UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

VINÍCIUS CAVALCANTI NOGUEIRA DE SÁ

DETECÇÃO DE MÃOS ATRAVÉS DA COMBINAÇÃO DE TÉCNICAS DE DETECÇÃO DE TOM DE PELE E MOVIMENTO PARA BACKGROUND COMPLEXO

Recife

2018

VINÍCIUS CAVALCANTI NOGUEIRA DE SÁ

DETECÇÃO DE MÃOS ATRAVÉS DA COMBINAÇÃO DE TÉCNICAS DE DETECÇÃO DE TOM DE PELE E MOVIMENTO PARA BACKGROUND COMPLEXO

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Estatística e Informática

Curso de Bacharelado em Ciência da Computação

Orientador: Valmir Macário Filho

Recife

2018

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

FICHA DE APROVAÇÃO DO TRABALHO DE CONCLUSÃO DE CURSO

Trabalho defendido por Vinicius Cavalcanti Nogueira de Sá às 15 horas do dia 14 de agosto de 2018, no Auditório do CEAGRI-02 – Sala 07, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado **Detecção de Mãos Através da Combinação de Técnicas de Detecção de Tom de Pele e Movimento para Background Complexo**, orientado por Valmir Macário Filho e aprovado pela seguinte banca examinadora:

Valmir Macário Filho
DC/UFRPE

João Paulo Silva do Monte Lima
DC/UFRPE

À todos que me ajudaram no desenvolvimento deste estudo e aos que creram que era possível.

Agradecimentos

Agradeço a meus pais, Laís e Antonino, pelos anos de dedicação e influência que foram fundamentais para a minha formação. Agradeço a minha namorada Bianca, que me ajudou bastante e ofereceu muito amor e paciência comigo no tempo necessário para finalizar este trabalho. Agradeço a todos os professores responsáveis pela minha formação acadêmica, especialmente a meu orientador Valmir. Agradeço a todos os amigos que fiz em minha trajetória na Universidade Federal Rural de Pernambuco, pois todo tiveram sua parcela de participação nessa trajetória.

“Da próxima vez que alguém reclamar que você cometeu um erro, diga a essa pessoa que talvez isso seja uma boa coisa, porque sem imperfeição nem você nem eu existiríamos.”

(Stephen Hawking)

Resumo

A tecnologia tem como função social facilitar a vida de seus usuários. Com a evolução da mesma, e com o surgimento da globalização, o acesso à informação e a comunicação como um todo se tornaram muito mais disponíveis para população em geral. Ainda assim, grupos com necessidades especiais sofrem com a defasagem de produtos e sistemas que possam atender as suas necessidades. Este trabalho fará uso de tecnologias pré-existentes que possam ser usadas de modo a facilitar a vida desses usuários, mais especificamente surdos. Vivemos em um mundo onde nos deparamos com uma imensidão de dispositivos com câmeras, ou de equipamentos que podem ser conectados a uma. A visão computacional se torna muito importante ou senão essencial a partir dessa realidade. Diversas áreas utilizam imagens para automatizar ou auxiliar as suas atividades dentro de seus segmentos, sendo eles voltados para o entretenimento, indústria ou outros. Sendo assim, é possível perceber a importância do processamento de imagens como solução de problemas em áreas diversas. Neste trabalho foi utilizado o processamento de imagem para elaborar uma possível solução na área de reconhecimento de mãos. A utilização da mão como uma maneira de comunicação é evidente. Podemos vê-la como uma personagem principal não somente na comunicação cotidiana através de gestos, como também podemos utiliza-la no controle de interfaces computacionais, no auxílio na imersão em realidade virtual, na manipulação de objetos virtuais em uma realidade aumentada. Também podemos vê-la como facilitadora na acessibilidade a partir da comunicação por sinais, sendo este último exemplo o ponto chave deste trabalho, que visa facilitar a comunicação entre surdos e possíveis usuários interessados na língua de sinais através de uma nova abordagem. O reconhecimento de mão foi realizado por meio de uma abordagem híbrida envolvendo segmentação por tons de pele e movimento, esta abordagem foi escolhida para contornar as dificuldades que cada tipo de segmentação traz. A melhor taxa de acerto que tivemos com esta abordagem 76,4% em ambientes internos e 45,15% em ambientes externos.

Palavras-chave: tecnologia, sistema, surdos, processamento de imagem, mãos.

Abstract

Technology has a social function to facilitate the life of its users, with its evolution, and with the emergence of globalization, the access to information and communication in general have become much more accessible for the general population. Nevertheless, groups with special needs still suffer from the lack of products and systems that can meet their needs. This work will make use of pre-existing technologies that can be used to make life easier for these users, especially deaf users. We live in a world where we are faced with an immensity of devices with cameras, or of equipment that can be connected to one, the computer vision becomes very important or otherwise, essential from this reality. Many areas use images to automate or assist their activities within their segments, whether they are for entertainment, industry or others. Thus, it is possible to realize the importance of image processing as a solution of problems in different areas. In this work it was used image processing to elaborate a possible solution in the hand recognition area, the use of the hand as a way of communication is evident. We can see it as a main character not only in everyday communication through gestures, but we can also use it in the control of computational interfaces, in the aid of immersion in virtual reality, in the manipulation of virtual objects in augmented reality or even as facilitator in the accessibility from the communication by signals, being this last example the key point of this work, that aims to facilitate the communication between deaf and possible users interested in the sign language through a new approach. Hand recognition was performed through a hybrid approach involving skin tone segmentation and movement, this approach was chosen to overcome the difficulties that each type of segmentation brings. The best hit rate we had with this approach was 76.4% indoors and 45.15% in outdoor environment.

Keywords: technology, system, deaf, image processing, hands.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

IHM	Iteração Humano-computador
RGB	*Red, Green, Blue* (Vermelho, Verde e Azul, respectivamente. Espaço de cor)
YCbCr	*Luminance, Croma Blue, Croma Red* (Luminância, Crominância azul, Crominância Vermelha, respectivamente. Espaço de cor)
IHLS	*Improved Hue, Luminance, Saturation* (Tonalidade melhorada, Luminância, Saturação, respectivamente. Espaço de cor)
HLS	*Hue, Luminance, Saturation* (Matiz, Luminância, Saturação, respectivamente. Espaço de cor)
HSV	*Hue, Saturation, Value* (Matiz, Saturação, Valor, respectivamente. Espaço de cor)
HSI	*Hue, Saturation, Intensity* (Matiz, Saturação, Intensidade, respectivamente. Espaço de cor)
CIE-XYZ	*Commission internationale de l'éclairage* - XYZ (variação do RGB que não possui valores negativos. Espaço de cor)
FPS	*Frames* Por Segundo
ROI	Region of Interest (Região de Interesse)
GPU	*Graphics Processing Unit* (Unidade de processamento gráfico)
IDE	*Integrated Development Environment* (Ambiente de Desenvolvimento Integrado)
px	Pixel
IOU	*Intersection over Union* (Intersecção Sobre União)

Sumário

1 Introdução

1.1 Visão Geral

Comunicação entre pessoas está presente em todos os meios sociais, afinal todos os seres se comunicam. A comunicação pode ser representada oralmente, através da escrita, de maneira verbal ou não verbal, e é nesta última que este projeto estará situado. As pessoas naturalmente fazem uso de gestos como um auxiliador na comunicação, e nas pessoas com deficiência auditiva podemos ver os gestos como protagonistas, já que eles se utilizam da linguagem de sinais. Sendo assim, com o avanço da tecnologia como facilitadora na comunicação, nada mais natural do que ver esse intercâmbio acontecendo através de dispositivos como Kinect[1], Leap Motion[2], OptiTrack[3], câmera comuns[4] e 3D[5], entre outros. O desenvolvimento dos respectivos dispositivos se torna possível pela existência das técnicas de processamento de imagens.

A área de processamento de imagens visa extrair, analisar e utilizar informações relevantes a partir de imagens. Atualmente temos diversos dispositivos com câmeras, como celulares inteligentes, videogames, televisores inteligentes, entre outros. Isto sem citar os diversos tipos de câmeras existentes. Estes dispositivos estão em todos os lugares, tanto no meio industrial para o controle de qualidade e segurança, como para o uso cotidiano. A área de interação humano-computador (IHC), possivelmente, uma das áreas mais pesquisadas da atualidade visa estudar e facilitar a interação entre humanos e computadores (RAUTARAY; AGRAWAL, 2015). Entende-se que imagens são importantíssimas para o mundo moderno, e que áreas como medicina, segurança, entretenimento, comunicação e indústria são algumas das que utilizam processamento de imagem para auxiliar ou resolver problemas.

Um ponto importante que devemos citar, é o papel do vídeo no processamento de imagem, o mesmo trata-se de um conjunto de várias imagens reproduzidas em sequência. Então, a utilização de vídeos abre uma gama de possibilidades de sistemas, em diferentes áreas. Com vídeos, aplicações mais interessantes podem ser desenvolvidas, como o reconhecimento de gestos, controle de interfaces, sistemas e tradução de linguagens de sinais.

Os sistemas para controle de interfaces, acessibilidade, tradução de linguagem

[1] https://developer.microsoft.com/pt-br/windows/kinect
[2] https://www.leapmotion.com/
[3] http://optitrack.com/
[4] Dispositivos para captura de imagens em duas dimensões
[5] Dispositivos para captura de imagens em três dimensões

de sinais entre outros, dependem de métodos de reconhecimento de gestos. No caso, uma câmera irá captar imagens que serão processadas e então classificadas conseguindo assim, determinar qual gesto foi realizado. O reconhecimento de mãos é extremamente importante para estes sistemas, pois normalmente *softwares* deste segmento utilizam as mãos como principal foco de gestos, e ao realizar a etapa de reconhecimento de mãos de modo eficiente e robusto esses sistemas conseguirão trabalhar de maneira correta.

Este trabalho visa analisar técnicas de processamento de imagens para o reconhecimento de mão. Existem diversas técnicas que auxiliam o mesmo, como por exemplo a detecção de cor, rastreamento de movimento, casamento de formas, remoção de *background* ou a utilização de imagens de profundidade (RAUTARAY; AGRAWAL, 2015).

"*Gesture Control Device*" tem se mostrado uma tecnologia emergente nos dias atuais, como pode ser observado em gráficos desenvolvidos anualmente pela Gartner, empresa essa, dedicada à pesquisa da inserção da tecnologia no mercado (STAMFORD, 2017).

> O relatório Hype Cycle for Emerging Technologies é o Ciclo de Hype Gartner anual mais longo, fornecendo uma perspectiva inter-indústria sobre as tecnologias e tendências que estrategistas de negócios, líderes em inovação, líderes de R & D, empreendedores, desenvolvedores de mercado global e equipes de tecnologia emergente devem considerar o desenvolvimento de portfólios de tecnologia emergente (STAMFORD, 2017).

É possível perceber que "*Gesture control device*" está na fase chamada "*Peak of Inflated Expectations*" o que significa que a respectiva tecnologia se mostra promissora, e que levará cerca de 5 a 10 anos para se tornar *mainstream*. *Mainstream* expressa uma tendência, ou algo que está na moda, isto indica que este tipo de tecnologia poderá ser muito importante futuramente, e que possivelmente será largamente utilizada por todos tipos de dispositivos e sistemas possíveis. Tendo como exemplo o *touchscreen*, tecnologia esta que se tornou *mainstream*, e acabou mudando o modo de como utilizamos telefones celulares ajudando a criar o que chamamos hoje de *smartphone*. Sendo assim podemos concluir que o controle de dispositivos por meio de gestos poderá ser algo tão grandioso quanto foi o *touchscreen*. Caso o *Hype Cycle* esteja correto, teremos uma grande mudança na forma como nos comunicamos com os computadores ou com outras pessoas.

Dito isso, podemos perceber a importância que este projeto, depois de pronto, promoverá no desenvolvimento de novos softwares que também façam uso do proces-

samento de imagem, além de promover um grande impacto social com a facilitação da comunicação entre os falantes de Libras.

Neste trabalho será encontrada uma abordagem implementada pelo autor propondo um modelo híbrido de reconhecimento de mãos que poderá ser utilizado como etapa previa de sistemas de reconhecimento de gestos. Para determinar a eficácia foram codificadas duas abordagens distintas, mas que têm propósito parecido.

Deste modo será possível visualizar diversas abordagens de processamento de imagens, sendo estas, conversão de espaço de cor de RGB para YC_bC_r, conversão para tons de cinza, binarização, remoção de ruídos, correção de iluminação, segmentação de tons de pele e segmentação de movimento. Estes métodos foram combinados visando a aplicação em ambientes dinâmicos, com mudança de iluminação e movimentação de *background*. A base de dados escolhida cumpre os pontos citados acima, os vídeos contêm estes aspectos e foram utilizados por todas as implementações.

O modo de avaliação utilizado foi o *Intersection Over Union* ou IOU, que consiste em descobrir o grau de semelhança de uma imagem segmentada. Esta técnica foi utilizada em todos as implementações mantendo assim uma consistência nos dados.

1.2 Problemas de Pesquisa

Este projeto visa conseguir determinar um conjunto de técnicas que sejam mais eficientes e robustas na abordagem da detecção e rastreamento de mãos, porém descartando a necessidade de instrumentos, como luvas ou marcadores, evitando assim que a utilização se torne desconfortável ou pouco funcional. As aplicações que utilizam estes instrumentos são baseadas em reconhecimento de gestos. Evidentemente é necessário fazer a captura dos gestos realizados pelas mãos, e para ser eficiente estes sistemas necessitam de abordagens que sejam robustas na segmentação. Como dito anteriormente, o foco deste trabalho é determinar um conjunto de técnicas para a resolução desta problemática, e assim auxiliar os sistemas de reconhecimento de gestos, pois o que será visto neste projeto é a primeira etapa deste tipo de sistema.

Existem diversas abordagens a serem seguidas, como por exemplo, a segmentação por tons de pele, rastreamento de movimento, casamento de formas, utilização de características de profundidade, entre outras, como podemos observar na pesquisa de Rautaray e Agrawal (RAUTARAY; AGRAWAL, 2015).

Há não muito tempo, as técnicas anteriormente citadas eram comumente utilizadas de forma isolada, porém, com o desenvolvimento de recursos computacionais, tais técnicas passaram a ser combinadas, criando assim, novos tratamentos. Estas combinações surgem para suprir as deficiências que os algoritmos possam ter ao ser

aplicados em particular, deixando-o assim mais confiável.

Neste projeto tentaremos utilizar deste conceito, contudo, esta fusão torna a implementação mais complexa, e também consumirá mais recursos de hardware. Os algoritmos a serem utilizados devem ter como características principais serem rápidos e eficientes, para que possam ser aplicados em sistemas de tempo real e sem o uso de utensílios.

1.3 Objetivos

1.3.1 Objetivo Geral

Este trabalho tem como objetivo geral avaliar um conjunto de técnicas para localização de mãos robusta e em tempo real, que faça uso de técnicas de reconhecimento de tons de pele e movimento.

1.3.2 Objetivos Específicos

a) Determinar qual sistema de cor será utilizado para detecção de tom de pele das mãos.

b) Utilizar técnicas de detecção de movimentos para localização de mãos.

c) Utilizar combinação de técnicas de detecção de tom de pele em conjunto com técnicas de detecção de movimentos para localização de mãos.

d) Analisar a qualidade de segmentação da mão com as técnicas aplicadas.

e) Comparar resultados com outras abordagens já existentes.

1.4 Estrutura do Trabalho

Este trabalho está estruturado em sete capítulos, seguindo uma sequência lógica para facilitar o entendimento do problema a ser abordado e também como o problema pode ser resolvido.

No capitulo 1 foi apresentado uma visão geral do trabalho, os objetivos gerais e específicos, e como este trabalho foi desenvolvido.

Todo conhecimento necessário para o entendimento da abordagem proposta será disseminado no capitulo 2, fundamentação teórica, nele estará todas as definições acerca do tema.

O capítulo 3 é composto dos trabalhos relacionados a esta pesquisa.

O capítulo 4 explanará a análise dos algoritmos implementados e proposto, sendo executados na base de dados escolhida, e então será discutido o que poderá ser feito para melhorar os resultados obtidos.

No capítulo 5 será demostrada a metodologia aplicada a este trabalho, será também informado o ambiente experimental, a linguagem utilizada e o ambiente de desenvolvimento. Ele também contará com informações do banco de dados utilizado, das implementações que foram realizadas e trará a abordagem proposta pelo autor deste trabalho, e por fim as métricas de análise e experimentação.

O capítulo 6 é composto da análise dos resultados, sendo estes obtidos a partir das implementações apresentadas no capítulo 4.

Por fim o capítulo 7, onde estará presente a conclusão deste trabalho, as considerações finais e o que poderá ser feito futuramente.

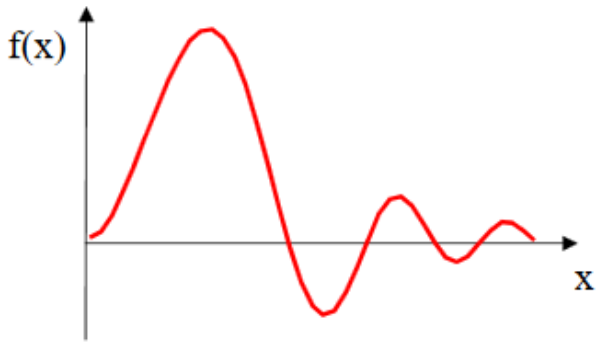
2 Fundamentação Teórica

Neste Capítulo serão descritos alguns conceitos referentes à área de processamento de imagens digitais. Serão descritos os conceitos de Imagens Digitais, Pixel, Vídeos Digitais, Espaço de cor, Histograma, Filtros, Operadores Morfológicos e Limiarização.

2.1 Imagens Digitais

Para entendermos o que são imagens digitais, temos que pensar no que são imagens, no caso, imagem nada mais é que uma representação da luz que está sendo capturada num determinado momento por um sensor de uma máquina fotográfica. Assim, a imagem é uma representação analógica de um momento, podendo assim, matematicamente, ser representada por uma função contínua como mostrado na Figura 1.

Figura 1 – Gráfico de uma função contínua

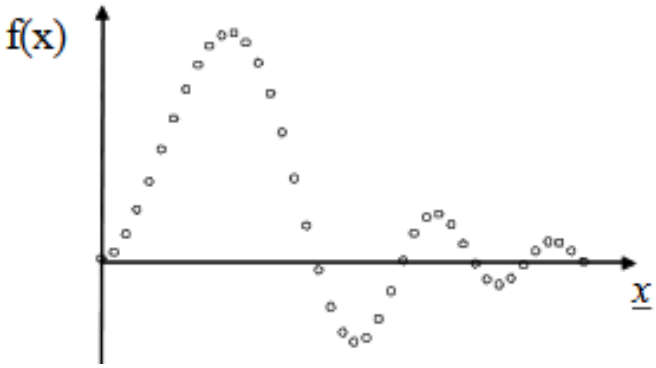


Fonte: (SCURI, 1999)

Como estamos em um meio digital, o computador, que trabalha com valores discretos, é preciso discretizar a função contínua. Isso é a obtenção de valores pontuais de $f(x)$ em x, como dito por Scuri (SCURI, 1999), então obtemos um gráfico da seguinte maneira, conforme mostrado na Figura 2.

Dito isto, pode-se definir que imagens digitais nada mais são do que imagens analógicas discretizadas. Normalmente, a imagem é representada na forma de uma matriz, e cada célula da matriz é um pixel.

Figura 2 – Gráfico de uma função discretizada

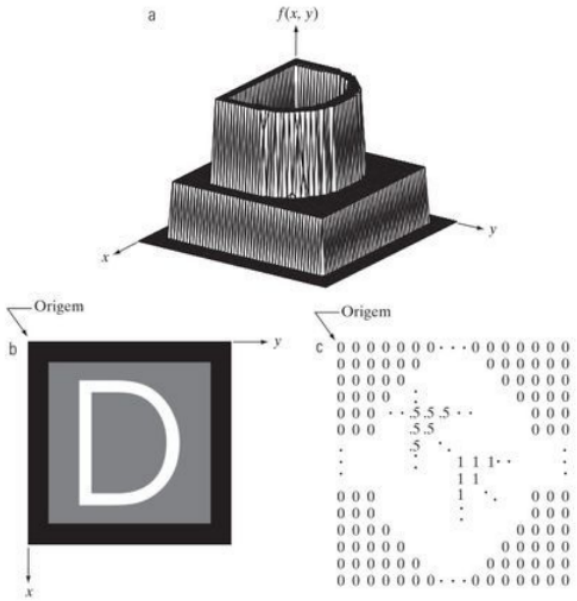


Fonte: (SCURI, 1999)

2.2 Pixel

As imagens digitais podem ser apresentadas como matrizes, então pela definição de matriz, temos que a imagem pode ser definida como sendo um conjunto de linhas(i) e colunas(j), sendo i e j números inteiros, assim vemos que cada elemento (i,j) da matriz é um pixel. Estes, dependendo do espaço de cor considerado, terá um valor ou um conjunto de valores que determinará a intensidade da cor, este conceito pode ser visualizado na Figura 3.

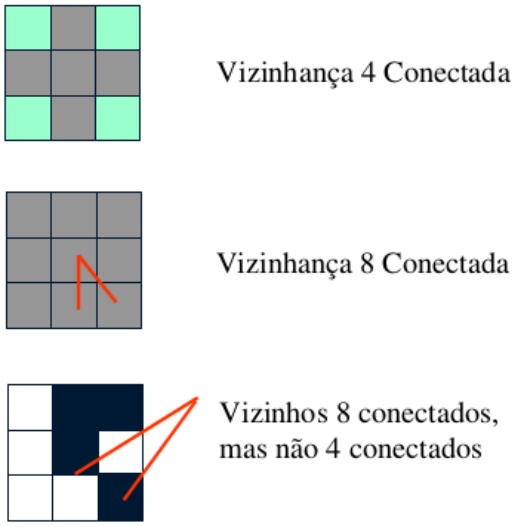
Figura 3 – (a) Imagem representada graficamente como uma superfície. (b) Representação como imagem digital. (c) Imagem representada como matriz numérica, cada valor representa um pixel de intensidade 0, 5 e 1, preto, cinza e branco, respectivamente.



Fonte: (GONZALEZ; WOODS, 2011)

Os pixeis possuem uma topologia, que pode ser definida como 4-conectada ou 8-conectada, a topologia de um pixel consiste em considerar os pixeis que são vizinhos a ele, no caso da 4-conectada o pixel central possui 4 pixeis a sua volta, estando eles posicionados a esquerda e direita e acima e abaixo, já o 8-conectado considera os mesmo vizinhos da 4-conectada e os pixeis das diagonais. A Figura 4 representa este conceito.

Figura 4 – Representação da vizinhança de um pixel.



Fonte: (SCURI, 1999)

2.3 Vídeos Digitais

Vídeos digitais, segundo Telkap (TEKALP, 2015) nada mais são do que imagens digitais que variam a intensidade da cor dos pixeis ao longo do tempo. Partindo deste princípio pode-se perceber que operações que são aplicadas a imagens digitais, podem também ser aplicadas a vídeos digitais.

2.4 Espaço de cores

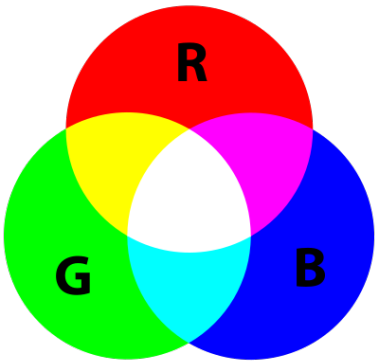
Antes de falar de espaço de cores, que é nossa próxima definição, teremos que falar sobre cor. No mundo real, a cor é a representação visual dos diferentes comprimentos de onda dos raios de fótons refletidos pelos objetos, diferentes comprimentos de onda farão o olho humano perceber diferentes cores. Já os espaços de cores são, modelos matemáticos para representar as informações das cores, com três ou quatro componentes de cores diferentes, segundo Shaik (SHAIK et al., 2015). Cada espaço de cor, pode ser utilizado para diferentes aplicações, dependendo somente da necessidade de cada problema.

Naturalmente os vídeos obtidos por meio de câmeras digitais, e que serão utilizados neste projeto, estão no espaço de cor RGB, no qual cada pixel possui três canais de cor, Vermelho (R), Verde (G) e Azul (B). Neste trabalho, todos os vídeos serão convertidos a cada *frame* para o espaço de cor YC_bC_r, que possue também três componentes. Y refere-se a luminância, e os canais C_b e C_r, referem-se a crominância. Por definição, a luminância está relacionado ao brilho da imagem, que varia do preto ao branco. A crominância, se refere ao valor de cada cor, no caso do YC_bC_r os canais C_b e C_r, as cores são azul e vermelho.

2.4.1 RGB

Segundo Rautaray, Agrawal (RAUTARAY; AGRAWAL, 2015) e Shaik (SHAIK et al., 2015) os espaços de cores que conseguem realizar uma separação entre os componentes de crominância e luminância são os mais indicados para a segmentação por tons de pele. O espaço de cor RGB não possui essa característica, pois ele é um sistema baseado em cores, em que a mistura das três cores presentes no sistema, podem gerar outras cores. Sendo assim o RGB não é indicado para tarefas que necessitem análise de cores ou detecções baseadas em cores, porque as informações de crominância e luminância não são uniformes. A Figura 5 ilusta este sistema de cor.

Figura 5 – Sistema de cor RGB



Fonte: (HORVATH, 2006)

2.4.2 YCbCr

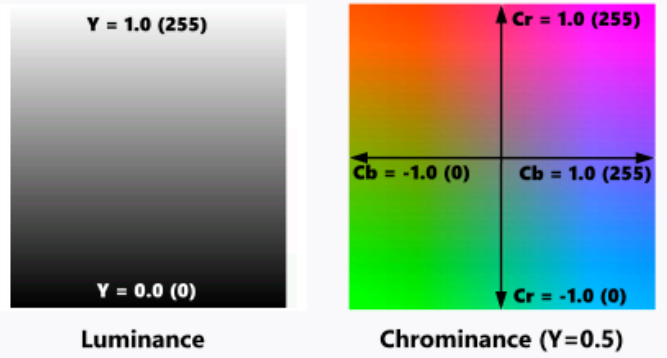
O sistema de cor YC_bC_r é considerado um espaço de cor ortogonal, então seus componentes são estatisticamente independentes. Estes componentes são separados em duas categorias, luminância e crominância, a luminância é definida pelo componente Y, e a crominância é subdivida em C_b e C_r, conforme dito por Kakumanu, Praveen and Makrogiannis, Sokratis and Bourbakis, Nikolaos em (KAKUMANU; MAKRO-

GIANNIS; BOURBAKIS, 2007). O autor também comenta que o YC_bC_r é a escolha mais popular para detecção de tons de pele.

Kakumanu, Praveen and Makrogiannis, Sokratis and Bourbakis, Nikolaos em (KAKUMANU; MAKROGIANNIS; BOURBAKIS, 2007) define que o componente Y é obtido a partir da soma ponderada de valores RGB, a crominância é calculada subtraindo o componente de luminância dos canais B e R do sistema RGB.

Outros pontos que podemos observar é o fato dele ser facilmente obtido, pois a conversão das imagens em RGB para YC_bC_r é realizada de maneira rápida e eficiente, em comparação com outros espaços de cores, ele ainda é ideal para ser trabalhado em imagens complexas sob mudanças de iluminação conforme afirmados por Shaik (SHAIK et al., 2015). O espaço de cor YC_bC_r pode ser exemplificado pela Figura 6.

Figura 6 – Sistema de cor YC_bC_r



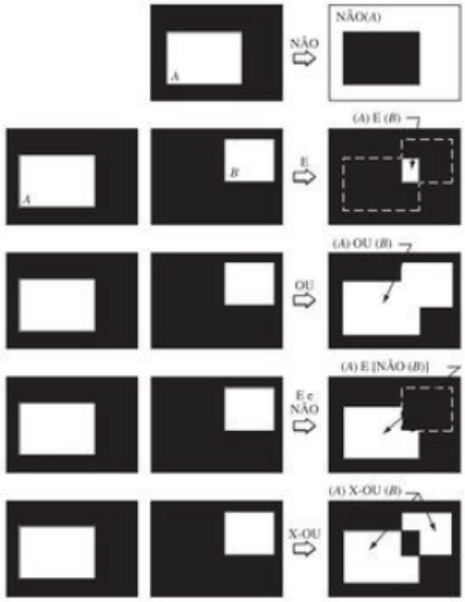
Fonte: (SLACKERPRIME, 2015)

2.5 Operadores Lógicos

Pela definição de Gonzales e Woods (GONZALEZ; WOODS, 2011), operações de conjuntos numéricos podem ser aplicados a imagens, então ao considerar cada imagem como um conjunto, pode-se aplicar as operações de união, subconjunto, interseção, conjuntos disjuntos e complemento. Isto é verdade para imagens coloridas que estejam no mesmo espaço de cor e também para imagens em preto e branco, também conhecidas como imagens binárias.

As imagens em preto e branco são de suma importância para este trabalho, pois serão aplicados operadores lógicos que como Gonzales e Woods (GONZALEZ; WOODS, 2011) define, são as operações aplicadas a conjunto, porem é costuma-se referir a elas como operadores lógicos, ”OU”(OR), ”E”(AND), ”NÃO”(NOT) e ”OU Exclusivo”(XOR). Considerando duas imagens binárias A e B, temos as situações apresentadas na Figura 7.

Figura 7 – Ilustração de operações logicas aplicadas a imagens binárias

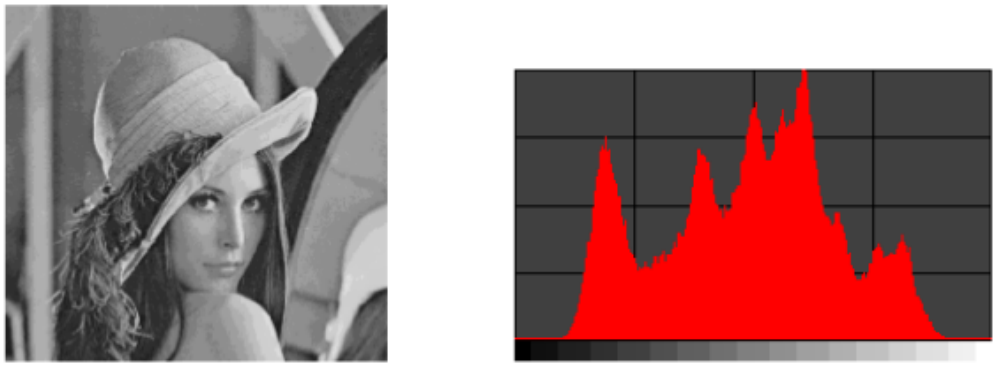


Fonte: (GONZALEZ; WOODS, 2011)

2.6 Histograma

A definição de histograma dada por Scuri (SCURI, 1999) é simples, ela é definida por uma função estatística da imagem digital, no qual para cada tonalidade existente na imagem, se tem a quantidade de pixels na imagem com aquele respectivo valor.

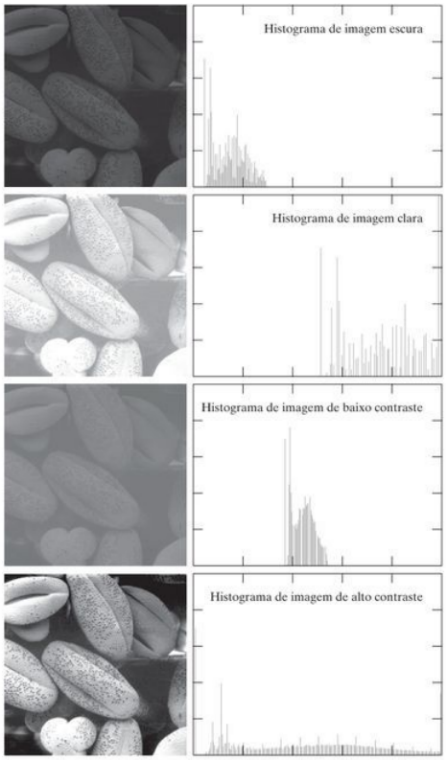
Figura 8 – Histograma de uma imagem em tons de cinza.



Fonte: (SCURI, 1999)

Imagens em diferentes situações, possuem diferentes histogramas, e isto pode ser um problema ao trabalhar com vídeos, pois cada *frame* poderá ter mudanças na iluminação, o que acarreta diversos problemas, devido a mudança de intensidade dos pixeis, deste modo, não se tem uma consistência, a Figura 9 exemplifica este problema.

Figura 9 – Quatro tipos básicos de imagens, escura, clara, baixo contraste, alto contraste.



Fonte: (GONZALEZ; WOODS, 2011)

Quando trabalhamos com histogramas, pode-se aplicar operações como a equalização, no qual os resultados podem ajudar outras etapas do processamento de imagem, pois podemos ajustar o histograma, aplicando algumas funções como a equalização, para obter uma maior diferenciação dos objetos de uma imagem.

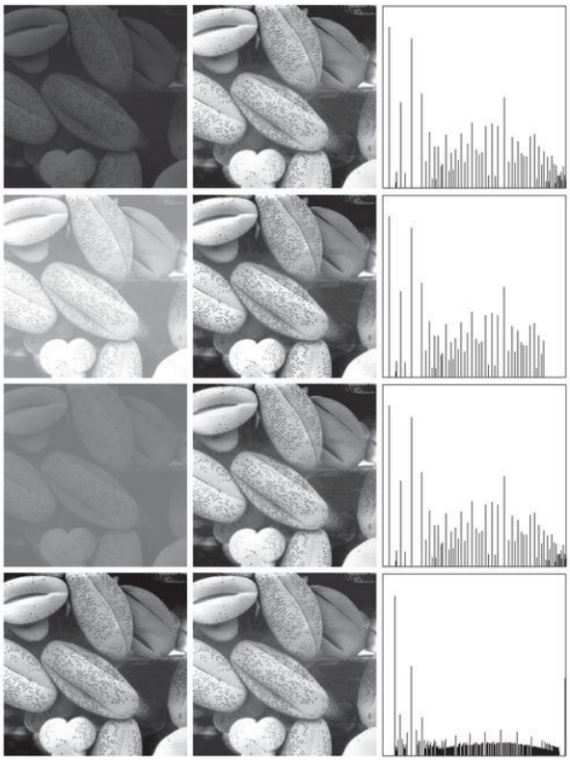
Equalização do histograma é segundo Szeliski (SZELISKI, 2010), uma função que tem como resultado um histograma mais “plano”, no qual, é realizado um mapeamento das intensidades e então os valores dos pixeis mais escuros serão mais iluminados, e os pixeis mais claros serão escurecidos. A Figura 10 exemplificará esta aplicação.

É possível perceber pela Figura 10 que após a aplicação da equalização de histograma, as imagens com diferentes valores de histograma ficaram com histograma mais equilibrado e então as imagens resultantes, em todos os casos serão equivalente.

2.7 Filtragem espacial

Segundo Segundo Gonzalez e Woods (GONZALEZ; WOODS, 2011) um filtro consiste em aceitar ou rejeitar a passagem de certos componentes da imagem. Como vimos anteriormente, a imagem pode ser vista como uma função que possui uma

Figura 10 – A coluna da esquerda é relacionado as imagens da Figura 9, a coluna central é das imagens após equalização do histograma e a coluna da direita é relativa ao novo histograma equalizado.



Fonte: (GONZALEZ; WOODS, 2011)

frequência. Então, filtragem espacial consiste em aceitar ou rejeitar componentes da frequência da imagem.

A filtragem espacial pode ser utilizada para suavizar ou borrar componentes na imagem, identificar bordas, restaurar, entre outras aplicações. O funcionamento dos filtros consiste em aplicação de um operador predefinido na vizinhança de um pixel, então esta filtragem criará um novo valor para ele. Caso esta operação seja aplicada de maneira linear então temos uma filtragem espacial linear, caso não, teremos uma filtragem espacial não linear.

2.8 Filtro Gaussiano

Os filtros espaciais Gaussianos (GONZALEZ; WOODS, 2011) são usados para borramento da imagem, causando assim um remoção de ruído. Os ruídos são pequenos detalhes da imagem que não são necessários. O filtro Gaussiano é um filtro passa-baixa, pois permite somente passagem de baixas frequência, atenuando assim as altas frequências. Este é um filtro utilizado no tratamento de imagens, como podemos observar em Yeo, Lee, Lim (YEO; LEE; LIM, 2015), Neiva e Zanchettin (NEIVA; ZAN-

CHETTIN, 2016).

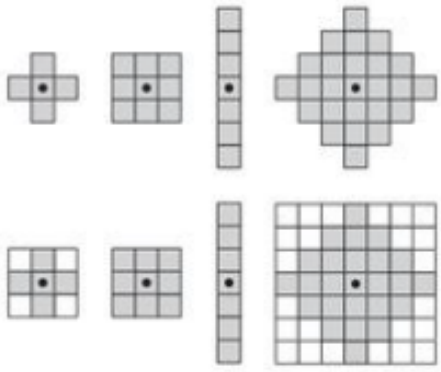
2.9 Filtros Morfológico

Para explicar o que são filtros morfológicos, devemos entender o que é morfologia. Morfologia refere-se na biologia como o ramo que estuda as formas e estrutura dos animais e plantas. Porém, no contexto de processamento de imagens, a morfologia passa a ter um contexto matemático, sendo assim, vemos que podemos defini-la como conjuntos matemáticos, e estes conjuntos são a representação de objetos em imagens digitais, conforme dito por Gonzalez e Woods (GONZALEZ; WOODS, 2011). Desta maneira vemos que em imagens os filtros morfológicos servem para ajudar a extração ou remoção de componentes necessários ou desnecessários numa imagem, respectivamente (GONZALEZ; WOODS, 2011).

Existem alguns tipos de operações morfológicas: erosão e dilatação. Erosão é definida por uma operação em conjuntos numéricos, no caso, como visto anteriormente, imagens podem ser compreendidas como conjuntos, então ao consideramos A como imagem e B como um elemento estruturante, teremos a Equação 2.1.

Elemento estruturante são pequenos conjuntos que são utilizados para analisar uma imagem em busca de características. Ao utilizarmos elementos estruturantes em imagens, eles devem ser um conjuntos matriciais regulares, podemos visualizar o que são elementos estruturantes observando a Figura 11 e conforme definido por Gonzalez e Woods (GONZALEZ; WOODS, 2011) os blocos em cinza presentes na Figura 11 são regiões de interesse que serão utilizados para aplicar alguma operação morfológica como a Erosão.

Figura 11 – Exemplo de elementos estruturantes. Os blocos em cinza representam as regiões de interesse.



Fonte: (GONZALEZ; WOODS, 2011)

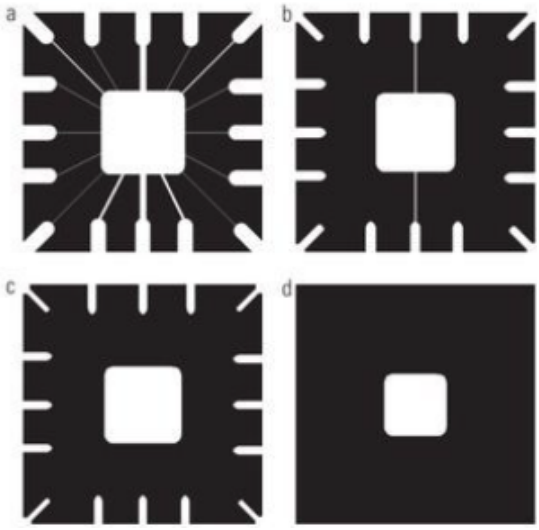
Desta maneira percebe-se que elementos estruturantes são conjuntos matemáticos e imagens digitais também, logo podemos aplicar operações de conjuntos nas

imagens, conforme a Equação 2.1.

$$A \ominus B = \{z | (B)\, \boldsymbol{z} \subseteq A\} \tag{2.1}$$

Assim, nas palavras de Gonzalez e Woods (GONZALEZ; WOODS, 2011), erosão de A por B é o conjunto de todos os pontos z de forma que B, transladado por z, está contido em A, a próxima imagem representa o efeito da aplicação de erosão em imagens.

Figura 12 – Exemplo da aplicação do filtro morfológico erosão, (a) imagem original com elementos desnecessários, (b) a (c) sucessivas aplicações de erosão, chegando ao resultado final em (d)



Fonte: (GONZALEZ; WOODS, 2011)

A aplicação deste filtro morfológico pode remover ruídos na imagem, deixando como resultado uma imagem mais simples de ser trabalhada. Como pode ser observado na Figura 12.

Dilatação, assim como a erosão, é definida também por uma operação em conjuntos numéricos, porém tem diferente aplicação da erosão. Esta operação tem como utilidade fazer a conexão de elementos desconexos, ou seja, preencher lacunas existentes na imagem (GONZALEZ; WOODS, 2011). A dilatação pode ser entendida na Equação 2.2, considerando A e B novamente como imagem e elemento estruturante, respectivamente.

$$A \oplus B = \{z | (B)\, \boldsymbol{z} \cap A \neq \varnothing\} \tag{2.2}$$

2.10 Segmentação

A segmentação é a ação de isolar regiões de pixeis (SCURI, 1999). Essas regiões isoladas podem possuir características em comum, assim destacam-se regiões que se tem interesse em trabalhar. Nesta seção serão descritos dois tipos de segmentação, por tons de pele, e por movimento.

2.10.1 Segmentação por tons de pele

A segmentação por tons de pele tem como objetivo separar da imagem as regiões que possuem pixels presentes em um intervalo que define os valores para tom de pele. Todos os valores de intensidade de pixeis que estiverem fora do intervalo determinado para a pele será definido como zero.

Este método de segmentação possui desafios como variação de iluminação, características das câmeras, etnia, características individuais e outros fatores (KAKUMANU; MAKROGIANNIS; BOURBAKIS, 2007).

A variação de iluminação é um dos pontos mais críticos, qualquer alteração na iluminação da pele afetará como a cor será recebida e isto resultará em mudança no tom da pele. Câmeras possuem características que resultam diferentes resultados sob mesmas condições de ambiente, iluminação. Os seres humanos possuem uma variedade grande de tons de pele, isto é, devido à variedade de etnias que temos, como os negros, asiáticos ou brancos. A pele pode assim variar de tons mais escuros até tons mais claros como amarelo. Segundo (KAKUMANU; MAKROGIANNIS; BOURBAKIS, 2007) as características individuais, como idade, sexo e partes do corpo também podem afetar a segmentação. E os outros fatores como *background*, sombras, movimento, maquiagens também afetarão a segmentação.

A segmentação por tons de pele pode ser conseguida por um tipo especial de segmentação, chamado de limiarização.

2.10.2 Segmentação de movimento

Uma imagem é composta de *foreground* e *background*. O objeto de interesse na imagem é chamado de *foreground* e o fundo desta imagem, é o *background*. Partindo deste princípio, o que deseja-se obter com o uso desta técnica é o *foreground*, eliminando o fundo. Usualmente, técnicas de segmentação de movimento são utilizadas com câmeras estáticas, pois assim, o fundo não tem movimento. Dessa forma, quando o objeto se movimenta pelo vídeo, produz um rastro na sequência dos *frames*, permitindo que seja isolado do fundo.

2.10.2.1 Diferença de *Frames*

Este é um método baseado na diferenciação de *frames* (OJHA; SAKHARE, 2015). Com o resultado da diferença entre os *frames*, é possível fazer a segmentação do movimento. Ainda, segundo Ojha e Sakhare (OJHA; SAKHARE, 2015), este método é altamente adaptativo para ambientes dinâmicos e é computacionalmente econômico, porém como característica, ele geralmente não consegue extrair formas completas de certos tipos de objetos em movimento.

2.11 Limiarização

O trabalho de Gonzales e Woods (GONZALEZ; WOODS,) demonstra que a limiarização é uma técnica utilizada quando queremos segmentar objetos do *background*, desta maneira será necessário visualizar o histograma de uma imagem, e então definir uma "linha" de corte, chamada de limiar. Este limiar, se definido corretamente, dividirá o histograma em duas partes, uma que será o objeto desejado e outra que será o *background*, e por meio da Equação 2.3 definirá os valores dos pixels maiores que o limiar para um e os menores para zero. O resultado obtido da aplicação de uma limiarização é uma imagem binaria, em preto e branco.

Para a Equação 2.3 considere $f(x, y)$ com uma imagem e "T" como o limiar aplicado a esta imagem.

$$g(x,y) = \begin{cases} 1 & se\ f(x,y) > T \\ 0 & se\ f(x,y) \leq T \end{cases} \tag{2.3}$$

A limiarização pode ser aplicada de maneira local ou global em uma imagem, neste projeto, usaremos somente a técnica global, no qual é definida por aplicar o limiar em toda a imagem.

A limiarização possui algumas vantagens como velocidade computacional e pode ser facilmente implementada. Existem diversas técnicas para diversos problemas diferentes, cada uma possui uma limitação, característica e desempenho distintos.

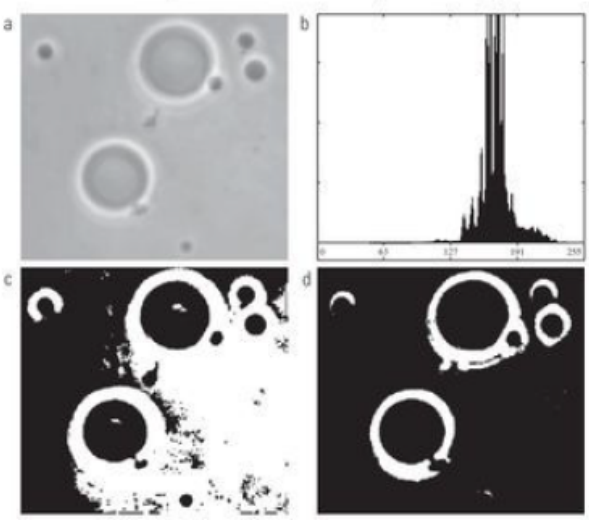
Um desses métodos é *método de Otsu* (OTSU, 1979), este método possui uma definição complexa, porem conforme os autores Gonzales e Woods (GONZALEZ; WOODS,) sumarizaram a definição nas palavras a baixo.

> O método é ótimo no sentido de que maximiza a *variância entre classes*, uma medida bem conhecida utilizada na análise estatística discriminante. A ideia básica é que as classes como limiares bem estabelecidos devem ser distintas em relação aos valores de intensidade de seus pixels e, inversa-

> mente, que um limiar que oferece a melhor separação entre as classes em temos de valores de intensidade seria melhor limiar (limiar ótimo). Além do componente ótimo, o método de Otsu tem a importante peculiaridade de se basear inteiramente em cálculos realizados no histograma de uma imagem, um arranjo 1-D obtido facilmente. (GONZALEZ; WOODS,).

A Figura 13 representa a aplicação de uma limiarização global tradicional e Otsu, conforme podemos observar o método de Otsu apresenta melhores resultados do que a maneira tradicional.

Figura 13 – (a) imagem original, em tons de cinza, (b) histograma da imagem "a", (c) aplicação do método tradicional de limiarização global e (d) aplicação do método de Ostu para limiarização global.



Fonte: (GONZALEZ; WOODS,)

3 Trabalhos Relacionados

Para a elaboração deste trabalho, foi feita uma análise bibliográfica de alguns trabalhos que abordam diferentes técnicas necessárias ao desenvolvimento do tema. Algumas das técnicas que serão utilizadas para a elaboração deste projeto, é a segmentação por tons de pele e detecção de movimento.

O trabalho de Kakumanu *et al.* (KAKUMANU; MAKROGIANNIS; BOURBAKIS, 2007) traz uma pesquisa relacionando diversos espaços de cores, como o RGB, $CIE-XYZ$, HSV, YC_bC_r, entre outros, para métodos de detecção. Aplicações como detecção e rastreamento de faces, rastreamento de mãos, controle de interfaces computacionais, e tudo que necessite detecção de cores tem como base, algum espaço de cor, e a escolha do espaço de cores não é uma atividade trivial.

Este projeto necessita da escolha correta de um espaço de cor que seja robusto para detecção de tons de pele. Um bom sistema de cores é um que seja eficiente em diferentes lugares, que detecte diferentes tons de pele, sendo ela da pele mais clara a mais escura e que não sofra muito com diferença de iluminação. O artigo, faz uma relação dos espaços de cores com vários classificadores, então, nele estarão presentes tabelas relacionando cada espaço de cor com diferentes classificadores e em diferentes situações de iluminação, plano de fundo e pele. Este trabalho apresenta uma revisão sobre o tema, mas não apresenta uma conclusão sobre qual melhor espaço de cor a ser utilizado, a escolha dependerá dos requisitos da aplicação, por exemplo espaço de cores lineares poderão afetar desempenho em sistemas em tempo real.

Basilio *et al.* (BASILIO et al., 2011) utiliza somente o sistema YC_bC_r para detectar tons de pele, tendo seu trabalho atingido 88,8% de acerto e cerca de 5% de falsos positivos. Em sua pesquisa, o autor tem como objetivo detectar imagens de teor sexual a partir da porcentagem de pele detectada na imagem. O autor em sua pesquisa também define valores para limiares.

No trabalho de Khamar Basha Shaik *et al.* (SHAIK et al., 2015) também são encontradas informações importantes sobre espaço de cores. Este trabalho faz um estudo comparativo entre dois sistemas de cores, o HSV e o YC_bC_r. Eles não utilizaram o sistema de cor RGB, pois este não possui muitas informações acerca da do tom das cores (crominância) e intensidade, então ele não é o preferido para realizar detecções baseadas em cores (SHAIK et al., 2015). O autor conclui que a transformação de RGB para HSV não é linear, sendo assim, nada eficiente, o que não seria ideal para sistemas de tempo real. Outro fator observado por ele, é que o HSV funciona melhor com imagens de fundo simples. Entretanto a transformação de RGB para YC_bC_r é eficiente,

pois é linear, logo poderia ser aplicado a sistemas de tempo real, neste a extração de informações de iluminação se torna mais fácil, logo, poderá ser aplicado a imagens com fundos complexos.

Dadgostar e Sarrafzadeh (DADGOSTAR; SARRAFZADEH, 2006) trazem uma abordagem para detecção de tons de pele em tempo real, baseado no limiar da Matiz. Esse trabalho realiza a detecção de tons de pele em ambiente de condições normais, com variação de fundo e mudança na iluminação. O tom da pele é considerado uma característica estática, porém em um ambiente real ela não se torna uma boa característica para ser utilizada, pois será afetada pela mudança de iluminação e *background*, reduzindo a robustez da detecção.

O artigo utiliza-se de características dinâmicas da cor da pele, propondo um modelo de detecção de tons de pele adaptativo, baseado no limiar da Matiz. Já na avaliação, foram utilizadas duas técnicas de detecção de movimento. Também foi realizado uma comparação da técnica apresentada na pesquisa com o modelo tradicional utilizando como característica o tom de pele estático. A utilização da abordagem adaptativa não afetou o desempenho do algoritmo em relação a abordagem estática. A classificação foi realizada utilizando-se de duas abordagens distintas de detecção de movimento, *frame-subtraction* e *optical-flow*. Segundo os autores a melhor abordagem a ser seguida é a de subtração de *frames*, causando um aumento médio de 20% nas detecções corretas. Este artigo apresenta ideias interessantes, as quais possibilitam o uso de uma abordagem hibrida, utilizando a detecção de tons de pele adaptativa e detecção de movimento em sistema de tempo real. O artigo realiza essa estratégia para reconhecimento de faces e neste trabalho iremos utilizar uma abordagem híbrida para o reconhecimento da mão.

Continuando no aspecto da detecção da pele, o trabalho de Khan, Hanbury e Stoettinger (KHAN; HANBURY; STOETTINGER, 2010) discorre sobre um classificador *Random Forrest* para a detecção dos tons de pele. Foi escolhido este algoritmo pela sua capacidade de generalização e por ser de rápido treinamento, ele foi avaliado e comparado com diversos outros algoritmos como *Bayesian Network*, *Multilayer Perceptron*, SVM, *AdaBoost*, *Naive Bayes* e RBF. O autor obteve melhor velocidade de treinamento comparado a os outros algoritmos. O melhor resultado utilizando *F-score* foi obtido Com 10 árvores. O experimento foi realizado em 25 vídeos, neles é possível perceber a presença de múltiplas pessoas, em ambientes internos e externos, com câmeras estáticas e móveis, sob diferentes tipos de iluminação. Os autores utilizam um outro sistema de cor, o $IHLS$, no qual é dito ser o melhor para este tipo de detecção em relação aos modelos angulares como HLS, HSV, HSI, entre outros. Para a classificação foi realizado validação cruzada com 10 *folds* e em todos os cenários o *Random Forrest* foi melhor.

O trabalho de Liu *et al.* (LIU et al., 2017) trata de um sistema em tempo real que realiza rastreamento de mãos. O autor utiliza técnicas que extraem características de movimento e segmentação de cor, que serão utilizadas no algoritmo chamado *Feature Fusion Hand Tracker* que pode ser entendida como uma melhoria do algoritmo de rastreamento *mean shift tracking algorithm*. Sendo este último, um algoritmo de rastreamento muito utilizado devido a sua simplicidade e eficiência, entretanto esta abordagem utiliza somente cores para rastrear o alvo, e o estudo Liu *et al.* (LIU et al., 2017) utiliza a técnica de fusão de características para modificar o *mean shift tracking algorithm*, que passa a considerar também movimento como característica para rastrear o alvo. Esta melhoria faz com que seja possível rastrear objetos que tenham uma cor parecida com o fundo.

O trabalho de Neiva e Zanchettin (NEIVA; ZANCHETTIN, 2016) propõe um sistema para reconhecimento de gestos dinâmicos, com a finalidade de traduzir linguagens de sinais em *backgrounds* complexos. O sistema é uma aplicação móvel com *web server*, que tem módulos para adicionar novas imagens com gestos, realizar a remoção de *background*, módulos de treinamentos e classificação. Eles obtiveram cerca de 84% de acerto dos gestos em imagens de *background* simples, e para os complexos a taxa baixa para 58%. Essa diferença demonstra como o *background* influencia na classificação dos gestos, logo, uma boa segmentação da mão poderá melhorar bastante a taxa de acerto.

Como diferencial ao trabalho anteriormente citado, nosso projeto de pesquisa visa focar no reconhecimento das mãos e futuramente trabalhar na classificação de gestos estáticos, *frame* a *frame*. Ao contrário do trabalho de Neiva e Zanchettin (NEIVA; ZANCHETTIN, 2016) que foca em gestos dinâmicos.

O trabalho de Wu-Chih Hu *et al.* (HU et al., 2015) é resultado de uma extensão de um trabalho anterior dos autores, o método proposto neste trabalho utiliza como método para o rastreamento, o detector de características *Harris Corner* modificado, e então, classifica os pontos detectados como *foreground* ou *background*. As regiões de *foreground* ou de primeiro plano, são obtidas a partir da diferenciação entre o *frame* atual e o anterior, no qual foi realizado uma transformação de perspectiva utilizando uma matriz de homografia[1]. Matriz esta, que mapeia os pontos em comum de diferentes imagens, neste caso diferentes *frames*. Um fator interessante deste trabalho é o fato da abordagem anteriormente citada poder ser utilizada em imagens captadas por câmeras em movimento. O rastreamento do objeto é realizado utilizando o algoritmo Filtro de Kalman baseado no centro dos objetos em movimento.

O trabalho de Ojha e Sakhare (OJHA; SAKHARE, 2015) traz diversas aborda-

[1] https://www.ic.unicamp.br/ rocha/teaching/2012s1/mc949/aulas/additional-material-revision-of-concepts-homography-and-related-topics.pdf

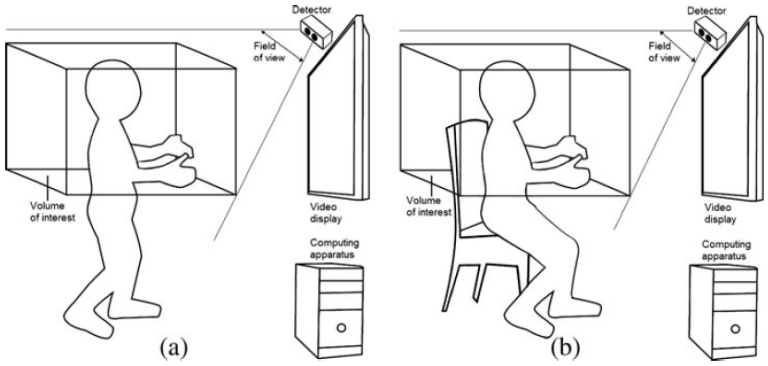
gens de como rastrear objetos em vídeos de segurança. Uma das abordagens é o rastreamento por movimento, ele exibe os pontos positivos e negativos de cada método. Focando na segmentação de movimento, o trabalho cita três diferentes técnicas. Subtração de *background*, diferenciação temporal e fluxo ótico. A subtração de *background* é o método mais utilizado para detecção de objetos com câmeras estáticas, é realizado uma diferenciação dos *frames* com o *frame* detectado como fundo e então o objeto móvel poderá ser detectado. A técnica de diferenciação temporal também trabalha com a diferenciação dos *frames*, esta técnica é mais utilizada em ambientes que sejam bastante dinâmicos, pois ela é altamente adaptativa, porém, não faz um bom trabalho caso seja necessário extrair as formas completas dos objetos em movimento. Por último, temos a técnica de fluxo ótico, que é um método baseado em vetores. Com ela, é feita a previsão do movimento considerando pontos encontrados em múltiplos *frames*. O uso desta técnica permite que a detecção aconteça em quais quer movimentos que ocorram nas imagens de vídeo. O artigo traz uma abordagem de detecção de objeto com auto adaptação dos limiares.

A pesquisa de Song *et al.* (SONG et al., 2017) traz uma abordagem híbrida para detecção da pele como região de interesse baseada em movimento. No projeto é utilizado uma detecção em tempo real empregando algoritmos de componentes conectados, que rodam em paralelo, e utilizam uma abordagem com limiar adaptativo para detecção dos tons de pele. A segmentação da região de interesse (ROI) foi definida no espaço de cor RGB, utilizando a distribuição do histograma e um limiar especifico para tons de pele, então foram detectados os componentes conectados, sendo assim definido o ROI. A detecção do movimento é feita por diferenciação de *frames*. O projeto de Song *et al.* (SONG et al., 2017) utiliza uma abordagem baseada em GPU para realizar a etapa de *clustering* em tempo real. Os resultados obtidos pela pesquisa do autor concluem que os métodos tradicionais como os de detecção de cores ou por limar da Matiz são computados eficientemente, porem são afetadas por mudanças no ambiente e *background*. O método baseado em GPU proposto, consegue trabalhar com algoritmos confiáveis, mas que pecam na eficiência. Os algoritmos se comportam de forma mais lenta, visto que são realizados mais cálculos, ou cálculos mais complexos, consequentemente apresentando resultados mais confiáveis. O respectivo método detecta diferentes cores da pele em tempo real, a média de *frames* por segundo obtida pelo sistema, no hardware exibido, foi de 18,40 FPS.

O trabalho de Yeo *et al.* (YEO; LEE; LIM, 2015) traz uma abordagem interessante para esta pesquisa pois ele propõe utilizar *hardware* de baixo custo para realizar o rastreamento da mão e o reconhecimento dos gestos. Para realização da sua pesquisa, foi criado um ambiente composto de três módulos que podemos observa-lo na Figura 14, sendo eles, módulo da câmera, responsável por capturar as imagens no qual serão aplicados técnicas de processamento de imagens, ele tem como objetivo

realizar a detecção da mão, o modulo de detecção é responsável por rastrear a mão e os dedos, ele tem como resultado a localização 2D dos objetos detectados, para isto ele usou uma máquina de estados finitos e o filtro Kalman, responsáveis por melhorar a precisão e o rastreamento, por fim é utilizado o módulo de interface, sendo este o responsável por traduzir os resultados obtidos dos módulos anteriores e mapeados em uma aplicação, permitindo assim o seu controle.

Figura 14 – Ambiente utilizado por Yeo *et al.* (YEO; LEE; LIM, 2015)



Fonte: (YEO; LEE; LIM, 2015)

Um ponto importante deste trabalho de Yeo *et al.* (YEO; LEE; LIM, 2015) é limiar que foi definido para realizar a segmentação por tons de pele e o fato de ser utilizado uma câmera comum ou *Kinect* para realizar o rastreamento da mão, porém um dos pontos que difere da abordagem proposta neste trabalho de pesquisa é o fato da câmera está estática, a utilização de imagens com movimento de *background* poderá afetar o resultado apresentado em sua pesquisa.

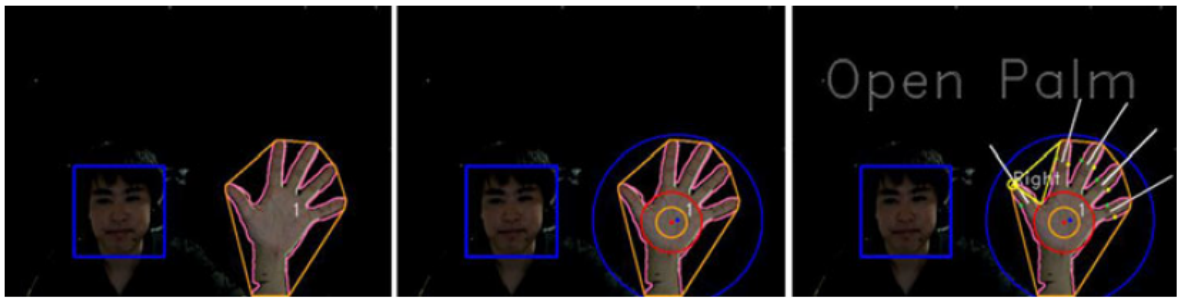
As Figuras 15, 16 representam alguns resultados obtidos por Yeo *et al.* (YEO; LEE; LIM, 2015).

Figura 15 – Resultados apresentados por Yeo *et al.* (YEO; LEE; LIM, 2015)

Fonte: (YEO; LEE; LIM, 2015)

A outra abordagem também vista foi a de Thabet *et al.* (THABET et al., 2017). Eles propõem a utilização de uma fusão de características e a utilização do *Fast mar-*

Figura 16 – Resultados apresentados por Yeo *et al.* (YEO; LEE; LIM, 2015)

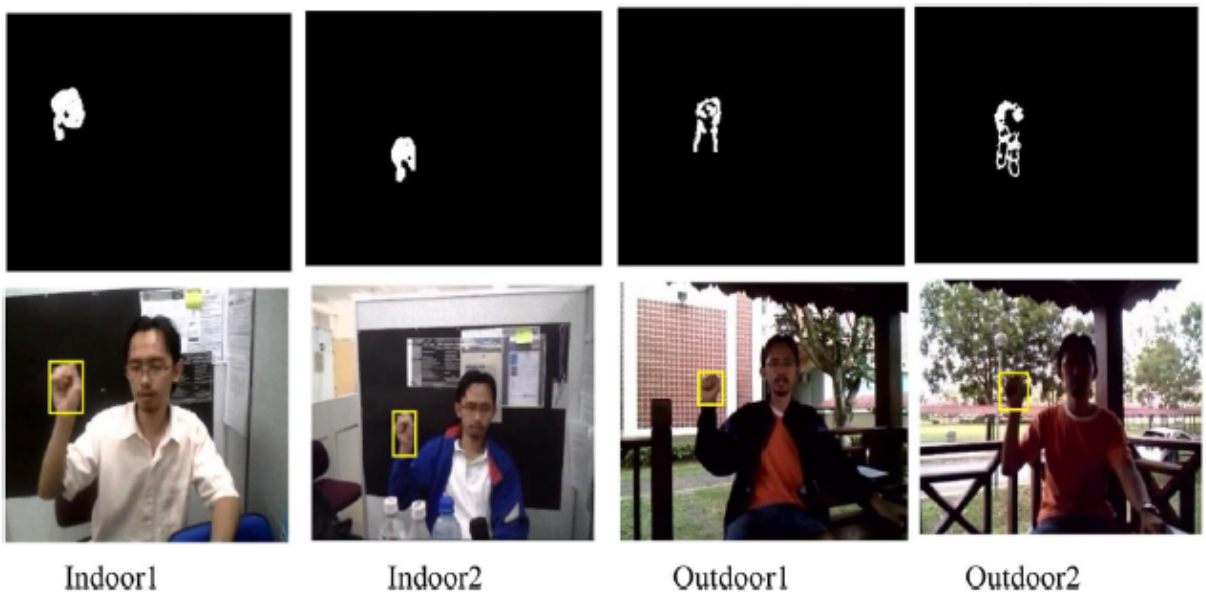


Fonte: (YEO; LEE; LIM, 2015)

ching method, para realizar segmentação e detecção da mão em *background* complexo. Este é um trabalho recente e utiliza abordagens que são pouco comuns, quando comparado a os outras pesquisas aqui relacionadas. O método *Fast marching method* realiza o a detecção de tons de pele, segmentação de contornos e segmentação por movimento. Outro ponto importante em seu trabalho foi a utilização de um limiar adaptativo, obtido da segmentação da face e então feito um treinamento para definir os melhores valores de limiar a serem utilizados.

Esta abordagem também possui a mesma limitação da abordagem de Yeo *et al.* (YEO; LEE; LIM, 2015), pode não ser possível a utilização deste método em uma abordagem com movimentação de *background* pois, foi realizado toda a experimentação com vídeos estáticos. A Figura 17 exibirá resultados apresentados por Thabet *et al.* (THABET et al., 2017) em sua pesquisa em ambientes externos e internos.

Figura 17 – Resultados apresentados por Thabet *et al.* (THABET et al., 2017)



Fonte: (THABET et al., 2017)

A escolha do espaço de cor YC_bC_r foi definida pela quantidade de trabalhos observados que utilizam este espaço de cor para detecção de tons de pele, e um trabalho como o de Basilio *et al.* (BASILIO et al., 2011) que foca em realizar detecção de tons de pele foi importante para esta pesquisa, outro fator é que o YC_bC_r é linear, sendo possível sua utilização em sistemas em tempo real.

No trabalho de Wu-Chih Hu *et al.* (HU et al., 2015) pode ser observado a utilização do método de rastreamento modificado *Harris Corner* que possibilita a utilização de uma câmera móvel para rastrear objetos, mas deverá ser considerado o custo computacional deste método para execução em tempo real. O trabalho de Song *et al.* (SONG et al., 2017) é um exemplo de técnicas que são computacionalmente custosas e não poderão ser utilizados em hardware de baixo poder computacional, porem esta técnica é muito robusta, talvez com a evolução do hardware esta técnica torne-se mais comum.

O trabalho de Khan e Borji (KHAN; BORJI, 2018) utiliza vídeos com visão egocêntrica, no qual a câmera está gravando em primeira pessoa. Este trabalho trata de pontos diferentes de reconhecimento de mãos, utilizando a técnica de aprendizagem profunda para realizar a segmentação das mãos. Khan e Borji (KHAN; BORJI, 2018) utilizam RefineNet que é uma abordagem de inteligência artificial no qual é aplicada uma rede de refinamento multi-caminho para segmentação semântica de alta resolução. Essa é uma abordagem diferente das citadas anteriormente inclusive a proposta presente neste trabalho, pois não são utilizadas técnicas de inteligência artificial. O trabalho foi aplicado a diversos bancos de dados distintos e obteve bons resultados.

Outro trabalho que utiliza técnicas de inteligência artificial é o de Bambach *et al.* (BAMBACH et al., 2015), ele utiliza banco de dados com vídeos em primeira pessoa e *deep learning*. Como técnica principal para segmentação das mãos é utilizado Redes Neurais Convulacionais que oferecem boa performance em tarefas de classificação e neste trabalho é utilizado mais especificamente o CaffeNet. Modelo baseado em tons de pele foi escolhido para realizar o treinamento dessa rede. O espaço de cor utilizado no trabalho de Bambach *et al.* (BAMBACH et al., 2015) é o YUV.

A escolha de uma abordagem híbrida utilizando segmentação de tons de pele e movimento foi determinada por observação na literatura, ambos tipos de segmentação são utilizadas em sistemas de tempo real e isto possibilita que sejam também utilizados em hardware como celulares, podemos observar isto no trabalho de Yeo *et al.* (YEO; LEE; LIM, 2015) que não utiliza celulares, mas o custo computacional citado em seu trabalho é menor do que o poder computacional dos *smatphones* atuais.

Este trabalho visou utilizar os pontos fortes observados nos trabalhos com a finalidade que seja possível realizar um rastreamento da mão, de maneira eficaz em tempo real, e foi também estudado os pontos fracos para que fosse possível evita-

los da melhor maneira. Como contribuição para a área, este trabalho visa demostrar que é possível por meio das técnicas processamento de imagem utilizar método que sejam eficazes computacionalmente e que traga um bom reconhecimento da mão em ambientes diversos, podendo ser utilizado como pré-processamento nos sistemas de reconhecimento de gestos.

4 Detecção de Mãos Através de Detecção de Pele e Movimento

Detecção de mão é um assunto explorado no universo da visão computacional, o estudo desta área tem como intuito criar sistemas que facilitem a interação humano-computador, acessibilidade, imersão em ambientes virtuais entre outros tópicos. Rastrear a mão permite que gestos ou sinais sejam reconhecidos por sistemas e assim é possível realizar ações. Existem diversas maneiras de realizar a detecção das mãos, seja por segmentação de cores, movimentos, formas, abordagens híbridas ou usando câmeras de profundidade, cada uma dessa maneira tem particularidades e neste projeto escolhemos utilizar uma abordagem híbrida.

O principal problema que existe acerca deste assunto é o grande número de variáveis que devem ser tratadas como por exemplo, os diferentes ambientes, a mudança na iluminação, os muitos tons de pele existentes e a limitação dos hardwares, estes são os principais pontos que afetam a detecção da mão. Conseguir uma abordagem que seja genérica suficiente para abranger todos estes pontos é muito complexo e talvez atualmente ainda não seja algo viável.

Neste capítulo serão descritas algumas abordagens da literatura pra detecção de mãos e a proposta de uma abordagem de uma detecção capaz de segmentar mãos em vídeos com pouca ou nenhuma movimentação de câmera, utilizando uma abordagem híbrida, composta de segmentação de cores e movimento em tempo real.

4.1 Abordagens de Detecção de Mãos

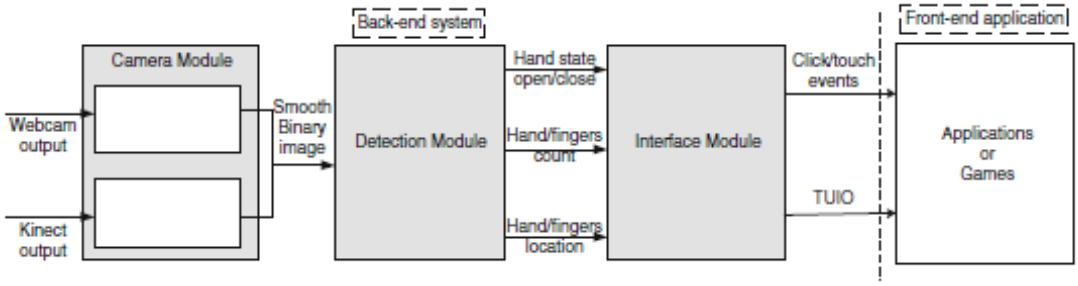
Nesta seção serão descritos duas abordagens de trabalhos da literatura que realizam a detecção de mãos. São os trabalhos *Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware* e *Fast marching method and modified features fusion in enhanced dynamic hand gesture segmentation and detection method under complicated background*, dos seguintes autores Yeo *et al.* (YEO; LEE; LIM, 2015) e Thabet *et al.* (THABET et al., 2017) respectivamente.

4.1.1 Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware

O trabalho de Yeo *et al.* (YEO; LEE; LIM, 2015), tem como proposta realizar o rastreamento da mão e o reconhecimento dos gestos utilizando *hardware* de baixo

custo. Ele possui os módulos de câmera, detecção e interface, com a seguinte arquitetura, mostrada na Figura 18.

Figura 18 – Arquitetura do trabalho de Yeo *et al.* (YEO; LEE; LIM, 2015).



Fonte: (YEO; LEE; LIM, 2015)

O módulo de câmera utiliza como entrada imagens obtidas por uma *webcam* comum. Então é utilizado uma etapa de detecção de face, esta etapa utiliza o algoritmo clássico *Haar-like feature*[1] desenvolvido por Paul Viola e Michael Jones. Após isto, os *frames* de entrada são convertidos para o espaço YC_rC_b, e o *background* é removido utilizando uma subtração entre os *frames* e uma imagem estática do ambiente. A imagem estática foi obtida capturando o primeiro *frame* do vídeo analisado, todos os vídeos presentes na base de dados é iniciada com o fundo sem objetos de interesse.

O resultado do processamento anterior será dividido três canais Y, C_r e C_b, e em cada canal foi aplicado em sequência limiarização, binarização e operadores morfológicos, estes operadores são erosão e dilatação.

Como limiar para segmentação de tons de pele utilizado neste trabalho temos os valores apresentados na Equação 4.1.

$$\begin{aligned} 54 \leq Y \leq 163 \\ 131 \leq Cr \leq 157 \\ 110 \leq Cb \leq 135 \end{aligned}$$
$$Onde\, Y, Cb, Cr = [0, 255] \quad (4.1)$$

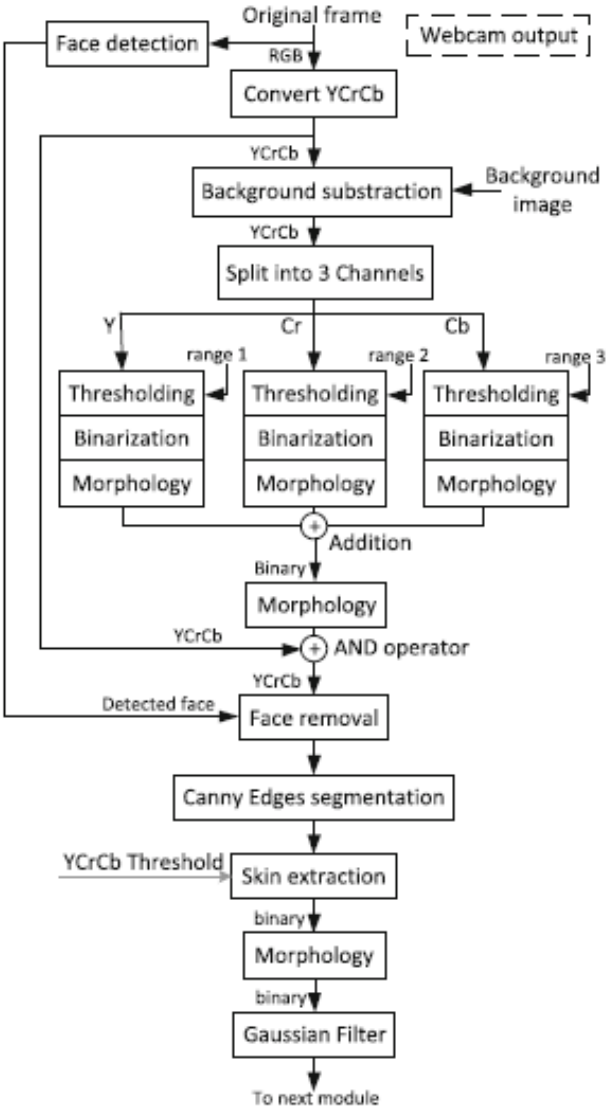
No final destes processos os canais foram mesclados, a imagem resultante é submetida a aplicação de novas operações morfológicas, e é realizado a operação logica *AND* entre este *frame* processado e o *frame* inicial convertido em YC_rC_b.

A próxima etapa consta em pegar a imagem resultante de todas as etapas anteriores e aplicar a remoção da face que foi detectada.

[1] <https://docs.opencv.org/3.3.1/d7/d8b/tutorial_py_face_detection.html>

O autor utiliza como próxima etapa a imagem resultante após a remoção da face, nesta parte do algoritmo é utilizado o detector de bordas Canny (YEO; LEE; LIM, 2015), então será aplicado limiares para segmentar a pele, seguido de mais operações morfológicas e por fim o filtro gaussiano. O algoritmo descrito pode ser visualizado no fluxograma da Figura 19.

Figura 19 – Fluxograma do modulo da câmera do trabalho de Yeo *et al.* (YEO; LEE; LIM, 2015).



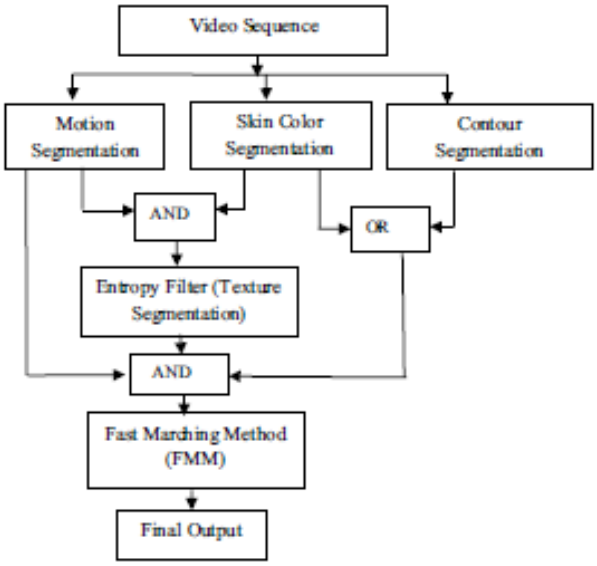
Fonte: (YEO; LEE; LIM, 2015)

Este modulo também tem como objetivo realizar o rastreamento da mãos em *background* complexo.

4.1.2 Abordagem Híbrida com *Fast Marching*

Thabet *et al.* (THABET et al., 2017) utilizam uma abordagem híbrida no qual são utilizados 3 métodos diferentes de segmentação de mãos, conforme sua arquitetura presente na Figura 20.

Figura 20 – Arquitetura do trabalho de Thabet *et al.* (THABET et al., 2017).



Fonte: (THABET et al., 2017)

O módulo de segmentação utiliza diferenciação de *frames*, binarização da imagem, filtro mediana e o método base do trabalho, o *Fast Marching*, conforme a Figura 21.

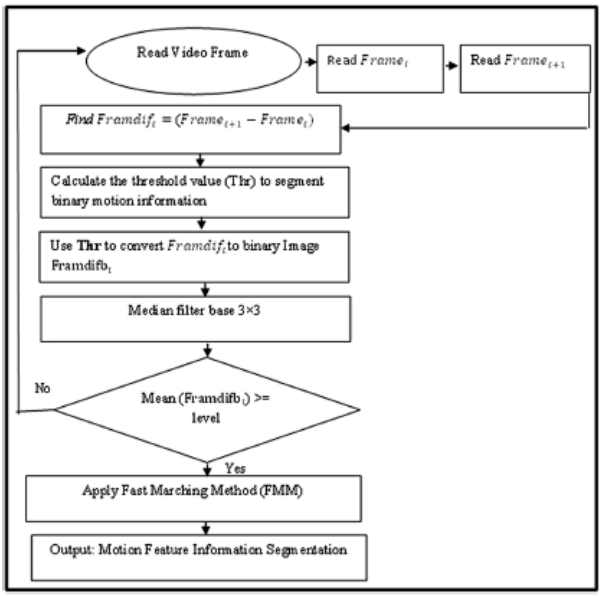
No módulo de segmentação de tons de pele, foi utilizado um detector de face, baseado em no algoritmo de Viola-Jones para obter os valores limiares ótimos de C_b e C_r para segmentação da pele, conforme o fluxograma da Figura 22.

Por fim, o módulo de segmentação de contorno, no qual Thabet *et al.* (THABET et al.) utilizam um modelo que segmenta somente contornos em comum entre dois *frames* com o algoritmo Canny. Segue a Figura 23 contendo o fluxograma utilizado pelo autor.

A técnica de *Fast Matching* é utilizada em cada etapa final dos módulos de segmentação de movimentação e tons de pele. Esta técnica tem como intuito obter a segmentação dos gestos realizado, o autor relata que esta etapa é importante para conseguir a segmentação durante a movimentação. O *Fast Matching Method* é baseado na região de crescimento e no rastreamento de fronteiras do objeto.

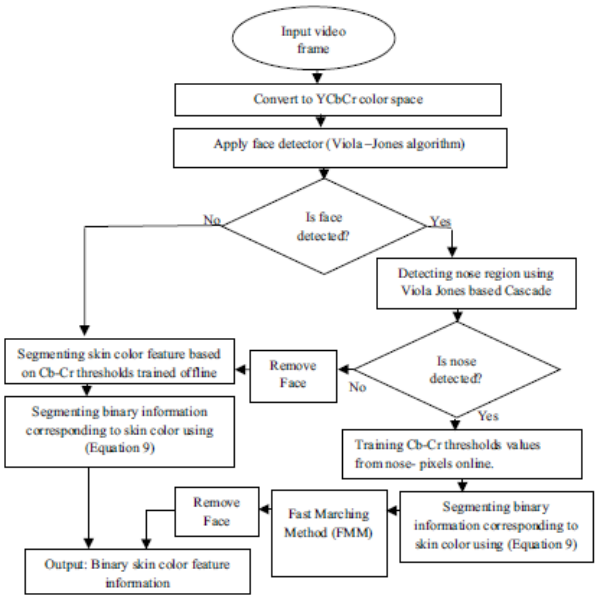
Ele funciona conforme a Equação 4.2, no qual BW se refere a uma imagem binária, W é uma matriz de peso definida pela diferença de intensidade ou gradiente

Figura 21 – Módulo de segmentação de movimento do trabalho de Thabet *et al.* (THABET et al., 2017).



Fonte: (THABET et al., 2017)

Figura 22 – Módulo de segmentação de tons de pele do trabalho de Thabet *et al.* (THABET et al., 2017).

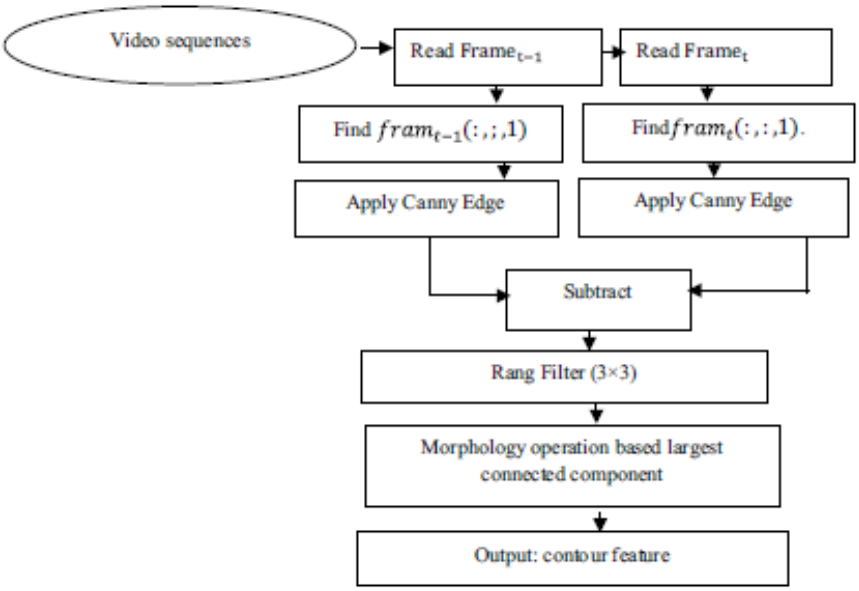


Fonte: (THABET et al., 2017)

da imagem, $MASK$ é a mascara das sementes de pixel da região de crescimento com tamanho igual a W e $THRESH$ é o limiar para obter a imagem binária.

$$BW = FMM(W, MASK, THRESH) \tag{4.2}$$

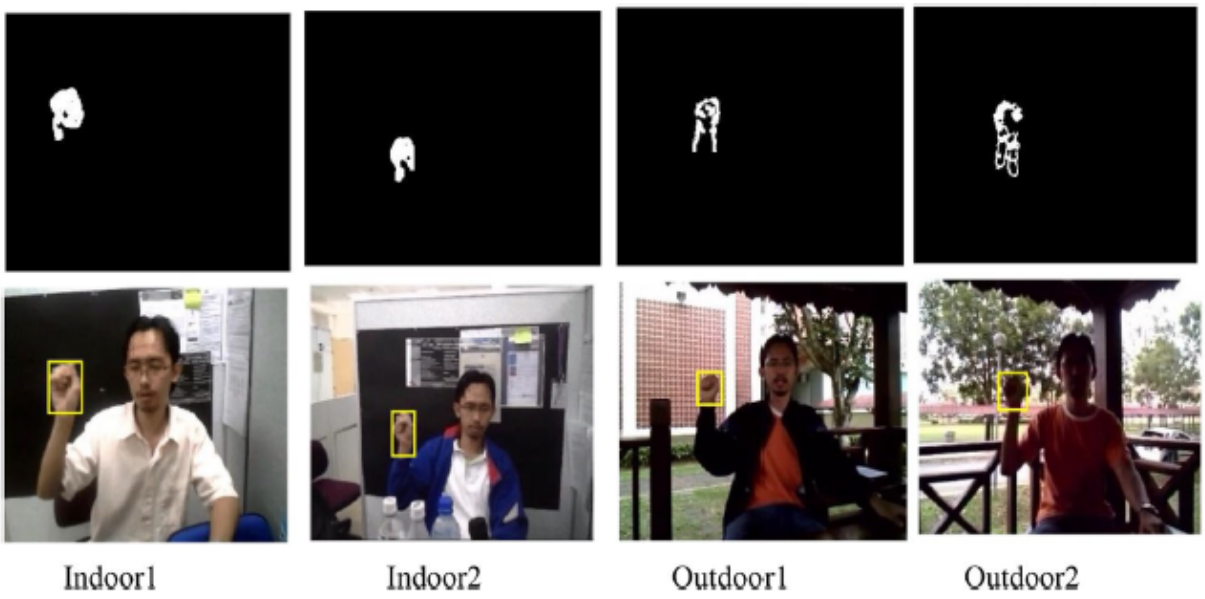
Figura 23 – Módulo de segmentação de contorno do trabalho de Thabet *et al.* (THABET et al., 2017).



Fonte: (THABET et al., 2017)

O trabalho de Thabet *et al.* (THABET et al., 2017) não utiliza imagens com *background* dinâmico, como podemos observar na Figura 24 a câmera é estática e o fundo é onde está presente a mão não possui movimentação.

Figura 24 – Resultados apresentados no trabalho de Thabet *et al.* (THABET et al., 2017).



Fonte: (THABET et al., 2017)

Nos ambientes *indoor1*, *indoor2* e *outdoor1* o fundo é considerado simples, já

o fundo do ambiente *outdoor2* é um pouco mais complexo, conforme a Figura 24.

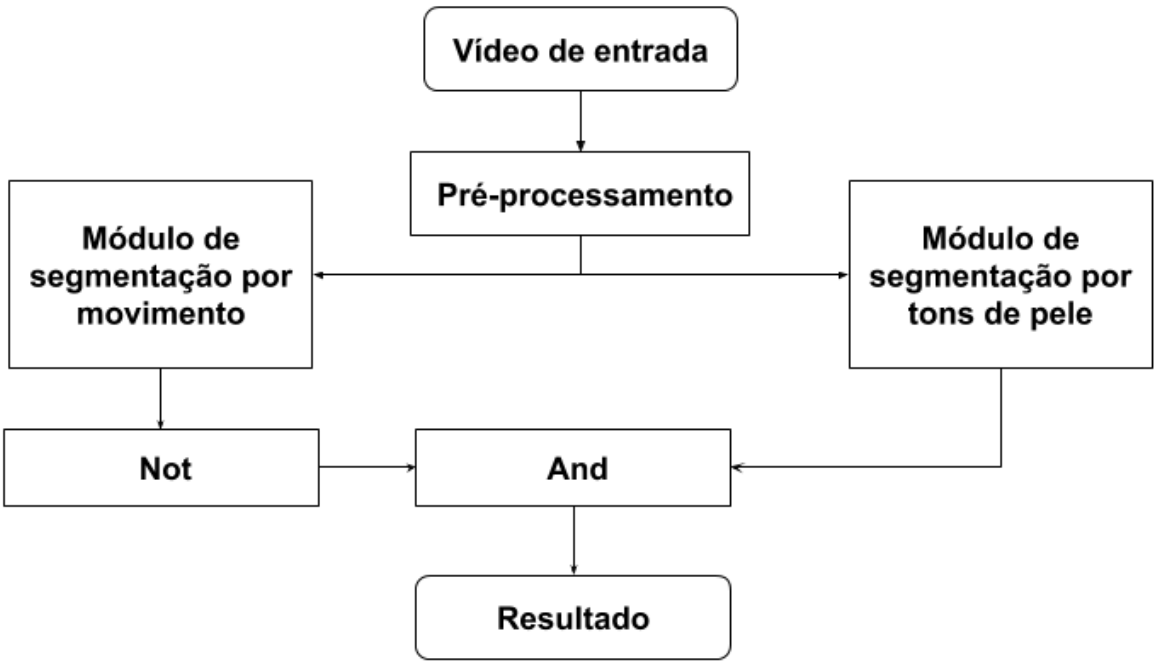
4.2 Algoritmo Proposto

O algoritmo proposto neste trabalho consta de uma abordagem híbrida com etapas de segmentação por tons de pele e segmentação de movimento.

As técnicas utilizadas para desenvolver este trabalho tem como diferencial serem abordagens simples a nível computacional, para que o algoritmo final fique com bom desempenho. Foi escolhida a utilização destas duas abordagens, pois uma ajudará a complementar a outra quando certas situações ocorrerem. Como por exemplo, imagine que um objeto do *background* tenha tom parecido com a pele, ele será removido pela segmentação de movimentos, mas caso não seja possível identificar tons de pele na imagem, o modulo de movimento ajudará a determinar qual será a região de interesse. Deste modo vemos como os módulos podem ser complementar, aumentando assim a robustez desta abordagem.

A arquitetura geral é definida por dois módulos, um para segmentação de cor e outro para movimento, conforme a Figura 25.

Figura 25 – Arquitetura geral do algoritmo proposto.



Fonte: O autor

Na primeira etapa, o pré-processamento consiste em aplicação de métodos que serão comuns tanto no módulo de segmentação de tons de pele, como no módulo de

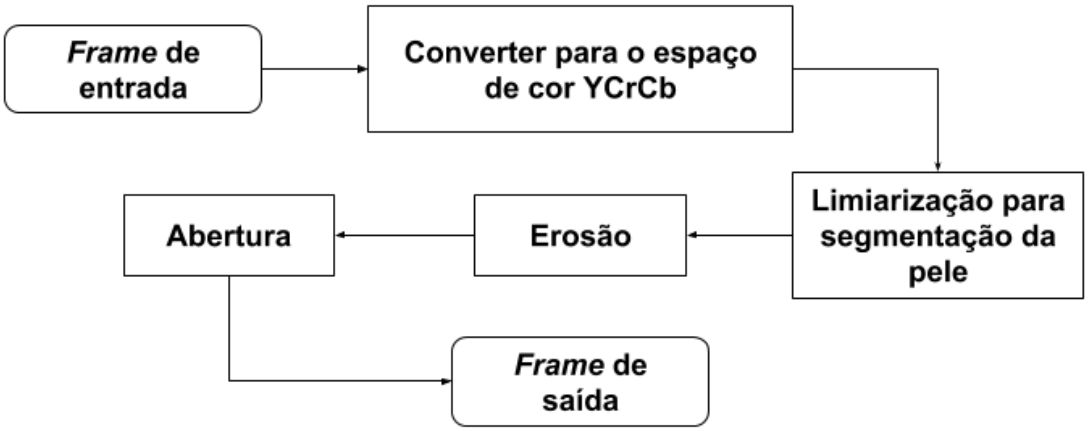
segmentação por movimentos. No pré-processamento é feito a separação dos *frames* em atual e anterior, então ambos serão convertidos para tons de cinza.

O módulo de segmentação por tons de pele começa convertendo os *frames* para o canal YC_rC_b, vide Figura 27a, a sua arquitetura pode ser observada na Figura 26, e os resultados parciais na Figura 27. Em seguida, é aplicado o método de limiarização para os canais C_r e C_b, a imagem resultante está na Figura 27b, com os valores máximos e mínimos para detecção dos tons da pele, os valores foram baseados no trabalho de Basilio *et al.* (BASILIO et al., 2011), conforme a Equação 4.3:

$$133 \leq Cr \leq 177 \ \ e \ \ 77 \leq Cb \leq 121 \tag{4.3}$$

Após essas etapas, aplica-se os filtros morfológicos de erosão, Figura 27c e abertura, Imagem 27d para reduzir a quantidade de ruídos na imagem resultante.

Figura 26 – Módulo de segmentação por tons de pele.

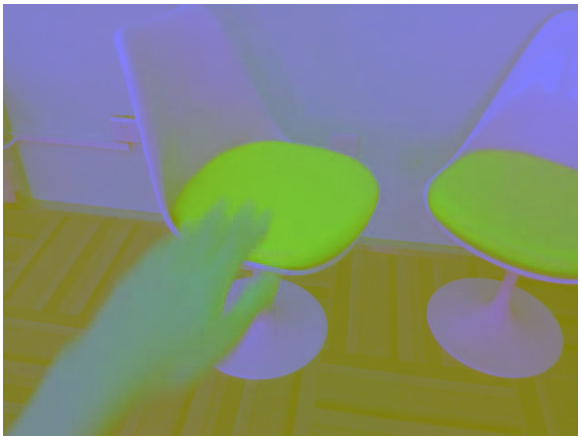


Fonte: O autor

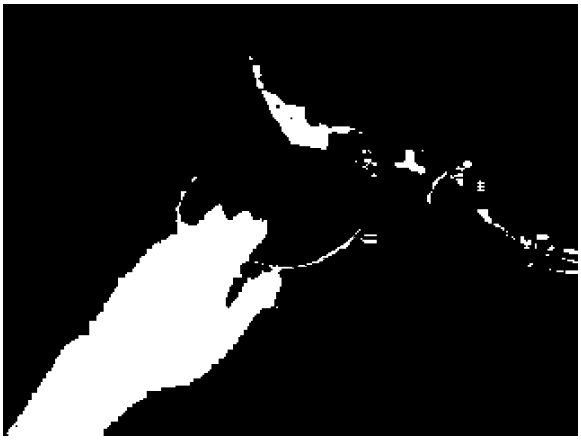
O módulo de segmentação por movimento é computacionalmente simples, sua arquitetura pode ser conferida na Figura 28 e os resultados parciais na Figura 29. Nessa etapa, é possível determinar quais objetos estão se movendo na imagem, e aplicado em conjunto com a segmentação por tons de pele, é possível determinar objetos que se movam e que tenham tons de pele, excluindo assim ruídos que tenham tom de pele, mas sejam estáticos. A saída do módulo de movimento é negativado para que seja obtido a região de interesse, no caso, a mão.

Este módulo é iniciado recebendo os *frames* já convertidos para tons de cinza e então é aplicado um filtro gaussiano. Depois é empregado o uso da diferenciação de *frames*, vide Figura 29a e, em seguida, a imagem resultante é binarizada com a limiarização global e Otsu combinadas, Figura 29b. Então, é aplicado uma erosão com o mesmo elemento estruturante do módulo anterior, Figura 29c. A diferenciação de

Figura 27 – Resultados parciais da segmentação de tons de pele.

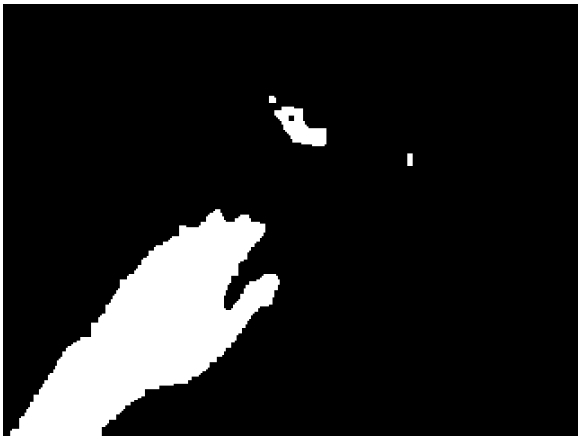


(a) Imagem resultante pós conversão para o YC_rC_b.



(b) Imagem resultante pós aplicação do limiar de tons de pele.

(c) Imagem resultante pós erosão.



(d) Imagem resultante pós abertura.

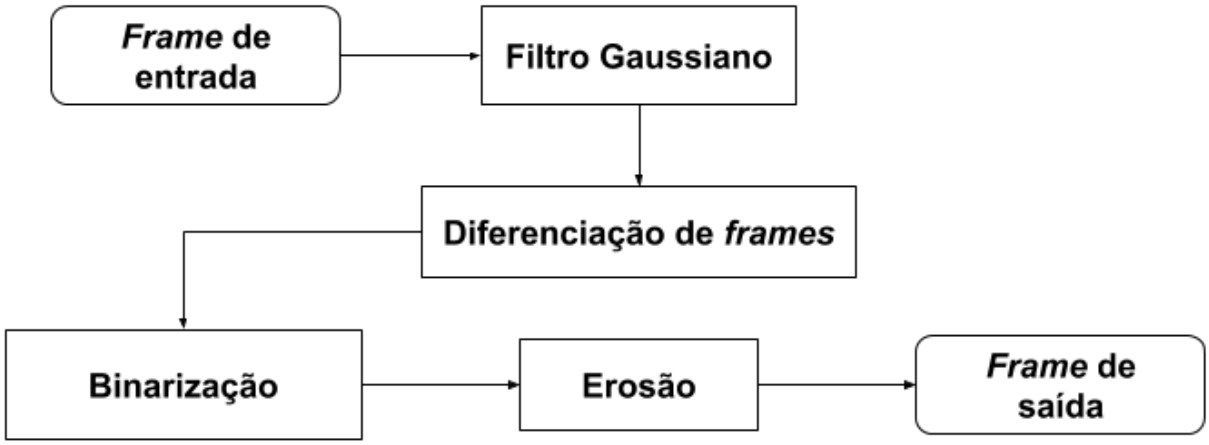
Fonte: o autor

frames tem como entrada dois *frames*, um atual e um anterior, o resultado dela é a exibição de uma imagem contendo *frames* do *frame* atual que não está presente no *frame* anterior. A imagem resultante negativada pode ser observada na Figura 29d

Conforme informado na arquitetura do algoritmo proposto, na Figura 25, vemos que são aplicados operadores lógicos após os módulos, estes operadores tem como finalidade tentar garantir que o algoritmo tenha um robustez e adaptabilidade, pois a movimentação do *background* pode causar um grande número de ruídos. Neste caso, robustez é a capacidade de conseguir segmentar a mão corretamente em ambientes de *background* complexo, e à adaptabilidade seria a capacidade de ajustar a captura da mão.

O algoritmo proposto possui capacidade de processamento em tempo real, devido a simplicidade das técnicas de processamento de imagens utilizadas. Um dos

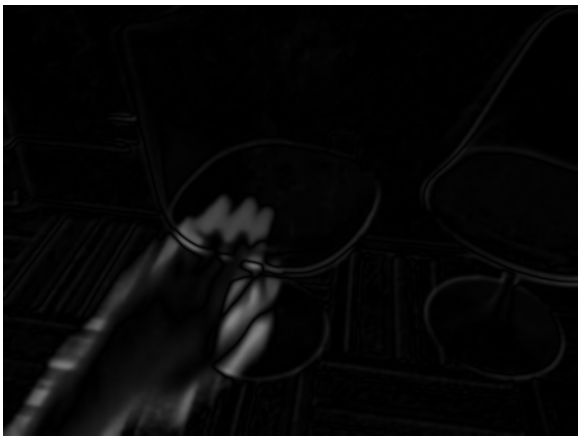
Figura 28 – Módulo de segmentação de movimento.



Fonte: O autor

pontos de destaque do módulo de segmentação por tons de pele é o fato dele conseguir reduzir a variação da iluminação, e esta característica torna esta segmentação mais robusta. O módulo de segmentação de movimento traz a este projeto um ponto importante que é remover os elementos do *background* e que possam ter tons parecidos com a pele. Desta maneira usando estes dois módulos e técnicas é possível obter imagens resultantes que removem o *background frame* a *frame*, deixando poucos ruídos e segmentando a região da mão.

Figura 29 – Resultados parciais da segmentação de movimento.



(a) Imagem resultante da diferenciação de *frames*.

(b) Imagem resultante binarização

(c) Imagem resultante pós erosão.

(d) Imagem final do módulo de segmentação de movimento negativado.

Fonte: o autor

5 Metodologia

Este capítulo apresentará as ferramentas de desenvolvimento e o ambiente no qual foi realizado a implementação dos algoritmos, o banco de dados, as implementações e as métricas para a análise.

5.1 Avaliação Experimental

Os experimentos nesse trabalho são para avaliar a capacidade do algoritmo proposto de segmentação mãos corretamento mesmo em ambientes complexos. Os detalhes experimentais serão explicados nas próximas subseções.

5.1.1 Ambiente Experimental

As experimentações deste trabalho de pesquisa foram realizados em um *notebook* com sistema operacional Microsoft Windows 10 Home X64, 8,00 GB de memoria RAM DDR3 1600MHz, um processador Intel I5-7200U CPU, 2.5Ghz com 3MB de memória cache.

Todas as implementações foram realizadas utilizando a linguagem de desenvolvimento Python versão 3.6.1, e como biblioteca de processamento de imagens foi utilizado OpenCV versão 3.3.0. O ambiente de desenvolvimento escolhido para este trabalho foi o JetBrains PyCharm Edu 4.0 Community[1].

5.1.2 Base de Dados

A base de dados utilizada para este projeto foi o EgoGesture (CAO et al., 2017). Essa base de dados contém 2.081 vídeos, 24.161 amostras de gestos e 2.953.224 *frames*. Os gestos são realizados por 50 indivíduos distintos.

Segundo o autor deste banco de dados, os vídeos foram coletados em seis diferentes ambientes, internos e externos. Também foram considerados gestos no qual pessoas estão andando. Dos seis ambientes, quatro são internos com a seguinte características:

- sujeito em modo estático, com *background* com desordem estática[2];
- sujeito em modo estático com *background* dinâmico;

[1] https://www.jetbrains.com/pycharm/

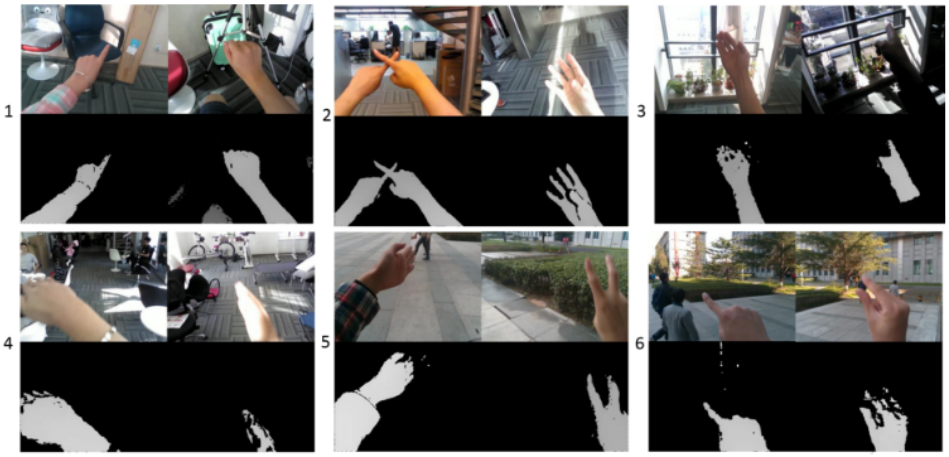
[2] Desordem estática pode ser exemplificada por um vídeo com *background* quase estático,pois contém uma pequena movimentação.

- sujeito em modo estático sob variação de iluminação
- sujeito realizando caminhada.

E os dois seguintes ambientes externos possuem as características seguintes:

- sujeito em modo estático com *background* dinâmico;
- sujeito realizando caminhada com *background* dinâmico.

Figura 30 – Alguns exemplos de cada um dos seis ambientes:



Fonte: Cao *et al.* (CAO et al., 2017)

Para este trabalho, foram consideradas quatro cenas de cada sujeito, e foram utilizados os dez primeiros sujeitos. Das quatro cenas escolhidas, duas foram de ambientes internos e duas de ambientes externos e cada uma possui seis vídeos.

Todos os vídeos selecionados possuem certo grau de movimentação no *background*. As cenas que possuem fundo com desordem estática podem ser compreendidas como vídeos em que o sujeito está gravando a cena e o *background* está estático, entretanto ainda existe um pequeno grau de movimentação em sua filmagem.

Neste projeto, foram consideradas cenas em ambiente interno com *background* estático com sujeito em modo estático e *background* dinâmico com sujeito em modo dinâmico, para as cenas em ambientes externos foi considerado os sujeitos nos dois modos, e com o *background* dinâmico.

Nas Figuras 31, 32, 33 e 34, são apresentados *frames* originais e rotulados de todos cenários usados neste trabalho.

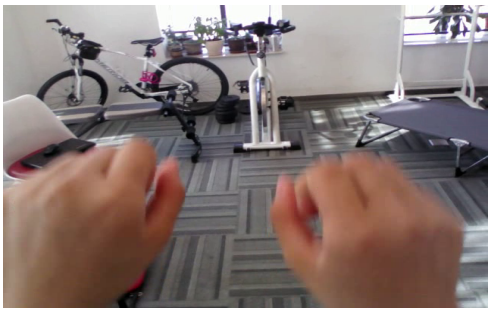
Figura 31 – Conjunto de imagens referentes a cena 1.

(a) Imagem original

(b) Imagem rotulada

Fonte: Cao *et al.* (CAO et al., 2017)

Figura 32 – Conjunto de imagens referentes a cena 4.



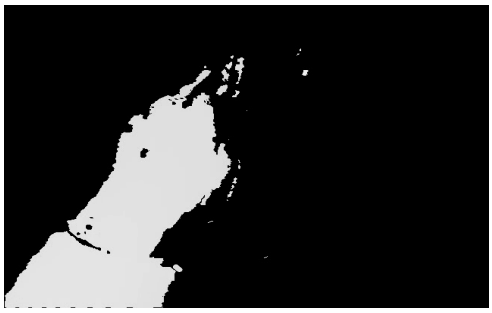
(a) Imagem original

(b) Imagem rotulada

Fonte: Cao *et al.* (CAO et al., 2017)

Figura 33 – Conjunto de imagens referentes a cena 5.

(a) Imagem original



(b) Imagem rotulada

Fonte: Cao *et al.* (CAO et al., 2017)

5.2 Métrica de Análise

Como métrica para avaliação da qualidade dos algoritmos de segmentação será utilizada a abordagem IOU, mostrada pelo trabalho de Fleet, Tomas e Pajdla (FLEET TOMAS PAJDLA, 2014). Este método determina o grau de similaridade entre o *frame*

Figura 34 – Conjunto de imagens referentes a cena 6.

(a) Imagem original (b) Imagem rotulada

Fonte: Cao *et al.* (CAO et al., 2017)

capturado e o *frame* rotulado. A Equação 5.1 representa o cálculo do IOU.

$$IOU = \frac{Área\ de\ Intersecão}{Área\ de\ União} \tag{5.1}$$

Definido o valor de IOU entre o *frame* capturado e o *frame* rotulado obteremos um valor entre 0 e 1, segundo Fleet, Tomas e Pajdla (FLEET TOMAS PAJDLA, 2014), o valor igual ou acima de 0.5 é considerado correto, abaixo disto incorreto. Para um vídeo avaliado teríamos vários valores de IOU, pois é gerado um IOU para cada *frame* que foi capturado, então, calcula-se a quantidade de *frames* classificados corretamente e divide pelo total de *frame* capturados segundo a Equação 5.2.

$$Taxa\ de\ Acerto = \frac{\sum Número\ de\ frames\ corretos}{Número\ de\ frames\ capturados} \tag{5.2}$$

Neste trabalho, foi realizada a captura de um *frame* a cada dez *frames*. O *frame* capturado é avaliado pelo método Intersecção Sobre União ou IOU.

5.3 Experimento

A obtenção de alguns parâmetros utilizados neste projeto baseou-se em na observação as imagens resultantes de cada iteração do algoritmo. O primeiro ponto que tem suma importância neste trabalho é o limiar de segmentação de tons de pele e teve seus valores baseados no limiar definidos por Basilio *et al.* (BASILIO et al., 2011) na Equação 5.3. Estes valores foram ajustados para melhorar a segmentação da pele e remoção do *background* este limiar foi apresentado na Equação 4.3 no capítulo anterior.

$$133 \leq Cr \leq 173\ \ e\ \ 80 \leq Cb \leq 120 \tag{5.3}$$

O próximo fator que foi analisado foi o tamanho da janela deslizante utilizado nos operadores morfológicos, para o tipo de imagem que o banco de dados proporciona o valor utilizado foi de $(5x5)$. Este valor foi obtido por meio da observação das imagens resultantes.

Para trabalhar com o módulo de segmentação de movimentação, é necessário fazer duas conversões de sistema de cor, do YC_rC_b para BGR e de BGR para tons de cinza. Outro fator que foi determinado por experimentação e observação foi o tamanho da janela deslizante utilizado no filtro Gaussiano aplicado para remoção de ruídos, o tamanho de $(11x11)$ apresentou os melhores resultados. Foram testados tamanhos de janela deslizantes menores como $(3x3)$, $(5x5)$, $(9x9)$, também foram testados janelas maiores de tamanho $(13x13)$ e $(15x15)$, porem para as janelas menores o nível de ruído apresentado foi grande, e para as maiores perdeu-se muita informação relevante.

Foram utilizadas duas técnicas de limiarização, a técnica de Otsu e limiarização global, combinadas, trazendo mais segurança a esta etapa, nela será aplicada os operadores morfológicos com intuito remover os ruídos das etapas anteriores. Logo em seguida foi realizado duas iterações de erosão com a janela deslizante definida anteriormente, o número de iterações também foi obtido por meio de experimentação.

A etapa de segmentação de pele utiliza o limiar baseado no trabalho de Basilio *et al.* (BASILIO et al., 2011) conforme já apresentado na Equação 4.3. Por fim os operadores lógicos "*and*" e "*or*" foram utilizados para suprir as falhas que possam ser encontradas nos módulos de segmentação por tons de pele ou movimento. O resultado obtido será apresentado no próximo capitulo.

As experimentações foram realizadas em dez sujeitos e cada um possui quatro tipos de cenas, duas internas e duas externas. Cada cena contém quatro vídeos. Logo foram analisados 144 vídeos, pois alguns sujeitos não possuíam os arquivos para as cenas. A média de *frames* por vídeos calculada foi de 1300, o algoritmo aplicou suas etapas em aproximadamente 187.200 *frames*, e destes 18.720 foram avaliados pela métrica do IOU.

Foram realizadas diversas iterações do algoritmo proposto com finalidade de determinar os melhores valores para as janelas deslizantes e limiar para segmentação da pele.

6 Resultados

Neste capitulo serão apresentados os resultados gerados por três técnicas: o trabalho de Yeo *et al.* (YEO; LEE; LIM, 2015), Thabet *et al.* (THABET et al., 2017) e o algoritmo proposto, no experimento com a base de dados EgoGesture. Todos os experimentos foram executados na mesma base de dados e foi utilizado o mesmo número de sujeitos e cenas, conforme explicado no capítulo anterior.

Os resultados serão apresentados por meio de tabelas, em que será exibido a porcentagem de acerto de cada cena em relação ao sujeito e temos o respectivo desvio padrão. Lembrando que cada cena possui quatro vídeos, então a porcentagem de acerto foi obtida por meio de uma média entre os vídeos da cena. Também serão exibidos gráficos respectivos a cada tabela apresentada. Também são exibidas imagens com a finalidade de exemplificar os resultados obtidos em cada algoritmo.

Por fim será apresentado uma tabela geral e gráficos de cada cena, considerando a porcentagem média de acerto por cena de cada algoritmo e o desvio padrão.

A Cena 1 é relacionado a ambiente interno com fundo de desordem estática e sujeito estático. A Cena 2 é composta de um ambiente interno com fundo dinâmico e sujeito dinâmico. As Cenas 3 e 4 constam de *background* dinâmico com sujeitos estáticos e dinâmicos, respectivamente.

6.1 Análise da Segmentação do Algoritmo de Thabet *et al.* (THABET et al., 2017)

Nesta seção serão apresentados os resultados e análise da segmentação de mãos realizada pelo algoritmos proposto por Thabet *et al.* (THABET et al., 2017).

A Tabela 1 apresenta os resultados que nessa base de dados não conseguiu segmentar as mãos nas imagens. A Tabela 1 mostra que os resultados obtidos com a métrica de IOU não foram satisfatórios, todos os *frames* considerados tiveram menos que 50% de similaridade com o respectivo *frame* rotulado, logo foram classificados como errado. Valores com ”nd”na tabela não foram encontrados, pois faltaram arquivos referentes as cenas e a os sujeitos no banco de dados de Cao *et al.* (CAO et al., 2017).

A Figura 35 apresenta os resultados obtidos pela abordagem de Thabet *et al.* (THABET et al., 2017) para quatro cenas diferentes, consistindo de duas em ambiente interno e duas em ambiente externos. As Figuras 35a e 35b representam o resultado da aplicação em ambiente interno e as Figuras 35c e 35d representam o resultado em

Tabela 1 – Porcentagem de acerto da cena (desvio padrão) $\times$ sujeito - Algoritmo de Thabet *et al.* (THABET et al., 2017)

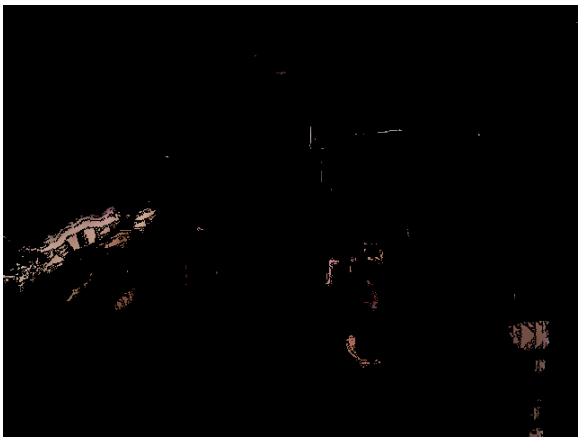
Cena / Sujeito	Sujeito 01	Sujeito 02	Sujeito 03	Sujeito 04	Sujeito 05	Sujeito 06	Sujeito 07	Sujeito 08	Sujeito 09	Sujeito 10
1 (Interno)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
4 (Interno)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
5 (Externo)	0 (0)	0 (0)	nd	0 (0)	0 (0)	0 (0)	nd	0 (0)	0 (0)	0 (0)
6 (Externo)	0 (0)	0 (0)	nd	0 (0)	0 (0)	0 (0)	nd	0 (0)	0 (0)	0 (0)

Fonte: O autor.

ambiente externo após aplicação do método proposto por Thabet *et al.* (THABET et al., 2017).

Figura 35 – Conjunto de imagens referentes aplicação da abordagem de Thabet *et al.* (THABET et al., 2017).

(a) Imagem resultante da abordagem de Thabet *et al.* (THABET et al., 2017) na cena 1.



(b) Imagem resultante da abordagem de Thabet *et al.* (THABET et al., 2017) na cena 4.

(c) Imagem resultante da abordagem de Thabet *et al.* (THABET et al., 2017) na cena 5.

(d) Imagem resultante da abordagem de Thabet *et al.* (THABET et al., 2017) na cena 6.

Fonte: o autor

Como pode-se observar, esta abordagem não foi eficiente, e o uso do método *Fast marching method* torna o algoritmo computacionalmente custoso. Também pode-se reparar que o *background* foi completamente removido em todas as figuras, entretanto a segmentação da mão foi gravemente afetada.

6.2 Análise da Segmentação do Algoritmo de Yeo *et al.* (YEO; LEE; LIM, 2015)

A abordagem apresentada por Yeo *et al.* (YEO; LEE; LIM, 2015) apresentou resultados mais relevantes, entretanto não conseguiu atingir bons resultados no banco de dados utilizado. Fica evidente a dificuldade de implementar um algoritmo ao ser consideradas as variáveis da mudança de iluminação a mudança de *background*, ambas causadas pela movimentação do sujeito ou da câmera.

Observando a Figura 36 vemos uma melhora em comparação com Thabet *et al.* (THABET et al., 2017). Apesar de conter mais ruídos a mão está mais presente conseguindo, deste modo, resultados um pouco melhores que a abordagem anterior na métrica IOU. O método apresentado por Yeo *et al.* (YEO; LEE; LIM, 2015) utiliza hardware de baixo custo, demostrando ser possível realizar segmentação da mão em ambientes complexos, porém requer alguns ajustes quando o ambiente se torna dinâmico. Talvez os resultados do estudo de Yeo *et al.* (YEO; LEE; LIM, 2015) apresentasse resultados melhores ao ser incorporado o módulo do Kinect, pois seriam consideradas também imagens de profundidade, entretanto o intuito desta pesquisa é considerar a utilização somente de imagens obtidas por câmeras tradicionais.

A Tabela 2 apresenta os resultados da aplicação do IOU nos *frames* da abordagem de Yeo *et al.* (YEO; LEE; LIM, 2015), podemos perceber que a maior porcentagem de acerto está na Cena 1 do Sujeito 8, apresentando 67,94% dos *frames* com grau de similaridade maior que 50%, e a porcentagem mais alta de acerto para uma cena em ambiente externo é de apenas 20,53% do Sujeito 6.

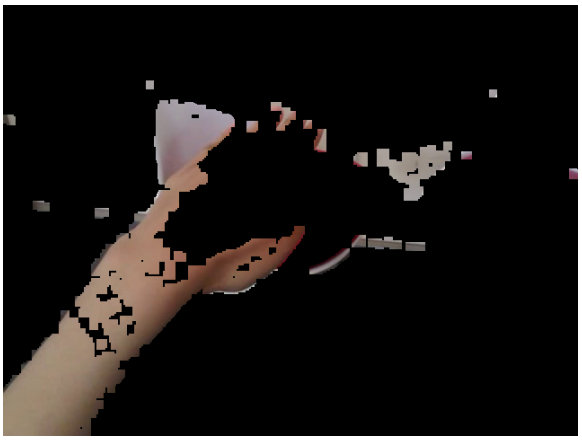
Tabela 2 – Porcentagem de acerto da cena (desvio padrão) X Sujeito - Algoritmo de Yeo *et al.* (YEO; LEE; LIM, 2015)

Cena / Sujeito	Sujeito 01	Sujeito 02	Sujeito 03	Sujeito 04	Sujeito 05	Sujeito 06	Sujeito 07	Sujeito 08	Sujeito 09	Sujeito 10
1 (Interno)	26,82 (10,19)	52,51 (8,87)	0 (0)	0 (0)	3,29 (3,73)	59,95 (8,99)	58,74 (12,72)	67,94 (6,29)	22,97 (6,04)	0 (0)
4 (Interno)	12,74 (7,31)	24,73 (16,71)	1,26 (1,26)	3,81 (4,07)	0 (0)	18,26 (12,05)	16,13 (7,38)	3,12 (3,8)	13,38 (12,23)	0 (0)
5 (Externo)	1,2 (2,11)	10,97 (4,96)	nd	0 (0)	3,36 (1,14)	8,96 (5,24)	nd	0 (0)	0 (0)	0,8 (1,39)
6 (Externo)	11,3 (10,57)	18,07 (15,88)	nd	0 (0)	2,7 (1,86)	20,53 (9,5)	nd	8,5 (10,75)	7,83 (5,82)	15,45 (7,04)

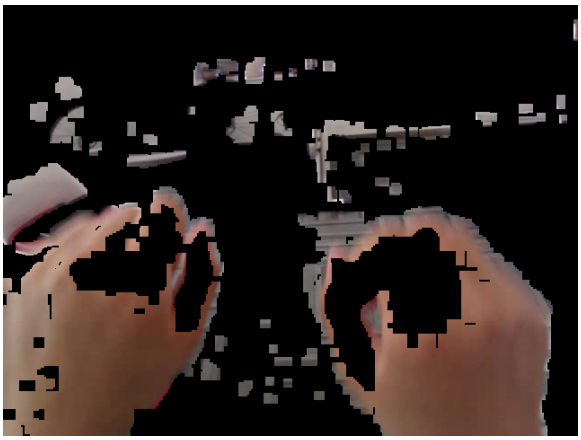
Fonte: O autor.

Como podemos ver no gráfico da Figura 37 referente a implementação de Yeo *et al.* (YEO; LEE; LIM, 2015), as cenas internas apresentaram melhores resultados do

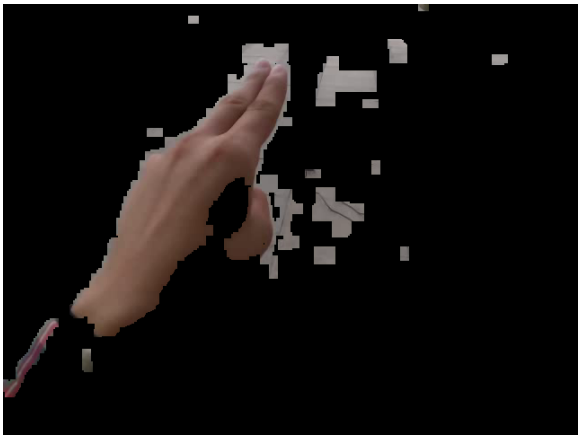
Figura 36 – Conjunto de imagens referentes aplicação da abordagem de Yeo *et al.* (YEO; LEE; LIM, 2015).



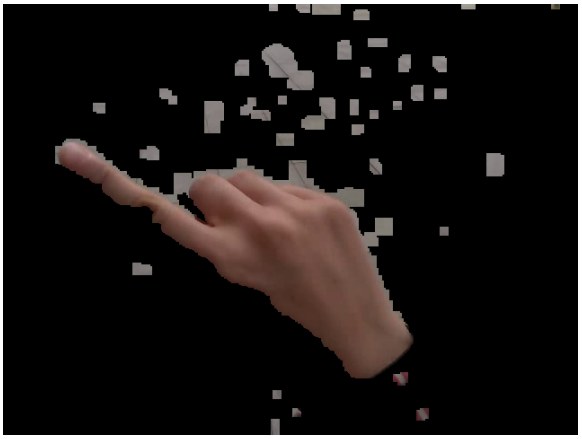
(a) Imagem resultante da abordagem de Yeo *et al.* (YEO; LEE; LIM, 2015) na cena 1.



(b) Imagem resultante da abordagem de Yeo *et al.* (YEO; LEE; LIM, 2015) na cena 4.



(c) Imagem resultante da abordagem de Yeo *et al.* (YEO; LEE; LIM, 2015) na cena 5.



(d) Imagem resultante da abordagem de Yeo *et al.* (YEO; LEE; LIM, 2015) na cena 6.

Fonte: o autor

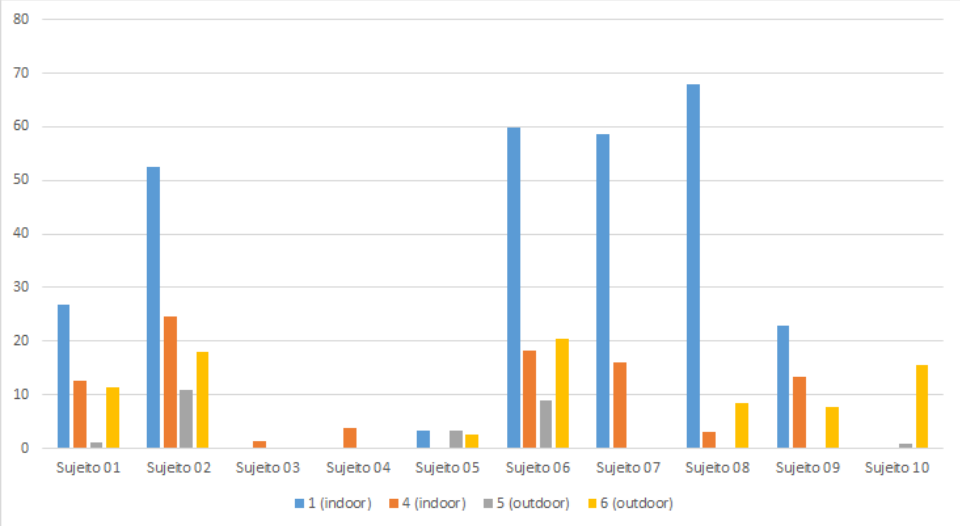
que as cenas externas e os Sujeitos 6, 7 e 8 tiveram bons resultados. Observamos que esta abordagem não apresentou bons resultados em ambientes externos.

6.2.1 Análise do Algoritmo Proposto

A abordagem proposta neste estudo apresentou bons resultados para este experimentos, considerando que foi utilizado somente imagens de câmeras RGB e uma abordagem híbrida composta por segmentação por tons de pele e movimentação. Os resultados apresentados na Figura 38 demonstra que o *background* foi quase totalmente removido. O *Frame* da Figura 38a apresenta um nível aceitável de ruídos, e nos demais *frames* apresentados não exibem ruídos.

Ao utilizar um banco de dados complexo como o *EgoGesture* do trabalho de

Figura 37 – Gráfico da porcentagem de acerto do algoritmo de Yeo *et al.* (YEO; LEE; LIM, 2015) X Sujeito



Fonte: O autor

Cao *et al.* (CAO et al., 2017), vemos que a abordagem proposta tem bons resultado, usando somente os vídeos capturados por câmera comum. Caso o algoritmo proposto neste projeto utilize um ambiente mais controlado, sem movimentação de *background*, é possível obter os melhores resultados.

A Tabela 3 exibe os resultados obtidos para os dez sujeitos que foram avaliados e suas quatro cenas, o resultado é procedido do respectivo desvio padrão, nela observamos que o melhor resultado está presente na Cena 1 do Sujeito 1, obtendo aproximadamente 76,4% dos *frames* avaliados como correto pelo IOU. Ao visualizar o gráfico na Figura 39 percebemos melhor estes resultados.

Tabela 3 – Porcentagem de acerto da cena (desvio padrão) X Sujeito - Algoritmo proposto pelo autor

Cena / Sujeito	Sujeito 01	Sujeito 02	Sujeito 03	Sujeito 04	Sujeito 05	Sujeito 06	Sujeito 07	Sujeito 08	Sujeito 09	Sujeito 10
1 (Interno)	76,4 (5,66)	39,34 (14,89)	25,71 (10,18)	52,6 (7,41)	36,66 (15,24)	50,71 (6,9)	71,52 (3,56)	62,43 (5,58)	38,92 (5,87)	27,33 (14,11)
4 (Interno)	39,46 (13,53)	27,6 (18,77)	5,53 (8,10)	20,65 (13,91)	22,68 (13,44)	47,55 (10,64)	37,21 (17,8)	28,52 (21,44)	63,65 (22,75)	18,39 (8,44)
5 (Externo)	31,14 (12,03)	0,5 (0,96)	nd	1,67 (1,68)	16,61 (3,06)	32,94 (6,45)	nd	14,79 (11,21)	2 (1,73)	45,15 (9,9)
6 (Externo)	29,85 (7,16)	10,93 (6,44)	nd	10,2 (7,23)	4,26 (3,13)	33,39 (9,71)	nd	20,59 (16,24)	30,73 (1,53)	37,64 (7,65)

Fonte: O autor.

Para estes cenários apresentados nas Figuras 40, 41, 42, 43 o algoritmo falhou, apresentando diversos ruídos nas imagens. Como o ambiente é dinâmico, as imagens apresentadas podem ter sido afetadas por diversos fatores, mas o principal está relacionado a iluminação, que afetará o tom da pele.

Figura 38 – Conjunto de imagens referentes aplicação da abordagem proposta pelo autor.

(a) Imagem resultante da abordagem proposta na cena 1.

(b) Imagem resultante da abordagem proposta na cena 4.

(c) Imagem resultante da abordagem proposta na cena 5.

(d) Imagem resultante da abordagem proposta na cena 6.

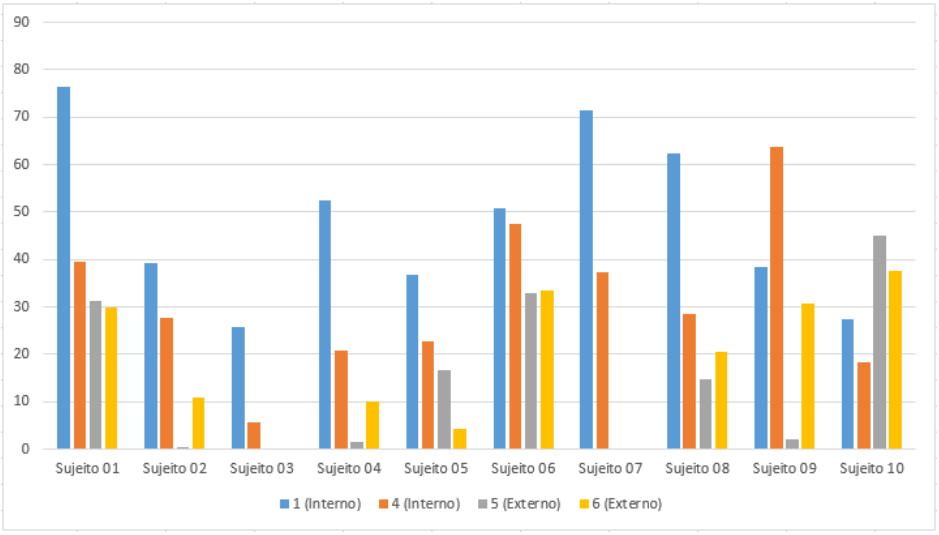
Fonte: o autor

6.2.2 Análise Geral

Nesta etapa será feita uma análise geral das três abordagens avaliadas neste trabalho na base de dados EgoGestures. Podemos notar que não foram obtidos resultados com similaridade acima de 50% para a abordagem de Thabet *et al.* (THABET et al., 2017). Desse modo, para esta base de dados, esta abordagem não demonstrou ser efetiva.

A abordagem de Yeo *et al.* (YEO; LEE; LIM, 2015) possui *frames* com maior similaridade a imagem rotulada, apresentando alguns bons resultados, porém conforme visto nas imagens presentes na Figura 36, o módulo da câmera, sozinho, não é tão eficiente em remover o *background* e deixar a região das mãos intacta, e talvez por isso, em sua abordagem o autor resolveu utilizar um módulo com câmera de profundi-

Figura 39 – Gráfico da porcentagem de acerto do algoritmo proposto X Sujeito



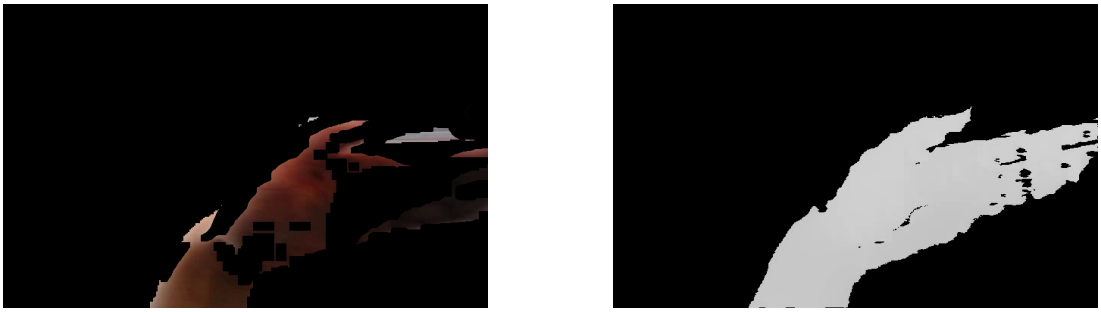
Fonte: O autor

Figura 40 – Exemplo de falha do algoritmo proposto na cena 1.

(a) Imagem com falha (b) Imagem rotulada

Fonte: O autor e Cao *et al.* (CAO et al., 2017)

Figura 41 – Exemplo de falha do algoritmo proposto na cena 4.



(a) Imagem com falha (b) Imagem rotulada

Fonte: O autor e Cao *et al.* (CAO et al., 2017)

dade, entretanto este é um ponto que não será avaliado neste estudo que possui como objetivo ter boas segmentações apenas em câmeras comuns.

Figura 42 – Exemplo de falha do algoritmo proposto na cena 5.

(a) Imagem com falha (b) Imagem rotulada

Fonte: O autor e Cao *et al.* (CAO et al., 2017)

Figura 43 – Exemplo de falha do algoritmo proposto na cena 6.

(a) Imagem com falha (b) Imagem rotulada

Fonte: O autor e Cao *et al.* (CAO et al., 2017)

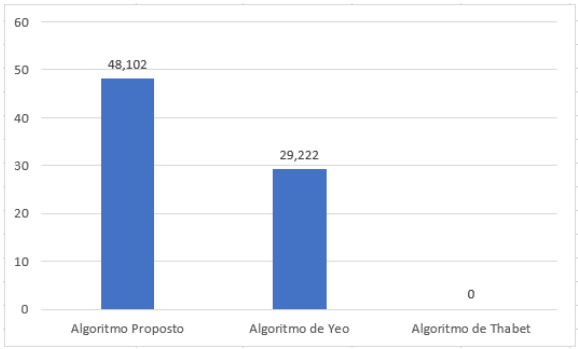
Tabela 4 – Porcentagem média de acerto da cena (desvio padrão) x Algoritmo

Cena / Algoritmo	Proposto	Yeo	Thabet
1 (indoor)	55,61 (17,22)	29,22 (28,15)	0 (0)
4 (indoor)	41,89 (16,34)	9,34 (8,81)	0 (0)
5 (outdoor)	21,61 (17,18)	3,16 (3,37)	0 (0)
6 (outdoor)	25,21 (14,68)	10,54 (7,22)	0 (0)

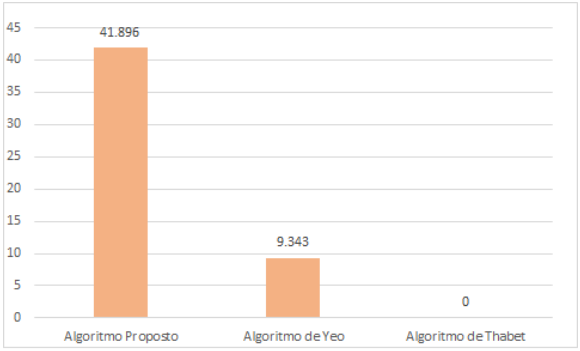
A Tabela 4 demonstra a média dos resultados obtidos pelo uso do *Intersection Over Union* em cada cena, e seu respectivo desvio padrão. Os resultados para cada cena são demonstrados nos gráficos presentes na Figura 44. Em todos os cenários a abordagem aqui proposta obteve resultados superiores aos demais algoritmos avaliados.

O trabalho de Thabet *et al.* (THABET et al., 2017) é um trabalho recente, e foi considerando o módulo da câmera, entretanto não apresentou bons resultados, mas foi possível, ainda assim, extrair informações relevantes do seu estudo. A técnica de *Fast marching method* que é utilizada no final de cada etapa da sua abordagem deixa

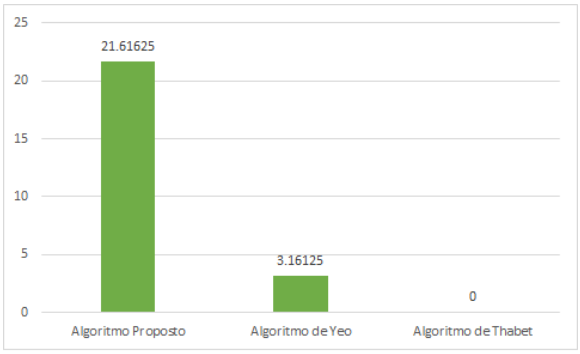
Figura 44 – Conjunto de gráficos referente as cenas avaliadas.



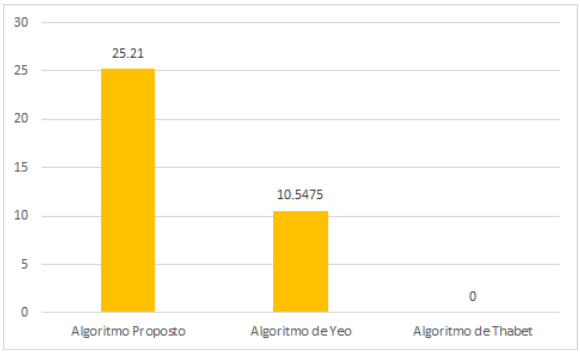
(a) Gráfico da cena 01 e os respectivos resultados dos algoritmos implementados.



(b) Gráfico da cena 04 e os respectivos resultados dos algoritmos implementados.



(c) Gráfico da cena 05 e os respectivos resultados dos algoritmos implementados.



(d) Gráfico da cena 06 e os respectivos resultados dos algoritmos implementados.

Fonte: o autor

o algoritmo pesado computacionalmente, necessitando de muito mais tempo para ser executado do que as outras abordagens.

O estudo de Yeo *et al.* (YEO; LEE; LIM, 2015) é uma abordagem mais parecida com este estudo aqui proposto. Ele conseguiu obter resultados intermediários em imagens com o fundo mais estático e resultados não tão bons em cenários com fundos dinâmicos.

A abordagem proposta neste estudo conseguiu os melhores resultados do que as implementações anteriores, demonstrando que esse algoritmo possui a capacidade de segmentar mãos mesmo em ambientes com *background* complexos e em ambientes dinâmicos como os avaliados na base de dados EgoGesture e utilizando câmeras 2D comuns.

7 Conclusão

7.1 Considerações Finais

Este trabalho de pesquisa teve o objetivo de realizar o reconhecimento de mãos em *background* complexo, propondo uma nova abordagem híbrida. No início, foi realizada uma revisão bibliográfica para trabalhos de detecção de mãos. A maioria das abordagens tratam somente a detecção da mão pelo tom de pele, porém, estes necessitam de um ambiente bem controlado. O algoritmo aqui proposto é híbrido, pois utiliza método de segmentação de tons de pele combinada com segmentação de movimentação, esta fusão de métodos é realizada por meio de operadores lógicos. Foi realizada uma comparação entre a abordagem aqui proposta e os trabalhos da literatura *Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware* de Yeo *et al.* (YEO; LEE; LIM, 2015) e *Fast marching method and modified features fusion in enhanced dynamic hand gesture segmentation and detection method under complicated background* de Thabet *et al.* (THABET et al., 2017). O algoritmo proposto obteve os melhores resultados nos experimentos realizados com a base de dados EgoGesture.

Realizar a segmentação das mãos em *background* complexo não é um trabalho trivial, a proposta desta abordagem teve o intuito de tornar o algoritmo mais robusto às variáveis apresentadas por um ambiente não controlado. Com a análise dos artigos estudados, foi possível apresentar uma proposta de algoritmo que conseguiu realizar este trabalho de maneira eficaz, não sendo necessários hardwares caros e complexos.

Ficou evidente, nas imagens geradas pelo algoritmo proposto, que com a perspectiva híbrida é possível suprir as dificuldades dos módulos de segmentação por tons de pele e de movimento. Separadamente, no entanto, estes métodos teriam problemas devido ao ambiente não ser controlado. Deste modo, pode-se inferir pelos experimentos na base de dados EgoGesture que a utilização da abordagem proposta consegue apresentar bons resultados na detecção de mãos em ambientes complexos.

7.2 Trabalhos Futuros

Uma alternativa que poderá ser futuramente adicionada é a utilização de uma abordagem adaptativa, que poderá determinar um melhor limiar para segmentação do tom de pele, conforme a iluminação muda, o tom da pele também irá mudar, tornando-se uma variável difícil de controlar.

Outro ponto que poderá ser aprimorado é a adição de um estabilizador de vídeo, pois o uso de um módulo de segmentação de movimento requer que o vídeo seja gravado por uma câmera estática, fazendo com que seja mais fácil e robusta a segmentação do *foreground*

As abordagens que utilizam aprendizagem profunda também devem ser consideradas e seria interessante como trabalho futuro realizar uma comparação destas técnicas com as abordagens hibridas aqui presente. Deste modo também podemos considerar a utilização do *deep learning* em conjunto com as técnicas hibridas, sendo que, na classificação dos gestos.

Referências

BAMBACH, S. et al. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2015. p. 1949–1957. Citado na página 37.

BASILIO, J. A. M. et al. Explicit image detection using ycbcr space color model as skin detection. *Applications of Mathematics and Computer Engineering*, p. 123–128, 2011. Citado 5 vezes nas páginas 31, 37, 46, 53 e 54.

CAO, C. et al. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In: *The IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017. Citado 8 vezes nas páginas 50, 51, 52, 53, 55, 59, 61 e 62.

DADGOSTAR, F.; SARRAFZADEH, A. An adaptive real-time skin detector based on hue thresholding: A comparison on two motion tracking methods. *Pattern recognition letters*, Elsevier, v. 27, n. 12, p. 1342–1352, 2006. Citado na página 32.

FLEET TOMAS PAJDLA, B. S. T. T. e. D. pdf. *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. [S.l.: s.n.], 2014. Citado 2 vezes nas páginas 52 e 53.

GONZALEZ, R.; WOODS, R. *Processamento Digital De Imagens*. ADDISON WESLEY BRA. ISBN 9788576054016. Disponível em: <https://books.google.com.br/books?id=r5f0RgAACAAJ>. Citado 2 vezes nas páginas 29 e 30.

GONZALEZ, R.; WOODS, R. *Processamento Digital De Imagens*. [S.l.]: ADDISON WESLEY BRA, 2011. Citado 7 vezes nas páginas 19, 22, 23, 24, 25, 26 e 27.

HORVATH, M. *Additive color mixing: adding red to green yields yellow; adding red to blue yields magenta; adding green to blue yields cyan; adding all three primary colors together yields white.* 2006. Disponível em: <https://en.wikipedia.org/wiki/RGB_color_model#/media/File:AdditiveColor.svg>. Citado na página 21.

HU, W.-C. et al. Moving object detection and tracking from video captured by moving camera. *Journal of Visual Communication and Image Representation*, v. 30, p. 164 – 180, 2015. ISSN 1047-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S104732031500053X>. Citado 2 vezes nas páginas 33 e 37.

KAKUMANU, P.; MAKROGIANNIS, S.; BOURBAKIS, N. A survey of skin-color modeling and detection methods. *Pattern recognition*, Elsevier, v. 40, n. 3, p. 1106–1122, 2007. Citado 3 vezes nas páginas 22, 28 e 31.

KHAN, A. U.; BORJI, A. Analysis of hand segmentation in the wild. *arXiv preprint arXiv:1803.03317*, 2018. Citado na página 37.

KHAN, R.; HANBURY, A.; STOETTINGER, J. Skin detection: A random forest approach. In: IEEE. *Image Processing (ICIP), 2010 17th IEEE International Conference on*. [S.l.], 2010. p. 4613–4616. Citado na página 32.

LIU, L. et al. A real-time and low-cost hand tracking system. In: *2017 IEEE International Conference on Consumer Electronics (ICCE)*. [S.l.: s.n.], 2017. p. 430–431. Citado na página 33.

NEIVA, D. H.; ZANCHETTIN, C. A dynamic gesture recognition system to translate between sign languages in complex backgrounds. In: IEEE. *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*. [S.l.], 2016. p. 421–426. Citado 2 vezes nas páginas 26 e 33.

OJHA, S.; SAKHARE, S. Image processing techniques for object tracking in video surveillance-a survey. In: IEEE. *Pervasive Computing (ICPC), 2015 International Conference on*. [S.l.], 2015. p. 1–6. Citado 2 vezes nas páginas 29 e 33.

OTSU, N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, IEEE, v. 9, n. 1, p. 62–66, 1979. Citado na página 29.

RAUTARAY, S. S.; AGRAWAL, A. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, Springer, v. 43, n. 1, p. 1–54, 2015. Citado 4 vezes nas páginas 13, 14, 15 e 21.

SCURI, A. E. Fundamentos da imagem digital. *Pontifícia Universidade Católica do Rio de Janeiro*, 1999. Citado 5 vezes nas páginas 18, 19, 20, 23 e 28.

SHAIK, K. B. et al. Comparative study of skin color detection and segmentation in hsv and ycbcr color space. *Procedia Computer Science*, v. 57, n. Supplement C, p. 41 – 48, 2015. ISSN 1877-0509. 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015). Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050915018918>. Citado 4 vezes nas páginas 20, 21, 22 e 31.

SLACKERPRIME. *Per-Pixel Scroller pt. 3 - Now in color!* 2015. Disponível em: <http://petitcomputer.wikia.com/wiki/User_blog:SlackerPrime/Per-Pixel_Scroller_pt._3_-_Now_in_color!> Citado na página 22.

SONG, W. et al. Motion-based skin region of interest detection with a real-time connected component labeling algorithm. *Multimedia Tools and Applications*, Springer, v. 76, n. 9, p. 11199–11214, 2017. Citado 2 vezes nas páginas 34 e 37.

STAMFORD, C. *Gartner's 2016 Hype Cycle for Emerging Technologies Identifies Three Key Trends That Organizations Must Track to Gain Competitive Advantage*. 2017. Disponível em: <Gartner's2016HypeCycleforEmergingTechnologiesIdentifiesThreeKeyTrendsThatOrganizationsMustTrackto Citado na página 14.

SZELISKI, R. *Computer Vision: Algorithms and Applications*. 1st. ed. New York, NY, USA: Springer-Verlag New York, Inc., 2010. ISBN 1848829345, 9781848829343. Citado na página 24.

TEKALP, A. M. *Digital Video Processing*. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2015. ISBN 0133991008, 9780133991000. Citado na página 20.

THABET, E. et al. Fast marching method and modified features fusion in enhanced dynamic hand gesture segmentation and detection method under complicated background. *Journal of Ambient Intelligence and Humanized Computing*, Springer, p. 1–15, 2017. Citado 16 vezes nas páginas 7, 8, 9, 12, 35, 36, 39, 42, 43, 44, 55, 56, 57, 60, 62 e 64.

YEO, H.-S.; LEE, B.-G.; LIM, H. Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware. *Multimedia Tools and Applications*, Springer, v. 74, n. 8, p. 2687–2715, 2015. Citado 19 vezes nas páginas 7, 8, 9, 12, 25, 34, 35, 36, 37, 39, 40, 41, 55, 57, 58, 59, 60, 63 e 64.