UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ANDERSON PRIEBE FERRUGEM

Visão Computacional: Indexação Automatizada de Imagens

Dissertação apresentada como requisito parcial
para a obtenção do grau de
Mestre em Ciência da Computação

Prof. Dr. Dante Augusto Couto Barone
Orientador

Porto Alegre, dezembro de 2004

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Ferrugem, Anderson Priebe

Visão Computacional:
Indexação Automatizada de Imagens / Anderson Priebe Ferrugem. – Porto Alegre: PPGC da UFRGS, 2004.

91 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2004. Orientador: Dante Augusto Couto Barone.

1. Recuperação de Imagens. 2. Visão Computacional. 3. Mapas Auto-Organizáveis. 4. SOM. I. Barone, Dante Augusto Couto. II. Título.

“Na ciência, existem questões ingênuas, questões entediantes, questões apresentadas de modo inadequado. Mas cada questão é um grito para entender o mundo. Não existe pergunta estúpida.”

— CARL SAGAN

AGRADECIMENTOS

Agradeço aos meus familiares:

THEO, que fez com que a vida valesse mais a pena.

Valeu filhão !!!

Ao meu Pai HÉLIO que me mostrou o caminho e me deu o exemplo de integridade e persistência (além de me introduzir no ramo da ciência, engenharia e outros).

Obrigado FERRUGEM !!!

A minha mãe LORACI que sempre me aparou e me mostrou o valor da emoção e do amor na minha vida.

Teu "guri"te agradece MÃE !!!

Aos meus irmãos ROBINSON e MICHELE com quem aprendi a compartilhar (e carregar o fardo de ser o exemplo...).

É isso aí manos !!!

A minha grande companheira e amiga ANDRÉIA, que na hora que precisei esteve ao meu lado.

Valeu minha amada "DEINHA".

'Brigadão por tudo !!!

Ao meu amigos: GIL que sempre acreditou no meu potencial. Meu orientador DANTE que impulsionou minha pesquisa e que também conto como amigo.

E a todos que de forma positiva vieram para acrescentaram "sabor"nesta longa e nunca completa busca pelo conhecimento que é a vida. Valeu !!!

Plagiando Carl Sagan "É um privilégio compartilhar o mesmo tempo e planeta com vocês, dada a infinitude do tempo e a imensidão do espaço".

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

BDI	*Beliefs, Desires, Intentions*
BI	*Banco de Imagens*
CBIR	*Content-Based Image Retrieval*
CIE	Comissão Internacional em Iluminação *Commission Internationale de l'Eclairage)*
CGV	*Cognitive Vision*
CogVis	*Cognitive Vision Systems*
CV	*Computer Vision*
FIRA	Federation International Robot-soccer Association
H-SOM	*Hierarchical Self-Organizing Map*
GH-SOM	*Grow Hierarchical Self-Organizing Map*
IA	Inteligência Artificial
IU	*Image Understanding*
KBR	*Keyword-based retrieval*
LRI	Laboratório de Robótica Inteligente
MV	*Maquine Vision*
NTSC	National Television Systems Committee
RN	Rede Neural
RNA	Rede Neural Artificial
SBR	*Similarity- Based Retrieval*
SOM	*Self-Organizing Map*
SOFM	*Self-Organizing Feature Map*
TS-SOM	*Tree Structured Self-Organizing Map*
UFRGS	Universidade Federal do Rio Grande do Sul
VA	Visão Artificial
VC	Visão Computacional
VM	Visão de Máquina

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

O avanço tecnológico atual está permitindo que as pessoas recebam cada vez mais informações visuais dos mais diferentes tipos, nas mais variadas mídias. Esse aumento fantástico está obrigando os pesquisadores e as indústrias a imaginar soluções para o armazenamento e recuperação deste tipo de informação, pois nossos computadores ainda utilizam, apesar dos grandes avanços nessa área, um sistema de arquivos imaginado há décadas, quando era natural trabalhar com informações meramente textuais. Agora, nos deparamos com novos problemas: Como encontrar uma paisagem específica em um banco de imagens, em que trecho de um filme aparece um cavalo sobre uma colina, em que parte da fotografia existe um gato, como fazer um robô localizar um objeto em uma cena, entre outras necessidades.

O objetivo desse trabalho é propor uma arquitetura de rede neural artificial que permita o reconhecimento de objetos genéricos e de categorias em banco de imagens digitais, de forma que se possa recuperar imagens específicas a partir da descrição da cena fornecida pelo usuário. Para que esse objetivo fosse alcançado, foram utilizadas técnicas de Visão Computacional e Processamento de Imagens na etapa de extração de feições de baixo nível e de Redes Neurais(Mapas Auto-Organizáveis de Kohonen) na etapa de agrupamento de classes de objetos.

O resultado final desse trabalho pretende ser um embrião para um sistema de reconhecimento de objetos mais genérico, que possa ser estendido para a criação de indices de forma automática ou semi-automática em grandes bancos de imagens.

Palavras-chave: Recuperação de Imagens, Visão Computacional, Mapas Auto-Organizáveis, SOM.

Computer Vision: Automated Indexing of Images

ABSTRACT

The current technological progress allows people to receive more and more visual information of the most different types, in different medias. This huge augmentation of image availability forces researchers and industries to propose efficient solutions for image storage and recovery. Despite the extraordinary advances in computational power, the data files system remain the same for decades, when it was natural to deal only with textual information. Nowadays, new problems are in front of us in this field. For instance, how can we find an specific landscape in a image database, in which place of a movie there is a horse on a hill, in which part of a photographic picture there is a cat, how can a robot find an object in a scene, among other queries.

The objective of this work is to propose an Artificial Neural Network (ANN) architecture that performs the recognition of generic objects and object's categories in a digital image database. With this implementation, it becomes possible to do image retrieval through the user´s scene description. To achieve our goal, we have used Computer Vision and Image Processing techniques in low level features extraction and Neural Networks (namely Kohonen's Self-Organizing Maps) in the phase of object classes clustering.

The main result of this work aims to be a seed for a more generic object recognition system, which can be extended to the automatic or semi-automatic index creation in huge image databases.

Keywords: Image Retrieval, Computer Vision, Self-Organizing Maps, SOM.

1 INTRODUÇÃO

1.1 Introdução

O mundo moderno cada vez mais exige que se automatizem processos antes realizados apenas pelo homem, principalmente no campo de reconhecimento de imagens. A inspeção visual de objetos manufaturados, o reconhecimento de rostos e impressões digitais são alguns exemplos de aplicações deste tipo de tecnologia. Em geral, a técnica implementada para essas tarefas utiliza conhecimento *a priori* de características bem estabelecidas do objeto a ser reconhecido em ambientes controlados ou com pouca variação, como no caso de controle de qualidade em ambientes industriais, aplicações essa do campo conhecido como Visão de Máquina.

Reconhecimento de um número pequeno de objetos, bem definidos, em ambientes controlados, em geral, não é problemático, pois, técnicas tradicionais resolvem bem essa questão. O problema ocorre quando se tenta implementar sistemas de reconhecimento mais genéricos. Este tipo de abordagem esbarra em questões como: Que métrica utilizar para medir o grau de similaridade entre os objetos, de que forma deve-se agrupar as classes de objetos, o quão robusto o sistema deve ser. Todos essas questões são também dependentes do contexto em que serão utilizadas, pois de acordo com a situação, um mesmo objeto pode ser classificado de várias formas. Por exemplo, um objeto banana, pode ser associado a uma categoria chamada frutas, ou se estivermos avaliando a origem do objeto ele pode ser agrupado com a categoria país tropical.

Este trabalho aborda o problema da extração de conhecimento semântico básico, de imagens digitais pertencentes a bancos de imagens, através do reconhecimento de alguns componentes que pertencem a essas imagens, através de um técnica proposta, envolvendo redes neurais artificiais. Essas informações semânticas podem ser utilizadas, posteriormente, na criação de descritores de imagens, de forma automática ou semi-automática, para uso na recuperação de imagens.

1.2 Motivação

Uma boa parte das informações digitais é composta por dados visuais. Basta observar o avanço tecnológico impressionante nas últimas décadas, que se acelera em ritmo vertiginoso, para percebermos como ele tem ficado cada vez mais rico em tecnologias de armazenamento e transmissão de imagens digitais. Infelizmente, uma imagem, apesar de ser extremamente rica de significados para os seres humanos, ao ser transformada em dados digitais, não passa de uma coleção de valores numéricos sem nenhum significado aparente. É difícil relacionar esses conjunto de dados (*pixeis*) com informações do tipo: pessoas em um piquenique, acidente aéreo, chegada do homem a lua. Uma solução para

esse problema é fazer anotações manuais, e realizar uma busca através dos métodos tradicionais de recuperação de texto. É isso que se faz em catálogos comerciais, mas essa solução tem dois pequenos problemas (LüNEBURG, 2002):

- Depende do contexto pessoal de quem faz as anotações (que sofre desde influências psicológicas até profissionais),
- É extremamente custosa em termos de tempo

Em geral, essa abordagem é feita por empresas que necessitam localizar rapidamente figuras de um BI (banco de imagens). Por exemplo, produtos utilizados na decoração de interiores (tapetes, quadros, acabamentos...).

Freqüentemente torna-se complicado localizar uma determinada imagem de interesse em um computador privado. Como localizá-la então, em um imenso acervo como na Internet, ou em um volumoso banco de imagens? No caso da Internet, existem soluções híbridas, que combinam a informação textual da própria página com informações de textura, forma e cor da(s) imagem(s). Em bancos de imagens, raramente existe alguma informação textual ou alguma outra forma de descrição da cena. Para essa situação, existem técnicas de recuperação de imagens que exigem maior ou menor nível de interação com o usuário. Uma das técnicas é disponibilizar um conjunto limitado de imagens com características representativas das demais existentes no BI, por exemplo, uma cena contendo um carro pode representar um conjunto de outras cenas que contém veículos. Dado esse conjunto inicial de imagens, o usuário seleciona as que possuem maior semelhança com a desejada. Desta forma, usando algoritmos de similaridade baseados em características como cor, forma e textura, o sistema fornece um novo grupo de imagens. Esse processo se repete de forma interativa, até que o usuário se dê por satisfeito com alguma imagem. Outra técnica existente é a extração de características de textura e forma, que são usadas como base para descrições mais detalhadas da cena em alto nível. Infelizmente, existe um abismo entre essas características e a descrição da cena em alto nível, chamado em inglês pelo termo *semantic gap* (WANG; LI; WIEDERHOLD, 2001). Pois não fica claro como informações específicas de um *pixel*, ou de um conjunto de *pixeis*, podem se relacionar com uma descrição mais elaborada.

Apesar destes problemas, a recuperação de imagens digitais pode ser usada em diversas áreas, tais como (LüNEBURG, 2002) :

- Catálogos de galerias de arte, museus e pinturas arqueológicas;
- Catálogo de projetos em arquitetura e engenharia;
- Sistemas de informação geográfica, previsão do tempo, classificação de imagens aéreas ou astronômicas;
- Imagens médicas;
- Publicidade (localização de imagens e/ou ilustrações);
- Base de marcas;
- Investigações criminais, violação de direitos autorais na internet;
- Arquivamento, recuperação e busca por duplicações de imagens em um BI;
- Arquivamento e recuperação de imagens em vídeos;

- Busca de imagens na internet.

A busca por imagens em um BI, onde não existe interação com o usuário, acaba implicando no problema típico da Visão Computacional de identificar objetos. Este não é um problema fácil de ser abordado, pois envolve questões de métricas, qualidade da imagem, condições de iluminação, variância da cor, entre outros. Quanto menor a possibilidade de controle sobre as condições ambientais, maior a dificuldade desta tarefa. Mesmo considerando neste trabalho a situação de recuperação de imagens, as técnicas abordadas podem ser estendidas para outras situações, tais como a Robótica Móvel, com as devidas adaptações, onde pode ser útil para navegação e manipulação de objetos (RUSSELL; NORVIG, 1995). Esses são desafios verdadeiros dentro da Visão Computacional, que segundo (WINSTANLEY, 1991), "é uma área de estudos verdadeiramente multidisciplinar, permitindo associar sistemas computacionais, engenharia e robótica com o universo da visão biológica, entendimento de imagens e inteligência artificial".

Outra motivação deriva do sucesso de modelos computacionais hierárquicos, utilizados no reconhecimento de objetos, tais como o HMAX(RIESENHUBER; POGGIO, 2000), o NeoCognitron (FUKUSHIMA, 1982) e o SIMPLIcity(WANG; LI; WIEDERHOLD, 2001). Esses modelos comprovam que uma estrutura hierárquica é a melhor solução para grandes espaços dimensionais, como imagens, pois essas estruturas fornecem os níveis de detalhamento necessários para a classificação das características.

1.3 Contribuições Esperadas

Nesta dissertação, é descrito um sistema para o reconhecimento de objetos/categorias em imagens variadas de um banco de imagens. Esta arquitetura pode ser parte integrante de um sistema de recuperação de imagens, ou de um sistema de anotações automáticas, ou semi-automáticas. Um diferencial da arquitetura proposta é a organização hierárquica dos mapas auto-organizáveis de Kohonen (em inglês, Self-Organizing Maps ou SOM), utilizada para armazenar categorias em um nível crescente de complexidade. Outra contribuição é o uso da GH-SOM (em inglês, *Grow Hierarchical Self-Organizing Map*), que é uma rede neural composta de SOM´s independentes, organizados de forma hierárquica. Nesta rede a camada superior possui uma representação rudimentar do espaço de entrada, e à medida que se desce na hierarquia, a granularidade aumenta, permitindo visualizar as similaridades do espaço de entrada em vários níveis.

Como se está trabalhando com imagens, é necessário verificar a similaridade entre as amostras. Evitando usar no treinamento de objetos diferentes amostragens com características semelhantes. Isso pode ser feito através de alguns métodos estatísticos e de medidas de similaridade, mas a GH-SOM se mostrou bem mais simples de se utilizar, permitindo visualizar as amostras consideradas "ruins", em contra-partida aos outros métodos citados, que geralmente fornecem análises numéricas.

1.4 Organização da dissertação

O capítulo 2 trata da visão computacional; são definidos alguns termos relevantes e, principalmente, é tratada a escolha do sistema de representação de cores. Igualmente, são apresentadas as técnicas de extração de texturas relevantes para esse trabalho.

O capítulo 3 introduz as Redes Neurais Artificiais, especificando mais detalhadamente os Mapas Auto-Organizáveis de Kohonen (SOM) e suas versões hierárquicas, estudadas para a implementação da rede neural proposta.

O capítulo 4 fornece uma panorâmica do estado da arte na área de recuperação de imagens, mostrando os principais sistemas desenvolvidos.

O sistema propriamente dito e a etapas envolvidas em sua elaboração, são apresentados no capítulo 5. E finalmente, as conclusões, juntamente com as propostas de trabalhos futuros são discutidas no capítulo 6.

2 FEIÇÕES DE UMA IMAGEM

As feições de uma imagem referem-se às características que a descreve. No caso de um banco de imagens, tanto podem ser textuais, como visuais. As informações textuais, em geral, são textos que descrevem o conteúdo da imagem, enquanto que, as feições visuais trabalham com características perceptíveis na própria imagem, tais como: cor, textura, forma e estrutura. Como este trabalho não tem como intuito usar informações textuais, este capítulo se dedica apenas às feições visuais da imagem.

Além de ser discutidos os métodos estudados de extração de feições (cor, textura, forma e estrutura), também é dada uma breve introdução aos conceitos de Visão Computacional e à representação formal da imagem. Desta forma, espera-se esclarecer alguns conceitos das áreas onde se insere o trabalho. O caso específico da área de recuperação de imagens é tratado no próximo capítulo.

2.1 Conceitos de Visão Computacional

Segundo Jähne (JäHNE; HAUßECKER, 2000), a área de Visão Computacional(VC) é entendida como um conjunto de técnicas utilizadas para aquisição,processamento, análise e entendimento de dados complexos e com alta dimensionalidade, extraídos de nosso ambiente para exploração científica e técnica . A meta da VC é modelar e automatizar o processo de reconhecimento visual (FORSYTH; PONCE, 2002).

Dentro do universo de sistemas para automatização e reconhecimento de imagens é importante definir alguns termos que às vezes causam alguma confusão dentro da literatura, tais como Visão de Máquina e Visão Computacional, que não são considerados pela grande maioria dos pesquisadores como termos equivalentes (BATCHELOR; WALTZ, 2001). Desta forma, para evidenciar e esclarecer as diferenças entre os principais termos relacionados com a área abordada neste trabalho, se segue uma breve definição dos mesmos:

Visão Artificial Campo que se concentra na análise e projeto de sistemas opto-eletrônico-mecânicos que percebem o ambiente ao seu redor através da detecção de padrões espaço-temporais da radiação eletromagnética e processa essa informações (BATCHELOR; WALTZ, 2001). Ou seja, essa área concentra-se na construção de dispositivos capazes de substituir um sistema de visão natural. Um exemplo é o projeto de retinas artificiais, capazes de substituir uma retina biológica danificada.

Visão Computacional (VC) A Visão Computacional pode ser vista como uma área da ciência, visto que se concentra em aspectos como formalismo matemático e modelagem de sistemas de visão (BATCHELOR; WALTZ, 2001).

Visão de Máquina (VM) Do inglês *Maquine Vision*(MV), A Visão de Máquina possui um aspecto prático, mais voltado para a engenharia, que para a ciência. Para exemplificar, sistemas de automação industrial que envolvem inspeção visual, de modo geral são objetos de estudo da área de Visão de Máquina. Nesses casos, o ambiente é relativamente controlado, permitindo se utilizar heurísticas e algoritmos robustos "*ad hoc*"para resolver um problema especifico, como, por exemplo, detectar uma borda defeituosa em uma peça.

Visão Cognitiva Um sistema de visão cognitiva envolve entendimento, conhecimento e aprendizagem. Entendimento compreende tanto reconhecimento como categorização de objetos e eventos, através de rótulos semânticos dos dados da cena. Interpretação, a compreensão e reação aos modelos semânticos do ambiente. Já conhecimento, implica a necessidade de considerar a memória como uma base comum para a representação e manutenção da informação, incluindo métodos para acesso associativo.

Processamento Digital de Imagens É o estudo de algoritmos aplicados em imagens digitais. Essa área envolve extração de informação através do reconhecimento de padrões. Problemas típicos deste campo incluem transformações geométricas (rotação, redução, etc), correção de cores, brilho e contraste, quantificação, conversão entre espaços de cores, filtragem, segmentação, edição, redução de ruído, detecção de bordas, síntese de imagens entre outros. Esta área do conhecimento também cobre o tratamento de sinais tridimensionais como vídeo digital e tomografia.

2.2 Representação de imagens

Para representar e manipular imagens em um computador é necessário definir um modelo matemático apropriado (GOMES; VELHO, 1997). Como uma imagem é resultado de um estimulo de luz, é possível estabelecer um universo matemático no qual se possa definir modelos abstratos de imagens de forma que permitam sua representação discreta com o propósito de possibilitar uma codificação da mesma em um computador (GOMES; VELHO, 1997).

Considere um escala de cinza, tal como mostrado na figura 2.1, para representar uma imagem monocromática, *i* e *j* são dois números inteiros tais que $1 \leq i \leq m$ e $1 \leq j \leq n$. $f(i, j)$ é a função inteira tal que $1 \leq f(i, j) \leq W$, onde W indica o valor branco em uma escala de cinza. Nesta situação uma matriz F (figura 2.1) é chamada de imagem digital (BATCHELOR; WALTZ, 2001).

O endereço (i,j) define uma posição em F, chamado *pixel* (do inglês, *Picture Element*). Os elementos de F denotam a intensidade dentro da região definida por um pixel dentro da matriz que representa a imagem conforme a figura 2.2. A matriz F possui um total de m x n elementos e seu produto é chamado "resolução espacial de F". Isso significa que dado um conjunto de valores de intensidade, pode-se associá-los a tonalidades de cinza, dependendo do mapeamento. A figura 2.1 exemplifica este tipo de associação. A associação entre os valores e a tonalidade depende da capacidade de representá-los de forma discreta(número de bits) e do mapeamento feito pela função f(i,j). Como exemplo deste mapeamento, pode-se observar a figura 2.3, que mostra a imagem de um grampeador, representada através de 256 tons de cinza. O detalhe destacado e ampliado da figura 2.3 é mostrado em 2.4, e seus respectivos valores de intensidade estão discretizados na matriz da figura 2.3.

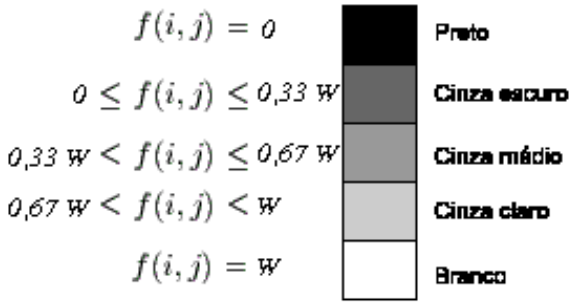

Figura 2.1: Exemplo de um mapeamento entre os valores armazenados na matriz F e as suas respectivas representações em tons de cinza (BATCHELOR; WALTZ, 2001).

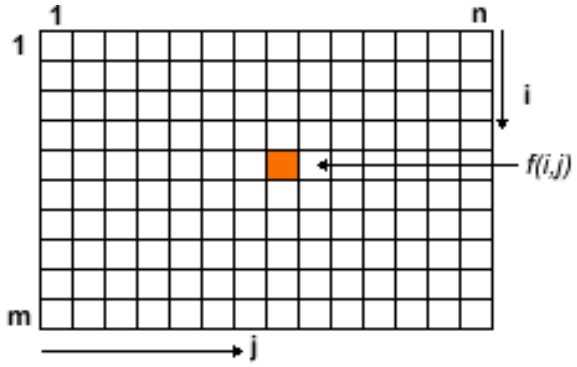

Figura 2.2: Imagem digital consistindo de um array de pixels mxn. O pixel na posição (i,j) tem intensidade igual a f(i,j) (BATCHELOR; WALTZ, 2001).

$$F = \begin{bmatrix} \text{f(1,1)} & \text{f(1,2)} & \text{...} & \text{f(1,n)} \\ \text{f(2,1)} & \text{f(2,2)} & \text{...} & \text{f(2,n)} \\ \text{f(3,1)} & \text{f(3,2)} & \text{...} & \text{f(3,n)} \\ \text{...} & \text{...} & \text{...} & \text{...} \\ \text{f(m,1)} & \text{f(1,2)} & \text{...} & \text{f(m,n)} \end{bmatrix} \tag{2.1}$$

A representação de uma imagem em escalas de cinza necessita de $log_2(1 + W)$ bits para cada pixel. Considera-se que $(1 + W)$ é uma potência inteira de dois; caso contrário $log_2(1 + W)$ deve ser arredondado para o mais próximo valor inteiro. Esta função de arredondamento é representada por $arrend()$.

Portanto uma imagem em escala de cinza m x n precisa de $arrend(log_2(1 + W))$ bits para cada pixel. Dessa forma, para armazenar toda a imagem F é necessário m x n $arrend(log_2(1 + W))$ bits, onde W representa os níveis de cinza.

Imagens coloridas trabalham de forma semelhante. Por exemplo, no caso do sistema RGB, existem três componentes, um para cada cor primária aditiva, R (vermelho), G(Verde) e B(Azul) onde cada componente pode ser representada de forma semelhante à matriz F de intensidade. Assim a matriz R representa a intensidade de cor em cada pixel do componente vermelho da imagem e assim por diante. Desta forma, tem-se então que: $R = r(i, j)$, $G = g(i, j)$ e $B = b(i, j)$, sendo que o vetor $r(i, j), g(i, j), b(i, j)$ define a intensidade e cor do pixel no ponto (i,j) em uma imagem colorida, conforme a figura 2.6. Neste caso, para armazenar uma imagem colorida é necessário m x n x

Figura 2.3: Imagem em tons de cinza de um grampeador sobre uma mesa (RUSSELL; NORVIG, 1995).

Figura 2.4: Detalhe ampliado da figura 2.3 (RUSSELL; NORVIG, 1995).

195	209	221	235	249	251	254	255	250	241	247	248
210	236	249	254	255	254	225	226	212	204	236	211
164	172	180	192	241	251	255	255	255	255	235	190
167	164	171	170	179	189	208	244	254	255	251	234
162	167	166	169	169	170	176	185	196	232	249	254
153	157	160	162	169	170	168	169	171	176	185	218
126	135	143	147	156	157	160	166	167	171	168	170
103	107	118	125	133	145	151	156	158	159	163	164
095	095	097	101	115	124	132	142	117	122	124	161
093	093	093	093	095	099	105	118	125	135	143	119
093	093	093	093	093	093	095	097	101	109	119	132
095	093	093	093	093	093	093	093	093	093	093	119

Figura 2.5: Valores da intensidade dos pixeis da figura 2.4 (RUSSELL; NORVIG, 1995).

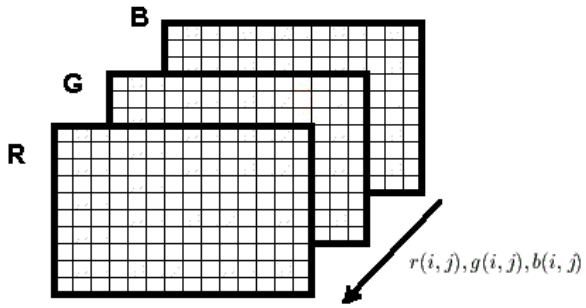

Figura 2.6: Componentes de uma imagem representada no formato RGB.

$rarrend(log_2(1+W))$ bits, onde *r* representa o número de componentes utilizados e W é o valor máximo de intensidade que cada cor pode alcançar em um canal.

2.3 Estudo dos sistemas de representação de cores

Para que se possa trabalhar com cores em um sistema gráfico é necessário construir uma representação que possa codificar a cor desejada. Existem vários sistemas de representação de cores. Nesta seção, serão tratados os sistemas pesquisados para a arquitetura desenvolvida,

2.3.1 Fundamentos da cor

O que é chamado de cor é uma faixa do espectro eletromagnético (figura 2.9)com comprimento de onda entre 380 nm e 780 nm. Esta é a faixa para a qual os olhos humanos são sensíveis. Nos seres humanos uma imagem é percebida através de dois tipos de "sensores": os bastonetes e os cones. Os bastonetes percebem variações de intensidade e são úteis durante o período noturno quando a quantidade de fótons é bastante reduzida; os cones são mais utilizados durante o período diurno e são responsáveis pela percepção da cor.

Através de experimentos psicológicos e fisiológicos, sabe-se que, o olho humano possui três diferentes tipos de cones, responsáveis pela sensação de cor, cuja sensibilidade é variável de acordo com o comprimento de onda (BATCHELOR; WALTZ, 2001). Para que esta sensação seja possível, os cones utilizam o fenômeno da tricromacidade, que permite reproduzir qualquer sensação de cor adicionando-se em diferentes proporções luz vermelha, azul e verde (cores aditivas primárias). Convém observar que este não é o único tipo de possibilidade de sistema visual encontrado na natureza. Existem animais dicromáticos (esquilos, coelhos), tetracromáticos(alguns pássaros) e pentacromáticos(pombo),que possuem quantidades de fotoreceptores para cor diferentes do ser humano (figura 2.7).

2.3.2 Teoria da Tricromacidade

A teoria da tricromacidade começa através da compreensão da estrutura da luz visível (figura 2.10), com Newton em 1666, pela separação dos componentes da luz branca com um prisma. Em 1801 Thomas Young sugere que é possível gerar qualquer cor através de três cores primárias aditivas. Essa teoria foi expandida por Hermann von Helmholtz que propõe que o olho humano percebe as cores com três graus de liberdade, sendo to-

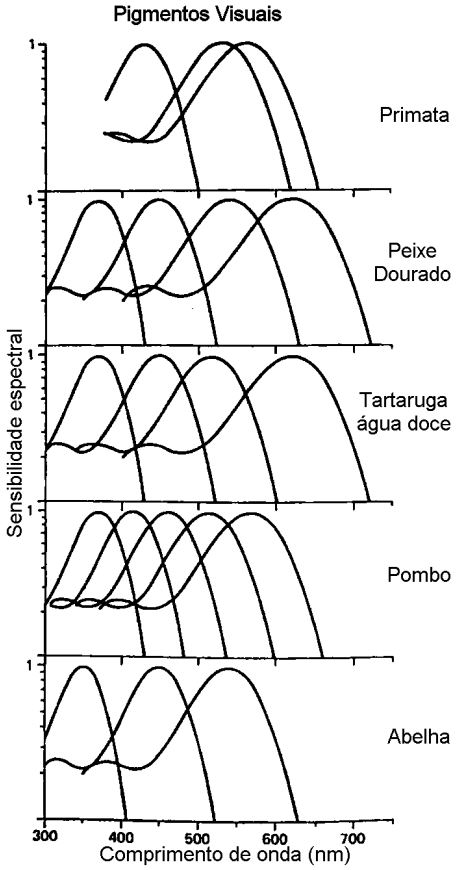

Figura 2.7: Mecanismos de percepção de cores através de diferentes pigmentos retinais em vários animais(VARELA; THOMPSON; ROSCH, 1997).

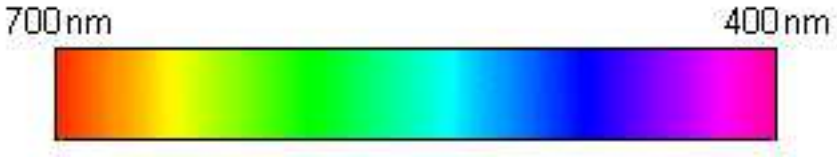

Figura 2.8: Espectro visível pelo olho humano.

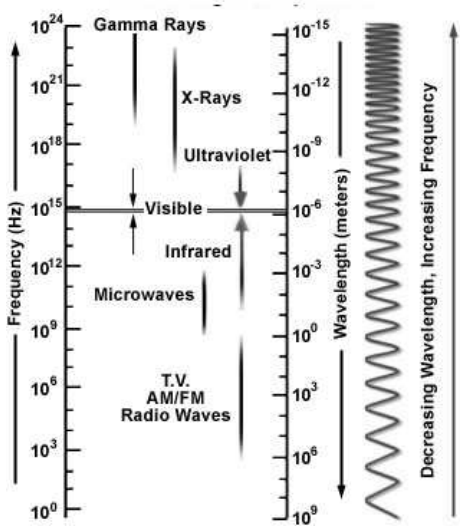

Figura 2.9: Espectro eletromagnético (DAVIDSON; NEAVES; ABRAMOWITZ, 2003).

das geradas pela mistura das primárias com suas correspondentes intensidades (escala de cores)(LEW, 2001).

Essa teoria foi confirmada em 1960 quando os três tipos de receptores foram encontrados na retina. A máxima resposta desses receptores (cones) correspondem ao azul (próximo de 440 nm), verde (próximo de 540 nm) e vermelho (próximo de 560 nm). Na figura 2.10 estão descritos as respectivas curvas de resposta de cada tipo de cone aos comprimentos de onda dentro do espectro visível.

2.3.3 Teoria do Oponente

A Teoria da Cor Oponente começa com Leonardo da Vinci que concluiu que as cores são produzidas pela mistura de amarelo e verde, azul e vermelho, preto e branco (LEW, 2001). A origem científica dessa teoria começa pela pesquisa do fisiologista do século XIX Edwald Hering que concluiu que a percepção do olho é baseada em três tipos de cores opostas. A forma moderna desta teoria é a proposta por Leo Hurvich e Dorothea Jameson em 1957 (VARELA; THOMPSON; ROSCH, 1997).

De acordo com a Teoria do Oponente, existem três canais de cores no sistema visual: um canal acromático[1] usado para perceber diferenças no brilho, e outros dois canais percebem variações na matiz de cores, e são portanto cromáticos. Esses canais são o vermelho-verde e o azul-amarelo. Essa teoria diz que vermelho, verde, azul e amarelo são as quatro matizes fundamentais ou as quatro psicologicamente únicas, que se combinam para formar matizes binárias complexas ou psicologicamente binárias.

[1]um canal acromático é aquele que não possui uma cor ou matiz. Na variação entre preto e branco existem várias escalas de cinza que são cores acromáticas, sem matiz

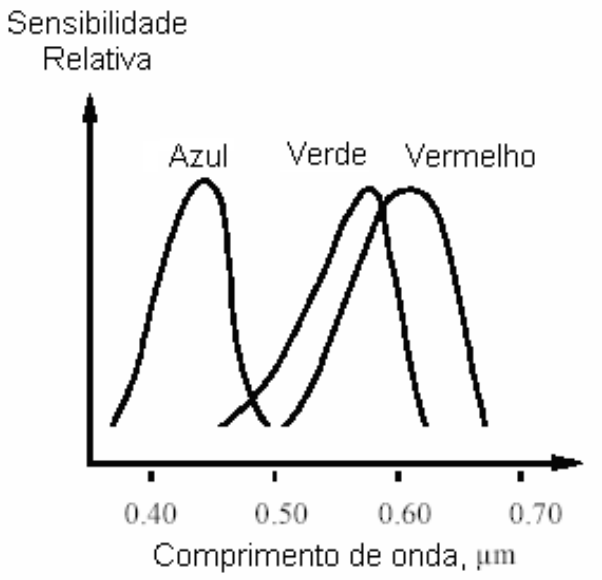

Figura 2.10: Curva de sensibilidade espectral dos cones do olho humano adaptado de (BATCHELOR; WHELAN, 2002).

De acordo com essa teoria, vermelho e amarelo podem formar cores amarelo-avermelhadas (laranjas), azul e vermelho podem formar cores púrpuras, entre outras combinações possíveis. Para cada matiz única, existe uma outra matiz única que não pode coexistir para formar uma matiz binária. Amarelo não pode coexistir com o azul (canal azul-amarelo) e vermelho não pode coexistir com verde (canal vermelho-verde), já que essas cores são **oponentes**.

A teoria das cores opostas explica a estrutura da aparência da cor através das diferentes respostas dos canais cromáticos e do acromático. Os seres humanos nunca experimentam nenhuma cor que seja uma combinação entre o vermelho e o verde ou azul e amarelo, pois um canal não pode ser simultaneamente vermelho e verde ou azul e amarelo (VARELA; THOMPSON; ROSCH, 1997). Essa teoria explica o porque de algumas matizes serem únicas e outras serem binárias. Matizes únicas são aquelas que provem de um canal cromático enquanto o outro canal é neutro ou balanceado. Matizes binárias provém da combinação de dois canais cromáticos.

Esses canais são observados em experimentos psicofísicos e não em neurofisiológicos (VARELA; THOMPSON; ROSCH, 1997). A teoria das cores oponentes foi confirmada em 1950 nas conexões ópticas entre o olho e o cérebro (LEW, 2001).

Atualmente a percepção das cores é entendida como a combinação de das teorias da cor oponente e da tricromaticidade. Desta forma, a luz é captada nos cones da retina (estímulo tricromático) e processada como cores oponentes no caminho para o cérebro. A diferença entre os cones que percebem ondas longas (em torno de 560 nm) e os canais de ondas médias (em torno de 530 nm) geram o canal vermelho-verde, e a diferença entre a soma dos canais de ondas longas e médias e o canal de ondas curtas (em torno de 440 nm) formam o canal amarelo-azul. O canal acromático é gerado pela soma de todos os sinais gerados pelos cones.

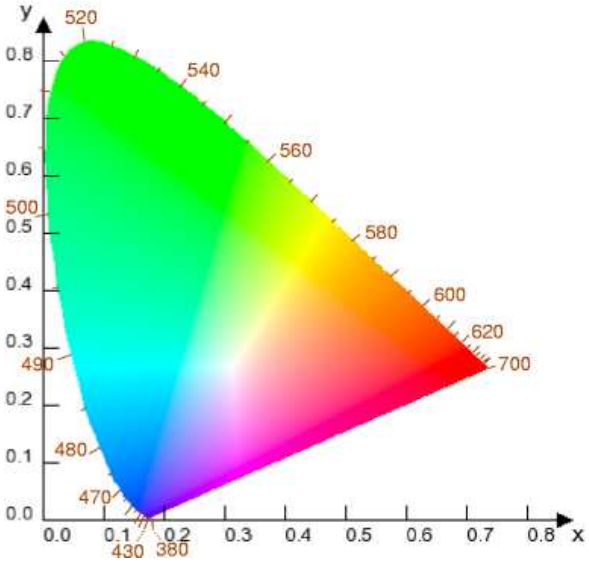

Figura 2.11: Diagrama de cromaticidade (GONZALEZ; WOODS, 1992)

2.4 Diagrama de Cromaticidade

As características normalmente utilizadas para distinguir uma cor de outra são o brilho, a matiz e a saturação (GONZALEZ; WOODS, 1992). O brilho é um descritor subjetivo que incorpora a noção cromática de intensidade; a matiz é o atributo associado com o comprimento de onda dominante em uma mistura de ondas de luz, e a saturação indica a quantidade de luz branca misturada com uma matiz.

A matiz e a saturação quando tomadas juntas são chamadas de cromaticidade. Desta forma, pode-se caracterizar uma cor por seu brilho e cromaticidade. O diagrama de cromaticidade (figura 2.11) é usado como referência padrão na definição de cores e de outros espaços de cores. Sendo assim, dado um conjunto de valores de triestímulos X,Y e Z necessários para formar uma cor; esta pode ser especificada por seus coeficientes tricromáticos ou coordenadas de cromaticidade de acordo com as equações 2.2, 2.3,2.4 e 2.5.

$$x = \frac{X}{X+Y+Z} \tag{2.2}$$

$$y = \frac{Y}{X+Y+Z} \tag{2.3}$$

$$z = \frac{Z}{X+Y+Z} \tag{2.4}$$

$$1 = x + y + z. \tag{2.5}$$

Como a soma das coordenadas é igual a uma unidade, apenas duas das três quantidades são necessárias para definir uma cor. Desta forma, quando os valores x(vermelho) e y(verde) são representados em um plano tem-se o diagrama de cromaticidade, a componente z (azul) pode ser calculada facilmente visto que $z = 1 - (x + y)$. Observe que a cor branca neutra está representa nas coordenadas $x = 0,3333$ e $y = 0,3333$ que corresponde a temperatura de 6.000 K (LEW, 2001).

2.5 Sistemas de representação de cores

As pessoas conseguem perceber uma grande gama de cores (estimada em 10 milhões de cores), com uma pequena variação de indivíduo para indivíduo. Apesar disso a quantidade de nomes para cores é bem limitada. O que leva muitas vezes, seja devido a características pessoais ou culturais, um mesmo nome designar cores diferentes, ou, a agrupar várias matizes. Portanto para que se possa trabalhar com cores é necessário definir um modelo matemático para representá-las as cores. Um modelo de cor é um modelo matemático abstrato que descreve a forma como as cores podem ser representadas como tuplas (geralmente três ou quatro componentes de cor). O Conjunto de cores compostas por estas tuplas é chamando espaço de cores[2].

A criação de padrões para descrever cores é de grande importância em sistemas de reconhecimento de objetos, pois de acordo com o método empregado para medir a similaridade entre as cores pode se escolher um sistema que enfatize propriedades particulares ou espaços de cores uniformes que capturem a significância das diferenças entre as cores (JäHNE; HAUßECKER, 2000). Abaixo, são apresentados os sistemas estudados para a arquitetura proposta, ressaltando que existem outros sistemas além destes.

2.5.1 Espaço de Cores XYZ

Foi um dos primeiros a ser definido formalmente. Este sistema (tabela 2.1) é baseado nas percepções de um observador padrão de um objeto pintado com uma determinada cor, iluminado por uma fonte de luz D65 (LEW, 2001). Desta forma os valores do triestímulo XYZ são computados através da adição do produto de luz, objeto e funções de casamento para cada comprimento de onda.

Os componentes X, Y e Z por possuírem valores saturados não podem ser vistos pelo olho humano ou produzidos artificialmente; desta forma, essas cores primárias são cores imaginárias. Isso não influencia na representatividade do sistema, pois, qualquer cor percebida pode ser descrita. Outra propriedade interessante é que a luminância é determinada apenas pelo valor de Y.

Tabela 2.1: Espaço de Cores XYZ

Modelos de cor	X,Y,Z
Características	Independente do dispositivo. Perceptualmente não uniforme. Transformação Linear. Não intuitivo. Dependente do ângulo de visão, geometria do objeto, direção da iluminação, intensidade e cor da iluminação.
Conversão [3]	$X = 0,607R + 0,174G + 0,200B$, $Y = 0,299R + 0,587G + 0,114B$, $Z = 0,000R + 0,066G + 1,116B$.
Observação	Todas cores percebidas são descritas matematicamente pelas três cores primárias, a luminância é baseada apenas no eixo Y

[2] o modelo que descreve um espaço de cores também é chamado de sistema de cor

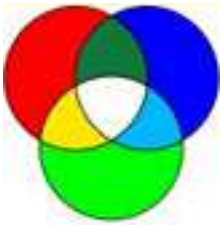

Figura 2.12: Resultado da composição aditiva no sistema RGB (GOMES; VELHO, 1997).

2.5.2 Sistema de Cor GRAY

Este modelo (tabela 2.2) também conhecido como escala de cinza ou intensidade. Ele é calculado a partir dos valores de RGB fornecidos por uma câmera digital; o que o torna dependente do dispositivo de aquisição de imagens. Este sistema não é perceptualmente uniforme, pois uma diferença do brilho entre dois valores não significa necessariamente que haja diferença entre dois sucessivos valores de cinza, sendo o sistema influenciado pelas condições de aquisição da imagem (LEW, 2001).

Tabela 2.2: Espaço de Cores em escala de cinza (LEW, 2001)

Modelos de Cor	GRAY
Características	Dependente do dispositivo. Perceptualmente não uniforme. Linear. Intuitivo. Dependente do ângulo de visão, geometria do objeto, direção intensidade e cor da iluminação.
Conversão	$GRAY = 0,299R + 0,587G + 0,144B.$
Observação	Informação em escalas de cinza.

2.5.3 Sistema RGB

O sistema RGB (do inglês *Red,Green,Blue*) (tabela 2.3) assemelha-se bastante a codificação de cores do olho humano, pois é baseado na teoria da tricomaticidade. Ele possui três sensores de cores geralmente associados ao vermelho (R), azul(B) e verde(G), mas que na verdade são sensíveis não a cor descrita em si, mas a certas faixas de comprimento de onda, sendo mais correto falar em sensores sensíveis a freqüências baixas(R), médias(G) e altas(B) do espectro da luz visível. O sistema RGB é um sistema de representação de cores aditiva, uma vez que, para se construir uma nova cor adicionamos as cores primarias aditivas, conforme pode-se observar na figura 2.12.

O sistema RGB geralmente é representado como um cubo, onde os eixos cartesianos representam as componentes primárias R,G e B (figura 2.13). As componentes podem ser somadas de forma a produzir qualquer cor dentro do cubo (RUSS, 1998). O sistema RGB pode também ser visto como um triângulo de cores (também conhecido como triângulo de Maxwell). Neste sistema, proposto por James Clerk Maxwell, três luzes coloridas altamente saturadas (vermelho, azul, verde) são vistas como vértices de um triângulo equilátero onde qualquer outro ponto neste representa uma mistura destas três cores primárias conforme mostra a figura 2.14 (BATCHELOR; WALTZ, 2001).

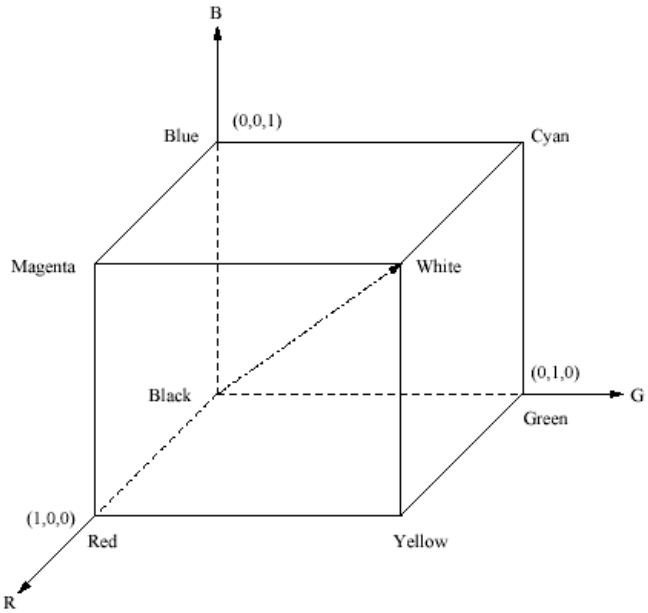

Figura 2.13: Cubo RGB (SHAPIRO; STOCKMAN, 2001)

Figura 2.14: Triângulo de Maxwell (BATCHELOR; WALTZ, 2001)

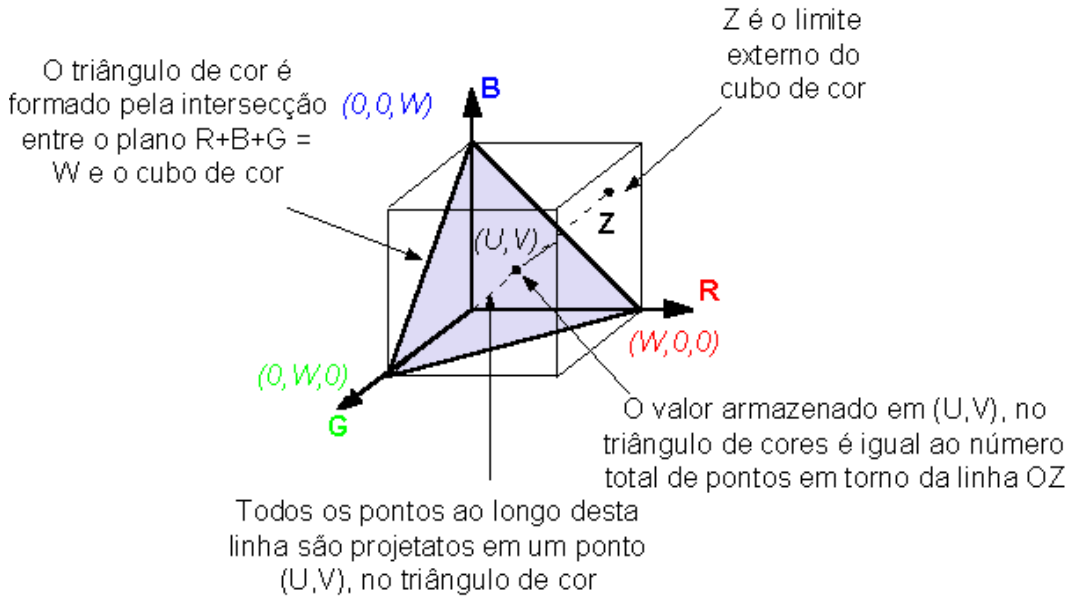

Figura 2.15: Espaço RGB - Cubo de cor e Triângulo de Cor (BATCHELOR; WALTZ, 2001)

A figura 2.15 mostra o triângulo de Maxwell dentro da representação do cubo RGB. Neste cubo (SHAPIRO; STOCKMAN, 2001) a cor e a saturação podem ser identificadas como um ponto localizado no triângulo, mas a intensidade não pode ser posicionada no mesmo triângulo pois ela é representada como um deslocamento ortogonal.

Tabela 2.3: Espaço de Cores RGB (LEW, 2001)

Sistema de Cor	RGB
Características	Dependente do dispositivo. Perceptualmente não uniforme. Linear. Não Intuitivo. Dependente do ângulo de visão, geometria do objeto, direção, intensidade e cor da iluminação.
Conversão	Identidade
Observação	Não é necessário conversão, visto que é o espaço de cores usado como referência.

2.5.4 Sistema de Cores CMY

A luz branca é uma radiação eletromagnética cujo comprimento de onda λ se localiza na faixa entre 400 e 700 nanometros causando a sensação de cor. A luz ao incidir sobre um objeto pode ser refletida quase que totalmente se for um objeto branco. Se for colorido apenas parte do espectro visível será refletido criando a sensação de cor conforme a freqüência refletida. Desta forma um objeto é branco (ideal) se reflete praticamente toda a luz incidente sobre eles e um objeto é negro (ideal) se absorve toda a luz incidente sobre

Figura 2.16: Resultado da composição subtrativa no sistema CYM (GOMES; VELHO, 1997).

ele, não refletindo nenhuma comprimento de onda. Neste caso, um objeto vermelho é aquele que reflete apenas o comprimento de onda relativo a esta cor e assim por diante. Ao misturar-se pigmentos com cores diferentes, a cor obtida é resultante da subtração dos comprimentos de onda possíveis de serem refletidos por pigmentos originais, conforme a figura 2.16. O sistema de representação de cores CMYK (*Cyan, Mangenta,Yellow, Key - Black*),descrito na tabela 2.4,funciona desta forma; ao contrário do sistema RGB, esta é uma representação de cores subtrativa. O CMYK é mais utilizado em impressões em papel, pois neste caso as cores são formadas pela combinação de tintas (BATCHELOR; WALTZ, 2001).

O sistema de cores CMY é complementar ao sistema RGB, ou seja, cyan é complementar ao vermelho, magenta ao verde e amarelo ao azul (FOLEY; VANDAM, 1982). O subespaço ocupado pelo CMY é o mesmo do RGB, com diferença da origem do centro de coordenadas, que para o CMY é o branco.

Como a representação do preto, através do CMY não é muito eficaz, no processo com tintas foi acrescentado mais um componente somente para especificar o preto (canal K),produzindo uma reprodução mais fiel da imagem.

Tabela 2.4: Espaço de Cores CMY

Sistema de Cor	CMY
Características	Dependente de dispositivo. Perceptualmente não uniforme. Linear. Não Intuitivo. Dependente do ângulo de visão, geometria do objeto,brilho, direção, intensidade e cor da iluminação.
Conversão	$C = 1 - R.$ $M = 1 - G.$ $Y = 1 - B.$
Observação	Para representa o no caso do CMYK preto utiliza-se o valor mínimo entre as três componentes básicas. $K = Min(C, M, Y).$

2.5.5 Sistema de Cor YIQ e YUV

Esse sistema (tabela 2.5) foi desenvolvido pela National Television Systems Committee (NTSC), visando a eficiência da transmissão de TV em cores. O componente Y

corresponde à luminância, o I para *in-phase* e o Q para a quatratura Q. Pode-se pensar I e Q como correspondentes próximos ao componentes Hue(Matiz) e S (Saturação). Nos sistemas PAL e SECAM o sistema de cor utilizado é o YUV. Ambos são iguais. A única diferença entre eles é que o plano I-Q difere do U-V por uma simples rotação de 33 graus sobre o eixo.

Tabela 2.5: Espaço de Cores YIQ (LEW, 2001)

Sistema de Cor	YIQ
Características	Independente de dispositivo. Perceptualmente não uniforme. Transformação Linear. Não Intuitivo. Dependente do ângulo de visão, geometria do objeto,brilho, direção, intensidade e cor da iluminação.
Conversão	$Y = 0,299R + 0,587G + 0,144B$. $I = 0,596R - 0,274G - 0,312B$. $Q = 0,211R - 0,523G + 0,312B$.
Observação	Y é a luminância da cor.

2.5.6 Sistema U*V*W*

Este sistema de coordenadas possui três características de cores: U*, V* e W*. O modelo de cor W* é baseado em uma escala de luminância que possui valores entre 0(preto) e 100 (branco), sendo que o método de escalonamento começa com preto e seleciona o valor de cinza mais perceptível. A partir deste valor,procura-se o próximo valor mais perceptível até chegar no branco (LEW, 2001). Este sistema fornece uma cor sólida quando ocorrem mudanças unitárias perceptíveis na luminância e na crominância. O U*V*W se tornou obsoleto em virtude da criação dos sistemas L*a*b e do L*u*v (PRATT, 2001).

2.5.7 Sistema L*a*b*

Este sistema (tabela 2.7) é baseado no sistema XYZ (RUSS, 1998). O sistema de coordenadas é construído de acordo com a teoria da cor oponente: L* é o canal preto-branco, a* o vermelho-verde e b* o amarelo-azul. Este sistema foi projetado para ser visualmente uniforme, ou seja as distâncias numéricas do sistema são relacionadas às diferenças perceptuais humanas (LEW, 2001). Essa é a grande vantagem deste sistema para recuperação de imagens, pois se percebemos que uma cor é semelhante a outra o mesmo ocorre no L*a*b. Cores perceptualmente semelhantes são próximas nesse espaço de cores. Infelizmente, tanto o L*a*b quanto o RGB são dependentes das condições da imagem.

2.5.8 Sistema L*u*v*

Também é um sistema visualmente uniforme baseado no XYZ. O componente L* define a luminância e o u* x v* definem a crominância. A tabela 2.8 traz considerações gerais desse sistema.

Tabela 2.6: Espaço de Cores U*V*W* (LEW, 2001)

Modelos de Cor	U*, V*, W*
Características	Independente de dispositivo. Perceptualmente uniforme. Transformação não-Linear: Instável quando a intensidade é pequena. Não Intuitivo. Dependente do ângulo de visão, geometria do objeto,brilho, direção da iluminação, intensidade e cor da iluminação.
Conversão	$U* = 13W * (u - u_0)$. $V* = 13W * (v - v_0)$. $W* = \begin{cases} 116(\frac{Y}{Y_0})^{\frac{1}{3}} - 16 \text{ se } \frac{Y}{Y_0} > 0,008856. \\ 903.3(\frac{Y}{Y_0}) \text{ se } \frac{Y}{Y_0} \leq 0,008856i. \end{cases}$ $u = \frac{4X}{X+15Y+3Z}$, $v = \frac{6Y}{X+15Y+3Z}$ $u_0 = \frac{4X_0}{X_0+15Y_0+3Z_0}$ $v_0 = \frac{6Y_0}{X_0+15Y_0+3Z_0}$
Observação	O modelo é visualmente uniforme.

Tabela 2.7: Espaço de Cores L*a*b (LEW, 2001)

Modelo de Cor	L*a*b*
Características	Independente de dispositivo. Perceptualmente não uniforme. Transformação não linear. Instável quando a intensidade é pequena Não Intuitivo. Dependente do ângulo de visão, geometria do objeto,brilho, direção da iluminação, intensidade e cor da iluminação.
Conversão	$L* = 116(\frac{Y}{Y_0})^{\frac{1}{3}} - 16 \text{ se } \frac{Y}{Y_0} > 0,008856,$ $903.3(\frac{Y}{Y_0}) \text{ se } \frac{Y}{Y_0} \leq 0,008856.$ $a* = 500[(\frac{X}{X_0})^{\frac{1}{3}} - (\frac{Y}{Y_0})^{\frac{1}{3}}]$ $b* = 200[(\frac{Y}{Y_0})^{\frac{1}{3}} - (\frac{Z}{Z_0})^{\frac{1}{3}}]$

2.5.9 Sistema de Cores HSI

A visão humana pode ser considerada como uma amostragem baseada em parâmetros, que não mede diretamente o fluxo radiante espectral [4], mas algumas propriedades da distribuição espectral como o fluxo radiante total (intensidade),o comprimento médio de onda(cor) e a largura da distribuição espectral (saturação da cor). Se a largura da distribuição espectral é estreita, então se tem uma cor pura com saturação elevada. Caso a distribuição espectral for larga a cor tem baixa saturação e se for baixa, nenhuma cor é percebida (JäHNE; HAUßECKER, 2000). O sistema de cor HSI é baseado nessas informações (Tabela 2.9), sendo H a matiz (*HUE*), S a saturação (*Saturation*) e I a Intensidade (*Intensity*).

O problema deste sistema é o fato do canal H tornar-se instável quando S é próximo de

[4]Fluxo Radiante é a potência da radiação, ou seja é energia por unidade de tempo, que descreve o total de energia emitida por uma fonte de luz por unidade e tempo

Tabela 2.8: Espaço de Cores L*u*v*

Modelo de Cor	L*u*v*
Características	Independente de dispositivo. Perceptualmente não uniforme. Transformação não linear. Instável quando a intensidade é pequena Não Intuitivo. Dependente do ângulo de visão, geometria do objeto,brilho, direção da iluminação, intensidade e cor da iluminação.
Conversão	$L* = \begin{cases} 116(\frac{Y}{Y_0})^{\frac{1}{3}} - 16 \text{ se } \frac{Y}{Y_0} > 0,008856. \\ 903.3(\frac{Y}{Y_0}) \text{ se } \frac{Y}{Y_0} \leq 0,008856. \end{cases}$ $u* = 13L * (u' - u'n)$ $v* = 13L * (v' - v'n)$ $u' = \frac{4X}{X+15Y+3Z}$, $v' = \frac{9Y}{X+15Y+3Z}$ $u'n = \frac{4Xn}{Xn+15Yn+3Zn}$, $v'n = \frac{9Yn}{Xn+15Yn+3Zn}$

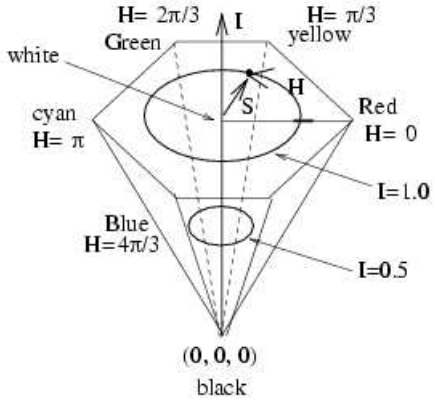

Figura 2.17: Espaço HSI - Hexacubo (SHAPIRO; STOCKMAN, 2001)

zero, o que se deve a singularidade não removíveis na transformação não-linear. Esta particularidade permite que uma pequena variação no sinal de entrada cause grandes saltos nos valores convertidos.

O HSI pode ser representado como um "hexacone", conforme a figura 2.17, onde a intensidade é o eixo vertical, a cor é o ângulo de 0 a 2 π; sendo que o ângulo zero demarca a cor vermelha. A faixa de saturação é entre 0 e 1 sendo denotada por S na figura. Quanto mais pura a cor, mais próximo de 1 seu valor de saturação. Nesta representação, o branco corresponde ao eixo I (saturação = 0).

2.6 Considerações finais sobre os sistemas apresentados

Cada sistema de recuperação de imagens pode utilizar um espaço de representação de cores de acordo com suas necessidades. Portanto, na hora de escolher o sistema é importante ter uma visão abrangente das principais características de cada um. Geralmente as descrições se limitam ao RGB e o HSI, mas o L*a*b e o L*u*v* estão sendo utiliza-

Tabela 2.9: Espaço de Cores HSI (LEW, 2001)

Sistema de Cor	HSI
Características	Dependente de dispositivo. Perceptualmente não uniforme. Intensidade é linear. Saturação é não-linear, se torna instável quando a intensidade é próxima de zero Matiz (Hue) é não linear,se torna instável quando a intensidade e a saturação são próximas de zero Intuitivo. Intensidade I: Dependente do ângulo de visão, geometria do objeto, direção da iluminação, intensidade e cor da iluminação. Saturação S : Dependente dos pontos mais iluminados em uma imagem e da cor da iluminação. Matiz H : Dependente da cor da iluminação.
Conversão	$H = \cos^{-1}\left(\frac{(2R-G-B)}{2\sqrt{(R-G)^2+(R-B)(G-B)}}\right)$. $S = 1 - \frac{3\min(R,G,B)}{(R+G+B)}$. $I = \frac{(R+G+B)}{3}$.
Observação	A intensidade do brilho de um ponto em uma cena, está relacionada com a quantidade de cada valor de R, G e B.

dos em vários sistemas de recuperação de imagens, e muitas vezes passa despercebido o motivo da escolha (no caso a relação de variação x percepção humana).

2.7 Textura

Segundo (GONZALEZ; WOODS, 1992), apesar de não existir uma definição formal de textura, ela pode ser entendida intuitivamente como uma característica que fornece informações de suavidade, rugosidade e regularidade de uma região. A textura de uma imagem pode ser avaliada em termos de freqüência espacial. A freqüência espacial de uma imagem informa a taxa de mudança da intensidade dos *pixels*. Por exemplo, regiões cuja a intensidade dos *pixels* é aproximadamente constante, caracterizam locais de baixa freqüência, já regiões em que existe um grande variação dos valores de intensidade dos*pixels* são locais de alta freqüência espacial. A figura 2.18 exemplifica ambas situações. A grade, que forma a textura da imagem, é mais estreita em direção ao centro, formando dois eixos que dividem a imagem em quatro quadrantes. Nesses eixos se observa uma alta freqüência espacial, visto que a taxa de mudança da intensidade dos pixels é maior nessa região. A medida que se distância desse eixo de maior freqüência, a variação dos pixels se torna menor, fazendo com que nos cantos da imagem a freqüência espacial seja mais baixa em relação ao resto da imagem.

A textura caracteriza muitos objetos, e em alguns casos é a principal maneira de diferenciá-los. No caso da figura 2.19 estão dois animais de espécies diferentes,um cav-

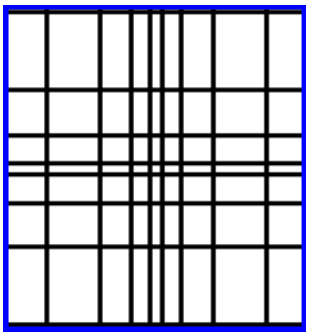

Figura 2.18: Textura composta por várias freqüências espaciais.

Figura 2.19: Imagem de uma zebra e um cavalo sem a componente cromática.

alo e uma zebra, mas com características morfológicas muito semelhantes. Apesar de existirem detalhes visuais que diferenciam cada um dos animais, a característica inicial utilizada para diferenciar os animais, é a textura do dorso de cada animal. A zebra possui geralmente o corpo coberto por listras pretas e brancas, enquanto o cavalo possui um corpo malhado ou liso. Claramente, pode-se observar que na figura 2.19 a zebra possui uma freqüência espacial maior que o cavalo da mesma figura.

Existem várias possibilidades para análise de texturas em imagens. As técnicas a seguir apresentadas, são mencionadas por sua relevância e utilização na área de recuperação de imagens.

2.7.1 Transformada de Fourier

O princípio de funcionamento da Transformada de Fourier consiste, essencialmente, em decompor ou separar uma forma de onda ou função em várias senóides de diferentes freqüências, cujo resultado do somatório de todas elas é igual à forma de onda original. Em outras palavras, é possível reconstruir qualquer função unidimensional $f(x)$ a partir da soma de termos senos e cossenos com incrementos de freqüência (WALKER, 1999). A Transformada de Fourier é ótima para tratar de sinais estacionários ou de processos estacionários [5], e citando o exemplo dado em (FEICHTINGER; STROHMER, 2001), a transformada de Fourier pode nos dar muita informação sobre um trecho musical, tais como a transição das notas e as notas prevalentes em termos de freqüência, mas informações como o momento de execução e duração ficam mascaradas (FEICHTINGER; STROHMER, 2001).

A transformada de Fourier, em termos de imagem (Transformada de Fourier Bi-Dimensional), aponta periodicidades espaciais na intensidade de uma imagem, ou seja, é possível encontrar freqüências dominantes que caracterizam uma imagem (BATCHELOR; WHELAN, 2002).

A transformada discreta de Fourier bidimensional, para uma imagem NxN,dita $f(x, y)$,

[5] Processos que possuem propriedades estatisticamente invariantes no tempo

Figura 2.20: Transformada de Fourier aplicada a uma imagem monocromática (PRATT, 2001).

sendo x e y as coordenadas dos *pixels* da imagem é definida como (equação 2.6):

$$F(u,v) = \frac{1}{N}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\exp[-j2\pi\frac{(ux+vy)}{N}] \tag{2.6}$$

onde $0 \leq u, v, \leq N-1$.

A resultante dessa equação contém dois componentes, o real e o imaginário, ou, fase e magnitude. A figura 2.20 mostra a componente magnitude, com suas freqüências devidamente ordenadas, da transformada de Fourier aplicada sobre a imagem da esquerda.

A transformada de Fourier é uma técnica de análise de freqüência global, que não armazena a informação espacial de cada freqüência. Como para a busca de imagens é relevante a disposição espacial, a transformada de Fourier não é utilizada nos sistemas de recuperação de imagens atuais. Nestes sistemas, a preferência tem sido por técnicas que preservem a informação espacial.

2.7.2 Filtros de Gabor

Como a transformada de Fourier perde a informação espacial, ou seja, como ela modela amplitude por freqüência, a informação do instante em que ocorreu um evento não é guardada. A saída proposta por Gabor, que ele chamou de *Short-Time Fourier Transform*, é bem simples. Imagine um sinal unidimensional cuja amplitude varie com o tempo, o que Gabor propôs foi aplicar a transformada de Fourier a apenas uma porção deste sinal, através de uma janela de dimensões pré-determinadas. Desta forma, o sinal é mapeado para dentro de uma função de tempo e freqüência. Em outras palavras, dado um sinal qualquer, é aplicada a Transformada de Fourier em intervalos de tempo determinados por uma janela que se desloca ao longo do sinal. Essa janela é uma função bidimensional que mapeia o sinal em função do tempo e da freqüência. No caso de uma imagem, se divide a mesma em várias regiões, e a cada uma dessas é aplicada a Transformada de Fourier. Desta forma, a informação resultante seria uma imagem com várias transformadas divididas em quadrantes. Com essa estratégia, mantém-se a informação espacial. Claro que a granularidade e precisão dessa informação depende do tamanho escolhido para a janela.

Quando a janela escolhida é uma Gaussiana, toda essa estrutura passa a se chamar Filtro de Gabor (LEW, 2001).

Os filtros de Gabor são filtros passa banda sensíveis a orientação e freqüência (LAMPINEN; SMOLANDER, 1996), permitindo a extração de informações referentes a feições da imagem dependentes da orientação, tais como contornos e texturas.

O kernel do filtro de Gabor é uma sinosóide localizada dentro de uma janela gaussiana,

Figura 2.21: Exemplo do filtro de Gabor aplicado a uma imagem. $\sigma = 2$, $\theta = 3$ e $f = 2$.

operando diretamente no domínio espacial da imagem. Formalmente, o filtro de Gabor é definido pela equação 2.7:

$$\psi(f, \theta, x, y) = exp(i(f_x x + f_y y) - \frac{f^2(x^2 + y^2)}{2\sigma^2}) \tag{2.7}$$

onde x e y são as coordenadas do pixel na imagem, f é a freqüência central da banda passante, sendo a freqüência x da janela determinada por $f_x = f\cos\theta$ e freqüência y determinada por $f_y = f\sin\theta$ e $i = \sqrt{-1}$. A variável θ determina a orientação espacial, e σ determina a largura da banda passante do filtro.

Os filtros de Gabor são freqüentemente utilizados em sistemas de recuperação de imagens baseados na similaridade de textura (IQBAL; AGGARWAL, 2002). Um dos principais problemas é determinar a resolução, a faixa de freqüência e orientação, que serão utilizadas para determinar a textura. Mas, para isso existem outras técnicas baseadas em filtros de Gabor, tais como Gabor Wavelets (MANJUNATH; MA, 1996) e Gabor Frames (FEICHTINGER; STROHMER, 2001).

2.7.3 Wavelets

As Wavelets constituem-se em uma ferramenta extremamente útil para a análise de uma imagem e para a extração de feições da mesma. Dentro do ponto de vista dos sistemas de recuperação de imagens, as wavelets são em geral utilizadas para extrair as informações de textura, algumas vezes com algumas adaptações.

As wavelets são classes de formas de ondas de duração limitada, cujo valor médio é zero. Como elas são utilizadas em várias escalas e posições conseguem capturar informações de um sinal com várias resoluções, mantendo a informação espacial. Ao contrário das senóides usadas na Transformada de Fourier, que são simétricas e suaves, as Wavelets tendem a ser assimétricas e irregulares. Uma família de Wavelets, de forma simplista, é um conjunto de funções continuas, com momentos nulo, com rápido decréscimo quanto x tende para infinito, ou é nula em segmentos fora do conjunto dos números Reais (R).

Basicamente, para realizar a análise via Wavelets, procede-se da seguinte forma: escolhida a Wavelet a ser usada, a mesma é aplicada sobre um sinal,considerando varias escalas e deslocamentos. Assim sendo, para um sinal de entrada, são gerados uma matriz de coeficientes wavelets, que representam o grau de similaridade entre o sinal avaliado e a wavelet utilizada. Quanto maior for o valor do coeficiente, maior é a similaridade entre os sinais. Cada escala escolhida para a wavelet representa uma resolução distinta na análise. Desta forma, diferentes escalas, representam níveis diferentes de granularidade na avaliação do sinal.

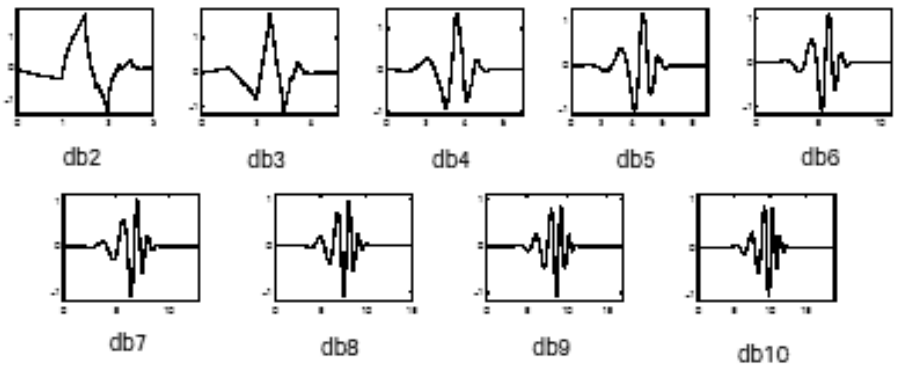

Figura 2.22: Exemplo de classe de wavelets: Daubechies (MISITI et al., 2001).

2.8 Formas

A descrição da forma de um objeto é uma importante feição que pode ser usada no reconhecimento do mesmo. Infelizmente, a recuperação baseada na forma de um objeto, só pode ser feita em ambientes muito restritos (LüNEBURG, 2002), pois existem alguns problemas na segmentação automática de uma imagem, que impedem a extração da forma correta. Apesar disso, bons trabalhos têm sido feitos na área, como o Blobworld(CARSON et al., 1999) que possui um método bem interessante. Outra possibilidade é utilizar métodos que não necessitem de segmentação, mas que computem dados estatísticos levantados das propriedades da forma de toda a imagem, tais como o histograma de direção de ângulos (SAMI BRANDT JORMA LAAKSONEN, 2000).

Como a extração da forma não foi tratada na arquitetura aqui desenvolvida, pois o trabalho lida com imagens em vários formatos e com objetos de formas variáveis (corpos não rígidos como o humano) ou com um formato não definíveis (montanhas, paisagens, florestas), fica como sugestão de um trabalho futuro, agregar a arquitetura de segmentação proposta no BlobWorld(CARSON et al., 1999) como entrada deste sistema.

3 REDES NEURAIS E MAPAS AUTO-ORGANIZÁVEIS

3.1 Introdução

Durante o século XIX, os fisiologistas descobriram que o cérebro é um sistema complexo, composto por células individuais capazes de receber e transmitir sinais através de uma pequena corrente elétrica (MAINZER, 1997).Estas células, denominadas posteriormente de neurônios, são as responsáveis pelo macrocomportamento do cérebro.

Basicamente, um neurônio transmite um sinal para outros neurônios via conexões sinápticas[1]. Dada descoberta deste comportamento alguns modelos matemáticos do funcionamento destas redes de neurônios começaram a surgir. Cada um desses modelos, por serem apenas abstrações do neurônio biológico, acabam por enfatizar algumas características e negligenciar outras. Não necessariamente um modelo de rede neural possui uma grande plausibilidade biológica, pois alguns destes modelos, têm objetivos específicos que não implicam em uma fidelidade ao sistema biológico, apenas utilizam genericamente o conceito de neurônio artificial.

Portanto, uma Rede Neural Artificial ou simplesmente Rede Neural é um conjunto de processadores distribuídos trabalhando de forma paralela constituídos de unidades de processamento simples, que tem a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso (HAYKIN, 2001). A unidade básica de uma rede neural artificial é um modelo simplificado que se assume ter o mesmo comportamento funcional de um neurônio biológico (VEELENTURF, 1995). As redes neurais são também conhecidas como sistemas conexionistas ou de processamento paralelamente distribuído (*parallel distributed processing*) (LUGER; STUBBLEFIELD, 1998) cuja a unidade básica de processamento é o neurônio artificial.

3.1.1 Neurônio Biológico

O funcionamento exato do cérebro ainda permanece um mistério, mas algumas funções e estruturas básicas foram esclarecidas no último século. Desta maneira, existem algumas pistas sobre o funcionamento das unidades básicas de processamento do cérebro, mas ainda estamos distantes da compreensão da origem da consciência e da inteligência.

A unidade fundamental de todo sistema nervoso é o neurônio (figura 3.1). O cérebro humano possui cerca de aproximadamente 100 bilhões de neurônios, alguns com milhares de conexões o que nos dá em torno de 10^4 conexões somente no córtex cerebral (SAGAN, 1992).

[1]Sinapse é a região eletroquimicamente ativa compreendidas entre duas membranas celulares de dois neurônios (KOVáCS, 1996). Esta região é o ponto de contato entre o dendrito da célula nervosa, que tem função de receber os impulsos nervosos oriundos de outras células, e o axônio responsável pela transmissão (CARVALHO; BRAGA; LUDEMIR, 1998)

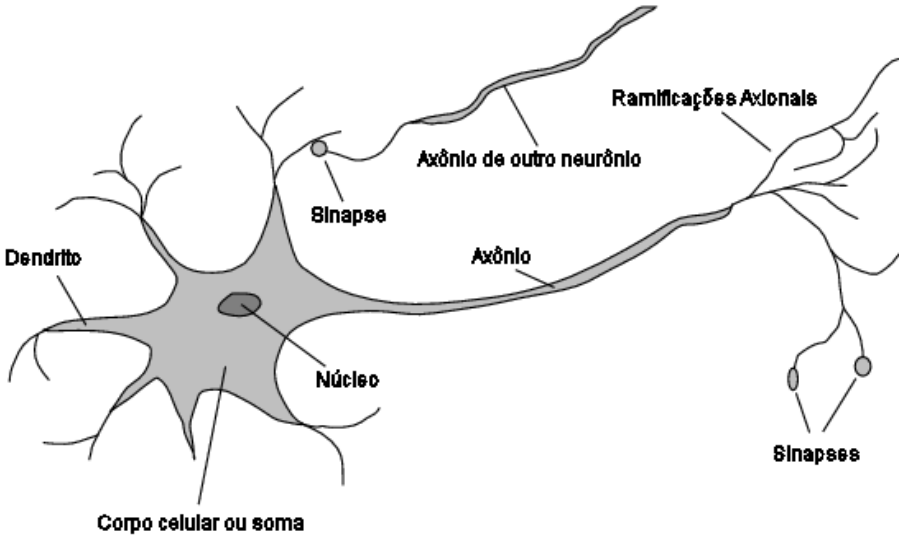

Figura 3.1: Neurônio ou célula nervosa (RUSSELL; NORVIG, 1995).

O neurônio é basicamente constituído pelas seguintes unidades:

- Corpo celular, também chamado soma, que contém o núcleo celular;
- Dendritos, que são responsáveis pela entrada dos sinais oriundos de outros neurônios;
- Axônio, que é responsável pela transmissão do sinal.

As sinapses são formadas pela junção axônio e dendrito conforme a figura 3.1. A transmissão do sinal se dá através de uma reação eletroquímica que dispara os neurotransmissores que estão armazenados dentro das vesículas do axônio. Cada neurotransmissor tem um receptor apropriado no dendrito. Esses processos químicos aumentam ou diminuem o potencial elétrico do corpo celular. Quando o potencial alcança um limiar, um pulso elétrico ou potencial de ação é enviado pelo axônio para os dendritos de outro neurônio. As sinapses que aumentam o potencial são chamadas excitatórias e as que diminuem o potencial são chamadas inibitórias. Como um neurônio recebe sinais inibitórios e excitatórios em pontos espacialmentes diferentes ao longo do tempo, ele é portanto um somador espacial e temporal (BARONE, 2003).

3.1.2 Neurônio Artificial

O neurofisiologista Warren McCulloch e o lógico Walter Pitts desenvolveram em 1943 o primeiro modelo matemático de um neurônio biológico. De natureza binária, o modelo de McCulloch e Pitts funcionava da seguinte forma: Se a soma ponderada dos sinais de entrada (valores binários) ultrapassassem um determinado valor de disparo, então a saída se tornava um; se não ultrapassasse, o valor da saída era zero (BITTENCOURT, 1998).

A figura 3.2 representa o modelo geral de neurônio atualmente utilizado. Este é uma generalização do modelo McCulloch e Pitts. O modelo é composto por:

Camada de entrada composta pelo vetor de valores de ativação de entrada ($\mathbf{a_i}$) para a unidade/neurônio *i*, onde a_j é valor da entrada *j*. $W_{j,i}$ é o peso da entrada *j* da unidade *i* e, o vetor composto pelos pesos de entrada do neurônio/unidade *i* é representado pela notação $\mathbf{W_i}$.

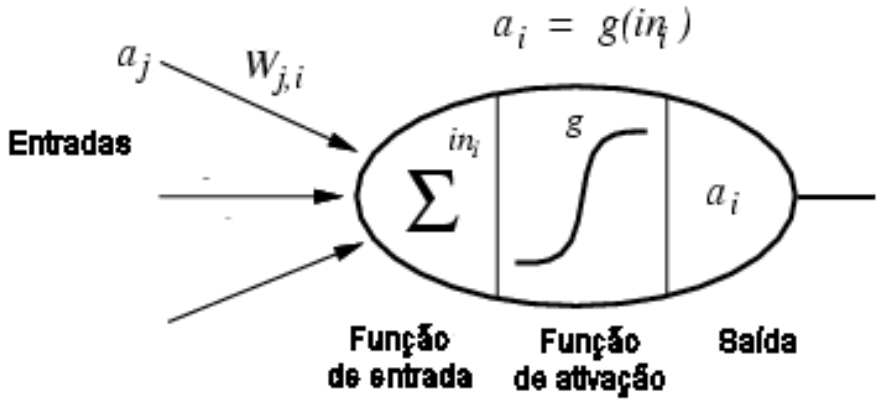

Figura 3.2: Neurônio Artificial. Adaptado de (RUSSELL; NORVIG, 1995).

Função de Entrada $\sum^{in_i}$ é a soma ponderada das entradas da unidade/neurônio *i*.

Função de ativação $g(in_i)$ é a função usada para restringir a amplitude da saída de um neurônio.

Saída a_i é valor de saída para a unidade *i*.

A computação do modelo geral é definida pela equação 3.1. Existem vários modelos de redes neurais. A tabela 3.1 fornece uma visão simplificada dos principais com suas aplicações e vantagens/desvantagens.

$$in_i = \sum W_{j,i} a_j = \mathbf{W}_i . \mathbf{a}_i \tag{3.1}$$

3.2 Mapas auto-organizavéis

Neste ítem é apresentado o modelo de rede neural criado por Teuvo Kohonen (KOHONEN, 1989), chamado de Mapa de características auto-organizáveis (self-organizing Features Map) ou mapa auto-organizável (SOM - self-organizing map) também conhecido como Mapa de Kohonen. Além do SOM básico proposto por Kohonen, também são descritas algumas extensões hierárquicas do mesmo, pois este tipo de estruturas foi desenvolvido nesta dissertação.

3.2.1 Mapas Auto-Organizáveis

Os mapas auto-organizáveis constituem uma classe especial de redes neurais cujos neurônios são dispostos geralmente em uma grade unidimensional ou bidimensional (figura 3.3).Deve-se ressaltar, que também é possível encontrar grades com dimensionalidade maior(HAYKIN, 2001). Nesta grade, os neurônios de saída competem entre si para serem ativados, sendo selecionado apenas um neurônio de saída ou um neurônio por grupo, ao qual será aplicado o algoritmo de ajuste.

O Mapa auto-organizável, ou SOM(Self Organizing Map), pertence a classe de redes neurais artificiais com aprendizagem não supervisionada. Este modelo é baseado no

Tabela 3.1: Modelos de redes neuronais

Perceptrons	
Aplicações	Reconhecimento de caracteres
Vantagem	Rede neuronal mais antiga
Desvantagem	Não reconhece padrões complexos, sensível a mudanças
Backpropagation	
Aplicações	Larga aplicação
Vantagem	Rede mais utilizada, simples e eficiente
Desvantagem	Treinamento supervisionado, exige muitos exemplos
Counterpropagation	
Aplicações	Reconhecimento de padrões. Análise estatística
Vantagem	Rapidez do treinamento
Desvantagem	Topologia complexa
Hopfield	
Aplicações	Recuperação de dados e fragmentos de imagens
Vantagem	Implementação em larga escala
Desvantagem	Sem aprendizado, pesos preestabelecidos
Bidirecional Associative Memories (BAM)	
Aplicações	Reconhecimento de padrões
Vantagem	Estável
Desvantagem	Pouco eficiente
Kohonen	
Aplicações	Reconhecimento de padrões não especificados
Vantagem	Auto-organização
Desvantagem	Pouco eficiente

fato, de que muitas redes neurais biológicas são camadas bidimensionais de unidades de processamento que podem ser células ou módulos celulares (MAINZER, 1997).

O principal objetivo do mapa auto-organizável é transformar um padrão de entrada com uma dimensionalidade qualquer em um mapa discreto unidimensional ou bidimensional, realizando esta transformação de uma forma topologicamente ordenada(HAYKIN, 2001).

Neste tipo de rede neural, após ser apresentado um padrão de entrada, o neurônio com o mais alto padrão de ativação e sua vizinhança são escolhidos para aprendizado. Os pesos são modificados de acordo com uma vizinhança de determinado raio, centrada no neurônio com o maior padrão de ativação para a entrada apresentada. Formalmente, este mapa considera uma projeção não linear de *P* do espaço *V* de sinais de entrada *v* para um mapa bidimensional *A*, conforme pode-se ver na figura 3.3. Observa-se nesta mesma figura,que a aprendizagem do mapa se dá da seguinte forma: o valor de entrada v seleciona um centro s. Na vizinhança deste centro, todos os neurônios têm seus pesos (w_i) modificados em direção de v. O grau desta mudança diminui em sentido inverso da distância ao centro s (MAINZER, 1997).

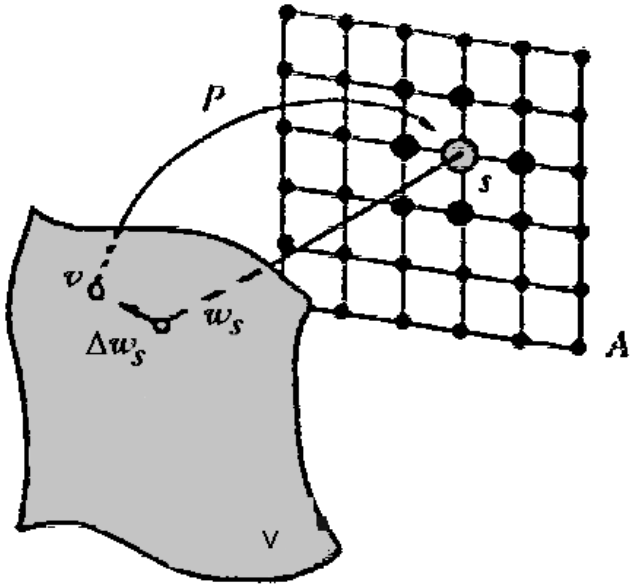

Figura 3.3: Modelo do Mapa Auto-Organizável de Kohonen (MAINZER, 1997)

3.2.2 Propriedades do Mapa de Características

As principais propriedades apresentadas pelos mapas de Kohonen são:

1. Aproximação do espaço de entrada - O mapa de características gerado pelo algoritmo SOM e representado pelo conjunto de vetores de pesos sinápticos na saída da rede, fornece uma boa aproximação do espaço de entrada.

2. Ordenação topológica - Neurônios vizinhos correspondem a padrões de entradas similares.

3. Casamento de Densidade - A densidade das unidades de saída correspondem qualitativamente a função de densidade de entrada. A rede SOM tende a representar excessivamente regiões de baixa densidade de entrada, e insuficientemente regiões de alta densidade de entrada.

4. Seleção de características. A partir de dados do espaço de entrada com uma distribuição não linear, o mapa auto-organizável é capaz de selecionar um conjunto das melhores características para aproximar a distribuição subjacente. Em outras palavras, um mapa auto-organizável pode ser visto como uma generalização não linear da Análise dos Componentes Principais.

3.2.3 Algoritmo SOM

Resumidamente, o algoritmo SOM pode ser descrito pelos passos a seguir (HAYKIN, 2001):

Início Definir o tipo de vizinhança e o raio. Começar com valores apropriados para os pesos sinápticos $w_j(0)$. Os $w_j(0)$ devem ser diferentes para j=1,2,..,*l*, onde *l* é o número de neurônios na grade. Na ausência de qualquer informação a priori, os pesos $w_j(0)$ devem ser escolhidos de forma aleatória. Pode ser desejável manter a magnitude dos pesos pequena.

Amostragem Retire uma amostra **x** do espaço de entrada com uma certa probabilidade, o vetor **x** representa o padrão de ativação que é aplicado a grade. A dimensão do vetor **x** é igual a *m*.

Casamento por similaridade Encontre o neurônio com o melhor casamento (vencedor) *i*(**x**) no passo de tempo *n* usando o critério da mínima distância euclidiana (equação 3.2):

$$i(x) = \arg\min_j \parallel \mathbf{x}(n) - \mathbf{w}_j \parallel, j = 1, 2, ..., l \tag{3.2}$$

Atualização : Ajuste os vetores de peso sináptico de todos os neurônios usando a fórmula de atualização (equação 3.3).

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n) h_{j,i(x)}(n)(\mathbf{x}(n) - \mathbf{w}_j(n)) \tag{3.3}$$

Onde $\eta(n)$ é o parâmetro da taxa de aprendizagem e $h_{j,i(x)}(n)$ é a função de vizinhaça centrada em torno do neurônio vencedor $i(\mathbf{x})$; ambos ($\eta(n)$ e $h_{j,i(x)}(n)$)são variados dinamicamente durante a aprendizagem para obter melhores resultados.

Continuação Continue com o passo "Amostragem"até que não sejam observadas modificações significativas no mapa de características.

3.2.4 TS-SOM Tree Self Organizing Map

Um dos modelos observados para implementação deste trabalho foi a TS-SOM (KOSKELA et al., 2000). A TS-SOM é utilizada para representar uma banco de imagens. Conforme a figura 3.4, nota-se que a mesma é composta de várias SOM´s bidimensionais, organizadas de forma hierárquica em vários níveis. A TS-SOM reduz o tempo de busca do neurônio com os pesos mais próximos do vetor de entrada apresentado, também conhecido como *best-matching unit* (BMU), de $O(N)$ para $O(\log N)$. Como a busca do BMU domina o tempo de computação do SOM, esta característica é vital na na classificação de grandes vetores de características. A busca do neurônio vencedor na camada abaixo, é limitada a uma pré-definida porção da rede SOM que está ligada com o neurônio vencedor da camada acima.

O algoritmo TS-SOM é baseado fracamente nos algoritmos tradicionais de busca em árvore. Por ser uma estrutura em árvore, o número de mapas aumenta à medida que se desloca um nível abaixo. O espaço de busca pelo BMU (Best Match Unit), ou neurônio vencedor no nível abaixo é restrito a uma porção de tamanho fixo abaixo do neurônio pai. A busca não precisa necessariamente ser limitada apenas aos filhos diretos, podem-se definir quais nodos da vizinhança também devem ser pesquisados.

Os vetores de treinamento são apresentados para o primeiro nível (topo), e o mapa é treinado usando o algoritmo SOM padrão. Após o mapa ser organizado, os vetores de peso são congelados, passando-se para o próximo nível, onde é realizado o mesmo processo.

3.2.5 H-SOM - Hierarquical Self-Organizing Maps

A idéia chave desse tipo de rede é usar um conjunto de múltiplas camadas, onde cada uma dessas camadas são SOM's independentes, que armazenam feições de acordo com a similaridade entre elas.

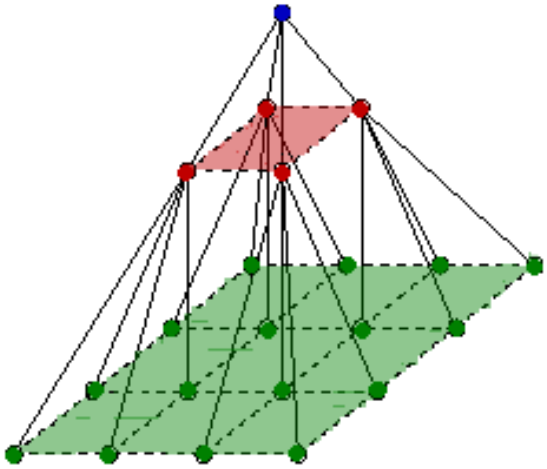

Figura 3.4: Estrutura de uma TS-SOM com três níveis bidimensionais. (KOSKELA et al., 2000)

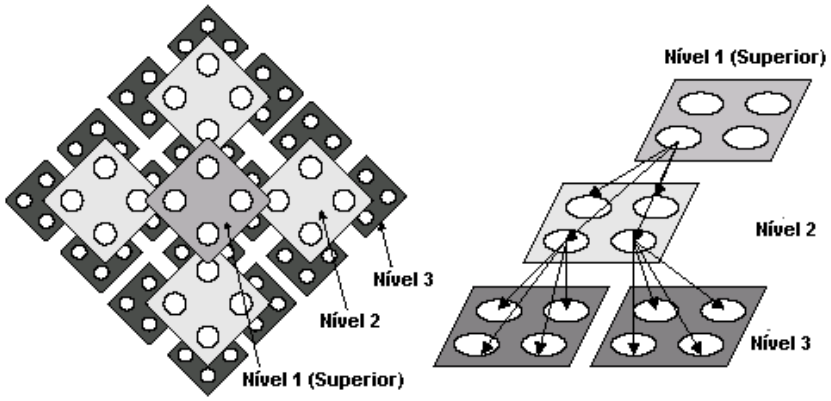

Figura 3.5: Vista superior e secção de uma H-SOM (CHAN; SPRACKLEN, 2000)

A primeira camada determina o tamanho de toda as restantes, pois para toda unidade no primeiro mapa, são adicionados um novo mapa de Kohonen (figura 3.6). Sendo esse processo repetido sucessivamente até a profundidade desejada (KOIKKALAINEN; OJA, 2000).

A H-SOM é uma extensão da SOM. A H-SOM é construída como uma árvore estruturada, onde cada nodo é uma rede SOM treinada com um determinado conjunto de dados.

O mapa do nível superior usa o conjunto completo de dados de treinamento de acordo com a quantização de cada neurônio. Os mapas filhos são treinados com subgrupos do conjunto de dados vencedor do nível superior; ou seja, ele é treinado com os dados do neurônio pai, que por sua vez é o vencedor do primeiro nível ou do mapa raiz.

São necessários definir os seguintes parâmetros na construção de uma H-SOM:

1. Número máximo de níveis,
2. número de neurônios em cada nível ou regra de produção da árvore,
3. tamanho da vizinhança,

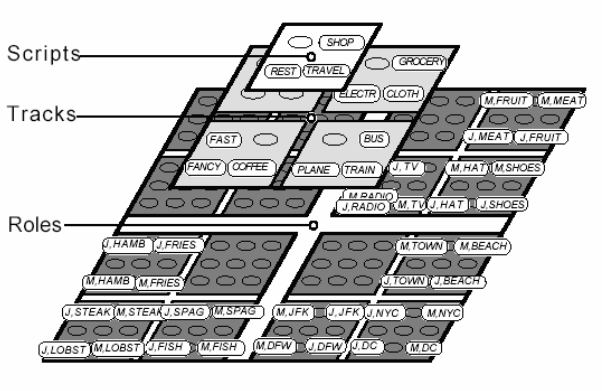

Figura 3.6: Exemplo de uma H-SOM, com cada unidade gerando novas SOM's independentes na camada abaixo(MIIKKULAINEN, 1990)

4. métodos de iniciação dos pesos em cada sub-mapa.

O treinamento da H-SOM é feito de forma seqüencial, do alto da árvore até os nodos mais a baixo, da seguinte forma:

1. Comecar com nível atual $= 1$ (superior);
2. Escolher de forma aleatória todos os conjuntos de treinamento ($\mathbf{x_n}$). O número de vetores de treinamento é dado pela variável $\mathbf{n}$.
3. Apresentar o vetor de treinamento $\mathbf{x_i}$, sendo $\mathbf{x_i} \in R^n$;
4. Usar como função de ativação a distância Euclidiana dos pesos e vetores de entrada.
5. Selecionar o neurônio vencedor(i) do nível atual(j)
6. Atualizar os pesos sinápticos do neurônio vencedor e de seus vizinhos de acordo com as regras básicas da SOM.
7. Repita os passos 3 à 6 até completar a fase de organização do mapa no nível atual.
8. Particionar o conjunto de dados de entrada, que serão usados para treinar o mapa do próximo nível. Essa partição irá gerar o subgrupo de vetores de treinamento S_i para as entradas do i_{esimo} neurônio vencedor;
9. Os filhos do neurônio vencedor obtido no passo anterior são treinados de acordo com as regras da SOM.
10. Repita os passos do 5 ao 8 até não existir mais camadas (níveis) a treinar.

3.2.6 GHSOM - Grow Hierarquical Self-Organizing Maps

Um dos maiores problemas da H-SOM é o fato de sua arquitetura ser definida *a priori* .Isso pode levar a uma acomodação dos mapas não condizente com um verdadeiro agrupamento por afinidade. Desta forma, foi desenvolvida a GHSOM's (figura 3.9). Esse arquitetura permite tanto o crescimento hierárquico, como o crescimento de tamanho da

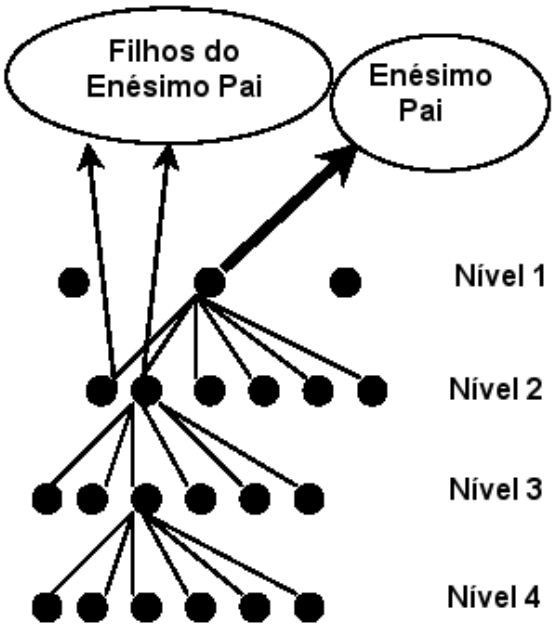

Figura 3.7: Exemplo de uma H-SOM, com cada unidade gerando novas SOM's independentes na camada abaixo(MIIKKULAINEN, 1990)

própria SOM de acordo com a distribuição dos dados. Caso ocorra um crescimento hierárquico, isso significa que o dado pode ser decomposto em sub-partes. Caso o crescimento seja horizontal, significa que o mapa esta se adaptando às necessidades do espaço de entradas (DITTENBACH; MERKL; RAUBER, 2000a).

3.2.7 Treinamento e funcionamento da GH-SOM

A GH-SOM é composta de vários SOM's independentes. Cada SOM possui capacidade de crescimento dinâmico. Em cada mapa os neurônios são agrupados por similaridade de resposta. Caso o espaço de entrada exija mais neurônios, o tamanho é dinamicamente incrementado, adaptando-se as necessidades do espaço (crescimento horizontal). Se dados de entrada similares (que excitam o mesmo neurônio) permitem a decomposição hierárquica, ou seja, é possível subdividi-los em grupos mais específicos, a GH-SOM permite o que é chamado crescimento vertical. Mantém-se o neurônio pai responsável por mapear as características em comum no mapa atual, e cria-se mapas "filhos"mais específicos no nível abaixo. O crescimento vertical e horizontal permitem a GH-SOM adaptar-se ao tamanho do espaço de dados da entrada, suplantando o principal problema da H-SOM, que possui tamanho pré-definido.

O ponto de partida para o crescimento da rede é o calculo da divergência global dos dados de entrada para uma única unidade (neurônio) da SOM em uma hipotética camada 0. A esse neurônio é associado um vetor de peso m_0, que é calculado como a média de todos valores de entrada (equação 3.4).

$$m_0 = [\mu_{0_1}, \mu_{0_2}, ..., \mu_{0_n}]^T \tag{3.4}$$

A divergência dos dados de entrada é fornecida através do cálculo do erro de quantização médio de um neurônio chamada **mqe** (*mean quantization error*). O $\mathbf{mqe_i}$ é computado como a distância média entre vetor de pesos m_i no neurônio *i* e o padrões de

entrada mapeados, conforme a equação 3.5. Nesta equação d é o número de dados de entrada.

$$\mathbf{mqe_i} = \frac{1}{d}.\|m_i - x\|. \tag{3.5}$$

Após o calculo do **mqe**, o treinamento da GH-SOM começa na primeira camada, que em geral possui tamanho $2x2$.

A cada neurônio i é associado um vetor de pesos n-dimensional m_i (equação 3.6), inicializados com valores aleatórios.

$$m_i = [\mu_{i_1}, \mu_{i_2}, ..., \mu_{i_n}]^T, m_i \in \Re \tag{3.6}$$

;

O processo de aprendizagem ocorre como uma competição entre os neurônios. O neurônio com o vetor de pesos mais próximo do padrão de entrada é declarado vencedor. Ele e sua vizinhança são ajustados de acordo com uma taxa de aprendizado α. Tanto α quando o tamanho da vizinhança são decrescentes ao longo do tempo. O tamanho da vizinhança é determinado por uma função h_{ci} baseada na distância ao neurônio vencedor c em um número de interações t. Desta forma a equação 3.7 representa a regra de aprendizagem, dado o padrão de entrada x.

$$m_i(t+1) = m_i(t) + \alpha(t).h_{ci}(t).[x(t) - m_i(t)] \tag{3.7}$$

;

O tamanho de cada mapa m é calculado baseado no erro de quantização média do mapa ($\mathbf{MQE_m}$), após um número fixo de interações (λ). O $\mathbf{MQE_m}$ é calculado de acordo com a equação 3.8, sendo *u* o número de unidades/neurônio existentes na SOM.

$$MQE_m = \frac{1}{u}\sum_i mqe_i. \tag{3.8}$$

Cada camada da GH-SOM é responsável por expandir uma porção dos dados entradas divergentes da camada anterior. Isto é feito acrescentando novos neurônios até se alcançar um tamanho adequado. O crescimento é definido pela relação entre o erro de quantização médio e uma porcentagem fixa τ_m. Quanto menor esta porcentagem, maior será o tamanho do mapa emergente. Desta forma se $MQE_m \geq \tau_m.mqe_0$ uma nova linha ou coluna são adicionadas no mapa *m*, após λ interações de treinamento na vizinhança do neurônio com maior mqe_i, chamado de unidade de erro e. A nova linha ou coluna é inserida entre a unidade de erro e seu vizinho com menor similaridade $\mathbf{d}$, conforme a figura 3.8.

Após finalizado o crescimento do mapa ($MQE_m < \tau_m.mqe_0$) são identificados os neurônios que serão expandidos para uma nova camada, ou seja, os que possuem um *mqe* alto. O critério de seleção é baseado no mqe_0. Um parâmetro τ_u define o grau de granularidade desejada. Desta forma o crescimento hierárquico é dado pela equação 3.9.

$$mqe_i > \tau_u mqe_0 \tag{3.9}$$

O processo de crescimento do mapa explicado anteriormente é aplicado a essa nova SOM criada. A principal diferença para o processo de treinamento anterior é que agora só uma fração dos dados de entrada é selecionada. O processo de treinamento da GHSOM termina quando mais nenhuma unidade requerer expansão adicional.

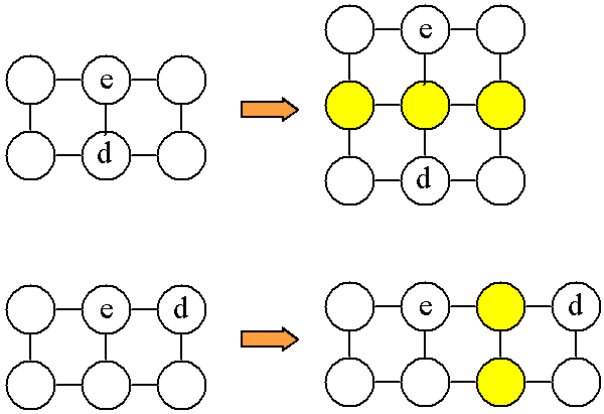

Figura 3.8: Crescimento horizontal da GH-SOM

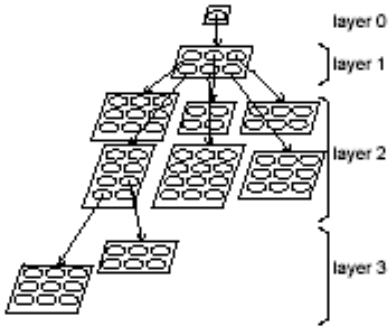

Figura 3.9: GH-SOM Treinada - A GH-SOM evolui para um estrutura que reflete a estrutura hierárquica dos dados de entrada(RAUBER; MERKL; DITTENBACH, 2002)

4 SISTEMAS DE RECUPERAÇÃO DE IMAGENS

Este capítulo aborda os sistemas de recuperação de imagens, seus paradigmas e destaca algumas características principais.

4.1 Introdução

No mundo pós-moderno, a explosão de tecnologias de manipulação de imagens tem cada vez mais facilitado o trabalho em várias profissões. Nesse contexto, inúmeras aplicações surgem aos olhos de agentes de publicidade e pessoas dos meios de comunicação em geral, que necessitam acessar no menor tempo possível imagens que possam ser anexadas a reportagens e matérias jornalísticas. Hoje, um jornalista tem à sua disposição milhares de fotos sobre um evento, mesmo os não cobertos por fotógrafos de seu veículo. Médicos podem ter acesso a arquivos enormes de raios-x e outros exames que envolvem inspeção visual, designers possuem uma quantidade quase que ilimitada de fontes visuais na web e em outras mídias. Mas, se por um lado a profusão de imagens facilita o trabalho criativo e o acesso a informações vitais, com é o caso da medicina, por outro induz a dificuldades crescentes para a implementação de sistemas automáticos de recuperação e classificação dessas imagens, abrindo campo de trabalho e investigação para profissionais de Computação que trabalhem com Visão Computacional, Processamento de Imagens, Banco de Dados e outras áreas. Mas apesar destes esforços ainda não foi descoberto uma forma correta de localizar dados visuais. Para tentar resolver esse problema, a área de recuperação de imagens tem despontado com várias alternativas.

Há duas técnicas principais de recuperação de imagens que são baseadas em acesso dirigido pelo conteúdo (CBIR- *content-based image retrieval*): recuperação por palavras-chave ou Índices, também conhecida como KBR (*Keyword-based retrieval*)e recuperação por Similaridade, também conhecida como SBR(*similarity - based retrieval*)(SETHI; COMAN, 1999).

Ao se construir um sistema de recuperação de informações e, portanto de imagens, devem ser considerados os seguintes fatores:

- Entendimento do conteúdo dos objetos dentro da base de dados;
- Extração da informação de interesse;
- Qualidade desejada da busca retornada.

As arquiteturas de CBIR utilizam os mais variados métodos de busca. Algumas usam a similaridade entre as cenas como critério. Neste caso, o usuário fornece uma imagem, ou uma descrição visual, que condiz com sua busca. As fotos selecionadas são escolhidas

através de algoritmos que avaliam quantitativamente a distância de similaridade. Nos sistemas de recuperação KBR as imagens são localizadas através de palavras-chave que descrevem a cena desejada. Convém lembrar que a maioria destes sistemas opera em bancos de imagens com anotações manuais.

4.2 Sistema de Recuperação de Imagens por Índices (KBR)

O KBR é um sistema que localiza uma imagem baseada em palavras-chave. O maior problema existente neste tipo de arquitetura é associar significado semântico a cada imagem. Ao se extrair as características de uma imagem o que se obtém são informações de textura, cor, forma, etc., ou seja, a estrutura básica da mesma. Destas características simples deve-se construir uma semântica complexa que descreva a figura. É extremamente difícil fazer este tipo de associação entre características básicas e significados abstratos. Tal dificuldade é chamada de "gap semântico".

Apesar disso, é extremamente interessante que os sistemas de recuperação de imagens possuam essa capacidade, pois a mesma pode aumentar muito a utilização de banco de imagens (LI; WANG, 2003), criando capacidade de indexação automática de acordo com critérios do usuário.

Um exemplo deste tipo de aplicativo é o ALIP (*Automatic Linguistic Indexing of Pictures*) da Universidade de Stanford. Este sistema constrói índices (palavras-chave), para cada uma das imagens.

O Alip cria modelos de cada tipo de imagens e os agrupa através de métodos estatísticos (figura 4.1). Primeiramente, são extraídas as informações de textura e cor. Estas informações são usadas para categorizar a cena de acordo com um dicionário de termos com as classes semânticas existentes.

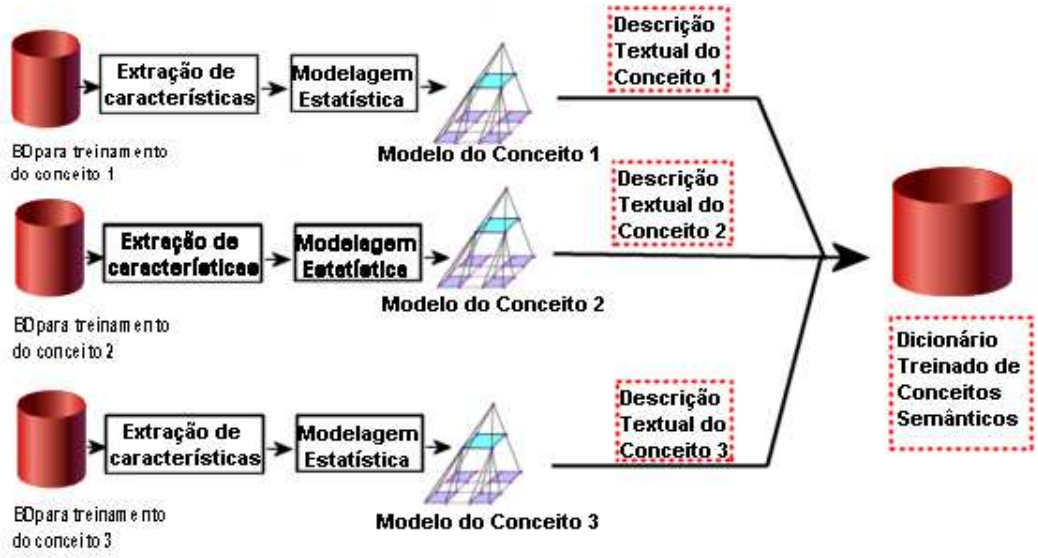

Figura 4.1: Processo de modelagem estatística do Alip (LI; WANG, 2003).

O Alip gera categorias de imagens, correspondendo a conceitos. Cada categoria possui um perfil modelado estatisticamente através do Modelo de Markov Oculto bidimensional e multiresolução, conhecido como 2D MHMM (de *Two-dimensional Multiresolution Hidden Markov Model*). A modelagem é feita através de uma coleção de vetores de

características extraídos em múltiplas resoluções e espacialmente arranjadas sob a forma de pirâmide. A 2D MHMM é estimada separadamente para cada categoria, permitindo a inserção de novas categorias facilmente. Esse sistema é composto de três principais componentes: o processo de extração de características, a modelagem por 2D MHMM e o processo de indexação lingüística.

Para a extração de características, a imagem é dividida em pequenos blocos de tamanho $4x4$. De cada bloco é extraída a cor média por canal (RGB), gerando três valores, e as características de textura representadas pela energia nas bandas de alta-freqüência (HL,LH e HH) da transformada Wavelet usada (Haar ou Daubechies-4), gerado um vetor com dimensionalidade igual a seis.

Após aplicar o primeiro nível da transformada Wavelet, o bloco $4x4$ é decomposto em quatro bandas de freqüência (LL, HL,LH e HH), onde cada banda possui uma matriz $2x2$ de coeficientes. Supondo os coeficientes da banda HL como $c_{k,l}, c_{k,l+1}, c_{k+1,l}, c_{k+1,l+1}$, a característica deste bloco é computada de acordo com a equação 4.1. O mesmo é feito para as bandas LH e HL. A banda HL reflete atividade na direção horizontal; a LH reflete mudanças na direção vertical e a HH na diagonal. A banda LL não é utilizada.

$$f = \frac{1}{2}\sqrt{\sum_{i=0}^{1}\sum_{j=0}^{1} c_{k+i,l+j}^2} \tag{4.1}$$

A modelagem estatística é feita conforme a figura 4.1. Primeiramente é criada manualmente uma série de conceitos que serão aprendidos. As características de cada conceito são extraídas das imagens em vários níveis de resolução. Após, é gerado um modelo estatístico transversal entre escalas, caracterizado por uma coleção de características em várias resoluções. É associado a cada modelo um grupamento de palavras que definem a imagem.

A indexação lingüística é feita baseada no grupamento de palavras oriundo da fase anterior. Para quantificar a similaridade estatística entre uma imagem e um conceito são computadas as características da imagem sobre os modelos existentes. De acordo com a distância de cada modelo é selecionado um conjunto de palavras que serão associadas à imagem e farão parte de seu índice. O resultado do Alip é sumarizado pela tabela 4.1.

4.3 Sistema de Recuperação de Imagens por Similaridade (SBR)

Esse método pode ser dividido basicamente em três técnicas (LEW, 2000):

- Busca por similaridade;
- Esboço;
- Ícones.

4.3.1 Busca por similaridade

Segundo (SETHI; COMAN, 1999) a abordagem SBR segue o ditado popular que a melhor representação de uma imagem é ela própria. Dessa forma, ao invés de associar palavras-chaves que descrevam cada imagem, extrai-se um vetor de características de cada imagem no momento em que ela é catalogada. Procede-se a uma busca na base de imagens, onde o usuário escolhe a partir de um conjunto de imagens iniciais qual a que se assemelha mais à imagem desejada. Este tipo de CBIR busca as imagens que são mais

Figura 4.2: Exemplo de classificação gerada pelo Alip.

semelhantes através de refinamentos sucessivos. O PicSOM (KOSKELA, 1999) é um exemplo deste tipo de aplicação (figura 4.3) .

Tabela 4.1: Resultado dos experimentos de categorização automática de imagens do Alip. Cada linha lista, em porcentagem, a quantidade de vezes que uma imagem de determinada categoria foi classificada como pertencente a cada uma da dez categorias existentes (LI; WANG, 2003)

%	África	Praia	Construções	Ônibus	Dinossauros	Elefantes	Flores	Cavalos	Montanhas	Comida
África	52	2	4	0	8	16	10	0	6	2
Praia	0	32	6	0	0	0	2	2	58	0
Construções	8	4	64	0	8	6	0	0	6	4
Ônibus	0	18	6	46	2	8	0	0	16	4
Dinossauros	0	0	0	0	100	0	0	0	0	0
Elefantes	8	0	2	0	8	40	0	8	34	0
Flores	0	0	2	0	0	0	90	0	2	6
Cavalos	0	2	0	0	0	4	24	60	4	6
Montanhas	0	6	6	0	2	2	0	0	84	0
Comida	6	4	0	2	6	0	8	0	6	68

O sistema de recuperação de imagens PicSOM é baseado na busca por imagem exemplo, conhecido também como QBPE (Querying by pictorial Example). O PicSOM usa uma TS-SOM (descrita no capítulo de Redes Neurais).

Na PicSOM a busca começa com uma seleção de imagens representativas uniformemente retiradas do nível superior da TS-SOM. A cada rodada, a seleção se torna mais apurada, de acordo com as seleções do usuário. O sistema marca as imagens selecionadas

pelo usuário com um valor positivo e as não selecionadas com um valor negativo. Estes valores são usados para fornecer um novo conjunto de imagens. Se as imagens selecionadas possuem um conjunto de características próximas a um mapa da TS-SOM isto significa que tanto as características presentes na busca quanto o seu peso relativo devem ser incrementados. Isto é feito marcando nos mapas os locais que correspondem às imagens selecionadas e às rejeitadas com valores positivos e negativos respectivamente. As respostas são normalizadas para que sua soma seja igual a zero. Cada nível SOM é tratado com uma matriz bidimensional formada pelas respostas dadas pelo usuário ao conteúdo apresentado. Essas matrizes passam por um filtro passa-baixa com máscara de convolução simétrica. Esta filtragem é feita para estender as respostas (valores) a toda vizinhança, que presumidamente deve conter imagens semelhantes as selecionadas do mapa ou não, e portanto devem influenciar de forma positiva ou negativa na seleção do próximo conjunto de imagens a ser apresentado. Este processo é feito através de todos os níveis da TS-SOM, o que acarreta mudança dos valores vizinhos conforme pode ser visto na figura 4.4, onde, os valores negativos são representados como regiões escuras e valores positivos como regiões claras. Nesta pode-se observar três níveis da TS-SOM. À esquerda está o mapa gerado pela seleção do usuário e à direita está o mapa após a convolução.

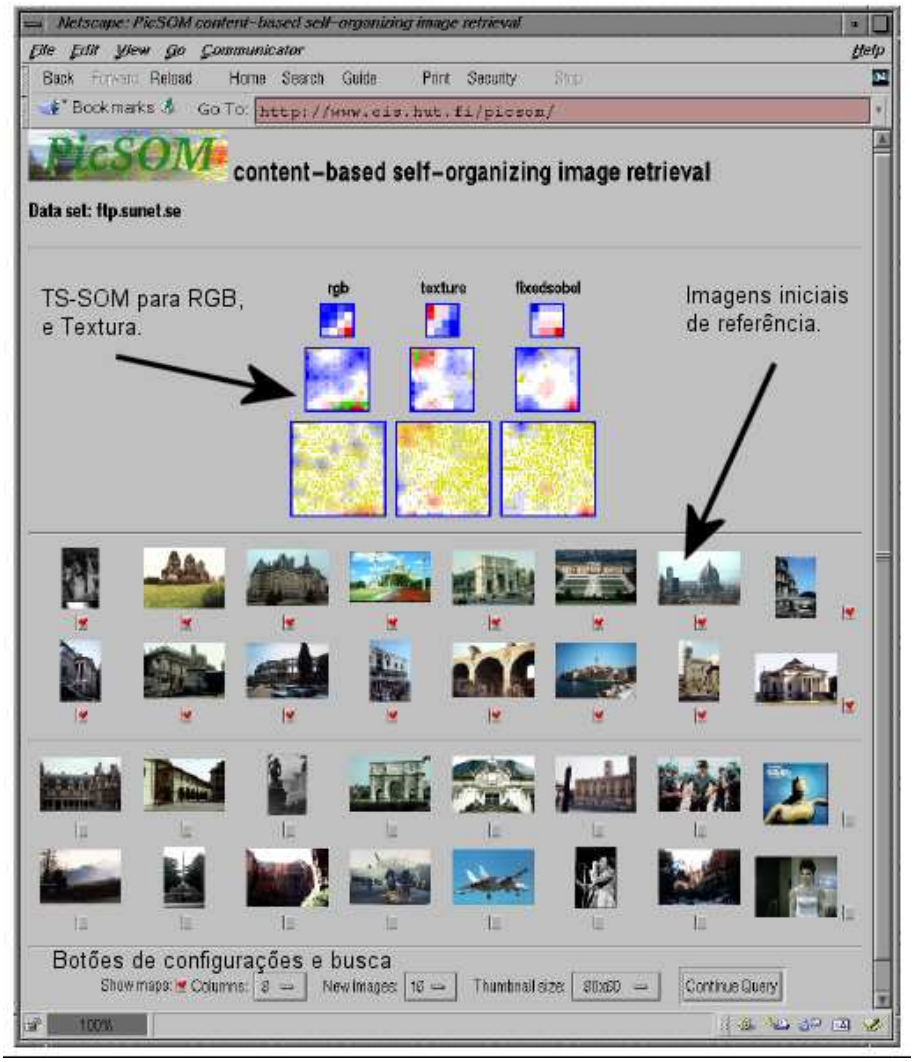

Figura 4.3: Exemplo da tela inicial do PicSOM.

Resumindo, no funcionamento da PicSOM, em uma primeira fase , apresentam-se as imagens do banco de imagens para a rede, que vai realizando classificações por características similares em um mesmo nível de hierarquia de atributos. A partir dessa classificação inicial constrói-se outros níveis filhos de redes SOM em número menor. Desta forma, o nível superior da hierarquia possui atributos mais gerais. Para localizar uma imagem que contivesse um leão, por exemplo, o sistema apresentaria várias imagens, e caso, nesse nível não fosse localizada a imagem desejada, o usuário escolheria outras imagens associadas para permitir que a rede mostre um mapa mais específico. Portanto, pode-se selecionar imagens associadas ao habitat desses animais,como savanas por exemplo. Assim sendo,uma vez apresentada uma imagem de savana, seriam ativados neurônios da rede SOM associados com os atributos desta imagem, desativando neurônios associados a imagens urbanas contendo prédios. Convém observar que a própria rede SOM se torna o índice do banco de imagens.

4.3.2 Busca por esboço

Neste tipo de CBIR, o usuário faz um esboço do que deseja buscar. Ou seja, o usuário cria uma busca através de um rascunho da imagem que ele gostaria de recuperar. A partir do rascunho do usuário o sistema retorna imagens com formas que correspondem aproximadamente aos ângulos e contornos desenhados (figura 4.5). Este tipo de busca permite que o usuário especifique diretamente que parte da imagem tem maior relevância para sua busca (LEW, 2000).

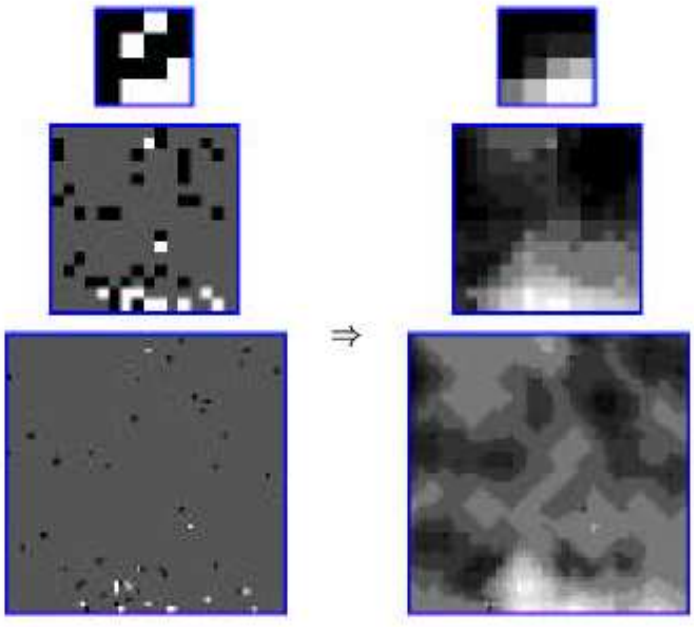

Figura 4.4: Exemplo da convolução das unidades positivas e negativas de um mapa no terceiro nível (KOSKELA, 1999)

4.3.3 Ícones

Neste paradigma, o usuário coloca ícones em uma área específica, da mesmo forma como eles devem aparecer na imagem objetivo. Isso permite ao usuário explicitar a disposição dos objetos da imagem desejada. Essas informações são casadas com posições

nas imagens disponíveis no banco de imagens considerando o conceito agregado a cada ícone. O ImagemScape (LEW, 2000)é um exemplo de sistema de recuperação baseado em ícones (figura **??**)

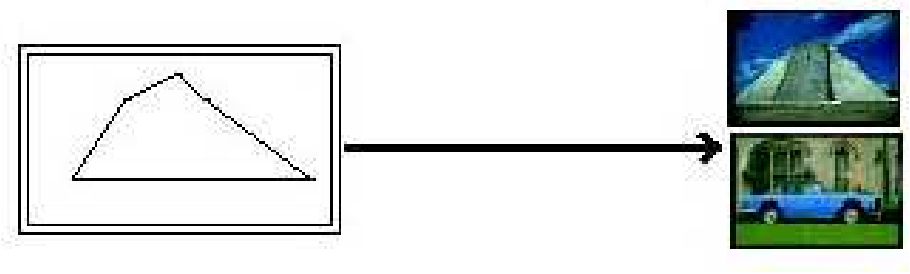

Figura 4.5: Exemplo de uma busca baseada no esboço. Adaptado de (LEW, 2000)

5 MODELO PROPOSTO

5.1 Introdução

Este trabalho descreve uma arquitetura para a criação de índices (indexação) e posterior recuperação de imagens, ambas de forma automática e baseadas na descrição da cena (conteúdo). Este método de recuperação, conforme dito anteriormente é conhecido como KBR (*keyword-based retrieval*), onde é gerado um conjunto de palavras-chaves, para cada imagem, que descrevem diretamente os objetos presentes que compõem a cena observada. A principal contribuição deste trabalho é demonstrar que os Mapas Auto-Organizáveis de Kohonen Hierárquicos podem ser versáteis e bem precisos neste tipo de aplicação, uma vez que podem se adaptar facilmente a novos contextos, e portanto novas classificações. Uma vez que, dependendo do contexto, um objeto pode ser descrito por diferentes termos, é vital que esse tipo de sistema possa ser adaptado para um novo contexto de forma dinâmica. Essa dinâmica é conseguida neste trabalho através de um novo treinamento da rede com os novos termos.

Neste trabalho, cada objeto é descrito por suas características de cor e textura. Os Mapas Auto-Organizáveis de Kohonen Hierárquicos foram usados para a redução de dimensionalidade e classificação. Em virtude de se estar trabalhando com imagens em que muitos dos objetos que compõem a cena a ser descrita não possuem exatamente uma forma, como por exemplo rocha e água, esse tipo de característica não é tratada neste sistema. Na abordagem utilizada, os objetos são classificados em categorias, onde cada categoria pode descrever ou um objeto ou uma cena, de acordo com uma estrutura hierárquica. Dessa forma, em um nível menos abstrato, existem categorias como vegetação e areia, sendo que, a combinação desses elementos básicos, leva a categorias mais abstratas como praia, campo ou cadeia de montanhas.

De forma geral, os sistemas de recuperação de imagens, ou buscam agrupar imagens com feições semelhantes para permitir que o usuário interativamente localize sua cena, ou, no caso da indexação semântica, geram modelos das cenas, ordenando hierarquicamente as características que fazem parte destas. Nesses sistemas, é avaliada a cena como um todo, o que é organizado de forma hierárquica é a distribuição de características de baixo nível de cada uma das cenas, como por exemplo, no caso das categorias propostas em (WANG; LI; WIEDERHOLD, 2001), (conforme mostrado na tabela 5.1). Nesta, os autores concentraram-se em criar hierarquias das feições de cada uma das categorias propostas, sem se preocupar com a consistência ou intersecção entre elas. Cada categoria representa uma espécie ou tipo de **cena** e não de objeto. Neste caso, independentemente da cena poder existir dentro de cada categoria os mais variados tipos de objetos. Já no modelo proposto cada super-categoria é composta de outras subcategorias, ou seja, cada categoria mais complexa é considerada um conjunto de categorias menos complexas, onde

cada nível hierárquico da Rede SOM utilizada, representa um nível de categorias e, portanto, um superconjunto de categorias mais simples.

No modelo aqui proposto, cada categoria mais complexa é considerada um conjunto de categorias menos complexas, onde cada nível hierárquico da Rede SOM utilizada, representa um nível de categorias e, portanto, um superconjunto de categorias mais simples.

Para evitar que as categorias apresentadas fossem dependentes de uma área específica, ou de critérios subjuntivos, foi inferido um conjunto de classes e categorias baseado em um sistema internacional de classificação.

As amostras foram selecionadas a partir de um subconjunto das imagens do banco de imagens utilizado. Como a seleção das amostras que caracterizam uma classe de objetos (através de suas informações de cor e/ou textura), fica de acordo com critérios empíricos de cada usuário, essas escolhas acabam sendo muito subjetivas, podendo levar a escolhas inadequadas (com sobreposição dos espaços amostrais). Uma amostra inadequada seria aquela cujas feições de cor e textura a caracterizassem como pertencente a uma outra categoria distinta, e não à categoria a qual deveria representar. Essa situação pode levar a problemas na generalização e aprendizagem das categorias. Um exemplo extremo desse tipo de erro seria escolher uma porção de uma montanha, onde só existissem árvores, e passar essa amostra para o conjunto de treinamento da classe montanha. Isso implica em uma sobreposição dos espaços amostrais(nesse caso entre montanha e vegetação), gerando erros de classificação. De forma a minimizar esse problema, para detectar sobreposições das amostras foi utilizada uma outra estrutura hierárquica de SOM's, que possibilita visualizar a estrutura hierárquica das feições fornecidas. Desta forma, se percebe o quanto uma amostra é realmente representativa de sua classe, permitindo eliminar amostras ditas inadequadas.

Neste capítulo, estão descritas tanto a arquitetura desenvolvida quanto as dificuldades observadas. Desta maneira, as etapas de soluções adotadas estão inclusas no desenvolvimento apresentado.

Tabela 5.1: Exemplos de algumas categorias propostas em (LI; WANG, 2003)

ID	Descrição da categoria
0	África, pessoa, paisagem, animal
10	Inglaterra, paisagem, montanha, lago, Europeu, pessoa, prédio histórico
20	Mônaco, Oceano, prédio histórico, comida, Europeu, pessoa
30	Guarda Real, Inglaterra, Europeu, pessoa

5.2 Caracterização do problema

O problema abordado consiste na criação de descrições textuais simples do conteúdo das imagens fotográficas armazenadas, de forma digital, em um banco de imagens. Essas descrições podem ser utilizadas na construção de índices para cada imagem, que seriam utilizados posteriormente na recuperação de imagens específicas.

A busca e recuperação de imagens em bancos de imagens digitais, ainda é um problema em aberto, pois, nenhuma das técnicas existentes é tão robusta quanto seria desejável. A situação ideal seria aquela em que todas imagens contivessem descrições de seu conteúdo. Salvo em alguns catálogos de imagens dentro de contextos específicos, isso não é comum. O grande problema de descrever uma imagem, é o fato desta descrição

poder ser muito subjetiva. Uma mesma imagem pode ser descrita de várias formas por uma pessoa em instantes diferentes, sendo que certos detalhes podem ser omitidos em uma descrição e ressaltados em outra. Em bancos de imagens genéricas, criar índices é ainda mais difícil, pois, não existem certos aspectos fixos para guiar a busca. Por exemplo, em um catálogo de imagens arquitetônicas (contexto específico), pode-se usar o estilo arquitetônico como um dos índices, já em um banco de imagens mais abrangente, essa informação pode ser totalmente inútil, visto que a busca deve permitir descrições mais genéricas.

Uma forma de contornar o problema é construir um sistema de indexação automática, que descreva o conteúdo da imagem a partir de categorias pré-estabelecidas. O grande problema de se construir um sistema de notação automática, é o abismo existente entre o conjunto de *pixels* que formam uma imagem e o seu significado semântico (SMEULDERS et al., 2000), como, por exemplo, o que em termos de disposição espacial de textura e cor caracterizam uma cena como praia ou campo.

Convém lembrar que o reconhecimento de classes genéricas de objetos é um problema em aberto para a visão computacional (LI et al., 2004). Segundo (LI; WANG, 2003) "Várias décadas de pesquisa no campo de visão computacional e de recuperação de informação mostraram que algoritmos genéricos que possam aprender conceitos a partir de uma imagem e automaticamente traduzir o conteúdo das imagens em termos lingüísticos é extremamente difícil". Apesar dessa dificuldade, um sistema desse tipo tem aplicação nas mais diversas áreas, desde a tomada de decisão em um sistema robótico, até a localização de imagens em um vídeo e ou banco de imagens.

Alguns erros no reconhecimento são inerentes ao domínio deste trabalho, pois, como as imagens bi-dimensionais são registros de um mundo tri-dimensional, podem ocorrer oclusões e variações de iluminação e cor que comprometem o reconhecimento. Desta forma, a precisão do sistema também depende do pré-processamento realizado nas imagens fornecidas.

5.2.1 Descrição do domínio

O banco de imagens digitais escolhido é composto de imagens oriundas de um subconjunto de imagens pertencentes a empresa COREL®, disponíveis para download na internet[1]. Esse subconjunto possui 10.000 imagems, sendo que três exemplos são mostrados na figura 5.1. Além dessas imagens, foram acrescidas também oriundas da internet.

Como é necessário realizar notações manuais para que se possa realizar métricas comparativas com a notação automática, foi utilizado um subconjunto de imagens composto por 910 imagens significativas.

As imagens que compõem o banco de imagens são todas coloridas, visto que a arquitetura apresentada depende da característica cor para realizar a classificação das cenas. Com relação ao formato e tamanho, as imagens utilizadas possuem o formato jpeg/jpg, com tamanhos variados.

Nos estudos iniciais, se considerou utilizar como objetos básicos de uma cena, formas trigonométricas simples, que seriam agrupadas para formar formas mais complexas. Essa abordagem, após os testes iniciais, mostrou-se extremamente ineficaz, dada a complexidade de achar formas trigonométricas fixas nas imagens utilizadas.

Para se definir uma categoria básica de objetos pertencentes a uma imagem, determinou-se que, uma categoria básica seria composta por objetos discerníveis visualmente como

[1] http://www-db.stanford.edu/ wangz/image.vary.jpg.tar

Figura 5.1: Exemplos de imagens que compõem o banco de imagens.

partes de uma cena. Desta maneira ficou determinado que uma categoria mais complexa seria constituída por categorias posicionadas em níveis hierárquicos inferiores. A figura 5.2 demonstra esses conceitos. Em um nível superior está a categoria praia, que é composta pelas categorias do nível 1: areia, céu e água. A própria categoria praia pode ser parte de uma super-categoria paisagem, que engloba cenas de praia, campo e cidade.

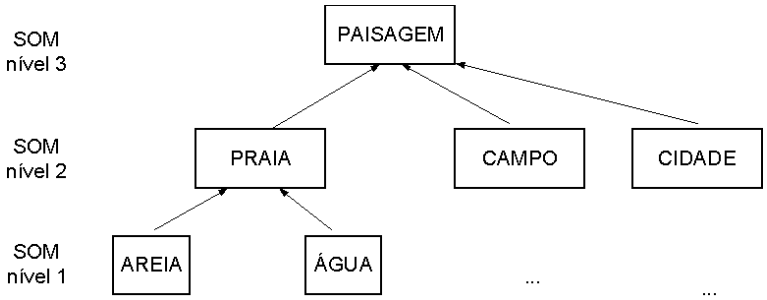

Figura 5.2: Exemplo da organização hierárquica para a categoria praia.

Um dos problemas em se criar definições de categorias de objetos, é a falta de uma coleção de palavras (em inglês *thesaurus*), que permita uma classificação hierárquica de cenas em geral, mesmo porque essas coleções de palavras, quando existem, são condicionadas a um domínio específico e em geral não observam a estrutura da cena através dos objetos que a compõem, mas sim, características gerais. Por exemplo, no caso de imagens da área arquitetônica, os estilos das construções são usados na criação das coleções de palavras, enquanto a distribuição espacial de janelas, portas e outros elementos de um prédio possuem um significado secundário. A forma utilizada para evitar esse problema, foi usar como base a Classificação de Vienna. A Classificação de Vienna ou *International Classification of the Figurative Elements of Marks* [2] é utilizada para classificar elementos figurativos em marcas. Apesar desta classificação não ser utilizada para imagens, mas sim para logomarcas, ela possui um estrutura hierárquica adequada *top-down*, que começa por uma classificação genérica e chega a particularidades de uma imagem, dividindo todos elementos figurativos em categorias, divisões e seções, num total de 29 categorias, 144 divisões e 1887 seções. Com isso, as categorias usadas nessa arquitetura são formalizadas segundo um padrão internacional, evitando uma certa aleatoriedade na definição destas.

Na classificação foi considerado parte do subconjunto número seis de categorias da Classificação de Vienna[3], que é descrito a seguir:

6.1 Montanhas, Rochas e Grutas,

[2]Essa classificação esta disponível na internet no endereço http://www.wipo.int/classifications/fulltext/vienna5/enmain0.htm
[3]A Categoria 6 utilizada neste trabalho e seu pormenores encontra-se no Anexo I.

6.3 Paisagens com água, rio ou córrego,

6.6 Paisagens tipo desérticas ou tropical,

6.7 Paisagens urbanas ou vilarejos.

Essa estrutura serviu como base para a criação das categorias usadas. Para os objetos básicos foram criadas as seguintes categorias:

1 Vegetação,

2 Céu,

3 Areia,

4 Rocha,

5 Prédio (arranha-céus),

6 Casa (edificações baixas),

7 Água,

8 Rua,

9 ? (Valor indeterminado).

O próximo nível de abstração agrega dez categorias, que podem ser compostas por alguns desses objetos básicos:

1 Montanhas,

2 Praia,

3 Deserto,

4 Oásis,

5 Paisagem com água, rio ou córrego,

6 Floresta (ou mata ou qualquer outro tipo de vegetação),

7 Rural(Paisagem Rural),

8 Cidade,

9 Vilarejo,

10 ? (Valor indeterminado).

Todas essas categorias compõem uma subcategoria chamada paisagem. Em ambos os níveis existe a categoria indeterminada, a qual foi criada como índice de cenas em que sua descrição é desconhecida, dúbia ou mesmo de difícil caracterização.

5.3 Arquitetura do Sistema

Basicamente, o sistema é composto por uma unidade de entrada, que recebe a imagem, uma unidade de pré-processamento, que extrai as características de cor e textura e de uma unidade neural composta de redes SOM organizadas em diversos níveis. Essas redes possuem como saídas as categorias que compõem a imagem (figura 5.3). Neste projeto a saída é um arquivo texto com o nome do arquivo de imagem e a categoria na qual a mesma foi classificada. A seguir cada unidade é descrita em seus pormenores.

Figura 5.3: Arquitetura do sistema.

5.3.1 Unidade de pré-processamento

Nesta unidade é feita a extração e a redução de dimensionalidade das características de cor e textura. Um dos principais problemas nesta etapa é identificar cada objeto individualmente. Algumas soluções passam pela segmentação, mas a menos que ela seja realizada de forma automática, não faz sentido trabalhar com essa idéia. De forma a minimizar o problema de localizar objetos em uma imagem (o que por si só já identifica um novo trabalho), cada imagem de entrada é dividida em nove regiões de tamanhos iguais, conforme a figura 5.4, onde é identificada uma categoria básica por região, portanto, o sistema retorna no máximo nove categorias por imagem. A divisão em nove regiões foi uma escolha "*ad hoc*", o ideal seria trabalhar com uma janela móvel com tamanho variável que se ajustasse ao tamanho de cada objeto/cena, mas essa etapa, por ser bastante complexa, também já constitui um novo trabalho. Dessa forma a pesquisa com janelamento móvel se constitui em proposta para trabalhos futuros.

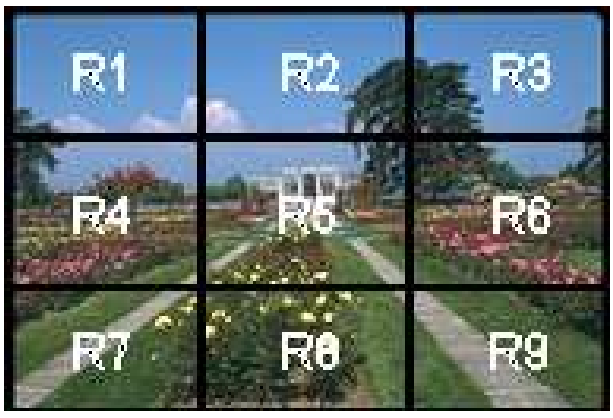

Figura 5.4: Regiões de um imagem.

Dividida a imagem nessas nove regiões, realiza-se a extração das feições de cor e textura para cada uma delas.

Nos primeiros testes, cor e a textura foram tratadas como características isoladas, cada uma com seu próprio mapa, sendo agrupadas em uma nova SOM em um nível hierárquico superior. Esta abordagem se mostrou ineficiente dentro da arquitetura proposta, pois, tanto a cor como a textura podem ser dependentes das condições de captura da imagem

(câmera, hora do dia, posição do sol, condições climáticas). A feição de textura foi a que produziu maior variação e erros de classificação. Para tornar mais robusta a classificação, utilizou-se um vetor em que as informações de textura e cor foram agrupadas em um único vetor. Esta abordagem foi baseada nas técnicas utilizadas no sistema SIMPLIcity (WANG; LI; WIEDERHOLD, 2001)(CHEN; WANG, 2004), e em outros sistemas como o apresentado em (TSAI; MCGARRY; TAIT, 2003).

Para caracterizar cada objeto foram usados histogramas de cores de forma que seus tamanhos fossem constantes e portanto independentes do tamanho da imagem de entrada.

O principal critério para determinar o tamanho de cada vetor foi a necessidade do vetor que descreve a textura de um objeto ter o mesmo tamanho do que descreve a cor. Como não seria viável trabalhar com 10.000 imagens, pois as mesmas deveriam ter notações manuais, foi reduzido para testes, o conjunto para 910 imagens significativas do conjunto de características. De posse deste subconjunto do banco de imagens levantou-se as estatísticas (tabela 5.2) a respeito dos tamanhos das imagens de forma a maximizar a possibilidade de balancear textura e cor.

Tabela 5.2: Estatística do Banco de Imagens

Tamanho	Máximo	Mínimo	Média	Mediana	Desvio Padrão	Variância
Largura	1600	224	$457,86$	384	$157,05$	$24665,81$
Altura	1200	148	$351,47$	338	$110,36$	$12180,14$

Baseado no tamanho médio (largura = 457,86 e altura = 351,47), determinou-se que o vetor completo teria 8192 posições no total, sendo 4096 para cada característica. Para se chegar a esse número trabalhou-se primeiramente na característica textura. Considerando-se que o tamanho de cada bloco é em média $152,62$ x $117,16$. Arredondando para 150x120 tem-se uma matriz de 75 x 60 de coeficientes ao se aplicar a Wavelet Haar. Sendo três blocos para cada freqüência tem-se um vetor de 13.500 posições (75 x 60 x 3) representando a característica textura. Em geral as imagens podem ser trabalhadas como se tivesse 256 cores. No caso de três canais teria-se $16.777.216$ posições, um número bem alto. De forma a balancear os vetores, optou-se por usar apenas 1/4 do número de característica de textura (3375 posições). Como o sistema de representação de cores usa um número de bits múltiplos de dois para representar cada cor ($2^8 = 256$ cores). O número da base 2 mais próximo é 4096 (2^{12}). Desta forma ficou estabelecido esse valor para cada característica. Partindo dele determinou-se o número de bins de cores e de coeficientes wavelets para cada situação testada. A seguir, são dados detalhes da construção de cada vetor de características.

5.3.2 Extração da Característica Cor

Para a extração da característica de cor, a imagem de entrada foi testada em três espaços de cores, RGB, HSI e L*a*b*. O sistema RGB foi escolhido por ser o mais comum, sendo desta forma utilizado como referência entre os demais. O sistema HSI foi selecionado por representar a percepção humana de cor. O L*a*b* foi escolhido por ser visualmente uniforme e baseado no sistema perceptual humano, além de ser bastante utilizado em sistemas de busca baseados em cores. A grande vantagem do L*a*b* é de ser perceptualmente uniforme, ou seja, mudanças em seus eixos traduzem em modificações proporcionalmente perceptíveis. Grandes deslocamentos resultam em grandes modificações perceptíveis, pequenos deslocamentos são percebidos como pequenas mudanças.

As características de cores foram extraídas e testadas através da combinação dos canais de cada um dos espaços de cores utilizados, conforme é detalhado mais adiante.

Conforme dito anteriormente, para caracterizar cada objeto foram usados histogramas de cores com tamanho constante de 4096 valores (figura 5.5).

No caso de apenas dois canais o número de bins[4], é 64 (64 x 64 = 4096). Desta forma, o gráfico bidimensional (figura 5.6) gerado é visto como uma matriz de valores G, onde $g_{m,n}$ representa o valor da enésima coluna da enésima linha da matriz G, sendo m o número total de colunas e n o número total de linhas (equação 5.1).

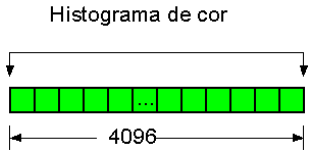

Figura 5.5: Vetor contendo o histograma de cor.

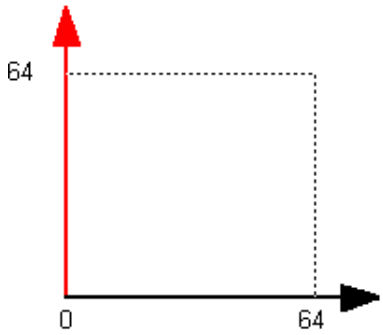

Figura 5.6: Gráfico bidimensional para dois canais.

$$G = \begin{bmatrix} \text{g(1,1)} & \text{g(1,2)} & \text{...} & \text{g(1,m)} \\ \text{g(2,1)} & \text{g(2,2)} & \text{...} & \text{g(2,m)} \\ \text{g(3,1)} & \text{g(3,2)} & \text{...} & \text{g(3,m)} \\ \text{...} & \text{...} & \text{...} & \text{...} \\ \text{g(n,1)} & \text{g(1,2)} & \text{...} & \text{g(l,m)} \end{bmatrix} \tag{5.1}$$

Cada coluna $K(i)$ da matriz G é representada de acordo com a equação 5.2.

$$K(i) = \begin{bmatrix} \text{g(1,i)} \\ \text{g(2,i)} \\ \text{g(3,i)} \\ \text{...} \\ \text{g(n,i)} \end{bmatrix} \tag{5.2}$$

Sendo sua transposta (equação 5.3):

$$K'(i) = \begin{bmatrix} \text{g(1,i) g(2,i) g(3,i) ... g(n,i)} \end{bmatrix} \tag{5.3}$$

[4]cada bin armazena o total de *pixels* com valores dentro de uma determinada faixa, por exemplo, dado uma imagem com 256 valores de tons de cinza, esses valores podem ser agrupados em 128 bins onde cada bin recebe dois valores, o primeiro bin armazena o total de *pixels* dessa imagem com valores na faixa de 0 à 1, o segundo bin armazena o total de *pixels* com valores na faixa de 2 à 3, e assim por diante.

Desta forma, para gerar o vetor unidimensional V, as transpostas de cada coluna são colocadas lado a lado segundo a equação 5.4, para formar o vetor de características V.

$$V = K'(1), K'(2), ..., K'(m). \tag{5.4}$$

Para o caso de três canais de cor, foi usado um histograma tridimensional de 16 x 16 x 16 bins (4096 valores) conforme a figura 5.7. Para gerar o vetor de entradas, os valores são transpostos de forma ordenada conforme o sistema bidimensional. Cada bin do novo eixo possui um vetor de características que serão agrupadas de forma a montar o vetor de características (equação 5.5).

$$V = V(1), V(2), ..., V(16). \tag{5.5}$$

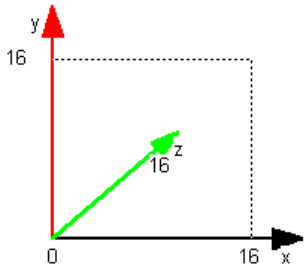

Figura 5.7: Gráfico tridimensional para três canais.

A exceção a esta regra é a situação de usar apenas um canal de cor para tipificar o objeto. Neste caso, é usado um número menor de bins. Considerou-se que em se tratando de características cromáticas e pelas imagens usadas, 256 cores tornam-se suficientes para tratar o problema. Desta forma, a única saída é diminuir o vetor textura para se criar um vetor de características balanceado com apenas 512 valores possíveis. Este histograma depois é transformado em um único vetor, ao qual será agregada a informação de textura, constituindo dessa forma a entrada da primeira camada SOM, conforme a figura 5.11.

Foram usados dois tipos de histogramas, com/e sem ponderação. O histograma de cor ponderado utilizado nesse modelo leva em consideração a informação espacial, ou seja a posição do *pixel* é usada para ponderar o peso da cor no histograma (FERRUGEM et al., 2004). *Pixels* mais próximos do centro da imagem/região possuem pesos maiores que *pixels* mais distantes.

Essa informação é importante no caso de um quadrante possuir mais de um tipo de objeto. Desta forma o objeto mais próximo do centro do quadrante é o que tem maiores chances de ser reconhecido como o objeto dominante ou característico deste quadrante, reduzindo-se assim a ambigüidade de algumas cenas.

O histograma ponderado é construído da seguinte forma(COMANICIU; RAMESH; MEER, 2000):É escolhido um raio h e o ponto central (x_c, y_c) da região a ser processada. Para cada *pixel*, localizado dentro deste raio na região a ser processada, é extraído um vetor de características de cor. Sendo $b(x, y)$ uma função que retorna o número do bin referente a cor localizada na posição (x, y) da região escolhida, o *pixel* é associado a um índice u, sendo $u = b(x, y)$. Considerando m o número de bins que formam o histograma resultante da operação identificado como q, q_u é o valor armazenado no histograma no índice u. Cada característica de cor adiciona uma fração no resultado final, de acordo com a equação 5.6, tendo peso maior as características mais próximas do cen-

tro da região. Nesta equação δ é o delta de Kronecker [5], e a função $k(x, y)$ determina a importância (valor) da feição de cor desse histograma. Essa função deve ser um núcleo (*kernel*) isotrópico com um contorno convexo e de decrescimento monotônico de acordo com (CHENG, 1995). O núcleo utilizado neste modelo foi o de Epanechnikov (equação 5.7) mostrado na figura 5.8. Na equação 5.7, d é a dimensão utilizada (neste caso $d = 2$) e Cd é o volume da unidade d-dimensional, que no espaço bidimensional é uma esfera, e, portanto $Cd = 4/3\pi$.

$$q_u = \frac{\sum_{i=1}^{n} k(|(x_i, y_i) - (x_c, y_c)|/h)\delta(b(x_i, y_i)), u)}{\sum_{i=1}^{n} k(|(x_i, y_i) - (x_c, y_c))|/h)} \tag{5.6}$$

$$k(x, y) = \begin{cases} \frac{1}{2}C_d^{-1}(d + 2)(1 - (x, y)^2) \text{ se } (x, y) \leq h \\ 0 \text{ caso contrário.} \end{cases} \tag{5.7}$$

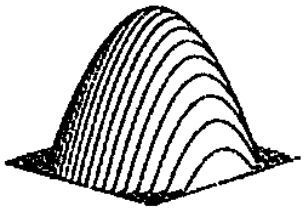

Figura 5.8: Kernel de Epanechnikov (CHENG, 1995).

5.3.3 Extração de Textura

Para a extração da feição de textura, a imagem é convertida para tons de cinza, visto que as componentes de cor são desnecessárias.

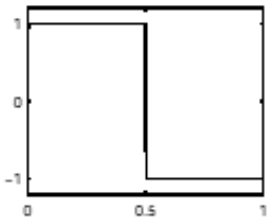

Figura 5.9: Wavelet Haar (MISITI et al., 2001).

As características de textura são obtidas através da aplicação da wavelet Haar na imagem convertida em tons de cinza. Conforme se nota na figura 5.9, esta wavelet é uma função em forma de degrau. Basicamente, essa wavelet, foi escolhida devido ser a mais simples das wavelets. Ao se aplicar a wavelet na imagem, são gerados os coeficientes wavelets da imagem. Um coeficiente wavelet, nada mais é que, o grau de similaridade entre o trecho analisado e a wavelet utilizada (dentro de uma determinada escala). Quanto maior o coeficiente, mais semelhante é a wavelet em relação ao sinal. Em função da necessidade de processamento e do número fixo de entradas da rede SOM, é usado apenas o quarto nível de decomposição wavelet. Em cada nível é gerado quatro bandas de

[5] Essa função retorna 1 caso os indices sejam iguais e 0, caso contrário

freqüência (LL,HL,LH,HH), conforme a figura 5.10, sendo uma de aproximação à imagem original(LL) e três referentes aos detalhes da imagem (HL,LH,HH). Cada banda é o produto cruzado de um filtro passa-baixa(L) e de um filtro passa-alta(H), sendo que, as abreviações usadas nas bandas vem do inglês: L de *low*(baixo) e H de *hight*(alto). Cada uma das bandas é referente aos detalhes, que são utilizadas para caracterizar a textura de uma imagem, mostrando atividades em uma direção (WANG; LI; WIEDERHOLD, 2001). A banda HL mostra atividade no sentido horizontal, a LH no sentido vertical e HH na diagonal. A banda LL (aproximação) possui informações dos componentes de baixa freqüência da imagem, em uma alta escala.

Deve-se ressaltar que, devido a características das wavelets, para cada bloco de quatro *pixels* da imagem original é criado um coeficiente de freqüência.

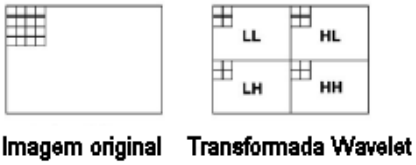

Figura 5.10: Decomposição de uma imagem em bandas de freqüência pela transformada Wavelet (LI; WANG, 2003).

Cada imagem é redimensionada para o tamanho $512x256$ por ser o valor múltiplo de 2 mais próximo do valor médio das imagens$(457,86x351,47)$. Ao ser aplicado o primeiro nível de wavelet, a banda LL armazena a imagem redimensionada pela metade $(256x128)$. É aplicada a wavelet sobre a banda LL gerando quatro novas bandas de tamanho $128x64$. Finalmente se aplica o terceiro nível de wavetet sobre a nova banda LL, gerando 4 bandas de tamanho $64x32$. Os resultados deste nível são quatro bandas de tamanho 2048. Desta decomposição são usadas as bandas HL e LH para descrever a textura da imagem e formar o vetor com 4096 valores. Este vetor será agregado ao vetor gerado na extração das informações de cor.

5.3.4 Vetor de características

Feita a extração das características conforme visto nas seções anteriores, é construído o vetor de características que é constituído pelas informações de cor e de textura mais o rótulo da imagem, de acordo com a figura 5.11

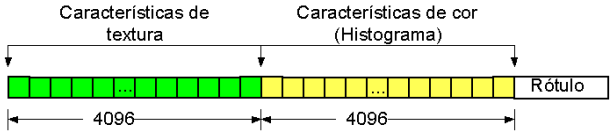

Figura 5.11: Vetor de Feições de Entrada.

O rótulo deste vetor só é utilizado na fase de treinamento, pois o mesmo apenas informa qual é o objeto representado na imagem. Na fase de reconhecimento, ao se construir o vetor de feições da imagem a ser reconhecida, o rótulo não possui nenhuma informação, e mesmo que, supostamente, possuísse alguma informação, ela não seria considerada.

5.3.5 Topologia da Rede

A determinação da rede neural, seguiu os seguintes critérios:

- A rede deve possuir camadas, onde cada camada é relacionada a um nível de generalização dos objetos.

- A rede deve permitir redução de dimensionalidade, pois os dados de entrada possuem alta dimensionalidade, dado que o vetor de feições, usado como entrada, tem tamanho 8192.

Dado esses critérios iniciais, os estudos preliminares apontaram para a rede SOM, pelos seguintes motivos:

Redução dimensionalidade Uma imagem pode ser vista como uma matriz, desta forma é importante reduzi-la para um formato mais compacto sem perder as informações importantes.

Informação hierárquica Como uma imagem pode ser decomposta em vários objetos, as redes hierárquicas como a H-SOM (KOIKKALAINEN; OJA, 2000), TS-SOM (KOSKELA et al., 2000), GH-SOM (DITTENBACH; MERKL; RAUBER, 2000b) e a HOSOM (SUGANTHAN, 1999) são arquiteturas adequadas para o tipo de representação desejada.

Arquiteturas dinâmicas Possibilidade de crescimento da rede através de arquiteturas já existentes como a empregada na H-SOM (MIIKKULAINEN, 1990).

Desta forma os Mapas Auto-Organizáveis tornaram-se uma opção quase que natural.

Existem vários modelos hierárquicos para as redes SOM, mas a grande maioria, conforme foi constatado posteriormente, prioriza a classificação de forma hierárquica das características, mas não permitem organizar de forma hierárquica as categorias como é necessário. Essa conclusão só foi possível após observar algumas redes. Entres as estudadas é interessante observar a TS-SOM(Tree Structured Self-Organizing Map) (KOSKELA et al., 2000) e a GH-SOM, onde ambas estruturas funcionam bem na construção de hierarquias de feições e em geral são usadas como o próprio índice em CBIR's que trabalham com similaridade das imagens. Mas nenhuma das estruturas consegue manter a estrutura de categorias desejadas. Outros problemas são: a GH-SOM pode crescer de forma desbalanceada; na TS-SOM não é possível garantir a organização em árvore que ela implementa, e não é claro como inserir rótulos em camadas intermediárias, visto que os Mapas trabalham com refinamentos sucessivos das características de entrada

Sendo assim, a rede utilizada foi construída de acordo com as necessidades do sistema. A hierarquia é constituída por uma estrutura de redes SOM em camadas. Cada camada corresponde a um nível de categorias, com maior ou menor complexidade, que servirão como descritores da cena, de acordo com a figura 5.15. Dessa forma, na primeira camada são tratados os objetos básicos, tais como: rocha, grama, céu e outros. Uma das possibilidades de objetos básicos que foi avaliada, era a de usar formas trigonométricas simples, mas esse tipo de estrutura aplicada a imagens reais não é trivial, além do que este tipo de abordagem não corresponde às necessidades do sistema, que trabalha com imagens bastante complexas em relação a formas trigonométricas simples.

Uma vez treinada essa camada, ao ser fornecida uma imagem na entrada do sistema, serão selecionados os neurônios vencedores para cada região, das nove, em que é dividida a imagem. O rótulo de cada neurônio vencedor é usado para identificar o objeto reconhecido. A figura 5.13 exemplifica essa situação. Dada a imagem de entrada (figura 5.12), ela é dividia em nove regiões. Cada região é individualmente apresentada para a primeira

Figura 5.12: Exemplo de imagem a ser classificada

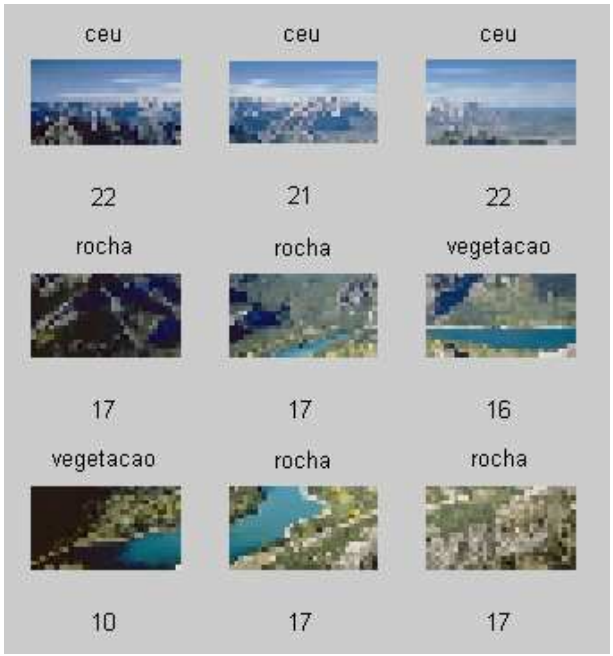

Figura 5.13: Classificação gerada pela primeira camada de SOM, os números indicam o neurônio vencedor para cada região

camada, e para cada neurônio vencedor, identificado por seu número na figura 5.13, é usado seu rótulo para identificar o objeto localizado na região.

A segunda camada faz a ligação entre as categorias da primeira, considerando que, objetos básicos compõem categorias que descrevem cenas mais complexas. Para exemplificar, considerando uma cena como a praia da figura 5.14, esta pode ser composta das seguintes categorias básicas: água (geralmente nas partes inferiores da imagem), céu (geralmente localizado na partes superiores da imagem) e areia. Assim, para identificar uma cena mais complexa, é avaliada a saída da primeira camada, cujos, neurônios vencedores formam o vetor de entrada para a segunda camada.

Ainda é possível acrescentar uma terceira camada, para realizar a separação entre imagem internas e externas (paisagens).

Figura 5.14: Imagem de uma categoria praia composta por objetos básicos.

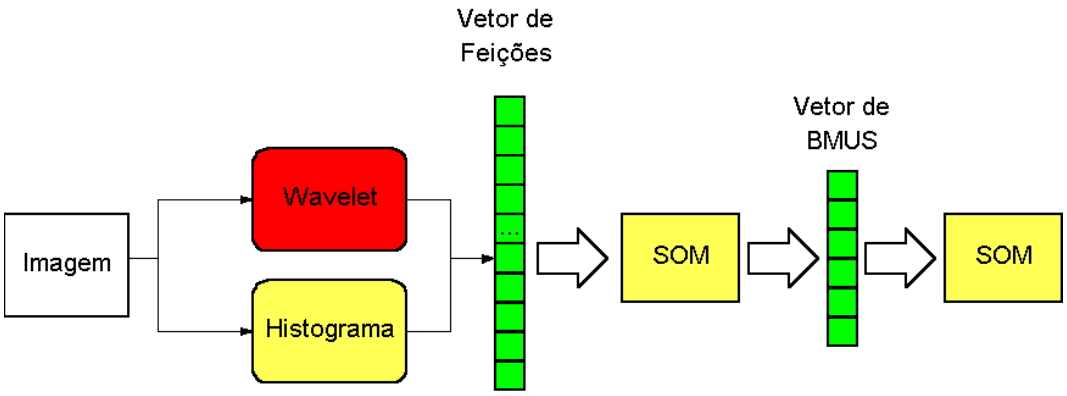

Figura 5.15: Processamento dos dados no modelo proposto.

5.4 Aprendizagem do sistema

A etapa de aprendizagem é feita de forma bem simples. São selecionadas partes das imagens que representam cada um dos objetos básicos previamente descritos. Para a textura cada amostra é redimensionada para as dimensões $512x256$ *pixels*, conforme referenciado nas seções anteriores. Essas amostras são porções das categorias básicas, criadas manualmente, onde, por exemplo, uma porção retirada de uma rocha, serve como amostra desse tipo de categoria. Feita a coleção de amostras, que neste caso se constituiu de vinte

amostras para cada tipo básico, é gerado um conjunto de treinamento, com os vetores de cada amostra, constituídos das feições de cor e textura, mais o rótulo que é extraído do nome do arquivo, conforme a figura 5.11 Esses vetores formam o conjunto de treinamento da primeira camada. Interessante observar que o tamanho médio das amostras (tabela 5.3) não diferem muito do tamanho médio de cada um dos nove setores em que a imagem será dividida.

Tabela 5.3: Estatística das Amostras dos Objetos Básicos

Tamanho	Máximo	Mínimo	Média	Mediana	Desvio Padrão	Variância
Largura	459	76	180, 92	172	58, 45	3.415, 85
Altura	387	56	132, 72	119	56, 44	3.185, 99

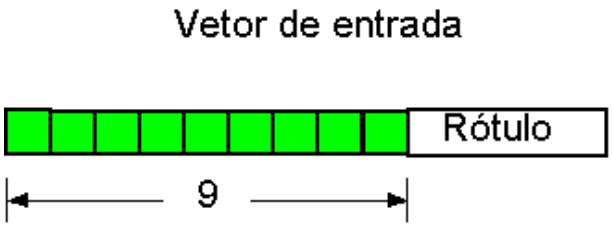

Figura 5.16: Vetor de treinamento.

Após a primeira camada ser treinada com os objetos básicos são apresentadas a ela imagens sem restrições de tamanho, que representam as categorias da próxima camada (praia, cadeia de montanhas). Essas imagens são divididas em nove regiões de tamanho fixo. Destas regiões são extraídos os neurônios vencedores que vão formar o vetor de treinamento da próxima camada conforme a figura 5.16. Esses vetores são utilizados para construir o conjunto de treinamento da segunda camada. Esse processo pode ser realizado sucessivamente de acordo com as necessidades das categorias. A figura 5.17 mostra o resultado de uma das redes geradas para testes. A primeira camada possui os neurônios devidamente treinados com as categorias básicas, enquanto a segunda camada está com as categorias praia e montanhas treinadas a partir da primeira camada.

5.4.1 Seleção das amostras

Um grande problema encontrado foi a superposição de amostras, já que elas são selecionadas de forma empírica pelo usuário, não existindo nenhuma forma de verificar a similaridade das mesmas e eliminar as amostras inadequadas. A solução encontrada foi utilizar as capacidades da GH-SOM para avaliar as amostras com similaridade e substituí-las por outras mais significativas . A GH-SOM foi extremamente útil neste sentido e deve ser utilizada como uma característica a mais neste sistema. A figura 5.18 destaca uma amostra ruim. Na seleção, pode-se notar que as amostras de água, areia e céu estão na mesma célula da GH-SOM, o que demonstra que de acordo com os critérios usados essas amostras são semelhantes.

5.5 Funcionamento

O funcionamento da rede ocorre de forma bem simples:Dada uma imagem de entrada, essa é redimensionada para o padrão de entrada da rede (textura). Feito isso, os

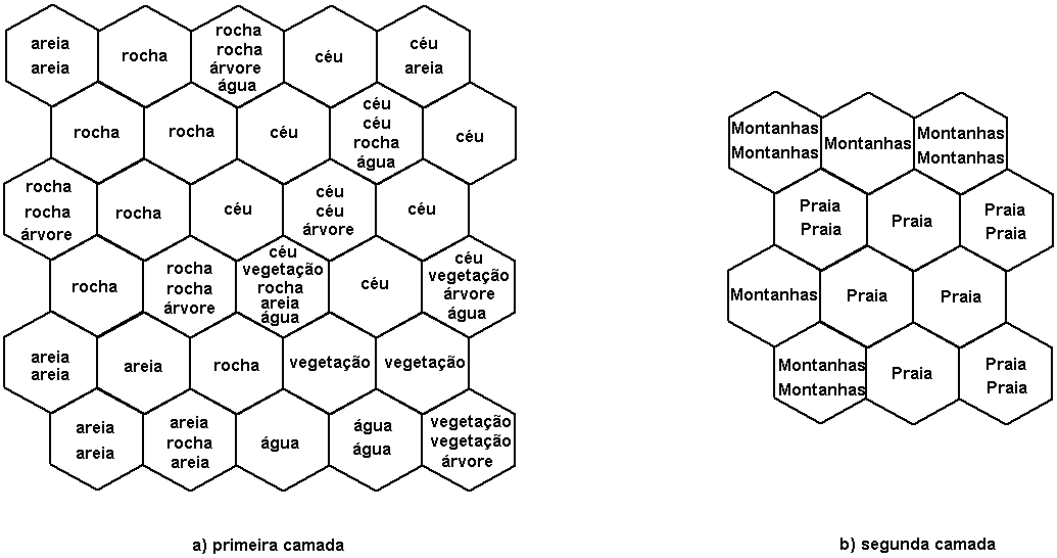

Figura 5.17: Exemplo de mapa gerado usando a técnica proposta . a) SOM da primeira camada para os objetos básicos: rocha, areia, árvore, vegetação e água, b) SOM da segunda camada para categorias praia e montanhas.

rótulos dos neurônios vencedores são utilizados para a geração da descrição da imagem no primeiro nível. Em seqüência, os nove neurônios vencedores formam um vetor que alimenta a segunda camada, onde o neurônio vencedor vai gerar o rótulo da categoria referente a segunda camada SOM, que comporta categorias mais complexas, e assim sucessivamente. O interessante dessa rede é que um novo índice com esses rótulos pode ser gerado através do treinamento de novas formas, permitindo que, apesar de se uma informação textual extraída das imagens em questão, o sistema tem uma certa plasticidade visto que o mesmo pode reorganizar todas as imagens do banco de dados a partir de um novo conceito aprendido, bastando que se faça um novo catálogo baseado nestes novos conceitos aprendidos.

Desta forma, para recuperar uma imagem, é fornecida uma descrição textual baseada nas categorias de imagens pré-definidas e a rede fornece as imagens que casam com a descrição.

A implementação da rede foi feita no Matlab 6.1, utilizando toolboxes de redes neurais do próprio MatLab junto com a SOM toolbox [6] e a GH-Som Toolbox.

5.6 Testes

Os testes realizados medem a acurácia e a precisão do sistema. Foram feitos os seguintes testes:

1. Dois testes usando somente a característica cor.
2. um teste usando somente a característica textura.
3. um teste usando ambas características.

A acurácia do sistema (equação 5.8) é medida da seguinte forma: Ao se realizar uma busca, são retornadas n figuras, considerando-se que destas n figuras, ter-se-é x figuras

[6] Disponível no site http://www.cis.hut.fi/projects/somtoolbox/

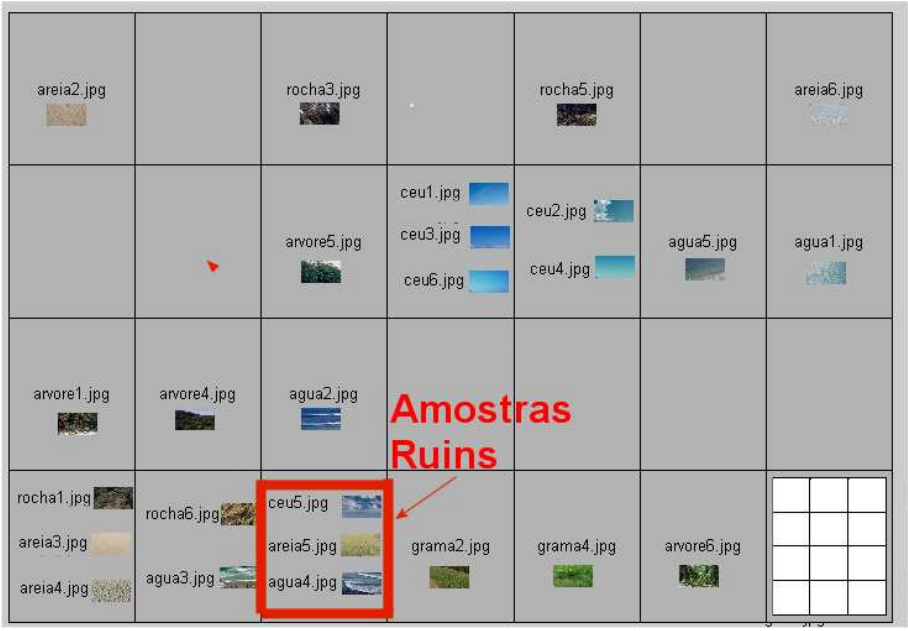

Figura 5.18: Amostras classificadas pela GH-SOM.

corretas. Estas *x* imagens corretas retornadas são divididas pelo número total de imagens existentes no banco de imagens que casam com a descrição. Já a precisão (equação 5.9) usa o número total de figuras corretas retornadas dividido pelo número **total** de figuras retornadas na busca.

Por exemplo, de uma busca são retornadas 16 figuras, sendo 5 corretas. Considerando-se que o número de imagens do banco de imagens que casam com a busca são 6 e apenas 5 foram retornadas, a acurácia do sistema é de $5/6$ ou $0,83$ ou 83%. Já a precisão fica sendo $5/16 = 0,31$ ou 31% ou seja de um total de 16 figuras apenas 31% são condizentes com a busca realizada.

$$\text{Acurácia} = \frac{\text{número de figuras retornadas condizentes com a busca}}{\text{número total de figuras condizentes com a busca}} \tag{5.8}$$

$$\text{Precisão} = \frac{\text{número de figuras retornadas condizentes com a busca}}{\text{número de figuras retornadas}} \tag{5.9}$$

Os mapas auto-organizáveis gerados em todos testes são formados por uma grade bidimensional de tamanho 12x12 e formato hexagonal. Para determinar a vizinhança, é usada a função "bubble"(equação 5.10), sendo o algoritmo de treinamento da rede em lote. Neste algoritmo, todo o conjunto de treinamento é apresentado a rede, um vetor por vez, antes que se façam os ajustes dos pesos. Na equação 5.10, σ_t é o raio da vizinhança no tempo t, $d_{ci} = \|r_c - r_i\|$ é a distância entre as unidades/neurônios c e i do mapa, e $1(x)$ é a função de passo, onde, $1(x) = 0$ se $x < 0$ e $1(x) = 1$ se $x > 0$. A figura 5.19 apresenta a função "*bubble*"com raio de vizinhança igual a 2.

$$h_{ci}(t) = 1(\sigma_t - d_{ci}) \tag{5.10}$$

Para evitar muitas ambigüidades, não foram consideradas as categorias: paisagem com água, floresta e vilarejo. Um subconjunto de 100 imagens do banco de imagens inicial foram classificadas automáticamente. Essa redução foi feita para viabilizar os testes em menor tempo.

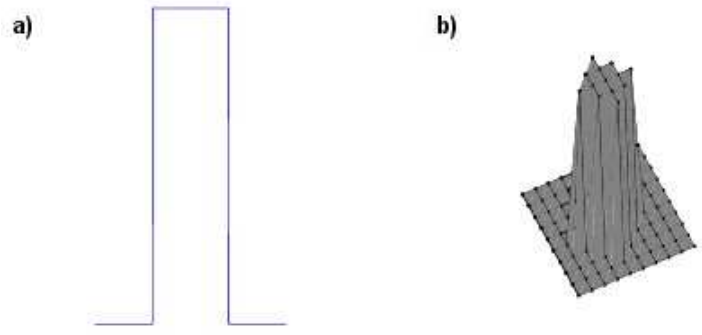

Figura 5.19: a) Representação unidimensional da função "*bubble*"; b) Representação bidimensional da função "*bubble*".

O primeiro teste utilizou os canais R e G; o segundo teste utilizou os canais: R, G e B, o terceiro teste trabalhou com a wavelet Haar no terceiro nível de decomposição. O último teste combinou as características do segundo e terceiro em um só vetor.

5.6.1 Resultados

Os resultados de acurácia e precisão para vetor de características baseado na informação dos canais R e G são mostrados na tabela 5.4. Nesta tabela, é possível notar que a categoria Oásis tem a melhor acurácia com retorno de 100% de todas imagens contendo-a. Por outro lado a categoria Cidade tem maior precisão, visto que 100% das figuras retornadas pertenciam a essa categoria.

Tabela 5.4: Acurácia e Precisão das categorias para os canais R e G

	Acurácia	Precisão
Deserto	0,76	0,36
Montanha	0,30	0,60
Oasis	1,00	0,27
Rural	0,55	0,50
Praia	0,40	0,50
Cidade	0,15	1,00

Para os objetos básicos a melhor acurácia foi Vegetação, com 53%. Enquanto, a melhor precisão foi de 59% pertencente ao objeto Céu (tabela 5.5).

Para fins comparativo, foi criada uma tabela contendo os resultados da classificação. Nesta tabela, cada linha indica em porcentagem a quantidade de vezes que um objeto/categoria foi classificado como o objeto/categoria da coluna respectiva. Por exemplo, na tabela 5.6, a categoria Deserto foi reconhecida como:

Deserto 76% das vezes (acurácia);

Montanha 6%;

Oásis 0%;

Rural 0%;

Tabela 5.5: Acurácia e Precisão dos objetos básicos para os canais R e G

	Acurácia	Precisão
Vegetação	0,53	0,19
Céu	0,36	0,59
Areia	0,15	0,41
Rocha	0,25	0,31
Prédio	0,10	0,52
Casa	0,33	0,09
Água	0,21	0,12
Rua	0,14	0,09

Praia 18%. Esta foi a categoria onde houve o maior índice de erros;

Cidade 0%.

Nota-se que a diagonal superior da esquerda para a direita registra a acúracia de cada categoria. O maior índice de erros para Deserto deu-se com a categoria Praia (18%), sendo natural a confusão considerando-se a semelhança.

Tabela 5.6: Classificação das categorias para os canais R e G

%	Deserto	Montanha	Oasis	Rural	Praia	Cidade
Deserto	76	6	0	0	18	0
Montanha	5	30	55	5	5	0
Oasis	0	0	100	0	0	0
Rural	36	0	0	55	9	0
Praia	47	0	7	7	40	0
Cidade	41	11	15	15	4	15

Já para os objetos básicos houve maior índice de erros dada a possibilidade de confusão entre suas distribuições de cores. Com exceção da coluna Vegetação que aparentemente atraiu o maior índice de classificação para si, em geral, a coluna com maior valor é a da própria categoria.

Tabela 5.7: Classificação dos objetos básicos para os canais R e G

%	Vegetação	Céu	Areia	Rocha	Prédio	Casa	Água	Rua
Vegetação	53	13	0	5	0	12	15	2
Céu	25	36	2	4	2	12	17	2
Areia	73	2	15	1	1	7	1	0
Rocha	45	0	6	25	1	20	1	1
Prédio	8	2	5	25	10	32	3	16
Casa	35	2	7	19	5	33	0	0
Água	54	18	0	5	3	0	21	0
Rua	0	0	10	43	19	10	5	14

Para três canais (R, G e B) os índices melhoram. Apenas as categorias Oásis e Praia possuem menor acurácia em relação ao teste anterior. Oásis diminui sua acurácia, mas aumenta sua precisão, enquanto Praia apresenta rendimento inferior, tanto em acurácia quanto precisão.

Tabela 5.8: Acurácia e Precisão das categorias para os canais R, G e B

	Acurácia	Precisão
Deserto	0,81	0,62
Montanha	0,41	0,56
Oasis	0,75	0,20
Rural	0,83	0,59
Praia	0,20	0,30
Cidade	0,59	0,94

Tabela 5.9: Acurácia e Precisão dos objetos básicos para os canais R, G e B

	Acurácia	Precisão
Vegetação	0,77	0,60
Céu	0,28	0,87
Areia	0,76	0,43
Rocha	0,45	0,35
Prédio	0,08	0,31
Casa	0,14	0,08
Água	0,38	0,16
Rua	0,33	0,10

Esse teste possui um bom desempenho na classificação das categorias (tabela 5.10).

Tabela 5.10: Classificação das categorias para os canais R, G e B

%	Deserto	Montanha	Oasis	Rural	Praia	Cidade
Deserto	81	6	6	0	6	0
Montanha	14	41	14	18	14	0
Oasis	0	0	75	0	0	25
Rural	0	0	8	83	8	0
Praia	7	40	20	13	20	0
Cidade	15	0	15	4	7	59

A distorção que ocorria no teste anterior para a categoria Vegetação é eliminada (tabela 5.11). Os erros apresentados são mais justificáveis dada a semelhança entre os objetos. É o que ocorre com Céu e Água.

Ao utilizar-se apenas a informação de textura, o rendimento do sistema cai de forma significativa (tabelas 5.12, 5.13). Isso deve-se ao fato de trabalhar-se com imagens heterogêneas, onde, é difícil achar padrões estáticos de textura para objetos e categorias.

Tabela 5.11: Classificação dos objetos básicos para os canais R, G e B

%	Vegetação	Céu	Areia	Rocha	Prédio	Casa	Água	Rua
Vegetação	77	1	10	7	1	3	1	1
Céu	2	28	22	6	4	9	28	1
Areia	2	0	76	10	3	2	4	3
Rocha	6	1	28	45	1	9	6	4
Prédio	6	1	3	24	8	21	8	29
Casa	19	0	21	33	7	14	0	7
Água	18	3	18	21	0	3	38	0
Rua	14	0	10	14	0	14	14	33

No caso da precisão dos objetos básicos (tabela 5.13), o sinal - indica que não foi retornado nenhuma imagem referente ao objeto associado, portanto não é possível calcular sua precisão.

Tabela 5.12: Acurácia e Precisão das categorias usando a característica textura

	Acurácia	Precisão
Deserto	0,24	0,50
Montanha	0,30	0,29
Oasis	0,33	0,13
Rural	0,45	0,33
Praia	0,53	0,30
Cidade	0,22	0,67

Tabela 5.13: Acurácia e Precisão dos objetos básicos usando a característica textura

	Acurácia	Precisão
Vegetação	0,00	0,00
Céu	0,84	0,20
Areia	0,22	0,12
Rocha	0,00	-
Prédio	0,00	-
Casa	0,00	-
Água	0,03	0,07
Rua	0,05	0,50

A melhor acurácia do teste envolvendo somente a textura foi da categoria Praia, com 53%, seguido por Rural, 45% (tabela 5.14). Nos objetos básicos, Céu teve a melhor acurácia, seguido por Areia (tabela 5.15). Esses dois objetos possuem uma textura bem mais definida que outros objetos. Convém observar que ambos participam da classificação de Praia.

Ao combinar a informação de textura com cor, houve uma melhora significativa em relação ao teste com somente textura. O teste combinado teve um bom desempenho para quase todas as categorias, com exceção de Cidade onde a acurácia e precisão foi de

Tabela 5.14: Classificação das categorias para textura

%	Deserto	Montanha	Oasis	Rural	Praia	Cidade
Deserto	24	35	0	24	12	6
Montanha	24	35	0	24	12	6
Oasis	0	17	33	17	17	17
Rural	0	0	18	45	36	0
Praia	0	20	13	7	53	7
Cidade	11	19	26	4	19	22

Tabela 5.15: Classificação dos objetos básicos para textura

%	Vegetação	Céu	Areia	Rocha	Prédio	Casa	Água	Rua
Vegetação	0	78	21	0	0	0	1	0
Céu	1	84	15	0	0	0	1	0
Areia	0	76	22	0	0	0	2	0
Rocha	0	72	26	0	0	0	2	0
Prédio	2	58	38	0	0	0	2	0
Casa	2	49	42	0	0	0	5	2
Água	0	82	15	0	0	0	3	0
Rua	0	67	24	0	0	0	5	5

0%. Nos objetos básicos (tabela 5.19), Prédio teve o pior índice, 1%. Comparando-se as tabelas 5.11 e 5.19, observa-se que as informações sem textura possuem índices de erro menor.

Tabela 5.16: Acurácia e Precisão das categorias usando a característica textura e os canais R, G e B

	Acurácia	Precisão
Deserto	0,59	0,48
Montanha	0,30	0,50
Oasis	1,00	0,24
Rural	0,64	0,70
Praia	0,40	0,25
Cidade	0,00	0,00

As figuras 5.20 e 5.21 mostram um quadro comparativo da precisão e da acurácia para a classificação das categorias. O teste que utiliza somente cor (R, G e B) teve um desempenho médio superior aos outros (tabelas 5.20 e 5.21). O teste com textura e cor não superou o desempenho médio do teste com três canais de cor.

Pelos gráficos das figuras 5.22 e 5.23, observa-se que o objeto Céu teve a melhor acurácia para o teste com somente textura, e , a melhor precisão para o teste evolvendo os canais R, G e B.

Tabela 5.17: Acurácia e Precisão dos objetos básicos usando a característica textura e os canais R, G e B

	Acurácia	Precisão
Vegetação	0,72	0,50
Céu	0,42	0,73
Areia	0,77	0,31
Rocha	0,19	0,32
Prédio	0,01	0,33
Casa	0,12	0,09
Água	0,21	0,10
Rua	0,52	0,16

Tabela 5.18: Classificação das categorias para os canais R, G, B e textura

%	Deserto	Montanha	Oasis	Rural	Praia	Cidade
Deserto	59	6	35	0	0	0
Montanha	15	30	40	5	10	0
Oasis	0	0	100	0	0	0
Rural	0	0	0	64	0	36
Praia	13	20	13	13	40	0
Cidade	22	7	11	0	59	0

Tabela 5.19: Classificação dos objetos básicos para os canais R, G, B e textura

%	Vegetação	Céu	Areia	Rocha	Prédio	Casa	Água	Rua
Vegetação	72	1	13	10	0	1	1	2
Céu	3	42	24	0	1	8	22	0
Areia	14	2	77	3	0	2	2	0
Rocha	10	5	53	19	1	2	8	2
Prédio	8	2	11	16	1	20	10	32
Casa	28	0	47	2	2	12	0	9
Água	5	13	49	8	0	3	21	3
Rua	10	0	19	19	0	0	0	52

Tabela 5.20: Acurácia e precisão média da classificação de categorias

%	Acurácia	Precisão
Textura e RGB	49	36
Textura	35	37
RGB	60	54
RG	53	54

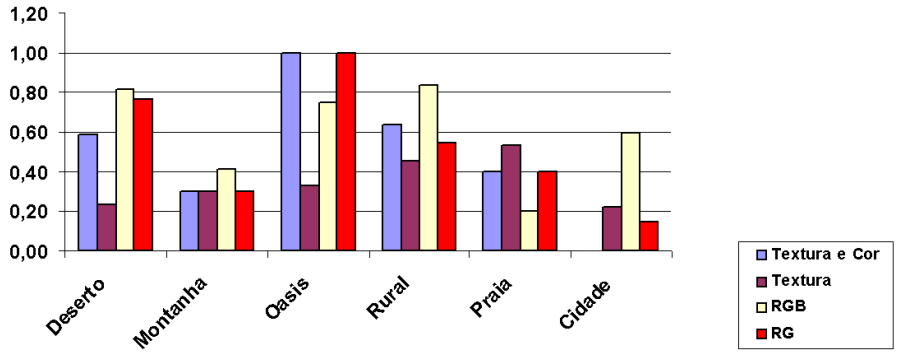

Figura 5.20: Acurácia da classificação das categorias.

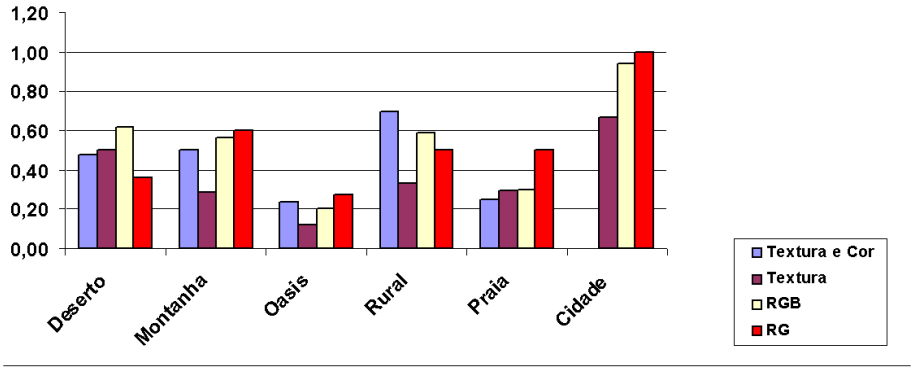

Figura 5.21: Precisão da classificação das categorias.

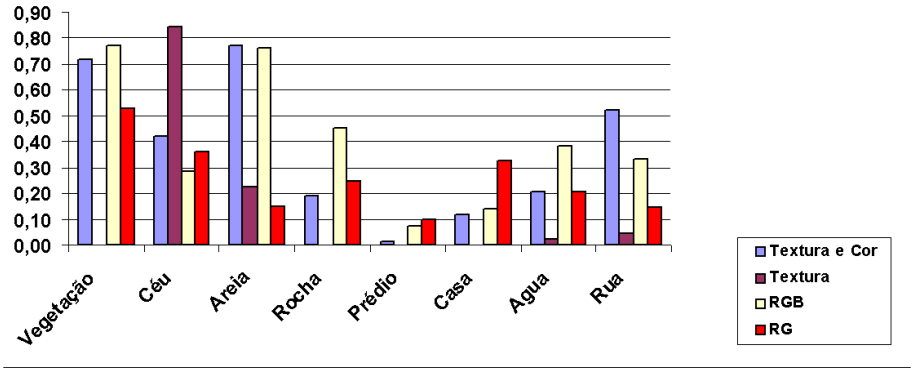

Figura 5.22: Acurácia da classificação dos objetos básicos.

Tabela 5.21: Acurácia e precisão média da classificação de objetos básicos

%	Acurácia	Precisão
Textura e RGB	37	32
Textura	14	18
RGB	40	36
RG	26	29

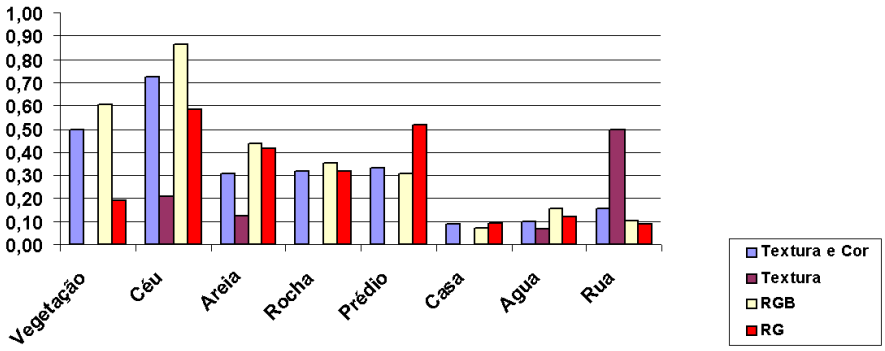

Figura 5.23: Precisão da classificação dos objetos básicos.

6 CONCLUSÕES E TRABALHOS FUTUROS

A criação de um sistema capaz de montar índices que descrevam uma imagem de forma automática é essencial para a busca e recuperação de imagens e no reconhecimento de objetos através do computador (WANG; LI; WIEDERHOLD, 2001). Apesar de muitos sistemas de CBIR apostarem em uma maior interatividade com o usuário, para a busca de imagens em um banco de imagens digitais, é necessário uma maior compreensão da cena que compõe a imagem, de forma que exista realmente uma busca mais simplificada e menos exaustiva. Além disso, sistemas iguais ou com o mesmo enfoque do apresentado neste trabalho, possuem uma potencialidade de aplicação em várias áreas, como comércio, educação, bibliotecas digitais entre outros.

Criar índices de forma automática não é uma tarefa simples, pois, além do problema de relacionar as características de baixo nível, como textura e cor, com informações semânticas de alto nível, tais como "homem surfando", que associam a um conjunto de *pixels* da imagem um significado, existem os problemas relacionados com a robustez do pré-processamento.

Este trabalho teve bons resultados. Os testes demonstraram que quanto maior a quantidade e melhor a qualidade das características extraídas, maior a possibilidade de se produzirem boas classificações. A melhora significativa nos testes envolvendo o espaço de cor RGB, entre o teste com apenas os canais R e G e o com os canais R, G e B demonstra essa situação. Apesar dos dados envolvendo o teste com três canais terem menor precisão, pois cada canal é comprimido em 16 *bins*, ao contrário do teste com dois canais onde é usado 64 *bins* por canal, o acréscimo do canal B melhorou a qualidade da classificação.

A textura mostrou-se menos eficaz na classificação em relação as cores para imagens heterogêneas. Isso se dá em virtude da impossibilidade de extrair texturas padrão das imagens apresentadas. A categoria Cidade, cuja as imagens variam o ângulo, tipo de edificações e podem apresentar ou não multidões é o melhor exemplo deste problema. Outro fator é a possibilidade de encontrar-se no mesmo quadro duas categorias diferentes, podendo gerar um padrão de textura híbrido. Portanto, observa-se a necessidade de tratar a textura de forma local (textura de pele, formações rochosas, multidões, floresta, etc.).

O sistema mostrou-se promissor e com indices compatíveis com os gerados pelo Alip , cujo resultados foram descritos no capítulo 4.

O trabalho aqui apresentado representa uma parcela de um sistema maior de reconhecimento e busca. Existem vários trabalhos futuros, que podem ser feitos com base na arquitetura descrita, mas que, devido à abrangência do assunto, não puderam ser tratadas nesta dissertação.

Um dos temas que pode ser tratado em trabalhos futuros, é a questão do pré - processamento da imagem; a variância da cor em imagens é um problema existente e que influencia a busca. Uma das possibilidades de tratar esse problema, é explorando por al-

goritmos como o Retinex (FUNT; CIUREA; MCCANN, 2004). Esse algoritmo corrige as cores de acordo com a vizinhança dos *pixels*, baseado em um processo semelhante ao que os seres humanos fazem. Outro trabalho futuro seria pesquisar qual Wavelet é mais adequada para tratar textura em imagens coloridas, além de estender sua aplicação para as outras componentes de cor.

O estudo de como será inserido a característica de forma no sistema também implica em um outro trabalho. Um dos principais problemas com esse tipo de feição é a segmentação automática. O BlobWorld (CARSON et al., 1999) parece fornecer bons resultados utilizando o Algoritmo EM (Maximização de Expectativa), sendo extremamente proveitoso utilizar a informação de região como parte da informação necessária para caracterizar uma cena. A abordagem de janelas móveis e escaláveis, em contrapartida ao conjunto de quadrantes fixos é outro ponto de exploração em aberto, visto que permite uma classificação mais flexível.

A principal métrica para avaliar o desempenho dentro da área de busca e recuperação de imagens é averiguar a precisão e acurácia do sistema de indexação. Para que isso seja possível, é necessário ter um bom número de imagens anotadas pelo processo manual, para que se possa averiguar a precisão do sistema. Todo esse processo demanda um bom tempo de trabalho manual. Para citar um exemplo, o Alip (LI; WANG, 2003) de Stanford tem 60.000 imagens anotadas manualmente, consistindo um conjunto extremamente significativo de dados, o sistema deste trabalho possui 1.000 anotações manuais. Resta ainda realizar, como trabalho futuro, mais estatísticas e métricas , para avaliação de desempenho no sistema apresentado em outros espaços de cor.

Dados os objetivos iniciais, o sistema correspondeu de forma adequada aos problemas apresentados e gerou também um novo conjunto de demandas de tratamento multidisciplinar para trabalhos futuros.

Os testes realizados nas classificações das imagens mostraram-se bastante promissores. Mas uma das principais colaborações deste sistema, é a abordagem voltada para a semântica das imagens, organizando as descrições das cenas de forma hierárquica. Em geral, a preocupação dos sistemas que possuem tratamento hierárquico está voltada para a criação de modelos de objetos através de uma hierarquia de feições. Esse tipo de modelagem acaba reduzindo a descrição das imagens em um número infinito de modelos de cenas *ad hoc*, sem se preocupar com a dependência entre os objetos mais simples na construção de uma descrição.

Outra contribuição deste trabalho, é o de usar um sistema de classificação de cores que leva em consideração a disposição espacial dos objetos em questão, permitindo que algumas ambigüidades sejam parcialmente corrigidas. Assim, objetos com quantidades de cores semelhantes são distanciados, a partir da disposição dessas cores.

A GH-SOM em geral tem sido utilizada para classificar livros e outros objetos de forma hierárquica de acordo com suas feições de entrada. Usá-la para visualizar a distribuição hierárquica das amostras e assim eliminar parte dos problemas gerados por amostras não significativas das categorias, é um aproveitamento aparentemente inédito dessa arquitetura e que se mostrou bastante útil dentro deste trabalho.

Esse trabalho possui um grande potencial de aplicação. Considerando os pontos em aberto, comuns a toda área, pode-se dizer que essa arquitetura encerra uma abordagem diferenciada das adotadas até então, permitindo observar o problema do *gap* semântico a partir de um novo ponto de vista.

REFERÊNCIAS

BARONE, D. A. C. **Sociedades Artificiais**: a nova fronteira da inteligência nas máquinas. Porto Alegre: Bookman, 2003. 332p.

BATCHELOR, B.; WHELAN, P. **Intelligente Vision Systems for Industry**. Londres, UK: Springer, 2002. 473p.

BATCHELOR; WALTZ. **Intelligente Machine Vision**: techniques, implementations and applications. Londres, UK: Springer, 2001. 392p.

BITTENCOURT, G. **Inteligência Artificial**: ferramentas e teorias. Florianopólis: [s.n.], 1998. 362p.

CARSON, C.; THOMAS, M.; BELONGIE, S.; HELLERSTEIN, J. M.; MALIK, J. Blobworld: a system for region-based image indexing and retrieval. In: INTERNATIONAL CONFERENCE ON VISUAL INFORMATION AND INFORMATION SYSTEMS, 3, 1999, Amsterdam. **Proceedings...** Berlin: Springer-Verlag, 1999. p.509–516. (Lecture Notes in Computer Science, v.1614).

CARVALHO, A. C. P. de L. F. de; BRAGA, A. doP.; LUDEMIR, T. B. **Fundamentos de Redes Neurais Artificiais**. Rio de Janeiro: DCC/IM, 1998. 246p.

CHAN, A.; SPRACKLEN, T. Object Recovery Using Hierarchical Self-Organizing Maps. In: INTERNATIONAL CONFERENCE ON ENGINEERING APPLICATIONS OF NEURAL NETWORKS, 2000, Kingston Upon Thames, Reino Unido. **Proceedings...** [S.l.: s.n.], 2000.

CHEN, Y.; WANG, J. Image Categorization by Learning and Reasoning with Regions. **Journal of Machine Learning Research**, [S.l.], n.5, p.913–939, Aug. 2004.

CHENG, Y. Mean Shift, Mode Seeking and Clustering. **IEEE Transaction on Pattern Analysis and Maquine Inteligence**, [S.l.], v.17, n.8, p.790–799, Aug. 1995.

COMANICIU; RAMESH; MEER. Real-time tracking of non-rigid objects using mean shift. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2000. **Proceedings...** [S.l.: s.n.], 2000. v.2, p.142–149.

DAVIDSON, M. W.; NEAVES, S. H.; ABRAMOWITZ, M. **Molecular Expressions**: science, optics and you. Disponível em: <http://micro.magnet.fsu.edu/optics/activities/students/prismsstudent.html>. Acesso em: jan. 2003.

DITTENBACH, M.; MERKL, D.; RAUBER, A. Using Growing Hierarchical Self-Organizing Maps for Document Classification. In: EUROPEAN SYMPOSIUM ON ARTIFICIAL NEURAL NETWORKS, ESANN, 2000, Bruges, Belgium. **Proceedings...** [S.l.: s.n.], 2000. p.7–12.

DITTENBACH, M.; MERKL, D.; RAUBER, A. The Growing Hierarchical Self-Organizing Map. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS,IJCNN, 2000, Como, Italy. **Proceedings...** [S.l.: s.n.], 2000. v.6, p.15–19.

FEICHTINGER, H.; STROHMER, T. **Numerical Harmonic Analysis and Image Processing.** Berlin, Alemanha: Springer, 2001. 7p.

FERRUGEM, A.; PRESTES, E.; IDIART, M.; BARONE, D. A Perceptual User Interface Using Mean Shift. In: ADVANCES IN ARTIFICIAL INTELLIGENCE, IBERO-AMERICAN CONFERENCE ON ARTIFICIAL INTELLIGENCE, IBERAMIA, 9, 2004, Puebla, Mexico. **Proceedings...** Berlin: Springer-Verlag, 2004. p.590–599. (Lecture Notes in Computer Science, v.3315).

FOLEY, J. D.; VANDAM, A. **Fundamentals of Interactive Computer Graphics**. Massachusetts, USA: Addison-Wesley, 1982. 612p.

FORSYTH, D.; PONCE, J. **Computer Vision**: a modern approach. Upper Saddle River: Prentice Hall PTR, 2002. 693p.

FUKUSHIMA. Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. **Pattern Recognition**, [S.l.], v.15, n.6, p.455–469, 1982.

FUNT, B.; CIUREA, F.; MCCANN, J. Retinex in MATLAB. **Journal of Electronic Imaging**, [S.l.], v.13, n.1, p.48–57, Jan. 2004.

GOMES, J.; VELHO, L. **Image Processing for Computer Graphics**. New York: Springer, 1997. 352p.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. California, USA: Addison-Wesley, 1992. 716p.

HAYKIN, S. **Redes Neurais**: príncipios e prática. 2.ed. Porto Alegre: Bookman, 2001. 900p.

IQBAL, Q.; AGGARWAL, J. K. COMBINING STRUCTURE, COLOR AND TEXTURE FOR IMAGE RETRIEVAL:a performance evaluation. In: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, ICPR, 2002, Quebec City, Canada. **Proceedings...** [S.l.: s.n.], 2002.

JäHNE, B.; HAUßECKER, H. **Computer Vision and Applications**: a guide for students and practitioners. California, USA: Academic Press, 2000. 678p.

KOHONEN, T. **Self-Organization and Associative Memory**. New York: Springer, 1989.

KOIKKALAINEN, P.; OJA, E. Self-organizing hierarchical feature maps. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, IJCNN, 2000. **Proceedings...** Los Alamintos: IEEE Computer Society, 2000. v.2, p.279–284.

KOSKELA; LAAKSONEN; LAAKSO; OJA. The PicSOM Retrieval System: description and evaluations. In: CHALLENGE OF IMAGE RETRIEVAL, CIR, 2000. **Proceedings...** [S.l.: s.n.], 2000.

KOSKELA, M. **Interactive Image Retrieval using Self-Organizing Maps**. 1999. 107p. Dissertação (Mestrado em Ciência da Computação) — Helsinki University of Technology, Espoo, Finland.

KOVáCS, Z. L. **Redes Neurais**: fundamentos e aplicações. 2.ed. Rio de Janeiro: Collegium Cognitio, 1996. 174p.

LAMPINEN, J.; SMOLANDER, S. Self-organizing feature extractions in recognigion of wood surface defects and color imagens. **International Journal of Pattern Recognition and Artificial Inteligence**, [S.l.], v.10, n.2, p.97–113, 1996.

LEW, M. S. Next-Generation Web Searches for Visual Content. **IEEE Computer**, [S.l.], v.33, n.11, p.46–53, 2000.

LEW, M. S. **Principles of Visual Information Retrieval**. Londres, UK: Springer-Verlag, 2001. 356p.

LI, J.; WANG, J. Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v.25, n.9, p.1075–1088, Sept. 2003.

LI, Y.; ; BILMES, J. A.; SHAPIRO., L. G. Object Class Recognition using Images of Abstract Regions. **International Conference on Pattern Recognition, ICPR**, [S.l.], v.1, p.40–43, Dec. 2004.

LüNEBURG, S. S. aus. **Feature Histograms for Content-Based Image Retrieval**. 2002. 140p. Dissertação (Mestrado em Ciência da Computação) — Albert-Ludwigs-Universität Freiburg,, Freiburg, Germany.

LUGER, G. F.; STUBBLEFIELD, W. A. **Artificial Intelligence**:structures and strategies for complex problem solving. 3rd.ed. Reading, USA: Addison-Wesley, 1998. 824p.

MAINZER, K. **Thinking in complexity**: the complex dynamics of matter, mind, and mankind. 2 nd.ed. New York: Springer-Verlag, 1997. 357p.

MANJUNATH, B.; MA, W. Texture Features for Brownsing and Retrieval of Image Data. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v.18, n.8, p.837–842, Aug. 1996.

MIIKKULAINEN, R. Script recognition with hierarchical feature maps. **Connection Science**, [S.l.], v.2, p.83–101, 1990.

MISITI, M.; MISITI, Y.; OPPENHEIM, G.; POGGI, J.-M. **Wavelet ToolBox for use with Matlab - Users Guide**. [S.l.: s.n.], 2001. Disponível em: <www.mathworks.com>. . Acesso em: jan 2004.

PRATT, W. K. **Digital Image Processing**: piks inside. 3rd.ed. California, USA: John Wiley & Sons, 2001. 738p.

RAUBER, A.; MERKL, D.; DITTENBACH, M. The growing hierarchical self-organizing maps: exploratory analysis of high-dimensional data. **IEEE Transactions on Neural Networks**, [S.l.], v.6, n.13, p.1331–1341, 2002.

RIESENHUBER, M.; POGGIO, T. Models of object recognition. **Nature neuroscience supplement**, [S.l.], v.3, p.1199–1204, nov. 2000.

RUSS, J. C. **The Image Processing Handbook**. 3rd.ed. California, USA: CRS Press, 1998. 750p.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence**:a modern approach. Upper Saddle River: Prentice Hall PTR, 1995. 932p.

SAGAN, C. **Cosmos**. [S.l.]: Villa Rica, 1992.

SAMI BRANDT JORMA LAAKSONEN, E. O. Statistical Shape Features in Content-Based Image Retrieval. In: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, ICPR, 2000, Barcelona, Spain. **Proceedings...** [S.l.: s.n.], 2000.

SETHI, I. K.; COMAN, I. Image Retrieval using hierarchical self-organizing feature maps. **Pattern Recognition Letters**, [S.l.], n.20, p.1337–1345, 1999.

SHAPIRO, L.; STOCKMAN, G. **Computer Vision**. Upper Saddle River: Prentice Hall PTR, 2001. 580p.

SMEULDERS, A. W. M.; WORRING, M.; SANTINI, S.; GUPTA, A.; JAIN., R. Content-Based Image Retrieval at the End of the Early Years. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v.22, n.12, p.1349–1380, Dec. 2000.

SUGANTHAN, P. N. Hierarchical Overlapped SOM's for Pattern Classification. **IEEE Transactions on Neural Networks**, [S.l.], v.10, n.1, p.193–196, Jan. 1999.

TSAI, C.-F.; MCGARRY, K.; TAIT, J. Image Classification Using Hybrid Neural Networks. In: ACM, SIGIR, 2003, Toronto, Canadá. **Proceedings...** New york: ACM Press, 2003. p.431–432.

VARELA, F.; THOMPSON, E.; ROSCH, E. **The Embodied Mind**: cognitive science and humam experience. 6.ed. Cambridge, USA: MIT Press, 1997. 308p.

VEELENTURF, L. **Analysis and Application of Artificial Neural Network**. Londres, UK: [s.n.], 1995. 270p.

WALKER, J. **A primer on Wavelets and their scientific applications**. New York: CRS Press, 1999.

WANG, J.; LI, J.; WIEDERHOLD, G. SIMPLIcity: semantics-sensitive integrated matching for picture libraries. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v.23, n.9, p.947–963, Sept. 2001.

WINSTANLEY, G. **Artificial Intelligence in engineering**. Londres, UK: John Wiley & Sons, 1991. 432p.

APÊNDICE A CLASSIFICAÇÃO DE VIENNA

A.1 Classificação de Vienna Categoria 6

6.1 Montanhas, Rochas, Grutas.

- **6.1.1** rochedos, rocha, muros de rochas.// Nota: não incluindo items da referência (6.3.1)
- **6.1.2** Montanhas, Paisagens de montanhas.
- **6.1.3** Vulcões.
- **6.1.4** Montanhas ou vulcões estilizados.
- **6.1.7** Grutas.

6.3 Paisagens com água, rio ou córrego.

- **6.3.1** Paisagens lacustres ou marítimas.
- **6.3.2** Praias, Costas, baía.
- **6.3.3** Ilha, Recife
- **6.3.4** Mar aberto, parte de mar sem costa.
- **6.3.5** Lagos ou oceanos com montanhas em volta ou ao fundo.
- **6.3.6** Porto.
- **6.3.10** Outras cenas lacustres ou marítimas.
- **6.3.11** Paisagens com água corrente.
- **6.3.12** Fendas, Paisagem com fenda(s)
- **6.3.13** Queda d'água, Paisagem com queda d'água(s).
- **6.3.14** Rios, córregos,corredeiras, correntes, com ou sem paisagens.
- **6.3.20** Outras paisagens com água corrente.

6.6 Deserto ou paisagens tipo tropical

- **6.6.1** Paisagens de deserto ou paisagens com vegetação muito esparsa.
- **6.6.2** Oásis
- **6.6.3** Outras paisagens com palmeiras
- **6.6.25** Outras paisagens do tipo tropical.

6.6.3 Outras paisagens com palmeiras

6.7 Paisagens urbanas ou de vilarejos

6.7.1 Ruas

6.7.2 Praças

6.7.4 Áreas Construídas

6.7.5 Áreas Construídas compostas de arranhas-céu.

6.7.6 Áreas Construídas compostas de casas baixas.

6.7.7 Áreas Construídas compostas de cabanas.

6.7.8 Áreas Construídas compostas de construções

6.7.11 Paisagens urbanas ou de vilarejos com água, rio ou córrego.

6.7.25 Outras paisagens urbanas ou cenas de vilarejos.

6.19 Outras paisagens

6.19.5 Floresta, vegetação rasteira [1].

6.19.7 Vinhedos.

6.19.9 Outras áreas cultivadas.

6.19.11 Gramados, Pastagens.

6.19.13 Pastagens com montanhas entorno ou ao fundo.

6.19.15 Paisagens com moinho(s) de vento.

6.19.16 Paisagens com casa(s)

6.19.17 Paisagens com fábrica(s) ou outra(s) construção(ões) industrial(is))

6.19.19 Paisagem polar.

[1] Um grupo de três ou mais árvores pode ser encaixada em uma das seções da 5.1.1 à 5.1.4 se a cena não representar uma floresta propriamente dita