UNIVERSIDADE FEDERAL DE SERGIPE
CAMPUS PROF. ALBERTO CARVALHO
DEPARTAMENTO DE SISTEMAS DE INFORMAÇÃO

NATHANAEL OLIVEIRA VASCONCELOS

Modelo de Interação Natural com Imagem Projetada: Proposta Alternativa ao uso de Sensores Reais, Baseado em Algoritmos de IA e Processamento de Imagens

ITABAIANA
2015

UNIVERSIDADE FEDERAL DE SERGIPE

CAMPUS PROF. ALBERTO CARVALHO

DEPARTAMENTO DE SISTEMAS DE INFORMAÇÃO

NATHANAEL OLIVEIRA VASCONCELOS

Modelo de Interação Natural com Imagem Projetada: Proposta Alternativa ao uso de Sensores Reais, Baseado em Algoritmos de IA e Processamento de Imagens

Trabalho de Conclusão de Curso submetido ao Departamento de Sistemas de Informação da Universidade Federal de Sergipe - DSII-TA/UFS, como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Alcides Xavier Benicasa

ITABAIANA

2015

NATHANAEL OLIVEIRA VASCONCELOS

Modelo de Interação Natural com Imagem Projetada: Proposta Alternativa ao uso de Sensores Reais, Baseado em Algoritmos de IA e Processamento de Imagens

Trabalho de Conclusão de Curso submetido ao Departamento de Sistemas de Informação da Universidade Federal de Sergipe - DSIITA/UFS, como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

Itabaiana, 24 de Fevereiro de 2015.

BANCA EXAMINADORA:

Prof. Dr. Alcides Xavier Benicasa
Orientador
DSIITA/UFS

Prof. Dr. Methanias Colaço Júnior
DSIITA/UFS

Profa. Msc. Mai-Ly Vanessa A.S. Faro
DSIITA/UFS

Dedico este tabalho à meus pais,
Maria Nair de Oliveira Vasconcelos e Antonio Barbosa de Vasconcelos
por estarem sempre presentes e serem
motivos para minha existência

AGRADECIMENTOS

Este trabalho não seria possível sem as oportunidades que tive. Nenhuma dessas oportunidades seriam possíveis se não fosse pelo esforço e o apoio dos meus pais Nair e Antonio, que me ensinaram a viver com dignidade, e se dedicaram ao máximo para garantir o melhor a mim, a vocês, faltam palavras para agradecer.

À minha irmã Náyra, por sempre torcer por mim. Ao meu irmão Ramon, pelo incentivo e por ser um exemplo a ser seguido pela dedicação e persistência. Minha prima, irmã de criação e um pouco mãe Edna, sempre presente, foram tantos ensinamentos, incentivos, não há palavras para agradecer você. Meu cunhado de consideração Givaldo, muito obrigado pela força, conselhos e incentivo.

À meu orientador Prof. Dr. Alcides Benicasa, pela confiança depositada em mim, e em especialmente, pela oportunidade de aprendizado que o acompanhamento de suas orientações me possibilitou, não somente para a realização deste trabalho, mas também durante toda a graduação. Fica aqui o meu agradecimento e a minha admiração pelo seu exemplo de competência.

Agradeço aos demais professores do Departamento de Sistemas de Informação da Universidade Federal De Sergipe - Campus Prof. Alberto Carvalho pelo ensino de qualidade, e, em especial ao Prof. Msc. Marcos Dósea, pelos ensinamentos e conselhos durante o tempo em que fiz parte da Empresa Júnior do Departamento.

À todos os amigos de graduação, com quais pude compartilhar momentos e experiência, em especial os companheiros de laboratório (Breno, Charles e Thiago), que contribuíram para este trabalho.

À COPES, pelo apoio financeiro no projeto de pesquisa.

Por fim, à todos que deram apoio moral e incentivo que me motivaram a não me desviar do meu objetivo e acreditar que conseguiria alcança-ló.

RESUMO

A interação intuitiva entre humanos e computadores é um campo de pesquisa que tem sido bastante investigada nos últimos anos. O uso de diferentes tipos de sensores, como por exemplo, sensores de movimento, toque, etc, tem proporcionado grandes avanços nesta área. Entretanto, de acordo com a especificidade de cada sensor, o custo para o desenvolvimento de determinadas aplicações torna-se muitas vezes inviável. Atualmente, com maior acesso a dispositivos de captura de imagens e computadores com capacidades suficientes de processamento, é possível identificar interação sem a necessidade de dispositivos especiais. Sendo assim, este trabalho possui como principal objetivo a proposta de um modelo como alternativa ao uso de sensores reais, neste caso, sensores de toque para interação homem-máquina. Como validação do modelo proposto foi desenvolvida uma aplicação que torna possível a interação entre o usuário e a imagem projetada por um projetor de imagem (*datashow*). Para isso, foram utilizadas técnicas de inteligência artificial e processamento de imagens, que permitiram identificar informações relevantes da cena (alvos), isolando-a através do uso de segmentação de imagens e, por fim, a utilização do conceito de histogramas para identificar se houve ou não interação.

PALAVRAS-CHAVES: Processamento de imagens; inteligência artificial; interação humano-computador; imagem projetada; segmentação de imagens; histograma.

ABSTRACT

The intuitive interaction between humans and computers is a field of research that has been investigated in recent years. The use of different types of sensors, such as motion sensors, touch, etc., have provided great strides in this area. However, according to the specificity of each sensor, the cost for developing some applications it is often infeasible. Currently with greater access to image capture devices and computers with enough processing capability, you can identify interaction without the need for special devices. Thus, this work has as main objective the proposal for a model as an alternative to using real sensors, in this case, touch sensors for human-machine interaction. As validation of the proposed model was developed an application that makes possible the interaction between the user and the image projected by an image projector. For this, we used the techniques of artificial intelligence and processing images, which allowed to identify relevant information from the scene (targets), isolating them through the use of image segmentation, and finally, the use of the term histogram to identify whether or not there was interaction.

KEYWORDS: Image processing; artificial intelligence; human-computer interaction; projected image; image segmentation; histogram.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos os computadores têm se tornado dispositivos comuns ao cotidiano das pessoas. De acordo com Meirelles (2014), coordenador da 25ª Pesquisa Anual do Uso de Tecnologia da Informação no Mercado Brasileiro, divulgada em abril de 2014 pela Fundação Getúlio Vargas, existem 136 milhões de computadores em uso no Brasil, uma densidade de 67% per capita ou 2 computadores para cada 3 habitantes. Os avanços na miniaturização de dispositivos, aliados ao surgimento de ferramentas para a comunicação sem fio, processadores portáteis e novas tecnologias sensitivas, abriram as portas para pesquisas sobre novas formas de interação e ainda, de acordo com Chen et al. (2005), novas tecnologias surgem dia após dia, ocupando novos espaços, transformando a realidade.

Com o objetivo de aproximar e facilitar a interação entre o homem e a máquina, atualmente, diversos consoles, computadores e *smartphones* apresentam funcionalidades de interatividade baseadas em comandos por voz, gestos, ou ainda, pela captura de movimentos do globo ocular. De acordo com Rusnak (2012), os sistemas presentes nestes dispositivos, os quais permitem este tipo de interação, são denominados por sistemas interativos.

O uso de gestos está em ascensão, sendo uma forma de interação natural que, para Valli (2007), consiste em sistemas que entendem ações naturalmente utilizadas pelas pessoas para se comunicar, permitindo aos usuários interagir entre eles e com o ambiente a seu redor. Com isso, o uso de gestos para a comunicação com dispositivos computacionais diminui a carga cognitiva, resultando um modo bastante natural e intuitivo de interagir.

Dispositivos baseados em gestos podem ser construídos utilizando diferentes mecanismos de rastreamento óptico, magnético ou mecânico ligados ao computador e/ou colocados no corpo do usuário. Muitos desses sistemas utilizam equipamentos sofisticados (dispositivos de rastreamento, luvas, câmeras especiais, etc.). Outra forma de interação através de gestos é baseada na visão computacional, que para Jain e Dorai (1997), tem como objetivo a interpretação automática de cenas complexas, para isso, são utilizadas técnicas de reconhecimento de padrões nas imagens capturadas, ao invés de dispositivos de rastreamento, além de algumas restrições do ambiente, como por exemplo, o fundo da cena, cores dos objetos a serem reconhecidos, e condições de iluminação, construindo assim, ambientes bem controlados, de maneira a facilitar operabilidade do sistema.

É fato que a utilização de dispositivos tais como: luvas de dados, sensores eletromagnéticos, entre outros, simplificam a identificação de gestos, mas como afirmam Erol et al. (2005), aumentam o custo do sistema e, em alguns casos, requerem calibração, tornando assim uma tecnologia ainda não muito acessível à maioria da população. Ao

contrário desses dispositivos, interfaces baseadas em visão computacional oferecem diversas vantagens, consistindo numa interação natural entre humanos e computadores, sem a necessidade de instalações especiais, nem da utilização de qualquer dispositivo (mecânico, óptico ou magnético) que o usuário deva vestir ou manipular. Possibilitando assim, uma maior liberdade e facilidade no uso do sistema.

Em paralelo ao que foi descrito e, de acordo com Pereira (2007), ao longo do tempo a ciência criou novos ramos de estudo que envolvem a simulação de processos que ocorrem no corpo humano e buscam também criar modelos ou máquinas que simulem determinadas características e comportamentos humanos. Algumas áreas têm se destacado em tal empreendimento, como é o caso da Inteligência Artificial (IA). A IA consiste de esforços intelectuais e tecnológicos relacionados à construção de máquinas inteligentes, à formalização do conhecimento, à mecanização do raciocínio, e ao uso de modelos computacionais para compreender a psicologia e o comportamento de pessoas e animais (DOYLE; DEAN, 1996).

Várias áreas da Ciência da Computação utilizam conhecimentos da IA com o intuito de automatizar processos. É o que ocorre com a visão computacional, área na qual este trabalho se enquadra. Para Pereira (2007), a atenção visual é a habilidade que o sistema visual dos vertebrados superiores utiliza para selecionar e processar somente as regiões mais relevantes em uma cena visual. Podendo ser entendida assim, como um mecanismo para lidar com a incapacidade de tratar de uma só vez uma grande quantidade de informação visual, tanto em sistemas biológicos, quanto em sistemas computacionais. Com isso, somente as regiões mais importantes em uma cena são escolhidas para processamento.

Sendo assim, este trabalho possui como principal objetivo propor um modelo que, utilizando técnicas de Inteligência Artificial e Processamento de Imagem, torne possível a detecção, reconhecimento da ação e interação do usuário com o computador, a partir de informações providas por uma câmera de vídeo, de modo que não seja necessária a utilização de sensores especiais. Para validação do modelo proposto, foi desenvolvido uma aplicação que possibilita a interação intuitiva entre o usuário e a imagem projetada por um projetor de imagem (*datashow*).

Este trabalho está dividido em cinco capítulos. No Capítulo 2, é realizada a revisão da literatura necessária para o embasamento teórico a ser utilizada durante o desenvolvimento deste trabalho, com o estudo sobre algoritmos de processamento de imagens para o tratamento e obtenção de informações relevantes da cena que possam identificar regiões propensas à possíveis alvos que, de maneira intuitiva, é seguida pelo estudo sobre algoritmos de segmentação de imagens para o isolamento de possíveis objetos alvos, a partir das regiões identificadas. Na revisão da literatura também são apresentados os trabalhos relacionados. A aplicação proposta é descrita no Capítulo 3, no qual é apresentada a lin-

guagem de programação utilizada e todo o processo de forma detalhada, iniciando com a captura de imagens, passando pelo pré-processamento, identificação das regiões salientes, segmentação da imagem e, por fim, a identificação de interação. Os experimentos e seus resultados são apresentados e analisados no Capítulo 4. Foram realizados experimentos em ambientes heterogêneos, divididos em duas etapas, a primeira para identificação das regiões das opções e a segunda com o objetivo de identificar a interação. O Capítulo 5 conclui o trabalho apresentando um resumo dos principais resultados obtidos, as contribuições da pesquisa desenvolvida e algumas sugestões de trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo serão apresentados os embasamentos teóricos que foram utilizados para o desenvolvimento deste trabalho e, por último, os trabalhos relacionados.

2.1 Evolução da Interação

Interação Homem-Computador (IHC) é a disciplina que se preocupa com o planejamento, avaliação e implementação de sistemas computacionais interativos para o uso humano e com o estudo dos fenômenos mais importantes que os rodeiam (HEWETT et al., 2009).

Ao longo das décadas, as tecnologias passaram a assumir tarefas mais elaboradas, e como consequência, as formas de interação entre humanos e máquinas passaram a requerer mais treinamento. Os primeiros modelos de computadores sequer possuíam interfaces gráficas, e mesmo com o surgimento dos computadores pessoais, também chamados de PCs, acompanhados de monitores, o alto grau de complexidade envolvendo interação permaneceu evidente (FIGUEIREDO et al., 2012).

Em 1980, após empresas como Intel, Apple, IBM e outras haverem lançado os primeiros PCs, houve uma mudança significativa em relação ao seu uso de uma forma geral. Estes passaram a estar presentes nos mais diversos setores, sendo usados, por exemplo, como simuladores na indústria, ferramentas de escritório e fontes de entretenimento dentro de domicílios (KANELLOS, 2002). Segundo Figueiredo et al. (2012), a demanda de uso desses equipamentos aflorou a necessidade de novas pesquisas com foco nos usuários e em como estes percebiam e reagiam às máquinas. Iniciaram-se estudos voltados a interfaces gráficas mais intuitivas, dispositivos de entrada que permitiam uma interação mais rápida e precisa, entre outros tópicos.

Para Raymond e Landley (2004), a área da IHC divide-se em três eras: computação em lote, interface de linhas de comando e interface gráfica de usuário (GUI - Graphic User Interface). A interação na computação em lotes deu-se através do uso de cartões perfurados como entrada ou fitas K7 e saída por meio de impressora. Em relação à utilização das interfaces de linhas de comando, a interação dava-se por meio de transações de pedido e resposta, com os pedidos expressados em comandos de texto que eram convertidos em funções do sistema. Com os computadores capazes de exibir gráficos e, com o advento do mouse, mais um método de entrada surgiu, as Interfaces Gráficas.

2.1.1 Interação Natural

Nos últimos anos, a evolução tecnológica permitiu a criação de vários dispositivos, os quais permitiram o surgimento de uma nova modalidade de interface, disponível em aplicações como telas sensíveis ao toque, reconhecimento de comandos de voz, reconhecimento de gestos, entre outras. Esse novo paradigma é conhecido como Interfaces Naturais de Usuário (INU). Para Blake (2012), INU podem ser definidas como interfaces projetadas para reusar habilidades existentes para a interação direta com o conteúdo.

Essa definição evidencia três importantes conceitos a respeito de INU, destacados a seguir:

- INU são projetadas, ou seja, requerem que sejam premeditadas e que sejam feitos esforços prévios para sua concepção. É preciso assegurar que as interações em uma INU sejam apropriadas tanto para o usuário quanto para o conteúdo e o contexto (BLAKE, 2012).

- Reutilizam habilidades existentes, durante vários anos, os usuários têm praticado a comunicação, verbal ou não verbal, além de interações com o ambiente. O poder computacional e tecnológico evoluíram ao ponto em que é possível tirar vantagem dessas habilidades. INU fazem isso ao permitir que usuários interajam com computadores por meio de ações intuitivas como tocar, gesticular e falar (BLAKE, 2012).

- Tem interação direta com o conteúdo, ou seja, o foco da interação está no conteúdo e na interação direta com ele. Isso não significa que a interface não possa ter controles, como botões ou caixas de seleção, se necessário. Significa apenas que esses controles deveriam ser considerados secundários, comparados ao conteúdo, e que a direta manipulação do conteúdo deveria ser o método de interação primário (BLAKE, 2012).

Sendo assim, como afirma Dias Diego R. C. (2013), uma interface INU exige apenas que o usuário seja capaz de interagir com o ambiente por meio de interações previamente já conhecidas pelo mesmo. Esse tipo de interface exige aprendizagem, porém é facilitada, pois não exige que o usuário seja apresentado a um novo dispositivo.

2.1.1.1 Rastreamento do corpo humano

Para Figueiredo et al. (2012), o conceito de interação natural permite abordar todas as formas de comunicação que um ser humano é capaz de executar, como a fala, toque, gestos, expressões faciais, olhares entre outras. O mesmo se aplica aos sentidos: visão, audição, tato, etc. Idealmente, estes tópicos deveriam ser tratados de forma integrada, combinando os dados para gerar soluções completas. No entanto, os atuais avanços

em tecnologias sensitivas e técnicas de processamento de sinais ainda não permitem uma abordagem com este nível de abrangência.

Uma das relevantes linhas de pesquisa estudadas na área de interação natural, que é a área que se enquadra este trabalho, é a interpretação de movimentos corporais. Segundo Aggarwal e Cai (1999), a relevância da comunicação corporal é percebida desde as primeiras investigações da área de IHC, setores de pesquisa se dedicam a estudos na área de rastreamento do corpo humano desde o início da década de 1980. Desde então, os métodos que se propõem a tal tarefa passaram a abranger diferentes formas de perceber o corpo do usuário.

Uma das possibilidades exploradas é o uso de dispositivos e acessórios anexos ao corpo para auxiliar o rastreamento. Neste sentido, há dispositivos que possuem sensores inerciais associados como acelerômetros e giroscópios e assim conseguem medir o deslocamento e rotação de partes do corpo. Por outro lado também existem acessórios com cores específicas e/ou emissores de luz que fornecem as informações necessárias a uma unidade de captura.

Há também métodos que dispensam o auxílio de acessórios, deixando o corpo do usuário mais livre. Dentro deste escopo existem técnicas monoculares e mais recentemente algumas que realizam o rastreamento através de sensores de profundidade. Desta forma, em termos de uso e aplicabilidade, as técnicas de rastreamento do corpo humano podem ser classificadas como apresentado na Figura 1.

Figura 1 – Classificação de técnicas de rastreamento do corpo humano a partir de critérios de uso e cenários de aplicabilidade, obtida de Figueiredo et al. (2012)

Uma das formas de capturar os movimentos corporais é através do uso de acessórios ou dispositivos anexos ao corpo. Luvas táteis, e sensores, são capazes de identificar informações sobre a posição e a orientação de partes do corpo. Estes equipamentos podem estar ligados a uma unidade de processamento através ou não de fios. De forma geral, fornecem informações precisas e em alta taxa de atualização.

Outra forma de capturar os movimentos do corpo é através do rastreamento de acessórios detectáveis por câmeras comuns ou infravermelhas. Neste caso, no lugar de sensores presos ao corpo do usuário, sinais visuais são fornecidos e capturados, facilitando o rastreamento, como por exemplo, luvas com cores destacadas, roupas cobertas de luzes

infravermelhas. Para Figueiredo et al. (2012), os sinais visuais funcionam como dicas para o processo do rastreamento, pois ao se introduzir acessórios de aspecto diferenciado no ambiente, se torna mais fácil detectá-lo e distinguir as partes interativas da cena.

Existem alternativas que não recorrem a acessórios ou dispositivos extras ligados ao usuário. Técnicas de rastreamento neste sentido podem usar câmeras comuns, tecnicamente chamadas de monoculares, ou outros sensores mais elaborados, capazes de recuperar as informações de profundidade do ambiente.

A utilização de dispositivos tais como: luvas de dados, sensores eletromagnéticos, roupas com marcadores, entre outros, simplificam a identificação de gestos, por outro lado, o usuário precisa de um dispositivo, que vai contra o conceito de liberdade em interação natural. Nestes casos, o usuário precisa constantemente dedicar parte de sua atenção em como está segurando o aparelho, se está apontando-o na direção correta e até se preocupar em não lançá-lo contra outros objetos. Limitando o usuário durante a interação, tanto em relação ao aspecto físico quanto ao intelectual.

O uso de sensores de profundidade, apesar de ser algo inovador e com diversas vantagens, demonstram limitações em relação à iluminação devido ao uso de luz infravermelha. O reconhecimento do padrão emitido é bastante prejudicado caso haja incidência de luz solar através de janelas ou vãos abertos no ambiente (FIGUEIREDO et al., 2012).

De um modo geral, como afirma Júnior (2010), embora haja um grande interesse e volume de trabalho envolvendo diversas áreas de estudo de interação através de gestos, não se trata de um problema único, de forma que um grande número de soluções coexistem e estão em constante evolução, em domínios, aplicações e com limitações distintas, como restrições aos tipos de gestos que se pode usar, ao conjunto de gestos possíveis ou às condições de uso.

2.2 Processamento de Imagens

O uso de câmeras para detecção da interação do usuário com uma aplicação tem sido muito comum em sistemas interativos, em especial nas aplicações baseadas em sensoriamento sem toque, onde, além do uso da câmera, algoritmos de processamento de imagens são utilizadas para o tratamento e obtenção de informações relevantes da cena.

A área de processamento de imagens digitais tem atraído grande interesse nas últimas décadas. A evolução da tecnologia digital aliada ao desenvolvimento de novos algoritmos, capazes de processar sinais bidimensionais, vem permitindo uma gama de aplicações cada vez maior (MORALES; CENTENO; MORALES, 2003), como por exemplo, na medicina principalmente na ajuda de diagnósticos, cirurgia guiada por computador, em geo-processamento, radares de trânsito, sensoriamento remoto na visualização do clima

de uma determinada região, na arquitetura e nas engenharias (elétrica, civil, mecânica), entre outros (MORGAN, 2008).

Para Gonzalez R. C. e Woods (2010), o interesse em métodos de processamento de imagens digitais decorre de duas áreas principais de aplicação: melhoria da informação visual para a interpretação humana e processamento de dados para percepção automática através de máquinas. Segundo Grando (2005), na abordagem de Gonzáles e Woods, a primeira categoria concentra-se em técnicas para melhora de contraste, realce e restauração de imagens danificadas. A segunda categoria concentra-se em procedimentos para extrair de uma imagem informação de forma adequada, para o posterior processamento computacional. É na segunda categoria, ou seja, na percepção automática por máquinas, que se enquadra o trabalho aqui descrito.

2.2.1 Passos fundamentais em processamento de imagens

Uma imagem pode ser definida como uma forma compacta de representar muitas informações. Em um sistema de processamento de imagens, essas informações podem passar por diversas etapas, as quais descrevem o fluxo das informações com um dado objetivo definido pela aplicação (GONZALEZ R. C. E WOODS, 2010). A estrutura funcional completa de um sistema de processamento de imagens é ilustrada na Figura 2.

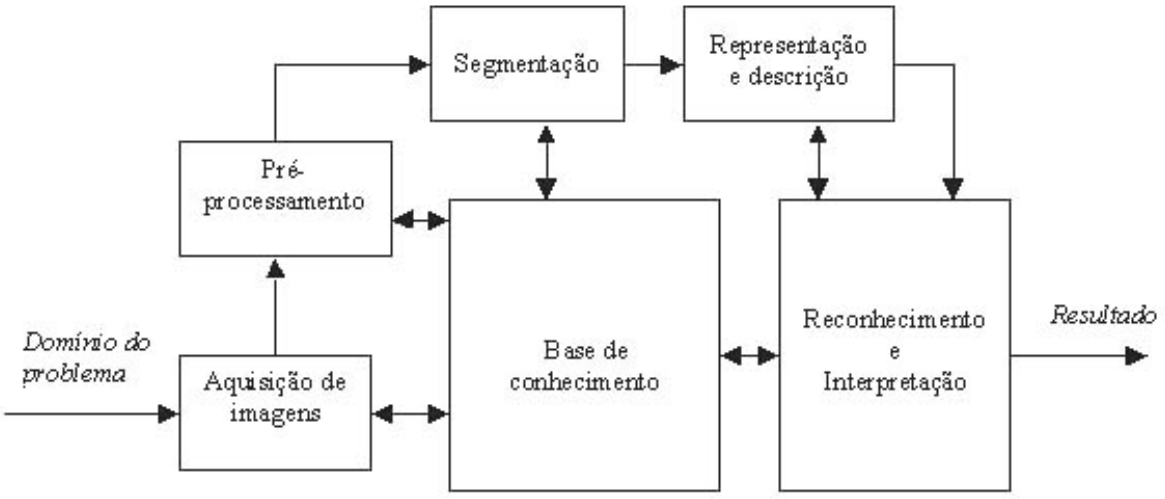


Figura 2 – Estrutura funcional completa de um sistema de processamento de imagens, adaptada de Gonzalez R. C. e Woods (2010).

Como afirma Gonzalez R. C. e Woods (2010), o diagrama não significa que todo processo se aplique a uma imagem, pois as metodologias podem ser aplicadas a imagens para diferentes propósitos, com diferentes objetivos. Sendo assim, segue uma descrição das etapas mais comuns. As descrições destas etapas foram norteadas principalmente por Facon (2002) e Gonzalez R. C. e Woods (2010), sendo as seguintes:

- Aquisição da imagem: consiste em adquirir uma imagem através de um sensor e transformá-la em uma imagem digital, sobre a forma de uma tabela de valores discretos inteiros chamados de *pixel* (FACON, 2002). Dentre os aspectos envolvidos neste passo pode-se mencionar: a escolha do tipo do sensor, o conjunto de lentes a utilizar, as condições de iluminação da cena, os requisitos de velocidade da aquisição, a resolução e o número de níveis de cinza da imagem digitalizada, entre outros (GONZALEZ R. C. E WOODS, 2010);

- Pré-processamento: a imagem resultante do passo anterior pode apresentar diversas imperfeições, tais como presença de *pixels* ruidosos, contraste e/ou brilho inadequado, regiões interrompidas ou indevidamente conectadas, entre outras. Assim, a função do pré-processamento é melhorar a imagem de forma a aumentar as chances para o sucesso dos processos seguintes. Este passo envolve técnicas para filtragem e realce, remoção de ruído, compressão, e etc. (GONZALEZ R. C. E WOODS, 2010). O pré-processamento não é indispensável, mas necessário na maioria dos casos (FACON, 2002);

- Segmentação: consiste em dividir uma imagem em partes ou objetos constituintes, ou seja, nos objetos de interesse que compõem a imagem. A segmentação é efetuada pela detecção de descontinuidades (contornos) e/ou de similaridades (regiões) na imagem (FACON, 2002). Em geral, a segmentação automática é uma das tarefas mais difíceis no processamento de imagens digitais. Por um lado, um procedimento de segmentação de imagens bem-sucedido aumenta as chances de sucesso na solução de problemas que requerem que os objetos sejam individualmente identificados. Por outro lado, algoritmos de segmentação inconsistentes quase sempre acarretam falha no processamento (GONZALEZ R. C. E WOODS, 2010);

- Representação e descrição: o alvo da representação é elaborar uma estrutura adequada, agrupando os resultados das etapas precedentes (FACON, 2002). A representação pode ser por fronteira e/ou regiões. A representação por fronteira é adequada quando o interesse se concentra nas características externas (cantos ou pontos de inflexão). A representação por região é adequada quando o interesse se concentra nas propriedades internas (textura ou forma do esqueleto) (GONZALEZ R. C. E WOODS, 2010). O processo de descrição, também chamado de seleção de características, procura extrair características que resultam em informação quantitativa ou que sejam básicas para a discriminação entre classes de objetos (GONZALEZ R. C. E WOODS, 2010);

- Reconhecimento e interpretação: reconhecimento é o processo que atribui um rótulo a um objeto, baseado na informação fornecida pelo descritor. A interpretação envolve a atribuição de significado a um conjunto de objetos reconhecidos (GON-

ZALEZ R. C. E WOODS, 2010). É o passo mais elaborado do processamento de imagens digitais, pois permite obter a compreensão e a descrição final do domínio do problema, fazendo uso do conhecimento a priori e do conhecimento adquirido durante as fases precedentes (FACON, 2002);

- Base de conhecimento: o processamento de imagens digitais pressupõe a existência de conhecimento prévio sobre o domínio do problema, armazenado em uma base de conhecimento, cujo tamanho e complexidade variam dependendo da informação. Embora nem sempre presente, a base de conhecimento guia a operação de cada módulo do processamento, controlando a interação entre os módulos (GRANDO, 2005).

É possível perceber, à medida que se passa por níveis crescentes de abstração, que ocorre uma redução progressiva da quantidade de informações manipuladas. Na aquisição da imagem e no pré-processamento, os dados de entrada são *pixels* da imagem original e os dados de saída representam propriedades da imagem na forma de valores numéricos associados a cada *pixel*. Na segmentação, representação e descrição, esse conjunto de valores produz como resultado uma lista de características. O reconhecimento e a interpretação produzem, a partir dessas características, uma interpretação do conteúdo da imagem (FACON, 2002).

2.2.2 Fundamentos de imagens digitais

Para Gonzalez R. C. e Woods (2010), uma imagem pertencente a uma cena pode ser definida como uma função bidimensional $f(x, y)$, composta por um determinado número de *pixels*, de modo que cada *pixel* deva possuir coordenadas de localizações x e y, associadas a um valor específico. É importante notar que o processamento de uma imagem depende diretamente destes valores. O termo *pixel* é uma abreviatura do inglês *picture element* que significa elemento da figura, corresponde a menor unidade de uma imagem digital, onde são descritos a cor e o brilho específico de uma célula da imagem (MORGAN, 2008).

Normalmente, uma imagem capturada por uma câmera de vídeo é apresentada em cores, assim, cada *pixel* da imagem deve ser formado por um conjunto de valores, ou também conhecido como canais de cores, geralmente de tamanho três ou quatro, podendo pertencer aos padrões de cores RGB (*red*, *green* e *blue*) ou $CMYK$ (*cyan*, *magenta*, *yellow* e *key=black*), respectivamente.

A combinação dos canais de cores pode representar uma grande variedade de cores, no entanto, muitas vezes essa quantidade de informação pode ser desnecessária para os objetivos de determinadas aplicações, de modo que seu processamento possa levar a desperdício de recurso. Uma forma para a resolução deste problema é a utilização de

uma técnica de transformação para tons de cinza, considerado um processo simples sob os canais de cores. Aqui consideramos o padrão RGB, uma vez que este é o padrão de cores utilizado neste trabalho. A Equação 2.1, apresentada a seguir, descreve a transformação dos canais de um *pixel* i, pertencente ao padrão RGB, para tons de cinza, como segue:

$$I_i = \frac{R_i + B_i + B_i}{3} \tag{2.1}$$

na qual, é calculada a média aritmética dos três canais de cores, R, G, e B do *pixel* i. I_i é o valor do tom de cinza obtido, também conhecido como valor de intensidade, representando o *pixel* i por um único valor.

Como afirma Grando (2005), uma imagem digital $f(x, y)$ pode ser representada por uma matriz, cujos índices de linha e coluna identificam um ponto *(pixels)* da imagem e representam o conjunto de valores (canais de cores). Por exemplo, a Figura 3 representa uma imagem digital de 4 *pixels* de largura por 4 *pixels* de altura, e a representação matricial, cujos elementos são dados pelas intensidades dos *pixels* nas posições correspondentes.

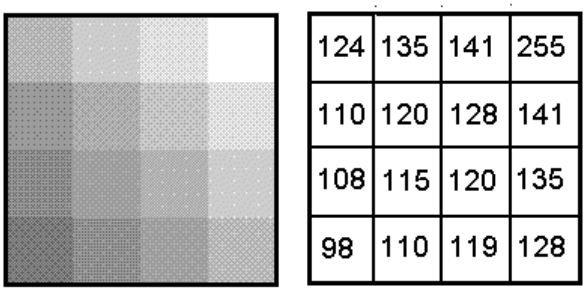


Figura 3 – Imagem digital e a representação matricial, obtida de Almeida, A. B. (1998).

2.2.3 Atenção Visual

A quantidade de informação visual e sonora disponível para ser processada pelos seres vivos é quase sempre muito grande. A capacidade de selecionar consciente ou inconscientemente determinados estímulos, sejam visuais, sonoros ou outros, dentre uma grande variedade de estímulos é essencial e intrínseca à maioria dos seres vivos. Essa capacidade biológica de atenção visual, ou sonora, nos faz capazes de reagir rapidamente a alterações no ambiente.

No caso de estímulos visuais, Benicasa (2013) afirma que alguns estímulos são naturalmente conspicuosos ou salientes em um determinado contexto. Como por exemplo, um observador dará atenção automaticamente e involuntariamente a uma jaqueta vermelha posicionada entre vários ternos pretos.

Tendo-se como base estudos em seres humanos e macacos, pode-se afirmar que o processo de seleção visual seleciona apenas um subconjunto da informação sensorial

disponível, na forma de uma região circular do campo visual, conhecida como foco de atenção (BENICASA, 2013). Desta forma, segundo Shic e Scassellati (2007), a atenção auxilia na redução da quantidade de informação, que resulta de todas as combinações possíveis dos estímulos sensoriais pertencentes a uma cena, pois apenas informações que estão dentro da área da atenção são processadas, enquanto que o restante é suprimido (CAROTA; INDIVERI; DANTE, 2004). Considerado isso, Itti (2005) define atenção visual como um eficiente mecanismo para reduzir tarefas complexas, como análise de uma cena, em um conjunto de sub-tarefas menores.

Wolfe e Horowitz (2004) demostraram que algumas características como cor, orientação ou tamanho dos objetos em uma imagem são responsáveis por guiar o mecanismo biológico de atenção visual. Para o entendimento do processo de atenção visual é importante observar que a busca por um ponto de maior atenção ou saliência pode ser simples e eficiente em alguns casos, porém não tão simples para outros (BENICASA, 2013).

A Figura 4 mostra algumas destas características. Na Figura 4(a), o contraste entre o azul e o vermelho ressalta a existência de um numeral 5 (cinco) de cor diferente dos demais. No entanto, perceber um número cinco azul e maior é um pouco mais complicado. A Figura 4(a) também é um exemplo da importância de conhecimento a priori para executar determinadas buscas visuais, pois dificilmente é possível identificar o número dois existente nesta imagem sem que alguém tenha dito que há um número dois. As Figuras 4(b) e 4(c) demonstram a importância da orientação e do contraste de cores para ressaltar objetos diferentes em imagens. Na Figura 4(b) é difícil encontrar os pares de triângulos horizontais, mas esta tarefa é simplificada devido ao contraste de cores entre os retângulos azuis e os retângulos rosas. Na Figura 4(d), a busca por cruzes é ineficiente devido ao fato de que aqui a informação de intersecção não guia a atenção.

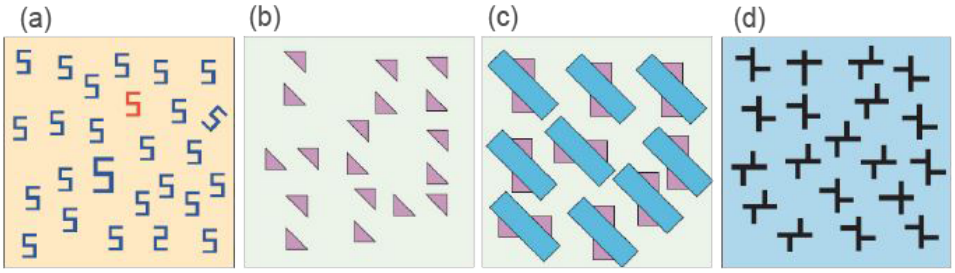


Figura 4 – Exemplos de busca visual, obtida de Wolfe e Horowitz (2004).

Como afirma Pereira (2007), em uma visão didática, podem ser identificados dois métodos principais para obtenção da atenção visual. Os métodos *top-down* e *bottom-up*. O método *top-down* usa conhecimentos obtidos a priori para detectar regiões de maior interesse numa imagem. Esses conhecimentos podem ser obtidos de várias formas. Geralmente, utilizam-se ferramentas de aprendizagem baseadas em modelos geométricos/relacionais (como redes semânticas ou grafos relacionais) ou modelos estatísticos (como redes

neurais e máquinas de vetores de suporte). Porém, esses conhecimentos também podem ser fornecidos por um ser humano, selecionando-se manualmente regiões de maior interesse numa imagem. A atenção visual *bottom-up* é guiada por características primitivas da imagem como cor, intensidade e orientação. Além disso, ela atua de modo inconsciente, ou seja, o observador é levado a fixar sua atenção em determinadas regiões da imagem devido aos estímulos causados pelos contrastes entre características visuais presentes na imagem.

O sistema de atenção visual *bottom-up* proposto por Itti (1998) é um dos mais conhecidos e utilizado atualmente para seleção de regiões salientes em imagens. A seguir será descrito os principais aspectos desse modelo.

2.2.3.1 Modelo de Saliência

Na Figura 5, uma adaptação de Itti (1998), é apresentado o modelo do mapa de saliências. O modelo pode ser descrito nas seguintes etapas: extração de características, filtragem linear, diferenças centro-vizinhança, normalização e soma dos mapas de características. A imagem de entrada é decomposta em três mapas de características: intensidade, cor e orientação. Esses mapas são criados através de pirâmides de Gauss e Gabor, através de sucessivas filtragens e sub-amostragens da imagem de entrada (ITTI, 1998).

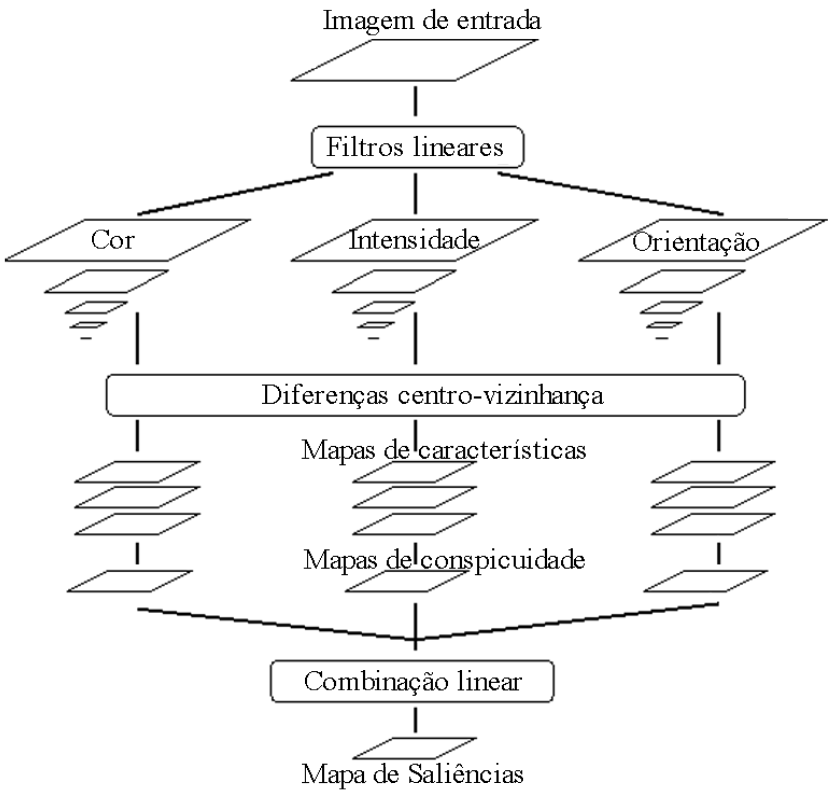


Figura 5 – Arquitetura do modelo do mapa de saliências, adaptada de Itti (1998).

Para o entendimento da geração de um mapa de saliência, será descrito a seguir

os principais aspectos do modelo proposto em Itti e Koch (2001).

2.2.3.2 Extração de Características Visuais Primitivas

Para gerar um mapa de saliência, três tipos de características visuais primitivas são extraídas: cor, intensidade e orientação. Em seguida, quatro canais de cores são criados (R para vermelho, G para verde, B para azul e Y para amarelo). Sendo r, g, b os canais vermelho, verde e azul da imagem de entrada, os canais de cores são representados por (BENICASA, 2013):

$$R = \frac{r - (g + b)}{2} \tag{2.2}$$

$$G = \frac{g - (r + b)}{2} \tag{2.3}$$

$$B = \frac{b - (r + g)}{2} \tag{2.4}$$

$$Y = \frac{(r + g)}{2} - \frac{|r - g|}{2} - b \tag{2.5}$$

A imagem de intensidades é representada por $I = (r+g+b)/3$, que define a imagem em tons de cinza. A Figura 6 apresenta um exemplo de extração das características.

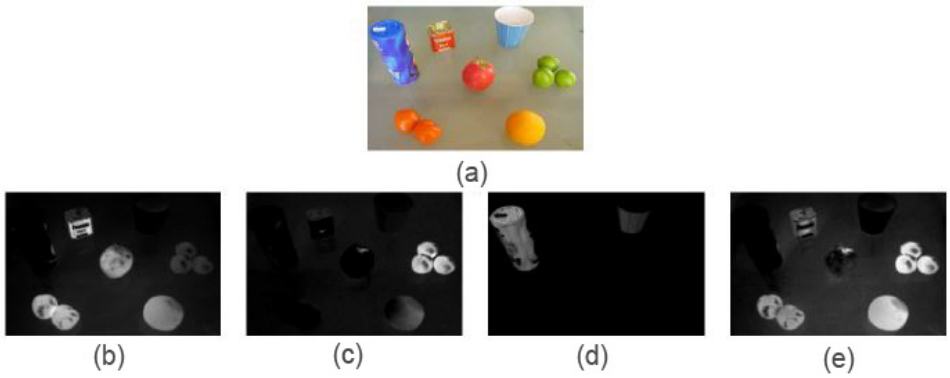


Figura 6 – Extração de 4 canais de cores. a) Imagem de Entrada, b) Extração do canal vermelho, c) Canal verde, d) Canal azul e e) Canal amarelo (SIKLOSSY, 2005 apud BENICASA, 2013)

Os canais de cores e a imagem de intensidades são submetidos a um processo de filtragem linear. Este processo é realizado por meio da geração de Pirâmides Gaussianas e Pirâmides Direcionais. A Pirâmide Gaussiana é composta por versões filtradas passa-

baixa da convolução [1] Gaussiana aplicada à imagem de entrada. A Pirâmide Direcional é uma decomposição multi-escala e multi-orientação de uma imagem. Nesta decomposição linear, uma imagem é subdividida em um conjunto de sub-bandas localizadas em escala e orientação. A representação piramidal é usada para a obtenção de amostras da imagem sem detalhes indesejáveis. A seguir, os processos de geração das Pirâmides Gaussianas e Direcionais são detalhados.

2.2.3.3 Pirâmide Gaussiana

As Pirâmides Gaussianas são geradas utilizando um algoritmo proposto por Burt e Adelson, a imagem de entrada é representada por uma matriz G_0, essa matriz contém c colunas e r linhas de *pixels*. Para cada nível da pirâmide é gerada uma imagem em uma escala menor que a escala no nível superior. A imagem de entrada é a base ou nível zero da Pirâmide Gaussiana. Cada nível inferior da pirâmide contém uma imagem que é uma redução ou uma versão filtrada passa-baixa da imagem da base da pirâmide (BENICASA, 2013). Uma pirâmide Gaussiana G_θ pode ser definida recursivamente como segue:

$$G_\theta(x,y) = \sum_{m=-2}^{2}\sum_{n=-2}^{2} w(m+2,n+2)\ G(x,y), \quad para\ \theta = 0 \tag{2.6}$$

$$G_\theta(x,y) = \sum_{m=-2}^{2}\sum_{n=-2}^{2} w(m+2,n+2)\ G_{\theta-1}(2x+m,2y+n), \quad para\ 0 < \theta \leq 8 \tag{2.7}$$

onde w(m; n) são os pesos gerados a partir de uma função Gaussiana, utilizados para gerar os níveis da pirâmide para todos os canais. A Figura 7 mostra um exemplo da Pirâmide Gaussiana.

Figura 7 – Representação piramidal, usada para a obtenção de amostras da imagem sem detalhes indesejáveis, obtida de Itti (2000)

[1] Filtragem de uma imagem de forma que o valor de cada *pixel* seja determinado por uma média ponderada dos valores dos *pixels* vizinhos, sendo o valor da ponderação determinado por um operador matricial, tendo como resultado um efeito de borramento na imagem.

2.2.3.4 Pirâmide Direcional

O modelo de (ITTI, 1998) também considera informações sobre orientações locais como uma característica importante para o desenvolvimento da atenção visual. Na Figura 8 é apresentado um exemplo em que o contraste na orientação pode guiar a atenção visual.

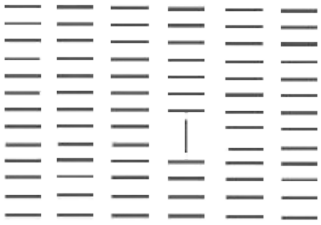


Figura 8 – Exemplo de orientação, barra vertical inserida em um ambiente com barras horizontais torna-se o elemento mais saliente devido a grande diferença de orientação (SIKLOSSY, 2005 apud BENICASA, 2013)

Os mapas de orientações $O_\theta(\theta)$ são criados através da convolução do mapa de intensidades I_θ, com filtros direcionais de Gabor para quatro orientações $\theta \in 0°, 45°, 90°, 135°$. A aplicação destes filtros visa identificar barras ou bordas em uma determinada direção, para isso utiliza-se de uma função gaussiana (BENICASA, 2013).

2.2.3.5 Diferenças Centro-Vizinhança

Os mapas de características são obtidos por meio da diferença entre canais de cores em diferentes escalas, este processo é conhecido como diferença centro-vizinhança. Nesta subtração de imagens, o centro é um *pixel* da imagem em uma escala $c \in \{2, 3, 4\}$ e a vizinhança é o *pixel* correspondente de outra imagem em uma escala $s = c + \sigma$ com $\sigma \in \{3, 4\}$ da pirâmide (PEREIRA, 2007).

A subtração destas duas imagens é obtida pela interpolação das imagens para a escala fina, seguida da subtração ponto a ponto (BENICASA, 2013). O primeiro conjunto de mapas é construído a partir do contraste de intensidades, definido como segue:

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)| \tag{2.8}$$

que apresenta inspiração biologicamente baseada nos mamíferos, onde o contraste de intensidade é detectado por neurônios sensíveis a centros escuros com vizinhança clara e por neurônios sensíveis a centros claros com vizinhança escura (ITTI; KOCH, 2001). O segundo conjunto de mapas é similarmente construído a partir dos canais de cores, definidos como:

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \tag{2.9}$$

$$\mathcal{BY}(c,s) = \left|(B(c) - Y(c)) \ominus (Y(s) - B(s))\right|, \tag{2.10}$$

a inspiração biológica para a construção desse conjunto de mapas é a existência, no córtex visual, do chamado sistema de cores oponentes: no centro de seus campos receptivos, neurônios são excitados por uma cor e inibidos por outra e vice-versa. Tal sistema existe para vermelho=verde, verde=vermelho, azul=amarelo, amarelo=azul (ITTI; KOCH, 2001). O terceiro conjunto de mapas é construído a partir de informações de orientação local, de acordo com as seguintes equações:

$$\mathcal{O}(c,s,\theta) = \left|O(c,\theta) \ominus O(s,\theta)\right|, \tag{2.11}$$

na qual, $\theta \in 2,3,4$. Neste caso, a inspiração biológica para a construção dos mapas de orientação é a propriedade de neurônios do sistema visual de responder apenas a uma determinada classe de estímulos, como por exemplo barras orientadas verticalmente (ITTI; KOCH, 2001).

2.2.3.6 Saliência

Segundo Benicasa (2013), a maioria dos modelos de atenção bottom-up inspirados biologicamente segue a hipótese de Koch and Ullman (1985), onde vários mapas de características alimentam um único mapa mestre ou mapa de saliência.

O mapa de saliência é um mapa escalar bidimensional de atividade representada topograficamente pela conspicuidade ou saliência visual (ITTI; KOCH, 2001). Uma região ativa em um mapa de saliência codifica o fato desta região ser saliente, não importando se esta corresponde, por exemplo, a uma bola vermelha meio a bolas verdes, ou a um objeto que se move para a esquerda enquanto outros se movem para a direita (BENICASA, 2013).

Para a construção de um único mapa de saliência, os mapas de características são individualmente somados ($\oplus$) nas diversas escalas, gerando três mapas de conspicuidades: $\widetilde{I}$ para intensidade, $\widetilde{C}$ para cor e $\widetilde{O}$ para orientação. Entretanto, um fator importante a ser notado é que, previamente à somatória dos mapas de cada característica, Itti (1998) propõem sua normalização, denotada por *N(.)*, com o objetivo de que uma região que apresente um nível de saliência contrastante com as demais seja amplificada e, por outro lado, regiões salientes não contrastantes sejam mutuamente inibidas. A Figura 9 demonstra a função da normalização *N(.)*.

Após o processo de normalização, os mapas de características são então combinados em três mapas de conspicuidades, conforme descrito anteriormente, definidos como segue:

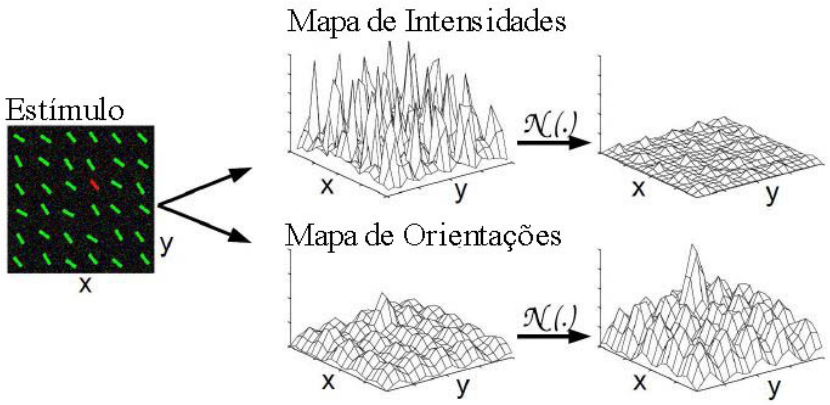


Figura 9 – Exemplo do comportamento do operador de normalização *N(.)*

$$\bar{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s)), \tag{2.12}$$

$$\bar{\mathcal{C}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))], \tag{2.13}$$

$$\bar{\mathcal{O}} = \sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c,s,\theta))\right) \tag{2.14}$$

De acordo com Itti (1998), a motivação para a criação dos três canais separados ($\widetilde{I}$,$\widetilde{C}$, $\widetilde{O}$) é a hipótese de que características similares competem pela saliência, enquanto modalidades diferentes contribuem independentemente para o mapa de saliência.Finalmente, os três mapas de conspicuidades são normalizados e somados, resultando em uma entrada final para o mapa de saliência S, como segue:

$$\mathcal{S} = \frac{1}{3}(\mathcal{N}(\bar{\mathcal{I}}) + \mathcal{N}(\bar{\mathcal{C}}) + \mathcal{N}(\bar{\mathcal{O}})) \tag{2.15}$$

2.3 Segmentação de Imagens

Segundo Morgan (2008), a segmentação de imagens representa um processo fundamental para extração e identificação de objetos ou áreas de interesse da imagem, bem como permite reduzir a informação da mesma a fim de manipular objetos simples de uma cena que geralmente correspondem a linhas ou regiões (grupos de pontos conectados) de interesse.

Para Gonzalez R. C. e Woods (2010), o alvo da segmentação é subdividir uma imagem, digitalizada e pré-processada, em partes ou objetos constituintes. O nível até o qual essa subdivisão deve ser realizada depende do predicado da imagem e/ou do problema a ser resolvido, ou seja, a segmentação deve parar quando os objetos de interesse

tiverem sido isolados. Sendo assim, o procedimento de segmentação deve se concentrar nas características do objeto, descartando o restante da imagem (RUSS, 2011).

Tarefas de segmentação de imagens são recorrentes em sistemas de visão computacional, na qual, deseja-se identificar regiões de acordo com certas características, constituindo geralmente uma etapa intermediária de um sistema maior, em que os resultados da segmentação serão utilizados para outros procedimentos (KöRBES, 2010).

Para Lourega (2006), a grande dificuldade da segmentação reside no fato de não se conhecer, inicialmente, o tipo de estrutura que se encontra numa imagem. Essas estruturas são identificadas a partir da natureza da imagem (iluminação, presença de ruídos, textura, contornos, oclusões), das primitivas a serem extraídas (contornos, segmentos retos, regiões, formas, texturas) e das limitações físicas (complexidade algorítmica, execução em tempo real, memória disponível). É com base nessas informações que são escolhidas, em princípio, estruturas que possibilitam a melhor aplicação.

Segundo Gonzalez R. C. e Woods (2010), os algoritmos de segmentação para imagens monocromáticas são baseados em duas propriedades básicas de níveis de cinza: descontinuidade e similaridade. A descontinuidade refere-se a partição da imagem segundo mudanças bruscas nos níveis de cinza, como na detecção de pontos, linhas e bordas. Já a similaridade baseia-se na na similaridade entre *pixels*, seguindo um determinado critério, como por exemplo, limiarização, crescimento de regiões, divisões e fusões de regiões. Para este trabalho, é utilizada a técnica de crescimento de regiões, na seção a seguir são destacadas algumas técnicas baseadas em similaridades e a justificativa da escolha.

2.3.1 Técnicas Baseadas em Similaridades

2.3.1.1 Limiarização - *Thresholding*

De acordo com Marques O. F. e Vieira (1999), o objetivo da limiarização é identificar duas classes distintas na imagem, por meio do uso de um limiar para dividir a imagem em dois conjuntos de *pixels*. Considerando um limiar T, qualquer ponto x, y na imagem, tal que $f(x, y) > T$, será chamado de ponto do objeto, caso contrário, o ponto será chamado ponto de fundo (GONZALEZ R. C. E WOODS, 2010). Para Marques O. F. e Vieira (1999), esse processo também é conhecido como binarização, pois tem como resultado uma imagem binária, composta por *pixels* brancos e pretos. O processo de limiarização é descrito como segue:

$$lim(x, y) = \begin{cases} 1 & f(x, y) \geq T \\ 0 & f(x, y) < T \end{cases} \tag{2.16}$$

Segundo Grando (2005), os métodos de limiarização têm duas abordagens distintas, uma global e outra local. O método global utiliza um único limiar T para toda imagem, já

o método local têm como princípio dividir a imagem em sub-regiões, cada região tem seu limiar. Além disto estes métodos são classificados em dois grupos: manual e automático. O método manual é baseado na disposição dos níveis de cinza no histograma, sendo a escolha do limiar feita de forma empírica por um operador humano. No método automático, também baseado no histograma, não há necessidade da escolha do valor de limiar, uma vez que os próprios algoritmos retornam esse valor. A Figura 10 ilustra a influência do valor do limiar sobre a qualidade da limiarização.

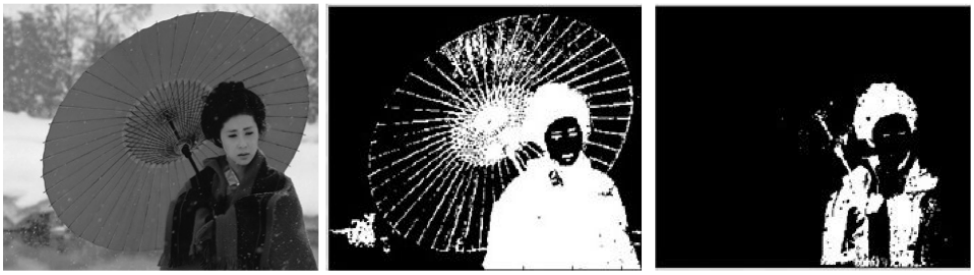


Figura 10 – Influência do valor do limiar sobre a qualidade da limiarização, da esquerda para a direita, a imagem original, seguida pela mesma imagem aplicando-se um limiar 30 e 10, respectivamente, obtida de Melo, N. (2009).

A desvantagem dessa técnica, que combina o histograma com a limiarização, é que a mesma não resolve todos os problemas de segmentação, pois não leva em consideração, por exemplo, a forma dos objetos na imagem, isto é, dois objetos de formatos diferentes podem ser indistinguíveis usando-se esta técnica.

2.3.1.2 Crescimento de Regiões - *Region Growing*

Esta técnica de segmentação encontra regiões diretamente na imagem agrupando *pixels* ou sub-regiões em regiões maiores baseado em critérios de crescimento pré-definidos. O procedimento parte de um conjunto de pontos, chamados de sementes, e, a partir destes pontos vai agrupando pontos utilizando uma vizinhança de influência, formando as regiões. Nesta vizinhança são analisadas propriedades e são medidas similaridades para determinar se o *pixel* faz parte ou não da região sendo considerada. As propriedades normalmente consideradas são: cor, intensidade de nível de cinza, textura, momentos, etc. (MARENGONI; STRINGHINI, 2009).

Os dois principais métodos de segmentação por crescimento de regiões são o *Watersheds* (divisor de águas) e o Algoritmo do Funcional de Mumford e Shah. Como afima Wangenheim (2011), destes dois, o *Watershed* é o mais rápido, permitindo que seja utilizado em aplicações interativas, mesmo quando as imagens a serem processadas são grandes e complexas. Por esse motivo esse método foi utilizado neste trabalho e será abordado na seção seguinte.

2.3.1.3 *Watershed*

Esse método baseia-se no princípio de "inundação de relevos topográficos" (RUSS, 2011). Ao considerar uma superfície topográfica e gotas de água caindo sobre tal superfície. Uma bacia de retenção é dada pela região tal que, quando gotas de água caem sobre tal região, estas percorrem o relevo da superfície, através de fluxos descendentes, até chegar a um mesmo mínimo da superfície. As linhas que separam diferentes bacias de retenção são chamadas de *watersheds* (KLAVA, 2000).

Uma formulação alternativa e que é facilmente expressa na forma de algoritmo para a transformação *watershed* é baseada na simulação de inundação: considerando a imagem de entrada em níveis de cinza como uma superfície topográfica, o objetivo é produzir linhas de divisão de águas nesta superfície. Para tal, um furo é feito em cada mínimo regional [2] m_k da superfície, que é, então, submersa a uma taxa constante, de modo que a água entre pelos mínimos regionais. Quando frentes de água, vindas de diferentes mínimos regionais, estão prestes a se encontrar, uma barreira é construída para evitar tal encontro. Em algum momento, o processo chega a um estado tal que somente os topos das barreiras estão visíveis acima do nível da água, correspondendo às linhas de *watershed* (KLAVA, 2000). Na Figura 11 é apresentada a analogia do funcionamento desse método.

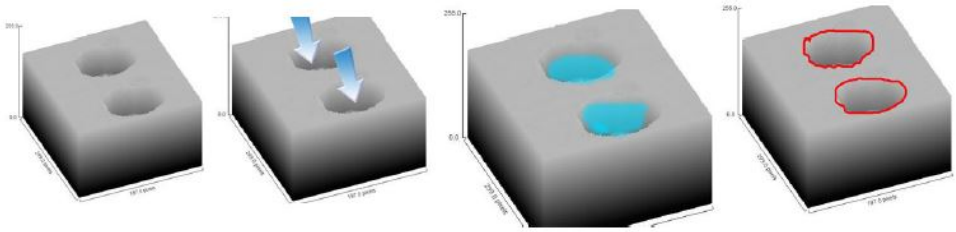


Figura 11 – Analogia do funcionamento da segmentação *Watershed*, obtida de Andrade (2011).

Como afirma Klava (2000), um problema do *watershed* clássico é a super segmentação da imagem, uma vez que o gradiente costuma apresentar muitos mínimos regionais, principalmente devido a ruído e texturas na imagem original. Uma abordagem alternativa para contornar esse problema é a utilização de marcadores, desta forma, haverá uma bacia de captação para cada marcador, podendo eliminar a ocorrência de super-segmentação (PECCINE, 2004).

No processo de inundação por marcadores, ilustrado na Figura 12, a água vai entrando pelos furos correspondentes aos diferentes marcadores, submergindo inicialmente as bacias primitivas, nas quais estão localizados os marcadores. Ao longo do processo de inundação, quando o nível da água ultrapassa a altura mínima das bordas com bacias

[2] Um mínimo regional em uma imagem em níveis de cinza é uma zona plana não adjacente a nenhuma outra zona plana com menor altitude (nível de cinza). Uma zona plana cujos *pixels* têm todos o mesmo nível de cinza.

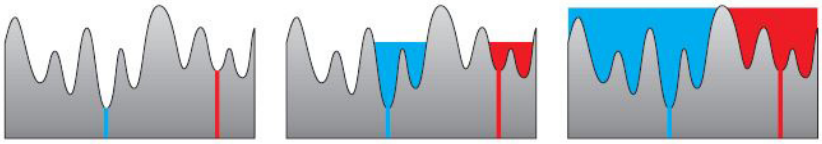


Figura 12 – Inundação por marcadores, obtida de Klava (2000).

primitivas vizinhas, estas são submersas. A inundação prossegue até que todo o relevo correspondente à imagem esteja submerso (KLAVA, 2000).

Conforme pode ser observado na Figura 13, para cada marcador obtém-se uma região de interesse na partição correspondente.

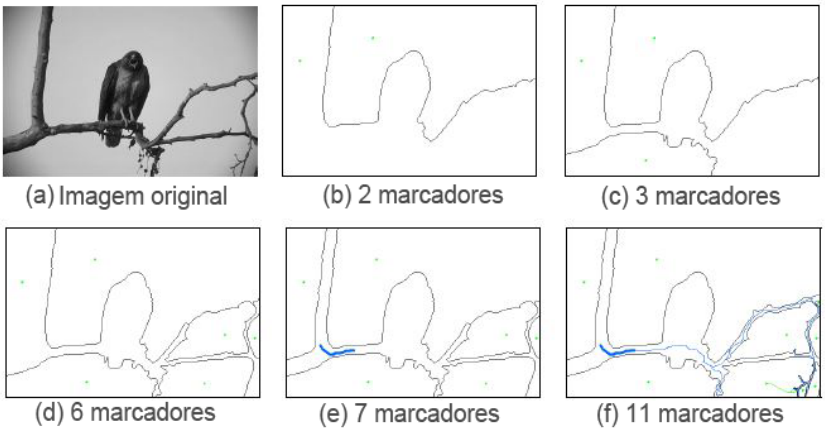


Figura 13 – Partições obtidas através da inundação por marcadores., obtida de Klava (2000).

2.4 Histograma

Para Marques O. F. e Vieira (1999), o histograma de uma imagem é composto por um conjunto de números, indicando o percentual de *pixels* contidos na imagem, que apresentam um determinado nível de cinza. Estes valores são normalmente representados por um gráfico de barras que fornece, para cada nível de cinza, o número, ou o percentual, de *pixels* correspondentes na imagem. A informação obtida através do cálculo do histograma de cores de uma imagem pode ser útil para a indicação de sua qualidade quanto ao nível de contraste, brilho médio, ou demais informações a serem utilizadas para fins específicos. De acordo com Gonzalez R. C. e Woods (2010), o cálculo do histograma de cores pode ser descrito como:

$$h(r_k) = n_k \tag{2.17}$$

em que, r_k é o k-ésimo nível de cinza e n_k é o número de *pixels* da imagem contendo o nível de cinza r_k. Na Figura 14 é apresentado um exemplo da aplicação das Equações 2.1

e 2.17.

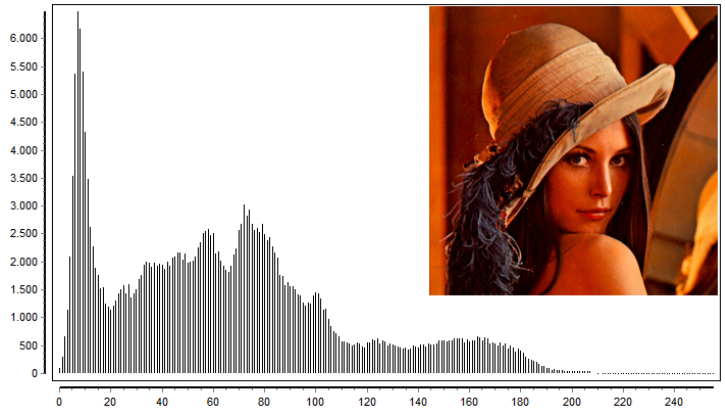


Figura 14 – Exemplo de transformação para tons de cinza e cálculo do histograma de cores. Imagem Lena (512x512*pixels*) obtida de Gonzalez R. C. e Woods (2010).

2.5 Trabalhos Relacionados

Segundo Aggarwal e Cai (1999), a relevância da comunicação corporal é percebida desde as primeiras investigações da área de IHC, setores de pesquisa se dedicam à estudos na área de rastreamento do corpo humano desde o início da década de 1980. Desde então, os métodos que se propõem a tal tarefa passaram a abranger diferentes formas de perceber o corpo do usuário.

Como visto na Seção 2.1.1, uma das possibilidades exploradas para identificação de interação é o uso de dispositivos especiais, ou acessórios anexos ao corpo para auxiliar o rastreamento. Nesse sentido, na Tabela 1 são destacados alguns trabalhos pertinentes.

Tabela 1 – Exemplos de trabalhos/dispositivos que utilizam sensores especiais ou acessórios anexos ao corpo para auxiliar na identificação de gestos.

Título do trabalho/dispositivo	Autor(es)/Fabricante	Ano
CHARADE	BAUDEL et al.	1993
Rutgers Master II-ND	POPESCU et al.	1999
EyeToy	Sony PlayStation 2	2003
Nitendo Wii	Nitendo	2006
PlayStation Eye	Sony PlayStation 3	2007
Prakash	RASKAR et al.	2007
Real-time hand-tracking with a color glove	WANG and POPOVIC	2009
PlayStation Move	Sony	2010
Kinect	MICROSOFT	2010
CyberGlove II	SYSTEMS	2010

Baudel e Beaudouin-Lafon (1993) propuseram uma aplicação denominada de CHARADE, que permite ao usuário controlar o computador durante uma apresentação com slides apenas utilizando gestos de mão. Para isto, os autores utilizaram uma luva especial, ligada a um controlador, responsável pela detecção dos movimentos, funcionamento da

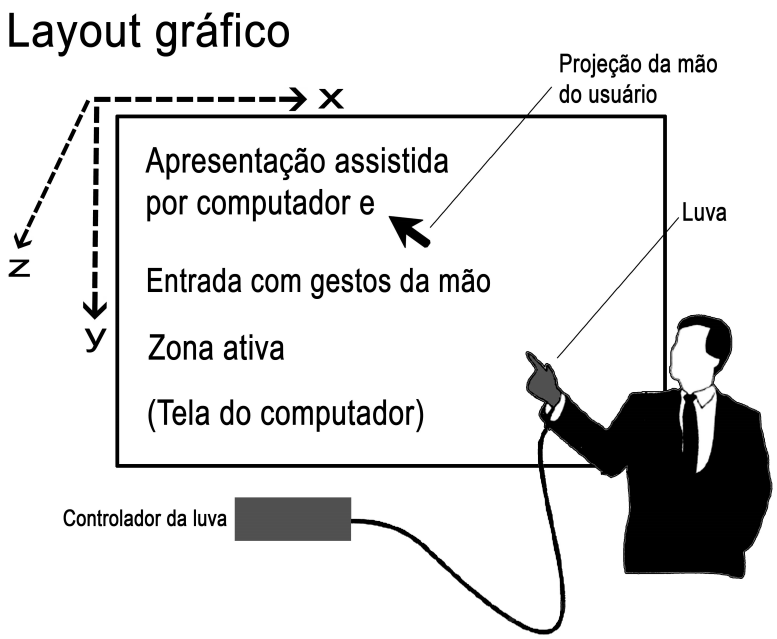


Figura 15 – Ilustração do funcionamento da aplicação CHARADE (BAUDEL; BEAUDOUIN-LAFON, 1993).

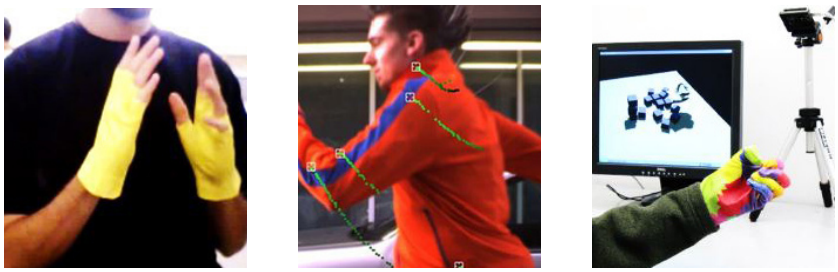
aplicação é ilustrado na Figura 15. A luva Rutgers Master II-ND (Figura 16(a)) proposta por Popescu, Burdea e Bouzit (1999), é uma interface tátil projetada para interações com ambientes virtuais, com o objetivo de aumentar o realismo da simulação durante a manipulação do objeto virtual. Um exemplo mais recente e sofisticado é a luva tátil CyberGlove II, ilustrada na Figura 16(b), possui 22 sensores espalhados entre os dedos e a palma da mão, podendo ser utilizada em uma ampla variedade de aplicações do mundo real, incluindo a avaliação digital de protótipo, realidade biomecânica virtual e animação (SYSTEMS, 2014).

(a) Luva Rutgers Master II-ND (b) Luva CyberGlove II

Figura 16 – Exemplos de luvas com sensores.

Outra forma de capturar os movimentos do corpo é através do rastreamento de acessórios detectáveis por câmeras comuns ou infravermelhas. Neste caso, no lugar de sensores presos ao corpo do usuário, sinais visuais são fornecidos e capturados, facilitando o rastreamento do corpo. Estes vão desde luvas com cores destacadas, como no trabalho

de Figueiredo et al. (2009), ilustrado na Figura 17(a), a roupas cobertas de luzes infravermelhas (Figura 17(b)) (RASKAR et al., 2007), ou a utilização de luvas coloridas, como por exemplo, no trabalho de Wang e Popovic (2009), em que uma câmera monitora a mão vestida com uma luva multi-colorida, sendo um dispositivo de entrada do usuário para aplicações de realidade virtual desktop (Figura 17(c)).



(a) (b) (c)

Figura 17 – a)Luvas com cores destacadas (FIGUEIREDO et al., 2009); b)Roupa coberta de luzes infravermelhas (RASKAR et al., 2007); c)Luvas coloridas (WANG; POPOVIC, 2009).

Como afirma Castro (2012), o mundo dos jogos eletrônicos causou grande popularização da Interação Natural com o lançamento de dispositivos que permitem a sua utilização nos jogos. Como um importante dispositivo, podemos citar o lançamento da Sony em 2003 do EyeToy para o PlayStation 2, que é um tipo de webcam usada para reconhecimento de gestos para os jogos, ilustrada na Figura 18(a). Em 2006 a Nitendo lançou o Wii, equipado com uma barra emissora de luz infravermelha e um joystick sem fio (Figura 18(b)), este último possui um acelerômetro e uma câmera de alta taxa de atualização que capta apenas sinais infravermelhos (LEE, 2008), essa nova experiência de se utilizar gestos em jogos fez o dispositivo tornar-se algo popular e desejável (EWALT, 2006).

Em 2007 a Sony lançou o PlayStation Eye, a versão do EyeToy para o seu novo console, o PlayStation 3, possuindo o dobro da sensibilidade do seu antecessor. Três anos depois a empresa lança o PlayStation Move (controles com sensores de movimentos), que utilizados juntamente com o PlayStation Eye, permitiu uma grande precisão no rastreamento dos movimentos (CASTRO, 2012). Ainda em 2010, Microsoft lança a para o seu console Xbox360, o Kinect (Figura 18(c)), que tem se destacado na área de sistemas interativos. De acordo com a MICROSOFT, com o Kinect é possível desenvolver aplicações que permitem aos usuários interagir com computadores por gestos ou fala.

(a) (b) (c)

Figura 18 – a)Sony do EyeToy, obtida de Castro (2012); b)Joystick sem fio do Wii, obtida de Castro (2012); c)Kinect, obtida de Microsoft (2013).

Atualmente, com maior acesso a dispositivos de captura de imagens e computadores com capacidades suficientes de processamento, é possível fazer uso de gestos com as mãos livres, sem a necessidade de luvas ou sensores presos ao usuário, mesmo em plataformas domésticas. Com isso, a utilização de apenas câmera para interação, que já ocorre há décadas, tem se intensificado (JúNIOR, 2010). Sendo assim, são destacadas na Tabela 2 algumas pesquisas nesse sentido.

Tabela 2 – Exemplos de trabalhos/aplicações que utilizam apenas câmera para auxiliar na interação.

Título do trabalho/Nome da aplicação	Autor(es)/Fabricante	Ano
Computer Vision for Interactive Computer Graphics	FREEMAN et al.	1998
Model Based Three Dimensional Hand Posture Recognition for Hand Tracking	ERKAN	2004
Uma Aplicação de Visão Computacional que Utiliza Gestos da Mão Para Interagir com o Computador	TRUYENQUE	2003
Câmera Kombat	PAULA et al.	2006
SixthSense: a wearable gestural interface	MISTRY et al.	2009
Modelo Abrangente e Reconhecimento de Gestos com as Mãos Livres para Ambientes 3D	BERNARDES JÚNIOR	2010
PIAI - Processamento de Imagens aplicado à Apresentação Interativa	ALMEIDA et al.	2013
Processamento de Imagens e IA aplicado à Apresentação Interativa: Uma Comparação entre um Método Interativo Tradicional e um Método Interativo Fuzzy	VASCONCELOS et al.	2014

Freeman, Anderson e Beardsley (1998) descrevem algumas técnicas simples para a interação através de visão computacional. Um desses exemplos mostra como a orientação da mão, que é o conjunto de *pixels* diferentes do fundo, é utilizada para dirigir um robô, ilustrado na Figura 19. Na pesquisa de Erkan (2004), são utilizadas duas câmeras para reconhecimento da postura da mão, construindo um modelo da mão 3D simples com formas geométricas básicas.

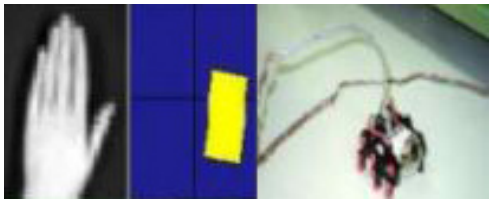


Figura 19 – Orientação da mão utilizada para dirigir um robô, obtida de Freeman, Anderson e Beardsley (1998).

No trabalho de Truyenque (2005), gestos da mão e as posições dos dedos são utilizados para simular algumas funções presentes em mouses e teclados, um exemplo é ilustrado na Figura 20(a). Paula, Bonini e Miranda (2006) propuseram o Câmera Kombat, um jogo de luta multi-jogador que, com a utilização de técnicas de visão computacional, permite a interação do jogador com o jogo sem a necessidade de utilizar dispositivos convencionais, como teclado, mouse, joysticks, entre outros. Para isto utilizam uma webcam posicionada na frente do jogador e, quando este utiliza uma sequência pré-determinada de movimentos, o personagem virtual dispara uma espécie de "magia" no outro competidor, é ilustrado na Figura 20(b) a interface da aplicação.

(a) (b)

Figura 20 – a) Utilização de gestos para controlar helicóptero em jogo, obtida de Truyenque (2005); b) Interface do Câmera Kombat, obtida de Paula, Bonini e Miranda (2006).

Na pesquisa de Mistry e Maes (2009), denominada de SixthSense, é utilizado uma câmera e um pequeno projetor, com uma aparência semelhante a um colar (Figura 21(a)), informações são projetadas sobre superfícies, paredes e objetos físicos, permitindo interagir com a informação projetada através de gestos naturais da mão. Em seu trabalho, Júnior (2010) possibilita o reconhecimento de gestos com as mãos livres, para uso de interação em ambientes 3D (jogos e aplicações para educação), com um domínio previamente determinado, permitindo que gestos sejam selecionados no domínio de cada aplicação. Ataide e Pimentel (2011) desenvolveram um jogo onde o usuário exerce controle sobre uma aeronave utilizando o próprio corpo, inclinando seus braços abertos, que representam as asas da aeronave, como mostra a Figura 21(b).

(a) (b)

Figura 21 – a) SixthSense, pesquisa de Mistry e Maes (2009); b) Jogo onde o usuário exerce controle sobre uma aeronave utilizando o próprio corpo, obtida de Ataide e Pimentel (2011).

De um modo geral, como afirma Júnior (2010), embora haja um grande interesse e volume de trabalho envolvendo diversas áreas de estudo de interação através de gestos, não se trata de um problema único, de forma que um grande número de soluções coexistam e estão em constante evolução, em domínios, aplicações e com limitações distintas, como restrições aos tipos de gestos que se pode usar, ao conjunto de gestos possíveis ou às condições de uso.

Considerado como uma pesquisa prévia a este trabalho, em Almeida et al. (2013) e Vasconcelos et al. (2014), foi desenvolvida uma aplicação inicial, porém contendo algumas limitações, as quais serviram de objetivos a serem alcançados neste trabalho. De maneira geral, ao iniciar a execução do aplicativo, faz-se necessário definir na projeção onde estão localizadas as opções de navegação interativa, sendo assim, o processamento da imagem não ocorre sob toda a imagem projetada, mas sim em regiões específicas, as quais devem ser delimitadas inicialmente.

Nesse trabalho é proposto um modelo que não é necessário informar previamente as opções de navegação interativa, sendo assim, inicialmente o processamento da imagem ocorrerá sob toda a imagem projetada. Utilizando técnicas de Inteligência Artificial e Processamento de Imagens foi possível obter informações relevantes da cena que puderam identificar regiões propensas a possíveis alvos e isolá-los, no próximo capítulo, o modelo de interação natural proposto será descrito em detalhes.

3 MODELO PROPOSTO DE INTERAÇÃO NATURAL COM IMAGEM PROJETADA

O modelo proposto neste trabalho possui como principal objetivo a interação do usuário com uma imagem projetada por um *datashow* comum, através de toques em regiões específicas da imagem projetada. Para validação do modelo proposto, foi desenvolvida uma aplicação utilizando a linguagem Java, a mesma foi escolhida por possuir uma vasta documentação disponível na internet, IDEs poderosas e gratuitas, bem como diversas bibliotecas consistentes. Este último citado é um ponto muito importante neste trabalho, pois com a necessidade de comunicação entre os dispositivos físicos utilizados, no que refere-se à projeção e captura dinâmica das imagens, foram utilizadas algumas bibliotecas de domínio público, consolidadas em suas áreas, descritos brevemente como segue:

- *leitor de arquivo*: para a leitura de arquivos .ppt[1] foi utilizada a biblioteca POI-HSLF, que fornece mecanismos para ler, criar ou modificar apresentações elaboradas neste formato, desenvolvida e publicada por Apache (2013), com o objetivo de criar e manter APIs Java para manipular arquivos do aplicativo Microsoft Powerpoint;
- *captura de imagens*: para a captura de imagens foi utilizada a biblioteca WEBCAM-CAPTURE de Bartosz (2013), que permite o acesso a câmeras de vídeo diretamente a partir do código Java, possibilitando a leitura de imagens e detecção de movimentos, oferecendo suporte à diversas plataformas;
- *interface*: para o desenvolvimento da interface do aplicativo proposto foi utilizada a biblioteca INFONODE, disponível publicamente em (NNL, 2013), sendo uma API Java Swing, baseado em encaixe de *frameworks*, permitindo a criação de GUI Swing com considerável redução de código. Outra biblioteca utilizada foi a INSUBSTANTIAL (GITHUB, 2013), com o objetivo de permitir ao usuário personalizar o *layout* do aplicativo, de acordo com suas preferências e necessidades.

Como pode ser observado na Figura 22, os dispositivos necessários para a utilização do aplicativo proposto são bastante comuns, sendo os seguintes: computador, *datashow* e câmera, que pode ser a própria câmera embutida no computador. A aplicação proposta também possui como objetivo possibilitar ao usuário, tanto a execução normal de uma apresentação utilizando *slides*, quanto a execução interativa. Entretanto, será dado ênfase à segunda opção.

[1] Formato comumente utilizado para a elaboração de *slides*.

A fim de sanar a limitação da aplicação anterior, proposta por Almeida et al. (2013) e Vasconcelos et al. (2014) (descrita na seção de trabalhos relacionados) na qual, ao iniciar a aplicação faz-se necessário definir na projeção a localização das opções de navegação interativa, é proposto um modelo que não é necessário informar previamente as opções de navegação. Para isso, inicialmente foi inserida uma máscara de reconhecimento, que possui as regiões das opções em vermelho. Com testes verificou-se que ao capturar a foto da projeção perde-se qualidade, interferindo no resultado da identificação das opções, por esse motivo, é necessário um pré-processamento, em que são destacados os pontos vermelhos (regiões das opções). Em seguida, é feito o tratamento e obtenção de informações relevantes da cena que possam identificar os alvos (opções), para isso, foi aplicado o modelo baseado em mapas de saliências proposto por Itti (1998).

Para permitir a interação do usuário com a imagem projetada foi necessário a inserção de algumas informações adicionais ao *slide*, ou imagem projetada, de forma que, intuitivamente, saiba-se onde e como a interação deva ocorrer. Para isto, foi criada uma máscara, como pode ser visto na Figura 22, que é incorporada à lateral direita da imagem projetada, composta, neste trabalho, por quatro funções f básicas: anterior, início, próximo e final.

Figura 22 – *Layout* do aplicativo proposto.

Sendo assim, ao capturar a imagem inicial e concluir o pré-processamento, é aplicado o modelo baseado em mapas de saliências para a identificação dos alvos (opções), a etapa seguinte é a segmentação e, por fim, a identificação de interação. As etapas desse processo são apresentadas na Figura 23 e descritos em detalhes nas próximas seções.

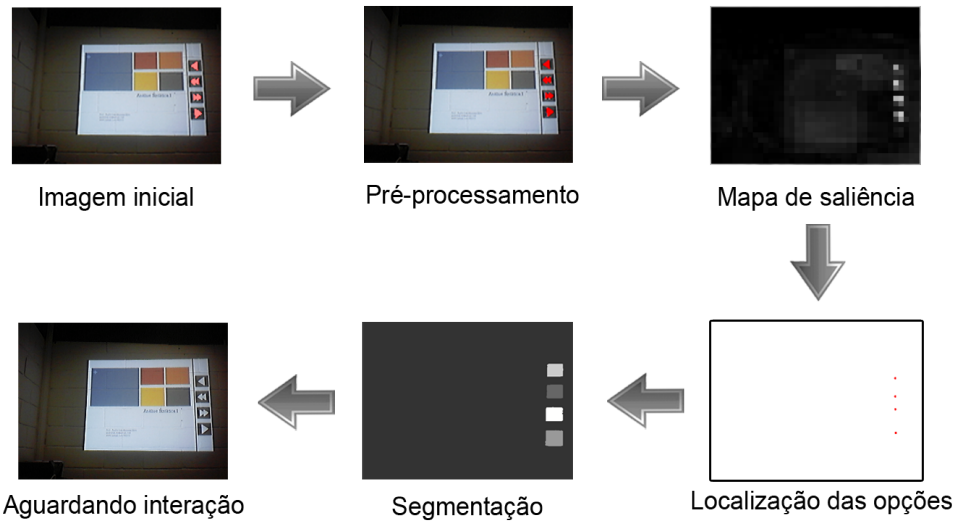


Figura 23 – Processo resumido para a identificação das opções.

3.1 Pré-processamento

Ao capturar a foto da projeção observou-se que os tons vermelhos (regiões das opções) ficam mais claros, interferindo no resultado da identificação das opções, por esse motivo, foi necessário o pré-processamento, no qual são destacados os pontos vermelhos. Para isso, foi utilizado um limiar V que, para qualquer ponto x, y na imagem, tal que $f(x,y)_r \geq V$, será chamado de ponto do objeto, caso contrário, o ponto será chamado ponto de fundo. Esse processo de limiarização é descrito como segue:

$$lim(x,y) = \begin{cases} (r = 255; g = 0; b = 0) & f(x,y)_r \geq V \\ (\text{r; g; b}) & f(x,y)_r < V \end{cases} \tag{3.1}$$

Na Figura 24(a), pode-se observar a imagem de entrada sem o pré-processamento e na Figura 24(b), a imagem com os pontos vermelhos destacados. O valor do limiar V utilizado aqui foi de 170.

3.2 Identificação das regiões salientes

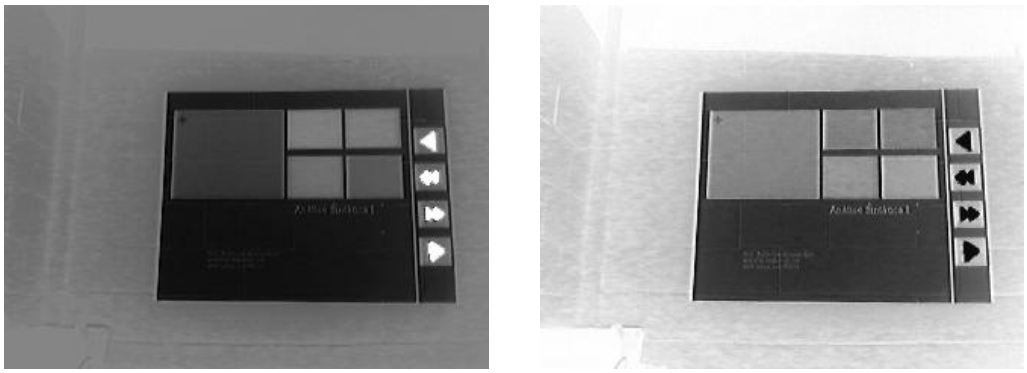
Para detectar os pontos mais salientes da imagem projetada, que são opções vermelhas com as quais o usuário deve interagir, foi implementado na linguagem Java o algoritmo de mapa de saliência proposto por Itti (1998), este foi explicado com maiores detalhes no capítulo anterior. Algumas adaptações foram realizadas no algoritmo implementado, para que este se encaixasse melhor às necessidades do trabalho corrente.

No modelo proposto por Itti (1998), inicialmente são criados quatro canais cores: R para vermelho, G para verde, B para azul e Y para amarelo, sendo r, g, b, respectivamente,

(a) Imagem de entrada (b) Imagem pré-processada

Figura 24 – Imagem projetada com opções vermelhas.

os canais vermelho, verde e azul da imagem de entrada, neste trabalho foram utilizados apenas os canais R e G. O canal R é utilizado pelo fato de serem vermelhas as cores das regiões que se procura como pontos salientes na cena (opções), já o canal G é utilizado pelo fato de no modelo proposto por Itti (1998) ser utilizado a combinação desses dois canais. Essa diferença entre o modelo proposto e o modelo de Itti (1998), permite obter maior velocidade de processamento, uma vez que a manipulação da imagem é diminuída pela metade ao desconsiderar dois dos quatro canais disponíveis. A extração dos canais R e G podem ser vistos na Figura 25.



(a) Canal R (b) Canal G

Figura 25 – Canais de cores R e G da Figura 24(b).

No processo de construção do mapa de cores é utilizada uma pirâmide gaussiana, nesta pesquisa possui 5 níveis (Equações 3.2 e 3.3), como podem ser vistos na Figura 26 e foi calculada utilizando-se uma matriz de pesos de tamanho 4x4, gerados por uma função gaussiana. Os valores da matriz de pesos utilizada podem ser visualizados na Tabela 3.

$$G_\theta(x,y) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} w(m+2, n+2)\ G(x,y), \quad para\ \theta = 0 \tag{3.2}$$

$$G_\theta(x,y) = \sum_{m=-2}^{2}\sum_{n=-2}^{2} w(m+2,n+2)\; G_{\theta-1}(2x+m,2y+n), \quad para\ 0<\theta\leq 4 \qquad (3.3)$$

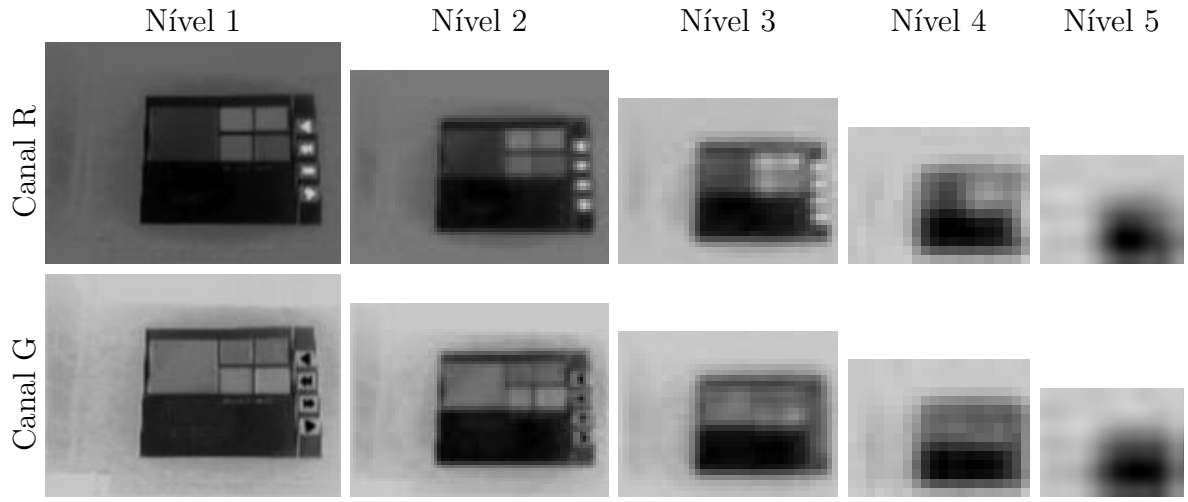


Figura 26 – Pirâmide Gaussiana com 5 níveis (esquerda para direita) das cores RG (cima para baixo).

Tabela 3 – Matriz de pesos utilizados para calcular a pirâmide gaussiana

1	1	1	1
1	10	10	1
1	10	10	1
1	1	1	1

Como observado no capítulo anterior, para cada nível da pirâmide é gerada uma imagem em uma escala menor que a escala no nível superior, essa representação piramidal é usada para a obtenção de amostras da imagem sem detalhes indesejáveis. Continuando o processo, a próxima etapa do modelo é a criação dos mapas de características, que são obtidos por meio da diferença entre os canais de cores em diferentes escalas, este processo é conhecido como diferença centro-vizinhança, neste trabalho a subtração é realizada apenas entre os canais RG (Equação 2.9), o resultado desta etapa pode ser visualizado na Figura 27.

Figura 27 – Mapas de características R e G da Figura 24(b).

Seguindo o modelo proposto, a próxima etapa é a criação do mapa de conspicuidade *RG*, que constitui na normalização dos mapas de características e soma dos mesmos. Como visto na Seção 2.2.3.7, a normalização tem o objetivo amplificar regiões que apresentam um nível de saliência contrastante e, por outro lado, inibir regiões salientes não contrastantes, a Figura 9 demonstra a função da normalização *N(.)*.

O modelo proposto por Itti (1998) é constituído por três mapas que, combinados, destacam pontos salientes aos olhos humanos, seguindo um comportamento semelhante ao biológico, estes são: mapa de orientações, mapa de cores e mapa de intensidade. No entanto, as regiões que são procuradas na imagem projetada se destacam apenas por cor, e não por orientação (vertical, horizontal, diagonal) ou por intensidade da região, por esse motivo, o mapa de saliência final utilizado nesta pesquisa é formado apenas pelo mapa de cores *RG*, definido como segue:

$$\bar{S} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c,s))] \tag{3.4}$$

Como pode ser observado na Figura 28, esse mapa de cores destaca as regiões vermelhas da imagem projetada, tornando possível a identificação das posições de cada opção, para em seguida serem submetidos à um algoritmo de segmentação, descrito na próxima subseção.

Figura 28 – Mapa de saliência da imagem da Figura 24.

3.3 Segmentação das regiões salientes

Conhecendo uma posição de cada opção, o próximo passo é identificar a região de cada opção, para isso, entra o conceito de segmentação. Na literatura exitem diversas aplicações com o objetivo de segmentar imagens, foi feito uma pesquisa e a aplicação

intitulada de *SegmentIt* proposta por Klava e Hirata (2013) teve o resultado ideal para o trabalho proposto, como pode ser visto na Figura 29.

Figura 29 – Segmentação das regiões das opções.

SegmentIt é uma ferramenta de segmentação interativa, que permite alternar entre as abordagens de Bacias Hidrográficas (bacias hidrográficas de marcadores e divisor de águas hierárquica) de modo que o usuário pode explorar os pontos fortes de ambos. Para este trabalho foi utilizado apenas a abordagem com marcadores, que são os pontos com maior saliência descrito na subseção anterior. Com o conhecimento das regiões das opções, a aplicação está pronta para identificar se houve ou não interação, processo o qual é descrito a seguir.

3.4 Identificação de interação

Um dos principais pontos deste trabalho está relacionado à identificação do momento no qual uma determinada interação tenha ocorrido, assim como a função de navegação correta a ser executada. Considerando o conjunto de técnicas de processamento de imagens apresentadas na seção anterior, foi possível diminuir, consideravelmente, a quantidade de informações a serem tratadas pela aplicação, uma vez que a diferença entre os histogramas de cores de cada imagem capturada, seja a principal informação para a identificação da interação. Apesar de parecer uma tarefa trivial, a aplicação proposta neste trabalho, em um ambiente real, pode torna-se instável, uma vez que duas imagens capturadas em um intervalo de segundos, ou até menos do que um segundo, podem apresentar diferenças sutis no histograma, geradas por interferências do ambiente, mesmo que imperceptíveis à visão humana. Sendo assim, para que a comparação entre duas imagens, aparentemente iguais, seja realizada de forma correta, foi necessário definir um valor acei-

tável de diferença entre duas imagens, ou seja, um valor máximo de mudança, tornando possível a diferenciação entre uma interação real e variações naturais do ambiente.

O valor de mudança aceitável, denominado aqui por σ, foi definido baseado no cálculo do desvio padrão dos valores dos histogramas para um conjunto de imagens capturadas, sequencialmente, em um intervalo curto de tempo. Inicialmente, o cálculo da média dos valores dos histogramas obtidos para um conjunto de imagens, referentes a uma determinada função, é descrito como segue:

$$m^f = \frac{1}{n^f} \sum_{i=1}^{n^f} \left(\frac{1}{256} \sum_{k=0}^{255} h_i^f(r_k) \right) \tag{3.5}$$

em que, n^f é o número de imagens capturadas referente à função f, k representa os níveis de cinza, h_i^f é o valor do histograma de cores da imagem i referente à função f e r_k, conforme descrito na seção anterior, representa o k-ésimo nível de cinza. De acordo com Correa (2003), o cálculo do desvio padrão é a medida mais usada para a comparação de diferenças entre conjuntos de dados, que consiste em determinar a dispersão dos valores em relação à media. Assim, o valor de mudança aceitável é definido como segue:

$$\sigma^f = \sqrt{\frac{1}{n^f} \sum_{i=1}^{n^f} (m_i^f - m_f)^2} \tag{3.6}$$

De forma resumida, após esse o cálculo de σ^f, a apresentação é iniciada em modo interativo e, em intervalos de 400ms, uma imagem da apresentação é capturada, transformada em seguida para tons de cinza, calculando-se em seguida a média do histograma referente a cada função. A seguir, a média obtida é comparada com a diferença da média dos histogramas do conjunto de imagens capturadas inicialmente (Equação 3.5) com o desvio padrão (Equação 3.6), de forma que, caso o valor da média do histograma obtido seja menor do que $m^f - \sigma^f$, pode-se concluir que houve uma interação, de forma que o comando relacionada à função que tenha apresentado a diferença seja executado. Para o caso de detecção simultânea de interação, nenhuma função deverá ser executada. Para um melhor entendimento em relação ao fluxo de identificação de interação o Algoritmo 1 mostra o fluxo descrito anteriormente, que inicia-se na fase de treinamento do conjunto

inicial de imagens, concluindo com a identificação da interação selecionada.

Algoritmo 1: Modelo de interação natural com imagem projetada

1. Calcular o desvio padrão σ :
   Pré-processamento;
   Identificar regiões salientes(opções);
   Segmentar;
2. Executar a cada intervalo de 400ms:
   Capturar imagem;
   Calcular a média do histograma de cada função;
   Para cada opção da imagem verificar:
   if $mediaHistograma < mediaConjuntoImagens - desvioPadrao$ **then**
   Houve interação ;
3. Se somente uma opção foi escolhida, função é executada;

Para um melhor entendimento do momento de identificação de interação, nas Figuras 30 e 31 são apresentadas do lado esquerdo exemplos de imagens projetadas e do lado direito os gráficos das respectivas médias do histograma de cada opção. Na Figura 30(a), é apresentada uma imagem projetada sem a tentativa de interação, observa-se na Figura 30(b), que a mão, ao ser sobreposta à opção final, faz variar a média de 220 para 156.

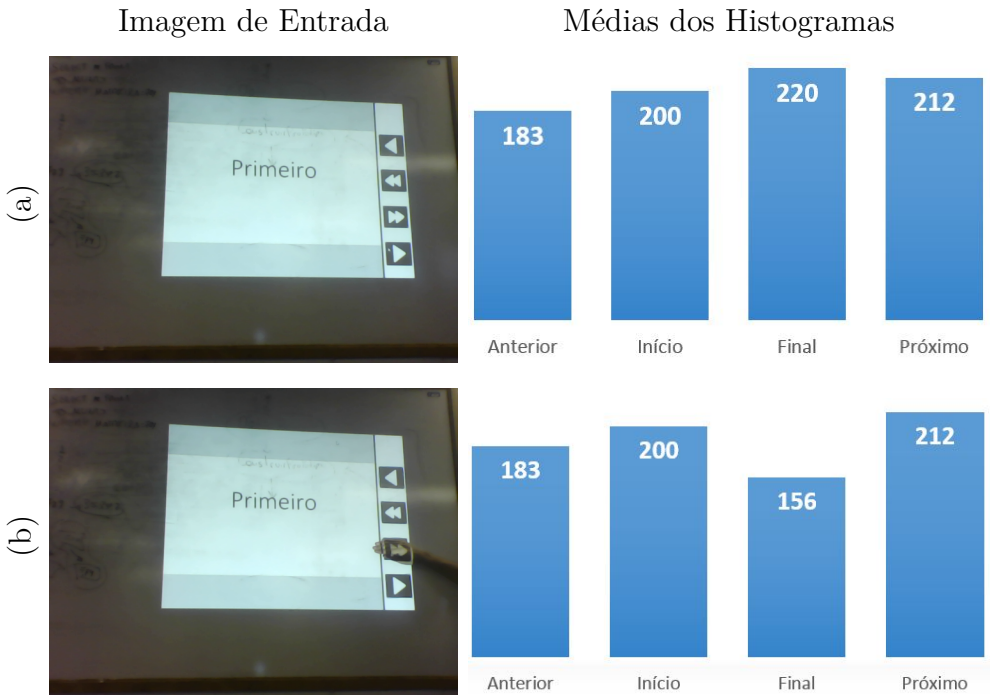


Figura 30 – Imagens projetadas e as respectivas médias dos histogramas (1-2). a)Imagem de Entrada; b)Mão sobreposta à opção ir para *slide* final;

Algo semelhante acontece na Figura 31(c), na qual, ao sobrepor a mão na opção

início, faz variar a respectiva média de 200 para 152. Na Figura 31(d), ao posicionar a mão sobre à opção próximo, a média varia de 212 para 146. Por fim, na Figura 31(e), sobrepondo a mão na opção anterior, a média varia de 183 para 139.

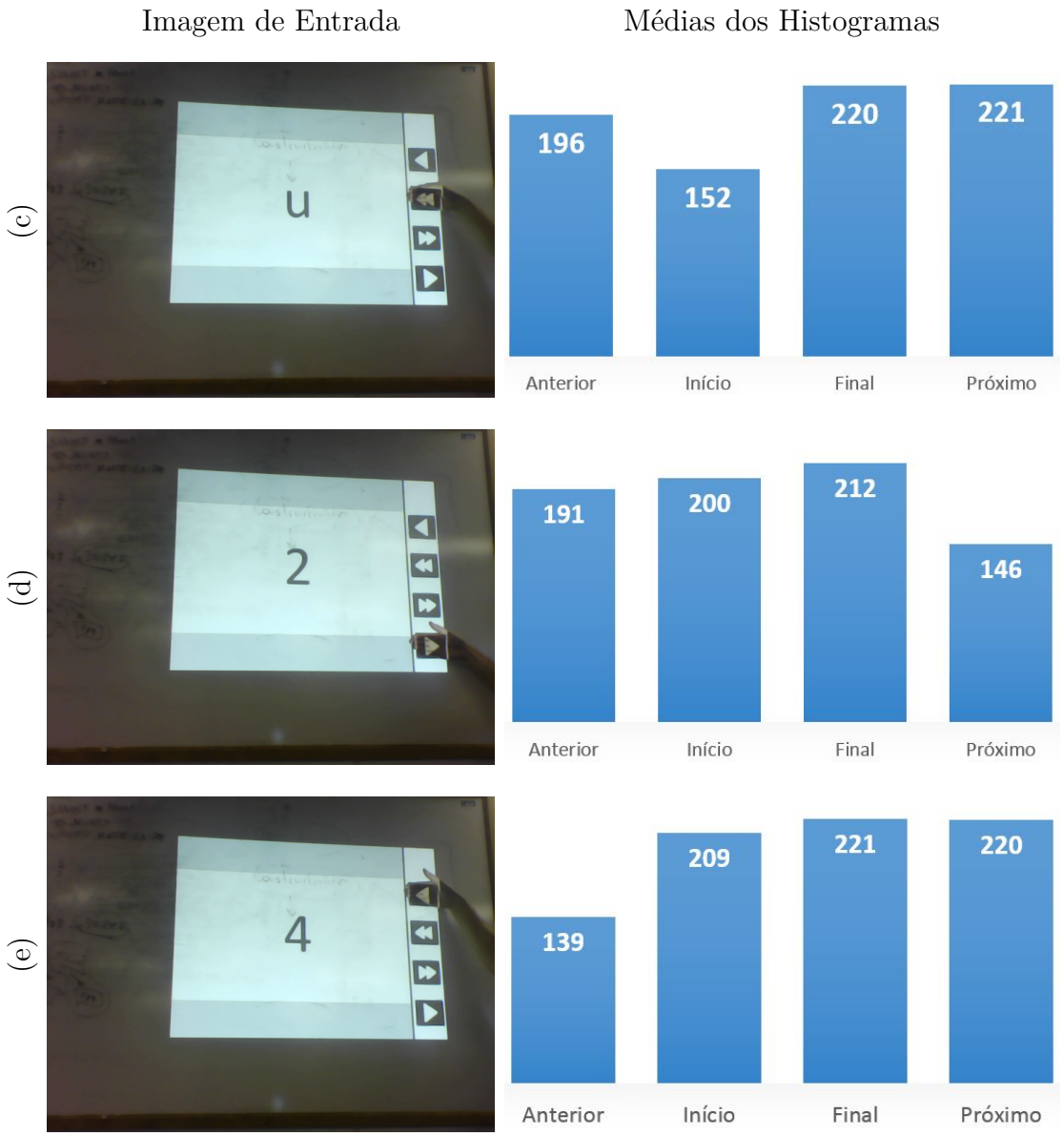


Figura 31 – Imagens projetadas e as respectivas médias dos histogramas (2-2). c)Mão sobreposta à opção ir para *slide* inicial; d)Mão sobreposta à opção próximo *slide*; e)Mão sobreposta à opção *slide* anterior;

Na seção seguinte serão apresentados os experimentos realizados, os mesmos foram divididos em duas etapas, a primeira para identificação das regiões das opções e a segunda com o objetivo de identificar a interação.

4 EXPERIMENTOS

Neste capítulo serão apresentados os experimentos, que inicialmente foram realizados para a identificação das regiões das opções, e em seguida para a identificação de interação.

4.1 Identificação das regiões das opções

Como o objetivo inicial é a identificação das regiões das opções, foram realizados experimentos considerando quatro situações de ambientes:

1. Luz alta sobre a apresentação, luz ambiente normal, fundo branco;
2. Luz alta sobre a apresentação, luz ambiente alta, fundo branco;
3. Luz alta sobre a apresentação, luz ambiente normal, fundo escuro;
4. Luz normal sobre a apresentação, luz ambiente normal, fundo branco (ambiente ideal).

A luz alta sobre a apresentação é consequência do reflexo do próprio *datashow* e/ou de lâmpadas, já em relação a luz ambiente alta, é consequência da alta claridade no ambiente, podendo ser causada pelas das lâmpadas e/ou grande quantidade de janelas.

Com o objetivo de avaliar os resultados, foram definidas três situações possíveis. Sendo a primeira *IC*, que significa a identificação correta de todas as regiões, a segunda, denominada por *NA*, que significa necessidade de ajuste, por não detectar inicialmente, mas com algumas alterações, a aplicação passa a identificar corretamente as regiões. Consideramos como ajustes, erguer um pouco a imagem projetada ou alterar o limiar de identificação dos pontos salientes. Por fim, a medida *NC*, que significa a não identificação de todas as regiões ou nenhuma, mesmo com ajustes. Na Tabela 4 são apresentados os resultados obtidos a partir de experimentos considerando todos os ambientes citados.

Tabela 4 – Experimentos para identificação das regiões das opções em Ambiente Heterogêneos

Ambiente	Identificação das regiões
1	NA
2	NC
3	NC
4	IC

Na Figura 32 são apresentados exemplos em ambientes ideais, com as respectivas imagens projetadas, mapas de saliência e as regiões das opções.

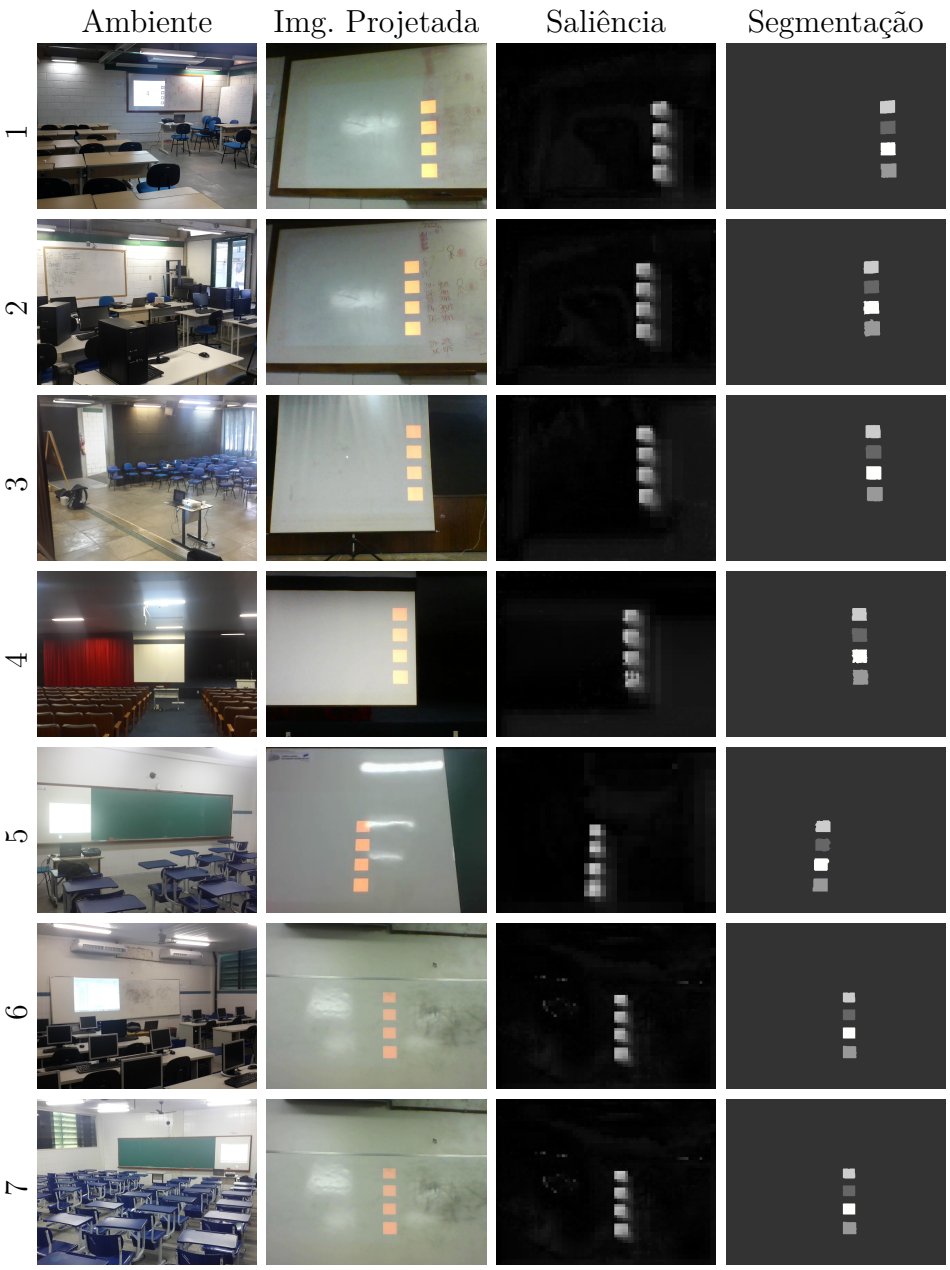


Figura 32 – Exemplos da identificação das regiões das opções em ambientes ideais, com as respectivas imagens projetadas, mapas de saliência e as regiões das opções.

Na Figura 33 são apresentados situações em que não foi possível identificar as regiões das opções por motivos citados anteriormente.

4.2 Experimentos para identificação da interação

Com o conhecimento do ambiente ideal foi possível detectar que a interação pode ser realizada com sucesso na maioria dos casos, sendo assim, foram feitos experimentos

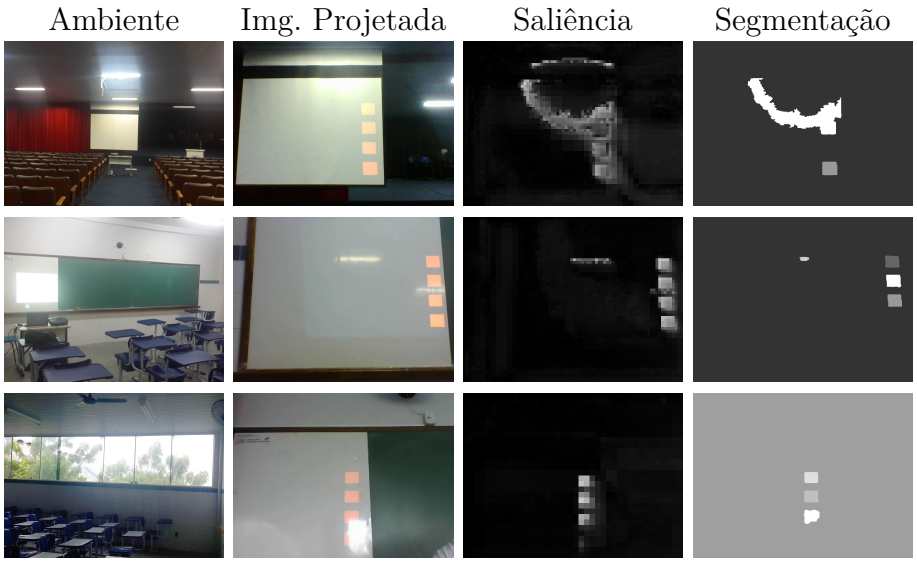


Figura 33 – Exemplos de situações em que não foi possível identificar as regiões das opções.

nos ambientes apresentados na Figura 32, com duração entre 15 e 20 minutos, nos quais foram contabilizados o número total de tentativas de interação - *TTDI*, a quantidade de interações bem sucedidas - *IBS* (identificação correta da interação), quantidade de identificação incorreta - *II*(opção acionada sem a solicitação do usuário), e o número de interações mal sucedidas - *IMS* (intenção de realizar alguma interação, mas não foi acionada pela aplicação).

No Ambiente 1, de um total de 177 tentativas de interação - TTDI, 162 foram bem sucedidas e em 15 momentos não foi realizada a interação solicitada, tendo uma taxa de IBS de 91,5%. No Ambiente 2, de 210 tentativas de interações houvem 194 de IBS e 16 mal sucedidas, possuindo uma taxa de IBS de 92,3%. No Ambiente 3, de 121 TTDI, 116 foram bem sucedidas e apenas 5 não, possuindo uma taxa de interação bem sucedida de 95,8%. No Ambiente 4, de 147 tentativas de interação, 135 foram bem sucedidas e 12 mal sucedidas, apresentando uma taxa de IBS de 91.8%. No Ambiente 5, de 158 tentativas de interação, 154 foram bem sucedidas e 4 mal sucedidas, com uma taxa de IBS de 97,4%. No Ambiente 6, de um total de 145 TTDI, foram bem sucedidas 138 e 7 mal sucedidas, tendo uma taxa de 95,2%. Por fim, no Ambiente 7, de 204 tentativas de interação, 190 foram bem sucedidas e 14 não, possuindo uma taxa de IBS de 93,1%. Os resultados completos dos experimentos são apresentados na Tabela 5.

Tabela 5 – Experimentos da identificação de interação nos ambientes da Figura 32.

Ambiente	TTDI	IBS	IMS	II	Porcentagem de IBS
1	177	162	15	2	91.5
2	210	194	16	3	92.3
3	121	116	5	0	95.8
4	147	135	12	3	91.8
5	158	154	4	0	97.4
6	145	138	7	0	95.2
7	204	190	14	3	93.1

5 CONCLUSÃO

A aplicação apresentou desempenho excelente em ambiente ideal, com uma taxa de interação bem sucedida acima de 90% em todos os ambientes testados, de modo que não houve interpretações errôneas de funções e um pequeno número de disparo de função na ausência de interação, este último, por consequência de variações do ambiente, como por exemplo, sombra, reflexo ou outros fatores. Outra consideração é que em alguns casos a aplicação pode não acionar a opção solicitada, isso pode ser causado três fatores: primeiro, por variações do ambiente citado anteriormente, pois a aplicação pode estar identificando alguma interação em uma das opções, e que, somada com a solicitada pelo usuário, a aplicação não aciona nenhuma, já que não tem como saber qual realmente é a desejada. Outra possibilidade é a de o usuário sobrepor a mão à opção de forma muito rápida, podendo não ocorrer corretamente a captura da imagem na qual a mão esteja posicionada sobre a opção desejada e, por fim, não sobrepor a região por completo, fazendo com que altere pouco o histograma da respectiva opção.

O ambiente ideal foi definido como uma projeção realizada em superfície de fundo claro, com iluminação ambiente normal, ou ainda, sem iluminação no ambiente, contanto que, não exista incidência direta de luz forte sobre a projeção. Observou-se também que, utilizando um intervalo de 300 à 600ms entre cada imagem capturada, foi possível uma interação natural com a projeção. Intervalos de tempo maiores causaram um atraso entre o tempo de tentativa de interação e a execução da função. Por outro lado, intervalos menores fizeram com o que a função fosse executada diversas vezes. Conclui-se então que, a aplicação proposta neste trabalho pode ser considerada como uma interessante alternativa para o desenvolvimento da interação entre o homem e a máquina, a qual, através de técnicas de processamento de imagens e inteligência artificial, foi possível dispensar o uso de dispositivos menos acessíveis.

5.1 Trabalhos futuros

Como trabalhos futuros destacam-se o reconhecimento automático das opções, sem a necessidade de uma máscara de treinamento, bem como a melhoria do algoritmo de identificação de interação para abranger ambientes diferentes do ambiente ideal, como por exemplo, projeções em superfícies escuras ou com incidência de luz forte sobre a projeção. Outra melhoria à vista é que a aplicação permita a interação com a projeção independente do programa de exibição.

Referências

AGGARWAL, J. K.; CAI, Q. Human motion analysis: A review. 1999. Citado 2 vezes nas páginas 17 e 34.

Almeida, A. B. *Usando o Computador para Processamento de Imagens Médicas*. 1998. Acesso em: 15 fev. 2015. Disponível em: <http://www.informaticamedica.org.br/informaticamedica/n0106/imagens.htm>. Citado 2 vezes nas páginas 7 e 22.

ALMEIDA, T. S. et al. Piai - processamento de imagens aplicado à apresentação interativa. III Semana de Informática da Universidade Federal de Sergipe (SEMINFO/UFSITA2013), 2013. Citado 2 vezes nas páginas 39 e 41.

ANDRADE, W. T. Segmentação baseada em textura e watershed aplicada a imagens de pólen. Campo Grande, MS, 2011. Citado 2 vezes nas páginas 7 e 32.

APACHE, S. F. Poi-hslf. disponível em http://poi.apache.org/slideshow/index.html. Acesso em abril de 2013, 2013. Citado na página 40.

ATAIDE, T. P.; PIMENTEL, R. C. Segmentação de imagens aplicada a jogos. 2011. Citado 3 vezes nas páginas 8, 38 e 39.

BARTOSZ, F. Java webcam capture. disponível em https://github.com/sarxos/webcam-capture. Acesso em abril de 2013, 2013. Citado na página 40.

BAUDEL, T.; BEAUDOUIN-LAFON, M. Charade. *Communications of the ACM*, v. 7, p. 28–35, 1993. Citado 3 vezes nas páginas 7, 34 e 35.

BENICASA, A. Sistemas computacionais para atenção visual top-down e bottom-up usando redes neurais artificiais. *USP São Carlos*, 2013. Citado 6 vezes nas páginas 22, 23, 25, 26, 27 e 28.

BLAKE, J. Natural user interfaces in .net. manning pubs co series. Manning Publications Company, 2012. Citado na página 16.

CAROTA, L.; INDIVERI, G.; DANTE, V. A softwarehardware selective attention system. Neurocomputing, 647653, 2004. Citado na página 23.

CASTRO, R. H. A. d. Desenvolvimento de aplicações com uso de interação natural: Um estudo de caso voltado para vídeo colaboração em saúde. Paraíba, 2012. Citado 3 vezes nas páginas 8, 36 e 37.

CHEN, C. et al. Visualizing the evolution of hci. In HCI05 Conference on People and Computers XIX, 2005. Citado na página 12.

CORREA, M. B. B. C. *Probabilidade de Estatística.* [S.l.]: 2a Edição. Belo Horizonte: PUC Minas Virtual, 2003. Citado na página 47.

DIAS DIEGO R. C., e. a. Desenvolvimento de aplicações com interface natural de usuário e dispositivos primesense como meio de interação para ambientes virtuais. Tendências e Técnicas em Realidade Virtual e Aumentada, 2013. Citado na página 16.

DOYLE, J.; DEAN, T. Strategic directions in artificial intelligence. ACM Computing Surveys, 1996. Citado na página 13.

ERKAN, A. N. Model based three dimensional hand posture recognition for hand tracking. 2004. Citado na página 37.

EROL, A. et al. A review on vision-based full dof hand motion estimation. In: IEEE Computer Society Conference on Computer and Pattern Recognition, 2005. Citado na página 12.

EWALT, D. M. Nintendo's wii is a revolution. disponivel em: http://www.forbes.com/2006/11/13/wii-review-ps3-tech-media-cx$_d$e$_1$113*wii.html*.15*desetembrode*2014, 2006.*Citadonapágina*36.

FACON, J. Processamento e analise de imagens. PUCP, 2002. Citado 3 vezes nas páginas 19, 20 e 21.

FIGUEIREDO, L. S. et al. Interação natural a partir de rastreamento de mãos. Tendências e Técnicas em RVA, 2012. Citado 5 vezes nas páginas 7, 15, 16, 17 e 18.

FIGUEIREDO, L. S. et al. An open-source framework for air guitar games. 2009. Citado 2 vezes nas páginas 7 e 36.

FREEMAN, W.; ANDERSON, D.; BEARDSLEY, P. Computer vision for interactive computer graphics. IEEE Computer Graphics and Applications, 1998. Citado 3 vezes nas páginas 8, 37 e 38.

GITHUB, I. Github - insubstantial. disponível em https://github.com/insubstantial/insubstantial. Acesso em abril de 2013, 2013. Citado na página 40.

GONZALEZ R. C. E WOODS, R. E. *Processamento Digital de Imagens - 3^a ed.* [S.l.]: São Paulo: Pearson, 2010. Citado 8 vezes nas páginas 7, 19, 20, 21, 29, 30, 33 e 34.

GRANDO, N. Segmentação de imagens tomográficas visando a construção de modelos médicos. Dissertação de Mestrado do Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. CEFET-PR Centro Federal de Educação Tecnológica do Paraná, 2005. Citado 4 vezes nas páginas 19, 21, 22 e 30.

HEWETT, T. et al. Chapter 2: Human-computer interaction. ACM SIGCHI Curricula for Human-Computer Interaction, 2009. Citado na página 15.

ITTI, L. Models of bottom-up and top-down visual attention. California Institute of Technology, 2000. Citado 2 vezes nas páginas 7 e 26.

ITTI, L. Models of bottom-up attention and saliency. Neurobiology of Attention, Chapter 94. Elsevier, Oxford, 2005. Citado na página 23.

ITTI, L.; KOCH, C. Computational modelling of visual attention. Nature Reviews Neuroscience 2, 194203, 2001. Citado 3 vezes nas páginas 25, 27 e 28.

ITTI, L. e. a. A model of saliency-based visual attention for rapid scene analysis. 1998. Citado 9 vezes nas páginas 7, 24, 27, 28, 29, 41, 42, 43 e 45.

JAIN, A.; DORAI, C. Practicing vision: Integration, evaluation and applications. 1997. Citado na página 12.

JúNIOR, J. L. B. Modelo abrangente e reconhecimento de gestos com as mãos livres para ambientes 3d. São Paulo, 2010. Citado 4 vezes nas páginas 18, 37, 38 e 39.

KANELLOS, M. More than 1 billion served. disponivel em: http://news.cnet.com/2100-1040-940713.html. Acesso dezembro de 2014, 2002. Citado na página 15.

KLAVA, B. Segmentação interativa de imagens via transformação watershed. São Paulo, 2000. Citado 3 vezes nas páginas 7, 32 e 33.

KLAVA, B.; HIRATA, N. S. T. Segmentit. disponivel em: *http : //segmentit.sourceforge.net/*. 23 de dezembro de 2014, 2013. Citado na página 46.

KöRBES, A. Análise de algoritmos da transformada watershed. Campinas, SP, 2010. Citado na página 30.

LEE, J. C. Hacking the nintendo wii remote. 2008. Citado na página 36.

LOUREGA, L. V. Meseghi: Um método de segmentação para o processamento linear e não-linear de imagens. Santa Maria: Programa de Pós-Graduação em Engenharia da Produção, 2006. Citado na página 30.

MARENGONI, M.; STRINGHINI, D. Tutorial: Introdução à visão computacional usando opencv. Universidade Presbiteriana Mackenzie, 2009. Citado na página 31.

MARQUES O. F. E VIEIRA, H. N. *Processamento Digital de Imagens*. [S.l.]: Rio de Janeiro: Brasport, 1999. Citado 2 vezes nas páginas 30 e 33.

MEIRELLES, F. S. 25ª pesquisa anual do uso de ti. FGV-EAESP-CIA, 2014. Citado na página 12.

Melo, N. *Abordagens do processo de Segmentação: Limiarização, Orientada a Regiões e Baseada em Bordas*. 2009. Acesso em: 15 fev. 2015. Disponível em: <http: //www.dsc.ufcg.edu.br/~pet/jornal/setembro2011/materias/recapitulando.html>. Citado 2 vezes nas páginas 7 e 31.

MICROSOFT, C. Kinect for windows. disponível em http://www.microsoft.com/en-us/kinectforwindows/. Acesso em setembro de 2013, 2013. Citado 2 vezes nas páginas 8 e 37.

MISTRY, P.; MAES, P. Sixthsense a wearable gestural interface. Proceedings of SIGGRAPH Asia, Emerging Technologies. Yokohama, Japan, 2009. Citado 3 vezes nas páginas 8, 38 e 39.

MORALES, D.; CENTENO, T. M.; MORALES, R. C. Extração automática de marcadores anatômicos no desenvolvimento de um sistema de auxílio ao diagnóstico postural por imagens. III Workshop de Informática aplicada à Saúde CBComp, 2003. Citado na página 18.

MORGAN, J. Técnicas de segmentação de imagens na geração de programas para máquinas de comando numérico. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal de Santa Maria, 2008. Citado 3 vezes nas páginas 19, 21 e 29.

NNL, T. Infonode docking windows. disponível em http://www.infonode.net/index.html?idw. Acesso em abril de 2013, 2013. Citado na página 40.

PAULA, L. P.; BONINI, N.; MIRANDA, R. Câmera kombat - interação livre para jogos. V Brazilian Symposium on Computer Games and Digital Entertainment, Recife, 2006. Citado 2 vezes nas páginas 8 e 38.

PECCINE, G. Segmentação de imagens por watersheds: Uma implementação utilizando a linguagem java. Santa Maria - RS, 2004. Citado na página 32.

PEREIRA, E. T. Atenção visual bottom-up guiada por otimização via algoritmos genéticos. Campina Grande, 2007. Citado 3 vezes nas páginas 13, 23 e 27.

POPESCU, V.; BURDEA, G.; BOUZIT, M. Virtual reality simulation modeling for a haptic glove. 1999. Citado na página 35.

RASKAR, R. et al. Prakash: Lighting aware motion capture using photosensing markers on multiplexed illuminators. 2007. Citado 2 vezes nas páginas 7 e 36.

RAYMOND, E. S.; LANDLEY, R. W. Chapter 2. history: A brief history of user interfaces. The Art of Unix Usability, 2004. Citado na página 15.

RUSNAK, V. Interaction methods for large high-resolution screens. disponível em http://is.muni.cz/th/172757/fi_r/dtp.pdf. Acesso em outubro de 2013, 2012. Citado na página 12.

RUSS, J. The image processing handbook. 6 ed. Boca Raton: CRC Press, 2011. Citado 2 vezes nas páginas 30 e 32.

SHIC, F.; SCASSELLATI, B. A behavioral analysis of computational models of visual attention. *International Journal of Computer Vision*, 2007. Citado na página 23.

SYSTEMS, C. Cyberglove ii. disponível em: http://www.cyberglovesystems.com/products/cyberglove-ii/overview. Acesso em 24 de Agosto de 2014, 2014. Citado na página 35.

TRUYENQUE, M. A. Q. Uma aplicação de visão computacional que utiliza gestos da mão para interagir com o computador. Rio de Janeiro, 2005. Citado 2 vezes nas páginas 8 e 38.

VALLI, A. Notes on natural interaction. 2007. Citado na página 12.

VASCONCELOS, N. O. et al. Processamento de imagens e ia aplicado à apresentação interativa: Uma comparação entre um método interativo tradicional e um método interativo fuzzy. IV Semana de Informática da Universidade Federal de Sergipe (SEMINFO/UFSITA2014), 2014. Citado 2 vezes nas páginas 39 e 41.

WANG, R. Y.; POPOVIC, J. Real-time hand-tracking with a color glove. 2009. Citado 2 vezes nas páginas 7 e 36.

WANGENHEIM, A. V. Introdução à visão computacional. Santa Catarina, 2011. Citado na página 31.

WOLFE, J. M.; HOROWITZ, T. S. What attributes guide the deployment of visual attention and how do they do it? Nature Review Neuroscience 5, 495501, 2004. Citado 2 vezes nas páginas 7 e 23.