da Computação Gráfica, bem como na de Visão Computacional, tem despertado nos pesquisadores grande interesse.

Tais movimentos, quando reconstruídos em 3D, podem ser usados para diversas finalidades, como, por exemplo, nas indústrias de jogos eletrônicos e cinematográfica para a animação de personagens virtuais com base em dados de movimentos coletados de seres reais. Nas áreas ligadas a saúde, estes sistemas podem ser usados em estudos do movimento corporal e também auxiliar especialistas no diagnóstico e tratamento de diversas doenças. Podem ser usados também, na correção e otimização de movimentos de atletas visando a obtenção de melhores desempenhos. Os militares utilizam estes sistemas no treinamento de seus soldados e no desenvolvimento de equipamentos para auxiliá-los (FLAM, 2009).

A acurácia dos dados de movimentos capturados geralmente depende do tipo de aplicação em que serão utilizados. Aplicações como a animação de avatares em videogames normalmente exigem um desempenho em tempo real e aceitam uma menor acurácia com relação aos movimentos capturados.

No entanto, se a animação de personagens virtuais diz respeito à geração de um vídeo para Cinema, Televisão, ou Web, não há necessidade de desempenho em tempo real, pois a inserção de tais personagens geralmente ocorrerá na etapa de pós-produção (ou seja, após as filmagens e capturas de movimentos). Porém, neste último caso, a acurácia exigida quanto aos dados de movimentos é bem maior.

Os principais fatores que impactam a qualidade e acurácia da captura de movimentos normalmente são:

- As tecnologias de rastreamento 3D utilizadas; e

- Os métodos de interpretação dos dados capturados usados na inferência dos movimentos.

## 1.1 Motivação

Um dos grandes problemas na captura dos movimento humanos é a visualização realística de movimentos complexos. Em sistemas onde exige maior acurácia como na medicina de reabilitação ou em cinema, exigem uma visualização realística dos movimentos, visto que os movimentos de um ser humano envolvem dezenas de músculos e articulações.

Geralmente sistemas com alta acurrácia tem alto custo, porém, o aparecimento

recente de diversos dispositivos RGB-D de baixo custo e boa resolução, tais como, o Kinect versão 1 e 2 da Microsoft e o RealSense da Intel, entre outros, disperta o interesse para a utilização destes tipos de dispositivos para a criação de métodos de captura com menor custo.

Contudo, o desenvolvimento de um método que emprega equipamentos de baixo custo e com alta acurácia em relação aos métodos que utilizam somente informações de profundidades, contribuem para o campo da pesquisa desenvolvida.

## 1.2 Objetivos

O projeto teve como principal objetivo a elaboração de um método de captura de movimentos humanos, obtidos a partir de um único dispositivo RGB-D, que utilize tanto as informações de textura da imagem, quanto de profundidade, visando o aumento da acurácia dos movimentos inferidos em relação aos métodos que utilizam apenas informações de profundidade.

O objetivo específico referiu-se a implementação do método desenvolvido, por meio de um protótipo em software, viabilizando a comparação com os métodos que utilizam exclusivamente as informações de profundidade, a fim de comprovar sua eficácia.

## 1.3 Organização da Dissertação

O restante desta dissertação está organizada da forma que segue:

No capítulo 2 são apresentados os principais sistemas que envolvem a captura de movimentos, além da apresentação dos tipos de tecnologias de rastreamento usadas em suas aplicações. Nesse capítulo também são apresentados alguns dos formatos de arquivos usados para armazenar informações de movimento.

São apresentadas no capítulo 3 algumas das tecnologias empregadas em dispositivos RGB-D, em especial os dispositivos Microsoft Kinect versões 1 e 2.

Os principais métodos para a captura de movimentos e detecção de pose encontrados na literatura são abordados no capítulo 4. Dentre estes são descritos em maior detalhe os trabalhos de Shotton et al. (2011) e Yang e Ramanan (2013), por estarem diretamente relacionados ao método desenvolvido.

No capítulo 5, tem-se a descrição detalhada do método de captura de movimentos desenvolvido neste trabalho.

O capítulo 5.2 apresenta a implementação do protótipo do método descrito no capítulo anterior.

No capítulo 6 são apresentados os experimentos realizados para validação do método, bem como uma análise de seus resultados.

Por fim, no capítulo 7 tem-se a conclusão baseada em todo o processo de desenvolvimento do método e nos resultados obtidos. Também são apresentadas algumas considerações sobre desdobramentos futuros da pesquisa.

# 2 CAPTURA DE MOVIMENTOS

Neste capítulo são apresentados os principais tipos de sistemas de captura de movimentos, também denominados de MoCap (*Motion Capture*), bem como as tecnologias de rastreamento 3D empregadas em suas implementações e os formatos de arquivos mais usados no armazenamento da hierarquia de juntas de um esqueleto virtual e o movimento associado.

## 2.1 Tipos de Sistemas de Captura

Um sistema de captura de movimentos, de acordo com Menache (2000), deve ser capaz detectá-los, gravá-los e transformá-los, em termos matemáticos, por meio do rastreamento de uma certa quantidade de pontos chaves no espaço ao longo do tempo, combinando esse dados para se obter uma única representação tridimensional de todo o movimento. De modo geral, a captura de movimentos tem como objetivo gravar movimentos reais, que são difíceis de serem simulados por meios de programação, podendo ser aplicados em um modelo virtual e serem visualizados em duas ou três dimensões. Este tipo de sistema é usado em diversas áreas, tais como: militar, médica, esportes, entretenimento, cinema, entre outras.

Existem diversas maneiras de realizar a capturar movimentos, alguns sistemas utilizam câmeras de vídeo para captar visões do evento, outros utilizam campos eletromagnéticos, ultrassom, ou são baseados em estruturas rígidas com potenciômetros ligadas ao corpo.

São encontradas na literatura diferentes formas de classificação dos métodos de MoCap: síncronos ou assíncronos, ativos ou passivos, com ou sem marcadores, ou conforme os princípios físicos adotados.

São descritos, nesta seção, alguns tipos de sistemas de captura de movimentos classificados de acordo com os princípios físicos empregados. ((FURNISS, 1999) e (MENACHE, 2000)).

### 2.1.1 Mecânicos

Sistemas MoCap mecânicos são muitas vezes referidos como sistemas com exoesqueleto, devido a estrutura que é fixada ao ator. Normalmente são estruturas construídas de metal ou plástico rígidos, possuem articulações com potenciômetros, sensores de dobra ou codificadores de eixos para fornecer dados de orientação e posição de cada uma delas. Esses tipos de sistemas oferecem altas taxas de amostragem, alto grau de precisão, baixa

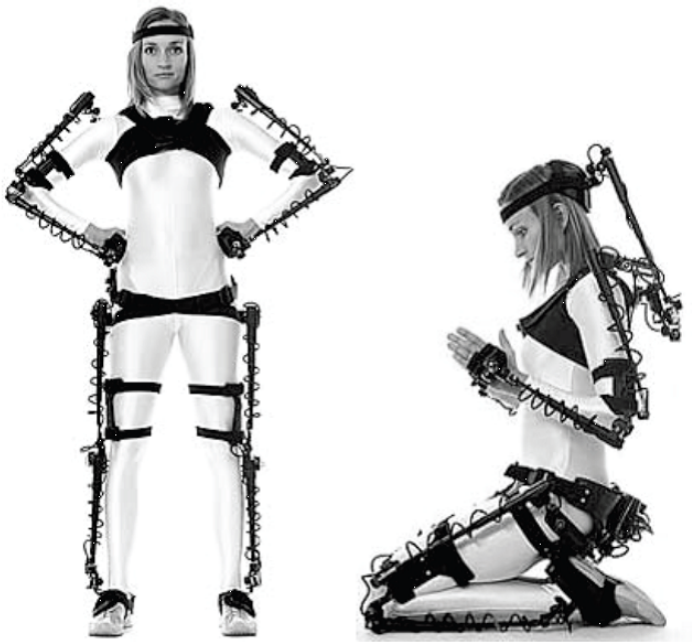
latência, exigem pouco ou quase nenhum processo de calibragem e não são afetados por interferências externas.

Além disso, alguns deles podem ser criados com a característica de gerar força inversa ao movimento ou *"force feedback"*, por exemplo, quando o ator segura um objeto virtual, é gerada uma força contrária ao movimento dos dedos, dando ao mesmo a sensação de estar segurando um objeto real, melhorando a imersão e o realismo.

Uma das desvantagens encontradas neste tipo de sistema é que o ator fica limitado aos graus de liberdade da estrutura. Se ele tentar realizar movimentos fora dos graus de liberdade, além de não conseguir, poderá danificar o equipamento. Outro problema é que um exoesqueleto pode ser volumoso ou até mesmo pesado, gerando desconforto e podendo ser visualmente perturbador. A mobilidade do ator também é prejudicada devido aos fios que compõem o sistema, assim ficando com um espaço pequeno para trabalhar.

Atualmente alguns modelos como o "Gypsy 5" da Meta Motion (2015) mostrado na Figura 1 são desenvolvidos utilizando tecnologia sem fio, desta forma, aumentando a mobilidade e a área de trabalho. Também são fabricados com materiais mais leves, fazendo com que o volume e o desconforto gerado ao ator seja menor.

Figura 1 – Exoesqueleto Gypsy da empresa Meta Motion



Fonte: Meta Motion (2015)

### 2.1.2 Magnéticos

Este tipo de tecnologia tem a função de medir o campo magnético gerado por um transmissor de baixa frequência. Os receptores são fixados nas articulações desejadas no corpo do usuário, sendo que cada um deles calcula sua posição e orientação no espaço em relação ao transmissor. Existem dois tipos de sistemas magnéticos: os de corrente contínua, que utilizam ondas quadradas; e os de corrente alternada, que utilizam ondas senoidais.

Sistemas magnéticos são caracterizados por sua velocidade de processamento, por sua alta precisão, e por serem comumente usados para aplicações de tempo real. No entanto, podem sofrer interferências causadas por objetos metálicos, campos magnéticos e fontes de energia elétrica, até mesmo por paredes e pisos que possuem estrutura de ferro. Os cabos que ligam os receptores aos transmissores limitam os movimentos e espaço de trabalho do usuário. Outra desvantagem é que o volume de captura é menor em comparação com outros sistemas.

Um grande usuário de sistemas MoCap com tecnologia magnética são os militares, na fabricação de equipamentos que exitem alta precisão. A medicina é outro campo que faz um uso considerável deste tipo de sistema, como por exemplo, equipamentos utilizados no tratamento minimamente invasivo de doenças vasculares, urológicas e em procedimentos cardíacos. Os sensores apresentados na Figura 2 criados por Ascension Technology Corporation (2015), são suficientemente pequenos para que possam ser inseridos dentro do corpo, ou serem anexados em instrumentos médicos. Além de seu tamanho reduzido estes sensores podem ser rastreados em tempo real.

Figura 2 – Rastreador Eletromagnético “Flock of Birds”


Fonte: Ascension Technology Corporation (2015)

### 2.1.3 Acústicos

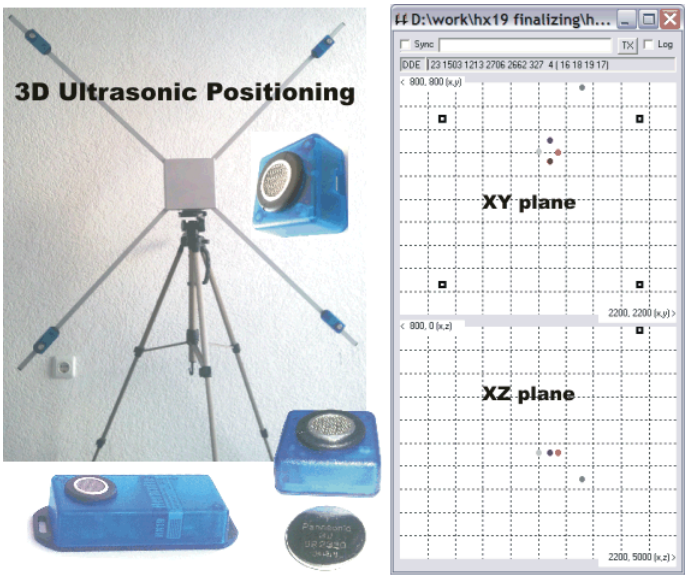
Esse tipo de tecnologia de rastreamento utiliza transmissão e recepção de ondas sonoras de alta frequência (ultrassônica). Cada transmissor emite um som característico

de forma sequencial e quando são captados pelos receptores, a posição no espaço pode ser calculada, medindo-se o tempo percorrido pelo som, e a orientação através de triangulação das distâncias.

Entre as vantagens deste tipo de sistema tem-se: a área de captura pode ser relativamente grande; o uso de equipamentos leves; a oclusão dos componentes não é um problema; não sofrem interferência magnética ou elétrica. E com relação às desvantagens tem-se: perda da precisão em ambientes que refletem as ondas sonoras ou por alguns sons gerados por outros objetos, apresenta uma baixa taxa de amostragem e suporta um número restrito de marcadores.

Sistemas exclusivamente acústicos podem ser encontrados na área de vigilância, tendo em vista, sua capacidade de monitorar grandes áreas e a utilização de um amplo número de marcadores não simultâneos. O sistema da empresa Hexamite LTD (2015), mostrado na Figura 3, limita seu espaço de cobertura pelo número de componentes empregados.

Figura 3 – Sistema de Posicionamento Ultassonico Hx19 da empresa Hexamite




Fonte: Hexamite LTD (2015)

### 2.1.4 Ópticos

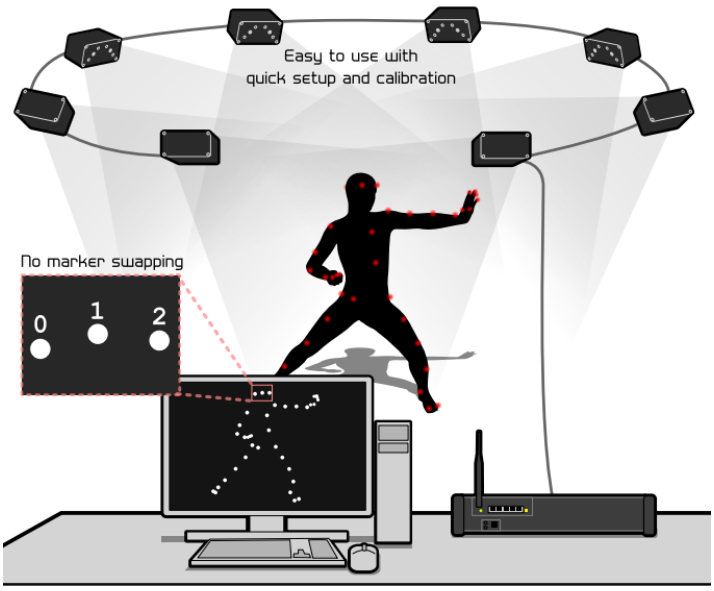
Os sistemas MoCap ópticos empregam um conjunto de câmeras que devem ser sincronizadas entre si. Pode-se usar marcadores, que deverão ficar presos nas articulações do usuário. Caso sejam marcadores passivos, são usadas câmeras que emitem luzes em uma determinada frequência que será refletida pelos marcadores. Já com marcadores ativos,

eles próprios emitem luz e as câmeras apenas a captam. Também existe a possibilidade de não usar marcadores, assim, os pontos de interesse podem ser definidos por meio de métodos de reconhecimento de silhueta ou componentes estruturais como cabeça, braços, perna, etc.

Sistemas ópticos exigem uma carga maior de processamento computacional, por este motivo, geralmente não são usados em aplicações em tempo real. Normalmente os dados são obtidos de maneira indireta, assim aliviando o processamento e permitindo uma maior amostragem para a gravação de movimentos extremamente rápidos. Uma grande vantagem desse tipo de sistema é o grau liberdade que proporciona ao usuário, pois não existem cabos conectados à ele.

Um ponto negativo é que a iluminação ambiente deve ser controlada, principalmente quando são empregados marcadores passivos. Porém, existem sistemas como o da Phase Space Inc. (2015), ilustrado na Figura 4, que conseguem realizar a captura em ambientes externos com luz natural, geralmente sendo necessário realizar uma limpeza dos ruídos encontrados nos dados gerados. Também existe o problema de oclusão, em que ao menos duas câmeras devem ter visão clara de um determinado marcador para possibilitar o cálculo da posição e orientação por meio de triangulação (HARTLEY; STURM, 1997). Outro ponto negativo é o custo elevado dos equipamentos, por consequência da alta tecnologia aplicada no desenvolvimento das câmeras, tendo como principais variáveis na elevação do custo: a quantidade de câmeras usadas, a resolução e o número de quadros por segundo (FPS - *frames per second*).

Figura 4 – Sistema de captura de movimentos óptico da empresa PhaseSpace




Fonte: Phase Space Inc. (2015)

### 2.1.5 Inerciais

Sistemas inerciais empregam marcadores ativos contendo acelerômetros e giroscópios posicionados nas articulações do usuário para obter as informações diretas de posição e rotação.

Estes sistemas são portáteis e podem ser usados em quase todos os tipos de ambientes, alguns até mesmo em ambientes subaquáticos, seus dados não sofrem interferência externa, estão livres do problema de oclusão, e pelo fato da grande maioria deles usarem tecnologia sem fio, permitem que o usuário tenha maior liberdade de movimento, além de fornecerem um grande volume de dados. No entanto, são sistemas que tem um custo relativamente alto, principalmente por necessitarem de componentes miniaturizados. O uso de baterias é outra desvantagem, pois a utilização do sistema fica limitada ao tempo de duração das mesmas, necessitando recarregá-las ou de baterias reservas.

Um exemplo de sistema inercial de captura de movimentos é o IGS-190 da Meta Motion (2015) mostrado na Figura 5.

Figura 5 – Sistema inercial IGS-190 da empresa Meta Motion

Fonte: Meta Motion (2015)

## 2.2 Classificação Baseada em Sensores e Fontes

Como já mencionado na seção anterior, os sistemas MoCap fazem uso de sensores (receptores) e fontes (transmissores). É importante mencionar a classificação baseada na localização destes sensores e fontes, que divide os sistemas em: *inside-out*, *inside-in* e *outside-in* (MEYER; APPLEWHITE; BIOCCA, 1992):

- Sistemas *Inside-in*: os sensores e fontes estão localizados no corpo do usuário. Por exemplo, em sistemas mecânicos, onde os sensores e fontes estão fixos entre partes rígidas do exoesqueleto usado pelo usuário;

- Sistemas *Inside-out*: os sensores estão fixados no corpo do ator e recebem estímulos de fontes externas. Por exemplo, em um sistema com sensores (bobinas) presos no usuário, com capacidade de medir o campo magnético gerado por outras bobinas (fontes) fixadas na área de rastreamento. Sistemas inerciais, magnéticos e alguns acústicos fazem parte desta categoria; e

- Sistemas *Outside-in*: os sensores ficam localizados no espaço de captura e a fonte emissora localizada no corpo do usuário. Um exemplo deste tipo de sistema pode ser, um sistema acústico, onde os transmissores (fonte de som) são fixados no usuário e os receptores (microfones) estão posicionados na área de rastreamento. Alguns sistemas ópticos ou magnéticos também podem ser enquadrados nesta categoria.

## 2.3 Formatos de Arquivos

Alguns sistemas de Mocap têm o propósito de animar, personagens virtuais tridimensionais para o cinema, ou analisar os movimentos de atletas para aplicações ligadas ao esporte, entre outros, necessitando, então, armazenar os dados gerados para serem importados por outros sistemas e serem reproduzidos sempre que necessário. Com isso, foram criados diversos formatos de arquivos, visando padronizar o armazenamento destas informações. Nas próximas subseções serão descritos brevemente os três formatos considerados mais importantes: o ASF/AMC *(Acclaim Skeleton File/Acclaim Motion Capture data)*, o BVA *(BioVision Animation)* e o BVH *(Biovision Hierarchy)* (MENACHE, 2000).

### 2.3.1 ASF/AMC

O formato ASF/AMC é constituído por dois tipos de arquivos, o primeiro (ASF) especifica o esqueleto e a hierarquia de seus ossos, o segundo (AMC) refere-se a reprodução do movimento. A Figura 6 exibe os dois arquivos.

Figura 6 – Fragmentos dos arquivos ASF e AMC.

Fragmento de um arquivo ASF

```
01 :version 1.10
02 :name BioSkeleton
03 :units
04    mass 1.0
05    length 1.0
06    angle deg
07 :documentation
08    Example of an Acclaim skeleton
09    To be used with "Walk.amc"
10 :root
11    axis XYZ
12    order TX TY TZ RZ RY RX
13    position 0.0 0.0 0.0
14    orientation 0.0 0.0 0.0
15 :bonedata
16 begin
17    id 1
18    name hips
19    direction 0.000000 1.000000 0.000000
20    length 0.000000
21    axis 0.00000 0.00000 0.00000 XYZ
22    dof rx ry rz
23    limits (-180.0 180.0)
24       (-180.0 180.0)
25       (-180.0 180.0)
26 end
...
27:hierarchy
28 begin
29    root hips
30    hips hips1 hips2 hips3
...
31 end
```

Fragmento de um arquivo AMC

```
01 :FULLY-SPECIFIED
02 :DEGREES
03 1
04 root -1.244205 36.710186 -1.148101 0.958161 4.190043 -18.282991
05 hips 0.000000 0.000000 0.000000
06 chest 15.511776 -2.804996 -0.725314
07 neck 48.559605 0.000000 0.014236
...
08 2
09 root -0.227361 37.620358 1.672587 0.204373 -4.264866 -12.155879
10 hips 0.000000 0.000000 0.000000
11 chest 14.747641 2.858763 -1.345236
12 neck 44.651531 0.000000 -0.099206
...
```

Fonte: Menache (2000)

O arquivo ASF é criado com o uso de seções, onde cada uma é definida por uma palavra-chave iniciada por ":". Como pode ser observado na Figura 6, as seções encontradas no arquivo são as seguintes:

**:version** - especifica a versão de arquivo atual.

**:name** - o nome dado ao esqueleto.

**:units** - define as unidades (massa, tamanha e ângulo) que serão usadas.

**:documentation** - onde pode ser armazenada informações relevantes ao projeto.

**:root** - define atributos que serão usados para iniciar a leitura dos dados de movimentos armazenados no arquivo AMC, tais como, a ordem de rotação dos eixos, quais canais com suas respectivas ordens serão usados e a posição e orientação iniciais.

**:bonedata** - define os atributos dos segmentos (ossos) do esqueleto, com dados de identificação, posicionamento, orientação, comprimento, eixos e canais referentes a cada um dos segmentos.

**:hierarchy** - define a hierarquia dos segmentos (ossos) do esqueleto.

O arquivo AMC contém as informações de movimento do esqueleto definidas no arquivo ASF, para cada *frame* capturado, indicando o número do *frame*, a identificação do segmento do esqueleto e coordenadas.

A divisão deste formato em dois arquivos, possibilita a existência de múltiplos arquivos AMC com informações de movimento, para um único arquivo ASF com informações de atributos do esqueleto.

### 2.3.2 BVA

O BVA foi elaborado pela BioVision e diversos *softwares* de animação oferecem suporte para este formato.(LANDER, 1998). A Figura 7 exibe um fragmento de arquivo BVA.

Figura 7 – Fragmento de arquivo BVA

```
01 Segment:   Hips
02 Frames:    2
03 Frame Time: 0.033333
04 XTRAN       YTRAN   ZTRAN   XROT    YROT    ZROT    XSCALE  YSCALE  ZSCALE
05 INCHES      INCHES  INCHES  DEGREES DEGREES DEGREES INCHES  INCHES  INCHES
06  8.03        35.01   88.36   14.78  -164.35 -3.41    5.21    5.21    5.21
07  7.81        35.10   86.47   12.94  -166.97 -3.78    5.21    5.21    5.21
08 Segment:   Chest
09 Frames:    2
10 Frame Time: 0.033333
11 XTRAN       YTRAN   ZTRAN   XROT    YROT    ZROT    XSCALE  YSCALE  ZSCALE
12 INCHES      INCHES  INCHES  DEGREES DEGREES DEGREES INCHES  INCHES  INCHES
13  8.33        40.04   89.69  -27.24   175.94 -2.88    18.65   18.65   18.65
14  8.15        40.16   87.63  -31.12   175.58 -4.08    18.65   18.65   18.65
... (para todos os segmentos)
```

Fonte: Menache (2000)

O arquivo ASF é criado com o uso de seções, onde cada uma é definida por uma palavra-chave seguida de ":". Como pode ser observado na Figura 7, as seções encontradas no arquivo são descritas a seguir.

**Segment:** - define qual o segmento (osso) do esqueleto que esta sendo armazenado.

**Frames:** - é a quantidade total de *frames* (quadros) referente ao movimento do segmento.

**Frame Time:** - determina a taxa de exibição dos *frames* em segundos.

Em seguida são armazenados os dados de coordenadas da parte corrente.

Pode ser notado que os dados capturados não são armazenados de forma hierárquica, assim, para cada *frame*, os segmentos do esqueleto podem ser definidos na ordem e posição desejada. Isto pode causar redundância de informações armazenadas no arquivo, como por exemplo, informações de orientação e posição que podem estar sendo estabelecidas sem necessidade e valor da taxa de *frames*.

### 2.3.3 BVH

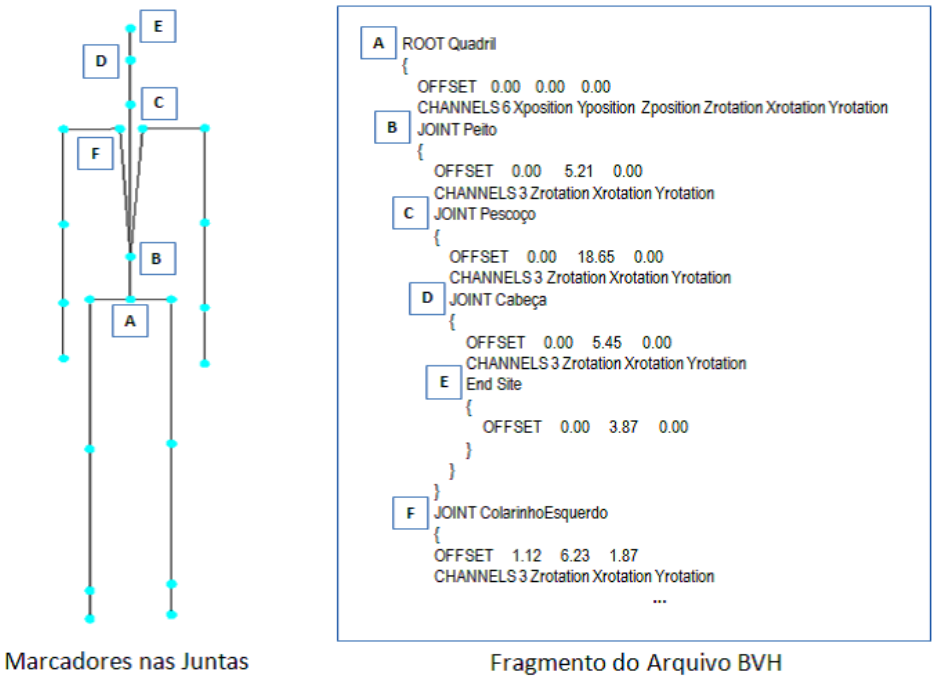
O formato BVH também foi desenvolvido pela BioVision, com o abjetivo de substituir o BVA e é suportado pela maioria dos *softwares* editores de animação. Porém, a

definição do esqueleto fica disposta de forma hierárquica, isto é, o movimento de um segmento (osso) filho é diretamente dependente do movimento do segmento (osso) pai.

Um arquivo BVH é separado em duas partes: a primeira define a estrutura do esqueleto, sua hierarquia e sua posição inicial; a segunda contém os dados dinâmicos do movimento (LANDER, 1998).

A imagem da esquerda na Figura 8 ilustra um esqueleto humanoide com marcadores em suas articulações, já a da direita, exibe um trecho da primeira parte de um arquivo BVH.

Figura 8 – Exemplo de um esqueleto humanoide com marcadores e representação hierárquica de arquivo BVH.




Fonte: Flam (2009)

Neste tipo de arquivo, a relação hierárquica é representada por blocos, sendo que a primeira informação de cada bloco é a semântica (Root/Joint) e o nome (Quadril/etc.) da articulação em questão. Em seguida, tem-se a informação do deslocamento relativo à articulação pai. No caso da 'Root', o deslocamento não existe. Por fim, encontram-se as grandezas que serão medidas no tempo (FLAM, 2009).

# 3 MÉTODOS E SENSORES RGB-D USADOS NA CAPTURA DE MOVIMENTOS

Este capítulo apresenta uma visão geral das tecnologias e métodos empregados na construção de sensores RGB-D, com ênfase nos equipamentos Kinect V1 e V2 da Microsoft.

## 3.1 Tecnologias Usadas em Dispositivos de Profundidade

Sensores de profundidade, são dispositivos criados com a função de calcularem o mapa de profundidade de uma cena. O mapa de profundidade, ou *depth map*, é formado por uma matriz bidimensional, onde são armazenados os valores para cada *pixel* da imagem capturada que representam as distâncias entre o dispositivo e os objetos existentes na cena. Em alguns casos, este mapa é incorporado na imagem RGB, em um canal adicional denominado como *Depth*, derivando RGB-D.

A seguir estão descritos os três principais métodos aplicados para a aquisição do mapa de profundidade. O primeiro é baseado em informações geradas por múltiplas câmeras RGB convencionais, o segundo emprega a tecnologia de 'luz estruturada' e o último a tecnologia *'time of flight'* (ToF).

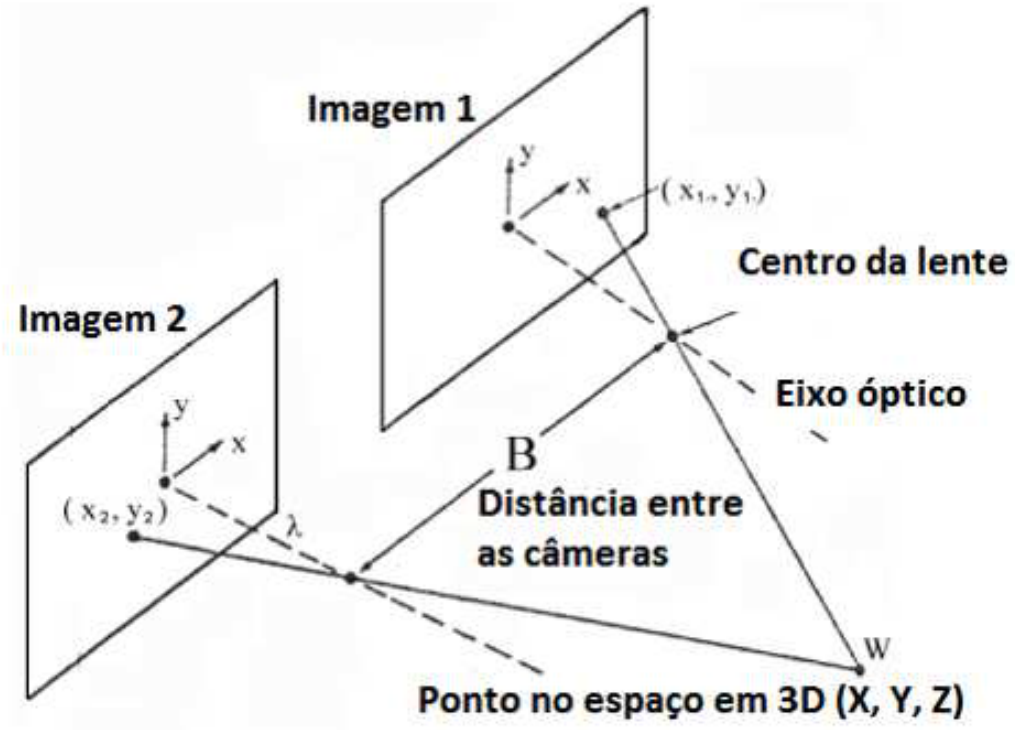
### 3.1.1 Imageamento Estéreo

Usando apenas câmeras RGB convencionais, pode-se obter o mapa de profundidade de uma cena aplicando diferentes técnicas, como por exemplo, utilizando imagens de uma única câmera, aplicando variações de iluminação no ambiente (BASRI; JACOBS; KEMELMACHER, 2007) ou variações na distância focal da câmera (ZHANG; NAYAR, 2006). Outra técnicas usada, é baseada em imagens estereoscópicas, onde são empregadas duas ou mais câmeras, cada câmera captura uma diferente perspectiva da mesma cena. O grande desafio para o funcionamento desta técnica é a identificação dos pontos chaves presentes em cada imagem que correspondem ao mesmo ponto da cena. Para a identificação destes pontos é necessário percorrer todos os *pixels* das imagens, fazendo com que seja uma tarefa mais pesada no que diz respeito a processamento necessário. Com a utilização de relações de geometria epipolar, pode-se reduzir as buscas por esses pontos (FAUGERAS, 1993).

A geometria epipolar para imageamento estéreo, encontrada também na literatura como restrição epipolar, determina a relação geométrica entre duas imagens capturadas com centros de projeções não coincidentes. Dado um ponto $P$ localizado na primeira ima-

gem, a busca realizada na segunda imagem por pontos correspondentes é limitada apenas para uma linha e não por toda a imagem. A identificação do mesmo ponto $P$ na segunda imagem e o conhecimento da distâncias entres as câmeras, viabiliza o calculo da triangulação dos pontos correspondentes no espaço tridimensional (FAUGERAS, 1993). A Figura 9 mostra duas imagens capturadas por meio de um par de câmeras, onde pode-se calcular um ponto no espaço 3D usando triangulação.

Figura 9 – Princípio do método de cálculo de triangulação de imagens para a obtenção do mapa de profundidade.




Fonte: Kabayama e Trabasso (2002)

O resultado dos cálculos de triangulação de todos os pontos 3D, constituem o mapa de profundidade da cena capturada (KABAYAMA; TRABASSO, 2002).

### 3.1.2 Luz Estruturada

Um dispositivo de luz estruturada é composto por pelo menos um projetor de luz infravervelha e uma câmera capaz de identificar a luz projetada.

Para calcular o mapa de profundidade, um padrão de luz de forma estruturada é projetado no ambiente e capturado por uma câmera. Em seguida, a deformação do padrão

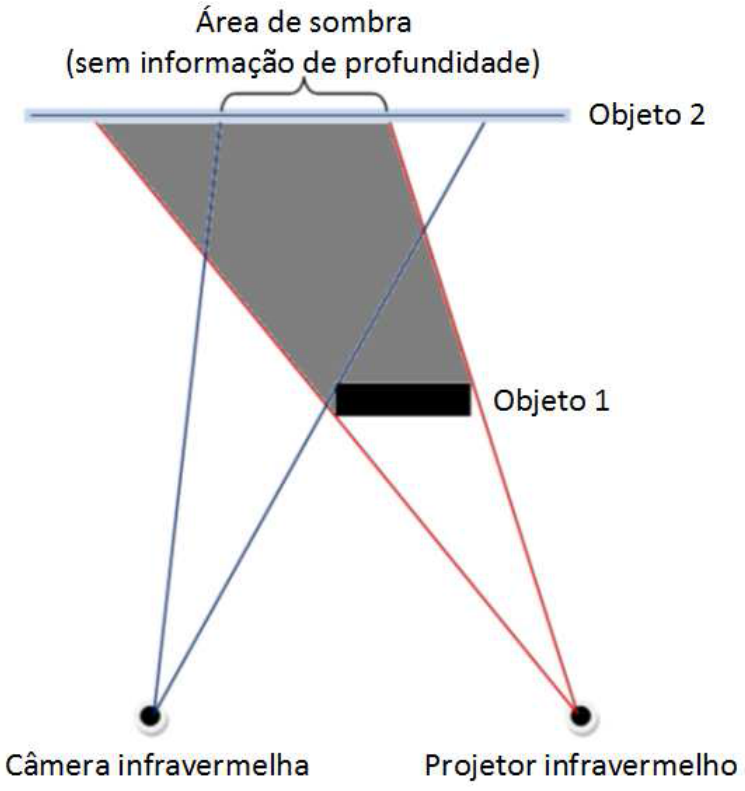
de luz é analisada, ou seja, com base nas distorções causadas no padrão de luz ao refletir nos objetos do ambiente, pode-se calcular a distância entre cada ponto de luz refletido em relação ao sensor (ZHANG, 2012).

Este tipo de sistema é considerado simples e não requer tratamento especial ao nível do sensor, sendo que qualquer disparidade na configuração pode ser calibrada a partir do conhecimento da estrutura projetada.

Sistemas com esta tecnologia são utilizados em ambientes fechados, onde a iluminação é controlada, pois sofrem interferência de iluminação vinda de outras fontes. São sistemas que dependem de disparidade óptica, portanto, são sensíveis a oclusões.

Estes sistemas também podem gerar áreas que não podem ser mapeadas, por duas razões: a primeira é o surgimento de regiões de sombra, que são causadas devido a distância entre o projetor e a câmera, com isto, algumas áreas visíveis à câmera podem não serem atingidas pelo padrão de luz projetado, por existir um objeto posicionados entre o projetor e a área que acomoda a sombra (DANCIU; BANU; CăLIMAN, 2012). A Figura 10 mostra como são geradas regiões de sombra.

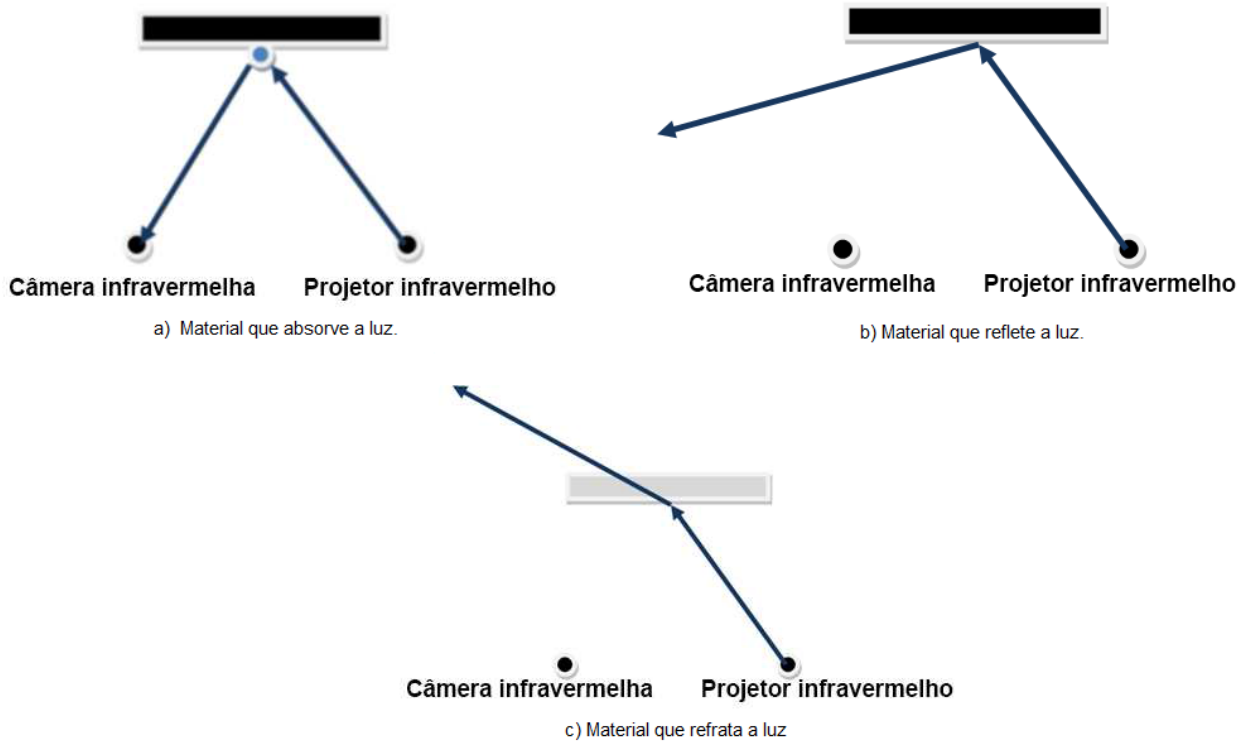
Figura 10 – Como são geradas regiões de sombras no mapa de profundidade utilizando luz estruturada




Fonte: Danciu, Banu e Căliman (2012)

A segunda razão, áreas que não podem ser mapeadas, é quando a câmera não é capaz de detectar a luz projetada. Alguns materiais como espelhos e louças podem causar reflexão da luz, materiais com vidros e plásticos translúcidos ou transparentes causam refração, com isso, a câmera não consegue capturar a luz emitida. Superfícies de materiais que absorvem a luz infravermelha como madeira, papel, etc, são mais adequados para que a captura da luz seja realizada (DANCIU; BANU; CăLIMAN, 2012). A Figura 11 mostra três comportamentos que podem ocorrer com a luz emitida pelo projetor, conforme o materiais de cada objeto.

Figura 11 – Comportamento dos feixes de luz infravermelha projetadas em superfícies de diferentes materiais;



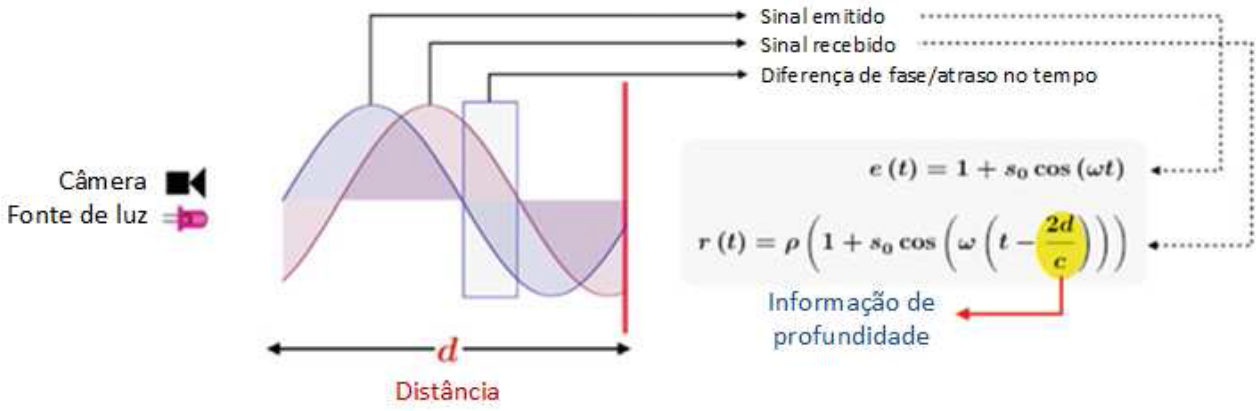

Fonte: Danciu, Banu e Căliman (2012)

### 3.1.3 ToF - *Time of Flight*

Sistemas que utilizam tecnologia *time of flight* contém um circuito que emite luz infravermelha acoplado à uma câmera que captura a luz emitida, tal como o nome sugere, calculando o "tempo de voo" da luz, ou seja, calculando o tempo de ida e volta da luz, desde sua emissão, reflexão em alguma superfície e retorno ao sensor. Desta maneira obtém-se a distância, ou profundidade, entre a superfície atingida e o sensor, sendo que este processo deve ser realizado para cada pixel do mapa de profundidade. Portanto,

quanto maior o "tempo de voo" da luz, maior será a distância do ponto em relação ao sensor (HANSARD et al., 2012). A Figura 12 ilustra uma visão geral do funcionamento de um dispositivo ToF.

Figura 12 – Visão geral do funcionamento do sensor ToF




Fonte: Danciu, Banu e Căliman (2012)

Uma grande vantagem de sistemas ToF é o uso de um único ponto de vista para calcular a profundidade, isso permite a diminuição de erros causados por oclusões, regiões de sombras e maior acurácia na preservação de bordas.

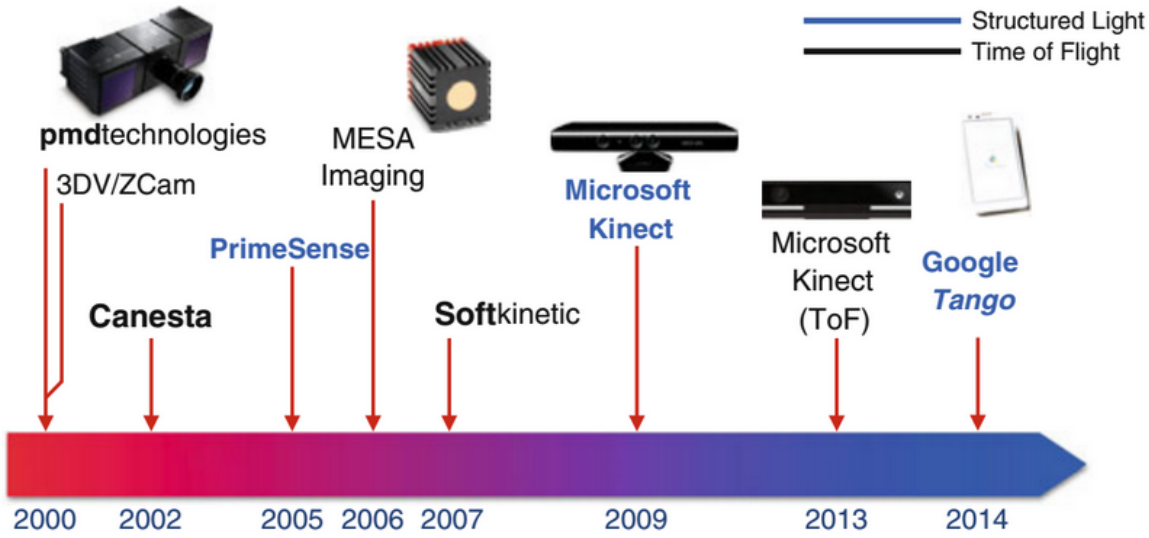
Outra vantagem é que a luz ambiente é rejeitada, uma vez que está em uma frequência diferente. Porém, as câmeras ToF normalmente apresentam baixa resolução devido ao processamento extra que acontece ao nível do sensor, e também, a iluminação irradiada de fontes externas podem saturar o sensor.

O princípio da tecnologia ToF não é recente. O mesmo tem sido utilizado durante décadas, por exemplo em sonar, espectrometria, espectroscopia e velocimetria, no entanto, os avanços recentes em hardware têm facilitado a inclusão desta tecnologia em equipamentos de baixo custo (HANSARD et al., 2012).

## 3.2 Dispositivos de Profundidade

Equipamentos que empregam a tecnologia de 'Luz Estruturada' ou 'ToF' estão sendo desenvolvidos por diversas empresas. A Figura 13 mostra alguns dos equipamentos fabricados a partir do ano 2000 que utilizam estas tecnologias. Entre eles estão a primeira e segunda geração do sensor Microsoft Kinect (KADAMBI; BHANDARI; RASKAR, 2014).

Figura 13 – Linha do tempo com sensores de luz estruturada (azul) e sensores ToF (preto).




Fonte: Kadambi, Bhandari e Raskar (2014)

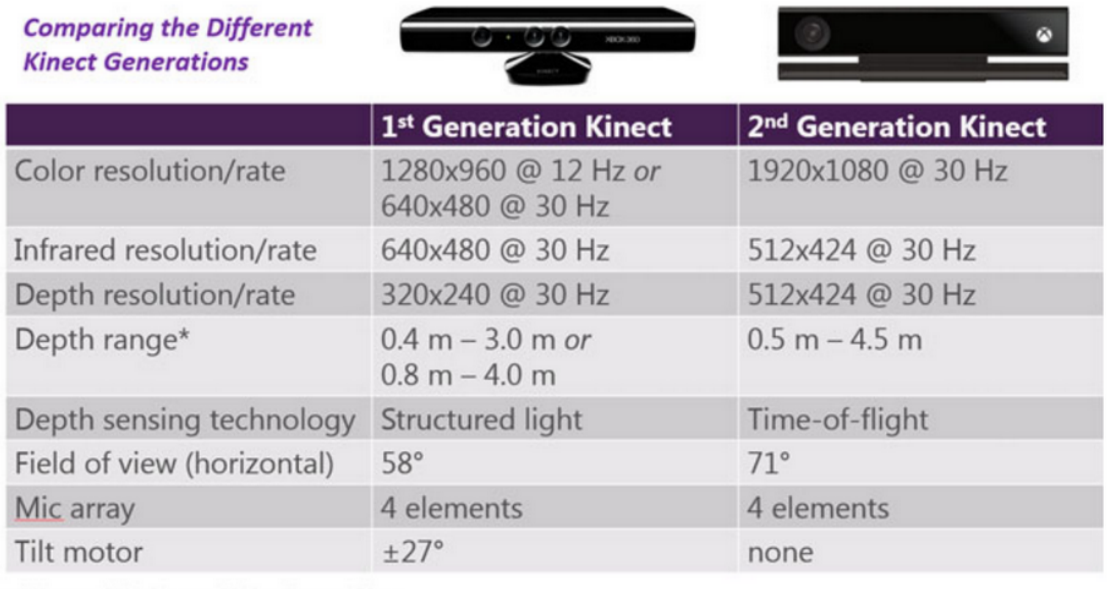
O Kinect é um acessório do console Xbox, que possibilita a interação com os jogos eletrônicos sem o uso de controle (*joystick*), capturando os movimentos dos jogadores em tempo real.

No ano de 2010 a empresa Microsoft lançou sua primeira geração do sensor Kinect. O sistema envolve uma câmera de vídeo RGB, uma câmera infravermelha, um laser infravermelho, quatro microfones e um motor de inclinação. O laser infravermelho emite um feixe de luz que, é dividido através do mecanismo de rede de difração, assim imitando o funcionamento de um projetor. Este dispositivo obtém o mapa de profundidade baseado em luz estruturada (KADAMBI; BHANDARI; RASKAR, 2014). Este dispositivo foi o primeiro exemplo de câmera de profundidade voltado para o mercado consumidor (ZENNARO et al., 2015).

O sensor da primeira geração foi substituído, em 2013, pela segunda geração, com características semelhantes. Analisando a Figura 14, pode-se notar que, ambas as versões trabalham em 30 quadros por segundo (fps), porém, a segunda possui vantagens de reso-

lução e ângulos das câmeras e, pelo fato de ter um campo de visão maior, não necessita de um motor de inclinação.

Figura 14 – Especificações de hardware dos modelos do sensor Kinect




| | 1st Generation Kinect | 2nd Generation Kinect |
|---|---|---|
| Color resolution/rate | 1280x960 @ 12 Hz *or* 640x480 @ 30 Hz | 1920x1080 @ 30 Hz |
| Infrared resolution/rate | 640x480 @ 30 Hz | 512x424 @ 30 Hz |
| Depth resolution/rate | 320x240 @ 30 Hz | 512x424 @ 30 Hz |
| Depth range* | 0.4 m – 3.0 m *or* 0.8 m – 4.0 m | 0.5 m – 4.5 m |
| Depth sensing technology | Structured light | Time-of-flight |
| Field of view (horizontal) | 58° | 71° |
| Mic array | 4 elements | 4 elements |
| Tilt motor | ±27° | none |

Fonte: Kadambi, Bhandari e Raskar (2014)

A principal diferença é que a segunda geração calcula a profundidade utilizando a tecnologia ToF. O Kinect versão 2 utiliza esta tecnologia, mas revelando novos elementos inovadores que superam algumas limitações críticas dos sensores ToF. Primeiro, a luz emitida é modulada em uma onda quadrada, em vez de senoidal como na maioria dos sensores deste tipo, o receptor de luz explora uma matriz de *pixels* diferente, isto é, cada *pixel* tem duas saídas e através dos fótons de entrada são carregados os valores nas duas saídas, de acordo com o estado do *clock*. Isto permite medir a diferença de fase e evita os problemas decorrentes da distorção harmônica (ZENNARO et al., 2015).

Outra questão crítica bem conhecida em sistemas ToF, é o erro devido à natureza periódica da medida de fase, o que limita a distância máxima mensurável. O Kinect lida com este problema usando múltiplas frequências de modulação. Por fim, o dispositivo também é capaz de adquirir duas imagens com tempo de exposição do obturador diferentes e a melhor imagem é selecionada em tempo real (SELL; O'CONNOR, 2014).

# 4 MÉTODOS DE CAPTURA DE MOVIMENTOS E DETECÇÃO DE POSE

Neste capítulo é apresentada a classificação de algoritmos de captura de movimentos humanos proposta por Yang (2014), bem como são descritos os principais trabalhos correlatos encontrados na literatura.

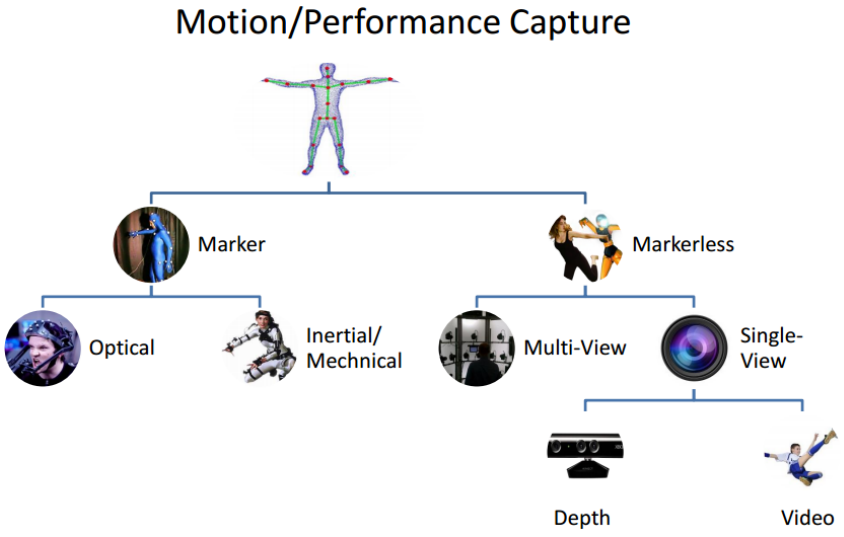
## 4.1 Classificação de Métodos de Captura de Movimentos Humanos

Yang (2014) propôs uma interessante classificação dos métodos de captura de movimentos humanos, a qual pode ser visualizada na Figura 15. Inicialmente os métodos são divididos entre os que utilizam marcadores e os que não utilizam marcadores.

Os métodos que fazem uso de marcadores foram subdivididos em duas sub-classes, os que usam dispositivos ópticos para detectar os marcadores e os que usam outras tecnologias como os mecânicos e inerciais.

Já entre os métodos que não utilizam marcadores, foi efetuada uma subdivisão entre aqueles que utilizam múltiplos pontos de vista e aqueles que utilizam apenas um ponto de vista. Com relação aos métodos que utilizam um único ponto de vista, estão os que usam informações de profundidade e os que usam informações de imagens (vídeo) RGB.

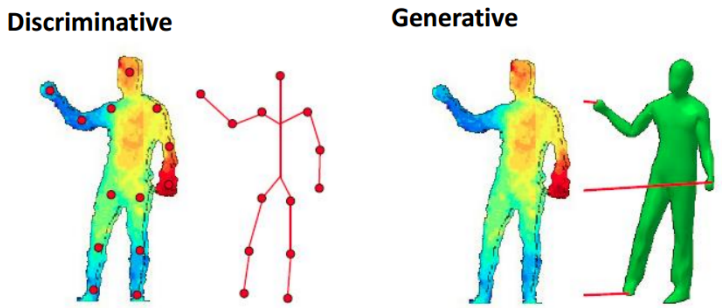
Figura 15 – Classificação dos métodos de captura de movimentos humanos




Fonte: Yang (2014)

Yang (2014) também divide os algoritmos de MoCap que utilizam único sensor de profundidade, entre discriminativo e generativo, conforme ilustrada na Figura 16.

Figura 16 – Divisão de métodos MoCap com único sensor de profundidade




Fonte: Yang (2014)

Os métodos generativos buscam aproximar o posicionamento de um modelo tridimensional ao posicionamento do usuário. Este modelo pode ser gerado através de escaneamento de pessoas reais ou a partir de softwares modeladores.

Os métodos discriminativos visam estimar a localização das articulações do corpo, gerando um esqueleto virtual sobre o usuário.

## 4.2 Trabalhos Correlatos

O trabalho de Ye e Yang (2014) apresenta um novo algoritmo para a estimativa simultânea de pose e forma em tempo real de objetos articulados, como seres humanos e animais. A principal característica da estimativa de pose é incorporar o modelo de deformação articulada com parametrização baseada em mapas exponenciais em um modelo de mistura gaussiana. Por meio do uso de um do modelo de medição probabilístico, o algoritmo não requer correspondências pontuais explícitas, ao contrário da maioria dos métodos existentes. Consequentemente, esta abordagem é menos sensível ao mínimo local, mas opera bem com movimentos rápidos e complexos. Este algoritmo também captura automaticamente a forma dos personagens durante o processo de estimativa de pose dinâmica. Resultados mostram que este método atinge acurácia comparável com o estado da arte, especialmente em caso de movimentos mais complexos, e não requer nenhum modelo paramétrico nem procedimento de calibração extra.

A captura do movimento do esqueleto e a geometria detalhada da superfície de múltiplas pessoas interagindo de perto é uma tarefa muito desafiadora, mesmo em uma configuração com múltiplas câmeras, devido a frequentes oclusões e ambiguidades em

atribuições de pessoa para pessoa. Para abordar esta tarefa, Liu et al. (2013) propõem uma estrutura que explora a segmentação de imagens de múltiplos pontos de vista. Para isso, um modelo probabilístico de forma e aparência é empregado para segmentar as imagens de entrada e para atribuir cada pixel exclusivamente a uma pessoa. Dado os modelos articulados de cada pessoa e os pixels rotulados, um esquema de otimização combinado, que divide o problema de otimização da pose de esqueleto em um único local e um dimensional global de menor dimensão, é aplicado um a um para cada indivíduo, seguido de estimativa de superfície para capturar deformações não-rígidas. É mostrado que em diversos casos, esta abordagem pode capturar o movimento 3D humano com precisão, mesmo quando os personagens se movem rapidamente, se usam vestuário largo, e se estão envolvidos em movimentos desafiadores como dança, luta-livre e abraços.

A reconstrução de representação tridimensional do movimento humano em tempo real é um importante campo de pesquisa com aplicações em ciências esportivas, indústria cinematográfica, ente outros. O trabalho de Helten et al. (2013) apresenta um algoritmo robusto para estimar um modelo de corpo humano personalizado a partir de apenas duas imagens de profundidade capturadas sequencialmente. O método utiliza um fluxo de imagens de profundidade para aplicação no algoritmo de rastreamento que combina a otimização do posicionamento local e uma pesquisa estabilizadora no banco de dados para rastrear a pose em tempo real. Experimentos realizados mostram que o algoritmo apresentado, é mais acurado e executa uma ordem de magnitude mais rápida do que vários algoritmos conhecidos na literatura atual.

Ye et al. (2011) apresentam um novo sistema para estimar a configuração de pose corporal a partir de um único mapa de profundidade. O método proposto combina a detecção e o refinamento de pose. O mapa de profundidade de entrada é comparado com um conjunto de exemplares de movimento pré-capturados, em seguida é gerada uma estimativa de configuração do corpo, bem como a rotulação semântica da nuvem de pontos de entrada. A estimativa inicial é então refinada, ajustando a configuração do corpo com a observação. Além de uma nova arquitetura do sistema, também foram incluídos no método: a modificação de uma técnica de suavização de nuvens de pontos para lidar com mapas de profundidade muito ruidosos, um alinhamento de nuvens de pontos e um algoritmo de pesquisa de posicionamentos independente. Experimentos em um conjunto de dados público mostram que esta abordagem atinge uma precisão compatível com os métodos encontrados na literatura.

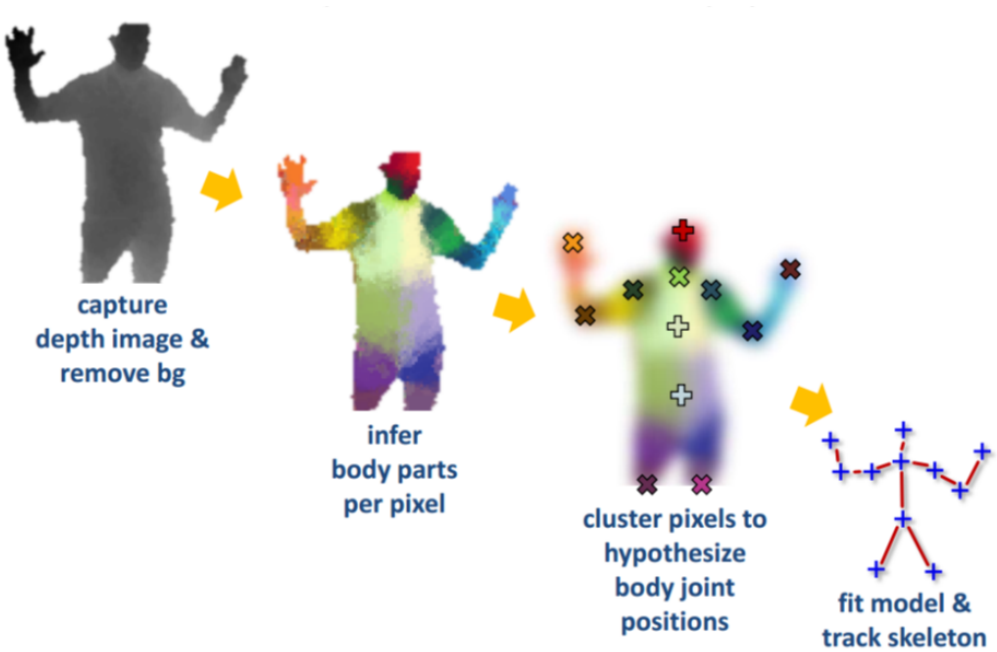
O trabalho de Gall et al. (2009) apresenta um método para capturar o movimento humano ou animal a partir de uma sequência de vídeo com múltiplos pontos de vistas. Dado um modelo articulado e silhuetas de uma sequência de imagens multi-view, esta abordagem recupera não apenas o movimento do esqueleto, mas também a deformação temporal não rígida da superfície 3D. Enquanto as deformações em grande escala ou movimentos rápidos são capturados pela pose do esqueleto e pela aproximação da geometria da

superfície, as deformações em pequena escala ou o movimento de roupas são capturados ajustando a superfície à silhueta. É mostrado que em diversas situações esta abordagem pode capturar o movimento 3D de animais ou seres humanos, mesmo quando ocorrem movimentos rápidos e com o uso de roupas largas como saias.

O próximo método que será descrito tem como caracteristica principal o fato de ter sido incorporado pela plataforma Microsoft Kinect. Conforme a classificação de Yang (2014) o método não utiliza marcadores, possui um único ponto de vista e utiliza unicamente informações de profundidade de forma discriminativa.

Shotton et al. (2011) treinaram um classificador *'Random Forests'* para segmentar partes do corpo a partir de uma a única imagem de profundidade, e então estimando o posicionamento das articulações de um modelo discriminativo, aplicaram técnicas de deslocamento médio. Este método foi apresentado por Shotton et al. (2011) visando prever o posicionamento tridimensional das articulações do corpo humano em uma única imagem de profundidade, utilizando reconhecimento de pose em partes, a partir de um classificador *'Random Forests'*. A Figura 17 ilustra a visão geral do funcionamento do método.

Figura 17 – *Pipeline* do método de MoCap usado no Kinect



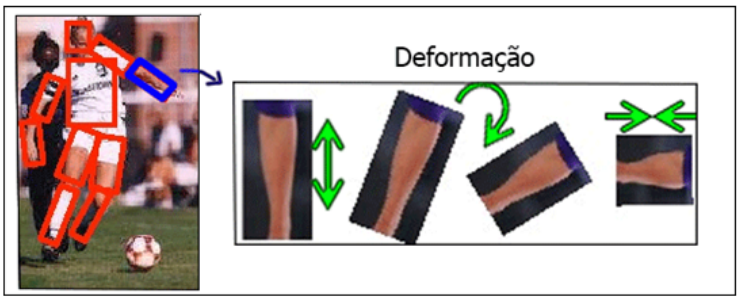

Fonte: Shotton et al. (2011)

Inicialmente uma imagem de profundidade da cena é capturada e o elemento de interesse (um ou mais humanos) é segmentado. Em seguida, é aplicado um algoritmo de classificação sobre os *pixels* da imagem, para se determinar a qual parte do corpo cada

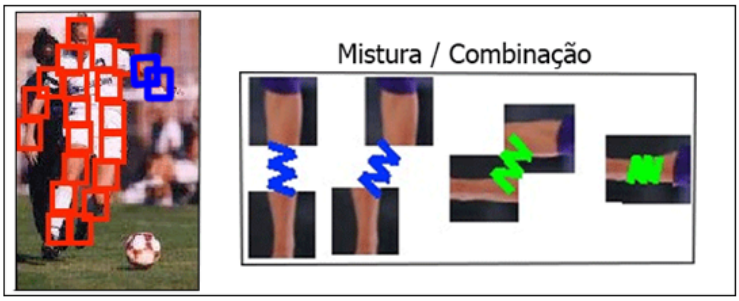
um pertence. Na próxima etapa, com base nos *pixels* agrupados por parte do corpo, é estimado o posicionamento das articulações. Por fim, é gerado um modelo esquelético ligando as articulações encontradas na etapa anterior, que será o resultado do rastreio do corpo humano. Deste modo, o método estima as articulações em três dimensões com restrições cinemáticas e coerência temporal, resultando em um esqueleto humanoide completo. Apesar de apresentar uma precisão que pode ser melhorada, ainda assim, o algoritmo apresenta uma boa aproximação do movimento realizado e com desempenho em tempo real (YANG, 2014).

O método sugerido por Yang e Ramanan (2013), tem como objetivo estimar poses, detectar articulações humanas em imagens estáticas, baseado em uma nova representação de modelos de partes flexíveis e gerar um esqueleto virtural 2D. O método difere de algumas abordagens clássicas que utilizam a deformação (giro e esforço) em um único modelo para cada membro como mostra a Figura 18 a). Neste novo método são combinados pequenos *templates* (modelos) não orientados para estimar a pose, ilistrado na Figura 18 b), o método captura conjuntamente as relações espaciais entre as partes locais e as relações de co-ocorrência entre a combinação das partes, aperfeiçoando o modelo por meio da conexão das partes com melhor relação como se tivessem molas, mantendo a dependência da geometria global com relação à aparência local.

Figura 18 – Diferença entre abordagem que utiliza deformações daquela que utiliza combinação de partes




a) Utiliza deformações em um único modelo.

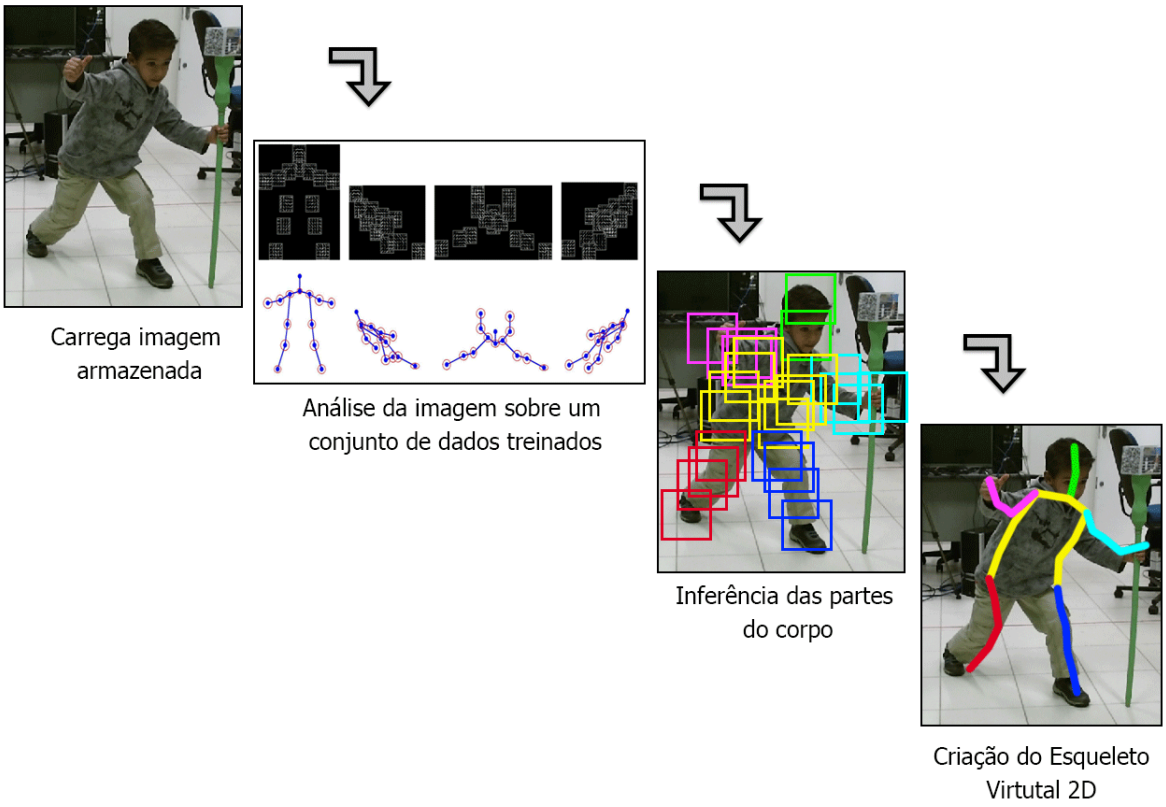



b) Mistura pequenos modelos pontuando sua relação.

Fonte: Yang e Ramanan (2013), adaptado pelo autor.

A Figura 19 ilustra uma visão geral do funcionamento do algoritmo de Yang e Ramanan (2013). O algoritmo não realiza a captura da imagem através de dispositivos como câmeras, as imagens de entrada devem estar armazenadas no disco rígido, cartão de memória ou similares.

Figura 19 – Pipeline do método de Yang e Ramanan (2013)



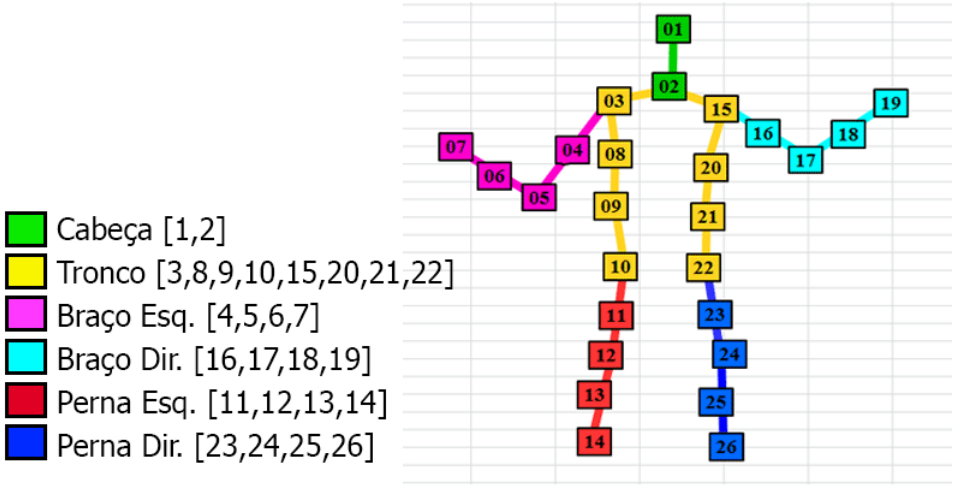

Fonte: Yang e Ramanan (2013), adaptado pelo autor.

A imagem é carregada, no estágio inicial do *pipeline*, e imediatamente submetida à análise de seus *pixels*, dando início a segunda etapa do processo mostrado na Figura 19. A representação de quatro modelos distintos treinados sobre o conjunto de dados, todavia, é permitida a composição entre qualquer tipo das partes entre si, sendo que a pontuação associada a cada combinação decompõe-se em uma árvore, assim mantendo a eficiência na pesquisa das pontuações geradas. O método faz uso da técnica de aprendizado de máquina conhecido como SVM (*Support Vector Machine*). Os parâmetros de treinamento incluem as relações espaciais locais e as relações de co-ocorrência entre as combinações das partes do corpo.

Ao término de toda a análise da imagem, são inferidas as partes do corpo que obtiveram maior pontuação, levanto em conta todas as combinações de todo o conjunto.

A Figura 20 mostra que o autor dividiu o corpo humano em vinte e seis partes, entre seis regiões, sendo duas partes na região da cabeça, oito na região do tronco, e mais quatro partes em cada uma das outra regiões (braço direito, braço esquerdo, perna direita e perna esquerda).

Figura 20 – Divisão das Partes do corpo (YANG; RAMANAN, 2013)



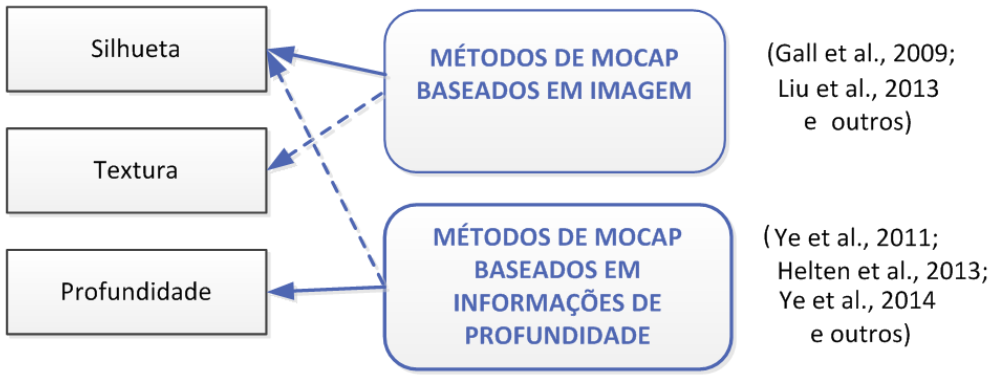

Fonte: Elaborado pelo autor.

Para finalizar, é gerado um esqueleto humanoide de duas dimensões a partir das vinte e seis partes da etapa anterior.

# 5 MÉTODO PARA A CAPTURA DE MOVIMENTOS HUMANOS USANDO UMA CÂMERA RGB-D

Como descrito no capítulo 4, a maior parte dos métodos de captura de movimentos humanos levantados a partir de consulta à literatura científica, no que diz respeito ao uso de uma única câmera (ou ponto de vista), podem ser agrupados em duas categorias: os que utilizam informações da imagem e os que utilizam as informações de profundidade, conforme mostra a Figura 21.

Figura 21 – Agrupamento dos métodos de MoCap de acordo com o tipo de informação utilizada.




Fonte: Yang (2014), adaptado pelo autor.

A divisão nestes dois grupos deve-se, normalmente, ao fato de usarem apenas câmeras RGB nos métodos que fazem uso de informações de imagem, enquanto que no segundo grupo, o equipamento de captura contém um sensor de profundidade, dando-se primazia aos dados gerados por este.

Quando o método de Mocap baseia-se na imagem, normalmente prioriza a identificação da silhueta do ator na etapa de eliminação do fundo (*background*) e extração do elemento de interesse (*foreground*). A informação de textura (características, como as cores, no interior da silhueta) pode ser usada na classificação dos *pixels* e associação às partes do corpo. Tais métodos geralmente exigem grande processamento devido à alta resolução das imagens.

Os métodos de Mocap que se baseiam em profundidade, trabalham com um menor volume de dados, uma vez que a resolução dos sensores de profundidade costuma ser bem menor do que a das câmeras RGB. Portanto, normalmente, oferecem um desempenho mais adequado para aplicações de tempo real, como é o caso de animação de avatares em videogames. A informação de profundidade é usada, neste caso, para a extração do elemento de interesse, ou seja, o ator (*foreground*), dos elementos de fundo (*background*), pois a distância da nuvem de pontos associadas ao ator estão mais próximas da câmera do que o fundo. A silhueta do ator, também, pode ser obtida a partir exclusivamente das informações de profundidade dos *pixels*.

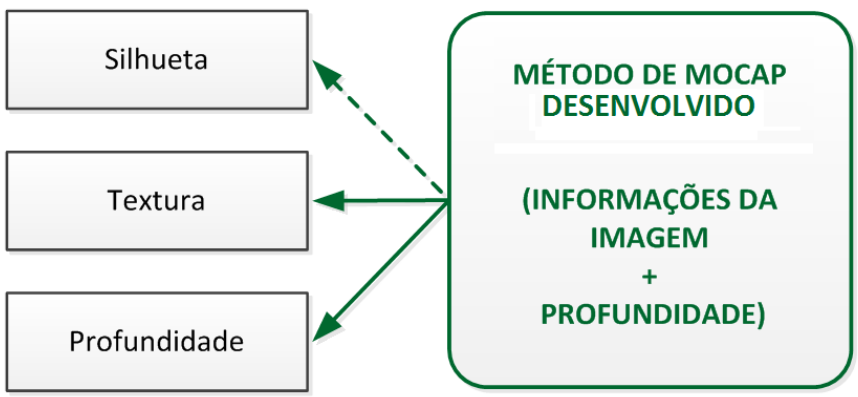
No entanto, equipamentos RGB-D como os Kinects v1 e v2 da Microsoft, por exemplo, possuem os dois tipos de câmeras: RGB e de profundidade. O método adotado especificamente nestes equipamentos, no que diz respeito à captura de movimentos e ao oferecimento de um esqueleto virtual, faz uso apenas da informação de profundidade e, portanto, pode apresentar erros, conforme mostraram os experimentos realizados e que serão apresentados na seção 6.3.

## 5.1 O Método de Captura de Movimentos Desenvolvido

Fazendo-se uso

Portanto, fazendo-se uso das informações de imagem e de profundidade oferecidas por equipamentos RGB-D como o Kinect, foi elaborado um método que oferece uma maior acurácia e, por isso, maior qualidade quanto aos movimentos inferidos a partir da captura, conforme ilustrado na Figura 22.

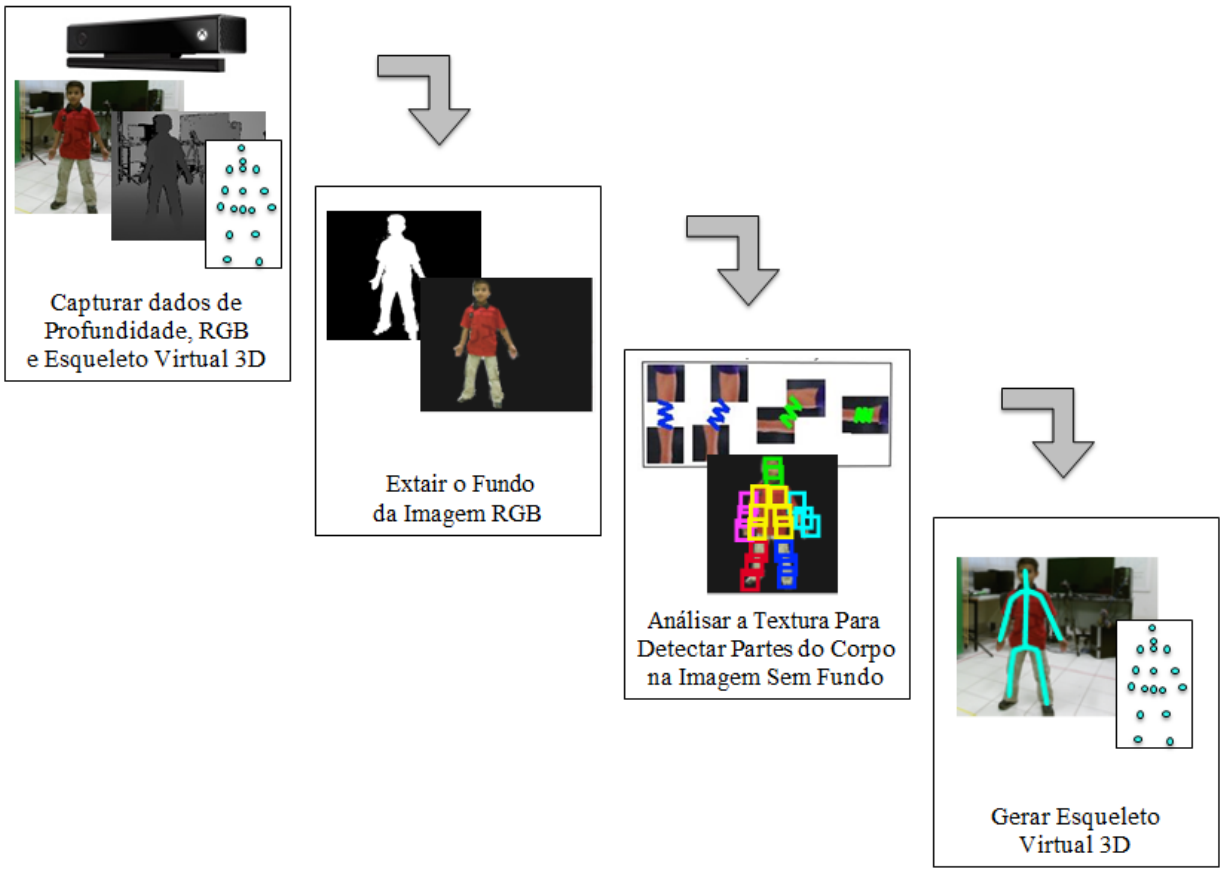
Figura 22 – Informações utilizadas pelo método desenvolvido.




Fonte: Elaborado pelo autor.

Como pode ser observado na Figura 22, as informações de profundidade foram utilizadas no método desenvolvido, para a separação do ator com relação ao fundo e também, na obtenção da silhueta, como nos métodos já existentes. Porém, a grande contribuição está na utilização dos dados de textura da imagem em conjunto com a de profundidade, no que tange a inferência das partes do corpo por *pixel*, o agrupamento (clusterização) dos *pixels* na estimação das posições das juntas e, por fim, a associação ao modelo do esqueleto virtual tridimensional. As etapas principais do método de Mocap desenvolvido são exibidas na Figura 23.

Figura 23 – Pipeline do Método Desenvolvido




Fonte: Elaborado pelo autor.

Como já mencionado, este método faz uso de informações de profundidade e de textura, portanto a primeira etapa tem a função de realizar a aquisição de tais informações a partir do dispositivo Microsoft Kinect versão 2, também é capturado o esqueleto gerado pelo próprio dispositivo.

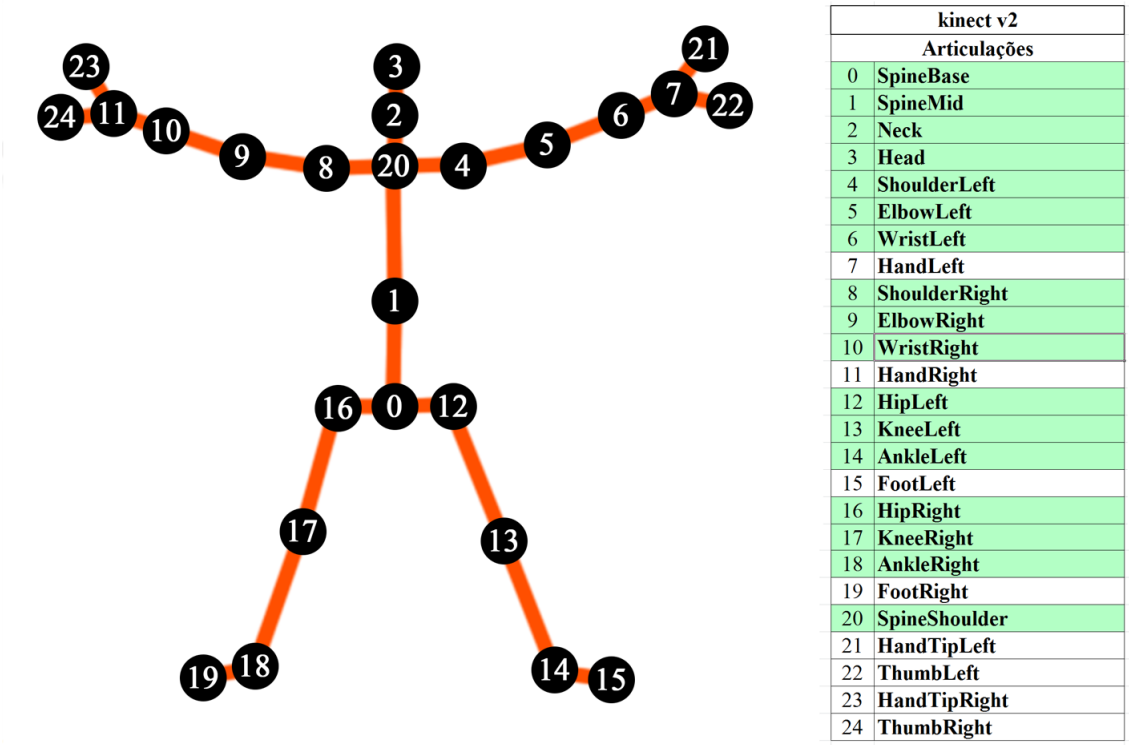
A segunda etapa, utiliza as informações de profundidade para obter uma silhueta do ator, logo após, em conjunto com a imagem RGB, é extraído o fundo da imagem

preservando-se apenas o objeto de interesse (ator).

Na terceira etapa, já com a imagem RGB contendo apenas o objeto de interesse, a mesma é submetida à análise de textura proposta por Yang e Ramanan (2013), descrita no capítulo 4, a qual foi modificada (detalhada na Seção 5.2.5) de forma a reconhecer as partes do corpo do ator, mas não inferir as articulações nem gerar um esqueleto virtual, tendo como propósito principal, somente determinar a localização bidimensional das regiões do corpo na imagem.

A quarta e última etapa, visou analisar as informações geradas pela etapa anterior e inferir as localizações das articulações do novo esqueleto virtual calculando as distâncias entre as partes. A estrutura (esqueleto virtual) gerada com as articulações foi baseada na estrutura gerada pelo algoritmo do Kinect v2, a fim de facilitar comparações futuras entre os dois esqueletos. À esquerda da Figura 24 estão todas as vinte e cinco articulações suportadas pelo Kinect v2, no lado direito estas mesmas estão listadas, mas apenas às dezessete que estão destacadas com a cor verde, são consideradas neste método, pois a fase de detecção das partes do corpo não oferece informações suficientes para a estimativa da posição dos pés, mãos e dedos.

Figura 24 – Esqueleto suportado pelo Kinect v2




| kinect v2 | |
|---|---|
| Articulações | |
| 0 | SpineBase |
| 1 | SpineMid |
| 2 | Neck |
| 3 | Head |
| 4 | ShoulderLeft |
| 5 | ElbowLeft |
| 6 | WristLeft |
| 7 | HandLeft |
| 8 | ShoulderRight |
| 9 | ElbowRight |
| 10 | WristRight |
| 11 | HandRight |
| 12 | HipLeft |
| 13 | KneeLeft |
| 14 | AnkleLeft |
| 15 | FootLeft |
| 16 | HipRight |
| 17 | KneeRight |
| 18 | AnkleRight |
| 19 | FootRight |
| 20 | SpineShoulder |
| 21 | HandTipLeft |
| 22 | ThumbLeft |
| 23 | HandTipRight |
| 24 | ThumbRight |

Fonte: Elaborado pelo autor.

Finalizando esta última etapa, as coordenadas da terceira dimensão (eixo $z$) de cada articulação foram obtidas através de uma equação (detalhada no próximo capítulo) que usa informações do tamanho de cada osso (obtida na Seção 5.2.2) e as coordenadas $x$ e $y$ conseguidas na etapa anterior, assim resultando em um esqueleto virtual tridimensional híbrido, gerado com informações de profundidade e de textura.

Tanto o esqueleto rastreado em tempo real pelo Kinect, quanto o esqueleto obtido pelo método desenvolvido são armazenados em arquivos, assim sendo, foi desenvolvido um algoritmo que lê o arquivo armazenado contendo os *frames* com o rastreio do esqueleto, que em seguida, realiza a conversão para um arquivo de formato BVH (descrito na Subseção 2.3.3), que é padrão das aplicações que fazem uso da captura de movimentos.

## 5.2 Implementação do Método

Neste capítulo é apresentada a implementação do método de captura de movimentos humanos descrito no capítulo anterior, isto é, quais foram os materiais usados, os detalhes de cada etapa e suas características, os algoritmos utilizados e as soluções praticadas durante a implementação. É importante ressaltar que, todos os algoritmos implementados neste projeto foram codificados no Matlab.

### 5.2.1 Materiais

A) **Software**

- Sistema Operacional: Windows 10 Professional - 64 bits versão 6.2.9200.

- Ambiente de desenvolvimento: Microsoft Visual Studio 2012; MATLAB R2015A (8.5.0.197613)64-bit(win64); DirectX DirectX 12.0.

- Bibliotecas: Kinect for Windows SDK (v2.0-1409); Image Acquisition Toolbox 4.9(Matlab); Kin2 (Kinect 2 Toolbox for MATLAB).

B) **Hardware**

- Computador com: Placa mãe Asus P6TD Deluxe; Processador QuadCore Intel Core i7 930, 2933 MHz (22 x 133); Memória RAM 6 GB DDR3 SDRAM; Disco rígido ATA Device (1500 GB, 7200 RPM, SATA-II); Adaptador gráfico NVIDIA GeForce GTX 650 (1024 MB); Controladora USB Renesas USB 3.0 eXtensible Host Controller - 1.0; 2 Monitores LG E2350 23" LCD.

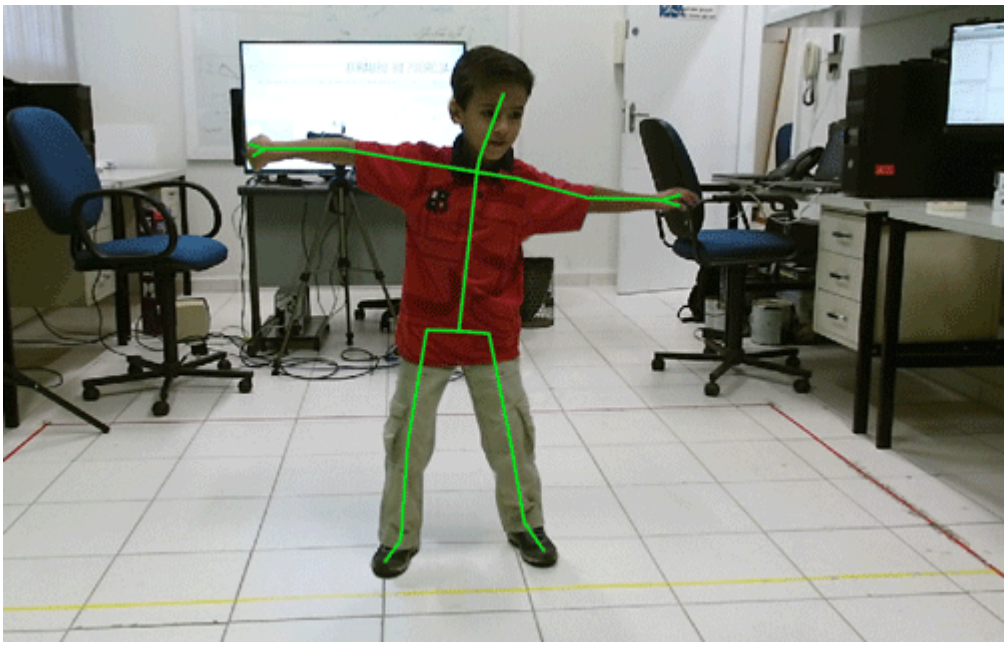
- Demais equipamentos: Kinect for Windows v2; tripé; trena; medidor de nível.

### 5.2.2 Processo de Calibração

Antes da iniciar as etapas do método devenvolvido, deve-se realizar o processo de calibração, que tem a finalidade de identificar e armazenar o tamanho de cada osso do ator usado no método desenvolvido, estas informações serão usadas em equações aplicadas na seção 5.2.6, fase que gera o esqueleto virtual tridimensional.

Para isso foi criado um algoritmo que captura uma imagem de corpo inteiro do ator, em que ele fica parado de frente para o kinect, com os braços abertos e com as pernas um pouco afastadas, conforme esta sendo mostrado na Figura 25.

Figura 25 – Imagem capturada no processo de calibração



Fonte: Elaborado pelo autor.

### 5.2.3 Captura dos dados

Para a aquisição dos dados foi utilizado o dispositivo Microsoft Kinect versão 2 (apresentado no capítulo 3), juntamente com o Kinect SDK *(Software Development Kit)* 2.0 para Windows do próprio fabricante (MICROSOFT, 2016).

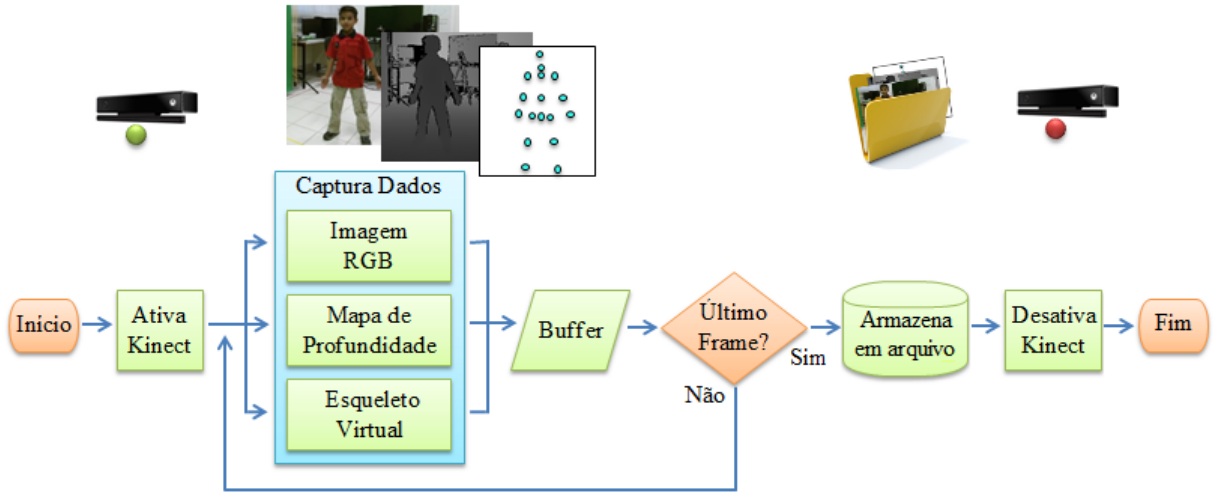
É importante lembrar que este dispositivo possui resolução de imagem RGB de 1920x1080 *pixels* e mapa de profundidade 512x424 *pixels* e utiliza a técnica de *Time of Flight* para obtenção do mapa de profundidade.

Para simplificar e agilizar a codificação em *software* e a manipulação dos recursos do Kinect no Matlab, foi utilizada a *Toolbox* "Kin2", a qual encapsula o Kinect SDK 2.0

e a escrita na linguagem C++. A versão usada contém duas classes e trinta funções para Matlab que fornecem acesso a cor, profundidade, infravermelho, imagens RGB, coordenar recursos de mapeamento, rastreamento em tempo real de seis corpos com 25 juntas, estado das mãos (aberta ou fechada), processamento de faces e gestos, entre outras. Terven e Córdova-Esparza (2016) sugerem que o desempenho de uma aplicação utilizando Kin2 é em média 30% melhor, em relação a uma aplicação em C++ nativo, produzindo códigos fontes com menor tamanho, assim, resultando em uma redução significativa no tempo de desenvolvimento de protótipos.

No processo de captura, para cada *frame* é necessário armazenar, tanto informações da imagem RGB, bem como o mapa de profundidade e todo o esqueleto gerado pelo algoritmo de Shotton et al. (2011) para que possam ser analisados e comparados posteriormente. A Figura 26 ilustra a maneira como este módulo de captura foi implementado.

Figura 26 – Fluxograma de captura de dados usando o Kinect v2.




Fonte: Elaborado pelo autor.

Esta é a única etapa que exige que o Kinect esteja conectado ao computador, uma vez que, todos os dados referentes a cena capturada estarão em armazenamento secundário. Para que não haja redução da taxa de *frames*, toda a cena fica na memória principal do computador e somente no final do processo ocorre a gravação em disco. A Figura 27 mostra a estrutura de dados usada no *Buffer*, contendo as variáveis: save_CountFrame (acumula a quantidade de *frames* capturados), save_Color (armazena a imagem RGB), save_DepthColor(guarda o mapa de profundidade referente a imagem RGB), save_Bodies_Position (armazena os dados de posicionamento do esqueleto rastreado pelo kinect), save_Bodies_Orientation (estão os dados de orientação do esqueleto) e a

save_BodyIndex (será usada para armazenar a silhueta baseada em profundidade).

Figura 27 – Estrutura de dados que armazena as informações capturadas pelo dispositivo kinect

```
framesKinect_yyyymmdd_HHMMSS.mat

   % quantidade de frames capturados
  save_CountFrame = 0;
   % dados da imagem RGB
  save_Color = zeros(color_height,color_width,3,MaxCountFrame,'uint8');
   % dados do mapa de profundidade
  save_DepthColor = zeros(depth_height,depth_width,3,MaxCountFrame,'uint8');
   % dados de posição do esqueleto do kinect (em quatérnios)
  save_Bodies_Position = zeros(3,25,MaxCountFrame,'uint8');
   % dados de orientação do esqueleto do kinect (em quatérnios)
  save_Bodies_Orientation = zeros(4,25,MaxCountFrame,'uint8');
   % dados da silhueta baseada em profundidade
  save_BodyIndex = zeros(depth_height,depth_width,MaxCountFrame,'uint8');
```

Fonte: Elaborado pelo autor.

### 5.2.4 Extração da Imagem de Fundo

Para cada *frame* tem-se informações de profundidade e da imagem RGB. A profundidade é utilizada para gerar uma silhueta do objeto de interesse (ator), posteriormente, a textura que está fora da silhueta é eliminada aplicando-se a equação (5.1) multiplicando cada ponto da matriz $A$ por cada ponto da matriz $B$.

A matriz $A$ contém os dados de um dos canais da imagem RGB, a matriz $B$ contém os dados da imagem da silhueta (0 ou 1), também conhecida como filtro de camada ou máscara, e a matriz $Z$ tem os dados de uma nova imagem RGB, ou seja, uma imagem que tem o ator com a extração do fundo, com uma quantidade reduzida de *pixels* com textura. Desta forma reduziu-se o esforço computacional aplicado na etapa seguinte, que tem o objetivo de analisar a textura da imagem para detectar as partes do corpo do ator.

$$A_{m,n} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ a_{3,1} & a_{3,2} & \cdots & a_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} B_{m,n} = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,n} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,n} \\ b_{3,1} & b_{3,2} & \cdots & b_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m,1} & b_{m,2} & \cdots & b_{m,n} \end{bmatrix} Z_{m,n} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,n} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,n} \\ z_{3,1} & z_{3,2} & \cdots & z_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{m,1} & z_{m,2} & \cdots & z_{m,n} \end{bmatrix}$$

$$z_{i,j} = a_{i,j} * b_{i,j} \mid i,j \in \mathbb{Z} \mid 1 \leq i \leq m \mid 1 \leq j \leq n \tag{5.1}$$

A Figura 28 mostra o fluxo de implementação feito para a extração do fundo.

Figura 28 – Fluxograma da extração de fundo da imagem RGB


Fonte: Elaborado pelo autor.

Como este processo é executado sobre informações que já estão armazenadas, não há necessidade de manter um *buffer* de memória e a imagem RGB contendo somente o ator, pode ser armazenada diretamente em arquivo sem haver perda de dados.

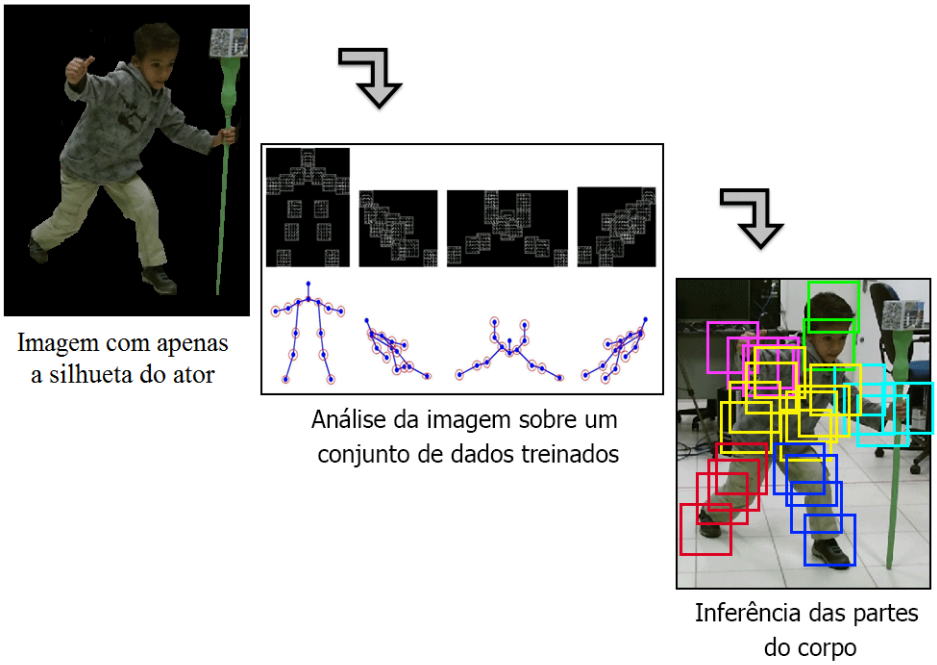
### 5.2.5 Detecção das partes do corpo do ator

A imagem RGB sem textura no fundo é usada como dado de entrada para esta etapa, o método sugerido por Yang e Ramanan (2013), descrito no capítulo 4, foi usado para a detecção das partes do corpo do ator.

Segundo esse método, uma imagem RGB é lida e, em seguida, os *pixels* desta imagem são analisados, e possíveis partes do corpo são combinadas levando em consideração as relações espaciais entre as partes locais e as relações de co-ocorrência entre a mistura das partes, desta maneira, se elege as partes com melhor relação, fazendo uso de uma estrutura SVM de aprendizado de máquina e de uma base de dados já treinada disponibilizada pelo autor em um de seus sites de pesquisa, e por fim gera um esqueleto virtual bidimensional (YANG; RAMANAN, 2013).

Este método foi adaptado para ler a imagem RGB com o fundo extraído, analisar os *pixels* e detectar as partes do corpo, mas não gerar o esqueleto virtual, pois a estrutura humanoide do esqueleto virtual gerada pelo método original é diferente da estrutura desejada, portanto, a próxima etapa deste processo visa gerar um esqueleto tridimensional conforme as características almejadas. A Figura 29 ilustra as etapas utilizadas do método de Yang e Ramanan (2013).

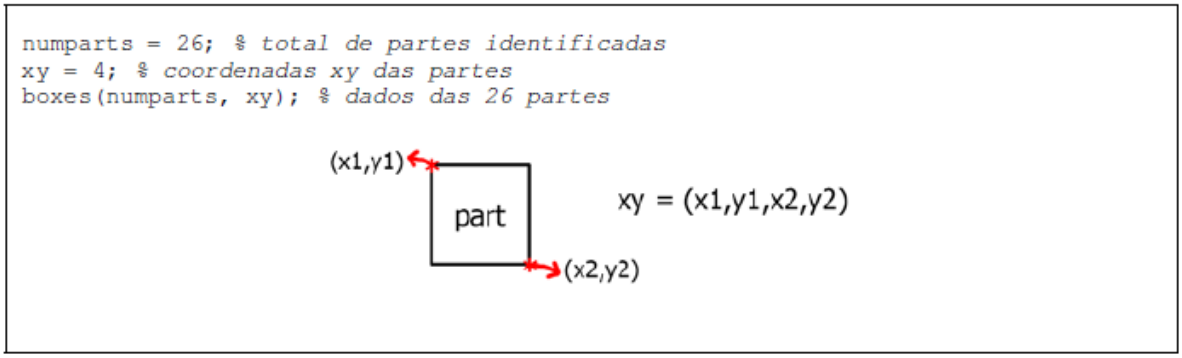
Figura 29 – Etapas do método de Yang e Ramanan (2013) utilizadas no método desenvolvido.




Fonte: Elaborado pelo autor.

O processo de análise de textura da imagem resulta na detecção de vinte e seis partes do corpo do ator, as mesmas são gravadas no computador para serem usadas posteriormente. Na Figura 30 pode-se visualizar um trecho de código criado no matlab, onde é realizada a geração da matriz que guarda os dados referentes as partes do corpo encontradas na imagem RGB.

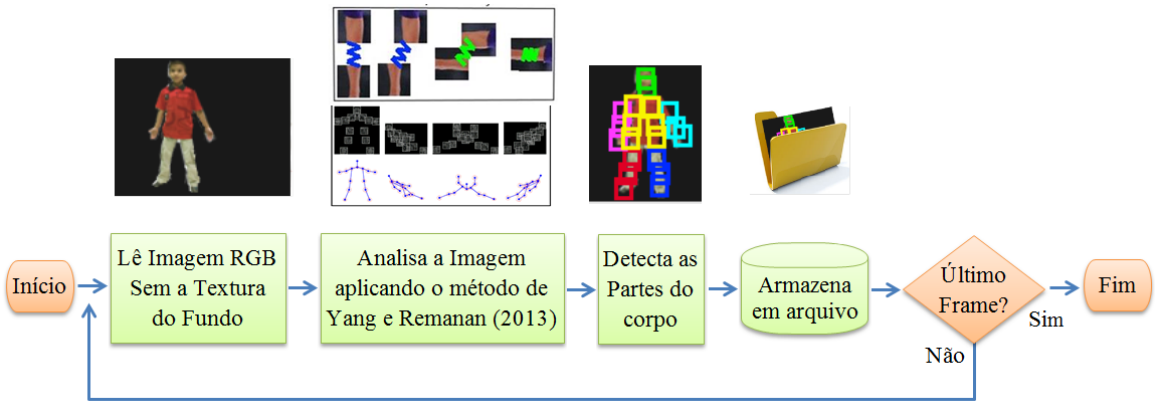
Figura 30 – Matriz que armazena dados das partes do corpo identificadas na imagem RGB.




Fonte: Elaborado pelo autor.

Na Figura 31 pode-se visualizar um fluxograma do processo de detecção das partes do corpo.

Figura 31 – Fluxograma da Detecção de partes do corpo




Fonte: Elaborado pelo autor.

### 5.2.6 Geração do Esqueleto Virtual Tridimensional

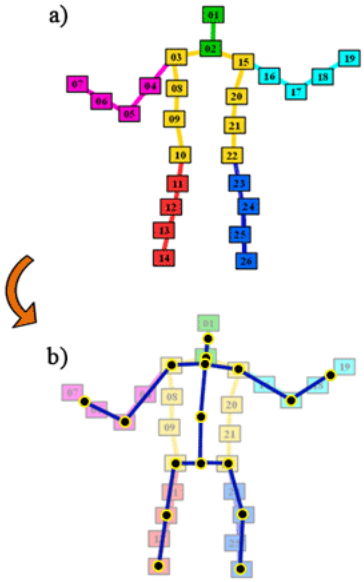
A última etapa tem o objetivo de gerar um esqueleto virtual tridimensional, para isso, foi realizada uma análise sobre as informações das partes do corpo reconhecidas na etapa anterior, e com base na estrutura do esqueleto gerado pelo kinect, as articulações do corpo são inferidas inicialmente em duas dimensões encontrando o centro de cada parte com a equação (5.2), onde $(x1, y1)$ e $(x2, y2)$ são coordenadas de uma parte do corpo. A mesma equação também é usada para encontrar o ponto central entre duas partes.

$$\begin{aligned} x &= (x1 + x2)/2 \\ y &= (y1 + y2)/2 \end{aligned} \tag{5.2}$$

A Figura 32 tem no seu lado esquerdo superior (a) uma ilustração do corpo dividido pelas partes reconhecidas, no lado esquerdo inferior (b) tem uma ilustração das articulações inferidas sobre as partes, e no lado direito (c) da figura, está uma tabela contendo os dados usados na análise das partes e inferência das articulações.

Figura 32 – Dados das articulações inferidas para o esqueleto virtual.

a)



b)



c)

**Dados das articulações definidas para o esqueleto virtual**

| Índice | Descrição | Nome | Posição em relação às partes |
|---|---|---|---|
| 1 | Base da coluna | SpineBase | Entre as partes 10 e 22 |
| 2 | Meio da coluna | SpineMid | Entre as partes (20 com 21) e (8 com 9) |
| 3 | Pescoço | Neck | No centro da parte 2 |
| 4 | Cabeça | Head | Entre as partes 1 e 2 |
| 5 | Ombro esquerdo | ShoulderLeft | No centro da parte 15 |
| 6 | Cotovelo esquerdo | ElbowLeft | No centro da parte 17 |
| 7 | Pulso esquerdo | WristLeft | Entre as partes 18 e 19 |
| 8 | Ombro direito | ShoulderRight | No centro da parte 3 |
| 9 | Cotovelo direito | ElbowRight | No centro da parte 5 |
| 10 | Pulso direito | WristRight | Entre as partes 6 e 7 |
| 11 | Quadril esquerdo | HipLeft | No centro da parte 22 |
| 12 | Joelho esquerdo | KneeLeft | No centro da parte 24 |
| 13 | Tornozelo esquerdo | AnkleLeft | No centro da parte 26 |
| 14 | Quadril direito | HipRight | No centro da parte 10 |
| 15 | Joelho direito | KneeRight | No centro da parte 12 |
| 16 | Tornozelo direito | AnkleRight | No centro da parte 14 |
| 17 | Espinha ao ombro | SpineShoulder | 15% da distancia entre (Neck e SpineMid) |

Fonte: Elaborado pelo autor.

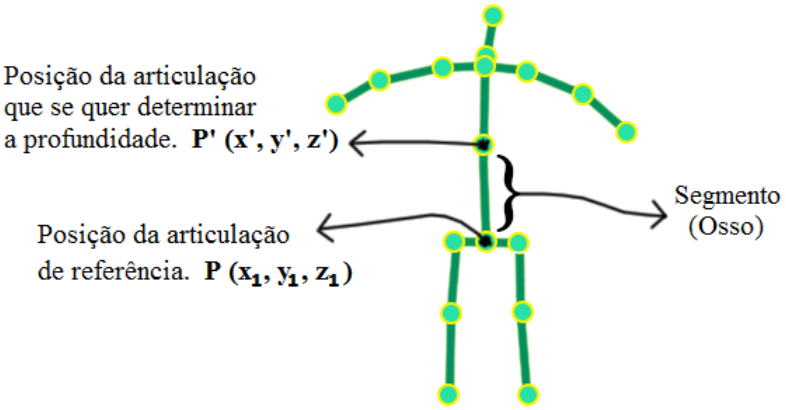
Para conseguir as coordenadas do eixo $z$ (profundidade) de cada articulação, necessárias para se chegar a um esqueleto virtual 3D, foi aplicada a equação (5.3), a qual permite encontrar o valor do eixo $z$ de uma articulação, partindo de uma articulação $P(x, y, z)$ já conhecida e também com o conhecimento prévio do tamanho do segmento (osso).

$$primeiraParte = \sqrt{(tamanhoSegmento^2) - ((x' - x_1)^2 + (y' - y_1)^2)}$$

$$z' = (z_1 - primeiraParte) \quad ou \quad z' = (z_1 + primeiraParte) \tag{5.3}$$

A Figura 33 mostra uma visão de como a equação é usada neste processo. Primeiramente, é preciso ter o tamanho de cada segmento ($tamanhoSegmento$), o qual foi descoberto pelo processo de calibração (descrito no final da seção 5.2.3). Também é preciso ter as coordenadas da "articulação de referência" (conhecido) $P(x_1, y_1, z_1)$, esta articulação é formada pelas coordenadas de $x$ e $y$ adquiridos na inferência das articulações em 2D, juntamente com a coordenada $z$ da 'Base da coluna' conseguido pelo kinect. Já as coordenadas da "posição da articulação que se quer determinar a profundidade" $P'(x', y', z')$, são usadas as coordenadas de $x$ e $y$ resultante da inferência das articulações em 2D, e o valor de $z'$ será o valor obtido no final deste processo.

Figura 33 – Informações para encontrar o eixo z das articulações



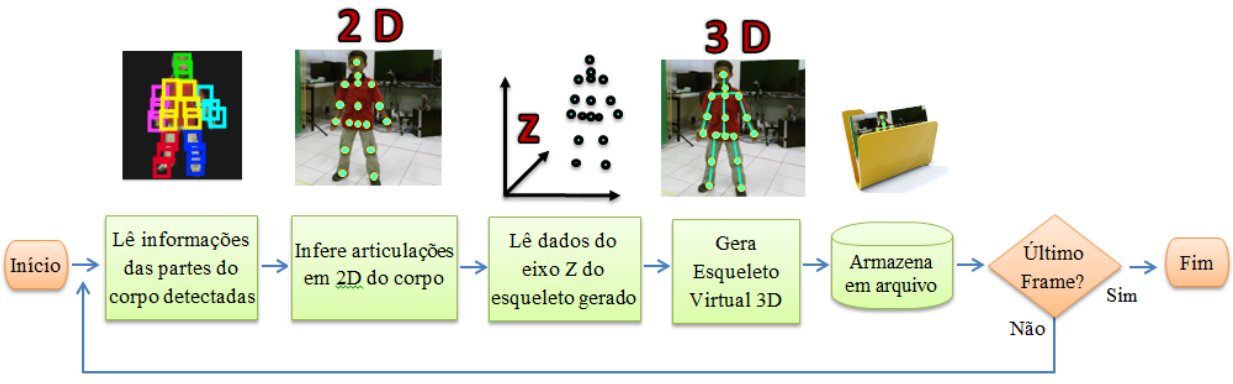

Fonte: Elaborado pelo autor.

Após aplicar a equação (5.3), existirá dois possíveis valores para $z'$, portanto, a escolha de qual dos dois usar é decidido verificando se o mesmo segmento gerado pelo kinect concorda ou não com a orientação positiva do eixo $z$.

Esta equação deve ser aplicada substituindo a "articulaçao de referência" $P(x_1, y_1, z_1)$ pela "articulação $P'(x', y', z')$ que é a articulação que já foi determinada a profundidade", e substituindo a articulação $P'(x', y', z')$ pela posição da próxima articulação que se quer determinar a profundidade, que será a articulação que está na ponta do próximo segmento, Isto se repete até que todo esqueleto seja coberto.

Como em todas as outras etapas do método descrito neste capítulo, também foi criada uma visão da implementação desta última, que pode ser vista na Figura 34.

Figura 34 – Fluxograma da extração de fundo da imagem RGB




Fonte: Elaborado pelo autor.

# 6 TESTES E ANÁLISE DE RESULTADOS

Neste capítulo está descrito o processo experimental, desde a configuração do cenário até a aplicação dos testes que geraram parâmetros comparativos.

Os testes foram iniciados com experimentos mais simples, como a aplicação do algoritmo do Kinect, que usam somente profundidade, visando capturar e validar as informações, bem como, identificar problemas que poderiam ocorrer com o uso de apenas informações de profundidade.

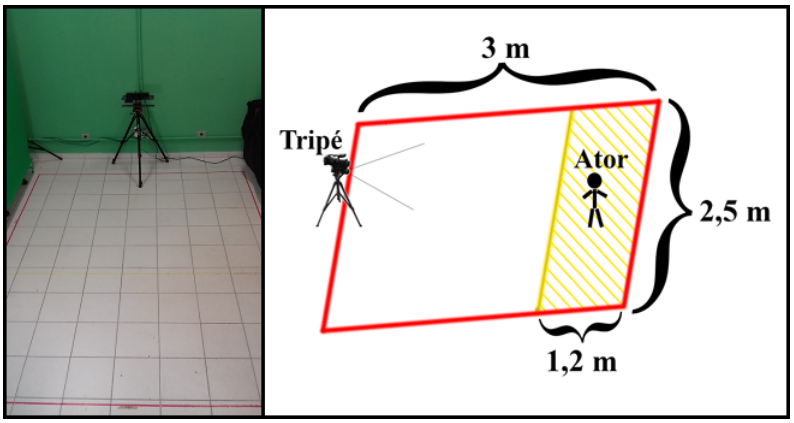
Os experimentos seguintes, foram realizados utilizando o método desenvolvido combinando informações de profundidade com texturas, visando comparar as informações resultantes com às oferecidas pelo método do Kinect.

## 6.1 Configuração do Ambiente Experimental

O ambiente utilizado nos experimentos tem como base um retângulo de 3,00 X 2,50 metros demarcado no chão. A área de atuação do ator foi restrita a um retângulo dentro da área base do cenário, medindo 1,20 X 2,50 metros.

O Tripé acompanhado com os dispositivos fixados a ele, foi posicionado em um dos lados menores do retângulo, ao centro, sobre a fita, a uma altura de 0,94 metros, considerando sua parte inferior (pé) até o chão, paralelo ao nível do solo. A Figura 35 ilustra o ambiente experimental.

Figura 35 – Ambiente Experimental




Fonte: Elaborado pelo autor.

Para a fixação do dispositivo Kinect ao tripé, foi criado um suporte de metal especialmente projetado para esta finalidade. A Figura 36 mostra o tripé, juntamente com o Kinect v2 acoplado ao suporte criado.

Figura 36 – Tripé com suporte especial para o Kinect v2.

Fonte: Elaborado pelo autor.

Todas as medidas e posições adotadas levam em conta a disponibilidade de espaço físico e as características do hardware de captura (Kinect), tais como, campo de visão horizontal (70 graus) e vertical (60 graus), distância mínima (∼0,5 metros) e máxima de profundidade (∼4,5 metros).

## 6.2 Captura e Validação das Informações

Foram realizados alguns testes buscando comprovar as funcionalidades do Kinect, no que tange a resolução espacial e temporal das informações RGB e de profundidade, bem como, do esqueleto virtual fornecido pelo equipamento. Encontra-se na tabela 1, as especificações dos dados referentes a captura das informações.

Tabela 1 – Especificações das informações obtidas por meio do Kinect v2.

| **Dados Capturados** | **Tamanho** | **Fps** |
|---|---|---|
| Imagem RGB | 1920 x 1080 pixels para cada cor | 30 |
| Mapa de profundidade | 512 x 424 pixels | 30 |
| Esqueleto virtual | 26 articulações (posição e orientação em 3D) | 30 |

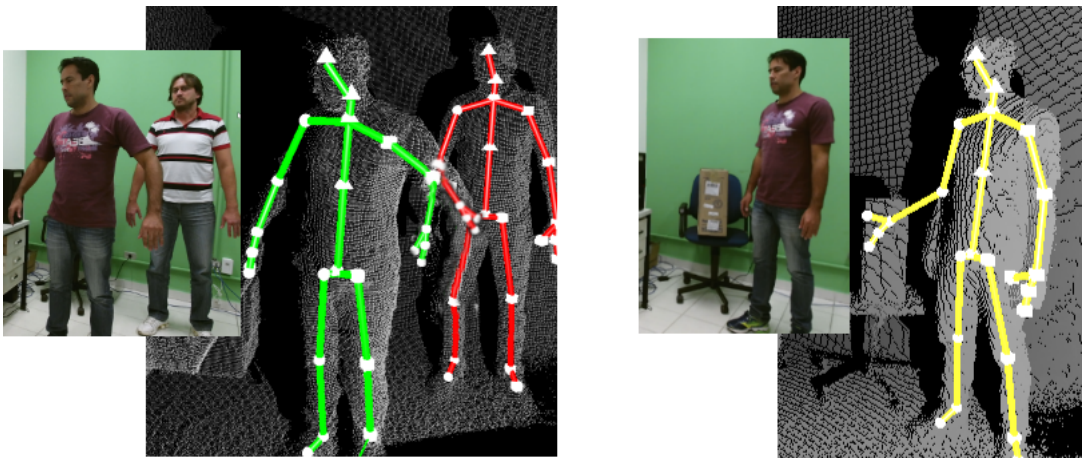
Fonte: Elaborado pelo autor.

Foi constatado que o equipamento junto com seu algoritmo, fornece informações de profundidade, RGB e esqueleto virtual com uma uma taxa de até 30 fps, possibilitando um retorno em tempo real para o ator.

## 6.3 Situações de Erro na Captura de Movimentos com o Kinect

Foram testadas situações em que apenas as informações de profundidade são insuficientes para a correta captura de movimentos, tais como, a oclusão de parte do corpo humano por outros objetos pessoas e a realização de movimentos fora do volume de captura da câmera de profundidade. Como esperado, nestas situações, a captura conteve erros. Também foram observados erros em algumas situações específicas como mostrada na Figura 37, com objetos na mão, aproximação de alguns objetos e mesmo com pequenas oclusões.

Figura 37 – Erros apresentados pelo Kinect



Fonte: Elaborado pelo autor.

## 6.4 Imagem de Entrada Com ou Sem Fundo

Este experimento visa comparar o desempenho do método desenvolvido, usando como entrada a imagem RGB da maneira como ela foi capturada e com o uso da mesma imagem após o processo de retirada do fundo.

Foram utilizadas 176 imagens com fundo e as mesmas sem o fundo, nas quais o ator faz movimentos sentando e levantando do chão, mantendo-se em pé segurando um bastão, realizando movimentos principalmente com os braços, e aproximando-se de objetos que estão na cena.

O experimento foi realizado aplicando o método desenvolvido nas imagens com o fundo, e posteriormente aplicando nas imagens sem o fundo, os resultados das duas aplicações foram comparados com o *ground truths* de cada imagem. O *ground truths* corresponde a imagem de referência, na qual a localização de cada junta foi posicionada manualmente. O processo manual de geração de *ground truths* é descrito em detalhes na seção 6.5.1.

Para a determinação da acurácia neste experimento foi calculada a média da distância euclidiana para cada articulação detectada em cada frame de uma sequência de entrada de 176 frames(com e sem extração do fundo) em comparação com o ground truth de cada frame. Os resultados obtidos são exibidos na Tabela 2.

Tabela 2 – Resultados da aplicação do método com relação as imagens com e sem extração de fundo.

| **Imagens** | **Imagens (qtde)** | **Acurácia** |
|---|---|---|
| Sem o fundo | 176 | 67,61% |
| Com o fundo | 176 | 32,39% |

Fonte: Elaborado pelo autor.

Como pode ser observado na Tabela 2, além das imagens sem o fundo atingirem maior acurácia, o tempo de execução do método com estas imagens foi em média 20% mais rápido que o tempo gasto usando as imagens com fundo. Sendo que, o tempo médio com o uso das imagens sem o fundo foram 20 segundos, e com a utilização das imagens com o fundo foram 25 segundos. Consequentemente, a extração do fundo foi adotada como uma etapa para o método desenvolvido nesta dissertação.

## 6.5 Método Desenvolvido e o Usado pelo Kinect

Segundo Sigal, Balan e Black (2010), a avaliação quantitativa da estimativa e rastreamento de postura humana é atualmente limitada devido à falta de conjuntos de dados (*datasets*) comuns contendo "*ground truths*" com os quais se pode testar e comparar algoritmos. Em vez disso, os testes qualitativos ainda são amplamente utilizados e a avaliação muitas vezes depende da inspeção visual dos resultados. Esta abordagem é a que foi utilizada para a validação dos resultados obtidos a partir do método desenvolvido. Portanto, o *ground truths*, no presente trabalho, foi gerado manualmente e não por meio da detecção de marcadores reais colocados sobre o corpo do ator.

### 6.5.1 Ground Truths

Para possibilitar a comparação entre o método desenvolvido e os dados obtidos a partir do uso do SDK do Kinect, foi necessário implementar uma ferramenta, onde são exibidas as imagens capturadas de cada *frame*, para a identificação e apontamento manual das articulações do corpo do ator, como resultado, as informações são armazenadas para serem usadas como *ground truths*, que consistem em referências da melhor localização possível do esqueleto.

Como pode ser visto na Figura 38, a ferramenta oferece uma única interface a qual exibe a imagem RGB de um *frame* capturado, ao posicionar o mouse sobre a imagem o cursor ficará em forma de um " + ", basta posicionar o cursor do mouse sobre a articulação e clicar com o botão esquerdo, assim, um ponto verde indicará o local marcado.

Figura 38 – Interface da ferramenta criada para o posicionamento manual das articulações sobre a imagem do ator.


Fonte: Elaborado pelo autor.

As articulações devem ser marcadas conforme a sequência indicada na tabela 3.

Tabela 3 – Sequência de marcação das articulações para o Ground Truth

| **Sequência** | **Articulações** |
|---|---|
| 1 | Base da coluna |
| 2 | Meio da coluna |
| 3 | Coluna na altura dos ombros |
| 4 | Pescoço |
| 5 | Cabeça |
| 6 | Ombro do lado esquerdo da imagem |
| 7 | Cotovelo do lado esquerdo da imagem |
| 8 | Pulso do lado esquerdo da imagem |
| 9 | Ombro do lado direito da imagem |
| 10 | Cotovelo do lado direito da imagem |
| 11 | Pulso do lado direito da imagem |
| 12 | Quadril do lado esquerdo da imagem |
| 13 | Joelho do lado esquerdo da imagem |
| 14 | Tornozelo do lado esquerdo da imagem |
| 15 | Quadril do lado direito da imagem |
| 16 | Joelho do lado direito da imagem |
| 17 | Tornozelo do lado direito da imagem |

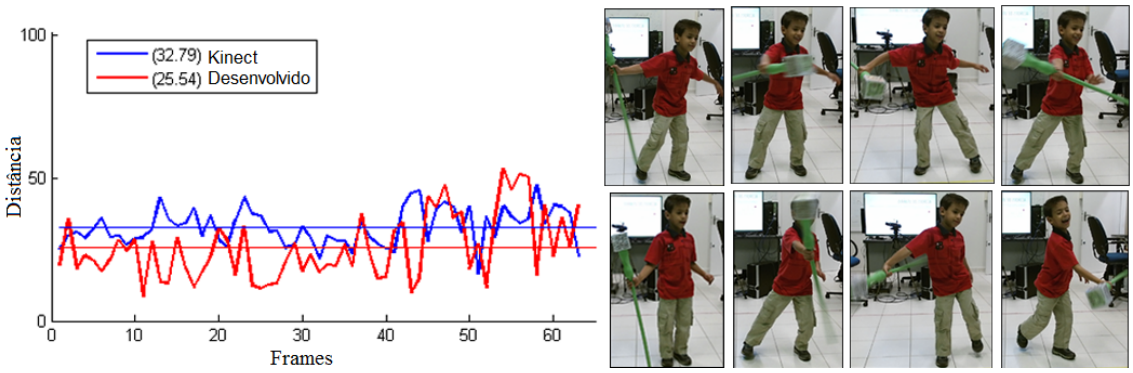
Fonte: Elaborado pelo autor.

### 6.5.2 Avaliação Qualitativa

Para uma avaliação qualitativa, foi comparada a acurácia do método desenvolvido com o melhor resultado obtido manualmente (*ground truth*), a mesma comparação foi feita usando-se exclusivamente o método usado pelo Kinect.

A acurácia foi medida calculando-se a distância euclidiana de cada articulação do esqueleto (no método desenvolvido e no método do Kinect) com o esqueleto (*ground truth*), em seguida foram calculadas as médias das distâncias. Analisando-se o gráfico da Figura 39, pode-se notar que o método desenvolvido se manteve com uma menor distância média, assim, foi 22,11% mais eficiente que o do Kinect para este caso. A amostra usada continha 63 *frames*, o ator se manteve em pé segurando um bastão e realizando movimentos principalmente com os braços. Os resultados obtidos podem ser visualizados no gráfico da Figura 39.

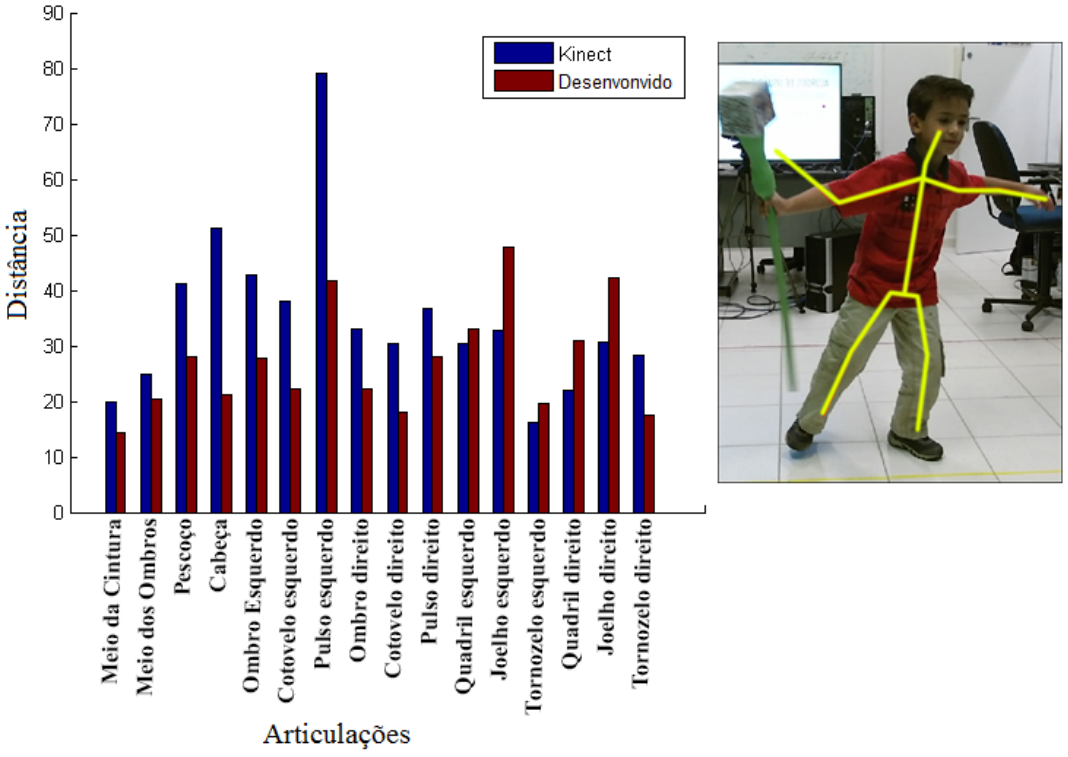
Figura 39 – Média da distância das articulações por *frame*




Fonte: Elaborado pelo autor.

O gráfico mostrado na Figura 40 foi gerado calculando-se a média da distância de cada uma das articulações entre todos os *frames*, dessa forma, pode-se perceber que a articulação que apresentou o pior resultado no esqueleto gerado pelo Kinect, é exatamente à do pulso do braço que está segurando o bastão, como pode-se observar na imagem que está do lado direito do gráfico, confirmando que o método desenvolvido foi capaz de superar alguns dos erros gerados pelo Kinect, os quais foram descritos na seção 6.3.

Figura 40 – Média das articulações por *frame*




Fonte: Elaborado pelo autor.

### 6.5.3 Avaliação Visual

Observando-se a Figura 41, nota-se que, em diversas situações, o método desenvolvido apresentou uma captura mais fiel ao movimento realizado do que quando foram consideradas apenas as informações de profundidade, como faz o método adotado pelo Kinect.

Figura 41 – Avaliação visual das articulações.


Fonte: Elaborado pelo autor.

# 7 CONCLUSÕES

Este trabalho assumiu como objetivo o desenvolvimento de um método de captura de movimentos humanos utilizando um único dispositivo RGB-D, empregando, tanto informações de textura da imagem, quanto de profundidade, visando aumentar a acurácia dos movimentos inferidos em relação ao método que usa somente informações de profundidade empregado no dispositivo Kinect.

Foram observados erros no posicionamento das articulações rastreadas pelo Kinect, em situações específicas, como por exemplo, quando o ator segura ou se aproxima de objetos, ao se abaixando e mesmo quando ocorre pequenas oclusões.

Após experimentos realizados com imagens contendo fundo e com imagens com o fundo extraído, concluiu-se que o uso das imagens sem o fundo alcançaram um melhor desempenho, tendo 67,61% mais acurácia, e também considerando o tempo de execução do método, foi em média 20% mais rápido. Portanto, uma fase de extração de fundo das imagens RGB foi incorporado ao método desenvolvido.

Na realização da avaliação quantitativa, foi comparada a acurácia do método desenvolvido com a do método do Kinect, usando como referência o *ground truths* gerado manualmente. Conclui-se que o método desenvolvido mostrou uma menor distância média de seu esqueleto com o esqueleto tomado como correto, assim, foi 22,11 % mais acurado que o do método do Kinect para o caso testado. Também foi calculada a média da distância de cada uma das articulações para todos os *frames*, com isso, notou-se que a articulação que apresentou a maior discrepância, é exatamente à do pulso do braço que está segurando o bastão, confirmando que o método desenvolvido foi capaz de diminuir os erros gerados com o método do Kinect.

Uma avaliação qualitativa foi realizada, mostrando que em diversos casos que ocorrem erros gerados pelo rastreio do Kinect, pode-se perceber até visualmente que o método desenvolvido conseguiu reduzir tais erros.

No entanto, o método desenvolvido, por utilizar informações de textura da imagem somente pode ser utilizado em situações com boa iluminação, ao contrário do Kinect que se baseia apenas nas informações de profundidade obtidas por meio do reconhecimento de sinais infravermelhos, os quais funcionam com pouca ou nenhuma iluminação. Um outro aspecto a ser considerado é que o Kinect é capaz de rastrear 25 articulações, enquanto o método desenvolvido faz uso de apenas 17 articulações, devido a quantidade de informações conseguidas na fase de análise dde textura para detectar as partes do corpo do ator.

Dada a importância do tema, considera-se ainda que existe muito a percorrer nesta

área, sendo assim, é um campo fértil de pesquisa.

Trabalhos futuros poderiam incluir o uso de câmeras RGB externas ao dispositivo Kinect, bem como o uso simultâneo de múltiplas câmeras RGB e múltiplos dispositivos como Kinect, além da possibilidade de aumentar-se a quantidade de articulações detectadas.

# Referências

ASCENSION TECHNOLOGY CORPORATION. **Manufacturer of DriveBAY™ and TrakSTAR™ electromagnetic tracking systems; Electromagnetic transmitters; 6DOF sensors.** 120 Graham Way, Suite 130, Shelburne, VT 05482 , USA, 2015. Disponível em: <http://www.ascension-tech.com>. Acesso em: 14.07.2015.

BASRI, R.; JACOBS, D.; KEMELMACHER, I. Photometric stereo with general, unknown lighting. **International Journal of Computer Vision**, Springer, v. 72, n. 3, p. 239–257, 2007. ISSN 1573-1405. Disponível em: <http://dx.doi.org/10.1007/s11263-006-8815-7>.

DANCIU, G.; BANU, S. M.; CăLIMAN, A. Shadow removal in depth images morphology-based for kinect cameras. In: IEEE. **System Theory, Control and Computing (ICSTCC), 2012 16th International Conference on**. 2012. p. 1–6. ISBN 978-1-4673-4534-7. Disponível em: <http://ieeexplore.ieee.org/document/6379195/>.

FAUGERAS, O. **Three-dimensional Computer Vision: A Geometric Viewpoint**. MIT Press, 1993. 663 p. (Artificial intelligence). ISBN 9780262061582. Disponível em: <https://books.google.com.br/books?id=Aa6TTW9dWy0C>.

FLAM, D. L. **OpenMoCap: uma aplicação de código livre para a captura óptica de movimento**. Dissertação (Mestrado) — Universidade Federal de Minas Gerais. Departamento de Ciência da Computação., 2009.

FURNISS, M. Motion capture. In: . Electronic document, Papers, MIT - Massachusetts Institute of Technology, 1999. Disponível em: <http://web.mit.edu/comm-forum/papers/furniss.html>. Acesso em: 18.07.2015.

GALL, J. et al. Motion capture using joint skeleton tracking and surface estimation. In: IEEE. **Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on**. [S.l.], 2009. p. 1746–1753. ISSN 1063-6919.

HANSARD, M. et al. **Time-of-Flight Cameras: Principles, Methods and Applications**. Springer London, 2012. (SpringerBriefs in Computer Science). ISBN 9781447146582. Disponível em: <https://books.google.com.br/books?id=PiF4narL1Z0C>.

HARTLEY, R. I.; STURM, P. Triangulation. **Computer vision and image understanding**, Elsevier, v. 68, n. 2, p. 146–157, 1997.

HELTEN, T. et al. Personalization and evaluation of a real-time depth-based full body tracker. In: IEEE. **2013 International Conference on 3D Vision - 3DV 2013**. [S.l.], 2013. p. 279–286. ISSN 1550-6185.

HEXAMITE LTD. **Ultrasonic Industrial Positioning Systems, and ranging - Hx19, VAT: BG147157269**. 55 Nessebar Fort Club, Sunny Beach 8240, European Union, Bulgaria, 2015. Disponível em: <http://www.hexamite.com/>. Acesso em: 13.07.2015.

KABAYAMA, A.; TRABASSO, L. Performance evaluation of 3d computer vision techniques. **Journal of the Brazilian Society of Mechanical Sciences**, SciELO Brasil, v. 24, n. 3, p. 234–238, 2002. Disponível em: <http://dx.doi.org/10.1590/S0100-73862002000300013>.

KADAMBI, A.; BHANDARI, A.; RASKAR, R. 3d depth cameras in vision: Benefits and limitations of the hardware. In: _____. **Computer Vision and Machine Learning with RGB-D Sensors**. Cham: Springer International Publishing, 2014. p. 3–26. ISBN 978-3-319-08651-4. Disponível em: <http://dx.doi.org/10.1007/978-3-319-08651-4_1>.

LANDER, J. Working with motion capture file formats. **Game Developer Magazine**, Manhasset, New York, USA, v. 5, n. 1, p. 30–37, 1998. Disponível em: <www.darwin3d.com/gamedev/articles/col0198.pdf>.

LIU, Y. et al. Markerless motion capture of multiple characters using multiview image segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 35, n. 11, p. 2720–2735, Nov 2013. ISSN 0162-8828.

MENACHE, A. **Understanding motion capture for computer animation and video games**. illustrated. [S.l.]: Morgan kaufmann, 2000. v. 1. 238 p.

META MOTION. **Manufacturer of Gypsy™, IGS-190™ and more**. 268 Bush, St. 1 San Francisco, CA 94104, USA, 2015. Disponível em: <http://www.metamotion.com>. Acesso em: 14.07.2015.

MEYER, K.; APPLEWHITE, H. L.; BIOCCA, F. A. A survey of position trackers. **Presence: Teleoperators & Virtual Environments**, MIT Press, v. 1, n. 2, p. 173–200, 1992. Disponível em: <http://dl.acm.org/citation.cfm?id=196564.196568>.

MICROSOFT. **Kinect for Windows Software Development Kit (SDK) 2.0**. 2016. Disponível em: <https://developer.microsoft.com/en-us/windows/kinect/develop>. Acesso em: 11 nov. 2016.

PHASE SPACE INC. **Optical Motion Capture System**. 1933 Davis Street, Suite 304, San Leandro, CA 94577, USA, 2015. Disponível em: <http://http://www.phasespace.com>. Acesso em: 15.07.2015.

SELL, J.; O'CONNOR, P. The xbox one system on a chip and kinect sensor. **IEEE Micro**, v. 34, n. 2, p. 44–53, Mar 2014. ISSN 0272-1732.

SHOTTON, J. et al. Real-time human pose recognition in parts from a single depth image. In: **CVPR**. IEEE, 2011. Disponível em: <https://www.microsoft.com/en-us/research/publication/real-time-human-pose-recognition-in-parts-from-a-single-depth-image/>.

SIGAL, L.; BALAN, A. O.; BLACK, M. J. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. **International Journal of Computer Vision**, Kluwer Academic Publishers, Hingham, MA, USA, v. 87, n. 1-2, p. 4–27, mar 2010. ISSN 0920-5691. Disponível em: <http://dx.doi.org/10.1007/s11263-009-0273-6>.

TERVEN, J. R.; CóRDOVA-ESPARZA, D. M. Kin2. a kinect 2 toolbox for matlab. **Science of Computer Programming**, Elsevier, v. 130, p. 97 – 106, 2016. ISSN 0167-6423. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167642316300569>.

YANG, R. Motion capture from rgb-d camera. In: . Electronic document, Slides, CVPR'2014, University of Kentucky, 2014. p. 44. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9B05E2CBED9FD36B738423C272F41148?doi=10.1.1.677.4564&rep=rep1&type=pdf>. Acesso em: 19.11.2016.

YANG, Y.; RAMANAN, D. Articulated human detection with flexible mixtures of parts. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 35, n. 12, p. 2878–2890, 2013.

YE, M. et al. Accurate 3d pose estimation from a single depth image. In: IEEE. **2011 International Conference on Computer Vision**. [S.l.], 2011. p. 731–738. ISSN 1550-5499.

YE, M.; YANG, R. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In: IEEE. **2014 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.], 2014. p. 2353–2360. ISSN 1063-6919.

ZENNARO, S. et al. Performance evaluation of the 1st and 2nd generation kinect for multimedia applications. In: **2015 IEEE International Conference on Multimedia and Expo (ICME)**. [S.l.: s.n.], 2015. p. 1–6. ISSN 1945-7871.

ZHANG, L.; NAYAR, S. Projection defocus analysis for scene capture and image display. In: **ACM SIGGRAPH 2006 Papers**. New York, NY, USA: ACM, 2006. (SIGGRAPH '06), p. 907–915. ISBN 1-59593-364-6. Disponível em: <http://doi.acm.org/10.1145/1179352.1141974>.

ZHANG, Z. Microsoft kinect sensor and its effect. **IEEE MultiMedia**, IEEE, v. 19, n. 2, p. 4–10, Feb 2012. ISSN 1070-986X. Disponível em: <http://ieeexplore.ieee.org/document/6190806/?reload=true&arnumber=6190806>.