Localização de Placas de Sinalização em Imagens de Ruas e Estradas Utilizando um Mecanismo de Atenção Visual*

Fabrício A. Rodrigues
Herman M. Gomes

Departamento de Sistemas e Computação – Universidade Federal da Paraíba
Av. Aprígio Veloso s/n, 58109-970, Campina Grande/PB
{fabre,hmg}@dsc.ufpb.br

Abstract

This paper presents the preliminary results of the application of a saliency-based attention mechanism to the problem of traffic sign recognition. This attention mechanism implements a Saliency Map that is constructed from three image characteristics: intensities, colors and orientations. These characteristics are extracted and represented by a Gaussian Pyramid (for intensities and colors) and by a Steerable Pyramid (for orientations). Subtraction between pyramid levels produces feature maps that are combined to form the Saliency Map. This map is used to guide the selection of the most interesting regions and an inhibition mask is used to prevent repeatedly analysing a same image region. The attention mechanism constitutes the Detection Module of a traffic sign recognition system. Results have shown that the attention machanism allowed a drastic reduction in the number of points to be analysed in each image frame.

Keywords: Visual Attention, Driver Support Systems, Traffic Sign Recognition

Resumo

Este Artigo apresenta os resultados preliminares da aplicação de um mecanismo de atenção visual ao problema do reconhecimento de sinais de tráfego. Este mecanismo implementa um Mapa de saliência que é construído a partir de três características da imagem: intensidades, cores e orientações. Estas características são extraídas e representadas por uma Pirâmide Gaussiana (para intensidades e cores) e por uma Pirâmide Direcional (para orientações). Subtração entre os níveis da pirâmide produz mapas de características que são combinados para formar o Mapa de Saliência. Este mapa é usado para guiar a seleção das regiões de maior interesse e uma máscara de inibição é usada para impedir que uma região seja analisada mais de uma vez. O mecanismo de atenção apresentado neste trabalho constitui o Módulo de Detecção de um sistema de reconhecimento de placas de sinalização. Os resultados mostraram que o mecanismo de atenção permitiu uma drástica redução no número de pontos a serem analisados em cada imagem.

Palavras-chave: Atenção Visual, Sistemas de Apoio ao Motorista, Reconhecimento de Sinais de Tráfego.

*Este trabalho recebeu suporte financeiro da CAPES

1 Introdução

Dirigir um veículo é uma tarefa intensiva de processamento de informação visual, na qual o reconhecimento de sinais de tráfego ocupa um papel fundamental. Relatos mostram que uma grande quantidade de colisões em cruzamentos e choques frontais de veículos seriam evitados se o motorista tivesse meio segundo adicional para reagir, e que a falta de atenção é a causa de muitos acidentes [9]. O trabalho apresentado neste artigo é parte de um projeto maior que objetiva construir um sistema para localizar e reconhecer placas de sinalização em imagens capturadas de um veículo em movimento.

Este trabalho está inserido no contexto dos Sistemas de Apoio ao Motorista (*Driver Support Systems* - DSS) [11]. A principal função de um DSS é aumentar a segurança e o conforto de dirigir. O sistema pode informar, por exemplo, a presença de animais na pista, o limite de velocidade permitido, condições anormais da estrada etc, aumentando efetivamente a segurança do motorista e dos passageiros. Podemos citar como tarefas de um DSS: a detecção das marcas da estrada (limites da estrada); a detecção e o reconhecimento de sinais de tráfego (palcas, semáforos, sinais pintados na pista, etc.); e a detecção de obstáculos (veículos, pedestres, animais, etc.). 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.

Uma característica fundamental dos sistemas visuais biológicos é a capacidade de detectar rapidamente partes interessantes no estímulo visual de entrada [10]. Esta habilidade, chamada de atenção visual, é uma maneira de reduzir a quantidade de informação visual de entrada para um tamanho manejável, de tal forma que tarefas com processamento complexo possam ser realizadas por recursos computacionais limitados [7]. Em sistemas computacionais baseados em visão, os mecanismos de atenção visual são responsáveis por alocar recursos de processamento para as regiões mais significativas da cena visual. Neste caso, apenas a informação essencial na realização da tarefa é analisada, e a grande quantidade de detalhes irrelevantes é ignorada. A principal vantagem na utilização deste tipo de mecanismo é o aumento significativo na eficiência do sistema.

Neste trabalho, foi utilizado um mecanismo de atenção visual para compor o Módulo de Detecção do sistema responsável por selecionar as regiões de interesse nas imagens de entrada. Este artigo aborda a construção deste mecanismo de atenção, além dos resultados preliminares da sua aplicação na localização de placas de sinalização em imagens de ruas e estrada. Na próxima seção descreveremos brevemente alguns trabalhos revisados na área de atenção visual. Na Seção 3 apresentamos a arquitetura geral do sistema, e na Seção 4 destacamos o Módulo de Detecção. Os experimenos com este módulo e os resultados alcançados são apresentados na Seção 5. Finalmente, na Seção 6, são apresentadas as conclusões e os próximos passos do projeto.

2 Trabalhos Correlatos

A maioria dos modelos de atenção *bottom-up* segue a hipótese de Koch e Ullman [8], onde vários mapas de características alimentam um único mapa mestre ou mapa de saliência. O que diferencia estes modelos são as estratégias utilizadas para representar e extrair a saliência. O mapa de saliência é um mapa escalar bi-dimensional cuja atividade representa topograficamente a saliência visual [4]. 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. A seguir descreveremos alguns trabalhos mais específicos na área de atenção visual.

Tsotsos e Culhane [2] 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 é compreendido de 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 "brilha" através de todos os níveis da hierarquia. 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* (WTA). Embora trabalhe com a idéia de saliência, usando medidas como: brilho, movimento, contraste, curvatura, cor, linhas retas longas e outras, o modelo não implementa efetivamente um mapa de saliência.

Uma outra arquitetura, proposta por Milanese e colegas [10], utiliza uma estratégia híbrida, integrando indícios *bottom-up* e *top-down*, e 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. Uma técnica de reconhecimento de objetos baseada em Memória Associativa Distribuída (*Distributed Associative Memory* - DAM) é usada 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 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 detectar 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 provoca um movimento atencional.

No trabalho de Sela e Levine [13] é apresentado um sistema de atenção visual *bottom-up* para guiar a fixação de um sensor inspirado na arquitetura da retina humana. 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. Para encontrar as linhas de simetria e suas orientações, é utilizada uma medida de simetria baseada nos centros das arestas co-circulares. Duas arestas são co-circulares se um círculo pode ser desenhado de tal forma que elas sejam tangentes. O centro da co-circularidade é definido como o ponto central deste círculo, assim como o raio da co-circularidade é definido como o raio do círculo. Para a validação desta abordagem, os pontos de intersecção das linhas de simetria são comparados com as descobertas psicofísicas apresentadas no trabalho de Kaufman e Richards [6]. Para cada possível ponto de fixação é atribuída uma magnitude, que depende de dois fatores: (1) Os ângulos que separam as arestas que contribuem com as linhas de simetria e (2) O grau de "fechamento" de cada ponto. Os pontos de alta saliência são agrupados, e um simples ponto central é determinado para a fixação. O algoritmo computa pontos de interesse tanto para uma imagem foveal mapeada em coordenados retangulares, quanto para uma imagem periférica mapeada em coordenadas *log-polar*. O sistema foi implementado em uma rede paralela de processadores, que facilita um desempenho próximo ao tempo real, e foi testado no reconhecimento de faces e com imagens de cenas ao ar livre.

Um modelo de atenção visual baseado em saliência para analise de cenas rápidas foi proposto por Itti e colegas [5]. Neste modelo, a entrada visual é decomposta em um conjunto de mapas de características (ex. 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 de Centro-Vizinhança (*center-surround*) semelhantes a campos receptivos visuais. As operações de Centro-Vizinhança são implementadas como diferenças entre escalas finas e grossas de uma representação piramidal. A normalização e a combinação através da escala dos mapas de características, produz, para cada característica, um mapa de conspicuidade. A diferença através da escala entre dois mapas é obtida pela interpolação para a escala mais fina e pela subtração ponto-a-ponto. Os mapas de conspicuidade são normalizados e somados, produzindo uma entrada final para o mapa de saliência. Este mapa é dotado de dinâmicas internas que provocam movimentos atencionais. Em qualquer dado tempo, uma rede neural *winner-take-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, tal que o sistema desloca o foco para a próxima região mais saliente, favorecendo o processo de busca.

3 Arquitetura Proposta

Para a construção do nosso sistema, nós propusemos uma arquitetura híbrida contedo dois módulos principais: um para localizar regiões de interesse na imagem, onde existem maiores probabilidades de se encontrar as placas de sinalização (Módulo de Detecção), e outro para o classificar as regiões selecionadas (Módulo de Reconhecimento), conforme mostra a Figura 1. Este artigo enfoca o Módulo de Deteção.

O Módulo de Detecção do sistema utiliza um mecanismo de atenção visual para localizar as regiões de interesse na imagem de entrada. Após a pesquisa bibliográfica decidimos adaptar o modelo posposto por Itti e colegas [5] 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 *Bottom-up*, um Mapa de Saliência é gerado e depois usado como base para a seleção das regiões de interesse. O ponto principal é que o mapa codifica o fato de um dada região ser saliente, não importando qual característica a torna saliente. Portanto, a utilização de tal mecanismo, torna possível uma futura adaptação do sistema para detectar e

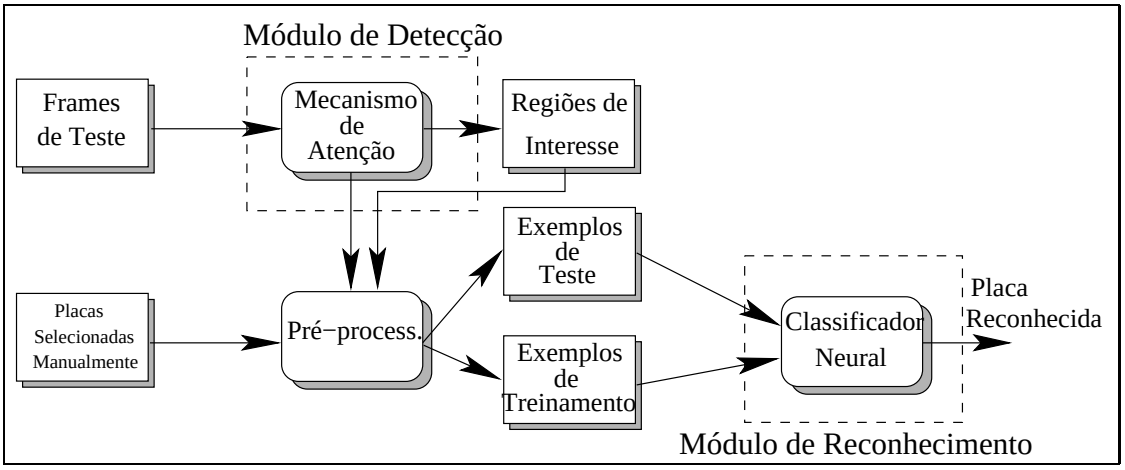

Figura 1: Arquitetura geral do sistema: os retângulos representam os dados e os retângulos arredondados representam os processos

reconhecer outros tipos de objetos.

O contexto do problema investigado envolve padrões previamente conhecidos (placas de sinalização) e a necessidade de um sistema capaz de aprender tais padrões. Sendo assim, a aplicação de uma técnica que utiliza uma aprendizagem supervisionada é possível. Diante disso, decidimos utilizar uma Rede Neural *Multilayer Perceptron* com algoritmo de treinamento *Backpropagation* (MLP-BP), para a tarefa de classificação (Módulo de Reconhecimento), por se tratar de uma técnica de classificação tradicional e de fácil utilização.

A Rede MLP-BP utilizada é formada por três camadas. O número de neurônios na camada de entrada é definido em função do tamanho das imagens (regiões selecionadas pelo Módulo de Detecção). Para a definir o número de neurônios na camada escondida foi realizado um experimento simples que determinou a melhor configuração. A camada de saída tem número de neurônios correspondente ao número de classes que formam os conjuntos de treinamento e teste. 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. O Módulo de Reconhecimento é formado por vários classificadores binários, onde cada classificador é uma Rede MLP-BP treinada para duas classes de imagens. A partir de uma estratégia de votação que utiliza as respostas dos classificadores binários, o Módulo de Reconhecimento indica a qual classe pertence a região selecionada pelo Módulo de Detecção. Maiores detalhes sobre o Módulo de Reconhecimento podem ser encontrados em [12].

Filtros de pré-processamento são aplicados, tanto nas imagens selecionadas manualmente e que foram utilizadas para treinar o classificador neural, quanto nas regiões selecionadas pelo Módulo de Detecção. Este pré-processamento segue a seguinte ordem: conversão das imagens para níveis de cinza, aplicação de um *blur* gaussiano e equalização de histograma. Na próxima seção serão discutidos os detalhes de implementação do Módulo de Detecção, o qual é o foco principal deste trabalho..

4 O Módulo de Detecção

Para diminuir a complexidade da busca visual, é essencial a utilização de uma estratégia capaz de localizar automaticamente objetos de interesse dentro de uma cena. Tsotsos [15] mostrou que a maior contribuição para a diminuição dessa complexidade, em sistemas baseados em visão, é a atenção visual. A Figura 2 mostra a arquitetura do Módulo de Detecção do nosso sistema, que é uma adaptação do modelo proposto por Itti e colegas [5].

A entrada para este módulo é provida na forma de imagens coloridas, digitalizadas em uma área de 352X240 pixels (Figura 3(a)). O primeiro processo realizado é uma filtragem linear, que extrai características visuais primitivas das imagens. Três tipos de características são extraídas: intensidades, cores e orientações. A imagem de intensidades, definida como I, é a imagem de entrada em níveis de cinza. A partir do sistema RGB, quatro imagens de canais de cores são extraídas: $R = r - (g + b)/2$ 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. A Figura 3 mostra o exemplo de uma imagem de intensidades e dos canais de cores extraídos a partir da imagem de entrada. Tanto para a imagem de intensidades quanto para as imagens de canais de cores são geradas pirâmides gaussianas com cinco níveis, fornecendo uma representação multi-escala. Para a geração das pirâmides utilizamos um

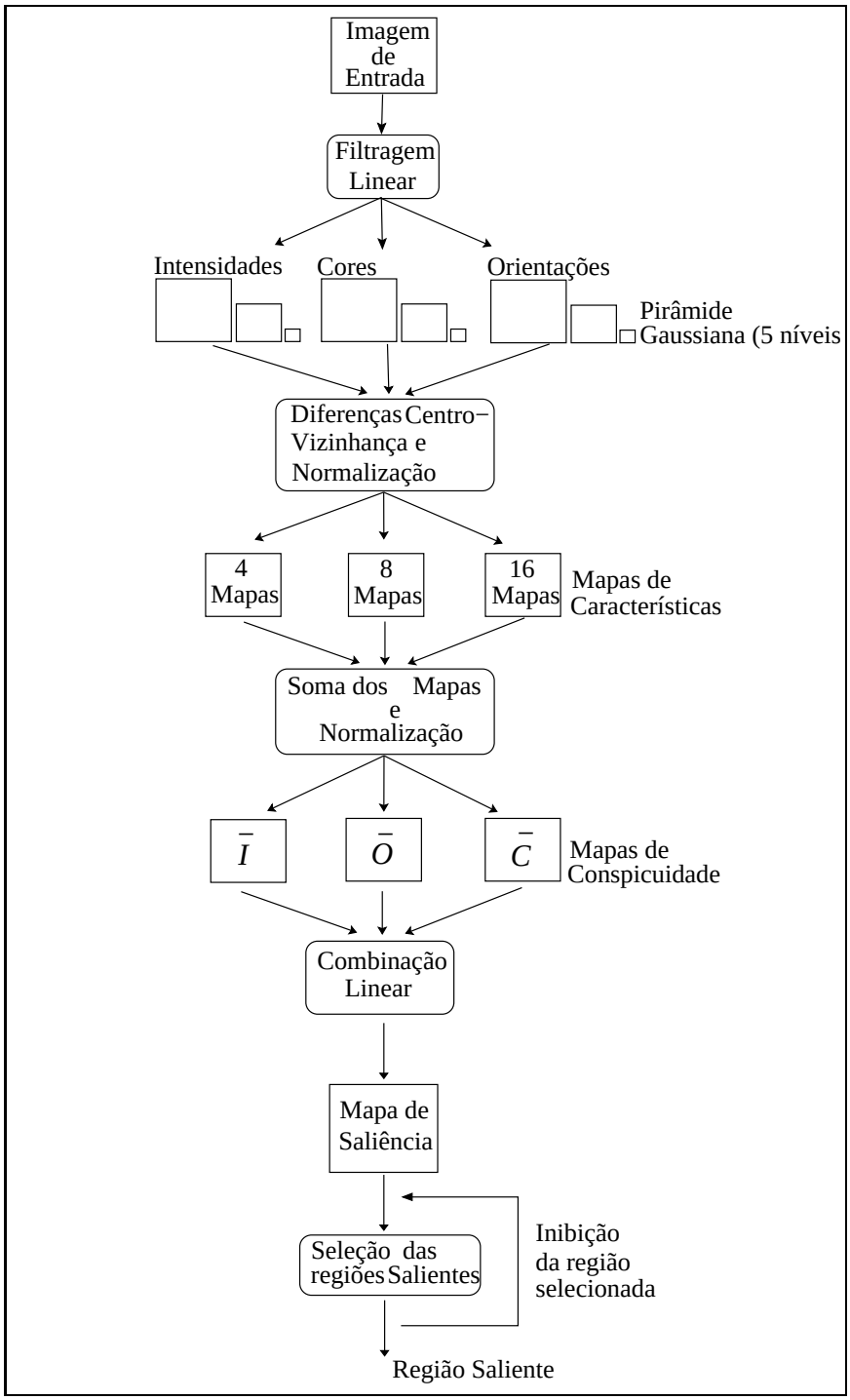

Figura 2: Arquitetura do Módulo de Detecção (adaptação do modelo de Itti e colegas [5]).

algoritimo clássico apresentado no trabalho de Burt e Adelson [1]. O mesmo algoritmo é utilizado depois para gerar a interpolação das imagens.

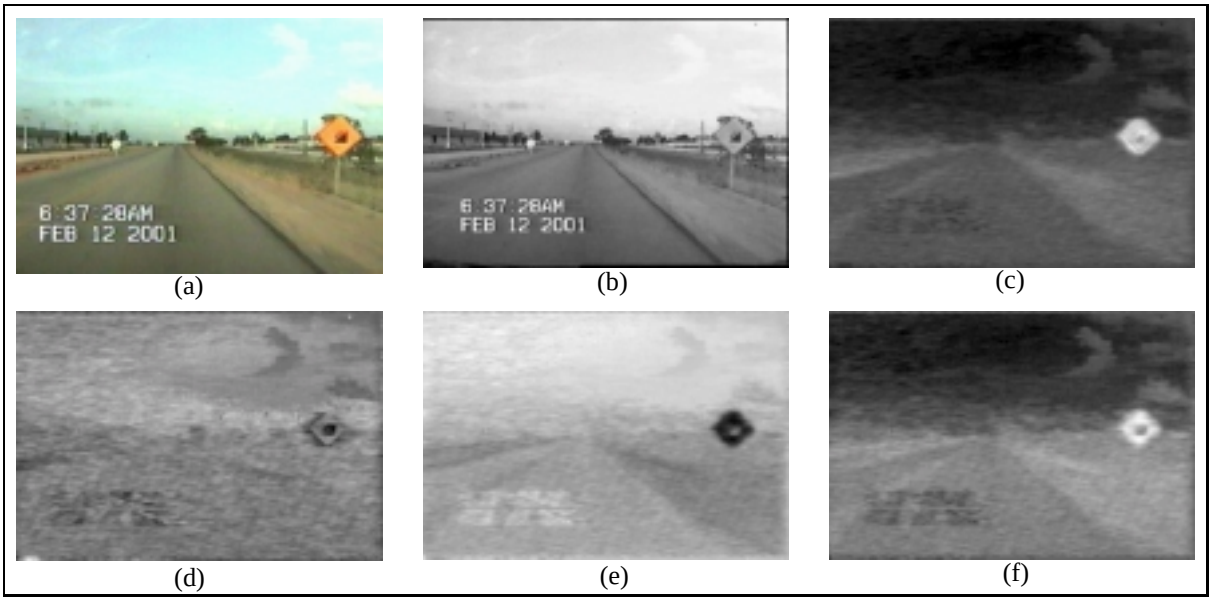

Figura 3: Exemplo da filtragem linear. (a) Imagem de entrada, (b) imagem de intensidade e respectivos canais de cores R (c), G (d), B (e) e Y (f).

As orientações são obtidas a partir da imagem I aplicando-se Filtros Direcionais (*Steerable Filters*) [3] em quatro orientações (0°, 45°, 90°, 135°), e gerando uma pirâmide direcional (*Steerable Pyramid*)[14][1] para cada orientação. A Figura 4 mostra o exemplo de uma Pirâmide Gaussiana e de uma Pirâmide Direcional.

O processo chamado de Diferenças Centro-Vizinhaça na Figura 2, é implementado como a diferença entre as escalas (denotado por $\ominus$), ou seja, o centro é um pixel da imagem na escala $c \in \{1, 2\}$ da pirâmide Gaussiana, e a vizinhaça é o pixel correspondente em outra imagem na escala $v \in \{3, 4\}$ da pirâmide. A diferença é obtida pela interpolação das imagens para a escala original (base da pirâmide) e subtração ponto a ponto (diferença entre os pixels). A utilização de várias escalas produz extração de características multiescala, resultando em 28 Mapas de Características (MC). O primeiro conjunto de mapas é construído a partir do contraste de intensidades, num total de 4 mapas (Equação (1)). 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 vizinhamça escura.

$$\mathcal{I}(c, v) = |I(c) \ominus I(v)| \tag{1}$$

O segundo conjunto de mapas é similarmente construído a partir dos canais de cores, num total de 8 mapas (Equações (2) e (3)). 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.

$$\mathcal{RG}(c, v) = |(R(c) - G(c)) \ominus (G(v) - R(v))| \tag{2}$$

$$\mathcal{BY}(c, v) = |(B(c) - Y(c)) \ominus (Y(v) - B(v))| \tag{3}$$

O terceiro conjunto de mapas é constituído a partir de informações de orientação local, num total de 16 mapas (Equação 4). Estes mapas são inspirados em neurônios do córtex visual sensíveis aos contornos de uma imagem.

$$\mathcal{O}(c, v, \theta) = |O(c, \theta) \ominus O(v, \theta)|\,, \tag{4}$$

Para construir o Mapa de Saliência(MS), os Mapas de Características nas diversas escalas são somados ponto-a-ponto, (denotado por $\bigoplus$), como mostra as Equações (5), (6) e (7), resultando em um Mapa de

[1]Para a contrução da Pirâmide Direcional, nós adaptamos o código fonte e os filtros Kernel desenvolvidos por Simoncelli e Freeman, os quais estão disponíveis via ftp anônimo em: ftp.cis.upenn.adu:pub/eero/steerpyr.tar.Z

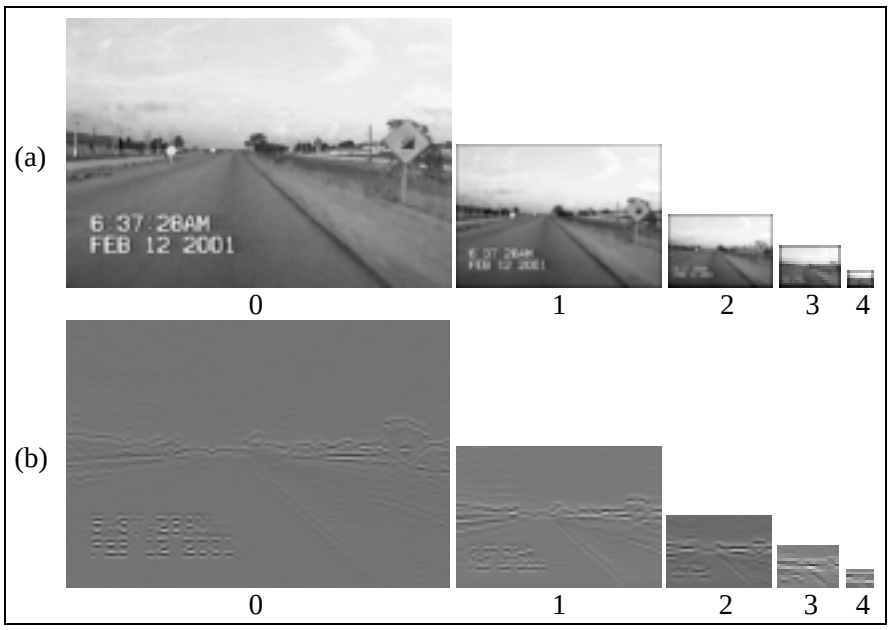

Figura 4: Exemplo da representação piramidal utilizada no modelo. Pirâmide Gaussiana (a) e Pirâmide Direcional (b), geradas a partir da imagem de intensidades (Figura 3(b)).

Conspicuidade (MC) para cada característica (intensidades, cores e orientações). A motivação para a criação dos três canais separados é a hipótese de que características similares competem pela saliência, enquanto que características diferentes contribuem independentemente para o Mapa de Saliência [5].

$$\overline{\mathcal{I}} = \bigoplus_{c=1}^{2} \bigoplus_{v=3}^{4} \mathcal{N}(\mathcal{I}(c,v)) \tag{5}$$

$$\overline{\mathcal{C}} = \bigoplus_{c=1}^{2} \bigoplus_{v=3}^{4} [\mathcal{N}(\mathcal{RG}(c,v)) + (\mathcal{BY}(c,v))] \tag{6}$$

$$\overline{\mathcal{O}} = \sum_{\theta \in (0°,45°,90°,135°)} \mathcal{N}\left(\bigoplus_{c=1}^{2} \bigoplus_{v=3}^{4} \mathcal{N}(\mathcal{O}(c,v,\theta)) \right) \tag{7}$$

O propósito do MS é representar a conspicuidade (saliência) em cada região do campo visual por uma quantidade escalar, e guiar a seleção das regiões atendidas, baseado em uma distribuição espacial. Os três mapas de conspicuidade são normalizados e somados, resultando em uma entrada final para o Mapa de Saliência (Equação 8). Um operador de normalização $\mathcal{N}(.)$ é utilizado para suprir a ausência de uma supervisão *top-down*. $\mathcal{N}(.)$ consiste em: 1) normalizar os valores no mapa para um intervalo fixo $[0..M]$, com o objetivo de eliminar diferenças de amplitude dependentes da modalidade (no nosso caso M=255); 2) encontrar a região de máximo global do mapa M e computar a média $\overline{m}$ de todas as outras máximas locais; 3) multiplicar globalmente o mapa por $(M - \overline{m})^2$.

$$\mathcal{S} = \frac{1}{3}(\mathcal{N}(\overline{\mathcal{I}}) + \mathcal{N}(\overline{\mathcal{C}}) + \mathcal{N}(\overline{\mathcal{O}})) \tag{8}$$

Os passos do algoritmo para implementação do Módulo de Detecção discutidos até agora, são baseados no modelo de Itti e colegas [5]. A principal modificação na nossa implementação está na estratégia utilizada para a seleção das regiões de interesse e na estratégia de inibição das regiões já atendidas. Inicialmente nós implementamos uma estratégia de ordenação dos pixels de maior valor e de inibição das regiões atendidas, com o objetivo de validar a utilização do MS na seleção das regiões de maior interesse. A cada passo, um ponto no mapa de saliência é selecionado e uma máscara de inibição circular preenchida com intensidades nulas é desenhada ao seu redor. Este procedimento inibe todos o pontos presentes no círculo, impedindo que esta região seja tratada novamente. O diametro do círculo (40 pixels) é baseado na dimensão dos objetos (placas de sinalização) presentes nas imagens do subconjunto utilizado, diminuido assim o risco de inibir uma placa vizinha a uma região anteriormente atendida. A Figura 5 mostra o Mapa de Saliência final (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). Os resultados dos experimentos com o mecanismo de atenção são apresentados na próxima seção.

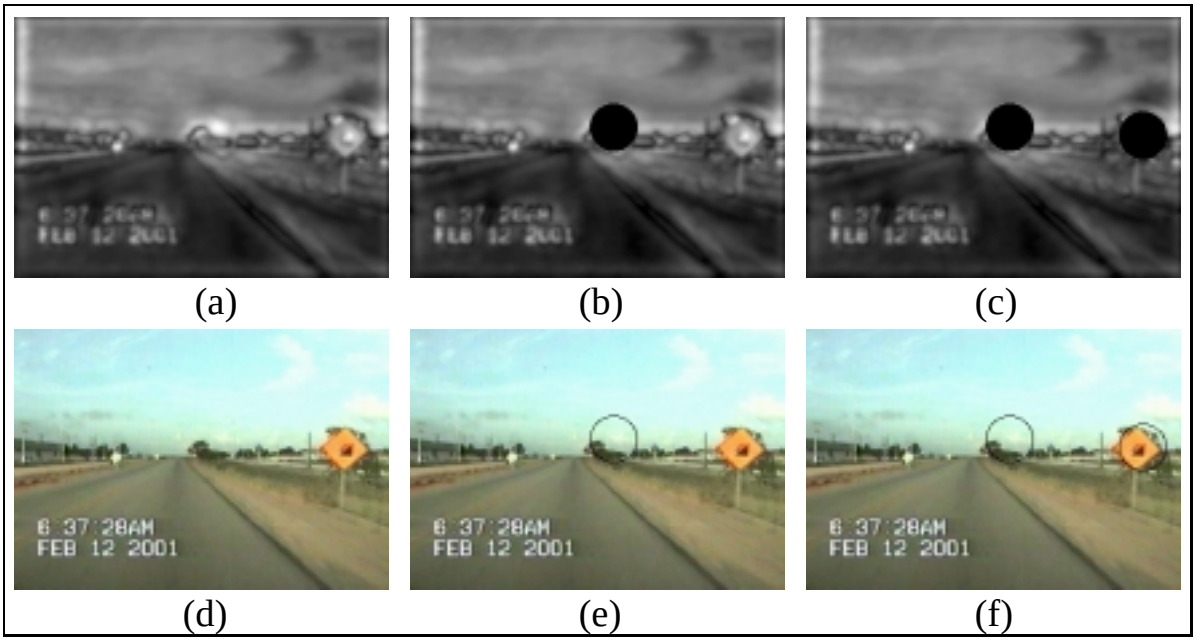

Figura 5: Exemplo da seleção de regiões de interesse. Mapa de Saliência (a), inibição das regiões atendidas (b) e (c), imagem de entrada (d) e seleção das regiões salientes na imagem de entrada.

5 Experimentos e Resultados

A base de dados utilizada nos experimentos é formada por imagens extraídas de um vídeo digital. Este vídeo foi filmado a partir de um veículo em movimento durante uma viagem com dia claro, no trecho de aproximadamente 120Km que separa as cidades de João Pessoa/PB e Campina Grande/PB, Brasil. Tais imagens incluem cenas com segmentos de estradas rurais e urbanas. O *hardware* de aquisição consistiu de uma câmera CCD comum em um tripé, montado na frente do assento direito do veículo, sem mecanismo de estabilização. Após a aquisição, o vídeo foi particionado em *frames* e cada um deu origem a uma nova imagem colorida, com dimensão 352X240 pixels. A partir do conjunto total de imagens, foram selecionadas apenas aquelas que continham sinais de tráfego (placas de trânsito). Para os experimentos que descreveremos a seguir foi utilizado um subconjunto com 15 imagens com exemplos variados de placas. Um total de 16 placas estão presentes no subconjunto, sendo 14 imagens com uma placa e uma imagem com duas placas.

Inicialmente, nós aplicamos o algoritmo 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 MS são selecionadas. Com esta primeira estratégia as regiões com placas de 12 imagens foram selecionadas, correspondendo a 75% do total de imagens. Dentre as imagens que tiveram placas selecionadas, 11 imagens correspondem a trechos de rodovias e uma imagem corresponde a um trecho de rua urbana. Em geral, essas imagens 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 salintes correspondem apenas a trechos de ruas urbanas, e são formadas por por muitas estruturas. Como já foi dito anteriormente, o MS 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 várias regiões mais salientes que as placas, implicando em um número maior de regiões analizadas para que a placa seja selecionada. O valor de K foi incrementado a partir de 1 até que as placas nessas imagens fossem selecionadas. O gráfico da Figura 6 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 de alta saliência na sua vizinhana inibiu a região saliente onde ela se encontrava.

Analisando 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 da imagem. Mesmo quando K asumiu seu maior valor (K=19),

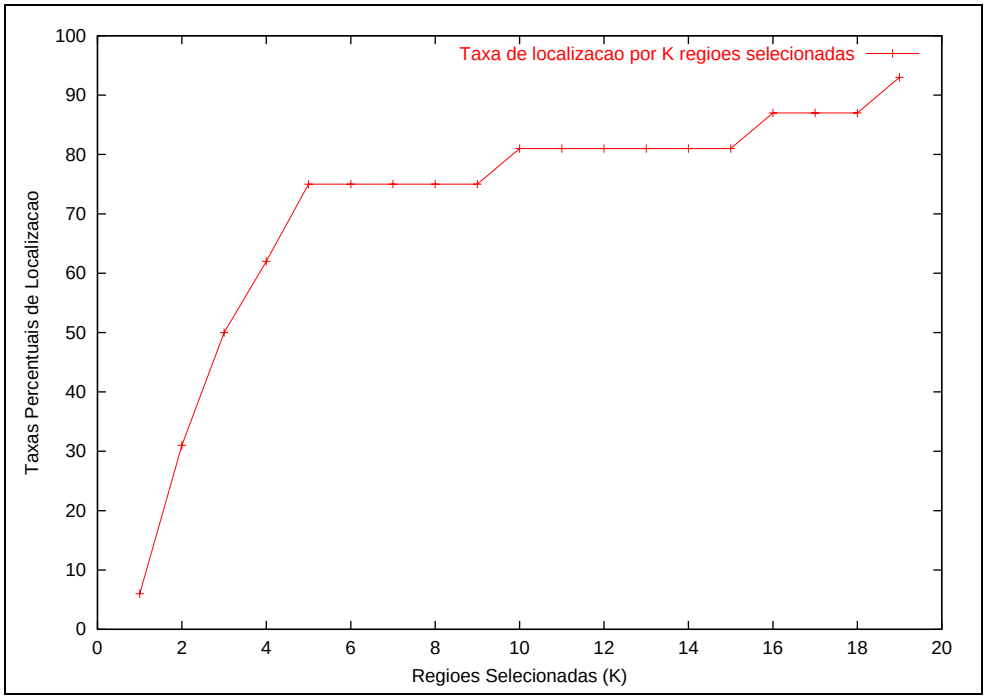

Figura 6: Taxas Percentuais de localização das placas no subconjunto de imagens quando consideramos um número K de regiões selecionadas.

o percentual de pontos examinados com relação ao total de pontos da imagem foi de apenas 0,0225%. O gráfico da Figura 7 mostra os percentuais de pontos selecionados nas imagens do subconjunto utilizado nos testes.

Com o objetivo de reforçar a validade dos resultados alcançados nos experimentos com o mecanismo de atenção, foi implementado um algoritmo para gerar pontos de interesse randômicos. Para cada imagem foram gerados dez pontos, e definida uma distância mínima entre o pontos igual a 20 pixels, simulando a máscara de inibição. Dentre todos os 150 pontos, apenas um coincidiu com uma região que contém uma placa, o que demonstra que o mecanismo de atenção está desempenhando um papel muito mais importante do que uma simples seleção aleatória de pontos de interesse.

6 Conclusões e Trabalhos Futuros

O presente trabalho discutiu detalhes da implementação de um mecanismo de atenção visual e os resultados preliminares da sua aplicação na localização de placas de sinalização em imagens de ruas e estradas. O mecanismo de atenção utiliza um Mapa de Saliência que representa topograficamente as regiões de interesse na cena visual. O Mapa de Saliência é utilizado para guiar a seleção dos pontos de maior valor. As regiões ao redor dos pontos selecionados são inibidas por uma máscara circular preenchida com intensidades nulas, com o objetivo de evitar que regiões salientes sejam atendidas mais de uma vez.

Os resultados alcaçados indicam que o mecanismo implementado é apropriado para a solução do problema investigado. Com a utilização deste mecanismo foi possível reduzir drasticamente o número de pontos atendidos. Uma característica importante na seleção das placas é a quantidade de estruturas presentes na imagem. Quanto maior o número de estruturas maior o número de regiões selecionadas até que uma placa seja localizada. Podemos concluir que o modelo é mais eficiente quando analisa imagens de estradas e menos eficiente quando analisa imagens de ruas urbanas. Um outro fator determinante para a seleção dos objetos de interesse na cena é natureza das regiões ao redor desses objetos. Tal fator é determinado pelas operações do processo chamado de Diferença Centro-vizinhança, que valorizam as diferenças entre um pixel e a região ao seu redor. Assim, as imagens de estradas tiveram as placas presentes em um número menor de regiões de interesse, já que na maioria delas existe uma região homogênea como *background* (vegetação e céu), por trás das placas.

O próximo passo do projeto será a integração dos dois principais módulos do sistema, Detecção e Reconhecimentos. As imagens previamente selecionadas pelo mecanismo de atenção, servirão de entrada para a Rede Neural que deverá classificá-las. Estão previstas novas aquisições de imagens, com o objetivo de

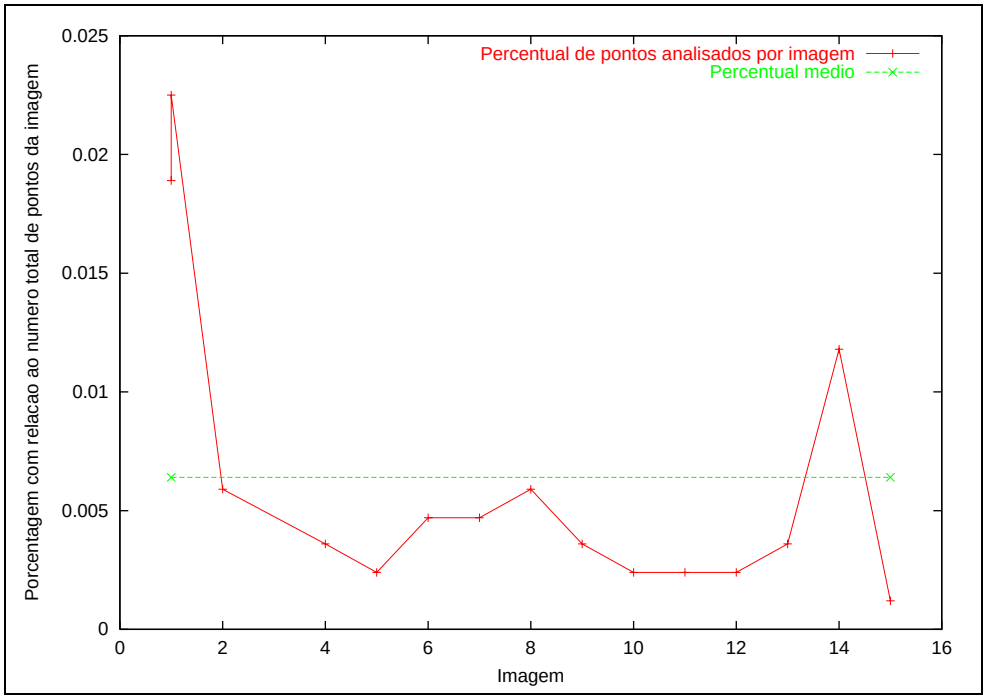

Figura 7: Percentuais de pontos analisados até que uma placa tenha sido selecionada, nas imagens do subconjunto de testes. Note que a imagem 3 não teve a placa selecionada, já que a seleção de um ponto vizinho muito próximo inibiu a região onde a placa estava localizada. A linha horizontal representa o percentual médio.

enriquecer os conjuntos de treinamento e teste da Rede Neural, bem como realizar novos experimentos com mecanismo de atenção.

Referências

[1] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, April 1983.

[2] S. M. Culhane and J. K. Tsotsos. 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, September 1992. Los Alamitos, California.

[3] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Patt. Anal. and Machine Intell.*, 13(9):891–906, Sept. 1991.

[4] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar. 2001.

[5] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov. 1998.

[6] L. Kaufman and W. Richards. Centre-of-gravity tendencies of visual forms. *Perception & psychophysics*, 5(2):85–88, 1969.

[7] C. Koch. Selective visual attention and computational models. CNS 186: Attention Christof Koch, 2000. March 2.

[8] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum. Neurobiol*, 4:219–227, 1985.

[9] C. Little. The intelligent vehicle initiative - advancing human centered smart vehicles. http://www.tfhrc.gov/pubrds/pr97-10/p18.htm.

[10] R. Milanese, H. Wechsler, S. Gill, J. M. Bost, and T. Pun. 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, 1994.

[11] H.-H. Nagel. Computer vision for support of road vehicles drivers. Institut für Algorithmen und Kognitive Systeme, Fakultät für Informatik der Universität Karlruhe, on-line: http://www.tfhrc.gov/pubrds/pr97-10/p18.htm.

[12] F. A. Rodrigues. Localização e reconhecimento de placas de sinalização utilizando um mecanismo de atenção visual e redes neurais artificiais. Dissertação de mestrado, Universidade Federal de Campina Grande, Sep. 2002.

[13] G. Sela and M. D. Levine. Real-time attention for robotic vision. *Real-Time Imaging*, 3:173–194, 1997.

[14] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *2nd Annual IEEE Intenational Conference on Image Processing*, October 1995. Washington, DC.

[15] J. K. Tsotsos. Analyzing vision at the complexity level. *The Behavioral and Brain Sciences*, 13:423–469, 1990.