RECONHECIMENTO ATIVO DE PEQUENOS OBJETOS PELA FUSÃO DE DADOS AUDIOVISUAIS

SAMUEL SÉRVULO JACINTO DE OLIVEIRA

RECONHECIMENTO ATIVO DE PEQUENOS OBJETOS PELA FUSÃO DE DADOS AUDIOVISUAIS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Mario Fernando Montenegro Campos
Coorientadora: Izabela Lyon Freire

Belo Horizonte
Abril de 2015

Sérvulo Jacinto de Oliveira, Samuel

O48r Reconhecimento ativo de pequenos objetos pela fusão de dados audiovisuais / Samuel Sérvulo Jacinto de Oliveira. — Belo Horizonte, 2015
xxii, 88 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Mario Fernando Montenegro Campos
Coorientadora: Izabela Lyon Freire

1. Computação - Teses. 2. Processamento de sinais - Teses. 3. Visão por computador - Teses. I. Orientador. II. Coorientadora. III. Título.

CDU 519.6*82.10(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Reconhecimento ativo de pequenos objetos pela fusão de dados audiovisuais

SAMUEL SÉRVULO JACINTO DE OLIVEIRA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. MÁRIO FERNANDO MONTENEGRO CAMPOS - Orientador
Departamento de Ciência da Computação - UFMG

Izabela Lyon Freire

PROFA. IZABELA LYON FREIRE - Coorientadora
Departamento de Ciência da Computação - UFMG

PROF. HANI CAMILLE YEHIA
Departamento de Engenharia Eletrônica - UFMG

Douglas Macharet

PROF. DOUGLAS GUIMARÃES MACHARET
Departamento de Ciência da Computação - UFMG

PROF. ERICKSON RANGEL DO NASCIMENTO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 24 de abril de 2015.

Agradecimentos

Muitas pessoas foram fundamentais para que eu tenha chegado até aqui.

À minha família, em especial minha mãe, agradeço o amor e apoio incondicionais ao longo de toda essa jornada. Sempre.

Ao meu orientador, prof. Mario, agradeço pela oportunidade e pelo aprendizado inestimáveis que me proporcionou.

À minha coorientadora, profa. Izabela, que foi uma feliz surpresa na reta final, pelas sugestões e ideias valiosas quando a luz no fim do túnel parecia um pontinho distante.

À Júlia (com acento e carinho), que tornou doces e leves esses anos, pelos sorrisos e planos B.

Aos meus colegas de apartamento: Veneza, Borges e Seu Francisco, que proporcionaram risadas, histórias e experiências que vou levar para a vida toda.

Aos amigos do Ceará, Rogério (meu grande companheiro de aventuras), Levi e Vladimir, pelas conversas, apoio emocional e desabafos na Guinness Pizza.

Aos colegas de laboratório, em ordem alfabética, pra não parecer que tenho favoritos ;) : Anderson, Balbino, David, Elerson, Hector, Igor, Omar, Paulo, Rafael, Ramon, Vinícius e professores Chaimo, Douglas e Erickson.

Por fim, agradeço aos meus companheiros das noites insones que sempre estiveram lá: obrigado, Red Bull e Milkyway.

Termino este agradecimento com um grande obrigado a todos e um desejo secreto de que tenha usado todas às crases corretamente. :)

"Se você está triste, não fique triste!"
(Vovó Sérvulo)

Resumo

Robôs frequentemente precisam reconhecer objetos de uso comum: em uso doméstico, tarefas de busca e salvamento ou sistemas de vigilância. Essa habilidade fundamentalmente requer que informações sensoriais sejam processadas e representadas da melhor forma possível, a fim de maximizar o desempenho do robô. Este trabalho apresenta uma abordagem de percepção ativa para reconhecimento de objetos utilizando estímulos de áudio e vídeo, captados por sensores montados em um robô, que utiliza uma haste articulada para interagir com o objeto.

Para fins experimentais, um conjunto estruturado de pequenos objetos foi adotado, em que geometrias simples e composição de único material são adotadas a fim de facilitar a compreensão das assinaturas de áudio. Para cada combinação de geometria e material, uma assinatura de audiovisual é desenvolvida em uma abordagem de aprendizado de máquina que implementa fusão de sensores.

O desempenho do reconhecimento é avaliado para os sinais originais e níveis de ruídos decrescentes nos sinais de áudio e vídeo, onde duas estratégias para a fusão de sensores são avaliadas comparativamente: fusão de decisões, em uma abordagem de meta-aprendizado, e fusão de atributos. É mostrado que a fusão de decisões tem o melhor desempenho e destaca-se quando comparada ao uso do individual de áudio ou vídeo, com taxas de acerto de 99,4%, 96,2%, e 91,6%, respectivamente, melhorando o reconhecimento e proporcionando estabilidade em cenários de alta interferência. Os descritores de áudio introduzidos são ordenados de acordo com seu poder discriminatório.

Contribuições deste trabalho incluem a avaliação de técnicas de representação de sinais impulsivos, um arcabouço para fusão audiovisual e a publicação da base de dados.

Palavras-chave: Reconhecimento de objetos, Fusão de Sensores, Visão Computacional, Processamento de Sinais.

Abstract

Robots routinely face the need to recognize common use objects: for domestic use, search and rescue tasks or surveillance systems. This ability fundamentally requires them to process sensory information and best represent it, in order to maximize performance. This work presents an active perception approach to object recognition using both audio and visual stimuli, acquired by sensors mounted on a robot, which uses an articulated rod to poke the object in order to actively generate audio signatures.

The object domain consists of a structured set of small objects, in which simple geometries and single-material compositions are adopted in order to make it easier to achieve an understanding of audio signatures. For each combination of geometry and material composition, an audiovisual signature is developed in a machine learning approach that implements the sensor fusion.

Performance of classification is evaluated for the original signals and for decreasing signal-to-noise ratio of the audio signals, where two strategies for sensor fusion are evaluated: decision fusion in a meta-learning manner, and feature fusion. Decision fusion achieves the highest accuracy and improves over audio- or video-only classification, with accuracies of 99,4%, 96,2%, and 91,6%, respectively, enhancing recognition and providing stability over high interference scenarios. The audio descriptors introduced are ranked according to their discriminatory power.

The contributions of this work include evaluation of techniques for representation of impulsive signals, a framework for audiovisual fusion and the publication of dataset.

Keywords: Object Recognition, Sensor Fusion, Computer Vision, Signal Processing.

Lista de Figuras

Lista de Tabelas

Lista de acrônimos

AUC	*Area Under ROC Curve*
BASE	*Binary Appearance and geometrical Shape Elements*
BER	*Band-Energy Ratio*
BFCC	*Bark Frequency Cepstral Coefficients*
BoW	*Bag-of-Words*
BRIEF	*Binary Robust Independent Elementary Features*
BRISK	*Binary Robust Invariant Scalable Keypoints*
DFT	*Discrete Fourier Transform*
CWT	*Continuous Wavelet Transform*
DNN-HMM	*Deep Neural Network - Hidden Markov Model*
FAST	*Features from Accelerated Segment Test*
FREAK	*Fast REtinA Keypoint*
FT	*Fourier Transform*
GMM-HMM	*Gaussian Mixture Model - Hidden Markov Model*
HMM	*Hidden Markov Model*
LDA	*Latent Dirichlet Allocation*
LPC	*Linear Predictive Coding*
MAP	Maximum a Posteriori
MFCC	*Mel Frequency Cepstral Coefficients*

MPE	*Minimum Phone Error*
ML	*Maximum Likelihood*
ORB	*Oriented FAST and Rotated BRIEF*
PAM	*Partition Around Medoids*
PCA	*Principal Component Analysis*
PLSA	*Probabilistic Latent Semantic Analysis*
RASTA-PLP	*Relative Spectral Transform - Perceptual Linear Prediction*
ROC	*Receiver Operating Characteristic*
SAC	*SAmpled Consensus initial alignment*
SIFT	*Scale Invariant Feature Transform*
SNR	*Signal-to-noise Ratio*
SOM	*Self Organizing Map*
SPM	*Spatial Pyramid Matching*
STFT	*Short-Time Fourier Transform*
SURF	*Speeded Up Robust Features*
SVM	*Support Vector Machine*
WT	*Wavelet Transform*
TSP	*Time-Stretched Pulse*

Sumário

Capítulo 1

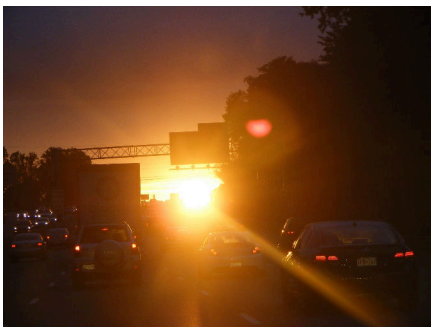
Introdução

Na natureza, interpretar, abstrair e gerar respostas a estímulos são algumas das tarefas mais importantes e complexas executadas pelo cérebro. Para tanto, uma gama de órgãos sensoriais observada em várias espécies se desenvolveu ao longo da evolução, tal que a percepção multissensorial configura-se como mecanismo presente em seres mais adaptados ao meio [Stein & Meredith, 1993].

A capacidade de perceber o ambiente através de diferentes modalidades sensoriais permite que haja comportamento complementar entre elas: se uma única modalidade não é suficiente para prover uma boa estimativa sobre algo, informações das demais podem ser combinadas em uma estimativa mais robusta [Ernst & Bülthoff, 2004]. Assim, é possível superar interferências ou condições do meio que afetam certos sentidos, mas que inalteram outra modalidade sensorial. Um exemplo desta complementaridade ocorre com o ser humano: se pouca informação for obtida usando apenas a visão, por problemas como como luminosidade baixa ou ofuscante, ainda é possível perceber o ambiente utilizando a audição, tato ou olfato, auxiliando na percepção do ambiente. Exemplos de interferências comumente encontradas na visão são ilustrados na Figura 1.1.

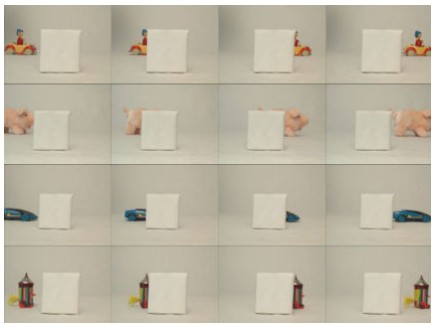
Se na natureza os órgãos sensoriais são responsáveis pela captação de informação presente no ambiente, essa operação é executada em robôs ou outros sistemas autônomos por sensores, dispositivos que mensuram uma grandeza física e a transformam em uma representação que possa ser interpretada por um observador ou sistema.

Robôs e sistemas autônomos frequentemente enfrentam a necessidade de reconhecer objetos no ambiente em que operam, seja esse o objetivo final ou para auxiliar em outras tarefas, como por exemplo navegação autônoma através de marcos [Thrun et al., 2005]. Essa habilidade é bastante relevante onde a compreensão do ambiente é fundamental para um bom desempenho do robô, como ambientes domésticos [Pineau



(a) Iluminação ofuscante

(b) Falta de iluminação



(c) Oclusão visual

Figura 1.1. Exemplos de interferência visual. Imagens retiradas de Somerville [2011]; Yates [2014]; Lacheze et al. [2009].

et al., 2003] ou cenários de interação humana [Thrun et al., 1999].

Para melhorar seu desempenho, a área de pesquisa em fusão de sensores tem papel fundamental na forma como a percepção e interpretação do ambiente ocorrem, ao permitir combinar as informações obtidas de múltiplas fontes, fazendo com que a informação resultante seja superior às informações individuais caso consideradas isoladamente.

A melhoria na informação final se dá através da diminuição de influência do ruído (natural do ambiente e o introduzido pelo próprio sensor), aumentando sua exatidão, ou da agregação de informação adicional. Por exemplo: imagens de uma mesma cena podem ser combinadas levando em consideração o deslocamento dos *pixels* entre elas para gerar uma imagem de maior resolução (Figura 1.2), bem como imagens obtidas de ângulos diferentes podem ser combinadas para prover informação de profundidade da cena.

É possível ainda combinar informações de modalidades sensoriais diferentes, algo efetivo para atividades que podem ser descritas por duas ou mais modalidades fortemente relacionadas, como a fala. Nesta, os fonemas, menor unidade fonológica da língua, estão fortemente associados à configuração visual da boca no momento da pronúncia, como formato dos lábios e o quanto é exibido da língua ou dentes. Essa

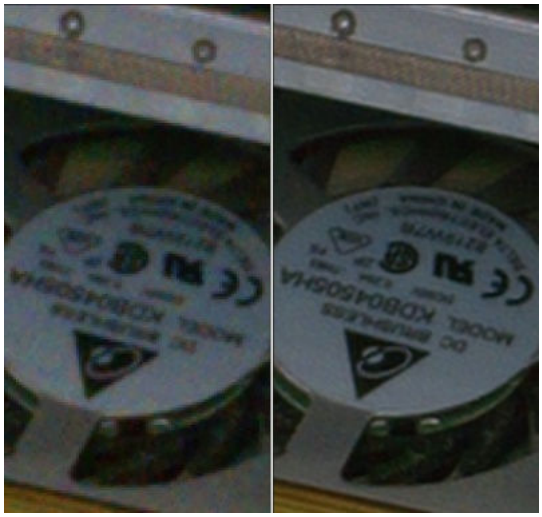
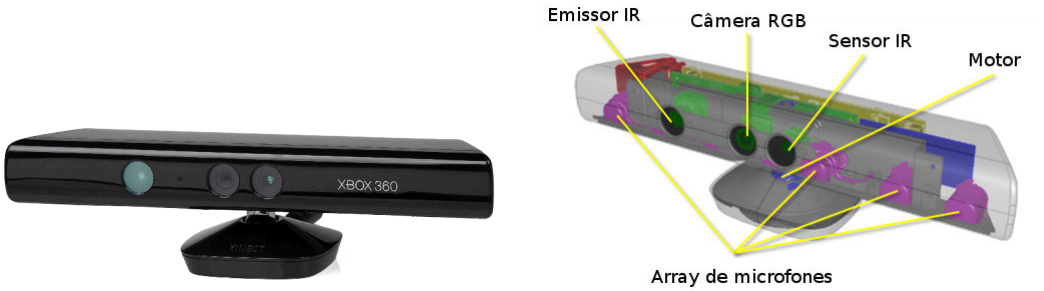


Figura 1.2. Superresolução, um exemplo de fusão. A imagem à direita, de maior definição, é obtida através da combinação de várias imagens de qualidade inferior (à esquerda). Imagem retirada de Almalence [2007].

característica implica em uma melhor estimativa de quais palavras estão sendo ditas por um interlocutor através de sistemas de reconhecimento [Shivappa et al., 2010].

Por fim, outra grande vantagem da fusão é a redundância de sensores: há tanto a diminuição do impacto em caso de falha de algum desses quanto a possibilidade de uso de sensores de baixo custo cujo uso conjunto pode equiparar o desempenho de um equipamento mais sofisticado, contribuindo para redução do custo total. Um exemplo de sensor integrado de baixo custo é o Kinect, usado nesta pesquisa e ilustrado na Figura 1.3.



(a) Sensor Kinect, visão externa (b) Sensor Kinect, componentes

Figura 1.3. Sensor integrado de baixo custo Kinect, que agrega uma câmera RGB, uma câmera de infravermelho e um *array* de microfones, atualmente usado em diversas aplicações comerciais e científicas [Walker, 2012]. Imagens retiradas de Amos [2010].

1.1 Aplicações de fusão audiovisual

Inicialmente usada para fins militares [Hall & Llinas, 1997], como reconhecimento de alvo, vigilância e controle de veículos, a fusão de sensores é uma área que vem crescendo, acompanhando o aumento do poder computacional dos computadores, disponibilidade de sensores e o desejo de uso de informações cada vez mais exatas e completas. Com o advento da internet e a crescente oferta de conteúdo multimídia, também há demanda por métodos automáticos para detecção e classificação de conteúdo, como detecção de conteúdo impróprio ou geração de *tags* automáticas para indexação.

Especificamente para fusão audiovisual há aplicações em duas categorias-chave:

- Classificação: biometria [Ortega-Garcia et al., 2004], classificação de conteúdo multimídia [Liu et al., 2011], reconhecimento automático de fala/interlocutor [Chen, 2001] e análise de reuniões [Shivappa et al., 2010];
- Localização: *tracking* de pessoas [Gehrig et al., 2005], localização autônoma [Strobel et al., 2001], entre outras.

No contexto de classificação de objetos, tais técnicas podem ser aplicadas em situações como:

- Vigilância e segurança em ambientes inteligentes [Menegatti et al., 2004];
- Exploração autônoma, através da identificação de marcos multimodais [Bohg et al., 2010; Ekvall et al., 2006];
- Coleta inteligente a partir da identificação de materiais e objetos [McGibney et al., 2012], ilustrado pela Figura 1.4.

1.2 Objetivo

Nesta seção são apresentados os objetivos gerais e específicos deste trabalho.

1.2.1 Objetivo geral

Nesta pesquisa deseja-se estudar o impacto do uso de diferentes modalidades sensoriais e como essas se complementam. Foi escolhida como aplicação uma tarefa comum na área de robótica: reconhecimento de objetos. Para isto, foram escolhidas como fontes

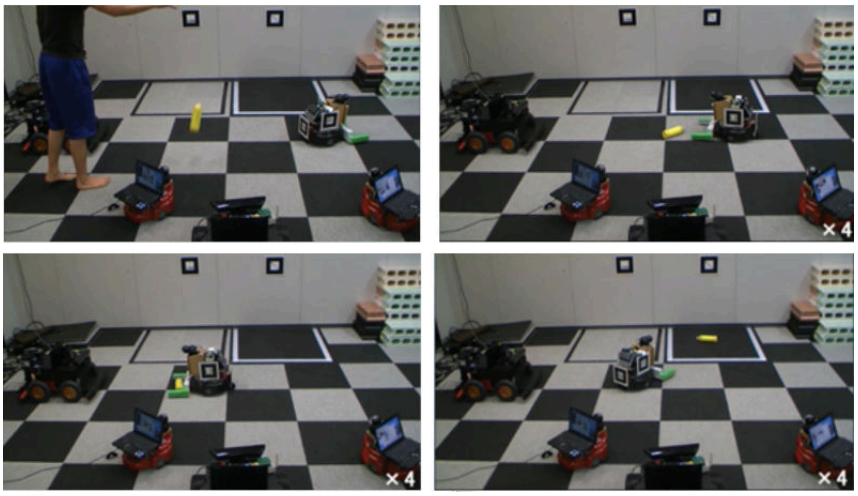


Figura 1.4. Exemplo de aplicação: coleta seletiva [McGibney et al., 2012]. Nesse trabalho, uma equipe de robôs detecta um objeto introduzido no ambiente (a garrafa amarela) e, com base em sua cor e material (inferido com base em sua assinatura sonora decorrente do impacto do objeto com o chão), o classifica em uma categoria, movendo o objeto para um local específico. Imagem retirada de McGibney et al. [2012].

de dados modalidades sensoriais facilmente atendidas pelos sensores modernos: audição e visão.

Não há como objetivo comparar os desempenhos de reconhecimento de objetos entre as modalidades ou qual supera a outra de forma geral, mas como seu uso conjunto pode melhorar o desempenho de um sistema. Assim, esta dissertação aborda o problema de extração de informações de som e imagem (de textura e de profundidade) e sua combinação com a finalidade de reconhecimento.

O problema relacionado a este trabalho pode ser descrito pela pergunta:

Como melhorar o reconhecimento de objetos combinando informações de som e imagem?

Há interesse em analisar qual o ganho proporcionado ao se usar conjuntamente informações auditivas e visuais e como essas podem se complementar em diferentes cenários.

1.2.2 Objetivos específicos

Objetivos específicos deste trabalho incluem:

- Execução rápida;
- Uso de equipamentos simples;

- Criação de base de dados própria.

Dado o objetivo de classificar objetos utilizando informações audiovisuais, algumas considerações devem ser feitas:

- Que tipos de objetos deseja-se reconhecer?
  - Esses geram som espontaneamente?
    - Em caso negativo, como gerar uma assinatura de áudio?
- Que tipo de assinatura de áudio/imagem serão utilizadas?
- Como essas informações serão combinadas de modo satisfatório?
- Quais as condições de ambiente a serem consideradas?

Essas questões delinearão a abordagem usada neste trabalho.

Como muitos dos objetos encontrados no dia-a-dia não emitem som por si só, há necessidade de algum tipo de interação com esses para geração de som, usado no reconhecimento. A base de dados criada para fins experimentais segue essa característica, e uma abordagem de reconhecimento ativo é usada, neste caso, para geração de som dos objetos a partir de uso de um manipulador robótico.

1.2.3 Contribuições

As principais contribuições deste trabalho são:

- Estudo sobre a fusão de atributos de auditivos e visuais para reconhecimento de objetos;
- Aplicação de um método de fusão que explora os atributos identificados;
- Criação de uma base de dados audiovisual estruturada de objetos com diferentes materiais e geometrias.

1.3 Organização do documento

Esta dissertação é estruturada da seguinte forma: O Capítulo 2 apresenta os principais trabalhos relacionados envolvendo reconhecimento de objetos. O Capítulo 3 contém referencial teórico relacionado ao reconhecimento e processamento audiovisual, bem como embasamento biológico.

No Capítulo 4 é descrita a metodologia usada para reconhecimento de objetos estáticos usando informações audiovisuais. O Capítulo 5 contém a descrição dos experimentos executados e análise dos resultados obtidos. Por fim, no Capítulo 6 são discutidas as conclusões desta pesquisa, suas limitações e trabalhos futuros.

Capítulo 2

Trabalhos relacionados

Neste capítulo são detalhados trabalhos de reconhecimento de objetos diretamente relacionados às três subáreas desta pesquisa: reconhecimento auditivo, visual audiovisual. Descritores e abordagens em outras aplicações relacionadas estão no próximo capítulo.

2.1 Reconhecimento visual

O problema de reconhecimento de objetos é um dos mais conhecidos na área de Robótica e Visão Computacional, importância traduzida na grande quantidade de pesquisas para este fim produzidas anualmente e na demanda por meios escaláveis e eficazes de fazê-lo, seja para indexação *online* por máquinas de busca ou melhor atuação de sistemas autônomos que operam em ambientes reais.

Trabalhos na área tiveram início na década de 1970 em Visão Computacional com pesquisas seminais como a de Agin & Binford [1976], que usa informação de profundidade para descrever partes dos objetos como cilindros. É possível observar a relação intrínseca entre técnicas para representação e reconhecimento de objetos e os sensores usados, como modelos geométricos baseados em contorno na década de 70 devido a ascensão de *range finders*, ou modelos baseados em textura na década de 80 após o surgimento de câmeras eletrônicas.

Se por um lado a popularidade de modelos geométricos se deve ao seu bom desempenho em condições variadas de iluminação e textura, além da eficácia no reconhecimento de objetos através da projeção do contorno do objeto em um plano, a riqueza de informação provida por modelos baseados em textura [Murase & Nayar, 1995] é um dos principais atrativos desse último, apesar de sua maior complexidade.

Métodos baseados em descritores locais surgiram posteriormente com o objetivo de superar interferências visuais como oclusão ou visão parcial, como no popular tra-

balho de Lowe [1999], onde foi introduzido o descritor SIFT (*Scale Invariant Feature Transform*), que usa informação do gradiente local em diferentes escalas para representar partes da imagem, provendo um descritor invariante a rotação.

De fato, muitas abordagens usam representações locais para gerar representações intermediárias para reconhecimento [Boureau et al., 2010], assim, muito trabalho tem sido investido em gerar representações locais robustas [Lowe, 1999; Bay et al., 2008] e boas abstrações dos dados, como BoW (*Bag-of-Words*) [Csurka et al., 2004] ou SPM (*Spatial Pyramid Matching*) [Lazebnik et al., 2006].

Em Csurka et al. [2004] são usados descritores SIFT [Lowe, 1999] agrupados por k-means [Kaufman & Rousseeuw, 1990] para geração do vocabulário (devido a métrica euclidiana) para o modelo BoW, um histograma do número de ocorrências de um padrão (o descritor) em uma imagem. A abordagem foi testada com uma SVM (*Support Vector Machine*) de *kernel* linear em uma base própria contendo 1776 imagens de sete categorias, obtendo uma taxa de acerto de 85%. Esse modelo tem como vantagens sua simplicidade, eficiência e invariância a transformações afins e robustez a variações nas condições de ambiente, além de ser flexível para explorar descritores com outros tipos de informação, como geométrica.

Em Lazebnik et al. [2006], histogramas hierárquicos de descritores SIFT são extraídos ao particionar a imagem progressivamente em sub-regiões, computando os histogramas em cada uma. Essa abordagem tem a vantagem de associar espacialmente a ocorrência dos descritores, ao contrário do BoW, melhorando a performance do reconhecimento de objetos em ambientes com vários objetos próximos uns dos outros, além da ausência de necessidade de segmentação do objeto de interesse. Essas características do SPM motivam seu uso em modelos mais recentes para reconhecimento de objetos [Boureau et al., 2010; Coates et al., 2011] baseados em *deep learning* (uma forma de *feature learning*) [Arel et al., 2010; Bengio et al., 2014], onde tenta-se modelar abstrações em alto nível dos dados através do aprendizado de diferentes representações.

Essa tendência de obtenção de descritores cada vez mais discriminativos nasceu da necessidade de armazenar e processar quantidades massivas de dados visuais para reconhecimento, onde a importância de representações compactas e de fácil comparação cresceu dramaticamente, encorajando o uso de descritores compactos [Calonder et al., 2010] e abstrações eficientes e representativas dos dados através das técnicas de *feature learning* e/ou codificação esparsa [Oliveira et al., 2012; Sudhakaran & Pappachen James, 2014], que reduz a informação redundante presente nos dados, provendo compactação ao custo de maior complexidade computacional.

Um exemplo desta abordagem é a de Krizhevsky et al. [2012], que usou redes neurais de convolução para generalizar a base de imagens ImageNet [ImageNet, 2014],

que contém mais de 15 milhões de imagens com milhares de subcategorias.

Nesse trabalho foram usadas redes em oito camadas, cinco de convolução e três totalmente conectadas, com o objetivo de maximizar uma regressão logística multinomial para as 1000 classes do desafio associado à base. Foram atingidos resultados excelentes, diminuindo o erro *top-1* e *top-5*[1] em teste para 37,5% e 17,0%. Esse trabalho foi estendido em Szegedy et al. [2014] ao aumentar o número de camadas para 24, além de diminuir em 12 vezes o número de parâmetros necessários, alcançando uma taxa de erro de 6,67%.

O bom desempenho em grande escala dessas abordagens se deve a sua capacidade de gerar representações dos dados em diferentes níveis de hierarquia, além do tamanho da base e a variedade das amostras. Sua execução local, no entanto, é inviável considerando seu objetivo, complexidade do modelo (60 mil parâmetros para as redes em Krizhevsky et al. [2012], por exemplo) e a necessidade de *hardware* robusto para execução distribuída.

2.2 Reconhecimento auditivo

Historicamente, o problema de reconhecimento de objetos pode ser considerado um problema de Reconhecimento de Padrões. Em processamento de áudio[2], trabalhos seminais para reconhecimento começaram a surgir no início da década de 50 com o objetivo de reconhecimento de voz, como o trabalho de Davis et al. [1952], que explora a ressonância espectral para o reconhecimento de números ditos por um único interlocutor. Desde então, muito foco tem sido dado a reconhecimento de voz [Rabiner & Schafer, 2007] e, em menor grau, música [Tzanetakis & Cook, 2002; Gold et al., 2011], com exemplos de aplicações comerciais, como assistentes de voz pessoais [Siri, 2015] e produtos para reconhecimento musical[Shazam, 2015; Soundhound, 2015].

Ainda que do mesmo domínio de aplicação, estes dois tipos de reconhecimento têm diferenças fundamentais em relação ao problema de reconhecimento de objetos: a estrutura do problema e a forma de aquisição dos dados.

Ao contrário de áreas de aplicação que lidam com a voz humana, a maioria dos dados não é estruturada, assim não é possível fazer suposições sobre repetições ou estrutura harmônica do sinal [Chu et al., 2009].

[1] Erro *top-n*: Erro de classificação onde n categorias previstas são sugeridas em ordem de prioridade para a amostra. A classificação é considerada correta se a categoria real se encontrar entre as n sugeridas.

[2] Representação de um som, seja ele sintetizado artificialmente ou capturado por transdutor, como um microfone [Rossing, 2007].

Sendo a fala estruturada e tendo algumas regras de construção conhecidas, estruturas mais complexas podem ser divididas em módulos menores, onde a fala contínua pode ser representada através de palavras ou fonemas. Ainda que haja grande variação na pronúncia, a estrutura permite adequação de modelos sequenciais clássicos como HMM (*Hidden Markov Model*) [Rabiner & Schafer, 2007].

A correlação estrutural imposta por restrições sintáticas da língua faz com que o reconhecimento seja mais fácil. Para sons genéricos, no entanto, não há estrutura conhecida. Assim esses modelos sequenciais podem ser inviáveis dependendo da escala de aplicação.

A aquisição de sinais de áudio vindos da voz ou de instrumentos musicais é feita de forma passiva, no entanto, dificilmente objetos geram sons por si só isoladamente, requerendo, na maioria das vezes, uma abordagem ativa para o problema. Especificamente para reconhecimento de objetos, relativamente poucos trabalhos lidam diretamente com o problema abordado usando áudio.

Um dos primeiros trabalhos na área foi o de Krotkov [1995], que explora a teoria de Wildes & Richards [1988] para reconhecimento de material, baseada na associação do ângulo de repouso de um material[3] e sua rigidez. Nesse trabalho o material de um objeto é estimado após objeto sofrer interação mecânica, onde o ângulo de repouso θ é estimado através do tempo de decaimento t_e que a vibração do objeto demora a decair a $1/e$ de seu valor original após o objeto ser golpeado com um pêndulo:

$$\theta = \frac{1}{\pi f t_e}, \tag{2.1}$$

onde f é a frequência de amostragem. Nesse trabalho a análise é feita através do espectrograma do sinal. Resultados experimentais para reconhecimento de material com objetos de madeira, alumínio, vidro, plástico e bronze apontaram que o áudio e sua análise via espectrograma podem ser usados com sucesso para discriminação de material. Há uso de informação visual, mas de maneira bastante simples: apenas para cálculo da distância percorrida pelo objeto após o choque mecânico, com o objetivo de auxiliar na estimação de parâmetros como o peso, considerado na etapa de reconhecimento.

Análises posteriores [Krotkov et al., 1997] verificaram que o tempo de decaimento é dependente da frequência, mas que ainda assim, a forma como um som decai é um importante fator para reconhecimento via áudio [Klatzky et al., 2000], inclusive para o ser humano. Este mesmo modelo foi usado para criar um modelo para sintetização de

[3]Maior ângulo que o talude (plano inclinado com o objetivo de prover estabilidade) do monte de um determinado material faz com o plano horizontal sem ocorrer deslizamento à medida que mais material é adicionado.

áudio baseado em interação mecânica através da generalização do espectrograma em múltiplas iterações [Richmond & Pai, 2000].

Uma limitação desses trabalhos é a pequena quantidade de objetos testados, apenas um por material, impedindo uma análise mais aprofundada se o reconhecimento pode ser generalizado para outras instâncias de mesmos materiais ou se há outros fatores implícitos necessários ao o modelo.

Mais recentemente, Sinapov et al. [2009] abordaram o problema usando *clustering* não supervisionado usando de SOM (*Self Organizing Map*) [Kohonen, 2000] para discriminar entre objetos recipientes (que armazenam conteúdo, como caixas de cereal, potes, latas de refrigerante, entre outros) e não recipientes. As capturas de áudio são feitas durante todas as ações de interação (como empurrar, agarrar, derrubar e chacoalhar) de um manipulador com o objeto, podendo influenciar negativamente o desempenho pela adição de ruído das juntas e servo motores. Em compensação, informação tátil, como "dureza" do objeto, é utilizada para classificação, tornando o método menos suscetível a ruído pela fusão de sensores.

Em McGibney et al. [2011] são usados descritores MFCC (*Mel Frequency Cepstral Coefficients*) em uma abordagem *nearest neighbor* para ambientes com níveis de ruído variados, onde os sinais de áudio são obtidos manualmente ao deixar o objeto cair no chão. No entanto, uma pequena base de testes foi usada, com apenas 4 objetos de diferentes materiais. A abordagem ainda é bastante suscetível a *outliers* e não escalável, pois é necessário computar a distância de cada amostra de teste para todas, isto pode ser contornado com técnicas de *hash*, para diminuição do espaço amostral, mas tal análise não foi feita. Testes estendidos com 10 objetos em um ambiente multiagente foram feitos em McGibney et al. [2012], obtendo 92% de acurácia, no entanto as limitações da abordagem permanecem.

2.3 Reconhecimento audiovisual

A área de reconhecimento de voz foi uma das pioneiras em reconhecimento audiovisual, através de trabalhos seminais como os de Petajan [1984] ou Yuhas et al. [1989], atraindo muito foco desde então [Shivappa et al., 2010]. Especificamente para o reconhecimento de objetos, relativamente poucos trabalhos tratam deste problema.

Em Arsenio & Fitzpatrick [2003], o reconhecimento de objetos é feito de forma passiva, detectando ritmo na utilização destes para alguma finalidade (assume-se que o uso do objeto gera som, como a movimentação feita ao se usar um martelo dá indícios de que o objeto usado é deste tipo), monitorando o eixo de trajetória principal do ob-

jeto (via vídeo) e bandas de frequência (via áudio) que estão oscilando conjuntamente. A fusão ocorre assumindo que a trajetória do objeto oscila na mesma frequência do som gerado, sendo esta feita com tolerância de até 60 ms para sincronização dos eventos. As limitações dessa abordagem consistem na restrição de associação rítmica entre as modalidades, que não é geral. É necessário ainda saber previamente como manusear os objetos, o que invalida o propósito de reconhecimento de objetos executado autonomamente por um robô.

Lacheze et al. [2009] propôs um esquema de reconhecimento passivo em dois estágios: uma etapa local, feita a cada *frame* (trecho do sinal de áudio), minimizando o erro entre as modalidades, seguida de uma decisão global (feita por votação simples, escolhendo a categoria mais escolhida entre os *frames* ou aquela que minimiza o erro global). A informação visual é tratada em uma abordagem BoW usando atributos visuais baseados em amostragem entrópica da textura dos objetos. O reconhecimento por áudio se dá através de redes neurais usando um modelo de cóclea, onde N filtros são aplicados ao sinal independentemente e a energia total dos *frames* em cada um dos N sinais é usada como atributo na etapa de classificação. Não são feitas comparações de desempenho com outros métodos, apenas entre as formas de fusão propostas com diferentes níveis de ruído, onde o esquema que combina fusão local e global tem melhor desempenho nos diferentes cenários testados. A base de dados usada para testes é limitada pois apenas objetos que emitem som ao serem movimentados são usados, como carros de fricção.

Outro uso audiovisual é feito em McGibney et al. [2012], onde as modalidades sensoriais são usadas em etapas diferentes para estimar informações independentes sobre o objeto. O áudio é usado para estimar o material do objeto e uma localização inicial, usada pelo módulo visual para foco de atenção. O módulo visual, por sua vez, classifica o objeto usando informações simples, como histograma de cor. A abordagem mantém características e problemas similares aos encontrados em [McGibney et al., 2011], se sobressaindo pelos diferentes problemas abordados no sistema (sistemas multiagente, manipulação, reconhecimento visual e de áudio, etc.).

Pieropan & Salvi [2014] usa uma abordagem baseada no uso de HMM para reconhecimento de ações relacionadas ao uso de objetos (abrir caixa de leite, pôr leite em copo, etc.), onde a fusão ocorre concatenando o vetor de características das duas modalidades. Para descrição do áudio, assim como em McGibney et al. [2012] também é utilizado MFCC em conjunto com primeiros e segundos derivativos para cada *frame*. O descritor visual usado é a posição relativa entre objetos envolvidos na ação.

Uma limitação das abordagens mencionadas é que todas são passivas: espera-se que o objeto tenha som gerado de alguma forma, seja por interação com humanos ou

gerado espontaneamente.

Os trabalhos mais próximos da abordagem desta pesquisa são os de Nakamura et al. [2007b] e o de Sinapov [2013], que fazem reconhecimento ativo dos objetos. Em ambos são processadas informações audiovisuais e hápticas.

Em Nakamura et al. [2007b] a abordagem *Bag-of-Words* é usada no processamento dos 3 tipos de informação: descritores SIFT para informação visual, descritores MFCC para áudio e descritores hápticos (voltagem dos sensores de pressão do manipulador, representando a dureza do objeto, e ângulo entre os dedos do manipulador).

O reconhecimento é feito de maneira não-supervisionada, analisando a co-ocorrência das palavras de cada modalidade e estimando propriedades de uma modalidade a partir de outra (se um objeto faria barulho após ser tocado ou se é duro somente o observando, por exemplo), essa modelagem ocorre usando distribuições multinomiais condicionalmente independentes com PLSA (*Probabilistic Latent Semantic Analysis*) [Hofmann, 1999] ou LDA (*Latent Dirichlet Allocation*), em [Nakamura et al., 2009].

Resultados experimentais com 40 objetos em 8 categorias diferentes corroboram o fato de que o uso de outras fontes sensoriais aliadas à visão reforça o reconhecimento, mas apontam a existência de uma correlação muito maior entre informações háptico-visuais do que audiovisuais.

Essa falta de correlação entre som e tato é possivelmente causada pela forma de interação escolhida (agito do objeto) e base de dados usada, constituída na maior parte por objetos que na prática emitem sons de baixa amplitude mesmo com interação, como bichos de pelúcia.

Sinapov [2013] demonstra que mesmo uma grande quantidade de objetos pode ser discriminada entre si quando múltiplas interações são executadas, como agarrar, levantar, derrubar, etc., onde as informações audiovisuais são integradas à informações proprioceptivas para reconhecimento de uma base contendo 100 objetos estáticos em 20 categorias.

Nesse trabalho é usada uma abordagem supervisionada para reconhecer objetos visualmente usando histogramas RGB, descritores SURF (*Speeded-Up Robust Features*) [Bay et al., 2008] codificados com uma abordagem BoW [Csurka et al., 2004], além do fluxo óptico das imagens[4] [Sun et al., 2010]. O áudio é representado por coeficientes da DFT (*Discrete Fourier Transform*) codificados em intervalos discretos ao longo do sinal. Os torques nas juntas e posição dos dedos foram usados como descritores proprioceptivos.

Os modelos (um para cada tipo de interação possível) são classificados separada-

[4]Padrão que dá sensação de movimento entre duas imagens consecutivas.

mente e combinados de acordo com a regra da combinação uniforme. Seja A um vetor de atributos onde cada $a_m \in A$ equivale à entrada para um modelo de interação, a saída desta combinação é

$$P(\hat{y} = y|A) = \alpha \sum_{a_m \in A} P(\hat{y} = y|a_m), \tag{2.2}$$

onde y representa a categoria do objeto e α é uma constante de normalização.

Esse trabalho se sobressai pelo uso robusto de mais de um tipo de interação mecânica e por investigar como seu uso sequencial pode ajudar a desambiguar as classes ao maximizar o ganho de informação para algumas modalidades através de determinada interação (como deixar o objeto cair, para o áudio). O uso de três informações sensoriais (visual, sonora e háptica) é complementar e contribui para estimação de várias propriedades do objeto, como peso, dimensão, material, além do comportamento quando manipulado (se tomba facilmente quando o robô tenta empurrá-lo, característica capturada pelo fluxo ótico), além de uma base sólida de experimentação.

O trabalho aqui apresentado se diferencia de Sinapov [2013] por dar mais ênfase aos descritores de áudio, com foco na dinâmica impulsiva do som capturado [Biondi et al., 2014; Dufaux, 2001; Freire & Apolinário, 2010] em vez dos descritores visuais, com o objetivo de demonstrar como o áudio auxilia no reconhecimento de objetos que sejam visualmente bastante similares. Uma abordagem ativa é usada para interação, onde um manipulador executa uma interação mecânica simples com os objetos. Experimentos considerando a geometria da base de dados e diferentes níveis de ruído também são incluídos.

De forma geral, algumas características dificultam uma avaliação qualitativa dos trabalhos relacionados a este problema, como a pequena quantidade de trabalhos, a carência de bases de dados padrão e consequente a comparabilidade de casos de teste e desempenho. Essa é uma das razões que motiva a liberação para interessados da base de dados criada, como uma contribuição para pesquisa neste segmento.

Capítulo 3

Referencial teórico

3.1 Descritores

Um descritor é uma propriedade mensurável de um fenômeno observado [Bishop, 2007]. Sua função é representar a informação de forma a facilitara comparação entre diferentes instâncias e oferecer compactação dos dados originais. A compactação é importante principalmente em tarefas que lidam com um grande volume de dados.

Na literatura, há muitos descritores disponíveis para reconhecimento, variando de acordo com o tipo de dados amostrados, aplicação desejada e equipamento utilizado. Desse modo, diferentes descritores são usados para representação de dados visuais e de áudio, entretanto, há alguns passos comuns na extração de qualquer descritor:

1. Segmentação da informação de interesse: aplicada com o intuito de remover ruído ou informação que não é de interesse para a aplicação. Em reconhecimento de objetos usando imagens, seleção apenas de regiões em que o objeto se encontra, por exemplo;

2. Detecção de características de interesse que se deseja representar, como pontos salientes (fáceis de discriminar) em uma imagem, ou no caso de áudio, o formato da onda ou sua frequência fundamental;

3. Extração de descritores dessas características.

Descritores são normalmente usados para reconhecimento de objetos em conjunto com técnicas de aprendizado de máquina, área que consiste no estudo de métodos automáticos para extrair modelos de generalização a partir de dados conhecidos, com o objetivo de reconhecer novos dados, mesmo sem conhecimento *a priori* sobre o processo que os gerou [Alpaydin, 2009].

Normalmente sua aplicação ocorre quando não há conhecimento suficiente do domínio de aplicação para gerar um modelo matemático que o represente com fidelidade, motivo pelo qual tenta-se aproximar este modelo ideal através de dados.

Um problema de aprendizado de máquina pode ser definido como um problema de mapeamento definido em um domínio de entrada X (atributos a serem considerados na classificação), um conjunto de saídas Y (categorias ou classes associadas aos atributos) e uma distribuição de probabilidade P sobre X e Y. Um classificador C é uma função de mapeamento entre os dois domínios, tal que $C : X \rightarrow Y$. Para maiores detalhes sobre métodos de aprendizado de máquina e classificadores, referir a Bishop [2007]; Alpaydin [2009].

As próximas seções contêm um panorama das técnicas relacionadas à representação de dados de áudio e vídeo por meio de descritores.

3.1.1 Descritores auditivos

O som é um fenômeno físico decorrente de uma perturbação mecânica em um meio elástico (no caso do ar, compressão e expansão dos gases que o formam) no espectro audível: 16Hz a 20KHz, no caso dos humanos [Blauert, 1996]. Sinais de áudio têm propriedades temporalmente variáveis e por isso são referidos como processos aleatórios não-estacionários [Richard et al., 2013], motivo pelo qual técnicas de análise tanto no domínio do tempo quanto da frequência são empregadas para representar sua dinâmica.

De maneira geral, descritores de áudio podem ser divididos de acordo com algumas características, como escopo do descritor (global ou local), tipo de representação usada para analisar o sinal considerada, que pode ser temporal, FT (*Fourier Transform*), WT (*Wavelet Transform*), etc. A partir dessas representações são extraídos descritores para capturar características temporais, espectrais e espectro-temporais de eventos sonoros [Peeters et al., 2011].

Alguns autores consideram ainda outros tipos descritores como categorias à parte, como descritores cepstrais[1] [Peltonen et al., 2002] ou cocleogramas (de Paterson, de Lyon, Gammatone, etc.) [Richard et al., 2013]), mas isto não é um consenso.

Dois descritores temporais comuns na literatura, por exemplo, são:

- *Short-time energy*:

$$E_j = \frac{1}{N} \sum_i s_j^2(i), \tag{3.1}$$

[1] *Cepstrum*: espectro do logaritmo do espectro de uma onda. Conceito introduzido por Bogert et al. [1963], que compara a relação entre este novo domínio, referido como "quefrência", e a frequência com a relação tempo-frequência. O termo *"cepstrum"* é um anagrama do termo com o qual tem relação, *"spectrum"* [Oppenheim & Schafer, 2004].

onde E_j é a energia do sinal s em uma janela j retangular de tamanho N. Este descritor fornece uma representação da variação de amplitude ao longo do tempo e permite distinguir segmentos de áudio com atividade (fala ou outros sons). Sua variação ao longo do tempo pode ajudar a identificar ritmo e periodicidade.

- *Short-time Average Zero-Crossing Rate*:

$$Z_j = \frac{1}{2} \sum_i sgn[s_j(i)s_j(i+1)], \tag{3.2}$$

onde

$$sgn[x] = \begin{cases} 1 & se \quad x \geq 0 \\ -1 & c.c. \end{cases} . \tag{3.3}$$

Este descritor contabiliza o número médio de inversões de sinal ao longo da amostra, podendo, de acordo com sua curva (considerando variância, estabilidade e regularidade) identificar segmentos de fala, som ambiente e música [Zhang & Kuo, 2001].

Descritores espectrais são obtidos ao se transformar o sinal para o domínio domínio da frequência, normalmente utilizando a TF. Devido à natureza não-estacionária do som e à limitação da TF original de representar a mudança temporal do sinal, a STFT (*Short-Time Fourier Transform*) pode ser aplicada, executando a transformada a uma janela de curta duração, de 10 a 30 ms[2], podendo assim representar o sinal por uma soma ponderada de senoidais, como proposto pela FT [Richard et al., 2013]. Esse procedimento permite acessar os componentes temporais e espectrais locais como uma sequência de observações que estimam propriedades locais do sinal.

A STFT é um meio termo entre a representação temporal e espectral do sinal, entretanto sua resolução depende do tipo e tamanho da janela adotada, que é a mesma para todas as frequências e intervalos de tempo. Uma abordagem mais flexível para o problema é o uso de Wavelets [Mallat, 1989], que oferece informação espectro-temporal e permite o uso de janelas maiores onde informações sobre as baixas frequências são desejadas[3], quanto janelas menores para altas frequências, além do uso de funções arbitrárias finitas que não uma senoide, como na TF.

[2]Intervalo devido à voz humana ser considerada estacionária neste período de tempo. Acabou tornando-se um valor de referência por este motivo.

[3]Nota: ainda que conceitualmente Wavelets sejam representadas em termos de escalas, o termo "frequência" é mantido para melhor entendimento.

Um compêndio com mais detalhes de outros descritores de áudio comumente usados pode ser encontrado em Peeters et al. [2011]. Uma iniciativa de padronizar descritores multimídia, incluindo descritores de áudio de baixo nível, culminou com a criação do padrão MPEG-7 [Casey, 2001]. Análises comparativas de descritores podem ser encontradas em Peltonen et al. [2002]; Cowling & Sitte [2003]; Mitrovic et al. [2007]; Chu et al. [2009].

A quantidade de descritores de áudio disponíveis e seu extenso uso em diferentes domínios de aplicação sugerem que uma etapa prévia de validação é necessária para verificar sua eficácia na tarefa desejada.

Especificamente para classificação de objetos, que produzem sons não-vocais, o estudo comparativo de Cowling & Sitte [2003] aponta o descritor MFCC ao lado de CWT (*Continuous Wavelet Transform*) como tendo o melhor desempenho entre os descritores testados, em sua maioria descritores espectrais, como coeficientes LPC (*Linear Predictive Coding*), BFCC (*Bark Frequency Cepstral Coefficients*), etc. O estudo de Chu et al. [2009] reforça o bom desempenho do MFCC para este tipo de tarefa, com ênfase no melhor desempenho alcançado através de seu uso conjunto com outros descritores.

O descritor MFCC, introduzido em 1980 por Davis & Mermelstein [1980], é um descritor que se destaca na literatura, tendo sido usado em diversas áreas, incluindo reconhecimento automático de fala e de locutor, tem exibido bom desempenho sistematicamente [Richard et al., 2013]. Este tem grande aceitação pelo fato de ser uma representação compacta das informações acusticamente relevantes para o sistema auditivo humano[4] e de suprimir variações espectrais irrelevantes em altas frequências devido ao uso de filtros logaritmicamente espaçados.

Juntamente com o descritor RASTA-PLP (*Relative Spectral Transform - Perceptual Linear Prediction*) [Hermansky, 1990] foi um dos descritores mais usados para reconhecimento de fala com modelos acústicos baseados em GMM-HMM (*Gaussian Mixture Model - Hidden Markov Model*) [Rabiner, 1989] para contornar os problemas de tamanho e variabilidade do sinal em si.

Esses modelos possibilitaram o surgimento de técnicas mais robustas para reconhecimento, seguindo a tendência do reconhecimento por vídeo no uso de abordagens baseadas em *deep learning* [Bengio et al., 2014], como o modelo DNN-HMM (*Deep Neural Network - Hidden Markov Model*) [Dahl et al., 2012; Yu & Deng, 2014], onde o HMM é usado para modelar a propriedade sequencial do sinal e a DNN para modelar as distribuições de probabilidade de emissão do HMM, que melhora substancialmente

[4]Diz-se que é um descritor perceptualmente inspirado devido ao espaçamento nas bandas do espectro dado aos filtros, aproximando o funcionamento dos capilares internos à cóclea humana.

o reconhecimento de longas sequências de palavras em relação a técnicas do estado da arte como modelos GMM-HMM que usam critérios como ML (*Maximum Likelihood*) ou MPE (*Minimum Phone Error*) para treinamento.

3.1.2 Descritores visuais

A primeira etapa, seleção da informação de interesse, pode ser executada usando uma máscara, esta podendo ser gerada de maneira simples, por um algoritmo de *flood-fill*, por exemplo, ou mais robustamente, através da projeção de uma nuvem de pontos segmentada. É necessário a seguir selecionar pontos representativos presentes nas regiões de interesse que possam abstrair a imagem a ser representada com menos ambiguidade. Em condições em que as regiões de interesse são pequenas com relação ao restante da imagem, amostragem densa dos pontos pode ser utilizada.

Há uma grande quantidade de trabalhos na literatura relacionados a detecção de pontos de interesse [Harris & Stephens, 1988; Lowe, 2004; Bay et al., 2008; Rosten et al., 2010]. De modo geral, tais detectores funcionam selecionando pontos de acordo com a distintividade de sua vizinhança.

Descritores de pontos de interesse podem ser divididos em duas grandes famílias, de acordo com o tipo de informação adquirida pelo sensor: os baseados em textura ou imagens de profundidade.

Na última década, a maioria das abordagens tende para para o primeiro grupo, representadas na literatura por métodos como SIFT [Lowe, 2004] e SURF [Bay et al., 2008], descritores muito populares graças a sua grande habilidade discriminativa. Mas ambos sendo baseados em histogramas de gradientes, os gradientes de cada pixel no *patch* precisam ser calculados, o que tem alto custo computacional, dificultando seu uso em tarefas que exigem processamento rápido, como tarefas em tempo real, ou uso em grande volume de dados.

Este é um dos motivos pelo qual tem aumentado o interesse em descritores binários nos últimos anos, como BRIEF (*Binary Robust Independent Elementary Features*) [Calonder et al., 2010], ORB (*Oriented FAST and Rotated BRIEF*) [Rublee et al., 2011], BRISK (*Binary Robust Invariant Scalable Keypoints*) [Leutenegger et al., 2011] e FREAK (*Fast REtinA Keypoint*) [Alahi et al., 2012], por exemplo.

Tais descritores têm diversas vantagens em relação às abordagens tradicionais citadas: são de baixo custo computacional (apenas as intensidades dos *pixels* no *patch* são comparadas, sem necessidade de computação do gradiente), tem armazenamento compacto (muito relevante para grandes bases de dados), e podem ser comparados

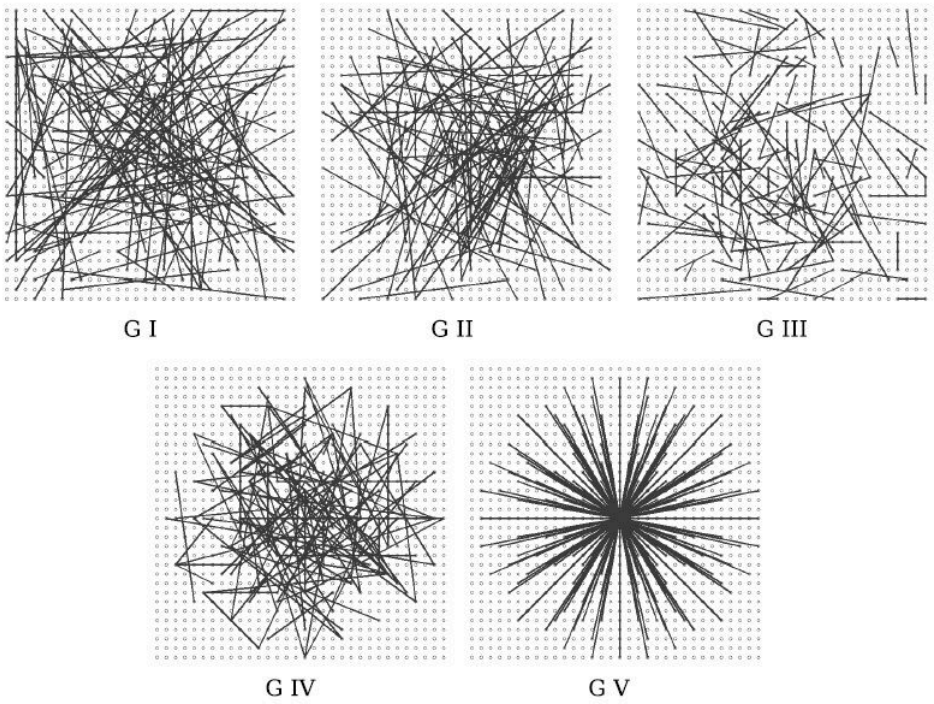


Figura 3.1. Diferentes padrões de amostragem testados para o descritor BRIEF. Cada linha representa um par no *patch* cujas intensidades são comparadas. Imagem retirada de [Calonder et al., 2010].

eficientemente, através de métricas como a distância de Hamming[5], por exemplo [Muja & Lowe, 2012].

Genericamente, descritores binários têm três características: padrão de amostragem da vizinhança do ponto de interesse, compensação de orientação (dependendo do descritor) e amostragem de pares. O padrão de amostragem se refere a um padrão geométrico de seleção dos pontos na vizinhança do *keypoint*, como ilustrado pela Figura 3.1.

Em alguns descritores, como o ORB, é feita ainda uma transformação para torná-los invariantes à rotação. No caso deste último, é calculado um vetor de orientação do centro do *patch* ao centroide de intensidade, a partir do qual é aplicada uma rotação para uma orientação canônica.

A seguir, pares de pontos no padrão geométrico têm suas intensidades comparadas (seguindo uma ordem específica), sendo atribuído um valor binário a essa comparação de acordo com qual elemento é maior. Esses valores são codificados em um vetor, gerando o descritor para o ponto de interesse.

O descritor extraído tem escopo local, pois representa apenas uma região da

[5]Número de posições em que duas *strings* binárias diferem entre si.

imagem, a vizinhança do *keypoint*, uma representação global da imagem é necessária para uma etapa seguinte de reconhecimento. Para representação global (para uma etapa posterior de classificação ou pareamento, por exemplo), uma abordagem comum [Galvez-López & Tardos, 2012] é utilizar a técnica *Bag-of-Words* [Csurka et al., 2004].

3.1.3 Bag-of-Words

Um BoW é um histograma do número de ocorrências de um padrão (o descritor) em uma imagem. Esta abordagem é composta de duas etapas distintas: *coding*, onde os descritores locais são agrupados de acordo com um vocabulário conhecido, formando categorias. E *pooling*, onde o vetor de características final da imagem é formado pela histograma de frequência dos *keypoints* em cada categoria.

Na área de pesquisa em Recuperação de Informação, onde o BoW normalmente é utilizado, o vocabulário é formado por palavras escolhidas que estão ou não presentes no texto. Como o espaço amostral dos descritores na área de Visão Computacional é muito vasto, técnicas de aproximação para atribuir descritores à certa "palavra" tem de ser aplicadas, assim a geração do vocabulário é executada através do *clustering* de descritores (neste caso, BASE) usados para esse propósito.

Particularmente para o caso de descritores binários, isto pode ser feito usando o algoritmo PAM (*Partition Around Medoids*), uma implementação do k-medoides, [Kaufman & Rousseeuw, 1990]. A escolha do k-medoides para fazer o agrupamento se deve ao fato deste ser mais robusto a ruído e *outliers* se comparado ao k-means, por usar medoides[6], além de permitir o uso de outras métricas de distâncias mais apropriadas, dependendo do contexto.

O k-means, por exemplo, que minimiza a distância euclidiana entre as amostras, não é apropriado para uso com descritores binários, que podem ser codificados como inteiros, mas não mantém uma relação direta entre os valores no domínio $\mathbb{R}$. No caso desses descritores, uma métrica de distância correta seria a de Hamming, que pode ser usada com o k-medoides. O alto custo computacional do o k-medoides pode ser impeditivo em alguns casos, dependendo da quantidade de amostras usadas para geração do dicionário, pois seu funcionamento na prática é um algoritmo de força bruta que tenta encontrar uma combinação de k amostras entre os dados (os medoides) que possam minimizar a distância global dos *clusters*. Entretanto, como a etapa de *coding* é executada antes da fase de treinamento, não há impacto no desempenho geral do sistema durante sua execução. O *pooling* é feito quantizando os descritores da imagem com as "palavras" do dicionário determinadas na etapa anterior.

[6] Amostra do próprio grupo que minimiza a distância a distância total do *cluster*.

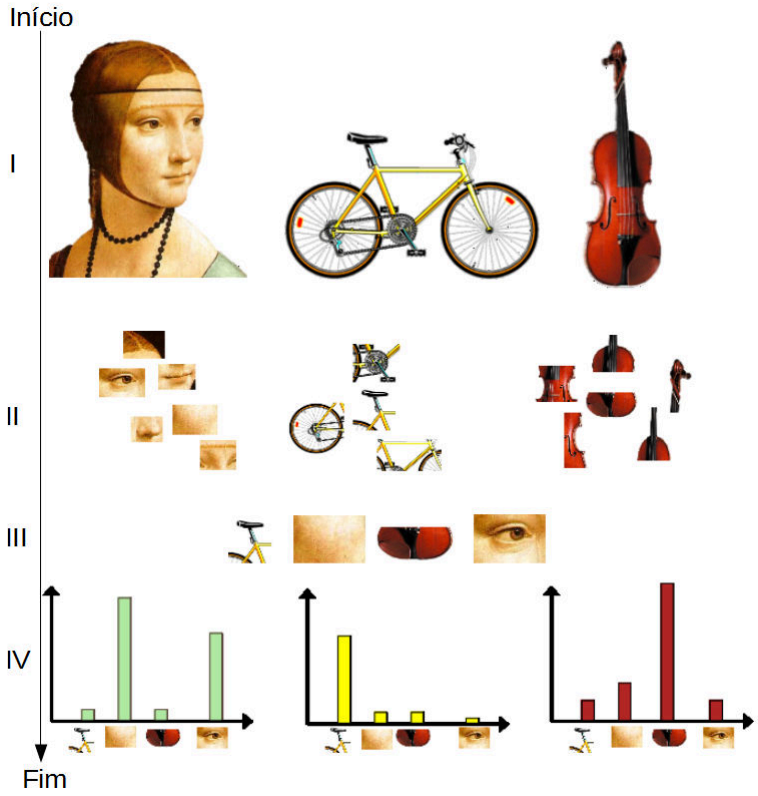


Figura 3.2. Exemplo em alto nível da execução do BoW para imagens. Adaptado de Fei-Fei [2005].

A Figura 3.2 contém um exemplo em alto nível da execução do BoW para uso em imagens:

I : Aquisição de imagens;

II : Extração dos descritores locais de cada imagem;

III : Geração de um vocabulário a partir de uma base de descritores (gerados na etapa anterior, por exemplo). Observe que cada um dos descritores escolhidos como "palavra" serve como centroide para representar um grupo de descritores;

IV : Histograma normalizado de cada "palavra" na imagem;

O descritor BoW final para a imagem, por exemplo, é o histograma obtido na etapa IV.

3.1.3.1 Definição do tamanho dicionário

Há algumas alternativas para definir o tamanho ideal para o dicionário usado com o BoW. Uma delas é a análise da silhueta s dos *clusters* [Rousseeuw, 1987; Kaufman

& Rousseeuw, 1990] do dicionário. Seja D um dicionário de tamanho arbitrário, a silhueta s de um *cluster* de D é definida como

$$s = \frac{b - a}{\max(a, b)}, \tag{3.4}$$

onde a é a distância média entre uma amostra e as restantes do mesmo *cluster* e b é a distância média entre uma amostra e as do *cluster* mais próximo.

Intuitivamente, a silhueta de um *cluster* define o quão bem representado este é pelo seu centroide/medoide: um valor alto de silhueta sugere que o *cluster* não é denso, ou seja, apesar do centroide/medoide ser o mais próximo de suas amostras, não é um representante ideal e não há uma separação bem definida entre o *cluster* e seus vizinhos.

A análise da silhueta consiste em analisar a silhueta média dos *clusters* de dicionários com tamanhos variados, onde um dicionário com menor silhueta média tem *clusters* mais bem definidos.

Outra abordagem para definição de tamanho do dicionário é verificar o desempenho de reconhecimento usando o dicionário em uma etapa de validação anterior ao treinamento. O dicionário com um melhor desempenho (medido por taxa de acerto, por exemplo) em validação é um bom candidato a para a fase de treinamento/teste.

Ambas as abordagens verificam o quão ajustado um dicionário é aos dados, a diferença reside em que escopo essa análise é feita. Na análise da silhueta ela ocorre verificando o quão conciso o dicionário usado é em uma abordagem não-supervisionada, na análise do desempenho do reconhecimento ela ocorre verificando o quão discriminatório o dicionário é diretamente nos dados usando uma abordagem supervisionada.

3.2 Fusão

Nesta seção se comenta algumas características da fusão audiovisual tal ocorre na natureza.

3.2.1 Embasamento biológico

A percepção pode ser considerada robusta se responde eficientemente a estímulos externos. Neste sentido, o cérebro humano é um dos melhores exemplos para ilustrar os conceitos relativos ao processamento e integração de estímulos: Esse processa e combina sinais dos cinco sentidos de maneira eficiente e em tempo real.

Pesquisas recentes sugerem que essa integração não ocorre de maneira rígida, mas que é sensível à situação, como quais sentidos estão sendo integrados ou as características dos estímulos recebidos, permitindo que estes sejam combinados eficientemente, tornando o cérebro bastante flexível quanto à integração de informações [Stein & Stanford, 2008].

Conhecida na biologia como integração multissensorial, esta é definida em nível neuronal por Stein & Stanford [2008] como:

Definição 1 *Integração multissensorial: diferença estatisticamente significativa em nível neuronal entre o número de impulsos nervosos decorrentes da combinação multimodal dos estímulos e o do estímulo individual mais efetivo entre estes.*

A integração multissensorial tem impacto direto na velocidade com que uma resposta do organismo é gerada (motora, por exemplo) ao induzir um limiar mínimo de estímulo inferior ao induzido pelos componentes unimodais [Stein & Stanford, 2008]. No ser humano, essa integração ocorre de forma tão suave que é percebida apenas pela ocorrência de estímulos muito discrepantes ou conflitantes entre os sentidos, gerando ilusões perceptivas, como o Efeito McGurk[7] [McGurk & MacDonald, 1976] ou o ventriloquismo[8] [Howard & Templeton, 1966].

Possíveis explicações para estes fenômenos são que alguns sentidos predominariam sobre outros, como a visão à audição, modificando sua percepção tal qual acontece no ambiente [Burr & Alais, 2006] ou que o sentido dominante é utilizado como referência para estruturar os estímulos dos demais sentidos [Shelton & Searle, 1980; Zahorik, 2001] e que por esse motivo teria tanta influência.

Isto é bastante plausível considerando que a percepção do ser humano é primariamente visual, de tal forma que a representação interna do mundo criada pelo homem é muito mais desenvolvida na visão que nos demais sentidos [Blauert, 1996]. Pode-se observar que conceitos e descrições são fortemente baseados em identidades visuais: diz-se em inglês *"the bell rings"* em vez de *"the sound bells"*, por exemplo. Ainda que o evento seja primariamente de natureza auditiva, a identidade visual do sino é mais forte.

[7] Evento ocorrido durante o reconhecimento de fala ao se associar informações visuais e auditivas muito conflitantes, podendo a informação final percebida diferir bastante de ambas. Por exemplo, quando uma pessoa ouve o som representado pelo fonema /ba/ mas vê o locutor pronunciando o fonema /ga/, a informação final não é /ba/ ou /ga/, mas sim algo próximo a /da/. Isto ocorre pois o cérebro integra a visão (movimento dos lábios) com o som da fala, modificando a atividade cerebral relacionada ao estímulo.

[8] Sensação de que o som parece vir dos lábios de um locutor, ao invés de sua real origem.

Mesmo com a aparente predominância da visão, seres humanos também extraem informações semânticas valiosas a partir do som, como sua fonte (um copo quebrando ou uma porta fechando, por exemplo), como foi produzido (um choque mecânico ou ressonância), além de simples propriedades do som em si, como o timbre [Gaver, 1993]. Outras pesquisas reforçam esse conceito ao comprovar que humanos conseguem distinguir entre materiais com bastante exatidão [Giordano & McAdams, 2006].

Ainda que se entenda em baixo nível como as informações de diferentes sentidos são transmitidas e para onde no cérebro, como ele as integra exatamente, principalmente entre modalidades diferentes, para formar uma versão unificada, coerente e semanticamente relevante do mundo permanece um problema em aberto [Burr & Alais, 2006].

Sabe-se, no entanto, que essa integração ocorre principalmente nos chamados *neurônios multissensoriais* (presentes em várias regiões do cérebro, mas existentes em maior concentração no mesencéfalo e córtex cerebral, especificamente no colículo superior[9]) e parece seguir duas regras gerais [Ernst & Bülthoff, 2004]:

1. Maximiza o ganho de informação recebido das diferentes modalidades;

2. Reduz a variância na estimação sensorial para aumentar sua confiança.

A hipótese mais aceita de como essas propriedades se comportam com relação à dominância de um sentido ou outro, como no caso do efeito McGurk e ventriloquismo, é que a percepção favorece a modalidade mais precisa ou apropriada de acordo com a situação [Boff et al., 1986]. Em tarefas como localização de eventos, a visão tem forte influência devido a sua capacidade de determinar distâncias e informações espaciais com maior precisão, ainda que a audição muitas vezes forneça uma estimativa inicial devido ao rápido tempo de resposta [Stein & Stanford, 2008]. Isso ocorre pois o processamento auditivo feito pelo cérebro humano é mais rápido que o visual, ainda que a velocidade de propagação do som (no ar) seja muito menor que a velocidade da luz (340 m/s contra 3 x 10^8 m/s), motivo pelo qual a audição teria dominância na percepção temporal de eventos.

A dominância é determinada não pelo sentido, mas sim pela informação fornecida e o quão confiável ela é dado o estímulo [Ernst & Bülthoff, 2004], o que explica as ilusões sensoriais comentadas anteriormente.

É possível observar que essas características da percepção são muito similares e compatíveis com modelos baseados em inferência Bayesiana e MAP (*Maximum a*

[9]Região do sistema nervoso central, especificamente no teto do mesencéfalo.

Posteriori) empregados em técnicas como Filtro de Bayes e derivados, utilizados amplamente em fusão de sensores.

3.2.2 Fusão de sensores

A fusão de sensores pode ser definida como [Dasarathy, 1997]:

Definição 2 *Estudo dos conceitos e técnicas desenvolvidas com objetivo de processar e integrar dados em um ambiente multissensor com o objetivo original de reverter a fissão de informação inerente ao processo de percepção de um ambiente físico por parte de um sensor.*

Ou seja, a percepção do ambiente é limitada pelos sensores utilizados, que capturam apenas parte da informação disponível (fissão), e não pela disponibilidade desta. A fusão é um meio de recombinar essas informações visando representar o mais fiel ou precisamente a informação sensoreada.

Com o aumento de popularidade nas últimas décadas e o maior acesso a novos tipos de sensores, atualmente existe uma grande variedade de métodos empregados para fusão nos mais diversos domínios de aplicação [Nakamura et al., 2007a]. Considerando a existência de confusão com relação a terminologia de fusão de sensores/informação/dados devido a sua aplicação em diversas áreas, será adotada a seguinte definição de Hall & Llinas [1997] para o termo "fusão" pelo restante deste documento:

Definição 3 *Combinação de dados de múltiplos sensores, e informações relacionadas providas por base de dados associadas, para alcançar maior exatidão e inferências mais específicas do que seria possível usando apenas um único sensor.*

Os tipos de fusão de sensores existentes podem ser classificados de acordo com várias características, como: a relação entre os dados de entrada, o nível de abstração em que é feita a combinação dos dados, os tipos de sensores envolvidos, bem como a teoria base por trás do processo de fusão: probabilidade, teoria da evidência, teoria da possibilidade ou lógica *fuzzy* [Bloch, 1996].

De acordo com a relação entre os dados a serem combinados, a fusão pode ser dividida nas seguintes categorias, como ilustrado pela Figura 3.3:

- Complementar: quando os dados a serem combinados representam diferentes aspectos de um todo, onde uma informação mais completa e não redundante pode ser obtida através da fusão (caso da fusão audiovisual para reconhecimento

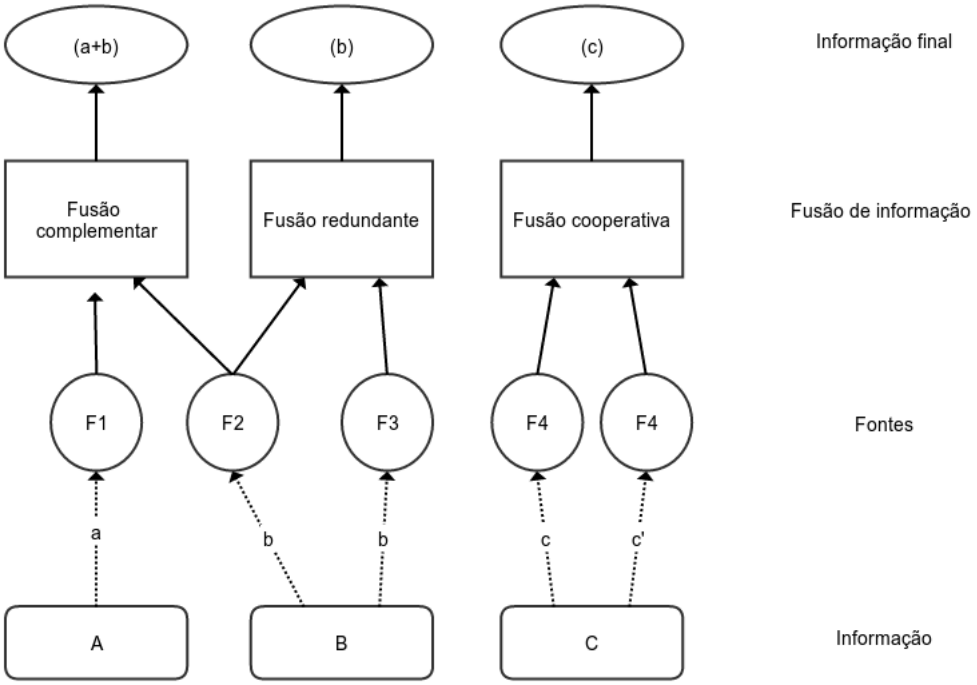


Figura 3.3. Categorias de fusão de acordo com a relação entre as fontes. Imagem traduzida de Nakamura et al. [2007a].

de voz, onde a voz e a configuração da boca de uma pessoa podem ser usadas, por exemplo).

- Redundante: quando dados similares obtidos de fontes diferentes são combinados a fim de torná-lo mais exato (como na localização via GPS, por exemplo, onde a posição é corrigida levando em conta o tempo de resposta de vários satélites);
- Cooperativa: quando dados podem ser combinados para gerar um mais complexo, que melhor representa a realidade, do que os dados individuais (como no caso da estereoscopia);

Outra classificação dos métodos de fusão é referente ao nível de abstração em que a integração dos dados ocorre:

- Fusão de dados (fusão de baixo nível): os dados são combinados diretamente se os sensores que os medem o façam para o mesmo fenômeno físico, como o som capturado por microfones, cujas diferenças de fase entre os sinais podem ser exploradas por um *beamformer*[10];

[10]Técnica de processamento de sinais em que dois ou mais sinais provenientes de sensores diferentes, cuja posição relativa é conhecida, são combinados, gerando interferência construtiva e destrutiva, resultando em um sinal que privilegia ondas vindas/emitidas de dada direção.

- Fusão de atributos (fusão de nível intermediário): envolve a extração de características representativas dos dados obtidos dos sensores e sua combinação (através de concatenação, por exemplo) em um único vetor de atributos. Para dados audiovisuais é utilizada a siga FAV (*Feature AV*) neste texto;

- Fusão de decisões (fusão de alto nível): combinação das informações de cada sensor após uma etapa de classificação ou estimação para geração de atributos de uma identidade, como localização, categoria, etc. Exemplos de técnicas de fusão desta classe incluem métodos de ponderação, técnicas baseadas em votação e variantes. Para dados audiovisuais é utilizada a sigla MAV (*Meta AV*) neste texto.

A fusão de atributos tem a vantagem de explorar as variações entre as características dos dados de cada sensor, provendo um microgerenciamento temporal dos dados. Por outro lado, a dimensão do vetor de características final se apresenta como uma possível dificuldade, motivo pelo qual técnicas de redução de dimensionalidade como PCA (*Principal Component Analysis*) [Bishop, 2007] são largamente utilizadas, beneficiando métodos que sofreriam desta limitação, como um HMM [Rabiner & Schafer, 2007]. Outras técnicas, como uma SVM [Bishop, 2007], que pode utilizar funções de *kernel*, dispensam tal tratamento, já que podem lidar com a redução de dimensão internamente.

A fusão neste nível de abstração tem a vantagem de ter um esquema de fácil implementação, todavia, há a restrição da necessidade de sincronização dos dados provenientes dos sensores, através de *downsampling* ou *upsampling*.

A fusão a nível de decisão envolve a combinação das probabilidades ou verossimilhanças (geradas por classificadores unimodais para cada sinal, por exemplo) baseada em algum esquema de ponderação associado à confiança dos dados. As vantagens deste modelo são ser capaz de ignorar características ruidosas ou pouco descritivas na etapa de fusão, dando maior peso às fontes de dados com menor incerteza, e utilizar amostras com vetor de atributos de menor dimensão, tornando a etapa de treinamento mais rápida.

Técnicas de meta-aprendizado e *ensemble* se encontram nesta categoria, como *bagging*, *boosting* e *stacking* (tipo de bagging). No *bagging* o treino é dividido em n conjuntos de treinamento amostrados uniformemente com reposição, cada um desses conjuntos é usado para treinar um classificador fraco[11], como uma árvore de decisão. No teste, a classificação final é feita aplicando a amostra a cada um desses classificadores

[11] Classificador cujo o desempenho é pouco melhor que um classificador aleatório.

fracos e combinando o resultado de cada um através de técnicas de votação ou maioria [Kittler et al., 1996], por exemplo. O classificador *Random Forest* [Breiman, 2001] pertence a esta categoria.

No *boosting*, são feitos treinos sucessivos com amostras classificadas erroneamente com o intuito de aprimorar seu resultado para amostras difíceis de classificar. O classificador AdaBoost [Alpaydin, 2009] pertence a esta categoria.

No *stacking* ou *stacked generalization*, o vetor de atributos de cada modalidade é classificado separadamente, gerando uma distribuição discreta. As distribuições são a seguir concatenadas em um novo vetor de atributos, servindo como entrada para uma nova etapa de classificação [Dzeroski & Zenko, 2004; Alpaydin, 2009].

Um critério importante de escolha entre as abordagens é a correlação entre as modalidades: uma fusão a nível de decisão é ideal quando as modalidades não são correlacionadas ou há grande variação na degradação das informações de cada sensor, assim cada entrada é processada independentemente da outra. De maneira inversa, uma fusão de atributos é efetiva quando as modalidades são fortemente correlacionadas. Entretanto, dado o problema da dimensionalidade, uma fusão de decisões pode superar uma de atributos mesmo se os dados forem muito correlacionados [Sargin & Yemez, 2007].

Dependendo da situação, nenhuma das duas opções pode oferecer uma solução ótima sozinha (especialmente quando as modalidades tem mistura de dados correlacionados e não-correlacionados), motivo pelo qual há a existência de estratégias híbridas, com o intuito de tirar proveito da velocidade de processamento de técnicas de baixo nível para dados com pouco ruído ou da capacidade seletiva de técnicas de alto nível para para dados pouco confiáveis, ao custo de tempo de processamento adicional [Stork et al., 1992; Neti et al., 2001].

Nesta pesquisa foram usados como descritores auditivos a FT, descritor espectral com bom poder discriminativo, BoW de MFCCs, uma representação compacta e discriminatória de sinais de áudio e um descritor construído a partir do tempo de decaimento do sinal de áudio, devido à grande diferença do tempo de decaimento das amostras dependendo do objeto.

Como descritor visual foi adotado o BASE [Nascimento et al., 2013], por ser um descritor cuja aquisição tem baixo custo computacional, codificado como um BoW, provendo uma representação compacta e intuitiva.

Na classificação foram usados dois classificadores, o *Random Forest* [Breiman, 2001] e a Regressão Logística [Bishop, 2007], o primeiro por ser um classificador *ensemble* robusto a *outliers* e estado-da-arte em classificação, o segundo por ser um classificador linear, necessário em uma das abordagens para fusão, com desempenho

superior a uma Regressão Linear simples.

Capítulo 4

Metodologia

Este capítulo aborda a metodologia adotada para identificação e extração de atributos de áudio e imagem e sua combinação. É descrita a integração e coordenação do módulo auditivo e módulo visual, além da etapa de fusão, que combina a saída dos anteriores.

4.1 Visão geral

O sistema recebe como entrada um sinal de áudio de duração limitada e uma imagem RGB acompanhada de sua respectiva imagem de profundidade, onde assume-se que as imagens são de mesma resolução e registradas[1]. A seguir, cada modalidade sensorial é processada separadamente, seguindo etapas similares, são elas:

1. Pré-processamento dos dados;
2. Extração de descritores;

Os descritores são combinados em um vetor de atributos na etapa de fusão, usado como entrada na etapa seguinte, de classificação. O diagrama da Figura 4.1 provê uma visão geral do sistema. Detalhes do funcionamento de cada módulo e a motivação por trás de sua modelagem se encontram nas próximas seções, com a descrição do módulo auditivo na Seção 4.2, módulo visual na Seção 4.3 e fusão audiovisual na Seção 4.4.

[1]Há uma relação biunívoca entre os *pixels* de cada imagem. Dado um pixel (x,y), sua intensidade é I(x,y) e sua profundidade D(x,y).

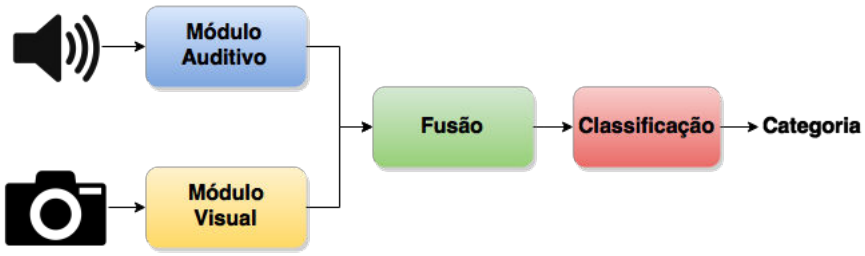


Figura 4.1. Diagrama de fluxo do sistema. Adquiridas amostras de imagem (RGB e de profundidade) e áudio do objeto, essas são processadas separadamente seguindo o fluxo de cada módulo.

4.2 Módulo Auditivo

O módulo auditivo é o responsável pela caracterização sonora dos objetos. Para isto é analisado um trecho de áudio onde se sabe que há som produzido pelo objeto, a partir do qual são extraídos os descritores. Este processo envolve duas etapas fundamentais: recorte do sinal e extração de descritores. A Figura 4.2 contém um diagrama do funcionamento do módulo auditivo.

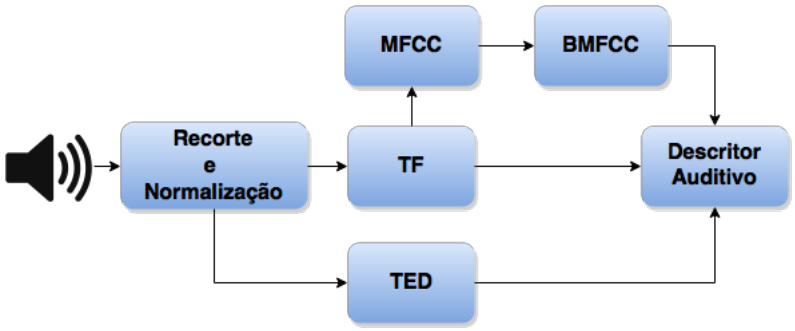


Figura 4.2. Diagrama de fluxo do módulo auditivo, contendo a etapa de pré-processamento (recorte e normalização do sinal) e extração dos descritores.

4.2.1 Recorte

A primeira fase, recorte do sinal, é feita de forma simples, através da variação de energia no domínio do tempo. Seja $s_j(i)$ o valor da amostra i do sinal s em uma janela deslizante retangular j de tamanho N, a energia E [Peeters et al., 2011] do sinal na janela é dada por

$$E_j = \sum_i s_j^2(i). \tag{4.1}$$

Dada a energia em cada *frame*[2], com duração de 10ms e passo de 1ms, o *frame* o em que a energia é máxima é considerada como o início do choque:

$$o = \arg\max_{j \in J}(E_j), \tag{4.2}$$

onde J é o conjunto ordenado de *frames* ao longo do sinal.

O último *frame* u considerado parte do recorte é dado pelo primeiro *frame* após o cuja energia tem um decaimento de 100 dB[3] em relação a o, assim a condição:

$$E_{u_{50}} \geq \left(\frac{E_o}{10^{50}}\right). \tag{4.3}$$

Por fim, as amostras no intervalo do recorte são a seguir normalizadas pelo máximo:

$$s = \frac{s}{\max s(i)}. \tag{4.4}$$

Um exemplo do procedimento é ilustrado na Figura 4.3.

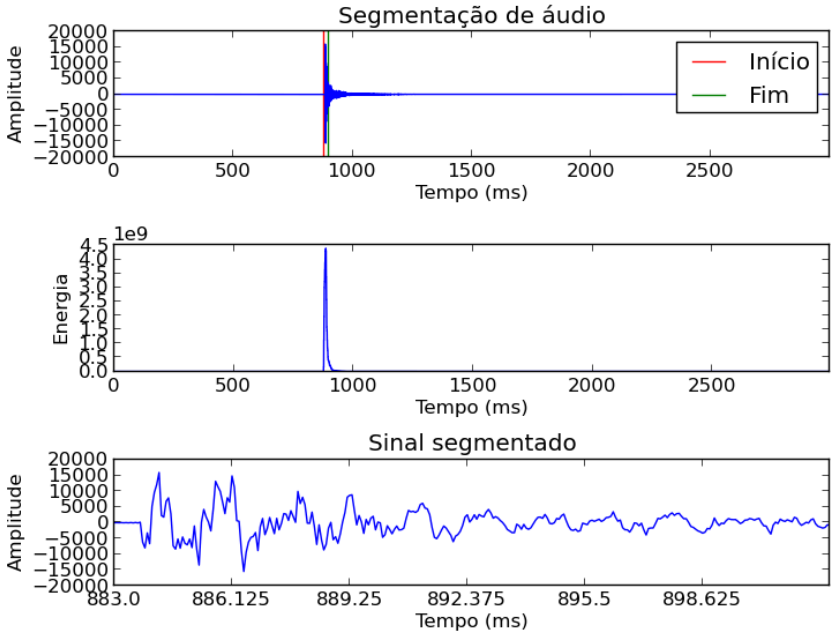


Figura 4.3. Exemplo do recorte de sinal de um trecho de áudio da base de dados coletada para este trabalho.

[2]Trecho do sinal, aqui usado como referência ao sinal na janela deslizante.
[3]Valor definido empiricamente após inspeção manual da base de dados usada.

4.2.2 Descritores

Três descritores foram escolhidos para a representação de áudio, um temporal e dois espectrais: medidas do envelope de energia do sinal, coeficientes MFCC codificados em uma abordagem *Bag-of-Words* e coeficientes da Transformada de Fourier.

4.2.2.1 *Time for Energy Decay* (TED)

Este descritor foi criado a partir do tempo de reverberação (RT) de Sabine [Sabine, 1906]. O descritor, a partir daqui chamado de TED(*Time for Energy Decay*), é assim referido pois outros fatores podem influenciar o envelope de energia do sinal além da reverberação interna dos objetos. Seja TED_N o tempo para que a energia de um sinal decaia NdB, seguindo a mesma estrutura da equação 4.3, o último *frame* para cada nível de decaimento N obedece a condição

$$E_{u_N} \geq \left(\frac{E_o}{10^{N/10}} \right). \tag{4.5}$$

O TED_N, em milissegundos, é então estimado:

$$TED_N = \frac{10^3 (I_{u_N} - I_o)}{f} ms, \tag{4.6}$$

onde I_u é o índice da primeira amostra do *frame* especificado (u, neste caso) e f é a frequência de amostragem do sinal.

Por fim o descritor final, TED, é formado pela concatenação dos diferentes TED_N em um vetor:

$$TED = [TED_5, TED_{10}, \cdots, TED_{50}]. \tag{4.7}$$

Foram utilizados valores de decaimento no intervalo [5,50] com passo de 5.

O uso deste descritor foi introduzido após verificar experimentalmente que, dependendo das características do objeto, há grande diferença no formato de seus envelopes de energia, de acordo com a premissa de Wildes & Richards [1988]. Um objeto feito de isopor, que gera um sinal de baixa amplitude, tem TED_{30} em torno de 5 a 6 vezes o de outros materiais, como madeira, por exemplo. Isto ocorre devido à SNR (*Signal-to-noise Ratio*): mesmo em seu pico, a energia gerada pelos objetos ainda é pouco maior em relação ao ruído ambiente, como o recorte do sinal depende desta relação, de acordo com a equação 4.5, análoga à 4.3, a tendência é que sinais com baixa amplitude tenham maior tempo de reverberação.

4.2.2.2 BoW MFCC (BMFCC)

Este descritor se baseia nos MFCC [Davis & Mermelstein, 1980] para criar uma representação em alto nível do sinal. Para cada *frame* obtido na etapa de recorte, um conjunto de coeficientes é gerado através do uso de filtros triangulares espaçados na escala Mel, cuja conversão para a escala Hertz é

$$M(f) = 1125 \ln(1 + f/700), \tag{4.8}$$

ilustrada na Figura 4.4.

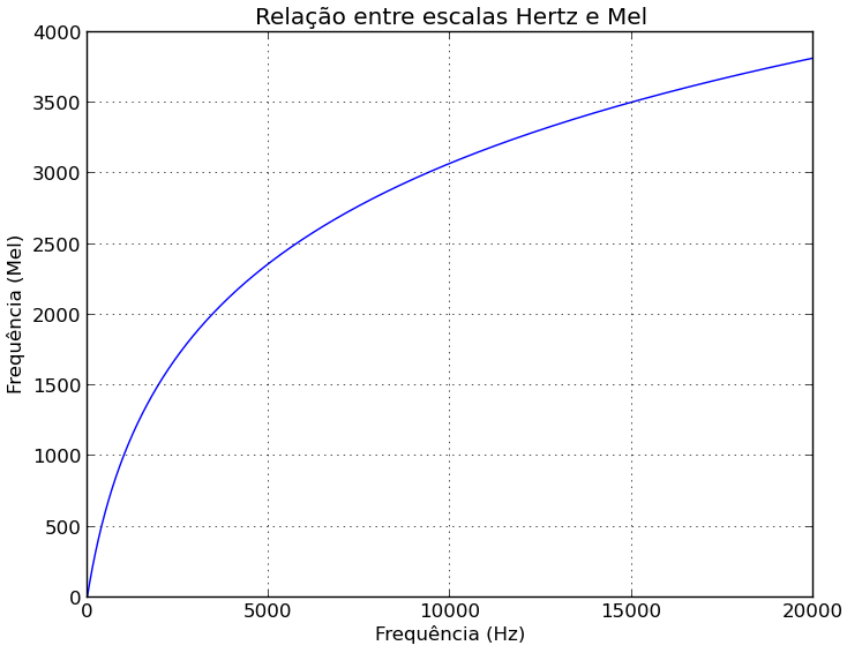


Figura 4.4. Relação entre as escalas Hertz e Mel.

Para geração dos coeficientes, o sinal em cada *frame* j é primeiramente convertido para o domínio da frequência através da Transformada de Fourier, DFT (*Discrete Fourier Transform*), neste caso:

$$S_j(k) = \sum_{n=1}^{N} s_j(n) e^{-2i\pi kn/N}, \tag{4.9}$$

onde $1 \leq k \leq K$ e K é o tamanho da transformada. A seguir são aplicados ao espectro filtros *bandpass* triangulares logaritmicamente espaçados (seguindo a escala Mel), a

partir dos quais são computados os MFCCs:

$$MFCC(i) = \sum_{k=1}^{K} X_K \cos\left[i\left(k - \frac{1}{2}\right)\frac{\pi}{20}\right], \tag{4.10}$$

onde i é o i-ésimo coeficiente MFCC do *frame* e X_k o logaritmo da energia do k-ésimo filtro. Aqui, 26 filtros e 13 coeficientes MFCC foram usados, valores estipulados após análise experimental.

Devido à falta de estrutura no sinal, ao contrário do contexto em que os MFCCs geralmente são aplicados conjuntamente com modelos sequenciais como HMM [Rabiner & Schafer, 2007], é necessário gerar uma representação global para a amostra. Ainda que seja possível usar operadores estatísticos como média e variância dos coeficientes [Bloch, 1996], há perda de informação relevante, optando-se por adotar uma estratégia BoW, modelo adequado e intuitivo para representação de sons ambiente [Aucouturier et al., 2007].

Os coeficientes dos *frames* de algumas amostras são utilizados em etapa prévia para geração do vocabulário através do algoritmo de clusterização *k-means* [Kaufman & Rousseeuw, 1990], dada a sua rápida convergência e suporte a valores de ponto flutuante, como os MFCCs.

Dado o vocabulário, cada conjunto de coeficientes MFCC i é atribuído à "palavra" que minimiza a distância euclidiana:

$$Palavra(MFCC_i) = \arg\min_{p \in P} \left(\sqrt{\sum_{k=1}^{13} (MFCC_i(k) - P_p(k))^2} \right), \tag{4.11}$$

onde $MFCC_i(k)$ e $P_p(k)$ se referem ao k-ésimo coeficiente do i-ésimo conjunto de MFFCs e p-ésima "palavra" do vocabulário.

Seja O o vetor com o número total de ocorrências das |P| "palavras"

$$O = O_1, O_2, \cdots, O_{|P|}, \tag{4.12}$$

este é normalizado pelo máximo, gerando o descritor BoW MFCC (BMFCC):

$$BMFCC = \frac{O}{max(O)}. \tag{4.13}$$

O uso dos MFCCs no descritor se deve ao seu bom desempenho em tarefas de classificação de sons gerados por objetos passivos [Nakamura et al., 2007b; McGibney et al., 2012] bem como na caracterização e classificação de áudio de uma forma genérica

em vários domínios de aplicação [Richard et al., 2013; Chu et al., 2009], além de sua representação compacta do espectro audível.

4.2.2.3 *Compact Fourier Transform* (CFT)

Seguindo a abordagem de Biondi et al. [2014] para caracterização de sinais de áudio impulsivos, como os usados neste trabalho, um descritor a partir dos coeficientes da Transformada de Fourier é utilizado. Dados os coeficientes F_k de cada frequência k, como os computados pela equação 4.9:

$$F_k = \sum_i S_i(k), \tag{4.14}$$

onde cada valor F_k corresponde à contribuição da frequência k ao longo dos *frames* i. Para computar o descritor estes valores são concatenados em um vetor referido como CFT:

$$CFT = F_0, F_1, .., F_n, \tag{4.15}$$

e normalizados pelo máximo:

$$CFT = \frac{CFT}{max(CFT)}. \tag{4.16}$$

A motivação para uso deste descritor é sua simplicidade, rapidez de extração e representação do espectro em uma escala diferente da usada pelos MFCCs, além de se mostrar como um bom descritor para a categoria de sinais usados neste trabalho.

4.2.2.4 Descritor final

Por fim o descritor final de áudio é obtido combinando os descritores globais citados anteriormente em um único vetor de atributos A_a:

$$A_a = [TED, BMFCC, CFT]. \tag{4.17}$$

4.3 Módulo visual

O módulo visual é responsável pela segmentação do objeto de interesse e extração de descritores visuais do objeto de interesse, descritas a seguir. A Figura 4.5 contém um diagrama do funcionamento do módulo visual.

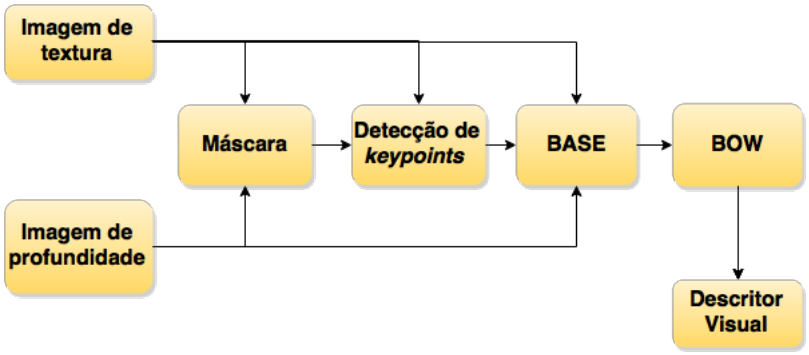
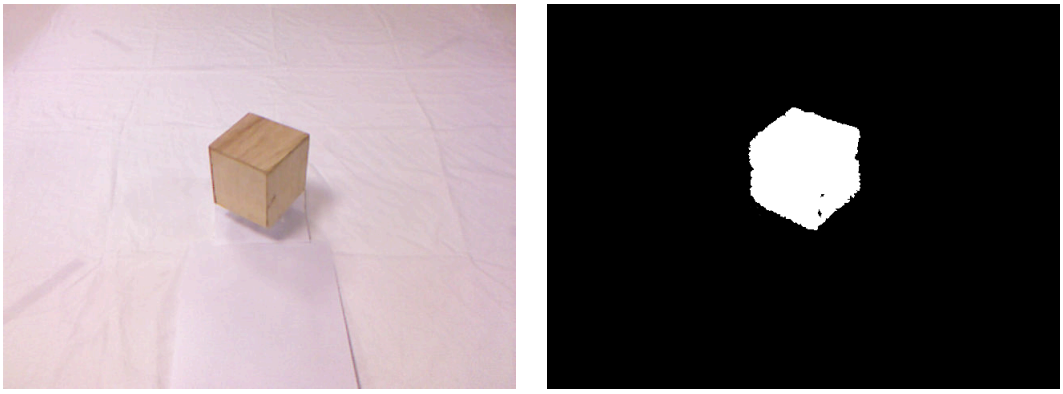


Figura 4.5. Diagrama de fluxo do módulo visual, contendo a etapa de pré-processamento (criação de máscara e detecção de *keypoints*) e extração dos descritores.

4.3.1 Segmentação

Nesta etapa deseja-se gerar uma região de interesse para que sejam detectados pontos de interesse do objeto, cujos descritores são extraídos na etapa seguinte. Tendo como objetivo a simplicidade e as configurações experimentais adotadas, optou-se por uma abordagem simplificada do trabalho de Bjorkman & Kragic [2010] baseada no SAC (*SAmpled Consensus initial alignment*) [Fischler & Bolles, 1981].

O plano da superfície no qual o objeto se encontra é reconhecido por consenso pelo modelo (planar), e tem sua cor alterada para preto. Os demais pontos da nuvem de pontos, pertencentes ao objeto, são pintados na cor branca. A seguir a nuvem é projetada em um plano a partir da origem do sistema de coordenadas, gerando uma máscara a ser usada na detecção de pontos de interesse. A Figura 4.6 contem um exemplo de máscara criada.



(a) Objeto (b) Máscara

Figura 4.6. Exemplo de máscara de uma amostra.

4.3.2 Descritor

Para detecção de pontos de interesse é utilizado o detector FAST (*Features from Accelerated Segment Test*)[Rosten et al., 2010], isto se deve a sua capacidade de extrair uma boa quantidade de pontos com baixo custo computacional, fator atrativo para tarefas executadas em tempo real, como desejado neste trabalho. O estudo comparativo feito por Nascimento et al. [2013] aponta o bom desempenho do método juntamente com descritores binários para a tarefa de reconhecimento.

A representação visual de cada objeto é feita utilizando o descritor BASE (*Binary Appearance and Shape Elements*) [Nascimento et al., 2013] para representar seus pontos de interesse. A escolha deste descritor é embasada em dois motivos: utiliza informações de textura e geometria, sendo robusto à falta de texturas e variações na iluminação, e, sendo um descritor binário, tem todas as vantagens dos descritores desta categoria mencionados na seção anterior: rapidez, representação compacta e baixo custo computacional.

A criação do descritor tem três etapas:

- Extração dos atributos de textura da vizinhança do ponto de interesse;
- Extração dos atributos geométricos da nuvem de pontos ao redor do ponto de interesse de acordo com suas normais;
- Combinação dos atributos em um vetor binário;

Seja k um *keypoint* da imagem RGB representada em escala de cinza e um *patch* P ao redor de k, são amostrados pares de *pixels* em P de acordo com o padrão de amostragem do descritor, ilustrado na Figura 4.7.

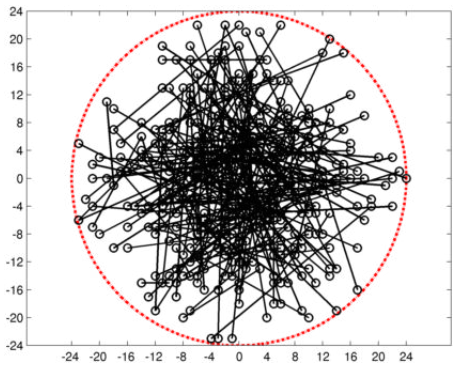


Figura 4.7. Padrão de amostragem do descritor BASE. Imagem retirada de Nascimento et al. [2013].

Seja $C = \{(x_i, y_i), i = 1, ..., 256\}$ o conjunto dos pares de *pixels* amostrados de P segundo o padrão apresentado. Para cada par de C, similarmente a Calonder et al. [2010], é comparada a intensidade dos *pixels*:

$$\tau_a(x_i, x_i) = \begin{cases} 1 & se \quad p_i(x_i) < p_i(y_i) \\ 0 & c.c. \end{cases}, \tag{4.18}$$

onde $p(x)$ é a intensidade do *pixel*.

A seguir, a extração da geometria $\tau_g(x_i, y_i)$ de cada par selecionado é baseada em duas propriedades:

- Deslocamento d da normal (calculado através do produto escalar entre as duas normais, verificando-se se excede determinado limiar l);
- Convexidade c da superfície entre os dois pontos, sendo o valor de c negativo caso a superfície entre os pontos seja convexa e positivo caso contrário.

$$\tau_g(x_i, x_i) = \begin{cases} 1 & se \quad d < l \wedge c < 0 \\ 0 & c.c. \end{cases}. \tag{4.19}$$

As informações de aparência e geometria de cada um dos pares $(x_i, y_i) \in C$ são então combinadas:

$$f(x_i, y_i) = \begin{cases} 1 & se \quad \tau_a(x_i, y_i) \vee \tau_g(x_i, y_i) \\ 0 & c.c. \end{cases}. \tag{4.20}$$

O descritor de cada ponto de interesse é obtido concatenando o valor binário de $f(x_i, y_i)$ para cada um dos 256 pares em C, podendo ser comprimido em inteiros de 8 bits (32, no caso), por exemplo.

Com os descritores locais computados, assim como no módulo de áudio, é criada uma representação global usando a técnica BoW [Csurka et al., 2004]. Os descritores de algumas amostras são utilizados em etapa prévia para geração do vocabulário através do algoritmo de clusterização *k-medoids* [Kaufman & Rousseeuw, 1990], devido à possibilidade de uso de diferentes métricas de distância, como a de Hamming, mais adequada à representação binária usada.

Dado o vocabulário, cada descritor d é atribuído à "palavra" p que minimiza a distância de Hamming:

$$Palavra(d) = \arg\min_{p \in P} \left(\sum_{i=1}^{256} (d(i) \oplus P_p(i) \wedge 1) \right), \tag{4.21}$$

onde $d(i)$ e $P_p(i)$ se referem ao i-ésimo bit do descritor BASE e da p-ésima "palavra" e $\oplus$ é o operador binário XOR.

A "palavra" com menor distância para o descritor é considerada como uma ocorrência. Seja O o vetor com o número de ocorrências das |P| "palavras"

$$O = O_1, O_2, \cdots, O_{|P|}, \tag{4.22}$$

este é normalizado pela norma L1, gerando o vetor de atributos visual A_v:

$$A_v = \frac{O}{\sum_{i=0}^{|P|} |O_i|}. \tag{4.23}$$

4.4 Fusão audiovisual

Sabendo que ambas as modalidades tem uma representação global para cada amostra, o processo de fusão pode ser visto como um problema análogo ao de combinação de atributos para classificação ou combinação de múltiplos classificadores através de técnicas de meta-aprendizado.

Para combinação das informações auditivas e visuais dos objetos, duas abordagens clássicas em fusão de sensores foram implementadas: fusão de atributos e fusão de decisões.

A primeira abordagem consiste na combinação do vetor de atributos proveniente de cada modalidade sensorial por concatenação. De modo genérico, seja A_i o vetor de cada modalidade i, o vetor de atributos final é

$$A_f = A_i, \cdots, A_n. \tag{4.24}$$

Logo, o vetor de atributos audiovisuais para esta abordagem, a partir daqui referida como FAV, é:

$$A_{av} = A_a, A_v. \tag{4.25}$$

Na segunda abordagem é usada a abordagem de meta-aprendizado *stacking* [Dzeroski & Zenko, 2004]. Seja $P(i)$ o vetor da distribuição de probabilidade da modalidade i, o vetor de atributos audiovisuais para esta abordagem, a partir daqui referida como MAV, é:

$$A_{av} = P(a), P(v). \tag{4.26}$$

A Figura 4.8 contém um diagrama do funcionamento da fusão audiovisual.

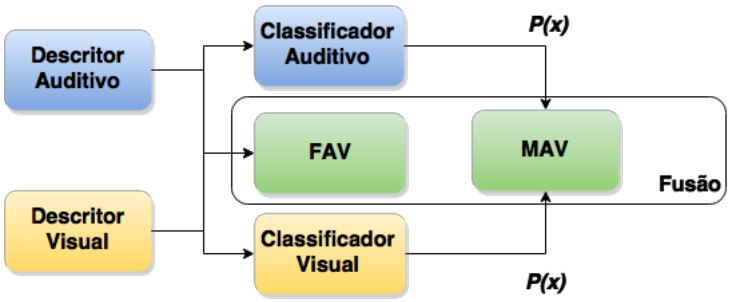


Figura 4.8. Diagrama de fluxo da fusão audiovisual. FAV e MAV correspondem a *Feature AV* e *Meta AV*, respectivamente, as duas abordagens usadas para fusão audiovisual.

Uma das motivações por trás do uso destas abordagens é a comparação de desempenho entre os dois principais conceitos em fusão de sensores, fusão de atributos e decisões, além da verificação da hipótese de que uma abordagem baseada em fusão de decisões se sobressai sobre uma de fusão de atributos em um cenário de independência entre as modalidades [Sargin & Yemez, 2007], como suposto devido ao método de captura adotado na fase de experimentos.

Outro motivo desta escolha é a discrepância entre os domínios de cada modalidade sensorial. Diferentemente de abordagens em que a representação dos dados se dá em um domínio comum, não há associação direta entre os componentes de cada modalidade sensorial, caso em que uma representação/fusão dos dados em baixo nível não é intuitiva ou prática.

Capítulo 5

Resultados experimentais

Este capítulo aborda os resultados experimentais desta pesquisa. Na Seção 5.1 detalhes da implementação são descritos. A Seção 5.2 contém detalhes da base de dados capturada para teste. Na Seção 5.3 é descrita a validação do modelo e, por fim, na Seção 5.4 cada experimento executado é descrito e analisado. Comentários gerais constam na Seção 5.5.

5.1 Implementação

O módulo auditivo foi desenvolvido em Python usando as bibliotecas Scipy [Jones et al., 2001] e Numpy [van der Walt et al., 2011] devido sua fácil prototipação, ampla disponibilidade de métodos para análise de sinais e rapidez no processamento devido ao uso de sub-rotinas compiladas em C.

No módulo visual foram utilizadas a implementação em C++ de Nascimento et al. [2013] para o descritor BASE e as bibliotecas OpenCV [Bradski, 2000] e PCL [Rusu & Cousins, 2011] para pré-processamento dos dados devido às implementações robustas de algoritmos e estruturas de dados em visão computacional, como detecção de pontos de interesse e segmentação de nuvens de pontos.

A fusão audiovisual e classificação foram desenvolvidas em Python e integradas com os módulos anteriores usando o sistema *opensource* ROS [Quigley et al., 2009]. Foi usada a biblioteca Scikit-Learn [Pedregosa et al., 2011], que contém implementações estáveis de classificadores, como o *Random Forest* [Breiman, 2001], usado nesta pesquisa.

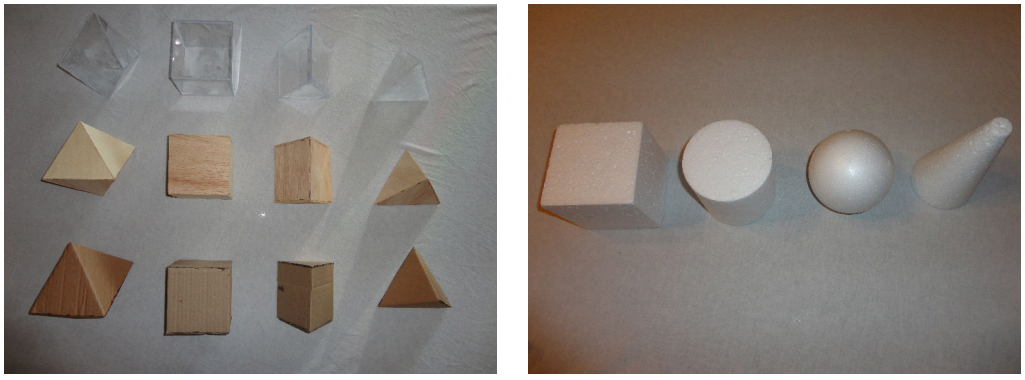
O sistema foi implementado usando ROS Hydro, Python 2.7.3 e compilador C++ g++ 4.6.3 em um ambiente Ubuntu 12.04 instalado em uma máquina com processador Intel Core i5 e 4GB de RAM.

5.2 Base de dados

Robôs e sistemas autônomos que operam em ambientes com presença humana idealmente devem ser capazes de reconhecer objetos de uso comum, cuja composição normalmente varia entre um certo conjunto de materiais, como plástico, madeira, papelão e metal.

Embora inúmeros *datasets* para reconhecimento de objetos estejam disponíveis para uso, como as bases Aloi [Geusebroek et al., 2005], Caltech-101 [Fei-Fei et al., 2004], Caltech-256 [Griffin et al., 2006], Coil [Nene et al., 1996], ImageNet [Deng et al., 2009] e *RGB-D Object Dataset* [Lai et al., 2011], não há conhecimento da existência de *datasets* audiovisuais disponíveis para reconhecimento de objetos. Um *dataset* audiovisual relacionado a esta tarefa foi disponibilizado por Pieropan & Salvi [2014], mas seu propósito se restringe à identificação de ações humanas associadas a objetos, não propriedades dos objetos em si. Por este motivo foi selecionado um subconjunto destes materiais e criada uma base contendo amostras de auditivas e visuais de 16 objetos com diferentes configurações de material e geometria.

A base, ilustrada na Figura 5.1, é composta por poliedros ocos feitos de madeira, plástico e papelão. Em cada material, 4 geometrias diferentes foram confeccionadas: cubo, prisma de base triangular, tetraedro e octaedro. Quatro objetos adicionais sólidos feitos de Isopor foram usados: cubo, cilindro, esfera e cone.



(a) Objetos ocos (b) Objetos sólidos

Figura 5.1. Objetos usados para análise de impacto de geometria e material no desempenho de cada modalidade no problema de reconhecimento.

O objetivo da configuração deste conjunto foi analisar o impacto da geometria e material separadamente e em conjunto no reconhecimento de objetos usando múltiplas modalidades sensoriais.

Foram confeccionados objetos nos seguintes materiais e características:

- Madeira de compensado com 4mm de espessura;
- Plástico isopropeno com 2mm de espessura;
- Papelão com 4mm de espessura.

A base, que totaliza 1600 amostras de cada modalidade sensorial, inclui:

- 100 imagens RGB-D registradas por objeto, com resolução de 640x480, capturadas com o sensor Kinect;
- 100 amostras de áudio por objeto, contendo 4 faixas amostradas a 16kHz com 32 bits de resolução capturada como o sensor Kinect.

Apesar de o problema de reconhecimento de objetos de uso cotidiano não ter sido abordado de forma direta e imediata neste estudo, a estruturação da base de dados permite um estudo mais sistemático e controlado de influências do material e da geometria sobre os sons gerados na interação com o robô, um dos trabalhos futuros desta pesquisa.

5.2.1 Aquisição de dados

As capturas foram feitas em laboratório silencioso, sem isolamento acústico ou controle de reverberação, sob condições estáveis de iluminação artificial. O ambiente de captura é ilustrado na Figura 5.2. As amostras foram capturadas com cada objeto em cima de uma base distante a $50\sqrt{2}$ cm do sensor, apontado diretamente para o objeto. A Figura 5.3 ilustra o procedimento de captura das amostras.

Figura 5.2. Ambiente de captura das amostras da base. As amostras foram capturadas na mesa ao centro do laboratório a uma distância fixa de 50 centímetros do sensor.

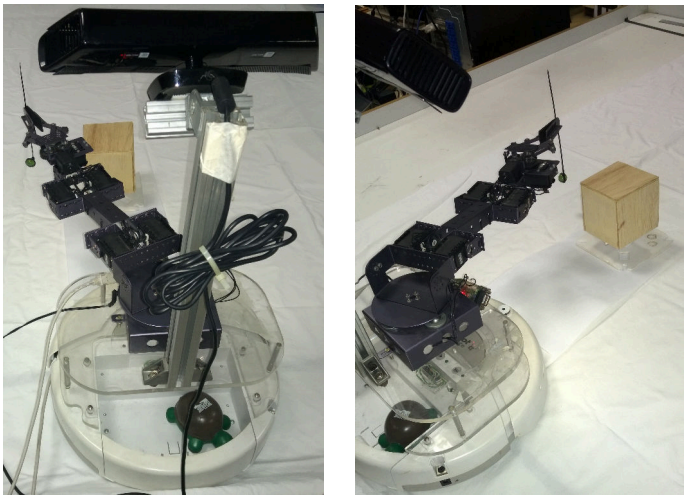


Figura 5.3. Aquisição de amostras. O sensor fica posicionado no topo do robô, com um braço robótico interagindo ativamente com o objeto para aquisição de áudio.

5.2.1.1 Aquisição auditiva

A maioria dos objetos do cotidiano não gera som por si só, por este motivo, a aquisição de auditiva é feita de forma ativa, neste caso, através da interação mecânica de um braço robótico com o objeto. O braço robótico usado é um *CrustCrawler AX-12 Smart Arm*, com três graus de liberdade e outros dois no manipulador controlado por sete servo motores Dynamixel AX-12.

Além da disponibilidade do equipamento em laboratório, o uso do braço foi projetado de acordo com o trabalho de Nakamura et al. [2007a]; Sinapov [2013], onde a aquisição é feita através da interação do robô com o objeto através de movimentos pré-configurados, ou McGibney et al. [2012], onde o áudio do objeto é gerado a partir de seu choque com o chão a partir de uma altura fixa.

Outras formas de aquisição são possíveis, como captura da resposta ecóica de um pulso sonoro ou TSP (*Time-Stretched Pulse*) [Suzuki et al., 1995] gerado para este fim. Entretanto, para redução do escopo da pesquisa, essa análise permanece como um trabalho futuro.

A interação é feita através de movimentos pré-configurados simples do braço com o objeto e base imóveis. Após sinalização de que uma captura de áudio deve ser executada, o manipulador é estendido até uma posição fixa p a 10 cm do objeto.

O sinal de áudio começa a ser capturado, enquanto o braço é deslocado para uma posição p' alterando-se o ângulo do servo da base do braço robótico. Por inércia, o contrapeso, ilustrado na Figura 5.4, oscila, batendo no objeto e voltando a uma posição de repouso. Após um tempo fixo de três segundos, a captura é interrompida.

Figura 5.4. Contrapeso anexado ao manipulador do robô.

O ruído gerado pelo manipulador é minimizado por duas razões:

1. Apenas um servo é utilizado para gerar o choque enquanto o som é capturado: o da base. Isso ocorre para aumentar a inércia do contrapeso, gerando um choque com mais energia, e diminuir o número de servo motores acionados, diminuindo o ruído sonoro gerado.

2. O padrão polar do microfone é hipercardióide, o que minimiza o ruído de fontes próximas, como o servo da base. Como o servo motor está localizado o mais longe possível do ponto de impacto, foco da captura, o ruído sonoro gerado pelo servo é muito baixo;

Ainda que a frequência máxima de captura[1] de áudio do Kinect seja 16 kHz, considerada baixa, testes com um microfone com maior taxa de amostragem revelaram que grande parte da energia dos sinais se concentra na faixa até 8 kHz, exemplificado pela Figura 5.5. Esse fato aliado aos bons resultados em experimentos similares usando o Kinect [McGibney et al., 2012; Pieropan & Salvi, 2014] e a possibilidade de uso dos recursos providos pelo equipamento, fizeram com que o sensor fosse usado.

As figuras 5.6 e 5.7 ilustram o espectro de algumas amostras capturadas com o Kinect.

[1]Segundo o Teorema de Nyquist, para que um sinal analógico seja representado discretamente com mínimo de perda, a frequência de amostragem deve ser, no mínimo, o dobro da maior frequência do espectro desse sinal.

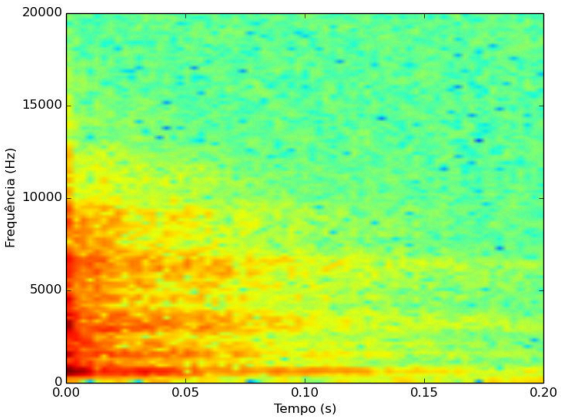
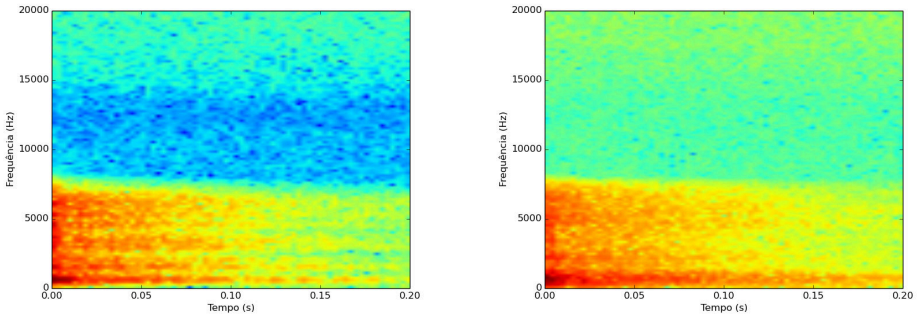
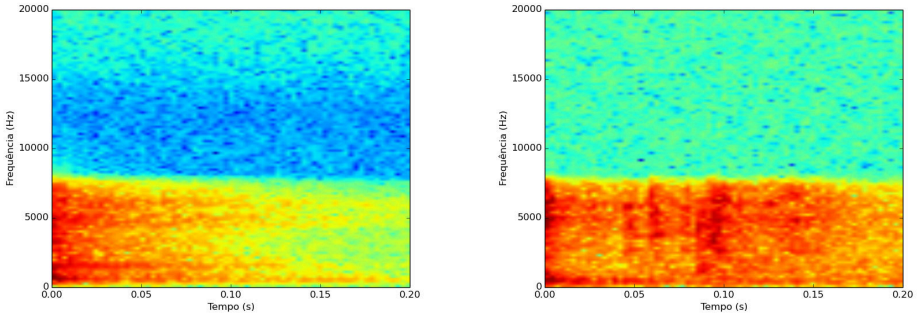


Figura 5.5. Espectrograma de um amostra do objeto cubo de madeira capturada com o microfone Audio Technica AT829, de padrão polar omnidirecional.



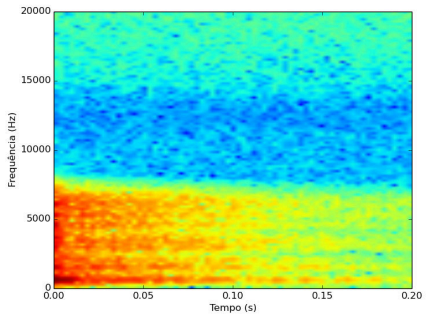
(a) Cubo de madeira (b) Cubo de papelão



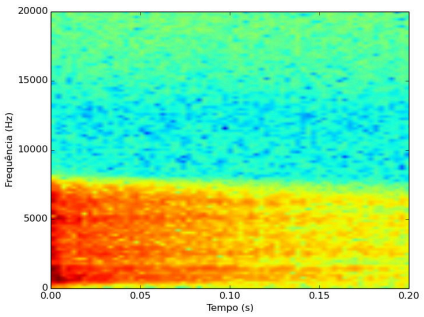
(a) Cubo de plástico (b) Cubo de Isopor

Figura 5.6. Espectrogramas de objetos de mesma geometria (cubo) e diferentes materiais.

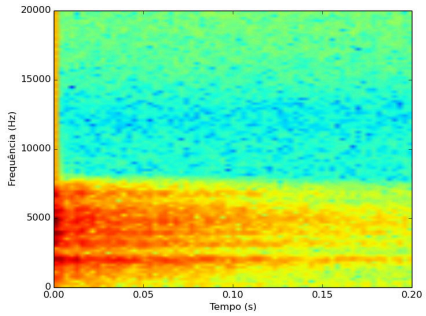
A diferença de energia ao longo da captura do objeto de Isopor comparada às demais se deve a baixa amplitude do sinal, que se confunde com o ruído ambiente, assim não há um pico de energia, como nas demais capturas.



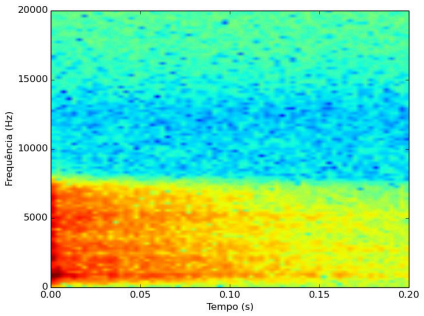
(a) Cubo de madeira



(b) Octaedro de madeira



(a) Tetraedro de madeira



(b) Prisma de madeira

Figura 5.7. Espectrogramas de objetos de mesmo material (madeira) e diferentes geometrias.

5.2.1.2 Aquisição visual

A aquisição visual, ao contrário da auditiva, não requereu interação mecânica, e foi executada com a base sob o objeto rotacionando a 59 rpm e o braço robótico em repouso fora do campo de visão do sensor. Foram capturados vídeos do objeto girando a 30 Hz, do qual foram ser extraídas as imagens de textura e profundidade (os *frames* do vídeo) usadas nos experimentos.

5.3 Validação dos experimentos

Preliminar aos experimentos, foi executada uma etapa de validação com dois objetivos: selecionar os atributos mais relevantes para cada modalidade e validar a etapa de classificação.

Os resultados exibidos foram produzidos usando o classificador *Random Forest* [Breiman, 2001], que exibiu os melhores resultados em análise prévia. Para a abordagem MAV, que executa duas etapas de classificação, a classificação de cada modalidade é executada por um *Random Forest* com 100 árvores de decisão e a classificação final é feita usando uma Regressão Logística [Bishop, 2007], um tipo de classificador linear, categoria sugerida nesta etapa por [Dzeroski & Zenko, 2004].

Para quantificar o desempenho, os experimentos de classificação foram realizados usando validação cruzada com 10 partições[2], onde os resultados exibidos são a média μ de acertos nas 10 partições e seu desvio desvio padrão σ ou sua respectiva AUC (*Area Under ROC Curve*) [Fawcett, 2006].

A AUC é uma medida com valor entre 0 e 1 e reflete a relação entre falsos positivos e falsos negativos na classificação, onde um classificador ideal obteria 1. Quanto mais similares os atributos interclasses são, maior a tendência de que a AUC fique em torno de 0.5, caso em que a abordagem não seria melhor que uma decisão aleatória.

Para geração do dicionário do BoW auditivo e visual e seleção de atributos, dez amostras de cada objeto (dez sinais de áudio e dez imagens de textura/profundidade) foram selecionadas, das quais descritores foram extraídos e os desempenhos de suas combinações avaliados.

A geração do dicionário auditivo ocorreu agrupando os descritores MFCC usando o algoritmo *k-means* [Kaufman & Rousseeuw, 1990]. O dicionário visual foi gerado agrupando os descritores BASE usando do algoritmo *k-medoids* [Kaufman & Rousseeuw, 1990], devido a natureza binária do descritor.

Os centroides e medoides resultantes cada algoritmo de *clustring* são as palavras do dicionário de cada modalidade. Foram testados tamanhos de dicionário no intervalo [2, 100] para a modalidade auditiva e tamanhos de 1000, 2000, 3000 e 4000 para a modalidade visual. A análise das silhuetas dos dicionários auditivos indicou que o que melhor representava os descritores MFCC era o dicionário com 20 "palavras". Os dicionários visuais passaram por outro processo de análise, onde verificou-se a capacidade de reconhecimento das amostras de validação por cada tamanho de dicionário, exibidas na forma de AUC na Tabela 5.1. Assim foram usado um dicionário auditivo com 20 "palavras" e um visual com 1000, respectivamente.

[2]A base de amostras é dividida em 10 partições disjuntas, com amostras selecionadas aleatoriamente sem reposição, das quais uma é utilizada como teste e o restante como treino para o classificador. O processo é executado 10 vezes, de tal forma que cada partição seja utilizada como teste exatamente uma vez.

Tamanho	AUC
1000	0.999
2000	0.984
3000	0.992
4000	0.972

Tabela 5.1. Valores AUC para dicionários visuais de diferentes tamanhos.

5.3.1 Seleção de atributos

Além dos descritores auditivos mencionados na metodologia, outros foram testados para verificar sua eficácia na tarefa, como coeficientes FT e MFCC sequenciais e WT [Mallat, 1989].

A motivação do teste com os dois primeiros descritores foi verificar o quanto a informação temporal implicitamente contida na sequência dos coeficientes tem impacto quando comparada com as versões compactas dessas representações, descritas na metodologia. Desejava-se ainda comparar o desempenho entre as duas escalas de representação, Mel e Hertz, e verificar se nesta tarefa os MFCCs se sobressaiam quando comparados aos coeficientes da FT, de mais rápida obtenção.

A motivação para teste com WT foi sua flexibilidade quanto à decomposição do sinal, através do uso de diferentes escalas e Wavelets mãe, além de bons resultados em tarefas envolvendo sons impulsivos, como no trabalho de Libal & Spyra [2014], feito no contexto de análise de reverberação balística. Nesta abordagem os coeficientes de aproximação da Wavelet em sete escalas diferentes são concatenados em um vetor de atributos, usado para classificação.

Na validação foram testadas duas Wavelets mãe: Daubechies com 1 (Haar) e 3 *vanishing points*. Os resultados de Wavelet aqui apresentados foram obtidos a partir desta última, apta a representar sinais com mais flexibilidade.

Como o tamanho do vetor depende da duração do recorte do sinal, o tamanho do maior vetor foi escolhido, sendo as demais amostras completadas com zero até este limite.

Para avaliar o desempenho de cada combinação, a AUC foi usada, cujos valores são apresentados na Tabela 5.2. A combinação destacada foi a adotada na metodologia.

Pelos valores AUC observa-se a facilidade dos descritores em discriminar as categorias. Dentre os descritores, apesar da eficiência, o uso dos descritores sequenciais aumentou consideravelmente o tempo de convergência dos modelos devido ao grande número de atributos, o que reforçou a escolha dos descritores TED, BMFCC e CFT para uso no descritor auditivo, além do desempenho desta combinação por si só.

Combinação	AUC	Combinação	AUC
C	0,98404	T + C	0,99240
F	0,99150	T + F	0,98963
F + C	0,98868	T + F + C	0,99309
W	0,97835	T + W	0,98288
W + C	0,98407	T + W + C	0,98851
W + F	0,99391	T + W + F	0,99201
W + F + C	0,99248	T + W + F + C	0,99203
B	0,92257	T + B	0,96364
B + C	0,98999	**T + B + C**	**0,99469**
B + F	0,99244	T + B + F	0,99203
B + F + C	0,99179	T + B + F + C	0,99199
B + W	0,97594	T + B + W	0,97761
B + W + C	0,98746	T + B + W + C	0,98876
B + W + F	0,99234	T + B + W + F	0,98563
B + W + F + C	0,99236	T + B + W + F + C	0,99136
M	0,99116	T + M	0,99043
M + C	0,99199	T + M + C	0,99294
M + F	0,99390	T + M + F	0,99368
M + F + C	0,99389	T + M + F + C	0,99227
M + W	0,99262	T + M + W	0,99224
M + W + C	0,99382	T + M + W + C	0,99238
M + W + F	0,99359	T + M + W + F	0,99201
M + W + F + C	0,99189	T + M + W + F + C	0,99269
M + B	0,99030	T + M + B	0,98727
M + B + C	0,99122	T + M + B + C	0,99166
M + B + F	0,99316	T + M + B + F	0,99183
M + B + F + C	0,99507	T + M + B + F + C	0,99103
M + B + W	0,99064	T + M + B + W	0,98885
M + B + W + C	0,99355	T + M + B + W + C	0,99441
M + B + W + F	0,99200	T + M + B + W + F	0,99281
M + B + W + F + C	0,99435	T + M + B + W + FS + C	0,99213
T	0,93977	-	-

Tabela 5.2. Valores AUC para diferentes combinações entre os descritores de áudio. Onde C = CFT, F = FT Sequencial, B = BMFCC, M = MFCC Sequencial, T = TED e W = Wavelet.

Com base no trabalho de Nascimento et al. [2013] e considerando a robustez do descritor usado, não foi feita seleção de atributos visuais, apenas definição do tamanho do dicionário.

5.3.2 Validação com rótulos aleatórios

Para validar implementação usada na etapa de classificação, foi executado um teste com rótulos aleatórios usando toda a base, onde um rótulo aleatório é atribuído a cada amostra independente de seu vetor de atributos ou classe real. As matrizes de confusão as abordagens auditiva, visual, FAV e MAV estão na Figura 5.8, cuja legenda consta na Tabela 5.3.

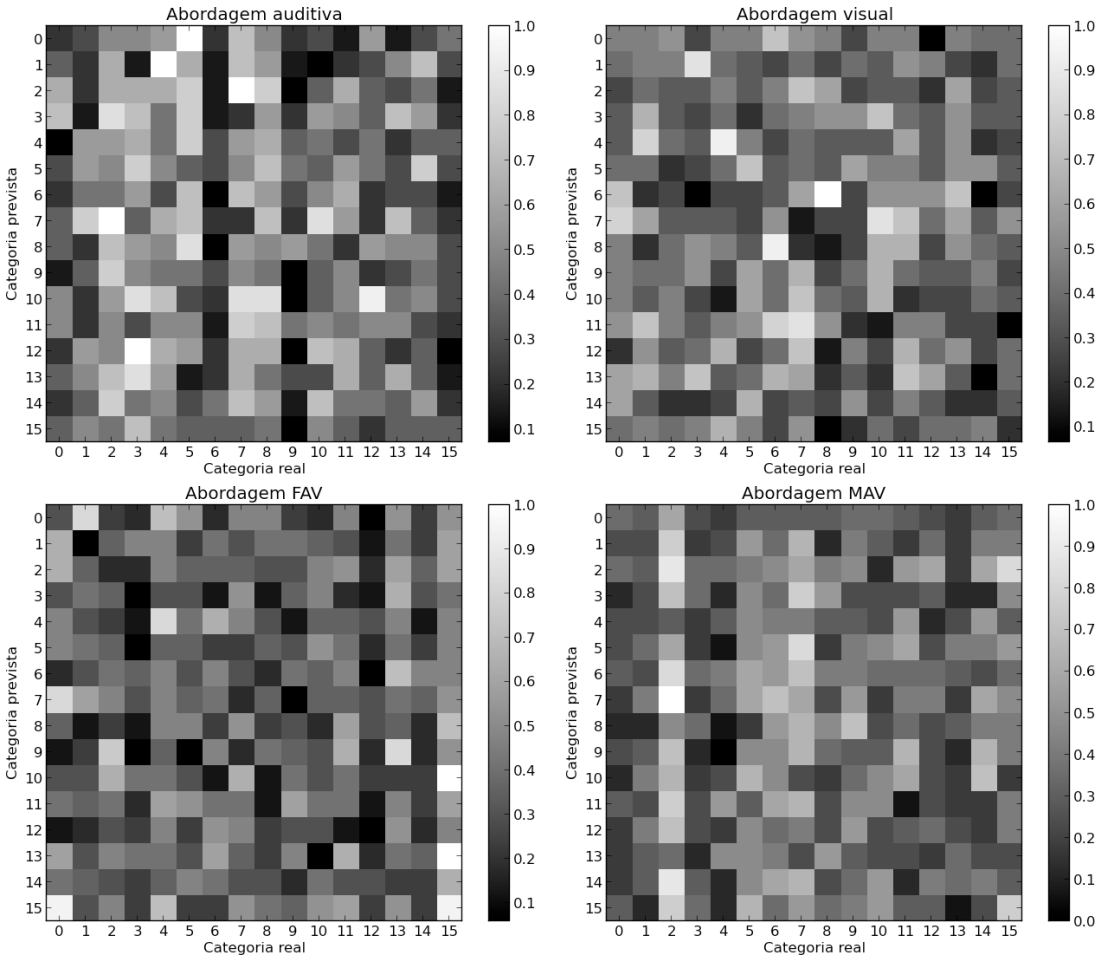


Figura 5.8. Matrizes de confusão das abordagens na validação com rótulos aleatórios.

Índice	Objeto	Índice	Objeto
0	Cubo de madeira	8	Cubo de papelão
1	Tetraedro de madeira	9	Tetraedro de papelão
2	Prisma de madeira	10	Prisma de papelão
3	Octaedro de madeira	11	Octaedro de papelão
4	Cubo de plástico	12	Cubo de Isopor
5	Tetraedro de plástico	13	Cilindro de Isopor
6	Prisma de plástico	14	Cone de Isopor
7	Octaedro de plástico	15	Esfera de Isopor

Tabela 5.3. Legenda de objetos para as matrizes de confusão.

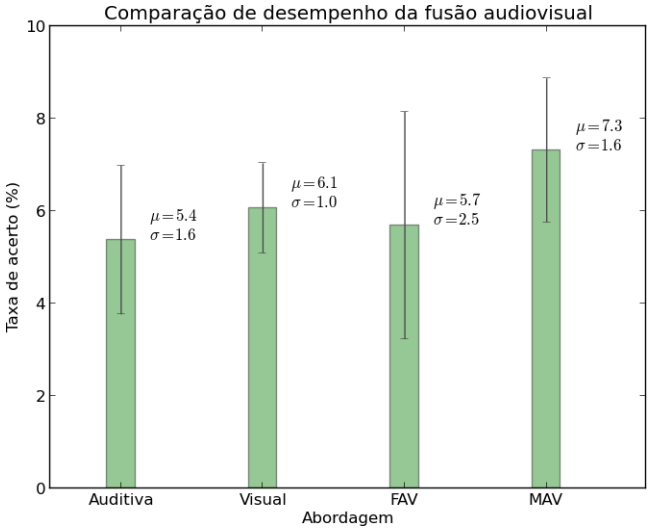


Figura 5.9. Taxa de acerto de cada abordagem na validação com rótulos aleatórios.

A distribuição da matriz de confusão mostra o baixo desempenho do reconhecimento, o que é esperado, já que não há relação direta entre os atributos de cada amostra e sua categoria, indicando não haver erro aparente na implementação da etapa de classificação ou *overfitting* claro dos dados.

Na Figura 5.9 estão as taxas de acerto para cada modalidade, onde é possível observar que, em média, o acerto de cada modalidade é em torno de 6,12%, próximo dos 6,25% esperados de um classificador aleatório para as 16 categorias.

5.4 Experimentos

As próximas seções detalham os experimentos executados para avaliar diferentes perspectivas da metodologia, são eles:

1. Verificação da capacidade descritiva do áudio para objetos com mesmo material (Seção 5.4.1);

2. Verificação da capacidade descritiva do áudio para objetos com mesma geometria (Seção 5.4.2);

3. Verificação do desempenho do reconhecimento audiovisual em ambiente livre de ruído (Seção 5.4.3);

4. Robustez a ruído auditivo. Foi introduzido ruído nas amostras de áudio em diferentes níveis de interferência, para verificar a robustez a dados ruidosos na modalidade auditiva (Seção 5.4.4);

5. Robustez a ruído visual. Foi introduzido ruído nas amostras visuais em diferentes níveis de interferência, para verificar a robustez a dados ruidosos na modalidade visual (Seção 5.4.5);

6. Robustez a ruído visual. Ruído aditivo em diferentes níveis foi introduzido em ambas as modalidades sensoriais para verificar a estabilidade a ruído provida pela fusão mesmo com dados ruidosos em ambas as modalidades (Seção 5.4.6);

7. Verificação da capacidade de reconhecimento de objetos externos ao treinamento. São incluídas amostras de objetos externos à base para verificar a estabilidade do sistema para classes não treinadas (Seção 5.4.7).

5.4.1 Reconhecimento de materiais pelo som

Para comparar a capacidade discriminativa do som para materiais, foram executados testes de classificação binária entre objetos de mesma geometria, onde a classe real de cada objeto é atribuída de acordo com seu material.

As tabelas 5.4, 5.5, 5.6 e 5.7 mostram os valores AUC para as comparações entre objetos de mesma geometria mas materiais diferentes: madeira e papelão, papelão e plástico, etc.

	Papelão	Plástico
Madeira	0,995	0,995
Papelão	-	0,989

Tabela 5.4. Valores AUC entre tetraedros de diferentes materiais.

Os valores AUC indicam que é possível discriminar entre materiais com relativa facilidade usando som. Estima-se que esta facilidade se deve a propriedades do material,

	Papelão	Plástico
Madeira	0,995	0,995
Papelão	-	0,991

Tabela 5.5. Valores AUC entre prismas de diferentes materiais.

	Papelão	Plástico
Madeira	0,995	0,995
Papelão	-	0,995

Tabela 5.6. Valores AUC entre cubos de diferentes materiais.

	Papelão	Plástico
Madeira	0,995	0,995
Papelão	-	0,995

Tabela 5.7. Valores AUC entre octaedros de diferentes materiais.

como frequência fundamental, heterogeneidade na reverberação do som[3] e coeficientes de reflexão nas paredes internas.

Para verificar o impacto que a composição de um objeto tem nos descritores extraídos, o cubo de madeira foi preenchido com espuma de polietileno expandido em todo o seu interior, como ilustrado na Figura 5.10.

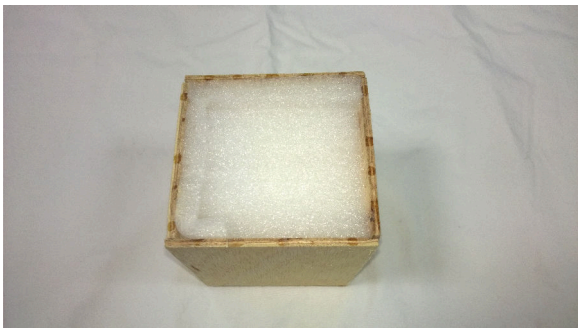


Figura 5.10. Espuma de polietileno expandido usada para preenchimento do cubo de madeira.

A adição da espuma tem alto impacto no sinal gerado, exemplificado pela Figura 5.11, ao serem comparados o TED_{30} do mesmo objeto oco e preenchido pela espuma.

[3] A velocidade de propagação do som ao longo da fibra da madeira é mais rápida do que na direção ortogonal [Rossing, 2007], por exemplo.

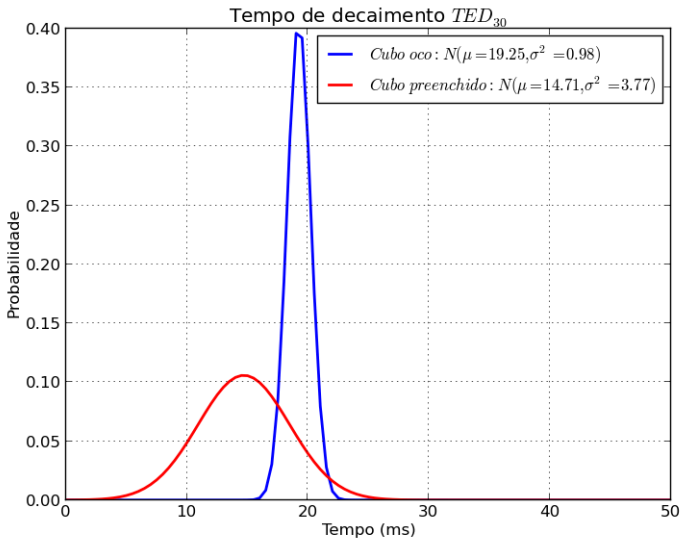


Figura 5.11. Comparação do TED_{30} do mesmo objeto quando oco e preenchido com espuma de polietileno expandido.

Ainda que a composição da espuma não seja para a finalidade de isolamento acústico, sua inclusão no interior do objeto reduz bastante o tempo TED_{30} do sinal ao absorver mais rapidamente a onda gerada. Assim, tanto o recorte de áudio da captura quanto os descritores extraídos são afetados pela composição do objeto.

5.4.2 Reconhecimento de geometrias pelo som

Para comparar a capacidade discriminativa do som para geometrias, neste experimento foram executados testes de classificação binária entre objetos de mesmo material mas geometrias diferentes, onde a classe real de cada objeto é atribuída de acordo com sua geometria.

As tabelas 5.8, 5.9, 5.10 e 5.11 mostram os valores AUC para as comparações entre objetos de mesmo material mas geometrias diferentes: cubo e prisma de madeira, prisma e tetraedro de papelão, etc.

	Prisma	Cubo	Octaedro
Tetraedro	0,995	0,995	0,995
Prisma	-	0,995	0,995
Cubo	-	-	0,995

Tabela 5.8. Valores AUC entre os objetos de madeira.

	Prisma	Cubo	Octaedro
Tetraedro	0,985	0,995	0,966
Prisma	-	0,854	0,995
Cubo	-	-	0,995

Tabela 5.9. Valores AUC entre os objetos de papelão.

	Prisma	Cubo	Octaedro
Tetraedro	0,995	0,995	0,971
Prisma	-	0,960	0,986
Cubo	-	-	0,995

Tabela 5.10. Valores AUC entre os objetos de plástico.

	Cilindro	Cone	Esfera
Cubo	0,698	0,872	0,987
Cilindro	-	0,740	0,994
Cone	-	-	0,971

Tabela 5.11. Valores AUC entre os objetos de Isopor.

Novamente, os valores AUC indicam que é possível discriminar entre geometrias com mesmo material com relativa facilidade usando som. Estima-se que a reverberação interna do som e sua reflexão nas paredes dos poliedros ajudam na discriminação auditiva.

Neste experimento o reconhecimento dos objetos de Isopor teve pior desempenho, quando comparado aos demais. Neste caso, geometria e material parecem influenciar os resultados: os testes binários com a esfera, de geometria totalmente curva, para este material tem os melhores resultados, mas os sinais gerados por todos os objetos deste material são de baixa amplitude, estando mais sujeitos a ruído, o que pode explicar o pior desempenho com relação ao reconhecimento dos demais objetos.

5.4.3 Reconhecimento de objetos pela fusão de dados audiovisuais

O objetivo deste experimento foi averiguar a capacidade discriminativa auditiva e visual combinadas, para isto testes envolvendo o reconhecimento de todos os objetos da base foram executados. Os resultados comparativos de cada abordagem são ilustrados pelas matrizes de confusão na Figura 5.12 e taxas de acerto na Figura 5.13. A Figura 5.14 mostra as curvas ROC (*Receiver Operating Characteristic*) e seus respectivos valores AUC.

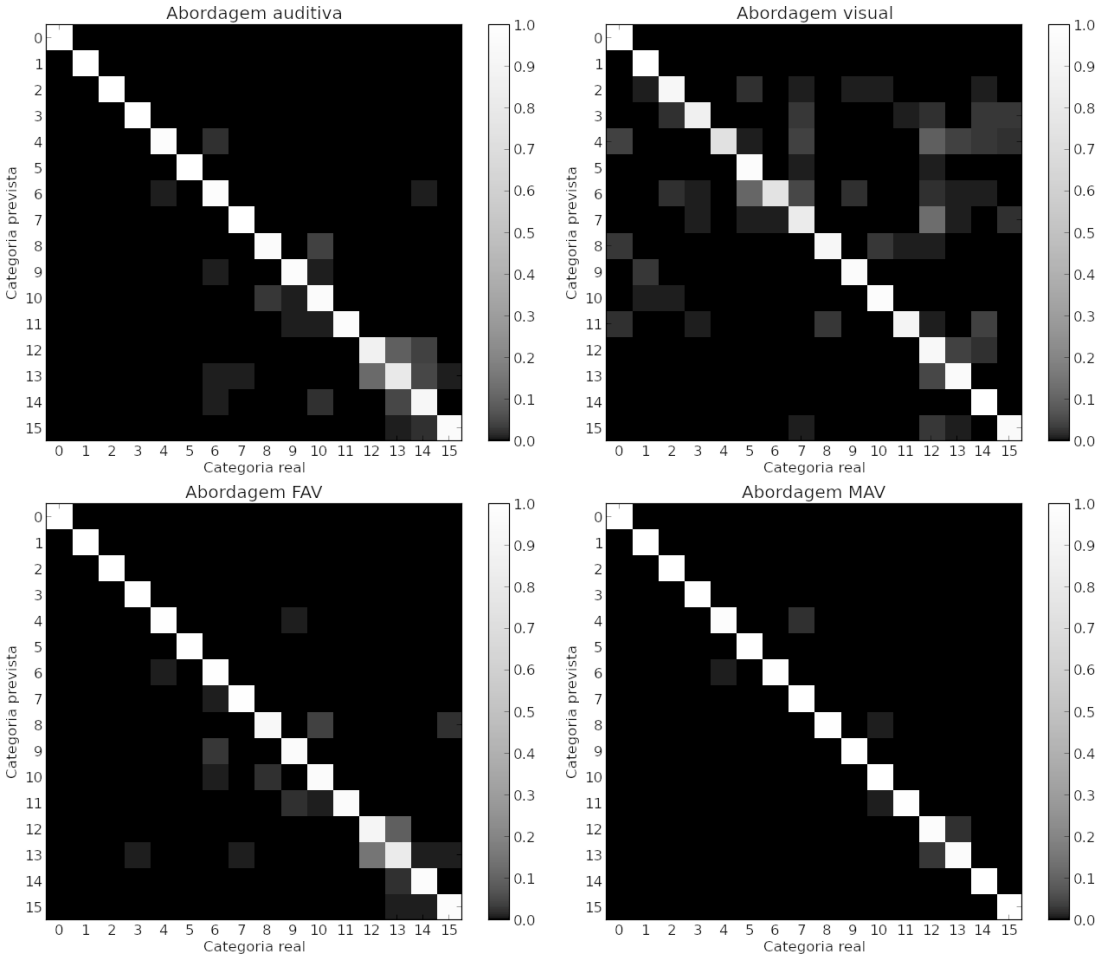


Figura 5.12. Matrizes de confusão das abordagens no experimento de reconhecimento contendo todos os objetos da base.

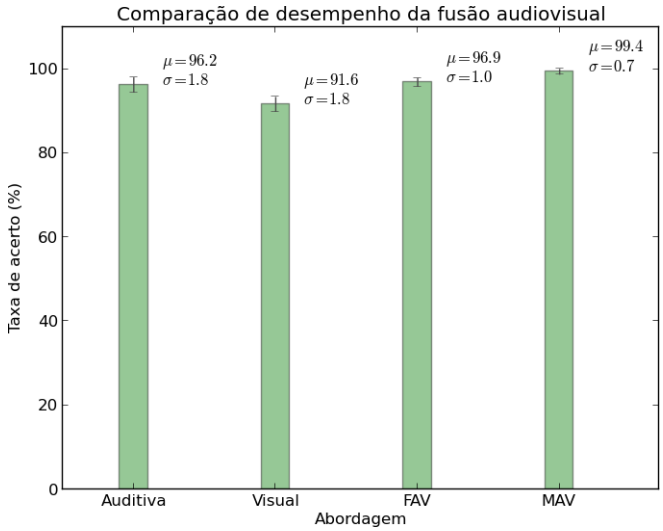


Figura 5.13. Taxas de acerto das abordagens no experimento de reconhecimento contendo todos os objetos da base.

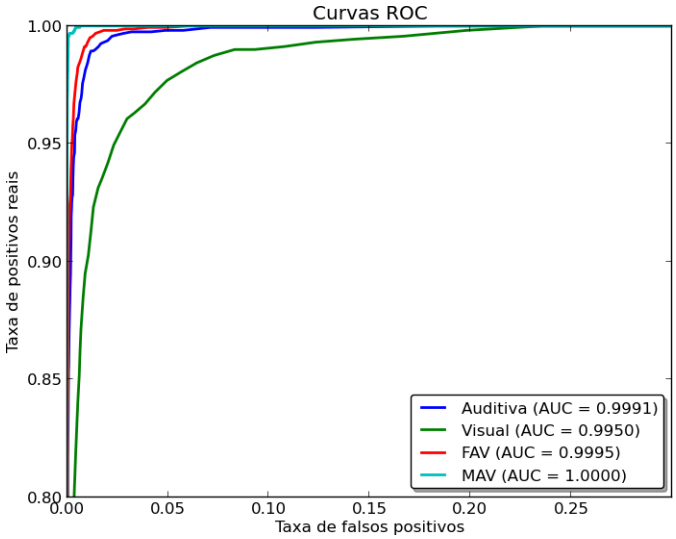


Figura 5.14. Curvas ROC das abordagens no experimento de reconhecimento contendo todos os objetos da base.

De acordo com as curvas da Figura 5.14, a abordagem para fusão audiovisual MAV tem o melhor desempenho neste experimento de reconhecimento, principalmente ao se observar as matrizes de confusão da Figura 5.12. Nessas, enquanto as abordagens auditiva e FAV têm dificuldade em desambiguar objetos feitos de isopor (devido aos sinais de baixa amplitude) e a visual tem problema similar com objetos de plástico (devido à transparência) e Isopor, a abordagem MAV consegue explorar a complementaridade das modalidades e integrá-las melhor, compensando os problemas de cada uma e melhorando o resultado de forma geral.

Em particular, este resultado tem conformidade com a premissa de Sargin & Yemez [2007] de que uma abordagem baseada em fusão de decisões é mais eficaz quando as modalidades são independentes entre si, além de estar mais alinhada com as características da fusão biológica.

5.4.4 Robustez a ruído auditivo

É difícil avaliar o impacto real da fusão quando as todas as modalidades já tem desempenho tão alto, assim para averiguar a robustez a ruído provida pela fusão audiovisual, ruído aditivo sintético foi utilizado nas amostras de áudio de modo a perturbar sua representação pelos descritores. Ruído em diferentes níveis de SNR (*Signal-to-noise Ratio*) foi acrescido ao sinais para avaliar a estabilidade do sistema em diferentes ce-

nários. Seja s o sinal e r um ruído sintético, a SNR, em decibéis, é dada por

$$SNR(dB) = 10 \log_{10} \left(\frac{E_s}{E_r} \right), \tag{5.1}$$

onde E_s e E_r são as energias totais do sinal e do ruído, respectivamente. A energia de cada um pode ser obtida através de sua variância, sabendo que o componente DC (média do sinal) foi removido, assim

$$s = s - \mu(s) \quad \text{e} \quad r = r - \mu(r). \tag{5.2}$$

Logo,

$$E_s = \sigma^2(s) \quad \text{e} \quad E_r = \sigma^2(r). \tag{5.3}$$

Assim,

$$SNR(dB) = 10 \log_{10} \left(\frac{\sigma^2(s)}{\sigma^2(r)} \right). \tag{5.4}$$

Dada uma SNR desejada, a partir da Equação 5.4, verifica-se a variância v necessária:

$$\sigma^2(r) = \frac{\sigma^2(s)}{10^{SNR/10}}. \tag{5.5}$$

Por fim, faz-se

$$r = r * \frac{\sqrt{v}}{\sqrt{\sigma^2(r)}}. \tag{5.6}$$

Adicionando ruído com a abordagem mencionada, foram testados diferentes níveis de SNR: $-\infty$ (apenas ruído), -20, -15, -10, -5, 0, 5, 10, 20, 30, 40, 50 e ∞ (sinal puro).

As figuras 5.15 e 5.16 contém os resultados para os níveis de ruídos mencionados para dois tipos de ruído: gaussiano, amostrado de uma distribuição $N(\mu = 0, \sigma^2 = 1)$, e rosa, baseado no algoritmo de Voss-McCartney [Voss & Clarke, 1975] provido por Johnson [2011]. Este último foi testado para simular melhor um ambiente real, onde dificilmente há quantidade igual de energia em todas as bandas do espectro.

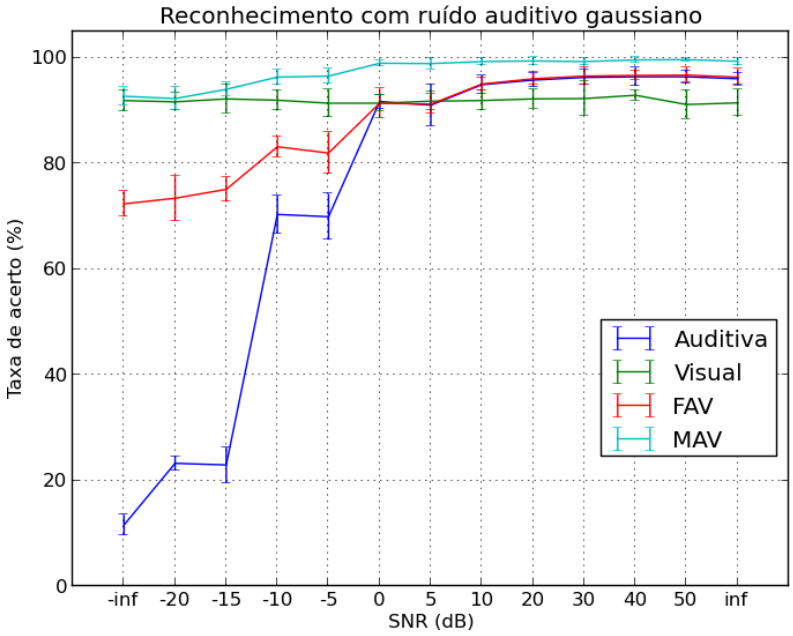


Figura 5.15. Desempenho das abordagens com adição de ruído auditivo gaussiano.

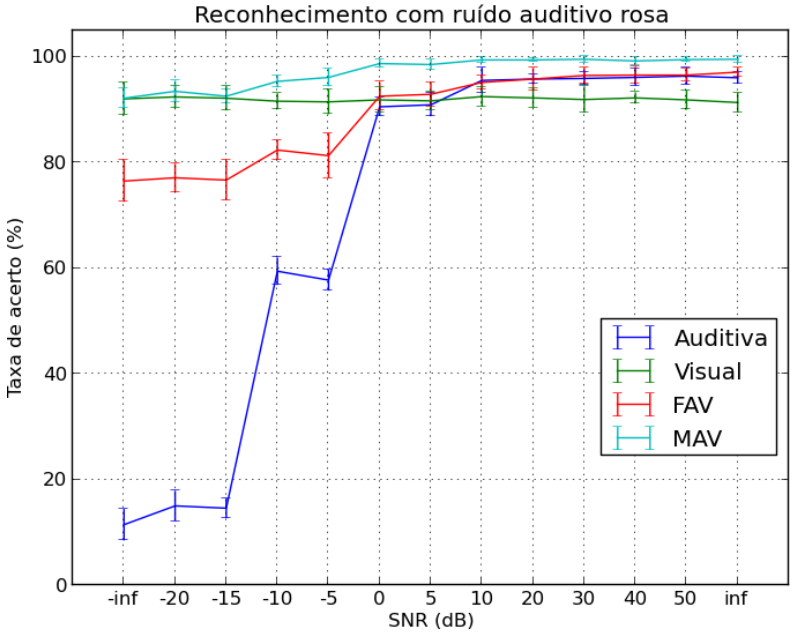


Figura 5.16. Desempenho das abordagens com adição de ruído auditivo rosa.

Para os dois tipos de ruído, observa-se que a fusão audiovisual provê estabilidade no reconhecimento, mesmo sob forte interferência: o reconhecimento audiovisual é degradado a apenas um fator do desempenho da modalidade ruidosa.

O nível de ruído necessário para degradar significativamente o sinal (de $-\infty$ a 0dB) se dá pela natureza da captura: a SNR é calculada levando em consideração o recorte do sinal como um todo, mas há alta concentração de energia nos momentos iniciais, onde há o pico do sinal, assim, apenas SNR baixas mascaram o trecho inicial do recorte.

É interessante notar ainda como a abordagem MAV supera a FAV em todos os casos, indicando o bom desempenho de fusão de decisões em cenários com SNR inversamente proporcionais em cada modalidade, e que mesmo no pior caso (modalidade auditiva com ruído infinito), o reconhecimento ainda é mantido com a mesma qualidade da modalidade visual, menos ruidosa.

As quedas de desempenho nos intervalos [-20,-15] e [-10,-5] no experimento com ruído rosa são atribuídas a desvio estatístico, considerando que os intervalos são próximos e as médias estão no desvio padrão, possivelmente tendo sido geradas pelo particionamento aleatório da validação cruzada, feita a cada SNR.

5.4.5 Robustez a ruído visual

Similarmente ao experimento anterior, ruído aditivo gaussiano foi introduzido nas amostras visuais de modo a perturbar sua representação pelos descritores. Ruído em diferentes níveis de SNR foi acrescido às imagens para avaliar a estabilidade do sistema em diferentes cenários de interferência. Os níveis de SNR testados foram 0, 10, 20, 30, 40, 50 e ∞ (amostra sem ruído), amostrados das distribuições normais indicadas na Tabela 5.12.

SNR	Distribuição
0	$N(\mu = 0, \sigma^2 = 250)$
10	$N(\mu = 0, \sigma^2 = 80)$
20	$N(\mu = 0, \sigma^2 = 20)$
30	$N(\mu = 0, \sigma^2 = 8)$
40	$N(\mu = 0, \sigma^2 = 2,5)$
50	$N(\mu = 0, \sigma^2 = 0,8)$
∞	$N(\mu = 0, \sigma^2 = 0)$

Tabela 5.12. SNR visual e distribuições amostradas.

Esses valores de σ^2 foram obtidos de acordo com a definição de PNSR (*Peak*

Signal-to-Noise), similar à Equação 5.1:

$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{MSE} \right), \tag{5.7}$$

onde MAX é o valor máximo atribuído a um *pixel* e MSE é o *Mean Squared Error*. Como as imagens são em escala de cinza, $MAX = 255$. O MSE pode ser definido como:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |o(x,y) - e(x,y)|^2 \tag{5.8}$$

onde m e n são largura e comprimento da imagem, respectivamente, e $i(x,y)$ e $e(x,y)$ *pixels* correspondentes na imagem original e na imagem degradada.

A Figura 5.17 contém as taxas de acerto do reconhecimento para os níveis de ruídos mencionados.

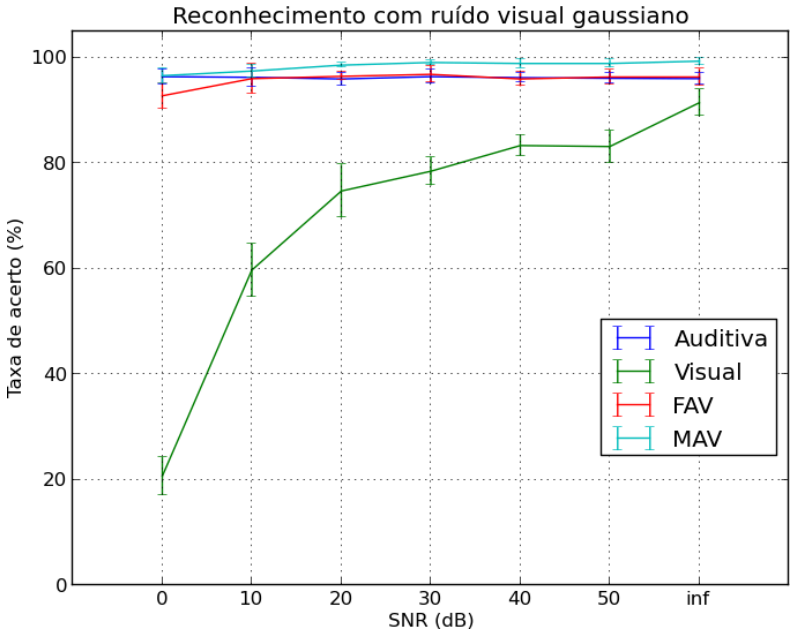


Figura 5.17. Desempenho das abordagens com adição de ruído visual gaussiano.

Assim como o experimento onde houve degradação das amostras auditivas, neste a fusão provê estabilidade mesmo com a modalidade visual bastante ruidosa. Neste cenário a abordagem MAV também supera a FAV e não é degradada abaixo da modalidade auditiva, reforçando que modalidades sensoriais mais aptas a perceberem o ambiente tem maior peso, assim como acontece na natureza.

5.4.6 Robustez a ruído auditivo e visual

Para verificar como a fusão se comporta em casos em que há ruído nas duas modalidades sensoriais de origem, foi realizado um último experimento envolvendo adição de ruído sintético. Neste, ruído em diferentes níveis foi acrescentado às duas modalidades sensoriais. Os níveis de SNR para a modalidade auditiva e visual foram $-\infty$ (apenas ruído), 0, 30 e ∞ e 0, 30 e ∞, respectivamente.

A Figura 5.18, 5.19 contém as taxas de acerto do reconhecimento para os níveis de ruídos mencionados para ruído auditivo gaussiano e rosa, respectivamente. Ruído visual gaussiano foi usado em ambos os casos.

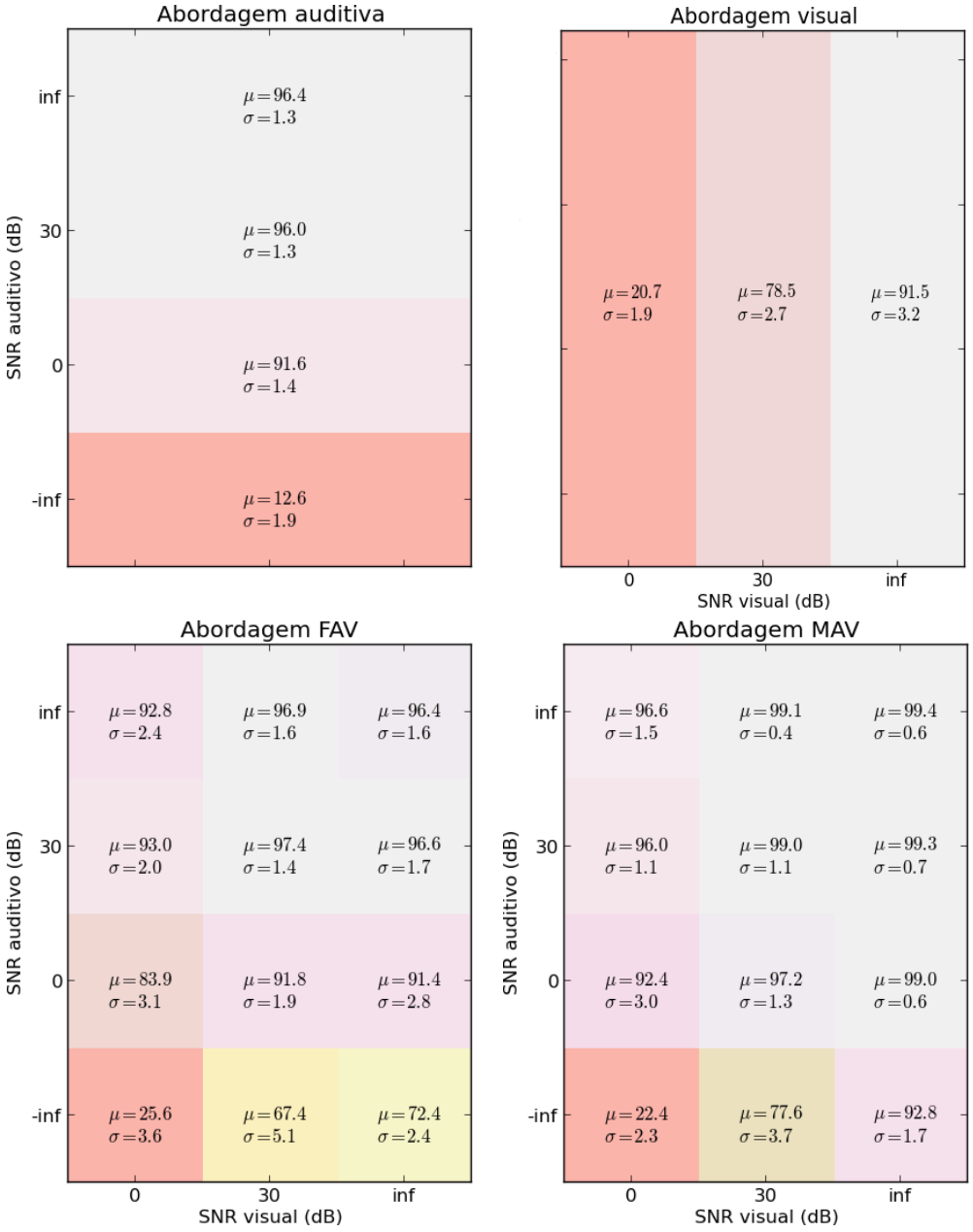


Figura 5.18. Taxa de acerto de cada abordagem para diferentes níveis de ruído auditivo gaussiano e ruído visual gaussiano.

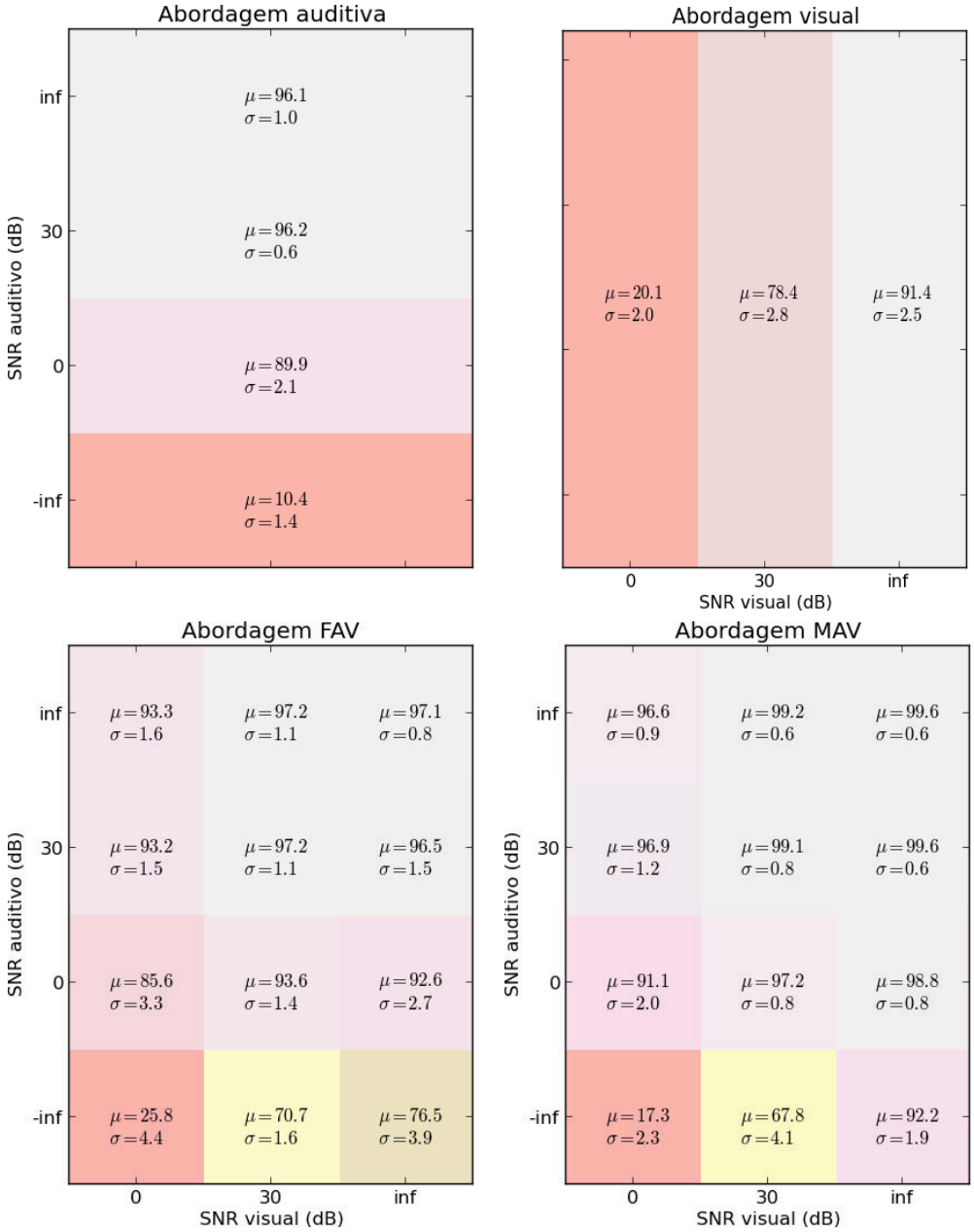


Figura 5.19. Taxa de acerto de cada abordagem para diferentes níveis de ruído auditivo rosa e ruído visual gaussiano.

Como esperado, a medida que os duas modalidades sensoriais vão sendo degradadas, há menor impacto do ruído nas abordagens FAV e MAV, onde o reconhecimento só é realmente degradado quando há muita interferência em ambas as modalidades. A abordagem MAV demonstra ter ainda mais estabilidade que a FAV, uma vez que nos cenários com ruído auditivo infinito a FAV sofre perturbação com mais facilidade: mesmo com informação visual sem ruído, o desempenho é degradado a aproximadamente 70% de acerto.

Nas 12 combinações de ruído testadas, MAV obteve melhor desempenho que FAV em 10, sendo inferior em apenas dois cenários de alta interferência (ruído auditivo infinito e visual com 0 dB e 30 dB de SNR), o que pode ser explicado pela dimensionalidade do vetor de atributos de cada abordagem: a abordagem FAV tem um vetor de atributos aproximadamente 40 vezes maior que a MAV, o que aumenta as chances do classificador encontrar um atributo que seja mais discriminante. Como há bastante ruído adicionado aos dados originais, a compactação da representação tem papel relevante no desempenho do classificador.

5.4.7 Reconhecimento de objetos externos ao treinamento

Este experimento foi executado com o objetivo de averiguar a robustez do sistema quanto ao reconhecimento de classes ainda não treinadas e analisar a interferência de amostras desta categoria. Para tanto, uma pequena base de dados contendo amostras objetos variados não presentes na base anterior foram incluídas no conjunto de teste dos classificadores. Esta nova base, ilustrada na Figura 5.20, contém 20 objetos de uso comum de diferentes materiais, geometrias e texturas.

Figura 5.20. Objetos variados para experimento de estabilidade do sistema ante classes não desconhecidas.

Cada objeto teve 10 amostras capturadas nas mesmas condições da base anterior, totalizando 200 amostras para cada modalidade sensorial. Como nenhum destes objetos passou pela etapa de treinamento, foi adotada uma abordagem *One vs Rest*

[Bishop, 2007]: é criado um classificador binário para cada classe (os 16 objetos ilustrados na Figura 5.1), usando parte das amostras desta classe como treinamento positivo e amostras das demais classes como treinamento negativo. Amostras que são consideradas negativas para todos os 16 classificadores, são rotuladas como "negativo".

Os resultados comparativos de fusão e de cada modalidade independente para este experimento são ilustrados pelas matrizes de confusão na Figura 5.21 taxas de acerto na Figura 5.22. Devido a falta de ocorrência de falsos positivos nas abordagens auditivas e MAV as curvas ROC comparando seu desempenho serão omitidas.

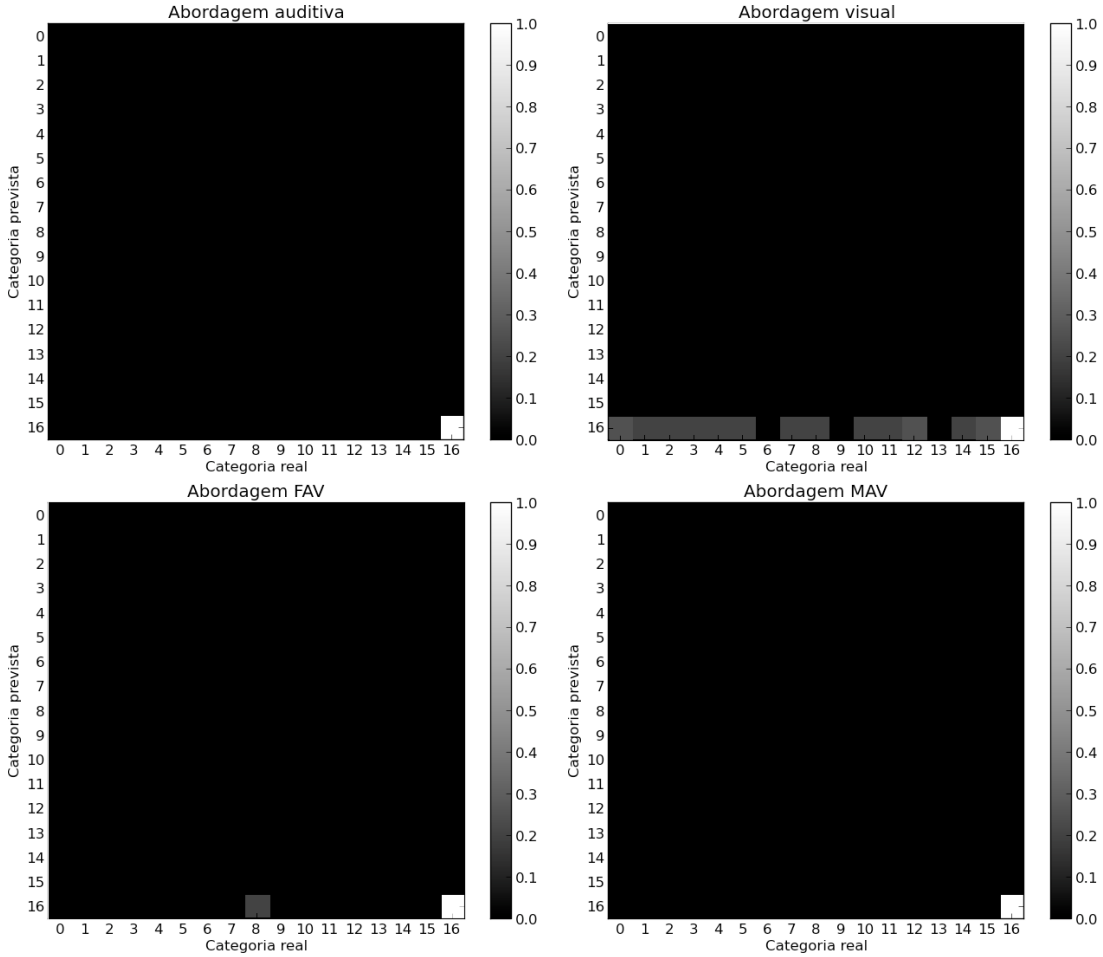


Figura 5.21. Matrizes de confusão das abordagens para fusão e modalidades sensoriais no experimento de reconhecimento com objetos externos à base.

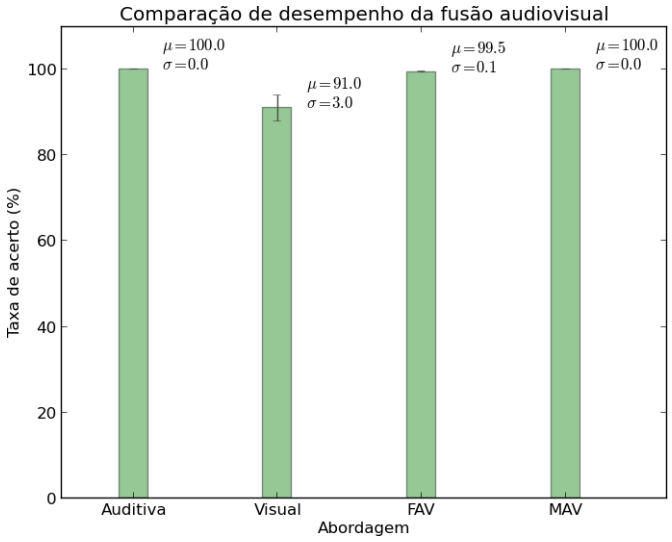


Figura 5.22. Taxas de acerto das abordagens para fusão e modalidades sensoriais no experimento de reconhecimento com objetos externos à base.

É possível observar que os resultados deste experimento são consistentes com o experimento de reconhecimento apenas de objetos da base original, dadas as taxas de reconhecimento da Figura 5.22, mais ainda, nenhuma amostra negativa foi classificada erroneamente pelas abordagens auditivas e MAV, sustentando o uso de informações auditivas como uma alternativa válida no reconhecimento de objetos.

5.5 Comentários gerais

Uma dificuldade inicial encontrada durante os experimentos foi a proximidade dos objetos a serem reconhecidos do sensor. Devido a necessidade do objeto estar próximo ao robô para que houvesse interação mecânica e a aquisição de áudio pudesse ocorrer, há possibilidade de ocorrência de buracos na imagem de profundidade[4], dependendo do alcance do manipulador, o que pode prejudicar o reconhecimento visual.

Com os experimentos executados foi validado o uso da fusão de sensores como opção para desambiguação no reconhecimento de objetos, no entanto, testes com uma base de dados mais complexa são necessários a fim de verificar se esse desempenho pode ser generalizado, pois supõe-se que o alto desempenho das modalidades se deve às características da base, como número de classes, baixa variedade de materiais e geometria, além de simetria nos objetos. Testes estendidos incluindo objetos assimétri-

[4] O alcance padrão do sensor Kinect, por exemplo, é de 800mm à 4000mm. Fonte: http://msdn.microsoft.com/en-us/library/hh438998.aspx

cos, formas côncavas e materiais variados são necessários para corroborar a abordagem usada de maneira mais ampla.

Capítulo 6

Conclusão e trabalhos futuros

6.1 Conclusão

Neste trabalho, foi investigado a fusão de dados audiovisuais para a tarefa de reconhecimento de objetos. Foram aplicadas com sucesso técnicas estabelecidas de reconhecimento por áudio e reconhecimento por imagem em uma abordagem de aprendizado de máquina, melhorando o desempenho do sistema ao aumentar sua taxa de reconhecimento e precisão.

Ainda que haja vasta literatura relacionada à tarefa proposta, principalmente na área de Visão Computacional, poucos trabalhos o fazem com uso de áudio e imagem, o que sugere que esta área carece de mais pesquisa, considerando as vantagens fornecidas. As vantagens no uso de diferentes modalidades perceptivas na representação dos dados residem na capacidade discriminativa provida pelo uso de ambas quando comparada ao seu uso individual, além de sua robustez a ruído.

Outra vantagem do método adotado é a independência de sensores, que, por não necessitarem de sincronização na fase de aquisição, podem ser usados de forma independente, assim, equipamentos mais adequados a certas situações podem ser utilizados.

Experimentos com dados reais foram executados para avaliar a metodologia, onde foi possível verificar que, para a base coletada, a fusão audiovisual melhorou a capacidade de desambiguação do classificador. A incorporação de atributos de diferentes modalidades sensoriais facilitou o reconhecimento de objetos complexos de serem reconhecidos por uma modalidade (como os de Isopor usando somente o som), acrescentando robustez ao modelo, fazendo com que o sistema sofresse pouca degradação mesmo sob alta interferência.

6.2 Limitações e trabalhos futuros

Ainda que tenham sido obtidos resultados muito interessantes para o problema abordado, há limitações referentes à metodologia adotada, discutidos como trabalhos futuros para esta pesquisa:

- **Extensão dos experimentos**: a utilização de uma base mais abrangente com objetos de uso comum, como panelas e copos, para avaliar o desempenho do método com objetos com maior variação textura, geometria e material;
- **Estudo sistemático**: estudo da influência de geometria e material sobre os sons produzidos sob a perspectiva de propriedades físicas, como coeficientes de reflexão;
- **SNR como fator na fusão**: Sabendo o nível de ruído em cada modalidade é possível diminuir sua influência na fusão de acordo com o ganho que pode se obter ao utilizá-la. A inclusão deste fator explicitamente pode amenizar a interferência destrutiva de modalidades ruidosas no reconhecimento;
- **Modo de aquisição de áudio**: O uso de outros meios de geração de sinal de áudio, como outras formas de interação mecânica, seguindo a abordagem de [Sinapov, 2013], ou executando captura de reflexão de um pulso sonoro, podem contribuir para averiguar propriedades do material do objeto. Esta última, além de aumentar o alcance das capturas, contribui também para redução de eventual formação de buracos nas imagens de profundidade;
- **Percepção ativa do ambiente**: a aquisição de informação e desempenho do sistema são bastante limitados pela pose do sensor em relação ao objeto de interesse. Através da percepção ativa do ambiente é possível amenizar estes problemas ao posicionar o sensor de forma a maximizar o ganho de informação;
- **Integração de semântica**: A semântica tem alto impacto na probabilidade *a priori* no reconhecimento de um objeto. Certos objetos têm maior probabilidade de ocorrência em determinados ambientes, como panelas em cozinhas ou computadores em laboratórios ou escritórios. Ao acessar este fator é possível reduzir o espaço amostral a ser considerado no reconhecimento;

Referências Bibliográficas

Agin, G. J. & Binford, T. O. (1976). Computer Description of Curved Objects. *Computers, IEEE Transactions on*, C-25(4):439--449. ISSN 0018-9340.

Alahi, A.; Ortiz, R. & Vandergheynst, P. (2012). FREAK: Fast Retina Keypoint. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 510--517. ISSN 1063-6919.

Almalence (2007). Photoacute. Url: `http://photoacute.com/studio/examples/mac_hdd/index.html`. Acessado em 12/05/2015.

Alpaydin, E. (2009). *Introduction to Machine Learning*. The MIT Press.

Amos, E. (2010). Kinect. Url: `http://en.wikipedia.org/wiki/Kinect`. Acessado em 12/05/2015.

Arel, I.; Rose, D. C. & Karnowski, T. P. (2010). Deep Machine Learning - A New Frontier in Artificial Intelligence Research. *IEEE Computational Intelligence Magazine*, 5(November):13--18. ISSN 1556-603X.

Arsenio, A. & Fitzpatrick, P. (2003). Exploiting cross-modal rhythm for robot perception of objects. MIT.

Aucouturier, J.-J.; Defreville, B. & Pachet, F. (2007). The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, 122(2):881--91. ISSN 1520-8524.

Bay, H.; Ess, A.; Tuytelaars, T. & Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346--359.

Bengio, Y.; Goodfellow, I. J. & Courville, A. (2014). Deep Learning.

Biondi, R.; Dys, G.; Ferone, G.; Renard, T. & Zysman, M. (2014). Low Cost Real Time Robust Identification of Impulsive Signals. *International Journal of Computer, Information, Systems and Control Engineering*, 8(9):1524--1527.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer. ISBN 978-0387310732.

Bjorkman, M. & Kragic, D. (2010). Active 3D scene segmentation and detection of unknown objects. Em *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3114--3120. ISSN 1050-4729.

Blauert, J. (1996). *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press. ISBN 0262024136.

Bloch, I. (1996). Information combination operators for data fusion: a comparative review with classification. *IEEE Transactions on Systems, Man and Cybernetics*, 26(1):52--67. ISSN 1083-4427.

Boff, K.; Kaufman, L. & Thomas, J. (1986). Intersensory interactions. Em *Handbook of Perception and Human Performance*, capítulo 6, pp. 25.1--25.36. Wiley.

Bogert, B.; Healy, M. & Tukey, J. (1963). The quefrency alanysis of time series for echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking. Em *Proc. Symp. on Time Series Analysis*, pp. 209--243.

Bohg, J.; Johnson-Roberson, M.; Bjorkman, M. r. & Kragic, D. (2010). Strategies for multi-modal scene exploration. Em *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4509--4515. ISSN 2153-0858.

Boureau, Y. L.; Bach, F.; LeCun, Y. & Ponce, J. (2010). Learning mid-level features for recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2559--2566. ISSN 10636919.

Bradski, G. (2000). The OpenCV Library. *Dr Dobbs Journal of Software Tools*, 25:120--125. ISSN 1044-789X.

Breiman, L. (2001). Random forests. *Machine learning*, pp. 5--32. ISSN 0885-6125.

Burr, D. & Alais, D. (2006). Combining visual and auditory information. *Progress in brain research*, 155(December 2005):243--58. ISSN 0079-6123.

Calonder, M.; Lepetit, V.; Strecha, C. & Fua, P. (2010). BRIEF: Binary Robust Independent Elementary Features. Em *Proceedings of the 11th European Conference on Computer Vision: Part IV*, Lecture Notes in Computer Science, pp. 778--792. Springer-Verlag. ISSN 0302-9743.

Casey, M. (2001). MPEG-7 sound-recognition tools. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):737--747. ISSN 10518215.

Chen, T. (2001). Audiovisual speech processing. *IEEE Signal Processing Magazine*, 18(1):9--21. ISSN 10535888.

Chu, S.; Narayanan, S. & Kuo, C.-C. (2009). Environmental Sound Recognition With Time-Frequency Audio Features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6). ISSN 1558-7916.

Coates, A.; Arbor, A. & Ng, A. Y. (2011). An Analysis of Single-Layer Networks in Unsupervised Feature Learning. *Aistats 2011*, pp. 215--223.

Cowling, M. & Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15):2895--2907. ISSN 01678655.

Csurka, G.; Dance, C. R.; Fan, L.; Willamowski, J. & Bray, C. (2004). Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*.

Dahl, G. E.; Yu, D.; Deng, L. & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 20(1):30--42. ISSN 15587916.

Dasarathy, B. (1997). Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*, 85(1).

Davis, K. H.; Biddulph, R. & Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal Of The Acoustic Society Of America*, 24(6):7.

Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4). ISSN 0096-3518.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K. & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. Em *IEEE Computer Vision and Pattern Recognition (CVPR)*. Url: `http://www.image-net.org`. Acessado em 12/05/2015.

Dufaux, A. (2001). *Detection and recognition of impulsive sounds signals*. Tese de doutorado.

Dzeroski, S. & Zenko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54:255--273. ISSN 08856125.

Ekvall, S.; Jensfelt, P. & Kragic, D. (2006). Integrating active mobile robot object recognition and SLAM in natural environments. Em *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5792--5797.

Ernst, M. O. & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in cognitive sciences*, 8(4):162--9. ISSN 1364-6613.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861--874. ISSN 01678655.

Fei-Fei, L. (2005). Recognizing and learning object categories. Url: `http://people.csail.mit.edu/torralba/shortCourseRLOC/`. Acessado em 12/05/2015.

Fei-Fei, L.; Fergus, R. & Perona, P. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Em *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, pp. 178–178. Url: `http://www.vision.caltech.edu/Image_Datasets/Caltech101`. Acessado em 12/05/2015.

Fischler, M. A. & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography.

Freire, I. L. & Apolinário, J. A. (2010). Gunshot detection in noisy environments. Em *International Telecommunications Symposium (ITS)*, p. 4.

Galvez-López, D. & Tardos, J. D. (2012). Bags of Binary Words for Fast Place Recognition in Image Sequences.

Gaver, W. W. (1993). What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception. *Ecological Psychology*, 5:1--29. ISSN 1040-7413.

Gehrig, T.; Nickel, K.; Ekenel, H. K.; Klee, U. & McDonough, J. (2005). Kalman filters for audio-video source localization. Em *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pp. 118--121.

Geusebroek, J. M.; Burghouts, G. J. & Smeulders, A. W. M. (2005). The Amsterdam library of object images. *International Journal of Computer Vision (IJCV)*, 61(1):103--112. ISSN 09205691. Url: `http://aloi.science.uva.nl`. Acessado em 12/05/2015.

Giordano, B. L. & McAdams, S. (2006). Material identification of real impact sounds: effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America*, 119:1171--1181. ISSN 00014966.

Gold, B.; Morgan, N. & Ellis, D. (2011). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Wiley-Interscience.

Griffin, G.; Holub, A. & Perona, P. (2006). Caltech-256 object category dataset. Relatório técnico 1. Url: `http://www.vision.caltech.edu/Image_Datasets/Caltech256`. Acessado em 12/05/2015.

Hall, D. & Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6--23. ISSN 0018-9219.

Harris, C. & Stephens, M. (1988). A Combined Corner and Edge Detector. *Procedings of the Alvey Vision Conference 1988*, pp. 147--151. ISSN 09639292.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustic Society of America*.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Uncertainity in Artifitial Intelligence - UAI'99*, p. 8. ISSN 15206882.

Howard, I. P. & Templeton, W. B. (1966). *Human spatial orientation*. John Wiley & Sons Ltd. ISBN 0471416622.

ImageNet (2014). Base de dados imagenet. Url: `http://www.image-net.org/`. Acessado em 12/05/2015.

Johnson, L. (2011). py-sound: A library of python code for manipulating sounds. Url: `https://github.com/lmjohns3/py-sound/`. Acessado em 12/05/2015.

Jones, E.; Oliphant, T.; Peterson, P. & Others (2001). {SciPy}: Open source scientific tools for {Python}.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 9.ª edição. ISBN 0471878766.

Kittler, J.; Hater, M. & Duin, R. P. W. (1996). Combining classifiers. *Proceedings - International Conference on Pattern Recognition*, 2(3):897--901. ISSN 10514651.

Klatzky, R. L.; Pai, D. K. & Krotkov, E. P. (2000). Perception of Material from Contact Sounds. *Presence: Teleoperators and Virtual Environments*, 9:399--410. ISSN 1054-7460.

Kohonen, T. (2000). *Self-Organizing Maps*. Springer, terceira edição.

Krizhevsky, A.; Sutskever, I. & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pp. 1--9. ISSN 10495258.

Krotkov, E. (1995). Robotic Perception of Material. Em *Proceedings of the IJCAI*, pp. 88--94.

Krotkov, E.; Klatzky, R. & Zumel, N. (1997). Robotic perception of material: Experiments with shape-invariant acoustic measures of material type. Em Khatib, O. & Salisbury, J., editores, *Experimental Robotics IV*, volume 223 of *Lecture Notes in Control and Information Sciences*, pp. 204--211. Springer Berlin Heidelberg.

Lacheze, L.; Guo, Y.; Benosman, R.; Gas, B. & Couverture, C. (2009). Audio/video fusion for objects recognition. Em *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 652--657.

Lai, K.; Bo, L.; Ren, X. & Fox, D. (2011). A large-scale hierarchical multi-view rgb-d object dataset. Em *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 1817–1824. ISSN 1050-4729. Url: `http://rgbd-dataset.cs.washington.edu`. Acessado em 12/05/2015.

Lazebnik, S.; Schmid, C. & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. Em *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pp. 2169--2178. ISSN 10636919.

Leutenegger, S.; Chli, M. & Siegwart, R. Y. (2011). BRISK: Binary Robust invariant scalable keypoints. *2011 International Conference on Computer Vision*, pp. 2548--2555.

Libal, U. & Spyra, K. (2014). Wavelet based shock wave and muzzle blast classification for different supersonic projectiles. *Expert Systems with Applications*, 41(11):5097--5104. ISSN 09574174.

Liu, N.; Zhao, Y.; Zhu, Z. & Lu, H. (2011). Exploiting Visual-Audio-Textual Characteristics for Automatic TV Commercial Block Detection and Segmentation. *IEEE Transactions on Multimedia*, 13(5):961--973. ISSN 1520-9210.

Lowe, D. (1999). Object Recognition from Local Scale-Invariant Features. *IEEE International Conference on Computer Vision*. ISSN 0-7695-0164-8.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91--110. ISSN 0920-5691.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the waveletrepresentation. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 11(7):674--693. ISSN 0162-8828.

McGibney, D.; Morioka, R.; Sekiyama, K. & andToshio Fukuda, H. M. (2012). A Multiple Robot Cognitive Sharing System using Audio and Video Sensors. Em *International Symposium on Distributed Autonomous Robotic Systems (DARS)*, pp. 246--257.

McGibney, D.; Umeda, T.; Sekiyama, K.; Mukai, H. & Fukuda, T. (2011). Cooperative Distributed Object Classification for Multiple Robots with Audio Features. Em *Micro-NanoMechatronics and Human Science (MHS), 2011 International Symposium on*, number 3, pp. 134--139.

McGurk, H. & MacDonald, J. W. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746--748.

Menegatti, E.; Mumolo, E.; Nolich, M. & Pagello, E. (2004). A Surveillance System based on Audio and Video Sensory Agents. Em *Intelligent Autonomous Systems*, pp. 335--343. IOS Press.

Mitrovic, D.; Zeppelzauer, M. & Eidenberger, H. (2007). Analysis of the data quality of audio descriptions of environmental sounds. *Journal of Digital Information Management*, 5(2):48--55. ISSN 0972-7272.

Muja, M. & Lowe, D. (2012). Fast matching of binary features. *Computer and Robot Vision (CRV)*, pp. 404--410.

Murase, H. & Nayar, S. (1995). Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14(1):5--24. ISSN 0920-5691.

Nakamura, E. F.; Loureiro, A. a. F. & Frery, A. C. (2007a). Information fusion for wireless sensor networks. *ACM Computing Surveys*, 39(3). ISSN 03600300.

Nakamura, T.; Araki, T.; Nagai, T. & Iwahashi, N. (2009). Grounding of Word Meanings in LDA-Based Multimodal Concepts. Em *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3943--3948.

Nakamura, T.; Nagai, T. & Iwahashi, N. (2007b). Multimodal object categorization by a robot. Em *IEEE International Conference on Intelligent Robots and Systems*, pp. 2415--2420.

Nascimento, E. R.; Oliveira, G. L.; Vieira, A. W. & Campos, M. F. (2013). On the development of a robust, fast and lightweight keypoint descriptor. *Neurocomputing*, 120:141--155. ISSN 09252312.

Nene, S. A.; Nayar, S. K. & Murase, H. (1996). Columbia university image library (coil-100). Relatório técnico. Url: `http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php`. Acessado em 12/05/2015.

Neti, C.; Potamianos, G.; Luettin, J.; Matthews, I.; Glotin, H. & Vergyri, D. (2001). Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins Summer 2000 Workshop. *2001 IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No.01TH8564)*.

Oliveira, G. L.; Nascimento, E. R.; Vieira, A. W. & Campos, M. F. M. (2012). Sparse Spatial Coding: A novel approach for efficient and accurate object recognition. Em *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2592--2598. IEEE.

Oppenheim, A. V. & Schafer, R. W. (2004). DSP history - From frequency to quefrency: a history of the cepstrum. *IEEE Signal Processing Magazine*, 21(5):95--106. ISSN 1053-5888.

Ortega-Garcia, J.; Bigun, J.; Reynolds, D. & Gonzalez-Rodriguez, J. (2004). Authentication gets personal with biometrics. *IEEE Signal Processing Magazine*, 21(2). ISSN 1053-5888.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825--2830.

Peeters, G.; Giordano, B. L.; Susini, P.; Misdariis, N. & McAdams, S. (2011). The Timbre Toolbox: extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902--16. ISSN 1520-8524.

Peltonen, V.; Tuomi, J.; Klapuri, A.; Huopaniemi, J. & Sorsa, T. (2002). Computational Auditory Scene Recognition. Em *IEEE International Conference on Audio, Speech and Signal Processing*, pp. 1941--1944.

Petajan, E. D. (1984). *Automatic Lipreading to Enhance Speech Recognition (Speech Reading)*. Tese de doutorado, University of Illinois, Champaign, IL, USA.

Pieropan, A. & Salvi, G. (2014). Audio-visual classification and detection of human manipulation actions. Em *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3045--3052.

Pineau, J.; Montemerlo, M.; Pollack, M.; Roy, N. & Thrun, S. (2003). Towards robotic assistants in nursing homes: Challenges and results. Em *Robotics and Autonomous Systems*, volume 42, pp. 271--281.

Quigley, M.; Conley, K.; Gerkey, B.; FAust, J.; Foote, T.; Leibs, J.; Berger, E.; Wheeler, R. & Mg, A. (2009). ROS: an open-source Robot Operating System. *IEEE International Conference on Robotics and Automation (ICRA)*, 3(Figure 1):5.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257--286.

Rabiner, L. R. & Schafer, R. W. (2007). Introduction to digital speech processing. *Foundations and trends in signal processing*, 1(1):1--194.

Richard, G.; Sundaram, S. & Narayanan, S. (2013). An Overview on Perceptually Motivated Audio Indexing and Classification. *Proceedings of the IEEE*, 101(9):1939--1954.

Richmond, J. & Pai, D. (2000). Active measurement of contact sounds. Em *IEEE International Conference on Robotics and Automation*, volume 3, p. 7. ISSN 1050-4729.

Rossing, T. D. (2007). *Springer Handbook of Acoustics*. Springer New York, New York, NY. ISBN 978-0-387-30446-5.

Rosten, E.; Porter, R. & Drummond, T. (2010). Faster and better: a machine learning approach to corner detection. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):105--119.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53--65. ISSN 03770427.

Rublee, E.; Rabaud, V.; Konolige, K. & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *2011 International Conference on Computer Vision*, pp. 2564--2571. ISSN 1550-5499.

Rusu, R. B. & Cousins, S. (2011). 3D is here: Point Cloud Library (PCL). Em *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China.

Sabine, W. C. (1906). Architectural acoustics. Em *Proceedings of the American Academy of Arts and Sciences*, pp. 51--84. JSTOR.

Sargin, M. & Yemez, Y. (2007). Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, 9(7):1396--1403.

Shazam (2015). Reconhecedor musical shazam. Url: `www.shazam.com`. Acessado em 12/05/2015.

Shelton, B. R. & Searle, C. L. (1980). The influence of vision on the absolute identification of sound-source position. *Perception & Psychophysics*, 28(6):589--596. ISSN 0031-5117.

Shivappa, S. T.; Trivedi, M. M. & Rao, B. D. (2010). Audiovisual Information Fusion in Human-Computer Interfaces and Intelligent Environments: A Survey. *Proceedings of the IEEE*, 98(10):1692--1715. ISSN 0018-9219.

Sinapov, J. (2013). *Behavior-grounded multi-sensory object perception and exploration by a humanoid robot*. Tese de doutorado, Iowa State University.

Sinapov, J.; Wiemer, M. & Stoytchev, A. (2009). Interactive learning of the acoustic properties of household objects. Em *IEEE International Conference on Robotics and Automation*, pp. 2518--2524. ISSN 10504729.

Siri (2015). Assistente de voz pessoal siri. Url: `www.apple.com/ios/siri`. Acessado em 12/05/2015.

Somerville, N. (2011). Sunset. Url: `http://stonehillfield.blogspot.com.br/2011/09/mcsunset.html`. Acessado em 12/05/2015.

Soundhound (2015). Reconhecedor musical shazam. Url: `www.soundhound.com`. Acessado em 12/05/2015.

Stein, B. E. & Meredith, M. A. (1993). *The merging of the senses.* The MIT Press, Cambridge, MA, US.

Stein, B. E. & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature reviews. Neuroscience*, 9(4):255--266. ISSN 1471-0048.

Stork, D.; Wolff, G. & Levine, E. (1992). Neural network lipreading system for improved speech recognition. Em *International Joint Conference on Neural Networks (IJCNN)*, volume 2, pp. 289--295. IEEE.

Strobel, N.; Spors, S. & Rabenstein, R. (2001). Joint audio-video object localization and tracking. *IEEE Signal Processing Magazine*, 18(1):22--31. ISSN 10535888.

Sudhakaran, S. & Pappachen James, A. (2014). Sparse distributed localized gradient fused features of objects. *Pattern Recognition*, 48(4):1538--1546. ISSN 00313203.

Sun, D.; Roth, S. & Black, M. J. (2010). Secrets of Optical Flow Estimation and Their Principles - Optical flow : motion of image pixels. Em *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2432--2439.

Suzuki, Y.; Futoshi, A.; Hack-Yoon, K. & Toshio, S. (1995). An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *The Journal of the Acoustical Society of America*, 97:1119. ISSN 00014966.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.; Hill, C.; Reed, S.; Anguelov, D.; Jia, Y.; Erhan, D.; Sermanet, P.; Vanhoucke, V. & Rabinovich, A. (2014). Going deeper with convolutions.

Thrun, S.; Bennewitz, M.; Burgard, W.; Cremers, A.; Dellaert, F.; Fox, D.; Hahnel, D.; Rosenberg, C.; Roy, N.; Schulte, J. & Schulz, D. (1999). MINERVA: a second-generation museum tour-guide robot. Em *IEEE International Conference on Robotics and Automation (ICRA)*, volume 3. ISSN 1050-4729.

Thrun, S.; Burgard, W. & Fox, D. (2005). *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents).* The MIT Press. ISBN 0262201623.

Tzanetakis, G. & Cook, P. (2002). Musical genre classification of audio signals. Em *IEEE Transactions on Speech and Audio Processing*, volume 10, pp. 293--302. ISSN 10636676.

van der Walt, S.; Colbert, S. C. & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2):22--30. ISSN 1521-9615.

Voss, R. F. & Clarke, J. (1975). '1/f noise' in music and speech. *Nature*, 258(5533):317--318.

Walker, R. (2012). Freaks, geeks and microsoft: How kinect spawned a commercial ecosystem. Url: `http://www.nytimes.com/2012/06/03/magazine/how-kinect-spawned-a-commercial-ecosystem.html`. Acessado em 12/05/2015.

Wildes, R. & Richards, W. (1988). Recovering material properties from sound. Em *International Conference on Natural Computation*.

Yates, H. G. (2014). F.e.a.r. Url: `http://lovebeinghere.com/2014/06/05/f-e-a-r/`. Acessado em 12/05/2015.

Yu, D. & Deng, L. (2014). *Automatic Speech Recognition: A Deep Learning Approach*. Springer.

Yuhas, B. P.; Goldstein, M. H. & Sejnowski, T. J. (1989). Speech Signals Using Neural Networks. *IEEE Communications Magazine*, (November):65--71.

Zahorik, P. (2001). Estimating sound source distance with and without vision. *Optometry and vision science : official publication of the American Academy of Optometry*, 78(5):270--275. ISSN 1040-5488.

Zhang, T. & Kuo, C.-C. J. (2001). Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, 9(4). ISSN 1063-6676.