

Localização e Reconhecimento de Placas de Sinalização Utilizando um Mecanismo de Atenção Visual e Redes Neurais Artificiais

Fabício Augusto Rodrigues

Dissertação submetida à Coordenação do Curso de Pós-Graduação em
Informática da Universidade Federal de Campina Grande como parte dos
requisitos necessários para obtenção do grau de Mestre em Informática.

Área de Concentração: Ciência da Computação

Herman Martins Gomes

Orientador

Campina Grande, Paraíba, Brasil

Fabício Augusto Rodrigues, Agosto 2002

Agradecimentos

Agradeço aos meus pais, João Rodrigues e Raquel Rodrigues, pelo apoio incondicional em todos os momentos desta caminhada, além do infinito amor que sempre me dedicaram desde o meu nascimento. Sem eles esta vitória não seria possível.

Agradeço à minha amada Elisângela, que sempre esteve ao meu lado, valorizando os momentos de êxito e me incentivando a superar as dificuldades dos momentos difíceis.

Aos professores participantes da banca examinadora, Joseana e Marcelo Barros, os meus agradecimentos pelas sugestões e críticas que contribuíram para o enriquecimento do trabalho.

Agradeço ao meu orientador Herman Martins Gomes, que sempre procurou conduzir o trabalho com paciência e dedicação.

Agradeço à Aninha, Vera e Zeneide, que sempre se mostraram dispostas a auxiliar os alunos da COPIN. À todas as pessoas que de alguma forma contribuíram para a conclusão deste trabalho, o meu muitíssimo obrigado.

Por fim, agradeço ao bondoso Deus por me conceder esta oportunidade rara, num país onde a maioria da população não tem o direito básico à educação.

Resumo

Esta dissertação tem como principal objetivo investigar o problema da detecção e reconhecimento de placas de sinalização utilizando dados de uma câmera de vídeo acoplada a um carro em movimento. Foi projetada a arquitetura de um protótipo de sistema de reconhecimento contendo dois módulos principais: um Módulo de Detecção para automaticamente localizar placas de trânsito dentro de cada quadro em uma sequência de imagens; e um Módulo de Reconhecimento para classificar as regiões localizadas em cada cena, com base num conjunto prévio de imagens treinadas. Para o Módulo de Detecção, nós utilizamos um mecanismo de atenção baseado em saliência (*bottom-up*), o qual é constituído a partir de uma Pirâmide Gaussiana, bem como a partir de operadores locais de orientação. Testes com este mecanismo apresentaram resultados promissores uma vez que, sinais estavam presentes na maioria das regiões salientes. Experimentos preliminares com o Módulo de Reconhecimento sozinho apresentaram bons resultados, com taxa média de reconhecimento em torno de 84%. Entretanto, ao utilizar as saídas do Módulo de Detecção, em que as imagens não necessariamente são centralizadas, o uso de um classificador neural monolítico apresentou, para todas as classes, resultados insatisfatórios. Devido a este problema, foram realizados alguns experimentos simples envolvendo redes neurais de classificação binária nos quais se demonstrou a viabilidade de utilização de uma estratégia de classificação combinando as saídas destes classificadores. Os resultados também indicaram que melhores taxas de reconhecimento poderiam ser atingidas através de um aumento no número de exemplos de treinamento.

Abstract

The main objective of this dissertation is to investigate the problem of the traffic signs detection and recognition, using data from a video camera attached to a moving car. We designed an architecture containing two main modules: a detection module to automatically locate traffic signs inside each frame; and a recognition module to classify the located regions based on a set of previously trained images. For the detection module we are using a saliency-based attention mechanism (bottom-up), which is constructed from a Gaussian Pyramid, and locally oriented neighborhood operators. Some initial tests of this mechanism showed promising results since signs were present in most image salient regions. Preliminary experiments with the recognition module presented good results, with 84.40% recognition average rate. The results also indicate that better rates could be reached if we increased the number of examples in the training set. However, when using the outputs of the Detection Module, in which sign images are not necessarily centred, the use of a monolithic neural classifier presented, for all classes, insatisfactory results. In face of this problem, we developed new simple experiments involving binary classification networks (discriminating between 2 classes at a time). These experiments have shown that it would be possible to employ a classification strategy combining the output of these binary networks. Results have also shown that better recognition rates could be achieved through an increase in the training set size.

Índice

1	Introdução	1
1.1	Visão e Reconhecimento	1
1.2	Técnicas para Visão	2
1.2.1	Aquisição	3
1.2.2	Pré-processamento	4
1.2.3	Segmentação	4
1.2.4	Representação e Descrição	4
1.2.5	Reconhecimento	5
1.2.6	Base de Conhecimento	6
1.3	O Problema do Reconhecimento de Sinais de Tráfego	6
1.4	Objetivos e Relevância	8
1.5	Estrutura da Dissertação	9
2	Sistemas de Apoio ao Motorista	11
2.1	Introdução	11
2.2	Principais Trabalhos em DSS	13
2.2.1	Detecção de Obstáculos	13
2.2.2	Detecção e Reconhecimento de Sinais de Tráfego	14
2.2.3	Detecção das Marcas da Estrada	15
2.2.4	Sistemas Integrados	16
2.2.5	Aspectos Importantes	18
2.3	Sumário	19

3	Atenção Visual	21
3.1	Atenção em Sistemas Biológicos	21
3.1.1	Sistema Visual Humano	22
3.2	Atenção Visual em Máquinas	27
3.2.1	Computação de Características Visuais Primitivas	28
3.2.2	Representação Piramidal	29
3.2.3	Mapa de Saliência	29
3.2.4	Inibição do Retorno	30
3.2.5	Alguns Modelos de Atenção <i>Bottom-up</i>	30
3.3	Sumário	35
4	Redes Neurais	37
4.1	Fundamentos de Redes Neurais	37
4.2	Notas Históricas	39
4.3	Arquiteturas	40
4.4	Processos de Aprendizagem	41
4.5	Alguns Modelos de Redes Neurais	42
4.5.1	Rede de Kohonen	42
4.5.2	Rede de Hopfield	44
4.5.3	<i>Perceptron</i> Multicamadas (<i>Multilayer Perceptron</i>)	46
4.6	Classificação de Padrões	47
4.7	A Ferramenta SNNS	48
4.8	Sumário	49
5	Arquitetura do Protótipo	51
5.1	Arquitetura Geral	51
5.2	Módulo de Detecção	53
5.2.1	Filtragem Linear	53
5.2.2	Diferenças Centro-Vizinhança (<i>Center-Surround Differences</i>)	60
5.2.3	O Mapa de Saliência	62
5.3	Módulo de Reconhecimento	66
5.4	Sumário	68

6	Experimentos e Resultados	69
6.1	Experimentos Iniciais	69
6.1.1	Resultados Iniciais	71
6.2	Experimentos para Definição da Arquitetura da Rede	73
6.3	Experimentos com o Módulo de Detecção	76
6.4	Experimentos Envolvendo a Integração dos Módulos	79
6.4.1	Módulo de Detecção	82
6.4.2	Módulo de Reconhecimento	84
6.4.3	Análise dos Resultados	90
6.5	Sumário	91
7	Conclusões e Propostas de Trabalhos Futuros	92
7.1	Sumário da Dissertação	93
7.2	Considerações Gerais	94
7.3	Propostas de Trabalhos Futuros	96
A	Base de Imagens	104
A.1	Imagens Utilizadas no Primeiro Experimento	104

Lista de Figuras

1.1	Elementos do proceso de análise da imagem	3
3.1	Esquema do cérebro que mostra o caminho retino-geniculado	22
3.2	Esquema do olho humano que mostra a localização da fóvea na retina.	23
3.3	Exemplo da hierarquia de processamento do modelo de Tsotsos e Culhane: (a) configuração inicial; (b) seleção dos itens de maior interesse - zona de passagem (linhas sólidas) e zona inibida (linhas tracejadas).	31
3.4	Visão geral da arquitetura proposta por Milanese e colegas.	32
3.5	Arquitetura geral do modelo proposto por Itti e colegas	33
3.6	Tendências de fixação utilizadas pelo modelo de Sela e Levine. (a) Resulta- dos (centros de fixação) de experimentos examinando a fixação do olhar de humanos adultos. (b) As mesmas formas com suas linhas de simetria (trace- jado) e suas interseções, demonstrando que o comportamento humano pode ser emulado desta forma.	34
4.1	Esquema do neurônio de McCulloch e Pitts.	39
4.2	Esquema de uma Rede Neural Artificial organizada em camadas.	41
4.3	Esquema de uma rede <i>Feedforward</i> (a) e de uma rede <i>Feedback</i>	41
4.4	Exemplo de uma Rede de Kohonen com topologia hexagonal.	43
4.5	Exemplo de uma Rede de Hopfield.	45
4.6	Exemplo hipotético de classificação em que uma reta (superfície de decisão) separa as duas classes de padrões no espaço de características.	48
5.1	Arquitetura geral do protótipo: os retângulos representam os dados e os retângulos arredondados representam os processos.	52

5.2	Arquitetura do módulo de detecção (adaptação do modelo de Itti e colegas).	53
5.3	Exemplo de imagens resultantes do processo de filtragem linear. (a) Imagem de entrada, (b) imagem de intensidade e respectivos canais de cores R (c), G (d), B (e) e Y (f).	55
5.4	Exemplo de uma pirâmide Gaussiana com cinco níveis, gerada a partir de uma imagem de intensidades (Figura 5.3(b)).	57
5.5	Exemplo de uma Pirâmide Direcional com cinco níveis, na orientação 90° . .	61
5.6	Soma dos Mapas de Conspicuidade (a), (b) e (c) para geração do Mapa de Saliência (d).	63
5.7	Exemplo da seleção das regiões de interesse. Mapa de Saliência (a), (b) e (c) e Imagem de entrada (d), (e) e (f). A cada região inibida no Mapa de Saliência, a região correspondente na imagem de entrada também é selecionada. .	65
5.8	Esquema ilustrando a estratégia adotada para gerar as micro-sacadas.	66
6.1	Exemplo do pré-processamento aplicado. (a) Imagem original, (b) aplicação do filtro <i>Gaussian blur</i> , (c) conversão para níveis de cinza e (d) equalização de histograma.	70
6.2	Gráfico das taxas de acerto da Rede Neural.	71
6.3	Gráfico dos valores do SSE para as várias arquiteturas, na milésima época. .	74
6.4	Arquitetura determinada pela estratégia baseada no SSE/Época de treinamento.	76
6.5	Taxas percentuais de localização das placas em todas as imagens, quando consideramos um número K de regiões selecionadas.	79
6.6	Curva de pontos analisados (em termos percentuais) até que uma placa tenha sido selecionada, nas imagens do subconjunto de teste. A linha horizontal representa o percentual médio.	80
6.7	Exemplos de placas utilizadas nos experimentos: classe (1) - placa pare, classe (2) - placa proibido ultrapassar, classe (3) - placa limite de velocidade 60Km, classe (4) - placa curva à direita, classe (5) placa faixa de pedestres e classe (6) placa indicação de lombada.	81
6.8	Exemplo da aplicação da estratégia de micro-sacadas. Como resultado temos 17 imagens para cada ocorrência de placa.	82

6.9	Representação da arquitetura das redes de classificação binária utilizadas nos experimentos.	86
6.10	Taxas de acerto das redes binárias em cada combinação de duas classes, nas duas análises realizadas.	88
6.11	(a) Mapa de Saliência com várias regiões salientes na fronteira entre o céu e a vegetação, resultando em um grande número de regiões vizinhas selecionadas.	90

Lista de Tabelas

6.1	Taxas de acerto da Rede Neural para cada combinação de tamanhos dos conjuntos: T padrões para treinamento e 14-T padrões para teste (em que $T=1,2,\dots,13$).	72
6.2	Valores das somas dos erros quadrados (SSE) das redes, na milésima época do treinamento, em relação ao número de neurônio na camada escondida. Neste exemplo, o conjunto de treinamento é formado por 7 classes.	74
6.3	Codificação das saídas desejadas para 7 classes, utilizando a estratégia <i>winner-takes-all</i>	75
6.4	Pontos de interesse selecionados em cada imagem até que uma placa tenha sido localizada. A imagem 1 aparece duas vezes na tabela por apresentar duas placas. A placa na imagem 3 não foi localizada, já que um ponto mais saliente na sua vizinhança causou sua inibição.	78
6.5	Relação entre os pontos de interesse selecionados e as placas localizadas.	83
6.6	Taxa de acerto da rede monolítica para as sete classes treinadas.	85
6.7	Matriz de confusão para as sete classes de imagens.	85
6.8	Taxas percentuais de acerto das redes neurais binárias para cada combinação de duas classes, a partir da análise por votação.	87
6.9	Taxas percentuais de acerto das redes neurais binárias para cada combinação de duas classes, a partir da análise absoluta.	89
A.1	Imagens de 1 a 9 utilizadas no primeiro experimento.	104
A.2	Imagens de 5 a 12 utilizadas no primeiro experimento.	105
A.3	Imagens de 13 a 15 utilizadas no primeiro experimento.	106
A.4	Imagens utilizadas no treinamento.	107

A.5	Imagens utilizadas nos testes.	108
-----	--	-----

Lista de Abreviaturas

- BP: *Backpropagation*.
- CCD: *Charge Coupled Device*.
- CSC: *Color Structure Code*.
- C-maps: *Conspicuity Maps*.
- DSS: *Driver Support System*.
- DAM: *Distributed Associative Memory*.
- ENIAC: *Eletronic Numerical Integrator and Computer*.
- EDVAC: *Eletronic Discrete Variable Automatic Computer*.
- FPGA: *Field Programmable Gate Arrays*.
- GOLD: *Generic Obstacle and Lane Detection*.
- GPS: *Global Positioning System*.
- IA: *Inteligência Artificial*.
- INNS: *International Neural Networks Society*.
- IEEE: *Institute of Electrical and Electronics Engineers*.
- IPVR: *Institut für parallele und verteilte höchstleistungsrechner*.
- IPM: *Inverse Perspective Mapping*.
- MLP: *Multilayer Perceptron*.

- PLD: *Programmable logic device*.
- RNA: Rede Neural Artificial.
- SNNS: *Stuttgart Neural Network Simulator*.
- SSE: *Sum of Squared Errors*.
- TIP: *Transputer Image Processing*.
- 3D: Três Dimensões.

Lista de Símbolos

- d_j : Distância Euclidiana.
- $\mu_i(t)$: Saída do neurônio i no tempo t (Rede de Hopfield).
- f_h : Função Degrau.
- X_p : Vetor de entradas.
- T_p : Vetor de saídas desejadas.
- $w_{ij}(t)$: Pesos a partir do nodo i até o nodo j no tempo t .
- η : Termo de ganho ou taxa de aprendizagem.
- δ_{pj} : Termo de erro para o padrão p no nodo j .
- R, G, B e Y : Imagens dos canais de cores criados a partir das três cores básicas r, g e b .
- $R(\sigma), G(\sigma), B(\sigma)$ e $Y(\sigma)$: Pirâmides Gaussianas criadas a partir das imagens dos canais de cores.
- I : Imagem de Intensidades.
- $I(\sigma)$: Pirâmide Gaussiana gerada a partir da imagem de intensidades.
- g_l : Imagem g no nível l da Pirâmide Gaussiana.
- w : Padrão de pesos utilizados na geração da Pirâmide Gaussiana.
- $g_{l,n}$: Imagem interpolada que tem o mesmo tamanho da imagem g_{l-1} da Pirâmide Gaussiana.

- $A(\theta)$: Porção angular da decomposição da Pirâmide Direcional.
- $B(\omega)$: Porção Radial da decomposição da Pirâmide Direcional.
- \ominus : Operação de diferença entre duas imagens da Pirâmide Gaussiana.
- $\mathcal{I}(c, v)$: Mapa de Característica construído a partir do contraste de intensidades nas escalas c e v .
- $\mathcal{RG}(c, v)$: Mapa de Característica construído a partir do contraste das cores vermelho e verde nas escalas c e v .
- $\mathcal{BY}(c, v)$: Mapa de Característica construído a partir do contraste das cores azul e amarelo nas escalas c e v .
- $\mathcal{O}(c, v, \theta)$: Mapa de Característica construído a partir de informações de orientação local nas escalas c e v e na orientação θ .
- $\bar{\mathcal{I}}$: Mapa de Conspicuidade para intensidades.
- $\bar{\mathcal{C}}$: Mapa de Conspicuidade para cores.
- $\bar{\mathcal{O}}$: Mapa de Conspicuidade para orientações.
- \oplus : Operação de soma dos Mapas de Características.
- \mathcal{N} : Operador de Normalização dos mapas.
- \mathcal{S} : Entrada final para o Mapa de Saliência.

Capítulo 1

Introdução

Construir máquinas capazes de perceber o ambiente ao seu redor é um dos grandes desafios da Inteligência Artificial (IA). As pesquisas nesta direção envolvem estudos sobre os cinco sentidos do ser humano: visão, audição, tato, paladar e olfato. Entretanto, os sentidos mais estudados são a visão e a audição. Uma característica da IA é o seu caráter multidisciplinar, já que busca inspiração em diversas áreas de conhecimento. Este trabalho de dissertação estuda a percepção do ponto de vista da visão, buscando nos sistemas biológicos a fonte principal de inspiração. Neste contexto, foi implementado um mecanismo de atenção visual para ser utilizado na localização de placas de sinalização, em imagens reais de ruas e estradas. Além disso, foi investigada a possibilidade de se utilizar Redes Neurais para classificar as placas localizadas. Soluções para os problemas da localização e do reconhecimento das placas podem ser tratados como módulos para um Sistema de Apoio ao Motorista.

Este capítulo apresenta motivações para o reconhecimento de objetos baseado em visão e para atenção visual. Além disso, discute alguns aspectos gerais dos Sistemas de Apoio ao Motorista, da detecção e do reconhecimento de placas de sinalização e define os objetivos do trabalho. Por último, apresenta a estrutura e um sumário dos capítulos presentes nesta dissertação.

1.1 Visão e Reconhecimento

Uma atividade essencial para todos os animais é a percepção dos estímulos provenientes do meio ambiente. Dentre as tarefas envolvidas na percepção desses estímulos, o reconheci-

to dos estímulos visuais tem importância fundamental no que diz respeito à sobrevivência. Em animais como os mamíferos, que possuem um sistema visual bem desenvolvido, a maior parte dos estímulos importantes são detectados através da visão que, portanto, desempenha um papel primordial nestes organismos [Gonçalves, 1999]. Uma das características mais marcantes dos sistemas visuais biológicos é a atenção visual, responsável por selecionar as informações mais relevantes do estímulo de entrada. A habilidade de fixar rapidamente a visão em pontos de interesse na cena e reconhecer possíveis presas, predadores ou rivais, é determinante para a perpetuação e evolução das espécies [Itti and Koch, 2001].

A percepção dos estímulos visuais em máquinas é estudada pela Visão Computacional. Podemos definir Visão Computacional como o conjunto de métodos e técnicas que dão suporte aos sistemas computacionais, na análise e interpretação de imagens. No contexto da Visão Computacional, o termo reconhecimento pode ser visto como o processo que classifica os objetos de uma imagem a partir de informações previamente conhecidas (ou aprendidas).

O processo de reconhecimento faz parte de um processo maior denominado análise de imagem. Análise de imagem é o processo de descoberta, identificação e compreensão dos padrões que são relevantes na realização de tarefas baseadas em imagens. Um dos principais objetivos na análise de imagem por computador é dotar a máquina, de alguma forma, de capacidade aproximada à capacidade humana, na análise visual [Gonzalez and Woods, 1992]. Sendo assim, idealmente, um sistema de análise de imagens deve ser capaz de demonstrar um certo grau de inteligência, ou seja: ser capaz de extrair informações importantes dentre o grande número de detalhes irrelevantes; ser capaz de aprender a partir de exemplos, generalizando seus conhecimentos para aplicar em uma nova e diferente circunstância; e ter a habilidade de inferir a partir de informações incompletas.

1.2 Técnicas para Visão

O reconhecimento dos estímulos visuais é uma tarefa central em qualquer sistema de visão. Alguns autores chegam a afirmar que qualquer problema de visão ou percepção em geral pode ser visto como um problema de reconhecimento. Como já foi dito anteriormente, o reconhecimento faz parte do processo de análise de imagem, e segundo Gonzales e Woods [Gonzalez and Woods, 1992], as técnicas em análise de imagem podem ser divididas em

três áreas básicas: (1) aquisição e processamento de baixo nível, com funções que podem ser vistas como reações automáticas, ou seja, reações que não requerem comportamento inteligente; (2) processamento de nível intermediário, com processos de extração e caracterização de componentes em uma imagem; e (3) processamento de alto nível, que envolve os processos de reconhecimento e interpretação. A Figura 1.1 mostra os processos de cada uma dessas áreas.

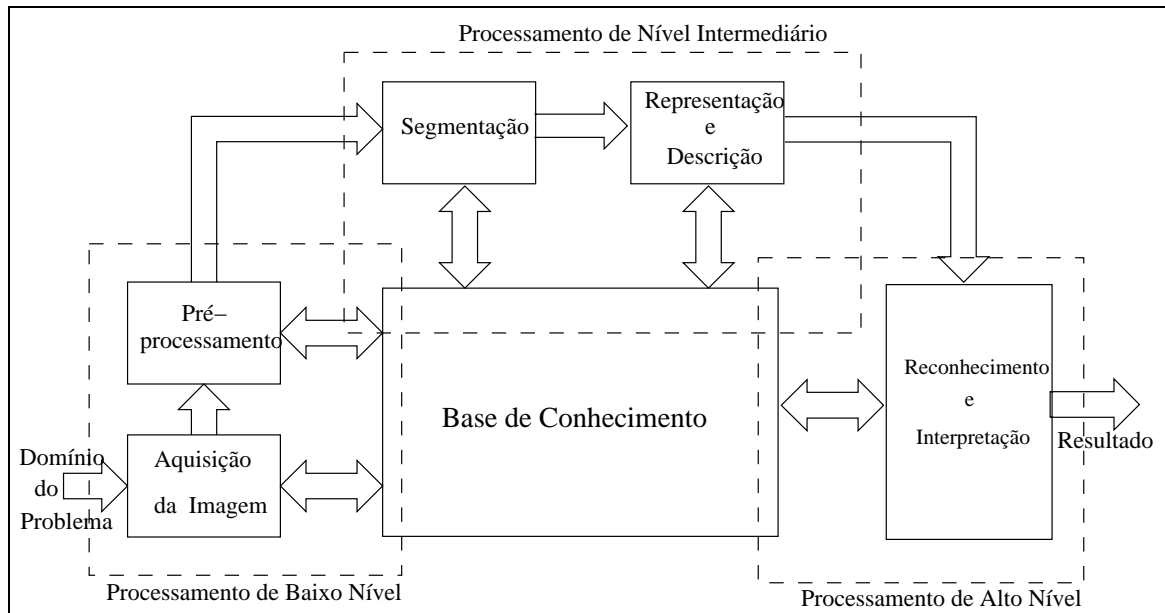


Figura 1.1: Elementos do proceso de análise da imagem. Adaptado de [Gonzalez and Woods, 1992].

1.2.1 Aquisição

O primeiro passo do processo de reconhecimento é a aquisição da imagem. Para tanto, é necessária a utilização de um sensor de luz. Um exemplo de dispositivo bastante utilizado para este fim é a câmera CCD (*Charge Coupled Device*). Este tipo de dispositivo utiliza uma matriz de células fotossensíveis que atuam como capacitores, armazenando carga elétrica na proporção da energia luminosa incidente. Um conjunto de prismas e filtros de cor, decompõem a imagem em seus componentes de cor RGB (*Red, Green e Blue*), onde cada componente é capturado por um CCD independente, gerando a imagem colorida. Como a imagem gerada por esses dispositivos é analógica, torna-se necessária a utilização de um outro dispositivo para converter o sinal analógico em digital, chamado digitalizador.

1.2.2 Pré-processamento

Após a aquisição e digitalização da imagem, o próximo passo é o pré-processamento. A função chave do pré-processamento é melhorar a imagem, com o objetivo de aumentar as chances de sucesso dos processos seguintes [Gonzalez and Woods, 1992]. Nesta etapa, são utilizadas técnicas para aumento de contraste, remoção de ruídos, realce, normalização etc., com o objetivo de converter os padrões para uma forma que possibilite uma simplificação do posterior processo de reconhecimento.

1.2.3 Segmentação

O próximo estágio é o processo chamado de segmentação. De um modo geral, a segmentação particiona uma imagem de entrada em suas partes constituintes ou objetos. Cada uma destas partes é uniforme e homogênea com respeito a algumas propriedades da imagem, como por exemplo cor e textura. Algoritmos de segmentação são geralmente baseados em duas propriedades básicas: descontinuidade e similaridade. Na primeira categoria, o particionamento da imagem é baseado no subconjunto de pontos de um objeto que o separa do restante da imagem. As técnicas de segmentação nesta categoria buscam evidenciar os limites entre os objetos, através da detecção de pontos isolados e da detecção de linhas e bordas na imagem. Na segunda categoria, a segmentação é baseada nas técnicas de limiarização¹, crescimento por regiões, união e divisão de regiões [Gonzalez and Woods, 1992].

1.2.4 Representação e Descrição

Geralmente, a saída do estágio de segmentação são dados brutos de pixel. Neste caso pode ser necessário converter os dados para uma forma conveniente, possibilitando o processamento por computador. Dois tipos de representação podem ser utilizados: representação limite ou representação regional. A representação limite é apropriada quando o foco está em características da forma externa, como por exemplo em cantos. Representação region-

¹O objetivo da limiarização é transformar a imagem de níveis de cinza para binário. O procedimento analisa todos os pixels da imagem comparando seu valor de intensidade com um limiar. Se este valor for acima de um limiar o valor deste pixel deve ser alterado para a cor branca ou preta dependendo do interesse em objetos claros com fundo escuro ou vice-versa

al é apropriada quando o foco está em propriedades internas, tais como textura ou forma esquelética. No entanto, em algumas aplicações estas representações coexistem. Escolher a representação é apenas parte da solução para a transformação de dados brutos em uma forma conveniente para o processamento computacional subsequente. Um método para descrever os dados tal que as características de interesse sejam realçadas, também deve ser utilizado. Descrição, também chamada de seleção de característica, lida com a extração de características que resultam em algumas informações quantitativas de interesse ou que são básicas para diferenciar uma classe de objetos de outra [Gonzalez and Woods, 1992].

1.2.5 Reconhecimento

O último estágio no processo de análise da imagem envolve reconhecimento e interpretação. Reconhecimento é o processo que fixa um rótulo a um objeto baseado na informação fornecida pelos seus descritores. Interpretação envolve a fixação de significado a um grupo de objetos reconhecidos. Para resolver problemas de reconhecimento pode-se partir de três abordagens [Gonzalez and Woods, 1992]:

- Estatística - conjuntos de medidas de características (na forma de n-tupla ou vetores) são extraídos das imagens e métodos estatísticos são utilizados para separar as classes. Dentre os métodos utilizados podemos citar classificadores bayesianos, métodos probabilísticos, regras de decisão etc.;
- Estrutural - padrões são representados em uma forma simbólica (tais como *strings* e árvores), e os métodos de reconhecimento são baseados em casamento de símbolos ou em modelos que tratam padrões de símbolos como sentenças, a partir de uma linguagem artificial;
- Neural - na abordagem neural, como o próprio nome diz, o reconhecimento é realizado utilizando-se Redes Neurais. Alguns autores consideram o reconhecimento via Redes Neurais como sendo um tipo particular de reconhecimento estatístico, já que as características também são na forma de n-tuplas ou vetores e existe uma equivalência entre alguns modelos de Redes Neurais e técnicas estatísticas fundamentais [Bishop, 1995].

Nesta dissertação, o problema do reconhecimento é tratado do ponto de vista da abordagem neural, já que um dos interesses do trabalho foi o estudo de modelos computacionais biologicamente inspirados.

1.2.6 Base de Conhecimento

Em um sistema de análise de imagens, o conhecimento sobre o domínio do problema pode ser codificado na forma de uma base de conhecimento. Este conhecimento pode ser representado tanto na forma de regiões simples quanto regiões detalhadas de uma imagem, onde a informação de interesse a ser localizada é conhecida. Dessa forma, a busca pela informação pode ser reduzida. A base de conhecimento também pode ser bastante complexa, tal como uma base de dados contendo imagens de alta resolução de uma determinada região, obtidas a partir de um satélite, em conexão com aplicações de detecção de mudanças. Além disso, para guiar a operação e cada módulo de processamento, a base de conhecimento também controla a interação entre eles. Isto significa que a comunicação entre os módulos de processamento geralmente é baseada no conhecimento prévio de qual deve ser um resultado. Um exemplo da utilização de conhecimento em tarefas de processamento de imagens, é o pedido repetido, através da base de conhecimento, para o processo de segmentação.

1.3 O Problema do Reconhecimento de Sinais de Tráfego

A crescente evolução tecnológica dos computadores e dispositivos de captura de imagem, a preços acessíveis, tem possibilitado a aplicação de Visão Computacional nas mais diversas áreas. Dentre outras aplicações podemos citar: sistemas de apoio ao diagnóstico médico [Barros et al., 1999], reconhecimento de assinaturas [Gomes et al., 1996], reconhecimento de impressões digitais, reconhecimento de faces [Rowley et al., 1998]. Em particular, este trabalho de dissertação tem como interesse os Sistemas de Apoio ao Motorista (*Driver Support Systems* - DSS).

Dirigir um veículo é uma tarefa que requer um processamento intensivo da informação visual. Relatos mostram que uma grande quantidade de colisões em cruzamentos e choques frontais de veículos poderiam ser evitados se o motorista tivesse meio segundo adicional para reagir, e que a falta de atenção é a causa de muitos acidentes [Little, 2001]. Dessa forma,

a utilização de sistemas que auxiliem o motorista na tarefa de dirigir é uma das áreas de interesse para investigação de técnicas de Visão Computacional.

A principal função de um DSS é auxiliar o motorista principalmente no tocante a conforto e segurança. O sistema pode informar, por exemplo, a presença de animais e pedestres na pista, o limite de velocidade permitido (e controlar a velocidade do veículo, no caso de veículos autônomos), condições anormais da estrada (tais como deslizamentos, buracos etc.), rotas de direção, condições do veículo etc. Portanto, a utilização de sistemas desse tipo pode fornecer ao motorista mais informação do que ele está normalmente acostumado, aumentando efetivamente a sua segurança e a dos passageiros. Algumas das principais tarefas de um DSS são:

- Detecção das marcas da estrada. Por exemplo, as delimitações da pista, cruzamentos, bifurcações;
- Detecção e reconhecimento de sinais de tráfego. Por exemplo, placas de sinalização, semáforos, sinais pintados na pista;
- Detecção de obstáculos. Por exemplo, veículos, pedestres, animais.

Fornecer informações sobre sinalização talvez seja uma das tarefas mais importantes de um DSS. No que diz respeito a segurança de tráfego, a sinalização desempenha um papel fundamental. Na sua grande maioria, as estruturas utilizadas para sinalizar ruas e estradas transmitem informações sobre possíveis riscos como por exemplo, placas que indicam limite de velocidade, placas que indicam a possível presença de animais na pista, indicação de faixa contínua etc. Muitas vezes, os motoristas não respeitam a sinalização por pura desatenção, ou por estarem em situações de tráfego intenso. Em momentos como este, um sistema de detecção e reconhecimento de sinais de tráfego pode funcionar como um co-piloto e fornecer informações que normalmente seriam ignoradas pelo motorista.

Além dos recursos computacionais necessários para a implementação de sistemas desta natureza, é fundamental a utilização de dispositivos sensores que capturem a informação necessária de uma forma eficiente. Os sensores utilizados por um DSS podem ser os mais diversos dependendo da tarefa a ser realizada. Entre outros podemos citar os radares, sensores infra-vermelho, GPS e, no caso de sistemas baseados em visão, câmeras de vídeo.

1.4 Objetivos e Relevância

No âmbito dos Sistemas de Apoio ao Motorista, nosso trabalho está inserido no subproblema da detecção e reconhecimento de sinais de tráfego, mais precisamente placas de sinalização. O Primeiro objetivo foi estudar e implementar um mecanismo de atenção visual *bottom-up* capaz de localizar as placas de sinalização em imagens extraídas de um vídeo previamente adquirido. A atenção visual é uma característica dos sistemas visuais biológicos, que permite aos seres vivos extrair do ambiente apenas as informações visuais mais relevantes. Esta habilidade serve de inspiração no desenvolvimento de mecanismos computacionais capazes de localizar objetos em uma imagem. No âmbito dos Sistemas de Apoio ao motorista, um mecanismo de atenção visual tem a função de localizar, na imagem de entrada, elementos que forneçam alguma informação importante ao motorista.

O segundo objetivo deste trabalho foi investigar a utilização de uma Rede Neural *Multilayer Perceptron* para a tarefa de reconhecimento das placas localizadas pelo mecanismo de atenção. O terceiro objetivo foi avaliar de forma preliminar a integração entre os dois módulos do protótipo. É importante destacar que não era objetivo deste trabalho gerar a integração final entre os módulos. Como objetivo geral, tivemos o estudo da atenção visual tanto do ponto de vista biológico quanto do ponto de vista computacional, visando ampliar as linhas de pesquisa do grupo de Modelos Computacionais e Cognitivos, ao qual este trabalho está vinculado.

Durante a pesquisa não foram descobertos grupos ou pesquisadores no Brasil interessados especificamente no domínio da detecção e reconhecimento de placas de sinalização, o que nos leva a crer que esta é uma área ainda pouco explorada em nosso país. Dessa forma, entendemos que uma das contribuições deste trabalho está na perspectiva de se gerar conhecimentos que auxiliem no desenvolvimento futuro de tecnologias nacionais, para a solução desse e de problemas correlatos. Outra contribuição é a proposta de utilizar um modelo híbrido, formado por um módulo de detecção, que utiliza um mecanismo de atenção biologicamente inspirado, e um módulo de reconhecimento, que utiliza uma Rede Neural, aplicado ao problema da detecção e reconhecimento de placas de sinalização. Os trabalhos na área de Sistemas de Apoio ao Motorista revisados durante o projeto não utilizam mecanismos de atenção visual. Portanto, a implementação de um modelo atencional, sua aplicação

na localização de placas de sinalização em imagens reais e a análise percentual dos resultados alcançados são contribuições importantes para as áreas de atenção visual e DSS.

Devemos destacar ainda o interesse existente por parte de algumas fábricas de automóveis em financiar projetos de pesquisa desta natureza, como foi o caso da parceria entre o Laboratório de Reconhecimento de Imagem da Universidade de *Koblenz-Landau* na Alemanha e a *Daimler-Chrysler* [Priese et al., 1993]. Portanto, a continuidade da linha de pesquisa iniciada neste trabalho pode gerar, no futuro, perspectivas de parceria industrial-científica entre alguma montadora nacional e a Universidade, para o desenvolvimento de um produto comercial.

1.5 Estrutura da Dissertação

A dissertação está organizada conforme a seguinte estrutura: o Capítulo 2 apresenta um estudo sobre Sistemas de Apoio ao Motorista e uma revisão de alguns dos principais trabalhos da área, divididos de acordo com as tarefas a que se propõem a solucionar. Além de trabalhos nas áreas de detecção de obstáculos, detecção e reconhecimento de sinais de tráfego e detecção das marcas da estrada, é apresentada uma revisão de um trabalho que integra algumas tarefas de DSS com o objetivo de construir um veículo autônomo.

O Capítulo 3 trata da atenção visual e está dividido em duas seções principais. Na primeira é apresentado um estudo geral dos aspectos neurofisiológicos do sistema visual humano, com destaque para as estruturas envolvidas com a atenção. Na segunda seção são apresentados os principais conceitos e técnicas utilizadas na implementação de mecanismos de atenção visual em máquinas, com destaque para a abordagem *bottom-up*, além de uma descrição resumida dos trabalhos mais relevantes para a nossa pesquisa.

O Capítulo 4 apresenta um estudo sobre os principais fundamentos das Redes Neurais Artificiais. Apresenta ainda um breve histórico que destaca os fatos mais importantes da área. Discute alguns dos principais modelos de Redes Neurais e seus respectivos algoritmos de treinamento. Além disso, faz uma pequena introdução ao SNNS, simulador de Redes Neurais desenvolvido na Universidade de Stuttgart e que foi utilizado neste trabalho de dissertação.

O Capítulo 5 descreve a arquitetura do modelo proposto, apresentando inicialmente a arquitetura geral e em seguida detalhando os dois principais módulos. O primeiro é o Mó-

dulo de Detecção, que utiliza um mecanismo de atenção visual para selecionar as regiões de interesse na cena. O segundo é o Módulo de Reconhecimento, constituído por uma Rede Neural Artificial, responsável pela tarefa de classificação das regiões selecionadas. O Capítulo 6 descreve os experimentos realizados com um protótipo de implementação da arquitetura proposta e apresenta os resultados desses experimentos. Por fim, o Capítulo 7 apresenta as conclusões e as propostas de trabalhos futuros.

Capítulo 2

Sistemas de Apoio ao Motorista

Este capítulo apresenta um estudo sobre os Sistemas de Apoio ao Motorista (*Driver Support Systems* - DSS), introduzindo alguns conceitos importantes e uma revisão dos principais trabalhos da área. A revisão aborda sistemas de detecção de obstáculos, sistemas de detecção e reconhecimento de sinais de tráfego e sistemas de detecção das marcas da estrada. Além disso, é apresentado um trabalho que integra algumas das tarefas de um DSS com o objetivo de construir um veículo autônomo.

2.1 Introdução

Nas últimas décadas, a área de sistemas de transporte tem dado grande importância a questões como: aumento das condições de segurança, otimização do uso da malha rodoviária, redução no consumo de energia, preservação do meio ambiente com respeito a poluição etc. Esforços em resolver esses problemas têm desencadeado o interesse em novos campos de pesquisa, nas quais diversas técnicas são investigadas para a automação completa ou parcial de tarefas relacionadas com a direção de veículos. Estas tarefas incluem: seguir uma rota mantendo o veículo dentro da pista correta, evitar obstáculos, permanecer a uma distância segura dos outros veículos, regular a velocidade de acordo com as condições de tráfego e características da estrada, encontrar a menor rota para um destino, estacionar dentro de ambientes urbanos, fornecer ao motorista informações sobre sinalização, entre outras. O desenvolvimento de DSS's baseados em visão é um campo de pesquisa multidisciplinar que envolve desde estudos na área de transportes até técnicas computacionais.

Visão é o principal sentido que nós usamos para perceber a estrutura do ambiente que nos cerca. Em virtude da grande quantidade de informação que uma imagem carrega, a visão é uma forma extremamente poderosa para sentir o que está ao redor. Além disto, é uma modalidade sensorial passiva, diferentemente de outras formas de percepção ativa, como radares, lasers, sonares, *bumpers* etc, os quais além de geralmente carregarem uma quantidade inferior de informações, adquirem dados de uma forma invasiva, alterando assim o ambiente e podendo gerar alguma forma de poluição. Devido à importância da informação visual para o motorista humano, torna-se claro o papel eminente da Visão Computacional para projetos de DSS. Dessa forma, Visão Computacional constitui o maior tipo de canal sensorial requerido para oferecer informações úteis e coompreensíveis ao motorista nas suas atividades.

Conceitualmente, um DSS deve comportar-se como um co-piloto humano cooperativo, ou seja, deve ajudar o motorista em atividades de rotina que geralmente provocam desatenção. O objetivo principal de um DSS é aumentar a segurança do tráfego. Para isso, deve fornecer informações que facilitem o raciocínio humano nas decisões tomadas sobre as ações imediatas no trânsito. Visando entender o papel que a Visão Computacional tem neste contexto, podemos identificar algumas das funções de um DSS:

1. O DSS deve **comunicar-se com o motorista** visando aprender sobre suas metas correntes e informá-lo sobre qualquer mensagem que julgar importante. Aqui estão relacionados sistemas de detecção e reconhecimento de sinais de tráfego, detecção de luzes de tráfego (semáforos) etc.;
2. O DSS deve **monitorar a situação de tráfego** no ambiente imediato do veículo, bem como nos próximos segmentos da estrada ao longo da rota desejada. Incluem-se aqui sistemas de detecção de obstáculos (pedestres, veículos, animais etc.);
3. O DSS deve **monitorar o veículo** continuamente e suas reações aos comandos que o motorista transmite quando opera atuadores como por exemplo: o volante e os pedais de freio e aceleração. Visando avaliar apropriadamente tais atividades do motorista, o DSS deve ter a capacidade básica de operar estes e outros atuadores sob certas condições pré-definidas. Estão relacionados com este item os sistemas de direção

automática que dependem de tarefas como: detecção das marcas da estrada, detecção de obstáculos etc.

As pesquisas em DSS realizadas até hoje pelos diversos grupos no mundo têm gerado resultados significativos e, neste aspecto, alguns países se destacam. Podemos citar como países líderes em pesquisa e desenvolvimento de DSS o Canadá, Estados Unidos, França, Japão e principalmente Itália e Alemanha. Um grupo da Universidade de Parma na Itália, alcançou resultados interessantes no desenvolvimento de um veículo autônomo [ARGO, 2001; Bertozzi et al., 1999; Broggi et al., 1999], que foi capaz de trafegar autonomamente ao longo de 2000 km. Alguns detalhes deste trabalho serão apresentados na Seção 2.2.4. No caso da Alemanha, pesquisas resultaram em um produto comercial, o *Mobile Vision System* da empresa *Aglaia GmbH* [AGLAIA, 2002]. Este sistema pode ser instalado em qualquer veículo de passeio e é capaz de reconhecer sinais de tráfego (placas de sinalização), detectar a pista e auxiliar no estacionamento do veículo. Por se tratar de um produto comercial, detalhes técnicos não estão disponíveis.

2.2 Principais Trabalhos em DSS

Nesta seção serão apresentados alguns trabalhos da área, divididos em seções de acordo com a tarefa que o sistema se propõe a realizar, mais especificamente: detecção de obstáculos, detecção e reconhecimento de sinais de tráfego, detecção das marcas da estrada e sistemas integrados.

2.2.1 Detecção de Obstáculos

Na área de detecção de obstáculos, Gavrilin e Philomin [Gavrilin and Philomin, 1999] usaram uma técnica estatística chamada *Simulated Annealing*¹ [Duda et al., 2000] para gerar uma hierarquia de padrões *off-line*, que é usada para fazer a correspondência com padrões desconhecidos. A detecção é feita com base na forma dos objetos, baseada no cálculo de

¹*Simulated Annealing* é uma técnica de otimização estocástica em que durante os estágios iniciais do procedimento de busca, movimentos que aumentam a função objetivo podem ser aceitos. A idéia é fazer exploração suficiente do espaço de busca visando evitar mínimos locais.

uma transformação de distância e o reconhecimento é realizado por uma técnica de agrupamento do tipo *K-means*. O sistema é descrito para uso em tempo real a bordo de veículos de passeio. Para a realização de testes, foram utilizadas imagens em 256 tons de cinza com um tamanho de 360x288 pixels. A implementação foi realizada em um dual-Pentium MMx 450Mhz e alcançou desempenho entre 1 e 5 quadros por segundo na detecção de pedestres. Além disso, a taxa de detecção ficou entre 75% e 85%. Também foram realizados experimentos na detecção de sinais de tráfego, que alcançaram melhores resultados. A taxa de reconhecimento foi de 95% e o desempenho foi entre 10 e 15 quadros por segundo.

Outro trabalho que utilizou um algoritmo de *Simulated Annealing* foi apresentado por Betke e Makris [Betke and Makris, 1995]. O problema do reconhecimento de objetos é tratado como o problema de descrever, da melhor maneira, a correspondência entre um objeto hipotético e uma imagem modelo. Como medida de correspondência é utilizado um coeficiente de correlação normalizado. *Simulated Annealing* é usado para diminuir o tempo de busca com relação a uma busca exaustiva. O algoritmo é aplicado no reconhecimento de marcos por um robô de navegação. Seu desempenho é ilustrado com imagens de cenas complicadas do mundo real, contendo sinais de tráfego que se comportam como os marcos ou obstáculos a serem detectados. O algoritmo apresentado é capaz de reconhecer os sinais em imagens com ruído, que contêm grande conteúdo de informação. Falsas detecções podem ocorrer quando o modelo apresenta pequeno conteúdo de informação.

2.2.2 Detecção e Reconhecimento de Sinais de Tráfego

Piccioli e colegas [Piccioli et al., 1996] desenvolveram um método de detecção e reconhecimento de sinais de tráfego, tanto para imagens em níveis de cinza quanto para imagens coloridas. O método trabalha em três estágios: primeiro, a busca pelos sinais é reduzida a uma região específica da imagem, usando algum conhecimento *a priori* ou indício de cor na cena; segundo, é realizada uma análise geométrica das arestas extraídas da imagem, que gera candidatos a sinais triangulares e circulares; terceiro, um estágio de reconhecimento testa, através de técnicas de correlação cruzada (*cross-correlation*), cada candidato que, se validado, é classificado de acordo com uma base de dados de sinais existente. As imagens utilizadas nos experimentos foram adquiridas a partir de uma câmera montada em um veículo. Experimentos com 600 imagens contendo um ou mais sinais triangulares resultaram em

uma taxa com cerca de 92% de acerto. Uma experimentação extensiva mostrou que o método é robusto quanto à detecção de arestas corrompidas por baixo nível de ruído e trabalha tanto com imagens de ruas urbanas quanto com imagens de estradas rurais e rodovias. Um progresso adicional no esquema de detecção e reconhecimento foi obtido por meio de integração temporal, baseada em métodos de filtragem *Kalman*, das informações extraídas.

A partir de uma cooperação entre a empresa Daimler-Benz AG, a universidade de Paderborn e um grupo da universidade Koblenz-Landau, dentro do projeto *European PROMETHEUS*, foi desenvolvido um sistema com o objetivo específico de reconhecer sinais de tráfego. O grupo Paderborn implementou a análise da imagem em níveis de cinza, a Daimler-Benz AG implementou a identificação de pictogramas² e o grupo Koblenz-Landau implementou análise da imagem colorida [Priese et al., 1993]. Esta última abordagem é baseada em uma segmentação com base na informação de cor, o CSC (*Color Structure Code*). Este código foi desenvolvido pelo próprio grupo e o resultado desta segmentação é uma estrutura de dados hierárquica. Num trabalho posterior foi acrescentado um sistema de controle *Fuzzy* para a tarefa final de decidir se um objeto é ou não um sinal de tráfego [Priese et al., 1993]. Para um comportamento em tempo real foram testados diferentes componentes de hardware (C 40, Motorola PC601, T 805) em um sistema TIP (*Transputer Image Processing*). Usando um processador Motorola PC601 o sistema alcançou uma taxa de reconhecimento de 98%. Um protótipo foi instalado em um veículo de testes da Daimler-Benz AG, tendo por objetivo servir como ferramenta de apoio ao motorista [Rehrmann et al., 1995].

2.2.3 Detecção das Marcas da Estrada

Na área de detecção das marcas da estrada, Broggi [Broggi, 1995a; Broggi, 1995b] utilizou transformação geométrica e processamento morfológico para desenvolver um sistema capaz de detectar as marcas, mesmo em condições de sombra extremamente severas. Para a implementação do algoritmo assumiu-se uma rodovia vazia e estruturada, e um conjunto completo

²Pictograma - símbolo gráfico cuja própria forma expressa o significado, que deve ser entendido imediatamente e internacionalmente sem prévio conhecimento. Diferente de um símbolo puro, um pictograma é baseado na forma do objeto que ele representa. Placas indicando travessia de pedestres, possível presença de animais na pista, são exemplos da utilização de pictogramas.

de parâmetros de aquisição conhecidos (posição da câmera, orientação, parâmetros ópticos etc.). O uso de uma arquitetura de hardware massivamente paralela (PAPRICA) [Broggi et al., 1994] permitiu alcançar uma taxa de processamento, em tempo real, de aproximadamente 17 frames por segundo. Essa abordagem é baseada em duas suposições: existência de uma pista plana e marcas da estrada visíveis. No caso de pista irregular, a detecção só é possível se for assumida uma largura fixa. No caso de marcas obstruídas por outros veículos, é utilizado um processo de reorganização a partir do emparelhamento de imagens estéreo. Entretanto, as duas imagens só serão idênticas se novamente a pista for plana. Neste caso, a diferença entre as duas imagens pode ser utilizada para detectar obstáculos.

2.2.4 Sistemas Integrados

Os trabalhos acima descritos dizem respeito, em sua maioria, a módulos específicos de um DSS. É fácil notar que esses módulos podem ser integrados com o objetivo de implementar um sistema móvel autônomo. Um projeto com essas características foi desenvolvido na Universidade de Parma, Itália [ARGO, 2001; Bertozzi et al., 1999; Broggi et al., 1999].

O protótipo desse projeto, denominado Veículo Autônomo ARGO, consiste de um veículo de passeio *Lancia Thema* equipado com um sistema baseado em visão chamado GOLD (*Generic Obstacle and Lane Detection*), que funciona como piloto automático do veículo. O sistema permite extrair informações do ambiente e da rodovia a partir de cenas adquiridas. Através de visão estéreo, obstáculos na rodovia são detectados e localizados, enquanto que o processamento de uma única imagem monocular permite extrair a geometria da rodovia na frente do veículo. Por ser genérica, a técnica permite detectar obstáculos sem limitações de forma, cor ou simetria e detectar marcas na pista em condições severas de sombra e até mesmo no escuro. As imagens são adquiridas a partir de um sistema de visão estereoscópica, que consiste de duas câmeras sincronizadas capazes de adquirir pares de imagem em níveis de cinza. O resultado do processamento (posição dos obstáculos e geometria da rodovia) é usado para guiar um atuador no volante e informações depuradas são apresentadas para o motorista em um monitor e um painel de controle a bordo do veículo.

No processamento das imagens foi introduzida uma transformação geométrica chamada IPM (*Inverse Perspective Mapping*). O ângulo de visão através do qual a imagem é adquirida e o efeito perspectiva contribuem para associar conteúdo de informação diferente a cada

pixel da imagem. O uso do IPM permite remover o efeito perspectiva da imagem adquirida, remapeando-a em um novo domínio bidimensional em que o conteúdo da informação está distribuído de forma homogênea entre todos os pixels. A aplicação da transformação IPM requer o conhecimento das condições específicas de aquisição (posição da câmera, parâmetros ópticos, orientação etc.) e algum conhecimento *a priori*.

A tarefa de detecção da rodovia é reduzida à tarefa de detectar as marcas pintadas na pista. O conhecimento *a priori* é a suposição de uma rodovia plana na frente do veículo. A vantagem do uso da transformação IPM é que, na imagem remapeada, a largura das marcas da pista é quase invariante. Isto simplifica o passo de detecção seguinte e permite sua implementação com uma técnica de casamento de padrões tradicional. A detecção é baseada na busca por padrões horizontais escuro-claro-escuro com um determinado tamanho, já que as marcas são representadas, na imagem remapeada, por linhas claras quase verticais, com largura constante e rodeadas por um fundo mais escuro. Desde que o modelo assumido para o ambiente externo (rodovia plana) permita determinar o relacionamento espacial entre pixels da imagem e o mundo 3D, é possível derivar a geometria da rodovia e a posição do veículo dentro da pista. Foi observado que em duas situações a detecção de pista falha:

1. Quando a rodovia possui uma curvatura e uma das marcas não é visível - o remapeamento produzido é incompleto;
2. Quando a rodovia não é plana - o remapeamento produzido é deformado.

A tarefa de detecção de obstáculos é tratada como uma mera localização de objetos que podem obstruir o caminho do veículo, para isso IPM estéreo é usado em conjunto com um modelo geométrico da rodovia. Devido aos diferentes ângulos de visão das câmeras estéreo, um obstáculo quadrado homogêneo ideal produz dois grupos de pixels com uma forma triangular, que correspondem as suas arestas verticais. O processo de detecção é baseado na localização desses triângulos. Uma complicação neste processo é a possibilidade da presença de dois ou mais obstáculos na frente do veículo ao mesmo tempo, produzindo mais de um par de triângulos, como também a possibilidade da presença de obstáculos parcialmente visíveis, produzindo um único triângulo. Para localizar os triângulos é usado um histograma polar [Borenstein and Koren, 1991]. O histograma apresenta picos que correspondem a cada triângulo e a posição do pico determina o ângulo de visão sob o qual a aresta do obstáculo é

vista. A distância do obstáculo pode ser estimada através de uma análise adicional da diferença da imagem ao longo das direções ressaltadas pela máxima do histograma polar, visando detectar pontas do triângulo. Alguns pontos críticos são notados na detecção de obstáculos, o que deixa claro que a confiança na detecção depende do tamanho, distância e forma dos obstáculos. Os pontos críticos, ou seja, situações em que o sistema pode falhar, são:

- Quando o obstáculo está muito longe das câmeras (geralmente entre 45 e 50 m);
- Quando o obstáculo está próximo do *guard-rail* e um único grande obstáculo é detectado;
- Quando um obstáculo é parcialmente visível, assim apenas uma de suas arestas pode ser detectada;
- A detecção de obstáculos distantes algumas vezes falha quando seu brilho é similar ao brilho da rodovia.

Visando testar o veículo sob diferentes situações de tráfego, ambiente de rodovia e condições de tempo, uma excursão de 2000 Km foi realizada em junho de 1998. Durante este teste, o veículo ARGO dirigiu-se autonomamente ao longo da rede de estradas e auto-estradas da Itália, e o sistema foi posto a prova também em estradas rurais estruturadas. Utilizando-se um processador *Pentium MMX* (200 MHz) o sistema alcançou desempenho de 5.1ms para detecção de obstáculos e 4.6ms para detecção de pista. Testes não oficiais mostraram que o sistema detectou corretamente a posição da pista em 95% das situações consideradas.

2.2.5 Aspectos Importantes

Além de trabalhos que tratam especificamente de módulos de um DSS, são encontrados na literatura trabalhos que discutem aspectos que podem ser determinantes para um desempenho eficiente dos sistemas.

No trabalho de Nayar e Narasimhan [Nayar and Narasimhan, 1999] são discutidas questões relativas às condições de tempo (névoa, chuva, granizo, neve etc.) em sistemas que operam ao ar livre. Primeiro são estudadas as manifestações visuais das diferentes condições

de tempo. Para isso é identificado o que já se conhece sobre fenômenos ópticos da atmosfera. Depois, são estudados os efeitos causados pelo mau tempo que podem se tornar vantagens. Baseados nessas observações, desenvolveram métodos para recuperar propriedades da cena em imagens obtidas sob condições de mau tempo. Este trabalho é apenas uma tentativa inicial de entender e explorar as manifestações do tempo.

No trabalho de Salgin e Ballard [Salgian and Ballard, 1998] são descritas algumas rotinas visuais baseadas em modelos de cores e formas, para serem utilizadas em veículos autônomos. Além disso, são discutidas questões cruciais que envolvem o planejamento de tais rotinas. As rotinas descritas no trabalho são, detecção de luzes de semáforo, detecção de placas pare (*stop sign*), detecção de interseções, detecção de veículos, obstáculos e pista. Testes foram realizados tanto com imagens geradas a partir da direção do veículo em um mundo simulado, quanto com imagens geradas a partir da direção no mundo real. Um aspecto interessante do trabalho é a discussão em torno da relação entre a detecção dos objetos e a velocidade do veículo. O tempo entre o aparecimento do objeto no vídeo e a resposta do sistema à sua detecção é utilizado para calcular a taxa de sincronização. Esta taxa decresce à medida que a velocidade do veículo aumenta, dificultando assim a detecção.

2.3 Sumário

Neste Capítulo, estudamos algumas questões gerais relacionadas aos sistemas de apoio ao motorista (DSS) baseados em Visão Computacional, com o objetivo de contextualizar o nosso trabalho. Além disso, foi apresentada uma revisão de alguns dos principais trabalhos da área. Em sua maioria, os trabalhos apresentados utilizam técnicas com pouca ou nenhuma inspiração biológica. Isso nos motiva a investigar a utilidade de técnicas que tenham esta inspiração, com o objetivo de gerar novos conhecimentos que forneçam suporte para trabalhos futuros. Em geral, os trabalhos apresentam bons resultados, muito embora sejam apenas protótipos e talvez estejam longe de se tornarem comuns no nosso cotidiano.

As discussões em torno do lançamento de produtos para o mercado envolvem questões polêmicas como legislação e questões técnicas como ruas e estradas bem conservadas, bem sinalizadas etc. Com relação à legislação, podemos fazer a seguinte pergunta: no caso de acidentes causados/influenciados por falhas no sistema (no caso de veículos autônomos), de

quem seria a responsabilidade? Analisando as questões discutidas neste capítulo, podemos observar que os sistemas dependem essencialmente de ruas e estradas bem conservadas e bem sinalizadas, sob pena de ter seu desempenho reduzido. Podemos tomar como exemplo o trabalho de Broggi [Broggi, 1995a; Broggi, 1995b] que depende da existência das marcas na estrada para conseguir determinar a rota e detectar objetos. Na ausência das marcas o sistema perde totalmente seu poder. Entretanto, o desenvolvimento de DSS's aliado a uma boa infraestrutura de ruas e estradas, pode resultar na redução dos índices de acidentes e consequentemente de mortes no trânsito, justificando plenamente investimentos nessa área.

Capítulo 3

Atenção Visual

Este capítulo está dividido em duas seções principais. Na primeira seção é apresentado um estudo geral de alguns aspectos neurofisiológicos do sistema visual humano, destacando as principais estruturas neurais que estão ligadas diretamente com a atenção visual. Na segunda seção, os principais conceitos e técnicas utilizados na implementação de atenção visual em máquinas são introduzidos.

3.1 Atenção em Sistemas Biológicos

Atenção Visual é a capacidade que os sistemas visuais biológicos têm de detectar rapidamente partes interessantes no estímulo visual de entrada [Milanese et al., 1994]. Pode ser vista como um método para reduzir a quantidade de informação visual de entrada para um tamanho manejável, de tal forma que tarefas com processamento complexo possam ser tratadas pelos recursos computacionais limitados do cérebro [Koch, 2000]. Estudos psicofísicos tentam revelar as consequências comportamentais da atenção, já estudos neurofisiológicos tentam revelar os mecanismos neurais e as áreas do cérebro envolvidas na atenção. Com base nesses estudos alguns modelos computacionais foram propostos. Nas subseções seguintes trataremos de alguns aspectos neurofisiológicos do sistema visual humano, envolvidos na atenção.

3.1.1 Sistema Visual Humano

Os recentes progressos da neurociência têm fornecido uma idéia clara da estrutura e organização das regiões do cérebro que colaboram com as funções visuais. Especialização, modularidade e organização hierárquica parecem ser princípios importantes que ajudam a entender o funcionamento desse sistema paralelo extremamente complexo [Milanese, 1993]. A informação visual captada pelos fotorreceptores da retina é transmitida através do nervo óptico e alcança dois centros no cérebro: o núcleo geniculado lateral, também chamado de corpo geniculado lateral e que é parte do tálamo, e o colículo superior.

O primeiro caminho, chamado de retino-geniculado, é o mais estudado por ser o único que alcança o córtex visual (Figura 3.1), onde aproximadamente 90% da informação visual é processada. O córtex visual é uma estrutura organizada hierarquicamente em áreas, cada qual especializada em um aspecto específico da informação visual, tais como disparidade espacial, temporal, cromática e binocular. O segundo caminho, chamado de colicular, embora muitas vezes esquecido em descrições gerais do sistema visual humano, é particularmente importante no que diz respeito aos movimentos oculares. No sistema visual humano, características visuais são computadas na retina, colículo superior, corpo geniculado lateral e em áreas corticais visuais [Guyton and Hall, 1997].

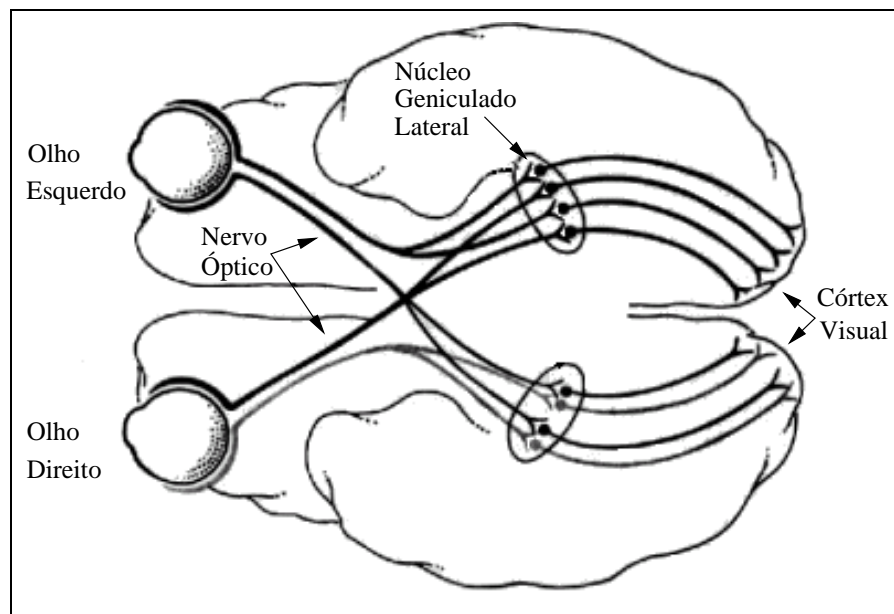


Figura 3.1: Esquema do cérebro que mostra o caminho retino-geniculado. Adaptado de [Shepherd, 1994].

A Retina

A retina é a porção do olho sensível à luz e possui dois tipos de células fotossensíveis: os cones e os bastonetes. Os cones são responsáveis pela visão em cores e estão associados à visão diurna, já os bastonetes são responsáveis pela visão no escuro, ou seja, visão noturna. Quando essas células são excitadas, sinais são transmitidos por todos os neurônios sucessivos da própria retina e, finalmente, para as fibras nervosas ópticas do córtex cerebral [Churchland and Sejnowski, 1992]. A informação capturada pelos fotorreceptores é processada por várias camadas de células retinais, principalmente representadas por células bipolares e células ganglionares. Estas células realizam primeiro um processamento espacial, fornecendo dois canais (*On* e *Off*) de contraste de luz, depois realizam processamento temporal, detectando gradientes temporais [Milanese, 1993].

Uma característica importante da retina é a distribuição não uniforme de fotorreceptores. Existe uma região diminuta no centro da retina chamada de fóvea (Figura 3.2). A fóvea ocupa uma área total de pouco mais de 1mm^2 , sendo especialmente responsável pela visão acurada e detalhada. A porção central, chamada de fóvea central, possui diâmetro de apenas $0,3\text{mm}$ e é composta inteiramente por cones, com densidade espacial máxima [Guyton and Hall, 1997]. Ao redor da fóvea, a densidade decresce de forma radial [Milanese, 1993].

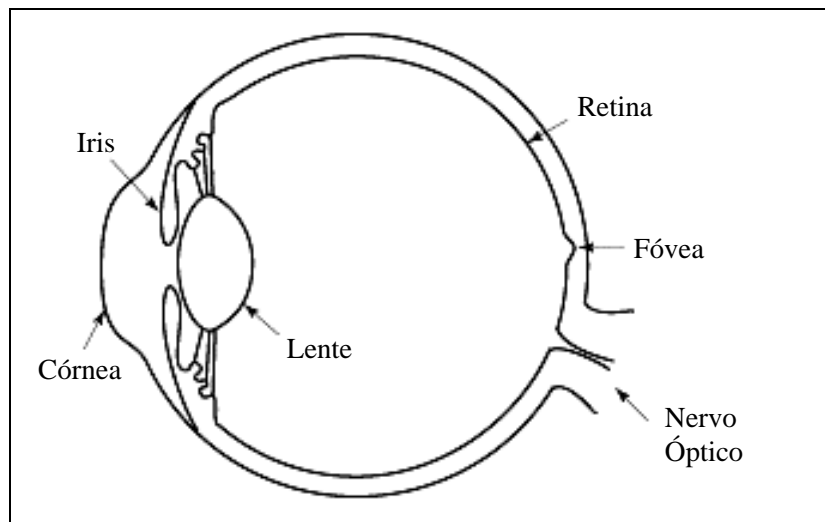


Figura 3.2: Esquema do olho humano que mostra a localização da fóvea na retina.

Movimentos Oculares

Talvez os movimentos oculares mais importantes sejam aqueles que fazem os olhos se fixarem sobre uma porção discreta do campo visual [Guyton and Hall, 1997]. Esses movimentos são controlados por dois mecanismos neuronais. O primeiro, chamado de mecanismo de fixação voluntária, permite que os olhos se movam voluntariamente para encontrar o objeto sobre o qual se quer fixar a visão. O segundo, chamado de mecanismo de fixação involuntária, mantém os olhos firmemente sobre o objeto depois de encontrado.

Um pequeno campo cortical localizado bilateralmente nas regiões corticais pré-motoras dos lobos frontais, é responsável pelo controle dos movimentos de fixação voluntária. Uma pessoa com disfunção ou destruição destas áreas tem dificuldade ou torna-se incapaz de desligar os olhos de um ponto de fixação e depois movê-los para outro ponto. Áreas visuais secundárias do córtex occipital são responsáveis por controlar os movimentos de fixação involuntária. Quando essas áreas são destruídas, o indivíduo tem dificuldade em manter os olhos dirigidos para um determinado ponto de fixação ou torna-se incapaz de fazê-lo. A este movimento de fixação também dá-se o nome de sacada.

A principal razão biológica para a existência dos movimentos oculares é a necessidade de localizar a fóvea nas regiões mais importantes, ou mais informativas de uma cena visual, tendo em vista que na fóvea há maior densidade de fotorreceptores [Milanese, 1993]. Quando uma cena visual está se movendo continuamente diante dos olhos, quando uma pessoa está num carro em movimento, ou quando uma análise mais detalhada de um objeto é necessária (por exemplo ao se prestar atenção involuntariamente às características faciais, a fim de reconhecer uma pessoa), os olhos se fixam em um ponto importante após o outro no campo visual, saltando de um para o outro duas ou três vezes por segundo. Estes saltos são chamados de micro-sacadas. Os movimentos sacádicos são tão rápidos que não mais de 10% do tempo total necessário para a fixação do olhar é gasto no movimento propriamente dito, com os 90% restantes dedicados aos sítios de fixação (como por exemplo, focalizando o objeto. Algumas outras características dos movimentos sacádicos são:

- Atraso: o tempo entre o começo do estímulo até a preparação da ação motora varia entre 200 e 300 ms;
- Duração: depende da amplitude da sacada (geralmente menor que 15 graus) e varia de

10 ms (1 grau) até 70 ms (20 graus);

- Laço aberto: depois de iniciada, a trajetória de uma sacada não pode ser modificada;
- Efeito global: quando uma sacada é produzida por dois estímulos vizinhos, ela aporta em algum lugar no meio, e não em um deles.

Vias Visuais

Depois que os impulsos nervosos abandonam as retinas, dirigem-se para trás pelos nervos ópticos. No quiasma óptico, todas as fibras das metades nasais da retina cruzam para o lado oposto, onde se juntam às fibras das metades temporais da retina do outro lado para formar os tratos ópticos [Shepherd, 1994]. As fibras de cada trato óptico fazem sinapse no corpo geniculado lateral e, a partir daí, as fibras geniculocalcarinas vão para o córtex visual primário por meio das radiações ópticas.

Além de alcançar o córtex visual primário, as fibras visuais se encaminham para áreas mais antigas do cérebro: (1) para o núcleo supraquiasmático, provavelmente para o controle dos ritmos circadianos¹; (2) para os núcleos pré-tectais, para provocar os movimentos reflexos dos olhos na focalização de objetos importantes e para a ativação do reflexo pupilar à luz; (3) para o colículo superior, para o controle dos movimentos direcionais rápidos dos olhos; (4) para o núcleo geniculado lateral ventral do tálamo e, depois, para as regiões basais circunjacentes do cérebro, provavelmente para ajudar a controlar algumas das funções comportamentais corporais [Guyton and Hall, 1997].

Dessa forma, as vias visuais podem ser divididas em um sistema novo, para a transmissão direta ao córtex visual, e um sistema antigo, para transmissão a áreas mais antigas do cérebro. O sistema novo é responsável, no homem, pela percepção de virtualmente todos os aspectos das formas visuais, cor e outros da visão consciente [Guyton and Hall, 1997].

O Tálamo e o Corpo Geniculado Lateral

O tálamo é organizado em vários núcleos, alguns dos quais estão conectados diretamente ao córtex, e está localizado no centro do cérebro. O maior núcleo do tálamo é o pulvinar,

¹Ritmos circadianos são processos fisiológicos e comportamentais como por exemplo, dormir/acordar, digestão, secreção hormonal, etc., que oscilam em torno de um período de 24 horas.

que tem conexões recíprocas com todas as áreas corticais que preservam o mapeamento retinotópico, mas não tem entrada direta do nervo óptico. As conexões com o córtex formam vários mapas retinotópicos, alguns dos quais são ligados a áreas corticais únicas, e alguns outros projetam mais de uma área.

O núcleo mais estudado é o corpo geniculado lateral [Milanese, 1993], que está localizado na extremidade dorsal do tálamo. Ele representa um estágio de transmissão intermediário entre a retina (através do nervo óptico) e o córtex visual. As células do corpo geniculado lateral têm correspondência 1:1 com células ganglionares da retina, e são organizadas respeitando a topologia retinal. As fibras do nervo óptico, do sistema visual novo, terminam todas no corpo geniculado lateral. As duas principais funções do corpo geniculado lateral são: (1) retransmitir informações visuais do trato óptico para o córtex visual por meio das radiações ópticas: essa função de retransmissão é precisa, tanto que há uma transmissão ex-ata ponto a ponto com um alto grau de fidelidade espacial por todo trajeto da retina ao córtex visual; (2) servir de portal à transmissão de sinais para o córtex visual: isto é, controlar a quantidade de sinais que têm permissão para chegar até o córtex.

O Córtex Visual

O córtex visual é a parte do córtex que responde aos estímulos visuais, e está localizado primariamente nos lobos occipitais [Shepherd, 1994]. É dividido em córtex visual primário e áreas visuais secundárias. O córtex visual primário é a região onde terminam os sinais visuais diretos a partir dos olhos. A estrutura organizacional do córtex visual se dá em vários milhões de colunas verticais de células neuronais, onde cada coluna representa uma unidade funcional. A mesma organização colunar é encontrada por todo córtex [Shepherd, 1994]. Pode-se calcular, aproximadamente, que o número de neurônios em cada uma das colunas verticais visuais seja de 1.000, ou talvez mais [Guyton and Hall, 1997].

Misturadas entre as colunas visuais primárias bem como entre as colunas de algumas das áreas visuais secundárias, existem áreas especiais, semelhantes a colunas, chamadas de manchas de cor. Elas recebem sinais laterais das colunas visuais adjacentes e respondem especificamente a sinais coloridos. Logo, presume-se que estas manchas sejam as áreas primárias responsáveis por decifrar a cor [Guyton and Hall, 1997].

Após deixar o córtex visual primário, a informação visual é analisada em duas vias prin-

cipais nas áreas visuais secundárias. Uma das vias analisa as posições tridimensionais dos objetos visuais no espaço. Desta informação esta via também analisa a forma global da cena visual, bem como o movimento na cena, isto é, ela diz onde está cada objeto a cada instante e se este está se movendo. A outra via é responsável pela análise dos detalhes visuais. Porções separadas desta via também analisam especificamente as cores. Portanto, esta via está implicada em tarefas visuais como o reconhecimento de letras, a leitura, a determinação da textura das superfícies, a determinação das cores detalhadas dos objetos e na interpretação, a partir de todas essas informações, a identidade do objeto e o seu significado.

Sumário

Nesta seção, estudamos algumas estruturas importantes do sistema visual humano, envolvidas na atenção visual. Discutiremos a seguir algumas técnicas utilizadas na maioria dos modelos computacionais de atenção visual biologicamente inspirados.

3.2 Atenção Visual em Máquinas

Processos atencionais podem ser considerados mecanismos de alocação de recursos de processamento a uma dada região da cena visual, permitindo o processamento desta região com maior eficiência, enquanto outras regiões são deixadas em segundo plano ou ignoradas [Gonçalves, 1999]. Muitas evidências sugerem que mecanismos atencionais são necessários para realizar com sucesso várias tarefas de visão. Para assegurar tratabilidade, sistemas de visão computacional devem localizar e analisar apenas a informação essencial da tarefa corrente e ignorar o vasto fluxo de detalhes irrelevantes [Culhane and Tsotsos, 1992]. Tsotsos [Tsotsos, 1990] analisou a complexidade computacional da busca visual, confirmando que atenção visual seletiva é uma contribuição maior na redução da quantidade de computação em qualquer sistema de visão.

Tem surgido recentemente a idéia de um *framework* com dois tipos de componentes para desenvolvimento atencional [Itti and Koch, 2001]. Este framework sugere que indivíduos direcionam a atenção para objetos em uma cena usando, tanto indícios *bottom-up*, quanto indícios *top-down*. Atenção *bottom-up* baseia-se em informações da imagem, com alto grau de paralelismo, rapidez e independe de esforço consciente, também chamada de processos

pré-atencionais. Atenção *top-down* é baseada em um conjunto de modelos armazenados; baseia-se na memória ou tem natureza cognitiva, dependem do contexto ou da tarefa e operam de forma sequencial, mais lenta e com esforço consciente. Pouco se sabe sobre as instâncias neurais dos componentes da atenção *top-down* e este aspecto da atenção visual não tem sido modelado em seus detalhes [Itti and Koch, 2001]. Dessa forma, concentraremos nossos estudos primeiramente nos mecanismos de atenção *bottom-up*. Segundo Itti e Koch [Itti and Koch, 2001] cinco importantes tendências, que enfatizam o controle atencional *bottom-up*, têm emergido em recentes trabalhos:

1. A percepção da saliência do estímulo de entrada depende criticamente do contexto ao seu redor;
2. Um único “Mapa de Saliência”, que codifica topograficamente estímulos eminentes na cena visual, tem provado ser uma estratégia de controle *bottom-up* plausível e eficiente;
3. Inibição do retorno: o processo que impede que uma região tratada anteriormente seja tratada novamente, é um elemento crucial no desenvolvimento atencional;
4. A forte interação entre atenção e movimento dos olhos tem sido um dos desafios computacionais no que diz respeito ao sistema de coordenadas utilizado para o controle da atenção;
5. A integração entre atenção e reconhecimento de objetos limita vigorosamente a seleção das regiões para uma análise mais detalhada, otimizando o processo.

Percepções a partir dessas cinco áreas fornecem um *framework* para o entendimento computacional e neurobiológico da atenção visual. Nas seções seguintes discutiremos algumas dessas questões.

3.2.1 Computação de Características Visuais Primitivas

O primeiro estágio de processamento em qualquer modelo de atenção *bottom-up* é a computação de características visuais primitivas. Estas características são computadas pré-atencionalmente de uma maneira massivamente paralela através de todo campo visual [Itti and Koch, 2001]. Processos visuais primitivos computam propriedades elementares das

imagens, tais como: brilho, textura, cor, fluxo de movimento e disparidade estéreo. O gradiente dessas grandezas é então usado para segmentar a imagem. Na maioria dos sistemas de atenção visual, esta segmentação produz vários mapas de características, que são combinados dando origem a um mapa de saliência.

3.2.2 Representação Piramidal

Uma forma bastante utilizada para representar características visuais primitivas em diferentes escalas é a representação piramidal. Uma pirâmide de uma imagem pode ser vista como uma sequência de cópias desta imagem, em que a densidade e a resolução são reduzidas em cada nível. O mais baixo nível da pirâmide (nível 0) é a imagem original e cada nível é obtido a partir do nível anterior. A Pirâmide Gaussiana [Burt and Adelson, 1983] é o tipo mais comum de representação piramidal. Esta representação é formada por versões filtradas passa-baixa de convolução Gaussiana da imagem de entrada. Um filtro passa-baixa atenua as altas frequências espaciais de uma imagem e acentua as baixas frequências. A representação piramidal é utilizada com o objetivo de obter amostras da imagem onde detalhes indesejados são suprimidos, ruídos são eliminados, características grosseiras são realçadas etc. No Capítulo 4 é apresentado um algoritmo clássico para geração da Pirâmide Gaussiana.

3.2.3 Mapa de Saliência

A maioria dos modelos de atenção *bottom-up* segue a hipótese de Koch e Ullman [Koch and Ullman, 1985], onde vários mapas de características alimentam um único mapa mestre ou mapa de saliência. O mapa de saliência é um mapa escalar bi-dimensional cuja atividade representa topograficamente a saliência visual [Itti and Koch, 2001]. Uma região ativa em um mapa de saliência codifica o fato desta região ser saliente, não importando se ela corresponde, por exemplo, a uma bola vermelha no meio de bolas verdes, ou se corresponde a um objeto se movendo para a esquerda enquanto outros se movem para a direita.

O mapa de saliência codifica um conjunto de medidas de saliência que não é dependente de nenhuma dimensão de característica particular, fornecendo, deste modo, uma estratégia de controle eficiente para focalizar a atenção em regiões salientes na cena visual [Itti and Koch, 2001]. Uma região na cena visual é definida como saliente se ela vence a competição

espacial em uma ou mais dimensões de características dentre várias escalas espaciais. Uma dificuldade em combinar diferentes mapas de características é que eles não representam, a priori, modalidades comparáveis, com escalas variadas e mecanismos de extração diferentes [Itti et al., 1998]. Outra dificuldade se deve ao fato de que, como todos os mapas são combinados, objetos salientes que aparecem em um pequeno número de mapas podem ser mascarados por ruído ou objetos menos salientes podem estar presentes em um número maior de mapas [Itti et al., 1998].

3.2.4 Inibição do Retorno

Uma rede neural do tipo *winner-takes-all* é uma arquitetura neural plausível para descobrir a localização mais saliente no mapa de saliência, uma vez que é capaz de determinar um ponto de interesse representado por um neurônio vencedor [Itti and Koch, 2001]. Porém, o uso deste mecanismo causa outro problema computacional: como prevenir que a atenção esteja voltada permanentemente para a região mais ativa no mapa de saliência. Uma estratégia computacional eficiente, que recebeu apoio experimental, consiste em inibir brevemente neurônios no mapa de saliência na região atualmente assistida. Após a região atualmente assistida ser suprimida, a rede *winner-takes-all* converge naturalmente para a próxima região mais saliente e, ao repetir este processo, geram-se movimentos sacádicos, os quais definem os caminhos atencionais [Tsotsos et al., 1995].

3.2.5 Alguns Modelos de Atenção *Bottom-up*

Como citado acima, os modelos de atenção *bottom-up*, em sua maioria, baseiam-se em um mapa de saliência. O que diferencia esses modelos é a estratégia usada para “podar” o estímulo de entrada e extrair a saliência. A seguir, é apresentada uma revisão de alguns trabalhos na área de atenção visual. São apresentados apenas os trabalhos que consideramos mais importantes para esta dissertação.

Tsotsos e Culhane [Culhane and Tsotsos, 1992] propuseram um modelo composto de uma hierarquia de processamento e um “raio de atenção” que guia a seleção das regiões de maior interesse. O raio atravessa a hierarquia, passando através das regiões de maior interesse e inibindo as regiões que não são relevantes. No nível mais baixo, o protótipo

compreende uma representação hierárquica do estímulo visual de entrada. Cada unidade da hierarquia computa uma resposta de soma de pesos a partir de suas entradas no nível abaixo. Uma zona inibida e uma zona de passagem são delineadas por um raio que se destaca através de todos os níveis da hierarquia (Figura 3.3). A zona de passagem atravessa o vencedor em cada nível e a zona inibida cerca esses elementos, que competem em um processo *winner-takes-all*. Embora trabalhe com a idéia de saliência, o modelo não utiliza um mapa de saliência para representar a entrada visual. É utilizada uma representação hierárquica ou piramidal do estímulo de entrada, baseada nas bordas orientadas da imagem.

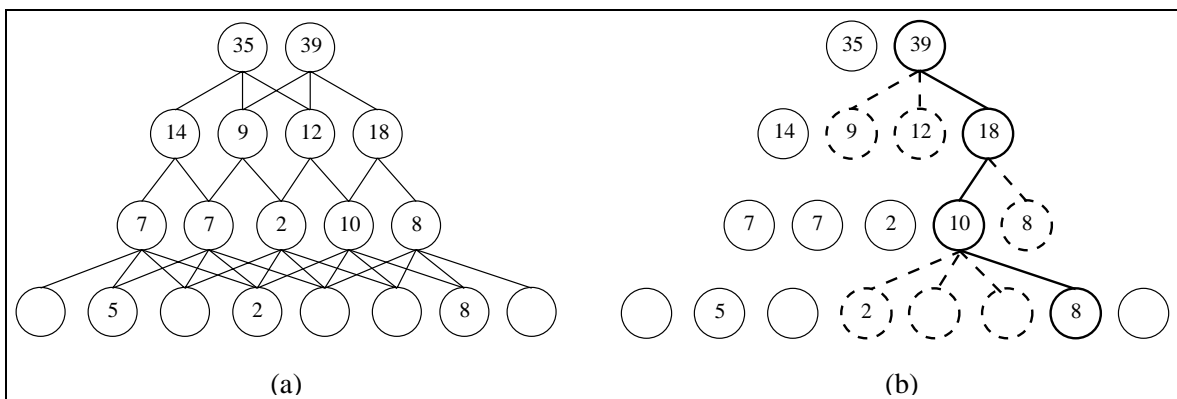


Figura 3.3: Exemplo da hierarquia de processamento do modelo de Tsotsos e Culhane: (a) configuração inicial; (b) seleção dos itens de maior interesse - zona de passagem (linhas sólidas) e zona inibida (linhas tracejadas). Adaptado de [Culhane and Tsotsos, 1992].

A Figura 3.4 mostra uma visão geral da arquitetura proposta por Milanese e colegas [Milanese et al., 1994], que utiliza uma estratégia híbrida, integrando indícios *bottom-up* e *top-down*. O modelo considera tanto imagens estáticas quanto seqüências de imagens. No caso de imagens estáticas, o subsistema bottom-up analisa a estrutura de cor RGB corrente e extrai a saliência. Isto é feito em dois estágios: no primeiro estágio são extraídos mapas de características (orientação, curvatura, contraste de cor) e um número correspondente de mapas de conspicuidade (*conspicuity maps* - *C-maps*), que realça regiões de pixels amplamente diferentes das regiões ao seu redor; o segundo estágio é representado por um processo de integração que une os *C-maps* em um simples mapa de saliência. Isto é obtido através de um processo de *relaxation*, que modifica os valores dos *C-maps* até que seja identificado um pequeno número de regiões de interesse.

Uma fonte de informação adicional baseada em reconhecimento é gerada pelo subsistema

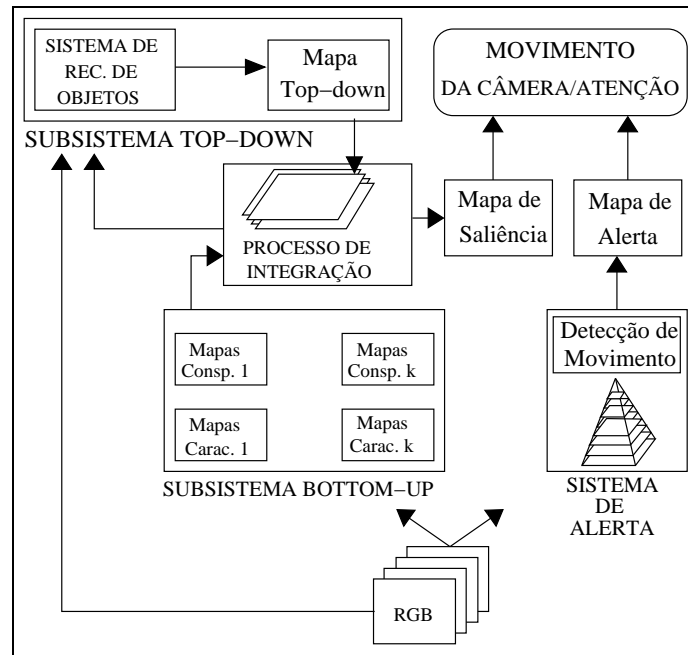


Figura 3.4: Visão geral da arquitetura proposta por Milanese e colegas. Adaptado de [Milanese et al., 1994].

top-down. Um módulo de reconhecimento baseado em uma Memória Associativa Distribuída (*Distributed Associative Memory* - DAM) é usado para detectar regiões da imagem que casam com alguns modelos armazenados. A saída da DAM, chamada de mapa de atenção *top-down*, representa uma entrada adicional para o processo de relaxamento que define o mapa de saliência. No caso da variação de tempo, a seqüência de imagens é analisada por um subsistema de alerta, que usa uma representação piramidal do estímulo de entrada para fornecer uma detecção aproximada de objetos em movimento contra um background estático. Este caminho é normalmente ineficaz, até que um objeto eventualmente entre no campo de visão. Quando isto ocorre o subsistema toma o controle sobre o resto do sistema e diretamente provoca um movimento atencional.

Um outro modelo de atenção visual baseado em saliência (Figura 3.5) e inspirado no comportamento e na arquitetura neuronal do sistema visual dos primatas foi proposto por Itti e colegas [Itti et al., 1998]. Neste modelo, a entrada visual é decomposta em um conjunto de mapas de características (por exemplo cores, intensidades, orientações etc.) e as diferentes regiões espaciais competem pela saliência dentro de cada mapa. Cada característica é computada por um conjunto de operações lineares do tipo centro-vizinhança (*center-surround*)

semelhantes a campos receptivos visuais.

As operações centro-vizinhança são implementadas como diferenças entre escalas finas e grossas. O centro é um pixel na escala $c = \{2, 3, 4\}$ e a região ao redor é o pixel correspondente na escala $s = c + \delta$, com $\delta = \{3, 4\}$. Um processo de normalização juntamente com combinações entre escalas dos mapas de características, produzem, para cada característica, um mapa de conspicuidade. A diferença entre as escalas de dois mapas é obtida através de uma interpolação das escalas mais grossas para a escala mais fina e de uma subtração ponto-a-ponto entre as escalas.

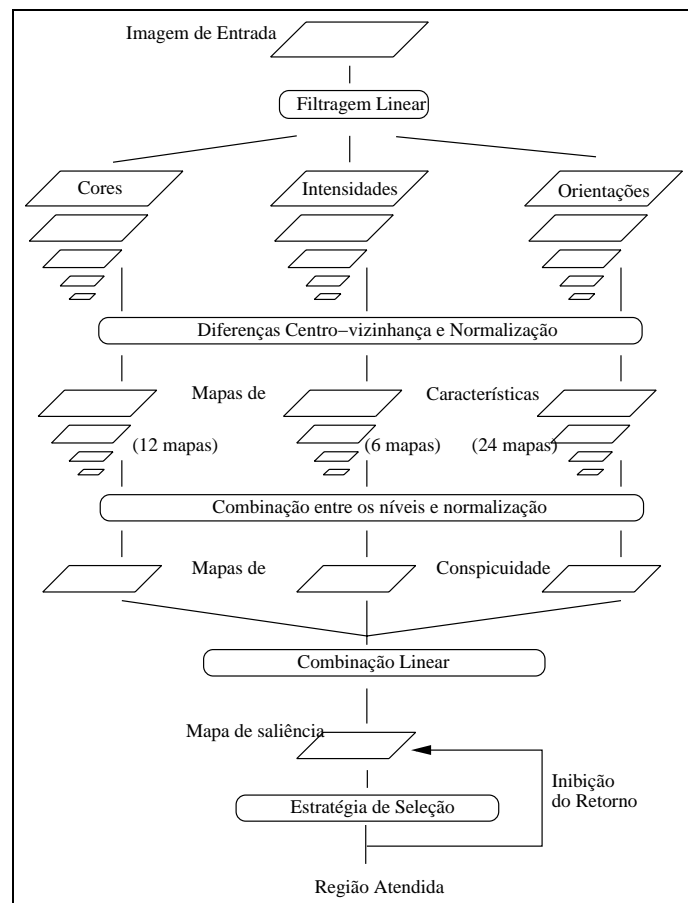


Figura 3.5: Arquitetura geral do modelo proposto por Itti e colegas. Adaptado de [Itti et al., 1998].

Os mapas de conspicuidade são normalizados e somados, produzindo uma entrada final para o mapa de saliência, o qual codifica a saliência local sobre a cena visual inteira. O mapa de saliência do modelo é dotado de dinâmicas internas que provocam movimentos atencionais. Em qualquer instante de tempo, uma rede neural *winner-takes-all* seleciona as

regiões mais ativas no mapa de saliência e atrai o foco atencional para a região mais saliente. Subseqüentemente, a região selecionada é inibida no mapa de saliência, de forma tal que o sistema desloca o foco para a próxima região mais saliente, otimizando assim o processo de busca.

No trabalho de Sela e Levine [Sela and Levine, 1997] é apresentado um sistema de atenção visual *bottom-up* para guiar a fixação de um sensor retinal móvel (baseado na estrutura não uniforme da retina). Os pontos de fixação, ou pontos de interesse, são definidos como os centros de regiões simetricamente cercadas, com base nos contornos da imagem. Esses pontos são modelados como as interseções das linhas de simetria em uma imagem, como mostra a Figura 3.6. Para encontrar as linhas de simetria e suas orientações, é utilizada uma medida de simetria baseada nas regiões centrais das arestas co-circulares. Duas arestas são ditas co-circulares se um círculo pode ser desenhado de tal forma que as arestas sejam tangentes. O centro da co-circularidade é definido como o ponto central deste círculo, da mesma forma, o raio da co-circularidade é definido como o raio do círculo.

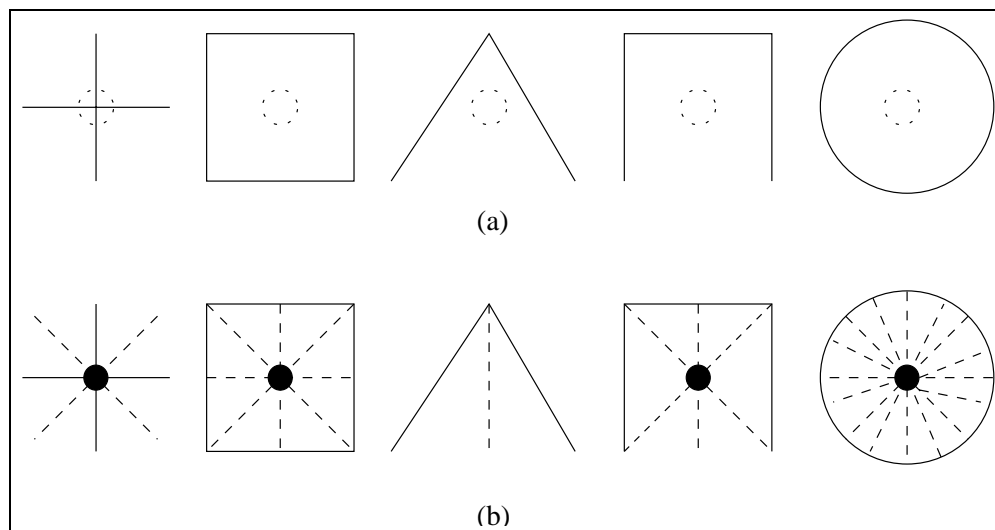


Figura 3.6: Tendências de fixação utilizadas pelo modelo de Sela e Levine. (a) Resultados (centros de fixação) de experimentos examinando a fixação do olhar de humanos adultos. (b) As mesmas formas com suas linhas de simetria (tracejado) e suas interseções, demonstrando que o comportamento humano pode ser emulado desta forma. Adaptado de [Sela and Levine, 1997].

Para a validação desta abordagem, os pontos de interseção das linhas de simetria foram comparados com as descobertas psicofísicas sobre a atenção humana. Para cada possível

ponto de fixação é atribuída uma magnitude. Este valor depende de dois fatores:

1. Os ângulos que separam as arestas que contribuem com as linhas de simetria - arestas que formam ângulos agudos atraem mais a fixação do que arestas que formam ângulo obtusos;
2. O grau de “fechamento” de cada ponto - isto afeta o significado perceptivo do ponto.

Os pontos de alta saliência que são fechados também são agrupados, e um simples ponto central é determinado para a fixação. Todas essas funções são realizadas sobre o campo visual inteiro, em cada *frame*. O algoritmo computa pontos de interesse tanto para uma imagem foveal mapeada em coordenadas retangulares, quanto para uma imagem periférica mapeada em coordenadas *log-polar*. O algoritmo foi implementado em uma rede paralela de processadores (Texas Instruments TMS320C40). Esta configuração facilita um desempenho próximo ao tempo real, em imagens que cobrem um grande campo de visão. O algoritmo foi testado no reconhecimento de faces e com imagens de cenas ao ar livre.

3.3 Sumário

Neste capítulo foi apresentado um estudo sobre a atenção visual, tanto do ponto de vista biológico quanto do ponto de vista computacional. No que diz respeito aos aspectos biológicos, o texto destaca as principais estruturas responsáveis pela atenção, como retina, núcleo geniculado lateral e córtex visual e colículo superior. Quanto aos aspectos computacionais, a atenção visual é estudada do ponto de vista da abordagem *bottom-up* (atenção baseada em saliência), seguindo principalmente a idéia de Koch e Ullman [Koch and Ullman, 1985] que propõem a utilização de um mapa de saliência para codificar as regiões de interesse na cena visual. A principal razão para a escolha da abordagem *bottom-up* em detrimento da abordagem *top-down* é o pouco conhecimento que se tem sobre as estruturas neurais responsáveis pelo controle atencional *top-down*.

Analizando os modelos de atenção visual estudados neste capítulo, nós decidimos investigar a possibilidade de adaptar o modelo de Itti e colegas [Itti et al., 1998] para formar o módulo de detecção do nosso sistema, que será discutido em mais detalhes no próximo capítulo. A primeira razão para esta escolha é o nosso interesse particular na utilização, sempre

que possível, de modelos e técnicas biologicamente inspirados. Partindo deste princípio, nós observamos que o modelo escolhido segue os princípios do sistema visual dos primatas, na implementação de todos os processos. Além disso, os autores acima apresentam alguns experimentos rudimentares utilizando imagens contendo placas de sinalização, que embora fossem bem comportadas (com poucas estruturas e *background* homogêneo, diferentemente das nossas) nos deram uma pista inicial de que seria possível localizar as placas nas nossas imagens com esta abordagem. Embora os outros trabalhos também apresentem inspiração biológica em certos aspectos, os critérios utilizados para seleção das regiões de interesse aliados aos dados utilizados nos experimentos, influenciaram na escolha do modelo do Itti, uma vez que este foi inserido no problema da localização de placas de sinalização, enquanto os outros modelos não trataram deste aspecto especificamente. Entretanto, uma comparação entre os modelos no domínio do problema investigado pode trazer resultados importantes. Por fugir do escopo do trabalho, esta atividade será apresentada como proposta de trabalho futuro no Capítulo 7.

Capítulo 4

Redes Neurais

Este capítulo apresenta um estudo sobre os principais fundamentos das Redes Neurais Artificiais. Apresenta ainda um breve histórico que destaca os fatos mais importantes da área. Discute alguns dos principais modelos de Redes Neurais e seus respectivos algoritmos de treinamento. Além disso, faz uma pequena introdução ao SNNS, simulador de Redes Neurais desenvolvido na Universidade de Stuttgart e que foi utilizado neste trabalho de dissertação.

4.1 Fundamentos de Redes Neurais

Redes Neurais Artificiais são técnicas computacionais que propõem um modelo matemático baseado na estrutura neural de organismos inteligentes, mais especificamente o cérebro humano [Tafner et al., 1995]. Segundo Haykin [Haykin, 1999], Uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso. A principal característica das Redes neurais é a capacidade de aprender a partir de exemplos e, assim, classificar novos padrões. Desde que ressurgiram no início da década de 80, Redes Neurais têm sido aplicadas com sucesso a uma vasta gama de problemas. Dentre as aplicações que utilizam Redes Neurais podemos citar: reconhecimento de objetos [Tu and Li, 1999], mineração de dados [Craven and Shavlik, 1997], reconhecimento de fala [Yuk and Flanagan, 1999], robótica [de A. Barreto et al., 2001] etc.

Desde o começo, os trabalhos em Redes Neurais têm sido motivados pelo reconheci-

to de que o cérebro humano processa informações de forma mais eficiente, se comparado com os computadores convencionais. O cérebro é um sistema de processamento de informação altamente **complexo, não-linear e paralelo**, que tem a capacidade de organizar seus componentes estruturais (neurônios) de forma a realizar certos tipos de processamento muito mais rapidamente que o mais rápido computador digital hoje existente [Haykin, 1999]. Uma rede neural se assemelha ao cérebro em dois aspectos [Haykin, 1999]:

1. O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem;
2. Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

O primeiro modelo de um neurônio artificial foi proposto pelo neurofisiologista Warren McCulloch e pelo matemático Walter Pitts em 1943 [McCulloch and Pitts, 1943]. O trabalho de McCulloch e Pitts descreve um cálculo lógico das Redes Neurais que unifica os estudos de neurofisiologia e da lógica matemática. A Figura 4.1 mostra um esquema do neurônio artificial proposto por eles. A operação dessa unidade de processamento pode ser resumida da seguinte forma:

- Sinais são apresentados às entradas (X_1, X_2, \dots, X_P);
- Cada sinal é multiplicado por um número, ou peso, que indica sua influência na saída da unidade (W_1, W_2, \dots, W_p);
- É realizada uma soma ponderada dos sinais, produzindo um nível de atividade (\sum);
- Se este nível de atividade exceder um certo limite (threshold) a unidade produz uma determinada resposta de saída (Y). Senão o neurônio permanece inativo.

As entradas do neurônio artificial podem ser comparadas com os estímulos do neurônio natural. Os pesos são valores que representam o grau de importância que cada entrada possui em relação àquele determinado neurônio. A função de ativação é uma função de ordem interna, cuja atribuição é decidir o que fazer com o valor resultante do somatório das entradas ponderadas.

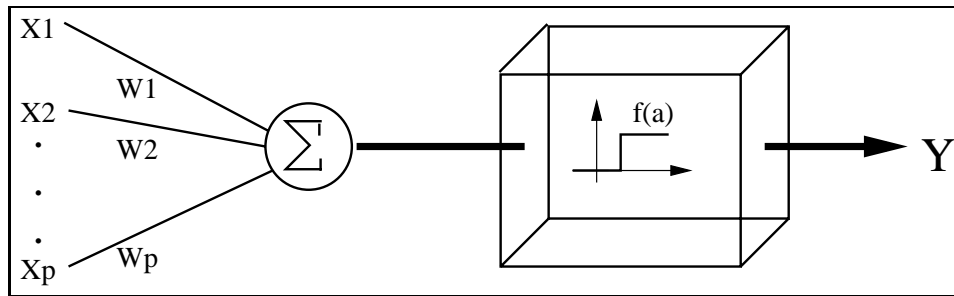


Figura 4.1: Esquema do neurônio de McCulloch e Pitts.

4.2 Notas Históricas

Como vimos na seção anterior, o primeiro modelo de neurônio artificial foi apresentado em 1943 pelos pesquisadores Warren McCulloch e Walter Pitts. O artigo de McCulloch e Pitts teve bastante repercussão no meio científico, tendo influenciado von Neumann a usar chaves de atraso idealizadas, derivadas deste modelo de neurônio artificial, na construção do ED-VAC (*Electronic Discrete Variable Automatic Computer*) que foi desenvolvido a partir do ENIAC (*Electronic Numerical Integrator and Computer*). O ENIAC foi o primeiro computador eletrônico de propósito geral.

Em 1949 Donald Hebb escreveu um livro intitulado “*The Organization of Behavior*”, suas idéias não eram completamente novas, mas Hebb foi o primeiro a propor uma lei de aprendizagem específica para as sinapses dos neurônios. Também proveniente deste período foi a construção do primeiro neuro computador, denominado *Snark*, por Marvin Minsky, em 1951. Embora nunca tenha executado qualquer função de processamento de informação interessante, o *Snark* serviu de inspiração para trabalhos futuros.

Em 1956, durante a 1ª Conferência Internacional de Inteligência Artificial, foi apresentado um modelo de rede neural artificial pelo pesquisador Nathaniel Rochester, da IBM. Seu trabalho consistia numa simulação de centenas de neurônios interconectados, através da construção de uma sistema para verificar como a rede responderia aos estímulos ambientais. O primeiro neuro computador a obter sucesso (*Mark I Perceptron*) surgiu em 1957/58, criado por Frank Rosenblatt, Charles Wightman e outros. Em 1969, Marvin Minsky e Seymour Papert publicaram um livro chamado “*PERCEPTRON*”, onde resumiram e criticaram seriamente a pesquisa sobre Redes Neurais. Devido a importância e o respeito que a comunidade científica tinha por Minsky e Papert, uma retração nos investimentos e programas de

pesquisa para essa tecnologia foi inevitável.

Um período de pesquisa silenciosa seguiu-se, quando poucos resultados foram publicados. Até que em 1982, o físico e biólogo do Instituto de Tecnologia da Califórnia John Hopfield deu um novo impulso às Redes Neurais, contestando, com sucesso, as teses matemáticas de Minsky e Papert. Com o trabalho de Hopfield e com a criação do algoritmo *Backpropagation* para o treinamento de redes com múltiplas camadas em 1986, as Redes Neurais ganharam novamente credibilidade. Houve, em termos de publicação, uma avalanche de trabalhos, deixando claro que as pesquisas nunca pararam, pois nem todos esses trabalhos poderiam ter sido produzidos em tão pouco tempo.

Em 1987 ocorreu em São Francisco a primeira conferência de Redes Neurais em tempos modernos, a IEEE International Conference on Neural Networks, e também foi formada a International Neural Networks Society (INNS). A partir destes acontecimentos decorreram a fundação do INNS Journal em 1989, seguido do Neural Computation e do IEEE Transactions on Neural Networks em 1990. Desde então, muitas universidades anunciaram a formação de institutos de pesquisa e programas de educação em neuro computação.

4.3 Arquiteturas

A maioria das arquiteturas neurais são tipicamente organizadas em camadas, como mostra a Figura 4.2, onde suas unidades podem estar conectadas tanto às unidades das camadas posteriores quanto das camadas anteriores. Normalmente as camadas são classificadas em três grupos:

Camada de Entrada: a única função dessa camada é receber os padrões de entrada e repassá-los à camada seguinte;

Camadas Escondidas: onde a maior parte do processamento é feita. Também chamadas extratoras de características;

Camada de Saída: onde o resultado final é concluído e apresentado.

As conexões entre as camadas podem gerar várias estruturas diferentes (arquiteturas). Do ponto de vista dessas conexões, a arquitetura da rede pode ser caracterizada por dois aspectos, ilustrados na Figura 4.3:

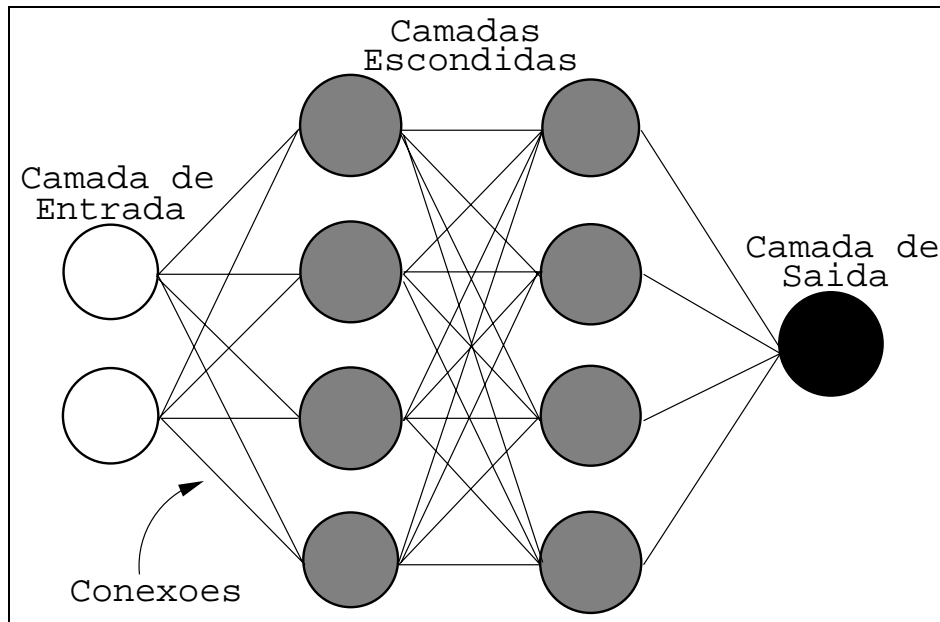


Figura 4.2: Esquema de uma Rede Neural Artificial organizada em camadas.

- *Feedforward*: quando o sinal de saída de um neurônio não é utilizado como entrada para os neurônios das camadas anteriores, ou para os neurônios da própria camada;
- *Feedback*: quando o sinal de saída de um neurônio serve de entrada para neurônios da mesma camada, ou de camadas anteriores.

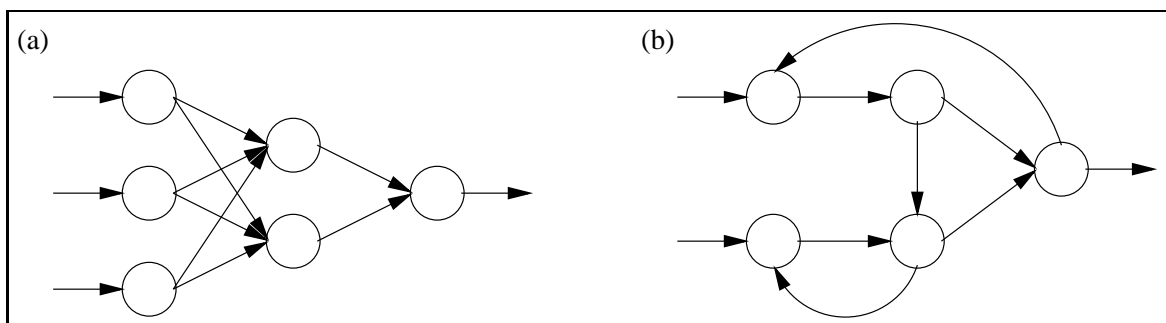


Figura 4.3: Esquema de uma rede *Feedforward* (a) e de uma rede *Feedback*.

4.4 Processos de Aprendizagem

A característica mais importante das Redes Neurais é a capacidade de aprender a partir de exemplos. A aprendizagem ocorre quando a rede neural atinge uma solução generalizada para

uma classe de problemas. Segundo Haykin [Haykin, 1999], a aprendizagem é um processo pelo qual os parâmetros livres de uma rede neural são adaptados através de um processo de estimulação pelo ambiente no qual a rede está inserida.

O conjunto de regras bem definidas para a solução de um problema de aprendizagem é denominado algoritmo de aprendizagem. Existem vários tipos de algoritmo de aprendizagem específicos para determinados modelos de Redes Neurais, que diferem entre si principalmente pelo modo como os pesos são modificados. Outro fator importante é a maneira pela qual uma rede neural se relaciona com o ambiente. Os principais paradigmas de aprendizagem são:

Aprendizagem supervisionada: um agente externo é utilizado para indicar à rede qual a resposta desejada para o padrão de entrada;

Aprendizagem não Supervisionada (auto-organização): não existe agente externo indicando a resposta desejada para os padrões de entrada, é também conhecida como aprendizagem auto-supervisionada;

Nas próximas seções serão estudados alguns dos principais modelos de redes neurais e seus algoritmos de aprendizagem.

4.5 Alguns Modelos de Redes Neurais

Vários são os modelos de Redes Neurais já publicados em revistas especializadas. Dentre esses modelos podemos destacar como clássicos: Kohonen, Hopfield e *Perceptron*. Os principais aspectos que diferem os diversos tipos de Redes Neurais são: O tipo de conexão entre os neurônios, o número de camadas da rede e o algoritmo de treinamento utilizado. A seguir serão apresentados os aspectos principais dos modelos Kohonen, Hopfield e *Multilayer Perceptron*.

4.5.1 Rede de Kohonen

O modelo de Kohonen é uma rede neural do tipo *feedforward* de treinamento não supervisionado. Ao ser considerado uma rede de duas dimensões, o modelo de Kohonen não impõe

nenhuma forma topológica, que pode ser triangular, retangular, hexagonal etc. A Figura 4.4 mostra um exemplo de uma Rede de Kohonen com topologia hexagonal.

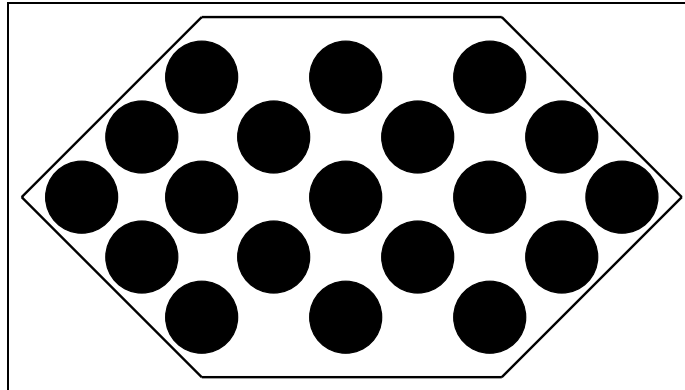


Figura 4.4: Exemplo de uma Rede de Kohonen com topologia hexagonal.

O esquema básico deste modelo tem a propriedade de modificar a si próprio. Os neurônios da camada competem entre si para serem os vencedores a cada modificação. O neurônio vencedor é aquele que gerar a menor distância Euclidiana entre o vetor de pesos e o vetor de entrada. Cada neurônio da rede representa uma saída. Todas as entradas são conectadas a todos os neurônios da rede. Os pesos são inicializados com valores aleatórios baixos. Uma entrada X é apresentada à rede sem especificar a saída desejada. De acordo com a entrada, um neurônio Y deverá responder melhor e este será o vencedor. Os pesos do neurônio vencedor e de seus vizinhos serão ajustados. Após todo o conjunto de treinamento ter sido apresentado à rede e os critérios de treinamento terem sido satisfeitos, a rede é considerada treinada. Na fase de teste, um conjunto de entradas é apresentado à rede sem haver alterações nos pesos. O vetor de entrada X é um conjunto ordenado de sinais valorados que devem possuir uma inter-relação. Em alguns casos é necessário realizar um pré-processamento nos dados. O algoritmo de aprendizagem é apresentado abaixo [Beale and Jackson, 1990]:

1. Inicializar os pesos com valores aleatórios baixo sem relação com os valores de entrada;
2. Setar o raio de vizinhança de cada neurônio (inicialmente pode ser igual ao tamanho da rede);
3. Apresentar uma entrada à rede $x_0(t), x_1(t), \dots, x_{n-1}(t)$, em que $x_i(t)$ é a entrada para o neurônio i no tempo t ;

4. Calcular a distância Euclidiana d_j entre a entrada e os pesos para cada neurônio de saída j ;

$$d_j = \sum_{i=0}^{n-1} (x_i(t) - w_{ij}(t))^2 \quad (4.1)$$

5. Selecionar o neurônio com menor distância Euclidiana como neurônio vencedor;
6. Atualizar os pesos do neurônio vencedor e seus vizinhos, definidos pelo tamanho da vizinhança $N_j(t)$;

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t)) \quad (4.2)$$

em que $j \in N_j(t)$; $0 \leq i \leq n-1$

O termo $\eta(t)$ é um termo de ganho ($0 < \eta(t) < 1$) que diminui com o tempo, fazendo a adaptação de pesos lentamente.

7. Se necessário, modificar o raio de vizinhança de todos os neurônios (decréscimo do raio de $N(t)$;
8. Repetir o passo 3.

Este paradigma é baseado na teoria de que as células nervosas corticais estão arranjadas anatomicamente em relação aos estímulos que recebem dos sensores às quais estão ligadas. O comportamento da rede tem como objetivo fazer com que a rede simule uma atividade cerebral [Tafner et al., 1995]. Cada neurônio representa uma saída da rede. Isto significa que, se a rede possuir 10 neurônios, conseqüentemente, haverá 10 saídas possíveis para qualquer número de entradas que a rede estiver sendo submetida. Além disso, cada neurônio está amplamente conectado com as entradas.

4.5.2 Rede de Hopfield

A Rede de Hopfield é uma rede autoassociativa, que tem algumas similaridades com as redes perceptron, porém com algumas diferenças importantes. Este modelo consiste de um número de neurônios totalmente conectados, ou seja, cada neurônio está conectado a todos os outros, como mostra a Figura 4.5.

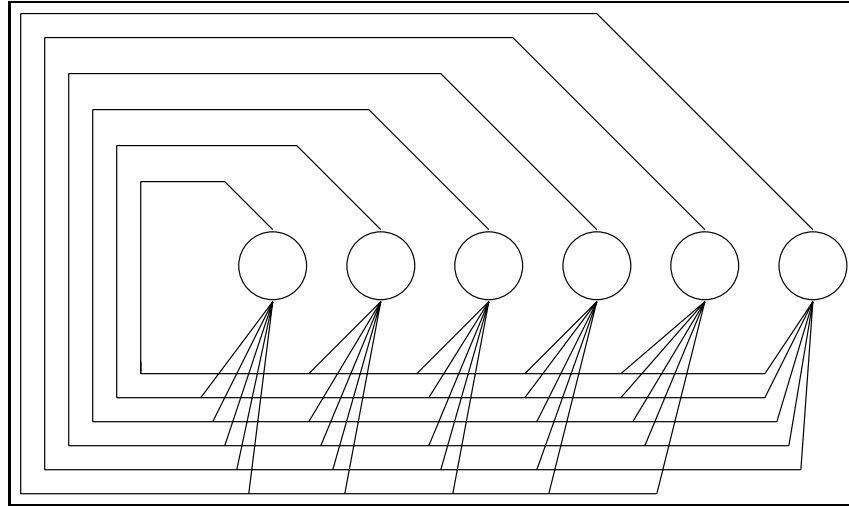


Figura 4.5: Exemplo de uma Rede de Hopfield.

A Rede de Hopfield também tem como característica importante o fato de ser uma rede simetricamente "pesada", isto é, os pesos das conexões de um neurônio para outro são os mesmos em ambas as direções. Cada neurônio tem um limiar e uma função-progresso, e o neurônio calcula a soma dos pesos de entrada e subtrai do valor limiar, passando este resultado para a função progresso que determina o estado de saída. A rede aceita somente dois estados de entrada, que pode ser binário (0, 1) ou bipolar (-1, +1). A principal característica da Rede de Hopfield é que não existem conexões de entrada ou saída óbvias, ou seja, todos os neurônios são iguais. Isto significa que a rede opera de uma forma diferente. As entradas da rede são aplicadas para todos os neurônios de uma só vez. A rede procede em ciclo através de uma sucessão de estados, até convergir em uma solução estável, que ocorre quando os valores dos neurônios não alteram muito. Como cada neurônio está conectado a todos os outros, o valor de um neurônio afeta o valor de todos os outros. Nas Redes de Hopfield, a primeira saída é tomada como a nova entrada, que produz uma nova saída, e assim sucessivamente. A solução ocorre quando não há mudanças significativas de um ciclo para o outro. O algoritmo de treinamento é apresentado abaixo [Beale and Jackson, 1990]:

1. Determinar os pesos das conexões:

$$w_{ij} = \begin{cases} \sum_{s=0}^{M-1} x_i^s x_j^s & i \neq j \\ 0 & i = j, 0 \leq i, j \leq M - 1. \end{cases} \quad (4.3)$$

em que w_{ij} é o peso da conexão entre o neurônio i e o neurônio j , e x_i^s é o elemento

i padrão exemplar para a classe s , e é +1 ou -1. No total existem M padrões, de 0 até $M - 1$. Os limiares dos neurônios são iguais a zero.

2. Inicializar com padrões desconhecidos:

$$\mu_i(0) = x_i \quad 0 \leq i \leq N - 1 \quad (4.4)$$

em que $\mu_i(t)$ é a saída do neurônio i no tempo t .

3. Repetir até convergir:

$$\mu_i(t+1) = f_h \left[\sum_{j=0}^{N-1} w_{ij} \mu_j(t) \right] \quad 0 \leq i \leq N - 1 \quad (4.5)$$

A função f_h é a função degrau (*hard-limiting non-linearity*). Repetir a iteração até as saídas dos neurônios ficarem imutáveis.

4.5.3 Perceptron Multicamadas (*Multilayer Perceptron*)

A rede *Perceptron Multicamadas* é formada por neurônios do tipo *Perceptron* derivados do modelo de McCulloch e Pitts. Para o treinamento dessas redes utiliza-se o algoritmo *Back-propagation*, desenvolvido por Rumelhart e colegas em 1986 [D. E. Rumelhart and Williams, 1986]. Durante o treinamento com o algoritmo *Backpropagation*, a rede opera em uma sequência de dois passos. Primeiro, um padrão é apresentado à camada de entrada da rede. A atividade resultante flui através da rede, camada por camada, até que a resposta seja produzida pela camada de saída. No segundo passo, a saída obtida é comparada à saída desejada para esse padrão particular. Se esta não estiver correta, o erro é calculado e propagado a partir da camada de saída até a camada de entrada. Os pesos das conexões entre as unidades das camadas internas vão sendo modificados conforme o erro é retropropagado. O ciclo é repetido até que o erro esteja dentro de um limite aceitável. O número de vezes que o conjunto de treinamento completo é apresentado à rede é chamado de épocas do treinamento, ou seja, cada vez que o conjunto de treinamento completo é apresentado conta-se um época. O algoritmo *Backpropagation* é apresentado a seguir [Beale and Jackson, 1990]:

1. Inicialize os pesos e os limiares com valores randômicos pequenos;

2. Apresente a entrada $X_p = x_0, x_1, x_2, \dots, x_{n-1}$ e a saída desejada $T_p = t_0, t_1, t_2, \dots, t_{m-1}$, onde n é o número de nodos de entrada e m é o número de nodos de saída;

3. Calcule a saída real. Cada camada calcula

$$y_{pj} = f \left[\sum_{i=0}^{n-1} w_i x_i \right] \quad (4.6)$$

e passa como entrada para a próxima camada. O valor final da camada de saída é o_{pj} ;

4. Adapte os pesos, começando com a camada de saída, e trabalhando em direção das camadas anteriores:

$$w_{ij}(t+1) = w_{ij}(t) + \eta \delta_{pj} o_{pj} \quad (4.7)$$

$w_{ij}(t)$ representa os pesos a partir do nodo i até o nodo j no tempo t , η é um termo de ganho ou taxa de aprendizagem, e δ_{pj} é um termo de erro para o padrão p no nodo j .

Para as unidades de saída

$$\delta_{pj} = k o_{pj} (1 - o_{pj}) (t_{pj} - o_{pj}) \quad (4.8)$$

Para as unidades escondidas

$$\delta_{pj} = k o_{pj} (1 - o_{pj}) \sum_k \delta_{pk} w_{jk} \quad (4.9)$$

onde \sum é o somatório sobre todos os k nodos na camada acima do nodo j .

Depois que a rede estiver treinada e o erro estiver em um nível satisfatório, ela poderá ser utilizada como uma ferramenta para classificação de novos dados. Para isto, a rede deverá ser utilizada apenas no modo progressivo (*feed-forward*). Ou seja, novas entradas são apresentadas à camada de entrada, são processadas nas camadas intermediárias e os resultados são apresentados na camada de saída, como no treinamento, mas sem a retropropagação do erro. A saída corresponde à interpretação da rede para a nova entrada apresentada.

4.6 Classificação de Padrões

Em problemas de classificação, a tarefa da rede é determinar a qual classe o novo padrão pertence. Neste caso, é necessário mapear as diferenças entre as classes no espaço de características. Isto pode ser realizado traçando uma reta que separa as duas classes, chamada

superfície de decisão, como mostra a Figura 4.6. A importância das Redes Neurais no contexto dos problemas de classificação está no grande poder que elas oferecem para representar o mapeamento não linear entre diversas variáveis de entrada e diversas variáveis de saídas [Bishop, 1995].

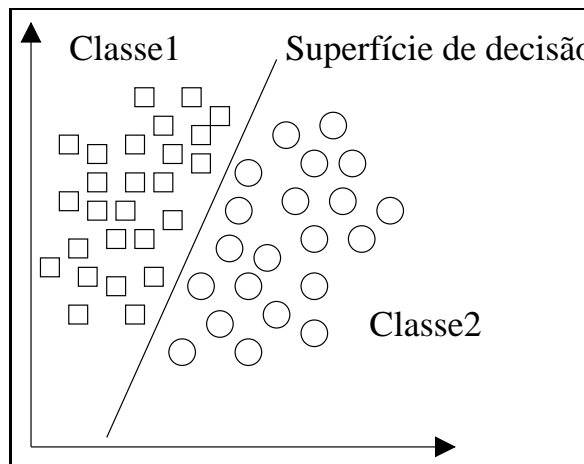


Figura 4.6: Exemplo hipotético de classificação em que uma reta (superfície de decisão) separa as duas classes de padrões no espaço de características.

Nos últimos anos, um número de melhoramentos tem sido propostos para o treinamento de *Perceptrons* de múltiplas camadas, como por exemplo o uso de taxas de aprendizagem variáveis e funções de ajuste de pesos otimizadas. Maiores detalhes sobre estes melhoramentos podem ser encontrados no livro do Haykin [Haykin, 1999].

4.7 A Ferramenta SNNS

Para a implementação, treinamento e teste das redes neurais foi utilizada a ferramenta SNNS (*Stuttgart Neural Network Simulator*) v4.2 em ambiente Linux. Esta ferramenta vêm sendo desenvolvida no Institute for Parallel and Distributed High Performance Systems (*Institut für Parallele und Verteilte Höchstleistungsrechner* - IPVR) na universidade de Stuttgart, desde 1989. O projeto desta ferramenta tem como objetivo alcançar eficiência e alta flexibilidade no projeto e aplicação de Redes Neurais, integrados em um só ambiente de simulação.

O simulador SNNS é formado por quatro componentes principais: o kernel do simulador, a interface gráfica com o usuário, a interface de execução em bach (bachman) e o compilador de redes snns2c. O kernel opera sobre uma representação interna das redes neurais e é

responsável por todas as operações e pelas estruturas de dados que as compõe. A interface gráfica trabalha sobre o kernel, fornecendo uma representação gráfica das redes neurais e controla o kernel durante a simulação. Esta ferramenta permite gerenciar a implementação de uma rede neural através de um painel principal chamado SNNS Manager. Este painel possibilita o acesso à todas as funcionalidades disponíveis no SNNS. O SNNS suporta cinco tipos de arquivos, dos quais os 3 mais importantes são:

1. **.net**: arquivos contendo informações sobre a topologia da rede e regras de aprendizado;
2. **.pat**: arquivos que contêm os padrões de treinamento e de teste;
3. **.res**: arquivos de resultados, os quais permitem que o usuário observe a saída da rede para cada tipo de padrão testado.

Vários são os Algoritmos de aprendizagem que estão disponíveis no SNNS, que podem atender a um número significativo de alternativas topológicas de redes neurais.

4.8 Sumário

Este capítulo apresentou um estudo dos principais fundamentos das Redes Neurais Artificiais, tais como, o neurônio artificial, arquiteturas das redes, suas topologias e os paradigmas de treinamento. Além disso, discutiu os aspectos mais importantes de três dos principais modelos de Redes Neurais, a Rede de Kohonen, a Rede de Hopfield e o *Multilayer Perceptron*. A partir desse estudo introdutório, foi possível optar pela utilização do modelo *Multilayer Perceptron* na investigação inicial de uma arquitetura para o Módulo de Reconhecimento do protótipo apresentado no próximo capítulo. A principal razão para esta escolha é a facilidade na utilização do modelo, por se tratar de um modelo clássico e que têm sido largamente utilizado e explorado.

As Redes Neurais Artificiais fornecem uma alternativa para a solução de problemas em que as abordagens numéricas e simbólicas não são muito adequadas. Ao contrário dos computadores digitais convencionais, as Redes Neurais executam suas tarefas utilizando simultaneamente um número de unidades processadoras (neurônios) que, embora sejam muito simples, possuem um elevado grau de interconexão. Dessa forma, operam em paralelo e o

conhecimento está distribuído por todas as conexões entre os neurônios. Embora as Redes Neurais sejam muitas vezes modelos grosseiros de sistemas nervosos biológicos, o conhecimento mais aprofundado sobre estes sistemas pode ajudar a melhorar os modelos artificiais existentes, minimizando suas limitações.

Capítulo 5

Arquitetura do Protótipo

A partir dos estudos nas áreas de Sistemas de Apoio ao Motorista, Atenção Visual e Redes Neurais apresentados nos Capítulos 2, 3 e 4 respectivamente, foi derivada uma arquitetura híbrida, que é formada por um mecanismo de atenção visual e uma rede neural, para localizar e reconhecer placas de sinalização. Neste capítulo é apresentada esta arquitetura, além dos detalhes da implementação do módulo de detecção. Em seguida são discutidos aspectos importantes na definição e implementação das arquiteturas neurais e que serão apresentados com mais detalhes no Capítulo 6.

5.1 Arquitetura Geral

Com base no que já foi discutido sobre o problema investigado, podemos derivar uma arquitetura geral para o protótipo, contendo dois módulos principais: um para localizar regiões de interesse na imagem, onde existe maior probabilidade de se encontrar as placas de sinalização (Módulo de Detecção), e outro para classificar as regiões localizadas (Módulo de Reconhecimento). Além disso, fica evidente a necessidade de uma etapa inicial de pré-processamento das imagens de entrada, conforme ilustrado na Figura 5.1. No próprio Módulo de Detecção é aplicado um pré-processamento, com o objetivo principal de realçar e extrair características da imagem. Já no pré-processamento aplicado após o Módulo de Detecção, o objetivo é padronizar os exemplos dos conjuntos de treinamento e teste com respeito à algumas características, como por exemplo o *background*, além de reduzir o excesso de ruído nas imagens. Para isto, utiliza-se por exemplo, equalização de histograma e aplicação de um

filtro do tipo *blur* Gaussiano.

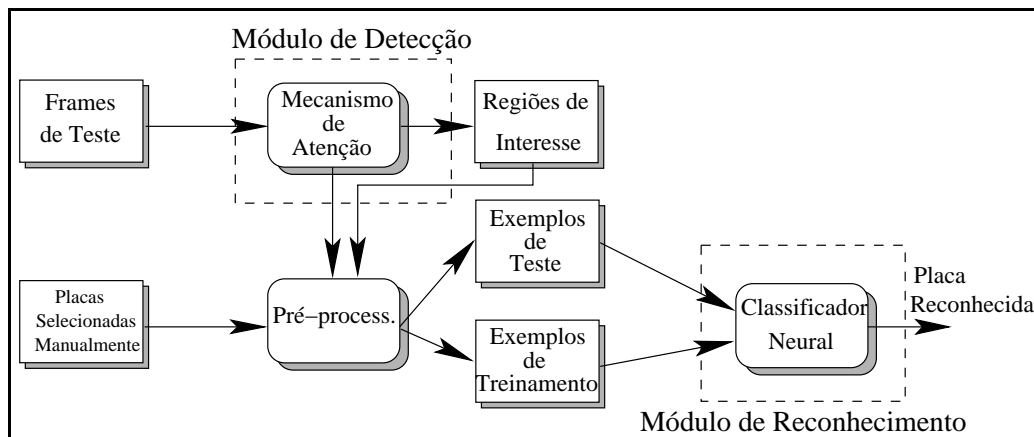


Figura 5.1: Arquitetura geral do protótipo: os retângulos representam os dados e os retângulos arredondados representam os processos.

O contexto de nosso problema envolve padrões previamente conhecidos (placas de trânsito) e a necessidade de um sistema capaz de aprender tais padrões. Sendo assim, a aplicação de uma técnica que utiliza aprendizagem supervisionada é possível. Diante do exposto, decidimos investigar a utilização de uma Rede Neural MLP-BP (*Multilayer Perceptron* com algoritmo de treinamento *Backpropagation*) para a tarefa de classificação (módulo de reconhecimento).

Outro módulo do protótipo que apresentaremos neste capítulo envolve um mecanismo de atenção para localizar as regiões de interesse na imagem de entrada, uma vez que trata-se de um mecanismo computacional biologicamente inspirado, além de ser amplamente difundido e investigado na literatura. Após a pesquisa bibliográfica decidimos investigar a adaptação do modelo posposto por Itti e colegas [Itti et al., 1998] ao nosso problema. Uma das principais vantagens deste modelo é o fato dele trabalhar com o conceito de “Saliência Visual”. Neste caso particular de modelo atencional baseado em saliência (*bottom-up*), um Mapa de Saliência é gerado. O ponto principal é que o Mapa de Saliência codifica o fato de uma dada região ser saliente, não importando qual característica a torna saliente. Podemos concluir então que o protótipo poderá ser adaptado futuramente para detectar e reconhecer outros tipos de objetos. Nas próximas seções discutiremos mais detalhadamente os dois principais módulos do protótipo. O pré-processamento aplicado foi muito simples, consistindo basicamente na redução de ruídos através da aplicação de um filtro *blur* Gaussiano e da equalização do

histograma. Além disso, as imagens foram segmentadas em regiões com placas e regiões sem placas. Este pré-processamento será discutido com mais detalhes no Capítulo 6.

5.2 Módulo de Detecção

A principal razão para a existência de um módulo capaz de detectar os objetos de interesse dentro da cena visual é a necessidade de diminuir a complexidade da tarefa de busca. Tsotsos [Tsotsos, 1990] mostrou que a maior contribuição para a diminuição dessa complexidade é a atenção visual. Partindo desta premissa, o módulo de detecção foi implementado a partir da adaptação do modelo proposto por Itti e colegas [Itti et al., 1998]. A Figura 5.2 ilustra a arquitetura implementada. Na próxima seção é descrito o processo de filtragem linear, que extrai características primitivas da imagem e as representam na forma piramidal. Em seguida, é descrito o processo que implementa diferenças entre os níveis da pirâmide, chamado de Diferenças Centro-Vizinhança. Por fim, são descritos os processos de soma dos mapas e combinação linear que geram como resultado o Mapa de Saliência. Esta seção é finalizada com a apresentação dos pontos distintos entre o modelo implementado neste trabalho e o modelo de Itti e colegas.

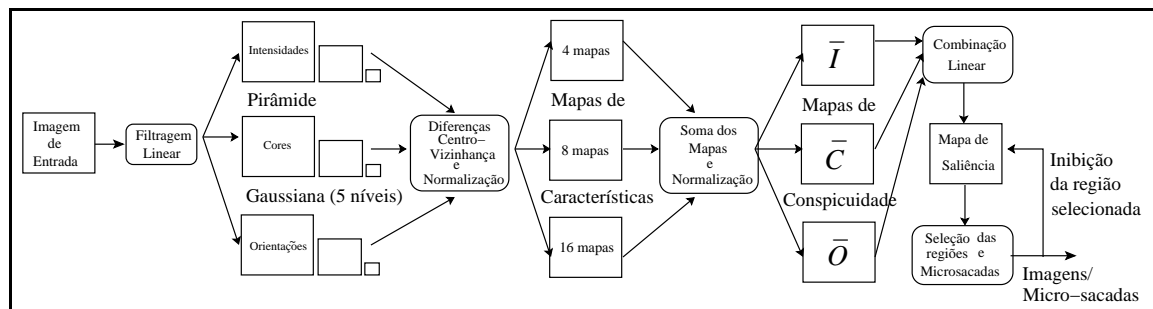


Figura 5.2: Arquitetura do módulo de detecção (adaptação do modelo de Itti e colegas).

5.2.1 Filtragem Linear

A entrada para o módulo de detecção é provida na forma de imagens coloridas, digitalizadas em uma resolução de 352X240. Uma maior resolução não foi possível devido a limitações no hardware de aquisição. O primeiro processo realizado é uma filtragem linear, que extrai

características visuais primitivas das imagens. Três tipos de características são extraídas: cor, intensidade e orientação. Com r , g e b sendo os canais vermelho, verde e azul da imagem de entrada, quatro imagens ou canais de cores são criados: $R = r - (g + b)/2$ para o vermelho, $G = g - (r + b)/2$ para o verde, $B = b - (r + g)/2$ para o azul e $Y = (r + g)/2 - |r - g|/2 - b$ para o amarelo, e para cada canal é gerada uma Pirâmide Gaussiana com cinco níveis: $R(\sigma)$, $G(\sigma)$, $B(\sigma)$ e $Y(\sigma)$ onde $\sigma \in \{0,1,2,3,4\}$. A imagem de intensidades é definida como $I = (r + g + b)/3$, ou seja, é a própria imagem de entrada em níveis de cinza. Assim como para os canais de cores, uma Pirâmide Gaussiana com cinco níveis $I(\sigma)$ também é gerada, fornecendo uma representação multi-escala e multi-resolução. Para a geração das pirâmides utilizamos um algoritmo clássico apresentado no trabalho de Burt e Adelson [Burt and Adelson, 1983] que será descrito a seguir. O mesmo algoritmo é utilizado depois para gerar a interpolação das imagens. Informação de orientação local é obtida a partir de I aplicando-se Filtros Direcionais (*Steerable Filters*) [Freeman and Adelson, 1991] em quatro orientações (0° , 45° , 90° , 135°), e gerando uma Pirâmide Direcional (*Steerable Pyramid*) [Greenspan et al., 1994; Simoncelli and Freeman, 1995; Karasiridis and Simoncelli, 1996] para cada orientação. Foram escolhidas apenas quatro orientações para estes filtros como intuito de tornar o processamento mais rápido. Seria recomendável a utilização de mais orientações em um sistema real, uma vez que as descrições das imagens apresentariam um nível de detalhes maior. A Figura 5.3 mostra exemplos de imagens de canais de cores e de uma imagem de intensidades resultantes da filtragem linear. Exemplos da Pirâmide Gaussiana e da Pirâmide Direcional serão apresentados nas próximas seções.

Geração da Pirâmide Gaussiana

Para a geração da Pirâmide Gaussiana, utiliza-se um algoritmo clássico apresentado em [Burt and Adelson, 1983]. O mesmo algoritmo é utilizado mais tarde para gerar a interpolação das imagens.

Supondo que a imagem é representada inicialmente por uma matriz g_0 que contém C colunas e R linhas de pixels. Cada pixel representa a intensidade de luz no ponto correspondente da imagem, representado por um inteiro I entre 0 e $K - 1$, em que K representa os níveis da pirâmide. Esta imagem torna-se a base ou nível zero da pirâmide Gaussiana. O nível

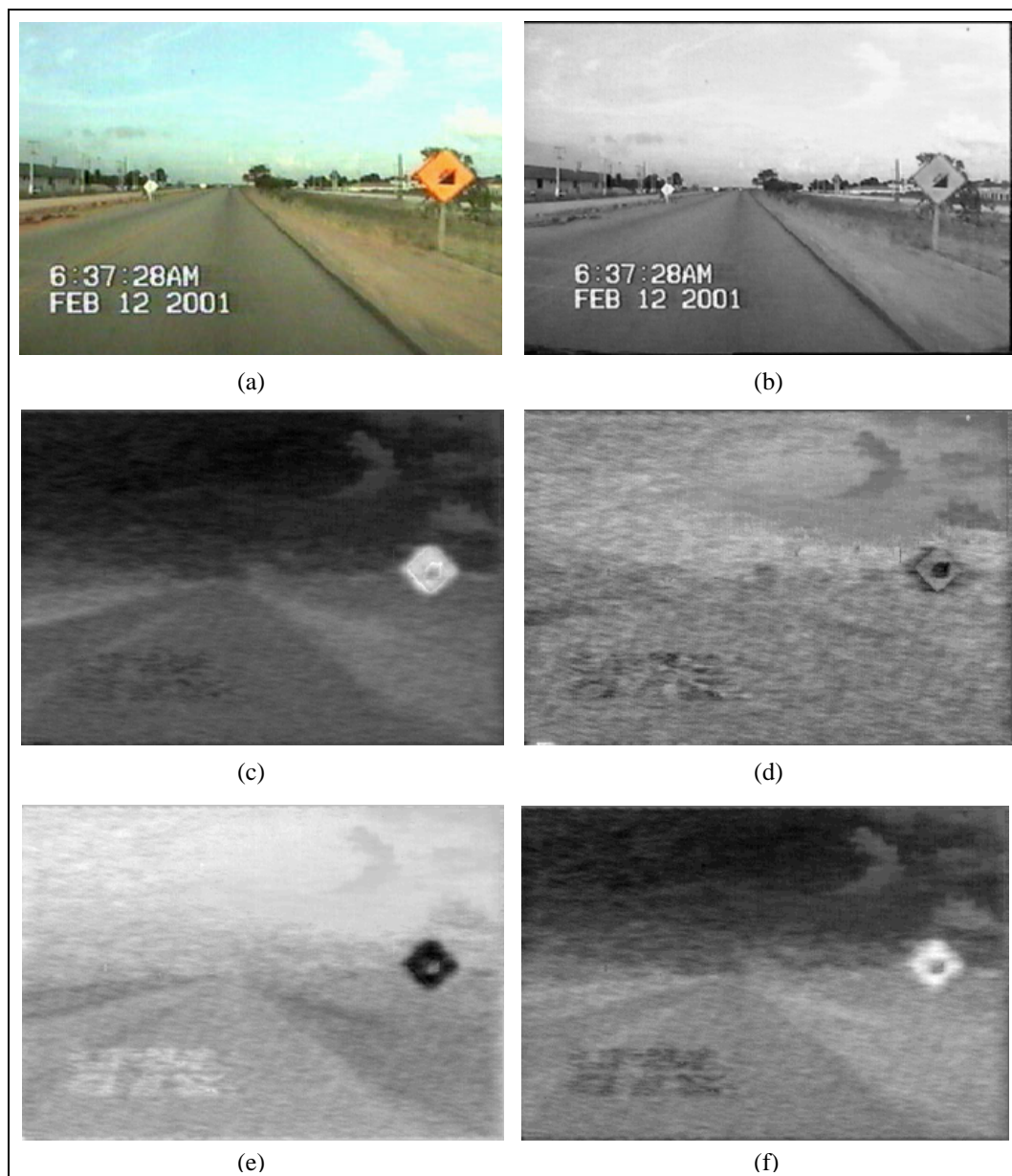


Figura 5.3: Exemplo de imagens resultantes do processo de filtragem linear. (a) Imagem de entrada, (b) imagem de intensidade e respectivos canais de cores R (c), G (d), B (e) e Y (f).

el 1 da pirâmide contém a imagem g_1 , que é uma redução ou uma versão filtrada passa-baixa de g_0 . Cada valor dentro do nível 1 é computado como uma média ponderada dos valores no nível 0, dentro de uma janela 5x5. Cada valor dentro do nível 2, representado por g_2 , é então obtido a partir dos valores dentro do nível 1, aplicando-se o mesmo padrão de pesos, e assim por diante. Para este processo é utilizada a função REDUZ.

$$g_k = REDUZ(g_{k-1}) \quad (5.1)$$

que significa, para níveis $0 < l < N$ e nodos i, j , $0 \leq i < C_l$, $0 \leq j < R_l$,

$$g_l(i, j) = \sum_{m=-2}^2 \sum_{n=-2}^2 w(m, n) g_{l-1}(2i + m, 2j + n) \quad (5.2)$$

O N se refere ao número de níveis na pirâmide, enquanto que C_l e R_l são as dimensões do l -ésimo nível. As dimensões da imagem original são apropriadas para a construção da pirâmide se os inteiros M_C , M_R e N existem, tal que $C = M_C 2^N + 1$ e $R = M_R 2^N + 1$. As dimensões de g_1 são $C_1 = M_C 2^{N-1} + 1$ e $R_1 = M_R 2^{N-1} + 1$. O mesmo padrão de pesos w é usado para gerar cada matriz da pirâmide, a partir de seu predecessor. Este padrão de pesos é chamado de *Generating Kernel* e para simplificar ele é construído separadamente:

$$w(m, n) = \hat{W}(m) \hat{W}(n). \quad (5.3)$$

A função unidimensional de tamanho 5 é normalizada

$$\sum_{m=-2}^2 \hat{W}(m) = 1 \quad (5.4)$$

e simétrica

$$\hat{W}(i) = \hat{W}(-i) \text{ para } i = 0, 1, 2. \quad (5.5)$$

Uma limitação adicional chamada de **contribuição igual**, estipula que todos os nodos em um dado nível devem contribuir com o mesmo peso total ($= 1/4$), para os nodos no próximo nível mais alto. Seja $\hat{W}(0) = a$, $\hat{W}(-1) = \hat{W}(1) = b$ e $\hat{W}(-2) = \hat{W}(2) = c$, contribuição igual, neste caso, requer que $a + 2c = 2b$. As três limitações são satisfeitas quando:

$$\hat{W}(0) = a \quad (5.6)$$

$$\hat{W}(-1) = \hat{W}(1) = 1/4 \quad (5.7)$$

$$\hat{W}(-2) = \hat{W}(2) = 1/4 - a/2. \quad (5.8)$$

A geração da pirâmide é equivalente a convolução da imagem g_0 com um conjunto de funções de peso equivalente h_l :

$$g_l = h_l \otimes g_0 \quad (5.9)$$

ou

$$g_l(i, j) = \sum_{m=-M_l}^{M_l} \sum_{n=-M_l}^{M_l} h_l(m, n) g_0(i2^l + m, j2^l + n). \quad (5.10)$$

A forma da função equivalente converge rapidamente para uma forma característica com níveis da pirâmide sucessivamente maiores, tal que somente sua escala muda. A forma da função de peso equivalente depende da escolha do parâmetro a . A função assume a forma de uma Gaussiana particularmente quando $a = 0.4$. O efeito de convoluir uma imagem com uma das funções de peso equivalente h_l é como aplicar um *blur* ou um filtro passa-baixa na imagem. A Figura 5.4 mostra um exemplo de uma pirâmide Gaussiana gerada a partir de uma imagem de intensidades.



Figura 5.4: Exemplo de uma pirâmide Gaussiana com cinco níveis, gerada a partir de uma imagem de intensidades (Figura 5.3(b)).

Interpolação da Pirâmide Gaussiana

Para a interpolação é definida a função EXPANDE como a reversa de REDUZ. Seu efeito é expandir uma matriz $[M + 1][N + 1]$ em uma matriz $[2M + 1][2N + 1]$ interpolando novos valores do nodo entre os valores dados. Assim, EXPANDE aplicada a matriz g_l da pirâmide gaussiana produz uma matriz $g_{l,1}$ que tem o mesmo tamanho de g_{l-1} .

Seja $g_{l,n}$ o resultado da expansão de g_l n vezes. Então

$$g_{l,0} = g_l \quad (5.11)$$

e

$$g_{l,n} = EXPANDE(g_{l,n-1}) \quad (5.12)$$

Por EXPANDE entende-se, para níveis $0 < l \leq N$ e $0 \leq n$ e os nodos i, j , $0 \leq i < C_{l-n}$, $0 \leq j < R_{l,n}$,

$$g_{l,n}(i, j) = 4 \sum_{m=-2}^2 \sum_{n=-2}^2 w(m, n) \bullet g_{l,n-1} \left(\frac{i-m}{2}, \frac{j-n}{2} \right). \quad (5.13)$$

Apenas os termos em que $(i-m)/2$ e $(j-n)/2$ são inteiros, serão incluídos nesta soma. Se aplicarmos a função EXPANDE l vezes à imagem g_l , obteremos $g_{l,l}$, que tem o mesmo tamanho que a imagem original g_0 .

Pirâmide Direcional (*Steerable Pyramid*)

A Pirâmide Direcional é uma decomposição multi-escala e multi-orientações de uma imagem, que fornece uma representação útil em muitas aplicações de processamento de imagem e Visão Computacional. Nesta decomposição linear, uma imagem é subdividida em um conjunto de subbandas localizadas em escala e orientação. O conjunto de filtros utilizados para a decomposição das orientações é chamado de conjunto de filtros direcionais (*Steerable Filters*). Um conjunto de filtros forma uma base direcional se: (1) eles são cópias rotacionadas uns dos outros e (2) uma cópia do filtro em qualquer orientação pode ser computada como uma combinação linear de filtros base [Freeman and Adelson, 1991; Greenspan et al., 1994; Simoncelli and Freeman, 1995; Karasiridis and Simoncelli, 1996].

A decomposição é mais facilmente definida no domínio de Fourier, onde ela é polar-separável. A Figura 1 contém um diagrama da resposta de frequência idealizada das sub-bandas, para $k = 4$. A magnitude do i -ésimo filtro passa banda orientado é escrito na forma polar-separável:

$$B_i(\vec{\omega}) = A(\theta - \theta_i)B(\omega), \quad (5.14)$$

em que $\theta = \tan^{-1}(\omega_y/\omega_x)$, $\theta_i = \frac{2\pi}{k}$ e $\omega = |\vec{\omega}|$. A seguir, são descritas as limitações nos dois componentes $A(\theta)$ e $B(\omega)$.

DECOMPOSIÇÃO ANGULAR

A porção angular da decomposição, $A(\theta)$, é determinada pela ordem derivativa desejada. Uma operação derivativa direcional no domínio espacial corresponde a multiplicação por uma função rampa linear no domínio de Fourier, que é reescrita em coordenadas polares conforme segue:

$$-j\omega_x = -j\omega \cos(\theta) \quad (5.15)$$

Note que o operador derivativo é descrito na direção x . A constante imaginária, e o fator de ω , que é absorvido na porção radial da função, são ignorados. Portanto, a porção angular relevante do primeiro operador derivativo (na direção x) é $\cos(\theta)$.

Derivativas direcionais de mais alta ordem correspondem a multiplicação no domínio de Fourier pela rampa salientada para um domínio, e assim a porção angular do filtro é $\cos(\theta)^N$ para uma derivativa direcional de N -ésima ordem.

DECOMPOSIÇÃO RADIAL

A função radial, $B(\omega)$, é limitada pelo desejo de construir a decomposição recursivamente (isto é, usando um algoritmo “pirâmide”) e pela necessidade de prevenir discontinuidades que ocorrem durante operações de sub-amostragem. Os filtros $H_0(\omega)$ e $L_0(\omega)$ são necessários para pré-processar a imagem em preparação para a recursão. Este subsistema decompõe um sinal em duas porções (passa-baixa e passa-alta). A porção passa-baixa é sub-amostrada e a recursão é executada pela aplicação repetidamente da transformação recursiva para o sinal passa-baixa.

As limitações dos filtros no diagrama são os seguintes:

1. Limitação de banda (para prevenir descontinuidade na operação de sub-amostragem):

$$L_1(\omega) = 0 \quad \text{para } |\omega| > \pi/2. \quad (5.16)$$

2. Resposta do Sistema Flat:

$$|H_0(\omega)|^2 + |L_0(\omega)|^2 [|L_1(\omega)|^2 + |B(\omega)|^2] = 1. \quad (5.17)$$

3. Recursão:

$$|L_1(\omega/2)|^2 = |L_1(\omega/2)|^2 [|L_1(\omega)|^2 + |B(\omega)|^2]. \quad (5.18)$$

É escolhido $L_0(\omega) = L_1(\omega/2)$, tal que a forma passa-baixa inicial é a mesma que a usada dentro da recursão. No nosso caso são geradas quatro pirâmides com cinco níveis. Para cada pirâmide utiliza-se uma orientação (0° , 45° , 90° , 135°). Para a construção da Pirâmide Direcional, nós adaptamos o código fonte e utilizamos os filtros base desenvolvidos por Simoncelli e Freeman¹ [Simoncelli and Freeman, 1995]. A Figura 5.5 mostra um exemplo de uma Pirâmide Direcional gerada a partir de um imagem de intensidades. A utilização de filtros direcionais lineares foi definida a partir do modelo original de Itti [Itti et al., 1998], que tem como inspiração a organização colunar do córtex visual. Para o propósito de implementar um Mapa de Saliência que guia a seleção de regiões de interesse, o conjunto de filtros se mostrou suficiente. Entretanto, a utilização de outro tipos de filtro, separadamente ou em conjunto, podem fornecer informações mais completas, influenciando de forma positiva a tarefa posterior de reconhecimento. Uma investigação segundo esta linha será apresentada como proposta de trabalho futuro no Capítulo 7.

5.2.2 Diferenças Centro-Vizinhança (*Center-Surround Differences*)

Diferença Centro-Vizinhança é implementada como a diferença entre escalas finas e grossas, ou seja, o centro é um pixel da imagem na escala $c \in \{1, 2\}$ e a vizinhança é o pixel correspondente em outra imagem na escala $v \in \{3, 4\}$ da representação piramidal. A diferença entre duas imagens, denotada por \ominus , é obtida pela interpolação das imagens para a escala

¹Tanto o código fonte quanto os filtros estão disponíveis via ftp anônimo em: <ftp.cis.upenn.edu/pub/eero/steerpyr.tar.Z>

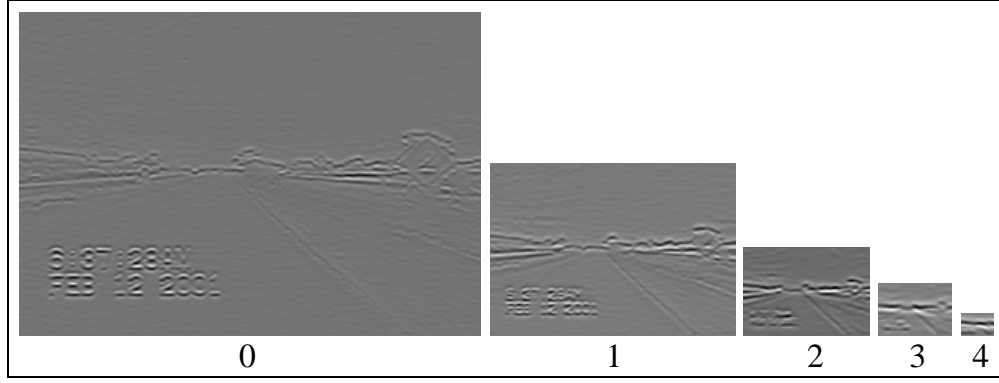


Figura 5.5: Exemplo de uma Pirâmide Direcional com cinco níveis, na orientação 90° .

finas e subtração ponto a ponto. A utilização de várias escalas produz extração de características multiescala, resultando em 28 Mapas de Características. O número de mapas é definido pela combinação das escalas c e v e pela orientação θ , como podemos observar nas equações 5.19, 5.20, 5.21 e 5.22. O primeiro conjunto de mapas é construído a partir do contraste de intensidades, num total de 4 mapas:

$$\mathcal{I}(c, v) = |I(c) \ominus I(v)| \quad (5.19)$$

Nos mamíferos, o contraste de intensidade é detectado por neurônios sensíveis a centros escuros com vizinhança clara, e por neurônios sensíveis a centros claros com vizinhança escura [Shepherd, 1994; Itti and Koch, 2001]. O segundo conjunto de mapas é similarmente construído a partir dos canais de cores, num total de 8 mapas:

$$\mathcal{RG}(c, v) = |(R(c) - G(c)) \ominus (G(v) - R(v))| \quad (5.20)$$

$$\mathcal{BY}(c, v) = |(B(c) - Y(c)) \ominus (Y(v) - B(v))| \quad (5.21)$$

A inspiração biológica para a construção desse conjunto de mapas é a existência, no córtex visual, do chamado Sistema de Cores Oponentes: no centro de seus campos receptivos, neurônios são excitados por uma cor e inibidos por outra e vice-versa. Tal sistema existe para vermelho/verde, verde/vermelho, azul/amarelo, amarelo/azul [Shepherd, 1994; Itti and Koch, 2001]. O terceiro conjunto de mapas é construído a partir de informações de orientação local, num total de 16 mapas:

$$\mathcal{O}(c, v, \theta) = |O(c, \theta) \ominus O(v, \theta)| \quad (5.22)$$

com $\theta \in (0^\circ, 45^\circ, 90^\circ, 135^\circ)$.

A inspiração biológica para a construção dos mapas de orientação é a propriedade de certos neurônios do sistema visual, de responder apenas a uma determinada classe de estímulos, como por exemplo barras orientadas verticalmente [Churchland and Sejnowski, 1992; Itti and Koch, 2001]

5.2.3 O Mapa de Saliência

Para a construção do Mapa de Saliência, os Mapas de Características nas diversas escalas são somados (\oplus), resultando em três Mapas de Conspicuidade: $\bar{\mathcal{I}}$ para intensidade, $\bar{\mathcal{C}}$ para cor e $\bar{\mathcal{O}}$ para orientação, na escala $\sigma = 4$:

$$\bar{\mathcal{I}} = \bigoplus_{c=1}^2 \bigoplus_{v=3}^4 \mathcal{N}(\mathcal{I}(c, v)) \quad (5.23)$$

$$\bar{\mathcal{C}} = \bigoplus_{c=1}^2 \bigoplus_{v=3}^4 [\mathcal{N}(\mathcal{RG}(c, v)) + [\mathcal{N}(BY(c, v))]] \quad (5.24)$$

$$\bar{\mathcal{O}} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \mathcal{N}\left(\bigoplus_{c=1}^2 \bigoplus_{v=3}^4 \mathcal{N}(\mathcal{O}(c, v, \theta))\right) \quad (5.25)$$

A motivação para a criação dos três canais separados ($\bar{\mathcal{I}}$, $\bar{\mathcal{C}}$ e $\bar{\mathcal{O}}$) é a hipótese de que características similares competem pela saliência, enquanto modalidades diferentes contribuem independentemente para o Mapa de Saliência [Itti and Koch, 2001]. O propósito do Mapa de Saliência é representar a conspicuidade - ou saliência - em cada região no campo visual por uma quantidade escalar, e guiar a seleção das regiões atendidas, com base na distribuição espacial da saliência. Os três Mapas de Conspicuidade são normalizados e somados, resultando em uma entrada final para o Mapa de Saliência (Equação 5.26). Para a normalização dos mapas (\mathcal{N}) é utilizado um intervalo com valores de 0 a 255. A Figura 5.6 mostra exemplos dos Mapas de Conspicuidade e do Mapa de Saliência resultante para uma imagem real de estrada.

$$s = \frac{1}{3}(\mathcal{N}(\bar{J}) + \mathcal{N}(\bar{C}) + \mathcal{N}(\bar{O})) \quad (5.26)$$

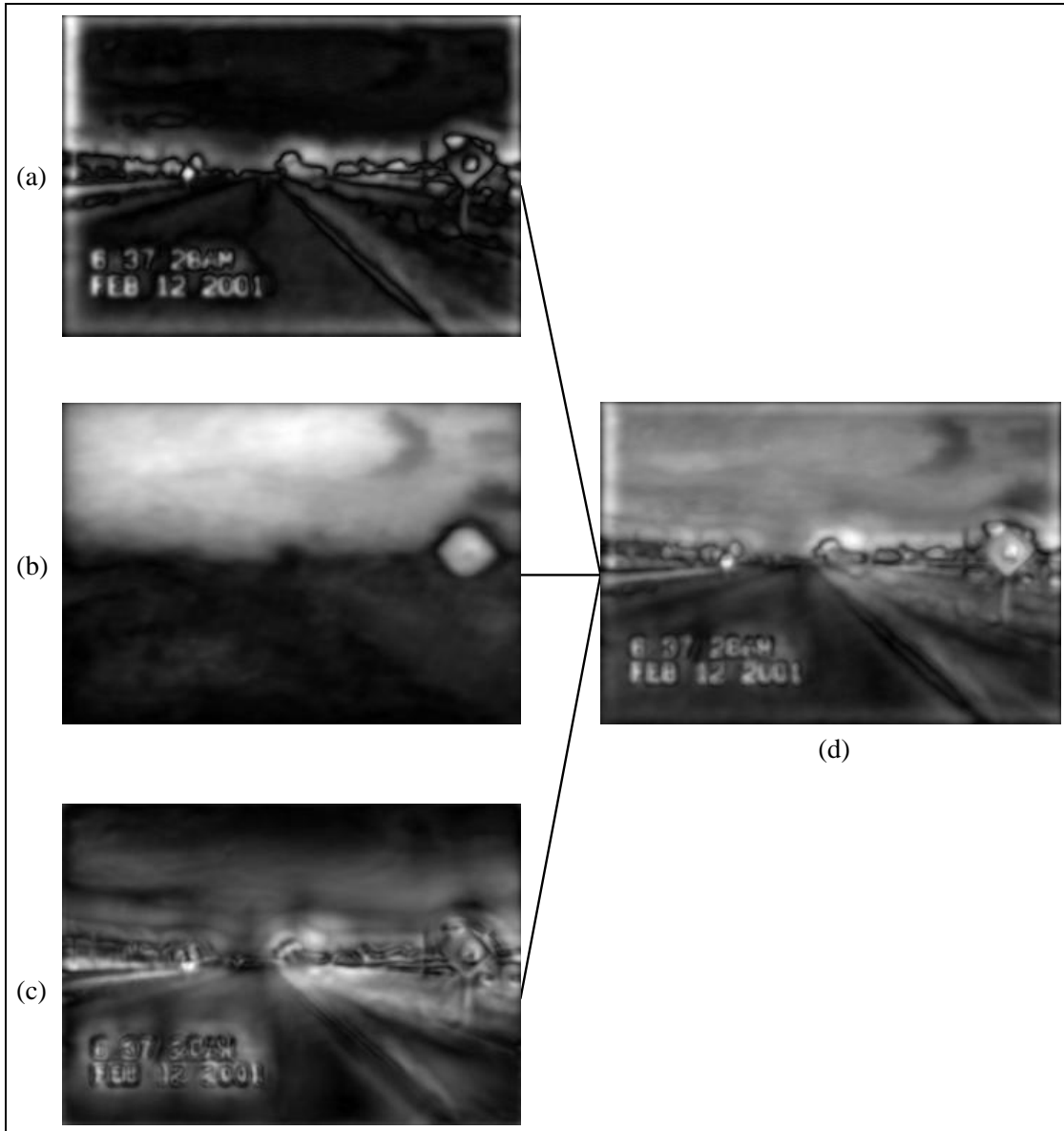


Figura 5.6: Soma dos Mapas de Conspicuidade (a), (b) e (c) para geração do Mapa de Sal-iência (d).

Os passos do algoritmo para implementação do módulo de detecção, discutidos até agora, são baseados no modelo de Itti e colegas [Itti et al., 1998]. Entretanto, nossa implementação apresenta algumas modificações importantes. Primeiramente, as nossas imagens de entrada têm tamanho 352x240 pixels, resultando em uma pirâmide gaussiana de 5 níveis. Este

tamanho foi definido a partir das limitações do hardware de aquisição. Entretanto, esta configuração de pirâmide se mostrou suficiente para o propósito de representar as regiões mais interessantes no mapa de saliência (como será mostrado no Capítulo 5). O modelo original recebe como entrada imagens de tamanho 640x480, tendo como resultado uma pirâmide com nove níveis e requerendo assim mais poder computacional.

Outra modificação na nossa implementação está na estratégia utilizada para a codificação das regiões de interesse e na estratégia de inibição das regiões já atendidas. Para selecionar as regiões de interesse, nós implementamos uma estratégia de ordenação dos pixels de maior valor. Para cada coordenada de interesse (que corresponde ao pixel de maior valor) no Mapa de Saliência, uma região ao redor da coordenada correspondente na imagem de entrada, é selecionada. O tamanho da região selecionada depende do tamanho das placas nas imagens de entrada. Isto significa que, quanto maior for o tamanho das placas nas imagens maior será o tamanho da região selecionada. Neste caso, deve-se levar em consideração um tamanho mínimo capaz de comportar a maior placa existente no conjunto de imagens. Além de selecionar a região de interesse na imagem de entrada, uma máscara de inibição circular preenchida com intensidades nulas é desenhada ao redor da coordenada no Mapa de Saliência. Este procedimento inibe todos os pontos presentes no círculo, impedindo que esta região seja tratada novamente. O diâmetro do círculo também é baseado no tamanho das placas presentes nas imagens, levando-se em conta o risco de inibir uma placa vizinha a uma região atendida. Isto será melhor discutido no próximo capítulo. A Figura 5.7 mostra um exemplo de um Mapa de Saliência (a) gerado a partir da imagem de entrada (d), a inibição das regiões atendidas (b) e (c) e a seleção das respectivas regiões na imagem de entrada (e) e (f).

A seleção na abordagem original de Itti e colegas é realizada por uma Rede Neural (*winner-takes-all*). A Rede recebe como entrada o Mapa de Saliência e seleciona as regiões de interesse com base nos neurônios vencedores, que depois são inibidos. Embora seja uma estratégia biologicamente plausível, ela não implementa deslocamentos em torno da região saliente. Isto pode causar problemas como, a seleção de apenas parte dos objetos. Normalmente, isto ocorre quando o foco da atenção está localizado em regiões de fronteira dos objetos. Para solucionar isto, foi utilizada uma estratégia que simula uma característica importante dos movimentos oculares, as micro-sacadas. Para cada região de interesse são executados deslocamentos, de tal forma que o foco da atenção se localiza em vários outros

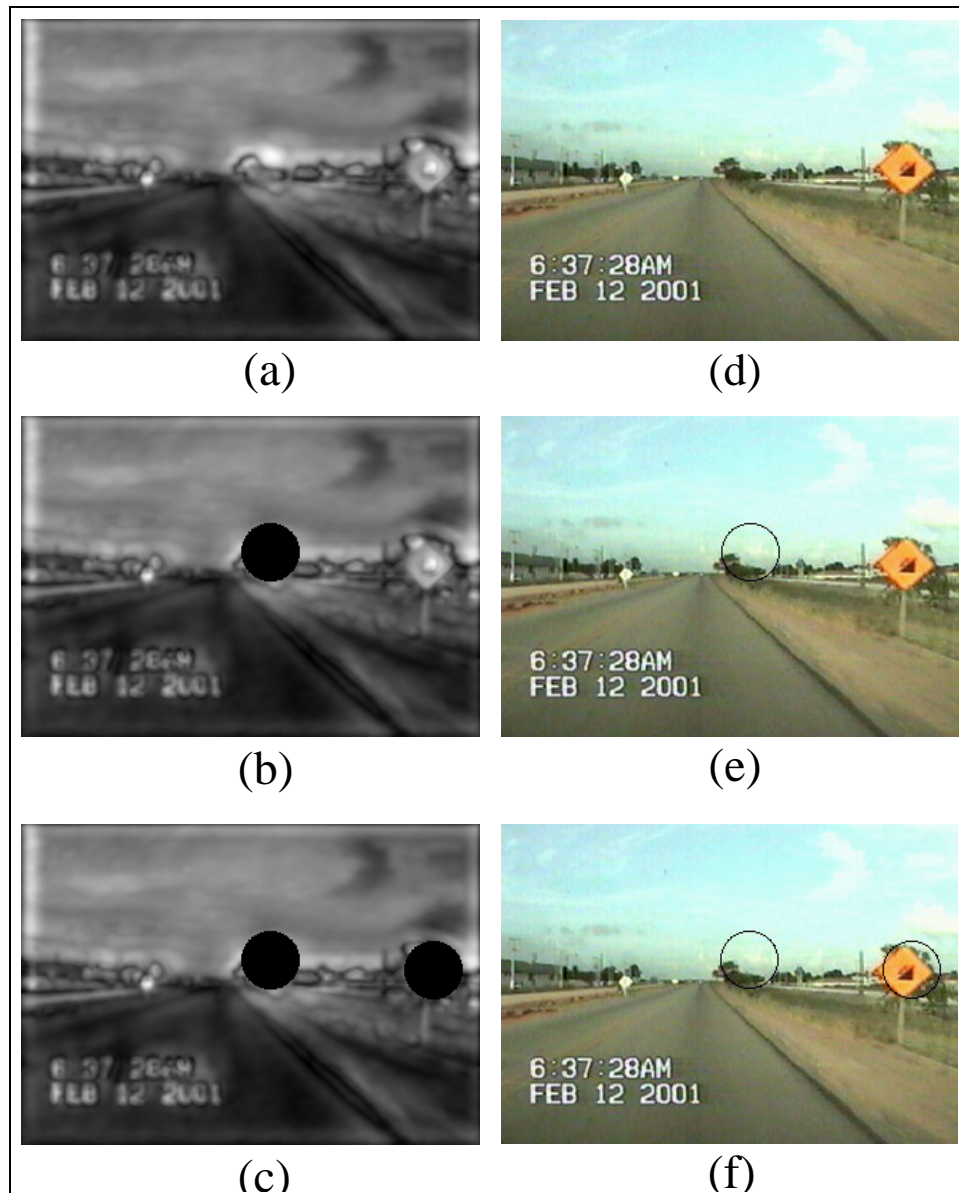


Figura 5.7: Exemplo da seleção das regiões de interesse. Mapa de Saliência (a), (b) e (c) e Imagem de entrada (d), (e) e (f). A cada região inibida no Mapa de Saliência, a região correspondente na imagem de entrada também é selecionada.

pontos vizinhos. Eventualmente, o foco da atenção estará voltado para uma região próxima ao centro do objeto. Para extrair imagens com micro-sacadas, os eixos da coordenada do ponto atendido são variados de cinco em cinco pixels e de dez em dez pixels, gerando 16 novas imagens, perfazendo assim um total de 17 imagens. A Figura 5.8 ilustra a estratégia adotada para gerar as micro-sacadas.

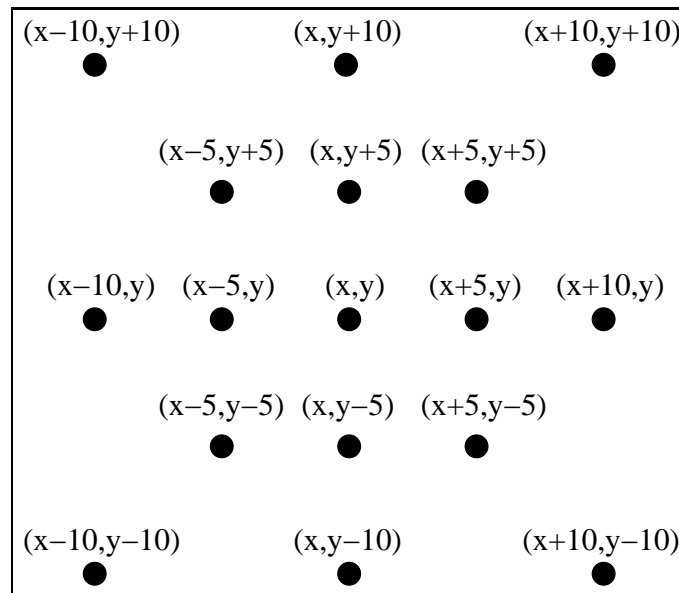


Figura 5.8: Esquema ilustrando a estratégia adotada para gerar as micro-sacadas.

5.3 Módulo de Reconhecimento

Utilizar a técnica de Redes Neurais para uma determinada tarefa envolve, em primeiro lugar, a definição do modelo e da arquitetura da rede. Esta definição está ligada diretamente a dois aspectos principais: o tipo de problema a ser resolvido e o tipo de aprendizado mais apropriado para a solução deste problema. Em problemas de classificação, a rede deve ser capaz de fazer um mapeamento das entradas em categorias discretas. Neste caso, a rede geralmente tem um neurônio de saída para representar cada categoria ou classe e ela deverá ser treinada para responder com apenas um neurônio ativo a cada novo padrão apresentado. O tipo de aprendizado depende de um fator principal: o conhecimento prévio do padrões. Quando os padrões são conhecidos pode-se utilizar o aprendizado supervisionado, onde para cada entrada é associada uma saída desejada. Neste tipo de aprendizado, o conjunto de

treinamento deve ser formado por pares de dados constituídos por padrão de entrada e uma saída desejada para este padrão.

O principal objetivo do módulo de reconhecimento, no escopo deste trabalho, é responder a seguinte pergunta: é possível reconhecer as placas selecionadas pelo mecanismo de atenção? Por envolver padrões previamente conhecidos (placas de sinalização), é possível aplicar uma técnica que utiliza uma aprendizagem supervisionada. Portanto, decidimos investigar a utilização de uma Rede Neural *Multilayer Perceptron* com algoritmo de treinamento *Backpropagation* (MLP-BP) para esta tarefa, por se tratar de uma técnica de classificação tradicional, de fácil utilização e com inspiração biológica.

A arquitetura da rede é definida contendo três camadas. O número de neurônios na camada de entrada é definido em função do tamanho das imagens, onde cada neurônio de entrada corresponde a um ponto da imagem. Para definir o número de neurônios na camada escondida foi realizado um experimento simplificado que determinou uma configuração mais compacta, porém eficiente. Este experimento será detalhado no próximo capítulo. A camada de saída terá número de neurônios correspondente ao número de classes que formam os conjunto de treinamento e teste. Isto significa que, se tivermos duas classes formando os conjuntos, teremos dois neurônios na camada de saída, cada um representando uma classe. Da mesma forma, se tivermos três classe, a camada de saída será formada por três neurônios e assim por diante. Estes três aspectos da arquitetura serão apresentados com mais detalhes no próximo capítulo. A partir da arquitetura da camada de saída, é utilizada como regra de classificação a estratégia *winner-takes-all*. Nesta estratégia, o neurônio que responde com o maior valor de saída corresponde à classe cujo padrão apresentado pertence, na interpretação da rede.

Para interpretar esta saída são utilizados dois tipos de análise: análise por votação e análise absoluta. Na primeira, a taxa de acerto da rede é calculada com base na classificação geral de um subconjunto de padrões da mesma classe, ou seja, se o neurônio que vencer mais vezes em todo subconjunto corresponder a esta classe conta-se um voto. Na segunda, a taxa de acerto é calculada com base no acerto absoluto para cada padrão apresentado, ou seja, para cada padrão classificado corretamente conta-se um acerto da rede.

Os conjuntos de treinamento e teste são formados a partir das imagens geradas pela estratégia de micro-sacadas. No caso do conjunto de treinamento as coordenadas utilizadas

para gerar as micro-sacadas são conhecidas *a priori*. O conjunto de testes é formado pelas imagens geradas pelo módulo de detecção. Todos os aspectos descritos acima serão tratados com mais detalhes no próximo capítulo.

5.4 Sumário

Neste capítulo foi apresentada a arquitetura geral do protótipo de detecção e reconhecimento de placas de sinalização. A implementação do módulo de detecção inspirado no modelo de Itti e colegas foi descrita em detalhes, com destaque para os processos de filtragem linear, diferenças centro-vizinhança, soma dos mapas de características e dos mapas de conspicuidade e a seleção das regiões de interesse. Além disso, foram discutidos os pontos distintos entre o modelo original e este trabalho, com destaque para a diferença no número de níveis da pirâmide e a estratégia de micro-sacadas. Foram discutidos também, conceitos importantes para a definição da arquitetura apropriada da rede neural utilizada nos experimentos, que serão descritos em mais detalhes no próximo capítulo.

O mecanismo de atenção visual que forma o módulo de detecção é inspirado no sistema visual dos primatas. Ao simular as sacadas e micro-sacadas, ele busca inspiração nos movimentos oculares. A extração de características visuais primitivas e as diferenças centro-vizinhança são baseadas na existência de neurônios especializados em perceber cada uma das características (intensidades, cores e bordas). No caso das Redes Neurais, a inspiração é a própria organização do sistema nervoso e o funcionamento de suas unidades (os neurônios). No Capítulo 6 serão descritos os experimentos com os dois módulos do protótipo.

Capítulo 6

Experimentos e Resultados

Este capítulo descreve os experimentos realizados com os módulos apresentados no Capítulo 5 e apresenta os resultados alcançados. Primeiramente, são descritos experimentos preliminares com uma Rede Neural, treinada para classificar imagens de placas selecionadas manualmente. Em seguida, são apresentados os experimentos com o módulo de atenção visual. Os resultados desses experimentos comprovam a aplicabilidade do modelo à tarefa de localização das placas. Por último, são descritos experimentos que avaliam preliminarmente a integração dos dois principais módulos do protótipo.

6.1 Experimentos Iniciais

Com o objetivo de entender melhor o problema investigado e de traçar as diretrizes para a construção do protótipo de detecção e reconhecimento das placas de sinalização, alguns experimentos iniciais com o módulo de reconhecimento foram realizados.

Para a construção da base de dados utilizada nos experimentos optamos por adquirir nossas próprias imagens, já que não encontramos bases consistentes com os objetivos deste trabalho de dissertação. As imagens foram extraídas de um vídeo filmado a partir de um veículo em movimento, durante uma viagem com dia claro entre as cidades de João Pessoa e Campina Grande. Tais imagens incluem cenas com segmentos da rodovia BR-230, que liga as duas cidades, e trechos de ruas urbanas da capital João Pessoa. O *hardware* de aquisição consistiu de uma câmera CCD comum em um tripé, montado na frente do assento direito do carro (assento do passageiro), sem mecanismo de estabilização.

Após a aquisição, o vídeo foi particionado em quadros e cada um deu origem a uma nova imagem colorida, com resolução 352X240 *pixels*. O próximo passo foi selecionar as imagens que continham sinais de trânsito (placas de sinalização). Dessas imagens, foram extraídas manualmente as regiões em que as placas apareciam (com tamanho 20x20 *pixels*), para formar os conjuntos de treinamento e teste. Como o módulo de detecção ainda não havia sido implementado e o objetivo principal era avaliar o desempenho da técnica neural de uma maneira independente do mecanismo de atenção utilizado, a seleção dessas regiões foi manual.

Os conjuntos de treinamento e teste foram compostos de três classes de imagens: **placas pare**, **placas proibido ultrapassar** e **imagens sem placas**, cada classe contendo 14 imagens. A classe de **imagens sem placas** foi construída a partir dos quadros que não continham placas, sendo formada por regiões da imagem com vegetação, asfalto, outros veículos, etc.

Com o objetivo de padronizar as imagens, três passos de pré-processamento foram aplicados: o primeiro passo foi a aplicação de um filtro do tipo *Gaussian Blur* [Waltz and Miller, 1998], de raio 1.0; o segundo passo foi a transformação das imagens, originalmente coloridas, para imagens em níveis de cinza; o terceiro e último passo do pré-processamento foi a equalização do histograma das imagens. A Figura 6.1 exemplifica este pré-processamento. Por se tratar de experimentos preliminares anteriores à implementação do mecanismo de atenção, este pré processamento difere daquele aplicado no Módulo de Detecção.

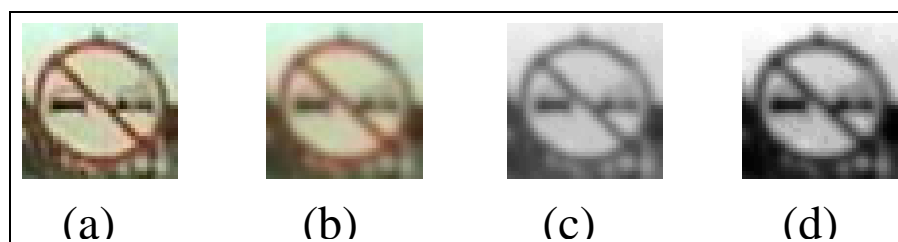


Figura 6.1: Exemplo do pré-processamento aplicado. (a) Imagem original, (b) aplicação do filtro *Gaussian blur*, (c) conversão para níveis de cinza e (d) equalização de histograma.

Para a tarefa de classificação das imagens, foi utilizada uma Rede Neural *Multilayer Perceptron* com a seguinte arquitetura: 400 neurônios na camada de entrada (imagens de entrada de tamanho 20x20), 200 neurônios na camada escondida (empiricamente definida como tendo metade dos neurônios da camada de entrada) e 3 neurônios na camada de saída

(representando cada uma das três classes de imagens escolhidas). Por se tratar de experimentos preliminares com o objetivo de traçar diretrizes, foi determinado que a definição de uma arquitetura ideal para a rede, se daria em um momento posterior. Dessa forma, a escolha da arquitetura se deu de forma empírica. Uma estratégia *winner-takes-all* foi usada para codificar a saída da rede, com um neurônio de saída para cada classe. Uma forma simplificada de treinamento e teste, na qual os conjuntos tinham tamanho (número de padrões) variado, foi usada durante os experimentos: treinando com T padrões e testando com $14 - T$ padrões por classe, em que $T \in \{1, 2, 3, \dots, 13\}$. Na fase de treinamento, todos os conjuntos foram treinados 100%.

6.1.1 Resultados Iniciais

A partir da estratégia de treinamento e teste utilizada, a taxa média de reconhecimento foi de 84,40%. Entretanto, esta taxa média não reflete o comportamento geral dos resultados para todas as combinações de conjuntos (treinamento/teste) já que a melhor taxa foi de 100% (para $T=11, 12, 13$) e a taxa mais baixa foi de 56,41% (para $T=3$). A Tabela 6.1 e o gráfico na Figura 6.2 mostram as taxas de acerto da rede para cada T padrões de teste.

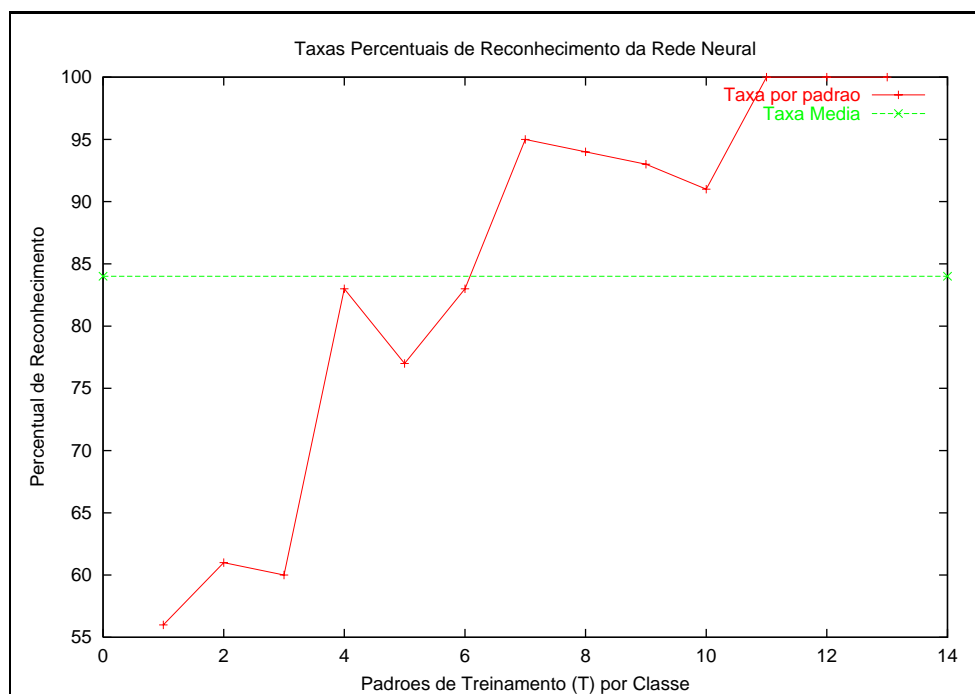


Figura 6.2: Gráfico das taxas de acerto da Rede Neural.

Padrões de Treinamento	Padrões de Teste	Taxa de Acerto (%)
1	13	56,41
2	12	61,11
3	11	60,60
4	10	83,33
5	9	77,78
6	8	83,33
7	7	95,24
8	6	94,44
9	5	93,33
10	4	91,67
11	3	100
12	2	100
13	1	100

Tabela 6.1: Taxas de acerto da Rede Neural para cada combinação de tamanhos dos conjuntos: T padrões para treinamento e 14-T padrões para teste (em que $T=1,2,\dots,13$).

Os resultados iniciais deram uma indicação que taxas de reconhecimento maiores são possíveis, mesmo para imagens em níveis de cinza, com tamanho pequeno (20x20 pixels). Os experimentos também mostraram que o número de padrões de treinamento tem um papel fundamental na tarefa de classificação, muito embora não seja possível fazer uma análise mais detalhada dos resultados uma vez que, foram aplicadas variações no número de padrões de ambos os conjuntos (treinamento e teste). Na próxima seção apresentaremos os experimentos relacionados à definição da arquitetura da rede neural.

6.2 Experimentos para Definição da Arquitetura da Rede

A camada de entrada é formada por 400 neurônios, que correspondem ao número de pixels das imagens que formam o conjunto de treinamento. Para determinar o número de neurônios na camada escondida foi utilizada uma estratégia simples. Esta estratégia consistiu em treinar redes com várias arquiteturas e com o mesmo conjunto de treinamento, escolhendo a arquitetura que obteve melhor resultado. A análise do melhor resultado se deu sobre o seguinte aspecto: comparação da soma dos erros quadrados (*Sum of Squared Errors* - SSE) das redes em uma determinada época do treinamento. A arquitetura considerada melhor foi a que obteve o menor SSE no número pré-fixado de épocas. O número de épocas corresponde ao número de vezes que o conjunto de treinamento completo é apresentado à rede durante o treinamento. Em todos os treinamentos, o parâmetro taxa de aprendizagem não foi variado, sendo definido heurísticamente como 0,02.

O número de neurônios na camada escondida das redes foi variado em múltiplos do número de classes. Por exemplo, diante de um conjunto de treinamento formado por 7 classes, a primeira rede treinada contém 7 neurônios na camada escondida, a segunda 14 e assim por diante. Os valores dos SSE's das redes foram comparados, tomando como base a milésima época do treinamento. O melhor resultado foi alcançado pela rede que tem número de neurônios na camada escondida igual a cinco vezes o número de classes, como mostra a Tabela 6.2 e o gráfico na Figura 6.3.

A saída da Rede Neural foi definida a partir de uma estratégia *winner-takes-all*, onde cada neurônio de saída é codificado para responder a uma determinada classe de padrões. Dessa forma, o número de neurônios na camada de saída corresponde ao número de classes

N° de Neurônios	SSE/1000 Épocas
7 neurônios	29.8216
14 neurônios	3.2632
21 neurônios	1.0610
28 neurônios	0.1326
35 neurônios	0.1261
42 neurônios	1.0976
49 neurônios	1.0945
56 neurônios	0.9165
63 neurônios	0.7991
70 neurônios	1.0807

Tabela 6.2: Valores das somas dos erros quadrados (SSE) das redes, na milésima época do treinamento, em relação ao número de neurônio na camada escondida. Neste exemplo, o conjunto de treinamento é formado por 7 classes.

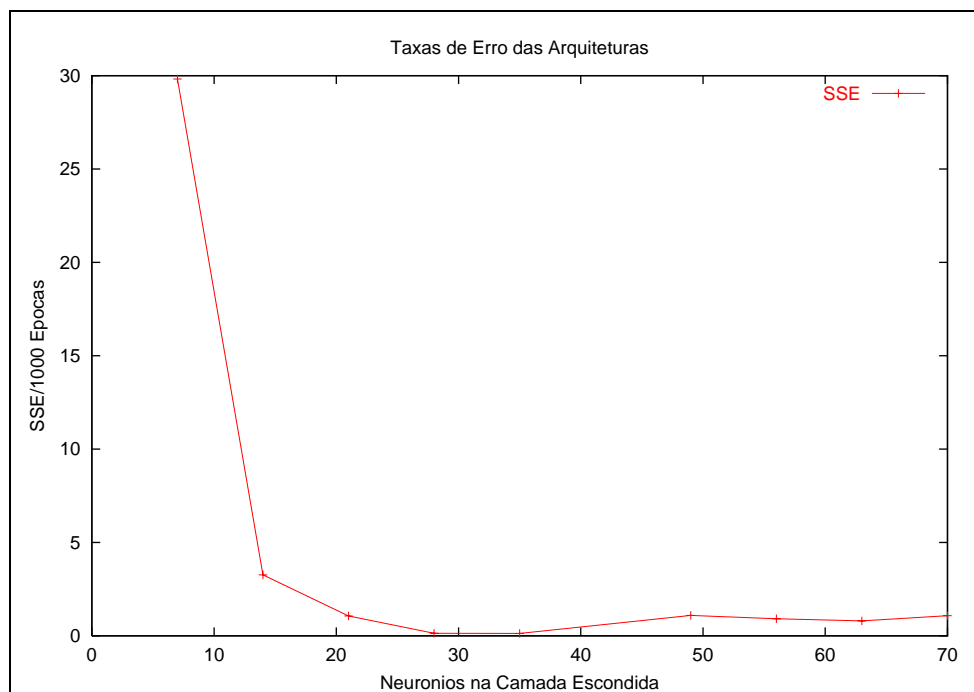


Figura 6.3: Gráfico dos valores do SSE para as várias arquiteturas, na milésima época.

no conjunto de treinamento. A Tabela 6.3 mostra um exemplo da codificação das saídas desejadas para 7 classes, utilizando esta estratégia.

Classes	Saídas Desejadas
Classe 1	1 0 0 0 0 0 0
Classe 2	0 1 0 0 0 0 0
Classe 3	0 0 1 0 0 0 0
Classe 4	0 0 0 1 0 0 0
Classe 5	0 0 0 0 1 0 0
Classe 6	0 0 0 0 0 1 0
Classe 7	0 0 0 0 0 0 1

Tabela 6.3: Codificação das saídas desejadas para 7 classes, utilizando a estratégia *winner-takes-all*.

Embora muito simplificada, a estratégia acima foi capaz de determinar uma arquitetura que convergia mais rápido para o tipo de dado utilizado, dentre um subconjunto de arquiteturas possíveis. Podemos encontrar na literatura, estratégias mais sofisticadas para este propósito como por exemplo, a de Validação Cruzada (*Cross Validation*) [Chen and Hagan, 1999]. Entretanto, Andersen e Martinez [Andersen and Martinez, 1999] mostraram que, ao testar muitas arquiteturas diferentes, o desempenho da Validação Cruzada é levemente melhor se comparado com a seleção randômica da arquitetura. Neste caso, eles afirmam que é melhor usar a arquitetura mais simples disponível. A Figura 6.4 ilustra a arquitetura determinada pela estratégia baseada no SSE/Época de treinamento. Para a implementação da rede foi utilizado um simulador de redes neurais de distribuição gratuita¹, desenvolvido na Universidade de Stuttgart, Alemanha, chamado SNNS (*Stuttgart Neural Network Simulator*). Este simulador possui vários algoritmos de aprendizado, dentre os quais diversas variações do *Backpropagation*. Todas as implementações de Redes Neurais deste trabalho de dissertação foram realizadas utilizando-se o SNNS. Uma descrição sucinta deste simulador pode ser encontrada na Seção 4.7.

¹Disponível para download no endereço: <http://www-ra.informatik.uni-tuebingen.de/SNNS/>

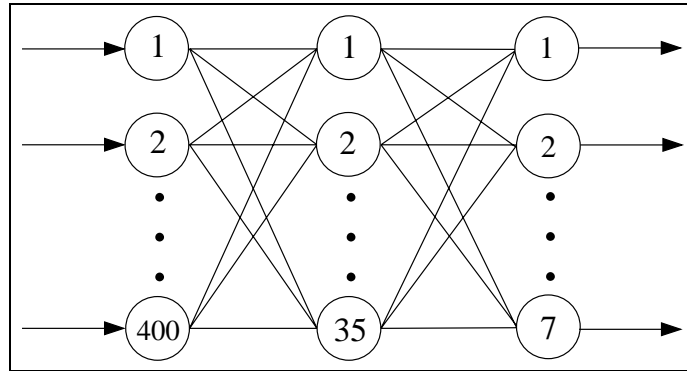


Figura 6.4: Arquitetura determinada pela estratégia baseada no SSE/Época de treinamento.

6.3 Experimentos com o Módulo de Detecção

Para os experimentos iniciais com o Módulo de Detecção, foi selecionado um subconjunto de imagens a partir da base de imagens extraídas do vídeo. Apenas imagens com placas foram selecionadas, num total de 15 imagens com 16 placas, sendo 14 imagens com uma placa e uma imagem com duas placas. Dentre as 15 imagens deste subconjunto, 12 são de trechos de rodovia e 3 são de trechos de ruas urbanas.

Inicialmente, o Módulo de Detecção foi aplicado ao subconjunto de imagens, fixando um número de regiões (K) a serem selecionadas. No primeiro experimento assumimos $K=5$, ou seja, as cinco regiões mais salientes no Mapa de Saliência são selecionadas. Com este primeiro valor de K, 12 das 16 regiões com placas foram localizadas, correspondendo a 75% de localização das placas. Dentre as imagens que tiveram placas selecionadas, 11 correspondem a trechos de rodovia e uma corresponde a um trecho de rua urbana. Em geral, as imagens com trechos de rodovias são formadas por vegetação, estrada, céu, a placa e algumas vezes por outros veículos, isto é, imagens com poucas estruturas presentes. Já as imagens que não tiveram placas selecionadas em nenhuma das cinco regiões mais salientes correspondem a trechos de ruas urbanas, e são formadas por muitas estruturas. Como já foi dito anteriormente, o Mapa de Saliência do modelo representa as regiões mais salientes da imagem, não importando qual característica torna essas regiões salientes. Devido a essa propriedade, as imagens formadas por muitas estruturas poderão conter um número maior de regiões mais salientes que as placas, implicando em um número maior de regiões analisadas para que a placa seja selecionada.

Em seguida, O valor de K foi sendo incrementado até que as placas nas demais imagens fossem selecionadas. O gráfico na Figura 6.5 mostra as taxas percentuais de localização das placas, com relação ao número de regiões de interesse selecionadas. Apenas em uma imagem a placa não foi selecionada, pois um ponto mais saliente na sua vizinhança foi atendido, causando sua inibição no Mapa de Saliência após a aplicação da máscara de intensidades nulas e raio 20. O raio da máscara é definido com base no tamanho das regiões com placas. No subconjunto utilizado nos testes, as placas têm tamanho variado, devido ao uso de quadros adjacentes da sequência de vídeo. Por esta razão, foi necessário definir um tamanho ideal para as regiões selecionadas. Foi definido um tamanho 40x40 pixels observando-se que este era o tamanho limite que suportaria as maiores placas (diferentemente dos primeiros experimentos com o Módulo de Reconhecimento que utilizou apenas um quadro, o que resultou em um tamanho menor - 20x20 - porém suficiente para suportar as placas neste caso específico). Nos experimentos seguintes, o problema da inibição de regiões com placas foi minimizado com a redução do raio da máscara. É importante destacar que no caso de uma região saliente na vizinhança, parte da placa aparece na região selecionada. Entretanto, não podemos considerar como placa localizada porque o ponto saliente não faz parte do objeto, ou seja, a seleção desta região não é determinada por suas características.

Observando os resultados do ponto de vista da complexidade computacional associada ao número de pontos a serem analisados na imagem, podemos notar que foi possível localizar 75% das placas examinando-se apenas 0,0059% ($K=5$) de todos os pontos das imagens (84480 pontos). Mesmo quando K assumiu seu maior valor ($K=19$), o percentual de pontos examinados em relação ao total de pontos da imagem foi de apenas 0,0225%, com 93,75% de placas localizadas. Fica claro que a utilização do mecanismo reduz drasticamente o espaço de busca, em relação a uma busca exaustiva. A Tabela 6.4 e o gráfico na Figura 6.6 mostram a quantidade de pontos selecionados (em termos percentuais) até que uma placa tenha sido localizada, em cada imagem do subconjunto utilizado nos testes.

Com o objetivo de reforçar a validade dos resultados alcançados nesses experimentos preliminares com o Módulo de Detecção, nós implementamos um algoritmo para gerar pontos de interesse randômicos. Para cada imagem foram gerados vinte pontos e definida uma distância mínima entre os pontos igual a 20 pixels, simulando a máscara de inibição. Dentre todos os pontos gerados randomicamente, apenas um coincidiu com uma região que contém

Imagem	Pontos de interesse selecionados	Percentual em relação ao total de pontos da imagem
Imagem 1	16	0,0189
Imagem 1	19	0,0225
Imagem 2	5	0,0059
Imagem 4	3	0,0036
Imagem 5	2	0,0024
Imagem 6	4	0,0047
Imagem 7	4	0,0047
Imagem 8	5	0,0059
Imagem 9	3	0,0036
Imagem 10	2	0,0024
Imagem 11	2	0,0024
Imagem 12	2	0,0024
Imagem 13	3	0,0036
Imagem 14	10	0,0118
Imagem 15	1	0,0012

Tabela 6.4: Pontos de interesse selecionados em cada imagem até que uma placa tenha sido localizada. A imagem 1 aparece duas vezes na tabela por apresentar duas placas. A placa na imagem 3 não foi localizada, já que um ponto mais saliente na sua vizinhança causou sua inibição.

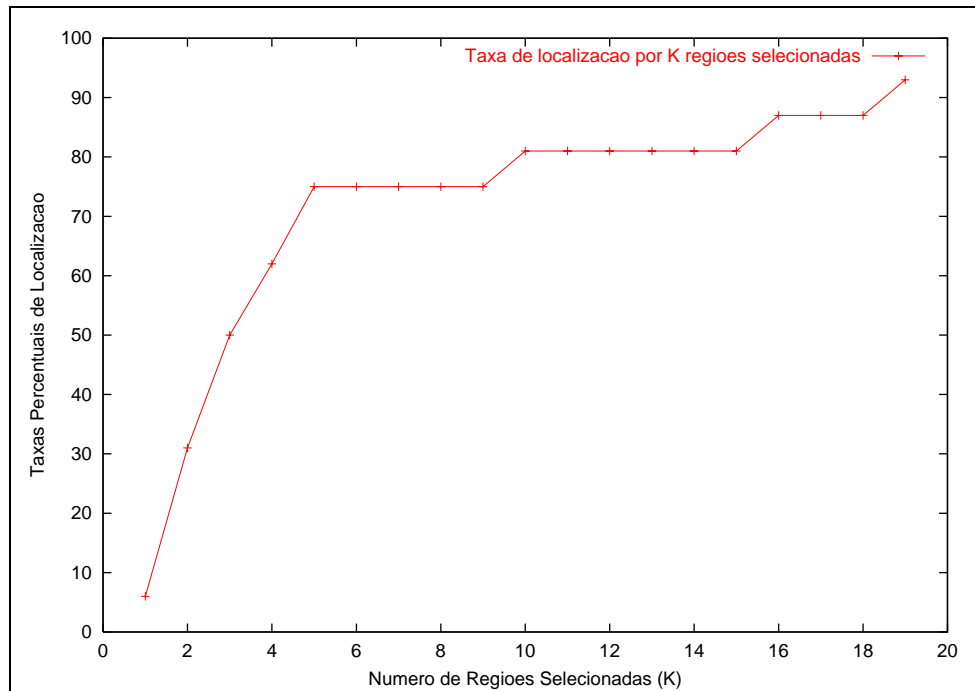


Figura 6.5: Taxas percentuais de localização das placas em todas as imagens, quando consideramos um número K de regiões seleccionadas.

uma placa, indicando que o mecanismo de atenção realiza uma tarefa muito mais sofisticada e de melhor desempenho do que uma simples busca aleatória. O subconjunto de imagens utilizadas pode ser encontrado no Apêndice A. A seguir são apresentados experimentos preliminares com os módulos de detecção e reconhecimento.

6.4 Experimentos Envolvendo a Integração dos Módulos

Com o Mecanismo de Atenção funcionando, o próximo passo foi realizar experimentos para tentar simular a integração dos dois módulos principais. Estes experimentos tiveram como objetivo responder a pergunta: é possível classificar as placas seleccionadas pelo Mecanismo de Atenção? Com base nos primeiros resultados na utilização da Rede Neural, a resposta para esta pergunta é positiva. No entanto, as imagens com placas utilizadas naquele momento foram seleccionadas manualmente. Agora, as imagens deveriam ser fornecidas pelo Módulo de Detecção.

Novas imagens foram escolhidas para formar dois subconjuntos: um para treinamento da

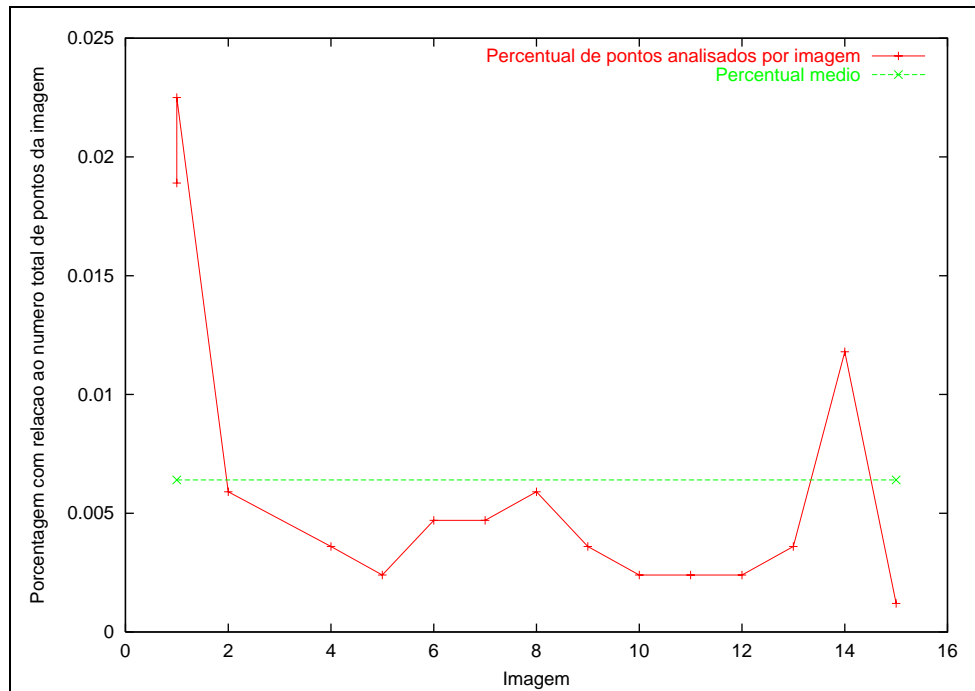


Figura 6.6: Curva de pontos analisados (em termos percentuais) até que uma placa tenha sido selecionada, nas imagens do subconjunto de teste. A linha horizontal representa o percentual médio.

Rede Neural e outro para testes com ambos módulos. A escolha das imagens teve como base a ocorrência das placas no vídeo, ou seja, cada vez que aparece uma placa ao longo do vídeo conta-se uma ocorrência. Cada ocorrência corresponde a um objeto no mundo real, não importando em quantos quadros do vídeo ele aparece. O vídeo utilizado para extrair as imagens contém poucas ocorrências de placas da mesma classe. Devido a isto, não foi possível formar conjuntos com um número maior de padrões. Diante desta limitação, foram escolhidas duas ocorrências de placas por classe, sendo uma para formar o conjunto de treinamento e outra para formar o conjunto de teste. Além disso, cada ocorrência deu origem a cinco imagens, que correspondem a uma sequência de quadros do vídeo. Cada sequência foi escolhida com base no último quadro em que a placa aparece e os quatro quadros anteriores completam o subconjunto. Os conjuntos de treinamento e teste são formados por 7 classes, sendo 6 classes de placas e uma classe de imagens sem placas (contra-exemplo). As classes de placas foram formadas por 5 imagens, num total de 30 imagens por conjunto. Entretanto, a estratégia de micro-sacadas utilizada (discutida no Capítulo 4) minimiza este problema.

Maiores detalhes sobre as dificuldades na aquisição das imagens e formação dos conjuntos utilizados nos experimentos serão discutidos na Capítulo 6. As classes de placas utilizadas foram: **pare**, **proibido ultrapassar**, **limite de velocidade 60Km**, **curva à direita**, **faixa de pedestres** e **indicação de lombada**. A Figura 6.7 apresenta exemplos das seis classes utilizadas (os números entre parêntesis são identificadores utilizados para representar cada uma das classes nos experimentos).

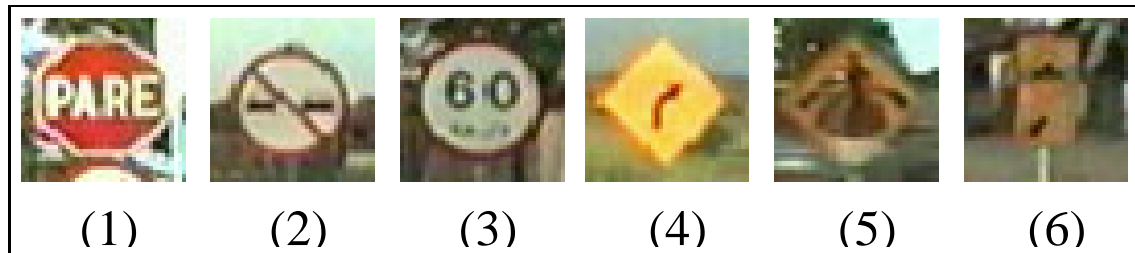


Figura 6.7: Exemplos de placas utilizadas nos experimentos: classe (1) - placa pare, classe (2) - placa proibido ultrapassar, classe (3) - placa limite de velocidade 60Km, classe (4) - placa curva à direita, classe (5) placa faixa de pedestres e classe (6) placa indicação de lombada.

Nos experimentos preliminares, descritos no início deste capítulo, todas as placas estavam localizadas no centro das imagens (regiões selecionadas manualmente). Como a estratégia de micro-sacadas, implementada no Módulo de Detecção (ver Seção 5.2.3 do Capítulo 5), fornece 16 variações na posição da placa nas imagens selecionadas, é necessário adotar a mesma estratégia na formação do conjunto de treinamento, com o objetivo de tentar gerar invariância.

No caso do conjunto de treinamento, as coordenadas centrais de cada placa são conhecidas *a priori*. Com o uso desta estratégia, cada classe no conjunto de treinamento é formada por 85 imagens (cinco quadros consecutivos em que aparece vezes 17 micro-sacadas), aumentando assim a quantidade de padrões de treinamento. A Figura 6.8 mostra o resultado deste procedimento. O tamanho dessas imagens foi baseado no tamanho das placas no último quadro escolhido, descrito no parágrafo anterior. Ao definir uma janela de tamanho 40x40 foi possível representar os diversos tamanhos das placas.

O conjunto de teste foi formado por imagens cujas coordenadas centrais foram selecionadas pelo Mecanismo de Atenção. Como os critérios adotados para a execução das



Figura 6.8: Exemplo da aplicação da estratégia de micro-sacadas. Como resultado temos 17 imagens para cada ocorrência de placa.

micro-sacadas foram iguais para ambos os conjuntos, o número de padrões de treinamento e teste é o mesmo. Diferentemente dos primeiros experimentos com a Rede Neural, os conjuntos de treinamento e teste são fixos. Nas seções seguintes serão apresentados os resultados dos experimentos em que, as saídas do Módulo de Detecção são usadas como entradas para o Módulo de Reconhecimento.

6.4.1 Módulo de Detecção

Nos aspectos gerais, o método utilizado nos novos testes com o módulo de detecção, segue a mesma linha dos primeiros experimentos. Foi definido um número de regiões de interesse (K) para serem selecionadas, que foi sendo incrementado até que as placas fossem

selecionadas. Novamente os resultados mostraram que o mecanismo implementado é bem apropriado para localizar as placas de sinalização nas imagens. Uma mudança em relação ao experimento anterior foi a escolha do raio da máscara de inibição. Assumindo um raio menor diminui-se o risco de inibir uma placa quando uma região saliente está localizada na sua vizinhança. Outra consequência desta diminuição no raio é o aumento no número de pontos analisados, até que o mecanismo encontre uma placa. Entretanto, este aumento não é tão significativo se novamente levarmos em conta o número total de pontos na imagem. Para esses novos experimentos foi utilizada uma máscara de raio=5, o que permitiu localizar 100% das placas no conjunto de testes analisando-se apenas 33 pontos da imagem. Isto corresponde a somente 0,039% dos 84480 pontos de uma imagem do conjunto (cada imagem tem tamanho 352X240). A Tabela 6.5 mostra a relação entre o número de pontos analisados e o percentual de localização das placas em todo conjunto de imagens. É importante ressaltar que em cada imagem do conjunto de testes, existe apenas uma placa.

Pontos de interesse selecionados	Percentual de placas localizadas em todo conjunto de Imagens
1	3,33%
2	26,66%
3	36,66%
4	43,33%
5	50%
6	63,33%
7	70%
9	73,33%
11	76,66%
14	83,33%
23	86,66%
25	93,33%
28	96,66%
33	100%

Tabela 6.5: Relação entre os pontos de interesse selecionados e as placas localizadas.

O conjunto de teste para a Rede Neural foi construído a partir das imagens com placas, selecionadas pelo Módulo de Detecção. Na próxima seção serão apresentados os resultados desses experimentos.

6.4.2 Módulo de Reconhecimento

Inicialmente, os experimentos tiveram como ponto de partida o treinamento de uma rede para classificar as 7 classes de imagens. A arquitetura foi definida através dos experimentos apresentados na Seção 6.2 deste capítulo. Durante o treinamento, a rede convergiu para um erro satisfatório (SSE da ordem de 0,02) e treinou todos os padrões, ou seja, obteve taxa de acerto para o conjunto de treinamento igual a 100%. Entretanto, ao aplicarmos o conjunto de teste, as taxas de acerto alcançadas foram muito baixas, com taxa média igual a 17,64%. A Tabela 6.6 apresenta as taxas de acerto para todas as classes. Foi possível identificar três motivos principais para esse comportamento:

1. A falta de um pré-processamento mais robusto e de uma representação mais compacta dos padrões;
2. A dimensionalidade do espaço de características (1600 neurônios de entrada) requer um número muito grande de exemplos de treinamento, para que a rede tenha poder de generalização. Isto é conhecido como a praga da dimensionalidade [Bishop, 1995];
3. A limitação da arquitetura MLP-BP em relação a translação dos objetos na imagem [Kröner, 1996]. Uma característica dos padrões utilizados nos experimentos é a variedade de possíveis localizações da placa dentro da imagem, em decorrência da estratégia de micro-sacadas. A descoberta do impacto negativo desta limitação da arquitetura se deu nos momentos finais do trabalho.

A Tabela 6.7 apresenta a matriz de confusão para todas as sete classes. A partir desta matriz é possível perceber que classes causam uma maior quantidade de falsas respostas da rede neural, com relação a classe desejada. A classe que causou maior interferência em outra foi a classe 6, que ocasionou 56 respostas erradas da rede (equivalente a 65,88%). Mesmo quando a rede respondeu com uma taxa de acerto maior, a forte influência de outra classe

Classe	Taxa de acerto
1	8,23%
2	28,23%
3	35,29%
4	10,59%
5	27,05%
6	25,88%
7	11,76%

Tabela 6.6: Taxa de acerto da rede monolítica para as sete classes treinadas.

pôde ser notada, como é o caso das classes 3 e 5 que tiveram forte influência das classes 5 e 7 respectivamente, ocasionando as baixas taxas de acerto.

		Resposta da Rede Neural						
		1	2	3	4	5	6	7
C L A S S E S	1	7	0	32	12	17	11	6
	2	1	24	24	3	24	9	0
	3	13	6	30	0	28	6	2
	4	8	3	9	9	8	38	10
	5	2	8	8	7	23	16	21
	6	1	0	56	0	6	22	0
	7	6	11	17	0	16	25	10

Tabela 6.7: Matriz de confusão para as sete classes de imagens.

Devido a este insucesso, decidimos investigar a possibilidade de utilizar um conjunto de redes neurais de classificação binária (reconhecendo duas classes apenas), para uma posterior classificação final utilizando um esquema de votação entre os múltiplos classificadores binários. Esta estratégia foi adotada levando-se em consideração que, quanto menor o número de classes mais facilidade a rede neural terá para construir a superfície de decisão. A seguir será apresentada a implementação inicial desta estratégia.

Implementação Inicial dos Classificadores Binários

As seis classes de placas foram combinadas duas a duas, resultando em 15 combinações. Para cada combinação de duas classes é utilizada uma rede. No treinamento das redes foram considerados os valores dos pixels das imagens, normalizados entre 0 e 1. As imagens foram convertidas para níveis de cinza e redimensionadas para um tamanho 20x20. Este pré-processamento teve como objetivo obter uma representação mais compacta da imagem, otimizando o treinamento das redes, uma vez que o número de neurônios de entrada foi reduzido de 1600 para 400. Isto foi realizado de forma empírica, observando que era possível, para um ser humano, reconhecer as placas mesmo após esta redução no tamanho. A camada escondida foi formada por 10 neurônios, com base na experiência apresentada na Seção 6.2, em que a melhor arquitetura foi a que teve número de neurônios na camada escondida igual a cinco vezes o número de classes. Ao utilizarmos a estratégia *winner-takes-all* para duas classes, a camada de saída foi formada por dois neurônios. A Figura 6.9 ilustra a arquitetura das redes de classificação binária utilizadas nos experimentos.

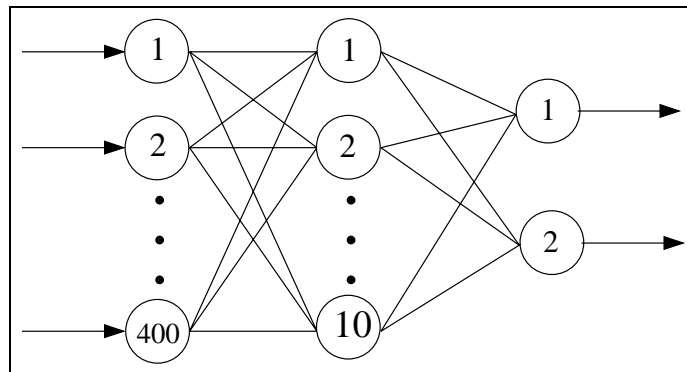


Figura 6.9: Representação da arquitetura das redes de classificação binária utilizadas nos experimentos.

Durante o treinamento, todas as redes atingiram SSE menor que 0.02 e os padrões foram 100% treinados. O maior número de épocas de treinamento foi 1184 para a rede treinada com as classes 3 e 5 e o menor número foi 286 para a rede treinada com as classes 1 e 2. Durante os testes, foram adotadas duas formas diferentes para analisar os resultados. Na primeira, leva-se em consideração que cada grupo de 17 imagens (micro-sacadas) corresponde ao mesmo objeto, o que na prática é verdade. Neste caso utiliza-se um critério de votação para que seja

contabilizado um acerto. Na segunda, são analisados os acertos absolutos da rede, ou seja, cada imagem é considerada como um objeto diferente.

Análise por Votação

Nesta análise, os votos são contabilizados a partir do seguinte critério: para cada resultado que a rede fornece é atribuído um voto à classe que corresponde ao neurônio vencedor. No final, compara-se o número de votos de todas as classes, elegendo como vencedora a classe que obtiver mais votos. Se a classe vencedora corresponder à classe dos padrões apresentados, considera-se que aquele objeto foi classificado. Caso contrário, se qualquer outra classe for a vencedora, considera-se que o objeto não foi classificado. As taxas percentuais de classificação para cada combinação de duas classes, são apresentadas na Tabela 6.8.

Índice do Classificador Binário	Classe - Classe	Taxa de acerto
1	1 - 2	90%
2	1 - 3	40%
3	1 - 4	60%
4	1 - 5	60%
5	1 - 6	60%
6	2 - 3	50%
7	2 - 4	80%
8	2 - 5	60%
9	2 - 6	100%
10	3 - 4	60%
11	3 - 5	60%
12	3 - 6	60%
13	4 - 5	40%
14	4 - 6	50%
15	5 - 6	80%

Tabela 6.8: Taxas percentuais de acerto das redes neurais binárias para cada combinação de duas classes, a partir da análise por votação.

Análise Absoluta

Nesta análise, cada uma das imagens que formam o conjunto foi tratada como um objeto diferente. Cada uma das seis classes é formada por 5 objetos dispostos em 17 regiões diferentes, resultando em 85 imagens por classe. Considerando cada imagem como um objeto diferente, teremos um conjunto total com 510 objetos. As taxas de acerto são calculadas a partir da resposta da rede para cada objeto. A Tabela 6.9 apresenta as taxas percentuais de acerto das redes a partir desta análise. O gráfico na Figura 6.10 mostra as taxas de acerto das redes para cada combinação de duas classes, a partir das duas análises realizadas. Mesmo aplicando metodologias de análise bem distintas, a comparação entre os resultados da análise absoluta e análise por votação nos mostra uma pequena vantagem da estratégia de análise por votação. A adoção desta metodologia é justificada pela utilização da estratégia de micro-sacadas, que faz com que cada imagem selecionada origine 16 novas imagens (num total de 17 imagens). Portanto, faz sentido tratar este bloco de 17 imagens como sendo o mesmo objeto.

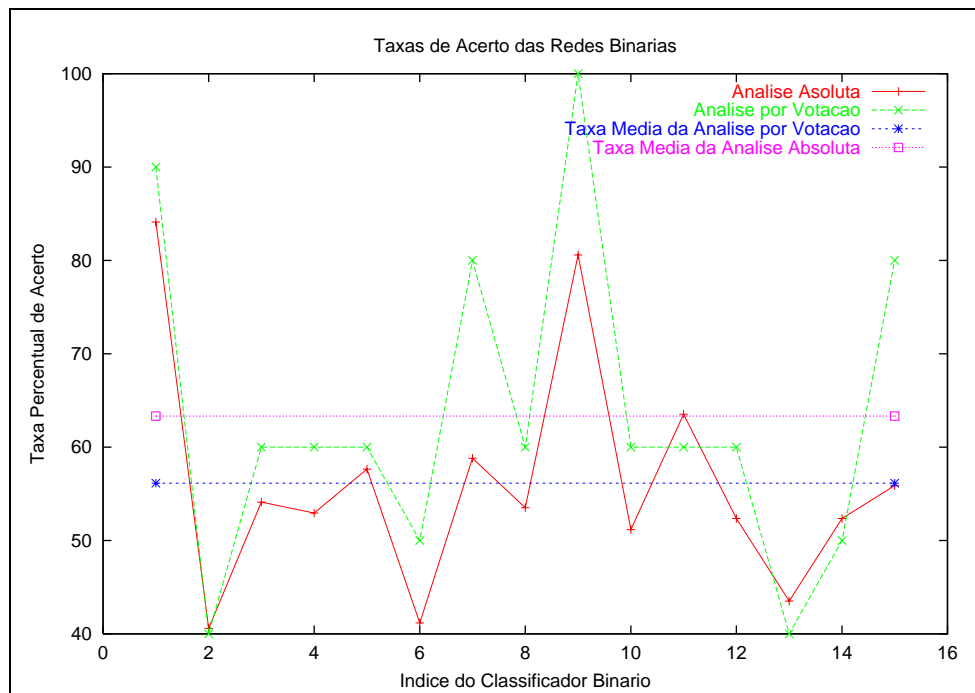


Figura 6.10: Taxas de acerto das redes binárias em cada combinação de duas classes, nas duas análises realizadas.

Índice do Classificador Binário	Classe - Classe	Taxa de acerto
1	1 - 2	84,12%
2	1 - 3	40,59%
3	1 - 4	54,12%
4	1 - 5	52,94%
5	1 - 6	57,64%
6	2 - 3	41,17%
7	2 - 4	58,82%
8	2 - 5	53,52%
9	2 - 6	80,58%
10	3 - 4	51,17%
11	3 - 5	63,53%
12	3 - 6	52,35%
13	4 - 5	43,52%
14	4 - 6	52,35%
15	5 - 6	55,88%

Tabela 6.9: Taxas percentuais de acerto das redes neurais binárias para cada combinação de duas classes, a partir da análise absoluta.

6.4.3 Análise dos Resultados

Os resultados alcançados demonstraram o desempenho do mecanismo de atenção *bottom-up* (ou baseado em saliências) que foi implementado neste trabalho, ao ser aplicado na localização de placas de sinalização em imagens de ruas e estradas. O problema ocorrido no primeiro experimento, em que uma placa foi inibida pela máscara de intensidades nulas, foi solucionado com uma mudança simples no raio da máscara. Esta mudança possibilitou localizar 100% das placas, nas imagens utilizadas no segundo experimento. Embora não tenha demonstrado grande influência no resultado final, a região de fronteira no horizonte (por exemplo, céu e vegetação) é, na maioria das imagens, bastante saliente. A redução no raio da máscara pode resultar em um grande número de pequenas regiões vizinhas atendidas. Neste caso, o número de regiões selecionadas até que o mecanismo localize uma placa é maior. A Figura 6.11 mostra um exemplo deste comportamento. Uma aquisição de imagens que conte com mais recursos materiais pode minimizar este problema. Um exemplo disto seria a utilização de câmeras que podem ser adaptadas em locais estratégicos do veículo. No nosso caso, a câmera foi instalada no interior do veículo, em um tripé na frente do assento dianteiro do passageiro. Além do problema das regiões de fronteira, também existem ruídos causados pelo pára-brisa do veículo, falta de estabilidade da câmera etc.

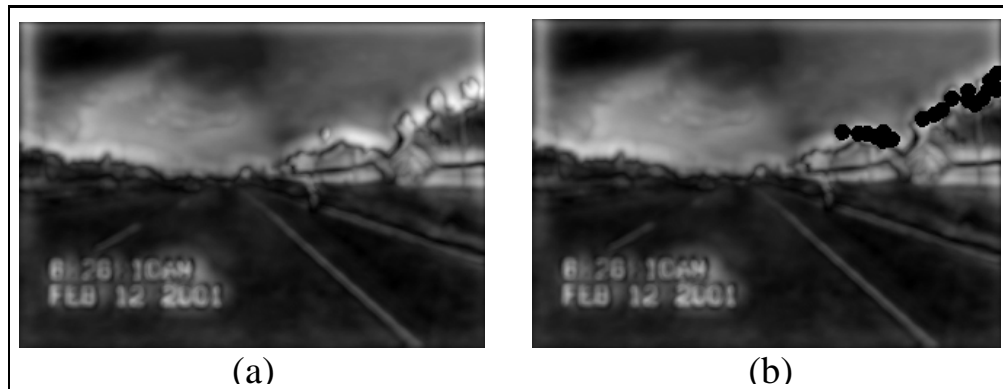


Figura 6.11: (a) Mapa de Saliência com várias regiões salientes na fronteira entre o céu e a vegetação, resultando em um grande número de regiões vizinhas selecionadas.

Embora não tenham apresentado resultados ótimos e a integração dos classificadores binários não tenha sido finalizada, os experimentos com o Módulo de Reconhecimento demonstraram que é possível classificar as regiões selecionadas pelo mecanismo de atenção

utilizando uma abordagem neural. Esperando-se que resultados mais expressivos possam ser alcançados com o aumento do número de padrões no conjunto de treinamento. Em consequência disto, teríamos um aumento no poder de generalização das redes. Além disso, a utilização de um pré-processamento mais completo e uma representação dos padrões através de um conjunto de características invariantes, pode também contribuir para o alcance de taxas de classificação melhores. A integração final dos classificadores binários é uma das propostas de trabalhos futuros, apresentadas no próximo capítulo.

6.5 Sumário

Os experimentos descritos neste capítulo tiveram como objetivo principal confirmar a aplicabilidade do mecanismo de atenção implementado à localização de placas de sinalização em imagens de estradas e ruas urbanas. Embora a base de imagens construída durante o trabalho apresente mais exemplos de imagens de estradas, foi possível observar que a quantidade de estruturas presentes nas imagens de ruas urbanas afeta o desempenho do Módulo de Detecção. Entretanto, ao analisarmos este desempenho do ponto de vista da quantidade de pontos selecionados em relação ao número total de pontos, notamos que o impacto no desempenho do Módulo de Detecção é bastante sensível. Os experimentos também objetivaram a investigação da viabilidade na utilização de Redes Neurais para a tarefa de classificação das placas de sinalização. Um artigo com os resultados da implementação do mecanismo de atenção foi aceito para publicação no SIBGRAPI'02 [Rodrigues and Gomes, 2002].

O desempenho do Módulo de Detecção foi demonstrado através dos resultados experimentais descritos neste capítulo. A inibição de uma placa em uma das imagens no primeiro experimento foi solucionada com uma simples modificação no raio da máscara de inibição. Embora esta mudança em geral cause um aumento no número de regiões atendidas antes da seleção de uma placa, os resultados dos experimentos mostram que o impacto sobre o desempenho final do Módulo de Detecção não é tão significativo. Mesmo não tendo alcançado resultados satisfatórios com o Módulo de Reconhecimento aplicado diretamente à saída do Módulo de Detecção, foi possível demonstrar a possibilidade de utilizar Redes Neurais como técnica de classificação para o tipo de padrão envolvido nos experimentos. Melhorias na arquitetura e propostas de trabalhos futuros serão discutidos no próximo capítulo.

Capítulo 7

Conclusões e Propostas de Trabalhos

Futuros

Esta dissertação apresentou um estudo e implementação de um mecanismo de atenção visual genérico e sua aplicação na localização de placas de sinalização em imagens adquiridas a partir de um veículo em movimento. Além disso, apresentou também uma investigação preliminar da aplicabilidade de Redes Neurais à tarefa de classificação das placas selecionadas, além de ter desenvolvido uma proposta de arquitetura para um sistema automático de detecção e reconhecimento das placas de sinalização. O caráter multidisciplinar do trabalho envolve áreas como Inteligência Artificial, Visão Computacional, Atenção Visual, Neurofisiologia etc. Estas áreas estão relacionadas com os modelos e técnicas estudados para a solução do problema da localização e do reconhecimento de placas de sinalização, que por sua vez está inserido no contexto dos Sistemas de Apoio ao Motorista.

Para formar o módulo responsável pela localização das placas dentro das imagens, foi construído um mecanismo de atenção visual baseado no conceito de saliência visual (características *bottom-up*). Este mecanismo implementa um Mapa de Saliência que codifica topograficamente as regiões salientes, ou regiões de interesse, em toda cena visual. A partir do mapa, uma estratégia de sacadas e micro-sacadas seleciona as regiões de interesse na imagem de entrada. A investigação de Redes Neurais para a tarefa de classificação das placas selecionadas se concentrou em torno de classificadores binários devido ao baixo desempenho de classificação obtido com uma rede monolítica multi-classe. Nesta abordagem, as classes são combinadas duas a duas, e para cada combinação, a tarefa de classificação é realizada por

uma rede diferente. A inspiração biológica foi uma das principais razões para a escolha dos modelos e técnicas estudados nesta dissertação.

7.1 Sumário da Dissertação

O Capítulo 2 apresentou um estudo sobre Sistemas de Apoio ao Motorista (DSS) e uma revisão de alguns dos principais trabalhos da área. Os trabalhos apresentados foram divididos de acordo com as tarefas que buscam solucionar, são eles: detecção de obstáculos, detecção e reconhecimento de sinais de tráfego e detecção das marcas da estrada. Além disso, o capítulo apresenta um trabalho desenvolvido na Universidade de Parma, Itália, em que alguns módulos inerentes aos Sistemas de Apoio ao Motorista são integrados para a construção de um veículo autônomo. Por fim, foram apresentados dois trabalhos que discutem alguns aspectos importantes: no primeiro são definidas rotinas visuais envolvidas na construção de DSS's e no segundo são discutidos aspectos relativos às condições de tempo que devem ser levados em conta durante a construção de sistemas baseados em visão que funcionam ao ar livre.

O Capítulo 3 apresentou um estudo sobre a atenção visual a partir dos pontos de vista neurofisiológico e computacional. Na primeira parte, foram estudadas as estruturas neurais envolvidas na atenção visual dentre elas: a retina que têm como uma de suas principais características a distribuição não uniforme de fotoreceptores, em que a região central, conhecida como fóvea, é responsável pela visão mais acurada; o córtex visual onde cerca de 90% da informação visual é processada; e o núcleo geniculado lateral que serve de ponte entre a retina e o córtex visual. Na segunda parte do capítulo, a atenção visual do ponto de vista computacional foi estudada. Para isto, foi seguida a abordagem *bottom-up*, que é baseada em informações da própria imagem de entrada. Além dos principais conceitos envolvidos, também foi apresentada uma descrição resumida dos trabalhos mais relevantes da área.

O Capítulo 4 apresentou um estudo sobre os principais fundamentos das Redes Neurais Artificiais. Apresentou ainda um breve histórico que destaca os fatos mais importantes da área. Discutiu alguns dos principais modelos de Redes Neurais e seus respectivos algoritmos de treinamento. Além disso, fez uma pequena introdução ao SNNS, simulador de Redes Neurais desenvolvido na Universidade de Stuttgart e que foi utilizado neste trabalho de dis-

sertação.

O Capítulo 5 descreveu a arquitetura do modelo proposto a partir do embasamento teórico adquirido nos capítulos anteriores. Inicialmente, foi apresentada a arquitetura geral do protótipo e em seguida foram detalhados os dois principais módulos. O primeiro é o Módulo de Detecção, que utiliza um mecanismo de atenção visual para selecionar as regiões de interesse na cena. Este mecanismo busca inspiração biológica para a implementação de todos os processos como, as sacadas e micro-sacadas que foram inspiradas nos movimentos oculares. Outro exemplo disto são os processos de extração de características visuais primitivas e de diferenças centro-vizinhança, que são inspirados na existência de neurônios do córtex visual especializados em perceber características semelhantes a estas. Neste trabalho de dissertação, focaliza-se o interesse nas seguintes características: intensidades, cores e bordas. O segundo módulo é o de Reconhecimento, constituído de Redes Neurais de classificação binária. Neste caso, a inspiração para utilizar a técnica de Redes Neurais é a própria organização do cérebro e o funcionamento dos neurônios.

O Capítulo 6 descreveu os experimentos realizados com os dois módulos do protótipo e apresentou os resultados destes experimentos. Os objetivos principais dos experimentos foram: (1) confirmar a aplicabilidade do Módulo de Detecção à localização de placas de sinalização em imagens reais de ruas e (2) investigar a viabilidade na utilização de Redes Neurais para a tarefa de classificação das placas localizadas. Ambos os objetivos foram alcançados, uma vez que os resultados apresentados demonstram o poder do Módulo de Detecção e a possibilidade de classificar as placas utilizando Redes Neurais. Entretanto, para conseguir melhores taxas de classificação, serão necessárias algumas mudanças discutidas a seguir.

7.2 Considerações Gerais

Diante dos resultados experimentais obtidos, principalmente com o Módulo de Detecção, consideramos que os objetivos principais deste trabalho de dissertação foram em sua maioria alcançados. É importante ressaltar que a implementação do protótipo não foi realizada rápida e facilmente, mas sim em um processo árduo e contínuo. Outro ponto importante é o nível experimental em que a área de atenção visual se encontra, o que resultou em dificuldades

durante a pesquisa.

No primeiro experimento realizado, o Módulo de Detecção foi capaz de localizar 93,75% das placas presentes nas imagens. Após um pequeno ajuste na configuração da máscara de inibição, o percentual de localização subiu para 100%. Entretanto, novas melhorias devem ser incorporadas para que o protótipo possa funcionar, por exemplo em tempo real. Algumas dessas melhorias serão apresentadas como propostas de trabalhos futuros na próxima seção.

Mesmo não tendo alcançado melhores taxas de classificação ao utilizar uma arquitetura monolítica, foi possível perceber a capacidade das Redes Neurais para realizarem tarefas desta natureza. Deve-se levar em conta a qualidade dos padrões de treinamento e teste, na análise dos resultados finais. Durante os experimentos, a intenção inicial foi utilizar conjuntos de padrões complicados para avaliar o impacto das características reais da imagem no desempenho da técnica. A complexidade do conjunto de padrões está relacionada com algumas características das imagens utilizadas nos experimentos, como por exemplo o seu tamanho. O fato de não extrairmos características invariantes das imagens resultou em um espaço de entrada muito grande, da ordem de 400 pixels, o que requer um número de padrões de treinamento proporcionalmente alto para cobrir adequadamente o todo o espaço de possibilidades de padrões. Entretanto, devido às limitações de recursos, uma aquisição exaustiva de imagens contendo placas de sinalização não foi possível. Outro problema que contribuiu para esta complexidade foi a translação das placas nas imagens. Para tratar este problema é necessário que a técnica tenha o poder de invariância quanto a esta característica. A própria arquitetura MLP-BP investigada neste trabalho enfrenta problemas de generalização dos padrões, com respeito a translação.

Uma dificuldade encontrada durante o trabalho foi a limitação dos recursos materiais disponíveis, principalmente no que diz respeito à aquisição das imagens. Durante a pesquisa bibliográfica foi possível notar que os grupos que apresentam os melhores resultados nas pesquisas em DSS são aqueles que possuem um sólido aporte financeiro. Em sua maioria, estes grupos contaram com uma infraestrutura básica desde o início do projeto. Contando com este apoio, eles dispuseram de recursos como: veículos, câmeras apropriadas, processadores paralelos, hardwares dedicados, computadores de bordo etc. Além disso, trabalham em países, cujas rodovias são bem conservadas, bem sinalizadas e com pouca poluição visual.

O desempenho e a eficiência de um Sistema de Apoio ao Motorista depende criticamente da sua capacidade de responder em tempo real. No caso do protótipo investigado neste trabalho de dissertação, o tempo real pode ser entendido como o tempo necessário para que o sistema forneça alguma informação ao motorista e este possa tomar alguma decisão. Um dos pontos que devem ser levados em conta é a velocidade do veículo. Dependendo desta velocidade, um sistema pode ser eficiente ou não, já que o espaço de tempo necessário para a tomada de decisão é inversamente proporcional a velocidade do veículo. Uma análise dos pontos críticos envolvidos nesta questão será apresentada como proposta de trabalho futuro.

7.3 Propostas de Trabalhos Futuros

A partir das questões discutidas anteriormente, algumas sugestões de trabalhos futuros são:

- Finalizar a integração dos classificadores binários, com o objetivo de concluir a implementação do Módulo de Reconhecimento;
- Adquirir mais imagens para enriquecer os conjuntos de padrões e realizar novos experimentos com o Módulo de Detecção utilizando estes novos padrões;
- Investigar a implementação do Módulo de Detecção em uma arquitetura de hardware dedicada, como por exemplo uma arquitetura baseada em circuitos reconfiguráveis (FPGA, PLD etc);
- Utilizar um pré-processamento mais completo e poderoso das imagens, bem como uma representação mais compacta, diminuindo a dimensionalidade do espaço de características dos padrões. Em seguida, realizar novos experimentos com o Módulo de Reconhecimento, utilizando estes padrões.
- Investigar novas arquiteturas neurais para a tarefa de classificação. Um ponto de partida pode ser o trabalho do Kröner [Kröner, 1996], que propôs uma arquitetura neural adaptativa para calcular características da imagem invariantes a rotação e translação;
- Realizar análise comparativa com uma ou mais técnicas utilizadas nos trabalhos estudados nesta dissertação como por exemplo, *Simulated Annealing*;

- Investigar a aplicabilidade do pré-processamento e da segmentação utilizados no Módulo de Detecção nos padrões de treinamento e teste do Módulo de Reconhecimento, para a eliminação de ruído, extração de características invariantes etc;
- Investigar o uso de Algoritmos Genéticos para a tarefa final de selecionar as regiões de interesse, a partir do Mapa de Saliência;
- Implementar o modelo a partir de uma metodologia de Co-design;
- Investigar a utilização de outros filtros (ex. mediana, adaptativa etc) que realçam bordas e homogenizam regiões entre as bordas, para gerar imagens com representação mais adequada para a tarefa de reconhecimento.
- Investigar a portabilidade do protótipo e realizar análise de custo de um produto final.
- Elaborar uma análise das principais limitações no uso do sistema em tempo real. Definir quais os pontos responsáveis pelas restrições de tempo.
- Realizar estudo comparativo entre os modelos de atenção visual apresentados neste trabalho, aplicados ao problema da localização de placas de sinalização.

Bibliografia

- [AGLAIA, 2002] AGLAIA (2002). Mobile vision system. Disponível on-line em: www.aglaia-gmbh.de/.
- [Andersen and Martinez, 1999] Andersen, T. and Martinez, T. (1999). Cross validation and mlp architecture. In *Proceedings of the IEEE International Joint Conference on Artificial Neural Networks and Genetic Algorithms (ICANNGA'99)*, pages 22–27.
- [ARGO, 2001] ARGO (2001). The argo projec. Università di Parma, Dip. di Ingegneria dell'Informazione. nannetta.ce.unipr.it/ARGO/english.
- [Barros et al., 1999] Barros, A. S., Soares, H. B., Dória, A. D., and Carvalho, M. A. G. (1999). Segmentation of abnormalities in digital mammograms. In *International Conference on Engeneering and Computer Education*. Rio de Janeiro/RJ.
- [Beale and Jackson, 1990] Beale, R. and Jackson, T. (1990). *Neural Computing: An Introduction*. Institute of Physics Publishing Bristol and Philadelphia.
- [Bertozzi et al., 1999] Bertozzi, M., Broggi, A., and Fascioli, A. (1999). The argo autonomous vehicle. In *Atti del 6 Congresso dell'Associazione Italiana per l'Intelligenza Artificiale*, pages 503–506.
- [Betke and Makris, 1995] Betke, M. and Makris, N. C. (1995). Fast object recognition in noisy images using simulated annealing. In *Proceedings of the Fifth International Conference on Computer Vision*, pages 523–530.
- [Bishop, 1995] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press. Oxford.

- [Borenstein and Koren, 1991] Borenstein, J. and Koren, Y. (1991). The vector field histogram - fast obstacle avoidance for mobile robots. *IEEE Transactions on Robotics and Automation*, 7(3):278–288.
- [Broggi, 1995a] Broggi, A. (1995a). A massively parallel approach to real-time vision-based road markings detection. In Masaky, I., editor, *Proceedings of IEEE Intelligent Vehicles'95*, pages 84–89. IEEE Computer Society. Detroit.
- [Broggi, 1995b] Broggi, A. (1995b). Robust real-time lane and road detection in critical shadow conditions. In *Proceedings of IEEE International Symposium on Computer Vision*, pages 353–358. IEEE Computer Society. Coral Gables, Florida.
- [Broggi et al., 1999] Broggi, A., Bertozzi, M., Fascioli, A., and Conte, G. (1999). Automatic vehicle guidance: the experience of the argo autonomous vehicle. *World Scientific Co. Publisher*. Singapore.
- [Broggi et al., 1994] Broggi, A., Conte, G., Gregoretti, F., Sansoè, C., and L.M.Reyneri (1994). The paprica massively parallel processor. In *Proceedings of IEEE Intl. Conf. on Massively Parallel Computing Systems*, pages 16–30.
- [Burt and Adelson, 1983] Burt, P. J. and Adelson, E. H. (1983). The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, COM-31(4):532–540.
- [Chen and Hagan, 1999] Chen, D. and Hagan, M. T. (1999). Optimal use of regularization and cross-validation in neural network. In *Proceedings of the 1999 International Joint Conference on Neural Networks*, pages 1275–1280. vol 2.
- [Churchland and Sejnowski, 1992] Churchland, P. S. and Sejnowski, T. J. (1992). *The Computational Brain*. The MIT Press. Cambridge, Massachusetts.
- [Craven and Shavlik, 1997] Craven, M. W. and Shavlik, J. W. (1997). Using neural networks for data mining. *Future Generation Computer Systems*, (2–3):211–229.
- [Culhane and Tsotsos, 1992] Culhane, S. M. and Tsotsos, J. K. (1992). A prototype for data-driven visual attention. In *Proceedings of the 11th IAPR International Conference on Pattern Recognition, The Hague, The Netherlands*, volume 1, pages 36–40. IEEE Computer Society Press. Los Alamitos, California.

- [D. E. Rumelhart and Williams, 1986] D. E. Rumelhart, H. and Williams, R. J. (1986). Learning internal representations by error propagation. *Parallel Distrib. Processing*, (1):318–62.
- [de A. Barreto et al., 2001] de A. Barreto, G., Araújo, A. F. R., Dücker, C., and Ritter, H. (2001). Implementation of a distributed robotic control system based on a temporal self-organizing neural network. In *Proceedings of the IEEE International Conference on System, Man and Cybern (SMC'01)*, pages 335–340. Tucson, Arizona.
- [Duda et al., 2000] Duda, R. O., Hart, P. E., and Storck, D. (2000). *Pattern Classification*. Wiley-Interscience. 2nd edition.
- [Freeman and Adelson, 1991] Freeman, W. T. and Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Trans. Patt. Anal. and Machine Intell.*, 13(9):891–906.
- [Gavrila and Philomin, 1999] Gavrila, D. M. and Philomin, V. (1999). Real-time object detection for smart vehicles. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 87–93. Computer Society Press.
- [Gomes et al., 1996] Gomes, H. M., Machado, P. D. L., and Filho, E. C. B. C. (1996). Investigation of techniques for off-line signature recognition. In *Proceedings of International Symposium on Systems Analysis and Synthesis (ISAS'96)*. Orlando, USA.
- [Gonçalves, 1999] Gonçalves, S. E. (1999). *Reconhecimento Visual Atencional*. PhD thesis, COPPE/UFRJ, D.Sc., Engenharia de Sistemas e Computação.
- [Gonzalez and Woods, 1992] Gonzalez, R. C. and Woods, R. E. (1992). *Digital Image Processing*. Addison-Wesley Publishing Company, Inc.
- [Greenspan et al., 1994] Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., and Anderson, C. H. (1994). Overcomplete steerable pyramid filters and rotation invariance. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 222–228. Seattle, Washington,.
- [Guyton and Hall, 1997] Guyton, A. C. and Hall, J. E. (1997). *Tratado de Fisiologia Médica*. Ed. Guanabara Koogan S.A. Rio de Janeiro/RJ.

- [Haykin, 1999] Haykin, S. (1999). *Neural Networks. A Comprehensive Foundation*. Prentice Hall. 2nd edition.
- [Itti and Koch, 2001] Itti, L. and Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203.
- [Itti et al., 1998] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.
- [Karasaridis and Simoncelli, 1996] Karasaridis, A. and Simoncelli, E. P. (1996). A filter design technique for steerable pyramid image transforms. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)*. Atlanta, GA.
- [Koch, 2000] Koch, C. (2000). Selective visual attention and computational models. *Computational Neuro-Science* 186. March 2.
- [Koch and Ullman, 1985] Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum. Neurobiol*, 4:219–227.
- [Kröner, 1996] Kröner, S. (1996). A neural network for calculating adaptive shift and rotation invariance image features. In *Proceedings of the European Signal Processing Conference*, volume II, pages 863–866. Trieste, Italy.
- [Little, 2001] Little, C. (2001). Disponível on-line em: <http://www.tfhr.gov/pubrds/pr97-10/p18.htm>.
- [McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bull Math Biophys*, (2):115–133.
- [Milanese, 1993] Milanese, R. (1993). *Detecting Salient Regions in an Image: from Biological Evidence to Computer Implementation*. PhD thesis, University of Genova.
- [Milanese et al., 1994] Milanese, R., Wechsler, H., Gill, S., Bost, J. M., and Pun, T. (1994). Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Proc. of ARPA Image Understanding Workshop*, pages 781–785.

- [Nayar and Narasimhan, 1999] Nayar, S. K. and Narasimhan, S. G. (1999). Vision in bad weather. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 820–827. Computer Society Press.
- [Piccioli et al., 1996] Piccioli, G., Micheli, E., Parodi, P., and Campani, M. (1996). Robust method for road sign detection and recognition. *IVC*, 14(3):209–259.
- [Priebe et al., 1993] Priebe, L., Rehrmann, V., Schian, R., and Lakmann, R. (1993). Traffic sign recognition based on color image evaluation. In *Proc. Intelligent Vehicles Symposium 93*, pages 95–100.
- [Rehrmann et al., 1995] Rehrmann, V., Lakmann, R., and Priebe, L. (1995). A parallel system for realtime traffic sign recognition. In *Proceedings International Workshop on Advanced Parallel Processing Technologies'95 (APPT)*, pages 72–78. Publishing House of Electronics Industry. Peking.
- [Rodrigues and Gomes, 2002] Rodrigues, F. A. and Gomes, H. M. (2002). Applying a visual attention mechanism to the problem of traffic sign recognition. Aceito para publicação no SIBGRAPI'02.
- [Rowley et al., 1998] Rowley, H., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38.
- [Salgian and Ballard, 1998] Salgian, G. and Ballard, D. H. (1998). Visual routines for autonomous driving. In *Proceedings of Sixth International Conference on Computer Vision*, pages 876–882. Norosa Publishing House. Bombay India.
- [Sela and Levine, 1997] Sela, G. and Levine, M. D. (1997). Real-time attention for robotic vision. *Real-Time Imaging*, 3:173–194.
- [Shepherd, 1994] Shepherd, G. M. (1994). *Neurobiology*. Oxford University Press. Third Edition.
- [Simoncelli and Freeman, 1995] Simoncelli, E. P. and Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *2nd Annual IEEE International Conference on Image Processing*. Washington, DC.

- [Tafner et al., 1995] Tafner, M. A., Xerez, M., and Filho, I. W. R. (1995). *Redes Neurais Artificiais: Introdução e Princípios de Neurocomputação*. Blumenau: EKO: Editora da FURB, 11^a edição edition.
- [Tsotsos, 1990] Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *The Behavioral and Brain Sciences*, 13:423–469.
- [Tsotsos et al., 1995] Tsotsos, J. K., Culhane, S., Wai, W., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, pages 507–547. 78(1-2).
- [Tu and Li, 1999] Tu, Z. and Li, R. (1999). A multilayer hopfield neural network for 3-d object recognition. In *Proceedings of International Mobile Mapping Workshop*, pages 7A.3.1–6. Bangkok, Thailand.
- [Waltz and Miller, 1998] Waltz, F. M. and Miller, J. W. V. (1998). An efficient algorithm for gaussian blur using finite-state machines. In *Conf. on Machine Vision Systems for Inspection and Metrology (SPIE VII)*, pages 3521–3537. Boston.
- [Yuk and Flanagan, 1999] Yuk, D. and Flanagan, J. (1999). Telephone speech recognition using neural network. In *Proceedings of the IEEE International Conference on Acoustics and Signal Processing*, volume 1, pages 157–160.

Apêndice A

Base de Imagens

A.1 Imagens Utilizadas no Primeiro Experimento

Conjunto de imagens utilizadas no primeiro experimento com o módulo de atenção:



Tabela A.1: Imagens de 1 a 9 utilizadas no primeiro experimento.









 <p>6:37:28AM FEB 12 2001</p>	 <p>6:38:43AM FEB 12 2001</p>
imagem 5	imagem 6
 <p>6:39:30AM FEB 12 2001</p>	 <p>6:54:57AM FEB 12 2001</p>
imagem 7	imagem 8
 <p>7:01:24AM FEB 12 2001</p>	 <p>7:03:56AM FEB 12 2001</p>
imagem 9	imagem 10
 <p>7:04:00AM FEB 12 2001</p>	 <p>7:04:01AM FEB 12 2001</p>
imagem 11	imagem 12

Tabela A.2: Imagens de 5 a 12 utilizadas no primeiro experimento.



Tabela A.3: Imagens de 13 a 15 utilizadas no primeiro experimento.

Exemplos de imagens utilizadas nos experimentos com a simulação dos módulos de imagem. Estão presentes no exemplo o último quadro de cada ocorrência.

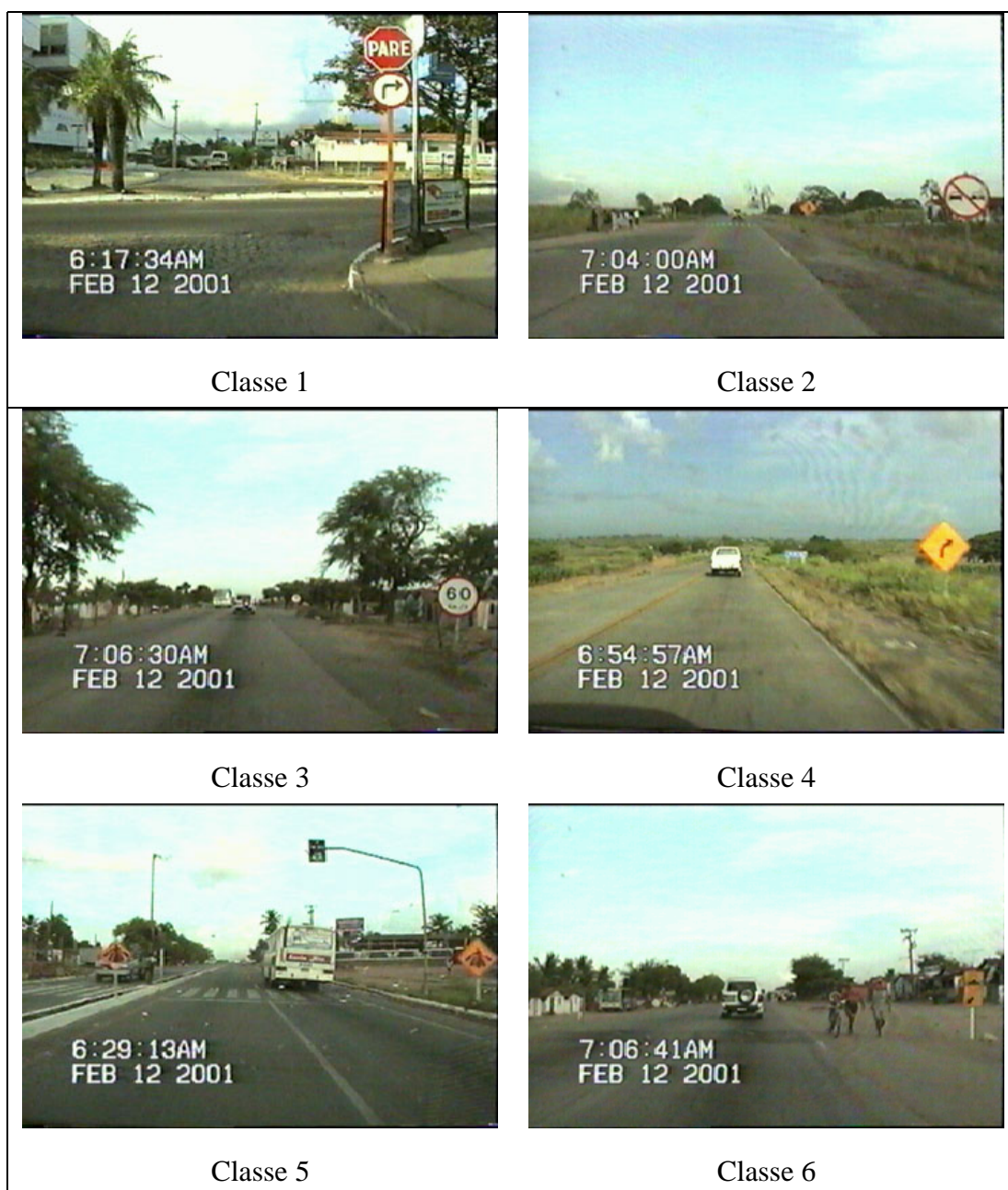


Tabela A.4: Imagens utilizadas no treinamento.







 <p>6:19:44AM FEB 12 2001</p>	 <p>7:01:24AM FEB 12 2001</p>
Classe 1	Classe 2
 <p>7:03:56AM FEB 12 2001</p>	 <p>7:01:27AM FEB 12 2001</p>
Classe 3	Classe 4
 <p>6:26:10AM FEB 12 2001</p>	 <p>7:06:27AM FEB 12 2001</p>
Classe 5	Classe 6

Tabela A.5: Imagens utilizadas nos testes.