UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS TECNOLÓGICAS
CURSO DE ENGENHARIA DA COMPUTAÇÃO

ARIANNE DOS SANTOS FERREIRA EVANGELISTA

SENSOR DE PROFUNDIDADE DO TIPO MICROSOFT KINECT COMO FERRAMENTA DE PESQUISA

São Luís
2017

UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS TECNOLÓGICAS
CURSO DE ENGENHARIA DA COMPUTAÇÃO

ARIANNE DOS SANTOS FERREIRA EVANGELISTA

SENSOR DE PROFUNDIDADE DO TIPO MICROSOFT KINECT COMO FERRAMENTA DE PESQUISA

Monografia apresentada ao Curso de Engenharia da Computação da UEMA, como registro para obtenção parcial do grau de Bacharel em Engenharia da Computação com ênfase em Automação e Controle.

São Luís
2017

Evangelista, Arianne dos Santos Ferreira.

Sensor de profundidade do tipo microsoft kinect como ferramenta de pesquisa / Arianne dos Santos Ferreira Evangelista. – São Luís, 2017.

58 f.

Monografia (Graduação) – Curso de Engenharia de Computação, Universidade Estadual do Maranhão, 2017.

Orientador: Prof. Denner Robert Rodrigues Guilhon.

1. Câmeras de profundidade. 2. Kinect. 3. Interferência. 4. Sistema multicâmera. 5. Calibração. I. Título.

CDU 004.4

ARIANNE DOS SANTOS FERREIRA EVANGELISTA

SENSOR DE PROFUNDIDADE DO TIPO MICROSOFT KINECT COMO FERRAMENTA DE PESQUISA

Monografia apresentada ao Curso de Engenharia da Computação na UEMA, como registro para obtenção parcial do grau de Bacharelado em Engenharia da Computação com ênfase em Automação e Controle.

Trabalho aprovado. São Luís, 30 de janeiro de 2017:

Msc. Denner Robert Rodrigues Guilhon
Orientador

PhD. Mauro Sergio Silva Pinto
Primeiro Membro da Banca

Msc. Simone Cristina Ferreira Neves
Segundo Membro da Banca

Dedico este trabalho àqueles que dedicaram a vida a cuidar de mim: meus pais!

AGRADECIMENTOS

Agradeço, primeiramente, a Deus porque nada seria possível sem Ele que demonstra sempre Sua misericórdia infinita.

Agradeço imensamente à minha família, de um modo especial, àqueles que fizeram e fazem de tudo diariamente para que nada me falte, meus pais: Josefa Maria Freitas Ferreira e Antonio Carvalho Evangelista, pelo imenso amor, cuidado e exemplo que sempre são.

Sou muito grata aos meus amigos que de algum modo me apoiaram, incentivaram e torceram (torcem) pelas minhas vitórias, mas aqui faço lembrança de alguns: Thaynara, Henrique, Juliana, Joel, Sthephane, Marcos, Wesley, Ricardo, Andressa, Vanessa,Rafael, Elaine, Ministério de Dança Shalom, Durcyanne e todos os amigos da Comunidade Católica Shalom. Porém, quero expressar minha gratidão, em particurlar a: Rayssa Lopes Campos, pelo companheirismo (do seu jeito), pelos estudos, incentivo, conselhos e pelas nossas conversas que só nós compreendemos; Paulo Renato, por toda disposição, ajuda e incentivo; Kalyanne, Adrielle e Mayara, pelos risos, orações, abraços, prantos, ensaios e lanches, pois me ajudaram a encarar os problemas com sorriso no rosto; Sofia Pizzato Scomazzon, por todo incentivo, por servir de modelo para minha vida, pelas conversas, conselhos, pela irmandade e porque mesmo do outro lado do mundo se faz presente diariamente.

Vorrei ringraziare anche a mia cara amica Jhoanna Climacosa per le preghiere e per essere sempre al mio fianco, anche se sia così lontana. In modo davvero speciale vorrei ringraziare al PhD. Aldo Genovesio per mi avere dato una grande opportunità di fare tirocinio in labratorio suo e perché mi ha insegnato a lavorare con l'amore.

Finalmente, meus mais sinceros agradecimentos ao meu orientador Denner Robert Rodrigues Guilhon, por ter me acolhido, por por todos os ensinamentos, pela extrema paciência (eu sei que precisou de muita!), per le chiacchierate troppo divertenti e le esperienze scambiate, mas acima de tudo, pela amizade, muito obrigada!

"Enquanto houver vontade de lutar haverá esperança de vencer."
(Santo Agostinho)

RESUMO

Os avanços nas técnicas de escaneamento mais acessíveis através de sensores 3D, tornaram possível que objetos sejam identificados, segmentados e medidos em tempo real, sem a necessidade de marcadores ou outros recursos especiais. Para se conseguir o máximo de informações de um cena qualquer, muitas vezes, torna-se necessário criar um sistema de múltiplos sensores. Atualmente, existem inúmeros dispositivos que podem realizar este tipo de aquisição. O sensor Kinect foi desenvolvido para jogos, porém passou a ser utilizado em diversas áreas de pesquisa por possuir várias funcionalidades, em se tratando de aquisição de imagem 3D e, também, pelo seu baixo custo. Assim, o presente trabalho discorrerá sobre as características do Kinect, suas vantagens e desvantagens, e fará comparações com dispositivos similares de captura de profundidade.

Palavras-chave: Câmeras de profundidade. Kinect. Interferência. Sistemas multicâmera. Calibração.

ABSTRACT

The progress in the scanning techniques by means of three dimensional sensors has made possible the identification, segmentation and real time measuring of objects without any special resource. In these days, there are multiples devices that can perform this acquisition. The Kinect sensor was developed primarily for games, however it has been popular in many other fields of research as it has numerous functionalities, e.g. three dimensional image acquisition and low-price. In this way, this text aims to present the main features of Kinect, its advantages and disadvantages as well as some comparisons with similar devices of three dimensional acquisition..

Keywords: Depth cameras. Kinect. interference. multicamera systems. calibration.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

SDK	*Software Development Kit* (Kit de desenvolvimento de software)
RGBD	*Red, Blue, Green e Depth,* (Imagem de Profundidade)
ToF	*Time of Flight* (Tempo de Voo - Cálculo de distância com base na diferença de fase entre um sinal modulado de onda emitido e o sinal demodulado recebido)
IR	*Infrared* (Infravermelho)
CCD	*Charge-Coupled Device* (Decomposição de Valores Singulares)
DSP	*Digital Signal Processor* (Processador Digital de Sinais)
ADC	*Analog to Digital Converter* (Conversor Analógico Digital)
SoC	*System on Chip* (Circuito Integrado com o maior número possível de componentes)
FPS	*Frames* por segundo

SUMÁRIO

1 INTRODUÇÃO

A visão computacional é uma área que tem prosperado com a evolução da tecnologia. Surgiu com os primeiros computadores e tem evoluído gradativamente à medida que mais recursos computacionais se tornam disponíveis. Um dos primeiros trabalhos nesta área foi desenvolvido por Roberts em 1963 (BACKES, 2016) propondo um dos primeiros detectores de borda. Atualmente, pesquisas na área de visão computacional são desenvolvidas para resolução de problemas específicos, como reconhecimento facial, análise de formas, segmentação de imagens médicas, análise de movimento, etc.(BACKES, 2016).

Dentro da visão computacional existem inúmeros desafios, um deles é fazer com que um sistema computacional "dotado de visão" através de uma câmera, tenha noção de profundidade. Para solucionar esse problema utiliza-se uma fonte de captura visual estéreo, ou seja, unir duas ou mais câmeras arranjadas de modo a possibilitar a triangulação dos pontos correlatos para cálculo de coordenada de profundidade.

Recentes avanços em câmeras de profundidade 3D, tais como o sensor Microsoft Kinect, criaram muitas oportunidades para computação multimídia. O Kinect foi criado pela Microsoft para ser usado em jogos de videogame, porém pode ser usado não somente como parte do console, mas também, com o computador através de um SDK (*Software Development Kit*) que a Microsoft disponibiliza para ser instalado no Windows e fazer uso das inúmeras aplicações disponíveis no Software, como por exemplo, reconhecimento facial ou reconhecimento do esqueleto humano. Por este motivo, ganhou grande popularidade na comunidade científica, onde pesquisadores desenvolveram uma enorme quantidade de aplicações inovadoras que estão relacionados a diferentes áreas, tais como a reconstrução 3D on-line, aplicações médicas, realidade aumentada (SARBOLANDI, 2015), dentre outras.

Pode-se interagir em jogos com o próprio corpo de uma forma natural quando se usa o sensor Kinect, este tipo de dispositivo é classificado como: Interface de Interação Natural (*Natural User Interface - NUI*). A ideia chave do funcionamento desta tecnologia é a compreensão da linguagem corporal humana: o computador deve primeiro entender o que o usuário está fazendo antes que ele possa responder. Este tem sido um campo de pesquisa ativo em visão computacional, mas mostrou-se formidavelmente difícil com câmeras de vídeo. O sensor Kinect permite que o computador detecte diretamente a terceira dimensão (profundidade) dos usuários e do ambiente, tornando a tarefa muito mais fácil. Também, entende quando os usuários conversam, sabe quem são quando andam em direção a ele, e pode interpretar seus movimentos e traduzi-los em um formato que os desenvolvedores possam usar para criar novas experiências (ZHANG, 2012). Este dispositivo é capaz de realizar tais capturas por incorporar vários hardwares avançados de detecção. Mais notavelmente, ele contém um sensor de profundidade que possibilita a

captura de movimento de corpo inteiro em 3D, uma câmera colorida que dá a possibilidade de fazer reconhecimento facial e uma matriz de quatro microfones que fornece recursos de reconhecimento de voz (ZHANG, 2012).

O presente trabalho discorrerá sobre o sensor Microsoft Kinect como ferramenta de pesquisa, levando em conta suas as características, aplicações (biomédicas, robótica, reconhecimento facial, etc), limitações, vantagens e desvantagens frente a outras técnicas e câmeras similares.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Desenvolver um estudo sobre o sensor Kinect, suas características, vantagens e desvantagens perante dispositivos similares, criando assim um material que sirva de referência para projetos futuros que utilize o Kinect como ferramenta de pesquisa.

2.2 OBJETIVOS ESPECÍFICOS

1. Levantar o estado da arte em relação ao uso de múltiplas câmeras para captura de imagem 3D ;
2. Levantar material bibliográfico sobre o uso do Kinect em pesquisas;
3. Utilizar bibliografia para estudar as características do Kinect;
4. Utilizar bibliografia para comparar o Kinect 1.0 com o Kinect 2.0;
5. Levantar uma discussão sobre as vantagens e desvantagens do Kinect comparados com outros dispositivos de captura 3D.

3 JUSTIFICATIVA E MOTIVAÇÃO

Nos últimos anos tem havido um desenvolvimento significativo de sensores ópticos, o que levou a um número crescente de aplicações, como rastreamento de objetos e pessoas, captura e análise de movimento, animação de personagens, reconstrução de cenas 3D, interfaces de usuário com base em gestos e aplicações científicas.

O Kinect é um sensor que permite novas interações durante jogos, com base no uso de gesto e voz. Desde sua apresentação em 2010, tem atraído pesquisadores de diferentes áreas (robótica (MARTíN; LORBACH; BROCK, 2014), engenharia biomédica (MAFRA, 2012) e visão computacional(ARAúJO, 2010)). Pouco depois do lançamento do Kinect, um SDK foi criado pela própria Microsoft, permitindo que o sensor fosse usado não apenas como um dispositivo de jogo, mas também como um sistema de medição. Por esses motivos o Kinect tem sido utilizado em diversas aplicações, que vai desde o entretenimento, exercícios, aplicações médicas simples e até mesmo em terapias físicas (SCHöNAUER, 2011).

A maioria das aplicações depende de apenas uma câmera devido à sua fácil implementação. No entanto, com aplicações mais complicadas como uma análise da caminhada humana, é necessário usar duas ou mais câmeras para capturar a pessoa de diversos ângulos em uma cena, pois a visualização de uma única câmera pode não ser capaz de capturar o usuário como um todo.

Como supracitado, o Kinect tem várias funcionalidades e pode ser utilizado em inúmeras aplicações, por esse motivo, vê-se a necessidade de um estudo sobre as características do Kinect, mostrando suas vantagens e desvantagens perante ferramentas similares que sirva de material de referência para trabalhos futuros.

4 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo explanar toda a teoria que serviu como base para a elaboração deste trabalho.

Inicialmente, serão abordados conceitos de imagem e visão computacional, em seguida os principais problemas de visão computacional serão expostos. Depois será falado sobre a construção da imagem 3D, seguido de um detalhamento sobre as características do Microsoft Kinect e por fim será levantada uma discussão a cerca do sensor Kinect como ferramenta de pesquisa.

4.1 IMAGEM

Pode-se definir uma imagem como uma função de duas dimensões, $f(x, y)$, onde x e y são coordenadas planas, e a amplitude de f em qualquer par de coordenadas (x, y) é chamada de intensidade ou nível de cinza da imagem naquele ponto, também conhecida como imagem monocromática.

Computadores não são capazes de processar imagens contínuas, mas apenas matrizes de números digitais (0 e 1), sendo assim, é necessário representar imagens como arranjos bidimensionais de pontos. O processo para trazer uma função contínua para o computador é discretizando-a (ou digitalizando-a), ou seja, tomando valores pontuais ao longo de x e guardando o valor de $f(x)$ correspondente (o eixo $f(x)$ também é contínuo, assim também precisa-se discretizá-lo) (SCURI, 2002). O processo de discretização do eixo x (o domínio) é chamado de Amostragem, o do eixo $f(x)$ (o contradomínio) é chamado de Quantização. Logo, tem-se uma imagme digital quando x, y, e os valores de amplitude de f estão todos em quantidades discretas e finitas. (PETROU, 1999 apud ARAúJO, 2010).

Uma imagem bidimensional é representada digitalmente por números binários codificados de forma que seja possível armazenar, transferir, reproduzir através de dispositivos eletrônicos. Existem dois tipos fundamentais de imagem digital, uma é do tipo vetorial, que é descrita por posição e tamanho de formas como linhas, curvas, círculos e retângulos e a outra é do tipo rastreio, ou bitmap, que apresenta uma matriz de pixels que correspondem ponto a ponto à imagem que está representando (ARAúJO, 2015).

Uma imagem colorida pode ser vista como a composição de três imagens monocromáticas. Um modelo comum é o RGB, em que um pixel é como um vetor cujas componentes representam as intensidades de vermelho, verde e azul (GOMES, 2014). Sendo que a soma dos valores resulta em um ponto colorido na imagem final (ARAúJO, 2015).

A captura de uma imagem por uma câmera digital moderna registra a impressão

luminosa que chega aos sensores internos dessa câmera. Esta captura consiste em uma projeção de uma cena tridimensional sobre um plano bidimensional (plano este que é o sensor da câmera). Já o mapeamento de uma cena tridimensional sobre um plano da imagem é uma transformação do tipo "muitos para um", isto é, um ponto na imagem não determina unicamente a posição de um ponto correspondente do mundo (GONZALEZ, 1992 apud ARAúJO, 2010).

Em algumas aplicações reais, tais como navegação robótica, mapeamento geográfico, engenharia reversa, etc, são utilizadas esses tipos de câmeras onde a informação da profundidade dos pontos na cena é perdida, logo, percebe-se que a imagem digital muitas vezes revela pouca informação ou contém ruídos e distorções. Por esse motivo, sistemas baseados em visão computacional precisam, na prática, de outras informações descritivas do ambiente, ou outros sensores para solucionar as limitações impostas pelo dispositivo de captura (KAEHLER, 2008).

4.2 VISÃO COMPUTACIONAL

Visão computacional é o estudo da extração de informação de uma imagem; mais especificamente, é a construção de descrições explícitas e claras dos objetos em uma imagem (BROWN, 1982). A Visão Computacional é a transformação de dados de uma câmera fixa ou de vídeo em uma decisão ou em uma nova representação e tais transformações são feitas para alcançar algum objetivo particular. Difere do processamento de imagens porque, enquanto ele se trata apenas da transformação de imagens em outras imagens, a Visão Computacional trata explicitamente da obtenção e manipulação dos dados de uma imagem e do uso deles para diferentes propósitos (RIOS, 2010).

4.3 PROBLEMAS EM VISÃO COMPUTACIONAL

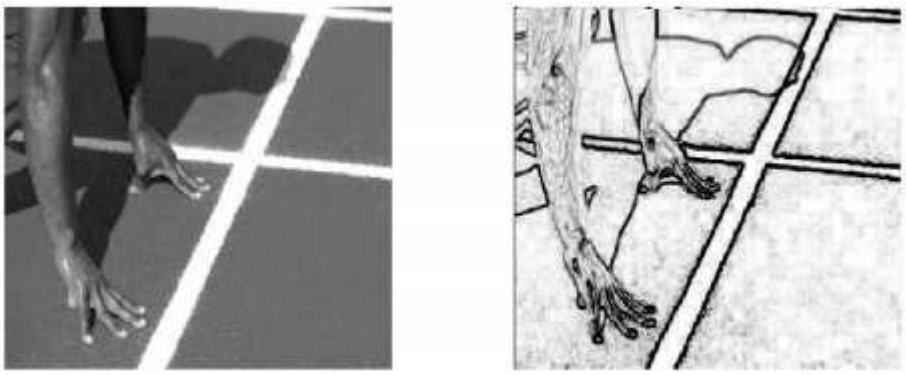
4.3.1 DETECÇÃO DE BORDAS

Um dos problemas mais importantes no reconhecimento de objetos em imagens é a detecção de bordas. Basicamente, consiste em detectar regiões da imagem nas quais ocorre uma mudança abrupta de brilho na imagem, que, geralmente, representam uma mudança das características do que está sendo visto. É um problema importante porque mudanças abruptas no brilho da imagem, sob a percepção natural humana da visão, podem representar: descontinuidade de profundidade – uma parede atrás de outra, por exemplo, geralmente é mais escura; descontinuidade da orientação da superfície – uma face da parede que está mais perpendicular à iluminação é mais clara que uma face que está paralela; mudanças nas propriedades do material – pedras pretas e brancas no chão são pedras diferentes – e variações na iluminação da cena – a pedra cinza do lado de fora

da casa e a pedra preta do lado de dentro são, na verdade, da mesma cor, além de outras características como reflexão e refração, por exemplo.

Em relação a esses aspectos, uma representação por bordas é bem fiel às propriedades físicas do mundo. Além disso, a representação do mundo bidimensional por linhas unidimensionais tem a vantagem de ser compacta, pois leva em conta apenas os detalhes relevantes da imagem, como exemplo tem-se a Figura 1 (RIOS, 2010).

Figura 1 – Exemplo de reconhecimento de borda: à esquerda, imagem original; à direita, bordas da imagem.



Fonte: Rios (2010)

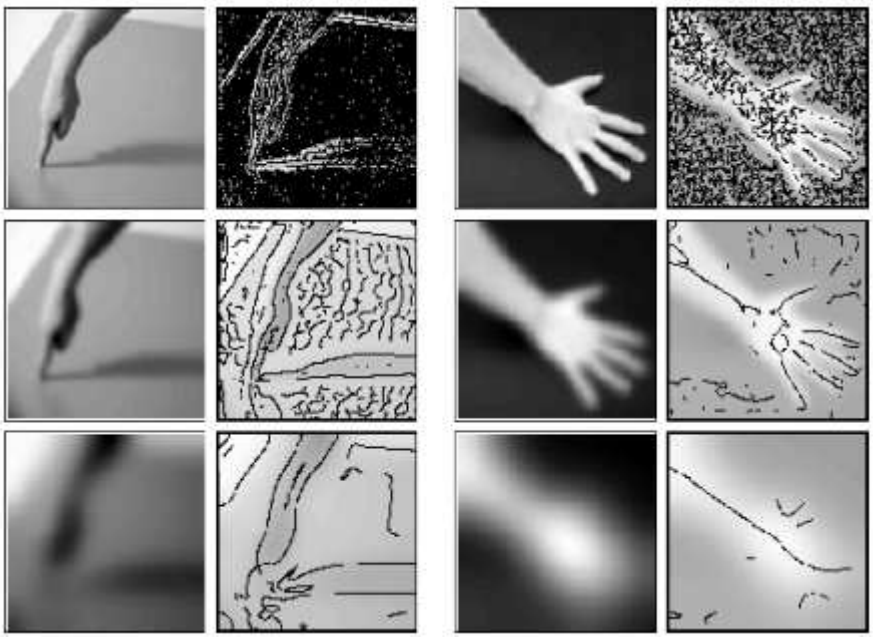
Entretanto, o reconhecimento de bordas claramente não é trivial. Uma imagem digital comum tirada do mundo é discreta. Logo, não há uma noção clara do que é borda ou não, já que borda significa descontinuidade da imagem e não existe uma definição natural do que é descontinuidade em uma função discreta. Um exemplo é a Figura 2 abaixo. À medida que as imagens são borradas, fica cada vez mais difícil definir suas bordas, já que imagens mais borradas são mais "contínuas" que imagens nítidas. Por causa disso, bordas extraídas de imagens reais têm problemas como a fragmentação, quando as bordas de uma mesma curva estão desconectadas, e bordas falsas, quando são criadas bordas para elementos irrelevantes. Esses problemas podem afetar a forma como os dados extraídos são analisados, frequentemente levando a interpretações falsas sobre o conteúdo.

4.3.2 RECONHECIMENTO DE OBJETOS

O reconhecimento de objetos é a tarefa de reconhecer, numa cena, um objeto predefinido na base de conhecimento ou um aprendido. Existem diversas formas de reconhecer objetos numa cena, mas, geralmente, os métodos empregados usam *templates* para gerar o conjunto de bordas do objeto requerido e, então, compara suas bordas com as bordas da imagem (métodos baseados em aparência). Outro conjunto de métodos bastante

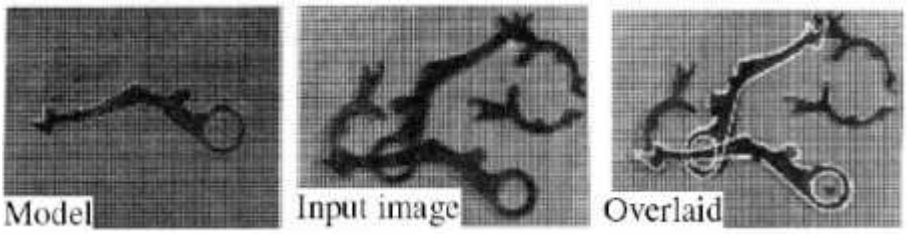
comum busca por semelhanças entre as características do modelo do objeto e as da imagem (métodos baseados em característica).

Figura 2 – Duas imagens diferentes borradas a diferentes intensidades e suas respectivas bordas.



Fonte: Rios (2010)

Figura 3 – Reconhecimento de objetos por meio de método baseado em característica.



Fonte: Rios (2010)

4.3.3 DETECÇÃO DE MOVIMENTO

Detectar que objetos ou áreas estão se movendo numa cena é importante para várias aplicações. Existem alguns métodos diferentes de se fazê-lo em relação a diversos referenciais, mas todos partem do mesmo princípio de observar quais pontos se moveram na imagem e com que velocidade. O fluxo óptico descreve a velocidade e a direção do movimento de cada pixel em relação à imagem anterior. Por medir um movimento na

imagem, não no mundo, é medido em pixels, não na unidade de velocidade do objeto – pois, não é possível saber a velocidade do mesmo, já que esse método não detecta o objeto. Com o fluxo óptico é possível inferir algumas informações úteis sobre a cena. Por exemplo, pontos mais distantes se movem mais lentamente na imagem que os mais próximos. Logo, é possível deduzir a velocidade de um objeto dependendo de sua velocidade da imagem. O conceito de fluxo óptico vem sendo bastante desenvolvido em áreas como a compactação de vídeo.

Figura 4 – Dois quadros de um vídeo mostrando uma tenista mexendo sua raquete e o fluxo óptico extraído das duas primeiras figuras em seguida.

Fonte:Rios (2010)

Uma maneira melhor para se seguir o movimento de um determinado objeto seria localizá-lo no vídeo, usando técnicas de reconhecimento de objetos a cada quadro que se passa nele. Esse método se chama vídeo *tracking*. Com o vídeo *tracking* é possível detectar movimentos de translação, bem como rotação, de um objeto em específico. Geralmente, um modelo bidimensional, ou tridimensional, se necessário, é usado para detectar o objeto nas imagens do vídeo, o que torna possível saber o deslocamento absoluto dele no ambiente. Entretanto, esse método é mais complicado para pontos que se movem muito rapidamente em relação à taxa de quadros por segundo. Além disso, um modelo tridimensional é obrigatório caso haja muitas mudanças no sentido do movimento.

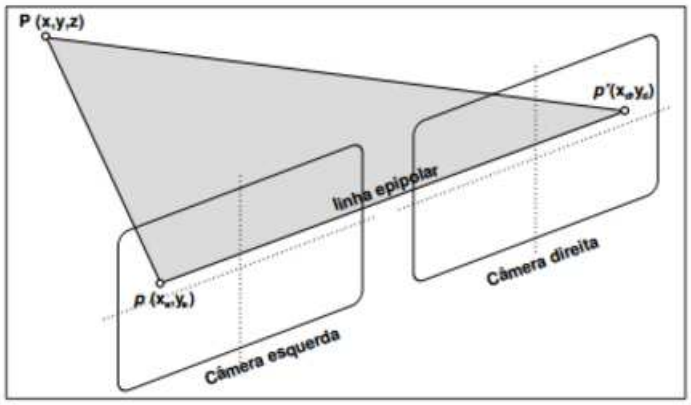
4.4 VISÃO ESTÉREO

Através de uma única imagem, apesar de toda informação que ela contém, não se consegue acessar diretamente a medida de profundidade de um ponto posicionado dentro da cena. São necessárias pelo menos duas imagens para que se possa calcular, através de triangulação, uma medida de profundidade do que está sendo visto. Isto é, provavelmente, o motivo porquê a maioria dos animais tem ao menos dois olhos e movimentam sua cabeça para direcionar sua visão para um alvo. Pelos mesmos motivos quando constrói-se robôs autônomos que necessitam tomar decisões baseando-se no que veem, são criados com um sistema de visão robótica composto por duas ou mais câmeras, assim como com algum

sistema de análise de movimento. Portanto, antes de tal sistema de visão estéreo ser construído, precisa-se entender como várias visões de uma mesma cena podem nos dar uma estrutura tridimensional do que está sendo visto, bem como as possíveis configurações de câmeras para tal reconstrução (ARAúJO, 2010).

A visão estéreo ou, estereoscopia, é o caso particular de processamento de visão computacional que utiliza por base imagens bidimensionais, adquiridas por um sistema de duas ou mais câmeras para perceber a dimensão de profundidade em uma cena (MARR, 1982). Neste caso, cada dispositivo vê a cena de dois referenciais diferentes, permitindo ao sistema computacional combinar as informações obtidas a partir do par de imagens, de modo a auferir uma representação tridimensional, como podemos observar na Figura 5. Este modelo inspira-se no funcionamento da visão humana.

Figura 5 – Típico problema de Visão Estéreo



Fonte: Araújo (2010)

A combinação de duas imagens bidimensionais, que tem por objetivo localizar os pontos correspondentes entre uma imagem e outra, não é uma tarefa trivial. Exige-se ajustes de parâmetros que são dependentes da cena, e tempo computacional extremamente elevado. Uma das principais técnicas utilizadas para encontrar tal correlação é conhecida como "técnica baseada em áreas" que utiliza a relação entre os valores de intensidade de uma janela na imagem da esquerda, e outra na imagem da direita, produzindo um mapa de disparidades denso. O tamanho da janela, bem como a área de busca na imagem, influenciam a exatidão da correspondência, e também, a complexidade do processamento (SUNYOTO; MARK; GAVRILA, 2004).

4.5 GEOMETRIA EPIPOLAR

A geometria epipolar entre duas vistas de uma mesma cena é essencialmente a geometria da intersecção dos dois planos de imagem I e I' com o feixe de planos que tem

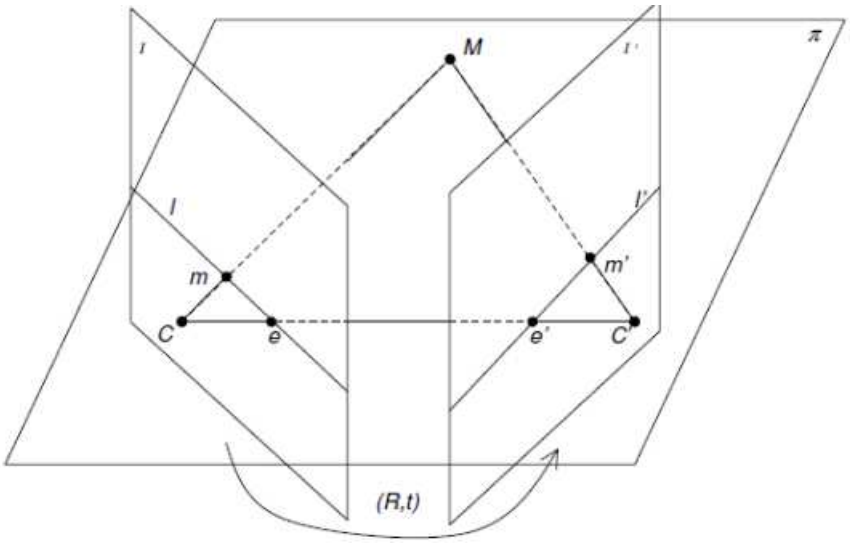
como eixo a linha de base (linha que une os dois centros ópticos) (KARLSTROEM, 2006).

Considere duas câmeras que capturam a mesma cena, cada uma com o seu centro de projeção, sendo estes não coincidentes. Cada par de imagem capturado por essas câmeras representa duas perspectivas diferentes de uma mesma cena estática. A geometria epipolar estabelece uma relação geométrica entre duas vistas capturadas nessas condições. Um ponto M do espaço tem projeções m e m' respectivamente na primeira e segunda imagens, e deseja-se saber a relação entre esses pontos correspondentes. A determinação dessa correspondência é o primeiro passo para se derivar informações do espaço tridimensional a partir de duas imagens em estéreo (ARAúJO, 2010).

Dados dois pontos correspondentes, m e m', referentes aos planos das imagens da câmera da esquerda e da direita, respectivamente, a relação entre estes pontos é dada pelo plano epipolar π, em que os centros de projeções das câmeras C e C', o ponto M no espaço 3D e os pontos m e m' nos planos da imagem são coplanares (ARAúJO, 2010).

Pode ser observado que o ponto m no plano da imagem pode ser projetado para o espaço 3D por um raio formado por m e o centro de projeção C. Este raio é visualizado como uma linha i' no plano da segunda vista. Assim, o ponto M do espaço 3D que é projetado para o ponto xm na primeira vista deve estar sobre este raio e também sobre a linha i' na segunda vista (ARAúJO, 2010).

Figura 6 – Geometria epipolar.



Fonte: Karlstroem (2006)

As entidades geométricas relacionadas à geometria epipolar são listadas a seguir:

- *Linha de base:* linha que passa pelos dois centros de projeções;

- *Epipolo:* ponto de intersecção da linha de base com o plano da imagem;
- *Plano epipolar:* plano definido pelo ponto 3D M e os centros de projeção C e C'. Note que, para cada ponto M, um plano epipolar é definido e que todas linhas epipolares interceptam o epipolo, como mostra a Figura 6. Além disso, um plano epipolar intersecta os planos das imagens da esquerda e da direita nas linhas epipolares e define a correspondência entre elas;
- *Linha epipolar:* é a linha determinada pela intersecção do plano da imagem com o plano epipolar.

4.6 RECONSTRUÇÃO ESTÉREO

Dado um equipamento de visão estéreo calibrado e dois pontos de imagem p e p', é do princípio direto para reconstruir a cena correspondente calcular a interseção dos dois raios $R = Op$ e $R' = O'p'$. No entanto, os raios R e R', na prática, nunca irão se cruzar devido aos erros naturais de calibração e localização, como notamos na Figura 7 (ARAúJO, 2010).

Neste contexto, várias abordagens razoáveis para o problema da reconstrução podem ser adotadas. Por exemplo, podemos optar por construir o segmento de reta perpendicular a R e R', que interceptam ambos os raios: o ponto médio deste segmento P é o ponto mais próximo de ambos os raios e pode ser tomado como a pré-imagem de p e p'.

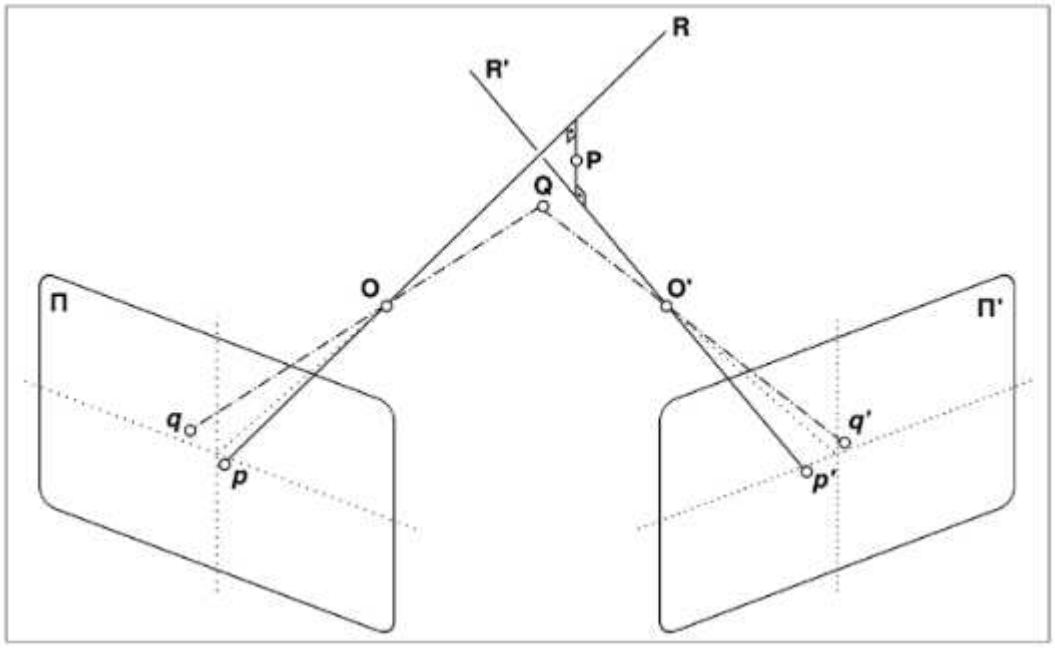
Alternativamente, pode-se reconstruir um ponto da cena usando uma abordagem puramente algébrica: dadas as matrizes de projeção M e M' e os pontos coincidentes p e p', podemos reescrever as restrições $zp = MP$ e $z'p' = MP$ como definido na Equação 4.1 .

$$\begin{cases} p \times MP = 0 \\ p' \times M'P = 0 \end{cases} \Rightarrow \begin{pmatrix} (p_x)M \\ (p'_x)M' \end{pmatrix} P = 0 \tag{4.1}$$

Este é um sistema restrito com quatro equações lineares independentes nas coordenadas homogêneas de P, que é facilmente resolvido usando as técnicas pelo método dos mínimos quadrados. Diferentemente da abordagem anterior, este método de reconstrução não tem uma interpretação geométrica óbvia, mas, generaliza facilmente para o caso de três ou mais câmeras, cada nova imagem simplesmente adicionando duas restrições adicionais.

Finalmente, nós podemos reconstruir a cena do ponto associado a p e p' como o ponto Q com imagens q e q' que minimizam $d^2(p, q) + d^2(p', q')$. Ao contrário dos outros dois métodos citados anteriormente, esta abordagem não permite o cálculo de forma fechada do ponto de reconstrução, que deve ser estimado através de técnicas dos mínimos quadrados não-lineares. A reconstrução obtida por qualquer um dos outros dois métodos

Figura 7 – Triangulação na presença de erros de medição.



Fonte: Araújo (2010)

pode ser usada como um palpite inicial para o processo de otimização. Esta abordagem não-linear também se generaliza rapidamente para o caso de várias imagens.

4.7 CÂMERAS DE PROFUNDIDADE

O dispositivo de captura que produz a imagem ou mapa de profundidade costuma ser chamado de câmera de profundidade ou câmera 3D. O mapa de profundidade também pode ser chamado de mapa de disparidade (ARAúJO, 2015).

4.7.1 CÂMERA ESTÉREO

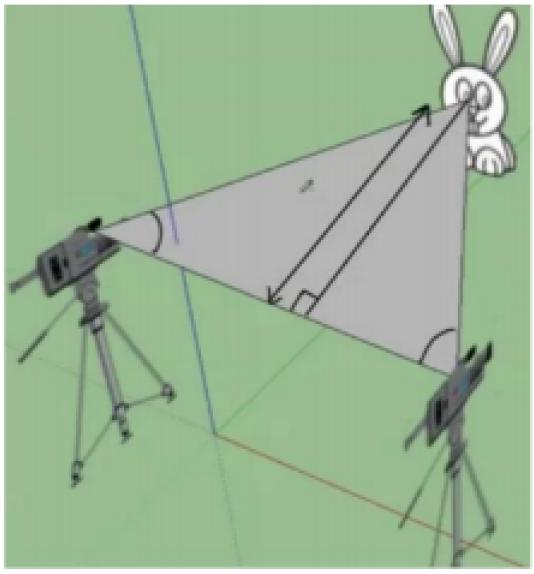
Com duas câmeras alinhadas horizontalmente, ou com uma câmera estéreo (Figura 8), é possível estimar a profundidade de um ponto a partir do triângulo (ilustrado na Figura 9) formado entre esse mesmo ponto e as duas lentes, simulando assim um fenômeno natural chamado estereoscopia, que ocorre com o ser humano quando uma cena qualquer é observada com os dois olhos.

Para comparar as imagens, precisa-se de um processo complexo de calibração, no qual é necessária uma amostra de pontos visíveis e corretamente correspondidos no espaço de interseção dos diferentes campos de visão, definidos pela posição e orientação das câmeras (ARAúJO, 2015).

Figura 8 – Exemplo de Câmera Estéreo (Bumblebee)

Fonte: Araújo (2015)

Figura 9 – Profundidade calculada por estereoscopia



Fonte: Araújo (2015)

4.7.2 CÂMERA TOF

As câmeras ToF (*Time of Flight*), também conhecidas por câmeras de distância ou câmeras de *range*, são capazes de medir a distância entre o sensor e vários pontos (pixels da imagem) da superfície dos objetos de uma cena, em um único instante. Junto com a distância, a amplitude e a intensidade do sinal refletido pela superfície também são

medidos, podendo a cena ser estática ou dinâmica (CENTENO, 2015).

Figura 10 – Exemplo de Câmera ToF Panasonic

Fonte: Araújo (2015)

A profundidade de um ponto na cena é calculada (Figura 10) a partir do tempo que a luz emitida pela câmera viaja até o destino e retorna ao sensor receptor. Essas câmeras geram imagens com resolução muito baixa, mas a uma alta taxa de quadros por segundos.

4.7.3 CÂMERA BASEADA EM LUZ ESTRUTURADA

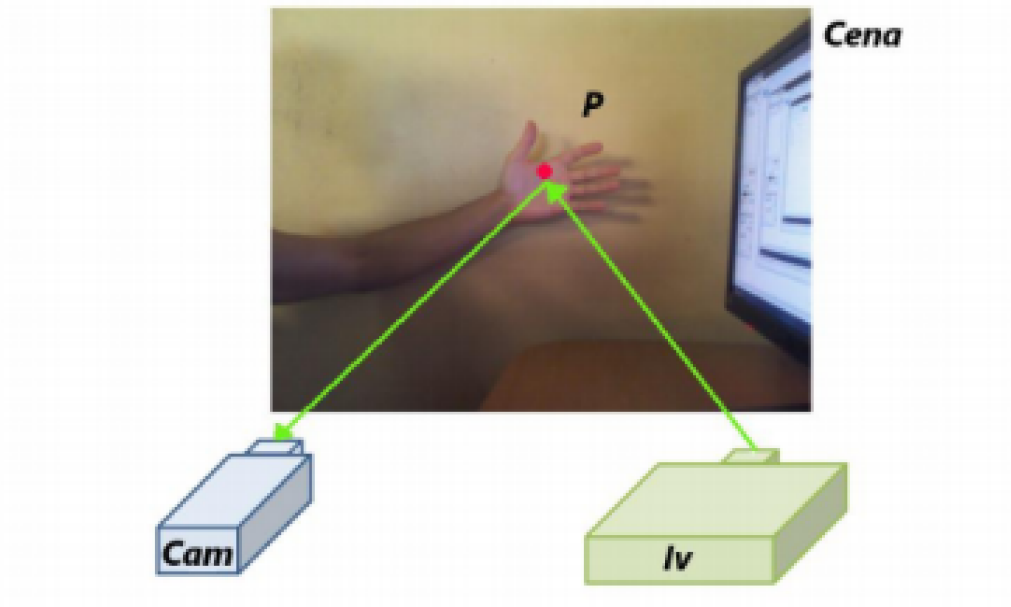
Esse tipo de câmera possui um projetor que ilumina a cena com um padrão estruturado de luz infravermelha. A câmera, alinhada horizontalmente em relação ao projetor (Figura 11), captura esses pontos que não são visíveis ao olho humano. Quando um objeto entra na cena, ele distorce o padrão desses pontos e essa variação é reconhecida como mais próximo ou distante (Figura 12). Essas câmeras possuem hardware e software específicos para efetuarem a tradução e correspondência da distância dos pontos do padrão que emitem em tempo real (ARAúJO, 2015).

Por utilizarem luz infravermelha, são considerados dispositivos para serem utilizados dentro de ambientes fechados. A luz do sol, por exemplo, poderia 'queimar' os pontos, deixando um conjunto deles muito claro, comprometendo assim a leitura da distância da área 'queimada'. Outro problema é que materiais transparentes, reflexivos e absorventes não refletem bem a luz infravermelha. O custo desse tipo de câmera é muito menor do que os tipos descritos anteriormente (ARAúJO, 2015).

4.8 MAPA DE DISPARIDADE

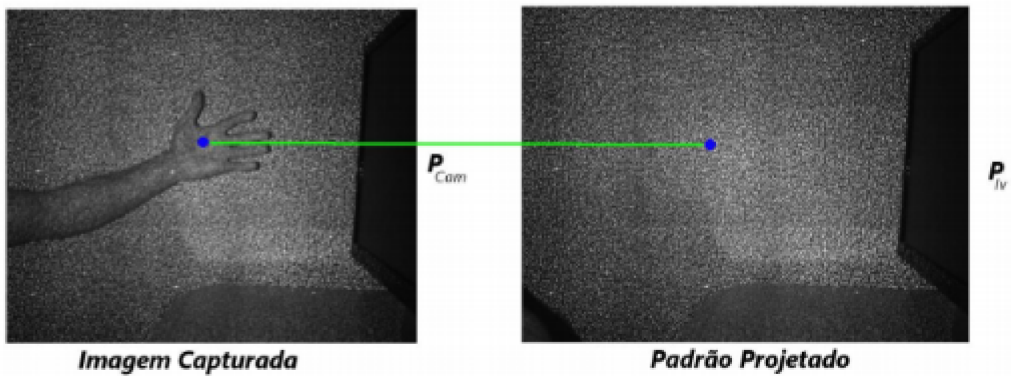
Para se recuperar a geometria 3D de um objeto, faz-se necessário mais informações do que aquilo que é projetado no plano de uma imagem 2D, como já foi dito. É necessário estimar a distância dos elementos da cena para o sistema de referência do sensor de captura que gerou a imagem.

Figura 11 – Fluxo de uma Câmera baseada em luz estruturada infravermelha



Fonte: Araújo (2015)

Figura 12 – Imagem de Luz Estruturada



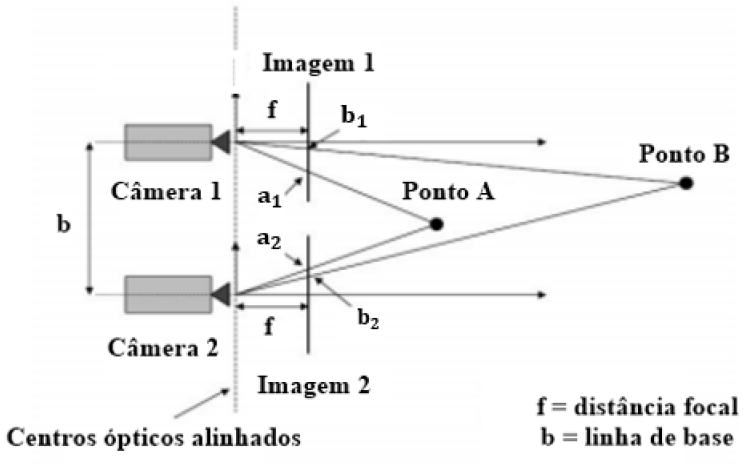
Fonte: Araújo (2015)

Mapa de disparidade, também conhecido como imagem de profundidade, mapa de profundidade ou imagem de disparidade, é uma imagem digital onde seu elemento mínimo, ao invés de conter um valor de intensidade (cor), carrega uma informação de distância em relação a um plano de coordenadas (usualmente o sistema de coordenadas discretas do dispositivo sensor) (JúNIOR, 2014).

Considere duas imagens de uma mesma cena produzida por um sistema de visão estereoscópica. Um mapa de disparidade pode ser visto como uma matriz M de inteiros de

tamanho W x H, em que W e H são, respectivamente, a largura e a altura da imagem.

Figura 13 – Sistema multicâmera



Fonte: Adaptado de Musatti (2013)

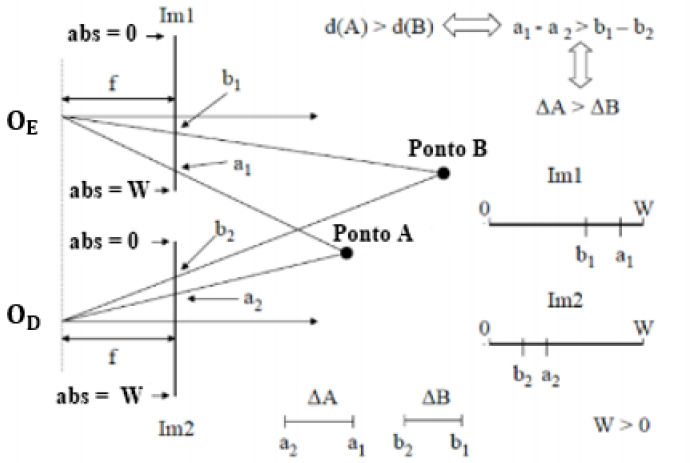
Cada número inteiro d presente nas células da matriz representa a distância, em pixels, entre o pixel de referência P, obtido a partir da primeira imagem, e os pixels correspondentes P', que pertencem à segunda imagem, no momento em que as duas imagens são sobrepostas. Quanto maior o valor da disparidade d, mais o ponto tridimensional, representado exclusivamente pelo par de pixels P e P', situa-se perto das duas câmeras. Consequentemente, a disparidade $d(A)$ calculada para o ponto A deve ser maior que a disparidade $d(B)$ do ponto B, já que A está definitivamente mais perto das duas câmeras que B. Os comentários acima, contudo, são verdadeiros apenas na condição em que os focos (centro óptico) das duas câmeras estejam alinhados e ligeiramente espaçados como se mostra na Figura 13. a_1 e a_2 são os pontos das projeções de A, respectivamente, sobre a primeira e a segunda imagem; da mesma forma b_1 e b_2 para o ponto B. Supõe-se que todos os pontos a_1, a_2, b_1 e b_2 têm o mesmo valor de ordenada, então estes pontos estão todos na mesma linha horizontal.

Calcularemos, portanto, os valores de desigualdade para os pontos A e B da seguinte forma:

- Disparidade para o ponto $A : d(A) = Xa_1 - Xa_2$
- Disparidade para o ponto $B : d(B) = Xb_1 - Xb_2$

Nota-se que $d(A) > d(B)$ sendo $Xa_1 > Xb_1$ e $Xa_2 < Xb_2$. É importante ressaltar que se deve considerar apenas os valores de abcissas dos quatro pontos, devido a hipótese de alinhamento vertical, tal como ilustrado na Figura 14:

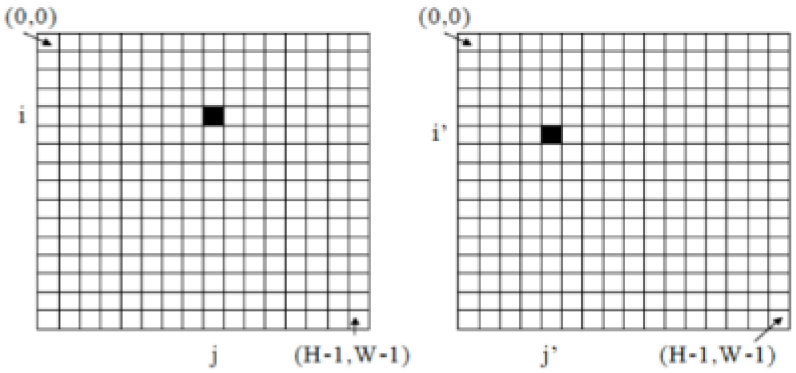
Figura 14 – Cálculo da disparidade



Fonte: Adaptado de Musatti (2013)

Se considerarmos as duas imagens como matrizes de elementos de imagem, os pontos a_1 e a_2 podem ser representados de acordo com os respectivos valores de linha e coluna. O pixel a_1 terá coordenadas de linha-coluna (i, j) e a_2 vai estar na posição (i', j').

Figura 15 – Matrizes de elementos de imagem



Fonte: Musatti (2013)

A disparidade entre a_1 e a_2 será a distância entre as localizações (i, j) e (i', j') quantificados por um número inteiro que indica a diferença de pixels entre as duas posições do par conjugado. Claramente, a disparidade pode ser calculada apenas para os pontos de cena que são visíveis em ambas as imagens; um ponto visível de uma imagem, mas não visível na outra deve ser ocluso. Uma vez que a matriz M tenha sido completamente preenchida, ou seja, uma vez que para cada pixel na imagem de referência foram encontrados

os correspondentes na segunda imagem e for calculada a disparidade entre os dois, teremos obtido o mapa de disparidade da cena que estamos analisando. Associando a informação contida na forma de M inteiros, variando de zero a D_{max}, onde D_{max} é a disparidade máxima admissível, a uma escala qualquer de cores, pode-se obter uma terceira imagem que representa os objetos na cena com cores diferentes de acordo com a distância entre as câmeras. Do que foi exposto acima, pode-se inferir que um mapa de disparidades pode ser utilizado como uma ferramenta para a avaliação ambiente tridimensional observado.

4.9 CALIBRAÇÃO DE CÂMERA

A calibração de câmera descreve uma correspondência entre as coordenadas do espaço objeto 3D (mundo real) e pontos da imagem (2D) (KONDRAT, 2011).

Essa correspondência se dá por uma função $R^3 \longrightarrow R^2$.

Essa função pode ser escrita como a multiplicação das propriedades intrínsecas e extrínsecas:

$$\begin{bmatrix} u \\ v \\ Z \end{bmatrix} = \begin{bmatrix} fx & 0 & cx \\ 0 & fy & cy \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{21} & r_{31} & t_1 \\ r_{21} & r_{22} & r_{32} & t_2 \\ r_{31} & r_{23} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{4.2}$$

Para fazer essa associação, devemos conhecer n pontos no mundo real e saber as coordenadas destes pontos, na imagem. Para conseguir fazer isso, usamos uma imagem com um padrão de pontos conhecidos. Quando realizada com sucesso, retorna três matrizes. A intrínseca, rotação e translação, que juntas formam a função de correspondência (Função 4.2)

4.9.1 PROPRIEDADES DE CAPTURA DE IMAGEM

Quando uma câmera captura uma foto, precisamos ter em mente que a imagem resultante é a projeção da imagem real, no sensor de captura dentro da câmera (KONDRAT, 2011).

As informações sobre como a imagem é projetada na câmera podem ser divididas em duas propriedades: intrínsecas e extrínsecas, que são obtidas através do processo de calibração da câmera.

4.9.1.1 PROPRIEDADES INTRÍNSECAS

São as propriedades sobre a geometria interna da câmera, uma vez medidos, não se alteram. São elas, distância focal e centro óptico.

Essas informações serão armazenadas numa matriz com o seguinte formato:

$$\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4.3}$$

Onde f_x, f_y são as distancias focais nos respectivos eixos e c_x, c_y definem o centro óptico da projeção (KONDRAT, 2011).

4.9.1.2 PROPRIEDADES EXTRÍNSECAS

São as propriedades da perspectiva que a imagem foi capturada, em relação a um referencial. Tem como finalidade estimar as posições e rotações da câmera em relação ao sistema de coordenadas do espaço objeto. Pode ser determinada por um modelo que relaciona os pontos na imagem com pontos no espaço objeto. Em geral, usa-se um padrão conhecido na imagem como referência, para fazer a relação entre os pontos (KONDRAT, 2011).

Essas propriedades variam conforme a mudança de posição da câmera em relação à imagem observada.

São usadas duas matrizes para armazenar respectivamente, as propriedades de rotação e translação da câmera em relação a uma referência. Como em geometria analítica, as matrizes representadas da seguinte forma (KONDRAT, 2011):

$$R = \begin{bmatrix} r_{11} & r_{21} & r_{31} & t_1 \\ r_{21} & r_{22} & r_{32} & t_2 \\ r_{31} & r_{23} & r_{33} & t_3 \end{bmatrix} \tag{4.4}$$

e

$$T = \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{4.5}$$

4.10 CONVERSÃO DE COORDENADAS DE UM PONTO 3D NO ESPAÇO REAL

As câmeras do sensor Kinect fazem uma transformação dos pontos reais no espaço 3D projetando-os no plano óptico formando uma imagem do mundo real no CCD (*Charge-*

Coupled Device - Dispositivo de carga acoplada). Esta transformação é dada pela fórmula 4.6:

$$\begin{bmatrix} u \\ v \\ Z \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{4.6}$$

Cada ponto com coordenadas reais no espaço X, Y, Z é projetada no ponto com coordenadas u, v da imagem 2D. A referência dos pontos no espaço tem origem no centro óptico da câmera e a distância Z é fornecida diretamente da câmera de profundidade como a distância entre o ponto e o plano óptico da câmera de infravermelhos. A matriz de transformação é chamada de matriz intrínseca e os valores f_x e f_y são, respectivamente, as distâncias focais do sistema de lentes usadas na câmera, enquanto c_x e c_y são as coordenadas em pixel do centro da imagem do sistema óptico em relação à origem colocado no canto superior esquerdo.

Desta transformação é possível obter a inversa que possibilita transformar os pontos da imagem com coordenadas u, v em pontos reais no espaço com coordenadas X, Y, Z. As unidades de medida em que são expressas as coordenadas reais dos pontos dependem dos valores utilizados na matriz intrínseca, no que diz respeito à forma como é realizada a calibração da câmera para determinar a matriz intrínseca.

Logo, a transformação inversa é conseguida utilizando a fórmula 4.7 :

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}^{-1} * \begin{bmatrix} u \\ v \\ Z \end{bmatrix} \tag{4.7}$$

O sistema óptico do sensor Kinect é equipado com duas câmeras diferentes e deslocadas em posições diferentes do sensor. A partir disto, percebe-se que as imagens fornecidas pelas duas câmaras não coincidem. A fim de transformar as coordenadas u, v dos pontos de uma câmera em coordenadas s, t dos pontos da outra câmera é necessário fazer a calibração extrínseca que permite obter a matriz de rotação e a de translação.

Para converter as coordenadas de um ponto u, v em s, t utiliza-se a fórmula 4.8:

$$\begin{bmatrix} s \\ t \\ W \end{bmatrix} = R * \begin{bmatrix} u \\ v \\ W \end{bmatrix} + T \tag{4.8}$$

Onde R é a matriz de rotação e T é o vetor de translação em relação às duas câmeras em questão.

5 KINECT

Inicialmente chamado de “Projeto Natal” (MUSATTI, 2013), o Kinect (Figura 16) foi desenvolvido pela companhia Israelense PrimeSense, que foca em sensores 3D para “interação natural” (WERBER, 2011). Posteriormente licenciado para a Microsoft, recebeu o nome de Kinect no lançamento oficial em 4 de novembro de 2010 na E3 (Electronic Entertainment Expo 2010) (WERBER, 2011).

Figura 16 – Sensor Microsoft Kinect 1.0

Fonte: Adaptado de Microsoft (2010)

O dispositivo tem uma variedade de sensores: uma câmera de vídeo (RGB) um sensor de profundidade IR, quatro microfones e um acelerômetro de 3 eixos. Contendo, também, uma variedade de chips de processadores e controladores, o mais relevante destes é o processador de imagem PrimeSense que pré-processa as duas saídas da câmera antes da transmissão através da interface USB.

5.1 ESTRUTURA INTERNA DO KINECT

O sensor Kinect é conectado por uma USB 2.0 com fonte de alimentação adicional e possui uma câmera RGB padrão, um sensor de profundidade e uma matriz de microfones de quatro canais (como anteriormente mencionado). Sua “cabeça” pode ser inclinada $\pm 27°$, um acelerômetro de três eixos mede a orientação e um LED de três cores pode ser usado para *feedback* visual (MUSATTI, 2013).

O dispositivo desmembrado é mostrado na Figura 17:

Tem-se os seguintes componentes:

Figura 17 – Kinect desmembrado

Fonte: Musatti (2013)

- Uma matriz de microfones (3 ao lado direito e um o lado esquerdo);
- Três aparelhos ópticos utilizados para o reconhecimento visual do corpo em movimento;
- Um ventilador para dissipação de calor; 64MB de memória flash DDR2;
- Um acelerômetro Kionix KXSD9 de três eixos;
- PrimeSense PS1080-A2 é o chip que representa o "coração" da tecnologia do Kinect.

Mais detalhadamente, o conjunto de aparelhos ópticos do Kinect é composto por uma câmara RGB e um sensor de profundidade de raios infravermelhos. Este sensor é composto por um projetor de infravermelhos e uma câmara sensível à mesma banda, que é usado para realiza leitura quando detecta raios infravermelhos (MUSATTI, 2013).

A câmera RGB tem uma resolução de 640 x 480 pixels, enquanto o infravermelho usa uma matriz de pixel 320 x 240. O conjunto de microfones é usado pelo sistema para a calibração do ambiente no qual está localizado, através da análise da reflexão de som nas paredes e na mobília. Assim, o ruído de fundo e os sons do jogo são eliminados, tornando possível um correto reconhecimento dos comandos vocais (MUSATTI, 2013).

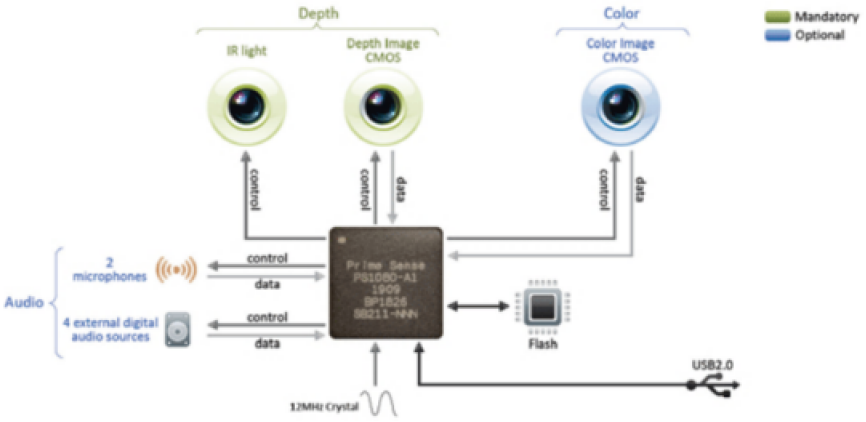
O Kinect é motorizado em torno do eixo horizontal e segue os movimentos dos jogadores, orientando-se na melhor posição para reconhecer os movimentos. Através de

estudos experimentais, foi determinado que o sensor Kinect processa o fluxo de vídeo para uma taxa de quadros (*frames*) de $30Hz$. O fluxo RGB usa uma resolução VGA de 8 bits de 640 x 480 pixels com um filtro de Bayer para as cores, enquanto a profundidade do sensor monocromático gera um fluxo de dados a uma resolução VGA (640 x 480 pixels) com profundidade de 11 bits, portanto 2048 valores possíveis. O sensor Kinect tem um alcance de utilização com um mínimo de 0,4 a um máximo de 4,0 metros. Possui um campo de visão angular de 57° horizontal e 43° vertical, enquanto que o ponto de apoio é motorizado e é capaz de um deslocamento de 27° graus para cima ou para baixo, controlável através de software. O campo horizontal a uma distância de 0,8 m tem aproximadamente 87 cm, enquanto a vertical aproximadamente 63cm, logo, resultando uma resolução de 1,3 mm por pixel. A resolução espacial $x = y$ a 2 metros de distância a partir do sensor é de 3 mm, enquanto que a resolução de profundidade z, ainda a dois metros de distância, é de 1 cm (MUSATTI, 2013).

Como já mencionado, o dispositivo permite que o usuário possa interagir com o console, sem a utilização de qualquer controlador que precise segurar, ou seja, utilizando apenas os movimentos do corpo, comandos de voz ou através de objetos presentes no ambiente. Segundo a Microsoft, o Kinect pode acompanhar os movimentos de até quatro jogadores, estejam em pé ou sentados.

A seguir um esquema de funcionamento do sistema:

Figura 18 – PrimeSense Design



Fonte: Kronlachner (2013)

A Figura 18 evidencia o chip PS1080 - A2 da PrimeSense que compreende o procedimento de análise de toda a cena, controlando adequadamente os dispositivos ópticos e de áudio a fim de colher as informações necessárias. Apenas para fins informativos, abaixo uma lista dos vários chips que compõem o Kinect:

- Wolfson Microelectronics WM8737G - Stereo ADC com microfone pré-amplificador;
- Fairchild Semiconductor FDS8984 - N – Channel PowerTrench MOSFET;
- NEC uPD720114 - USB 2.0 hub controlador;
- H1026567 XBOX1001 X851716 – 005 GEPP;
- Marvell AP102 - SoC (system-on-a-chip) com controlador de interface de câmera;
- Hynix H5PS5162FF 512 megabit DDR2 SDRAM;
- Analog Devices AD8694 ; Quad, baixo custo, baixo nível de ruído, o CMOS Output Rail-to-Rail Amplificador Operacional.
- TI ADS7830I - 8–Bit, 8 - amostragem de canal, conversor A/D com Interface I2C;
- Allegro Microsystems A3906 - Stepper de baixa tensão e Single/Dual DC Motor Driver;
- ST Microelectronics M29W800DB - 8 Mbit (1Mb x 8 or 512Kb x 16) NV memória Flash;
- PrimeSense PS1080 – A2 - Processador de sensor de imagem SoC
- TI TAS1020B USB Controlador de áudio frontal e central.

Finalmente para completar a lista de dispositivos de hardware um acelerômetro Kionix MEMS KXSD9, usado para controlar a inclinação e estabilizar a imagem.

5.2 CÂMERA RGB

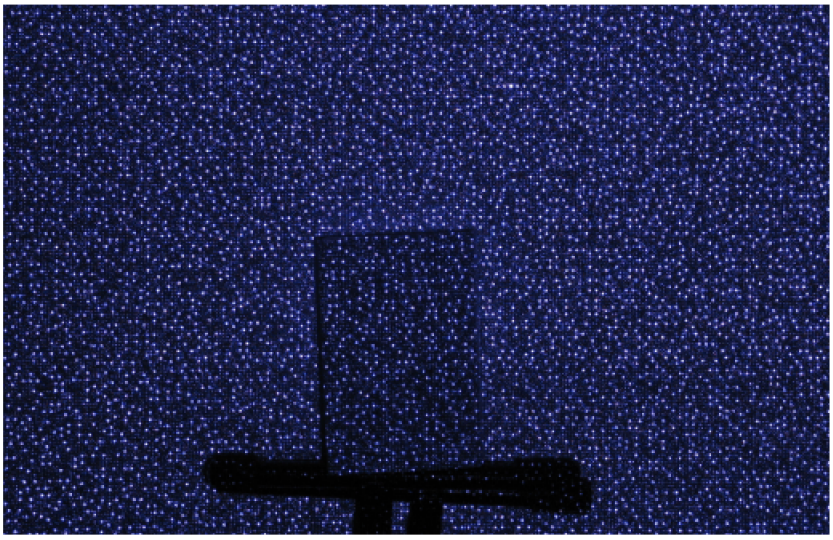
A câmera RGB do Kinect possui uma resolução padrão de 640 x 480 pixels operando em uma taxa de *frames* de 30Hz. Possui, também, um modo de alta resolução com 1280 x 1024 pixels. Mas a taxa de *frames* cai para cerca de 15Hz quando se utiliza o modo de alta resolução. A saída derivada da câmera RGB é codificada como imagem padrão Bayer, mas os *frameworks* disponíveis podem converter as informações brutas em uma imagem RGB padrão.

O *pipeline* do fluxo RGB também pode ser usado para produzir o fluxo bruto da câmera infravermelha (IR). Infelizmente, o fluxo RGB e o fluxo IR não podem ser usados em paralelo (KRONLACHNER, 2013).

5.3 SENSOR DE PROFUNDIDADE

O sensor de profundidade consiste em um *laser* infravermelho de 830 nm de comprimento de onda projetando um padrão específico de ponto no seu campo de visão 19. Uma câmera de infravermelho registra esses padrões nos objetos e um Processador de Sinal Digital (*Digital Signal Processor - DSP*) embarcado calcula a distância correlacionando a imagem ao vivo com os padrões de referência armazenados (KRONLACHNER, 2013).

Figura 19 – Padrão de pontos infravermelho em uma parede.



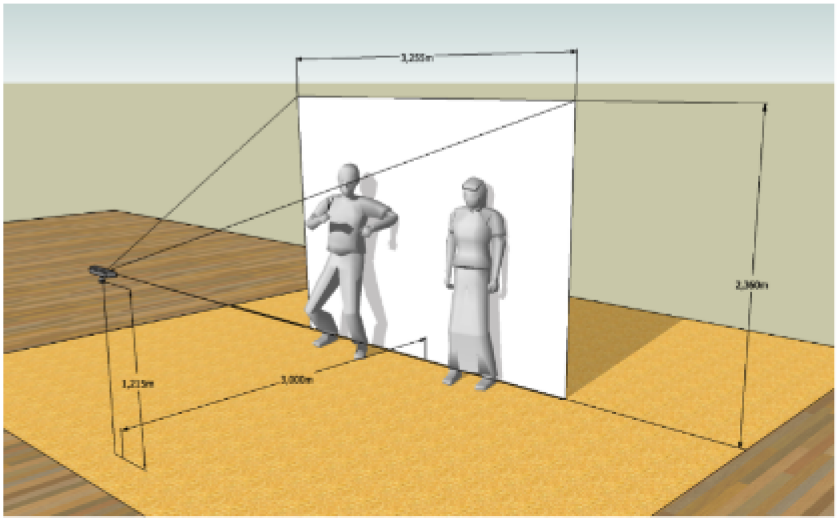
Fonte: Kronlachner (2013)

A saída do sensor de profundidade é um fluxo de vídeo de 640 x 480 pixels. Cada pixel possui 11 bits de informação de profundidade, portanto 2048 (2^{11}) valores diferentes para representar a distância são possíveis. Devido a razões computacionais internas, as oito colunas de pixels mais à direita na imagem não contêm dados. Portanto, a resolução utilizável é reduzida para 632 x 480. As bibliotecas para acessar o fluxo de profundidade do Kinect suportam a conversão de 11 bits de valores brutos de profundidade para coordenadas do mundo real em milímetros. Devido ao deslocamento horizontal do projetor e da câmera de infravermelho, uma sombra especialmente para objetos próximos à câmera é visível, pois tais dispositivos não compartilham exatamente o mesmo campo de visão.

5.3.1 ALCANCE DOS SENSORES DE PROFUNDIDADE

O alcance do sensor de profundidade do Kinect cobre aproximadamente 0,7 a 7 metros. O alcance ótimo é dado pelo fabricante como sendo de 1,2 a 3,5 metros. O campo de visão cobre 58° horizontal, 45° vertical e 70° diagonal.

Figura 20 – Alcance do Kinect



Fonte: Kronlachner (2013)

5.3.2 LATÊNCIA DO SENSOR DE PROFUNDIDADE

Como Kronlachner (2013) apontou em seu trabalho, a latência média do sensor de profundidade é 72,98 ms. Essa latência indica cerca de 2 quadros de atraso em um *framerate* de 30 imagens por segundo $\left(t = \frac{2}{30[Hz]} = 66.6[ms]\right)$. Este valor não inclui o cálculo de funções de alto nível fora da profundidade do vídeo.

5.4 MATRIZ DE MICROFONES

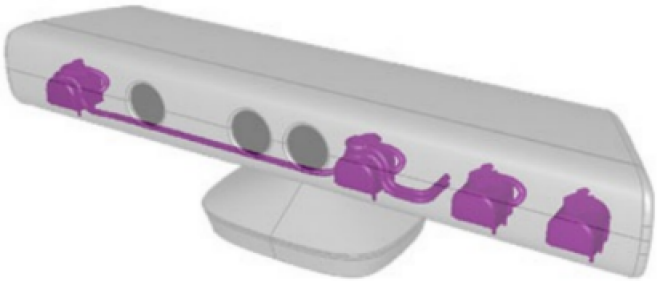
O sensor Kinect inclui uma matriz de microfone de 4 canais com ADCs (*Analog to Digital Converter*) a uma taxa de amostragem de 16 kHz e a resolução de 24 bits. Devido à sua taxa de amostragem bastante baixa e, portanto, frequência de corte de cerca de 8kHz, destina-se principalmente à gravação de fala. A localização básica da fonte de som e os algoritmos de formação de feixe podem ser usados para separar vários alto-falantes diante do sensor Kinect. No entanto, esses algoritmos não são fornecidos pelo próprio dispositivo.

Os quatro microfones de eletreto onidirecionais estão localizados atrás de uma grelha na parte inferior da barra do sensor (Figura 21). Seguindo o projeto de referência Primesense (Figura 18), o Kinect suporta receber um fluxo de áudio de quatro canais para executar o cancelamento de eco (KRONLACHNER, 2013).

5.5 FUNCIONAMENTO

O princípio de detecção luz gama estruturada é relativamente antiga (ZHANG, 2012), contudo, o lançamento do Microsoft Kinect foi, provavelmente, um dos grandes

Figura 21 – Matriz de microfones de 4 canais.



Fonte: Kronlachner (2013)

saltos em termos de hardware neste campo de pesquisa. Na verdade, o Kinect representa um bom substituto para câmeras-estéreo e até mesmo para sistemas multicâmera que fornecem um mapa de profundidade da cena.

A empresa PrimeSense superou o desafio de combinar um *patch* de superfície observado a uma parte do padrão projetado, produzindo padrão composto por 211 x 165 *speckles* (pontos) infravermelhos. Este padrão é repetido em uma grade 3 x 3 para formar o padrão de projeção geral do sensor. Uma câmera IR de alta resolução captura o padrão projetado que é processado para estimar um mapa de profundidade. Este mapa de profundidade é combinado com um sensor RGB para criar a imagem RGB-D (RGB + Depth - profundidade) resultante (MARTíN; LORBACH; BROCK, 2014).

O sistema PrimeSense PS1080 (*System on Chip - SoC*) fornece uma imagem de profundidade sincronizada, uma imagem colorida e fluxos de áudio. Todos os algoritmos de aquisição de profundidade são executados no PS1080 SoC, portanto nenhuma carga computacional é adicionada ao *host*. Funções de nível mais alto como análise da cena e rastreamento têm de ser feitas no *host* (KRONLACHNER, 2013).

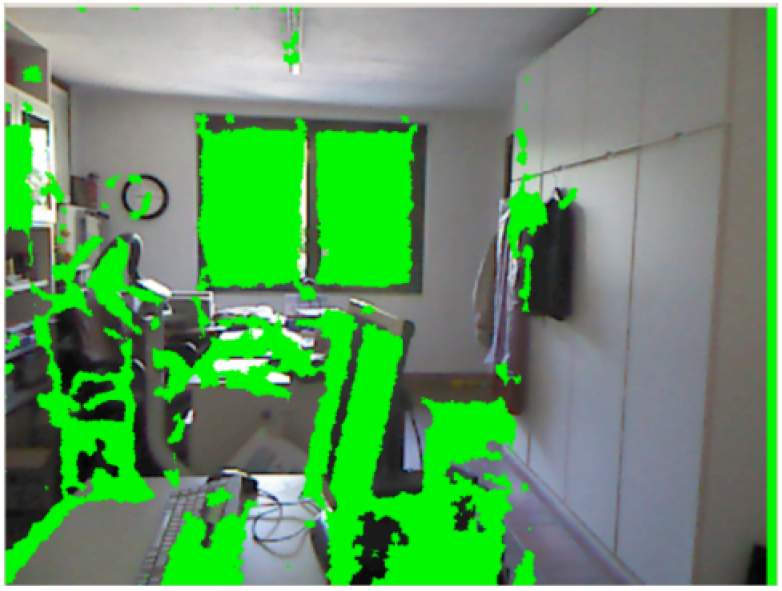
5.6 PROBLEMAS

5.6.1 DESALINHAMENTO ENTRE A IMAGEM RGB E A IMAGEM DE PROFUNDIDADE

A imagem de profundidade e RGB estão desalinhadas devido a diversos fatores, como a distância entre a câmera RGB e a câmera IR, pela rotação dos eixos ópticos e pelas distorções introduzidas pelas lentes (MUSATTI, 2013); estes fatores determinam para cada câmera um campo de vista diferente. O tamanho da imagem de profundidade é diminuída devido ao campo de vista inferior ao da câmera RGB.

Na Figura 22 pode-se observar o desalinhamento por meio da sobreposição da

Figura 22 – Desalinhamento entre a imagem RGB e a imagem de profundidade.



Fonte: Musatti (2013)

imagem de profundidade com a RGB, as áreas de imagem em verde representam todas as áreas de sombra na qual a distância entre o sensor e os pontos é zero, desta forma, evidencia como não há correspondência entre as áreas sombreadas e as bordas dos objetos.

5.6.2 RUÍDO NA IMAGEM DE PROFUNDIDADE

A profundidade da câmara de saída é afetada por um tipo particular de ruído que oscila o valor da distância dos pontos entre um valor válido e zero no tempo. Este tipo de fenômeno traduz-se em uma oscilação contínua do valor da distância dos pontos, principalmente nas bordas dos objetos visualizados, impedindo a detecção correta dos contornos.

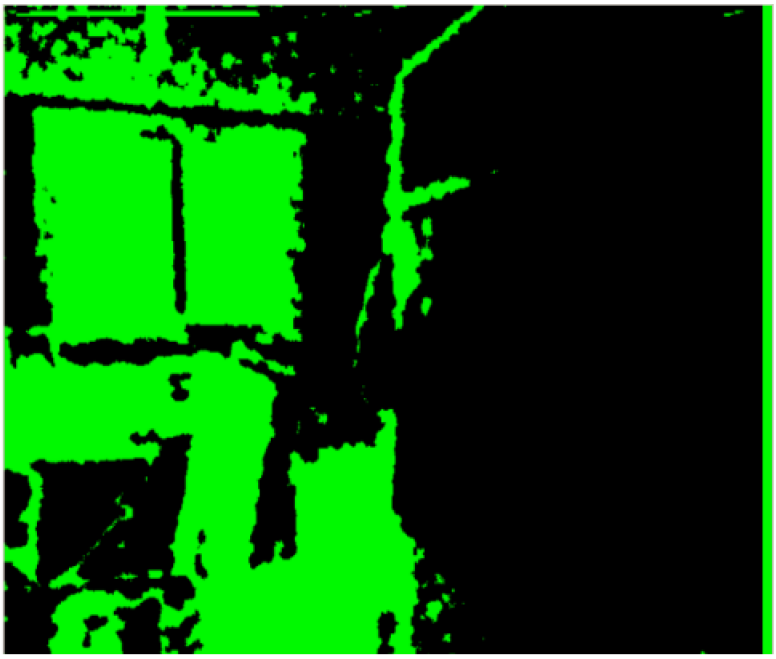
Na Figura 23 pode-se observar o efeito do ruído introduzido a partir da profundidade da câmara, nesta imagem a cor verde também representa os pontos com distância nula, de um modo particular as bordas dos objetos.

5.6.3 SOMBRAS NA IMAGEM DE PROFUNDIDADE

Devido a distância entre a câmera de infravermelho e a fonte de luz, formam-se sombras na imagem de profundidade. O sensor não é capaz de estimar a distância até essas áreas sombreadas e, portanto, o seu valor de profundidade está definido como zero, a Figura 24 ilustra o princípio no qual esse fenômeno ocorre:

A Figura 25 compara a imagem de profundidade e a imagem RGB, onde são clara-

Figura 23 – Ruido na imagem de profundidade.



Fonte: Musatti (2013)

mente visíveis no lado esquerdo dos objetos sombras produzidas pelo processo mencionado anteriormente.

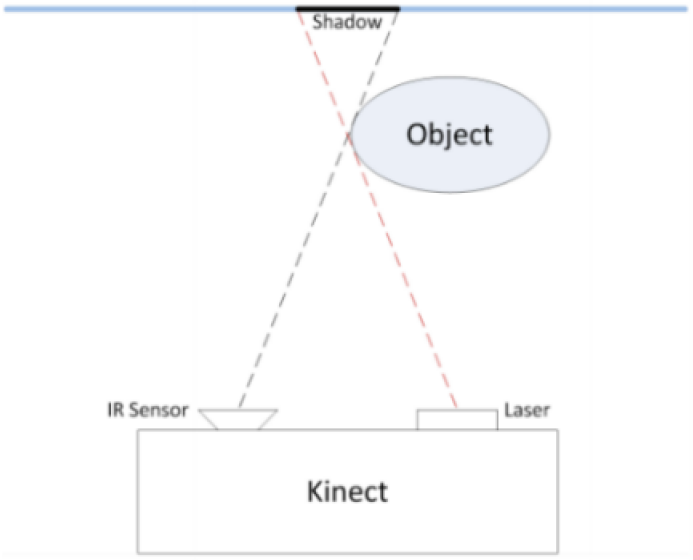
À esquerda tem-se uma imagem de profundidade utilizando uma escala de cores para representar os valores da distância, a cor preta representa o valor nulo.

5.7 USANDO VÁRIOS DISPOSITIVOS KINECT

O uso de vários dispositivos Kinect na mesma cena pode causar regiões não reconhecidas na imagem de profundidade devido a sobreposição de padrões de ponto infravermelho. Uma possível solução adicionando movimento independente a cada um dos sensores é proposta em Maimone e Fuchs (2012). A aplicação de vibração aos sensores Kinect resulta em um padrão de pontos embaçados dos sensores Kinect interferentes. Isso pode ser feito acoplando um motor com uma massa excêntrica na parte inferior do Kinect (Figura 26). A rotação do motor e sua massa acoplada induz uma pequena vibração no dispositivo. Como o laser e a câmera de infravermelho são unidos à mesma carcaça, veem seu próprio padrão claramente e sem distorção. O valor da profundidade pode ser estimado e o reconhecimento não é perturbado por outros sensores Kinect (Figura 27).

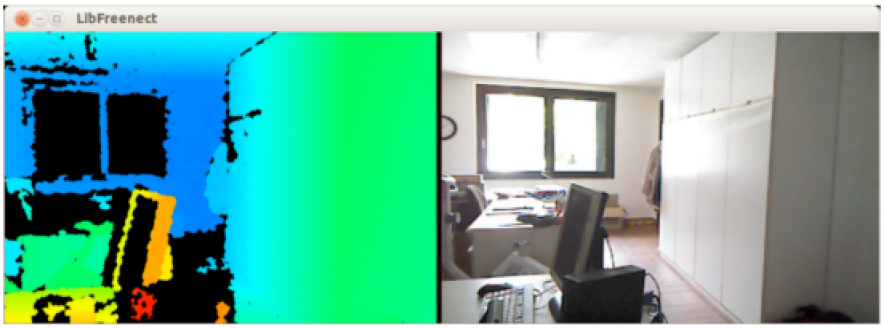
Além dos padrões de sobreposição, a largura de banda USB 2.0 deve ser considerada ao usar vários dispositivos Kinect com um computador. Testes mostraram que dois sensores Kinect com RGB ativado e fluxo de profundidade ocupam um barramento USB. Portanto,

Figura 24 – Fenômeno da sombra.



Fonte: Musatti (2013)

Figura 25 – Comparação entre a imagem de profundidade e a imagem RGB.



Fonte: Musatti (2013)

os controladores USB devem ser adicionados ao sistema para operar com mais de dois sensores Kinect.

Figura 26 – Motor DC ligado ao Kinect para indução de vibração

Fonte: Kronlachner (2013)

Figura 27 – Motor DC ligado ao Kinect para indução de vibração

Fonte: Kronlachner (2013)

6 DISCUSSÃO

O sensor Kinect é um dispositivo com inúmeros funcionalidades, dentre as quais tem-se a captura de movimento, captura de voz, medição de profundidade, etc. Por possuir tais funcionalidades, tem sido utilizado em estudos de inúmeras áreas e para diversos fins. O Kinect, como instrumento de medição de profundidade, sofre as mesmas limitações que outros dispositivos similares: medições são feitas apenas com objetos voltados para o dispositivo. Assim, para situações onde pretende-se reconstruir uma cena 3D completa, é necessário construir sistemas com múltiplos sensores para que se possa cobrir toda a cena e capturar o máximo possível de informações.

Uma área que tem ganhado com o uso do Kinect é a Reabilitação, na qual têm sido desenvolvidos sistemas de reabilitação computadorizados utilizando câmeras. Fernández-Caballero (2015) demonstra que um único Kinect é capaz de fazer "boas capturas de movimentos", porém quando se adiciona outro Kinect, perde-se qualidade na imagem de profundidade e aparecem ruídos causados pela sobreposição de raios transmitidos pelos sensores IR de ambos os Kinects. No entanto, alguns estudos mostram que os problemas de interferência podem ser resolvidos utilizando soluções de softwares ou de hardware.

Essmaeel (2012) elenca algumas soluções apresentadas em experimentos que existem na literatura. Algumas soluções de hardware podem trazer inconvenientes como a redução da taxa de fotogramas quando se usa multiplexação de tempo. As soluções de software oferecem uma maneira alternativa de lidar com a interferência e superar certas limitações das soluções de hardware, já que levam em consideração características da interferência. Um exemplo de solução é apresentado por Essmaeel (2012), que utiliza um filtro mediano que pode ser usado para preenchimento de orifícios causados por interferência, pois o Kinect não retorna dados quando interferências ocorrem gerando tais orifícios. Do mesmo modo, outros filtros podem ser usados para resolução de outros problemas causados por interferências. O uso de filtros de redução de ruídos em imagens, traz bons resultados, porém ao custo de remover todos os movimentos lentos realizados longe da câmera. Além disso, até agora nenhuma abordagem de ganho variável foi proposta para tratar este problema.

A calibração de múltiplas câmeras é um problema que tem sido bastante abordado com muitos métodos, fornecendo boas e confiáveis soluções. O método mais frequentemente adotado para calibrar várias câmeras é usando um objeto de referência visto por todas as câmeras. Esse método pode ser usado para calibrar o sensor RGB em cada Kinect. Entretanto, permanece o desafio de como calibrar todos os sensores de profundidade simultaneamente com os sensores RGB. A solução mais comum adotada até agora foi usar a capacidade do *driver* Kinect para registrar a imagem de profundidade na imagem RGB

Tabela 1 – Comparação do *range* de câmeras 3D

Modelo	**Xbox Kinect**	**D-Imarger EKI 3104**	**MESA SR-3000**
Resolução	640 x 480	160 x 120	176 x 144
Velocidade	30	20, 25, 30	25
Distância	0.8 m a 4 m	1.2 a 9 m	7.5 m
Campo de Visão			
Horizontal	57	60	47.5
Vertical	43	44	39.6
Inclinação física	(-27, +27)	não possui	não possui
Preço	$109,99	>$ 5000	>$ 5000

Fonte: Adaptado Rafibakhsh (2012)

e, em seguida, executar apenas a calibração nas câmeras RGB usando qualquer um dos métodos já conhecidos. Musatti (2013) realiza experimentos utilizando este método, mas, segundo Essmaeel (2012), apesar de fácil implementação, esta solução não é adequada para aplicações que exijam alto nível de precisão. Outra solução é utilizar um objeto de referência visto por todos os dispositivos (ESSMAEEL, 2012), (FABIAN et al., 2014), (MUSATTI, 2013).

6.1 COMPARAÇÃO COM OUTROS DISPOSITIVOS DE CAPTURA DE IMAGEM 3D

Existem diversos dispositivos para captura de imagem de profundidade. Rafibakhsh (2012) compara o desempenho do sensor Kinect com câmeras ToF (D-Imager EKL3104 e MESA SR-3000) em termos de resolução, velocidade, distância e campo de visão que é mostrada na Tabela através da tabela.

A Tabela 1 demonstra que o sensor Kinect, cujo custo é apenas uma fração de outros, representa uma melhoria considerável em relação às outras duas câmeras de profundidade. Ainda no trabalho de Rafibakhsh (2012) são feitos experimentos com Kinect e o *laser scanner* de alta definição *Faro Focus3D*. Os autores chegam a conclusão que o Kinect possui resolução e precisão de profundidade consideráveis em comparação com os *scanners de laser* terrestre de alta qualidade, e a interferência entre vários sensores Kinect é evidente em situações em que esses sensores não foram cuidadosamente posicionados. Atentam-se ao fato de o sensor Kinect custar bem menos que todos os outros dispositivos mencionados no trabalho, afirmam que o Kinect pode ser usado para o desenvolvimento de sistemas de segurança para obras, tendo somente que ser devidamente posicionados e calibrados. Martín, Lorbach e Brock (2014) utiliza o Kinect e o Asus Xtion para analisar os efeitos da interferência em sensores RGB-D baseados em luz estruturada, intercalando ambos em seus experimentos e obteve resultados idênticos.

Tabela 2 – Comparação entre Kinect 1.0 e Kinect 2.0: principais características.

	Kinect 1.0	**Kinect 2.0**
Câmera RGB (pixel)	1280 x 1024 ou 640 x 480	1920 x 1080
Câmera de Profundidade (pixel)	640 x 480	512 x 424
Distância máxima (m)	4.0	4.5
Distância mínima	0.8	0.5
Campo de Visão		
Horizontal	57	70
Vertical	43	60
Inclinação Física	Sim	Não
Reconhecimento de articulações (Esqueleto humano)	20	26
Reconhecimento de esqueleto completo	2	6
USB	2.0	3.0
Preço	$109,99	$249

Fonte: Adaptado de Pinto (2015)

6.2 COMPARAÇÃO ENTRE O KINECT 1.0 E KINECT 2.0

O Kinect 1.0 é um sensor de baixo custo que permite a medição em tempo real de informações de profundidade (por triangulação) e a aquisição de imagens RGB e IR a uma taxa de quadros de até 30 fps. Kinect 1.0 mede distâncias usando uma técnica de luz estruturada. O projetor de IR emite um padrão de *speckle* (pontos) na cena; A câmera IR capta o padrão refletido e calcula a profundidade correspondente para cada pixel de imagem (PINTO, 2015).

O principal inconveniente do sensor Kinect 1.0 é a baixa qualidade geométrica dos dados, ruído e baixa repetibilidade. O RGB tem qualidade ruim, comparável ao das *webcams*. Os dados de profundidade registrados pelo Kinect 1.0 também têm qualidade ruim, devido ao fato de que a abordagem de luz estruturada nem sempre é suficientemente robusta para fornecer um alto nível de completude da cena emoldurada.

O Kinect 2.0 surgiu como uma proposta de fornecer imagens de alta resolução, melhores medições de profundidade, realizar um rastreamento de esqueleto mais preciso e reconhecimento de gestos. Tem o mesmo número de sensores que o Kinect 1.0, no entanto, a profundidade é medida com um princípio de ToF, ou seja, a medição e realizada por um processo completamente diferente. As imagens RGB e IR adquiridas com o Kinect 2.0 parcialmente se sobrepõem, uma vez que a nova câmera colorida possui um campo de visão horizontal mais largo, enquanto a nova câmera IR tem um campo de visão vertical maior.

Pinto (2015) realizou um procedimento de calibração simples utilizando o Software PhotoModeler para avaliar a precisão e acurácia da estabilidade dos sensores (RGB e

IR) em ambos os casos, o desempenho superior do Kinect 2.0 foi bastante evidente. Foi analisado, também, o erro de profundidade, este erro foi definido como uma função da distância entre o dispositivo e o objeto. Tanto o erro produzido pelo Kinect 1.0 quanto o seu ruído podem ser descritos como funções polinomiais de segunda ordem. O Kinect 2.0 é caracterizado por um erro e uma precisão que aumentam linearmente. Segundo Pinto (2015) resultados promissores foram obtidos ao utilizar a biblioteca *Fusion* para correções de distorções, isto mostra a importância deste tipo de correção caso todo quadro de profundidade seja usado na reconstrução da cena.

7 CONSIDERAÇÕES FINAIS

O Kinect mostrou, desde seu lançamento no mercado, grande potencial para uso em pesquisa porque permite combinar dados visuais e de profundidade, atraindo o interesse de uma ampla variedade de áreas de pesquisa. Ele pode ser controlado remotamente por um PC e usado como um sistema de medição, através da captura de uma grande quantidade de dados em uma alta taxa de quadros.

Na Tabela 3, (Apêndice - A) tem-se alguns exemplos da variedade de aplicações que o Kinect pode estar envolvido. Na sua primeira geração, o Kinect já foi considerado um importante avanço no desenvolvimento de dispositivos de medição 3D mais eficientes. As vantagens mais notáveis do Kinect são sua acessibilidade, portabilidade e velocidade de medições. Além disso, vários Kinects podem ser configurados para trabalhar em conjunto, como outras publicações têm demonstrado (KAENCHAN et al., 2013), (STARANOWICZ; RAY; MARIOTTINI, 2015), (STONE; SKUBIC, 2011).

Por outro lado, existem algumas desvantagens relacionados a usabilidade do Kinect. Em Essmaeel (2012) é mostrado que as medidas retornadas pelo Kinect são tendenciosas. A calibração de múltiplos Kinects continua sendo um desafio. O procedimento de calibração geralmente segue o da calibração de câmera OpenCV, como no projeto RGB Demo (MUSATTI, 2013). No entanto, as características distintivas do Kinect devem ser consideradas, como o deslocamento entre a imagem infravermelha e de profundidade (MUSATTI, 2013) e o modelo geométrico (SMISEK, 2011). Além disso, em um sistema com múltiplos Kinects, a interferência se torna um problema sério. Considerando que as soluções de hardware são eficazes, mas causam uma diminuição no desempenho do sistema, soluções de software são menos eficazes, mas não diminuem o desempenho do sistema em termos de processamento de taxa de quadros. Outra desvantagem do Kinect é a profundidade relativamente curta, de 0,5 a 3 metros, o que a torna imprópria para uso em ambientes externos.

Portanto, a calibração, interferência e correção de polarização parecem ser os principais problemas a serem resolvidos na construção de sistemas Kinect confiáveis e precisos. Uma análise das soluções atuais sugere que, para conceber modelos de calibração com o mínimo de erros e sistemas multiKinect sem interferências, é necessário pesquisar ainda mais para investigar as características específicas da tecnologia de luz estruturada e adaptar as técnicas ao cenário de aplicação considerado.

7.1 TRABALHOS FUTUROS

Este trabalho mostrou que o sensor Kinect é um dispositivo multifuncional e que tem sido cada vez mais utilizado em pesquisas. As áreas médicas, apesar de também serem

beneficiadas com pesquisas com o Kinect, têm carência de tecnologias. Este estudo serve de suporte e abre portas para uma série de novas funcionalidade. Deste modo, pretende-se:

Utilizar o presente trabalho para decidir o melhor método de calibração de múltiplos Kinects;

Fazer testes com múltiplos Kinects e baseado nisto decidir quantos Kinects fazem melhor captura da marcha humana;

Unir o Kinect ao Matlab e desenvolver um sistema capaz de capturar e analisar a marcha humana;

Utilizar as capacidades de aquisição em tempo real do sensor Kinect, para desenvolver sistemas que possam dar suporte a essas áreas.

REFERÊNCIAS

ARAúJO, A. D. G. D. *Uma proposição para o cálculo de mapas de disparidade de imagens estéreo usando um interpolador neural baseado em funções de base radial*. Dissertação (Mestrado) — UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE, 2010. Disponível em: <http://repositorio.ufrn.br:8080/jspui/handle/123456789/15323>. Citado 7 vezes nas páginas 17, 18, 19, 23, 24, 25 e 26.

ARAúJO, L. M. F. M. *Desenvolvimento de um Sistema de Medições Livre de Marcadores Utilizando Sensores de Profundidade*. Dissertação (Mestrado) — UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE – UENF, 2015. Disponível em: <http://uenf.br/posgraduacao/engenharia-de-materiais/wp-content/uploads/sites/2/2013/07/Elisson_Dissertacao_Mestrado_v_Final-com-Corre%C3%A7%C3%B5es-da-Banca-3%C2%AA-c%C3%B3pia.pdf>. Citado 5 vezes nas páginas 18, 26, 27, 28 e 29.

BACKES, J. J. D. M. S. J. A. R. *Introdução à Visão Computacional Usando MATLAB*. [s.n.], 2016. v. 1. Disponível em: <https://books.google.com.br/books?id=m0YlDQAAQBAJ&pg=PA4&lpg=PA4&dq=surgimento+da+vis%C3%A3o+computacional&source=bl&ots=5C_LtxLxXp&sig=CzdZlhaDCZApAUpTMSzSJIhhVPs&hl=pt-BR&sa=X&ved=0ahUKEwjQsbDiks7RAhUIGpAKHRkOAHEQ6AEIVTAH#v=onepage&q=surgimento%20da%20vis%C3%A3o%20computacional&f=false>. Citado na página 14.

BROWN, D. B. C. M. *Computer Vision*. Prentice Hall, 1982. Disponível em: <http://homepages.inf.ed.ac.uk/rbf/BOOKS/BANDB/LIB/bandbpref.pdf>. Citado na página 19.

CENTENO, F. M. M. D. S. J. A. S. Modelagem do erro sistemático de distância nas medições realizadas com a câmara pmd camcube 3.0. *Boletim de Ciências Geodésicas*, v. 21, n. 1, p. 126–148, 2015. Disponível em: <http://www.scielo.br/pdf/bcg/v21n1/1982-2170-bcg-21-01-00126.pdf>. Citado na página 28.

ESSMAEEL, L. G. E. D. G. D. P. K. Multiple structured light-based depth sensors for human motion analysis: A review. *Ambient Assisted Living and Home Care*, 2012. Disponível em: <https://www.researchgate.net/publication/259215660_Multiple_Structured_Light-Based_Depth_Sensors_for_Human_Motion_Analysis_A_Review>. Citado 3 vezes nas páginas 46, 47 e 50.

FABIAN, J. et al. Integrating the microsoft kinect with simulink: Real-time object tracking example. *IEEE/ASME Transactions on Mechatronics*, v. 19, n. 1, p. 249–257, fev. 2014. ISSN 1083-4435. Citado na página 47.

FERNáNDEZ-CABALLERO, P. G. J. P. M. M. A. My kinect is looking at me - application to rehabilitation. *Ambient Intelligence-Software and Applications*, p. 233–241, 2015. Disponível em: <https://www.researchgate.net/publication/283099563_My_Kinect_Is_Looking_at_Me_-_Application_to_Rehabilitation>. Citado na página 46.

GOMES, J. E. R. de Q. . H. M. *Introdução ao Processamento Digital de Imagens*. [S.l.], 2014. Disponível em: <http://www.dsc.ufcg.edu.br/~hmg/disciplinas/graduacao/vc-2014.1/Rita-Tutorial-PDI.pdf>. Citado na página 18.

GONZALEZ, R. W. R. *Digital Image Processing*. [S.l.]: Pearson Education, 1992. Citado na página 19.

JúNIOR, J. P. D. S. *Alinhamento de Imagens de Profundidade com Aplicação no Reconhecimento da Língua de Sinais*. Dissertação (Mestrado) — Universidade de Brasília, 2014. Disponível em: <http://biblioteca.universia.net/html_bura/ficha/params/title/alinhamento-imagens-profundidade-com-aplica%C3%A7%C3%A3o-reconhecimento-da-lingua-sinais/id/60933787.html>. Citado na página 29.

KAEHLER, G. B. A. *Learning OpenCV*. O'Reilly Media, 2008. Disponível em: <http://www-cs.ccny.cuny.edu/~wolberg/capstone/opencv/LearningOpenCV.pdf>. Citado na página 19.

KAENCHAN, S. et al. Automatic multiple kinect cameras setting for simple walking posture analysis. In: *Proc. Int. Computer Science and Engineering Conf. (ICSEC)*. [S.l.: s.n.], 2013. p. 245–249. Citado na página 50.

KARLSTROEM, A. Correspondência entre imagens segundo geometria epipolar em projeção perspectiva. 2006. Disponível em: <http://monoceros.mcca.ep.usp.br/ESL/publications/rr2004-01.pdf/view>. Citado na página 24.

KONDRAT, E. *Scanner 3D: Aquisição de pontos 3D por raio Laser*. [S.l.], 2011. Citado 2 vezes nas páginas 32 e 33.

KRONLACHNER, M. *The Kinect distance sensor as human-machine-interface in audio-visual art projects*. [S.l.], 2013. Disponível em: <http://www.matthiaskronlachner.com/wp-content/uploads/2013/01/2013-01-07-Kronlachner-Kinect.pdf>. Citado 6 vezes nas páginas 37, 38, 39, 40, 41 e 45.

MAFRA, N. R. *Análise de Imagem na Avaliação Clínica da Marcha Humana*. Dissertação (Mestrado) — Universidade do Porto, 2012. Disponível em: <https://web.fe.up.pt/~tavares/downloads/publications/teses/MSc_NunoMafra.pdf>. Citado na página 17.

MAIMONE, A.; FUCHS, H. Reducing interference between multiple structured light depth sensors using motion. In: *Proc. IEEE Virtual Reality Workshops (VRW)*. [S.l.: s.n.], 2012. p. 51–54. ISSN 1087-8270. Citado na página 43.

MARR, D. *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. [S.l.]: MIT Press, 1982. Citado na página 23.

MARTíN, R. M.; LORBACH, M.; BROCK, O. Deterioration of depth measurements due to interference of multiple rgb-d sensors. In: *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*. [S.l.: s.n.], 2014. p. 4205–4212. ISSN 2153-0858. Citado 3 vezes nas páginas 17, 41 e 47.

MICROSOFT. *Kinect for Xbox 360*. [S.l.], 2010. Disponível em: <http://www.xbox.com/en-US/xbox-360/accessories/kinect>. Citado na página 35.

MUSATTI, C. P. A. *Introduzione all'utilizzo del sensore microsof Kinect*. Dissertação (Mestrado) — UNIVERSITÀ DI BRESCIA, 2013. Disponível em: <http://www.cassinis.it/Siti%20ex%20Uni/ARL/docs/projects/Sen_09.pdf>. Citado 11 vezes nas páginas 30, 31, 35, 36, 37, 41, 42, 43, 44, 47 e 50.

PETROU, P. B. M. *Image Processing: The Fundamentals*. [S.l.]: John Wiley & Sons, 1999. Citado na página 18.

PINTO, D. P. L. Calibration of kinect for xbox one and comparison betweenthe two generations of microsoft sensors. *sensores*, 2015. Disponível em: <https://www.researchgate.net/publication/283482095_Calibration_of_Kinect_for_Xbox_One_and_Comparison_between_the_Two_Generations_of_Microsoft_Sensors>. Citado 2 vezes nas páginas 48 e 49.

RAFIBAKHSH, J. G. M. K. S. C. G. H. F. L. N. Analysis of xbox kinect sensor data for use on construction sites: Depth accuracy and sensor interference assessment. *Construction Research Congress*, 2012. Disponível em: <https://www.researchgate.net/publication/268589647_Analysis_of_XBOX_Kinect_Sensor_Data_for_Use_on_Construction_Sites_Depth_Accuracy_and_Sensor_Interference_Assessment>. Citado na página 47.

RIOS, L. R. S. Visão computacional. 2010. Disponível em: <http://homes.dcc.ufba.br/~luizromario/Apresenta%C3%A7%C3%A3o%20de%20IA/Artigo%20(final).pdf>. Citado 4 vezes nas páginas 19, 20, 21 e 22.

SARBOLANDI, D. L. A. K. H. Kinect range sensing: Structured-light versus time-of-flight kinect. 2015. Disponível em: <https://arxiv.org/pdf/1505.05459.pdf>. Citado na página 14.

SCHöNAUER, T. P. H. K. C. Chronic pain rehabilitation with a serious game using multimodal input. *International Conference on Virtual Rehabilitation 2011*, 2011. Disponível em: <https://publik.tuwien.ac.at/files/PubDat_204331.pdf>. Citado na página 17.

SCURI, A. E. *Fundamentos da Imagem Digital*. Tecgraf/PUC-Rio, 2002. Disponível em: <https://webserver2.tecgraf.puc-rio.br/~scuri/download/fid.pdf>. Citado na página 18.

SMISEK, M. J. T. P. J. 3d with kinect. *IEEE International Conference on Computer Vision Workshops*, 2011. Citado na página 50.

STARANOWICZ, A. N.; RAY, C.; MARIOTTINI, G. L. Easy-to-use, general, and accurate multi-kinect calibration and its application to gait monitoring for fall prediction. In: *Proc. 37th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*. [S.l.: s.n.], 2015. p. 4994–4998. ISSN 1094-687X. Citado na página 50.

STONE, E. E.; SKUBIC, M. Passive in-home measurement of stride-to-stride gait variability comparing vision and kinect sensing. In: *Proc. Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*. [S.l.: s.n.], 2011. p. 6491–6494. ISSN 1094-687X. Citado na página 50.

SUNYOTO, H.; MARK, W. van der; GAVRILA, D. M. A comparative study of fast dense stereo vision algorithms. In: *Proc. IEEE Intelligent Vehicles Symp.* [S.l.: s.n.], 2004. p. 319–324. Citado na página 23.

WERBER, K. *Intuitive Human Robot Interaction and Workspace Surveillance by means of the Kinect Sensor*. Dissertação (Mestrado) — Lund University, 2011. Disponível em: <https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=2198971&fileOId=2214422>. Citado na página 35.

ZHANG, Z. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, v. 19, n. 2, p. 4–10, fev. 2012. ISSN 1070-986X. Citado 3 vezes nas páginas 14, 15 e 40.

Apêndices