
**Aplicação de um robô humanoide autônomo por
meio de reconhecimento de imagem e voz em
sessões pedagógicas interativas**

Daniel Carnieto Tozadore

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Daniel Carnieto Tozadore

**Aplicação de um robô humanoide autônomo por meio de
reconhecimento de imagem e voz em sessões pedagógicas
interativas**

Dissertação apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP, como
parte dos requisitos para obtenção do título
de Mestre em Ciências – Ciências de Computação
e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e
Matemática Computacional

Orientadora: Profa. Dra. Roseli Aparecida
Francelin Romero

USP – São Carlos
Maio de 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

T757a Tozadore, Daniel Carnieto
Aplicação de um robô humanoide autônomo
por meio de reconhecimento de imagem e
voz em sessões pedagógicas interativas /
Daniel Carnieto Tozadore; orientadora Roseli
Aparecida Francelin Romero. - São Carlos - SP,
2016.
133 p.

Dissertação (Mestrado – Programa de
Pós-Graduação em Ciências de Computação
e Matemática Computacional) – Instituto
de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2016.

1. Robótica Pedagógica. 2. *Human-Robot
Interaction* (HRI). 3. Reconhecimento Imagem
e Fala. I. Romero, Roseli Aparecida Francelin,
orient. II. Título.

Daniel Carnieto Tozadore

**Application of an autonomous humanoid robot by image and
voice recognition in interactive pedagogical sessions**

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Roseli Aparecida Francelin Romero

USP – São Carlos
May 2016

*Este trabalho é dedicado ao doce e sereno colo de minha mãe,
às sábias e firmes palavras de meu pai,
e aos sorrisos e abraços encantadores de minhas amadas irmãs.*

AGRADECIMENTOS

Ao Criador, por seu imenso amor por mim, por nossas conversas e silêncios e por nunca me deixar caminhar sozinho.

A meus primeiros e melhores professores: meus pais Marcos e Hermínia. Por todo investimento de qualquer natureza que fizeram em mim. Pelas noites de preocupações acordados. Por toda oração e entrega. Pelo incentivo de uma vida própria paga a preço de muita saudade. Por todas as broncas dadas e por aquelas que ainda virão.

As minhas lindas e amadas irmãs Juliana e Michele. Com elas que acalento minha alma e reencontro minha alegria. Meus queridos Avôs, principalmente o Vô Lelé que se juntou a Deus durante este trabalho. Meus tios, primos e familiares onde busco forças quando já não tenho mais.

A minha orientadora e amiga Prof^a Dra^a Roseli Romero, por acreditar em mim, pela confiança no meu trabalho, por sempre tirar o meu melhor, por sonhar comigo em meus projeto e me puxar de volta a realidade quando necessário.

A todos os professores e funcionários do ICMC que colaboraram para minha formação. Sobretudo aos funcionários da Seção de Pós-Graduação Carol, Alexandre e Leonardo, que acabaram tornando-se meus grandes colegas.

Aos colegas de laboratório, por todo companheirismo, ajuda e divertidas horas de trabalho no Laboratório de Computação Bio-inspirada, Biocom.

Aos companheiros de República, a Gato Preto, por todo carinho e respeito que construímos durante os 7 anos que morei lá. Também por me incentivarem nas madrugadas de estudo com violão, cajon e um convite pra ajudar na próxima música. Também aos ex-moradores, entiados e funcionários da casa.

A família Vôlei CAASO, que mais do que medalhas me trouxe amigos pra vida toda, conforto quando eu precisava e aprendizados de quadra que pude trazer para o resto da vida.

A todos os amigos que deixei de participar de suas vidas por estar concentrado neste trabalho. A todas as pessoas que me cumprimentam pela USP com calorosa saudação detalhada e eu ainda não sei quem são. Também àquelas que, desavisadas, perguntam quando vou me formar.

E aos inúmeros amigos que infelizmente não são aqui citados, pois fazem deste um lugar apertado para verbalizar toda a gratidão que sinto pelas amizades feitas durante este período.

*“As invenções são, sobretudo,
o resultado de um trabalho de teimoso.”*
(Santos Dumont)

RESUMO

TOZADORE, D. C.. **Aplicação de um robô humanoide autônomo por meio de reconhecimento de imagem e voz em sessões pedagógicas interativas.** 2016. 133 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

A Robótica Educacional consiste na utilização de robôs para aplicação prática dos conteúdos teóricos discutidos em sala de aula. Porém, os robôs mais usados apresentam uma carência de interação com os usuários, a qual pode ser melhorada com a inserção de robôs humanoides. Esta dissertação tem como objetivo a combinação de técnicas de visão computacional, robótica social e reconhecimento e síntese de fala para a construção de um sistema interativo que auxilie em sessões pedagógicas por meio de um robô humanoide. Diferentes conteúdos podem ser abordados pelos robôs de forma autônoma. Sua aplicação visa o uso do sistema como ferramenta de auxílio no ensino de matemática para crianças. Para uma primeira abordagem, o sistema foi treinado para interagir com crianças e reconhecer figuras geométricas 3D. O esquema proposto é baseado em módulos, no qual cada módulo é responsável por uma função específica e contém um grupo de funcionalidades. No total são 4 módulos: Módulo Central, Módulo de Diálogo, Módulo de Visão e Módulo Motor. O robô escolhido é o humanoide NAO. Para visão computacional, foram comparados a rede LEGION e o sistema VOCUS2 para detecção de objetos e SVM e MLP para classificação de imagens. O reconhecedor de fala *Google Speech Recognition* e o sintetizador de voz do *NAOqi API* são empregados para interações sonoras. Também foi conduzido um estudo de interação, por meio da técnica de Mágico-de-Oz, para analisar o comportamento das crianças e adequar os métodos para melhores resultados da aplicação. Testes do sistema completo mostraram que pequenas calibrações são suficientes para uma sessão de interação com poucos erros. Os resultados mostraram que crianças que tiveram contato com uma maior interatividade com o robô se sentiram mais engajadas e confortáveis nas interações, tanto nos experimentos quanto no estudo em casa para as próximas sessões, comparadas às crianças que tiveram contato com menor nível de interatividade. Intercalar comportamentos desafiadores e comportamentos incentivadores do robô trouxeram melhores resultados na interação com as crianças do que um comportamento constante.

Palavras-chave: Robótica Pedagógica, *Human-Robot Interaction* (HRI), Reconhecimento Imagem e Fala.

ABSTRACT

TOZADORE, D. C.. **Aplicação de um robô humanoide autônomo por meio de reconhecimento de imagem e voz em sessões pedagógicas interativas.** 2016. 133 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Educational Robotics is a growing area that uses robots to apply theoretical concepts discussed in class. However, robots usually present a lack of interaction with users that can be improved with humanoid robots. This dissertation presents a project that combines computer vision techniques, social robotics and speech synthesis and recognition to build an interactive system which leads educational sessions through a humanoid robot. This system can be trained with different content to be addressed autonomously to users by a robot. Its application covers the use of the system as a tool in the mathematics teaching for children. For a first approach, the system has been trained to interact with children and recognize 3D geometric figures. The proposed scheme is based on modules, wherein each module is responsible for a specific function and includes a group of features for this purpose. In total there are 4 modules: Central Module, Dialog Module, Vision Module and Motor Module. The chosen robot was the humanoid NAO. For the Vision Module, LEGION network and VOCUS2 system were compared for object detection and SVM and MLP for image classification. The Google Speech Recognition speech recognizer and Voice Synthesizer Naoqi API are used for sound interactions. An interaction study was conducted by Wizard-of-Oz technique to analyze the behavior of children and adapt the methods for better application results. Full system testing showed that small calibrations are sufficient for an interactive session with few errors. Children who had experienced greater interaction degrees from the robot felt more engaged and comfortable during interactions, both in the experiments and studying at home for the next sessions, compared to children who had contact with a lower level of interactivity. Interim challenging behaviors and support behaviors brought better results in interaction than a constant behavior.

Key-words: Pedagogical Robotics, Human-Robot Interaction (HRI), Image and Speech Recognition .

LISTA DE ILUSTRAÇÕES

Figura 1 – Produtos e a competições voltadas à Robótica Educacional.	27
Figura 2 – Crianças brincando com o NAO no experimento de Pinto <i>et al.</i> (2014a).	28
Figura 3 – Figuras do Tangram a esquerda e animais formados com essas figuras a direita.	32
Figura 4 – Sistema utilizado em Shukla, Mishra e Sharma (2013).	36
Figura 5 – Material utilizado em Tanaka e Ghosh (2011): (a) cartas e (b) tabela com formas geométricas de diferentes cores.	39
Figura 6 – Analise da qualidade de interação em Tanaka, Cicourel e Movellan (2007).	41
Figura 7 – Figuras escondidas em sobreposições (PINTO <i>et al.</i> , 2014b).	42
Figura 8 – Figuras geométricas formando objetos conhecidos como uma casa.	42
Figura 9 – Estrutura do modelo de saliência iNVT.	49
Figura 10 – Normalização dos mapas de conspicuidades.	54
Figura 11 – Processamento dos canais de características do sistema VOCUS2.	56
Figura 12 – Sistema de visão proposto em Benicasa (2013).	59
Figura 13 – Morfologia do Neurônio Biológico	60
Figura 14 – O Neurônio não-linear.	61
Figura 15 – Arquitetura da Rede Neural MLP.	62
Figura 16 – Modos do Oscilador de relaxamento: (a) Modo ativo e (b) modo excitável.	64
Figura 17 – A arquitetura da segmentação baseada em objetos.	65
Figura 18 – Exemplo de correspondência de imagens.	66
Figura 19 – Representação gráfica da função de diferença de Gaussiana.	67
Figura 20 – Construção do descritor do ponto-chave.	68
Figura 21 – Filtros de caixa que aproximam derivadas gaussianas de segunda ordem.	69
Figura 22 – Representação gráfica da vizinhança $3 \times 3 \times 3$.	70
Figura 23 – Filtro wavelet Haar.	71
Figura 24 – Entradas do descritor de uma sub-região.	72
Figura 25 – Bag-of-Features.	73
Figura 26 – Os vetores de suporte num caso de separação linear e o hiperplano separador.	76
Figura 27 – Formas geométricas 3D básicas utilizadas para reconhecimento.	80
Figura 28 – Dinâmica de comunicação entre os módulos.	81
Figura 29 – Processo de detecção com o VOCUS2.	82
Figura 30 – Variando o parâmetro de segmentação do VOCUS2.	83
Figura 31 – Silhueta dos objetos.	84
Figura 32 – Diferenças dos eixos: Imagem capturada e Visão do Robô NAO	85

Figura 33 – Imagens da base de treinamento.	86
Figura 34 – Distribuição corpórea do robô NAO.	88
Figura 35 – Interface gráfica do Coregraphe.	89
Figura 36 – Visão geral do funcionamento do software do NAO.	90
Figura 37 – Fluxo de interação para sessão de figuras geométricas 3D individuais.	92
Figura 38 – Exemplo de árvore de decisão para adivinhar a figura escolhida pela criança.	94
Figura 39 – Fluxo de interação para sessão de múltiplas figuras geométricas 3D.	95
Figura 40 – NAO permanecendo na mesma posição para o grupo de baixa interatividade.	100
Figura 41 – Robô sentado esperando antes de uma sessão para o grupo de alta interatividade.	101
Figura 42 – Juízes avaliando individualmente as sessões 42a e em grupo 42b.	105
Figura 43 – Formulário de pontos dos juízes preenchido.	105
Figura 44 – Média dos pontos dos juízes por tarefa de todos os vídeos para o grupo de baixa 44a e alta 44b interatividade.	107
Figura 45 – Decadência do engajamento da criança ao decorrer da sessão de alta interatividade.	108
Figura 46 – NAO esperando pelas crianças no jogo de perguntas com as figuras geométricas.	112
Figura 47 – Recall, Precision e F-measure agrupados por classe dos objetos.	117

LISTA DE TABELAS

Tabela 1 – Resultados de (TANAKA; GHOSH, 2011), sendo \mathcal{I} a média de interações por minuto e \mathcal{E} a média de interações por minuto resultantes em reforço de aprendizado.	39
Tabela 2 – Tabela de Avaliação MOS no trabalho de Pinto <i>et al.</i> (2014a).	43
Tabela 3 – Diretrizes propostas por Riek (2012) para estudos de HRI por meio de WoZ.	46
Tabela 4 – Principais características entre os sistemas iNVT e VOCUS2.	55
Tabela 5 – Resumo das etapa dos estudos com WoZ.	99
Tabela 6 – Média obtida pelo MOS	102
Tabela 7 – Acertos no questionário de avaliação de conteúdo.	103
Tabela 8 – Média de pontos dos juízes para as tarefas do grupo de baixa interatividade.	106
Tabela 9 – Média de pontos dos juízes para as tarefas do grupo de alta interatividade.	106
Tabela 10 – Média obtida com o MOS para o jogo de perguntas.	112
Tabela 11 – Porcentagem das respostas corretas das crianças.	113
Tabela 12 – Diretrizes propostas por Riek (2012) para estudos de HRI por meio de WoZ.	114
Tabela 13 – Repostas para as diretrizes de Riek (2012) para os experimentos deste estudo.	115
Tabela 14 – Matriz de confusão para o classificador.	116
Tabela 15 – Medidas do classificador.	116
Tabela 16 – Teste positivo por um usuário com bom conhecimento da implementação.	118
Tabela 17 – Teste positivo por um usuário com médio conhecimento da implementação.	118
Tabela 18 – Teste positivo por um usuário com nenhum conhecimento da implementação.	118
Tabela 19 – Teste negativo por um usuário com bom conhecimento da implementação.	118

SUMÁRIO

1	INTRODUÇÃO	23
1.1	Motivação	27
1.2	Objetivos	29
1.3	Justificativa	29
1.4	Organização da Monografia	29
2	REVISÃO BIBLIOGRÁFICA	31
2.1	Robótica Educacional	31
2.2	Modelos de Redes Neurais Artificiais aplicados a Visão Computacional	33
2.3	Visão Computacional	35
2.4	Classificação de Imagens com SVM	35
2.5	Trabalhos em HRI	37
2.6	Considerações Finais	45
3	FUNDAMENTAÇÃO TEÓRICA	47
3.1	Modelos de Atenção Visual	47
3.1.1	Modelo de atenção visual bottom-up - (iNVT)	48
3.1.2	Extração de Características Visuais Primitivas	49
3.1.3	Mapa de Saliência	51
3.1.4	Seleção da atenção e inibição de retorno	54
3.2	Sistema de atenção computacional (VOCUS2)	54
3.3	Modelo de Atenção Visual bottom-up e top-down	58
3.4	Redes Neurais Artificiais	58
3.4.1	Redes Mult Layer Perceptron e Algoritmo BackPropagation	61
3.4.2	Rede LEGION	63
3.5	Extração de características de imagens	65
3.5.1	SIFT	65
3.5.2	SURF	68
3.5.3	Invariância de escala	69
3.5.4	Classificação do ponto de interesse	70
3.5.5	Orientação do ponto do atributo de interesse	71
3.6	Bag-of-Features (BoF)	72

3.6.1	K-médias	73
3.6.2	Detecção de objetos	73
3.7	Support Vector Machines (SVM)	74
3.8	Considerações Finais	76
4	SISTEMA PROPOSTO	79
4.1	Descrição	79
4.2	Materiais e Métodos	82
4.2.1	Sistema de Visão	82
4.2.2	Sistema de Diálogo	86
4.2.3	Robô Humanoide NAO	87
4.2.3.1	Hardware	87
4.2.3.2	NAOqi	87
4.3	Sessões interativas propostas	90
4.3.1	Figuras geométricas 3D individuais: Robô propõe desafio para as crianças	91
4.3.2	Múltiplas figuras geométricas 3D: A criança desafia o robô	93
4.4	Considerações Finais	93
5	EXPERIMENTOS E RESULTADOS	97
5.1	Estudo de Interação com Crianças	97
5.1.1	Sessões interativas iniciais	100
5.1.1.1	MOS	101
5.1.1.2	Questionário de avaliação de conteúdo	102
5.1.1.3	Continuous Audience Response	103
5.1.1.4	Casos gerais e específicos	108
5.1.2	Sessões em Grupo	109
5.1.3	Jogo de Perguntas	110
5.1.3.1	MOS	111
5.1.3.2	Resposta por grupo	113
5.1.4	Considerações Finais	113
5.2	Testes dos Modulos	113
5.2.1	Testes em base de dados para o sistema de visão	114
5.2.2	Testes do sistema	116
5.2.3	Considerações Finais	118
6	CONCLUSÃO	119
6.1	Desafios e limitações	120
6.2	Trabalhos futuros	120
6.3	Lista de artigos gerados e publicados	121



INTRODUÇÃO

Por se tratar de uma aplicação mais dinâmica da computação, ou seja, de resultados mais sensíveis no mundo real, a robótica se destaca dentre as demais áreas, pois fornece com maior interatividade uma maneira de relacionar o mundo real com o mundo virtual. Para dar mais autonomia a todos esses processos em robôs, tornar esses sistemas inteligentes é cada vez mais recorrente em todo o mundo. Desde a Revolução Industrial ([BEAUCHAMP, 1998](#)), aproximar o comportamento das máquinas ao comportamento racional, seja de animais ou até mesmo derivado da mente complexa do ser humano, vem sendo alvo de pesquisas no campo da Inteligência Artificial (IA). Esta ciência passa por diversas fases ([LUGER, 2004](#)) e é definida por [Kurweil \(1990\)](#) como a arte de criar máquinas que executem comportamentos que exigem inteligência.

Dentre as principais subáreas da IA, temos a de reconhecimento de padrões que é definida por [Jain, Duin e Mao \(2000\)](#) como a forma em que a máquina observa o meio, distingue padrões (entidade, objeto, processo ou evento) de interesse e toma decisões. Nesse contexto, ao longo da literatura se estudou, entre outros, reconhecimento de faces ([GASPAR, 2006](#)), reconhecimento de cadeias de proteínas ([VAKSER; MATAR; LAM, 1999](#)). [Halkin \(2001\)](#), por exemplo, podemos ver que por meio de Redes Neurais Artificiais (RNA), além de entretenimento, o reconhecimento pode ser usado na área médica, para identificação de anomalias fisiológicas em imagens radiográficas, na indústria para detectar padrões de falha em peças ou na defesa, como uma importante ferramenta de localização de alvos em imagens de satélite e aerofotogrametria.

O estudo e aperfeiçoamento dessas técnicas é importante, pois o fato de reconhecer as informações do ambiente, processá-las e usar o resultado desse processamento no comportamento do robô, que é algo concreto, porém controlado por decisões do universo virtual, faz com que este processo seja melhor compreendido por usuários de todos os tipos ([KIRBY; FORLIZZI; SIMMONS, 2010](#)).

Desde da migração das fábricas no século passado para tomar lugar em nosso cotidiano a

robótica desperta interesse e curiosidade no que ela pode proporcionar para o benefício humano. A convivência de pessoas e máquinas tornou-se alvo de pesquisa e preocupação do ponto de vista ético (RICK; WATSON, 2010; MILLER, 2010), da engenharia (BREAZEAL *et al.*, 2005), da saúde (RICH; CROWSON; HARRIS, 1987) e da própria imaginação das pessoas. Por exemplo, um dos trabalhos mais notórios da ficção científica sobre robótica é o do escritor Isaac Asimov. Asimov apresentou as leis da robótica que são:

1^a lei: Um robô não pode ferir um ser humano ou, por inação, permitir que um ser humano sofra algum mal.

2^a lei: Um robô deve obedecer às ordens que lhe sejam dadas por seres humanos, exceto nos casos em que tais ordens contrariem a Primeira Lei.

3^a lei: Um robô deve proteger sua própria existência, desde que tal proteção não entre em conflito com a Primeira e Segunda Leis.

Apesar de saírem da ficção, as leis de Asimov têm sido uma das principais diretrizes da área de *Human-Robot Interaction* (HRI)¹, um dos campos emergentes que surgiu com o advento da robótica e que estuda a interação de humanos com robôs. É dedicado à compreensão, concepção e avaliação de sistemas robóticos para uso por ou com seres humanos. Intereração, por definição, requer a comunicação entre os robôs e os seres humanos. A comunicação entre um ser humano e um robô pode tomar várias formas. O problema em HRI é entender e moldar as interações entre um ou mais seres humanos e um ou mais robôs (GOODRICH; SCHULTZ, 2007). Interações entre seres humanos e robôs são inherentemente presentes em toda a robótica, mesmo para os chamados robôs autônomos, afinal de contas, os robôs ainda são usados pelos seres humanos e estão fazendo o trabalho deles (GOODRICH; SCHULTZ, 2007). Como resultado, avaliando as capacidades dos seres humanos e robôs, e projetar as tecnologias e formação que produzem interações desejáveis são componentes essenciais do HRI. Esse trabalho é inherentemente interdisciplinar por natureza, exigindo contribuições da ciência cognitiva, linguística, psicológica e de educação; desde a engenharia, matemática e ciência da computação; até a engenharia de fatores humanos.

Embora a análise dos padrões de interação existente ser essencial, é útil adotar uma perspectiva de projeto para quebrar o problema de HRI em suas partes constituintes. De acordo com Goodrich e Schultz (2007), em essência, um projeto deve contemplar cinco atributos que afetam as interações entre seres humanos e robôs:

- Nível e comportamento de autonomia
- Natureza da informação trocada
- Estrutura do time

¹ em português Interação Humano-Robô

- Adaptação, aprendizagem e treinamento das pessoas e robôs e
- Formato da tarefa

Interação, o processo de trabalhar em conjunto para alcançar um objetivo, emerge da confluência desses fatores. O *designer* tenta entender e modelar a interação em si, com o objetivo de fazer o intercâmbio entre humanos e robôs trazer benefícios em algum sentido (RIEK, 2012). Para Kelley (1984), interação é um componente chave de um projeto de HRI válido.

Apesar da autonomia ser um dos componentes necessários para fazer uma interação benéfica, um segundo componente é a maneira pela qual as informações são trocadas entre o humano e o robô. Medidas da eficiência de uma interação incluem o tempo de interação necessária para a intenção e/ou instruções serem comunicadas pelo robô (CRANDALL *et al.*, 2005), a carga de trabalho cognitivo ou mental de uma interação (SHERIDAN, 2002), a quantidade de conhecimento da situação produzida pela interação (ENDSLEY, 2011) (ou reduzida por causa das interrupções do robô), e a quantidade de entendimento comum ou um nível de conhecimento comum entre humanos e robôs (JOHNSTON *et al.*, 2002; KLEIN *et al.*, 2005).

Existem duas dimensões principais que determinam a forma como as informações são trocadas entre um humano e um robô: o meio de comunicação e o formato das comunicações. As mídias primárias são delineadas por três dos cinco sentidos: visão, audição e tato. Estes meios são manifestados em HRI como (DAUTENHAHN *et al.*, 2006):

- Gestos, incluindo mãos e movimentos faciais e pela sinalização à base de movimento intencionais,
- Fala e da linguagem natural, que incluem tanto a fala auditiva e respostas baseadas em texto, e que frequentemente enfatizam a interação de diálogo e uma mescla de iniciativa de ambas as partes,
- Interações físicas e hápticas, frequentemente utilizadas remotamente em realidade aumentada ou em teleoperação para fornecer uma sensação de presença, e também frequentemente usadas para promover proximidade emocional, social, e as trocas de assistência humano-robô.

É importante ressaltar que as trocas de informações baseadas na voz devem não só abordar o conteúdo das informações trocadas, mas também as regras de tal troca (GRICE, 1991), o que perguntar e em que medida o discurso é verídico, relevante, claro e informativo. A apresentação das informações hápticas podem incluir avisos por meio de vibrações e comunicar informações específicas através de ícones táteis. A apresentação das informações de áudio podem incluir alertas auditivos e troca de informações baseadas em consciência 3D (ver, por exemplo, (TRAFTON *et al.*, 2006)). As informações sociais podem incluir chamada de atenção, gestos,

compartilhamento de espaço físico, imitação, sons, expressão facial, fala e linguagem natural ([SCHEEFF et al., 2002; KANDA et al., 2004; NAKAUCHI; SIMMONS, 2002](#)).

Para [Goodrich e Schultz \(2007\)](#), a visão do robô é um dos meios do qual mais se consegue extrair informações do ambiente (como por exemplo rapidamente identificar obstáculos) e do usuário (como reconhecer um gesto de comando ou uma expressão facial), sendo essa característica a que terá mais ênfase neste trabalho. Como reconhecer imagens está dentro dos estudos de reconhecimento de padrões, após algumas extrações de informações e métricas da imagem é possível fazer sua classificação por meio de Redes Neurais Artificiais (RNA), *Support Vector Machines* (SVM), *Naïve Bayes*, ou algum outro classificador da literatura.

A Robótica Social se caracteriza pelo fato dos robôs assumirem comportamentos sociáveis e interagirem com humanos e/ou outros robôs por meio de sentidos naturais, como fala, visão e tato ([SATAKE et al., 2009](#)). A busca por respostas para esse aperfeiçoamento na biologia vem sendo realizada com sucesso desde a metade do século passado ([WALTER, 1950](#)) até os dias de hoje, em aplicações bem sucedidas, como por exemplo a Robótica Assistiva ([BROEKENS; HEERINK; ROSENDAL, 2009](#)). Tal fato tem chamado a atenção de educadores e também de pesquisadores da robótica que trabalham para que essa tecnologia contribua também no processo educacional de crianças. A parte da robótica social que nos interessa leva o nome de Robótica Educacional, ou Robótica Pedagógica, e é caracterizada pela utilização de robôs como ferramenta no ensino, oferecendo situações-problemas nos quais são abordados na prática conhecimentos abstratos, para que o aprendiz passe por um processo cognitivo mais completo, ou seja, a forma com a qual ele absorve o conhecimento ([CHELLA, 2005](#)). Essa área tem como objetivo promover estudo de conceitos multidisciplinares, como física, matemática, geografia, entre outros, por meio da robótica e introduzir novos conceitos e tecnologias no ambiente educacional de forma mais concreta.

Os resultados em trabalhos nesse âmbito chamaram a atenção também de algumas indústrias, nacionais e multinacionais, que começaram a investir em produtos específicos para o ensino. Um exemplo local é a empresa pETE, situada na cidade de São Carlos que fornece material para escolas de todo o Brasil. Outro exemplo é a multinacional norueguesa LEGO, que criou uma divisão exclusiva para a área educacional, a LEGO *Educational Division*. Esta área se concentra em pesquisar e desenvolver materiais e métodos que atuam de forma efetiva na educação infantil do mundo todo. Também podemos citar a Olimpíada Brasileira de Robótica [OBR \(2014\)](#), que oferece aos alunos participantes a oportunidade de usar todo conteúdo aprendido no ensino fundamental e médio aplicados à robótica. Nas provas teóricas os alunos respondem perguntas de física, matemática e química com aplicação em robôs ao passo que na prova prática, existem equipes formadas por alunos da mesma escola e do mesmo nível que devem construir, a livre escolha, robôs capazes de realizar diferentes tarefas de acordo com o nível dos alunos e as equipes são classificadas de acordo com a eficiência nessas tarefas. As empresas antes citadas, pETE e Lego são fornecedoras de material para essa etapa.

A Figura 1 mostra esses produtos e a competição voltadas à Robótica Educacional. Nas figuras 1a e 1b os kits de pETE e LEGO e na figura 1c uma das modalidades da Olimpíada Brasileira de Robótica.

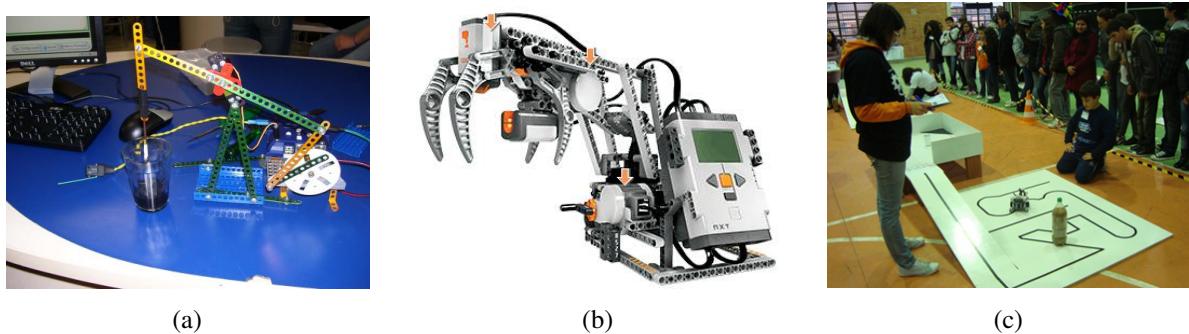


Figura 1 – Produtos e a competições voltadas à Robótica Educacional.

Fonte: [OBR \(2014\)](#).

Uma abordagem para uso da robótica educacional foi proposta por [Pinto et al. \(2014a\)](#) por meio de um modelo equipado com materiais e métodos desenvolvidos por alunos do Laboratório de Aprendizado de Robôs, onde este Mestrado foi realizado. Tal sistema é capaz de interagir com crianças utilizando o robô humanoide NAO ([ALDEBARAN, 2014](#)) integrado a um sistema de reconhecimento de figuras geométricas planas baseado na visão humana ([BENICASA et al., 2013](#)). Nesse trabalho (Figura 2), os alunos de 12 a 14 anos que participam do projeto Pequeno Cidadão, da USP de São Carlos, mostravam diversas figuras geométricas ao robô NAO que por meio de voz sintetizada, perguntava para a criança quantas formas geométricas similares - por exemplo quantos triângulos- ela conseguiu achar na figura mostrada. A criança, então, digita no teclado do computador conectado ao NAO o número de figuras por ela encontrado conforme a requisição do robô. O robô responde também por voz se a resposta estava certa ou errada e em caso negativo concede algumas dicas para uma outra tentativa. Foi notado que quando elas erram, o esforço para "ganhar" do robô na segunda chance se torna ainda maior.

1.1 Motivação

Atualmente, na era digital em que vivemos, não é novidade no contexto educacional o uso de tecnologias que auxiliam na promoção da interdisciplinaridade, no trabalho colaborativo, cooperativo, e compartilhado entre várias disciplinas. Isso tem se evidenciado no uso das redes sociais, Facebook, Youtube, blogs, dentre outros recursos digitais que a internet nos proporciona. Na medida em que se insere neste contexto, a Robótica Educacional torna-se mais efetiva e atraente, pois, além de se ampliar as possibilidades de recursos digitais associa-se a este processo o design, concepção, construção, e o controle via computador de dispositivos que os próprios alunos podem desenvolver e compartilhar com seus colegas via rede.

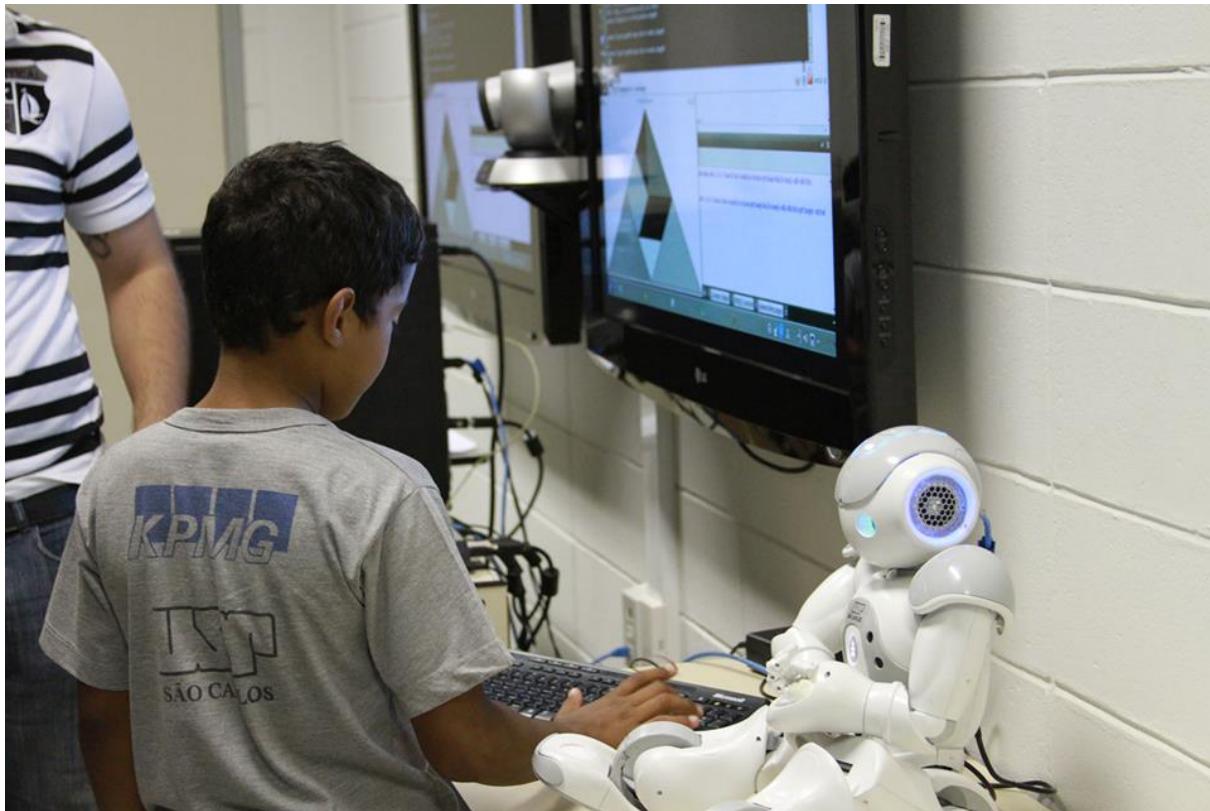


Figura 2 – Crianças brincando com o NAO no experimento de Pinto *et al.* (2014a).

Fonte: Pinto *et al.* (2014a).

Esta forma de aprender aumenta a autoestima dos alunos e o empoderamento deles quando percebem que conseguem construir e operar dispositivos robóticos cientificamente interessantes. Assim, os alunos percebem que deixam de ser meros usuários do computador e passam a atuar de maneira similar a especialista ao programar os robôs para executar tarefas propostas por eles, disponibilizando os seus intentos para que outros tenham acesso. Isso muda a forma como lidar com o conhecimento. Entretanto, para que isso se torne possível, é preciso que haja o engajamento da escola, dos professores, dos pais, da direção escolar, enfim, de toda uma comunidade educacional (GARCIA, 2002).

Dessa forma, os resultados obtidos pelo modelo de Pinto *et al.* (2014a), que empregou materiais e técnicas desenvolvidos pelo grupo de pesquisa do Laboratório de Aprendizado de Robôs, motivaram estudos para automatizar mais esse tipo de sistema, pois o mesmo encontra-se dependente de um teclado para a comunicação das crianças com o robô. Também ampliar o conteúdo abordado pelo robô é uma forma de contribuir para projetos similares futuros.

Como uma das principais ferramentas deste conjunto, o robô NAO tem capacidade de chamar a atenção com facilidade. Não apenas por sua similaridade com a aparência humana, mas também pelo seu repertório de recursos de hardware e software, discutidos na subseção (4.2.3).

1.2 Objetivos

Este projeto de mestrado tem como objetivo a combinação de técnicas de visão computacional, reconhecimento e síntese de fala e robótica social para construção de um sistema interativo que conduza sessões de interação por meio de um robô humanoide, podendo ser treinado com diferentes conteúdos a serem abordados com os usuários de forma autônoma. Para testes iniciais, o sistema foi treinado para brincar com crianças e reconhecer figuras geométricas 3D.

Problemas de pesquisa

Para explicitar como este projeto contribui para o estudo da arte foram propostas as seguintes problemas de pesquisa:

1. Como a autonomia e uma maior interatividade em um robô humanoide podem colaborar para a construção do conhecimento dos usuários?
2. Quais métodos a serem empregados viabilizam com eficiência a autonomia e as interações do sistema por meio de computação visual, reconhecimento e síntese de fala e recursos do próprio robô?

1.3 Justificativa

A maioria dos trabalhos encontrados em robótica educacional carecem de interação e de estudos centrados nos usuários, ou seja, que tiram suas métricas do público alvo ([BENITTI, 2012](#)). Acredita-se que uma maior autonomia do sistema pode aumentar o interesse e curiosidade das crianças mantendo seu foco por um período mais longo no conteúdo abordado pelo robô e, consequentemente, podendo gerar um ambiente favorável à construção do conhecimento. Como efeito secundário, este trabalho pode despertar nas crianças o interesse de seguirem na área tecnológica, haja visto um notório decréscimo de interesse na área chamada de *Science, Technology, Engineering, and Mathematics* (STEM) (Ciência, Tecnologia, Engenharia e Matemática) e suas carreiras, desde o advento da internet ([KUENZI, 2008](#)).

1.4 Organização da Monografia

No capítulo [2](#), é apresentada uma revisão bibliográfica sobre algum dos principais trabalhos relacionados a cada área deste projeto.

No capítulo [3](#), são explicados os métodos utilizados.

O capítulo [4](#) traz a descrição geral do sistema e o detalhamento da configuração dos métodos para a aplicação.

O capítulo 5 apresenta e discute os resultados dos experimentos desenvolvidos durante esta pesquisa e, finalmente, no capítulo 6, estão as considerações finais sobre o projeto e trabalhos futuros.



REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta a pesquisa dos trabalhos relacionados que forneceram a base científica para este projeto. Na seção 2.1, são apresentados os trabalhos de Robótica Educacional. Na seção 2.2, os trabalhos de redes Neurais Artificiais. A seção 2.3 traz os trabalhos de visão computacional. A seção 2.5 a evolução nos trabalhos de HRI. Finalmente a seção 2.6 apresenta um resumo deste capítulo.

2.1 Robótica Educacional

Para discutir o conceito de robótica educacional abordamos: [Mill et al. \(2008\)](#) e [Chella \(2005\)](#); para estudar as competências, [Perrenoud \(1999\)](#) ; para compreender o conceito de desenvolvimento cognitivo e construção do conhecimento, os autores estudados foram [Piaget \(1998\)](#), [Vigotsky \(1993\)](#),[Papert \(1994\)](#), [Coll \(1994\)](#) e [Frawley \(2000\)](#).

A respeito do desenvolvimento cognitivo, [Vigotsky \(1993\)](#) afirma que a verdadeira essência da memória humana (que a distingue dos animais) está no fato de os seres humanos serem capazes de lembrar, ativamente, com a ajuda de signos. Sendo assim, ajudar a criança associar coisas mais abstratas com objetos reais, como por exemplo formas geométricas, se torna uma forma divertida e eficiente de memorização, reforçando o aprendizado.

Há muito tempo se estuda a aproximação de objetos abstratos para a realidade. O Tangram, por exemplo, (Figura 3) é um quebra-cabeça chinês, inventado há quase mil anos. Até hoje ele encanta pessoas de todas as idades por ser um jogo simples de entender, porém com a dose certa de desafio. Seu objetivo é simples: formar as figuras pedidas usando todas as sete peças (conhecidas originalmente como tans). As peças são 2 triângulos grandes, 1 triângulo médio, 2 triângulos pequenos, 1 quadrado e 1 paralelogramo. Esse, como muito outros métodos, têm sido muito estudado por pesquisadores da área de educação para servir com elo entre o abstrato e o concreto de maneira divertida. Para [Bohning e Althouse \(1997\)](#) e [Lee, Lee e Collins \(2009\)](#), a

importância da noção espacial para alunos iniciantes de geometria é maior que a importância de saber seus nomes e diferenças, e o tangram apresentou bons resultados quando empregado para esse fim. Seu uso incluiu estudos para crianças com deficiência auditiva (GEMAQUE; SALES, 2013), e os resultados mostraram que o tangram aumentou o reforço no aprendizado também para esse grupo. Figueira-Sampaio *et al.* (2013) fizeram um estudo e mapeamento dos 29 principais objetos geométricos, dentre eles também o tangram, que auxiliam no ensino brasileiro e apontaram as melhores maneiras de utilizar cada um deles.

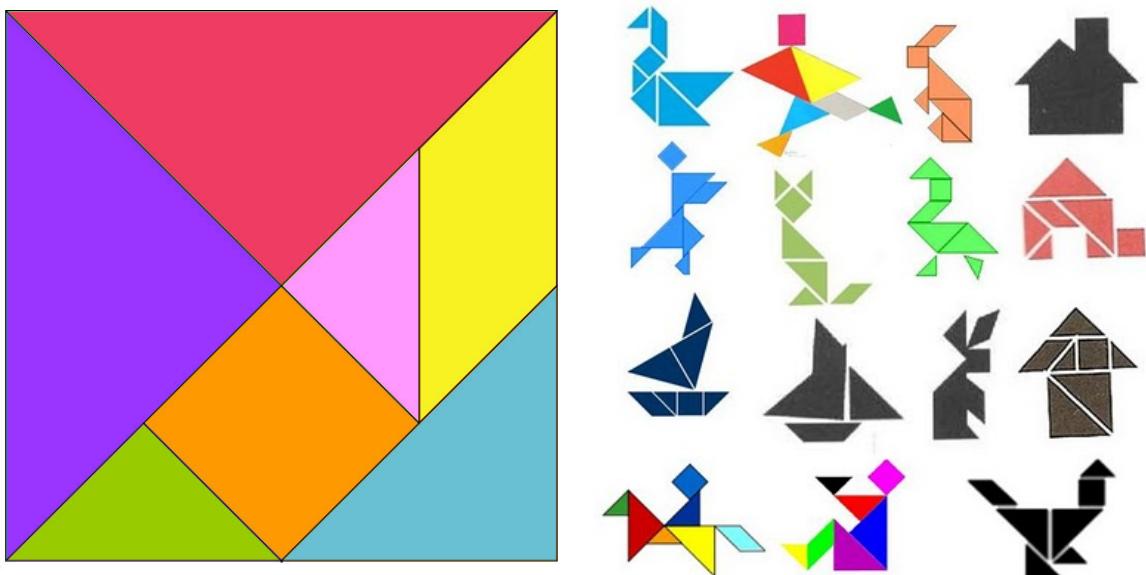


Figura 3 – Figuras do Tangram a esquerda e animais formados com essas figuras a direita.

Fonte: <http://ensinarevt.com/jogos/tangram> (Fevereiro de 2016).

Mais especificamente na Robótica Pedagógica, Chella (2005) afirma que o desenvolvimento do Ambiente de Robótica Educacional (ARE) foi fundamentado em princípios derivados da teoria de Piaget sobre o desenvolvimento cognitivo e revisados por Seymour Papert. Estas teorias sugerem que o centro do processo relacionado ao aprendizado é a participação ativa do aprendiz que amplia seus conhecimentos por meio da construção e manipulação de objetos significativos para o próprio aprendiz e a comunidade que o cerca.

Perrenoud (1999) mostrou que a robótica não tem o poder do aprendizado em si, mas sim serve como uma ferramenta poderosa, uma vez que a robótica e as tecnologias educacionais em geral não têm em si próprias o poder da construção dos conhecimentos, e sim o potencial para tal, porque fazem parte do contexto de vida das pessoas. Assim sendo, é necessário aprofundar os estudos sobre esse recurso para que sejam exploradas ao máximo as riquezas desse ambiente, que contribui para a formação de novas competências, aqui entendidas como a "capacidade de agir eficazmente em um determinado tipo de situação, apoiada em conhecimentos, mas sem limitar-se a eles".

No contexto nacional, d'Abreu e Bastos (2013) fizeram uma reflexão com professores de

uma escola estadual de Campinas sobre os métodos utilizados para o ensino *versus* os métodos tecnológicos disponíveis. Após a inserção do uso da robótica, foi notado um aumento na atenção dos alunos para questões mais difíceis de exatas. Ainda d'Abreu e Filho (2013) fazem um sumário dos 30 anos de atuação do Núcleo de Informática Aplicada à Educação (NIED), um instituto voltado especificamente para melhorias na educação através da informática. Nesse trabalho, são relatadas experiências bem sucedidas dessas aplicações em escolas de ensino fundamental. Na maior parte das experiências são utilizados os kits da *Lego Educational*, os quais possuem um processador conectado às peças de plástico reconfiguráveis, cujas interações são unidimensionais, ou seja, após a programação dos aparelhos, estes apenas seguem as instruções sem devolver nenhum tipo de *feedback*. Apesar disso, essa abordagem comprovou uma excelente eficiência na aplicação de conhecimentos teóricos para crianças e adolescentes.

2.2 Modelos de Redes Neurais Artificiais aplicados a Visão Computacional

Uma das técnicas mais utilizadas na tarefa de reconhecimento de padrões são as redes neurais artificiais (RNA) (CHAKRABORTY; PAL; CHATTERJEE, 2012). Wang (1995) destaca que, dentre os domínios de aplicações, é a tarefa na qual as RNA possuem maior potencial.

O uso das RNA aparece, segundo Iyengar e Kashyap (1991), como um bom modo de resolução, pois a construção dessas redes envolve um entendimento informal do comportamento, ao invés de criar procedimentos lógicos, resultando em uma implementação mais fácil.

Em Bezdek, Pal e K. (1992) destacam-se as quatro maiores vantagens do uso de redes neurais artificiais sobre muitas outras técnicas de reconhecimento de padrões.

- Adaptatividade: habilidade de se ajustar a novas informações;
- Velocidade: via o paralelismo massivo;
- Tolerância a falhas: capacidade de oferecer boas respostas mesmo com falta, confusão ou dados ruidosos;
- Optimalidade: visto a baixa taxa de erros em sistemas de classificação.

Redes neurais artificiais são métodos de modelagem recomendados para se lidar com sistemas abertos ou mais complexos, por serem adequadamente descritos por um conjunto de regras ou equações usualmente de baixo custo computacional. Como cores e formas geométricas são padrões bem conhecidos para se treinar a rede, sua escolha é justificável pela facilidade de implementação e rapidez de resposta, uma vez que a rede esteja treinada.

Egmont-Petersen, Ridder e Handels (2002) apontam diversos trabalhos - alguns deles listados abaixo - em que são utilizadas RNA no processamento de imagens, diretamente no processo de segmentação ou como ferramenta auxiliar.

Ong *et al.* (2002) propõe um método de segmentação de imagens coloridas baseado em classificação de cores usando mapas auto-organizáveis (SOM - *Self-Organizing Maps*). De forma não supervisionada, uma rede neural SOM 2D é utilizada para fazer um mapeamento e clusterização das cores nas imagens de treinamento e uma outra rede 1D extraía elementos representativos destes *clusters*. Essa segunda rede pode sofrer agrupamentos e divisões, para aglomerar cores muito próximas em um *cluster* ou dividir um outro em mais cores. Apesar de apresentar bons resultados, este trabalho não foi desenvolvido para aplicações de tempo real, sendo necessários ajustes para esta finalidade.

Waldherr, Thrun e Romero (2000) também discutem o processamento de cores para criar uma interface de reconhecimento de gestos para o controle de um robô. A partir da cor da camisa e do rosto era realizado o *tracking* do humano, reconhecendo-o para interpretar os gestos. Assim, a movimentação dos braços proporcionou o envio de comandos do tipo siga e pare, sendo que, no início do processo, a pessoa deveria ficar no centro de visão do robô. De modo parecido, em Tzafestas *et al.* (2009) foram classificados, também por meio de uma MLP, diversos gestos da mão para o controle de um robô móvel, obtendo uma acurácia de 98.5%.

Quiles (2004) propôs um sistema de visão criado para reconhecer objetos de cor e forma determinada, para o controle de robôs móveis. Os resultados para o reconhecimento de cores permitiram que o robô seguisse muito bem as cores determinadas (azul, vermelha e amarela) com o uso de Redes Neurais e aplicações de campos potenciais. Entretanto, o uso de apenas uma rede se mostrou ineficiente quando a tarefa era reconhecer cores e formas. Isso acontecia porque na etapa de pré-processamento, o reconhecimento de cores acabava por gerar uma imagem deformada do objeto, dependendo das condições de iluminação, escala, etc. Esse trabalho mostrou a importância de se utilizar as técnicas invariantes no tratamento de formas, para evitar esse tipo de erro.

Em Chavez (2002), foi também desenvolvido um modelo de reconhecimento de padrões invariantes mediante a extração das características, com uso de autômatos celulares e classificação com uso de RNA. Esse trabalho abordou principalmente a etapa de pré-processamento da imagem, retirando as características invariantes antes do resto do processamento. Essa abordagem superou as técnicas mais conhecidas na literatura (PCNN, Momentos de Hu e Zernick), quando a imagem era binarizada. Também foi construído um modelo de AC bi-dimensional, com as regras similares ao "Game of Life" para fazer a detecção de bordas em imagens binárias e em tons de cinza. O método apresentou uma maior eficiência computacional. Já Quiles, Romero e Zhao (2006) utilizaram uma PCNN (*Pulse Coupled Neural Network*) (LINDBLAD, 2005) para fazer um modelo aplicado diretamente à segmentação. Pulses sincronizados representavam os objetos da imagem, e grupos de neurônios fora de fase representavam os objetos distintos. Esse trabalho

conseguia separar objetos não-linearmente separáveis, tarefa bastante complicada para os demais algoritmos de segmentação

Da mesma forma, em [Muralidharan e Chandrasekar \(2010\)](#) foram usados os momentos de Hu para a retirada de características invariantes de imagens, classificando a rede com o uso do *K-nearest neighbors* (KNN) para a classificação dos objetos. Apesar de ser um método simples e de fácil implementação (já que classifica um ponto na rede segundo seus "k" vizinhos mais próximos), o KNN mostrou um bom resultado, com picos de 95% de acurácia.

2.3 Visão Computacional

Um sistema de visão computacional deve ser capaz de extrair apenas informações importantes nas imagens, fazer inferências a partir da incompletude de dados de entrada e conseguir identificar padrões com a maior independência possível em relação a mudanças de posição, orientação, tamanho e transformações geométricas. Essa é uma ideia compartilhada por autores como [Gonzalez e Woods \(2002\)](#), [Zhang e Lu \(2004\)](#), [Choksuriwong, Laurent e Emile \(2005\)](#), [Veltkamp e Latecki \(2006\)](#).

Em visão computacional, o que define um sistema de reconhecimento é que ele tenha as etapas de aquisição de imagens, pré-processamento, segmentação, extração de características e o reconhecimento ([LONG; ZHANG; FENG, 2005](#)). Assim, a maioria dos métodos propostos na literatura faz a simplificação da imagem, com a segmentação e detecção de bordas por exemplo, para em seguida extrair as características do padrão procurado. É sempre válido ressaltar o cuidado nas simplificações para que elas não sejam feita de forma demasiada, descartando algumas características importantes para o reconhecimento.

Para [Pedrini e Schwartz \(2008\)](#), [Liu e Yang \(2008\)](#), a cor é a propriedade mais importante para os humanos na análise de imagens. Por isso, é um dos aspectos mais relevantes de se extrair de uma imagem ([CONCI; AZEVEDO; LETA, 2008](#)). Assim, o reconhecimento de cores será a principal característica que guiará o sistema de visão para reconhecimento das formas propostas neste trabalho.

2.4 Classificação de Imagens com SVM

Em um trabalho revolucionário, [Chapelle, Haffner e Vapnik \(1999\)](#) mostrou que a classificação pode ser melhorada com base em histogramas das imagens utilizando Support Vector Machines (SVM). Antes disso, sabia-se que as abordagens de classificação generalizam mal para tarefas de classificação de imagens se a dimensionalidade do espaço das características fosse extremamente elevada, mas esta abordagem mostrou que o método de SVM pode realizar essa classificação facilmente se os únicos atributos fornecidos são histogramas de alta dimensionalidade. Eles usaram *kernels*, ou seja, uma transformada no conjunto de dados, do tipo RBF

heavy-tailed, mostrando uma diminuição no tempo de treinamento usando uma exponenciação e melhorando o desempenho do SVM linear que pode ser utilizada para substituir os *kernels* RBF.

Seguindo essa linha, [Bay, Tuytelaars e Gool \(2006a\)](#) utilizaram SVM de uma classe, de duas classes e multi-classe. Eles propuseram um *confidence-based dynamic ensemble* (CDE) - comitê dinâmico à base de confiança - de modo que puderam concluir quando é necessário retreinar o classificador e se novas características de baixo nível ou novos dados de treinamento podem ser incluídos. Um esquema de classificação de três níveis é proposto. No nível base, SVM são usados para calcular a previsão de um rótulo semântico. Um fator de confiança é dado para cada previsão empregando um algoritmo para SVM de uma classe que também usa uma distribuição de densidade do conjunto dados do treinamento. A nível multi-classe, os fatores de confiança dos classificadores são acumulados para dar apenas uma previsão. Mais uma vez um fator de confiança de nível de multi-classe é calculado para esta previsão. No nível geral, o CDE acumula as previsões das outras camadas para dar uma previsão agregada. Um fator confiança geral é dado neste nível. Se ele for alto, uma semântica é atribuída. Esta abordagem supera as desvantagens dos classificadores estatísticos tradicionais, uma vez que faz ajustes para incluir semânticas que conduzem à descoberta de características de baixo nível, melhorando a precisão.

[Shukla, Mishra e Sharma \(2013\)](#) utilizaram uma combinação de técnicas baseadas em características como mostrado na Figura 4 para anotação automática das imagens e o testaram com o conjunto de dados Social20 ([LI; SNOEK; WORRING, 2009](#)). Este conjunto possui 20 classes e um total de 19,972 imagens, sendo aproximadamente 1000 imagens pertencentes unicamente a uma classe. As classes são muito distintas como avião, praia, barco, ponte, ônibus, borboleta, carro, arquitetura da cidade, sala de aula, cão, flor, porto, cavalo, cozinha, leão, montanha, rinoceronte, ovelhas, rua, e tigre. Os autores concluíram que esta técnica é uma maneira eficiente de anotação automática de imagens, alcançando uma acurácia de 91% e um tempo de treinamento em 1.25 segundos, que é um bom indicador que esta abordagem pode ser facilmente integrada a um sistema que requer anotação automática de imagens.

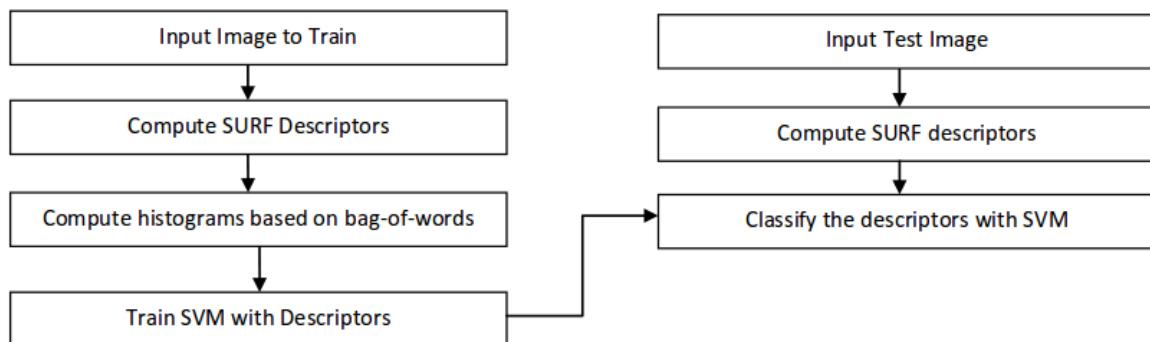


Figura 4 – Sistema utilizado em [Shukla, Mishra e Sharma \(2013\)](#).

Fonte: [Shukla, Mishra e Sharma \(2013\)](#).

[Balayil e Anees \(2014\)](#) apresentaram uma abordagem muito simples e intuitiva para

executar rotulação multi-classe de imagens usando modelo de representação *bag-of-words* e duas passagens pelo KNN para a classificação. A abordagem de classificação modificada considera a co-ocorrência de rótulos dentro do conjunto de dados, juntamente com a classificação obtida utilizando duas passagens pelo KNN, reduzindo assim as chances de atribuir esses rótulos que dificilmente co-ocorrem. Apesar de sua simplicidade, este método fornece um desempenho razoável em comparação com os métodos do estado-da-arte para resolver o problema. Ele dá um valor de *recall* melhor quando empregado SIFT, enquanto o uso de SURF oferece uma resposta mais rápida. A relação entre rótulos preditos são facilmente obtidos utilizando uma ontologia predefinida.

2.5 Trabalhos em HRI

Os trabalhos relacionados com o robô NAO apresentam diversos focos de pesquisa, muitos deles não relacionados diretamente com aspectos pedagógicos. Dentre tais abordagens, destacam-se as seguintes: [Suay e Chernova \(2011\)](#), que utilizaram o sensor *Kinect* junto com um humanoide NAO para permitir que o robô imitasse o movimento dos braços do usuário, tivesse seu caminhar ordenado e o olhar do robô fosse direcionado pela posição da mão direita para ordem de deslocamento. Essas funcionalidades foram obtidas pelo estabelecimento de alguns modos de controle com relação à posição de alguns membros do corpo do usuário.

Já [Veltrop \(2012\)](#) fez uso da fusão de vários sensores para teleoperar um humanoide NAO. Para tal, utilizou o sensor *Kinect* para estimativa das configurações espaciais do corpo do usuário; um controle *Wiimote* para melhor estimativa das configurações da mão/punho e ativação de determinados comportamentos no robô; uma esteira para permitir o robô caminhar conforme o caminhar do usuário na esteira evitando, assim, que o mesmo saísse do campo de visão do *Kinect*; e um *head-mounted display* servindo como a visão do robô no usuário.

[Koenemann e Bennewitz \(2012\)](#) utilizaram uma vestimenta de sensores inerciais no corpo de um usuário a fim de capturar as configurações das junções de seu corpo e mapeá-las para as junções do robô em tempo-real. Também, foi desenvolvido um módulo de balanço que permitisse o robô se comportar de forma estável durante a imitação. O robô humanoide desempenhou a imitação adequadamente durante a interação, permitindo também a estável imitação do movimento de ambas as pernas sem cair no chão.

[Zuher e Romero \(2012\)](#) usaram técnicas simples de matemática para tornar o humanoide NAO capaz de receber comandos, através da interação com o *Kinect*, em tempo real para caminhar, manipular objetos com suas mãos e realizar alguns comportamentos pré-definidos (sentar, se levantar e acenar com a mão) além de imitar os movimentos de braços e pernas de uma pessoa.

Em um estudo de interação social feito por [Tapus et al. \(2012\)](#), um experimento com quatro crianças foi proposto, no qual a interação com o robô NAO e com um humano era

comparada. Duas dessas crianças não mostraram qualquer efeito com a presença do robô, porém as outras duas prestaram mais atenção ao robô do que ao humano, sendo que uma delas mostrou uma grande interação com o robô. Já em outro teste, feito por [Csala, Nemeth e Zainko \(2012\)](#), crianças que são forçadas a ficar em pequenas caixas estreitas de 2x3 metros devido a uma cirurgia tiveram respostas bem positivas ao robô NAO, que as animava e convidava para fazer alguns exercícios.

[Kimberlee et al. \(2013\)](#) investigaram o uso de robôs para aumentar a comunicação e a atenção de adolescentes com transtorno de espectro autista. Para tal, adolescentes com autismo eram colocados para jogar juntamente com três outros adolescentes com outros distúrbios, recrutados de escolas de pessoas com necessidades especiais. Os testes foram feitos em três dias, consecutivos, onde os adolescentes autistas jogavam de três diferentes formas: a) com um robô humanoide, b) com uma *Smart Board* e c) com cartas. Apesar de apresentarem comportamentos individualistas nos três modos, o comportamento repetitivo foi diminuído quando utilizado tanto o robô como a *Smart Board*, mostrando que é possível o uso de robôs para ajudar no aprendizado. [Shamsuddin et al. \(2012a\)](#) também obtiveram ótimos resultados com crianças autistas quando em contato com o NAO. Além disso, a pesquisa do estado-da-arte feita em [Shamsuddin et al. \(2012b\)](#) mostra que o uso dos robôs vêm sendo testado por muitos pesquisadores, com bons resultados.

[Verner et al. \(2012\)](#) utilizaram diversos robôs para aprimorar a experiência de visitantes a um museu de tecnologia. Dentre eles, o robô NAO para interagir com o público explicando algumas sessões do museu e respondendo algumas perguntas pré-programadas dos visitantes. Algumas das conclusões que podem-se destacar foi o aumento do número de visitantes após a implementação dos robôs; o aumento do interesses das pessoas - medido pela média de perguntas por minuto; e o quanto as pessoas memorizavam em relação ao que foi ensinado, medido por um questionário ao final da visita. Todas essas características foram melhoradas com o apoio dos robôs.

[Smolar et al. \(2011\)](#) estudaram o processo cognitivo do próprio robô, ou seja, a forma de aprendizado por meio de sentidos como visão, audição, etc. Já [Tanaka e Ghosh \(2011\)](#) inverteram o papel do NAO no ensino. Nessa pesquisa, eles programaram o robô para participar das aulas de inglês para crianças japonesas (Figura 5), na universidade de Tsukuba, durante 3 dias e monitoraram as interações, gravando as 3 aulas pelas câmeras do NAO e posteriormente assistindo aos vídeos. O robô participava de todas as atividades junto com as crianças, como cantar cantigas em uma roda e responder perguntas de cores e animais feitas pela professora. No primeiro e segundo dia, ele foi programado para responder todas as respostas corretas e no terceiro dia para errar todas as respostas, e neste dia o professor não estava na sala. Dessa forma, no último dia as crianças advertiam o robô e explicavam o porquê de suas respostas estarem equivocadas. As crianças criavam maneiras particulares de explicar as coisas ao robô e, com isso, potencializavam seus próprios processos cognitivos. Mediram, também, o número de

interação por minuto e descobriram que o número de respostas corretas dadas pelas crianças, tanto individual quanto coletivamente, aumentavam proporcionalmente com a interação e que isso reforçava o aprendizado. Como mostrado na Tabela 1. Os autores, que nesse trabalho se referem a interação como *care-giving instance* - prestação de cuidados na tradução literal - concluíram que, apesar de cedo para se afirmar e que nem toda interação resultava em reforço de aprendizado, a inserção do robô potencializou o processo de reforço de aprendizado.

Ensaios	\mathcal{I}	\mathcal{E}
Dia 1	0,38	0,25
Dia 2	0,67	0,27
Dia 3	0,77	0,63

Tabela 1 – Resultados de (TANAKA; GHOSH, 2011), sendo \mathcal{I} a média de interações por minuto e \mathcal{E} a média de interações por minuto resultantes em reforço de aprendizado.

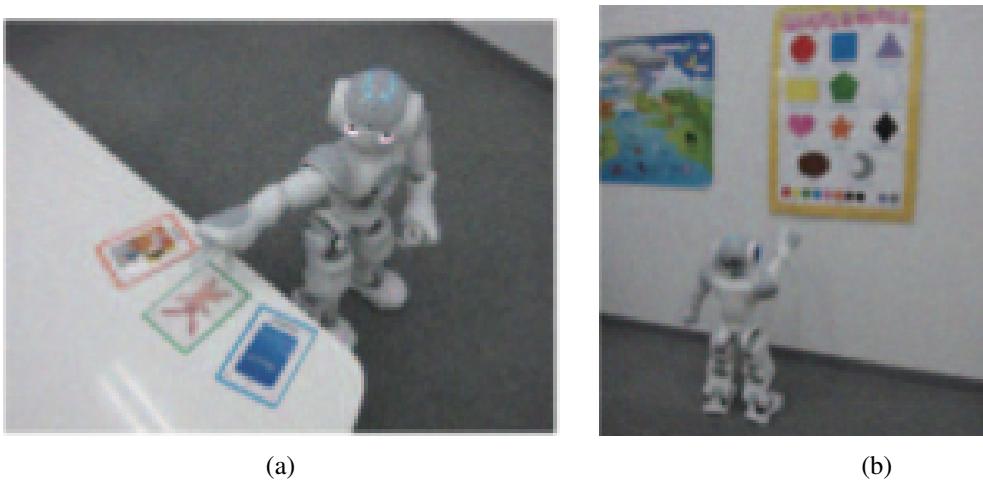


Figura 5 – Material utilizado em Tanaka e Ghosh (2011): (a) cartas e (b) tabela com formas geométricas de diferentes cores.

Fonte: [Tanaka e Ghosh \(2011\)](#).

Tanaka, Cicourel e Movellan (2007) utilizaram o humanoide QRIO, para estudar a interatividade com crianças de 1 ano e meio a 3 anos. O objetivo era desenvolver e avaliar métodos interativos para ajudar os professores na educação infantil. Fizeram um total de 45 sessões com as crianças, com duração média de 50 minutos cada. As sessões terminavam quando o robô ficava sem carga na bateria, ponto em que ele assumia uma postura de dormir. O estudo teve três fases: durante a fase I, que durou 27 sessões, o robô interagia com as crianças usando seu repertório comportamental completo. Durante a fase II, que durou 15 sessões, o robô foi programado para produzir comportamentos interessantes, porém altamente previsíveis. Durante a fase III, que durou três sessões, os robôs foram reprogramados para apresentar seu repertório completo. Todas as sessões de campo foram registradas por meio de duas câmeras de vídeo. Dois anos foram passados a estudar os vídeos e foram desenvolvidos métodos quantitativos para as suas análises. No começo da fase I as interações eram simples se limitando à alguma saudação

sem contato ou apenas seguindo alguma criança, porém elas foram aumentando e ao final desta fase eram mais complexas, como o robô rir ao ser tocado na cabeça, dançar quando detectar alguma música ou cutucar alguém como forma de chamar.

Uma das maiores preocupações dos autores foi encontrar uma maneira de avaliar a qualidade das interações, o tempo que elas duravam e a resposta, positiva ou negativa, das crianças. Descobriram que os métodos de resposta às audiências contínuas, originalmente utilizado para a pesquisa de marketing (FENWICK; RICE, 1991), foram particularmente úteis. Quinze sessões foram selecionadas aleatoriamente a partir das 45 sessões e independentemente codificadas *frame-by-frame* por cinco alunos de graduação que não sabiam do propósito do estudo. Os estudantes avaliaram as interações e deram nota para cada um delas, classificando-as de 0 a 2.2. Os resultados são mostrados na Figura (6) onde . (A) Vídeo assistido pelos avaliadores. (B) Pontos azuis são as médias das notas dadas pelos avaliadores nas 15 sessões escolhidas em relação a qualidade de interação. A reta vermelha é uma regressão linear dos pontos. (C) Confiabilidade inter-observador entre quatro avaliadores filtrada passa-baixo como constante de amortecimento. (D e E) Principais efeitos na qualidade de interação em relação ao tempo e numero da sessão respectivamente.

Os resultados mostraram que, ao passar do tempo, crianças começaram a tratar o robô mais como um amigo do que como um brinquedo devido às interações o aproximarem de sua realidade. Também o estudo constatou que as crianças prestavam mais atenção no robô inicialmente por ser uma novidade e por sua beleza estética, mas a longo prazo elas perdiam o interesse e aumentar a interação fazia com que se concentrarem por mais tempo. Concluíram também que a tecnologia atual da robótica é surpreendentemente perto de alcançar a socialização autônoma com crianças por períodos prolongados de tempo e que isso tem um grande potencial em contextos educativos que ajudam professores a enriquecer o ambiente de sala de aula.

Pinto *et al.* (2014a) utilizaram o robô humanoide NAO para o ensino de figuras geométricas simples, sem sobreposição, tais como, triangulo, quadrado e retângulo. A abordagem do experimento consistia em um jogo de perguntas, no qual o robô dava dicas sobre uma figura geométrica, tais como quantidade de lados, fórmula da área, fórmula do perímetro, que a criança deveria apresentar para ele. Após ouvir a dica do robô, o aluno escolhia uma das diversas figuras presentes na mesa e o robô indicava se a resposta estava correta. Caso contrário, o humanoide NAO dava uma nova dica e, ao final, sempre explicava e dava dicas sobre a figura que deveria ser reconhecida.

Em outro trabalho (PINTO *et al.*, 2014b), os mesmos autores realizaram um experimento no qual o robô NAO reconhecia figuras geométricas sobrepostas, como mostrado na Figura (7). Para entender melhor os resultados desse experimento, as crianças foram separadas em 2 grupos: o grupo que participaria do desafio com o robô e um segundo grupo chamado grupo de controle. Ambos os grupos eram acompanhados por especialistas na área de ensino infantil de matemática.

No primeiro grupo, as crianças entravam uma a uma na sala de aula em que o robô estava.

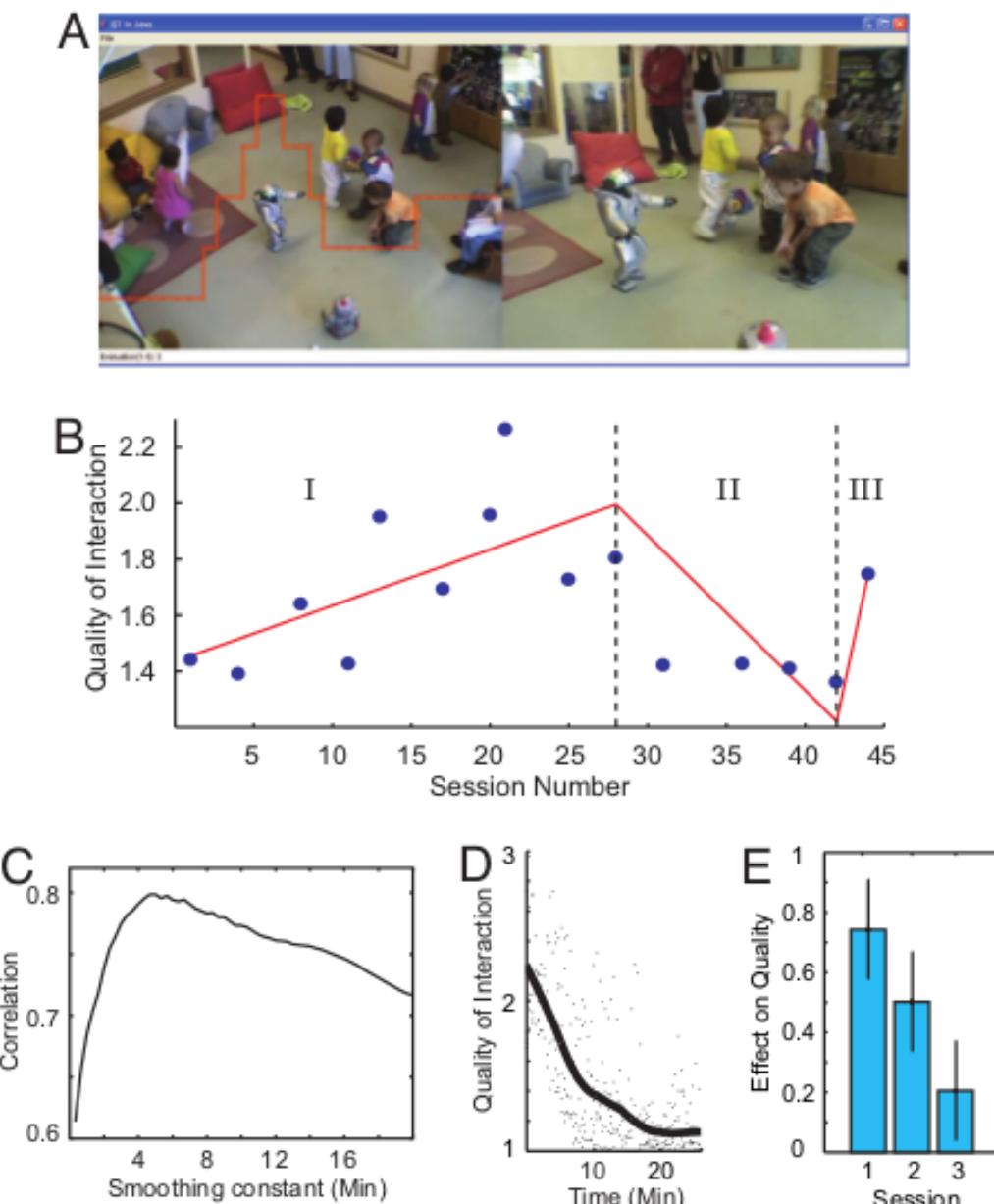


Figura 6 – Analise da qualidade de interação em Tanaka, Cicourel e Movellan (2007).

Fonte: Tanaka, Cicourel e Movellan (2007).

O robô fazia uma rápida apresentação utilizando alguns dos vários recursos nele presentes, como síntese de fala, *Led's* piscando, movimentação das mãos em forma de saudação, etc. Após isso, a criança apresentava um papel com uma figura conhecida, como um barco ou uma casa, formada por figuras geométricas do jogo Tangram (Figura 8). O NAO perguntava a criança quantas figuras ela conseguia reconhecer na imagem apresentada. Nenhuma das crianças encontrava na primeira tentativa todas as figuras presentes devido a sobreposição de algumas delas. O robô então respondia de forma desafiadora que havia encontrado mais formas que a criança e explicava que elas poderiam estar "escondidas" pelo fato de serem sobrepostas. Como consequência, todas as crianças que participaram desse experimento se empenharam mais na segunda tentativa para

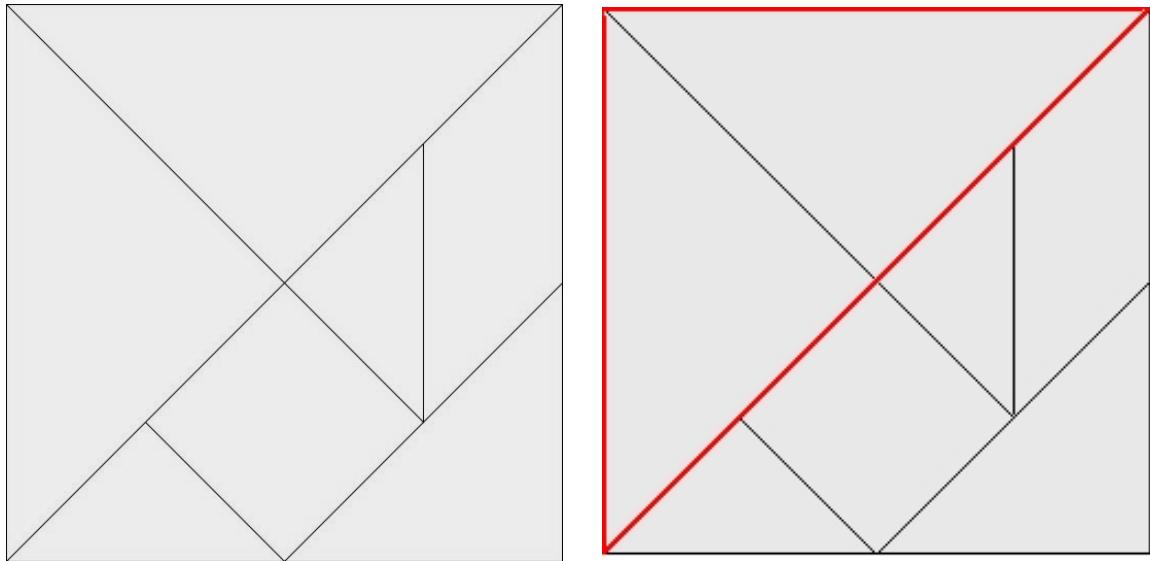


Figura 7 – Figuras escondidas em sobreposições (PINTO *et al.*, 2014b).

Fonte: Pinto *et al.* (2014b).

vencer o robô. Algumas até conseguiram achar mais figuras que ele.

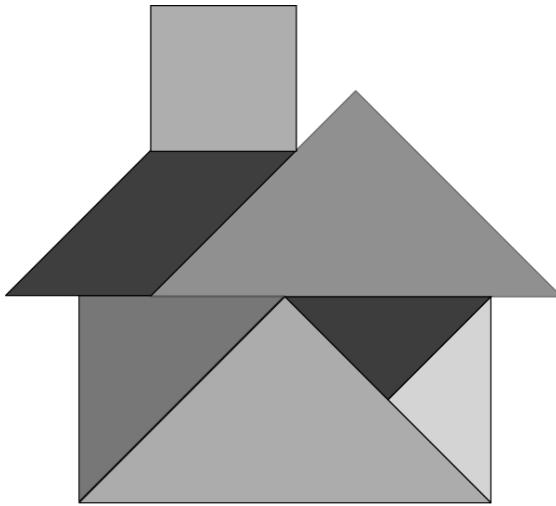


Figura 8 – Figuras geométricas formando objetos conhecidos como uma casa.

Fonte: Pinto *et al.* (2014b).

O educador aqui apenas intervia em caso de desvio do comportamento esperado, ou seja, alguma dúvida que a criança pudesse ter ou em alguma resposta equivocada do robô. Já no grupo de controle quem conduzia o a todo o processo era o próprio educador. Neste cenário, o professor conduzia o experimento no lugar do robô fazendo as perguntas e desafiando as crianças.

Ao final, os participantes dos dois grupos e os especialistas responderam a questionários de avaliação dos experimentos, o Mean Opinion Score (MOS). No MOS são tiradas as médias de pontuação de perguntas feitas aos usuários para tirar opiniões gerais dos ensaios. As perguntas para as crianças que participaram do teste com o robô são listadas a seguir e as respectivas

respostas estão apresentadas na Tabela 2:

1. O que você achou do tempo de resposta e ação do robô?
2. Você conseguiu entender tudo o que o robô falou?
3. O robô deu alguma resposta errada durante a atividade?
4. Agora que você finalizou as atividades, o que você acha sobre a robótica?
5. Você acha que o robô pode ajudar com seus estudos?
6. O que você acha de ter um robô como professor?
7. O que você acha de tentarmos outras atividades com os robôs?

Tabela 2 – Tabela de Avaliação MOS no trabalho de Pinto *et al.* (2014a).

MOS (%)	Excelente	Bom	Regular	Ruim	Péssimo
1	40.3	25.84	24.19	0	9.67
2	79	21	0	0	0
3	0	0	0	0	0
4	88.7	4.83	3.22	3.22	0
5	50	16.12	9.67	24.21	0
6	6.48	25.8	32.25	20.96	14.51
7	100	0	0	0	0

Fonte: Pinto *et al.* (2014a).

A conclusão foi de que todas as crianças do grupo sem contato com o robô se mostraram menos motivadas e após cerca de 3 minutos davam respostas aleatórias apenas para acabar o resto mais rápido, ao passo que no outro grupo a maioria delas levavam de 6 a 10 minutos tentando vencer o robô. Também importante destacar que todos os participantes do primeiro grupo responderam que gostariam de repetir o processo. Já os do segundo grupo apenas uma pequena porcentagem dos participantes afirmaram que repetiriam o teste.

Wizard of Oz

Uma técnica comumente empregada no repertório dos pesquisadores de HRI é a técnica *Wizard-of-Oz* (WoZ), proposta por Kelley (1984). Este modelo é baseada na famosa obra literária "O Maravilhoso Mágico de Oz"¹ no qual um personagem ilusionista utiliza de artimanhas para teleoperar um avatar que todos conhecem como o poderoso mágico, ou feiticeiro, da terra de Oz sem que os outros personagens saibam. Assim, WoZ refere-se a uma pessoa (geralmente quem está conduzindo o experimento, ou um parceiro) operando remotamente um robô, controlando

¹ *The Wonderful Wizard of Oz*

qualquer uma de uma série de coisas do repertório do robô, tais como o seu movimento, navegação, fala, gestos, etc. WoZ pode envolver qualquer quantidade de controle ao longo do espectro de autonomia, de totalmente autônoma para totalmente teleoperado, bem como a interação iniciativa mista. Por uma questão de conveniência, ficou estabelecido ao longo dos anos a taxonomia de Mágico - originalmente *Wizard* - para quem controla o robô, e Oz para o robô que está sendo controlado.

Os pesquisadores que empregam o WoZ argumentam que os robôs não são suficientemente avançados para interagir autonomamente com as pessoas de maneiras socialmente adequadas ou fisicamente seguras, assim, este tipo de manipulação permite aos participantes imaginar como a interação autônoma poderia ser no futuro. Eles também podem empregar este método para testar aspectos preliminares de sua concepção que ainda não foram totalmente implementados, como parte de um processo de design interativo de um projeto maior (RIEK, 2012).

No entanto, alguns pesquisadores têm levantado preocupações metodológicas a respeito do uso dessa técnica. Para a interação social, Weiss (2010) sugere que um robô controlado WoZ está servindo mais como um proxy para o ser humano e menos como uma entidade independente. Assim, não é de fato uma interação homem-robô, mas sim uma interação humano-humano por meio de um robô.

Na área de HRI, tradicionalmente não houveram critérios explícitos para guiar as pesquisas nesta tarefa, mas a comunidade *Natural Language Processing* (NLP), traduzido como Processamento de Linguagem Natural (PLN), criaram alguns. De acordo com Fraser e Gilbert (1991), para realizar uma simulação de WoZ válida, os seguintes requisitos devem ser atendidos:

1. Deve ser possível simular o sistema futuro, dadas as limitações humanas;
2. Deve ser possível especificar os futuros comportamentos do sistema;
3. Deve ser possível fazer a simulação convincente.

Fraser e Gilbert (1991) discutem o controle de vários aspectos sobre o comportamento do Mágico, tais como variáveis de reconhecimento (ou seja, o Mágico é perfeito? Como o erro é controlado?), variáveis de reação do Mágico (ou seja, o quanto rápido o Mágico deve responder?), e formação do Mágico (o Mágico teve tempo de treinamento suficiente para ser o mais semelhante possível ao sistema autônomo previsto?).

Green, Huttenrauch e Eklundh (2004) sugeriram a seguinte metodologia para construção de um cenário WoZ em HRI:

- **Instrução do usuário** fornece informações ao usuário para ajudá-lo a executar a tarefa pretendida. (O que o usuário deve fazer?)

- **Hipótese de comportamento** reflete as expectativas do designer sobre as ações de usuários dentro do cenário. (O que espera-se que o usuário faça?)
- **Comportamento do robô** especifica o que o robô vai fazer, para qualquer nível de autonomia. (O que o robô deve fazer?)

[Steinfeld, Jenkins e Scassellati \(2009\)](#) argumentam que é importante para os estudos WoZ também levar em conta como o ambiente pode afetar o robô e o humano, como o nível de autonomia do robô pode afetar a forma como os seres humanos o controlam, e, em alguns casos, como modelar e simular as ações do Mágico.

Assim, tendo em vista técnicas experimentais centradas em um ou mais humanos e/ou robô durante todo o ciclo de vida do projeto, incluindo não só *Wizard of WoZ*, mas também *Wizard with Oz* (método centrado no humano que utiliza a tecnologia real em um ambiente simulado), *Wizard and Oz* (método centrado no humano que utiliza a tecnologia real em um ambiente real), *Oz with Wizard* (método centrado no robô que inclui seres humanos, mas não se tiram métricas deles), *Oz of Wizard* (método centrado no robô no qual os seres humanos são simulados ou minimamente envolvidos) e *Wizard nor Oz* (todos os aspectos do sistema são simulados).

Finalmente, retornando ao trabalho original de [Kelley \(1984\)](#), que introduziu a técnica de WoZ para a comunidade, interação é um componente chave de um projeto válido. A simulação WoZ não se destina a ser o ponto final de uma questão de pesquisa, mas parte de um processo interativo em que o Mágico será gradualmente eliminado dos ensaios cedendo lugar à autonomia ([KELLEY, 1984](#)). Sua utilização é uma forma de mitigar as preocupações de engenharia, além das questões éticas e metodológicas que essa técnica traz e projetar experimentos que a empreguem de uma forma que permita uma transição suave para um sistema mais autônomo no futuro. Baseando-se nos problemas e abordagens apontadas nessa subseção, [Riek \(2012\)](#) propuseram algumas diretrizes para criação de um cenário favorável ao WoZ mostradas na Tabela 3.

2.6 Considerações Finais

Este capítulo apresentou os trabalhos que contribuíram para as decisões de implementação tomadas neste projeto. A seção 2.1, apresentou os trabalhos de Robótica Educacional. Na seção 2.2, foram vistos os trabalhos de redes Neurais Artificiais, mostrando sua aplicação e sua vantagem quando comparada a outros métodos. A seção 2.3 mostrou um pequeno histórico dos trabalhos iniciais sobre visão computacional. A seção 2.5 relatou alguns trabalhos de HRI, começando de uma forma mais genérica e depois convergindo para robôs humanoides. Também, nessa seção, foi discutido diferentes diretrizes para a técnica de WoZ.

Tabela 3 – Diretrizes propostas por Riek (2012) para estudos de HRI por meio de WoZ.

Fonte: Adaptada de Riek (2012).

Componente Experimental	Questões
Robô	Quantos robôs foram usados?
	Qual(ais) tipo(s) de robô(s)? (ex: Humanoide, zoomorfo, androide, etc)
	Qual o nível de autonomia? (Quais componentes eram do robô eram autônomos e quais eram controlados pelo Mágio?)
	Quais eram as habilidades do robô?
Usuário	Quais hipóteses os pesquisadores tem para com o robô?
	Quantos usuários participaram no total, e por bateria de experimentos?
	Qual a formação dos participantes?
	Quais instruções foram fornecidas aos usuários?
Mágio	Quais hipóteses os pesquisadores tem sobre os usuários?
	A simulação foi convincente para os usuários?
	Quais expectativas os usuários tiveram com relação ao robô antes e depois do experimento?
	Quantos Mágios foram utilizados?
Geral	Quais foram os dados demográficos do Mágio? (ex: o pesquisador, colegas de laboratório, terceiros?)
	O Mágio sabia das hipóteses comportamentais do experimento?
	Quais foram as variáveis de reação do Mágio e como elas foram controladas?
	Quais foram as variáveis de reconhecimento do Mágio e como elas foram controladas?
	Como foi tratado o controle de experimento para o erro do Mágio (deliberativo ou acidental?)
	Quanto e de qual tipo de treinamento o Mágio recebeu a priori para conduzir o experimento?
	Aonde o experimento foi realizado?
	Quais foram as variáveis do ambiente e como elas foram controladas?
	Quais cenários os cenários empregados pelos pesquisadores?
	Este experimento foi parte de um processo de design?
	Este estudo discute as limitações do WoZ?

CAPÍTULO
3

FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresenta a fundamentação teórica para os métodos utilizados neste trabalho. Na seção 3.1, os métodos de atenção visual. Na seção 3.2, o VOCUS2. Na seção 3.3, modelos *bottom-up* e *top-down*. Na seção 3.4, as redes neurais. Na seção 3.5, extrações de características. Na seção 3.6, o *bag-of-features*. Na seção 3.7 SVM e, finalmente, na seção 3.8 uma conclusão do capítulo.

3.1 Modelos de Atenção Visual

A atenção visual - alvo de muitas pesquisas nos últimos anos - apresenta vários modelos propostos que variam de acordo com a modelagem e a realização. [Tsotsos \(2011\)](#) e [Borji e Itti \(2013\)](#) apresentam um vasto estudo tomando essa abordagem.

Para este trabalho de mestrado, consideraremos trabalhos anteriores baseados nas seguintes hipóteses:

- Mapa de Saliência: A partir da extração de características específicas são calculados mapas da integração dos estímulos locais. A partir deste mapa é possível descobrir qual característica é mais saliente.
- Atenção Emergente: Interações competitivas entre uma grande quantidade de neurônios geram a definição da região de maior atenção, enviesada por um mecanismo *top-down*.
- Correlação Temporal: Sistemas compostos por grupos de neurônios com conexões excitatórias e inibitórias, cujo dinamismo é regido por equações diferenciais

[Itti e Koch \(2001\)](#) definem mapa de saliência como um processo rápido de busca pelo local de maior atenção e é dirigida predominantemente de forma *bottom-up*. Estímulos salientes,

acima de um *threshold*, se destacarão automaticamente na cena, o que faz essa abordagem sensível ao contexto (OGAWA; KOMATSU, 2004).

Já Koch e Ullman (1987) utilizaram mapas de saliência para direcionar a atenção a pontos de maior interesse na imagem como uma proposta de atenção seletiva, nos quais cada característica primitiva (orientação, cor ou intensidade) é calculada e representada por um único mapa. Nesse mapa representativo cada mapa de característica é calculado por pesos escolhidos pelo usuário. Uma rede *Winner-Take-All* é utilizada para encontrar a região do ponto mais saliente encaminha-lá para a representação central. De forma sequencial, a rede WTA indicará o segundo ponto de saliência, repetindo o processo até se esgotar as regiões salientes da cena.

A percepção de cores pelos humanos é uma combinação de três estímulos de cores primárias: vermelho, verde e azul. Esse espaço de cores é o chamado espaço RGB (*Red Green Blue*), porém ele não é considerado um bom espaço para a realização do processo de segmentação da imagem, pois existe uma alta correlação entre cada um dos componentes, dificultando a representação de cores nessa forma de representação (QUILES, 2004). Isso é facilmente observado quando ocorre alguma variação de intensidade da luz, pois os três componentes têm seus valores alterados. Porém, transformações lineares e não-lineares dos elementos RGB permitem a obtenção dos atributos de crominância (*hue*), brilho e saturação, que são geralmente utilizados para realizar a distinção de cores. As técnicas utilizadas para atenção visual neste trabalho são explicadas a seguir.

3.1.1 **Modelo de atenção visual bottom-up - (iNVT)**

O iNVT (*iLab Neuromorphic Vision C++ Toolkit*) é uma biblioteca de classes para desenvolvimento de modelos neuromórficos de atenção visual. Os modelos neuromórficos são inspirados nas funções biológicas do cérebro humano. De acordo com Itti e Koch (2001) existem cinco pontos importantes que devem ser considerados sobre modelos computacionais para atenção visual.

- A percepção de qualquer estímulo de entrada é totalmente dependente do contexto ao seu redor;
- Uma estratégia bastante eficiente de controle *bottom-up* dá-se na utilização de um único mapa de saliência, que codifica topograficamente os estímulos;
- Para o desenvolvimento da atenção, o processo de inibição é muito importante para evitar que uma região focada anteriormente seja novamente focada;
- A interação rígida entre atenção e movimentos oculares insere desafios computacionais ao sistema de coordenadas utilizado para controlar a atenção;
- A escolha dos locais de atenção é fortemente determinado pela compreensão de cena e o reconhecimento de objetos.

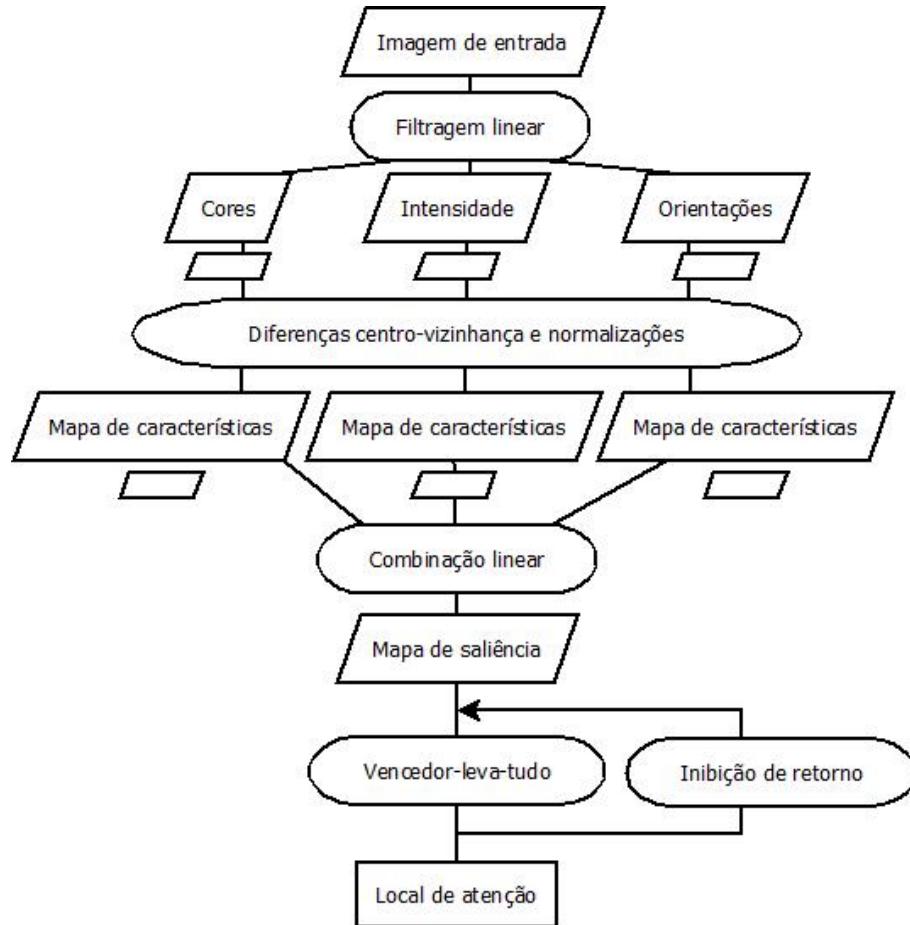


Figura 9 – Estrutura do modelo de saliência iNVT.

Fonte: Adaptada de [Itti e Koch \(2001\)](#).

O mapa de saliência é amplamente utilizado por modelos *bottom-up* e é formado pela composição de vários mapas com características visuais primitivas da imagem como, por exemplo, cor, intensidade e orientação. Qualquer que seja a dimensão característica, esta composição produz uma saliência independente. A cena visual com suas diversas regiões que disputam essa medida o fazem por muitas escalas especiais. As mais salientes são oriundas da região vencedora. A Figura 9 ilustra o modelo proposto por [Itti e Koch \(2001\)](#) para compreensão de como o mapa de saliência é gerado.

3.1.2 Extração de Características Visuais Primitivas

A decomposição de uma imagem em um conjunto de canais distintos é realizada através do modelo proposto por [Koch e Ullman \(1987\)](#). Nesse modelo é estabelecida a extração de características visuais primitivas ou características de baixo nível. Essas características são extraídas da imagem original em várias escalas espaciais, através de filtragens lineares. Qualquer modelo de atenção *bottom-up* tem como primeiro estágio de processamento a computação de características visuais primitivas. Há inspiração biológica considerando que a análise das

características da forma pré-atenção são paralelas a todo o campo visual (ITII; KOCH, 2001).

De um modo geral, a alimentação inicial do modelo por uma imagem estática dá-se por extração das seguintes características visuais: cor, intensidade e orientação. r , g e b são os canais vermelho, verde e azul da imagem de entrada. A imagem de intensidades é obtida por $I = (r + g + b)/3$. I será aplicado para gerar a pirâmide Gaussiana $I\sigma$, discutida a seguir. Para fins de normalização, os canais r , g e b são normalizados a partir de I . O objetivo desta normalização é inibir regiões que apresentem baixos valores de luminosidade (não salientes). Neste caso, r , g e b são normalizados somente quando I for maior do que 1/10 de seu valor máximo sobre toda imagem. Em seguida, quatro canais de cores são criados: $R = r - (g + b)/2$ para o vermelho, $G = g - (r + b)/2$ para o verde, $B = b - (r + g)/2$ para o azul e $Y = (r + g)/2 - |r - g|/2 - b$ para o amarelo. O valor zero é atribuído para sinais negativos (ITTI; KOCH; NIEBUR, 1998a). A obtenção de imagens sem ruídos e detalhes indesejados com realce para as características importantes dá-se pela geração de uma pirâmide Gaussiana que é composta de nove níveis para cada canal, sendo: $I\sigma$, $R\sigma$, $G\sigma$, $B\sigma$ e $Y\sigma$, onde $\sigma \in [0..8]$. A pirâmide Gaussiana é gerada de acordo com o algoritmo proposto em Burt e Adelson (1983), descrito na seção seguinte.

Pirâmide Gaussiana

Operações progressivas de filtragem passa-baixa e subamostragem são estabelecidas pela pirâmide Gaussiana de Burt e Adelson (1983). Um filtro Gaussiano com dimensão de 5×5 pixels foi usado no modelo de saliência de Itti, Koch e Niebur (1998a).

A representação em forma de pirâmide é utilizada com o objetivo de destacar características salientes e inibir demais regiões da imagem. Para a sua geração, um filtro Gaussiano é aplicado a cada nível da pirâmide previamente à geração do nível seguinte. Considerando uma imagem de entrada representada inicialmente por uma matriz G_0 , o nível zero da pirâmide, composta por linhas e colunas (x, y) , onde cada coordenada (pixel) representa um valor correspondente da imagem relacionada a cada característica. O nível 1 contém a imagem G_1 , que é a redução ou versão convolvida de G_0 . De forma similar aos canais considerados, uma pirâmide Gaussiana G_σ pode ser definida recursivamente como segue:

$$G_\sigma(x, y) = \sum_{m=-2}^2 \sum_{n=-2}^2 w(m+2, n+2) G(x, y), \quad \text{para } \sigma = 0 \quad (3.1)$$

$$G_\sigma(x, y) = \sum_{m=-2}^2 \sum_{n=-2}^2 w(m+2, n+2) G_{\sigma-1}(2x+m, 2y+n), \quad \text{para } 0 < \sigma \leq 8, \quad (3.2)$$

onde $w(m, n)$ são os pesos gerados a partir de uma função Gaussiana, empregados para gerar os níveis da pirâmide para todos os canais.

Pirâmide direcional

As informações sobre as orientações locais do modelo de Itti, Koch e Niebur (1998a) são importantes no desenvolvimento da atenção visual.

De acordo com Greenspan *et al.* (1994), a extração destas características pode ser obtida através da aplicação de filtros direcionais sobre a imagem. O perfil de sensibilidade do campo receptivo dos neurônios é aproximado ao de orientação seletiva presente no córtex visual primário.

Os mapas de orientações $O_\sigma(\theta)$ são criados através da convolução do mapa de intensidades I_σ , com filtros direcionais de Gabor para quatro orientações $\theta \in 0^\circ, 45^\circ, 90^\circ, 135^\circ$. Os filtros são usados para identificar as barras ou bordas em uma determinada direção.

Diferenças centro-vizinhança

A diferença centro-vizinhança é implementada como a diferença entre escalas, ou seja, o centro é um pixel da imagem na escala $c \in 2, 3, 4$ e a vizinhança é o pixel correspondente em outra escala $s = c + \delta$, com $\delta \in 3, 4$ da representação piramidal.

O contraste de intensidades é usado para construir o primeiro conjunto de mapas, definido como segue:

$$I(c, s) = |I(c) \ominus I(s)|. \quad (3.3)$$

O segundo conjunto de mapas é construído a partir dos canais de cores, definidos como:

$$RG(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \quad (3.4)$$

$$BY(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|, \quad (3.5)$$

O terceiro conjunto de mapas é gerado a partir de informações de orientação local, de acordo com as seguintes equações:

$$O(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|, \quad (3.6)$$

onde $\theta \in 0^\circ, 45^\circ, 90^\circ, 135^\circ$.

3.1.3 Mapa de Saliência

Esta pesquisa baseia-se no fato das células do córtex cerebral dos mamíferos organizarem-se de forma estruturada, resultando em macro regiões do cérebro capacitadas em processamentos específicos de tarefas como linguagem, controle e visão (KOHONEN, 2001). Computacionalmente, características primitivas da cena formarão mapas diferentes de saliência, por isso o mapa

precisa ser composto em um único, onde as regiões disputam a medida de saliência, deixando uma única vencedora, gerando um mapa independente de características.

De modo geral, o modelo é alimentado por um imagem e são extraídas as suas características de cor e orientação. As informações do canal Y , demonstradas na equação (3.10), são usadas para a normalização dos canais r , g e b , inibindo regiões cuja luminosidade (saliência) seja muito pequena, normalizando a região na qual Y for maior do que 1/10 do seu valor máximo sobre toda a imagem. Em seguida, quatro canais de cores são criados:

$$R = \frac{r - (g + b)}{2} \quad (3.7)$$

$$G = \frac{g - (r + b)}{2} \quad (3.8)$$

$$B = \frac{b - (r + g)}{2} \quad (3.9)$$

$$Y = (r + g)/2 - |r - g|/2 - b \quad (3.10)$$

Valores negativos de R , G , B e Y são atribuídos zero. Porém, a extração apenas desses canais de cores não retira todos os ruídos e informações indesejáveis de uma imagem. Para resolver esse problema foi utilizada a técnica da Piramide Gaussiana de [Burt e Adelson \(1983\)](#).

Nove níveis são considerados na criação dessa piramide, cujos os níveis mais altos representam convoluções dos níveis anteriores, além de um processamento de filtro passa-baixo e de sub-amostragem. No trabalho de [Itti, Koch e Niebur \(1998b\)](#), foi considerado um filtro Gaussiano com dimensões de 5x5 *pixels*, e essa abordagem destaca as saliências de uma imagem, ao mesmo tempo que inibe as demais regiões, retirando os ruídos do mapa.

A imagem de entrada é inicialmente representada pela matriz G_0 , representando o nível zero da piramide, composta pelas linhas e colunas (x, y) , que representam uma coordenada da imagem. A piramide pode ser recursivamente definida como:

$$G_\sigma(x, y) = \sum_{m=-2}^2 \sum_{n=-2}^2 w(m+2, n+2) G(x, y) \quad (3.11)$$

$$G_\sigma(x, y) = \sum_{m=-2}^2 \sum_{n=-2}^2 w(m+2, n+2) G_{\sigma-1}(2x+m, 2y+n) \quad (3.12)$$

sendo a equação (3.11) para $\sigma = 0$ e a equação (3.12) para $0 < \sigma \leq 8$. Os pesos gerados a partir de uma função Gaussiana são $w(m, n)$.

Os mapas de características são obtidos através das diferenças centro-vizinhança. O centro é um *pixel* da imagem em uma escala $c \in \{2, 3, 4\}$ e a sua vizinhança é um *pixel* correspondente em uma imagem em outra escala $s = c + \delta$ com $\delta \in \{3, 4\}$. O processo de geração desses mapas são inspirados biologicamente em neurônios do córtex visual dos mamíferos, e as equações que definem matematicamente as diferenças centro-vizinhança são:

$$\mathcal{I}(c, v) = |I(c) \ominus I(v)| \quad (3.13)$$

$$\mathcal{RG}(c, v) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \quad (3.14)$$

$$\mathcal{BY}(c, v) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))| \quad (3.15)$$

$$\mathcal{O}(c, v, \theta) = |(O(c, \theta) \ominus O(s, \theta))| \quad (3.16)$$

com $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$.

A maioria dos modelos de atenção seguem a hipótese de Koch e Ullman (1987), onde os mapas de cada característica alimentam o mapa de saliência. A região mais proeminente do mapa representa o ponto de maior saliência, independente se isso corresponde a um grande contraste de cor, intensidade ou orientação, pois nenhuma característica previamente conhecida é utilizada para guiar essa atenção, sendo portanto um modelo *bottom-up*.

Para formação desse mapa, os mapas de características são somados (\oplus), após terem sido normalizados ($\mathcal{N}(\cdot)$), gerando os mapas de intensidade ($\overline{\mathcal{I}}$) e cor ($\overline{\mathcal{C}}$). Essa normalização garante o descarte de alguns ruídos, ampliando o contraste das regiões salientes e inibindo regiões não contrastantes (como pode ser visto na Figura 10).

Os mapas de conspicuidades são formados como se segue:

$$\overline{\mathcal{I}} = \bigoplus_{c=2}^4 \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c, s)) \quad (3.17)$$

$$\overline{\mathcal{C}} = \bigoplus_{c=2}^4 \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c, s)) + \mathcal{N}(\mathcal{BY}(c, s))] \quad (3.18)$$

$$\overline{\mathcal{O}} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \bigoplus_{c=2}^4 \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c, s, \theta)) \quad (3.19)$$

$$S = \frac{1}{3} (\mathcal{N}(\overline{\mathcal{I}}) + \mathcal{N}(\overline{\mathcal{C}}) + \mathcal{N}(\overline{\mathcal{O}})) \quad (3.20)$$

sendo S o mapa de saliência resultante de todo o cálculo.

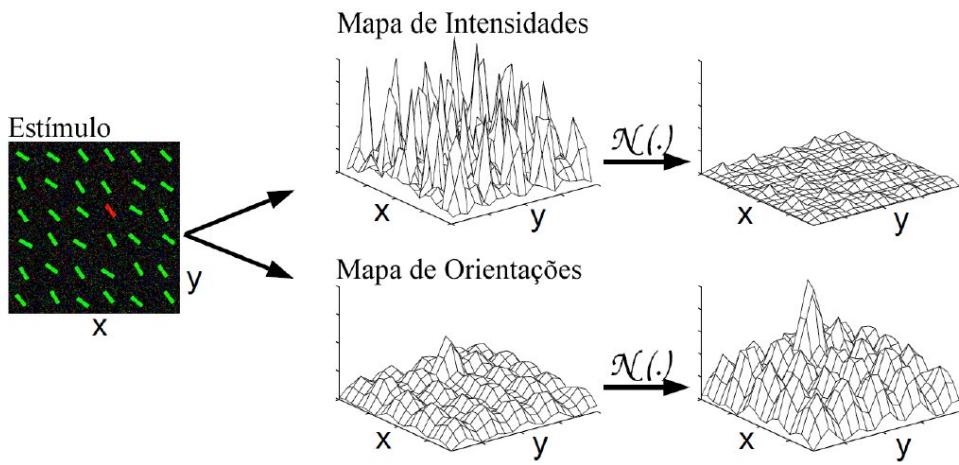


Figura 10 – Normalização dos mapas de conspicuidades.

Fonte: Benicasa *et al.* (2013).

3.1.4 Seleção da atenção e inibição de retorno

Para o desenvolvimento da seleção visual entre as regiões mais salientes do mapa de saliência, Itti, Koch e Niebur (1998a) empregam uma rede neural composta por neurônios do tipo Integra e Dispara, para representar o mapa de saliência S e fazer com que o estímulo de cada neurônio seja o valor de saliência dos pontos no mapa de saliência. A rede de neurônios Integra e Dispara alimenta uma rede neural do tipo *Winner-Takes-All* (WTA) (KOCH; ULLMAN, 1987) (TSOTSOS *et al.*, 1995). A rede é usada para localizar a região mais saliente no mapa de saliências, indicada pelo neurônio vencedor (ITTI; KOCH, 2000). Os neurônios da rede Integra e Dispara são empregados, neste caso, como integradores dos valores de S . Na rede, todos os neurônios recebem ativação de forma independente, até que o neurônio vencedor alcance o limite e dispare. Há, então, desencadeamento simultâneo de três mecanismos: primeiro, o foco da atenção é direcionado para a localização do neurônio vencedor; segundo, o inibidor global é acionado e todos os demais neurônios são inibidos; e por último, a região sob o foco da atenção é temporariamente inibida na rede de neurônios, permitindo que a próxima região saliente seja destacada. Assim, o foco da atenção não é mais redirecionado para a região anterior, que é caracterizada por um mecanismo de inibição de retorno (ITTI; KOCH; NIEBUR, 1998a).

3.2 Sistema de atenção computacional (VOCUS2)

O VOCUS2 (FRINTROP; WERNER; GARCIA, 2015) é um sistema de saliências no qual calcula-se um mapa de saliência a partir de uma entrada de imagem ou vídeo. Sua estrutura é baseada na tradicional abordagem proposta por Itti, Koch e Niebur (1998a) no qual são calculadas características paralelamente e o contraste no centro-vizinhança é calculado por diferenças de Gaussianas. Este algoritmo tem desempenho que está no estado da arte das tarefas

de segmentação de objetos salientes. O sistema é rápido, tem uma estrutura simples, coerente e produz mapas de saliência bem detalhados.

O VOCUS2 (FRINTROP; WERNER; GARCIA, 2015) é um sistema de obtenção de saliências cujos fundamentos são os mesmos encontrados no modelo de Itti-Koch (ITTI; KOCH; NIEBUR, 1998a): canais de características são computados em paralelo; as pirâmides computadas permitem processamento multi-escala e contrastes são obtidos através da Diferença de Gaussianas.

A estrutura mais importante é a de escala de espaço (que usa uma pirâmide gêmea), e sua taxa de centro-vizinhança, que se mostrou o parâmetro-pivô de sistemas de saliências. Como na abordagem de Itti, Koch e Niebur (1998a), o sistema resultante tem uma estrutura baseada em conceitos da percepção humana. Mas ao invés de produzir mapas de saliência baseados em segmentos, este sistema gera mapas baseado em *pixels*. Entretanto, para algumas tarefas, mapas de saliência baseados em segmentos são melhores para a determinação dos limites precisos do objeto. Para obter tais limites, o VOCUS2 integra um *framework* de geração de propostas de objeto.

O sistema VOCUS2 funciona como mostrado na Figura 11. A imagem de entrada é convertida em um espaço de cores-opONENTES com canais para a intensidade, verde-vermelho e azul-amarelo. Para cada canal, são computadas duas pirâmides de imagem (uma para o centro e uma para a vizinhança) nas quais o contraste entre o centro-vizinhança é calculado. A Tabela 4 indica as diferenças entre o sistema VOCUS2 e o iNVT.

Tabela 4 – Principais características entre os sistemas iNVT e VOCUS2.

Fonte: Adaptada de Frintrop, Werner e Garcia (2015).

	iNVT	VOCUS2
Características	intensidade (I), cor (C), orientação (O)	intensidade (I), cor (C)
Estrutura piramidal	uma pirâmide uma escala por camada	pirâmides gêmeas (diferença principal) múltiplas escalas por camada
Fusão de características	sub-amostragem ponderação pela unicidade prioridade para canal de cor e depois intensidade	interpolação média aritmética igualdade entre 3 canais

A maior parte do processamento ocorre pela extração de características de intensidade e de cor. O canal da característica orientação não é utilizado, pois atribui altos valores de saliência aos cantos do objeto e torna o método menos eficiente para segmentação. Para calcular as características cores, é utilizado o espaço de cores RGB. Os canais de característica intensidade (*I*) e características cores-opONENTES vermelho-verde (*RG*) e azul-amarelo (*BY*) são obtidos pelas Equações 3.21.

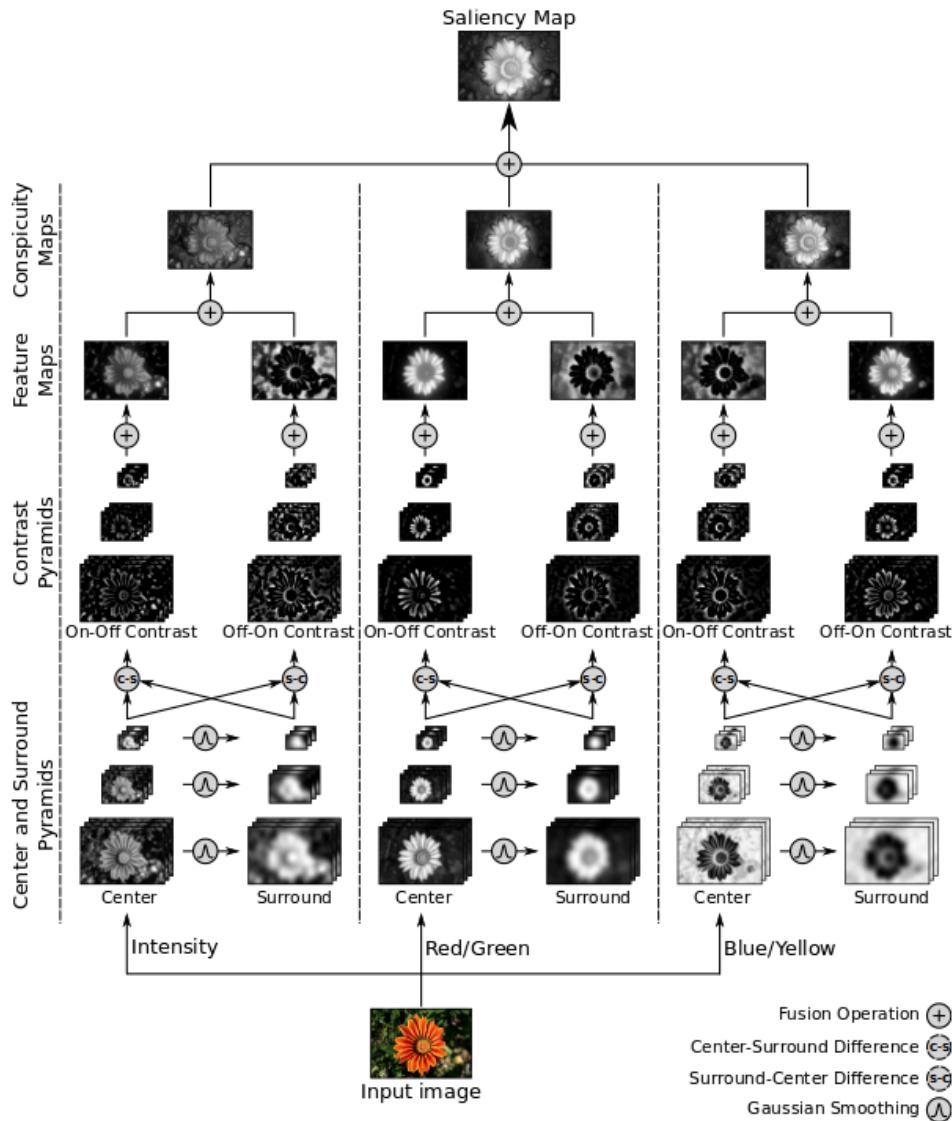


Figura 11 – Processamento dos canais de características do sistema VOCUS2.

Fonte: [Frintrop, Werner e Garcia \(2015\)](#).

$$\begin{aligned}
 I &= \left(\frac{R + G + B}{3} \right) \\
 RG &= R - G \\
 BY &= B - \frac{R + G}{2}
 \end{aligned} \tag{3.21}$$

Espaços de escala com pirâmides gêmeas

Duas diferenças entre o VOCUS2 e o iNVT no espaço de escala são que, ao invés de utilizar uma pirâmide Gaussiana simples, usa-se um espaço com escalas e oitavas, e usa-se pirâmides gêmeas ao invés da subtração de camadas da pirâmide. Essas pirâmides gêmeas possuem uma pirâmide central $C = (C_0, \dots, C_k)$ e uma pirâmide de vizinhança $S = (S_0, \dots, S_k)$. Cada imagem central C_i tem uma imagem de vizinhança correspondente S_i , que é a imagem C_i

suavizada por um fator sigma σ_x que corresponde à taxa de centro-vizinhança desejada. O fator sigma pode ser obtido através da Equação 3.22. O valor de σ_c é o utilizado para obter a imagem central C_i e σ_s é o fator de suavização efetivo para a imagem de vizinhança S_i .

$$\sigma_x = \sqrt{\sigma_s^2 - \sigma_c^2} \quad (3.22)$$

A vantagem desta abordagem é não estar limitada pelas taxas centro-vizinhança dadas pela pirâmide, podendo variar o valor de forma flexível (FRINTROP; WERNER; GARCIA, 2015).

Apesar de uma granularidade mais fina poder ser obtida utilizando um espaço de escala com várias escalas por camada, os mapas de escala usados para subtração ainda tem que ser escolhidos a partir do conjunto disponível da pirâmide e as taxas de centro-vizinhança não podem ser escolhidas arbitrariamente.

Contraste centro-vizinhança

Os contrastes de cor e de intensidade podem ser computados subtraindo o mapa do centro e o mapa de vizinhança. Para distinguir objetos com alto brilho em um ambiente escuro de objetos escuros em ambientes com alto brilho, o cálculo é separado em contrastes *off-on* e *on-off*, correspondendo às células da visão humana que respondem a somente um destes contrastes.

Isto oferece dois mapas de contraste para cada camada i das pirâmides: $X_i^f = C_i^f - S_i^f$ (para contrastes *on-off*) e $Y_i^f = S_i^f - C_i^f$ (para contrastes *off-on*), com $f \in I, RG, BY$. Em ambos os contrastes, valores abaixo de zero são truncados em zero. As pirâmides resultantes são chamadas de pirâmides de contraste.

Fusão dos canais de características

Para realizar a fusão dos canais de características, o algoritmo soma as pirâmides entre escalas para obter os mapas de características (F_1^f e F_2^f), como visto na Equação 3.23.

$$\begin{aligned} F_1^f &= \bigoplus_i X_i, \text{ with } i \in \{1, \dots, k\} \\ F_2^f &= \bigoplus_i Y_i, \text{ with } i \in \{1, \dots, k\} \end{aligned} \quad (3.23)$$

Diferentemente do iNVT, esta soma entre escalas (\oplus) interpola à escala mais fina, não a mais granulosa, antes de somar os mapas. Os dois mapas de características de cada canal são fundidos em mapas de conspicuidade (Equação 3.24) e os mesmos são combinados em um único mapa de saliência S (Equação 3.25).

$$C^f = f(F_1^f, F_2^f), \text{ com } f \in \{I, RG, BY\}, \quad (3.24)$$

$$S = g(C^I, C^{RG}, C^{BY}), \quad (3.25)$$

3.3 Modelo de Atenção Visual *bottom-up e top-down*

Muitos trabalhos baseiam-se na busca por características a partir da cena. [Benicasa \(2013\)](#) considera o sistema visual dos primatas que seleciona as características baseado em objetos. Esse mecanismo, chamado *top-down*, modula o sistema através do enviesamento da atenção de acordo não apenas com características primitivas (cor, orientação, intensidade, etc.), mas também com informações de memória, objetos pré-conhecidos e importantes para a realização da tarefa. Para correlacionar características dos objetos, neurônios específicos (chamados osciladores) representam um objeto, caso estejam sincronizados. Assim, cada conjunto de osciladores representa um dos objetos da cena, ficando em diferentes sincronias, tornando possível agrupá-los por similaridade, proximidade, etc.

A partir dessa correlação, foi proposta a utilização do modelo de rede neural *Locally Excitatory Globally Inhibitory Oscillator Network* (LEGION), o que possibilita a segmentação dos diferentes objetos da cena seguindo o modelo *top-down*, guiando a atenção, além do enviesamento *bottom-up* de busca por características primitivas, pelos objetos de interesse (figuras geométricas planas, no caso de ([BENICASA, 2013](#))). Todos os objetos competem pela atenção, fazendo uso de redes Integra & Dispara (I&D) e da inibição de retorno, para que o mesmo objeto não seja classificado mais de uma vez. Após a segmentação, o objeto é classificado por uma rede Perceptron multicamadas (MLP) e, quando um objeto é reconhecido, um peso maior é aplicado aos neurônios que o representam, aumentando a possibilidade de ser o neurônio vencedor ao passar pelo *self-organized map* (mapa auto-organizável) (SOM), que define o foco de atenção levando em consideração as características primitivas e os objetos reconhecidos. Esse modelo é particularmente interessante por conseguir representar objetos em cenas reais e foi utilizado para reconhecer figuras geométricas em sala de aula e também placas de sinalização de trânsito. A Figura 12 apresenta uma visão geral desse sistema.

3.4 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são soluções computacionais inspiradas no funcionamento do cérebro humano. A principal semelhança entre as RNAs e o cérebro está no extenso processamento paralelo apresentado e da independência de qualquer neurônio isolado, de forma que mesmo se um pequeno número de neurônios de uma rede forem danificados, ela continua a funcionar com pouca perda de desempenho. O neurônio, que representa a base do sistema nervoso ([KANDEL; SCHWARTZ; JESSELL, 1997](#)) é mostrado na Figura 13. O neurônio possui como função principal receber, processar e transmitir informações ([MACHADO, 2000](#)).

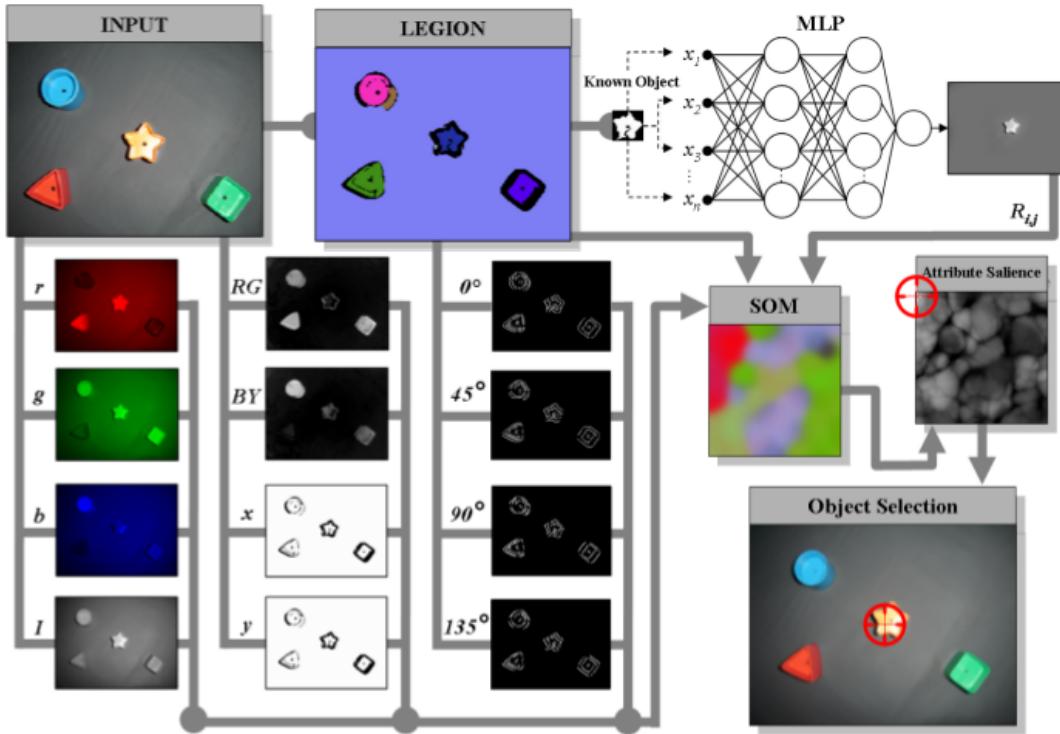


Figura 12 – Sistema de visão proposto em Benicasa (2013).

Fonte: Benicasa (2013).

A princípio, Warren McCulloch e Walter Pitts propuseram o primeiro modelo matemático de um neurônio. Esse Neurônio ficou restrito a representação computacional do neurônio biológico e não considerava o que ocorria no cérebro. Esse neurônio, mostrado na Figura 14, é composto por diversas entradas (os dentritos) que representam os estímulos recebidos e são ponderados pelos pesos (sinapses) e a saída (axônio) é obtida após um somatório de todas essas entradas. Na Equação 3.26, θ representa o limiar de ativação do neurônio. Matematicamente, a saída do neurônio é dada por:

$$\sum_{i=1}^n (x_i w_i) \geq \theta \quad (3.26)$$

$$v = (\sum_{i=1}^n x_i w_i) + bias \quad (3.27)$$

$$y = \varphi(v) \quad (3.28)$$

onde n representa o número de entradas (dendritos), x as entradas, w os pesos associados as sinapses (quando é w positivo a sinapse é excitatória, quando negativo é inibitória) e θ representa

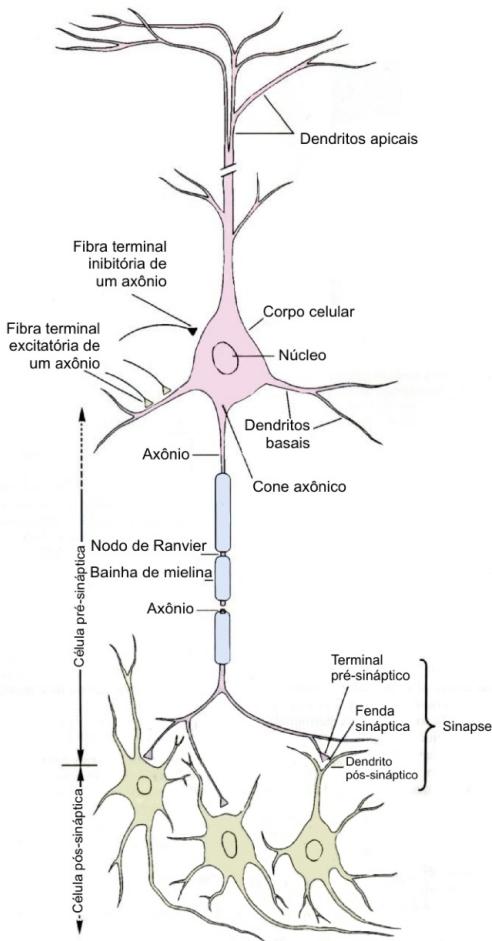


Figura 13 – Morfologia do Neurônio Biológico.

Fonte: Kandel, Schwartz e Jessell (1997).

o limiar de ativação. Matematicamente, a Equação 3.26 é o produto escalar (também conhecido como produto interno $x \cdot w$) do vetor de pesos w com o vetor de entradas x .

Devido à limitação dessa modelagem de neurônio, logo surgiram outras mais eficientes como o *percetron* por exemplo. Este modelo não-linear (Figura 14) é atualmente o mais utilizado pela comunidade de redes neurais. O resultado da extração de borda em sua modelagem x_i representa a entrada presente na sinapse i do neurônio, w_i o peso associado aquela sinapse. v é o somatório das entradas ponderadas pelos pesos acrescido do termo *bias* (Equação 3.27) que é o limiar de ativação. $\varphi(\cdot)$ é a função de ativação do neurônio e y a sua saída (Equação 3.28).

Existem diversas funções de ativação possíveis, dentre as quais as mais utilizadas são a função sigmóide logística (3.29) e a tangente hiperbólica (3.30). A principal vantagem da utilização destas funções, segundo Halkin (2001) está na sua garantia da derivação, o que permite a construção de algoritmos que dependem dessa derivada para o cálculo do gradiente, como é o caso do algoritmo de retropropagação do erro (subseção 3.4.1).

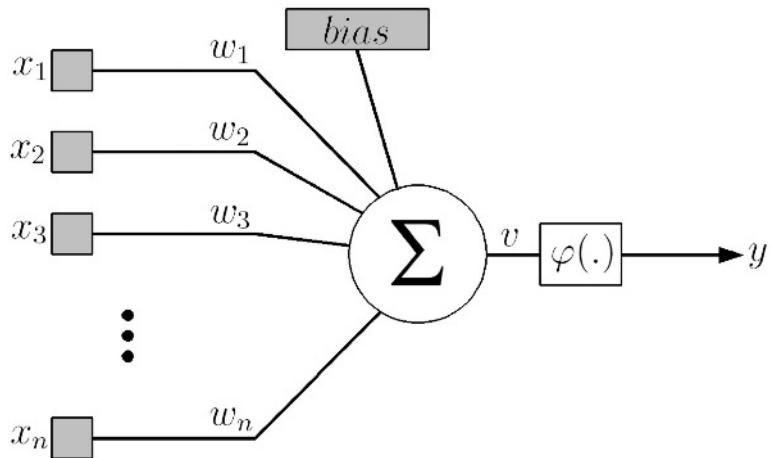


Figura 14 – O Neurônio não-linear.

Fonte: Halkin (2001).

$$\varphi(v) = (1 + \exp(-av))^{-1} \quad (3.29)$$

$$\varphi(v) = \tanh(v) \quad (3.30)$$

3.4.1 Redes Mult Layer Perceptron e Algoritmo BackPropagation

Apesar do *perceptron* apresentar uma melhor modelagem, ele ainda oferecia apenas uma única saída, sendo impossível resolver problemas mais complexos, ou seja, se uma resposta baseada em mais de uma variável fosse necessária o *perceptron* não era suficiente. As *Mult Layer Perceptron* (MLP)¹ (Figura 15) vieram como uma resposta para esse problema, sendo uma generalização do modelo de uma única camada. Sua arquitetura está apresentada na Figura 15, e mostra a organização por camadas da rede. A primeira camada, chamada camada de entrada, simplesmente recebe os sinais e transmite para a próxima camada, não sendo constituída de neurônios. O sinal é propagado através da rede (pelas camadas $l=0,1,\dots,L$), sofrendo a ponderação do peso sináptico de cada camada (w_{ij}). A soma de todas as entradas (y_j) ponderadas é chamada de campo local induzido (v) e é representada pela equação (3.31).

$$v_j^l = \left(\sum_{i=1}^m w_{ji}^l y_i^{l-1} \right) + bias_j^l \quad (3.31)$$

O sinal de saída do neurônio (N_i^l) é computado pela função de ativação $\varphi(\cdot)$ (Equação (3.32)).

¹ Redes neurais de múltiplas camadas

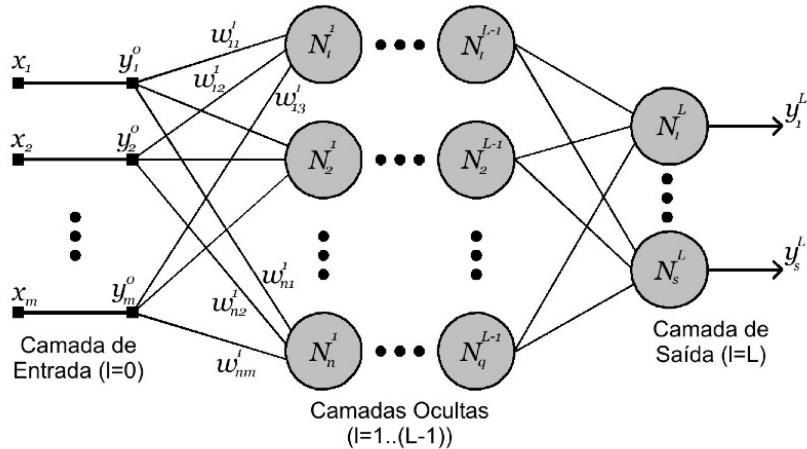


Figura 15 – Arquitetura da Rede Neural MLP.

Fonte: [Halkin \(2001\)](#).

$$y_j^l = \varphi(v_j^l) \quad (3.32)$$

O problema principal com as MLPs era o ajuste de pesos nas camadas ocultas, o que dificultava a convergência da rede. Para resolver esse problema, os pesquisadores Rumelhart, Hintno e Willians propuseram, em 1986, o algoritmo de retropropagação do erro (*backpropagation*).

Algoritmo de *BackPropagation*

O algoritmo de *backpropagation* é um dos primeiros algoritmos desenvolvidos para treinamento de redes neurais do tipo Multi-Camadas, tanto por seu valor histórico quanto pelo seu poder computacional. Esse algoritmo baseia-se na regra de aprendizado proposta por Widrow (chamada regra delta) e utiliza o gradiente descendente da superfície do erro, com a função de minimizar o erro quadrático.

A ideia principal é que o erro seja retropropagado, ou seja, o erro da camada atual é uma estimativa da somatória dos erros das camadas anteriores somada com o erro do bias da própria camada. Por esse motivo, o numero excessivo de camadas ocultas deve ser evitado, pois a precisão do erro é reduzida (**BRAGA; LUERMIR; CARVALHO, 2000**). O algoritmo é composto por duas fases: A fase *feedforward*, onde o sinal de entrada é propagado para frente, e a fase *backward*, onde o erro é propagado para trás.

Para o cálculo do erro, é necessária uma base de treinamento onde os valores esperados sejam conhecidos, assim o erro é obtido fazendo-se a comparação da resposta da rede com a resposta esperada (ou saída desejada), no tipo de aprendizado chamado Aprendizado Supervisionado. Esse tipo de aprendizado é recomendado quando o trabalho é facilmente rotulado, como no caso da definição de cores.

A Equação 3.33 apresenta a entrada da rede. A saída dos neurônios da camada $l = 1$ é gerada pelas Equações 3.31 e 3.32 respectivamente. Camada a camada, esse procedimento é executado até que o sinal atinja a camada de saída $l = L$ gerando o vetor $Y^L(p)$. Esse vetor é comparado ao vetor contendo os valores de saída desejados $D(p)$ para o padrão p apresentado (Equação 3.34) obtendo assim, o erro e_j para cada neurônio j da camada de saída L .

O erro quadrático total apresentado pela rede para o padrão de entrada p é definido por $E(p)$ (Equação 3.35) e o erro médio quadrático para todos os padrões é definido pela Equação (3.36). Os pesos w da rede são ajustados pela Equação 3.35. Para tal, camada a camada, o erro é retropropagado calculando-se os gradientes δ para cada neurônio (Equação 3.37). Os gradientes para os neurônios das camadas ocultas são calculados pela Equação 3.38. Após o cálculo do gradiente, o ajuste dos pesos é feito pela Equação (3.39) no qual η representa o coeficiente de aprendizagem. Vale ressaltar que um valor muito grande do η não garante a convergência do sistema.

$$y_j^0 = x_j \quad (3.33)$$

$$e_j(p) = d_j(p) - y_j^L(p) \quad (3.34)$$

$$E(p) = \frac{1}{2} \sum_{j \in L} e_j^2(p) \quad (3.35)$$

$$E_{med} = \frac{1}{P} \sum_{p=1}^P E(p) \quad (3.36)$$

$$\delta_j^l = e_j^L \varphi_j'(v_j^L) \quad (3.37)$$

$$\delta_j^l = \varphi_j'(v_j^l) \sum_k \delta_k^{l+1} w_{kj}^{l+1} \quad (3.38)$$

$$w_{ji}^l(t+1) = w_{ji}^l(t) + \eta \delta_j^l y_i^{l-1} \quad (3.39)$$

3.4.2 Rede LEGION

A *Locally Excitatory Globally Inhibitory Oscillator Network* (LEGION) (Rede Oscilatória Localmente Excitatórias Globalmente Inibitória ([WANG, 1995](#)) vem sendo muito utilizada nos últimos anos, principalmente pela sua alta velocidade de sincronismo e dessincronismo dos grupos de osciladores.

Sua arquitetura básica é composta por três componentes principais: um inibidor global, acoplamentos excitatórios locais e osciladores neurais. Os objetos da cena são representados pelos grupos de osciladores, que são sincronizados pelos acoplamentos excitatórios. Já o inibidor gera a dessincronização dos grupos distintos, que não representam o mesmo objeto na cena, gerando um mecanismo de cooperação local e inibição global. Em sua proposição inicial ([WANG, 1995](#)), o modelo é formado por uma rede de osciladores de relaxamento, sendo compostos pela variável excitatória x_i e pela variável inibitória y_i definidas como se segue:

$$\dot{x}_i = 3x_i - x_i^3 + 2 - y_i + \mathcal{I}_i + S_i + \rho \quad (3.40)$$

$$\dot{y}_i = \varepsilon(\alpha(1 + \tanh(x_i/\beta)) - y_i) \quad (3.41)$$

sendo I_i o estímulo externo ao oscilador i , S_i os acoplamentos, α, β e ε parâmetros do modelo (ε é normalmente uma constante positiva de valor pequeno), e ρ um sinal de ruído. Um oscilador de relaxamento típico quando I_i for constante e os termos S_i e ρ são eliminados.

As isóclinas nulas (quando $\dot{x}_i = \dot{y}_i = 0$) representam, respectivamente, uma função cúbica e uma função sigmóide. Quando $\mathcal{I}_i > 0$, o modelo representa um oscilador de tempo limite, oscilando em duas fases de valores de x um mais elevado e a segunda de valores baixos. Essas fases são denominadas ativa e silenciosa e o oscilador estará no modo *disparando*. A transição entre as fases acontece muito rapidamente em uma fase denominada *jumping*. Quando não há estímulo ($\mathcal{I}_i < 0$), o sistema está num ponto de equilíbrio estável, podendo ser induzido a oscilar por meio de estímulos recebidos pelos acoplamentos. Nesse caso, o oscilador é considerado *excitável*. A Figura 16 demonstra esses dois momentos do oscilador.

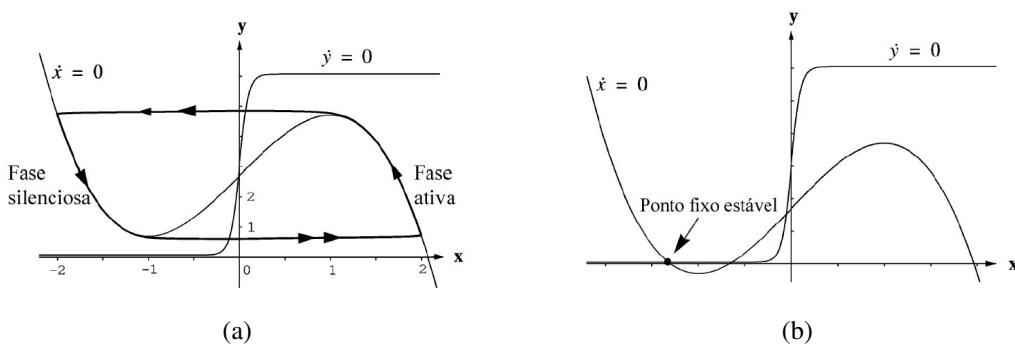


Figura 16 – Modos do Oscilador de relaxamento: (a) Modo ativo e (b) modo excitável.

Fonte: [Wang \(1995\)](#).

A cada pulso da rede LEGION, uma região homogênea é salientada, desconsiderando-se o resto da imagem. Essa região é enviada para a rede *MultLayer Perceptron* (MLP) - Redes de múltiplas camadas de *percéptrons* - que rotula se aquela região é conhecida (quadrado, triângulo,

retângulo, etc) ou não. Uma vez que a MLP conseguiu classificar aquela imagem salientada como uma figura geométrica, ela é contabilizada. Ao final da classificação, este processo é repetido novamente. A Figura 17 apresenta os passos da segmentação da imagem.

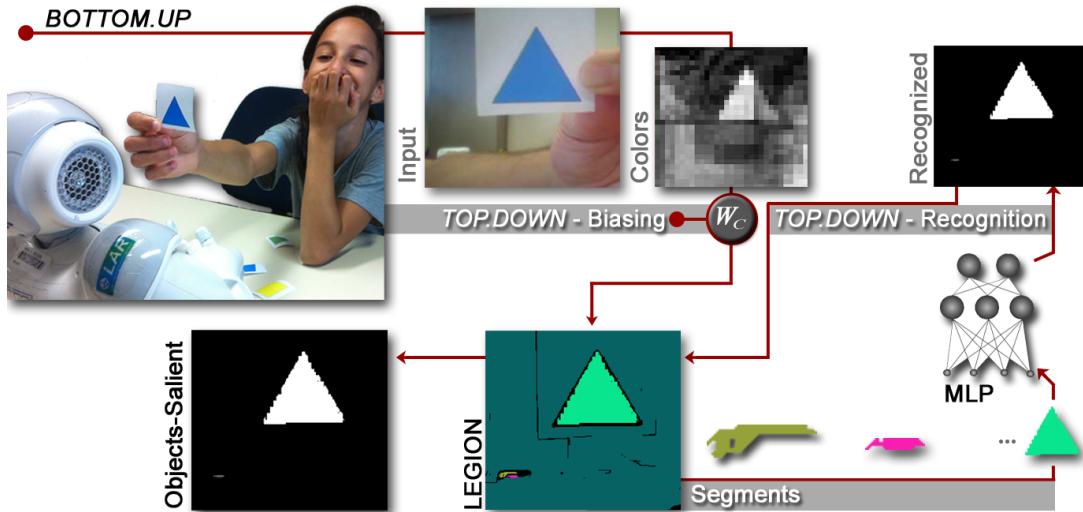


Figura 17 – A arquitetura da segmentação baseada em objetos.

Fonte: Pinto *et al.* (2014a).

3.5 Extração de características de imagens

Uma característica é uma propriedade que pode representar uma imagem ou parte dela. Pode ser um pixel, um círculo, uma linha, uma região com textura média dos níveis de cinza, etc. Apesar de não existir uma definição formal, características podem ser definidas como partes detectáveis da imagem com algum significado.

As características podem ser extraídas por diversos algoritmos, por exemplo, o *Scale-Invariant Feature Transform* (SIFT) (LOWE, 2004), o *Speeded-Up Robust Features* (SURF) (BAY; TUYTELAARS; GOOL, 2006b), o *Features from Accelerated Segment Test* (FAST) (ROSTEN; DRUMMOND, 2006) ou o *Oriented FAST and Rotated BRIEF* (ORB) (RUBLEE *et al.*, 2011).

3.5.1 SIFT

O SIFT é um algoritmo de visão computacional composto por duas partes distintas: o detector, usado para detectar os extremos da imagem e fazer a localização de pontos-chave, e o descritor, empregado na definição da orientação e descrição dos pontos-chave. O detector é baseado em cálculos de diferenças de Gaussianas e o descritor utiliza histogramas de gradientes orientados para descrever a vizinhança local dos pontos de interesse. O SIFT detecta a orientação do gradiente dominante em sua posição e registra o histograma do gradiente local resultante em

relação a essa orientação. Assim, as características do SIFT são relativamente bem comportadas em pequenas transformações.

O algoritmo SIFT para extração de características de imagens transforma uma imagem em uma "grande coleção de vetores de características locais" (LOWE, 2004). Cada um desses vetores de características é invariante à escala, rotação e translação da imagem. Em uma imagem são detectados e construídos vários pontos-chave e seus descritores. Esse conjunto de descritores pode ser usado para fazer correspondências entre imagens como exemplificado na Figura 18.



Figura 18 – Exemplo de correspondência de imagens.

Fonte: Lowe (2004).

Detector

O detector busca pontos invariantes à mudança de escala, possibilitando a detecção de pontos-chave em vários níveis de aproximação do objeto de interesse. Para isso, o detector procura por características estáveis em relação à escala aplicando uma função Gaussiana. Assim, o espaço escalar de uma imagem é dado pela operação de convolução entre a Gaussiana de escala-variável com uma imagem $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (3.42)$$

sendo x e y as coordenadas do centro do quadro, σ a escala aplicada na imagem e G a função Gaussiana definida como:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (3.43)$$

A detecção estável de pontos-chave é feita usando o espaço escalar extremo na função de Diferença de Gaussiana (DoG) que pode ser calculada pela diferença de duas escalas próximas (Figura 19).

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \quad (3.44)$$

Em que k é o fator constante de separação entre o espaço de escalas.

Os extremos são dados pelos máximos ou mínimos locais para cada DoG, que podem ser obtidos comparando a intensidade de cada ponto com a intensidade de seus vizinhos. A Figura 19 representa o modelo de detecção de pontos-chave em uma imagem. Para cada escala, a imagem é convolucionada com gaussianas para produzir um conjunto de imagens escaladas. As gaussianas de imagens adjacentes são subtraídas para produzir a diferença de gaussianas.

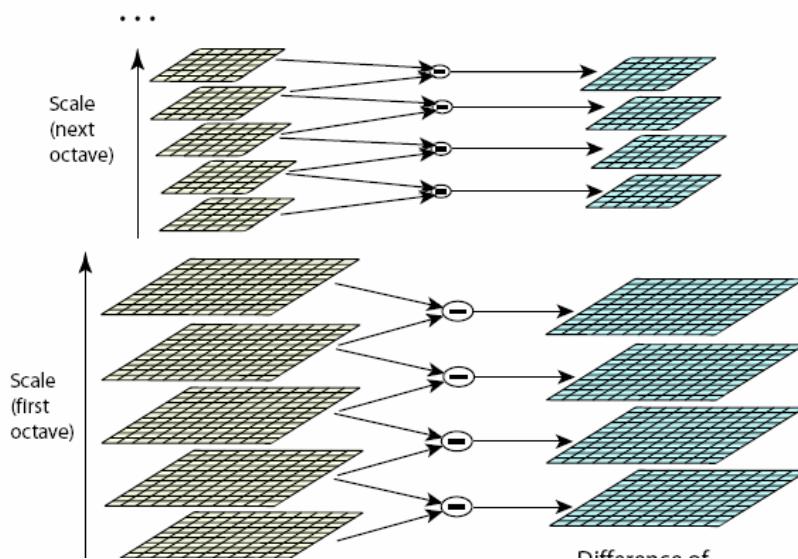


Figura 19 – Representação gráfica da função de diferença de Gaussiana.

Fonte: [Lowe \(2004\)](#).

Descriptor

Uma orientação é atribuída para cada ponto-chave para construir os descritores. Monta-se um histograma das orientações para uma região vizinha ao redor do ponto-chave. Cada ponto na vizinhança do ponto-chave é adicionado ao histograma com um determinado peso. Os picos no histograma de orientações correspondem às direções dominantes dos gradientes locais e são utilizados para definir a orientação do ponto-chave. Assim, cada ponto-chave tem quatro dimensões: sua posição x , y , magnitude e orientação. O descritor do ponto-chave é criado computando-se as magnitudes e orientações dos gradientes amostrados (Figura 20) ao redor da localização do ponto-chave.

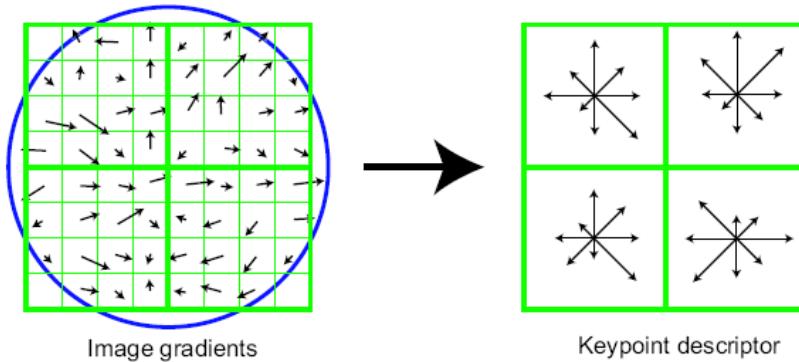


Figura 20 – Construção do descritor do ponto-chave.

Fonte: [Lowe \(2004\)](#).

Os vetores resultantes são chamados de características SIFT e, em conjunto com a técnica *k-nearest neighbours* (KNN) ([FIX; HODGES, 1951](#)), podem ser usados para identificar possíveis objetos em uma imagem. Quando existe um conjunto de características coincidentes em duas imagens, é altamente provável que as características se refiram a um mesmo objeto.

Devido ao grande número de características SIFT em uma imagem de um objeto, a técnica é capaz de reconhecer objetos mesmo que exista um considerável nível de oclusão do objeto na imagem.

3.5.2 SURF

O *Speeded-Up Robust Features* (SURF) ([BAY; TUYTELAARS; GOOL, 2006b](#)) é um detector e um descritor de pontos de interesse invariante às operações de escala e rotação. O detector SURF utiliza como base a matriz Hessiana, mas, no entanto, ao invés de utilizar uma medida para selecionar a localização e a escala, tal como é feito pelo detector Hessian-Laplace ([BAY; TUYTELAARS; GOOL, 2006b](#)), utiliza o determinante do Hessiano para ambos. Já o descritor SURF representa a distribuição das respostas da *wavelet Haar* na vizinhança do ponto de interesse.

Detector

Para a tarefa de detectar um objeto em uma imagem, ao invés de procurar o objeto como um todo, apenas os pontos de interesse do objeto são utilizados para identificação. Este tipo de abordagem é escolhida por várias razões, das quais as principais são: o custo computacional de procurar em dados com grande dimensionalidade, como armazenadas em imagens; e o alto nível de redundância incorporada, porque *pixels* não se movem de forma independente e possuem um

elevado grau de correlação. Existem vários métodos para definir e detectar pontos de interesse, que vão desde os que consideram cantos como pontos de interesse para os que consideram um *blob* de cada vez.

Detector rápido de Hessian

O detector de Hesse rápido detecta características *blob-like*. Ele baseia-se na matriz de Hesse, que a escala é definida como se segue:

$$\mathcal{H}(x, y, \sigma) = \begin{bmatrix} \frac{\delta^2}{\delta x^2} G(\sigma) * I(x, y) & \frac{\delta}{\delta x} \frac{\delta}{\delta y} G(\sigma) * I(x, y) \\ \frac{\delta}{\delta x} \frac{\delta}{\delta y} G(\sigma) * I(x, y) & \frac{\delta^2}{\delta y^2} G(\sigma) * I(x, y) \end{bmatrix} \quad (3.45)$$

Sabe-se que, no caso contínuo, gaussianas são adequadas para a análise do espaço de escala (KOENDERINK, 1984; LINDEBERG, 1990). No entanto, todos os detectores baseados em Hessian possuem um ponto fraco: quando se trabalha com imagens distintas, a gaussiana precisa ser também discretizada, como consequência, há uma perda de repetibilidade sob as rotações da imagem ao redor dos múltiplos ímpares de $\pi/4$. No entanto, a matriz de Hessian é escolhida porque a taxa de repetição é ainda muito alta em qualquer ângulo de rotação (BAY *et al.*, 2008). Como os filtros de gaussiana discretizados já não são ideais e as circunvoluções ainda são custosas, eles são aproximados por filtros de caixa (Figura 21). Desta forma, quando usado em conjunto com imagens integrais, os cálculos podem ser realizados em tempo constante. Embora esta seja uma aproximação ainda mais forte, o desempenho é comparável ou até melhor do que com o discreto e a gaussiana cortada (BAY *et al.*, 2008).

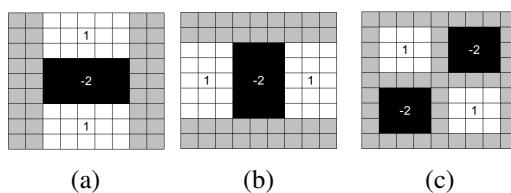


Figura 21 – Filtros de caixa que aproximam derivadas gaussianas de segunda ordem.

Fonte: Lowe (2004).

Para determinar quão "forte" é o ponto atual para classificá-lo como de interesse ou não é necessário calcular a matriz aproximada de Hessian utilizando o filtro de caixa.

3.5.3 Invariância de escala

Os cálculos anteriores são realizados em diferentes escalas, porque os pontos de interesse podem ser comparados entre as imagens, onde elas são vistas em diferentes escalas. O espaço de escala é implementado como uma imagem piramidal. Sem filtros de caixa, geralmente, a imagem piramidal é construída por sucessivos processos de suavização da imagem como uma subamostra da gaussiana e, em seguida, a fim de atingir um nível mais elevado da pirâmide. Com

os filtros de caixa e as imagens integrais, não há a necessidade de filtrar a imagem iterativamente e subamostrá-lo. Em vez disso, é o filtro que é o sobre-dimensionado e aplicada exatamente a mesma velocidade sobre a imagem original. Os níveis mais elevados da pirâmide são atingidos através da aplicação de filtros gradualmente maiores. O espaço escala é dividido em oitavas. Cada oitava é uma série de mapas de resposta do filtro, obtido por convolução da mesma imagem de entrada, com um filtro de tamanho crescente. Devido à natureza discreta dos filtros de caixa, o seu tamanho deve ser aumentado para um mínimo de dois *pixels*, a fim de manter a presença do *pixel* central.

3.5.4 Classificação do ponto de interesse

Para classificar um ponto de interesse de uma supressão não-máxima em um $3 \times 3 \times 3$ é aplicada em escala e imagem do espaço (Figura 22). Cada amostra é comparada com seus 8 vizinhos na imagem atual e os 18 vizinhos de escala superior e inferior no espaço de escala. Se ele tem a maior pontuação (determinante da matriz Hessian) de seus vizinhos, então ele é considerado um ponto de interesse, caso contrário, ele é descartado.

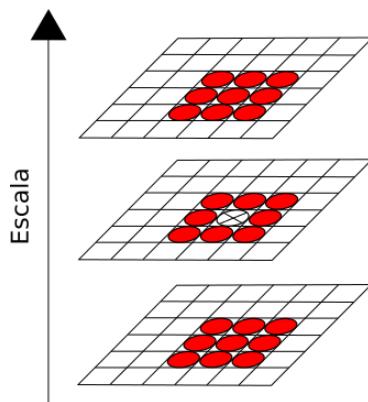


Figura 22 – Representação gráfica da vizinhança $3 \times 3 \times 3$.

Fonte: Adaptada de [Lowe \(2004\)](#).

Uma vez que é classificado como um ponto de interesse, a localização é refinada para a precisão de *subpixel*, ajustando uma parábola para o ponto de amostragem e seus vizinhos imediatos ([BROWN; LOWE, 2002](#)).

Descriptor

Após a detecção dos pontos de interesse, é necessário atribuir uma descrição de cada um, a fim de identificar e distingui-los um do outro e combiná-los entre as imagens. Idealmente, um bom descriptor irá fornecer uma descrição para cada ponto que é único, mas idêntico para todos os pontos de vista possíveis do mesmo ponto.

Este caso ideal é difícil de se alcançar, se não impossível. No entanto, os descriptores locais utilizam as informações sobre a textura dos pontos de interesse vizinhos para distingui-los,

tanto quanto possível um do outro. Especificamente, o descritor SURF, tende a funcionar muito bem nesta tarefa. O descritor SURF é construído em três etapas, que agora são descritas.

3.5.5 Orientação do ponto do atributo de interesse

Para atingir invariância de rotação, o primeiro passo consiste em atribuir uma orientação para o ponto, de modo que, quando é visto a partir de outra perspectiva, pode ser adequada. Para esta finalidade, e para tomar vantagem da utilização de imagens integrais, filtros *wavelet* Haar são utilizados (Figura 23). Sendo s a escala em que foi detectado o ponto de interesse, filtros de $4s$ de tamanho são usados e respostas *wavelet* em direções x e y com um passo de amostragem de s , são calculados em torno de um vizinho circular de raio $6s$.

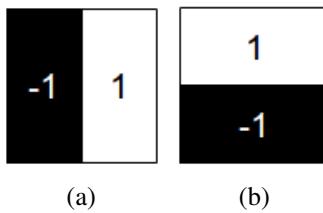


Figura 23 – Filtro wavelet Haar.

Fonte: [Bay, Tuytelaars e Gool \(2006b\)](#).

Vetor descritor

Embora haja algumas variações, o descritor SURF padrão consiste de um vetor com 64 entradas. Para construir este vetor, o primeiro passo é o de construir um quadrado $20s$ tamanho da região de f , centrado no ponto de interesse e com a orientação selecionada na etapa anterior. Esta região é dividida em 4×4 sub-regiões quadradas. Usando o filtro de Haar com tamanho $2S$, as respostas de filtro em 5×5 pontos de amostragem igualmente espaçadas são calculados na direção x e y . Note-se que estas instruções são definidas em relação à orientação da região quadrada. Mas, em vez de rotacionar a imagem em si, as respostas do filtro são calculadas na imagem sem rotação e então interpoladas. Após ponderação das respostas de filtro usando uma gaussiana com $\sigma = 3.3$ centrada no ponto de interesse, quatro somas em cada sub-região são calculadas: as somas de dx e dy e, para ter informações sobre a polaridade das mudanças de intensidade (Figura 24), as somas de $|dx|$ e $|dy|$.

Assim como o SIFT, o descritor SURF também consiste em transformar uma região em torno da característica em um vetor de componentes de frequência ([BOTTERILL; MILLS; GREEN, 2008](#)). De acordo com [Bay, Tuytelaars e Gool \(2006b\)](#), embora o SURF possa ser semelhante em conceito com o SIFT, o SURF é menos sensível ao ruído e supera o SIFT. Isso se dá por conta da integração global de informações obtidas a partir de gradiente da sub-região, em vez de gradientes individuais, como no caso do SIFT. Outra vantagem em relação ao SIFT é seu reduzido tempo de processamento para detecção e descrição dos pontos de interesse em imagens.

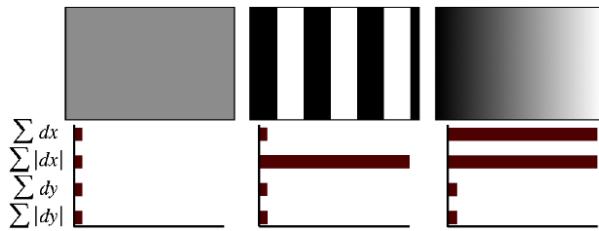


Figura 24 – Entradas do descriptor de uma sub-região.

Fonte: [Bay, Tuytelaars e Gool \(2006b\)](#).

Isso se dá pelo fato do descriptor SURF utilizar apenas 64 dimensões, reduzindo assim o esforço computacional dos operadores diferenciais no espaço de escala.

Vetor descriptor

Um ponto de interesse em uma imagem é considerado igual a outro ponto de interesse em outra imagem se eles estão próximos o suficiente no sentido do vizinho mais próximo. O algoritmo mais utilizado para determinar o vizinho mais próximo é o *kd-tree* ([MOORE, 1991](#)).

3.6 ***Bag-of-Features*** (BoF)

Bag-of-Features (BoF) é uma abordagem popular para classificação visual de objetos cujo interesse é devido à seu poder e simplicidade. Sua origem deriva do modelo *bag-of-words* proposto por [Salton e McGill \(1983\)](#). Esta abordagem é usada para diversas tarefas de visão computacional, tais como classificação de imagens, localização de robôs e reconhecimento de texturas. Os métodos que aplicam este modelo baseiam-se em coleções não ordenadas de descritores de imagens. Eles possuem a característica de descartar informações espaciais e são conceitualmente mais simples que os métodos alternativos. A ideia principal por trás desta abordagem é calcular características em uma imagem e, a partir de combinações com um conjunto de características, classificar a imagem. No modelo BoF, as características extraídas são agrupadas e as partições geradas são usadas para montar um dicionário de palavras visuais. Após quantizar as características usando o dicionário visual, as imagens são representadas pela frequência das palavras visuais.

Em um dicionário visual, cada imagem dentro do BoF representa um grupo encontrado pelo algoritmo de agrupamento. Se dada uma imagem, as características quantizadas forem similares às características encontradas no dicionário, pode se classificar a imagem usando o rótulo da imagem que está no dicionário. Por exemplo, no caso da Figura 25, se uma característica for uma imagem de um olho, pode-se associar a imagem à classe rosto, pois no dicionário visual existe uma característica olho com o rótulo rosto.

Para formar o dicionário visual do modelo, é necessário um algoritmo agrupador para separar as características em conjuntos semelhantes que possam ser comparadas futuramente por



Figura 25 – Bag-of-Features.

Fonte: [Salton e McGill \(1983\)](#).

um classificador. Os algoritmos mais utilizados para o agrupamento de características é a técnica K-médias e o classificador SVM que serão descritos a seguir.

3.6.1 K-médias

K-médias ([LLOYD, 1982](#)) é um método que segue o paradigma de aprendizado não supervisionado para fazer extração de conhecimento sem utilizar informações sobre as classes dos exemplos. Dessa forma, busca organizar um conjunto de objetos em grupos de acordo com alguma medida de similaridade ou dissimilaridade. O k-médias é um algoritmo agrupador que dada uma função de dissimilaridade retorna uma partição do conjunto de objetos. É o algoritmo mais conhecido para agrupamento de dados. Ele busca particionar o conjunto de objetos em k grupos, sendo cada objeto associado ao grupo mais próximo.

Seu funcionamento é simples, inicialmente seleciona-se k pontos aleatórios sobre o espaço, esses pontos são chamados de centroides. Para cada objeto computa-se o centroide mais próximo e o rotula como pertencente ao centroide. Em seguida, recalcula-se a posição dos centroides com base na posição de seus objetos associados. Para recalcular a posição dos centroides, considera-se a distância média de seus objetos relacionados. Assim, a cada passo do algoritmo os centroides são movidos em direção a seus objetos associados e o algoritmo é interrompido quando não houver mais variações nos seus centroides. No término da execução, se tem as coordenadas dos centroides que particionam o espaço.

3.6.2 Detecção de objetos

O primeiro passo do modelo é extrair características de imagens do objeto que se deseja detectar. Cada descritor de característica é composto por um vetor e representa uma palavra que

será usada para montar o dicionário visual. Em seguida, a matriz de descritores é submetida ao método k-médias, com número de grupos variando de 1 até o número de características. O próximo passo do algoritmo é gerar os histogramas das imagens de treinamento e validação. Para cada característica, localiza-se o centroide mais próximo e soma-se a quantidade de ocorrências para formar o histograma. Após formar o histograma da imagem, este é normalizado no intervalo de grupos. Depois de construídos os histogramas de todas as imagens, é construído um histograma médio para cada classe. Com o histograma de cada classe construído, pode-se fazer a classificação de novas imagens comparando o histograma de uma imagem desconhecida com os histogramas de classes conhecidas. O histograma que apresenta a menor dissimilaridade representa a classe da qual a imagem desconhecida será rotulada.

Qualquer algoritmo classificador pode ser usado para classificar as características. Os mais usados são o *Naive Bayes* e o *Support Vector Machines*.

3.7 ***Support Vector Machines (SVM)***

O método *Support Vector Machines* (SVM) consiste em encontrar uma função de predição para generalizar dados e é utilizado em várias pesquisas, como reconhecimento de padrões (CORTES; VAPNIK, 1995), reconhecimento de objetos (BLANZ *et al.*, 1996), identificação de fala, detecção de face, categorização de texto, entre outros. Na maioria dos testes, em comparação com valores de referência e outras abordagens clássicas, o desempenho na generalização do SVM (erro em classificar o conjuntos de teste) é semelhantes a eles ou significativamente melhor. Uma das razões para esses resultados é o treinamento com nenhum conhecimento prévio da distribuição de probabilidade de problema, isto é, presume-se que existe uma distribuição de probabilidade conjunta no espaço de dados e os exemplos de formação são amostrados independentes sobre esta distribuição.

Este pressuposto veio da *Statistical Learning Theory* (SLT)², que tenta garantir a existência de aprendizagem em algoritmos supervisionados (LUXBURG; SCHOLKOPF, 2008). Isso significa que o algoritmo irá encontrar a melhor função de generalização para o problema em questão, inferindo uma regra geral que explica os exemplos de treinamento já vistos e classifica novos exemplos desconhecidos com o menos "custo"(ou o menos risco de erro). Uma vez que não é possível saber o risco real de uma função (porque a distribuição de probabilidade subjacente é desconhecida na etapa de aprendizagem), é impossível calcular diretamente o melhor classificador (chamado de "classificador Bayes") (LUXBURG; SCHOLKOPF, 2008), assim o foco do SLT é minimizar o risco empírico, proporcionando uma garantia sobre a confiabilidade destas aproximações. Uma série de conveniências e métodos matemáticos são empregados para alcançar esta solução, como a lei dos grandes números, convergência uniforme, união de limites, o coeficiente de ruptura e a dimensão VC (em homenagem a Vapnik e Chervonenkis). Todos os

² Teoria da Aprendizagem Estatística

métodos, pressupostos e provas da SLT são discutidos em [Luxburg e Scholkopf \(2008\)](#).

A diferença entre o SVM e outros algoritmos clássicos de aprendizagem de máquina é o viés assumido. Assim como o ID3 ou MLP, o viés para o SVM é baseado em funções lineares. No entanto, são lineares no espaço de características, e podem proporcionar um classificador não-linear. Os dados podem ser não-separáveis num dado espaço de entrada, de modo que podem ser mapeados para outro espaço e outra dimensão com uma medida de similaridade. Esta função de mapeamento é chamado de *Kernel*. É importante para entender o seu espaço de entrada e a distribuição de seus dados para encontrar o melhor *kernel* que irá classificar com menos risco.

Outro tipo importante de medida de capacidade de função de classes é a margem, não considerada normalmente em outros algoritmos. Considerando-se um espaço bidimensional com dados de classes (ou rótulos) linearmente separáveis, estas classes podem ser separadas por uma linha reta (o hiperplano). Margem é a menor distância de qualquer ponto de treinamento para esta linha de separação. Assim, quanto maior for a margem de funções escolhidas, menor é a dimensão VC e melhor é a precisão do algoritmo. Um tratamento abrangente deste resultado pode ser encontrado em [Scholkopf e Smola \(2002\)](#).

O caso mais simples é o de SVM treinadas linearmente com dados bidimensionais separáveis. Isso significa que existe um "hiperplano separador" que separa as classes.

Seja $H_1 : \mathbf{x}_i * \mathbf{w} + b = 1$, \mathbf{w} um hiperplano normal, $|b|/\|\mathbf{w}\|$ a distância perpendicular do hiperplano à sua origem, e $\|\mathbf{w}\|$ a norma Euclideana do \mathbf{w} , os pontos x que pertencem ao hiperplano satisfazem $\mathbf{w} * \mathbf{x} + b = 0$. Considerando um hiperplano para cada classe (H_1 e H_2) pode-se notar que eles são paralelos (possuem a mesma norma) sem quaisquer pontos entre eles. Assim, para maximizar a margem, ou seja, achar a melhor margem, é preciso minimizar $\|\mathbf{w}\|^2$, sujeito à seguinte restrição desigualdade:

$$y_i(\mathbf{x}_i * \mathbf{w} + b) - 1 \geq 0 \quad \forall i \quad (3.46)$$

o que representa a combinação de margem separada para cada classe. pontos de treinamento esses que acabam deitado em um dos. Pontos de treinamento que pertencem à um dos hiperplanos H_1 e H_2 , cuja remoção pode alterar a solução encontrada são chamados de vetor de suporte (*support vector*), mostrado na Figura 26 para esse caso.

Para encontrar a solução ideal, dada uma função com restrições, é necessário formular o problema usando uma Lagrangeana, respeitando as condições Karush-Kuhn-Tucker (KKT). Isto satisfaçõz qualquer problema de restrição de otimização (convexa ou não), garantindo que o cruzamento das orientações confiáveis com orientações de descida coincide com a interseção do conjunto de orientações possíveis para restrições linearizadas das orientações de descida ([FLETCHER, 1987](#)). Em SVM como soluções para problemas convexos, as condições de KKT são necessárias e suficientes ([LUXBURG; SCHOLKOPF, 2008](#)).

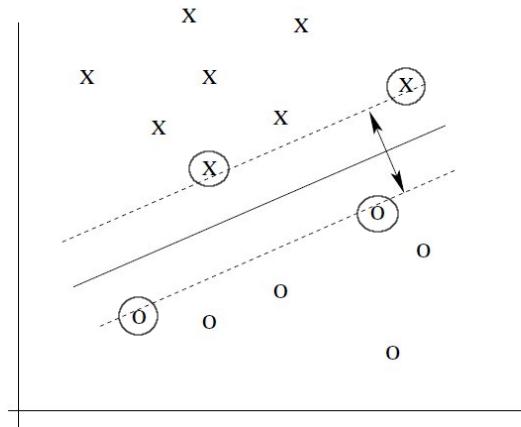


Figura 26 – Os vetores de suporte num caso de separação linear e o hiperplano separador.

Fonte: Adaptada de Luxburg e Scholkopf (2008).

SVM Multiclasse

Embora o método de SVM seja normalmente utilizado para a classificação binária, ele pode ser adaptado para problemas de múltiplas classes. Um classificador SVM multiclasse pode ser obtido através da formação de diversos classificadores e combinando os seus resultados. Duas das estratégias mais utilizadas para desenvolver SVM multiclasse são o "*one-against-one*" e "*one-against-all*". No primeiro caso, os classificadores combinados em um comitê e treinados para cada par de classe e a classe predominante fica sendo a predita. Já no segundo caso, os classificadores são treinados para cada classe contra todas as outras, ou seja, um classificador para cada classe, no qual separa binariamente esta das outras e a amostra de classificação é passada por todos esses classificadores, sendo rotulada com o rótulo do SVM que identificou aquela amostra como da sua classe correspondente.

Neste trabalho, classificadores SVM lineares são usados com entradas acima descritas. Uma vez que os recursos *bag-of-words* são obtidos das imagens de treinamento, eles são passados para os SVM que encontram um hiperplano que separa os dados de treinamento pela margem máxima. É empregada a abordagem "*one-against-all*", uma vez que alcança um desempenho de velocidade mais rápido comparado com a "*one-against-one*". Na execução *one-against-all* do SVM, n hiperplanos são implementados, em que n é o número de classes. Cada hiperplano pode ser usado para separar uma classe das outras.

3.8 Considerações Finais

Neste capítulo foram explicadas as técnicas de atenção visual na seção 3.1. O VOCUS2 (seção 3.2) é uma técnica de segmentação e atenção visual que segue os modelos de iNVT com uma variação nas pirâmides gaussianas, mas de melhor eficiência. Os modelos *bottom-up* e *top-down* (seção 3.3) só importantes para extrair as partes da imagem de interesse. As redes

neurais (seção 3.4) possuem algumas vantagens de implementação e velocidade e são usadas neste trabalho, para segmentação e classificação. As extrações de características (seção 3.5) é um passo essencial no processamento da imagem para uma melhor classificação. O modelo de *bag-of-features* (seção 3.6) agrupa as características das imagens de treinamento para classificar as imagens de teste segundo as semelhanças de características com as das imagens de treinamento. SVM (seção 3.7) é uma técnica da Teoria de Aprendizagem Estatística que possui artifícios robustos para classificação binário, e mais de um classificar pode ser combinado para formar um classificar multiclasse.



SISTEMA PROPOSTO

Tendo apresentado a pesquisa a priori e os métodos necessários para viabilizar o projeto proposto, este capítulo traz a descrição do sistema na seção 4.1, os detalhes de implementação do sistema de visão e de voz na seção 4.2 e as sessões interativas pedagógicas propostas na seção 4.3.

4.1 Descrição

O sistema proposto é um sistema interativo que conduza sessões de interação por meio de um robô humanoide, podendo ser treinado com diferentes conteúdos a serem abordados com os usuários de forma autônoma. Sua aplicação visa a utilização do sistema como ferramenta de auxílio no ensino de matemática para crianças.

O robô escolhido foi o robô NAO por já fazer parte do conjunto de robôs pertencentes ao Laboratório de Aprendizados de Robôs (LAR), do ICMC, e por apresentar ferramentas já fornecidas pelo fabricante. Entretanto, outros robôs ou até mesmo um avatar virtual podem ser agregados ao projeto com as alterações nos módulos afetados. Para os testes e avaliações iniciais, as formas geométricas 3D escolhidas são formas básicas, como mostra a Figura 27: pirâmides, cubos e esferas de cores diferentes. Essas peças foram feitas pela impressora 3D do mesmo laboratório. Após treinar os algoritmos e atingir resultados satisfatórios para essas figuras, o sistema encontra-se pronto para ser testado com usuários de maneira autônoma, segundo as sessões apresentadas na seção 4.3.

O esquema proposto é baseado em módulos, no qual cada módulo é responsável por uma função específica e contém um grupo de funcionalidades para tal. Além de ser favorável para um código mais limpo e organizado, esse tipo de implementação permite alterar facilmente um módulo - ou apenas uma parte dele - sem muito impacto nos outros, garantindo um menor acoplamento e uma maior coesão (JACOBSON, 1992).

Os módulos são brevemente descritos a seguir e a Figura 28 ilustra como funciona a

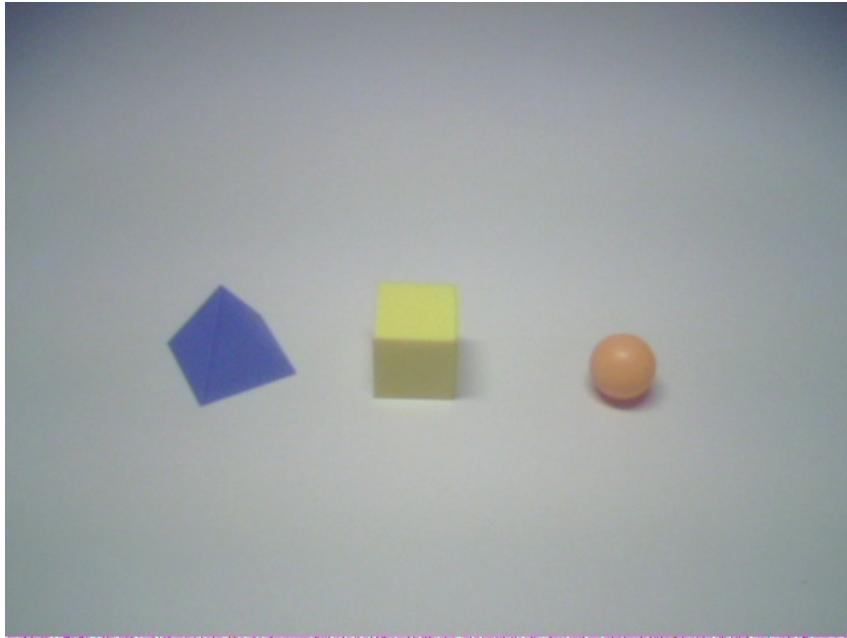


Figura 27 – Formas geométricas 3D básicas utilizadas para reconhecimento.

Fonte: Elaborada pelo autor.

dinâmica de comunicação entre eles.

Módulo Central

É o módulo principal que contém os outros módulos. Todo programado em *C++*, com objetos para a comunicação entre os módulos e também alguns mecanismos como sorteio de figuras geométricas e memória da posição dos objetos para ajudar e orientar o fluxo da interação. Ele se conecta com o robô por meio de proxies e com os seus módulos por chamadas de função.

Módulo de Diálogo

O Módulo de diálogo tem duas funções: interpretar o que o usuário diz, convertendo a fala em texto, e dar informações para o usuário, convertendo as frases dadas pelo Módulo Central na síntese de fala. Desta forma, é empregado o *Google Speech Recognition*, uma *Application Programming Interface (API)*¹ que envia um arquivo em formato *wave* para seu servidor web e tem como retorno a *string* correspondente. Esta comunicação é programada em *Python*, com a biblioteca *SpeechRecognition 3.1.3* ([FOUNDATION, 2014](#)), e este módulo se comunica com o Módulo Central por *stream* de arquivo. Para a síntese de voz foi utilizada a voz padrão do NAO, por ser bem aceita conforme mostrado da seção [5.1.1.1](#).

Módulo Motor

No módulo Motor estão implementadas funções que utilizam *NAOqi API*, um *framework* fornecido pela empresa fabricante para posicionar os motores do NAO. É sempre usado em conjunto com outro módulo, mas muito importe para acrescentar novidade às interações.

¹ Interface de Programação de Aplicação

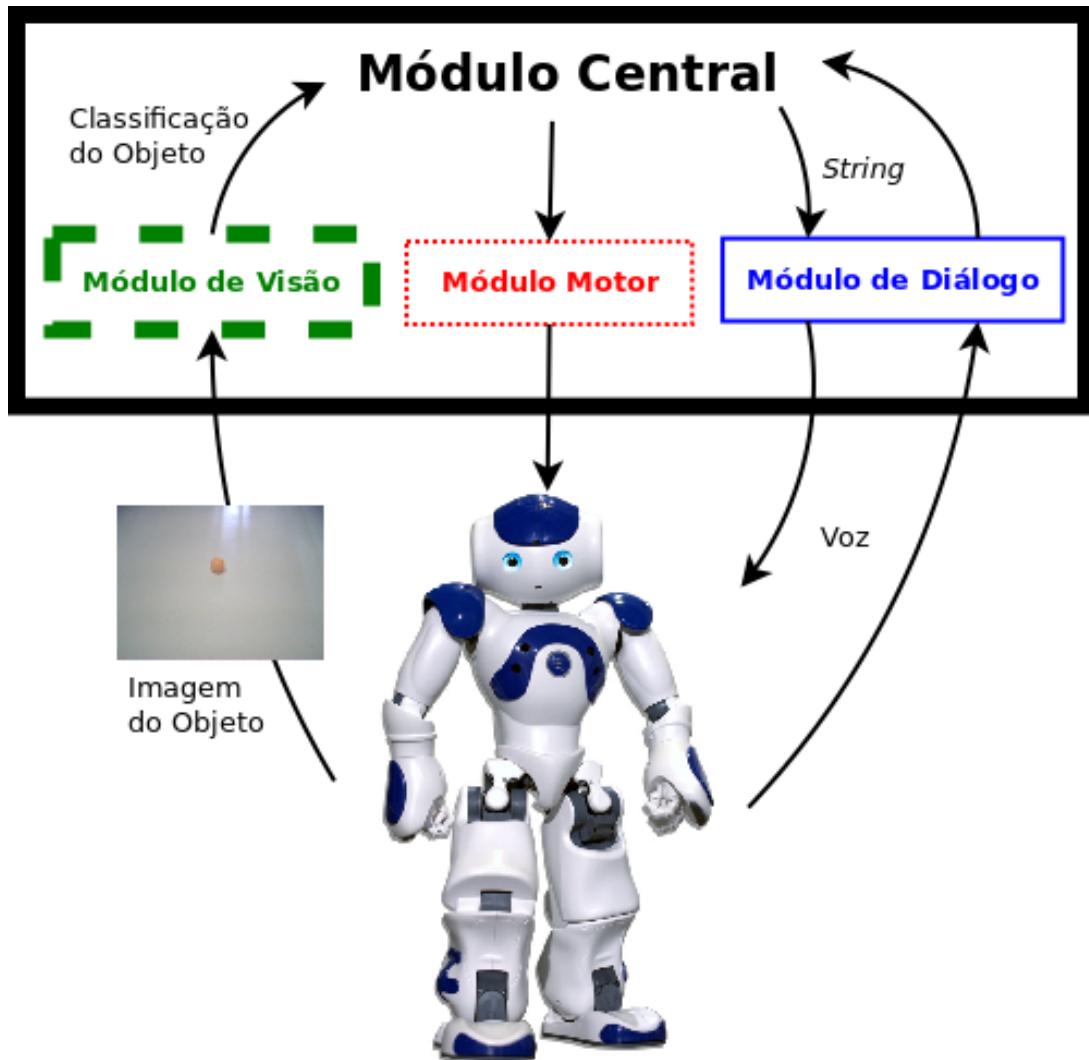


Figura 28 – Dinâmica de comunicação entre os módulos.

Fonte: Elaborada pelo autor.

Módulo de Visão

O Módulo de Visão está implementado com uma série de técnicas para detectar e reconhecer os objetos selecionados. Em primeiro lugar, o VOCUS2 (seção 3.2) para detecção e extração de fundo. Então, são computadas as características por SURF (subseção 3.5.2) e aplicado o método *bag-of-features* (seção 3.6) em seu histograma, e, finalmente, treinadas SVM multiclassas da forma *one-against-all* (seção 3.7) com estes descritores para predizer a classe de novas imagens.

4.2 Materiais e Métodos

4.2.1 Sistema de Visão

O sistema de visão empregado é composto por uma combinação de técnicas, que podem ser substituídas por outras similares, para detectar e classificar objetos. Entretanto, foram comparados apenas dois métodos para as funções de detecção e outros dois para a de classificação. Para detecção foram comparadas as técnicas de atenção visual utilizando uma rede LEGION e outra utilizando o VOCUS2. Já para a classificação, foram comparadas as técnicas de MLP e SVM. Os resultados de ambas as comparações são apresentados e discutidos a seguir.

Detecção

A detecção é o primeiro passo no processamento da imagem fornecida pela câmera do robô. A Figura 29 mostra um exemplo do processo de saliência para detecção. A primeira imagem (29a) é a imagem original de 400x400 pixels capturada pela câmera do NAO, a segunda imagem (29b) é o mapa de saliência gerado. A terceira imagem (29c) mostra o primeiro objeto mais saliente da imagem e, finalmente, a última imagem (29d) é a imagem recortada do objeto detectado.

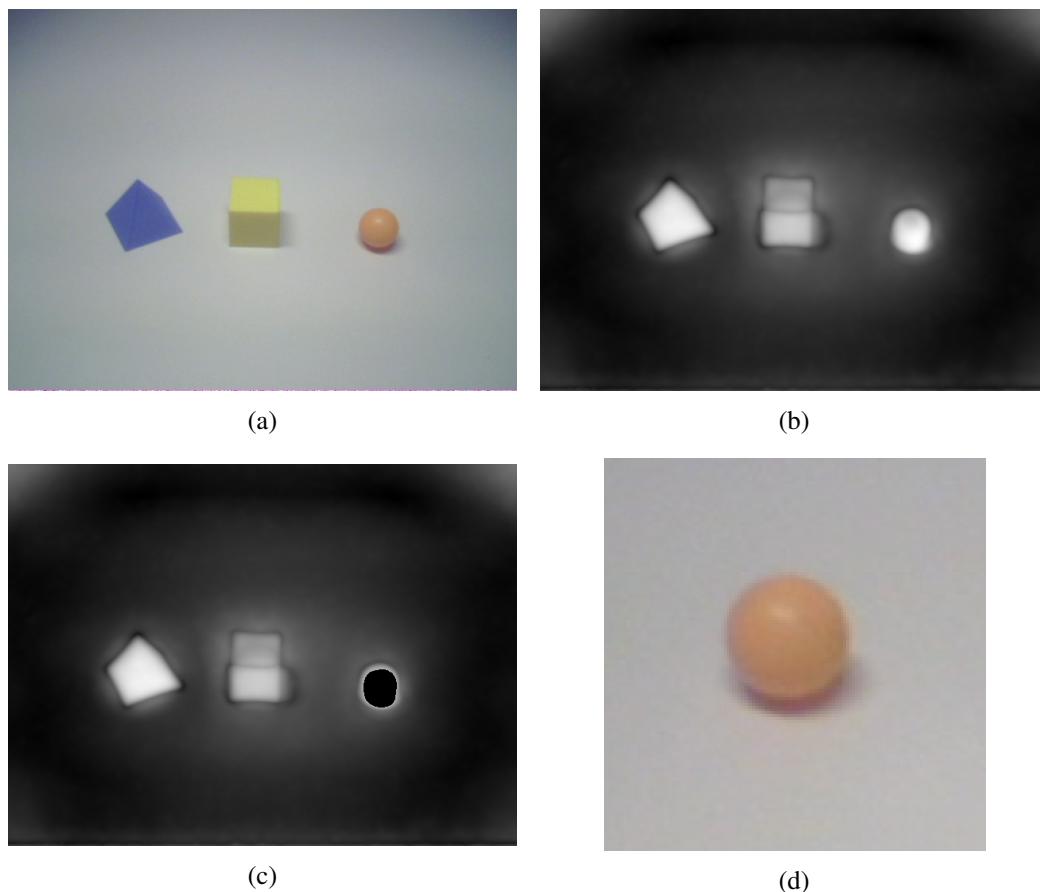


Figura 29 – Processo de detecção com o VOCUS2.

Fonte: Elaborada pelo autor.

A rede LEGION conseguiu identificar a posição de todos os objetos, porém, com um custo computacional maior que o VOCUS2. Além do mais, o fato da rede LEGION ser oscilatória, fazia com que as faces das figuras geométricas 3D fossem detectadas separadamente em uma das fases da oscilação, resultando numa segmentação não tão boa da imagem e desperdiçando processamento, uma vez que não são necessárias outras oscilações após encontrar uma das faces. Os parâmetros utilizados para a LEGION foram $\theta_p = 400$ (influência na formação de neurônios líderes) e $W_z = 5.0$ (peso do inibidor global).

Em contra partida, com os parâmetros utilizados, o VOCUS2 conseguiu uma segmentação do objeto como um todo e com um custo computacional melhor aproveitado quando comparado com a rede LEGION. Por esses motivos, foi feita a escolha de tirar as medidas de classificação finais e deixar o sistema equipado com o VOCUS2.

O principal parâmetro a ser ajustado na detecção é o limiar de centro-vizinhança, que resulta na junção de pixels vizinhos, formando a imagem salientada no último mapa. A versão final deste projeto está com este limiar configurado em 0.6. A Figura 30 mostra os resultados para diferentes valores deste parâmetro de 0.1 a 0.9 variando em 0.2 unidades.

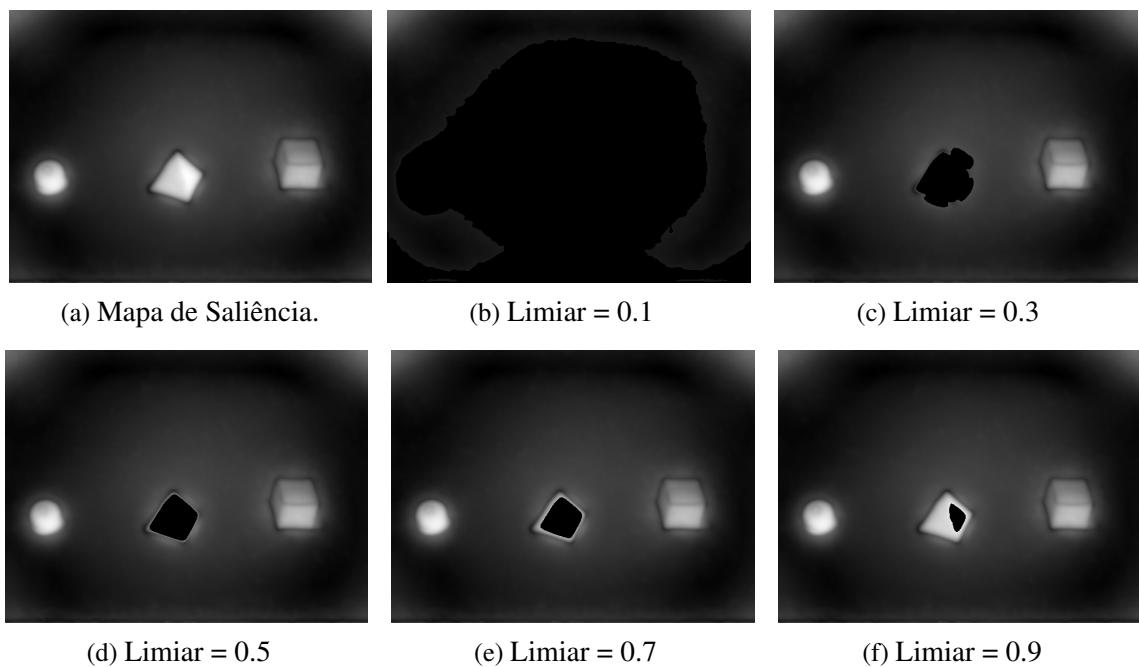


Figura 30 – Variando o parâmetro de segmentação do VOCUS2.

Fonte: Elaborada pelo autor.

Outro parâmetro que podemos variar é o número de objetos salientes desejados. Eles são apresentados um a um, em ordem decrescente de saliência. Na subseção 4.3.2 é apresentada uma variação desse parâmetro para 3 objetos.

Classificação

Inicialmente, para as entradas da MLP foram utilizados os perfis das formas geométricas

detectadas. Após o objeto ser detectado, a imagem era convertida para a escala cinza e os pixels pertencentes ao objeto eram alterados para 0 e o resto da imagem para 1, formando uma imagem binária com a silhueta do objeto, conforme mostrado na Figura 31. A MLP foi treinada e testada com essas transformações nas imagens da base de dados de treinamento e de teste, respectivamente.



Figura 31 – Silhueta dos objetos.

Fonte: Elaborada pelo autor.

A MLP foi treinada com 4480 neurônios na camada de entrada (tamanho da imagem de 70x64 pixels), 1 camada escondida de 5235 neurônios (adquirido empiricamente testando camadas entre 30 - 8000 neurônios), 2 neurônios na camada de saída, $\eta = 0.01$ (passo de aprendizado) e $e_j(p) = 0.001$ (Limiar de parada). Apesar de uma convergência rápida no treinamento, ou seja, alcançar a margem de erro desejada no algoritmo de *Backpropagation* em pouco tempo, e um conhecido custo computacional baixo, a acurácia obtida com essa técnica foi um pouco abaixo de 77%, que é uma margem considerada pouco satisfatória quando comparada com outras técnicas conhecidas, como o SVM por exemplo.

A solução escolhida para suprir a falta de informações sobre as imagens foi computar seus descritores SURF e tentar classificá-las por meio dessas características. Essa solução é acompanhada do modelo de BoF para detectar características a partir do histograma da imagem e agrupar essas características por semelhança com o algoritmo KNN. Porém, não foi possível alcançar uma convergência para o algoritmo de *backpropagation* (subseção 3.4.1).

A vantagem da MLP sobre o SVM é que, após treinada, seria necessário apenas um classificador para todas as classes. Entretanto, substituir o classificador MLP por um SVM foi uma alteração bem sucedida. A técnica multiclasse executada foi a de *one-against-all* (seção 3.7). Foi treinado um SVM da biblioteca openCV² de núcleo linear para detectar cada classe: esfera,

² <http://opencv.org/> Visitado em fevereiro de 2016

cubo e pirâmide, nesta ordem. Nesta abordagem, cada SVM prediz se a amostra pertence à sua respectiva classe ou qualquer uma das outras. A grosso modo, o SVM de uma classe é treinado com exemplos positivos e negativos da classe que está representando. Quando uma amostra desconhecida é apresentada, o SVM da primeira classe prediz se essa amostra também é um exemplo positivo de sua classe ou negativo. Se for negativo, ou seja, pertence à outra classe, a amostra é submetida à predição de SVM da classe seguinte. A amostra é rotulada se um dos classificadores SVM reconhece-la como pertencente à sua respectiva classe, caso contrário a amostra é classificada como desconhecida.

Por exemplo, se um fragmento que não representa nenhuma das formas geométricas do conjunto de treinamento for detectado erradamente pelo VOCUS2 e passado para o sistema e classificação, nenhum classificador irá predizer como de sua classe e esse objeto sairá como desconhecido do classificador multiclasse.

O conjunto de dados de treinamento foi composto por 60 amostras, sendo 20 de cada classe.

As imagens foram tiradas pela câmera do robô e seus tamanhos podem variar de acordo com a suas posições no campo de visão do robô, ou seja, a distância entre os objetos e a câmera. Na Figura 33 é possível ver algumas amostras do conjunto de dados de treinamento.

Importante lembrar que, como o robô ficava de frente para os objetos, colocar o objeto mais próximo ao robô diminuía seu posicionamento no eixo X da imagem, enquanto que posicioná-lo mais longe aumentava suas coordenadas nesse eixo. Para o eixo Y da imagem, quanto mais à esquerda do robô o objeto se encontrava no campo de visão, menor era sua coordenada nesse eixo, ao passo que quanto mais a direita, maior o valor de sua coordenada Y na imagem. A Figura 32 ilustra essa relação entre posicionamento no campo de visão do robô e coordenadas na imagem.

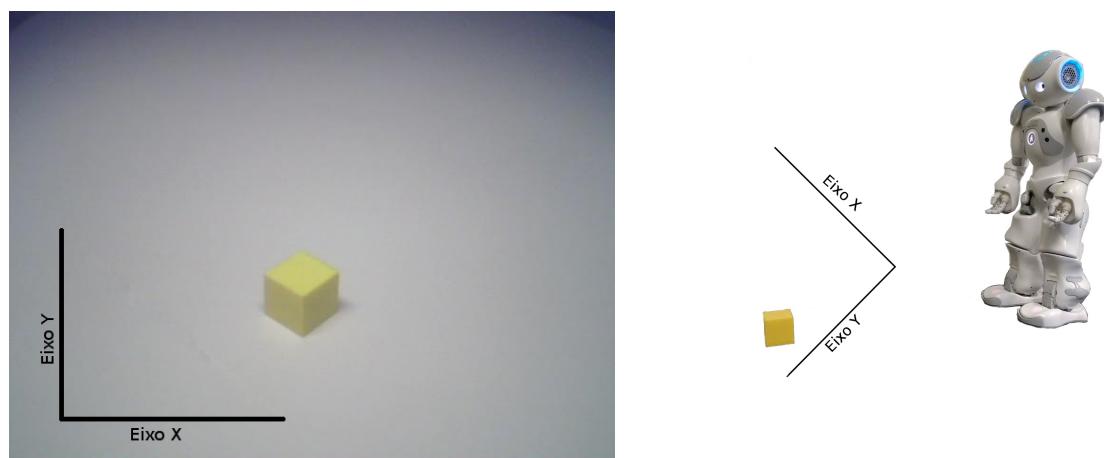


Figura 32 – Diferenças dos eixos: Imagem capturada e Visão do Robô NAO

Fonte: Elaborada pelo autor.

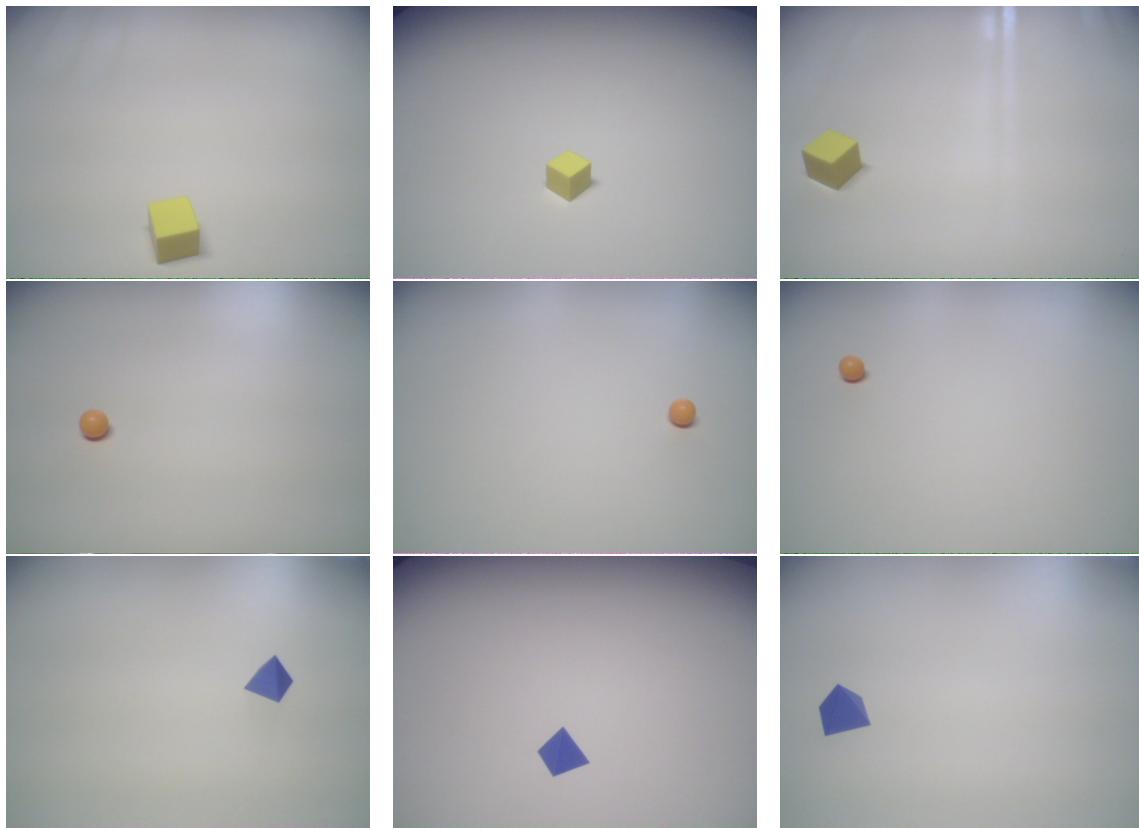


Figura 33 – Imagens da base de treinamento.

Fonte: Elaborada pelo autor.

Esta implementação mostrou-se adequada para o projeto, uma vez que foi obtida uma acurácia de 93% de precisão e tempo de treinamento em 0.9s para as figuras geométricas 3D escolhidas.

4.2.2 Sistema de Diálogo

As funcionalidades de voz do deste sistema são básicas. Elas devem viabilizar um diálogo padronizado do robô com o usuário, no qual o sistema deve apenas interpretar algumas entradas por comando de voz e responder algumas perguntas. Portanto, foram utilizadas plataformas prontas como o *Google Voice Recognizer*³ e o sintetizador de voz do fabricante do NAO.

O *Google voice recognizer* é uma *API* que se conecta a um sistema da própria Google pela Internet para traduzir fala em informação textual para uso em software. Ou seja, com um formato de áudio de entrada retorna como saída uma *string* correspondente ao texto do áudio de entrada.

Para a tarefa de síntese de voz, foi utilizado um software fornecido pela própria Aldebaran, uma *framework NAOqi*, melhor explicado na subseção 4.2.3.2. Apesar de possuir pouco suporte para o idioma português brasileiro - quando comparado ao inglês ou francês, por exemplo - esse

³ <https://www.google.com/intl/pt/chrome/demos/speech.html> Visitado em 10/02/2016.

mecanismo mostrou-se satisfatório para o diálogo desta aplicação, haja visto que no teste com crianças, na seção 5.1, elas alegaram entender com clareza as falas produzidas pelo robô.

Não houveram parâmetros para o reconhecimento de fala, pois a ferramenta utilizada não oferece nenhuma opção de alteração. Já os parâmetros de síntese de voz são: volume da voz do robô e velocidade da fala. Esses parâmetros podem ser configurados para cada frase e foram ajustados empiricamente para se ajustarem a cada tarefa.

Em todo o projeto, as perguntas foram elaboradas para induzir o usuário a dar uma única resposta, como "sim" ou "não". Porém, se o usuário fornecesse ao robô uma resposta maior, o Módulo de Diálogo procurava pela palavra esperada dentro da frase do usuário. Por exemplo, se a palavra esperada fosse "doze" e o usuário respondesse "Essa figura tem doze arestas", a palavra era procurada dentro da frase e a resposta validada por ser encontrada. Algumas palavras tinham preferência sobre as outras. Para a mesma palavra esperada, se a resposta dada fosse "Essa figura não possui doze arestas", essa resposta fazia com que o fluxo seguisse pela opção de resposta negativa a essa pergunta, pois continha uma palavra de negação na resposta.

4.2.3 Robô Humanoide NAO

A seguir, são apresentadas as principais características físicas e de desenvolvimento do robô NAO.

4.2.3.1 Hardware

Como características físicas, destacam-se seus 60 cm altura, 5 kg de peso, 25 graus de liberdade, 2 câmeras, 4 microfones, sintetizador de voz, entre outros itens não tão pertinentes à este trabalho. A Figura 34 mostra onde esses dispositivos se encontram ao longo do corpo do robô. Em questões de software, ele suporta as principais linguagens de programação, sendo a de maior suporte a de C++ ([ALDEBARAN, 2014](#)). O robô tem sido utilizado em mais de 550 universidades e laboratórios de pesquisa espalhados pelo mundo. Robótica, Inteligência Artificial, Ciência de Computação, divergindo para Sociologia e Medicina são algumas das áreas de pesquisa onde esse robô tem sido utilizado.

4.2.3.2 NAOqi

NAOqi é o conjunto de softwares disponibilizados pelo fabricante do NAO e é composto pelo Monitor, que apenas permite monitorar as câmeras do robô e o Choregraphe. O Choregraphe ([POT et al., 2009](#)) é um software distribuído pela empresa fabricante do NAO - Aldebaran - para programação fácil e rápida do robô. Quando o software é executado, a interface gráfica, mostrada na Figura 35, se encarrega de traduzir diagramas de funcionalidades do robô em código. Com isso, não é necessário que o usuário tenha conhecimento prévio de programação, pois apenas "arrastando e soltando" as caixas que representam as funcionalidades desejadas

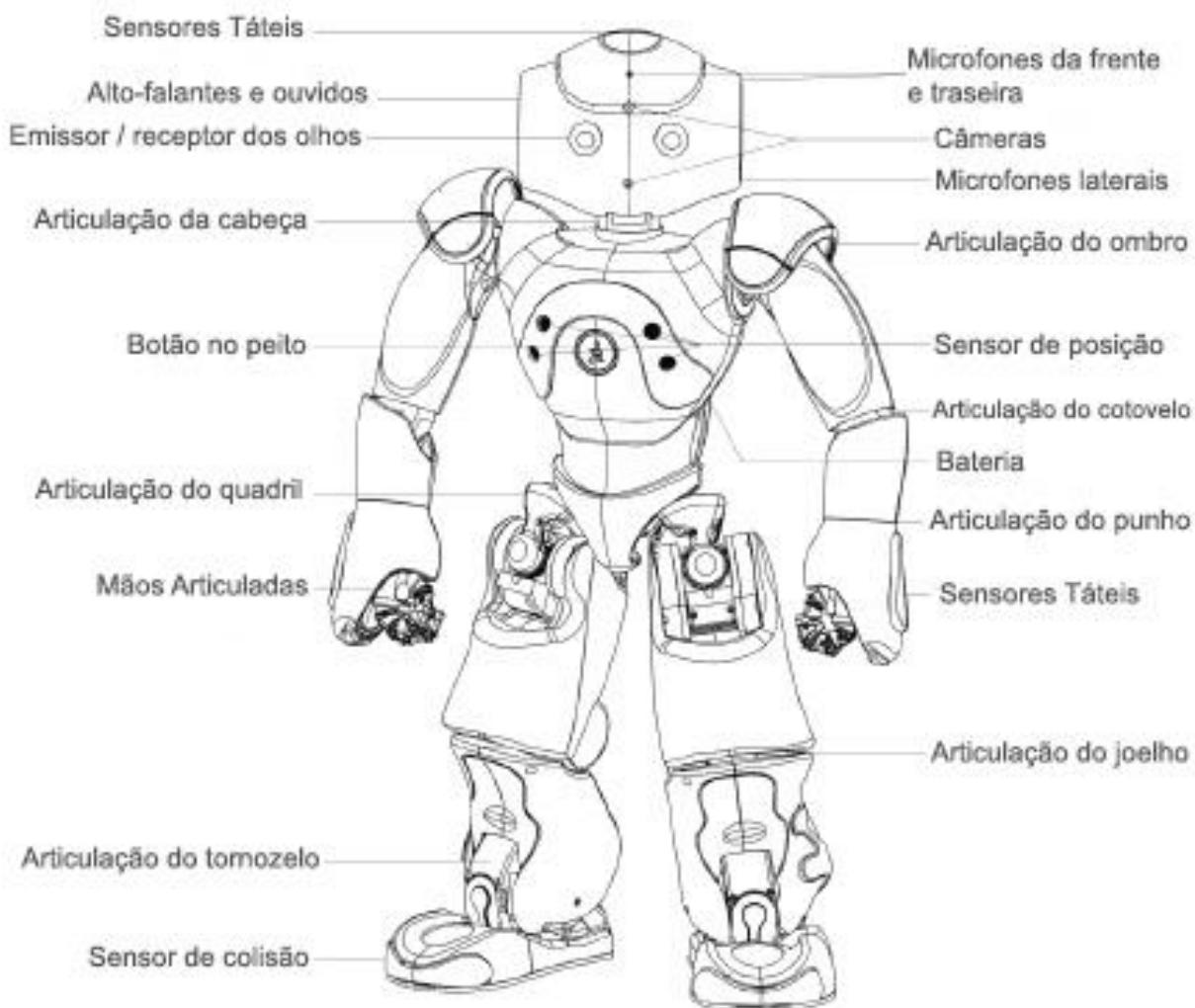


Figura 34 – Distribuição corpórea do robô NAO.

Fonte: Adaptada de [Aldebaran \(2014\)](#).

da zona 1 à zona 2, indicadas na figura, e conectando caixas é possível criar uma infinidade de comportamentos personalizados oriundos da combinação desses diagramas. A ordem de execução segue a ordem em que as caixas são conectadas, portanto, uma caixa só será executada quando a caixa antecessora for finalizada.

É possível utilizar o *Choregraphe* sem a presença de um robô. Dessa forma, as saídas são simuladas pelo robô NAO 3D na zona 3. Porém, quando o robô está conectado ao sistema, seja via cabo ou Wifi, a zona 3 serve de espelho do mesmo, ou seja, o comportamento é executado tanto pelo robô físico como pela simulação 3D. Na verdade, essa janela representa o estado atual do robô. A zona 4 mostra como é simples retirar o código de cada caixa bastando um duplo clique na caixa de interesse. O comportamento gerado - a combinação toda - pode ser igualmente recuperada em forma de código e ambos podem ser integrados à qualquer código. Apesar da Figura 36 mostrar código em C++, é possível utilizar variadas linguagens de programação,

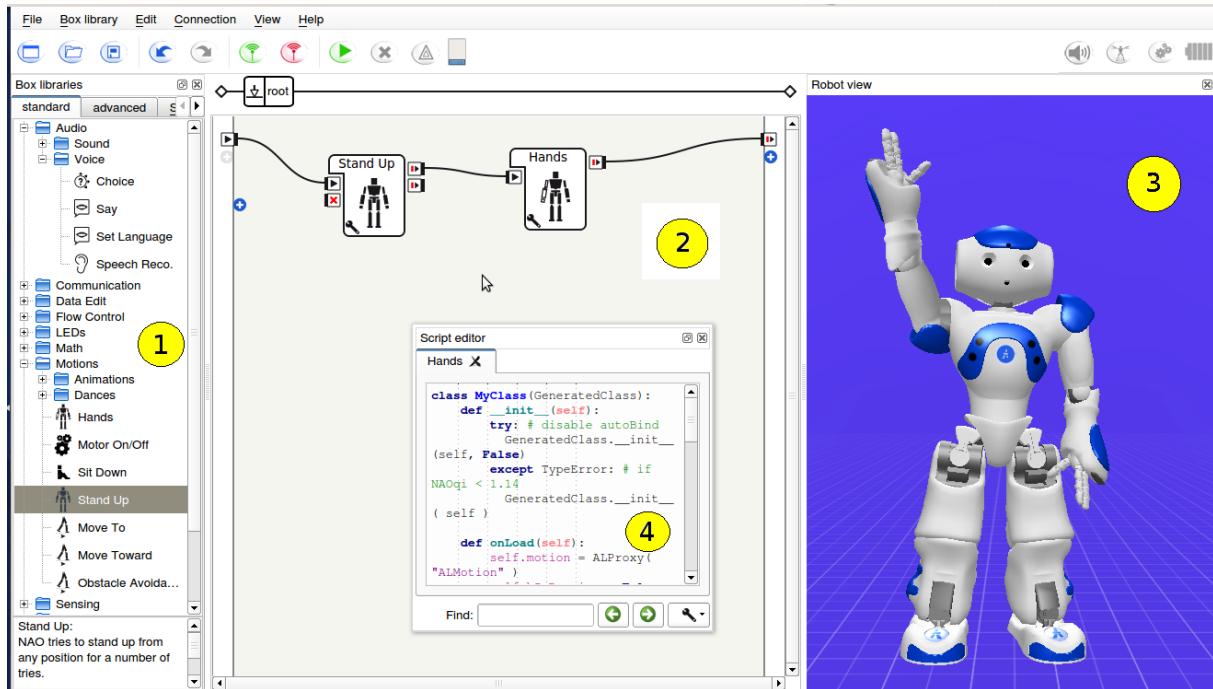


Figura 35 – Interface gráfica do Coregraphe.

Fonte: Elaborada pelo autor.

como Java e Python, porém a linguagem com maior suporte e inclusive indicada pela própria Aldebaran para desenvolvimento é C++. Portanto, toda a interação Homem-Robô será feita dentro do *Choregraphe*, como as falas, os movimentos de braços, pernas e cabeça, a coloração dos Leds (vermelha para o caso de erro e azul intermitente para o caso de acerto). Por meio de bibliotecas padrão proprietárias da Aldebaran, todos esses comportamentos salvos no robô podem ser acessados diretamente por programas escritos em C++.

Toda a implementação da atenção visual será feita em C++, e será necessário integrá-la ao robô NAO. Para isso, existe um *Software Development Kit* (SDK) de desenvolvimento de Software para ser utilizada com o sistema do NAOqi. A partir desta SDK, um novo módulo pode ser criado e salvo no robô. Essa SDK é chamada NAOqi API, disponível para computadores em 8 diferentes linguagens, porém apenas C++ e Python são interpretadas pelo robô. A *framework* C++ permite a escrita de código em tempo real, o que é importante para nossa abordagem.

Para compilar os códigos para o robô é recomendado pela própria Aldebaran a utilização do CMake (um sistema multiplataforma para realizar geração automatizada), juntamente com a *framework* qiBuild. O qiBuild gerencia as dependências entre os projetos e suporta o *cross-compilation*, permitindo a boa interpretação do código pelo OpenNAO. É obrigatório, para a versão do robô NAO utilizado, a compilação utilizando arquitetura de 32-bits. Apesar disso, o código resultante irá funcionar corretamente para qualquer ambiente, seja 32 ou 64 bits. Também considerando a versão utilizada do robô, a *framework* C++ SDK suporta apenas a versão 2.3.1 da biblioteca OpenCV, uma das maiores e mais utilizadas bibliotecas de visão

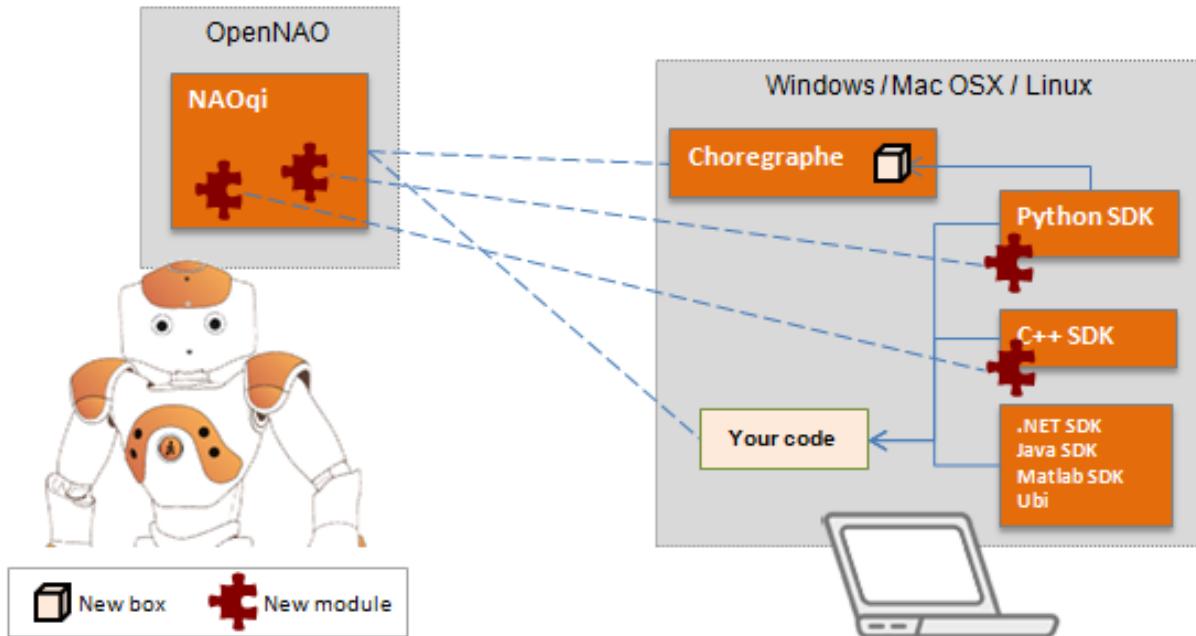


Figura 36 – Visão geral do funcionamento do software do NAO.

Fonte: [Pot et al. \(2009\)](#).

computacional de código aberto, responsável por tratar as imagens das câmeras do robô. A Aldebaran recomenda a utilização da *Integrated Development Environment* (IDE) - Ambiente de Programação) do QT Creator, pois algumas funções do QT estão inseridas no robô, tornando mais fácil a implementação de um novo módulo.

4.3 Sessões interativas propostas

Ao final deste mestrado, o sistema encontra-se pronto e configurado para realizar sessões de forma autônoma em duas fases de desafios: figuras geométricas 3D individuais, no qual o robô desafia a criança a encontrar o objeto que o robô escolher, e múltiplas figuras geométricas 3D, sendo o robô desafiado a acertar qual objeto a criança escolheu por meio de dicas.

O sistema completo está configurado para interagir com usuários, pedindo, reconhecendo e explicando as diferenças e conceitos sobre os três tipos escolhidos de figuras geométricas básicas 3D. Esses objetos foram feitos de plástico pela impressora 3D do Laboratório de Aprendizado de Robôs, exceto a esfera, que é uma bola de tênis de mesa. O motivo pelo qual esse conteúdo foi escolhido é que, além da simplicidade desses objetos, estudos preliminares ([PINTO; TOZADORE; ROMERO, 2015](#)) acusaram uma grande dificuldade por parte das crianças em diferenciar figuras geométricas 2D das figuras geométricas 3D, como chamar um cubo de quadrado ou uma pirâmide de triângulo. Assim, por questões pedagógicas, o Módulo de Diálogo foi configurado com frases para abordar o conceito de vértices, arestas, faces, bases e as coisas que caracterizam esses objetivos adotados. Vale a pena lembrar que, substituindo o conteúdo do

Módulo de Diálogo e de visão, é possível lidar com tantos assuntos quanto se queira.

O conjunto de interações entre o robô e o usuário para mais de uma tarefa é chamado de sessão e as combinações de técnicas já descritas são usadas para guiar o robô ao longo das sessões. Considerando o conteúdo selecionado, e com a ajuda de especialistas em educação matemática, foram sugeridas duas fases de desafios entre o robô e os usuários.

4.3.1 *Figuras geométricas 3D individuais: Robô propõe desafio para as crianças*

Nesta fase, as figuras geométricas 3D são apresentadas uma a uma ao robô conforme sua requisição. A configuração inicial é que não haja nada na mesa e o robô comece fazendo a requisição de algum objeto por meio de dicas, como por exemplo, quantas faces esse objeto tem, e a criança deve colocar o objeto certo na mesa dentro do campo de visão do robô. O robô deve reconhecer o objeto corretamente e esboçar expressão de satisfação se a criança acertar ou de decepção se errar. O nível de interação pouco varia para esse experimento, adicionando um pequeno dialogo inicial e final e colocando mais falas nos diálogos intermediários no grupo de alta interação. Esta sessão é representada pelo fluxograma ilustrado na Figura 37, que é composto por 8 passos, ou etapas, do fluxo principal e 2 passos complementares, devido às possíveis mudanças em algumas etapas. Para melhor discriminar o papel de cada módulo, a figura traz todas as etapas com bordas, de acordo com os módulos que atuam nesse passo, sendo alguns deles realizados por mais de um módulo.

No passo 1, o robô executa um pequeno diálogo de boas-vindas com a criança, aperta sua mão e pergunta seu nome para usar ao longo da sessão. É um passo importante, pois familiariza as crianças com o robô (PINTO; TOZADORE; ROMERO, 2015) fazendo com que elas percam a timidez logo no inicio da sessão. Além disso, o robô passa as instruções para o usuário sobre esta sessão. No passo 2, o Módulo Central escolhe aleatoriamente um dos três possíveis objetos para desafiar a criança a encontrá-lo. No passo 3, o Módulo de Diálogo solicita o objeto escolhido, mas sem dizer seu nome, apenas dando dicas e características. Um exemplo para o cubo poderia ser: "Você poderia, por favor, identificar e colocar na minha frente uma figura geométrica 3D com 8 vértices e 12 arestas?". No passo 4, o Módulo de Visão prediz o objeto e devolve a previsão para o Módulo Central, que no passo 5 verifica se o objeto escolhido pela criança coincide com a escolha no passo 2. Em caso negativo, o robô dá mais dicas e regressa ao passo 2, solicitando a figura novamente. Em caso positivo, o fluxo pela etapa 6, o robô felicita pela escolha certa e dá mais alguns conceitos complementares ou curiosidades sobre o objeto. No passo 7, o Módulo Central pode ser configurado para jogar um determinado número de vezes ou perguntar à criança, por meio do Módulo de Diálogo, se ela deseja jogar novamente. Se houver outra rodada, os passos são reproduzidos a partir do passo 2. Se não, a sessão se encerra na etapa 8 com um diálogo de despedida, no qual o robô encoraja a criança a, posteriormente, estudar os conceitos abordados nesta sessão.

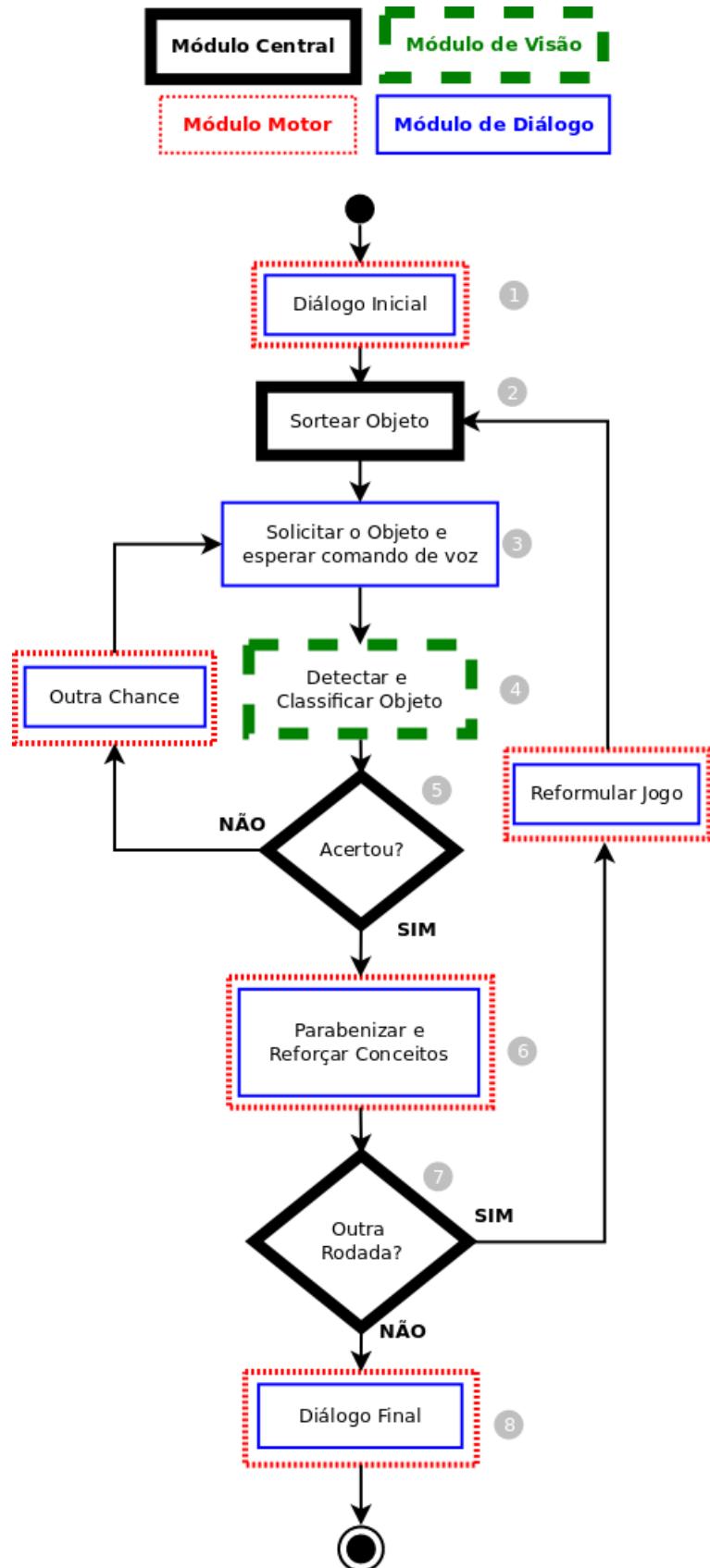


Figura 37 – Fluxo de interação para sessão de figuras geométricas 3D individuais.

Fonte: Elaborada pelo autor.

4.3.2 Múltiplas figuras geométricas 3D: A criança desafia o robô

A vez do usuário desafiar o NAO é feita da seguinte forma: o participante deve embaralhar as figuras geométricas em cima da mesa, de maneira que terminem alinhadas horizontalmente no campo de visão do robô. Em seguida, deve escolher mentalmente uma das três figuras e avisar verbalmente que já o fez. Após receber esse sinal, o NAO inicia uma série de perguntas conceituais ou relacionais para descobrir qual figura geométrica o usuário escolheu e assim identificar em qual posição a figura escolhida está: à direta do robô, ao centro ou à esquerda do robô. A Figura 38 mostra um exemplo de uma árvore de decisão que o Módulo de Diálogo cria para chegar à figura escolhida pelo usuário. O módulo de Dialogo é responsável por identificar qual foi a figura escolhida pelo participante, ao passo que o Módulo de Visão reconhece a posição de cada uma das figuras. Ambos passam as respectivas informações para o Módulo Central que as cruza para encontrar em qual posição a figura escolhida se encontra. O veredito é dado pelo Modulo Motor junto com o Módulo de Diálogo, no qual o robô esboça frases e expressões de felicidade, acusando que encontrou a figura escolhida e apontando para ela. Após a confirmação do usuário, o robô pergunta se gostaria de jogar mais uma vez, repetindo os passos já descritos se receber uma resposta positiva, ou encerrando com um pequeno diálogo final em caso contrário.

O fluxo de interação para esta fase de desafio é apresentado na Figura 39.

No passo 1, o robô executa um pequeno diálogo de boas-vindas com a criança, aperta sua mão e pergunta seu nome para usar ao longo da sessão. É um passo importante, pois familiariza as crianças com o robô (PINTO; TOZADORE; ROMERO, 2015) fazendo com que elas percam a timidez logo no inicio da sessão. Além disso, o robô passa as instruções para o usuário sobre esta sessão. No passo 2, o Módulo de Diálogo passa as instruções para o usuário embaralhar as figuras em cima da mesa e escolher mentalmente uma delas. No passo 3, o Módulo de Diálogo escolhe aleatoriamente um árvore de decisão (Figura 38) para descobrir o objeto escolhido. No passo 4, o Módulo de Visão reconhece os objetos e as respectivas posições e informa o Módulo Central, que no passo 5 verifica em qual posição está o objeto informado pelo Módulo de Diálogo. Em caso negativo, o robô pede outra chance e regressa ao passo 2. Em caso positivo, o fluxo segue pela etapa 6, no qual o robô expressa sinais de felicidade. No passo 7, o Módulo Central pode ser configurado para jogar um determinado número de vezes ou perguntar à criança, por meio do Módulo de Diálogo, se ela deseja jogar novamente. Se houver outra rodada, os passos são reproduzidos a partir do passo 2. Se não, a sessão se encerra na etapa 8 com um diálogo de despedida, no qual o robô encoraja a criança a, posteriormente, estudar os conceitos abordados nesta sessão.

4.4 Considerações Finais

Foi optado por deixar o Módulo de Visão equipado com o sistema VOCUS2 e com o classificador SVM, pois estes atendem à esta aplicação com maior eficiência e menor custo

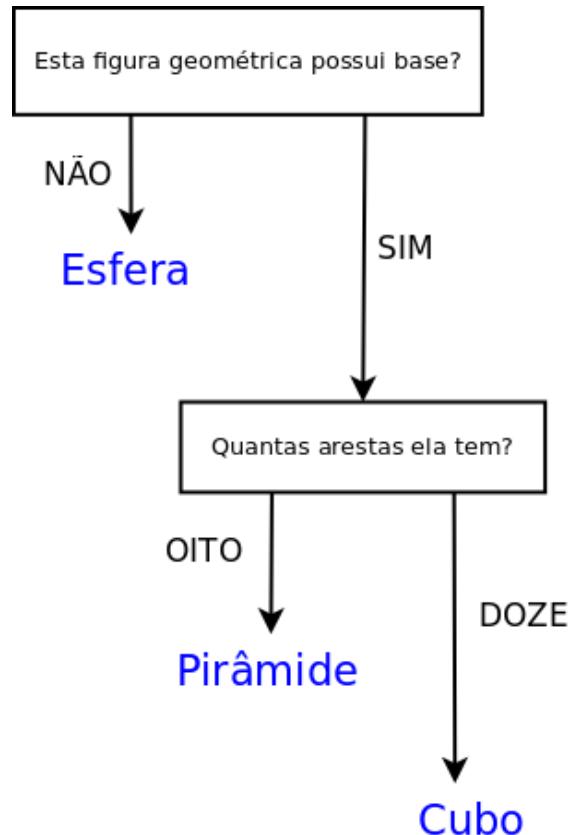


Figura 38 – Exemplo de árvore de decisão para adivinhar a figura escolhida pela criança.

Fonte: Elaborada pelo autor.

computacional. As sessões propostas encontram-se implementadas e prontas para serem testadas em usuários finais.

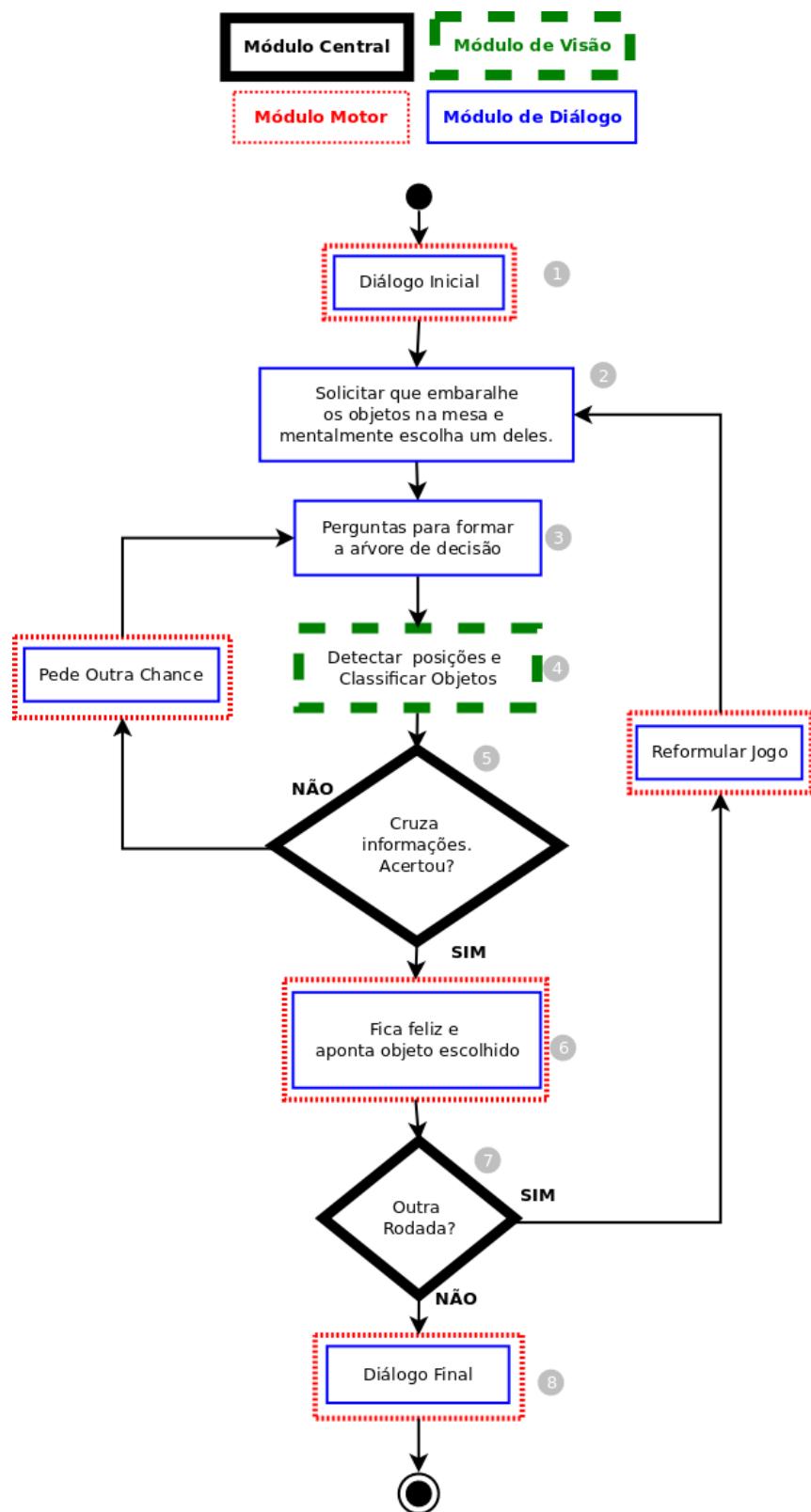


Figura 39 – Fluxo de interação para sessão de múltiplas figuras geométricas 3D.

Fonte: Elaborada pelo autor.

CAPÍTULO

5

EXPERIMENTOS E RESULTADOS

Neste capítulo são apresentados e discutidos os resultados dos estudos interativos com crianças (seção 5.1), bem como testes do módulo de visão e do sistema completo (seção 5.2).

5.1 Estudo de Interação com Crianças

Apesar de parecer um pouco óbvio que a inserção do robô aumenta a participação das crianças nas aulas, a forma com que a interação é realizada pode influenciar no empenho das crianças ao longo da sessão. Conseguir atingir esse potencial máximo de empenho por, meio dos comportamentos do robô, é um importante ponto de interesse nessa pesquisa, pois se trata da aplicação direta deste trabalho.

O Módulo Motor é responsável por toda a parte de mecânica do robô, conforme já descrito na subseção 4.1. Este módulo e a parte do Módulo de Diálogo responsável pela síntese de voz foram combinados para criar comportamentos pré-definidos para serem teleoperados em 3 séries de experimentos com crianças, a fim de estudar como os diferentes níveis de interação podem colaborar para o objetivo final deste projeto.

Os ensaios experimentais seguem a técnica de HRI centrada no usuário (seção 2.5) chamada *Wizard-of-Oz*, que consiste em um robô físico teleoperado que interage com seres humanos presenciais, sendo as métricas tiradas destes últimos. Por essa técnica, entende-se que o sistema ainda não se encontra com seu potencial máximo de autonomia, sendo este um passo intermediário para compreender comportamentos futuros dos usuários com relação às interações com o robô.

Segundo Grice (1991), a interação em HRI é definida como o processo de trabalhar em conjunto para alcançar um objetivo. Neste estudo, o objetivo na interação é a construção do conhecimento do usuário e a consolidação dos conceitos abordados por meio da interatividade do robô, tentando manter a atenção do usuário por um período mais longo. O nível de interatividade

é aqui tratado como o número de recursos e o tempo gasto do robô em cada interação para ajudar os usuários em suas tarefas. Por exemplo, se o robô utilizar discursos longos e com palavras mais animadoras, movimentar-se por meio de seus motores e expressar reações emocionais que acompanham o nível de acerto dos usuários, esta interação tem um nível de interatividade maior do que uma interação na qual o robô utiliza apenas discursos curtos.

Kelley (1984) afirma que uma forma viável de medir os experimentos de HRI centrados no usuário é recolher as impressões e sugestões dos participantes, por meio de questionários ou entrevistas a cada ensaio experimental, entender a correlação entre os pontos apontados pelos usuários e as limitações do sistema proposto e ajustá-los para os próximos experimentos. Sendo assim, durante as discussões dos experimentos realizados são destacadas as opiniões das crianças sobre cada etapa, sendo alguma delas colhidas por entrevista e outras por questionários, como o *Mean Opinion Score* (MOS) (REC, 2000). O MOS é um método de opinião subjetiva e oferece uma escala para medir a qualidade das interações com NAO, além dos seus movimentos, fala e a habilidade de entender e reconhecer as figuras apresentadas. Essa escala vai de 1 a 5, onde 1 significa muito pouco e 5 significa excelente, possuindo níveis intermediários no qual 3 é neutro. Esta técnica funciona como um *feedback* rápido de todos os comportamentos do robô em todas as sessões, permitindo mudanças para melhor interação robô-criança já nos próximos ensaios, como aumentar a velocidade da fala ou adicionar mais informações na requisição.

Este estudo durou em média 3 meses, de março a maio de 2015, e foi dividido em 3 etapas: sessões interativas iniciais, sessões em grupo e jogo de perguntas. A descrição das três etapas e seus resultados são apresentados a seguir. A tabela 5, traz um resumo cada etapa apontando o número de participantes por sessão, seu objetivo, a impressão dos usuários e a conclusão da etapa por grupo de interatividade.

Tabela 5 – Resumo das etapa dos estudos com WoZ.

Etapa	# Participantes	Objetivo	Grupo de Interatividade	Impressão dos Usuários	Conclusão
Sessões Interativas Iniciais	Individuais	Investigar o conhecimento das crianças sobre o assunto escolhido e comparar a qualidade de interação nas tarefas dos dois grupos.	Alta	Disseram que aprenderam mais e se sentiram mais a vontade.	Tiveram melhores interações com o robô e estiveram focadas por mais tempo.
Sessões em Grupo	Grupo de 3 Crianças	Esclarecer as dúvidas comuns dos alunos sobre o conteúdo escolhido e analisar o comportamento das crianças dos dois grupos em sessões coletivas.	Baixa	Acharam o robô rude e alegaram terem aprendido menos que as crianças do outro grupo.	Ficaram concentradas por menos tempo e não se mostraram tão empenhadas quanto as outras.
Jogo de Perguntas	Individuais	Medir o quanto as crianças de cada grupo conseguiram memorizar do conteúdo abordado nas etapas anteriores.	Alta	Não tiveram muito surpresa quanto ao novo comportamento do robô.	Tiveram maior participação durante as perguntas e atividades propostas.
			Baixa	Acharam o robô um pouco mais amigável.	Se mostraram menos participativas e mais tímidas.
				Já estavam acostumadas com os comportamentos do robô e se sentiram confiantes nas interações.	Tiveram um maior número de resposta corretas e estudaram mais em casa.
			Baixa	Ficaram surpresas que o robô "lembra" seus respectivos nomes.	Foram piores nas perguntas, mas mostraram uma boa evolução na relação com o robô.

5.1.1 Sessões interativas iniciais

Esta etapa dividiu as crianças em dois grupos: um que apresentava baixa interatividade no contato com o robô, enquanto o outro apresentava alta interatividade. As sessões eram individuais, compostas por um conjunto de tarefas interativas entre robô e crianças de acordo com o nível de interatividade do grupo em que estavam. No grupo de baixa interatividade, o robô permanecia todo o tempo sentado em suas próprias pernas com as mãos juntas, requisitando e esperando as formas 3D. No grupo de alta interatividade, o NAO tinha um grande repertório de interações para brincar com as crianças ao longo da sessão, porém fazendo - de forma diferente - as mesmas perguntas feitas no outro grupo.

Neste primeiro passo, foi importante identificar não só como as crianças reagiriam ao contato com o robô, mas também, o domínio do conteúdo escolhido que elas possuíam. Assim, as sessões foram montadas para também investigar o quanto as crianças sabiam sobre o conceito de formas geométricas, tais quais arestas, vértices, faces, bases, alturas, etc.

As crianças eram do projeto chamado *Projeto Pequeno Cidadão*, desenvolvido pela Universidade de São Paulo, campus de São Carlos. Neste projeto, crianças de 11-14 anos de idade têm acesso a uma espécie de programa pós-escola, com aulas de reforço, aulas de arte e atividades esportivas durante o período que não estão na escola tendo aulas regulares. É um grupo misto de todas as crianças de escolas municipais de São Carlos, permitindo uma margem aleatória de conhecimento entre os indivíduos. Foi especificado que o grupo de estudantes escolhidos para o projeto deveriam ter desempenhos em educação e habilidades sociais diferentes. Além disso, foi utilizado o mesmo número de meninos e meninas para comparar sua interação com o robô.

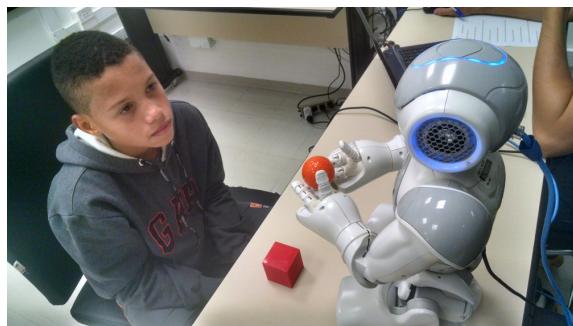


Figura 40 – NAO permanecendo na mesma posição para o grupo de baixa interatividade.

Fonte: Elaborada pelo autor.

Para as sessões do primeiro grupo (grupo de menor interatividade) o robô ficava inicialmente sentado em suas próprias pernas. Depois de uma mensagem de boas vindas (sem qualquer movimento), o robô requisitava um cubo com a mensagem: "Por favor, coloque na minha mão uma figura geométrica com 6 faces, 12 arestas iguais cuja base é um quadrado". A mesma abordagem aconteceu com a pirâmide e a esfera. O NAO pedia uma figura de 4 faces e base triangular (pirâmide) e figuras sem faces ou arestas, que tenham a forma de um planeta. Em todos os casos, os movimentos do NAO foram apenas para reconhecer a figura colocada

pelas crianças em suas mãos, como mostrado na Figura 40, dando uma segunda chance quando a criança errava.

Já para o segundo grupo, de maior interatividade, o robô iniciava as atividades sentado (Figura 41). Quando as crianças chegavam, o NAO se levantava, acenava com as mãos e perguntava pelo nome da criança. Embora fazendo as mesmas perguntas que fez ao primeiro grupo, o robô fazia outros movimentos, como coçar a cabeça ao reconhecer a figura geométrica, um aperto de mãos depois de perguntar o nome da criança e um "toca-aqui", se o aluno acertasse resposta. Em caso de resposta errada, o robô alterava o Led de seus olhos para vermelho, baixava a cabeça e devolvia a figura escolhida para a criança. Mais uma vez, o robô dava novas chances, repetindo dicas e ajudando a encontrar a figura geométrica certa. Por fim, o robô felicitava a participação e encorajava as crianças a estudarem este assunto para futuras sessões.



Figura 41 – Robô sentado esperando antes de uma sessão para o grupo de alta interatividade.

Fonte: Elaborada pelo autor.

Além do MOS, nesta etapa foi utilizado também um questionário de avaliação de conteúdo, proposto pelos pesquisadores envolvidos neste trabalho, e *Continuous Audience Response* (STEVENS *et al.*, 2009). Os métodos de avaliação escolhidos são apresentados e discutidos nas seguintes subseções.

5.1.1.1 MOS

Os resultados do MOS para esta etapa são apresentados na Tabela 6, e em seguida, uma breve discussão sobre comentários extras das crianças.

1. As perguntas do robô foram muito difíceis (5) ou muito fáceis (1)?
2. O robô conversou muito (5) ou pouco (1) com você?
3. Você entendeu muito bem (5) ou muito mal (1) o que o robô disse?

4. O questionário de pós-atividade é muito difícil (5) ou muito fácil (1)?
5. Você julga que aprendeu muito (5) ou pouco (1) com essa atividade?

Tabela 6 – Média obtida pelo MOS

# Pergunta	Nível de Interatividade Baixa (%)	Alta (%)
1	2.4	1.8
2	3.4	2.95
3	5	5
4	3.4	1.5
5	3.5	4.6

Fonte: Dados da pesquisa.

As crianças do grupo de baixa interatividade alegaram que o robô parecia rude no primeiro contato. O fato de o robô simplesmente começar solicitando formas 3D sem uma mensagem de boas-vindas ou uma explicação prévia da atividade não foi bem recebido pelas crianças. Além disso, elas consideraram o questionário de avaliação de conteúdo mais difícil e disseram sentir que aprenderam pouco na sessão.

Até mesmo afirmações ingênuas de ambos os grupos, como "O robô poderia jogar futebol com a gente", indicam que as crianças compreenderam o potencial do robô para interação e que esperavam um pouco mais de diversidade em seu comportamento. O grupo de alta interação disse que se sentiu mais seguro em suas decisões por ter criado uma atmosfera mais amigável com o robô.

5.1.1.2 Questionário de avaliação de conteúdo

A avaliação do conteúdo absorvido pelas crianças nas sessões foi elaborado, com o apoio de especialistas em educação matemática, um formulário de perguntas sobre conceitos do assunto abordado na sessão, mas utilizando situações do dia-a-dia. O questionário não repetia as mesmas perguntas que o robô fez, mas sim, generalizava o conhecimento tratado com as crianças, pedindo associações das formas 3D com objetos do cotidiano, pedindo desenho de alguma figura que o robô requisitou ou perguntando por comparações da diferença de algumas figuras. As questões e respostas esperadas foram:

1. Quantas faces tem um cubo?
R: 6 faces.
2. Qual é a figura geométrica 2D da face do cubo?
R: O quadrado.

3. Quais objetos na sua casa tem o formato de um cubo?

R: Um baú, uma caixa, um dado, ...

4. Desenhe uma pirâmide.

R: *Correção Subjetiva*.

5. Qual a diferença entre o cubo e a pirâmide?

R: Diferentes número de vértices, arestas, faces e figuras 2D de suas bases.

6. Cite 3 objetos presentes no seu dia que tem o formato de uma esfera.

R: Bolas, planetas, lustres ...

7. Porque a esfera é diferente das outras figuras?

R: A esfera não possui faces, arestas e vértices como as outras figuras.

O número de respostas corretas foi maior no grupo de alta interatividade do que no grupo de baixa. Isto é mais notável nas tarefas finais da sessão, quando a média de qualidade de interação tem uma diminuição significativa no grupo de baixa interatividade, ao passo que é mantida no grupo de alta. O desempenho das crianças são apresentados na Tabela 7, discriminadas por grupo.

Tabela 7 – Acertos no questionário de avaliação de conteúdo.

Questão	Acertos (%) por nível:	
	Baixa interatividade	Alta interatividade
1	72.5	92
2	68	87.5
3	68	85
4	90	100
5	25	77.5
6	64.5	85
7	10	75

Fonte: Dados da pesquisa.

Todas essas questões eram abordadas durante as sessões, mas não da mesma maneira com que foram perguntadas no questionário. No final de cada sessão, se houvesse qualquer indicação de que a criança daquela sessão ainda estava confusa sobre o assunto, a pessoa que estava conduzindo a interação (o Mágico) fazia com que o robô repetisse a explicação pertinente à dúvida da criança, e se ainda assim ela não entendesse, a pessoa (o Mágico) poderia intervir e dar alguma dica, mas explicar diretamente, deixando este papel para o robô.

5.1.1.3 Continuous Audience Response

Continuous Audience Response consiste em uma audiência, ou comitê de pessoas como juízes, fornecendo a eles o material áudio-visual das sessões para que avaliem a qualidade de

interação entre o robô e as crianças, como forma de medir quantitativamente algo qualitativo. Isto é, como a qualidade da interação do robô com as crianças é algo qualitativo, os juízes deram notas para as interações segundo seu julgamento a fim de se estudar o porquê de algumas interações serem melhor aceitas, ou produzirem melhores reações nas crianças ([VASU; GARSON, 1990](#)).

Dessa forma, foram separados cinco vídeos de cada grupo - baixa e alta interatividade - e os juízes tiveram que analisar a reação da criança em cada um deles, dando uma pontuação de 1 a 5. Cada vídeo corresponde a uma sessão experimental de uma criança interagindo com NAO robô. Apesar de não ser uma boa forma de medir o aprendizado, foi muito útil para compreender a relação entre as pontuações dadas pelos juízes e as respostas das crianças no questionário de avaliação de conteúdo. Na maioria dos casos, a má reação de uma criança em uma interação - o que leva a baixa pontuação - foi precedida por algum erro nas formas 3D solicitadas, ou alguma fala do robô que não foi muito bem compreendida pela criança .

Continuous Audience Response ([STEVENS et al., 2009](#)) é um método de avaliação usado na área de marketing, que vem sendo muito útil para aproximar ações qualitativas em pontuações quantitativas, como as reações das pessoas a algum estímulo ou o quanto boa era uma apresentação de dança. Este método foi utilizado por [Tanaka, Cicourel e Movellan \(2007\)](#) para associar a reação das crianças ao longo de uma sessão com um robô humanoide interativo sem fins pedagógicos. Neste trabalho, esta técnica foi empregada para entender a relação de alguns tipos de interação estarem ligados com um maior número de respostas corretas devido à motivação das crianças, medida pela questionário de avaliação de conteúdo, e o engajamento delas medido pelo MOS.

Duas fases de interação foram empregadas nesta etapa. A primeira é a "fase de boas-vindas"(ausente no grupo de baixa interatividade). A segunda consistiu de três repetições das seguintes tarefas: robô NAO requisitar uma figura geométrica 3D à criança, análise da escolha da criança (Mágico identifica o objeto escolhido) e, finalmente, dar uma resposta à criança (se sua classificação estava correta ou não). A primeira fase era constituída por: Levantar-se, reconhecer o nome das crianças e dar um aperto de mãos, enquanto a segunda é constituída por pedir uma figura, a figura Análise e Positivo/Negativo Resposta do robô.

Foi escolhido um conjunto de 11 pessoas de diferentes áreas, como psicologia, educação matemática, de graduação e pós-graduação, para desempenhar o papel de juízes, analisando e dando pontuações para cada tarefa que o robô fez com as crianças ao longo das sessões. Eles não foram informados sobre a finalidade deste experimento antes de avaliar as sessões, para preservar a imparcialidade sobre as reações das crianças. Como a presença dos juízes poderia intimidar as crianças e afetar seu desempenho nas sessões, foram gravados vídeos de todas as sessões e escolhidos aleatoriamente uma amostra de 5 vídeos de cada grupo.

Os juízes podiam analisar os vídeos em telas individuais ou em grupos com apenas um tela ([Figura 42](#)), de acordo com a sua disposição de tempo. Todos receberam um formulário de pontuação, conforme mostrado na [Figura 43](#), para preencher com suas notas as tarefas

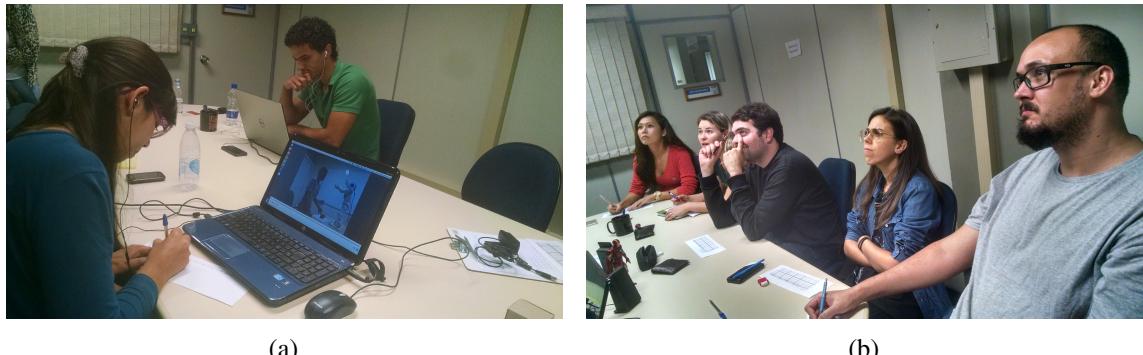


Figura 42 – Juízes avaliando individualmente as sessões 42a e em grupo 42b.

Fonte: Elaborada pelo autor.

de interação que estavam sendo exibidas. As regras para a avaliação foram: eles não podiam retroceder e assistir a qualquer ação novamente, nem falar com outro juiz durante a análise dos vídeo. Estes foram aspectos importantes a serem considerados durante a análise, pois o objetivo principal foi definir o ambiente como se estivessem assistindo as sessões em tempo real. Uma boa interação do robô com a criança significa que essa interação foi bem aceita pela criança, o que gerou uma reação favorável à próximas interações, enquanto uma reação ruim colocou a criança numa situação de desconforto ou fez com que ela perdesse o interesse e/ou empenho nas próximas interações.

Nome: Silvana

Intensidade\Tipo	Ficar em pé	Saudar/ Perguntar Nome	Aperto de Mão	1ª requisição	Análise	Reação sobre a resposta	2ª requisição	Análise	Reação sobre a resposta	3ª requisição	Análise	Reação sobre a resposta	Desvios de atenção
Baixa 1				5	3	4	2	3	1	3	3	4	1
Baixa 2				3	3	2	2	3	3	3	3	3	1
Baixa 3				4	4	4	3	3	2	3	3	4	1
Baixa 4				3	4	4	4	3	4	4	3	2	1
Baixa 5				4	4	5	3	4	3	2	3	4	1
<hr/>													
Alta 1	5	5	5	3	3	5	3	2	4	2	3	4	1
Alta 2	X	4	X	3	3	4	3	4	4	3	3	4	1
Alta 3	2	4	4	2	3	3	3	3	3	3	2	3	1
Alta 4	X	5	5	4	3	2	2	3	3	2	3	3	1
Alta 5	4	4	5	2	3	4	3	3	3	2	3	3	1

Figura 43 – Formulário de pontos dos juízes preenchido.

Fonte: Elaborada pelo autor.

A Tabela 8 mostra as médias de pontuações dos juízes para o grupo de baixa interatividade, enquanto a Tabela 9 mostra os resultados do grupo de alta. As pontuações vão de 1 (muito mau) a 5 (muito bom), no qual 3 é uma reação neutra da criança. É importante notar que, apesar dos juízes não terem sido informados previamente da finalidade do estudo, no final da análise dos vídeos, uma discussão foi realizada com eles esclarecendo os objetivos da pesquisa, comparando os questionários de avaliação de conteúdo das crianças e sugerindo melhorias no trabalho todo. Os valores apresentados nas células correspondem à média da pontuação dos juízes para cada tarefa - apresentada nas linhas - em cada sessão de vídeo gravado - apresentada nas colunas. Por

exemplo, o primeiro valor na Tabela 8, que é 4.11, é a média da pontuação de todos os 11 juízes para tarefa # 1 do primeiro vídeo selecionado para o grupo de baixa interatividade. O mesmo é válido para Tabela 9.

Tabela 8 – Média de pontos dos juízes para as tarefas do grupo de baixa interatividade.

Tarefa\Vídeo	1	2	3	4	5
Requisitar Figura #1	4.11	3.33	3.77	3.77	4.00
Analizar Figura #1	3.22	2.88	3.88	4.11	4.33
Resposta do Robô #1	3.77	2.77	3.44	4.00	4.22
Requisitar Figura #2	3.11	3.00	3.55	3.77	3.88
Analizar Figura #2	3.22	3.11	3.11	3.77	3.88
Resposta do Robô #2	2.55	3.11	3.44	4.00	3.55
Requisitar Figura #3	3.11	3.22	3.22	3.77	3.55
Analizar Figura #3	2.88	2.77	3.55	3.66	3.33
Resposta do Robô #3	3.44	3.33	3.66	3.00	3.66

Fonte: Dados da pesquisa.

Tabela 9 – Média de pontos dos juízes para as tarefas do grupo de alta interatividade.

Tarefa\Video	1	2	3	4	5
Robô Levantar-se	4.67	1.11	2.33	1.33	3.78
Reconhecimento de Nome	4.78	3.56	2.78	4.56	4.22
Aperto de mãos	4.78	1.67	2.78	4.44	4.00
Requisitar Figura #1	3.78	3.56	3.33	3.67	3.33
Analizar Figura #1	4.00	3.56	3.33	3.44	3.56
Resposta do Robô #1	5.00	4.11	3.67	3.00	3.89
Requisitar Figura #2	3.44	3.44	3.67	3.44	3.44
Analizar Figura #2	3.44	3.89	3.44	3.44	3.22
Resposta do Robô #2	4.33	4.44	3.67	3.22	3.56
Requisitar Figura #3	2.67	3.44	3.56	2.89	3.22
Analizar Figura #3	3.22	3.56	3.11	3.00	3.22
Resposta do Robô #3	4.22	4.44	3.67	2.00	3.22

Fonte: Dados da pesquisa.

Pode-se notar que a tarefa que recebeu a menor pontuação foi de 2.55 relativa à resposta do robô, em geral quando uma criança não gostava de receber a resposta de que a figura escolhida não era a requisitada pelo robô. O melhor resultado foi 4.78, na Tabela 9 quando o robô reconhecia o nome da criança mesmo sendo pré-programado com o nome, ou seja, não era um reconhecimento de forma automática como se encontra o sistema em seu estado atual.

As pontuações atribuídas pelos juízes foram tratadas em software e a média de cada tarefa para todos os 5 vídeos estão plotadas nos respectivos gráficos da Figura 44. Estes gráficos trazem no eixo X o número das tarefas e no eixo Y as médias de pontuação, que foi chamado de qualidade de interação das tarefas. O gráfico na Figura 44a representa as médias de tarefas da

Tabela 8, enquanto gráfico da Figura 44b é referente às médias da Tabela 9. Por exemplo, a média de todos os vídeos de alta de interatividade para a tarefa 2 (Figura 44b) é de aproximadamente 4.0, que é a maior pontuação. Os juízes concluíram que as crianças tinham uma melhor reação a tarefa de reconhecimento de nome por ser a tarefa mais pessoal, embora o robô tenha sido teleoperado.

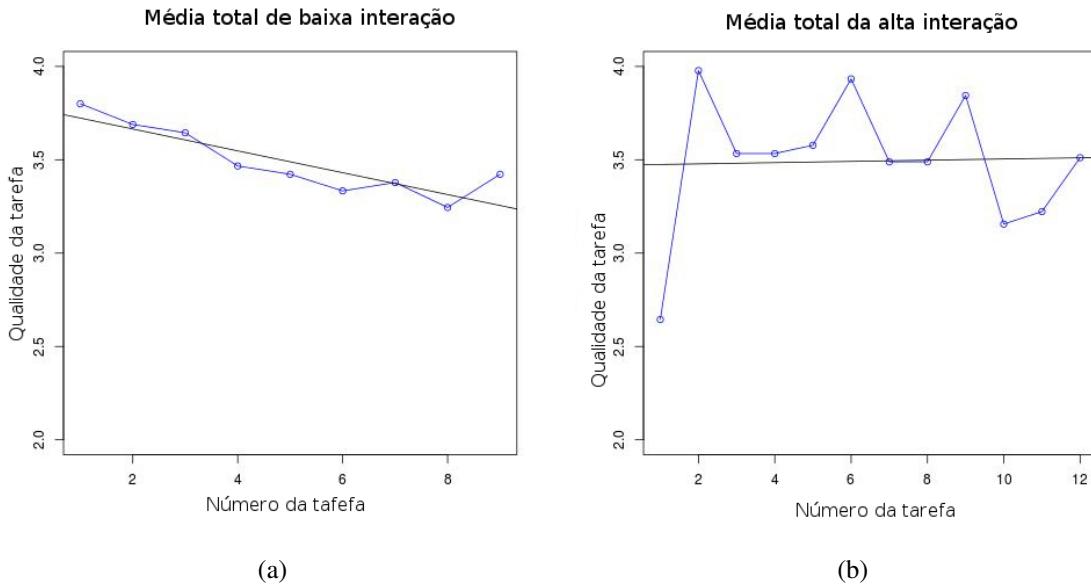


Figura 44 – Média dos pontos dos juízes por tarefa de todos os vídeos para o grupo de baixa 44a e alta 44b interatividade.

Fonte: Elaborada pelo autor.

Pode-se observar que a média diminuiu cerca de 3.8 - 3.5, no grupo de baixa interatividade, enquanto no outro grupo a média é superior a 3.5 em quase todas as tarefas. Os picos apresentados nesta Figura estão associados à mudança de tarefas. Os resultados mostraram que a atenção das crianças foi mantida durante um período mais longo quando a interação era mais dinâmica, o que de fato aconteceu nas sessões de alta interatividade.

Além disso, os juízes contaram quantas vezes eles acharam que houve um desvio de atenção da parte das crianças. Os juízes analisaram os sinais de desvio ocular e sinais de incômodo e tédio. A média total de desvio de atenção do grupo de baixa interatividade foi ligeiramente superior à do grupo de alta: 1.48 (média de vezes que aconteceu o desvio de atenção detectado pelos juízes) no grupo de baixa contra 1.21 no grupo de alta interatividade. Embora não seja uma diferença significativa, este fato encoraja a mais estudos nesta linha de estudos. Ainda que as perguntas para o grupo de alta interatividade fossem mais longas do que as perguntas do grupo de baixa, e mesmo com um conjunto de comportamentos mais dinâmicos, a diminuição na qualidade da interação foi inevitável, devido ao fato de que os comportamentos do robô ficaram mais previsíveis ao longo da sessão.

5.1.1.4 Casos gerais e específicos

No geral, os picos do gráfico na Figura 44b foram associados com respostas certas no questionário de avaliação de conteúdo. Parece óbvio que a presença do robô iria motivar as crianças no processo de aprendizagem, mas esta experiência mostrou que este aumento depende do cenário. No grupo de baixa interatividade, que tinha apenas interações de voz por exemplo, as crianças frequentemente olhavam para outros lugares além do robô e, muitas vezes elas perdiam o foco. Por outro lado, o grupo de alta interatividade mostrou uma dificuldade em se concentrar no assunto abordado pelo robô, devido à distração em outras interações realizadas ou a pela expectativa em relação ao próximo movimento do NAO. Em conclusão, há um equilíbrio entre interação e conteúdo que é a combinação desejável de se alcançar para atingir o tempo máximo de concentração das crianças por um período mais longo e de uma forma que contribua com eficiência para a construção do conhecimento das crianças.

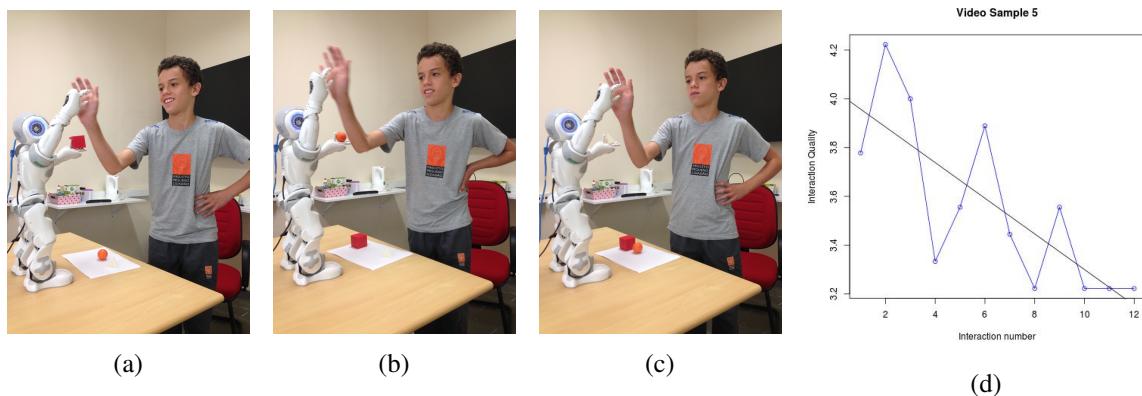


Figura 45 – Decadência do engajamento da criança ao decorrer da sessão de alta interatividade.

Fonte: Elaborada pelo autor.

Apesar dos resultados desta etapa mostrarem que uma maior interatividade aumenta a capacidade de concentração e de melhor assimilação do conteúdo, na Figura 45 é mostrado um exemplo de que devemos ter cuidado nesta abordagem. Na Figura 45a um novo comportamento gera uma boa reação do aluno. O mesmo comportamento na Figura 45b e Figura 45c mostram uma diminuição de concentração do estudante e na qualidade da interação com o robô para essas tarefas (Figura 45d). Outros aspectos de interação, como randomizar o próximo comportamento de robô pode conter a novidade necessária para manter a atenção e concentração das crianças.

Outro caso peculiar, que vale ser citado, é de uma menina do vídeo 2, que não conseguiu executar direito a tarefa de aperto de mão e ficou constrangida, gerando uma reação ruim. Esta tarefa teve uma média de 1.67 na Tabela 9, a média mais baixa entre as médias dessa tarefa. Isso foi detectado e gerou uma adaptação imediata nas interações das tarefas seguintes para corrigir a aversão ao robô criada nessa dessincronia. Os discursos e movimentos do robô foram alterados para mostrar para a menina que aquele erro foi interpretado pelo robô como algo natural, restabelecendo assim uma relação mais confortável no resto da sessão. Apesar desse erro

ter uma má influência nas respostas consequentes a ele, nas últimas tarefas as mudanças no fluxo foram mais eficazes e ela se sentiu mais confiante. Este fato chamou a atenção dos pesquisadores para importância da sensibilidade humana em tais casos. Acredita-se que, sem essas alterações, a garota poderia não ter sido capaz de restabelecer uma boa interação com o robô.

A conclusão para esta etapa é que, para a maioria dos casos, há um equilíbrio entre a interação robô-criança e as ações pedagógicas que detém a atenção das crianças por um período maior. Em casos específicos, algumas interações podem não causar a reação esperada, e devem ser tratadas com cuidado para evitar a distração ou mesmo um trauma do público alvo, como por exemplo uma ação repetitiva que poderia resultar num desvio de atenção, ou um erro no nome da criança que pode gerar uma quebra de expectativa.

Todas as crianças alegaram que a figura de um professor humano é indispensável. Por mais avançada que a tecnologia esteja, a sensibilidade humana na detecção de necessidades individuais ainda é fundamental. No entanto, os resultados obtidos sugerem que a robótica educacional pode ser uma boa ferramenta para auxiliar os professores, proporcionando alguns exercícios práticos, noções de lógica e trabalho em equipe. O aumento dos níveis de interatividade de robôs humanoides trouxe, para este estudo, resultados melhores do que aqueles apresentados por contatos visuais e auditivos com robôs, além de ter mostrado ser uma forma mais divertida para manter as crianças focadas.

5.1.2 Sessões em Grupo

Após o estudo e a análise das dificuldades enfrentadas pelas crianças no assunto escolhido na primeira etapa, foram montadas sessões para serem realizadas em grupo envolvendo todas as crianças para esclarecer seus erros comuns e dúvidas da primeira etapa. Esta etapa não discriminava as crianças dos dois grupos da primeira, mesclando-as em sessões iguais de 3 alunos por vez. Como a maior dificuldade detectada foi a de separar os conceitos de figuras geométricas 2D das 3D, algumas figuras 2D foram adicionadas ao inventário bem como outras figuras 3D também.

Cada sessão contava com 3 fases. Na primeira, o robô perguntava se as crianças lembravam o que tinham acertado e o que tinham errado na primeira etapa. Na segunda, o robô revisava os conceitos e apresentava as novas formas, explicando a diferença entre elas, que, diferentemente do que as crianças comumente pensavam, o triângulo e o quadrado não eram a mesma coisa que a pirâmide e o cubo, e explicava um pouco mais sobre como aumentar uma dimensão, de 2D para 3D, alterava algumas taxonomias e agregava mais medidas, como o volume e soma das áreas. Finalmente, na última das três fases o robô propunha 3 exercícios para as crianças: separar as figuras da mesa em grupos de dimensões semelhantes; separar as figuras da mesa por bases semelhantes e assimilar figuras com as mesmas formas, com medidas diferentes (por exemplo o paralelepípedo e o cubo, ou a pirâmide de base triangular e a de base quadrada. Como de costume, ao final da sessão, o robô parabenizava as crianças pelo esforço

delas e as avisava para se preparam para o próximo encontro, no qual as crianças seriam testadas pelo robô com um jogo de perguntas. A todo momento o robô incentivava e pedia pra que todas as crianças da sessão participassem.

As sessões foram gravadas para serem analisadas futuramente. No geral, as crianças tiveram um desempenho bom. Alegaram estarem mais tranquilas por sentirem que o robô estava tentando ajudá-las mais nessa etapa. Foi notório uma maior participação das crianças do grupo de alta interatividade da primeira etapa.

Como os resultados dessa etapa foram estritamente subjetivos, é inviável tirar conclusões sobre como elas reagiram ou o quanto ficaram atentas durante essas sessões. Porém, como será discutido na subseção seguinte, a avaliação da próxima etapa mostrou que essas sessões foram esclarecedoras para as crianças.

5.1.3 Jogo de Perguntas

Com o apoio dos mesmos especialistas em educação matemática da etapa de sessões interativas iniciais, foi proposto um jogo como uma metodologia de avaliação para investigar o quanto as crianças foram capazes de guardar do conteúdo das sessões anteriores. Chamamos isso "Acerte a pergunta para fazer o NAO feliz", sugerindo o que as crianças envolvidas deveriam fazer para ganhar: acertar as perguntas do robô para deixá-lo feliz.

Este jogo tinha 3 fases: Na primeira, o operador do robô (o Mágico) escolhia aleatoriamente uma pergunta para o robô fazer para as crianças de nível fácil. Em seguida, na segunda fase, uma questão de nível de dificuldade média, e, finalmente, para a última fase, uma questão de nível difícil. Em todas as fases, se a resposta das crianças estivesse correta, o robô se levantava um pouco mais e demonstrava alguns sinais de felicidade, como levantar as mãos pra cima para comemorar ou piscar os Leds dos olhos em azul mais rápido, ao passo que, em caso de resposta errada, o robô encolhia suas pernas e demonstrava algum sinal de decepção, como abaixar a cabeça ou piscar lentamente os Leds dos olhos em vermelho, mas em ambos os casos o jogo seguia para a próxima fase. Além disso, cada vez que acertavam a resposta, as crianças ganhavam um "ponto de felicidade" e cada resposta errada dada eles ganhavam um "ponto de tristeza", que anulava um "ponto de felicidade", podendo resultar em um saldo negativo na felicidade do robô.

Em outras palavras, quanto mais ereto e esticado o robô estivesse, mais feliz ele estaria e, consequentemente, mais pontuação a criança estaria alcançando naquele momento, enquanto mais deitado e encolhido o robô se encontrasse, menos pontuações a criança possuía naquele momento. Para as crianças, o objetivo principal do jogo era terminar as três fases com o robô em um estado feliz, isso significa, com um ou mais pontos de saldo de felicidade. Por exemplo, caso as crianças acertassem as duas primeiras questões - primeira e segunda fases consecutivas - e errassem a questão da última fase, as crianças ganhariam (2 pontos de alegria menos 1 ponto de tristeza, com saldo de 1). Assim, o saldo foi de 2 em estado feliz do robô e ganhou o jogo. No

entanto, errando duas fases quaisquer, mesmo acertando a outra, o seu saldo seria negativo, e assim, perderiam o jogo.

No jogo, o robô fazia questões em três níveis de dificuldade, sendo elas:

- Perguntas Fáceis:

1. Coloque na minha mão duas figuras geométricas 3D de bases diferentes.
2. Me dê um cubo e um paralelepípedo e me explique a diferença entre eles.
3. Coloque duas figuras geométricas 3D em minhas mãos e me diga seus respectivos nomes.

- Perguntas de dificuldade média:

1. Me dê 2 figuras geométricas 3D similares, porém com bases diferentes.
2. Me mostre uma figura geométrica com 15 arestas.
3. Me dê uma figura geométrica 3D e uma 2D e compare suas diferenças.

- Perguntas Difíceis:

1. Me dê um cubo e um quadrado e me explique a diferença entre eles.
2. Me dê um figura de 8 arestas e uma figura com 7 faces.
3. Explique as diferenças e relações entre figuras geométricas 2D e 3D e me dê um exemplo de cada.

O propósito desse experimento foi tentar medir o quanto as crianças tinham memorizado do conteúdo visto nas etapas anteriores por meio do jogo proposto, e verificar se houve alguma diferença para as crianças que eram do grupo de alta interatividade comparada com as outras. Além de o robô interagir com as crianças, o estado do NAO servia como um medidor para que elas soubessem como estava seu saldo. No final, dependendo do estado do NAO, a criança que estava jogando era consolada ou parabenizada pelo NAO com uma música triste ou feliz.

O NAO começava o jogo sentado com suas mãos em seus joelhos, como mostrado na Figura 46, e recebia as crianças com uma mensagem de boas-vindas, seguida de uma explicação sobre o jogo. Depois disso, o jogo seguia com as três perguntas conforme descrito anteriormente.

5.1.3.1 MOS

Nesta etapa também foi utilizado o MOS para uma avaliação subjetiva da opinião das crianças. As perguntas eram as mesmas das sessões interativas iniciais e a Tabela 10 apresenta os resultados.



Figura 46 – NAO esperando pelas crianças no jogo de perguntas com as figuras geométricas.

Fonte: Elaborada pelo autor.

1. As perguntas do robô foram muito difíceis (5) ou muito fáceis (1)?
2. O robô conversou muito (5) ou pouco (1) com você?
3. Você entendeu muito bem (5) ou muito mal (1) o que o robô disse?
4. Você julga que aprendeu muito (5) ou pouco (1) com essa atividade?

Tabela 10 – Média obtida com o MOS para o jogo de perguntas.

Questão	Média
1	4.2
2	4.5
3	3.4
4	5

Fonte: Dados da pesquisa.

Nas sessões interativas iniciais, as crianças do grupo de baixa interatividade alegaram que o robô parecia rude. Agora, o robô fingia reconhecer seus rostos, dizendo "Olá" e agitando suas mãos quando a criança chegava. Todas as crianças gostaram do "toque pessoal" desta etapa, e disseram terem-se sentido mais confortáveis para conversar com o robô.

A menina que teve uma má impressão no primeiro contato com o NAO, descrito na subseção 5.1.1.4, conseguiu acertar 2 das perguntas do jogo proposto. No final, ela queria tirar fotos com o robô, e pediu para voltar nas próximas vezes.

5.1.3.2 Resposta por grupo

Na Tabela 11 é mostrada a média de acertos para as perguntas do jogo, separando os grupos de alta e baixa interatividade da primeira etapa. Para o nível médio de dificuldade o número de acertos foi o mesmo, embora as perguntas tenham sido escolhidas aleatoriamente. Entretanto, o desempenho das crianças do grupo de alta interatividade nas perguntas fáceis e difíceis foi, respectivamente, 17% e 32% superior, comparado com o grupo de baixa interatividade. As crianças que interagiram mais com o robô se sentiram mais confortáveis e seguras em suas respostas com relação às crianças do outro grupo. Também afirmaram se sentir mais desafiadas por seu amigo robô (como eles chamavam o NAO), e 60% deles disseram que estudaram em casa para o jogo, contra 20% do outro grupo.

Tabela 11 – Porcentagem das respostas corretas das crianças.

Dificuldade	Grupo de Interatividade	
	Baixa (%)	Alta (%)
Fácil	65	82
Médio	40	40
Difícil	50	82

Fonte: Dados da pesquisa.

5.1.4 Considerações Finais

Essa seção apresentou os resultados dos experimentos envolvendo crianças e o robô NAO, utilizando técnica de WoZ. O estudo foi dividido em 3 etapas: sessões interativas iniciais, sessões em grupo e jogo de perguntas.

A Tabela 12 replica as diretrizes de Riek (2012) (Tabela 3) e, embora essas respostas estejam implícitas no texto da seção anterior, a Tabela 13 traz as respostas para essas diretrizes deste estudo.

5.2 Testes dos Modulos

Foi optado por não testar o Módulo de Voz, apenas o Módulo de Visão e o funcionamento do sistema completo foram testados. Os resultados para os testes do sistema de visão isolado encontram-se na subseção 5.2.1 e os do sistema completo na encontram-se na subseção 5.2.2.

Tabela 12 – Diretrizes propostas por Riek (2012) para estudos de HRI por meio de WoZ.

Fonte: Adaptada de Riek (2012).

Componente Experimental	Questões
A) Robô	1) Quantos robôs foram usados? 2) Qual(ais) tipo(s) de robô(s)? (ex: Humanoide, zoomorfo, androide, etc) 3) Qual o nível de autonomia? (Quais componentes eram do robô eram autônomos e quais eram controlados pelo Mágico?) 4) Quais eram as habilidades do robô? 5) Quais hipóteses os pesquisadores tem para com o robô?
B) Usuário	1) Quantos usuários participaram no total, e por bateria de experimentos? 2) Qual a formação dos participantes? 3) Quais instruções foram fornecidas aos usuários? 4) Quais hipóteses os pesquisadores tem sobre os usuários? 5) A simulação foi convincente para os usuários? 6) Quais expectativas os usuários tiveram com relação ao robô antes e depois do experimento?
C) Mágico	1) Quantos Mágicos foram utilizados? 2) Quais foram os dados demográficos do Mágico? (ex: o pesquisador, colegas de laboratório, terceiros?) 3) O Mágico sabia das hipóteses comportamentais do experimento? 4) Quais foram as variáveis de reação do Mágico e como elas foram controladas? 5) Quais foram as variáveis de reconhecimento do Mágico e como elas foram controladas? 6) Como foi tratado o controle de experimento para o erro do Mágico (deliberativo ou acidental?) 7) Quanto e de qual tipo de treinamento o Mágico recebeu a priori para conduzir o experimento?
D) Geral	1) Aonde o experimento foi realizado? 2) Quais foram as variáveis do ambiente e como elas foram controladas? 3) Quais cenários os cenários empregados pelos pesquisadores? 4) Este experimento foi parte de um processo de design? 5) Este estudo discute as limitações do WoZ?

5.2.1 Testes em base de dados para o sistema de visão

De modo a obter melhores medidas em relação à eficácia na classificação correta dos objetos, o módulo de visão foi submetido a uma série de testes. Na detecção, o VOCUS2 foi bem sucedido em detectar todas as amostras de teste, demonstrando que ele se adapta muito bem à esta aplicação, removendo o fundo e isolando o objeto.

Deve-se notar que, as imagens originais poderiam também servir como boas entradas para treinamento e teste, mas como o projeto final visa trabalhar em ambientes ruidosos e reconhecer objetos mais complexos do que os usados neste trabalho, foi decidido executar o sistema completo, mesmo com estes objetos simples, para estudar o seu comportamento como um todo. Foram treinadas três SVM, um para cada classe de figuras geométricas 3D, segundo

Tabela 13 – Repostas para as diretrizes de Riek (2012) para os experimentos deste estudo.

Fonte: Elaborada pelo autor.

Componente Experimental	Questões
A) Robô	1) Apenas 1 robô, o NAO 2) Humanoide 3) Nenhuma autonomia. Todos os componentes controlados pelo Mágico. 4) Interagir com as crianças reconhecendo figuras geométricas. 5) É possível criar um comportamento autônomo para as interações previstas.
B) Usuário	1) 30 no total, variando o numero por etapa nas sessões 2) Fundamental incompleta 3) Que respondessem à requisição do robô. 4) Usuários se motivam mais quando o sistema é mais interativo. 5) A simulação foi convincente para a maioria dos usuários. 6) As expectativas variavam de acordo com o grupo de interatividade.
C) Mágico	1) Apenas 1 Mágico 2) O pesquisador deste trabalho 3) O Mágico sabia das hipóteses comportamentais do experimento. 4) Variáveis foram controladas empiricamente. 5) Variáveis foram controladas empiricamente. 6) Erros apenas acidentais 7) Não houve treinamento para o Mágico
Geral	1) Na sala de reuniões do bloco 6 do ICMC. 2) Foram amenizadas com artifícios do mágico. 3) Interação para construção de conhecimento do usuário. 4) Este experimento foi parte de um projeto maior. 5) Este estudo não tem foco no WoZ.

descrito na subseção 4.2.1.

Para um teste de desempenho da técnica de classificação visual em um banco de dados conhecido, verifique (SHUKLA; MISHRA; SHARMA, 2013), para a classificação de um banco de dados de veículos, cheque (MONTANARI *et al.*, 2015). Já para esta aplicação, o sistema foi testado com um conjunto de dados de teste composto por 90 amostras distintas do conjunto de dados de treinamento, sendo 30 imagens de cada classe e variando as suas posições no campo de visão do robô. Na Tabela 14 é mostrada a matriz de confusão para a anotação automática do conjunto de dados de teste. As linhas são as classes que a amostra pertence e as colunas são as previsões feitas pela SVM. Por fim, a classe "desconhecida" na última coluna indica que nenhum dos SVM foram capazes de prever a amostra, ou seja, que a amostra não foi reconhecida como pertencente de nenhuma das classes da múltipla SVM.

Metade das figuras foram posicionadas na faixa intermediária do campo de visão do robô e a outra metade espalhadas pela periferia. Todos os objetos no meio foram previstos corretamente, enquanto que os erros de previsão foram feitos para os objetos mais distantes do centro do campo de visão do robô.

Tabela 14 – Matriz de confusão para o classificador.

	Cubo	Pirâmide	Esfera	Desconhecido
Cubo	26	0	0	4
Pirâmide	0	28	0	2
Esfera	1	0	29	0

Fonte: Dados da pesquisa.

A medida de *recall* é o número de previsões corretas dividido pelo número de ocorrências de previsões tidas como corretas, mesmo que equivocadamente. Em outras palavras, *recall* de qualquer classificador é calculado dividindo os pontos positivos corretamente classificados pela contagem total de positivos das imagens que estão sendo testadas (OLSON; DELEN, 2008). A medida de *precision* é o número de anotações corretas dividido pelo número de anotações previstas. Em outras palavras, é o número de imagens corretamente recuperadas dividido pelo número de imagens automaticamente recuperadas (JEON; LAVRENKO; MANMATHA, 2003). Para combinar *recall* e *precision* em uma única medida de eficiência, a média harmônica de precisão e recuperação é calculada. Ele é chamado de *F-measure* (Equação 5.1). Esta é uma das medidas de desempenho agregados.

$$F_measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \quad (5.1)$$

Estas três medidas são apresentadas na Tabela 15 e no gráfico da Figura 47.

Tabela 15 – Medidas do classificador.

	Recall	Precision	F-Measure
Cubo	0.96	0.86	0.91
Pirâmide	1	0.93	0.96
Esfera	1	0.96	0.98

Fonte: Dados da pesquisa.

5.2.2 Testes do sistema

Para testar o sistema completo, membros do laboratório de Aprendizado de Robôs foram convidados para desempenharem o papel de usuários do sistema em sessões de múltiplos objetos 3D (subseção 4.3.2). Embora os participantes já possuíssem domínio no conteúdo das sessões, estes testes serviram para diagnosticar as limitações imediatas na interação com o sistema autônomo.

Nestes testes, cada usuário participou de um bateria de 5 interações. Os respectivos resultados são apresentados nas Tabelas 16, 17, 18 e 19. A primeira coluna de cada tabela é o numero da sessão. Os tipos de erros são relativos aos erros dos participantes ou aos módulos

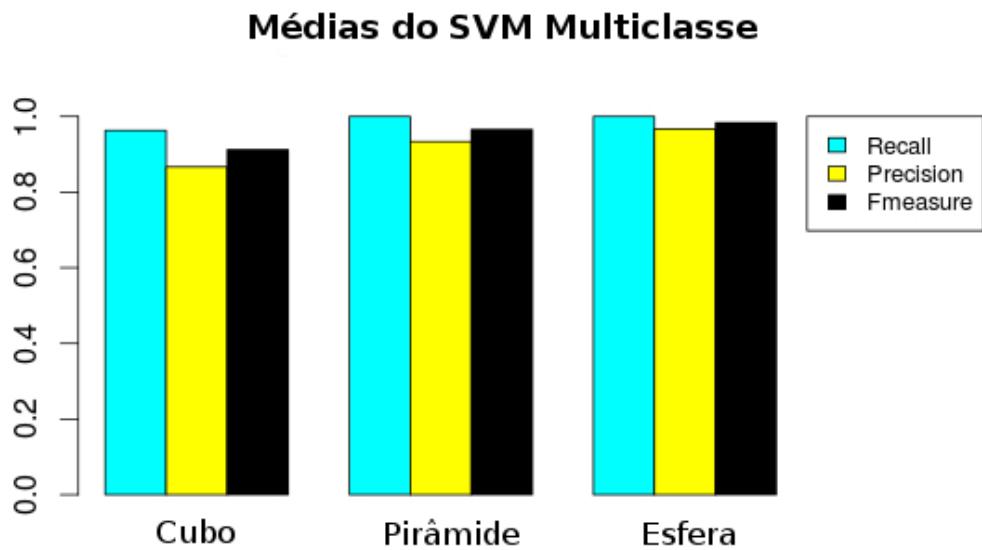


Figura 47 – Recall, Precision e F-measure agrupados por classe dos objetos.

Fonte: Elaborada pelo autor.

que falharam. Por exemplo, se o usuário perder o tempo de realizar alguma interação esse erro é computado na coluna "Usuário", se o Módulo de Diálogo falhar no reconhecimento de voz, esse erro é contabilizado na coluna do tipo de erro de voz e o mesmo serve para a coluna de tipo de erro de visão. A última linha corresponde ao total de erros por tipo de erro, a última coluna o total de erros por sessão e a última célula por sua vez corresponde ao total de erros da bateria de sessões para o respectivo usuário.

As três primeiras tabelas (16, 17 e 18) correspondem a testes positivos, que consistem em fornecer sempre entradas válidas para o sistema, enquanto que a última tabela corresponde a um teste negativo, no qual são fornecidas entradas inesperadas pelo sistema na primeira vez, porém como a sessão exige entradas corretas para se encerrar, as segundas tentativas para essa bateria de testes eram entradas esperadas.

O primeiro teste positivo e o teste negativo (Tabela 16 e 19) foram realizados por um usuário que conhecia bem a implementação do sistema e consequentemente suas limitações, como falar mais alto para o microfone capturar bem sua voz e posicionar as figuras bem ao centro do campo de visão do robô. O segundo teste positivo contava com um usuário que acompanhou parcialmente a implementação do sistema e o terceiro teste positivo por um usuário que desconhecia as funcionalidades do sistema a priori.

É válido lembrar que no teste negativo no qual as entradas eram fornecidas erradamente, o erro contabilizado ainda é referente ao sistema. Isto quer dizer que, mesmo o usuário colocando os objetos em ordem errada ou fornecendo ao sistema uma palavra fora de contexto e o sistema

Tipo de erro				
#	Usuário	Voz	Visão	Total
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
Total	0	0	0	0

Tabela 16 – Teste positivo por um usuário com bom conhecimento da implementação.

Tipo de erro				
#	Usuário	Voz	Visão	Total
1	1	1	0	2
2	0	0	0	0
3	1	0	0	1
4	0	0	0	0
5	0	0	0	0
Total	2	1	0	3

Tabela 17 – Teste positivo por um usuário com médio conhecimento da implementação.

Tipo de erro				
#	Usuário	Voz	Visão	Total
1	4	1	0	5
2	2	0	0	2
3	2	0	0	2
4	0	0	0	0
5	0	0	0	0
Total	8	1	0	9

Tabela 18 – Teste positivo por um usuário com nenhum conhecimento da implementação.

Tipo de erro				
#	Usuário	Voz	Visão	Total
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
Total	0	0	0	0

Tabela 19 – Teste negativo por um usuário com bom conhecimento da implementação.

Fonte: Dados da pesquisa.

reconhecer este erro, o sistema não falhou, pois a falha acontece quando essas entradas são interpretadas como válidas. Também, os erros de usuário não são computados nesse caso pois ele estava intencionado a errar para por o sistema a prova.

Como esperado, os resultados mostraram que um maior conhecimento da implementação do sistema gera menos erros durante uma sessão. Porém, a calibração do usuário é feita de forma rápida, conforme podemos notar que os erros diminuem ao longo das sessões para o usuário de nenhum conhecimento prévio das funcionalidades. Os erros foram cometidos em sua maioria por parte do usuário e não houveram erros para detecção e classificação de objetos nessa etapa de testes. Um resultado interessante foi que o sistema mostrou-se robusto para casos negativos, estando bem equipado para contornar casos como o mais comum do usuário falar muito baixo e o sistema não conseguir capturar sua voz. Quando isso acontecia o robô dizia que não tinha entendido e pedia para o usuário repetir um pouco mais alto e pausadamente.

5.2.3 Considerações Finais

O sistema de visão computacional implementado mostrou-se adequado para o projeto, uma vez que foi obtida uma acurácia de 93% de precisão e tempo de treinamento em 0.9s para as figuras geométricas 3D escolhidas. Já o sistema completo mostrou que, apesar de alguns ajustes poderem melhorar os erros dos usuários, encontra-se estável para testes com usuários finais, as crianças em fase de aprendizagem de figuras geométricas 3D.

CAPÍTULO

6

CONCLUSÃO

Este trabalho descreve a implementação de um sistema interativo que conduz sessões pedagógicas de forma autônoma, empregando técnicas de robótica social, detecção de objetos, classificação de imagens, reconhecimento e síntese de fala e recursos do robô humanoide NAO.

Foi realizada uma revisão bibliográfica com os trabalhos mais próximos para pesquisar os métodos que viabilizaram esta proposta e, em seguida, esses métodos foram estudados e encorporados ao projeto. Dentre tais métodos destacam-se um sistema de visão computacional para detecção de objetos, o VOCUS2; um classificador multiclasse SVM para classificação de imagens monoculares (anotação automática de imagens) baseado em *Bag-of-Features* e descritores SURF; o reconhecedor de fala *Google Voice Recognize*; o sintetizador de fala do robô humanoide NAO; e o método *Wizard-of-Oz* HRI para analisar e definir as melhores formas de interagir com os usuários ao longo das sessões.

Não houveram testes no reconhecedor e sintetizador de fala por serem softwares já consolidados e fornecidos por suas respectivas fabricantes. Porém, eles atenderam satisfatoriamente aos testes finais deste projeto.

Resultados na detecção e classificação de objetos

O sistema de visão computacional implementado mostrou-se adequado para o projeto, uma vez que foi obtida uma acurácia de 93% de precisão e tempo de treinamento em 0.9s para as figuras geométricas 3D escolhidas.

Resultados dos estudos interativos

As técnicas para estudo de interação dos usuários foram aplicadas em alunos de 10-14 anos das escolas públicas de São Carlos. As crianças foram divididas em dois grupos: um que apresentava baixa interatividade no contato com o robô, enquanto o outro apresentava alta interatividade. Esses estudos mostraram que a alternância no comportamento do robô segura a concentração dos alunos por mais tempo, pois gera mais expectativa nos participantes. Intercalar

conteúdos pedagógicos abordados pelo robô com interações físicas e mais descontraídas, fez com que as crianças memorizassem mais informações sobre o assunto da sessão. Quando comparados, os usuários que tiveram em contato com uma maior interatividade do robô se sentiram mais motivados a estudar para as próximas sessões em relação aos usuários que participaram de sessões com menor grau interatividade. Em outras palavras, o grupo de usuários que o robô utilizou de mais recursos próprios para interagir, por um período de tempo maior nas tarefas e simulando um comportamento mais humanizado (mostrando sinais de felicidade ou tristeza, perguntando pelo nome, cumprimentando, etc.) se sentiu mais a vontade e com um engajamento maior nos experimentos que o grupo em o robô não apresentava tal arsenal de interação.

Resultados do sistema completo

Um resultado interessante foi que o sistema mostrou-se robusto para casos negativos, estando bem equipado para contornar casos como o mais comum do usuário falar muito baixo e o sistema não conseguir capturar sua voz. Quando isso acontecia o robô dizia que não tinha compreendido e pedia para o usuário repetir um pouco mais alto e pausadamente.

6.1 Desafios e limitações

O maior desafio desta pesquisa foi a interdisciplinaridade que sua proposta traz. Empregar os principais métodos das duas áreas, pedagogia e tecnológica, que iniciaram sua fusão recentemente exige domínio e controle em ambas as áreas. Dessa forma, é essencial que especialistas de cada área estejam sempre presentes para bons resultados na aplicação final.

A parte de implementação também apresenta uma certa dificuldade, pois dispõe de muitos parâmetros em cada parte para serem regulados, oferecendo pouco controle do sistema como um todo em princípio. Porém, após os testes iniciais para regulagem dos módulos e da integração entre eles, o sistema mostrou-se estável para as sessões propostas.

O fator mais limitante deste projeto é o reconhecimento de voz, pois necessita de uma conexão com a Internet e foi onde os usuários mais tiveram dificuldades. O robô escolhido, o NAO, apesar de já fazer parte dos materiais de pesquisa do laboratório no qual esse trabalho foi desenvolvido, é um robô de alto custo financeiro, podendo ser um limitador para eventuais replicas dos experimentos aqui apresentados. Entretanto, a implementação em módulos visa a troca desse robô por qualquer outro, como uma maneira de resolver essa limitação.

6.2 Trabalhos futuros

Os trabalhos futuros para esse sistema podem ser separados para melhorias em duas formas: na implementação e na aplicação.

Implementação

A implementação pode ser melhorada comparando outras técnicas que ofereçam resposta mais rápidas no reconhecimento de voz, testar outros parâmetros para o classificador de imagens e buscar por robôs financeiramente mais acessíveis.

Aplicação

Os próximos passos para a aplicação deste trabalho é testar as sessões prontas no público alvo com o conteúdo já configurado e também testar o sistema para novos conteúdos, trocando sua base de dados e configurações do conteúdo nos módulos.

Outra sugestão é automatizar a contabilidade de erros, gerando relatórios detalhado das falhas do sistema completo, a fim de identificar quais pontos estão com maiores limitações e se for em casos de erros por parte dos usuários, diagnosticar suas raízes, sugerindo melhorias na abordagem do conteúdo de forma automática.

6.3 Lista de artigos gerados e publicados

Já publicados

TOZADORE, D. C. ; ROMERO, R. A. F. . Inserção de um Robô humanoide no Ensino de Figuras Geométricas 3D. In: I Workshop em pesquisas robóticas de São Carlos, 2015, São Carlos -SP.

PINTO, A. H. M.; TOZADORE, D. C. ; ROMERO, R. A. F. . A Question Game for Children Aiming the Geometrical Figures Learning by Using a Humanoid Robot.. In: : SBRLARS Robotics Symposium and Robot control (SBR LARS), 2015, Uberlândia - MG. Anais LARC, 2015.

MONTANARI, R. ; FRACCAROLI, E. S. ; TOZADORE, D. C. ; ROMERO, R. A. F. . Ground vehicle detection and classification by an unmanned aerial vehicle. In: : SBRLARS Robotics Symposium and Robot control (SBR LARS), 2015, Uberlândia - MG. Anais LARC, 2015.

MONTANARI, R. ; TOZADORE, D. C. ; FRACCAROLI, E. S. ; BENICASA, A.X. ; ROMERO, R. A. F. . A visual attention approach for the tracking of vehicles through UAV. In: XI Workshop de Visão Computacional (WVC2015), 2015, São Carlos. Anais do WVC2015, 2015. p. 1-6.

Submetidos

ROMERO, R. A. F.; MENEGHETTI, R. C. G.; TOZADORE, D. C.; PINTO, A. H. M. . A Pedagogical Approach for Teaching 3D Geometrical Figures incorporating a humanoid robot. Submitted to Computers & Education – Journal – Elsevier in October 2015.

TOZADORE, D. C. ; PINTO, A. H. M.; ROMERO, R. A. F. . Object Classification by SURF descriptor and SVM Applied in a Humanoid Robot for Pedagogical Approaches.

Submitted to The annual International Joint Conference on Neural Networks (IJCNN) 2016.

TOZADORE, D. C. ; PINTO, A. H. M.; ROMERO, R. A. F. . Variations in an Interactive Humanoid Robot to Analyze Changes in Children's Attention Span. To be submitted.

REFERÊNCIAS

- ALDEBARAN. 2014. Disponível em: <http://trends.directindustry.com/products/this-pint-sized-robot-plays-soccer-but-youll-never-guess-what-else-it-can-do/>. Citado 3 vezes nas páginas 27, 87 e 88.
- BALAYIL, G. S. K. M.; ANEES, V. M. Automatic multilabelling of images and semantic relation extraction. 2014. Citado na página 36.
- BAY, H.; ESS, A.; TUYTELAARS, T.; GOOL, L. van. Speeded-up robust features (surf). **Computer Vision and Image Understanding (CVIU)**, v. 110, n. 3, p. 346–359, June 2008. Citado na página 69.
- BAY, H.; TUYTELAARS, T.; GOOL, L. V. Surf: Speeded up robust features. In: **Computer vision–ECCV 2006**. [S.l.]: Springer, 2006. p. 404–417. Citado na página 36.
- BAY, H.; TUYTELAARS, T.; GOOL, L. V. Surf: Speeded up robust features. In: **In ECCV**. [S.l.: s.n.], 2006. p. 404–417. Citado 4 vezes nas páginas 65, 68, 71 e 72.
- BEAUCHAMP, C. Revolution industrielle et croissance économique au xix siecle. In: . [S.l.: s.n.], 1998. p. 70. Citado na página 23.
- BENICASA, A.; QUILES, M.; ZHAO, L.; ROMERO, R. Top-down biasing and modulation for object-based visual attention. In: LEE, M.; HIROSE, A.; HOU, Z.-G.; KIL, R. (Ed.). **Neural Information Processing**. Springer Berlin Heidelberg, 2013, (Lecture Notes in Computer Science, v. 8228). p. 325–332. ISBN 978-3-642-42050-4. Disponível em: <http://dx.doi.org/10.1007/978-3-642-42051-1_41>. Citado na página 54.
- BENICASA, A. X. **Sistemas computacionais para atenção visual Top-Down e Bottom-up usando redes neurais artificiais**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos, 9 2013. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-29042014-162209/>>. Citado 3 vezes nas páginas 15, 58 e 59.
- BENICASA, A. X.; QUILES, M.; ZHAO, L.; ROMERO, R. Top-down biasing and modulation for object-based visual attention. In: LEE, M.; HIROSE, A.; HOU, Z.-G.; KIL, R. (Ed.). **Neural Information Processing**. Springer Berlin Heidelberg, 2013, (Lecture Notes in Computer Science, v. 8228). p. 325–332. ISBN 978-3-642-42050-4. Disponível em: <http://dx.doi.org/10.1007/978-3-642-42051-1_41>. Citado na página 27.
- BENITTI, F. B. V. Exploring the educational potential of robotics in schools: A systematic review. **Computers & Education**, Elsevier, v. 58, n. 3, p. 978–988, 2012. Citado na página 29.
- BEZDEK, J. C.; PAL, D.; K., S. **Fuzzy models for pattern recognition: methods that search for structures in data**. 1992. New York: IEEE Press. Citado na página 33.
- BLANZ, V.; SCHOLKOPF, B.; BULTHOFF, H.; BURGES, C.; VAPNIK, V.; VETTER, T. **Comparison of view-based object recognition algorithms using realistic 3d models**. 1996.

In C. von der Malsburg, W. von Seelen, J. C. Vorbruggen, and B. Sendhoff, editors, *Artificial Neural Networks — ICANN*, pages 251 - 256, Berlin, 1996. Springer Lecture Notes in Computer Science, Vol. 1112. Citado na página 74.

BOHNING, G.; ALTHOUSE, J. K. Using tangrams to teach geometry to young children. **Early childhood education journal**, Springer, v. 24, n. 4, p. 239–242, 1997. Citado na página 31.

BORJI, A.; ITTI, L. **State-of-the-art in visual attention modeling**. 2013. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1), 185-207. Citado na página 47.

BOTTERILL, T.; MILLS, S.; GREEN, R. Speeded-up Bag-of-Words algorithm for robot localisation through scene recognition. In: **Image and Vision Computing New Zealand**. [S.l.: s.n.], 2008. p. 1–6. Citado na página 71.

BRAGA, A. de P.; LUDERMIR, T. B.; CARVALHO, A. C. P. de L. F. **Redes Neurais Artificiais: Teoria e aplicacoes**. LTC. 2000. Citado na página 62.

BREAZEAL, C.; KIDD, C. D.; THOMAZ, A. L.; HOFFMAN, G.; BERLIN, M. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In: **IEEE. Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on**. [S.l.], 2005. p. 708–713. Citado na página 24.

BROEKENS, J.; HEERINK, M.; ROSENDAL, H. Assistive social robots in elderly care: a review. **Gerontechnology**, v. 8, n. 2, p. 94–103, 2009. Citado na página 26.

BROWN, M.; LOWE, D. G. Invariant features from interest point groups. In: **BMVC**. [S.l.: s.n.], 2002. Citado na página 70.

BURT, P. J.; ADELSON, E. H. The laplacian pyramid as a compact image code. **Communications, IEEE Transactions on**, IEEE, v. 31, n. 4, p. 532–540, 1983. Citado 2 vezes nas páginas 50 e 52.

CHAKRABORTY, A. K.; PAL, D.; CHATTERJEE, P. **Fast Recognition of Mechanical Objects Using Neural Networks Under Robust Aspect**. 2012. Journal of The Institution of Engineers (India) vol. 93(1), pp. 55-62. Citado na página 33.

CHAPELLE, O.; HAFFNER, P.; VAPNIK, V. N. Support vector machines for histogram-based image classification. **Neural Networks, IEEE Transactions on**, IEEE, v. 10, n. 5, p. 1055–1064, 1999. Citado na página 35.

CHAVEZ, G. C. **Sistema Celular Evolutivo para Reconhecimento de Padrao Invariante**. 2002. Universidade de Sao Paulo. Citado na página 34.

CHELLA, M. T. **Ambiente de robótica educacional com Logo**. [s.n.], 2005. Disponível em: <www.Nied.unicamp.br/~siros/doc/artigo_sbc2002_wie_final.pdf>. Citado 3 vezes nas páginas 26, 31 e 32.

CHOKSURIWONG, A.; LAURENT, H.; EMILE, B. **Comparsion of invariant descriptors for object recognition**. 2005. In: IEE International Conference On Image Processing. pp. 377-380. Citado na página 35.

COLL, C. **Aprendizagem escolar e construção do conhecimento**. 1994. Porto Alegre: Artes Médicas Sul. Citado na página 31.

- CONCI, A.; AZEVEDO, E.; LETA, F. R. **Computacao Grafica: Teoria e Pratica. v2, Rio de Janeiro: Campus.** 2008. Citado na página 35.
- CORTES, S.; VAPNIK, V. **Support Vector Machines.** 1995. Machine Learning, 20:273–297. Citado na página 74.
- CRANDALL, J. W.; GOODRICH, M. A.; JR, D. R. O.; NIELSEN, C. W. Validating human-robot interaction schemes in multitasking environments. **Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on**, IEEE, v. 35, n. 4, p. 438–449, 2005. Citado na página 25.
- CSALA, E.; NEMETH, G.; ZAINKO, C. Application of the nao humanoid robot in the treatment of marrow-transplanted children. **International Conference on Cognitive Infocommunications (CogInfoCom)**, p. pp. 655–659, 2012. Citado na página 38.
- DAUTENHAHN, K.; WALTERS, M.; WOODS, S.; KOAY, K. L.; NEHANIV, C. L.; SISBOT, A.; ALAMI, R.; SIMÉON, T. How may i serve you?: a robot companion approaching a seated person in a helping context. In: ACM. **Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction.** [S.l.], 2006. p. 172–179. Citado na página 25.
- D'ABREU, J. V. V.; BASTOS, B. L. Robótica pedagógica: Uma reflexão sobre a apropriação de professores da escola elza maria pellegrini de aguiar. **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**, Campinas, SP, Brasil, 2013. Citado na página 32.
- D'ABREU, J. V. V.; FILHO, M. L. Núcleo de informática aplicada à educação (nied)-30 anos de atuação (1983–2013). In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação.** [S.l.: s.n.], 2013. v. 1, n. 1. Citado na página 33.
- EGMONT-PETERSEN, M.; RIDDER, D. de; HANDELS, H. **Image Processing with Neural Networks - a review.** 2002. Pattern Recognition (35), 2279-2301. Citado na página 34.
- ENDSLEY, M. R. **Designing for situation awareness: An approach to user-centered design.** [S.l.]: CRC press, 2011. Citado na página 25.
- FENWICK, I.; RICE, M. Reliability of continuos measurement copy-test- methods. **Journal of Advertising Research**, Journal of Advertising Research, v. 507, n. 28, p. 15422–1529, 1991. Citado na página 40.
- FIGUEIRA-SAMPAIO, A. da S.; SANTOS, E. E. F. dos; CARRIJO, G. A.; CARDOSO, A. Survey of mathematics practices with concrete materials used in brazilian schools. **Procedia-Social and Behavioral Sciences**, Elsevier, v. 93, p. 151–157, 2013. Citado na página 32.
- FIX, E.; HODGES, J. **Discriminatory analysis. Nonparametric discrimination: Consistency properties.** [S.l.], 1951. Citado na página 68.
- FLETCHER, R. **Practical Methods of Optimization.** 1987. John Wiley and Sons, Inc., 2nd edition. Citado na página 75.
- FOUNDATION, P. S. 2014. Available in: <https://pypi.python.org/pypi/SpeechRecognition/>. Citado na página 80.
- FRASER, N. M.; GILBERT, G. N. Simulating speech systems. **Computer Speech & Language**, Elsevier, v. 5, n. 1, p. 81–99, 1991. Citado na página 44.

- FRAWLEY, W. **Vigotsky e a ciência cognitiva: linguagem das mentes social e computacional.** 2000. Porto Alegre: Artes Médicas Sul. Citado na página 31.
- FRINTROP, S.; WERNER, T.; GARCIA, G. M. Traditional saliency reloaded: A good old model in new shape. In: . [S.l.: s.n.], 2015. p. 82–90. Citado 4 vezes nas páginas 54, 55, 56 e 57.
- GARCIA, M. F. **O Ensino por Meio da Pesquisa: O projeto Ciência na Escola.** Tese (Doutorado) — FE-UNICAMP, Campinas, SP, Brasil, 2002. Citado na página 28.
- GASPAR, T. L. **Reconhecimento de faces humanas usando redes neurais MLP.** 2006. Dissertacao, Escola de Engenharia de Sao Carlos, USP. Citado na página 23.
- GEMAQUE, R. M. L.; SALES, E. R. Geometria para alunos surdos por meio do tangram. In: **Congresso Internacional de Educação e Inclusão.** [S.l.: s.n.], 2013. p. 82–90. Citado na página 32.
- GONZALEZ, R. C.; WOODS, R. E. **Procesamento de Imagens Digitais.** 2002. Sao Paulo-SP: Edgard Blucher LTDA. Citado na página 35.
- GOODRICH, M. A.; SCHULTZ, A. C. Human-robot interaction: a survey. **Foundations and trends in human-computer interaction**, Now Publishers Inc., v. 1, n. 3, p. 203–275, 2007. Citado 2 vezes nas páginas 24 e 26.
- GREEN, A.; HUTTENRAUCH, H.; EKLUNDH, K. S. Applying the wizard-of-oz framework to cooperative service discovery and configuration. In: IEEE. **Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on.** [S.l.], 2004. p. 575–580. Citado na página 44.
- GREENSPAN, H.; BELONGIE, S.; GOODMAN, R.; PERONA, P.; RAKSHIT, S.; ANDERSON, C. H. Overcomplete steerable pyramid filters and rotation invariance. In: IEEE. **Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on.** [S.l.], 1994. p. 222–228. Citado na página 51.
- GRICE, H. P. **Studies in the Way of Words.** [S.l.]: Harvard University Press, 1991. Citado 2 vezes nas páginas 25 e 97.
- HALKIN, S. **Artifical Neural Networks: Methods and Application.** [S.l.]: Bookman, 2001. Citado 4 vezes nas páginas 23, 60, 61 e 62.
- HTTP://ENSINAREVT.COM/JOGOS/TANGRAM. **América do Sul.** Fevereiro de 2016. Disponível em: <<http://ensinarevt.com/jogos/tangram/>>. Acesso em: 10/02/2016. Citado na página 32.
- ITTI, L.; KOCH, C. A saliency-based search mechanism for overt and covert shifts of visual attention. **Vision research**, Elsevier, v. 40, n. 10, p. 1489–1506, 2000. Citado na página 54.
- ITTI, L.; KOCH, C.; NIEBUR, E. A model of saliency-based visual attention for rapid scene analysis. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE Computer Society, v. 20, n. 11, p. 1254–1259, 1998. Citado 4 vezes nas páginas 50, 51, 54 e 55.
- _____. _____. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, IEEE, n. 11, p. 1254–1259, 1998. Citado na página 52.

- ITTII, L.; KOCH, C. Computational modelling of visual attention. **Nature reviews neuroscience**, Nature Publishing Group, v. 2, n. 3, p. 194–203, 2001. Citado 4 vezes nas páginas 47, 48, 49 e 50.
- IYENGAR, S.; KASHYAP, R. **Neural networks: a computational perspective**. 1991. In ANTOGNETTI, Paolo, MILUTINOVIC, Veljko. Neural networks: concepts, applications, and implementations. Vol II, pp. 1-30. Englewood Cliffs: Prentice Hall. Citado na página 33.
- JACOBSON, I. Object oriented software engineering: a use case driven approach. Addison-Wesley Professional, 1992. Citado na página 79.
- JAIN, A.; DUIN, R.; MAO, J. **Statistical pattern recognition: A review**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v.22, n.1, p. 4-37. 2000. Citado na página 23.
- JEON, J.; LAVRENKO, V.; MANMATHA, R. Automatic image annotation and retrieval using cross-media relevance models. In: ACM. **Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval**. [S.I.], 2003. p. 119–126. Citado na página 116.
- JOHNSTON, J. H.; FIORE, S. M.; PARIS, C.; SMITH, C. **Application of cognitive load theory to developing a measure of team decision efficiency**. [S.I.], 2002. Citado na página 25.
- KANDA, T.; SATO, R.; SAIWAKI, N.; ISHIGURO, H. Friendly social robot that understands human's friendly relationships. In: IEEE. **Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on**. [S.I.], 2004. v. 3, p. 2215–2222. Citado na página 26.
- KANDEL, E. R.; SCHWARTZ, J. H.; JESSELL, T. M. **Fundamentos da Neurociencia e do Comportamento**. Guanabara Koogan. 1997. Citado 2 vezes nas páginas 58 e 60.
- KELLEY, J. F. An iterative design methodology for user-friendly natural language office information applications. **ACM Transactions on Information Systems (TOIS)**, ACM, v. 2, n. 1, p. 26–41, 1984. Citado 4 vezes nas páginas 25, 43, 45 e 98.
- KIMBERLEE, J.; KING, M.; HELLERSTETH, S.; WIREN, A.; MULLIGAN, H. Feasibility of using a humanoid robot enhancing attention and social skills in adolescents with autism spectrum disorder. **International Journal of Rehabilitation Research**, 2013. Citado na página 38.
- KIRBY, R.; FORLIZZI, J.; SIMMONS, R. Affective social robots. **Robotics and Autonomous Systems**, Elsevier, v. 58, n. 3, p. 322–332, 2010. Citado na página 23.
- KLEIN, G.; FELTOVICH, P. J.; BRADSHAW, J. M.; WOODS, D. D. Common ground and coordination in joint activity. **Organizational simulation**, Hoboken, NJ, USA: Wiley, v. 53, 2005. Citado na página 25.
- KOCH, C.; ULLMAN, S. Shifts in selective visual attention: towards the underlying neural circuitry. In: **Matters of intelligence**. [S.I.]: Springer, 1987. p. 115–141. Citado 4 vezes nas páginas 48, 49, 53 e 54.
- KOENDERINK, J. J. The structure of images. **Biological cybernetics**, Springer, v. 50, n. 5, p. 363–370, 1984. Citado na página 69.

- KOENEMANN, J.; BENNEWITZ, M. **Whole-body imitation of human motions with a nao humanoid.** 2012. Presented at the Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, Boston, Massachusetts, USA. Citado na página 37.
- KOHONEN, T. **Self-Organizing Maps (3th Edition ed.).** 2001. Citado na página 51.
- KUENZI, J. J. Science, technology, engineering, and mathematics (stem) education: Background, federal policy, and legislative action. 2008. Citado na página 29.
- KURWEIL, R. **The age of spiritual machines.** Dissertação (Mestrado) — MIT Press, 1990. Citado na página 23.
- LEE, J.; LEE, J. O.; COLLINS, D. Enhancing children's spatial sense using tangrams. **Childhood Education**, Taylor & Francis, v. 86, n. 2, p. 92–94, 2009. Citado na página 31.
- LI, X.; SNOEK, C. G.; WORRING, M. Learning social tag relevance by neighbor voting. **Multimedia, IEEE Transactions on**, IEEE, v. 11, n. 7, p. 1310–1322, 2009. Citado na página 36.
- LINDBLAD, T. K. J. **Image Processing Using Pulse-Coupled Neural Network.** 2005. Secaucus, NJ, USA: Springer. Citado na página 34.
- LINDEBERG, T. Scale-space for discrete signals. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, IEEE, v. 12, n. 3, p. 234–254, 1990. Citado na página 69.
- LIU, G.; YANG, J. **Image retrieval based on the texton co-occurrence matrix.** 2008. Pattern Recognition. Citado na página 35.
- LLOYD, S. P. Least squares quantization in pcm. **Information Theory, IEEE Transactions on**, IEEE, v. 28, n. 2, p. 129–137, 1982. Citado na página 73.
- LONG, F.; ZHANG, H.; FENG, D. **Fundamentals of Content-based Image Retrieval.** 2005. Multimedia Information Retrieval and Management: Technological Fundamentals and Applications (Signals and Communication Technology), pp.1-26. Citado na página 35.
- LOWE, D. G. Distinctive image features from scale-invariant keypoints. **Int. J. Comput. Vision**, Kluwer Academic Publishers, Hingham, MA, USA, v. 60, n. 2, p. 91–110, nov. 2004. ISSN 0920-5691. Disponível em: <<http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94>>. Citado 6 vezes nas páginas 65, 66, 67, 68, 69 e 70.
- LUGER, G. F. **Artificial Intelligence - Structures and Strategies for Complex Problem Solving.** [S.1.]: Pearson Education, 2004. Citado na página 23.
- LUXBURG, U.; SCHOLKOPF, B. **Statistical Learning Theory: Models, Concepts, and Results.** 2008. E-print, arxiv.org. URL <http://arxiv.org/abs/0810.4752>. Citado 3 vezes nas páginas 74, 75 e 76.
- MACHADO, A. **Neuroanatomia Funcional (2.Ed. ed.).** Rio de Janeiro: Atheneu. 2000. Citado na página 58.
- MILL, D.; LIMA, D. A. E.; LIMA, V. S.; TANCREDI, R. O desafio de uma interação de qualidade na educação a distância: o tutor e sua importância nesse processo. **Cadernos da Pedagogia, Ano**, v. 2, 2008. Citado na página 31.

- MILLER, K. W. It's not nice to fool humans. **IT professional**, IEEE, n. 1, p. 51–52, 2010. Citado na página 24.
- MONTANARI, R.; FRACCAROLI, E. C.; TOZADORE, D. C.; ROMERO, R. A. F. **Ground vehicle detection and classification by an unmanned aerial vehicle**. 2015. LARS/SBR 2015, 2015, Uberlândia - MG. Proceedings of LARS/SBR 2015, 2015. p. 253-257. Citado na página 115.
- MOORE, A. **A tutorial on kd-trees**. [S.l.], 1991. Citado na página 72.
- MURALIDHARAN, R.; CHANDRASEKAR, C. Scale invariant feature extraction for identifying an object in the image using moment invariants. **Proceedings of the International Conference on Communication and Computational Intelligence**, vol. 27, p. pp. 452–456, 2010. Citado na página 35.
- NAKAUCHI, Y.; SIMMONS, R. A social robot that stands in line. **Autonomous Robots**, Springer, v. 12, n. 3, p. 313–324, 2002. Citado na página 26.
- OBR. **Olimpíada Brasileira de Robótica**. 2014. Disponível em: <www.obr.org.br/>. Citado 2 vezes nas páginas 26 e 27.
- OGAWA, T.; KOMATSU, H. **Target selection in area v4 during a multidimensional visual search task**. 2004. Journal of Neuroscience 24(28), 6371-6382. Citado na página 48.
- OLSON, D. L.; DELEN, D. **Advanced data mining techniques**. [S.l.]: Springer Science & Business Media, 2008. Citado na página 116.
- ONG, S. H.; YEO, N. C.; LEE, K. H.; VENKATESH, Y. V.; CAO, D. M. **Segmentation of color images using a two-stage self-organizing network**. 2002. Image and Vision Computing (20), 63-68. Citado na página 34.
- PAPERT, S. **A máquina das crianças: repensando a escola na era da informática**. 1994. Porto Alegre: Artes Médicas. Citado na página 31.
- PEDRINI, H.; SCHWARTZ, W. R. **Analise de Imagens Digitais: Princípios, Algoritmos e Aplicações**. 2008. São Paulo: Ed. Thomson. Citado na página 35.
- PERRENOUD, P. **Construir as competências desde a escola**. 1999. Porto Alegre: Artes Médicas Sul. Citado 2 vezes nas páginas 31 e 32.
- PIAGET, J. **Seis Estudos de Psicologia**. [S.l.]: Forense Universitária, 1998. Citado na página 31.
- PINTO, A. H. M.; BENICASA, A. X.; OLIVEIRA, L. O.; MENEGUETTI, R. C. G.; ROMERO, R. A. F. **Attention Based Object Recognition applied to a Humanoid Robot**. 2014. LARS/SBR 2014, 2014, São Carlos. Proceedings of LARS/SBR 2014, 2014. p. 136-141. Citado 7 vezes nas páginas 15, 17, 27, 28, 40, 43 e 65.
- PINTO, A. H. M.; OLIVEIRA, L. O. de; BENICASA, A. X.; MENEGHETTI, R. C. G.; ROMERO, R. A. F. Inserção de um robô humanoide no ensino de objetos geométricos 2d sobrepostos. In: **Anais do Simpósio Brasileiro de Informática na Educação**. [S.l.: s.n.], 2014. v. 25, n. 1, p. 632–641. Citado 3 vezes nas páginas 15, 40 e 42.

PINTO, A. H. M.; TOZADORE, D. C.; ROMERO, R. A. F. **A Question Game for Children Aiming the Geometrical Figures Learning by Using a Humanoid Robot.** 2015. LARS/SBR 2015, 2015, Uberlândia - MG. Proceedings of LARS/SBR 2015, 2015. p. 253-257. Citado 3 vezes nas páginas 90, 91 e 93.

POT, E.; MONCEAUX, J.; GELIN, R.; MAISONNIER, B. Chorographe: a graphical tool for humanoid robot programming. In: IEEE. **Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on.** [S.I.], 2009. p. 46–51. Citado 2 vezes nas páginas 87 e 90.

QUILES, M. G. **Sistema de visão baseado em redes neurais para o controle de robôs móveis.** Dissertação (Mestrado) — Universidade de São Paulo, 2004. Citado 2 vezes nas páginas 34 e 48.

QUILES, M. G.; ROMERO, R. A. F.; ZHAO, L. **A pulse-coupled neural network as a simplified bottom-up visual attention model.** 2006. In IEEE Proceedings of the Ninth Brazilian Symposium on Artificial Neural Networks (SBRN'2006), pp. 1-6. Citado na página 34.

REC, I. **P.800: Methods for subjective determination of transmission quality. CCITT Recommendations; electronic version.** Geneva: IUT, 2000. (Recommendations). Citado na página 98.

RICH, E. C.; CROWSON, T. W.; HARRIS, I. B. The diagnostic value of the medical history: perceptions of internal medicine physicians. **Archives of internal medicine**, American Medical Association, v. 147, n. 11, p. 1957–1960, 1987. Citado na página 24.

RICK, L. D.; WATSON, R. N. The age of avatar realism when seeing shouldn't be believing. **IEEE robotics & automation magazine**, v. 17, n. 4, p. 37, 2010. Citado na página 24.

RIEK, L. D. Wizard of oz studies in hri: a systematic review and new reporting guidelines. **Journal of Human-Robot Interaction**, v. 1, n. 1, 2012. Citado 8 vezes nas páginas 17, 25, 44, 45, 46, 113, 114 e 115.

ROSTEN, E.; DRUMMOND, T. Machine learning for high-speed corner detection. In: **European Conference on Computer Vision.** [s.n.], 2006. v. 1, p. 430–443. Disponível em: <http://edwardrostens.com/work/rostens_2006_machine.pdf>. Citado na página 65.

RUBLEE, E.; RABAUD, V.; KONOLIGE, K.; BRADSKI, G. Orb: An efficient alternative to sift or surf. In: **International Conference on Computer Vision.** Barcelona: [s.n.], 2011. Citado na página 65.

SALTON, G.; MCGILL, M. **Introduction to modern information retrieval.** McGraw-Hill, 1983. (McGraw-Hill computer science series). ISBN 9780070544840. Disponível em: <<http://books.google.com.br/books?id=7f5TAAAAMAAJ>>. Citado 2 vezes nas páginas 72 e 73.

SATAKE, S.; KANDA, T.; GLAS, D. F.; IMAI, M.; ISHIGURO, H.; HAGITA, N. How to approach humans?-strategies for social robots to initiate interaction. In: IEEE. **Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on.** [S.I.], 2009. p. 109–116. Citado na página 26.

SCHEEFF, M.; PINTO, J.; RAHARDJA, K.; SNIBBE, S.; TOW, R. Experiences with sparky, a social robot. In: **Socially Intelligent Agents.** [S.I.]: Springer, 2002. p. 173–180. Citado na página 26.

- SCHOLKOPF, B.; SMOLA, A. **Learning with Kernels**. 2002. MIT Press, Cambridge, M. Citado na página 75.
- SHAMSUDDIN, S.; YUSSOF, H.; ISMAIL, L.; HANAPIAH, F. A.; MOHAMED, S.; PIAH, H. A.; ZAHARI, N. I. Initial response of autistic children in human-robot interaction therapy with humanoid robot nao. In: IEEE. **Signal Processing and its Applications (CSPA), 2012 IEEE 8th International Colloquium on**. [S.l.], 2012. p. 188–193. Citado na página 38.
- SHAMSUDDIN, S.; YUSSOF, H.; ISMAIL, L.; HANAPIAH, F. A.; MOHAMED, S.; PIAH, H. A.; ZAHARI, N. I. Initial response of autistic children in human-robot interaction therapy with humanoid robot nao. **International Colloquium on Signal Processing and its Applications**, p. pp. 188–193, 2012. Citado na página 38.
- SHERIDAN, T. B. **Humans and automation: System design and research issues**. [S.l.]: John Wiley & Sons, Inc., 2002. Citado na página 25.
- SHUKLA, T.; MISHRA, N.; SHARMA, S. Automatic image annotation using surf features. **International Journal of Computer Applications**, Foundation of Computer Science, v. 68, n. 4, 2013. Citado 3 vezes nas páginas 15, 36 e 115.
- SMOLAR, P.; TUHARSKY, J.; FEDOR, Z.; VIRCIKOVA, M.; SINCAK, P. Development of cognitive capabilities for robot nao in center for intelligent technologies in kosice. In: IEEE. **Cognitive Infocommunications (CogInfoCom), 2011 2nd International Conference on**. [S.l.], 2011. p. 1–5. Citado na página 38.
- STEINFELD, A.; JENKINS, O. C.; SCASSELLATI, B. The oz of wizard: simulating the human for interaction research. In: IEEE. **Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on**. [S.l.], 2009. p. 101–107. Citado na página 45.
- STEVENS, C. J.; SCHUBERT, E.; MORRIS, R. H.; FREAR, M.; CHEN, J.; HEALEY, S.; SCHOKNECHT, C.; HANSEN, S. Cognition and the temporal arts: Investigating audience response to dance using pdas that record continuous data during live performance. **International Journal of Human-Computer Studies**, Elsevier, v. 67, n. 9, p. 800–813, 2009. Citado 2 vezes nas páginas 101 e 104.
- SUAY, H. B.; CHERNOVA, S. **Humanoid robot control using depth camera**. 2011. Presented at the Proceedings of the 6th international conference on Human-robot interaction, Lausanne, Switzerland. Citado na página 37.
- TANAKA, F.; CICOUREL, A.; MOVELLAN, J. R. Socialization between toddlers and robots at an early childhood education center. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 104, n. 46, p. 17954–17958, 2007. Citado 4 vezes nas páginas 15, 39, 41 e 104.
- TANAKA, F.; GHOSH, M. The implementation of care-receiving robot at an english learning school for children. In: IEEE. **Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on**. [S.l.], 2011. p. 265–266. Citado 4 vezes nas páginas 15, 17, 38 e 39.
- TAPUS, A.; PECA, A.; ALY, A.; POP, C.; JISA, L.; PINTEA, S.; RUSU, A.; DAVID, D. Childrens with autism social engagement in interaction with nao, an imitative robot: A series of single case experiments. **John Benjamins Publishing Company**, v. 13, n. 3, p. 315–347, 2012. Citado na página 37.

- TRAFTON, J. G.; SCHULTZ, A. C.; PERZNOWSKI, D.; BUGAJSKA, M. D.; ADAMS, W.; CASSIMATIS, N. L.; BROCK, D. P. Children and robots learning to play hide and seek. In: ACM. **Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction.** [S.I.], 2006. p. 242–249. Citado na página 25.
- TSOTOS, J. K. **A computational perspective on visual attention.** 2011. MIT Press, Cambridge. Citado na página 47.
- TSOTSOS, J. K.; CULHANE, S. M.; WAI, W. Y. K.; LAI, Y.; DAVIS, N.; NUFLÓ, F. Modeling visual attention via selective tuning. **Artificial intelligence**, Elsevier, v. 78, n. 1, p. 507–545, 1995. Citado na página 54.
- TZAFESTAS, C.; MITSOU, N.; GEORGAKARAKOS, N.; DIAMANTI, O.; MARAGOS, P.; FOTINEA, S. E.; EFTHIMIOU, E. Gestural teleoperation of a mobile robot based on visual recognition of sign language static handshapes. **IEEE International Symposium on Robot and Human Interactive Communication**, p. pp. 1073–1079, 2009. Citado na página 34.
- VAKSER, I. A.; MATAR, O. G.; LAM, C. F. A systematic study of low-resolution recognition in protein–protein complexes. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 96, n. 15, p. 8477–8482, 1999. Citado na página 23.
- VASU, M. L.; GARSON, G. D. Computer-assisted survey research and continuous audience response technology for the political and social sciences. **Social Science Computer Review**, Sage Publications, v. 8, n. 4, p. 535–557, 1990. Citado na página 104.
- VELTKAMP, R. C.; LATECKI, L. J. **Properties and performance of shape similarity measures.** 2006. In: Internationa Conference on Data Science and Classification, pp. 1-9. Citado na página 35.
- VELTROP, T. 2012. Disponível em: <http://taylor.veltrop.com/>. Citado na página 37.
- VERNER, I. M.; POLISHUK, A.; KLEIN, Y.; CUPERMAN, D.; MIR, R.; WERTHEIM, I. A learning excellence program in a science museum as a pathway into robotics. **International Journal of Engineering Education**, v. 28, n. 3, p. 523, 2012. Citado na página 38.
- VIGOTTSKY, L. S. **Pensamento e linguagem.** 1993. Martins Fontes, São Paulo. Citado na página 31.
- WALDHERR, S.; THRUN, S.; ROMERO, R. A. F. **A gesture-based interface for human-robot interaction.** 2000. Autonomous Robots 9(2), 151-173. Citado na página 34.
- WALTER, W. G. An imitation of life. **Scientific American**, v. 182, p. 42 – 45, 1950 1950. Citado na página 26.
- WANG, S. **The Unpredictability of Standard Back Propagation Neural Networks in Classification Applications.** 1995. Management Science, vol. 41: pp. 555-559. Citado 3 vezes nas páginas 33, 63 e 64.
- WEISS, A. **Validation of an evaluation framework for human-robot interaction: the impact of usability, social acceptance, user experience, and societal impact on collaboration with humanoid robots.** [S.I.]: na, 2010. Citado na página 44.
- ZHANG, D.; LU, G. **Review of shape representation and description techniques.** 2004. Patter Recognition, v.37, n.1, p. 1-19. Citado na página 35.

ZUHER, F.; ROMERO, R. A. F. **Whole-body imitation and recognition of human motions through a humanoid robot.** 2012. In: SBR/LARS - Simpósio Brasileiro de Robótica/Latin American Robotics Symposium, 2012, Fortaleza - CE. Anais do SBR, 2012. p. 190-195. Citado na página [37](#).