Sistemas Computacionais para Atenção Visual *Top-Down* e *Bottom-UP* usando Redes Neurais Artificiais

Alcides Xavier Benicasa

Sistemas Computacionais para Atenção Visual *Top-Down* e *Bottom-UP* usando Redes Neurais Artificiais

Alcides Xavier Benicasa

Orientadora: **Profa. Dra. Roseli Aparecida Francelin Romero**
Co-orientador: **Prof. Dr. Zhao Liang**

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências - Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*.

USP – São Carlos
Setembro de 2013

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

B467s

Benicasa, Alcides Xavier
Sistemas Computacionais para Atenção Visual Top-Down e Bottom-Up usando Redes Neurais Artificiais / Alcides Xavier Benicasa; orientador Roseli Aparecida Francelin Romero; co-orientador Zhao Liang. -- São Carlos, 2013.
218 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2013.

1. Atenção Visual Bottom-Up e Top-Down. 2. Eviesamento Top-Down. 3. Atenção Baseada em Objetos. 4. Reconhecimento de Objetos. I. Romero, Roseli Aparecida Francelin, orient. II. Liang, Zhao, co-orient. III. Título.

Ao amor de minha vida Elen,
e ao amor de nossas vidas,
nossa querida filha Alícia.

Agradecimentos

Inicialmente agradeço aos amores de minha vida Elen e Alícia, pois sem este amor eu não existiria.

Agradeço especialmente à minha amada esposa Elen, pois seu apoio, paciência e confiança tornou possível a busca por forças para trabalhar "duro" e concluir mais esta fase de minha vida.

Agradeço também em especial à minha querida mãe Eunice que, apesar da distância, sempre esteve ao meu lado com muito amor, pensamentos de fé e carinho. Agradeço também ao meu pai Osvaldo (em memória), que à sua maneira, mostrou-me o caminho certo à percorrer.

Agradeço aos meus irmãos Vanderli, Marli, Walter, Luiz e Sueli, que sempre acreditaram em mim. Agradeço também a meu cunhado Edilson, pela ajuda incondicional sempre que solicitada.

Gostaria de expressar minha profunda gratidão à minha orientadora, Profa. Dra. Roseli Ap. Francelin Romero e ao meu co-orientador, Prof. Dr. Zhao Liang. Agradeço à Profa. Dra. Roseli pela oportunidade, orientação e inspiração para conduzir esta pesquisa. Agradeço ao Prof. Dr. Zhao por seu otimismo e disponibilidade demonstrados em todos os momentos desta jornada. Gostaria de agradecer a ambos, Profa. Dra. Roseli e Prof. Dr. Zhao pelas oportunidades do passado, presente e futuro que, não existiriam sem vossos apoios.

Agradeço também ao amigo Marcos Quiles, pelas discussões construtivas, parcerias e idéias surgidas durante minha permanência na USP. Agradeço também a todos os colegas do LAR, em especial ao amigo Jorge Kanda pelo companheirismo durante as disciplinas.

Agradecimentos aos funcionários do ICMC e em especial às secretárias da Pós-Graduação, por todo o apoio e atenção disponibilizada.

Finalmente, agradeço aos colegas e amigos do Departamento de Sistemas de Informação da Universidade Federal de Sergipe - Campus de Itabaiana, pela liberação parcial de carga horária, sem o qual a realização desta tese seria praticamente impossível. Agradeço também à Universidade Federal de Sergipe, pelo afastamento

concedido no ano de depósito da tese, sendo de suma importância para a escrita e conclusões finais deste trabalho. Ao programa de bolsas de pós-graduação para docentes e técnicos administrativos da Universidade Federal de Sergipe (THESIS), à CAPES, pelo apoio financeiro inicial e ao Instituto de Ciências Matemáticas e de Computação, pelo suporte e estrutura fornecidos para o desenvolvimento desta tese.

Resumo

A análise de cenas complexas por computadores não é uma tarefa trivial, entretanto, o cérebro humano pode realizar esta função de maneira eficiente. A evolução natural tem desenvolvido formas para otimizar nosso sistema visual de modo que apenas partes importantes da cena sejam analisadas a cada instante. Este mecanismo de seleção é denominado por atenção visual. A atenção visual opera sob dois aspectos: *bottom-up* e *top-down*. A atenção *bottom-up* é dirigida por conspicuidades baseadas na cena, como o contraste de cores, orientação, etc. Por outro lado, a atenção *top-down* é controlada por tarefas, memórias, etc. A atenção *top-down* pode ainda modular o mecanismo *bottom-up* através do enviesamento de determinadas características de acordo com a tarefa. Além do mecanismo de modulação considerado, o que é selecionado a partir da cena também representa uma importante parte para o processo de seleção. Neste cenário, diversas teorias têm sido propostas e podem ser agrupadas em duas linhas principais: atenção baseada no espaço e atenção baseada em objetos. Modelos baseados em objeto, ao invés de apenas direcionar a atenção para locais ou características específicas da cena, requerem que a seleção seja realizada a nível de objeto, significando que os objetos são a unidade básica da percepção. De modo a desenvolver modelos de acordo com a teoria baseada em objetos, deve-se considerar a integração de um módulo de organização perceptual. Este módulo pode segmentar os objetos do fundo da cena baseado em princípios de agrupamento tais como similaridade, proximidade, etc. Esses objetos competirão pela atenção. Diversos modelos de atenção visual baseados em objetos tem sido propostos nos últimos anos. Pesquisas em modelos de atenção visual têm sido desenvolvidas principalmente relacionadas à atenção *bottom-up* guiadas por características visuais primitivas, desconsiderando qualquer informação sobre os objetos. Por outro lado, trabalhos recentes têm sido realizados em relação ao uso do conhecimento sobre o alvo para influenciar a seleção da região mais saliente. Pesquisas nesta área são relativamente novas e os poucos modelos existentes encontram-se em suas fases iniciais. Aqui, nós propomos um novo modelo para atenção visual com modulações *bottom-up* e *top-down*. Comparações qualita-

tivas e quantitativas do modelo proposto são realizadas em relação aos mapas de fixação humana e demais modelos estado da arte propostos.

Abstract

Perceiving a complex scene is a quite demanding task for a computer albeit our brain does it efficiently. Evolution has developed ways to optimize our visual system in such a manner that only important parts of the scene undergo scrutiny at a given time. This selection mechanism is named visual attention. Visual attention operates in two modes: bottom-up and top-down. Bottom-up attention is driven by scene-based conspicuities, such as the contrast of colors, orientation, etc. On the other hand, top-down attention is controlled by task, memory, etc. Top-down attention can even modulate the bottom-up mechanism biasing features according to the task. In additional to modulation mechanism taken into account, what is selected from the scene also represents an important part of the selection process. In this scenario, several theories have been proposed and can be gathered in two main lines: space-based attention and object-based attention. Object-based models, instead of only delivering the attention to locations or specific features of the scene, claim that the selection it be performed on object level, it means that the objects are the basic unit of perception. In order to develop models following object-based theories, one needs to consider the integration of a perceptual organization module. This module might segment the objects from the background of the scene based on grouping principles, such as similarity, closeness, etc. Those objects will compete for attention. Several object-based models of visual attention have been proposed in recent years. Research in models of visual attention has mainly focused on the bottom-up guidance of early visual features, disregarding any information about objects. On the other hand, recently works have been conducted regarding the use of the knowledge of the target to influence the computation of the most salient region. The research in this area is rather new and the few existing models are in their early phases. Here, we propose a new visual attention model with both bottom-up and top-down modulations. We provide both qualitative and quantitative comparisons of the proposed model against an ground truth fixation maps and state-of-the-art proposed methods.

Sumário

Lista de Figuras

CAPÍTULO

1

Introdução

A atenção visual é uma característica importante para os seres vivos, de modo que os tornam capazes de interagir com o ambiente de forma rápida, possibilitando a identificação de áreas de maior interesse ou relevância em uma ambiente. De acordo com Itti and Koch (2001), a habilidade dos sistemas visuais biológicos de fixar rapidamente a visão em pontos de interesse e reconhecer possíveis presas, predadores ou rivais, é determinante para a perpetuação e a evolução das espécies. Alguns estímulos visuais são intrinsecamente conspicuosos ou salientes em um determinado contexto. Por exemplo, considere uma jaqueta vermelha posicionada entre vários ternos de cor preta, automaticamente e involuntariamente a região colorida com vermelho receberá a atenção.

Tendo-se como base estudos em seres humanos e macacos, pode-se afirmar que o processo de seleção visual seleciona apenas um subconjunto da informação sensorial disponível, na forma de uma região circular do campo visual, conhecida como foco de atenção. Desta forma, a atenção auxilia na redução da explosão combinatória resultante da análise de todas as informações sensoriais disponíveis e de todas as possíveis relações presentes em uma cena (Shic and Scassellati, 2007; Tsotsos, 1992), pois apenas informações que estão dentro da área da atenção são processadas, enquanto que o restante é suprimido (Carota et al., 2004). Além disso, a atenção visual se apresenta como um eficiente mecanismo para reduzir tarefas complexas, como análise de uma cena, em um conjunto de sub-tarefas menores (Itti, 2005).

Segundo Rossini and Galera (2006), pesquisadores da área de psicologia têm realizado nos últimos trinta anos trabalhos relacionados aos processos cognitivos da atenção visual humana, investigando a arquitetura cognitiva sobre os mecanismos de seleção e integração da informação relevante contida no ambiente. Talvez, o

primeiro pesquisador a investigar de maneira sistemática este fenômeno tenha sido Herman von Helmholtz (Wright and Ward, 1998), que conduziu um experimento considerado por muitos como a primeira demonstração científica da capacidade do sistema visual humano em direcionar a atenção para uma determinada área do campo visual, na ausência de movimentos sacádicos[1] (*covert attention*). Para demonstrar esta capacidade, Helmholtz fixou seu olhar em um pequeno ponto iluminado no centro de um campo com letras impressas não iluminadas. A seguir, este campo visual era iluminado rapidamente pelo clarão de uma faísca elétrica. Entretanto, a iluminação era tão breve que não permitia nenhum movimento ocular durante a apresentação dos estímulos. Nesta condição, ele foi incapaz de distinguir as letras ao redor do ponto de fixação. No entanto, se antecipadamente fosse direcionada sua atenção para uma parte específica do campo visual, o reconhecimento das letras ali representadas, tornava-se possível. Helmholtz interpretou este resultado como um reflexo da capacidade de direcionamento interno dos recursos atentivos para uma área específica do campo visual (Pashler, 1998). Este mecanismo também é caracterizado pela habilidade que possuímos, por exemplo, em detectar movimentos periféricos ou de identificar o próprio nome em uma lista (Frintrop et al., 2010). Desta forma, acredita-se que a atenção pode ser direcionada para uma região de interesse, previamente aos movimentos dos olhos. Entretanto, existem casos em que os movimentos dos olhos não são precedidos pela atenção *covert*, como por exemplo, durante o processo de leitura, onde a freqüência de movimentos sacádicos (*overt attention*) não permite à atenção *covert* realizar o escaneamento prévio da cena (Findlay and Gilchrist, 2001). Frintrop et al. (2010) apresenta uma visão conclusiva sobre o assunto, afirmando que ambas as atenções, *covert* e *overt*, geralmente trabalham juntas, baseado no fato de que a atenção é direcionada para uma região de interesse, seguida por um movimento sacádico responsável por focar a região, permitindo sua percepção de forma mais definida.

Nos últimos anos, a atenção visual tem sido estudada através de dois modelos gerais: o primeiro é baseado na localização dos estímulos no espaço, relacionado diretamente à atenção *covert*, uma vez que os próprios estímulos do espaço são responsáveis por guiar a atenção, por sua vez, o segundo é baseado nas características intrínsecas do objeto (Desimone and Duncan, 1995; Itti and Koch, 2001), onde movimentos sacádicos são necessários para a busca no campo visual pelo alvo desejado. No primeiro modelo, a atenção visual é gerada pela combinação de informações provenientes da retina e de regiões primárias do córtex visual (processos de baixa ordem ou processos *bottom-up* - dependente da cena) (Nothdurft, 2005). No segundo modelo, a atenção visual é proveniente de regiões superiores do córtex visual e de outras áreas corticais fora do córtex visual (processos de alta ordem ou processos *top-down* - dependente da tarefa) (Egeth and Yantis, 1997).

[1] Definido por Rayner (1998) como o movimento contínuo dos olhos à procura de informações visuais como, por exemplo, objetos em uma cena.

Um importante ponto a ser notado é a característica de integração entre os modelos de atenção visual *bottom-up* e *top-down*. Estudos demonstram que informações provenientes de regiões primárias e superiores do córtex visual são conduzidas, não por uma só via hierárquica ou fluxo de processamento, mas por pelo menos três vias de processamento paralelas e interativas no cérebro (Kandel et al., 1997). A existência de vias paralelas de processamento, por sua vez, infere em uma outra questão sobre a atenção visual, decorrente do processo cognitivo que ocorre em região cortical específica responsável pela integração dos diversos estímulos presentes na cena. Esta questão é conhecida como o problema da integração[2]. De maneira geral, tanto processos *bottom-up*, quanto processos *top-down*, atuam na seleção dos estímulos mais relevantes no campo receptivo. Nesta condição, a atenção visual pode ser considerada como um processo intermediário que integra coerentemente estes estímulos.

1.1 Motivação

Uma das principais questões relacionadas à atenção visual está em determinar qual tipo de informação é a mais relevante para guiar o desenvolvimento da atenção visual. A resposta pode variar de acordo com o ambiente, que pode nos indicar "onde" a saliência está localizada, ou ainda, a partir de características do alvo desejado, ou seja, "o que", de forma que a atenção seja direcionada para informações previamente conhecidas.

Nos últimos anos têm sido propostos diversos modelos computacionais para a atenção visual. Baseados na teoria da integração de características de Treisman and Gelade (1980), proposto inicialmente por Koch and Ullman (1985), modelos baseados em mapas de saliência são caracterizados pela integração de estímulos locais, extraídos a partir de diferentes localizações espaciais, de modo que a saliência seja baseada na localização de regiões salientes da cena, evidenciando a principal preocupação destes modelos em identificar *onde* se encontra a saliência da cena.

Modelos baseados em mapas de saliência também têm sido propostos considerando o enviesamento da atenção visual, com o objetivo de ativar regiões ou pontos específicos do mapa de saliência, baseado em informações prévias sobre o alvo. Neste caso, a preocupação está no "o que" se procura. Apesar de o mapa de saliência ter sido mantido como principal componente para o desenvolvimento da visual baseada no espaço, o tipo de tarefa destes modelos foram alterados de busca livre de tarefas, para uma busca visual direcionada, tornando-os especificamente modelos *top-down*. Como conseqüência do uso de enviesamento *top-down*, características salientes destacadas inicialmente no mapa de saliência são desconsideradas, caso estas não estejam associadas às características previamente conhecidas sobre o alvo.

[2]do inglês *Binding Problem*

Observamos ainda que, para garantir o desempenho de modelos de atenção visual baseados em mapas de saliência, mesmo apresentando características baseadas em objetos, espaço, ou ainda, com ou sem enviesamento *top-down*, faz-se necessária a existência de características salientes relacionadas aos objetos presentes na cena. Entretanto, a assimetria atribuída à presença ou ausência de uma determinada característica, também deve ser considerada como uma importante informação para o desenvolvimento da atenção visual.

De um modo geral, modelos de atenção visual baseados em mapas de saliência seguem um propósito comum, a detecção de regiões salientes da cena. Entretanto, de acordo com os propósitos iniciais de um modelo de saliência básico, nenhuma consideração é direcionada à estrutura dos objetos. Esta limitação dificulta, por exemplo, a detecção ou reconhecimento de objetos específicos, contudo, modelos com estas características têm sido utilizados como um importante componente para modelos mais abrangentes, de forma a auxiliar, em uma fase inicial, à entrega da atenção para regiões de maior interesse.

1.2 Objetivos

Este tese apresenta alguns novos modelos para atenção visual com características *bottom-up* e *top-down* utilizados para o desenvolvimento da atenção. O trabalho é baseado inicialmente na extração de características visuais primitivas da cena, seguido pela proposta de mecanismos baseados em objetos para a identificação de saliência a partir de características visuais primitivas e também de valores de reconhecimento de alvos específicos. Uma discussão detalhada em relação aos trabalhos relacionados será apresentada nos capítulos seguintes. Destacamos aqui um breve resumo dos principais objetivos pretendidos:

- Introdução de um mapa de saliência *bottom-up*, denominado aqui por mapa de atributo-saliência, gerado a partir da auto-organização de atributos primitivos da cena. Para esta finalidade propomos um novo modelo para o desenvolvimento da competição por atenção baseado em informações espaciais, de forma que o mapa de saliência seja representado por um mapa de atributo-saliência (Seções 5.1.1, 5.1.2 e 5.1.3, e também publicado em (Benicasa and Romero, 2010)).

- Proposta de um mapa de saliência *bottom-up* e *top-down*, contendo as mesmas características do mapa de atributo-saliência, entretanto, gerado a partir da auto-organização de atributos primitivos da cena e de informações cognitivas sobre objetos. Pretendemos assim neste novo modelo, viabilizar a competição pela atenção visual baseada nos valores de reconhecimento dos objetos e em suas características primitivas (Seções 5.2.1 e 5.2.2, e também publicados, respectivamente em (Benicasa et al., 2012) e (Benicasa et al., 012b)).

- Introdução de um mapa de objeto-saliente gerado a partir da competição pela atenção visual baseada estritamente em objetos. Consideramos para o desenvolvimento do modelo de atenção demais características, como por exemplo, o tamanho do objeto e a assimetria atribuída à presença ou ausência de uma determinada característica. O principal objetivo deste modelo é de possibilitar a redução do esforço computacional para o desenvolvimento da seleção visual e também de aumentar a taxa de predição do modelo (Seção 5.3.1, e submetido para publicação em (Benicasa et al., 2013)).

- Proposta de mecanismo para o enviesamento *top-down* prévio ao desenvolvimento da atenção. Considerando domínios onde se conheça previamente informações sobre o alvo desejado, propomos um novo mecanismo para que somente objetos candidatos a alvo participem da competição pela atenção visual. O principal objetivo desta proposta é de viabilizar a aplicabilidade do modelo em sistemas de tempo real (Seção 5.3.2, e aceito para publicação em (Benicasa et al., 013b)).

1.3 Organização do Texto

Este trabalho está organizada da seguinte maneira. No Capítulo 2 são apresentados conceitos de atenção visual considerados relevantes à inspiração biológica para o desenvolvimento desta tese. No Capítulo 3 são introduzidos alguns dos principais conceitos teóricos utilizados nesta tese. O Capítulo 4 apresentada a revisão de modelos de atenção visual e os principais pontos de investigação a serem abordados nesta tese. No Capítulo 5 são apresentados os modelos para atenção visual propostos, bem como os resultados obtidos durante os diversos estudos realizados. O Capítulo 6 apresenta uma análise qualitativa e quantitativa dos principais resultados obtidos. Finalmente, no Capítulo 7 são realizadas as discussões sobre os resultados, bem como as considerações finais deste trabalho e trabalhos futuros.

CAPÍTULO

2

Neurobiologia da Atenção Visual

A inspiração biológica dos modelos de atenção visual propostos dá-se, muitas vezes, em relação à estudos da neurobiologia do sistema visual. Sendo assim, neste capítulo serão apresentados conceitos de atenção visual considerados relevantes à inspiração biológica para o desenvolvimento desta tese.

2.1 O Fluxo da Informação Através do Córtex Visual

Iniciamos pelo fato de que a todo instante os olhos se deparam com uma carga de estímulos visuais enorme. Entretanto, seria impossível processar toda informação que chega aos olhos de uma só vez (Tsotsos et al., 1995). Diante disto, de acordo com Kandel et al. (1997), o sistema visual pode ser considerado como um sistema sensorial somático, onde existem camadas envolvidas com o processamento de diferentes aspectos das informações visuais.

Para Miller (2000), Itti and Koch (2001) e Kandel et al. (1997), a segregação das informações visuais inicia-se nos olhos, através da retina (Figura 2.1), onde células ganglionares com dois tamanhos, células grandes (magnocelulares), e células pequenas (parvocelulares), projetando-se para as camadas magnocelulares e parvocelulares do núcleo geniculado lateral localizado no tálamo, respectivamente, dando origem a três vias principais - duas a partir da camada parvocelular e uma a partir da camada magnocelular, que conduzirá a informação visual, de forma organizada, para o córtex visual primário, ou V1 (também chamado de córtex estriado), que conterá um mapa completo da retina. Por fora do córtex estriado localizam-se as áreas extra-estriadas, como por exemplo V2, V3, V4 e V5, sendo um conjunto de áreas visuais de ordem superior que também receberá informações provenientes da retina através de sinapses com o córtex visual primário. Sendo assim, a informação

contida em cada via apresenta implicação direta às seguintes percepções:

- *via 1* (parvocelular): implica na percepção de cor, com sinapses nas camadas superficiais do córtex visual primário, seguindo para V2, e se propagando para a área V4, uma área que possui muitas células que respondem à cor, terminando no córtex temporal inferior, área que diz respeito à percepção de cor e forma;

- *via 2* (parvocelular): implica na percepção das formas, com sinapses nas camadas mais profundas de V1, que também recebe uma pequena contribuição da via magnocelular. Assim como a *via 1*, esta via também se projeta para o córtex temporal inferior. Caracterizando-se como um sistema sensível ao contorno e à orientação das imagens, elementos importantes para a percepção da forma. Além de ser importante para a percepção da profundidade (e em certo grau da cor);

- *via 3* (magnocelular): especializada na detecção do movimento e das relações espaciais, também contribuindo para a percepção da profundidade. Faz sinapses com as camadas de V1, seguindo para V2 e então para V5, área concernente à profundidade e ao movimento. Esta via projeta-se para outras áreas do córtex parietal que dizem respeito à função visuoespacial. Caracterizando-se como um sistema relativamente insensível à cor e analisam mal os objetos estacionários.

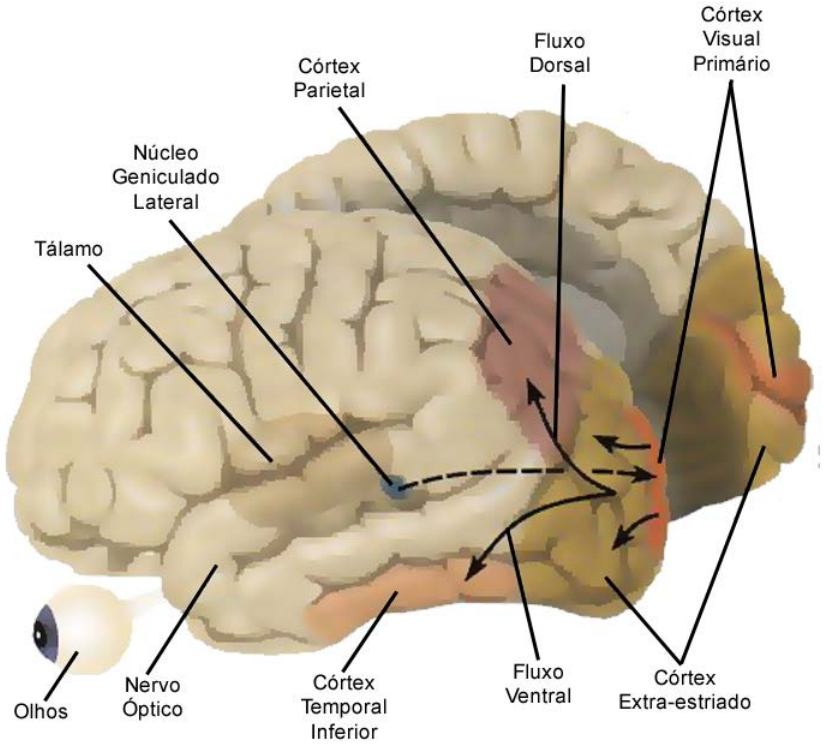


Figura 2.1: Visualização das projeções a partir da retina para o córtex visual primário. Adaptado de Lau (2013).

Como pode ser observado na Figura 2.2, as três vias especializadas pelo fluxo da informação podem ainda interagir entre os níveis. De maneira sucinta, o fluxo da informação, a partir do córtex visual primário, prossegue por dois caminhos paralelos, o primeiro passando por áreas corticais incluindo o córtex parietal (*fluxo dorsal*), responsáveis principalmente pela localização espacial e por direcionar a atenção para objetos de interesse na cena, e o segundo passando por áreas corticais incluindo o córtex temporal inferior (*fluxo ventral*), responsáveis principalmente pelo reconhecimento e identificação de estímulos visuais (Itti and Koch, 2001). Acredita-se que o controle do desenvolvimento da atenção ocorra principalmente no fluxo dorsal. Embora, provavelmente, as áreas do fluxo ventral não estejam diretamente relacionadas com o controle da atenção, estas têm sido responsáveis por receber o "*feedback*" da modulação atencional, estando envolvidas na representação de locais de atenção e objetos (Kandel et al., 1997). Além disso, acredita-se que várias áreas cerebrais de funções mais elevadas podem contribuir para o controle da atenção, uma vez que lesões nessas áreas podem causar uma situação de "negligência", na qual pacientes parecem ignorar partes de seu ambiente visual (Itti and Koch, 2001; Miller, 2000). Falaremos sobre o controle cognitivo da atenção na seção seguinte, onde abordaremos sobre o córtex pré-frontal que, de acordo com Cohen and Servan-Schreiber (1992); Miller (2000), desempenha uma importante função no controle da atenção.

De acordo com os estímulos apresentados e suas respectivas vias, torna-se evidente a principal preocupação relacionada a cada sistema, ou seja, o córtex temporal inferior se ocupa do *que* é visto, enquanto que o córtex parietal se interessa mais por *onde* estão os objetos.

Podemos concluir que a atenção visual é direcionada de forma involuntária (*bottom-up*) por estímulos visuais salientes que emergem dentre os possíveis alvos de uma cena. Entretanto, a atenção visual também pode ser direcionada para alvos específicos, de acordo com a importância atribuída pelo próprio observador (*top-down*). Consequentemente, para um comportamento de reação inteligente, deverão ser considerados ambos os estímulos. Sendo assim, a seguir será apresentado o controle cognitivo da atenção visual, a partir da interação entre os processos *bottom-up* e *top-down*.

2.2 Controle Cognitivo da Atenção Visual

O estudo da interação dinâmica e complexa existente entre as formas de atenção citadas (*bottom-up* e *top-down*), consideradas determinantes para guiar a atenção a cada momento, têm atraído a atenção de diversos pesquisadores ao longo dos últimos anos (Bacon and Egeth, 1994; Wolfe, 1994; Egeth and Yantis, 1997; Kim and Cave, 1999; Yan, 1999; Lamy et al., 2003; Connor et al., 2004).

Pesquisadores têm se preocupado em analisar o que ocorre no cérebro quanto

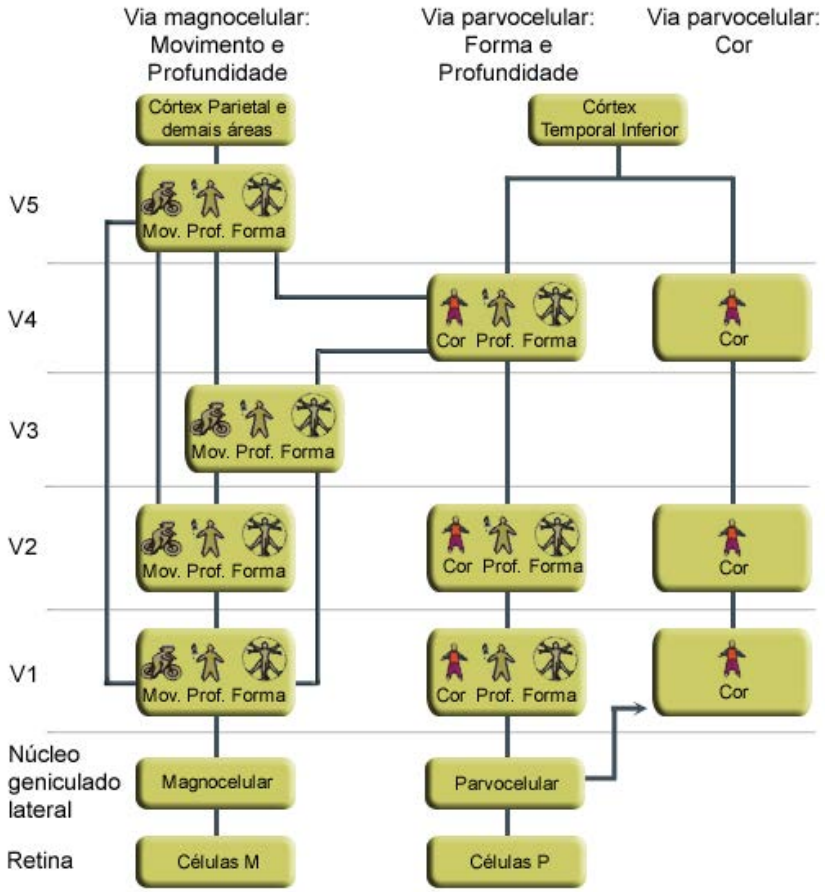


Figura 2.2: Vias visuais paralelas do sistema visual. Adaptado de Kandel et al. (1997).

à integração entre os estímulos que chegam ao córtex pré-frontal, considerado como parte fundamental para o controle da atenção (Cohen and Servan-Schreiber, 1992; Miller, 2000). A principal motivação está no fato de que seres humanos e outros animais podem reagir, não somente reflexivamente diante de informações sensoriais imediatas e salientes, mas também pela possibilidade de substituir ou aumentar reações habituais e reflexivas, com o objetivo de modular seu comportamento de acordo com intenções próprias, caracterizando assim sua natureza cognitiva, responsável pelo controle dos sensores, memória e operações motoras necessárias à propósitos comuns. De acordo com Corbetta (1998) e Frintrop et al. (2010), a atenção define a capacidade mental para selecionar estímulos, respostas, memórias, ou pensamentos que são comportamentalmente relevantes entre os muitos outros que são comportamentalmente irrelevantes.

Em (Treisman and Gelade, 1980), foi proposta a teoria de integração de características (*Feature Integration Theory*), baseada no fato de que um objeto que apresente características contrastantes com os demais objetos da cena, *pop-out*[1], recebendo a atenção. De acordo com Treisman and Gelade (1980), a teoria de integração de características propõe que características primitivas do campo visual,

[1]Impressão subjetiva de que o alvo "salte" da imagem e receba a atenção (Frintrop et al., 2010).

como por exemplo, informações sobre cores, orientações e intensidades, sejam registradas previamente, automaticamente, e de forma paralela através de todo campo visual, enquanto que objetos sejam identificados separadamente em um estágio posterior, o que exigirá o direcionamento da atenção. Na Figura 2.3 é apresentado o modelo proposto, composto por diversos mapas de características, neste caso, pelos mapas de cores e orientações, e um mapa de localização, responsável por codificar, em um único plano, as saliências dos mapas de características, assim como de manter informações necessárias para a representação e reconhecimento de objetos sob o foco de atenção. Para demonstrar este conceito, a Figura 2.4 apresenta um experimento composto por oito objetos, sendo sete distratores e um alvo (Murthy et al., 2001). Inicialmente, o objeto contrastante surge em uma determinada posição, recebendo o foco da atenção e, em seguida, sua posição é trocada com um distrator, sendo a atenção redirecionada para a nova posição. Neste experimento, é concluído que a atenção está intrinsecamente ligada às características da imagem e não a comandos sacádicos obrigatórios. O experimento é importante para a compreensão do processo de atenção visual, entretanto, somente um estímulo visual foi considerado, neste caso, a cor do alvo. É importante notar que o efeito *pop-out* ocorre somente na presença de distratores homogêneos, caso contrário, a busca baseada em uma pesquisa conjuntiva de característica poderá também ocorrer de forma eficiente, entretanto, sem apresentar *pop-outs* (Frintrop et al., 2010). Este comportamento pode ser observado no experimento proposto por Wolfe and Horowitz (2004), apresentado na Seção 2.4 (Figura 2.12(a)).

Para Miller (2000), a função do córtex pré-frontal para o controle cognitivo é enfatizada quando um mesmo estímulo pode levar a mais de uma resposta, conforme condicionado por um outro estímulo. Processo definido por Treisman and Gelade (1980) como busca conjuntiva de características. Por exemplo, a simples ação de atender ao telefone (A1) quando este estiver tocando (C1) está condicionada à premissa do telefone pertencer a você (C3), caso contrário (C2), o telefone não deverá ser atendido (A2). Pode-se entender com este exemplo, apresentado na Figura 2.5, que um determinado conjunto de estímulos (mostrados em vermelho) ativará uma saída apropriada (A1), devido aos padrões de associações que foram formados durante um período de aprendizagem, enquanto que informações distintas (mostrados em azul) serão representadas por diferentes padrões no córtex pré-frontal, levando a outras respostas, neste caso representado por (A2). Para Itti and Koch (2001), o córtex pré-frontal é responsável tanto pelo planejamento da ação, como a execução de movimentos sacádicos através do colículo superior, quanto pela modulação da atenção, via "*feedback*", do processamento das vias de entradas visuais (apresentadas na Seção 2.1). Assim, a atenção está envolvida no desencadeamento do comportamento relacionado ao reconhecimento, planejamento e controle motor (Miller, 2000; Itti and Koch, 2001).

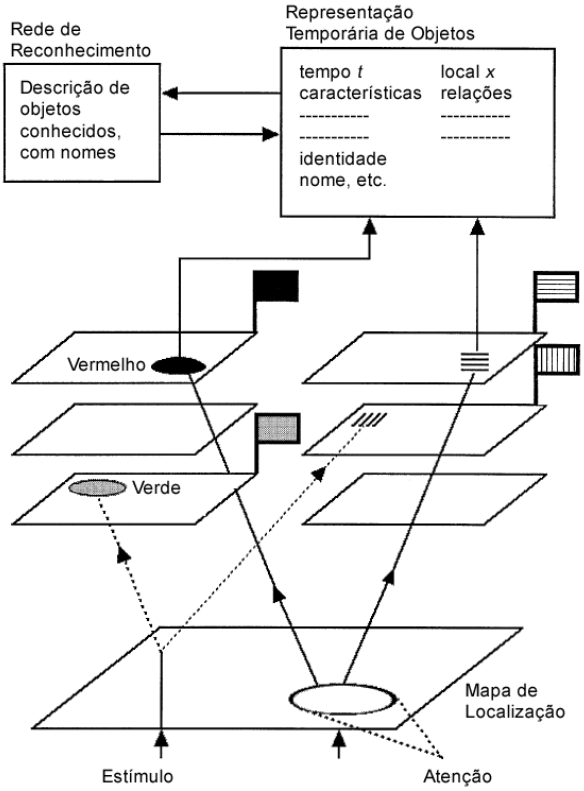


Figura 2.3: Diagrama do modelo da Teoria de Integração de Características proposto por (Treisman, 1998)

Conforme mencionado anteriormente, e de acordo com experimentos apresentados, um objeto que apresente um certo contraste em relação aos demais objetos da cena receberá a atenção. A informação que define o contraste entre objetos está relacionada tanto à características primitivas da imagem quanto à conhecimentos prévios sobre alvos específicos. Entretanto, é importante notar que, quando existem objetos com características primitivas similares, um mecanismo de atenção *top-down* será necessário para selecionar um dos objetos como alvo, baseando-se em suas próprias características primitivas ou, caso exista, em algum conhecimento prévio sobre o alvo. Em outras palavras, será necessária uma medida cognitiva, que possa salientar um dos objetos que, por sua vez, possuam características contrastantes com os demais, porém não contrastantes em si. De acordo com Ogawa and Komatsu (2004) e Egeth and Yantis (1997), se existir conhecimento prévio sobre um estímulo que possa diferenciar um relevante objeto em relação aos demais, torna-se possível direcionar a atenção para este objeto mais facilmente, uma vez que este se distingue de alguma maneira. Neste caso, apesar da eficiência da busca, *pop-outs* não devem ocorrer devido à natureza conjuntiva do alvo.

O processo de atenção visual, para exercer sua função de forma completa, deve contar com informações (estímulos), tanto involuntárias - *bottom-up* - quanto voluntárias - *top-down* - que possam direcionar a atenção para locais ou objetos de maior importância ou interesse na cena. Sendo assim, de acordo com a teo-

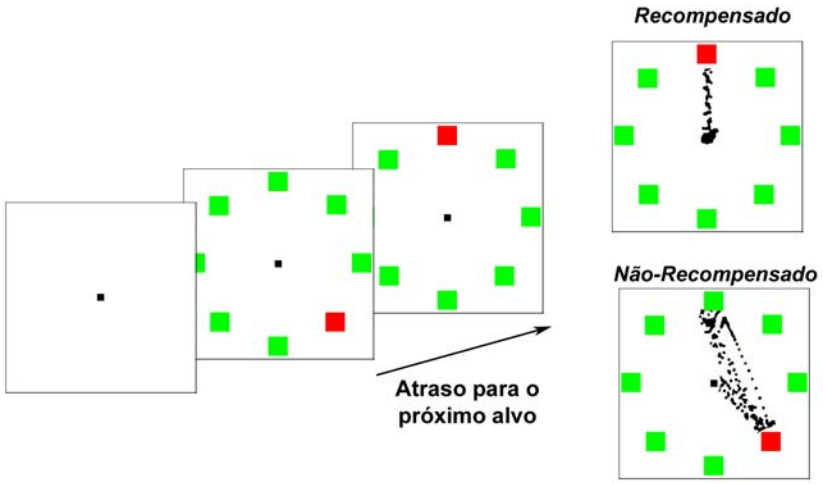


Figura 2.4: Seleção sequencial de características contrastantes. O alvo (mostrado em vermelho) se distingue dos distratores por sua cor. Após o surgimento da primeira imagem, o alvo troca de posição com um distrator. O observador, neste caso um primata, somente recebe a recompensa após o direcionamento da atenção para a posição final do alvo. Em outro experimento no qual não houve mudança de posição do alvo, a recompensa ocorreu imediatamente após o movimento sacádico correto. Para maiores detalhes veja Murthy et al. (2001).

ria de integração de características, podemos concluir que a integração dos sinais provenientes destas vias possa, coerentemente, estabelecer o foco da atenção visual e auxiliar em tarefas de reconhecimento, ou ainda, ser utilizada como informação para o enviesamento da atenção, assunto abordado inicialmente na Seção 2.3 e consequentemente nos modelos baseados nesta hipótese, apresentados na Seção 4.2.

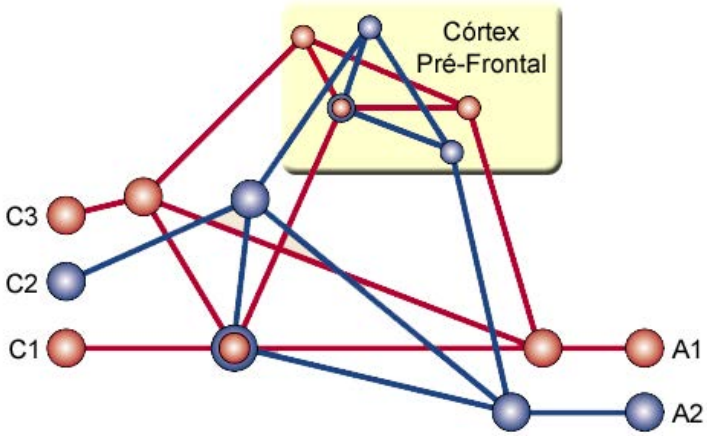


Figura 2.5: A função do Córtex Pré-Frontal no controle da cognição. Adaptado de Miller (2000).

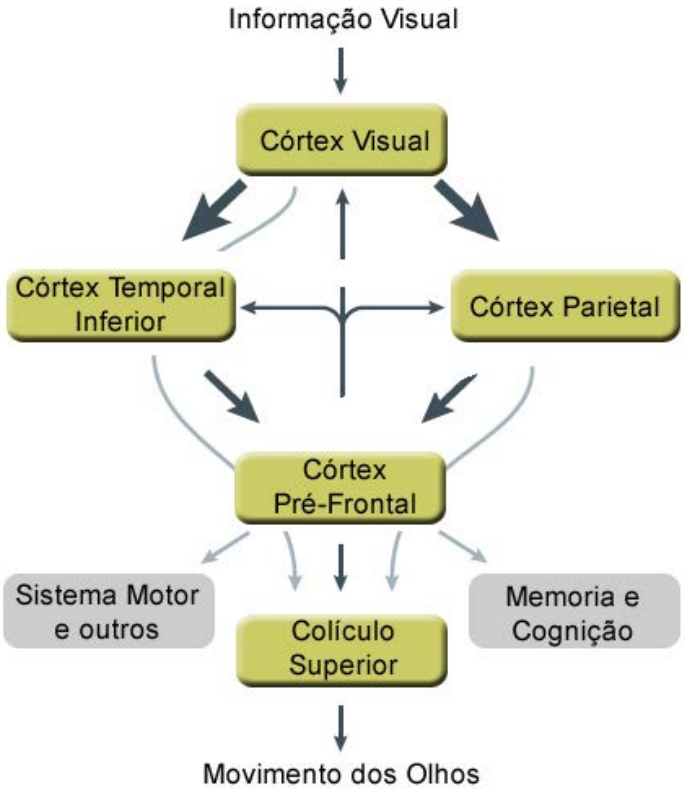


Figura 2.6: Mecanismo neural para o controle da atenção (Itti and Koch, 2001).

2.3 Enviesamento *Top-down*

Diversas pesquisas têm sido desenvolvidas com o objetivo de explicar o que ocorre no cérebro quanto à integração *bottom-up* e *top-down*, além de discussões sobre qual via exerce maior importância durante o processo de atenção visual.

Em (Theeuwes, 1992), são apresentados experimentos baseados na busca por um alvo específico, com características conhecidas previamente pelos observadores, ou seja, uma busca a partir de um enviesamento baseado em informações prévias - *top-down*. Conforme apresentado na Figura 2.7, em cada uma das imagens apresentadas, o alvo previamente conhecido possui a forma de um círculo na cor verde, contendo em seu interior um segmento de linha com orientação horizontal ou vertical. Os distratores podem variar em relação à forma, onde o alvo se encontra meio a quadrados verdes, podendo ainda existir um único distrator de cor vermelha. Outra variação conta com distradores que variam em relação à cor, neste caso, o alvo se encontra meio a distratores de mesma forma, porém de cor diferente, entretanto, entre os distratores, pode existir um com forma diferente. Foi possível concluir que a procura por um alvo que contraste com os demais elementos em relação à cor, a presença de um distrator de forma diferente basicamente não gera efeito, ou seja, não interfere no tempo necessário para a busca do alvo. Entretanto, em um ambiente no qual o alvo se difere dos distratores em relação à forma, a procura torna-se mais lenta na presença de um distrador de cor diferente. Este experimento corrobora com

Theeuwes (1991), e demonstra que a atenção visual baseada exclusivamente em uma característica *top-down*, neste caso a forma, não é suficiente para guiar a atenção.

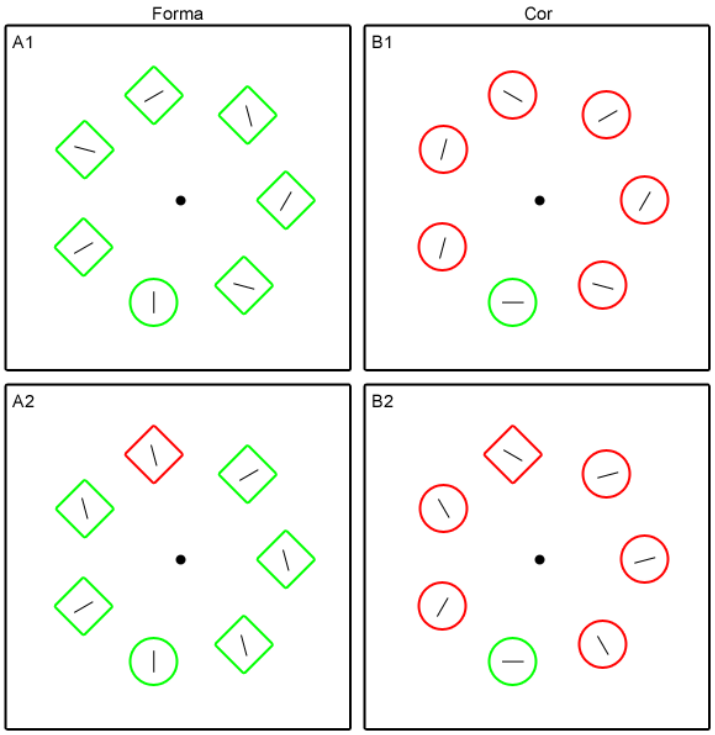


Figura 2.7: A influência da atenção visual baseada em características *top-down* e *bottom-up* (Theeuwes, 1992).

De forma análoga ao experimento apresentado por Theeuwes (1992), Ogawa and Komatsu (2004) apresenta o registro da atividade de neurônios individuais da área V4[2], durante a procura por um alvo previamente conhecido. As imagens foram compostas sempre por seis elementos, sendo dois destes contrastantes em cor e forma, respectivamente. As repostas neurais foram analisadas sob duas condições: a primeira caracterizada pela busca baseada na cor contrastante (duas caixas à direita), onde o alvo foi o elemento de cor diferente dos demais, e a segunda baseada na busca pela forma contrastante (duas caixas à esquerda), considerado alvo àquele elemento de forma diferenciada. As imagens ainda contaram com um estímulo inserido no campo visual (círculo cinza), com o objetivo de salientar um determinado objeto da imagem, podendo o estímulo estar sobre um elemento diferente em forma (caixas vermelhas) ou sobre um elemento diferente em cor (caixas verdes), representado sempre por uma cruz dentro do círculo cinza e, dependendo do tipo de busca, ser o elemento alvo (linhas contínuas) ou um distrator (linhas tracejadas). Em ambas as condições de busca, a taxa de disparo dos neurônios referente ao objeto estimulado foi maior quando o elemento estimulado fosse contrastante em cor. Assim, mesmo quando a estratégia *top-down* foi focar em buscas específicas (forma ou cor), elementos com características contrastantes receberam o foco da

[2]Conforme descrito na Seção 2.1, as células da área V4 são células que respondem à cor e forma.

atenção.

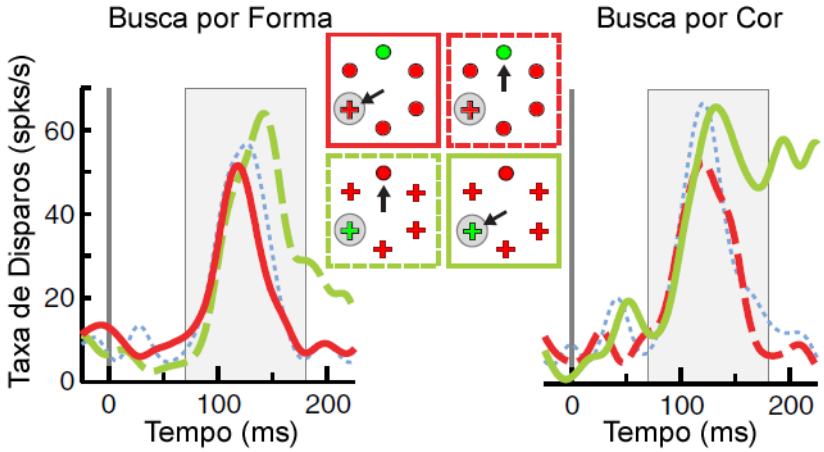


Figura 2.8: Atividade de neurônio da área V4 relacionada à atenção visual *top-down* e *bottom-up* (Ogawa and Komatsu, 2004).

Por outro lado, Bacon and Egeth (1994) demonstraram que, em experimentos onde há uma maior diversidade de elementos, a atenção visual *top-down* pode não sofrer com a presença de um distrator contrastante como, por exemplo, um distrator de cor diferente. A inspiração para esta hipótese foi tornar ineficaz o modo de detecção baseado apenas em características simples, forçando o observador a buscar por uma característica específica do alvo. A Figura 2.9 apresenta um experimento onde foram inseridos alvos redundantes, de modo que o alvo (indicado por uma seta) não fosse o único elemento de determinada forma. Com este experimento podemos concluir que não houveram efeitos de distração causado por elementos de cor contrastante, demonstrando a possibilidade do mecanismo *top-down* em guiar a atenção visual.

De maneira elucidativa, a prioridade do processo de atenção é determinada pela junção das atenções *top-down* e *bottom-up*. De acordo com Egeth and Yantis (1997) e Wolfe (1994), caso um observador esteja à procura de um alvo específico (*top-down*), por exemplo, uma linha vertical vermelha, todos os demais elementos presentes na cena que possuam a cor vermelha ou orientação vertical receberão maior prioridade de atenção *top-down* do que aos elementos que não possuam estas características, ou então, uma cena contendo um objeto vermelho meio a objetos verdes terá maior ativação *bottom-up* do que um objeto vermelho meio a objetos laranja. Conclui-se que a combinação destes dois mecanismos dará origem a um mapa de atenção, conhecido amplamente como mapa de saliência, responsável por determinar a ordem em que os objetos serão visitados durante a pesquisa visual. Proposto inicialmente por Koch and Ullman (1985), o mapa de saliência e modelos baseados nesta hipótese serão apresentados na Seção 4.1 e, dado sua importância para o desenvolvimento desta tese, seu desenvolvimento computacional será detalhado na Seção 3.1.

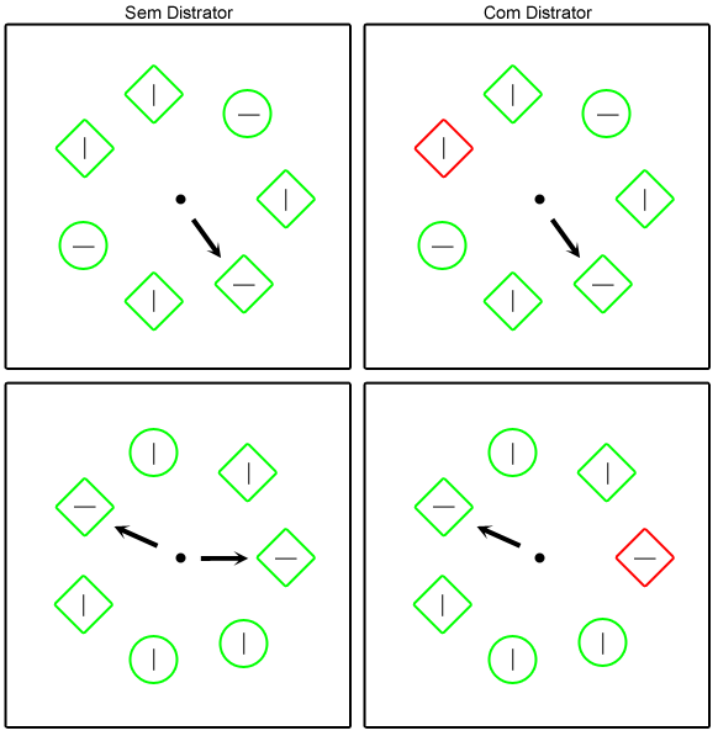


Figura 2.9: Atenção *top-down* meio à distradores (Bacon and Egeth, 1994).

Podemos concluir que a atenção visual depende do contraste relativo de cada característica *bottom-up*, de forma que a atenção *top-down* possua a função de modular a atenção *bottom-up* para elementos de maior importância na cena, ou seja, a atenção *bottom-up* tem como função alertar para itens salientes na imagem, enquanto que a atenção *top-down* torna possível a modulação dos sinais *bottom-up* quando existe a necessidade de direcionar a atenção para algo específico (Deco and Rolls, 2005; Connor et al., 2004).

2.4 Atributos no Comportamento da Atenção

Como observado nos experimentos apresentados, a atenção visual pode ser guiada por características específicas de um estímulo visual, conhecido também como atributo visual. Estas características são conhecidas como características pré-atentivas (por exemplo, um objeto "vermelho"), no sentido em que a informação visual sobre esta dimensão pré-atentiva (neste caso, "cor") deva estar disponível previamente ao processo de seleção para guiar a atenção (Wolfe, 2005). Sendo assim, quando um item de uma imagem difere dos demais em uma dimensão pré-atentiva, esta seleção está relacionada à atenção *bottom-up* e, caso o observador esteja a procura de um alvo com uma característica pré-atentiva específica, esta seleção diz respeito à atenção *top-down* ou enviesamento *top-down*.

Engel et al. (1997) demonstram uma importante característica relacionada à cor para o desenvolvimento da atenção visual. Segundo os autores, a retina humana

é composta por três classes de cones (*Long*, *Middle* e *Short*) que respondem, respectivamente, aos comprimentos de ondas de luz, sendo longos, médios e curtos, de modo que a aparência das cores são resultados dos processamentos destes cones dentro da retina e do cérebro. De acordo com imagens de ressonâncias magnéticas das áreas do córtex visual V1 e V2, Engel et al. (1997) concluem que a maior atividade neural está relacionada à estímulos compostos pelas cores "vermelho-verde", onde a atividade é medida a partir de neurônios recebendo entradas opostas a partir dos cones L e M. Outra forte resposta também é apresentada pelos estímulos "azul-amarelo", o que sugere que os canais de cores oponentes podem ser considerados como um relevante mecanismo para o desenvolvimento da atenção visual.

Outros atributos como, por exemplo, orientação e tamanho, também são responsáveis por guiar o mecanismo biológico de atenção visual (Wolfe and Horowitz, 2004). Para o entendimento do processo de atenção visual é importante observar que a busca por um ponto de maior atenção ou saliência pode ser simples e eficiente em alguns casos, porém não tão simples para outros. Considerando a Figura 2.10 (a), a tarefa de encontrar o alvo "vermelho", o maior alvo ou ainda o alvo inclinado, seria uma tarefa fácil. O atributo cor, orientação ou tamanho pode guiar, de forma eficiente, a atenção para o alvo. Na mesma Figura, entre os "5"*s* existe um alvo "2". Depois de ter sido encontrado, não há nenhuma dificuldade em encontrar novamente o alvo "2" entre os "5"*s*. No entanto, a atenção não pode ser guiada pela informação sobre a posição espacial que diferencia esses caracteres. Além disso, quanto maior a quantidade de "5"*s* (distratores) presente na cena, mais difícil se tornará a pesquisa. Resumindo, é fácil encontrar o "5 vermelho", inclinado ou grande. Entretanto, não é fácil encontrar o "2" entre os "5"*s*. É difícil encontrar os pares de triângulos na horizontal na mesma Figura em (b), mas em (c) é fácil, pois esta tarefa é simplificada devido ao contraste de cores entre os retângulos azuis e os retângulos rosas, em (d) a busca por cruzes é ineficiente, devido ao fato de que a informação de intersecção não guia a atenção (Wolfe and Horowitz, 2004).

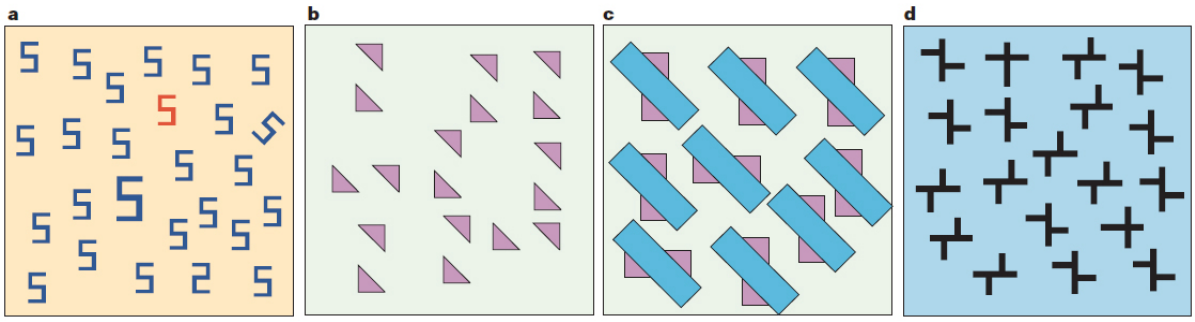


Figura 2.10: Exemplos fáceis e difíceis de busca visual (Wolfe and Horowitz, 2004).

De acordo com Wolfe and Horowitz (2004), propriedades em uma cena que levam à atenção são baseadas na teoria da integração de características proposta por Treisman and Gelade (1980) (veja Seção 2.2), onde a arquitetura para a visão

Figura 2.11: Modelo de Processamento da Atenção (Adaptado de Wolfe and Horowitz (2004)).

humana é composta por duas fases: a primeira responsável por identificar, de forma paralela, as características básicas da imagem, como por exemplo, cor, orientação, tamanho, movimento, profundidade, etc, e na segunda fase, passar por um "gargalo atencional", o que levará à processos responsáveis, como por exemplo, por integrar características primitivas, entregar a atenção e reconhecer os objetos (Figura 2.11).

A tarefa de identificação de regiões salientes através de características simples, ou da combinação de características simples, é geralmente eficiente. Por exemplo, como apresentado na Figura 2.12 (a), é fácil encontrar um "X" de cor preta devido a conjunção de forma e polaridade de luminância. Neste caso, a luminância, ou cor, direciona a atenção para o ponto preto e a forma para o item formado por linhas cruzadas. Estas características são suficientes para guiar a atenção rapidamente à interseção dos dois conjuntos de itens. Na Figura 2.12 (b) é apresentado uma situação interessante, onde a atenção é guiada, inicialmente, para os seguimentos de cor azul e orientação desconsiderada, deslocando o foco da atenção, em seguida, a atenção se volta para os segmentos com orientação horizontal e cor desconsiderada, nesta ordem. Na Figura 2.12 (c) o "L vermelho" e o "T" surgem rapidamente, porém em (d), devido a alterações na orientação, o "T" não é encontrado facilmente.

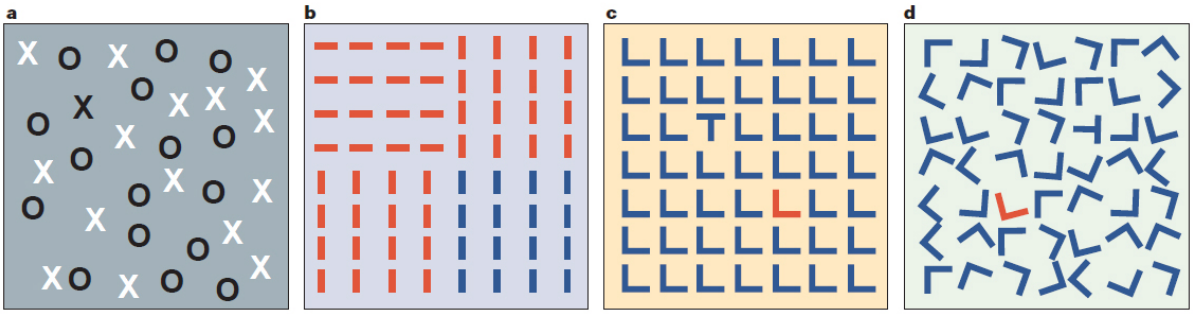


Figura 2.12: Pistas para a orientação (Wolfe and Horowitz, 2004).

Quando uma única característica é considerada para guiar a atenção, é possível que o aumento do número de distratores possa prejudicar a eficiência do processo de busca. A Figura 2.13 demonstra, gradativamente de (a) a (d), a eficiência da busca pelo alvo meio a distratores. Na Figura 2.13 (e) e (g), é uma tarefa simples encontrar

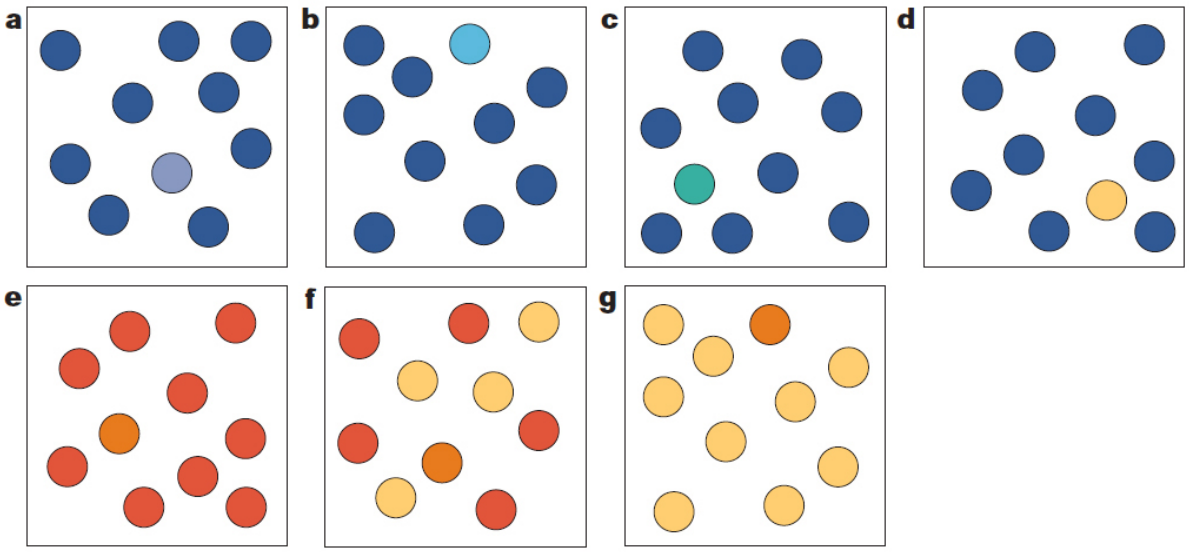


Figura 2.13: Busca meio a distratores (Wolfe and Horowitz, 2004).

o ponto de atenção, pois este se encontra inserido a distratores homogêneos, porém na Figura 2.13 (f), a tarefa de encontrar o ponto de atenção, meio a distratores heterogêneos, torna-se mais difícil.

Geralmente o número de itens (N_I) presentes em uma imagem influenciam diretamente no tempo de reação (TR) da seleção visual. Caso um alvo seja definido por um único atributo pré-atentivo, o declive da função TR x N_I deve ser próximo de zero (Wolfe, 2005). Assim, em buscas como, por exemplo, vermelho entre verdes, vertical entre horizontais e grande apresentarão valores de função próximo a zero, indicando que a cor, orientação e tamanho são dimensões pré-atentivas aptas a guiar a atenção visual (Wolfe, 2005; Treisman and Gormican, 1988; Treisman and Gelade, 1980).

É importante notar que, conforme apresentado na Seção 2.2, a possibilidade da existência de itens com características similares, sendo, neste caso, necessário realizar uma busca cognitiva baseada na conjunção de duas ou mais dimensões (Wolfe, 2005; Ogawa and Komatsu, 2004; Egeth and Yantis, 1997). Esta situação é apresentada na Figura 2.14, onde o grande quadrado branco é encontrado eficientemente, devido à atenção ser guiada tanto pelo atributo cor quanto pelo atributo tamanho (Wolfe, 2005, 1994; Wolfe et al., 1989).

A escolha por atributos a serem considerados para guiar a atenção visual é uma tarefa importante para a eficiência do processo de busca ao local de maior saliência da cena. Tradicionalmente, de acordo com a teoria da integração de características e estudos comportamentais (Treisman and Gelade, 1980; Theeuwes, 1992; Wolfe and Horowitz, 2004; Ogawa and Komatsu, 2004; Wolfe, 2005, 2007; Borji and Itti, 2013), destacamos aqui alguns atributos indiscutíveis à atenção visual, sendo os seguintes: cor, orientação, intensidade, tamanho e movimento. Em (Wolfe and Horowitz, 2004), é apresentado um estudo de outros prováveis e possíveis atributos

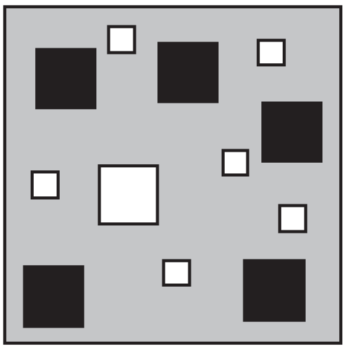


Figura 2.14: Em busca do grande quadrado branco (Wolfe, 2005).

que podem guiar o desenvolvimento da atenção visual. A descrição dos atributos citados será apresentada no Capítulo 4, de acordo com sua utilização nos modelos apresentados.

2.5 Considerações Finais

Neste capítulo foram apresentados conceitos de atenção visual considerados importantes para o desenvolvimento desta tese como, por exemplo, a distinção entre o desenvolvimento da atenção visual a partir de informações *bottom-up* e *top-down*.

De acordo com Kohonen (2001) e trabalhos apresentados neste capítulo, pode-se concluir que as células do córtex cerebral dos mamíferos organizam-se de forma altamente estruturada em suas funções, resultando em regiões do cérebro especificamente capacitadas no processamento sensorial de sinais como visão, audição, controle motor, linguagem, etc, podendo ser visto como um mapeamento estruturalmente organizado para fins específicos, de modo que neurônios apresentem uma ordenação física, onde estímulos semelhantes são processados por neurônios fisicamente próximos entre si no córtex cerebral.

Considerando o fluxo da informação através do córtex visual, apresentado na Seção 2.1, a percepção de cores, por exemplo, ocorre em uma região específica do córtex (V4) e, de acordo com Kohonen (2001), esta região representa um mapa de cores altamente estruturado e auto-organizado, além de considerar a existência de muitos outros tipos de mapas presentes no córtex.

Sendo assim, como um dos pontos de investigação desta tese, pretendemos utilizar como inspiração biológica o conceito de mapas auto-organizáveis representativos de regiões específicas do cérebro, no que refere-se aos conceitos de visão e atenção visual apresentados neste capítulo. Como embasamento teórico, na Seção 3.3 serão apresentados os principais conceitos utilizados para o desenvolvimento computacional deste tipo de mapa, baseado-se na proposta de Kohonen (2001).

No capítulo seguinte serão apresentados os principais fundamentos teóricos relacionados ao desenvolvimento desta tese.

CAPÍTULO

3

Fundamentos Teóricos

A extração de características primitivas, fator fundamental à geração do mapa de saliência, pode ser considerado um dos principais elementos deste trabalho. Sendo assim, iniciaremos a fundamentação teórica necessária ao entendimento desta tese a partir da descrição detalhada de um dos modelos de atenção visual mais referenciados e utilizada, proposto por Itti and Koch (2001). Em seguida serão apresentados os modelos de redes neurais artificiais que servirão como base para o desenvolvimento dos modelos de atenção visual propostos, destacando-se os modelos que empregam em sua formulação conceitos de correlação temporal e segmentação de imagens reais.

3.1 O Mapa de Saliência

Segundo Itti and Koch (2001), cinco pontos importantes devem ser considerados em trabalhos sobre modelos computacionais para atenção visual *bottom-up*. Primeiro, a percepção da saliência do estímulo de entrada depende criticamente do contexto ao seu redor; segundo, um único mapa de saliência que codifica topograficamente estímulos conspicuosos (salientes) sobre a imagem de entrada tem se mostrado uma estratégia de controle *bottom-up* eficiente; terceiro, o processo de inibição de retorno, que tem como função impedir que uma região focada anteriormente seja novamente focada, é um elemento muito importante para o desenvolvimento da atenção; quarto, a interação entre a atenção e os movimentos dos olhos têm inserido desafios computacionais com relação ao sistema de coordenadas usado para controlar a atenção; e por último, a compreensão de cena e reconhecimento de objetos condicionam fortemente a escolha dos locais de atenção.

O mapa de saliência é amplamente utilizado por modelos *bottom-up* e é for-

mado pela composição de vários mapas com características visuais primitivas da imagem como, por exemplo, cor, intensidade e orientação. Esta composição gera uma medida de saliência independente de qualquer dimensão de característica. As diversas regiões da cena visual disputam essa medida em várias escalas espaciais e a região vencedora é eleita como a mais saliente.

Para o entendimento da geração de um mapa de saliência, será descrito a seguir os principais aspectos do modelo proposto em (Itti and Koch, 2001).

3.1.1 Extração de Características Visuais Primitivas

O modelo proposto em (Koch and Ullman, 1985) é baseado na decomposição de uma imagem em um conjunto de canais distintos, o que denomina a extração de características visuais primitivas ou características de baixo nível. Estas características são extraídas da imagem original em várias escalas espaciais, através de filtragens lineares. Sendo assim, primeiro estágio de processamento em qualquer modelo de atenção *bottom-up* é a computação de características visuais primitivas, baseado na inspiração biológica de que estas características são analisadas, de forma pré-atencional, paralelamente por todo o campo visual (Itti and Koch, 2001).

De um modo geral, o modelo é alimentado inicialmente por uma imagem estática e, em seguida, são extraídas as seguintes características visuais: cor, intensidade e orientação. Com r, g e b sendo os canais vermelho, verde e azul da imagem de entrada, uma imagem de intensidades é obtida por $I = (r + g + b)/3$. I será utilizado para criar a pirâmide Gaussiana I_σ, discutida a seguir. Para fins de normalização, os canais r, g e b são normalizados a partir de I. O objetivo desta normalização é inibir regiões que apresentem baixos valores de luminosidade (não salientes). Neste caso, r, g e b são normalizados somente onde I for maior do que $1/10$ de seu valor máximo sobre toda imagem. Em seguida, quatro canais de cores são criados: $R = r - (g + b)/2$ para o vermelho, $G = g - (r + b)/2$ para o verde, $B = b - (r + g)/2$ para o azul e $Y = (r + g)/2 - |r - g|/2 - b$ para o amarelo. Valores negativos são atribuídos zero (Itti et al., 1998). A Figura 3.1 apresenta um exemplo de extração das características citadas.

Com o objetivo de obter amostras da imagem onde detalhes indesejados e ruídos sejam suprimidos e características importantes realçadas, uma pirâmide Gaussiana composta por nove níveis é gerada referente a cada canal, sendo: I_σ, R_σ, G_σ, B_σ e Y_σ, onde $\sigma \in [0..8]$. A pirâmide Gaussiana é gerada de acordo com o algoritmo proposto em (Burt et al., 1983), descrito na seção seguinte.

3.1.2 Pirâmide Gaussiana

A pirâmide gaussiana de Burt et al. (1983) é obtida através de operações progressivas de filtragem passa-baixas e sub-amostragem. No modelo de saliência de Itti et al. (1998), foi considerado um filtro Gaussiano com dimensões 5x5 *pixels*.

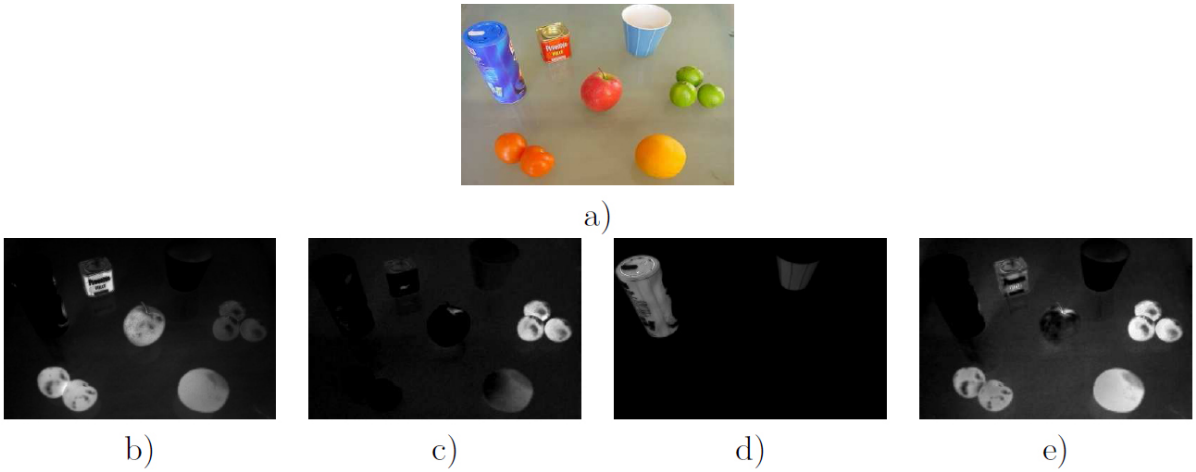


Figura 3.1: Extração de 4 canais de cores. a) Imagem de Entrada, b) Extração do canal vermelho, c) Canal verde, d) Canal azul e e) Canal amarelo (Siklossy, 2005).

A representação em forma de pirâmide é utilizada com o objetivo de destacar características salientes e inibir demais regiões da imagem.

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

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

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

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

3.1.3 Pirâmide Direcional

O modelo de Itti et al. (1998) também considera informações sobre orientações locais como uma característica importante para o desenvolvimento da atenção visual. De acordo com Greenspan et al. (1994), a extração destas características pode ser obtida através da aplicação de filtros direcionais sobre a imagem, tendo como principal objetivo aproximar o perfil de sensibilidade do campo receptivo dos neurônios de orientação seletiva presente no córtex visual primário. Na Figura 3.2 é apresentado um exemplo em que o contraste na orientação pode guiar a atenção

visual.

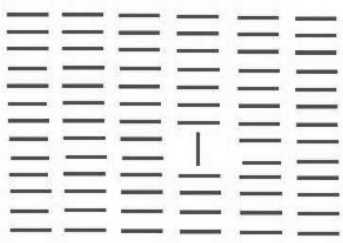


Figura 3.2: Exemplo de orientação. Barra vertical inserida em um ambiente com barras horizontais torna-se o elemento mais saliente devido a grande diferença de orientação (Siklossy, 2005).

Os mapas de orientações $O_\sigma(\theta)$ são criados através da convolução do mapa de intensidades I_σ, com filtros direcionais de Gabor para quatro orientações $\theta \in$ 0°,45°,90°,135°. A aplicação destes filtros visa identificar barras ou bordas em uma determinada direção, para isso utiliza-se de uma função gaussiana, como representada na Figura 3.3.

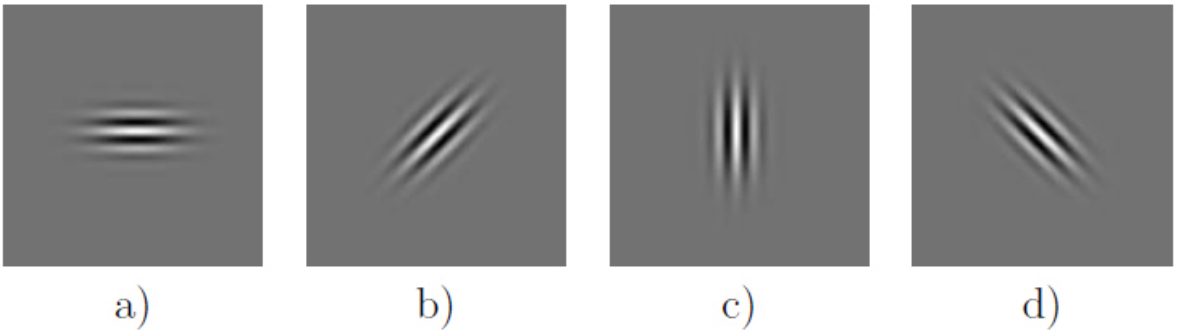


Figura 3.3: Imagem das intensidades dos quatro *kernels* de Gabor utilizados para determinar a informação da orientação local. a) 0°, b) 45°, c) 90° e d) 135° (Siklossy, 2005).

Para demonstrar o processo de filtragem linear, na Figura 3.4 é apresentado o resultado deste processo em uma imagem sintética para os quatro *kernels* considerados. Como pode ser observado na Figura 3.4 (c), onde aplica-se o filtro linear com orientação de 45°, a informação irrelevante é totalmente suprida, sendo destacada somente as barras com orientação de 45°.

3.1.4 Diferenças Centro-Vizinhança

A diferença centro-vizinhança ($\ominus$) é implementada como a diferença entre escalas finas e grossas, ou seja, o centro é um pixel da imagem na escala $c \in \{2, 3, 4\}$ e a vizinhança é o pixel correspondente em outra escala $s = c + \delta$, com $\delta \in \{3, 4\}$ da representação piramidal. A subtração destas duas imagens é obtida pela interpolação das imagens para a escala fina, seguida da subtração ponto a ponto. O primeiro conjunto de mapas é construído a partir do contraste de intensidades,

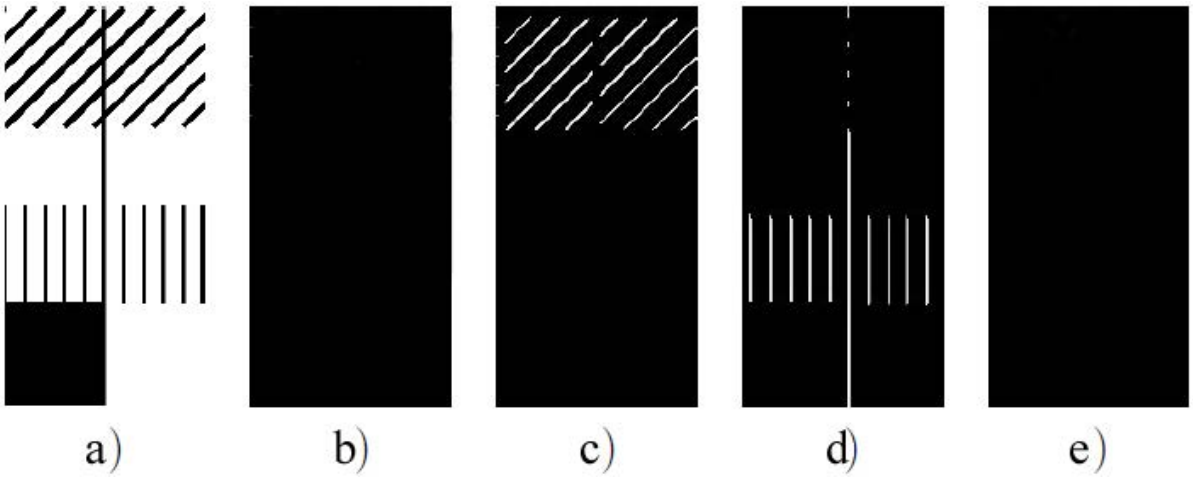


Figura 3.4: Extração de informação orientada utilizando filtragem linear de Gabor. a)Imagem de entrada, b)Informações filtradas com 0°, c)45°, d)90° e e)135°. Adaptado de Siklossy (2005).

definido como segue:

$$\mathcal{I}(c,s) = |I(c) \ominus I(s)|\,, \tag{3.3}$$

que apresenta inspiração biologicamente baseada nos mamíferos, onde 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 (Itti and Koch, 2001). O segundo conjunto de mapas é similarmente construído a partir dos canais de cores, definidos como:

$$\mathcal{RG}(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \tag{3.4}$$

$$\mathcal{BY}(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|\,, \tag{3.5}$$

onde 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 $vermelho/verde$, $verde/vermelho$, $azul/amarelo$, $amarelo/azul$ (Itti and Koch, 2001). O terceiro conjunto de mapas é construído a partir de informações de orientação local, de acordo com as seguintes equações:

$$\mathcal{O}(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)|\,, \tag{3.6}$$

onde $\theta \in 0°,45°,90°,135°$. Neste caso, a inspiração biológica para a construção dos mapas de orientação é a propriedade de neurônios do sistema visual de responder apenas a uma determinada classe de estímulos, como por exemplo barras orientadas verticalmente (Itti and Koch, 2001) (veja Figura 3.2).

3.1.5 Saliência

A maioria dos modelos de atenção *bottom-up* inspirados biologicamente segue a hipótese de 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 bidimensional de atividade representa topograficamente pela conspicuidade ou saliência visual (Itti and Koch, 2001). Uma região ativa em um mapa de saliência codifica o fato desta região ser saliente, não importando se esta corresponde, por exemplo, a uma bola vermelha meio a bolas verdes, ou a um objeto que se move para a esquerda enquanto outros se movem para a direita.

Para a construção de um único mapa de saliência, os mapas de características são individualmente somados ($\bigoplus$) nas diversas escalas, gerando três mapas de conspicuidades: $\bar{\mathcal{I}}$ para intensidade, $\bar{\mathcal{C}}$ para cor e $\bar{\mathcal{O}}$ para orientação. Entretanto, um fator importante a ser notado é que, previamente à somatória dos mapas de cada característica, Itti et al. (1998) propõem sua normalização, denotada por $\mathcal{N}(.)$, com o objetivo de que uma região que apresente um nível de saliência contrastante com as demais seja amplificada e, por outro lado, regiões salientes não contrastantes sejam mutuamente inibidas. A Figura 3.5 demonstra a função da normalização $\mathcal{N}(.)$

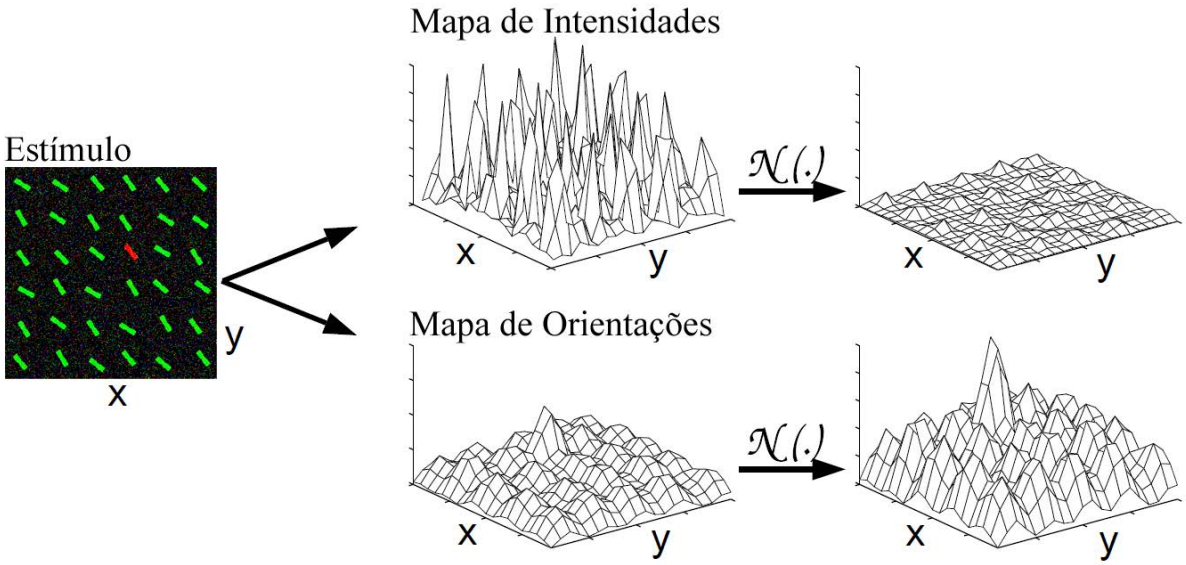


Figura 3.5: Exemplo do comportamento do operador de normalização $\mathcal{N}(.)$.

Após o processo de normalização, os mapas de características são então combinados em três mapas de conspicuidades, conforme descrito anteriormente, definidos como segue:

$$\bar{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s)) \tag{3.7}$$

$$\bar{\mathcal{C}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))] \tag{3.8}$$

$$\bar{\mathcal{O}} = \sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c,s,\theta)) \right) \tag{3.9}$$

De acordo com Itti et al. (1998), a motivação para a criação dos três canais separados ($\bar{\mathcal{I}}$,$\bar{\mathcal{C}}$,$\bar{\mathcal{O}}$) é a hipótese de que características similares competem pela saliência, enquanto modalidades diferentes contribuem independentemente para o mapa de saliência.

Finalmente, os três mapas de conspicuidades são normalizados e somados, resultando em uma entrada final para o mapa de saliência $\mathcal{S}$, como segue:

$$\mathcal{S} = \frac{1}{3}(\mathcal{N}(\bar{\mathcal{J}}) + \mathcal{N}(\bar{\mathcal{C}}) + \mathcal{N}(\bar{\mathcal{O}})) \tag{3.10}$$

3.1.6 Seleção da Atenção e Inibição de Retorno

Para o desenvolvimento da seleção visual entre as regiões mais salientes do mapa de saliência, Itti et al. (1998) utilizam uma rede neural composta por neurônios do tipo Integra e Dispara (apresentado em detalhes na Seção 3.2.2), que tem como função representar o mapa de saliência $\mathcal{S}$, de modo que a estimulação externa de cada neurônio é definida pelo valor de saliência dos respectivos pontos no mapa de saliência.

Por sua vez, a rede de neurônios Integra e Dispara alimenta uma rede neural do tipo WTA (*Winner-Takes-All*) (Koch and Ullman, 1985; Tsotsos et al., 1995), considerada 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). É importante notar que os neurônios da rede Integra e Dispara são utilizados, neste caso, somente como integradores dos valores de $\mathcal{S}$, onde cada neurônio tem como função ativar seu correspondente neurônio WTA.

Na WTA, todos os neurônios recebem ativação de forma independente, até que o neurônio vencedor (*winner*) alcança o limiar e dispara, desencadeando simultaneamente três mecanismos: primeiro, o foco da atenção é direcionado para a localização do neurônio vencedor; segundo, o inibidor global da WTA é acionado e todos os demais neurônios da WTA são inibidos; terceiro, a região sob o foco da atenção é temporariamente inibida na rede de neurônios Integra e Dispara, permitindo que a próxima região saliente seja destacada, garantindo também que o foco da atenção

não seja novamente direcionado para a região anterior, caracterizado um mecanismo de inibição de retorno (maiores detalhes podem ser encontrados em Itti et al. (1998)).

3.2 Sincronismo e Dessincronismo em Redes Neurais Pulsadas

As Redes Neurais Artificiais (RNAs) são sistemas de processamento paralelo e distribuído compostos por unidades de processamento simples (neurônios) que calculam determinadas funções matemáticas e são capazes de armazenar o conhecimento adquirido e torná-lo disponível para uso. Essas unidades de processamento são dispostas em uma ou mais camadas e interligadas por conexões sinápticas associadas a pesos que são utilizados para ponderar a entrada recebida de cada neurônio da rede e armazenar o conhecimento adquirido (Haykin, 2001).

O desenvolvimento de modelos de RNAs tem atraído a atenção de diversas áreas da ciência desde a década de 50, quando os pesquisadores McCulloch e Pitts desenvolveram o primeiro neurônio artificial (McCulloch and Pitts, 1943). Desde então, modelos de RNAs têm sido propostos e aplicados aos mais diversos problemas, dentre estes, a simulação de sistemas biológicos em neurociência computacional.

As RNAs podem ser divididas em três grandes grupos (Maass, 1997) em relação ao tipo de saída dos neurônios:

- *Saídas binárias*: o primeiro grupo consiste em redes nas quais os neurônios de saída produzem apenas dados binários, como por exemplo, o neurônio de McCulloch-Pitts (McCulloch and Pitts, 1943) e geralmente são utilizadas para desempenharem funções lógicas (Yegnanarayana, 2005).

- *Saídas contínuas*: o segundo grupo é formado por redes neurais em que suas saídas são representadas por funções contínuas, por exemplo, funções sigmóides. As redes da segunda geração são capazes de computar funções com entradas e saídas analógicas. Uma característica desta geração é que estes modelos suportam algoritmos de aprendizagem baseado na técnica do gradiente, como por exemplo, o *Backpropagation* (Haykin, 2001). Um exemplo de utilização dessas redes é a tarefa de aproximação de funções.

- *Saídas pulsadas*: o terceiro grupo é caracterizado pelas Redes Neurais Pulsadas (RNPs) [1]. Os neurônios de uma RNP, denominados *Spiking Neurons*, utilizam como forma de processamento pulsos ao invés de uma saída contínua ou binária, como os neurônios da primeira e segunda geração. Desta forma, além de integrar informação pelo nível de ativação do neurônio, estes modelos de rede são capazes de criar uma representação temporal através da distribuição dos pulsos gerados no tempo.

[1] do inglês *Spiking Neural Networks*

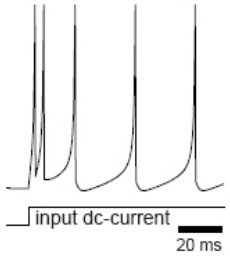


Figura 3.6: Propriedade neuro-computacional de neurônio pulsante biológico. (Izhikevich, 2004).

3.2.1 Redes Neurais Pulsadas

De acordo com Campbell et al. (1999), diferentes características visuais de objetos são processadas por diferentes áreas corticais. Conforme mencionado no Capítulo 2, a associação destas características, de forma a proporcionar a percepção coerente de um objeto, é conhecido como problema de integração de características.

Segundo Izhikevich (2004), nos últimos anos têm ocorrido uma maior ênfase, por parte da comunidade de redes neurais, em relação às RNPs. Motivado pelas descobertas biológicas, muitos estudos consideram as RNPs como um essencial componente no processamento da informação pelo cérebro. Um dos problemas cruciais para estudos envolvendo redes dinâmicas é a escolha do modelo de *spiking neuron* a ser utilizado. O comportamento do modelo condiz diretamente com a abrangência de suas características, o que o torna adequado para tipos específicos de problemas.

Izhikevich (2004) apresenta conjunto de características neurais biológicas relevantes a serem consideradas, de forma que o modelo computacional a ser utilizado implemente a característica mais adequada. Tomamos como exemplo a característica *tonic spiking*. Neste tipo de *spiking neuron*, para testar esta característica, neurofisiologistas injetam pulsos de corrente *dc* via um eletrodo preso ao neurônio e registram o potencial da membrana. A corrente de entrada e a resposta do neurônio são normalmente plotadas uma sob as outras (Figura 3.6). Enquanto a entrada estiver ativa, o neurônio dispara continuamente uma sucessão de pulsos. O disparo constante de tais neurônios indica que existe uma entrada persistente.

De acordo com Izhikevich (2004) e Campbell et al. (1999), diversos modelos matemáticos propostos para modelar a dinâmica neural têm sido propostos, porém, dado sua simplicidade e eficiência computacional, o modelo Integra e Dispara (*Integrate-and-fire*-I&F) é um dos modelos mais utilizados na neurociência computacional, sendo descrito a seguir.

3.2.2 Sincronização em Rede de Osciladores I&F

De acordo com Campbell et al. (1999), para a sincronização de uma RNP composta por neurônios do tipo I&F, o potencial de cada oscilador é representado por

uma simples variável. Quando esta variável atingir um determinado limiar, ocorrerá o disparo, retornando imediatamente seu valor para zero. Quando um oscilador dispara, um sinal de ativação é enviado para seus vizinhos. Pesquisas constatam a habilidade de sincronização da rede (Mirollo and Strogatz, 1990; Álvaro Corral et al., 1995; Hopfield and Herz, 1995). Em termos computacionais, dado o tamanho da rede e o ajuste dos parâmetros, pode-se obter rapidamente o sincronismo (Campbell et al., 1999).

A rede de osciladores I&F é definida como:

$$x_i(t + \Delta t) = x_i(t) + ((-x_i(t) + I) * \Delta t), i = 1, .., n, \tag{3.11}$$

onde x_i representa o potencial do oscilador i, n é o número de osciladores, o parâmetro I controla o período de um oscilador dessincronizado e Δt representa a discretização do tempo de disparo. O limiar de disparo do oscilador é configurado como sendo igual a 1. Quando $x_i = 1$ o oscilador dispara, x_i retorna a zero e envia-se a ativação para seus vizinhos.

Dado o disparo do oscilador i, a atualização do potencial dos osciladores vizinhos é definido como:

$$x_{ij} = x_{ij} + J_{ij}, i = 1, .., Z_i \tag{3.12}$$

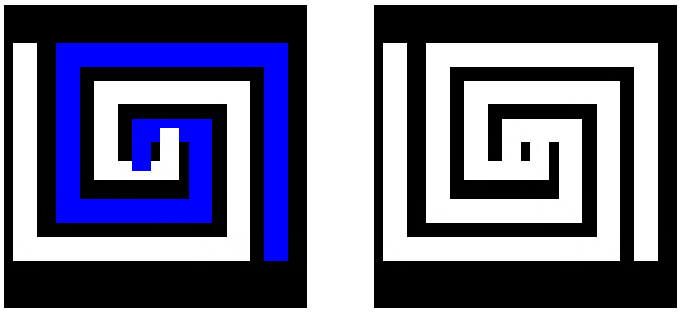
onde x_{ij} é o j-*ésimo* oscilador vizinho de i. Caso x_{ij} exceder o limiar de disparo, também irá disparar. É importante notar que esta informação é transmitida entre os osciladores, podendo se propagar por toda a rede.

O número de acoplamentos com osciladores vizinhos podem ser considerados dois ou quadro, dependendo da estrutura da rede, 1D ou 2D, respectivamente. De maneira geral, a força de acoplamento a partir do oscilador i para j é normalizado como segue:

$$J_{ij} = \frac{\alpha_s}{Z_i}, \tag{3.13}$$

onde Z_i é o número de vizinhos do oscilador i, por exemplo, $Z_i = 2$ para um oscilador i localizado no canto de um sistema 2D. A constante α_s é o fator que define a força do acoplamento. De acordo com Wang and Terman (1995), a normalização garante que todos os osciladores receberão o mesmo potencial de estímulo, auxiliando o processo de sincronização. Note que existem somente dois parâmetros no sistema: α_s e I_i.

Conforme a Figura 3.7, após o processo de sincronização, todos os osciladores pulsam de forma sincronizada, porém sem segmentação da imagem. De acordo com os princípios da correlação oscilatória, faz-se necessário o desenvolvimento do mecanismo que possibilite a dessincronização dos diferentes grupos de osciladores. De acordo com Campbell et al. (1999), o mecanismo de dessincronização utilizado é baseado no modelo proposto por Wang and Terman (1995), onde a principal di-

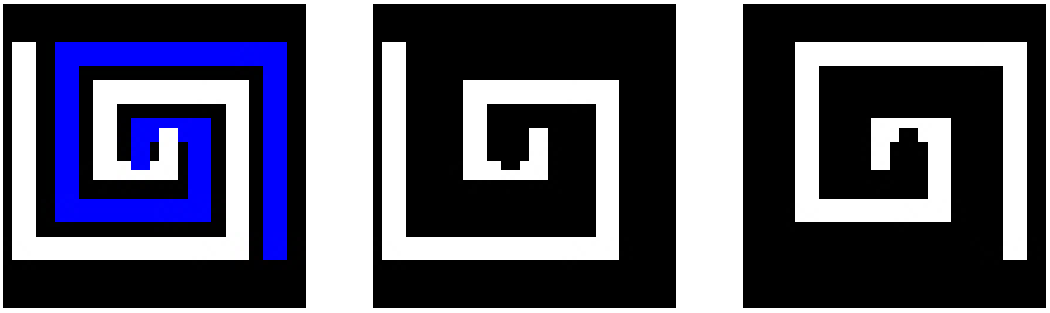


(a) Entrada **(b)** Sincronização Total

Figura 3.7: Osciladores não Segmentados.

ferença é o uso de osciladores do tipo I&F ao invés dos osciladores de relaxamento utilizados na LEGION. Sendo assim, discutiremos o processo de dessincronização na Seção 3.2.3.

Entretanto, nos trabalhos iniciais apresentados nesta tese (Capítulo 5), consideramos a utilização do atributo cor como característica para selecionar a vizinhança. Na Figura 3.8, por exemplo, pode-se obter cada objeto da imagem representado por um trem de pulsos em sincronia, enquanto que objetos distintos são representados por grupos de osciladores fora de fase.



(a) Entrada **(b)** Osciladores da espiral 1 **(c)** Osciladores da espiral 2

Figura 3.8: Osciladores Segmentados.

3.2.3 Rede LEGION

A capacidade de agrupar elementos (objetos) de uma determinada cena ou campo visual é um aspecto fundamental da percepção. Um outro modelo que tem se destacado na tarefa da sincronização é a Rede Osciladora Localmente Excitatória, Globalmente inibitória-*Locally Excitatory, Globally Inhibitory Oscillatory Network* (LEGION). O modelo LEGION, proposto originalmente por Wang and Terman (1995), têm sido uma arquitetura para modelos de correlação oscilatória muito utilizada nos últimos anos. Wang and Terman (1995) demonstram que o modelo é uma eficiente metodologia na implementação de modelos de correlação oscilatória, principalmente pela alta velocidade de aquisição de sincronismo entre os osciladores representando

objetos e dessincronia entre os grupos distintos de osciladores.

A rede LEGION têm sido utilizada em diversas aplicações, demonstrando um comportamento adequado para tarefas de segmentação. Em (Shareef et al., 1999) é apresentado a rede LEGION para a segmentação de imagens médicas, onde comparações dos resultados obtidos com outros métodos sugerem a rede LEGION como um eficiente *framework* computacional para segmentação de imagens. Em um outro trabalho bastante referenciado, Liu et al. (2001) propõem a extração de características geográficas a partir de imagens de sensoriamento remoto baseada na rede LEGION que, comparado com demais métodos tradicionais, demonstrou-se computacionalmente eficiente.

A arquitetura LEGION (Figura 3.9), em sua forma básica, é composta de três elementos principais: osciladores neurais, acoplamentos excitatórios locais e um inibidor global. Os acoplamentos excitatórios locais têm por finalidade sincronizar os grupos de osciladores representando cada um dos objeto presentes na cena visual. Por outro lado, o inibidor global tem como função gerar a dessincronização entre os grupos de osciladores. Desta forma a rede cria um mecanismo de cooperação local e competição global que são os dois requisitos necessários para a implementação da correlação oscilatória.

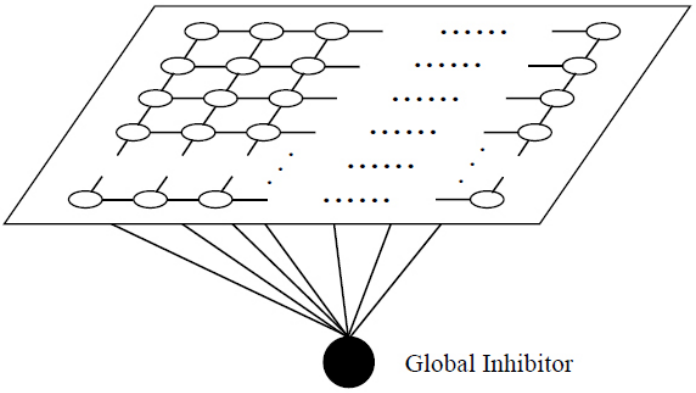


Figura 3.9: Arquitetura LEGION de duas dimensões. O inibidor global é representado pelo círculo preto (Wang and Terman, 1995).

Para melhor entendimento do comportamento da rede LEGION, iniciaremos sua descrição a partir das equações diferencias que descrevem os osciladores de relaxamento. De acordo com Wang (2005), os osciladores de relaxamento foram inicialmente estudados por van der Pol (1926) e desde então, têm sidos amplamente estudados em diversas áreas como, por exemplo, mecânica, biologia, química e engenharia. Um oscilador de relaxamento i é tipicamente definido pela conectividade recíproca entre unidades excitatórias x_i e inibitórias y_i (Terman and Wang, 1995; Wang, 2005). Definida como segue:

$$\dot{x}_i = 3x_i - x_i^3 + 2 - y_i + I_i \tag{3.14}$$

$$\dot{y}_i = \epsilon\left(\gamma\left(1 + \tanh\left(\frac{x_i}{\beta}\right)\right) - y_i\right), \tag{3.15}$$

onde I_i representa um estímulo externo do oscilador i e γ, β e ϵ são parâmetros do modelo. As isóclinas nulas $\dot{x}_i = 0$ e $\dot{y}_i = 0$ do sistema representam, respectivamente, uma função cúbica e uma função sigmóide. Quando $I_i > 0$, o modelo representa um oscilador de ciclo limite que oscila entre uma fase com valores de x mais elevados e uma segunda fase com valores de x mais baixos (Figura 3.10). Estas fases são denominadas respectivamente de fase ativa e fase silenciosa. Nestas condições, o oscilador é denominado *disparando*.

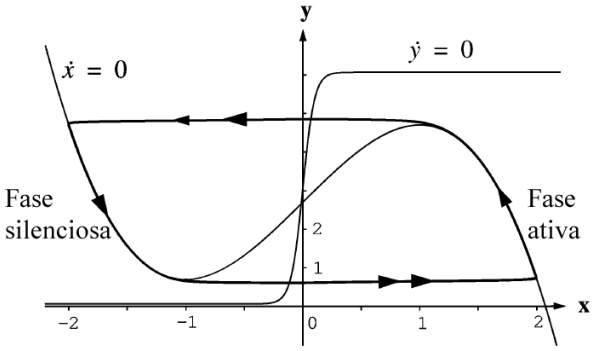


Figura 3.10: Dinâmica de ciclo limite de um oscilador de relaxamento quando $I_i > 0$ (Wang and Terman, 1995).

Quando $I_i < 0$ o comportamento deixa de ser oscilatório e o sistema passa a ser representado por uma dinâmica estacionária com ponto de equilíbrio estável (Figura 3.11). Neste caso, o oscilador é chamado de *excitável*. Sendo assim, a condição de um oscilador estar disparando ou excitável depende diretamente de um estímulo externo.

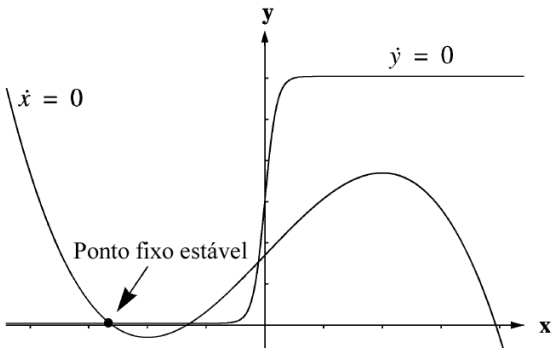


Figura 3.11: Dinâmica de ciclo limite de um oscilador quando $I_i < 0$ (Wang and Terman, 1995).

O parâmetro γ é responsável pelo controle do tempo em que $\dot{x}_i$ permanece em cada umas das fases, de modo que, quanto maior o valor de γ, menor será o tempo

em que $\dot{x}_i$ irá permanecer na fase ativa. O parâmetro β especifica a inclinação da sigmóide e deve ter um valor adequado para que ocorra um único ponto de interseção entre as isóclinas nulas. A Figura 3.12 apresenta as saídas $\dot{x}_i$ de quatro osciladores de relaxamento no tempo t. Consideramos os valores dos parâmetros $I_i = 1$, $\epsilon = 0.01$ e $\beta = 0.2$ para todos os osciladores. Analisando as variações do parâmetro γ, pode-se perceber que o tempo entre as fases ativas são iguais para todos os valores de γ.

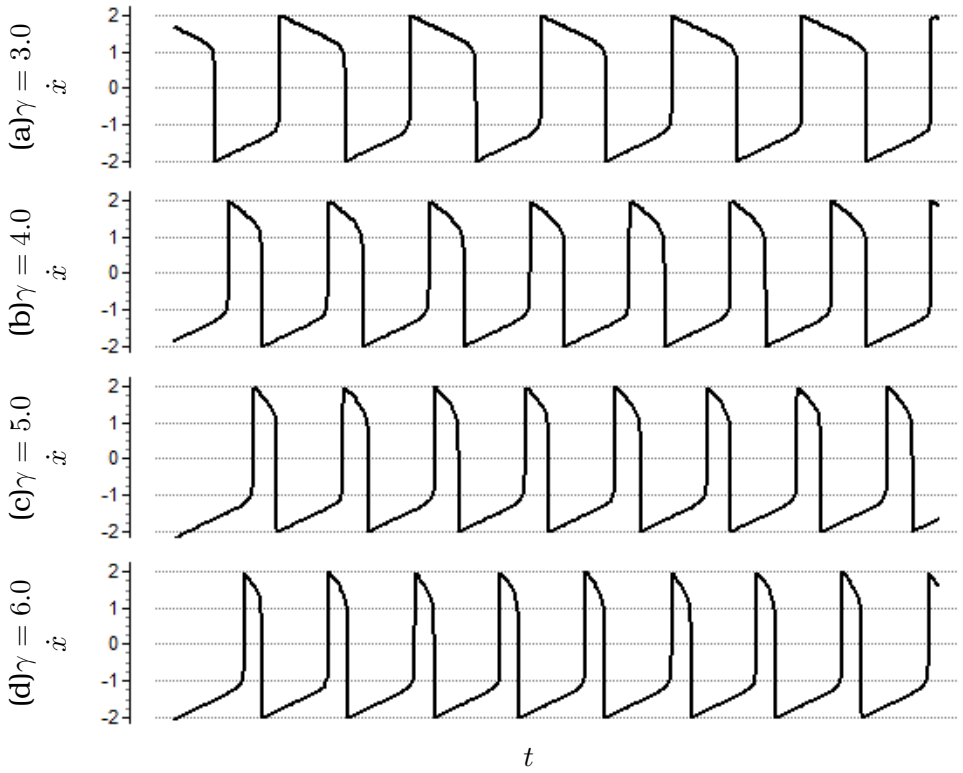


Figura 3.12: Atividade $\dot{x}_i$ de quatro osciladores no tempo t. (a) $\gamma = 3.0$, (b) $\gamma = 4.0$, (c) $\gamma = 5.0$ e (d) $\gamma = 6.0$. Para todos os osciladores utilizou-se os seguintes valores de parâmetros: $I_i = 1.0$, $\epsilon = 0.01$ e $\beta = 0.2$.

O parâmetro ϵ é uma constante positiva com valor pequeno, que define a evolução de $\dot{y}_i$, determinando o tempo de permanência nas fases ativa e silenciosa. Na Figura 3.13 são também apresentadas as saídas $\dot{x}_i$ de quatro osciladores no tempo t. Entretanto, analisamos neste caso variações do parâmetro ϵ. Com $I_i = 1.0$, $\gamma = 3.0$ e $\beta = 0.2$ para todos os osciladores, uma pequena variação em ϵ faz com que as frequência sejam modificadas consideravelmente, porém, o tempo de permanências nas fases ativa e silenciosa permanecem as mesmas em cada oscilador.

Conforme descrito nesta seção, uma rede LEGION é composta por osciladores neurais, acoplamentos excitatórios locais e um inibidor global. De acordo com Wang and Terman (1997), cada oscilador é composto por uma variável excitatória x_i e uma variável inibitória y_i, definidas pelas seguintes equações:

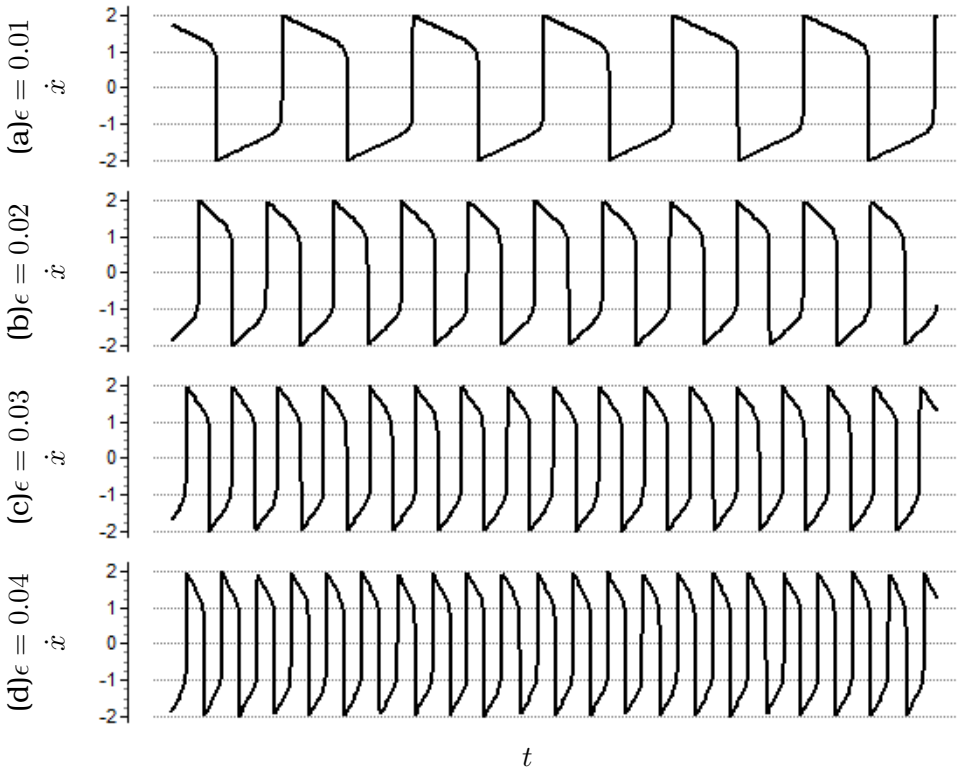


Figura 3.13: Influência de ϵ na atividade $\dot{x}_i$ de quatro osciladores no tempo t. (a) $\epsilon = 0.01$, (b) $\epsilon = 0.02$, (c) $\epsilon = 0.03$ e (d) $\epsilon = 0.04$. Para todos os osciladores utilizou-se os seguintes valores de parâmetros: $I_i = 1.0$, $\gamma = 3.0$ e $\beta = 0.2$.

$$\dot{x}_i = 3x_i - x_i^3 + 2 - y_i + I_i H(p_i + \exp^{-\alpha t} - \theta) + S_i + \rho \tag{3.16}$$

$$\dot{y}_i = \epsilon \left(\gamma \left(1 + \tanh \left(\frac{x_i}{\beta} \right) \right) - y_i \right), \tag{3.17}$$

onde S_i representa o acoplamento a partir dos osciladores vizinhos na rede e ρ é um ruído Gaussiano de pequena amplitude. A função do ruído gaussiano adicionado à entrada de cada oscilador é de evitar que condições iniciais da rede impliquem em estados de estabilidade não desejados e também evitar o possível sincronismo entre diferentes blocos de osciladores (Wang, 1999). Os parâmetros ϵ, γ e β mantêm as mesmas características apresentadas.

Considerando os propósitos de utilização da rede LEGION nesta tese, será descrito o modelo LEGION utilizado para segmentação de imagens proposto por Wang and Terman (1997), visto como uma evolução do modelo LEGION original proposto em (Wang and Terman, 1995). De acordo com Wang and Terman (1997), a principal diferença em relação ao modelo LEGION original é a introdução da função de Heaviside, associada ao estímulo externo I_i, definida como $H(a) = 1$ se $a \geq 0$ e $H(a) = 0$ se $a < 0$. Assim, se $I > 0$ e $H = 1$, o oscilador é denominado *disparando*. Por outro lado, se $I < 0$ e $H = 1$, o oscilador é denominado excitável, indicando que pode se tornar oscilatório, caso receba de seus osciladores vizinhos valores

suficientes do termo de acoplamento S. O parâmetro α é um fator de decaimento positivo definido de mesma ordem de magnitude do parâmetro ϵ, com a função de habilitar como líderes osciladores estimulados durante um determinado período de tempo. De maneira geral, a função de Heaviside é de possibilitar a distinção entre regiões maiores e fragmentos de ruídos presentes na imagem (Wang and Terman, 1997), onde regiões maiores devem conter ao menos um oscilador, denominado *lider*, posicionado no centro de uma região homogênea, que recebe grande estimulação a partir da vizinhança. Por outro lado, o oscilador corresponde a um ruído ou um pequeno fragmento isolado na imagem não recebe uma grande estimulação e não se torna um líder. A variável p_i representa o potencial lateral do oscilador i e determina se um oscilador será ou não um líder, definida pela seguinte equação:

$$\dot{p}_i = \lambda(1 - p_i)H\left[\sum_{k \in N(i)} T_{ik}H(x_k - \theta_x) - \theta_p\right] - \mu p_i, \tag{3.18}$$

onde $\lambda > 0$ é uma constante de tempo, T_{ik} é o peso de conexão permanente entre os osciladores k e i e $N(i)$ é denominado como a *vizinhança* de i. Se o somatório dos osciladores vizinhos $N(i)$ de i for superior a θ_p, p_i aproxima 1. Caso contrário, p_i aproxima zero com velocidade determinada por μ, definido na mesma ordem do parâmetro ϵ. Neste caso, p_i somente irá exceder o limiar θ (Equação (3.16)) se o oscilador i receber estímulo lateral suficiente. É importante notar que, os osciladores vizinhos de i também devem apresentar um certo grau de sincronia, de modo que suas oscilações devam exceder o limiar θ_x.

O termo de acoplamento S_i (Equação (3.16)) é definido pela seguinte equação:

$$S_i = \sum_{k \in N_i} W_{ik}H(x_k, \theta_x) - W_z H(z, \theta_{zx}) \tag{3.19}$$

onde W_{ik} define o peso de acoplamento dinâmico entre os osciladores i e k. De acordo com Wang and Terman (1997), as conexões dinâmicas W_{ik} são formadas com base nas conexões permanentes T_{ik}, seguindo um processo de normalização cujo objetivo é assegurar que todos os osciladores recebam a mesma quantidade de sinal de sua vizinhança. Entretanto, de acordo com o algoritmo computacional para segmentação e imagens reais, também proposto em (Wang and Terman, 1997), o peso dinâmico W_{ik} pode ser definido intuitivamente, baseado na similaridade entre os *pixels* i e k. Por exemplo, considerando imagens em tons de cinza, o peso dinâmico pode ser definido como:

$$W_{ik} = I_M/(1 + |I_i - I_k|), \tag{3.20}$$

onde I_M é o máximo valor do canal de intensidades I. O termo N_i define a vizinhança de interação do oscilador i, representada pelos osciladores que fazem conexão direta com este. O parâmetro θ_x é um limiar que indica quando um oscilador pode afetar

seus vizinhos e W_z define a forca de ligação entre o oscilador i e o inibidor global definido por z. A dinâmica do inibidor global z é definida por:

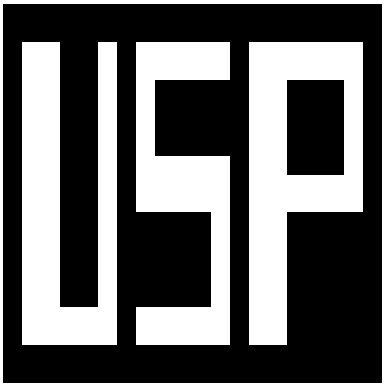
$$\dot{z} = \phi(\sigma_\infty - z), \tag{3.21}$$

onde $\sigma_\infty \equiv 0$ se para todo oscilador $x_i < \theta_{zx}$ e $\sigma_\infty \equiv 1$ se para todo oscilador i pelo menos em um $x_i \geq \theta_{zx}$ sendo θ_{zx} um limiar. Neste caso, se pelo menos um oscilador estiver acima do limiar, o inibidor global z receberá um estímulo e todos os osciladores da rede para o qual z superar o limiar θ_{zx} receberá um sinal de inibição. O parâmetro ϕ determina a taxa de decaimento do sinal de inibição.

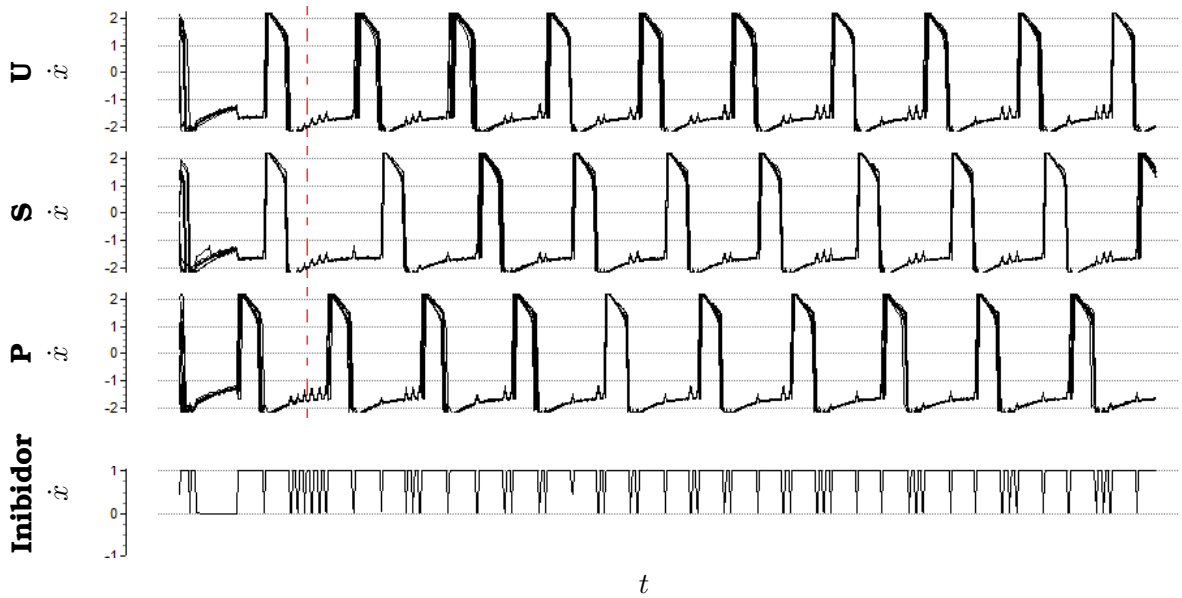
O processo pode ser resumido da seguinte forma: uma vez que um oscilador está na fase ativa, o inibidor global é acionado que por sua vez envia um sinal de inibição para toda a rede. Além disso, o oscilador que entra na fase ativa, também propaga o seu sinal para os seus respectivos vizinhos que por sua vez continuam o processo. Desta forma, a rede apresenta uma forma cooperativa de ativação local, enquanto o inibidor é responsável pela competição global. O inibidor global pode ser interpretado como uma espécie de mecanismo de atenção no qual, uma vez que um segmento está ativo, os demais são inibidos. Devido aos mecanismos de cooperação responsáveis por sincronizar osciladores vizinhos alimentados por sinais semelhantes e de competição cujo objetivo está na separação temporal dos grupos sincronizados, a rede LEGION se torna uma interessante abordagem ao problema da segmentação de imagens. Neste caso, cada ponto da imagem é representada por um oscilador, a sincronização entre estes definem os objetos/segmentos e a segmentação é realizada pela separação temporal das fases dos osciladores. É importante notar que, de acordo com Wang and Terman (1997), o modelo LEGION para segmentação necessita de no máximo $T + 1$ ciclos para segmentar a imagem, onde T representa o número de segmentos. Para maiores detalhes veja Terman and Wang (1995) e Wang and Terman (1997).

Para ilustrar o dinamismo da rede LEGION, a Figura 3.14 apresenta uma simulação computacional de uma rede LEGION 20x20, onde cada oscilador é conectado a 4 vizinhos imediatos (veja Figura 3.9). Para esta ilustração, os osciladores são integrados numericamente utilizando o método Runge-Kutta de quarta ordem. A imagem de entrada é uma imagem binária 20x20 *pixels* apresentada na Figura 3.14 (a), com três objetos (letras U, S e P). Os *pixels* da imagem de entrada foram mapeados para os osciladores com correspondência um-para-um, de forma que a rede de conexões preservaram a adjacência dos *pixels*. Como pode ser observado, a sincronização e dessincronização ocorrem após segundo ciclo oscilatório, denotado pela linha vermelha tracejada.

De acordo com Wang and Terman (1997), a dinâmica da rede LEGION, baseada em equações diferenciais aplicada à segmentação de imagens reais com um grande número de *pixels*, requer um grande esforço computacional. Com base



(a)



(b)

Figura 3.14: Simulação computacional de uma rede LEGION 20x20 para segmentação de uma imagem binária. (a) Imagem de entrada com 3 objetos (letras U, S e P). (b) Atividade temporal $\dot{x}_i$ dos osciladores para as primeiras 15000 integrações. Os parâmetros utilizados foram: $\epsilon = 0.02$, $\alpha = 0.005$, $\beta = 0.1$, $\gamma = 6.0$, $\theta = 0.9$, $\lambda = 0.1$, $\theta_x = -1.1$, $\theta_p = 5.0$, $W_z = 1.5$, $\mu = 0.01$, $\phi = 3.0$, $\rho = 0.02$, $T_{ik} = 2.0$ e $\theta_z = 0.1$.

nisto, os autores desenvolveram um algoritmo computacional seguindo os principais passos da simulação numérica das Equações ((3.16))-((3.21)), definido a seguir.

Algoritmo 1 LEGION

1. Inicialização

   Definir $z(0) = 0$;

   Determinar os pesos de acoplamentos dinâmicos:

   $W_{ik} = I_M/(1 + |I_i - I_k|)$, $k \in N(i)$;

   Encontrar os líderes:

   $p_i = H\left[\sum_{k \in N(i)} W_{ik} - \theta_p\right]$;

   Definir randomicamente todas as saídas $x_i(0)$ para a fase silenciosa:

   $-2 < x_i(0) < -1$.

2. Encontrar o oscilador j mais próximo da fase ativa e dispará-lo, de forma que $x_j(t) \geq x_k(t)$, onde k está correntemente na fase silenciosa e definir $p_j = 1$ e, em seguida, atualiza-se o potencial x_i de todos os demais osciladores:

   $x_j(t+1) = 1; \quad z(t+1) = 1; \quad$ {*disparando*}

   $x_k(t+1) = x_k(t) + (-1 - x_j(t)); \quad$,para $k \neq j$.

3. Iteragir até parar

   Se $(x_i(t) = 1$ e $z(t) > z(t-1))$ então

   $x_i(t+1) = x_i(t) \quad$ {permanece na fase ativa}

   senão

   Se $(x_i(t) = 1$ e $z(t) \leq z(t-1))$ então

   $x_i(t) = -2; \quad z(t+1) = z(t) - 1; \quad$ {*não disparando*}

   se $(z(t+1) = 0)$ então vá para o passo 2

   senão

   $S_i(t+1) = \sum_{k \in N(i)} W_{ik} H(x_k(t) - 1.0) - W_z H(z(t) - 0.5)$;

   Se $(S_i(t+1) > 0)$ então

   $x_i(t+1) = 1; \quad z(t+1) = z(t) + 1; \quad$ {*disparando*}

   senão $x_i(t+1) = x_i(t); \quad$ {*permanece na fase silenciosa*}

De acordo com o Algoritmo 1 apresentado por Wang and Terman (1997), existem dois parâmetros a serem considerados: W_z, que define o peso do inibidor global e θ_p, limiar responsável pela formação de osciladores líderes com altos potenciais laterais.

3.3 Mapas Auto-Organizáveis

A formação de mapeamentos topologicamente corretos é atribuída a uma diversidade de mecanismos, dos quais um em particular, a auto-organização, recebeu bastante atenção da comunidade acadêmica devido a suas fortes evidências

biológicas. Isto levou à proposição de vários modelos de mapas topográficos, ou mapas topologicamente corretos.

De acordo com Haykin (2001) o objetivo principal de um Mapa Auto-Organizado (*Self-Organizing Maps*-SOM) é transformar um padrão de sinal de entrada de dimensão arbitrária em um mapa de uma ou duas dimensões, e realizar esta transformação de maneira topologicamente ordenada. Existem três processos essenciais envolvidos na formação dos mapas auto-organizados:

- *Competição:* para cada padrão de entrada, os neurônios da rede computam seus respectivos valores de uma função discriminante. Esta função discriminante provê a base para a competição entre os neurônios. O neurônio com o maior valor desta função é declarado o vencedor da competição.

- *Cooperação:* o neurônio vencedor determina a localização espacial de uma vizinhança topológica de neurônios excitados.

- *Adaptação Sináptica:* este último mecanismo habilita os neurônios excitados a atualizarem seus valores individuais da função discriminante em relação à entrada, através de ajustes em seus pesos sinápticos. Os ajustes são feitos de forma que a resposta do vencedor neurônio à aplicação subseqüente de uma entrada similar seja aumentada.

De acordo com Kohonen (2001), o mapa SOM, apresentado na Figura 3.15, é definido a partir do mapeamento de um conjunto de dados de entrada, representados em um espaço $\Re^n$, em um arranjo bidimensional de neurônios. Para cada neurônio i do SOM é associado um vetor de referência ou peso $m_i = [\mu_{i1}, \mu_{i2}, \ldots, \mu_{in}] \in \Re^n$, onde μ_{in} representa a dimensão n do neurônio i. Inicialmente todos os componentes de m_i são inicializados aleatoriamente.

A organização topológica dos neurônio do SOM pode ser definida de diferentes formas, por exemplo, retangular, hexagonal ou irregular (Figura 3.16). Entretanto, é importante notar que, um vetor de entrada $p = [\xi_{i1}, \xi_{i2}, \ldots, \xi_{in}] \in \Re^n$ é conectado paralelamente a todos os neurônios do SOM através de m_i. De acordo com Hulle (2000) e Zuchini (2003), para cada vetor p apresentado a rede, deverá ocorrer a competição entre todos os neurônios m_i, de forma que o neurônio que possua o vetor de pesos mais próximo ao vetor de entrada, segundo alguma medida de similaridade, vencerá a competição.

Conforme apresentado, o vetor p pode ser comparado com todos os m_i a partir de qualquer métrica, contudo, sugere-se que a menor distância Euclidiana $\|p - m_i\|$, pode ser utilizada para encontrar o neurônio mais representativo, descrito como:

$$d(p, m_i) = \| p - m_i \| = \sqrt{\sum_{j=1}^{n} (p_j - m_{ij})^2}, \tag{3.22}$$

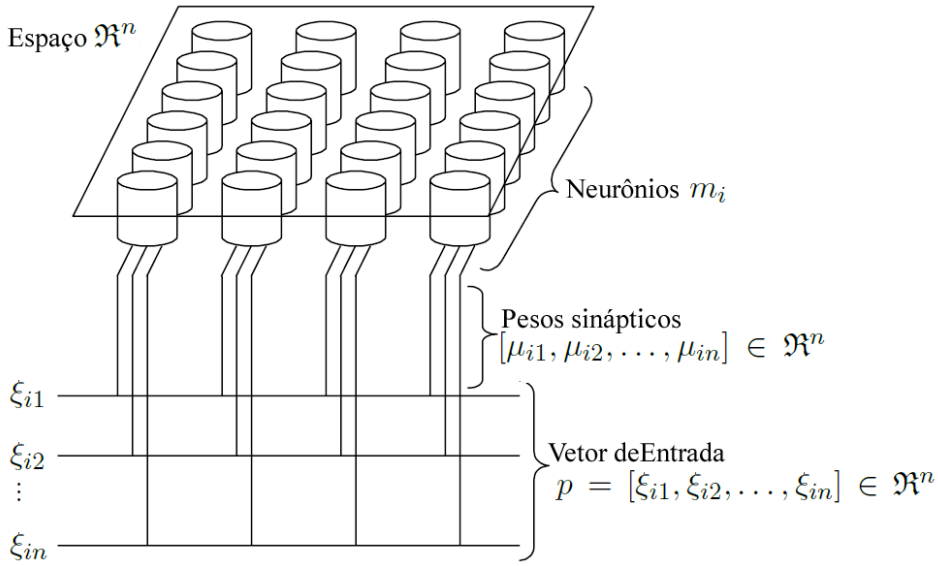
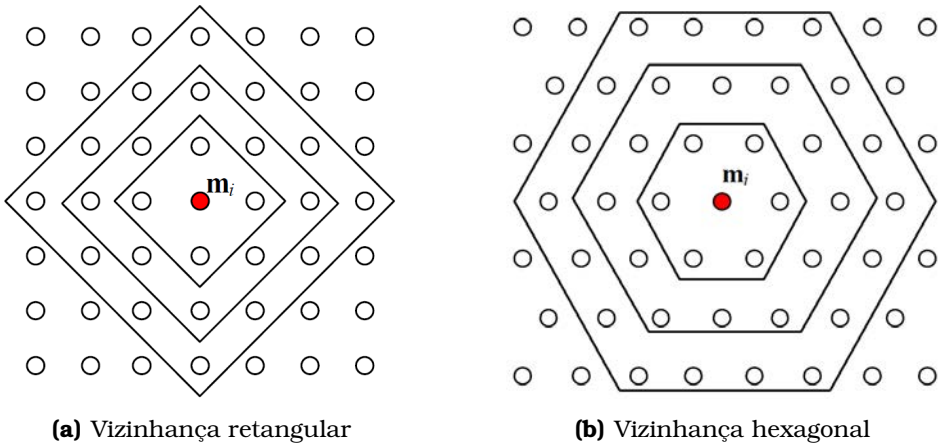


Figura 3.15: Arranjo dos neurônios do SOM e definição das variáveis. Adaptado de (Zuchini, 2003).



(a) Vizinhança retangular **(b)** Vizinhança hexagonal

Figura 3.16: Dois exemplos de topologias dos neurônios do SOM (Zuchini, 2003).

onde n é o número de dimensões consideradas. Consequentemente, a menor distância estará relacionada ao neurônio vencedor, definido como:

$$c = \arg\min_i \|p - m_i\| \tag{3.23}$$

Durante o período de aprendizagem da rede, o neurônio vencedor e os neurônios que encontram-se topograficamente próximos ao vencedor sofrerão a atualização de seus vetores de pesos. Este comportamento resulta em um relaxamento local ou efeito de suavização do vetor de pesos dos neurônio da vizinhança, que promove, em um aprendizado contínuo, a auto-organização da rede (Kohonen, 2001). A atualização dos pesos sinápticos é definido como segue:

$$m_i(t+1) = m_i(t) + \alpha_k(t)\ h_{ci}(t)\ [p(t) - m_i(t)], \tag{3.24}$$

onde $t = 0, 1, 2, ...$ é um número inteiro, representando uma coordenada discreta de tempo e $\alpha_k(t)$ define a taxa de aprendizado. Sobre o processo de relaxamento, a função $h_{ci}(t)$ exerce uma função de vizinhança, responsável pela adaptação do neurônio vencedor e de seus vizinhos de acordo com o grau de vizinhança ou proximidades. Para a convergência do processo de aprendizagem, é necessário que $h_{ci}(t) \to 0$ e $\alpha_k(t) \to 0$ quando $t \to \infty$. De acordo com Kohonen (2001), $h_{ci} = h(\| r_c - r_i \|, t)$, onde $r_c \in \Re^2$ e $r_i \in \Re^2$ representam as posições dos neurônios de índices c e i dentro do arranjo bidimensional. É importante notar que, a medida que $\| r_c - r_i \|$ aumenta, $h_{ci} \to 0$. Assim, o raio e a forma de h_{ci} definem a "elasticidade" do mapa. A função $h_{ci}(t)$ é descrita nos termos de uma função Gaussiana, como segue:

$$h_{ci}(t) = \exp\left(-\frac{\| r_c - r_i \|^2}{2\sigma^2(t)}\right), \tag{3.25}$$

onde o parâmetro $\sigma(t)$ define a largura da região de vizinhança, chamada raio de vizinhança. Normalmente $\sigma(t) \to 0$ quando $t \to \infty$ (Kohonen, 2001).

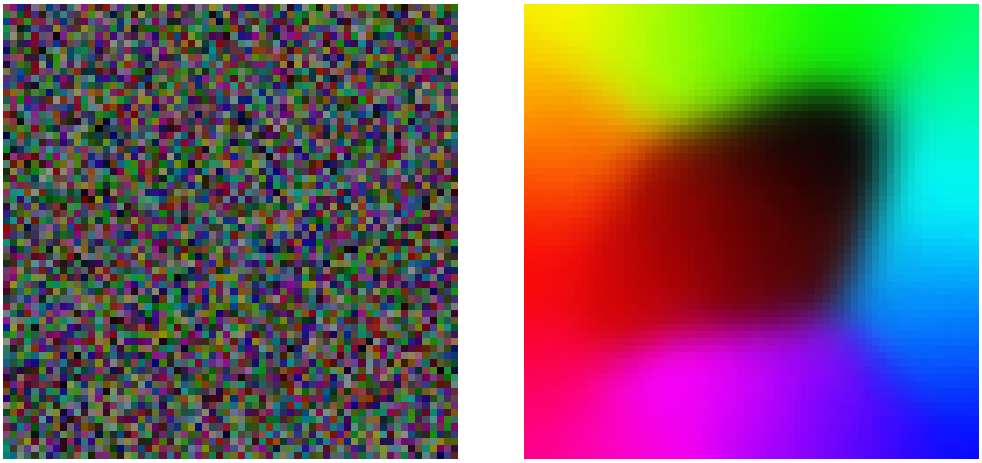
Ainda de acordo com (Kohonen, 2001), a precisão da função de tempo sobre $\alpha_k = \alpha_k(t)$ e $\sigma = \sigma(t)$, pode ser definida como:

$$\alpha_k(t+1) = \alpha_k(t)\ 0.9\ (1 - (t/n_{it})) \tag{3.26}$$

$$\sigma(t+1) = \sigma(t)\ 0.9\ (1 - (t/n_{it})) \tag{3.27}$$

onde n_{it} é número de iterações necessárias para decrementar α_k e σ até zero.

Para demonstrar o processo de treinamento da rede SOM, a Figura 3.17 apresenta uma simulação realizada, onde todos os neurônios da rede foram inicializados com pesos aleatórios (Figura 3.17 (a)). O processo de aprendizado não supervisionado é executado também a partir de padrões com pesos aleatórios. A convergência do mapa é apresentado na Figura 3.17 (b).



(a) Neurônios incializados aleatórios **(b)** Mapa SOM convergido

Figura 3.17: Exemplo de treinamento de um mapa SOM.

3.4 Considerações Finais

Neste capítulo foram apresentados os fundamentos teóricos considerados como base para o desenvolvimento dos modelos propostos para atenção visual nesta tese.

No próximo capítulo será apresentada a revisão dos principais modelos computacionais de atenção visual propostos baseados em diversas hipóteses neurobiológicas, onde serão enfatizados modelos baseados em mapas de saliência, enviesamento por mecanismos *top-down* e finalmente, porém não menos importante, baseados na teoria da correlação temporal, hipótese norteadora desta tese.

CAPÍTULO

4

Modelos Computacionais para Atenção Visual

Nos últimos anos têm sido propostos uma grande quantidade de modelos computacionais para a atenção visual. Um modelo computacional de atenção visual pode ser definido como um tipo específico de modelo de atenção visual, devendo contar com a descrição de como a atenção visual será modelada, assim como a realização de testes, de forma que os resultados gerados possam ser analisados e comparados com resultados experimentais de modelos biológicos de atenção visual (Tsotsos, 2011), ou ainda através de inspeção visual humana, conhecidos como mapas de fixação.

Considerando o grande número de modelos computacionais propostos relacionados à atenção visual, um dos objetivos deste capítulo é de apresentar trabalhos inspirados biologicamente relacionados aos modelos a serem propostos nesta tese, dentre os quais, teoria e modelo, apresentem propósitos ou hipóteses comuns. Em (Tsotsos, 2011) e (Borji and Itti, 2013) é apresentada uma revisão detalhada de trabalhos relacionados à atenção visual. Abordaremos neste capítulo a revisão de trabalhos baseados nas seguintes hipóteses:

- Mapa de Saliência: com base na teoria da integração de características de Treisman and Gelade (1980), proposto inicialmente por Koch and Ullman (1985). Modelos baseados nesta hipótese são caracterizados pela integração de estímulos locais, extraídos a partir de diferentes localizações espaciais;

- Atenção Emergente: modelos baseados nesta hipótese apresentam, comumente, um grande número de neurônios representativos, envolvidos em interações competitivas, sendo a atenção resultado do processamento neural, enviesado

por um mecanismo *top-down*;

- Correlação Temporal: baseado na teoria da correlação temporal, proposta por von der Malsburg (1981), tem como objetivo expressar a correlação entre células ou neurônios, ativados paralelamente por diferentes estímulos que representam partes de uma mesma entidade. Modelos baseados nesta hipótese, consistem tipicamente de grupos de neurônios, contendo conexões excitatórias e inibitórias, cujo dinamismo é regido por um conjunto de equações diferenciais, que compõe o sistema (Tsotsos, 2011).

Consideramos também nesta revisão, modelos de atenção relacionados à inter-relação de hipóteses, sendo referenciados de acordo com sua relevância para esta tese. O diagrama de Venn, apresentado na Figura 4.1, demonstra as possíveis combinações de hipóteses, os trabalhos citados nas seções seguintes (representados por uma estrela de cor branca) e os trabalhos propostos nesta tese (representados por uma estrela de cor vermelha).

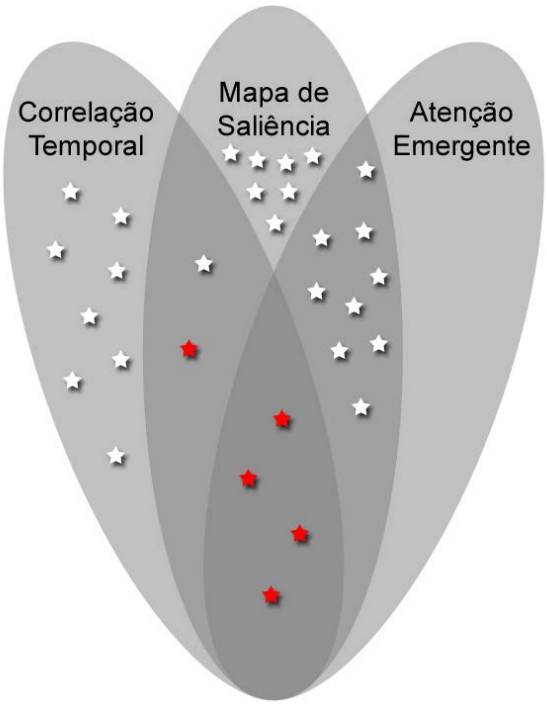


Figura 4.1: Diagrama de Venn para as três hipóteses descritas neste capítulo e suas combinações, somando um total de 6 possibilidades. Adaptado de Tsotsos (2011).

4.1 Modelos Baseados em Mapas de Saliência

De acordo com Itti and Koch (2001), a saliência é independente da natureza de um processo de busca pelo local de maior atenção, realiza-se de forma rápida e é dirigida principalmente de forma *bottom-up*. Entretanto, pode ser influenciada pelo contexto, como efeitos de plano de fundo, de modo que, se um estímulo é suficientemente saliente, se destacará automaticamente na cena (Ogawa and Komatsu,

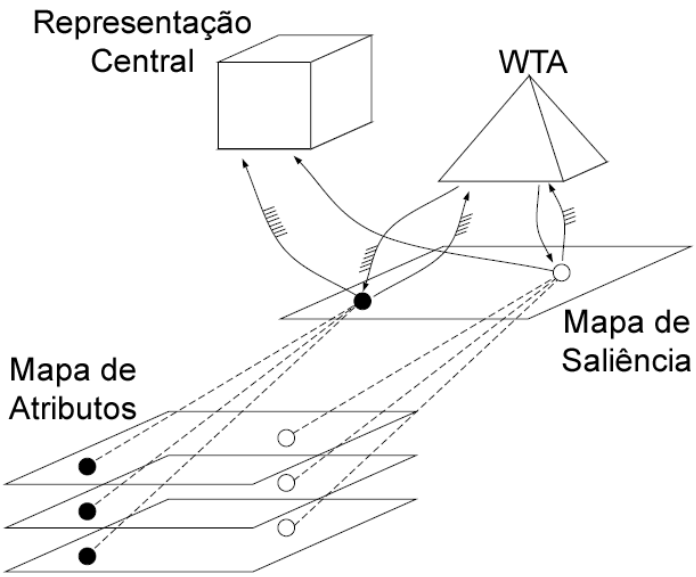


Figura 4.2: Modelo de atenção visual baseado em mapa de saliência proposto por Koch and Ullman (1985).

2004; Egeth and Yantis, 1997). Do ponto de vista neurobiológico, a saliência de todo um campo visual é alcançada de maneira pré-atentiva, tornando-a um rápido mecanismo para atenção visual.

Proposto por Koch and Ullman (1985), o principal propósito do mapa de saliência é a modelagem da atenção visual seletiva que, através de movimentos sacádicos, torna-se possível direcionar a atenção visual para regiões de maior interesse na cena. Segundo Koch and Ullman (1985), a atenção visual seletiva é composta por três etapas diferentes, apresentadas na Figura 4.2, sendo as seguintes:

- Etapa 1: onde um conjunto de características primitivas é calculada paralelamente através do campo visual, representados por um conjunto de mapas topográficos, ou seja, mapas representativos de cada característica. Assim, determinados locais do campo visual, que apresente contraste com os demais em relação a um determinado atributo, como por exemplo, cor ou orientação, são destacados no correspondente mapa. Posteriormente, esses mapas são combinados em um único mapa, denominado mapa de saliência, responsável por representar a conspicuidade de toda a cena visual;

- Etapa 2: uma rede *Winner-Take-All*[1] (WTA), é utilizada para localizar o ponto mais ativo do mapa de saliência;

- Etapa 3: propriedades da localização selecionada são encaminhadas para uma representação central. Logo em seguida, a rede WTA identifica, de forma sequencial, a próxima localização mais saliente.

[1]Termo introduzido por Feldman (1982)

Desde sua proposta, o modelo de Koch and Ullman (1985) tem sido aplicado para o desenvolvimento de trabalhos em diversas áreas do conhecimento. Itti et al. (1998), com a proposta de um modelo de atenção visual baseado diretamente na arquitetura proposta por Koch and Ullman (1985), propuseram um dos mais importantes modelos computacionais para atenção visual baseado em mapas de saliência. Composto das etapas descritas anteriormente, seu fluxo é resumido conforme o diagrama apresentado na Figura 4.3. Inicialmente, os atributos visuais da imagem de entrada passam por uma filtragem linear composta por 8 escalas espaciais, seguido pelo cálculo da diferença de centro-vizinhança, enfatizando o contraste espacial local para cada um dos atributos, processo que gera um total de 42 mapas. Estes mapas são então combinados para cada atributo, dando origem aos respectivos mapas de conspicuidades. Finalmente, os 7 mapas de conspicuidades são somados, gerando um único mapa de saliência. Uma WTA detecta e direciona a atenção para o ponto mais saliente que, logo em seguida é suprimido por um mecanismo de inibição de retorno. Consequentemente, a atenção é direcionada para o próximo ponto mais saliente. Em um trabalho posterior, Itti and Koch (2000) apresentam a implementação computacional detalhada deste modelo, tendo como foco principal o desenvolvimento de um modelo estritamente *bottom-up*, gerado a partir da integração das seguintes características primitivas: cor, intensidades e orientação. Dado a fundamental importância deste modelo para o desenvolvimento desta tese, sua descrição detalhada é apresentada na Seção 3.1.

Com a possibilidade de identificação do ponto de maior saliência em uma imagem, Walther et al. (2002) propuseram sua utilização como mecanismo prévio ao reconhecimento de objetos. Neste trabalho, dado o ponto destacado no mapa de saliência pela WTA, é realizado o retorno imediato à dimensão correta para identificar o mapa de característica que mais contribua para a saliência do local destacado, segmentado em seguida por um algoritmo de inundação (*flooding algorithm*). O mapa de características segmentado é então utilizado como *template* para o mecanismo de inibição do retorno do mapa de saliência. Os autores mencionam o termo inibição do retorno baseada em objetos, entretanto não se é possível afirmar que o segmento gerado a partir de um mapa de conspicuidade seja realmente um objeto. O segmento também é processado em uma máscara binária e convolvido através de um filtro gaussiano para a resolução da imagem original. Finalmente, esta região é submetida ao sistema de reconhecimento proposto por Riesenhuber and Poggio (1999). O modelo foi aplicado em cenas sintéticas, representando clipes de papel amassados, e também em cenas naturais. Entretanto, resultados satisfatórios foram obtidos somente a partir de cenas sintéticas. O diagrama do modelo é apresentado na Figura 4.4.

Em Rutishauser et al. (2004), considerado como uma extensão do trabalho de Walther et al. (2002), todo o mecanismo de saliência é mantido conforme trabalho

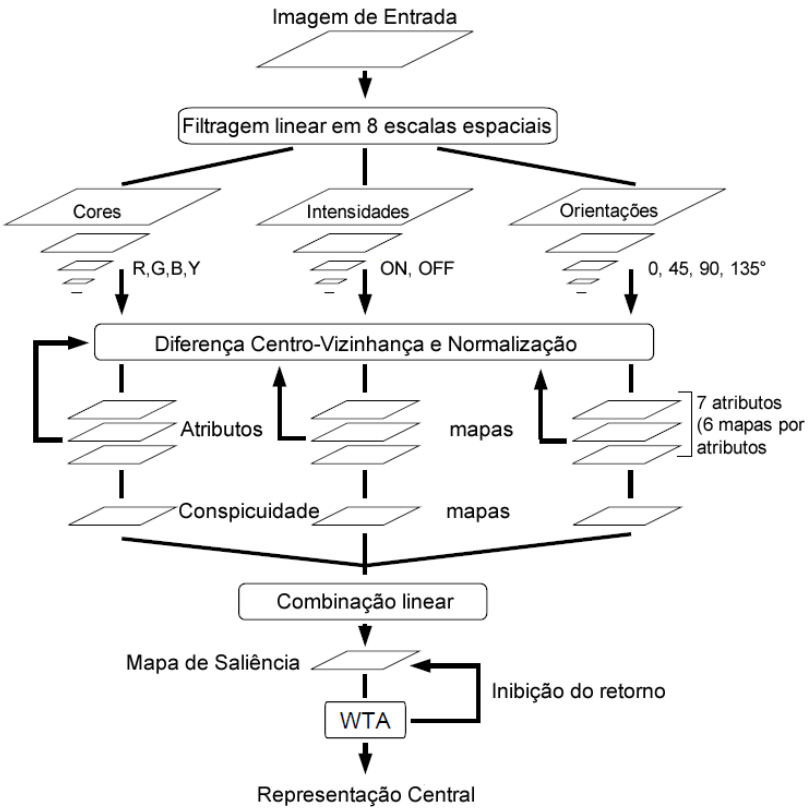


Figura 4.3: Modelo de atenção visual baseado em mapa de saliência proposto por Itti et al. (1998).

anterior. Entretanto, para sistema de reconhecimento foi utilizado o algoritmo de reconhecimento de objeto proposto por Lowe (1999). De acordo com Rutishauser et al. (2004), este algoritmo utiliza uma pirâmide gaussiana gerada a partir da imagem original para a extração de características locais ("pontos-chaves"), calculados através da diferença entre os níveis da pirâmide gaussiana (para maiores detalhes veja Lowe (1999)). O reconhecimento é então realizado por meio da combinação entre os pontos dos modelos de objetos treinados previamente (aprendizado supervisionado) e a região saliente destacada pela máscara gaussiana. Em um trabalho seguinte, Walther et al. (2005) estendem este modelo mantendo similarmente as técnicas de saliência e reconhecimento, porém introduzem um mecanismo de aprendizado não-supervisionado, baseado no mecanismo utilizado para o reconhecimento. Em (Walther and Koch, 2006) é abordado o conceito de proto-objetos, definido como uma unidade de informação visual que pode ser acessada pela atenção seletiva, podendo ser posteriormente validado como um objeto real. Neste trabalho foram mantidas as principais características dos modelos anteriores, porém com maior ênfase à plausibilidade biológica do modelo. De maneira geral, a partir da propagação da atenção encontrada no mapa de saliência, é gerado um mapa binário (neste caso não se utiliza segmentação), utilizado como uma máscara para a obtenção da região do proto-objeto, servindo também ao mecanismo responsável pela inibição do retorno.

Em (Harel et al., 2006), foi introduzido um modelo de saliência baseado em

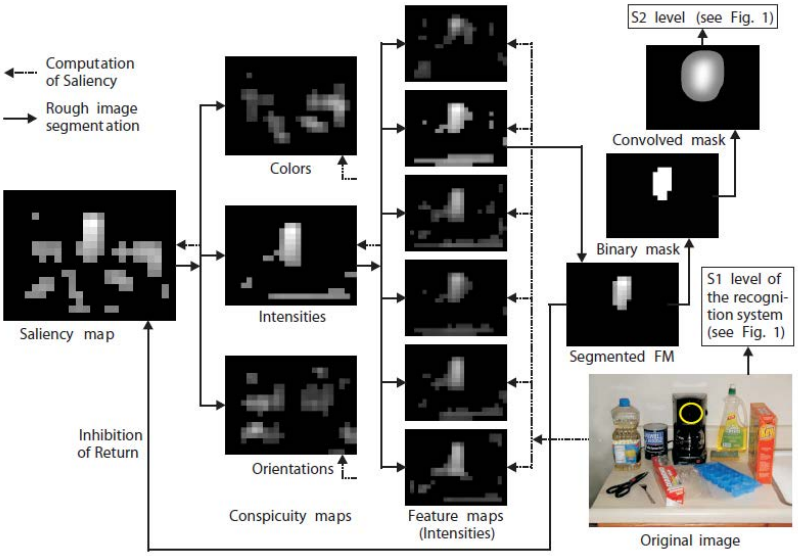


Figura 4.4: Modelo de atenção visual baseado em mapa de saliência para o reconhecimento de objeto proposto por Walther et al. (2002).

grafos (GBVS - *Graph-Based Visual Saliency*). No GBVS, os mapas de características (intensidades, cores e orientações) são extraídos de maneira similar à apresentada em (Itti et al., 1998), onde, para cada mapa de característica, um grafo totalmente conectado é utilizado, de modo que nodos são responsáveis pela representação de todas as localizações. Pesos entre dois nodos são associados proporcionalmente à similaridade em relação às suas características e pesadas por suas distâncias espaciais. Em um processo de normalização, nodos que apresentem altos valores de dissimilaridades em relação à vizinhança são definidos com altos valores de saliência. Os mapas são finalmente normalizados para enfatizar detalhes conspicuosos, e combinados em um único mapa de saliência

Embora umas das principais motivações para trabalhos baseados em mapas de saliência seja, analogicamente, resolver o problema do "ovo e da galinha", é importante notar que, para o direcionamento da atenção visual a objetos sem previamente conhecê-los, caso os objetos da cena não apresentem características salientes, ou ainda, caso as informações de fundo apresentem maior saliência do que os objetos existentes de fato, a acurácia destes modelos poderá ser afetada.

Diversos outros modelos e aplicações baseados em mapas de saliência têm sido propostos, entretanto pesquisadores têm direcionado seus trabalhos para a questão sobre "*o que*" se procura, uma vez que os mapas de saliência se encarregam de resolver o problema sobre a questão de "*onde*" deve ser iniciada a busca (para outros trabalhos baseados em mapas de saliência veja (Borji and Itti, 2013)).

Podemos concluir que modelos de atenção visual baseados no conceito de mapa de saliência possuem, como principal motivação, o aumento da eficiência do sistema, permitindo que recursos computacionais disponíveis sejam utilizados para processar apenas regiões de interesse na cena (Carota et al., 2004; Walther et al.,

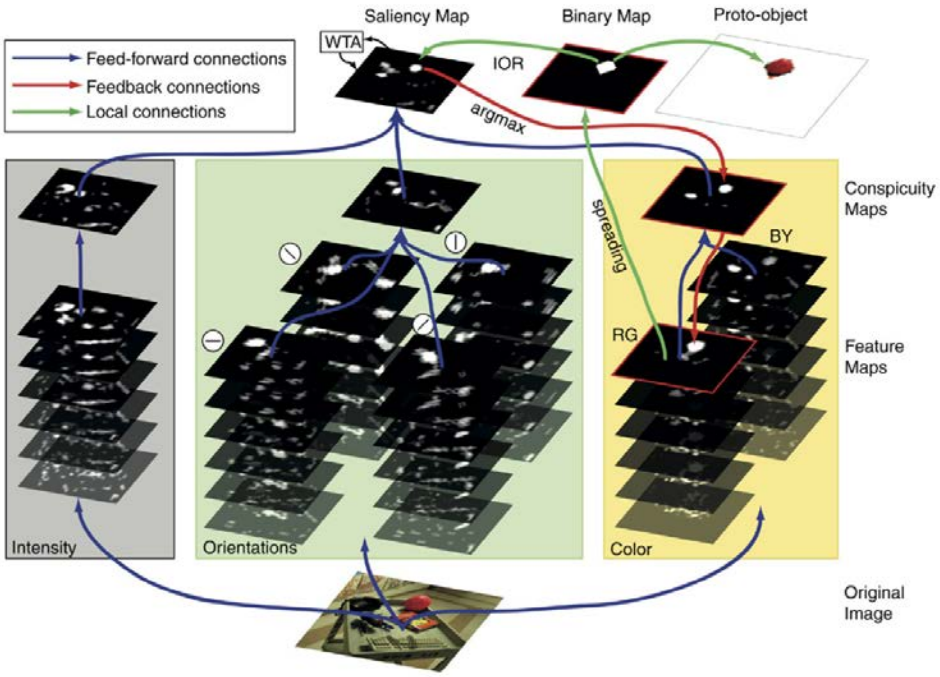


Figura 4.5: Modelo de atenção visual baseado em proto-objetos proposto por Walther and Koch (2006).

2005; Bonaiuto and Itti, 2006; Siagian and Itti, 2007).

4.2 Modelos com Enviesamento *Top-down*

De acordo com Elazary and Itti (2008) e Bruce and Tsotsos (2009), a seleção de objetos em uma cena é altamente influenciada por processos *bottom-up*, onde apenas as regiões mais salientes da imagem são analisadas, descartando-se as demais. De acordo com os trabalhos apresentados na seção anterior, modelos de atenção visual baseados em mapas de saliência são basicamente dirigidos por características primitivas da imagem, desconsiderando qualquer informação prévia a cerca dos objetos. Entretanto, conforme conceitos neurobiológicos apresentados no Capítulo 2, a atenção visual pode ser fortemente influenciada por informações *top-down*, de forma que, de acordo com Desimone and Duncan (1995), a atenção possa ser direcionada para informações de maior importância para o observador, ou seja, voluntariamente. Para Borji and Itti (2013), ambos os mecanismos não podem ser considerados mutuamente exclusivos, uma vez que a atenção visual pode ser direcionada individualmente por cada dos mecanismos ou através da combinação das atenções *bottom-up* e *top-down*. Dado a importância do assunto para o desenvolvimento desta tese, esta seção tem como objetivo apresentar modelos de atenção visual que utilizem algum tipo de enviesamento *top-down*, os quais foram utilizados como referência para o desenvolvimento dos alguns modelos aqui propostos.

Destacamos inicialmente o trabalho proposto por Clark and Ferrier (1988, 1989) que, baseado no modelo *bottom-up* de Koch and Ullman (1985), propõem a

implementação do enviesamento *top-down* do sistema de controle dos movimentos de um robô móvel (Figura 4.6). O sistema foi composto por dois estágios, sendo o primeiro responsável pela captura da imagem $I(x, y, t)$ através de câmeras, onde características visuais primitivas são extraídas de forma paralela. O resultado deste estágio é um conjunto de mapas de características $Y_i(x, y, t)$, que indica a presença ou ausência de uma característica i em cada localização x, y da imagem I no tempo t. Neste trabalho, um mapa de característica pode indicar, por exemplo, a presença de uma cor ou intensidade específica, ou ainda de uma linha de orientação. O segundo estágio do modelo possui como função realizar a combinação dos mapas de características, aplicando um determinado peso $K_i(t)$ a cada mapa Y_i, somando-os em seguida para gerar o mapa de saliência $S(x, y, t)$. Assim, o local do mapa com maior valor de saliência será o novo destino do robô móvel. A possibilidade de atribuição de pesos específicos a cada um dos atributos considerados foi uma característica interessante neste trabalho, despertando a possibilidade de modulação da atenção para características primitivas de maior interesse na cena, ou seja, o enviesamento *top-down*.

Também baseado na arquitetura proposta por Koch and Ullman (1985), um popular modelo de atenção visual foi proposto por Wolfe (1994), no qual, além da características *bottom-up*, comumente encontradas em modelos baseados em mapa de saliência, também utiliza como mecanismo para guiar a atenção, o enviesamento *top-down*. Neste trabalho, o enviesamento *top-down* é realizado através da atribuição de pesos associados aos mapas de características. Partindo da premissa que se conheça características do alvo a ser encontrado, por exemplo, uma barra vermelha meio a barras azuis, o mapa de característica sensível a cor vermelha receberá o maior peso, o que tornará a barra vermelha mais saliente. Entretanto, não é definido como os pesos são escolhidos no modelo, presumimos que ocorra empiricamente de acordo com características do alvo desejado. A arquitetura deste modelo é apresentada na Figura 4.7.

Em um modelo proposto posteriormente por Navalpakkam and Itti (2005), quatro aspectos importantes em relação à visão biológica foram considerados, sendo estes: a determinação da importância da busca por uma determinada entidade, o enviesamento da atenção a partir de características visuais de baixo nível do alvo desejado, o reconhecimento destes alvos a partir das mesmas características de baixo nível e, incrementalmente, a construção de um mapa de saliência visual que apresente a relevância de busca por toda a cena. De acordo com a arquitetura deste modelo, apresentada na Figura 4.8, assim como seus aspectos biológicos, podem ser observados propósitos comuns relacionados ao modelo proposto por Wolfe (1994). Entretanto, os princípios relacionados ao enviesamento *top-down* são bastante distintos. De acordo com os autores, a principal diferença está na forma com que o conhecimento *top-down* do alvo é representado. Neste caso, durante uma fase

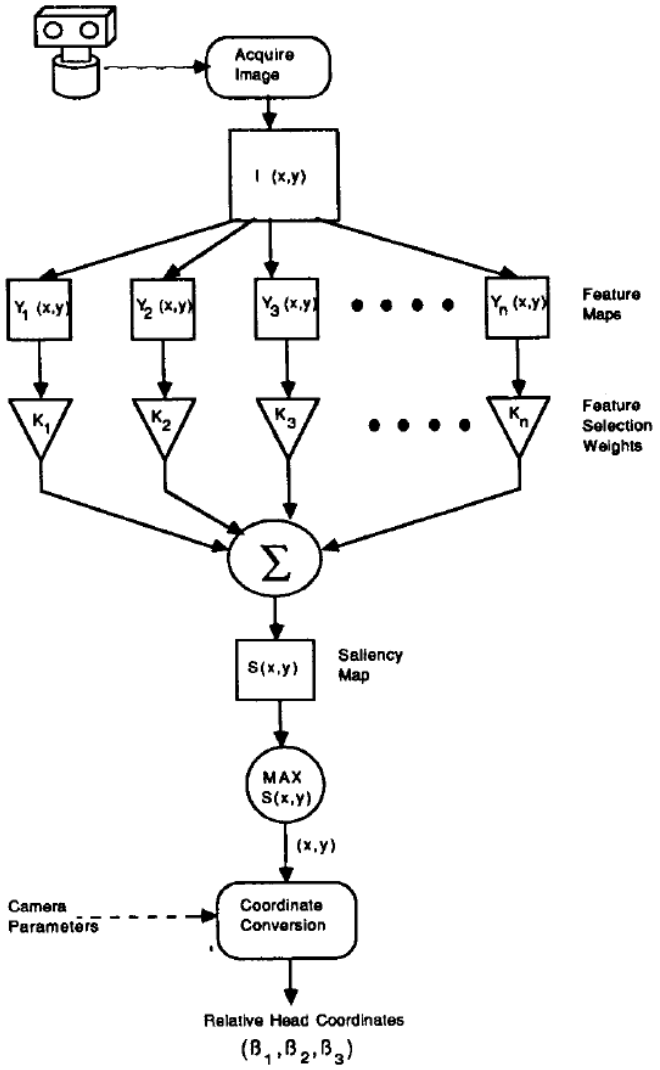


Figura 4.6: Modelo de seleção de atenção proposto por Clark and Ferrier (1989).

prévia de aprendizado, um vetor de pesos para cada característica é treinado a partir de imagens contendo o alvo, o qual será utilizado tanto para o enviesamento *top-down*, durante a geração do mapa de saliência, quanto para o reconhecimento do objeto. Os autores mencionam as limitações do modelo em reconhecer objetos complexos, entretanto sugerem sua utilização como um filtro prévio à sistemas de reconhecimento mais complexos. Em (Bonaiuto and Itti, 2006) este mecanismo foi utilizado para o direcionamento da atenção visual em um ambiente dinâmico, com o objetivo de detectar e reconhecer de faces previamente armazenadas em uma base de dados, através de sua comparação com informações extraídas da região saliente.

Frintrop (2006) apresenta o modelo denominado VOCUS (*Visual Object detection with a CompUtational attention System*), baseado nos princípios do mapa de saliência de Koch and Ullman (1985) e na implementação apresentada por Itti et al. (1998) e Itti and Koch (2000). De modo geral, o primeiro módulo (*bottom-up*) deste sistema se assemelha aos trabalhos apresentados por Wolfe (1994) e Navalpakkam

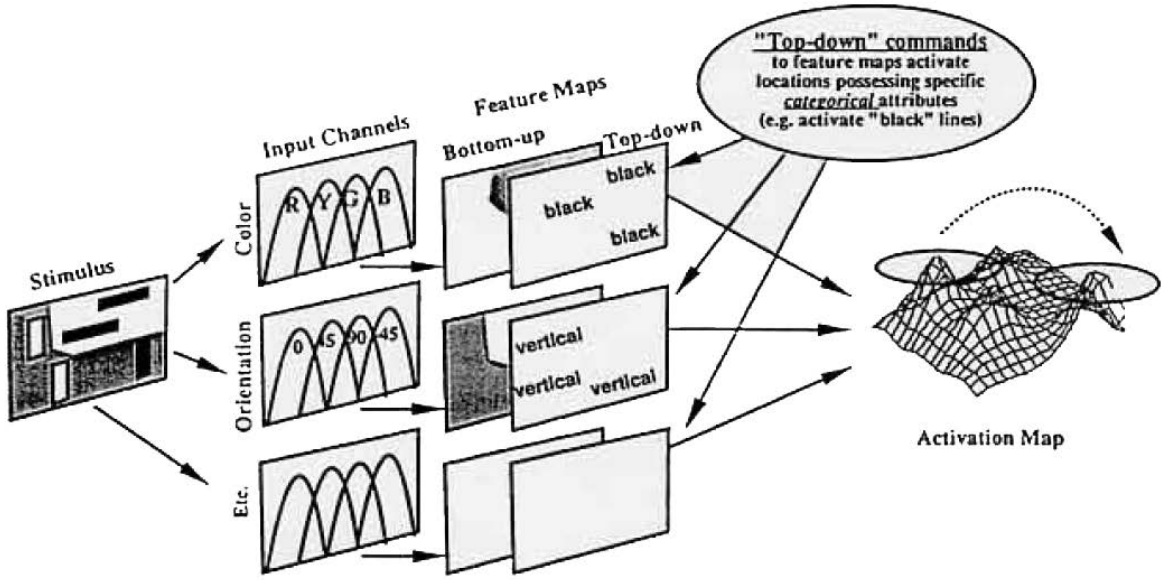


Figura 4.7: Arquitetura do modelo de atenção visual proposto por Wolfe (1994).

and Itti (2005), uma vez que apresentam as mesmas referências teóricas. Entretanto, como estamos enfatizando modelos emergentes baseados em enviesamentos *top-down*, acreditamos que contribuições neste sentido sejam importantes para a revisão teórica apresentada nesta tese. Neste modelo, a partir da identificação da região de interesse (ROI - *Region of Interest*) realizada manualmente pelo usuário, destacada na Figura 4.9 por um retângulo amarelo, o modelo VOCUS, em uma fase de aprendizagem, é encarregado de identificar a região mais saliente (MSR - *Most Salient Region*) da ROI, baseado em um mapa de saliência *bottom-up*. Nesta etapa, o sistema deverá aprender quais características primitivas melhor representam o alvo contido na ROI, e um vetor de pesos é aprendido para cada uma das características. Este processo é repetido de acordo com a quantidade de alvos que se deseja treinar. O vetor de pesos será posteriormente utilizado durante uma fase de busca para o enviesamento *top-down* do modelo, de acordo com o alvo desejado. Diferente do modelo proposto por Wolfe (1994), o vetor de pesos associados às características consideradas no VOCUS é gerado de forma automática, com base nas características encontradas na MSR, destacada em vermelho (Figura 4.9). Outra diferença que pode ser destacada refere-se ao módulo de busca. De forma resumida, os autores consideram neste modelo, assim como os demais aqui apresentados, um único mapa saliência para o desenvolvimento da atenção visual, entretanto, neste modelo é gerado a partir da composição de dois outros mapas de saliência, o primeiro, um mapa de saliência *bottom-up*, amplamente utilizado nos modelos apresentados neste capítulo, e o segundo, um mapa de saliência *top-down*, enviesado conforme o vetor de pesos aprendido. Pode-se concluir que com a composição dos mapas *bottom-up* e *top-down* torna-se possível enfatizar a saliência do objeto alvo mantendo a saliência de outras regiões da cena.

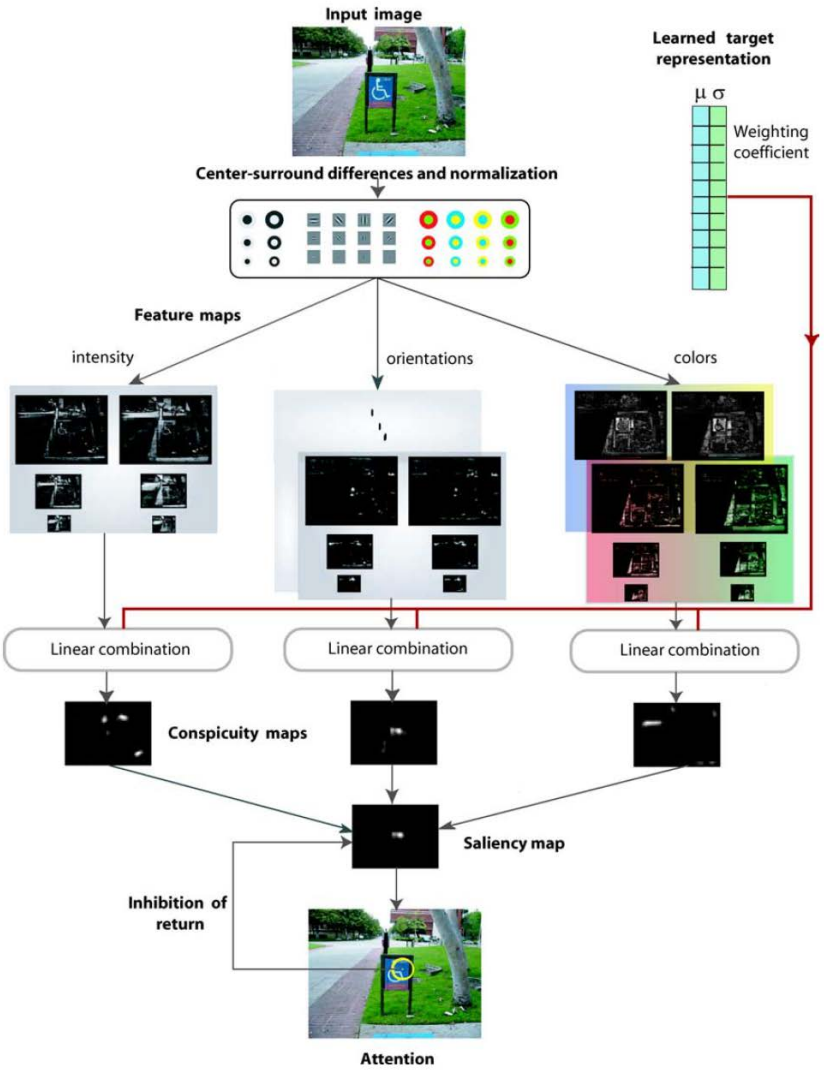


Figura 4.8: Arquitetura do modelo de atenção visual proposto por Navalpakkam and Itti (2005).

Como uma extensão do trabalho anterior apresentado em (Navalpakkam and Itti, 2005), Navalpakkam and Itti (2006a) propõem um modelo de atenção visual onde o enviesamento *top-down* também é gerado a partir de características visuais de baixo nível, entretanto, com uma significante modificação, tanto alvo quanto distratores são considerados. De acordo com Braithwaite and Humphreys (2003), considerações sobre características locais de fundo também podem facilitar o processo de busca. Assim, com o objetivo de aumentar o desempenho do modelo em relação ao tempo necessário para a detecção do alvo, um vetor de pesos é aprendido para cada uma das características a partir de imagens com a presença e a ausência do alvo, ou seja, a partir de alvos e distratores. Conforme apresentado na Figura 4.10, os mapas de saliência *bottom-up* $s_{ij}(A)$ são gerados para cada i-ésima característica pertencente à j-ésima dimensão (por exemplo, a característica RG do canal de cores) de uma imagem A, onde $i \in \{1 \ldots n\}$ e $j \in \{1 \ldots N\}$. O conhecimento prévio em relação ao alvo e distratores é usado para calcular os pesos *top-down* g_{ij} e g_j. Os mapas

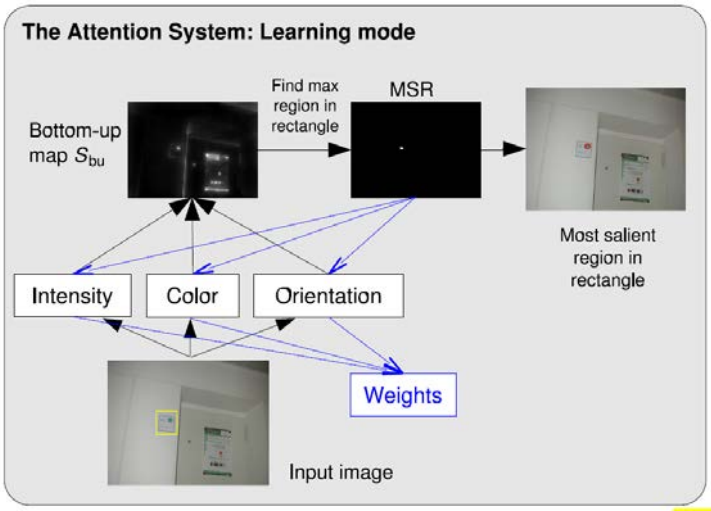


Figura 4.9: Arquitetura do módulo de aprendizado do modelo de atenção visual VOCUS, proposto por Frintrop (2006).

bottom-up $s_{ij}(A)$ são então multiplicados pelos respectivos enviesamentos *top-down* g_{ij} e, em seguida, somados para a geração o mapa de saliência $S_j(A)$ referente à j-ésima dimensão. Finalmente, os mapas de saliência de cada dimensão $S_j(A)$ são novamente enviesados pelos pesos *top-down* e somados para a geração do mapa de saliência global $S(A)$. O modelo apresentou *pop-outs* em ambientes com distratores homogêneos e, meio a distratores heterogêneos, também foi possível identificar o alvo, em ambientes com pouco contraste ou pesquisas conjuntivas, o modelo não apresentou resultados satisfatórios. Este trabalho foi estendido em (Navalpakkam and Itti, 2006b), onde os mapas de características foram divididos em sub-bandas com granularidades mais finas, aumentando sua capacidade representativa.

Nos modelos apresentados até este momento, a escolha do mapa de características correto a ser utilizado na pesquisa visual, assim como a decisão de como aplicar o enviesamento *top-down*, são características específicas de cada modelo, tendo como objetivo comum alcançar o melhor desempenho possível durante o processo de atenção visual. Entretanto, de acordo com Elazary and Itti (2010), a associação de pesos à características específicas nem sempre podem acelerar a busca por um determinado alvo, principalmente em mapas que salientam características, ou sub-bandas granulares de características, compartilhadas por ambos, alvos e distratores. Como tentativa de solução, Elazary and Itti (2010) propuseram a utilização de um conjunto de características representativas para o enviesamento *top-down* durante as fases de geração do mapa de saliência e treinamento do classificador, ambas baseadas em funções de densidade de probabilidade. De forma resumida, em um estágio prévio, funções de densidade de probabilidade são aprendidas para cada mapa característica, de acordo com as características visuais salientes do objeto, de modo que o valor máximo de cada mapa de característica seja utilizado para o treinamento do objeto selecionado. Adicionalmente, as mesmas informações

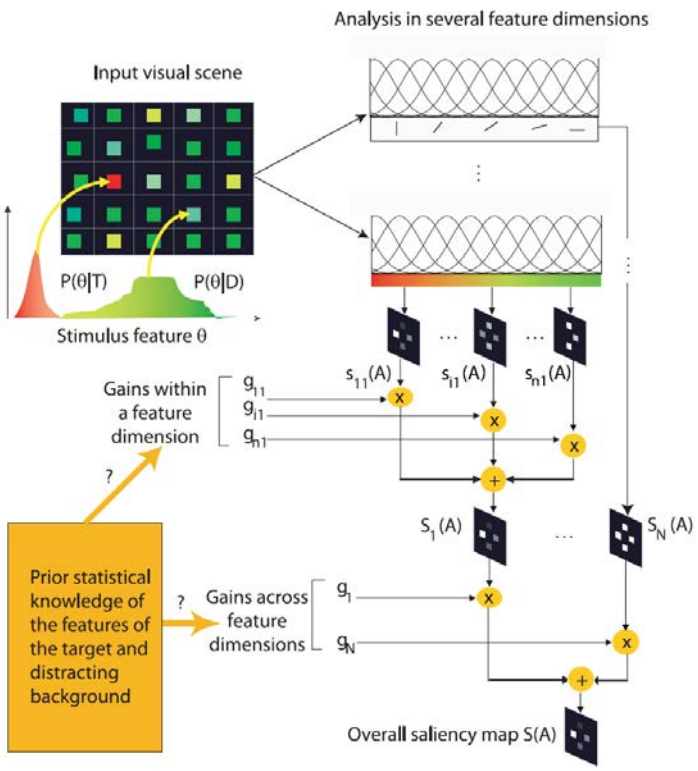


Figura 4.10: Arquitetura do modelo de atenção visual proposto por Navalpakkam and Itti (2006a).

aprendidas durante a fase de treinamento serão utilizadas para guiar a atenção. Os autores propõem a utilização do modelo proposto de duas maneiras: na primeira, a partir do local destacado no mapa de saliência, baseado na comparação das funções de densidade de probabilidade deste local em relação aos objetos previamente conhecidos, o modelo retorna uma lista ordenada de objetos que, provavelmente, possam ser encontrados neste local; na segunda forma, dado um determinado alvo, o modelo retorna uma lista de possíveis locais do mapa de saliência onde, provavelmente, possa ser encontrado o objeto procurado. Os autores sugerem que, a partir destes resultados, classificadores mais robustos sejam utilizados para a busca e reconhecimento de objetos.

Em (Borji et al., 2011) é apresentado um modelo de atenção com o objetivo de proporcionar o aumento na taxa de detecção de alvos. Em contraste com o modelos apresentados, que utilizam como enviesamento *top-down* informações do mapa de característica que apresente maior representatividade do alvo, neste trabalho os autores também incorporam o custo de processamento para cada mapa de características, de forma a forçar o processo de otimização para a definição de um vetor de pesos com maior taxa de detecção por um menor custo. É importante observar que o modelo de saliência proposto por Itti et al. (1998) e Itti and Koch (2000) não permite a seleção de uma determinada escala da pirâmide gaussiana, pois estas são utilizadas para o cálculo da diferença centro-vizinhança (mencionada inicialmente

na Seção 4.1 deste capítulo e apresentada em detalhes na 3.1). Assim, os autores revisaram o modelo de saliência básico utilizado (Itti et al., 1998; Itti and Koch, 2000), objetivando a seleção de um mapa de característica, em uma específica escala, que mais contribua para a saliência de um local desejado. O modelo mostrou-se apto não somente a atribuir peso a uma escala específica, mas também em inibir demais escalas não importantes para a detecção do alvo, obtendo portanto, o aumento na taxa de detecção e redução de custo computacional.

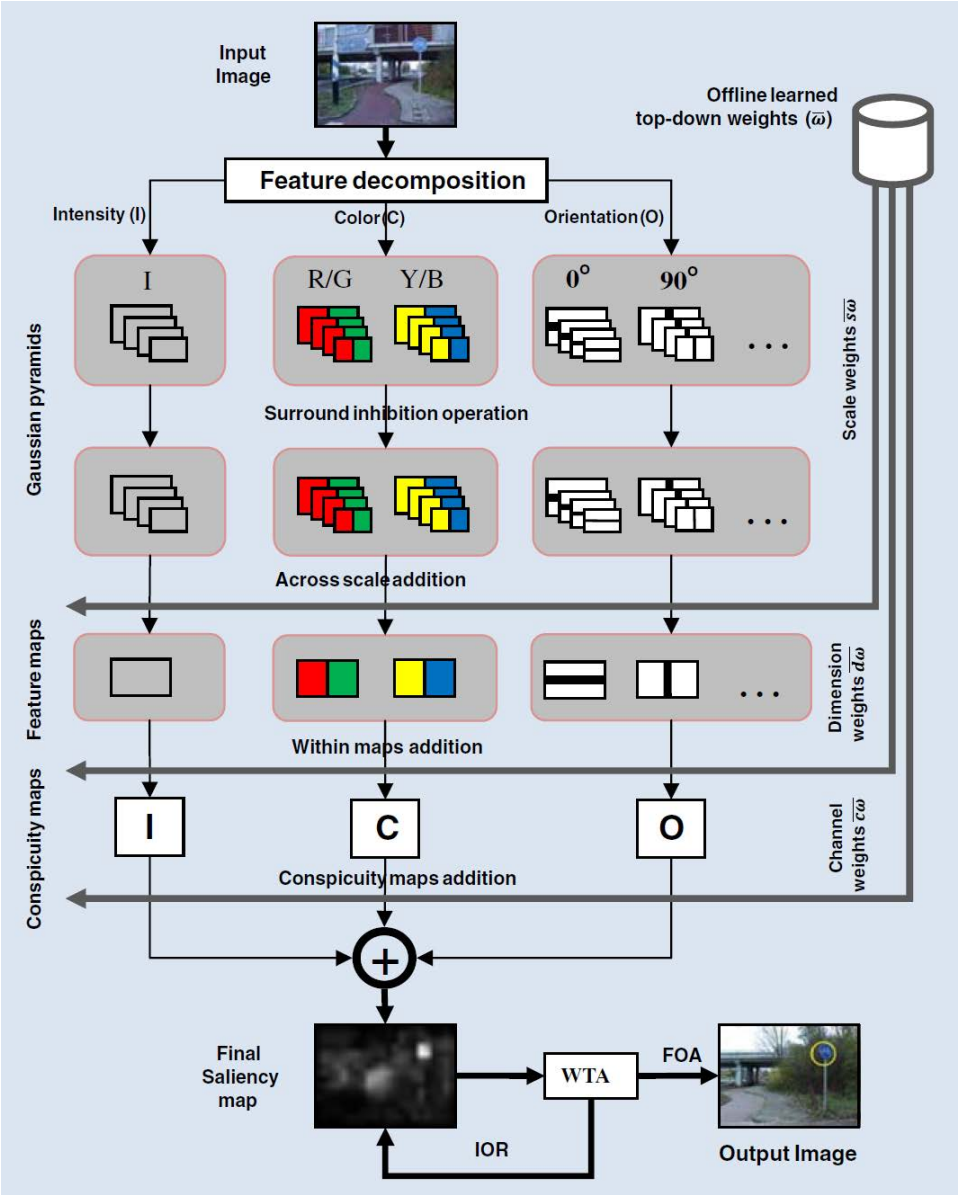


Figura 4.11: Arquitetura do modelo de atenção visual proposto por Borji et al. (2011).

Como pode ser visto em Borji and Itti (2013), diversos modelos e aplicações baseados em mapas de saliência têm sido propostos recentemente, nos quais o mapa de saliência é o principal componente para o desenvolvimento da atenção visual. Entretanto, a maioria dos trabalhos apresentados por Borji and Itti (2013) e outros aqui apresentados (Clark and Ferrier, 1989; Wolfe, 1994; Itti et al., 1998; Itti and

Koch, 2000; Navalpakkam and Itti, 2005; Bonaiuto and Itti, 2006; Frintrop, 2006; Navalpakkam and Itti, 2006a,b; Elazary and Itti, 2010; Borji et al., 2011), desenvolvem a seleção visual a partir de características primitivas da cena (*space-based*). Embora alguns trabalhos apresentem seleção baseadas em objetos (*object-based*) como, por exemplo, em (Walther et al., 2002; Rutishauser et al., 2004; Walther et al., 2005; Walther and Koch, 2006), o mecanismo também se baseia em um mapa de saliência, o que inviabiliza, por exemplo, a seleção de objetos não contrastantes, ou ainda, a seleção de falsos objetos, devido à características semelhantes ao alvo. Contudo, a seleção baseada em objetos é considerada um importante fator para o desenvolvimento da atenção visual (Desimone and Duncan, 1995; Roelfsema et al., 1998; O'Craven et al., 1999; Wang, 2005), que será abordada na próxima seção sob a hipótese da correlação temporal.

4.3 Modelos Baseados na Correlação Temporal

Nas Seções 4.1 e 4.2 foram apresentados dois mecanismos para atenção visual, o primeiro puramente *bottom-up*, baseado na extração de características primitivas da imagem, e o segundo com enviesamento *top-down*, associado à informações prévias sobre o alvo. Entretanto, evidências comportamentais e neurofisiológicas têm demonstrado que um terceiro mecanismo de seleção, baseado em objetos[2], pode ser empregado no sistema visual dos primatas (O'Craven et al., 1999; Roelfsema et al., 1998; Desimone and Duncan, 1995). A atenção baseada em objetos propõe que um mecanismo pré-atentivo realize, inicialmente, a segmentação do campo visual em um conjunto de objetos, grupos, ou superfícies, que possam servir de alvo para a atenção visual (O'Craven et al., 1999), de forma que a atenção seja considerada como uma propriedade emergente da correlação de vários mecanismos neurais, trabalhando paralelamente para a entrega da atenção (Desimone and Duncan, 1995).

O modelo de Itti et al. (1998), apresentado inicialmente na Seção 4.1, assim como os demais modelos de atenção visual *bottom-up* e com enviesamento *top-down* apresentados, utilizam mecanismos de inibição de retorno para a identificação do próximo ponto de maior saliência na cena, de forma que, quando um neurônio é selecionado como vencedor, este ponto, ou região ao redor, é considerado como o local de maior atenção (Koch and Ullman, 1985; Itti et al., 1998; Itti and Koch, 2001). Entretanto, de acordo com Wang (2005), estudos demonstram que uma característica fundamental da atenção visual está na habilidade de integrar diversas características de uma imagem, correlacionadas aos objetos coerentemente. Sendo assim, para realizar a seleção a nível de objetos deve-se, inicialmente, considerar a questão de como agrupar características ou atributos presentes em uma cena.

Desimone and Duncan (1995) e Wang (2005) definem esta questão diretamente relacionada ao problema da integração, abordado inicialmente na Seção 2.2.

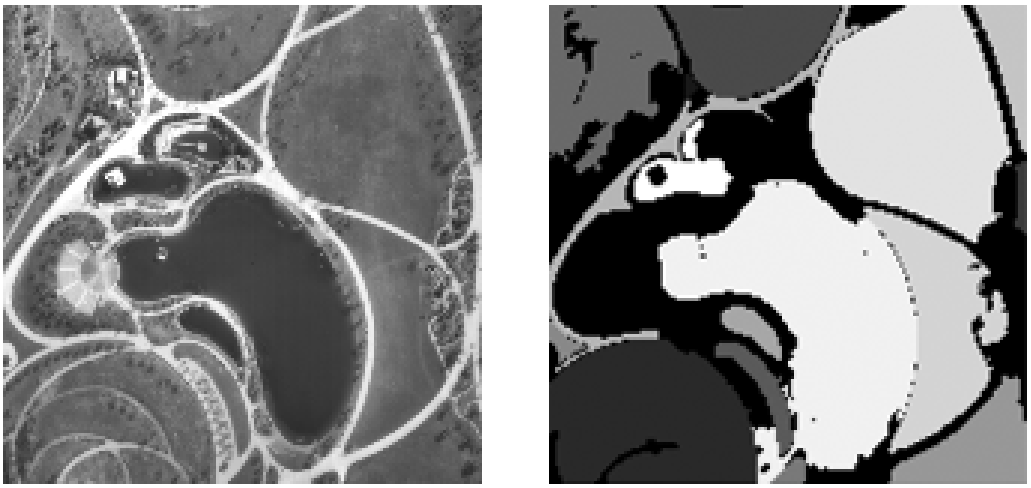
[2]do inglês *Object-Based Attention*

Este assunto vem sendo pesquisado na psicologia sob o título de organização perceptual ou agrupamento perceptual que, por sua vez, está diretamente relacionado à teoria da correlação temporal, proposta por von der Malsburg (1981) que, segundo Wang (2005), pode ser considerada uma interessante alternativa para o problema da integração. De maneira geral, a teoria da correlação temporal propõe que objetos sejam representados pela correlação temporal obtida através da sincronização e dessincronização das atividades de neurônios espacialmente distribuídos, codificando diferentes características de um objeto.

Segundo Izhikevich (2006), o sincronismo não deve ocorrer por simples casualidade, sendo assim, quando ocorre, mesmo que transitório em um subconjunto pequeno da rede, significa algo importante, algo significante. O uso da sincronização na modelagem de processos cognitivos, como a atenção, tem recebido um grande suporte biológico (Sejnowski and Paulsen, 2006; Jermakowicz and Casagrande, 2007). Descobertas neurobiológicas têm demonstrado que a atenção está ligada a sincronização entre neurônios e, através de experimentos, tem sido demonstrado que a atenção visual aumenta a coerência entre neurônios representando um mesmo estímulo, o que sugere que a sincronização é um importante mecanismo na seleção (Niebur and Koch, 1994; Fries et al., 2001; Buia and Tiesinga, 2006; Kim et al., 2007). Ainda, de acordo com Singer and Gray (1995), a teoria da correlação temporal demonstra a possibilidade do agrupamento e sincronização de característica representativas de objetos específicos, possibilitando a futura competição pela atenção auditiva (von der Malsburg and Schneider, 1986), ou pela atenção visual, que é foco de investigação desta tese.

Baseado tanto na teoria da correlação temporal, quanto em evidências neurobiológicas de que a sincronização é um processo fundamental do cérebro, Terman and Wang (1995) desenvolveram a correlação oscilatória (Campbell and Wang (1996); Wang and Terman (1997); Wang and Brown (1999)). Neste caso, características são representadas por osciladores e a integração é realizada através da sincronia destes (Wang, 2005). Desta forma, cada oscilador pode representar um conjunto de características (atributos) de tal forma que cada segmento (objeto) seja representado por um conjunto de osciladores com atividades síncronas, enquanto segmentos distintos são representados por grupos de osciladores fora de sincronia. De maneira geral, a correlação oscilatória pode ser descrita pela seguinte regra: neurônios que representam diferentes características de um mesmo objeto são sincronizados, enquanto que as atividades de grupos de neurônios representando diferentes objetos são dessincronizadas.

Como consequência do desenvolvimento da correlação oscilatória, Wang and Terman (1995) propuseram a rede LEGION (*Locally Excitatory Globally Inhibitory Oscillator Networks*), uma rede de osciladores localmente excitatórios e globalmente inibitórios, estabelecendo uma relação direta com a teoria proposta, além de pro-



(a) Imagem Original **(b)** Imagem Segmentada

Figura 4.12: "Mapa Cinza". Resultado da segmentação gerada pelo modelo de Wang and Terman (1997).

porcionarem um importante *framework* computacional para diversas áreas de aplicação, dentre estas, a segmentação de imagens e seleção de objetos, considerado um dos principais elementos pré-atentivos para o desenvolvimento de alguns modelos propostos nesta tese. A descrição detalhada do modelo LEGION encontra-se apresentada na Seção 3.2.3.

Diversos outros modelos baseados na correlação oscilatória têm sido propostos. Em Wang and Terman (1997), por exemplo, foi realizado o primeiro estudo sobre segmentação de imagens reais e seleção de objetos baseado na rede LEGION. Como o modelo proposto trata da segmentação de imagens reais, pode ser considerado como uma evolução do modelo original (Wang and Terman, 1995), que trabalha somente com imagens sintéticas. A Figura 4.12 apresenta o resultado do processo de segmentação proposto por Wang and Terman (1997), denominado por "mapa cinza", onde são apresentados todos os segmentos gerados em tons de cinza.

Em Campbell et al. (1999), o modelo LEGION é novamente modificado, substituindo-se os osciladores originais por neurônios do tipo Integra e Dispara (I&F), como objetivo de demonstrar o desempenho do modelo em relação à velocidade para a sincronização dos osciladores representando objetos específicos e dessincronização entre outros grupos de osciladores. O processo de sincronização e dessincronização de rede de neurônios do tipo I&F é abordada em detalhes na Seção 3.2.2, devido sua importância para processos de segmentação e competição por atenção desenvolvidos nesta tese.

A teoria da correlação temporal (von der Malsburg, 1981), assim como a correlação oscilatório (Terman and Wang, 1995), podem e estão sendo empregadas em diversas tarefas como, por exemplo, segmentação de imagens (Terman and Wang, 1995; Wang and Terman, 1995, 1997; Campbell et al., 1999; Liu et al., 2001), segregação de sinais sonoros ((von der Malsburg and Schneider, 1986; Wang, 1996;

Wang and Brown, 1999; Wrigley and Brown, 2004)), entre outros apresentados em Wang (2005).

No que refere-se à atenção visual, de acordo com Desimone and Duncan (1995), dados experimentais sugerem que objetos, durante a competição neural pela atenção, sejam representados como um todo, a partir da correlação paralela de diversas características. Sendo assim, considerando a correlação oscilatória como uma possível forma de tratamento para o problema da integração, destacamos aqui sua importância para a atenção visual baseada em objetos, propósito essencial para o desenvolvimento desta tese.

Neste contexto, diversos trabalhos têm sido propostos relacionados ao problema da integração (von der Malsburg, 1981; von der Malsburg and Schneider, 1986; Terman and Wang, 1995; Wang and Terman, 1995, 1997; Kazanovich and Borisyuk, 1999; Campbell et al., 1999; Wrigley and Brown, 2004), entre outros (ver (Tsotsos, 2011)), considerados trabalhos importantes para a modelagem da atenção. Conforme descrito anteriormente, uma das principais funções realizadas pela atenção é definir, a partir de diversas características, qual será a informação relavante que possibilitará o direcionamento da atenção. Assim, a atenção pode ser descrita como a integração de características de objetos, que se destacam sob o foco da atenção.

Considerando os princípios da seleção baseada em objetos, Wang (1999) propôs um modelo de atenção visual de objetos baseado no tamanho de cada segmento. Neste modelo, uma rede LEGION foi responsável pela segmentação da imagem e, para a seleção do maior segmento, um mecanismo de inibição lento foi introduzido, sendo responsável por armazenar o tamanho do maior segmento encontrado durante a oscilação da rede e, consequentemente, inibir a ativação de segmentos menores. Ao final do processo, somente o maior segmento está apto a oscilar. A Figura 4.13 apresenta o processo temporal da seleção. Neste trabalho, características referente à similaridade de *pixels* da imagem foram utilizadas durante o processo de integração da LEGION, enquanto que a característica relacionada ao tamanho foi utilizada para o direcionamento da atenção.

Em (Wang, 2002) foi proposto um modelo que utiliza como camada inicial uma rede LEGION para a segmentação cena, entretanto, com o objetivo de integrar diversos segmentos que fazem parte de um mesmo objeto através de características *top-down*, outras duas camadas relacionadas à memoria são utilizadas. De maneira geral, cada segmento sincronizado é apresentado à camada de reconhecimento, em seguida, segmentos reconhecidos referente ao mesmo padrão interagem entre si em uma camada específica, denominada por camada de memória de curto prazo, de forma que informações sejam retornadas à camada inicial, responsável por atualizar o sincronismo da rede LEGION de acordo com informações de alto nível sobre os segmentos (Figura 4.14). Embora experimentos tenham sido realizados com imagens

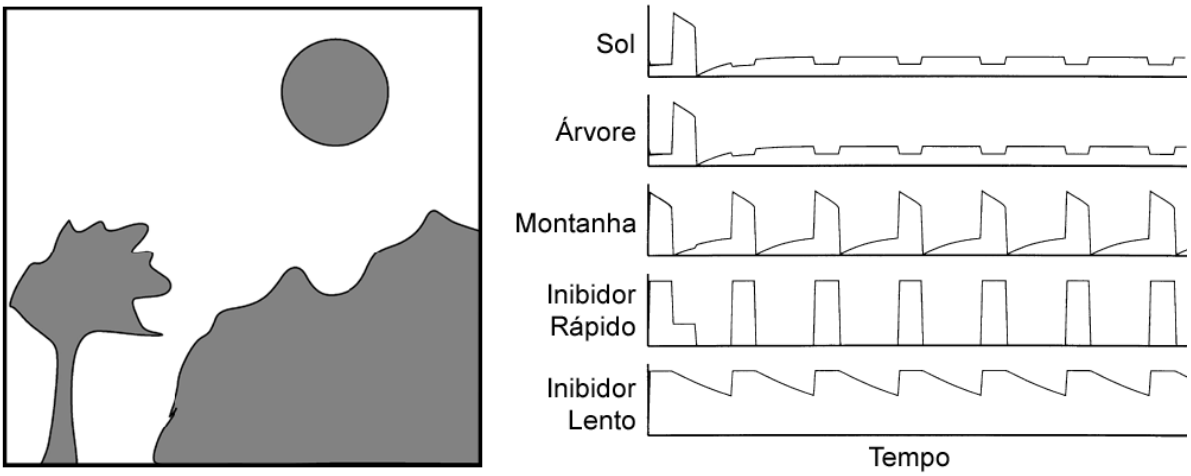


Figura 4.13: Modelo de atenção baseado no tamanho do objeto proposto por Wang (1999). Imagem de entrada (à esquerda) - Processo temporal de seleção (à direita).

sintéticas, o modelo demonstrou um interessante mecanismo para a integração de segmentos primitivos através de memória associativa.

Também de acordo com a seleção baseada em objetos, Kazanovich and Borisyuk (2002) demonstram que objetos simples podem ser selecionados, consecutivamente, através de uma rede neural oscilatória. Neste caso, a atenção é realizada através da sincronização de um oscilador central com um conjunto de osciladores que representam os objetos na imagem. De acordo com os autores, a função do oscilador central neste trabalho é similar à função do inibidor global do modelo LEGION proposto por Wang and Terman (1995). De acordo com a dinâmica do oscilador central, osciladores de objetos sincronizados e temporalmente ativos, têm seus valores de amplitude aumentado, enquanto que, para os demais osciladores, este valor é diminuído. A Figura 4.15 apresenta a arquitetura desta rede.

Baseado em um modelo direcionado para o processamento de imagens proposto por LindBlad (2005), denominado de Rede Neural de Pulso Acoplado (*Pulse-Coupled Neural Network* - PCNN), Quiles et al. (2006) propuseram um modelo de atenção visual aplicado à segmentação. Neste modelo, cada objeto da imagem é representado por um trem de pulsos em sincronia, enquanto que objetos distintos são representados por grupos de neurônios fora de fase. Nos experimentos, o modelo desenvolvido se mostrou adequado na segmentação de objetos não separáveis linearmente, como por exemplo, o problema da dupla-espiral (Cheng et al., 2001). Entretanto, o modelo proposto limita-se à segmentação temporal dos objetos presentes, não focalizando a atenção do modelo a um dos objetos segmentados. Em (Quiles et al., 2007) foram propostas modificações no modelo PCNN original, de forma a capacitar o modelo proposto a realizar a segmentação e seleção do objeto mais saliente de acordo com seu valor de intensidade. O modelo foi capaz de segmentar, selecionar um objeto em um instante do tempo e transferir o foco de atenção entre os diversos objetos presentes na imagem de entrada.

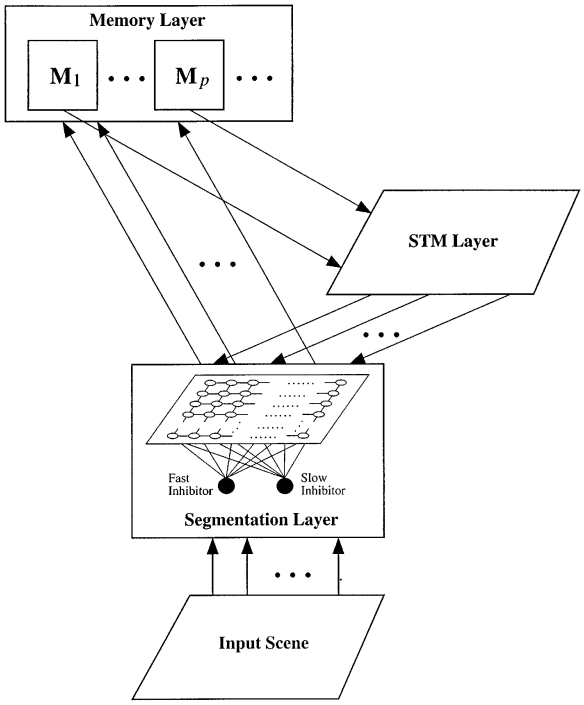


Figura 4.14: Modelo de atenção *bottom-up* e *top-down* baseado na análise de cenas proposto por Wang (2002).

No modelo de Wang (1999), apenas o tamanho do objeto é considerado como atributo saliente. Entretanto, sabe-se que o tamanho dos objetos é apenas um dos diversos atributos utilizados pelo sistema de visão ao selecionar um objeto (Wolfe and Horowitz, 2004). A saliência baseada no tamanho do objeto é uma característica a nível de objeto, enquanto que o mapa de saliência apresentado na Seção 4.1 é gerado a partir de informações locais. De acordo com Wang (2005), modelos de seleção de objetos são compatíveis com teorias baseadas em objetos, enquanto que modelos WTA são compatíveis com teorias baseadas na localização. Entretanto, em um trabalho recente, Quiles et al. (2011) propuseram um modelo de atenção visual baseado, tanto na correlação oscilatória, quanto no mapa de saliência. Conforme o

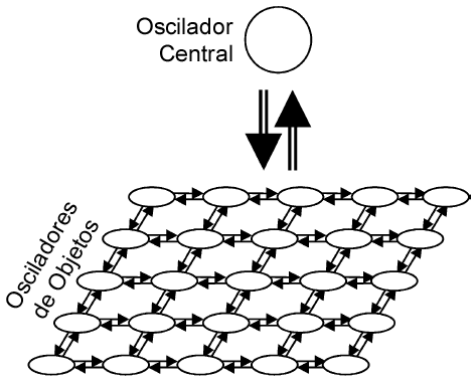


Figura 4.15: Arquitetura da rede de osciladores baseado em objetos proposta por Kazanovich and Borisyuk (2002).

diagrama apresentado na Figura 4.16, a imagem de entrada é segmentada por uma rede LEGION e, paralelamente, o mapa de saliência é gerado. Em seguida, um mapa de saliência de objetos é gerado através da integração dos resultados do mapa de saliência e da segmentação LEGION. Um mecanismo de inibição de retorno é utilizado para permitir o direcionamento da atenção entre os objetos destacados no mapa. De acordo com os autores, este é o primeiro modelo capaz de selecionar objetos em cenas reais baseado na saliência do objeto. Entretanto, nenhum enviesamento *top-down* foi considerado neste trabalho.

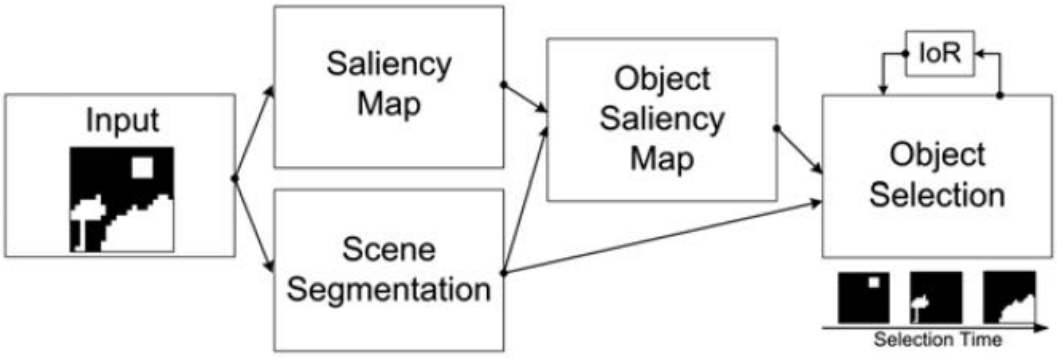


Figura 4.16: Diagrama de integração de módulos proposto por Quiles et al. (2011).

4.4 Pontos de Investigação

Embora exista uma grande quantidade de modelos propostos recentemente baseados nas hipóteses apresentadas neste capítulo, nos limitamos à revisão de modelos direcionados à atenção visual baseados principalmente em características *bottom-up*, enviesamento *top-down* e seleção baseada em objetos, assim como na inter-relação destas.

Mediante à constante evolução dos modelos de atenção e elementos necessário à sua composição, relacionamos aqui algum critérios, com o objetivo de apresentar uma visão ampla dos modelos revisados e também situar as principais contribuições relacionadas aos modelos propostos nesta tese. De acordo com o comportamento e estudos computacionais dos modelos apresentados, foram selecionados dezesseis critérios (`c`) para análise, que seguem:

- `c1` - *bottom-up*: saliência baseada em características primitivas da cena, de forma que a atenção seja direcionada, de maneira involuntária, para uma região desconhecida previamente;

- `c2` - *top-down*: saliência baseada em características do alvo a ser procurado na cena, de forma que a atenção seja direcionada para o alvo desejado de maneira voluntária;

- `c3` - mapa de saliência: utilização de um mapa representativo das características salientes da cena;
- `c4` - mapa de atributo-saliência: utilização de um mapa representativo das características salientes da cena gerado a partir da auto-organização de características em um plano bidimensional, proposto aqui como umas das contribuições desta tese;
- `c5` - enviesamento *top-down space-based*: saliência determinada por processos cognitivos, guiada através de características primitivas, denominadas por características *top-down* baseada no espaço. As características *top-down* são informação conhecidas previamente como, por exemplo, a cor ou orientação do alvo desejado, podendo ser definidas manualmente, de acordo com o que se deseja enfatizar, ou extraídas em um estágio prévio (treinamento) a partir de cenas contento o alvo, por meio de algoritmos de extração de características;
- `c6` - enviesamento *top-down object-based*: saliência também determinada através de processos cognitivos, porém guiada pelo fator de reconhecimento de objeto previamente conhecidos, obtido através de mecanismos de classificação;
- `c7` - *space-based*: mecanismo de seleção visual baseado no espaço, caracterizado pelo direcionamento da atenção entre pontos salientes;
- `c8` - *object-based*: mecanismo de seleção visual baseado em objetos, semelhante ao anterior, porém a seleção visual é sequencialmente direcionada para o próximo objeto, de acordo com uma escala de saliência;
- `c9` - correlação temporal: desenvolvimento da atenção baseada na seleção sequencial de segmentos ou objetos presentes na cena, desconsiderando qualquer informação sobre saliência;
- `c10` - competição por atenção baseada no espaço: entrega da atenção baseada na competição entre pontos ou neurônios que representam saliências locais, evitando a entrega sequencial da atenção entre pontos em uma escala de saliência;
- `c11` - competição por atenção baseada em objetos: semelhante ao critério anterior, considerando porém que a competição ocorra entre objetos;
- `c12` - reconhecimento do objeto: possibilidade de reconhecimento do objeto após o desenvolvimento da atenção;
- `c13` - cenas sintéticas: resultados experimentais obtidos da aplicação do modelo a partir cenas sintéticas;
- `c14` - cenas naturais: resultados experimentais obtidos da aplicação do modelo a partir de cenas naturais;

- `c15` - atributos: características utilizadas para guiar a atenção, podendo ser baseadas tanto no espaço, sendo: cor (c), orientação (o), intensidade(i); quanto baseadas no objeto, como: localização espacial na cena (l), tamanho (t) ou fator de reconhecimento (r);

- `c16` - tipo de tarefa: de acordo com Borji and Itti (2013) os tipos de tarefas podem ser classificados de três formas. O primeiro refere-se a busca livre de tarefas (f-*free*), onde a atenção do observador é direcionada involuntariamente para algum estímulo na cena, o segundo tipo de tarefa é denominado de busca visual (s-*source*), onde o observador direciona voluntariamente a atenção, através de movimentos sacádicos, à procura de características ou objetos específicos na cena. O terceiro tipo de tarefa (i-*interactive*), caracteriza-se pela composição dos dois anteriores.

Na Tabela 4.1 são apresentados os modelos de atenção visual descritos nas Seções 4.1, 4.2 e 4.3, e suas categorizações, de acordo com os critérios citados.

De acordo com os trabalhos apresentados, podemos notar que modelos baseados exclusivamente em mapas de saliência, a partir de meados da década passada, passaram a propor mecanismos para o enviesamento da atenção visual, com o objetivo de ativar regiões ou pontos específicos do mapa de saliência, de maneira manual ou por características salientes do próprio alvo. Embora o mapa de saliência tenha sido mantido como principal componente para a seleção visual baseada no espaço, o tipo de tarefa destes modelos foram alterados de busca livre de tarefas (f) para busca visual (s), tornando-os especificamente modelos *top-down*. Como conseqüência do uso de enviesamento *top-down*, características salientes destacadas inicialmente no mapa de saliência são desconsideradas, caso estas não estejam associadas às características previamente conhecidas sobre o alvo. Como um dos pontos de investigação e contribuição desta tese, pretendemos propor modelos de atenção visual com enviesamento *top-down*, entretanto, com a inserção de mecanismos que torne possível realizar modulações *top-down* baseadas no espaço e também a nível de objeto, permitindo que a atenção seja direcionada tanto para alvos desejados, quanto para demais regiões salientes. Diferente do modelo proposto em (Frintrop, 2006), onde o enviesamento *top-down* é baseado apenas em informações espaciais.

Podemos considerar que modelos de atenção visual baseados em mapas de saliência seguem um propósito comum, a detecção de regiões salientes da cena. Entretanto, de acordo com os propósitos iniciais de um modelo de saliência básico, nenhuma consideração é direcionada à estrutura dos objetos, de um modo geral. Esta limitação dificulta, por exemplo, a detecção ou reconhecimento de objetos específicos, contudo, modelos com estas características podem ser utilizados como um importante componente para modelos mais abrangentes, de forma a auxiliar, em uma fase inicial, à entrega da atenção para regiões de maior interesse. Consequentemente, como ponto de investigação e contribuição, pretendemos propor modelos

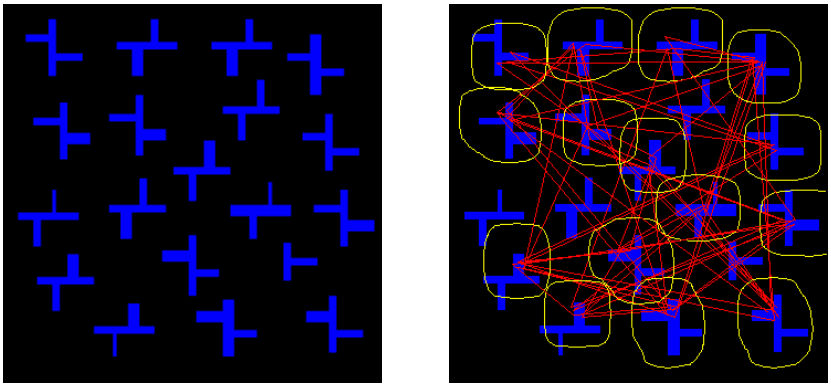
Tabela 4.1: Categorização dos modelos de atenção visual. Critérios em ordem: *bottom-up* (1), *top-down* (2), mapa de saliência (3), mapa de atributo-saliência (4), enviesamento *top-down space-based* (5), enviesamento *top-down object-based* (6), *space-based* (7), *object-based* (8), correlação temporal (9), competição por atenção baseada no espaço (10), competição por atenção baseada em objetos (11), reconhecimento do objeto (12), cenas sintéticas (13), cenas naturais (14), atributos (15) e tipo de tarefa (16).

Autor e Ano	Critérios Analisados															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Modelos baseados em Mapas de Saliência																
Itti et al. (1998)	+	-	+	-	-	-	+	-	-	-	-	-	-	+	COI	f
Itti and Koch (2000)	+	-	+	-	-	-	+	-	-	-	-	-	+	+	COI	f
Walther et al. (2002)	+	-	+	-	-	-	+	+	-	-	-	+	+	-	COI	f
Rutishauser et al. (2004)	+	-	+	-	-	-	+	+	-	-	-	+	-	+	COI	f
Walther et al. (2005)	+	-	+	-	-	-	+	+	-	-	-	+	-	+	COI	f
Walther and Koch (2006)	+	-	+	-	-	-	+	+	-	-	-	+	+	+	COI	f
Harel et al. (2006)	+	-	+	-	-	-	+	-	-	-	-	-	-	+	COI	f
Modelos com Enviesamento *top-down*																
Clark and Ferrier (1989)	-	+	+	-	+	-	+	-	-	-	-	-	-	+	COI	s
Wolfe (1994)	-	+	+	-	+	-	+	-	-	-	-	-	+	-	OI	s
Navalpakkam and Itti (2005)	-	+	+	-	+	-	+	-	-	-	-	+	+	+	COI	s
Bonaiuto and Itti (2006)	-	+	+	-	+	-	+	-	-	-	-	+	-	+	COI	s
Frintrop (2006)	+	+	+	-	+	-	+	+	-	-	-	+	+	+	COI	f/s
Navalpakkam and Itti (2006a)	-	+	+	-	+	-	+	-	-	-	-	+	+	+	COI	s
Navalpakkam and Itti (2006b)	-	+	+	-	+	-	+	-	-	-	-	+	+	+	COI	s
Elazary and Itti (2010)	-	+	+	-	+	-	+	-	-	-	-	+	-	+	COI	s
Borji et al. (2011)	-	+	+	-	+	-	+	-	-	-	-	+	+	+	COI	s
Modelos baseados na Correlação Temporal																
Wang and Terman (1995)	+	-	-	-	-	-	-	-	+	-	-	-	+	-	-	f
Wang and Terman (1997)	+	-	-	-	-	-	-	-	+	-	-	-	+	+	-	f
Campbell et al. (1999)	+	-	-	-	-	-	-	-	+	-	-	-	+	+	-	f
Wang (1999)	-	+	-	-	-	-	-	+	+	-	+	-	+	-	T	s
Wang (2002)	-	+	-	-	-	+	-	-	+	-	-	+	+	-	-	s
Kazanovich and Borisyuk (2002)	+	-	-	-	-	-	-	-	+	-	-	-	+	-	-	f
Quiles et al. (2006)	+	-	-	-	-	-	-	-	+	-	-	-	+	-	-	f
Quiles et al. (2007)	+	-	-	-	-	-	-	+	+	-	-	-	+	-	I	f
Quiles et al. (2011)	+	-	+	-	-	-	-	+	+	-	-	-	-	+	COIT	f
Modelos Propostos																
Benicasa and Romero (2010)	+	-	-	+	-	-	-	+	+	+	-	-	+	-	CP	f
Benicasa et al. (2012)	+	+	-	+	+	+	-	+	+	+	-	+	+	-	COIPR	f/s
Benicasa et al. (012b)	+	+	-	+	+	+	-	+	+	+	-	+	+	+	COIPR	f/s
Benicasa et al. (2013)	+	+	+	-	+	+	-	+	+	-	+	+	+	+	COIPRT	f/s
Benicasa et al. (013b)	-	+	+	-	+	+	-	+	+	-	+	+	-	+	COIPRT	s

de atenção baseados inicialmente na saliência da cena, porém, não exatamente para a identificação da saliência relacionada a um ponto ou região específica, mas sim à objetos, com o objetivo de tornar possível a análise de características associadas a cada objeto, auxiliando na competição por atenção baseada em objetos e também no reconhecimento de objetos mais complexos.

Ainda sobre os mapas de saliência, é importante notar que, para garantir o desempenho de modelos de atenção visual baseados em mapas de saliência, mesmo apresentando características baseadas em objetos, espaço, ou ainda, com ou sem enviesamento *top-down* de baixo nível, faz-se necessário a existência de características salientes relacionadas aos objetos presentes na cena. Assim, como ponto de investigação, temos como objetivo propor modelos de atenção visual para o direcionamento da atenção para objetos que apresentem padrões diferentes dos distratores, mesmo em situações onde a saliência não possa ser identificada a partir de informações espaciais. Na Figura 4.17 é apresentado um exemplo onde a saliência do objeto

contrastante ("cruz na posição horizontal") não está intrínseca nas características do objeto, mas em aspectos cognitivos de alto nível (critério $c4$).



(a) Imagem de Entrada **(b)** Sacadas sem sucesso

Figura 4.17: Exemplo de alvo ("cruz na posição horizontal") com saliência nula baseada no modelo proposto por Itti and Koch (2000).

Considerando também o comportamento evolutivo dos modelos computacionais de atenção visual baseados em mapas de saliência, pretendemos investigar o tratamento de experimentos que apresentem alvos, neste caso com características salientes, de modo que cada alvo seja composto por características não contrastantes entre si. Na Figura 4.18 é apresentado um experimento composto por duas imagens (Figura 4.18 (a) e (b)), nas quais os alvos (placas de sinalização), em termos de características primitivas, e de acordo com seus respectivos mapas de saliência (Figura 4.18 (c) e (d)) (Itti and Koch, 2000), são bastante semelhantes, embora seus significados cognitivos sejam distintos.

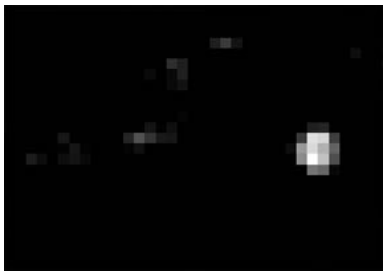
Outra fator a ser analisado é pertinente ao controle cognitivo da atenção. Uma questão norteadora para o desenvolvimento desta tese: podemos considerar como mais saliente o objeto que apresenta o maior valor de saliência em uma determinada escala? Consideremos o exemplo apresentado na Figura 4.19, onde características diferentes (cor e posição) precisam ser analisadas de maneira conjuntiva para a identificação do padrão contrastante (maça distante do agrupamento). Neste caso, a seleção visual baseada no espaço, regida somente por um mapa de saliência, tornaria difícil a entregar da atenção ao local contrastante. Como pode ser observado na Figura 4.19 (b), de acordo com o mapa de saliência proposto em (Itti et al., 1998), a maça foi a terceira região selecionada.

Na Figura 4.20 é apresentado outro experimento, onde o tamanho do objeto é a característica contrastante responsável por direcionar a atenção visual.

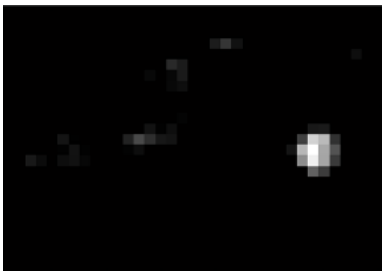
Considerando os experimentos apresentados nas Figuras 4.19 e 4.20, pretendemos viabilizar a competição pela atenção visual baseada em características do objeto. Para esta finalidade propomos um novo mecanismo para o desenvolvimento da competição por atenção baseado em mapas auto-organizáveis, de forma que, em alguns modelos propostos, o tradicional mapa de saliência será substituído por um

(a) Imagem 1 - Alvo - Curva Acentuada à Esquerda

(b) Imagem 2 - Alvo - Curva à Esquerda



(c) Mapa de Saliência - Imagem 1



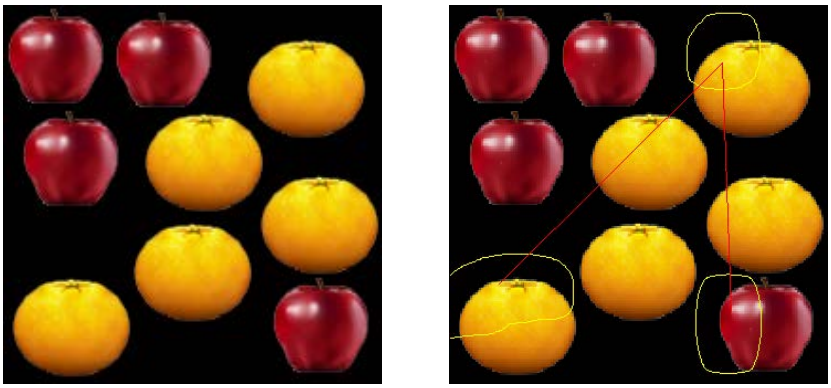
(d) Mapa de Saliência - Imagem 2

Figura 4.18: Exemplo de alvos cognitivamente distintos e características baseadas no espaço semelhantes.

mapa que denominamos de Mapa de Atributo-Saliência (MAS).

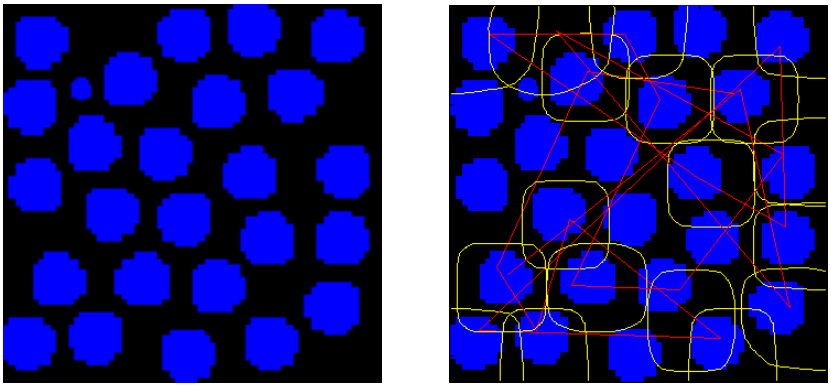
Como pode ser observado nos modelos baseados na correlação temporal, apresentados na Seção 4.3 e agrupados na Tabela 4.1, diversas contribuições nesta linha ainda podem ser propostas, corroborando com o objetivos desta tese. A categorização dos modelos propostos foi previamente apresentada na Tabela 4.1, de modo que destacamos a seguir os principais objetivos a serem pesquisados:

- Proposta de um mapa de saliência *bottom-up*, denominado aqui por mapa de atributo-saliência, gerado a partir da auto-organização de atributos primitivos da cena;
- Proposta de um mapa de saliência *bottom-up* e *top-down*, contendo as mesmas características do mapa de atributo-saliência, entretanto, gerado a partir da auto-organização de atributos primitivos da cena e informações cognitivas sobre objetos;
- Desenvolvimento da atenção atenção visual baseada em objetos, a partir das características de intensidade, cor, orientação, reconhecimento e tamanho;
- Proposta de mecanismo para competição por atenção baseada no espaço;
- Proposta de mecanismo para competição por atenção baseada em objetos;
- Enviesamento *top-down* baseado em objetos;
- Enviesamento *top-down* baseado no espaço;



(a) Imagem de Entrada **(b)** Sacadas sem sucesso

Figura 4.19: Exemplo de busca conjuntiva sem sucesso por meio de seleção visual baseada no espaço. De acordo com a seleção visual (Itti and Koch, 2000), a região contendo a maça distante do agrupamento foi selecionada após a terceira sacada.



(a) Imagem de Entrada **(b)** Sacadas sem sucesso

Figura 4.20: Exemplo de busca sem sucesso baseada em característica a nível de objeto.

- Seleção visual baseada em objetos;
- Reconhecimento de objetos sob o foco da atenção;
- Detecção de saliência a partir da presença ou ausência de características salientes.

4.5 Considerações Finais

De acordo com os modelos computacionais para atenção visual apresentados neste capítulo, pode-se observar que pesquisas computacionais na área de atenção visual estão relacionadas a diferentes áreas computação. Inicialmente foram apresentados modelos para atenção visual baseados em mapas de saliência, gerados a partir de informações contidas na própria cena. Em seguida, foram abordados modelos com propostas baseadas em informações previamente conhecidas sobre determinados alvos. Finalmente, com o objetivo de se obter a seleção visual baseada

em objetos, foram apresentados modelos baseados na correlação temporal, onde a segmentação pré-atentiva da cena em objetos deve ocorrer, possibilitando que cada objeto possa servir de alvo para a atenção visual. De acordo com as características dos modelos revisados, concluímos este capítulo apresentando os pontos de investigação e principais objetivos a serem realizados nesta tese.

No capítulo seguinte serão apresentadas as descrições detalhadas dos modelos propostos e os resultados obtidos através dos estudos realizados.

CAPÍTULO

5

Modelos Computacionais Propostos para Atenção Visuais

Nos capítulos anteriores foram apresentados diversos modelos computacionais de atenção visual, tendo sido discutido suas características e limitações. Neste capítulo serão apresentados os modelos propostos nesta tese para atenção visual.

5.1 Mapa de Atributo-Saliente e Localização de Objeto Saliente

Esta pesquisa é direcionada principalmente à criação de um mapa de saliência, diferente do mapa de saliência proposto em (Itti et al., 1998; Itti and Koch, 2000), denominado aqui por Mapa de Atributo-Saliência (MAS), gerado a partir de modelos de RNAs. Para sua modelagem, implementação, simulação e validação foi utilizada uma RNP composta por neurônios do tipo I&F (Seção 3.2.2) e um mapa SOM (Seção 3.3). Simulações foram desenvolvidas para verificar a viabilidade do modelo como um mecanismo de seleção de atributos. Foram realizadas ainda duas extensões nesta proposta, diferenciadas pelos padrões utilizados para o treinamento da rede SOM, descritas em detalhes nas seções seguintes.

5.1.1 Mapa de Atributo-Saliente

A inspiração biológica desta pesquisa está relacionada às evidências biológicas de que células do córtex cerebral dos mamíferos organizam-se de forma altamente estruturada em suas funções, o que resulta em regiões do cérebro especificamente capacitadas no processamento sensorial de sinais como, por exemplo, visão, audição, controle motor e linguagem (Kohonen, 2001). Considerando também o fluxo

da informação através do córtex visual, apresentado na Seção 2.1, podemos concluir que os neurônios exibem uma ordem física, tal que características semelhantes no espaço, ou campo visual, sejam processados por neurônios fisicamente próximos no córtex cerebral. A seguir será apresentada a descrição detalhada do modelo proposto.

Descrição do Modelo

Para alcançar os objetivos desta pesquisa e, para um melhor entendimento de seu desenvolvimento, o processo foi decomposto nas seguintes etapas: sincronização de osciladores e segmentação da imagem; geração de um mapa de cores; e a geração do mapa de atributo-saliência.

De acordo com os conceitos apresentados na Seção 3.2.2, sobre sincronização em RNPs com neurônios do tipo I&F e, com a utilização do atributo cor como característica para selecionar a vizinhança, obteve-se cada objeto da cena representado por um trem de pulsos em sincronia, enquanto que objetos distintos foram representados por grupos de osciladores fora de fase (Figura 5.1). Conforme descrito, a rede de neurônios do tipo I&F foi definida como:

$$x_i(t + \Delta t) = x_i(t) + ((-x_i(t) + I) * \Delta t), \qquad i = 1, .., n, \tag{5.1}$$

onde x_i representa o potencial do oscilador i, n é o número de osciladores, o parâmetro I controla o período de um oscilador dessincronizado e Δt representa a discretização do tempo de disparo. A atualização do potencial dos osciladores vizinhos é definido como:

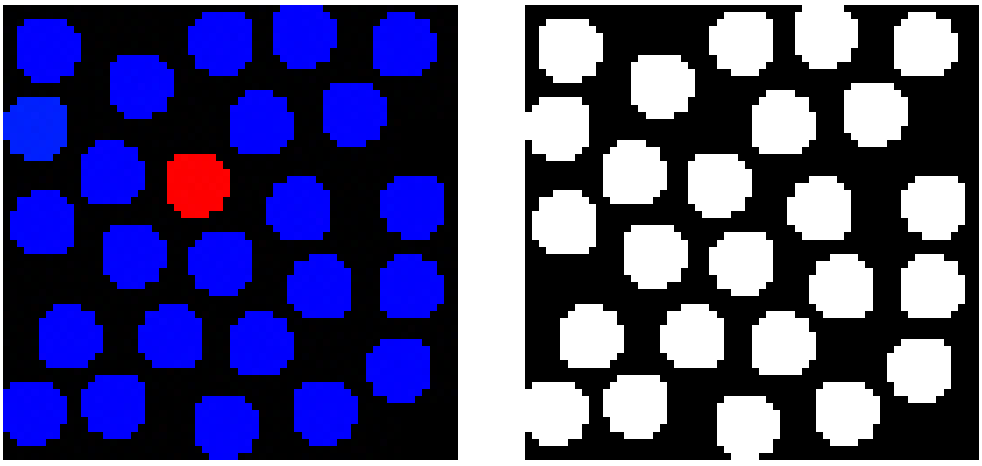
$$x_{ij} = x_{ij} + J_{ij}, \qquad i = 1, .., Z_i \tag{5.2}$$

$$J_{ij} = \frac{\alpha_s}{Z_i}, \tag{5.3}$$

onde x_{ij} é o j-*ésimo* oscilador vizinho de i, J_{ij} é força de acoplamento a partir do oscilador i para j, α_s é o fator que defina a força do sincronização e Z_i é o número de vizinhos do oscilador i. É importante notar que somente dois parâmetros precisam ser ajustados: α_s e I.

A geração do mapa SOM é baseada na descrição apresentada na Seção 3.3. De forma resumida, o mapa SOM é criado a partir padrões com valores aleatórios de atributos do tipo RGB (*vermelho, verde e azul*), organizado em um arranjo bidimensional. O cálculo das distâncias do vetor de pesos $m_s = [r, g, b]$ de cada neurônio s do SOM, em relação ao padrão aleatório de entrada $p = [r, g, b]$, é definido por:

$$d(p, m_s) = \| p - m_s \| = \sqrt{\sum_{j=1}^{3} (p_j - m_{sj})^2}, \qquad j \in \{r, g, b\}, \tag{5.4}$$



(a) Imagem Original **(b)** Imagem Segmentada

Figura 5.1: Processo de Sincronização e Segmentação.

de modo que o neurônio vencedor é selecionado de acordo com:

$$c = \arg\min_{i} \|p - m_s\| \,, \tag{5.5}$$

seguido pela atualização do peso sináptico do s-ésimo neurônio do SOM, em instante de tempo $(t+1)$, definido como segue:

$$m_s(t+1) = m_s(t) + \alpha_k(t)\ h_{ci}(t)\ [p(t) - m_s(t)], \tag{5.6}$$

onde $t = 0, 1, 2, ...$ é um número inteiro representando a coordenada discreta de tempo e $\alpha_k(t)$ define a taxa de aprendizado. O grau de adaptação do neurônio vencedor e de seus vizinhos depende, portanto, da função de vizinhança, $h_{ci}(t)$ e de $\alpha_k(t)$. Consideramos $h_{ci}(t)$ como:

$$h_{ci}(t) = \exp\left(-\frac{\| r_c - r_i \|^2}{2\sigma^2(t)}\right), \tag{5.7}$$

onde o parâmetro $\sigma(t)$ define a largura da região de vizinhança e r_c e r_i representam as posições dos neurônios de índices c e i dentro do arranjo. De acordo com Kohonen (2001), $\sigma(t) \to 0$ quando $t \to \infty$. Para a convergência do mapa, utilizou-se a seguinte função no decremento de $\alpha_k(t)$ e $\sigma(t)$:

$$\alpha_k(t+1) = \alpha_k(t)\ 0.9\ (1 - (t/n_{it})) \tag{5.8}$$

$$\sigma(t+1) = \sigma(t)\ 0.9\ (1 - (t/n_{it})) \tag{5.9}$$

onde n_{it} é número de iterações necessárias para decrementar α_k e σ até zero.

Após o treinamento, cores próximas encontram-se mapeadas em locais próximos na rede. A convergência do mapa é apresentado na Figura 5.2.

Com base no processo de sincronização e segmentação dos neurônios de

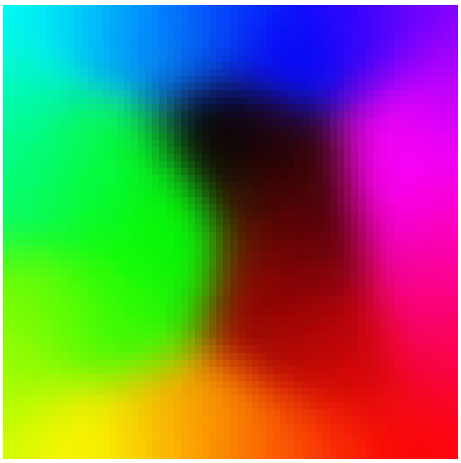


Figura 5.2: Mapa SOM de Cores. Valores de parâmetros utilizados: $\alpha_k = 0.5$, $\sigma = 300$ e $n_{it} = 10000$.

entrada e no mapa SOM de cores, um modelo de atenção visual baseada em um mapa de atributo-saliência foi proposto. O modelo é formado por uma rede neural composta por neurônios com dois tipos de conexões: conexões excitatórias e conexões inibitórias. As conexões excitatórias formam um mecanismo cooperativo responsável por sincronizar grupos de neurônios que representam padrões próximos no mapa SOM (neurônios com pesos similares), podendo estar relacionados a um mesmo segmento (objeto) ou segmentos com características semelhantes. Por outro lado, as conexões inibitórias têm como objetivo inibir regiões do SOM relacionadas a objetos de fundo da cena, permitindo que regiões do SOM, relacionadas aos objetos mais salientes, sejam selecionadas.

O mapa de atributo-saliência foi constituído pelo mesmo número de neurônios que a rede SOM de cores. Sua dinâmica foi definida da seguinte forma: considera-se uma imagem de entrada, sincronizada e segmentada. Quando um objeto (*neurônio*) x_i pulsar em um instante t, definido pela Equação 5.1, seu sinal é apresentado para à rede SOM. Cada neurônio do SOM estimula seu neurônio associado no mapa de atributo-saliência, que terá seu estado atualizado como segue:

$$\dot{v}_i = -v_i + E_i - W_Y Y_i, \qquad i = 1, .., n, \tag{5.10}$$

onde n é o número de neurônios do mapa de atributo-saliência. A Equação 5.10 representa um neurônio I&F. Considerando que o mapa de atributo-saliência é constituído pelo mesmo número de neurônios do mapa SOM, pode-se concluir que cada neurônio do SOM tem seu respectivo neurônio no mapa de atributo-saliência. A variável v_i representa o potencial de saliência do neurônio i e W_Y é o peso de inibição a partir do termo de acoplamento inibitório Y_i.

Consideramos l sendo um neurônio pertencente a um segmento ativo na rede I&F, e k seu respectivo índice, o padrão $l_k = [r, g, b]$ é apresentado ao SOM. A similaridade entre o padrão l_k, representando o determinado *pixel*, e cada neurônio

m_s do SOM é definido pela seguinte equação:

$$d(l,m) = \| l - m \| = \sqrt{\sum_{k=1}^{3}(l_k - m_k)^2}, \quad k \in \{r,g,b\} \tag{5.11}$$

Os termos de acoplamento excitatório E_i e inibitório Y_i são definidos pela seguinte equação:

$$E_i = Y_i = \exp^{-d(l,m_s)}, \qquad s = 1,..,n, \tag{5.12}$$

onde n representa o número de neurônios do SOM. É importante notar que, E_i será atualizado, se e somente se, o valor de E_i, em outro instante $t+1$, for superior t, ou seja, o termo E_i conterá o valor máximo excitatório do neurônio i, m_s representa cada neurônio do SOM e j é o índice de características.

As conexões inibitórias são determinadas com base no contraste entre atributos. Desta forma, se dois neurônios são alimentados por atributos semelhantes, ou seja, o contraste entre eles é pequeno ou zero, o termo Y_i da Equação 5.10 aproxima zero e, devido a função exponencial negativa da Equação 5.12, o peso de acoplamento inibitório assume um alto valor. Por outro lado, quando os sinais de alimentação de tais neurônios são definidos por atributos distintos, o peso da conexão inibitória entre eles é pequeno ou mesmo zero. Assim, objetos com características semelhantes são mutuamente inibidos, devido a competição gerada pelas conexões inibitórias. Podemos concluir que um objeto que apresenta um alto contraste com os demais não é inibido e permanece oscilando, salientando o atributo do objeto sob o foco de atenção. A Figura 5.3 apresenta, de forma ilustrativa, o processo descrito para a geração do mapa de atributo-saliência. É importante notar que, para a geração do mapa de atributo-saliência, não houve a necessidade de ajuste de parâmetro.

Simulações Computacionais

Apresentamos aqui os resultados de simulação obtidos a partir da aplicação do modelo descrito em imagens sintéticas. Para todas as simulações, o objeto saliente é definido como sendo àquele que apresenta um maior contraste com os demais objetos presentes na cena, denominados por objeto saliente e objetos de fundo, respectivamente. Esta suposição recebe suporte direto de experimentos biológicos que têm demonstrado que o contraste dos objetos que compõem uma determinada cena é mais importante que o nível absoluto de cada um dos atributos visuais em tarefas de inspeção visual (Wolfe and Horowitz, 2004; Yantis, 2000) (ver Cap.2).

As simulações têm como objetivo testar a capacidade de identificação do atributo saliente considerando o contraste de cores em cenas sintéticas. Na Figura 5.4 são apresentadas imagens contendo apenas um objeto saliente (*objeto vermelho*), ou seja, com alto contraste em relação aos objetos de fundo. As imagens apresen-

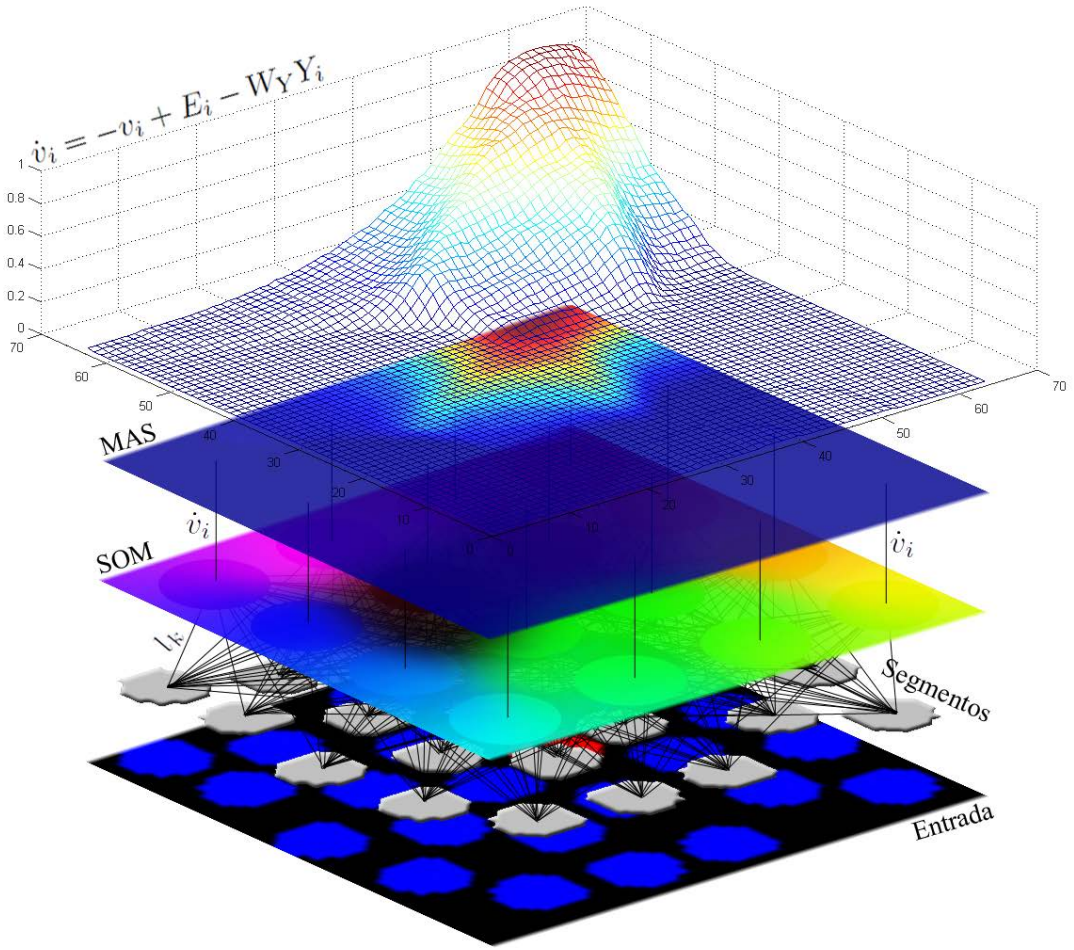


Figura 5.3: Processo de geração do Mapa de Atributo-Saliência.

tadas na coluna "Entrada" da Figura 5.4, são as respectivas imagens de entrada, com objetos de fundo, representando os distratores, que variam gradualmente de um padrão homogêneo até uma cena mais heterogênea. Os resultados obtidos utilizando estas imagens são apresentados na mesma linha, nas colunas à direita, sendo respectivamente o mapa SOM de cores, o Mapa de Atributo-Saliência com inibidor ativo, o Mapa de Atributo-Saliência com inibidor inativo e novamente o Mapa de Atributo-Saliência com inibidor ativo, porém com ênfase nas regiões salientes, possibilitando uma melhor visualização.

A partir destes resultados é possível observar que, mesmo estando o objeto saliente inserido em uma cena com distratores heterogêneos, o modelo apresenta um resultado coerente, no qual o atributo do objeto saliente (*vermelho*) é selecionado em todas as simulações. Este fenômeno coincide com os resultados obtidos em experimentos de seleção visual com humanos (Wolfe and Horowitz, 2004), assim como nos experimentos apresentados no Capítulo 2.

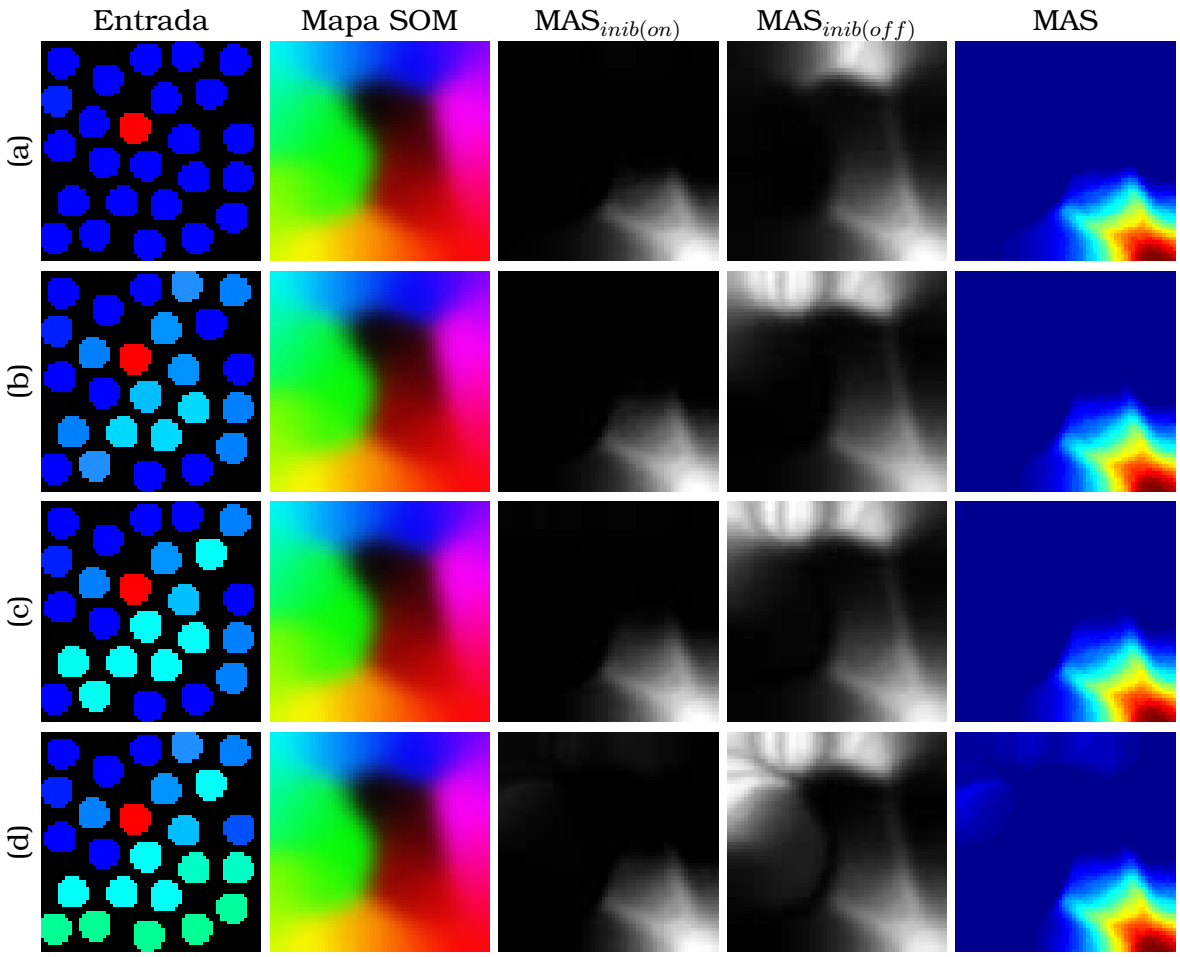


Figura 5.4: Mapa de Atributo-Saliente. Simulação variando a heterogeneidade dos distratores. (a), (b), (c) e (d) representam quatro níveis, variando de um fundo contendo objetos homogêneos a um padrão de objetos distratores com cores heterogêneas. Valores de parâmetros utilizados nas simulações: sincronização: $\alpha_s = 0.6$ e $I = 1.1$. Imagens com 64 x 64 *pixels* utilizadas em (Quiles et al., 2009)).

5.1.2 Treinamento Aleatório do SOM

Podemos considerar esta pesquisa como uma evolução natural da apresentada na Seção 5.1.1. Assim, serão focados somente as principais alterações propostas.

Na Seção 5.1.1, o MAS foi gerado a partir de um mapa SOM de cores e dos pulsos dos neurônios sincronizados e segmentadas pela RNP. O principal objetivo deste mapa foi salientar o atributo referente a cor do objeto mais saliente da imagem de entrada. Desta forma, não foi possível identificar a localização do objeto mais saliente especificamente, uma vez que informações sobre o atributo cor dos objetos estão relacionadas às localizações desconhecidas no espaço bidimensional do SOM, o que impossibilita qualquer futuro mecanismo de seleção visual selecionar, sequencialmente, cada um dos objetos salientes. Para demonstrar esta situação, consideramos a simulação apresentada na Figura 5.5, onde dois objetos com características similares (objetos *vermelhos*) são representados pela mesma região no

SOM. Para resolver este problema, de forma a identificar também a posição do objeto mais saliente, esta pesquisa possui como um de seus objetivos a utilização do mapa SOM, porém, gerado a partir de um conjunto aleatório de padrões compostos por atributos do tipo $rgbxy$ (*vermelho, verde, azul, posição* x *e posição* y), organizado em um arranjo bidimensional, onde cores com posições próximas sejam mapeadas em locais próximos no SOM.

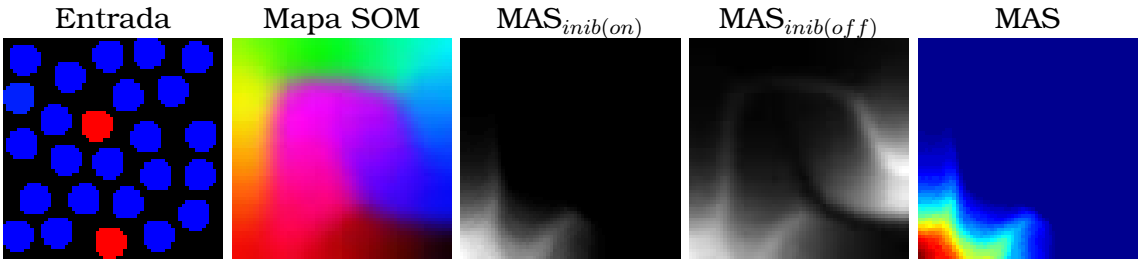


Figura 5.5: Mapa de Atributo-Saliência de objetos com o mesmo valor em relação à característica cor e posicionamento diferente. Imagem com 64 x 64 *pixels*.

Descrição do Modelo

Para o treinamento do SOM, considerou-se p um padrão de entrada tomado aleatoriamente. Consideramos aqui $p = [r, g, b, x, y]$ o padrão que será apresentado ao SOM. Dado a inserção dos atributos x e y no modelo, o cálculo da distância do vetor de pesos $m_s = [r, g, b, x, y]$ de cada neurônio s do SOM em relação ao padrão de entrada p foi definido como segue:

$$d(p, m_s) = \| p - m_s \| = \sqrt{\sum_{j=1}^{5} (p_j - m_{sj})^2}, \qquad j \in \{r, g, b, x, y\}, \tag{5.13}$$

onde todas as distâncias em relação à p são calculadas. Uma vez eleito o neurônio vencedor, realiza-se a atualização de m_s tanto do neurônio eleito, quanto de seus vizinhos, de forma a aumentar a representatividade desta região do SOM, em relação ao sinal de entrada. O processo de atualização do SOM é descrito conforme apresentado pela Equação 5.6.

A segmentação da cena é baseada nos conceitos apresentados na Seção 3.2.2, onde o atributo cor é a característica utilizada para a seleção da vizinhança. A sincronização da rede de neurônios I&F e a atualização do potencial dos osciladores vizinhos são definidas pelas Equações 5.1 e 5.2.

O mapa de atributo-saliência gerado nesta proposta tem como objetivo salientar a região do mapa relacionada aos atributos *cor* e *posição*, correspondente aos objetos saliente da cena. Embora o dinamismo para a geração do MAS tenha sido apresentado na proposta anterior (Seção 5.1.1), a medida de similaridade entre o padrão l_k e os neurônios m_s do SOM foram alterados para considerar os atributos x e y, necessários à geração do mapa de atributo-saliência proposto neste modelo,

descrito como segue:

$$d(p, m_s) = \| p - m_s \| = \sqrt{\sum_{j=1}^{5} (p_j - m_{sj})^2}, \qquad j \in \{r, g, b, x, y\}, \tag{5.14}$$

Desta forma, as posições x e y, saliente no mapa SOM deverão corresponder às posições dos objetos salientes da cena.

Simulações Computacionais

Os resultados das simulações foram obtidos a partir da aplicação do modelo proposto em imagens sintéticas. O principal objetivo das simulações foi analisar o comportamento do modelo em relação à precisão em identificar corretamente a cor predominante do objeto alvo no mapa de atributo saliência, assim como a posição correta do alvo.

A Figura 5.6 apresenta simulações com a mesma sequência de imagens sintéticas utilizadas no modelo proposto anterior (Figura 5.4), com variações de contraste entre alvo e distratores. Entretanto, temos como objetivo verificar a precisão dos resultados considerando, principalmente, os atributos x e y. Os resultados obtidos são apresentados nas colunas à direita da imagem de entrada, sendo respectivamente o mapa SOM, o Mapa de Atributo-Saliência com inibidor ativo, o Local de Maior Saliência (LMS) e novamente o Mapa de Atributo-Saliência com inibidor ativo, porém com ênfase nas regiões salientes.

Nas simulações apresentadas na Figura 5.6, a coluna LMS destaca, através do demarcador ("cruz"), o local de maior saliência. Em todas as simulações, o alvo foi precisamente localizado. É importante notar que, o aumento da heterogeneidade dos distratores faz com que outras regiões do mapa de atributo-saliência também tornem-se ativas. Entretanto, o modelo se manteve robusto quanto a localização do alvo. Como pode ser observado na coluna "MAS", a região de maior saliência permanece sempre no mesmo local (parte inferior esquerda do MAS).

Considerando o mesmo mapa SOM e parâmetros utilizados nas simulações apresentadas na Figura 5.6, na Figura 5.7 são apresentadas simulações com imagens contendo dois objetos salientes. Destacamos na coluna "LMS"o primeiro objeto saliente localizado. A localização do segundo alvo pode ser obtida através da verificação, em ordem decrescente, do potencial de saliência v_i (Equação 5.10) dos neurônios do MAS.

Na simulações seguintes, apresentadas na Figura 5.8 (a) e (b), utilizando ainda o mesmo mapa SOM, o modelo não apresentou precisão em relação às simulações anteriores, retornando uma localização aproximada do alvo (ver coluna "LMS"). Entretanto, para as mesmas imagens de entrada, analisamos o comportamento do modelo utilizando um segundo mapa SOM, gerado a partir do treinamento de padrões aleatórias compostos por atributos do tipo rxy (*vermelho*, *posição* x e *posição*

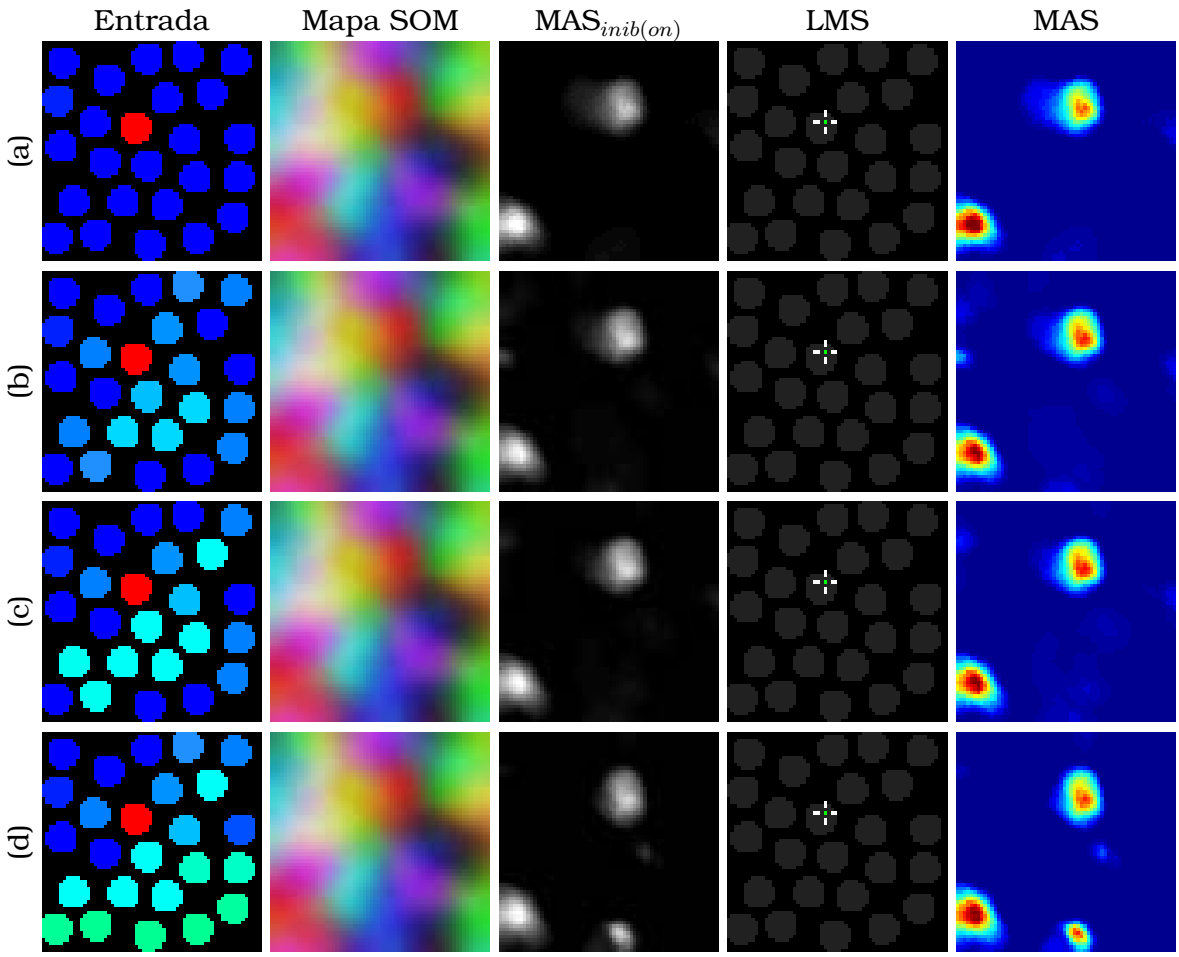


Figura 5.6: Localização de Objetos Salientes. Simulação variando a heterogeneidade dos distratores. (a), (b), (c) e (d) representam quatro níveis, variando de um fundo contendo objetos homogêneos a um padrão de objetos distratores com cores heterogêneas. Valores de parâmetros utilizados nas simulações: sincronização ($\alpha_s = 0.6$ e $I = 1.1$), rede SOM ($\alpha_k = 0.2$, $\sigma = 26$ e $n_{it} = 10000$). Imagens com 64 x 64 *pixels* utilizadas em (Quiles et al., 2009).

y). O objetivo foi de aumentar o número de representantes para os objetos salientes no MAS (Figura 5.8 (c) e (d)). Como pode ser observado na coluna "LMS", nas simulações (c) e (d), o modelo apresentou precisão satisfatória na localização dos alvos.

5.1.3 Treinamento Predefinido do SOM

Como uma evolução natural dos modelos anteriores, esta pesquisa é comumente baseada nas características apresentadas nas Seções 5.1.1 e 5.1.2. Entretanto, consideramos algumas alterações em relação à construção do mapa SOM que, consequentemente, implicam no comportamento do modelo de uma forma geral, apresentadas a seguir. Os resultados desta pesquisa, juntamente com as contribuições propostas nos modelos apresentados nas Seções 5.1.1 e 5.1.2, encontram-se publicados em (Benicasa and Romero, 2010).

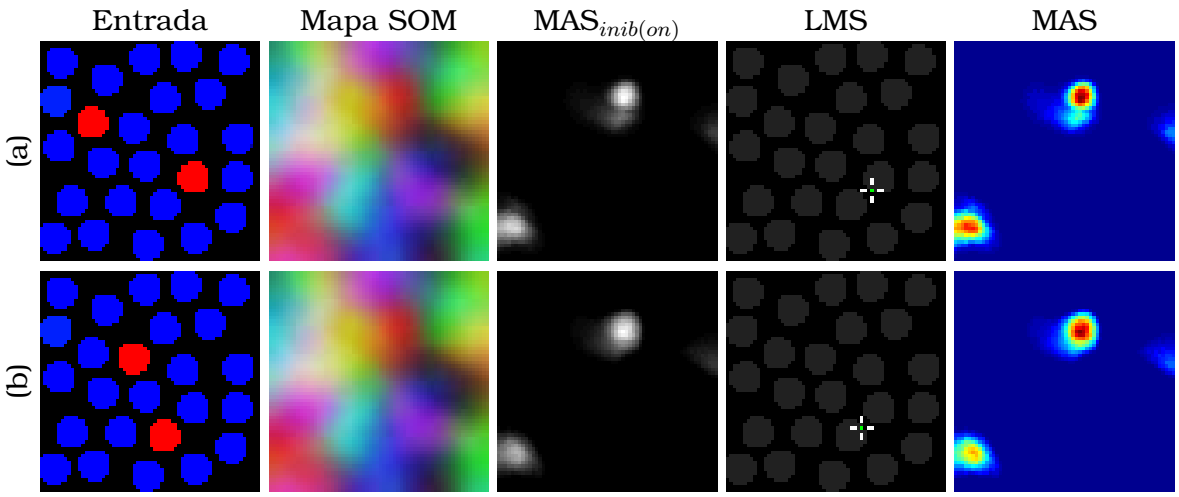


Figura 5.7: Localização de Objetos Salientes. Localização de dois objetos salientes. Imagens com 64 x 64 *pixels*.

Descrição do Modelo

De maneira resumida, para a geração do mapa de atributo-saliência, responsável por possibilitar a seleção visual entre objetos salientes, combinamos dois modelos de redes neurais, o primeiro, o modelo de Kohonen que, conforme mencionando, é responsável pelo agrupamento de cores e localizações espaciais dos objetos e, o segundo, uma rede neural pulsada, responsável pela sincronização e segmentação dos objetos presentes na cena. Na Figura 5.9 é apresentado o diagrama do modelo proposto, composto das seguintes etapas: geração do mapa de SOM; sincronização dos osciladores e segmentação da imagem; geração do mapa de atributo-saliência; e seleção dos objetos salientes.

A geração do mapa SOM é baseada na descrição apresentada na Seção 5.1.2, onde a distância entre o padrão de entrada p e o neurônio m_s do SOM é descrita pela Equação 5.13 e, encontrado o neurônio m_s vencedor, atualiza-se seu peso e dos neurônio vizinhos, de modo a estabelecer a interação local entre os neurônios da região. A atualização dos pesos ocorre de acordo com a Equação (5.6). Detalhes dos parâmetro são apresentados na Seção 3.3.

Neste trabalho, para aumentar a precisão do SOM, não adotamos padrões aleatórios para a fase de treinamento. Esta decisão foi tomada devido ao aumento do número de atributos mapeados em um único mapa bidimensional, influenciando diretamente a acurácia do modelo, principalmente em relação ao posicionamento dos objetos, devidos às constantes atualizações durante a fase de treinamento do SOM.

Sendo assim, para o treinamento do mapa SOM, inicialmente todos os neurônios da rede são inicializados com pesos aleatórios. Em seguida, os padrões utilizados para o treinamento são formados pelas características dos *pixels* presentes na própria cena, ou seja, as características $rgbxy$ (*red, green, blue* e as posições *x* e *y*),

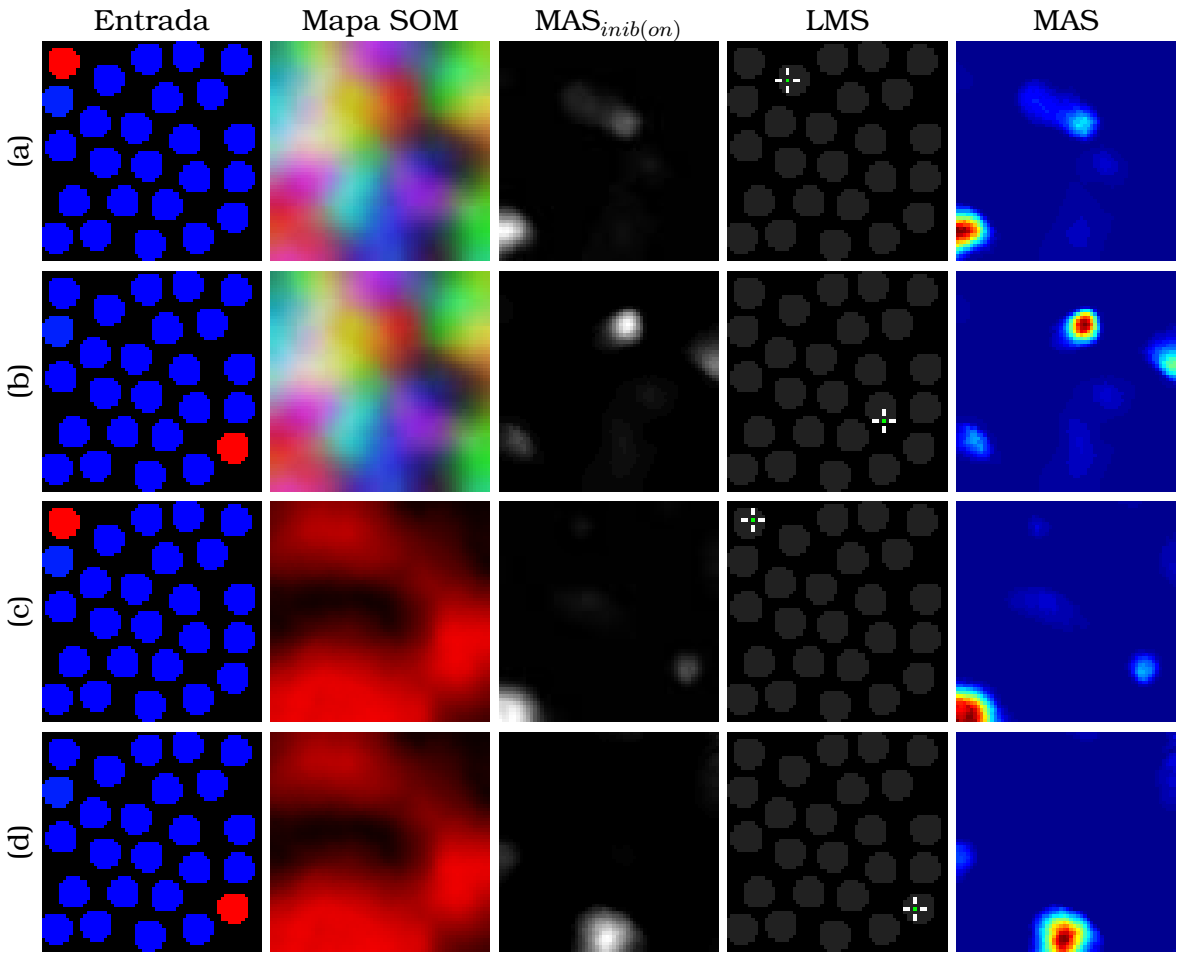


Figura 5.8: Localização de Objetos Salientes. Precisão na localização de alvos. Imagens com 64 x 64 *pixels*.

de modo que *pixels* com cores e posições próximas na imagem de entrada, sejam mapeados em cores e posições próximas no SOM. É importante notar a necessidade de normalização dos valores das características envolvidas. Na Figura 5.9 é apresentado um mapa SOM gerado a partir de uma imagem de entrada.

A sincronização dos osciladores e a segmentação da cena foi obtida por uma rede neural pulsada, composta por neurônios do tipo I&F, mantendo o atributo cor para a seleção da vizinhança, definida na Equação 5.1. É importante notar que, embora a seleção visual ocorra à nível de objetos, a competição pela atenção é baseada no espaço, onde cada neurônio x_i sincronizado da RNP, deverá participar da competição pela atenção.

Para identificar na cena o objeto saliente, foi considerado o termo v_i da Equação 5.10, que representa o potencial de saliência do neurônio i. No mapa de atributo-saliência, o atributo cor e posição do neurônio com maior valor de potencial v_i será o representante da cor e posição mais saliente da imagem de entrada. Com esta informação é possível identificar o objeto mais saliente da cena, devido ao fato desta estar previamente segmentada. Baseando-se nos valores de v_i, podemos identificar os próximos objetos salientes da cena.

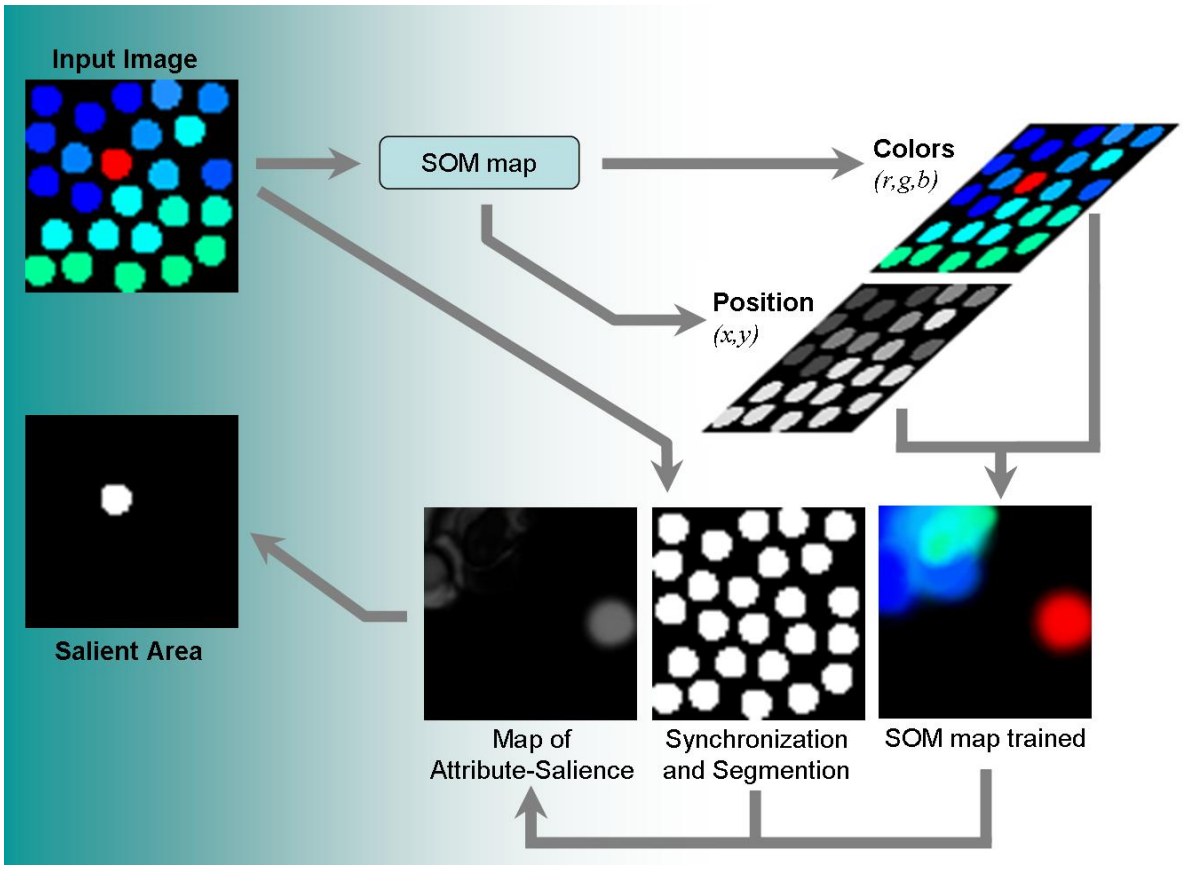


Figura 5.9: Diagrama do modelo proposto para a localização de objetos salientes II.

Simulações Computacionais

Os resultados das simulações foram obtidos a partir da aplicação do modelo, assim como nos experimentos anteriores, em imagens sintéticas. Os resultados obtidos são apresentados na Figura 5.10.

A partir dos resultados apresentados na Figura 5.10 foi possível concluir que, embora o objeto saliente esteja inserido em um cena com distratores heterogêneos, o modelos apresentou resultados coerentes[1], em que o objeto saliente (*vermelho*) foi selecionado em todas as simulações.

Para uma imagem contendo dois objetos salientes, o modelo proposto selecionou as duas regiões de maior saliência, como pode ser visto na Figura 5.11 (LMS).

Ainda em relação à simulação apresentada na Figura 5.11, pode-se notar no $\text{MAS}_{inib(on)}$ uma única região saliente, porém composta por duas regiões sobrepostas. Isto ocorre devido a auto-organização do SOM, onde os dois objetos de cores iguais e posições próximas foram mapeados para regiões próximas no mapa (veja Figura 5.12). Entretanto, estas regiões apresentam potenciais de ativação distintos, permitindo a seleção entre os dois objetos.

De acordo com Wolfe and Horowitz (2004), para a validação de um modelo de atenção visual, regiões selecionadas devem coincidir com os resultados obtidos a

[1] *Resultados de acordo com a atenção visual humana*

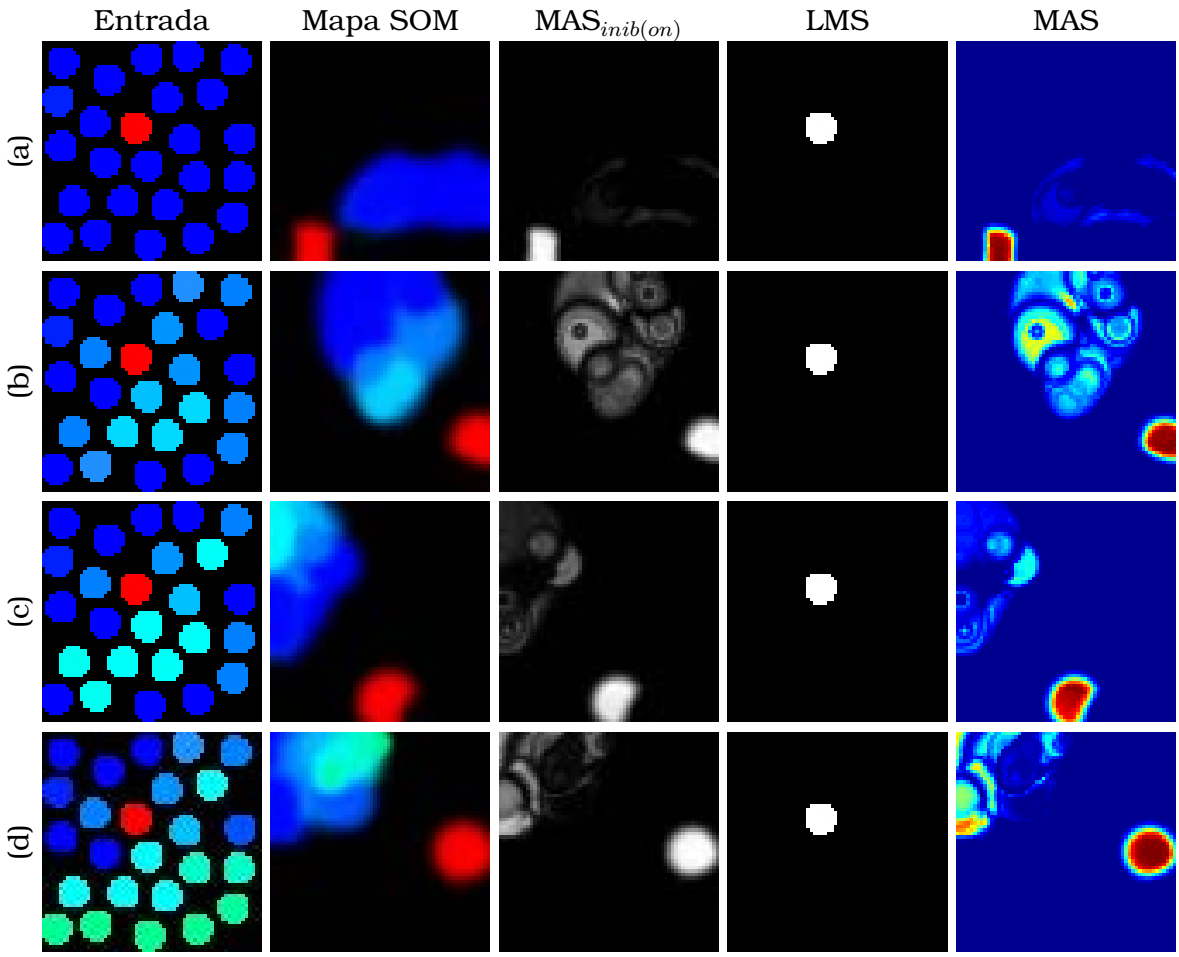


Figura 5.10: Localização de Objetos Salientes II. Simulação variando a heterogeneidade dos distratores. Valores de parâmetros utilizados nas simulações: sincronização ($\alpha_s = 0.6$ e $I = 1.1$), rede SOM ($\alpha_k = 0.2$, $\sigma = 26$ e $n_{it} = 5000$). Imagens com 64 x 64 *pixels* utilizadas em (Quiles et al., 2009).

partir da atenção visual humana.

5.2 Atenção *Top-Down* e *Bottom-UP*

Existem situações em que pode ser útil a presença de sistemas aptos à identificar alvos específicos e também demais objetos salientes, assim como suas localizações na cena, de maneira autônoma. Por exemplo, a navegação de um robô pode necessitar tanto de informações *top-down*, utilizadas para a detecção de pontos de referência e sinalizações em geral, quanto à detecção *bottom-up* de obstáculos inesperados (Navalpakkam and Itti, 2006a). Direcionar a atenção visual baseada em características específicas e informações primitivas de um objeto em um cena não é uma tarefa trivial para os modelos de atenção visual. De acordo com Oliva et al. (2003); Navalpakkam and Itti (2006a); Frintrop (2006); Borji and Itti (2013), diversos trabalhos têm sido desenvolvidos baseados na integração dos mecanismos de atenção *top-down* e *bottom-up*. Conforme apresentado no Capítulo 4, diversos modelos

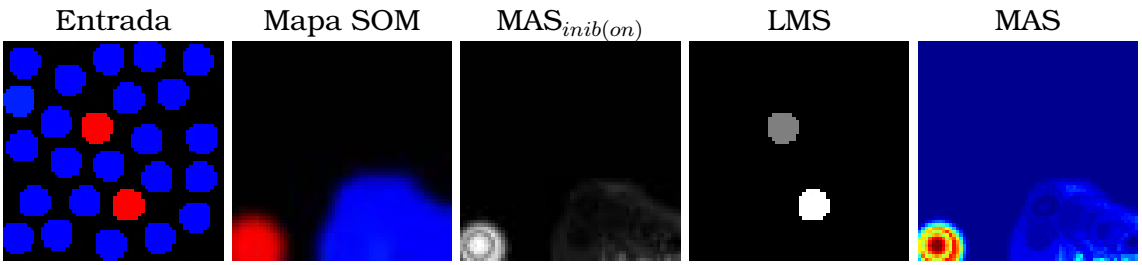


Figura 5.11: Localização de Objetos Salientes II. Dois objetos salientes. Valores de parâmetros utilizados de acordo com simulações apresentadas na Figura 5.10.

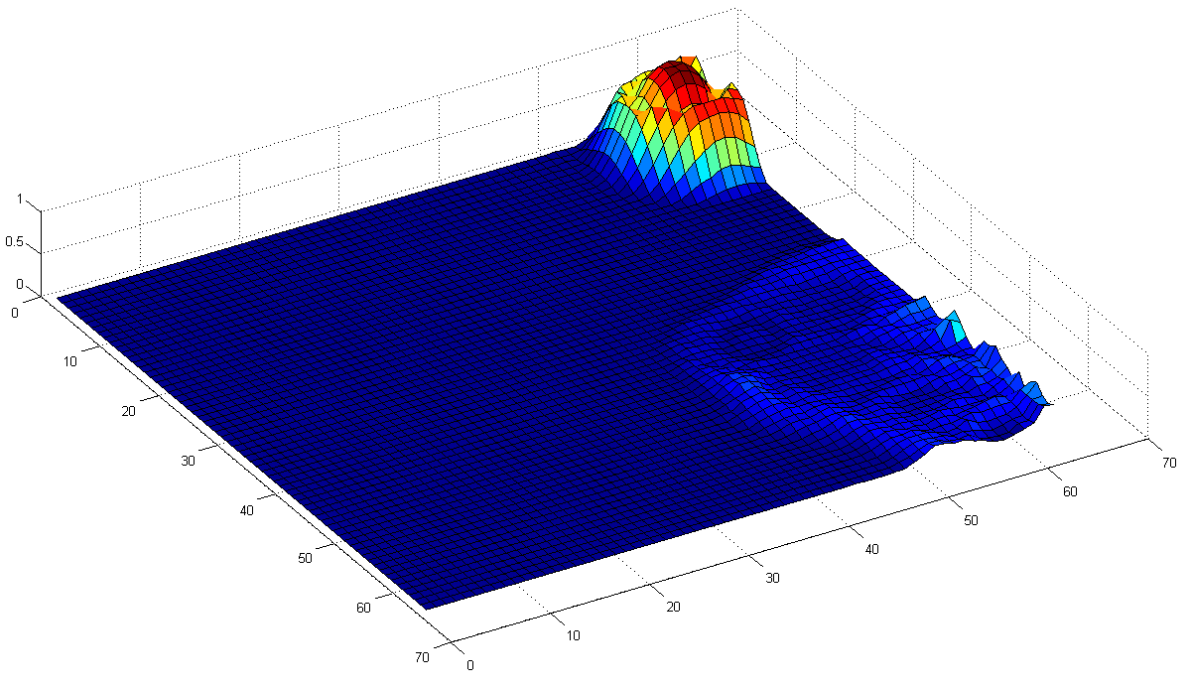


Figura 5.12: Localização de Objetos Salientes II. Regiões salientes sobrepostas.

de atenção visual têm sido propostos, entretanto, direcionados à atenção *bottom-up* ou à atenção *top-down*, de maneira isolada. Embora muitos modelos utilizem mapas de saliência que, como mencionado, é gerado a partir de informações primitivas da cena, são classificados como modelos puramente *top-down*. Isto pode ser visto na Tabela 4.1, referente aos modelos com enviesamento *top-down*. Além disso, devido ao enviesamento *top-down* do mapa de saliência, a partir de informações conhecidas sobre o alvo, outras regiões do mapa são excluídas automaticamente.

Nesta pesquisa, propomos um modelo de atenção *bottom-up* e *top-down* para a localização de objetos na cena. A competição pela atenção visual é definida pelo valor de reconhecimento do objeto e por suas características primitivas. A principal contribuição deste trabalho está relacionada à proposta de um modelo de atenção visual baseado na correlação temporal, de forma que os neurônio de segmentos sincronizados sejam previamente reconhecidos como objetos, onde, um fator de reconhecimento será também considerado como uma características para guiar a atenção visual.

O mapa de atributo-saliência, responsável pela competição por atenção, também representa um importante papel para a integração de características primitivas (córtex visual) e características cognitivas (córtex pré-frontal e memória.). O que recebe um forte embasamento biológico, pois, de acordo com Corbetta (1998) e Frintrop et al. (2010), a atenção define a capacidade mental para selecionar estímulos, respostas, memórias, ou pensamentos que são comportamentalmente relevantes, entre os muitos outros que são comportamentalmente irrelevantes. Características relacionadas à auto-organização de estímulos são mantidas neste trabalho, uma vez que estudos comprovam a organização biológica do cérebro em relação à estímulos visuais e a memória associativa (Kohonen, 2001).

5.2.1 Atenção *Top-Down* e *Bottom-UP* em Cenas Sintéticas

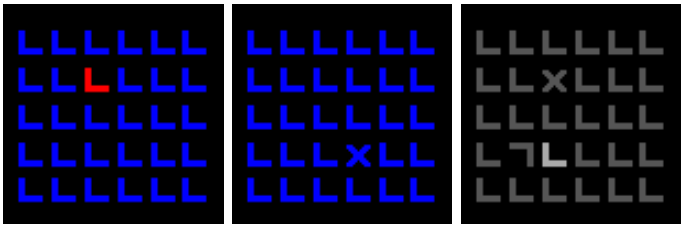
O primeiro modelo de atenção visual *top-down* e *bottom-up* proposto foi composto por quatro componentes. Primeiro, para o treinamento e reconhecimento de objetos (*top-down*), uma MLP (*multilayer perceptron*) é utilizada. Segundo, uma rede neural pulsada é utilizada para a segmentação dos objetos da cena. Esta etapa é também responsável pela classificação dos objetos segmentados, que será utilizada para definir a saliência do objeto. Terceiro, um mapa SOM é utilizado para organizar as informações *top-down* (memória associativa) e *bottom-up* (estímulos visuais) em um único mapa, de forma que características similares sejam mapeadas em localizações próximas no mapa. Quarto, uma rede neural com dois tipos de conexões: excitatórias e inibitórias, é utilizada para gerar o mapa de atributo-saliência e localizar o objeto saliente. Os resultados deste trabalho encontram-se publicados em (Benicasa et al., 2012). A seguir apresentaremos a descrição detalhada do modelo e as simulações realizadas.

Descrição do Modelo

Para Itti and Koch (2001), o primeiro estágio de processamento nos modelos de atenção *bottom-up* é a extração das características primitivas da cena, onde neurônios de estágios prévios da visão são sintonizados para simples atributos visuais. Sendo assim, este será a primeira etapa considerada nesta proposta.

Conforme apresentado na Seção 2.4, e também de acordo com Wolfe and Horowitz (2004), o ponto inicial para o entendimento do desenvolvimento da atenção visual está diretamente relacionado aos atributos visuais utilizados para guiar a atenção, podendo tornar a busca visual uma tarefa rápida e eficiente. Na Figura 5.13 (a) é possível identificar o objeto saliente facilmente, devido sua cor contrastante em relação aos demais objetos. Isto também ocorre na Figura 5.13 (b), de forma que a atenção seja guiada pela diferença na orientação dos objetos da cena. A diferença na característica intensidade entre os objetos da cena também pode guiar a atenção (Figura 5.13 (c)). Com base nos trabalhos revisados (Tabela 4.1), utilizaremos

nesta proposta os seguintes atributos primitivos para guiar a atenção visual: canais de cores, conforme propostas anteriores apresentadas neste capítulo, contraste de intensidades, cores oponentes e orientação local.



(a) (b) (c)

Figura 5.13: Exemplo de características contrastantes. (a) Cor, b) Orientação e (c) Intensidade.

O contraste de intensidades (Figura 5.13 (c)) é a diferença espacial relacionada à intensidade de luminosidade da cena (Itti and Koch, 2001). Com r, g e b sendo, respectivamente, os canais de cores vermelho, verde e azul de uma cena, (Figura 5.14 (a)), o mapa de contraste de intensidades I é definido pela seguinte equação:

$$I = \frac{r + g + b}{3} \tag{5.15}$$

Conforme apresentado na Seção 2.4, a partir do centro do campo visual, neurônios das áreas do córtex visual V1 e V2 apresentam maiores estímulos quando submetidos às cores vermelho/verde e azul/amarelo. Nesta proposta, baseada em (Itti et al., 1998), consideramos os mapas de diferenças espaciais em cores RG (*vermelho/verde*) e BY (azul/amarelo) (veja Figura 5.14 (c) e (d)), como informações para o desenvolvimento da atenção visual, definidos pelas seguintes equações:

$$RG = \frac{r - g}{\max(r, g, b)}, \tag{5.16}$$

e

$$BY = \frac{b - \min(r, g)}{\max(r, g, b)} \tag{5.17}$$

A orientação local (Figura 5.14 (e), (f), (g) e (h)) é obtida através da aplicação do filtro espacial (w_θ) de dimensão `3x3`, a partir do mapa de contraste de intensidades I.

$$w_0 = \begin{bmatrix} -1 & -1 & -1 \\ 2 & 2 & 2 \\ -1 & -1 & -1 \end{bmatrix} \quad w_{90} = \begin{bmatrix} -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \end{bmatrix} \tag{5.18}$$

$$w_{45} = \begin{bmatrix} -1 & -1 & 2 \\ -1 & 2 & -1 \\ 2 & -1 & -1 \end{bmatrix} \quad w_{135} = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}, \tag{5.19}$$

onde quatro orientações (O_θ) são usadas com $\theta \in \{0°, 45°, 90°, 135°\}$, definidos como segue:

$$O_\theta(x, y) = \sum_{s=-1}^{1} \sum_{t=-1}^{1} w_\theta(s+1, t+1) I(x+s, y+t) \tag{5.20}$$

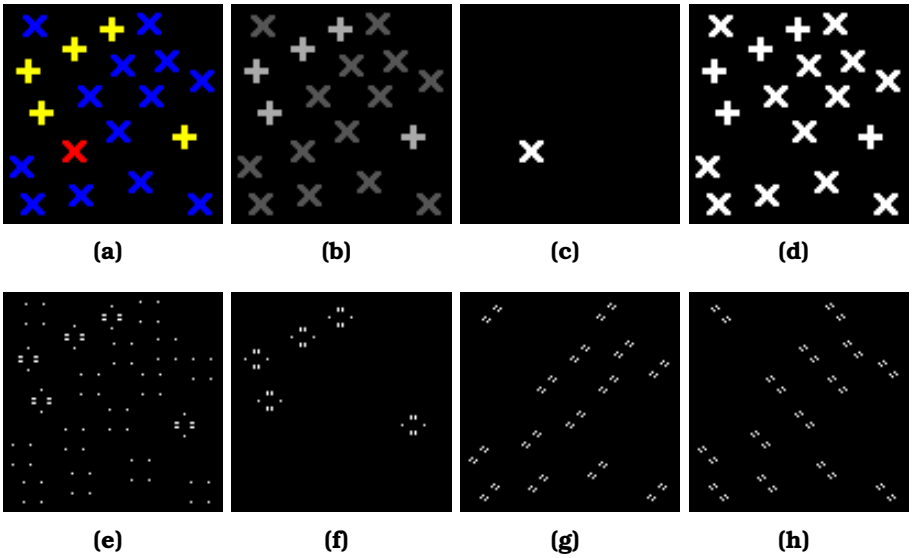


Figura 5.14: Exemplo de extração de características primitivas. (a) Imagem de Entrada, (b) Mapa de Intensidades, (c) Mapa de Cores Oponentes RG, (d) Mapa de Cores Oponentes BY e os Mapa de Orientações: (e) O_0, (f) O_{90}, (g) O_{45} e (h) O_{135}.

Como mencionado inicialmente nesta proposta, utilizamos uma MLP para o treinamento e reconhecimento de objetos, obtendo assim a informação *top-down* considerada neste modelo. Devido à MLP ser um modelo de rede neural amplamente conhecido, optamos por não descrevê-la em detalhes (para detalhes sobre este modelos veja Haykin (2001)).

De modo geral, para possibilitar ao modelo proposto direcionar a atenção visual para objetos conhecidos, a MLP deve ser treinada previamente. Os padrões para o treinamento da MLP considerados nesta proposta foram imagens de objetos sintéticos com dimensões de `16x16` *pixels*, o que define a dimensionalidade do vetor de características de entrada em $d = 256$. O número de neurônios da camada de entrada é igual a d, e o número de neurônios da camada de saída é definido pelo número de classes que o modelo de atenção deverá reconhecer. A Figura 5.15 apresenta alguns exemplos de objetos a serem reconhecidos pelo modelo de atenção.

Podendo ser considerado um modelo evolutivo às propostas anteriores apresentadas neste capítulo, o modelo de rede neural pulsada utilizada nesta proposta

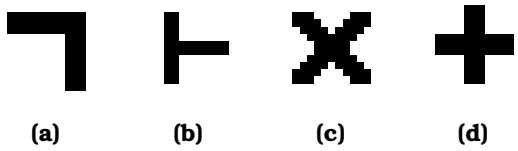


Figura 5.15: Exemplos de objetos para treinamento.

para a segmentação da cena também baseia-se na sincronização de osciladores do tipo I&F proposto por Campbell et al. (1999), definida pelas Equações 5.1 e 5.2. Consideramos nesta proposta o peso de acoplamento entre dois vizinhos osciladores Z_i definido de acordo com Wang and Terman (1997) e Quiles et al. (2011), de modo que o valor de similaridade entre dois osciladores possa ser definido tanto para imagens em tons de cinza, quanto para imagens coloridas, de acordo, respectivamente, com as seguintes equações:

$$W_{ij} = I_M/(1 + |I_i - I_j|), \tag{5.21}$$

$$W_{ij} = I_M/\left(1 + \sum_{h \in \{r,g,b\}} |h_i - h_j|\right), \tag{5.22}$$

onde I_M é o máximo valor dos canais I, r, g, e b. Nesta proposta, $I_M = 255$, I_i representa o valor do mapa de intensidades do neurônio i, h_i e h_j representam os valores dos canais de cores r, g e b dos osciladores vizinhos i e j.

Com os conceitos apresentados sobre sincronização e do uso da similaridade entre osciladores vizinhos, foi possível obter cada objeto da cena representado por uma rajada de pulso em sincronia. Com isto, um importante ponto desta proposta ocorre quanto todos os objetos estão segmentados e sincronizados. Considerando o dinamismo da rede de osciladores I&F, quando um objeto (grupo de neurônios sincronizados) pulsa, este será avaliado pela MLP. O valor de saída da rede é usada para a alimentação do atributo referente ao valor de reconhecimento R (Figura 5.16). Inicialmente, $R = 0$ para todos os neurônios. Após todos os objetos sincronizados terem pulsado pelo menos uma vez, todos os neurônios relacionados a cada objeto pulsante serão associados a um valor de reconhecimento, de forma a serem mapeados em posições próximas no mapa SOM.

Para a organização dos atributos em um único mapa bidimensional, também será considerado neste trabalho a utilização do SOM. Conforme mencionado, e de acordo com Haykin (2001), o principal objetivo do SOM é a transformação de um padrão de entrada, de dimensão arbitrária, em um mapa de uma ou duas dimensões, topologicamente organizado. Nesta proposta, o som foi criado a partir de um conjunto de treinamento constituído por padrões com η dimensões: canais de cores r, g e b, contraste de intensidades I, diferença espacial em cores RG e BY, orientações

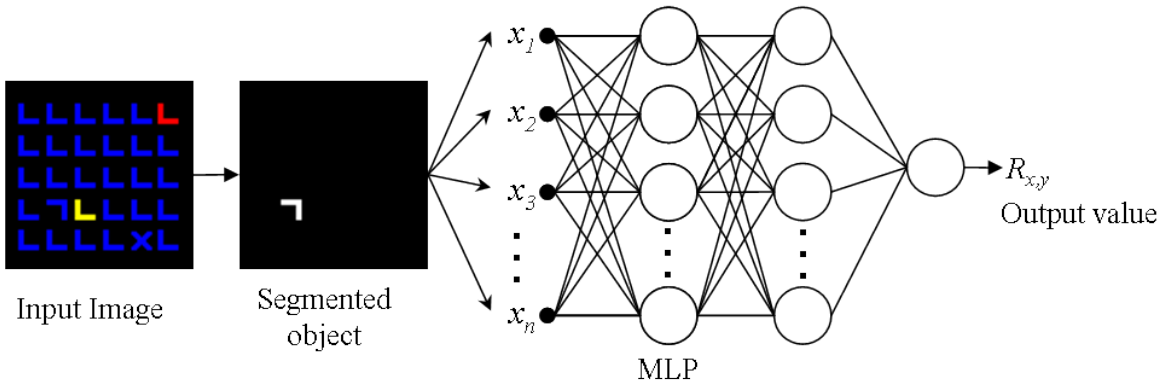


Figura 5.16: Segmentação e valor de reconhecimento.

O_θ com $\theta \in \{0°, 45°, 90°, 135°\}$, posições $[x, y]$ em um plano bidimensional de cada pixel da cena e, finalmente, o valor de reconhecimento R de cada pixel previamente reconhecido pela MLP. É importante notar que, a MLP tem como função classificar segmentos, composto por neurônios sincronizados, de forma que o valor de reconhecimento do segmento será atribuído igualmente os neurônios que o compõe.

Consideramos então $p = [r, g, b, I,\ RG,\ BY,\ O_0,\ O_{45},\ O_{90},\ O_{135},\ x,\ y,\ R]$ o padrão utilizado para o treinamento do SOM. Assim como na proposta apresentada na Seção 5.1.3, os padrões utilizados para o treinamento do SOM são formados por características dos *pixels* presentes na própria cena, com exceção do atributo R, obtido através da MLP, conforme descrito. Na Figura 5.17 é apresentado o SOM gerado a partir da imagem de entrada. A equação que descrevem o processo de geração do mapa SOM, de modo a considerar as novas características de p, é descrita como segue:

$$d(p, m_s) = \| p - m_s \| = \sqrt{\sum_{j=1}^{13} (p_j - m_{sj} -)^2}, \quad j \in \{r, g, b, I, RG, BY, O_0, O_{45}, O_{90}, O_{135}, x, y, R\}, \tag{5.23}$$

onde os neurônios m_s do SOM são atualizados de acordo com a Equação 5.6. Detalhes sobre os parâmetros utilizados encontram-se descritos na Seção 3.3.

De maneira similar aos modelos anteriores, o mapa de atributo-saliência foi obtido a partir de uma rede de neurônios composto por conexões excitatórias e inibitórias, definido pela seguinte Equação 5.10, reescrita a seguir:

$$\dot{v}_i = -v_i + E_i - W_Y Y_i, \qquad i = 1, .., n, \tag{5.24}$$

porém, os termos de acoplamento excitatórios E_i e inibitórios Y_i sofrem agora a influência do parâmetro W_j, associado a cada característica. É importante notar que, com o parâmetro W_j é possível realizar o ajuste dos pesos, de acordo com as características que se deseja salientar da cena. Considerando $l_k = [r, g, b, I,\ RG,\ BY,$

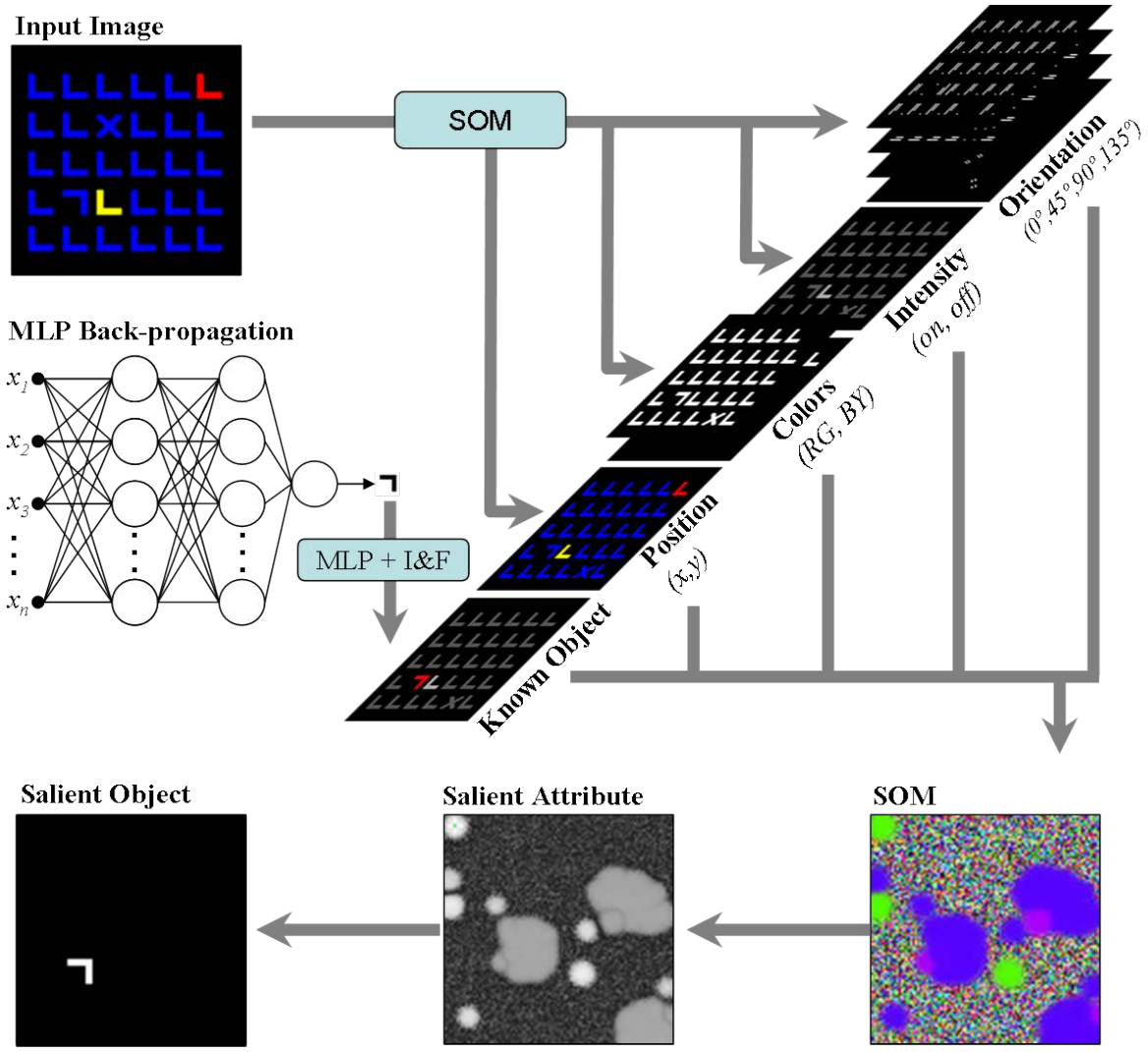


Figura 5.17: Diagrama do modelo de atenção *top-down* e *bottom-up*.

O_0, O_{45}, O_{90}, O_{135}, x, y, R] um padrão pertencente a um segmento ativo na rede I&F, a medida de similaridade entre o padrão l_k e os neurônios m_s do SOM foi alterada para considerar os demais atributos, descrito como segue:

$$d(l_k, m_s) = \| l_k - m_s \| = \sqrt{\sum_{j=1}^{13} W_j (l_{kj} - m_{sj})^2}, \quad j \in \{r, g, b, I, RG, BY, O_0, O_{45}, O_{90}, O_{135}, x, y, R\}, \tag{5.25}$$

onde os termos de acoplamento excitatório E_i e inibitório Y_i são definidos pela Equação 5.12.

Para identificar o objeto saliente, consideramos o termo v_i da Equação 5.25, que representa o potencial de saliência do neurônio i e, baseando-se nos valores de v_i, podemos identificar os próximos objetos salientes da cena.

Pode-se concluir que, com a proposta do parâmetro W_j para a geração do mapa de atributo-saliência, esta pesquisa apresenta uma importante evolução em relação às propostas anteriores, possibilitando o enviesamento *top-down* baseado no

espaço. Por outro lado, como a proposta de um mecanismo para a classificação dos segmentos sincronizados, possibilitamos também o enviesamento *top-down* baseado em objetos. A seguir serão apresentadas simulações para demonstrar o comportamento do modelo proposto.

Simulações Computacionais

Com o objetivo de analisar o comportamento do modelo de atenção *bottom-up / top-down* proposto, foram utilizadas como entrada para o modelo, imagens sintéticas (`100x100` *pixels*), representado as características consideradas neste trabalho para o desenvolvimento da atenção visual. Inicialmente, características são simuladas individualmente, com o objetivo de analisar o comportamento do modelo em condições específicas e, em seguida, apresentamos simulações envolvendo demais características, distratores e também o comportamento do modelo com variações do parâmetro de enviesamento *top-down* W_j. Para todas as simulações, os objetos conhecidos previamente, ou seja, treinados previamente pela MLP, são apresentados na Figura 5.18.

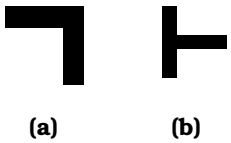


Figura 5.18: Objetos conhecidos.

De maneira geral, os valores de parâmetros utilizados para as simulações foram: para a sincronização e segmentação da cena, $\alpha = 0.7$ e $I = 1.1$; para geração do mapa de atributo saliência, $W_j = 1$. Contudo, para demonstrar o enviesamento *top-down* do modelo, o parâmetro W_j poderá ser modificado para fins comparativos.

Inicialmente, na simulação apresentada na Figura 5.19, consideramos uma cena contendo um objeto saliente (objeto "vermelho"), com alto contraste em relação aos demais objetos. O SOM, gerado a partir de treze dimensões da imagem de entrada é mostrado em (b). Em (c) e (e) são apresentados os mapas de atributo-saliência com os principais pontos de atenção, gerados a partir da combinação da RNP e do mapa SOM. Ainda na Figura 5.19, em (d) é apresentada a localização do objeto mais saliente. É importante notar que, nesta simulação, os atributos que podem guiar a atenção para o objeto mais saliente estão representadas nos mapas (f) e (j). Os demais mapas, devido sua homogeneidade, não influenciam a entrega da atenção. Por exemplo, no mapa de contraste de intensidades I, os neurônios são ativados igualmente por toda extensão do mapa. Neste caso, o termo de acoplamento inibitório Y_i da Equação (5.25) faz com que objetos com características similares sejam mutualmente inibidos, devido a dinâmica proposta relacionada à competição pela atenção. O mesmo ocorre com os mapas apresentados em (g), (h), (k), (l), (m),

(n), (o) e (p).

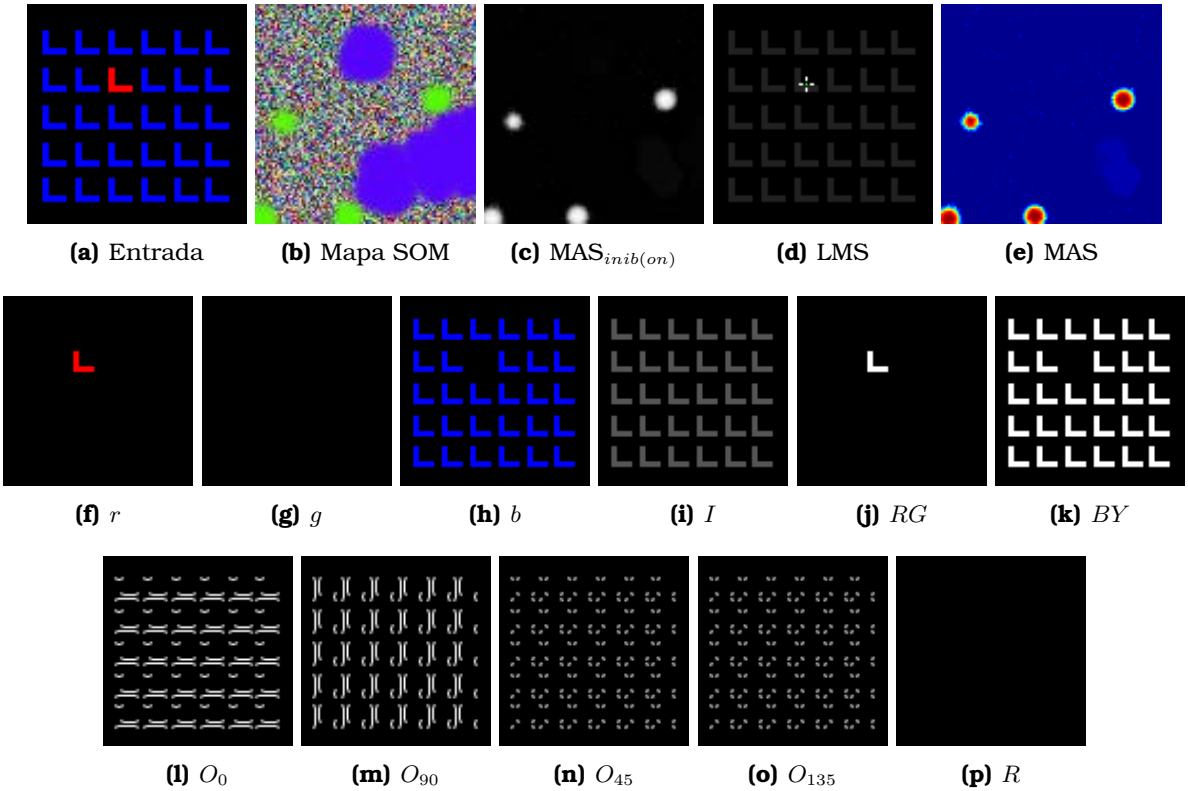


Figura 5.19: Modelo de atenção *top-down* e *bottom-up*. Simulação 1 - Contraste em Cores. (a) Imagem de entrada, (b) Mapa SOM, (c) Mapa de atributo-saliente com inibidor ativo, (d) Local de maior saliência, (e) Mapa de atributo-saliente com inibidor ativo com ênfase nas regiões salientes, (f) Canal red, (g) Canal $green$, (h) Canal $blue$, (i) Contraste de intensidades, (j) Cores oponentes $Red - Green$ e (k) $Blue - Yellow$, (l) Orientações O_0, (m) O_{90}, (n) O_{45} e (o) O_{135}, e (p) Reconhecimento dos objetos.

Na Figura 5.20 (a) é apresentada a cena contendo um objeto saliente que *pop-out*, devido ao contraste da característica orientação. Isto pode ser visto nos mapas (h) e (i) da mesma figura. Neste caso, as orientações O_0 e O_{90} não influenciam no direcionamento da atenção.

Como objetivo de identificar o objeto saliente na Figura 5.21 (a), faz-se necessário observar duas características: a diferença de cores e orientações entre os objetos. De acordo com Treisman and Gelade (1980) e Miller (2000), para a identificação do objeto saliente nestes casos, deve-se realizar uma busca conjuntiva por características que possam direcionar a atenção visual, baseado em um controle cognitivo realizado pelo córtex pré-frontal. De maneira elucidativa, de acordo com o modelo proposto, podemos descrever este processo da seguinte maneira. Quando um objeto pulsa na RNP, por exemplo uma "cruz vermelha", verifica-se qual a região SOM que melhor possa representá-la e, em seguida, está região é ativada no mapa de atributo-saliência. Objetos semelhantes à "cruz vermelha", provavelmente serão mapeados para a mesma região do SOM e, consequentemente, ativarão ainda mais a região do mapa de atributo-saliência. No caso do "X vermelho", embora este

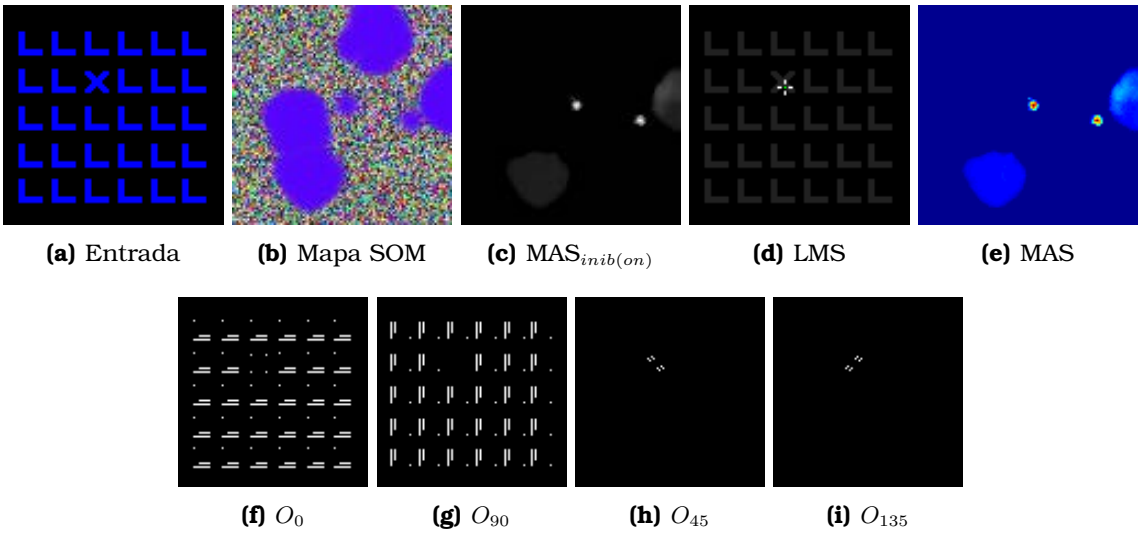


Figura 5.20: Modelo de atenção *top-down* e *bottom-up*. Simulação 2 - Contraste em orientações.

apresente a mesma cor do objeto "cruz vermelha", suas características relacionadas à orientação o levará a ser representado por outra região do SOM. Por outro lado, o acoplamento inibitório faz com que, quanto mais uma região for ativada, maior será sua inibição (veja Equação (5.25)). Assim, como existe somente um "X vermelho", este sofrerá o efeito da inibição gerado somente por si próprio, enquanto os demais objetos serão mutuamente inibidos. Entre as regiões salientes apresentadas no MAS ((c) e (e)), as duas regiões menores representam as características conjuntivas da busca realizada nesta simulação.

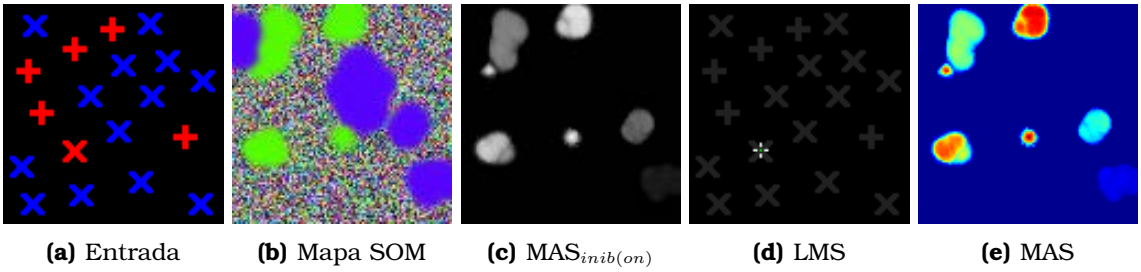
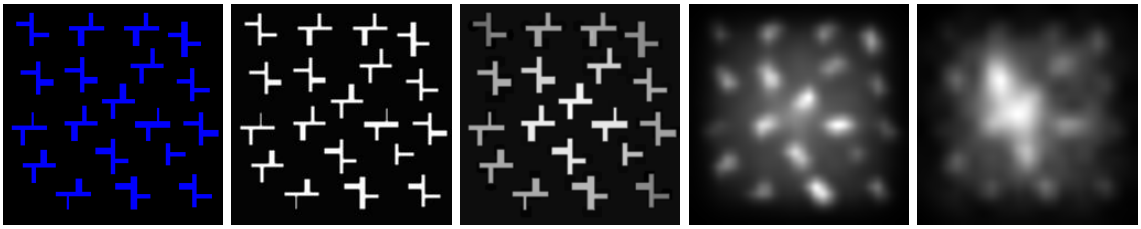


Figura 5.21: Modelo de atenção *top-down* e *bottom-up*. Simulação 3 - Busca conjuntiva baseada na cor e orientação.

De acordo com um dos pontos de investigação desta tese, na simulação seguinte, propomos uma cena onde não existam características *bottom-up* salientes relacionadas aos objetos presentes. Na Figura 5.22 (a) é apresentada a cena proposta contendo um objeto conhecido (Figura 5.18 (b)). Para demonstrar a saliência *bottom-up* nula deste objeto em relação aos distratores, são apresentados na Figura 5.22 (b), (c), (d) e (e), os mapas de saliência gerados, respectivamente, de acordo com os modelos propostos por Achanta et al. (2009) (AC), Cheng et al. (2011) (CH), Harel et al. (2006) (GBVS) e Itti et al. (1998) (ITTI). Como pode ser verificado

nos mapas de saliência apresentados, o objeto conhecido não apresenta valor de saliência contrastante com os distratores.



(a) Entrada **(b)** AC **(c)** CH **(d)** GBVS **(e)** ITTI

Figura 5.22: Exemplo de saliência nula de objeto conhecido.

Dando continuidade a esta simulação, a Figura 5.23 (a) apresenta a cena contendo o objeto conhecido que, neste caso, é responsável por guiar a atenção. Como pode ser observado nesta mesma figura, somente o mapa (m) apresenta características contrastantes com os distratores, permitindo a seleção visual. Do ponto de vista biológico, a inspiração para esta simulação foi obtida a partir dos experimentos apresentados na Seção 2.3 (Cap. 2) onde, de acordo com Theeuwes (1992), Wolfe (1994), Egeth and Yantis (1997) e Ogawa and Komatsu (2004), a atenção visual também deve ser baseada na busca por um alvo específico, com características conhecidas previamente pelos observadores, ou seja, uma busca a partir de um enviesamento baseado em informações prévias. Entretanto, demais regiões salientes continuam representadas no mapa de atributo-saliência, como pode ser visto na Figura 5.23 (c) ou (e).

Nas simulações apresentadas anteriormente nesta seção, foram utilizados os valores do parâmetro $W_j = 1.0$ para $j \in [1 \ldots 13]$. Entretanto, dependendo do tipo de informação contida na cena, e na diversidade de objetos salientes, é também possível realizar o enviesamento *top-down* através do parâmetro W_j (Equação (5.25)), de forma a associar pesos à características desejadas. A Figura 5.24 (Entrada) apresenta uma cena contendo vários objetos salientes e, para as simulações apresentadas nas linhas (a), (b), (c), (d), (e) e (f), o parâmetro W_j é ajustado de acordo com o valor informado na primeira coluna da Figura 5.24. É importante notar que, mediante variações do parâmetro W_j, regiões específicas do mapa de atributo saliência ($\text{MAS}_{inib(on)}$ e MAS) são ativados, permitindo a seleção visual de objetos específicos, apresentados em LMS.

5.2.2 Atenção *Top-Down* e *Bottom-UP* em Cenas Reais

Considerando as limitações apresentadas nas propostas anteriores, relacionadas ao tratamento de imagens reais e, como uma evolução da proposta apresentada na Seção 5.2.1, propomos aqui a utilização de uma rede LEGION (*Locally Excitatory Globally Inhibitory Oscillator Networks*) para a módulo responsável por segmentar da cena.

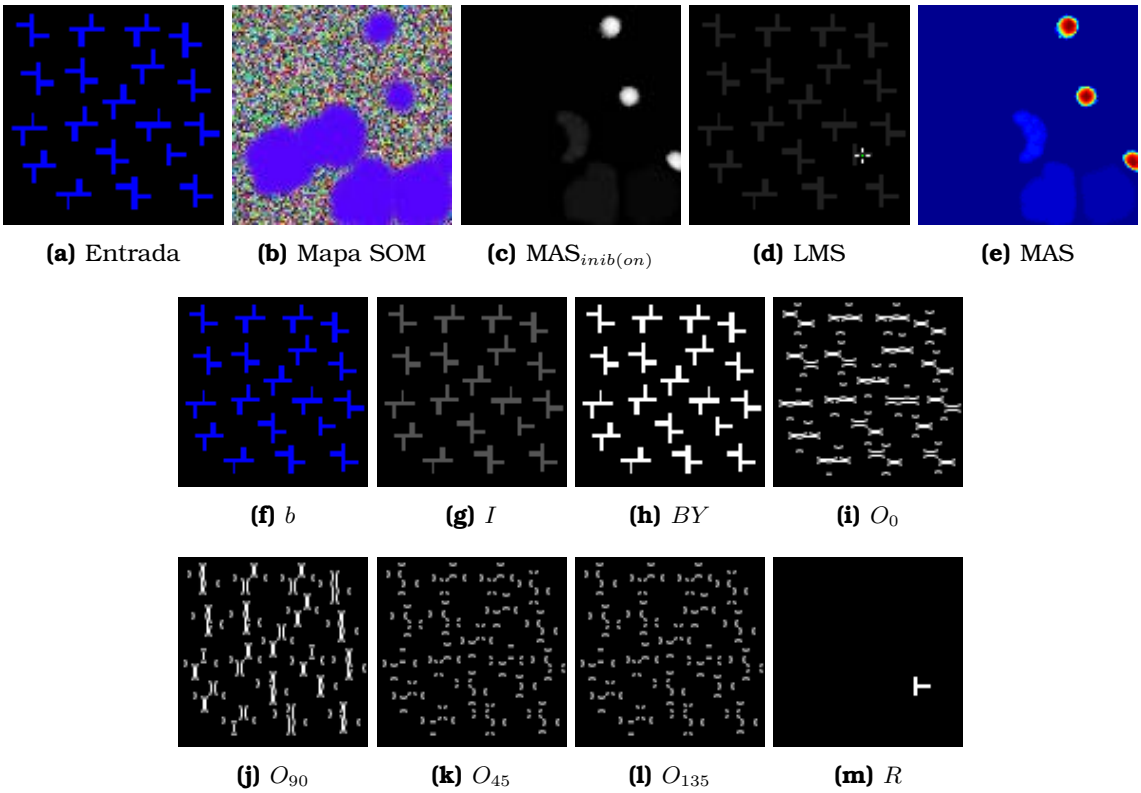


Figura 5.23: Modelo de atenção *top-down* e *bottom-up*. Simulação 4.

Diversos modelos de segmentação de imagens podem ser encontrados na literatura. Entretanto, baseado na teoria da correlação temporal, o modelo LEGION de Wang and Terman (1995) têm sido utilizado em diversos tipos de aplicações envolvendo tarefas de segmentação de imagens (Shareef et al., 1999; Liu et al., 2001; Quiles et al., 2011), se demonstrando um modelo adequado para este tipo de tarefa. Na Seção 3.2.3 apresentamos a descrição detalhada do modelo LEGION (Wang and Terman, 1995) e sua evolução algorítmica para segmentação de imagens reais, proposta por Wang and Terman (1997).

Baseado na arquitetura do modelo apresentado na Seção 5.2.1, este trabalho é composto pelos seguintes componentes: uma MLP, para o treinamento e reconhecimento de objetos; uma rede LEGION, para a segmentação dos objetos da cena; uma rede SOM, utilizada para o mapeamento bidimensional das características utilizadas para guiar a atenção visual; e finalmente, o mapa de atributo-saliência, responsável por destacar os objetos mais saliente na cena. Os resultados deste trabalho foram publicados em (Benicasa et al., 012b). A seguir apresentaremos a descrição do modelo e simulações realizadas a partir de imagens reais.

Descrição do Modelo

De acordo com o diagrama do deste modelo aqui proposto, apresentado na Figura 5.25, seu fluxo de informações pode ser descrito como segue. Inicialmente a cena é apresentada, paralelamente, para o módulo responsável pela extração das

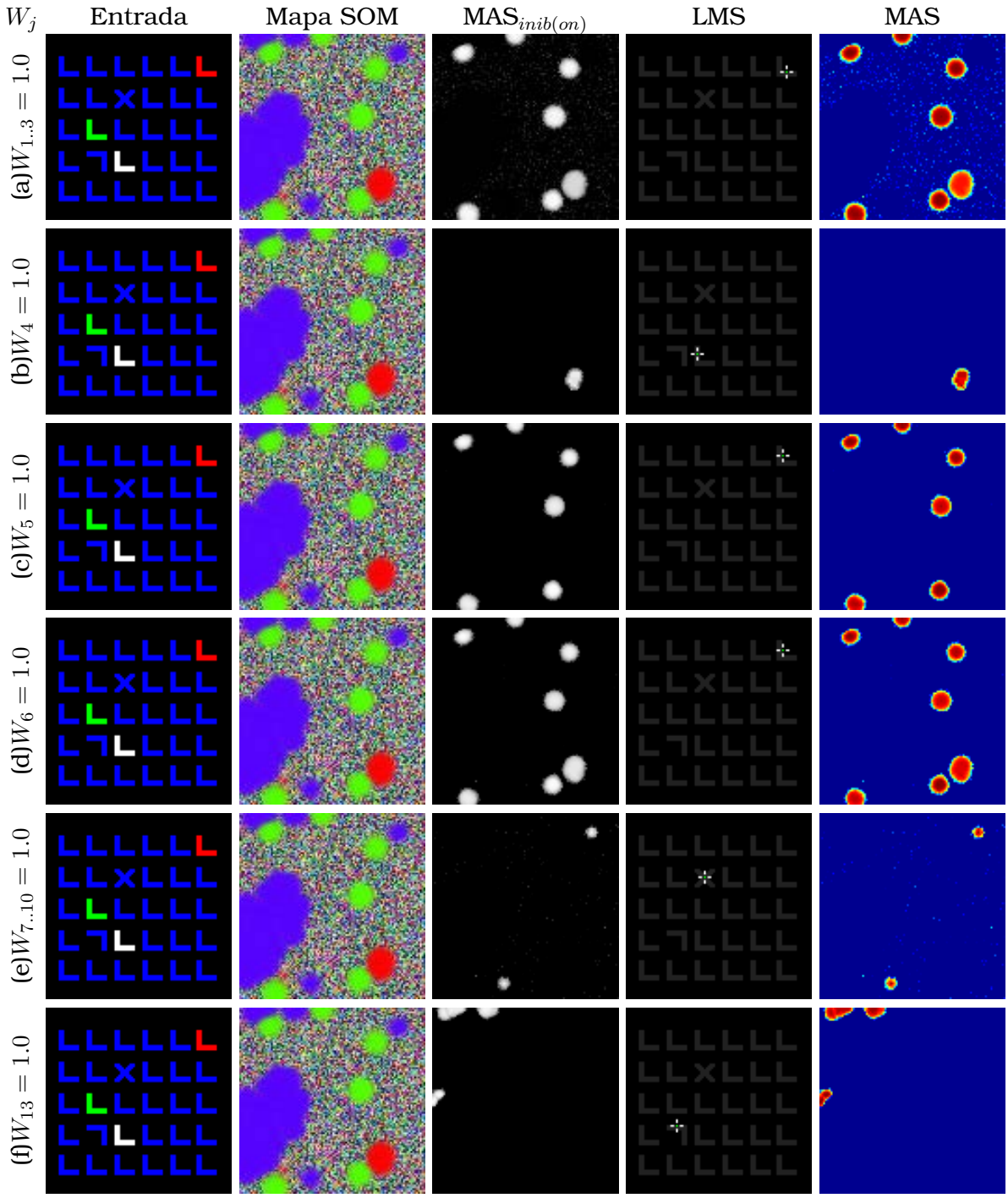


Figura 5.24: Modelo de atenção *top-down* e *bottom-up*. Simulação 5. Modulações do parâmetro W_j para o enviesamento *top-down* de características desejadas. O valor do parâmetro W_j encontra-se na primeira coluna. Para todas as simulações, foi utilizado $W_j = 0.0$ para todo j não informado, com exceção de $W_{11,12} = 1.0$.

características visuais primitivas da cena e para a rede LEGION. A saída gerada por estes módulos servirá de alimentação para os seguintes componentes: a rede MLP, para o reconhecimento do objeto e a rede SOM, responsável pela geração do mapa de atributo saliência, que tornará possível a localização dos objetos salientes.

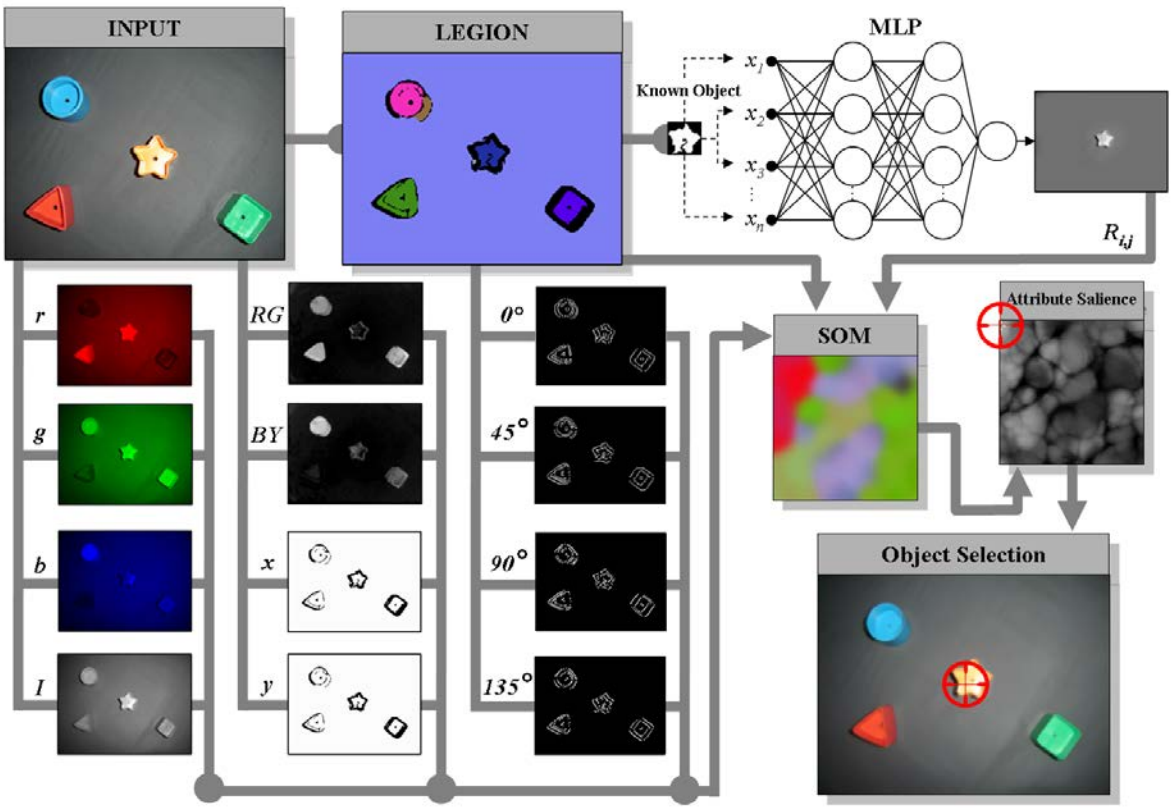


Figura 5.25: Diagrama do modelo de atenção *top-down* e *bottom-up* II.

A extração das características primitivas ocorre de acordo com os atributos visuais: intensidades (I), cores (r, g e b), cores oponentes (RG e BY) e orientações ($O_\theta \in \{0°, 45°, 90°, 135°\}$), definidas na Seção 5.2.1 pelas Equações (5.15–5.20).

Dado a proposta deste modelo em considerar a entrega da atenção baseada em objetos, um mecanismo de segmentação faz-se necessário. O processo de segmentação transforma uma imagem em um conjunto de segmentos que podem ser interpretados como objetos primitivos da cena. A segmentação da cena neste trabalho é desenvolvida por uma rede LEGION (Wang and Terman, 1997). A unidade básica da LEGION é um oscilador de van der Pol (1926), definido pela conectividade recíproca entre uma variável excitatória x_i e uma variável inibitória y_i (Wang and Terman, 1995):

$$\dot{x}_i = 3x_i - x_i^3 + 2 - y_i + I_i + S_i + \rho \tag{5.26}$$

$$\dot{y}_i = \epsilon \left(\gamma \left(1 + \tanh \left(\frac{x_i}{\beta} \right) \right) - y_i \right), \tag{5.27}$$

onde I_i representa um estímulo externo, S_i representa o acoplamento a partir dos osciladores vizinhos na rede e ρ é um ruído Gaussiano de pequena amplitude. ϵ, γ e β são parâmetros do oscilador.

O modelo LEGION proposto em (Wang and Terman, 1995), é apto à segmentação somente de imagens sintéticas. Com o objeto de realizar a segmentação de imagens reais, um termo de potencial lateral foi introduzido em (Wang and Terman, 1997). Este termo é utilizado para descriminar segmentos maiores de fragmentos de ruídos presentes na cena. De maneira geral, este mecanismo pode ser descrito como segue. Se um oscilador i está localizado no centro de um segmento, ou seja, em uma região homogênea da imagem, este oscilador é apto a receber um alto potencial lateral de seus vizinhos, podendo tornar-se o "líder" do segmento. Caso contrário, se um oscilador i representa um *pixel* isolado, este não receberá potenciais laterais de seus vizinhos e, portanto, não se tornará líder. Assim, o processo de segmentação é conduzido de forma que, somente segmentos que possuam pelo menos um líder, possam oscilar. A descrição do modelo LEGION para segmentação proposto em Wang and Terman (1997) encontra-se na Seção 3.2.3.

Entretanto, de acordo com Wang and Terman (1997), a dinâmica da rede LEGION, baseada em equações diferenciais aplicada à segmentação de imagens reais com um grande número de *pixels*, requer um grande esforço computacional. Como solução para esta limitação, os autores desenvolveram um algoritmo computacional seguindo os principais passos das equações originais. Utilizaremos nesta proposta o módulo de segmentação LEGION baseado no algoritmo proposto por Wang and Terman (1997) (veja Algoritmo 1). Uma importante característica deste algoritmo é a existência de somente dois parâmetros a serem ajustados, sendo eles: W_z, que define o peso do inibidor global e θ_p, limiar responsável pela formação de osciladores líderes com potenciais laterais regidos por θ_p. Na Seção 3.2.3 é apresentado um estudo sobre o comportamento do LEGION considerando variações dos parâmetros W_z e θ_p.

Com o objetivo de analisar o comportamento do algoritmo, na Figura 5.27 são apresentadas 18 simulações para uma imagem de ressonância magnética (MRI - *magnetic resonance imaging*) com variações dos parâmetros W_z e θ_p. Considerando variações de $W_z = [10, 20, 40]$ representadas, respectivamente, nas colunas da esquerda para a direita, pode-se perceber que, com valores altos de W_z torna-se mais difícil o agrupamento de *pixels* em uma região, de forma que o algoritmo demande de regiões mais homogêneas para o agrupamento. Para o parâmetro θ_p, foram consideradas variações entre $500 \leq \theta_p \leq 2000$. A utilização de valores altos para θ_p tendem a dificultar a seleção de líderes, uma vez que o algoritmo necessita de altos valores de potenciais laterais. Neste caso, quanto maior for θ_p, menor será o número de líderes e, consequentemente, menor será o número de regiões (segmentos). O gráfico apresentado na Figura 5.27 apresenta o número de segmentos gerados a

partir das variações de W_z e θ_p. Concluímos que altos valores de W_z (linha vermelha) e baixos baixos valores de θ_p geram uma maior quantidade de segmentos, enquanto que baixos valores de W_z (linha azul) e altos valores de θ_p geram um menor número de segmentos.

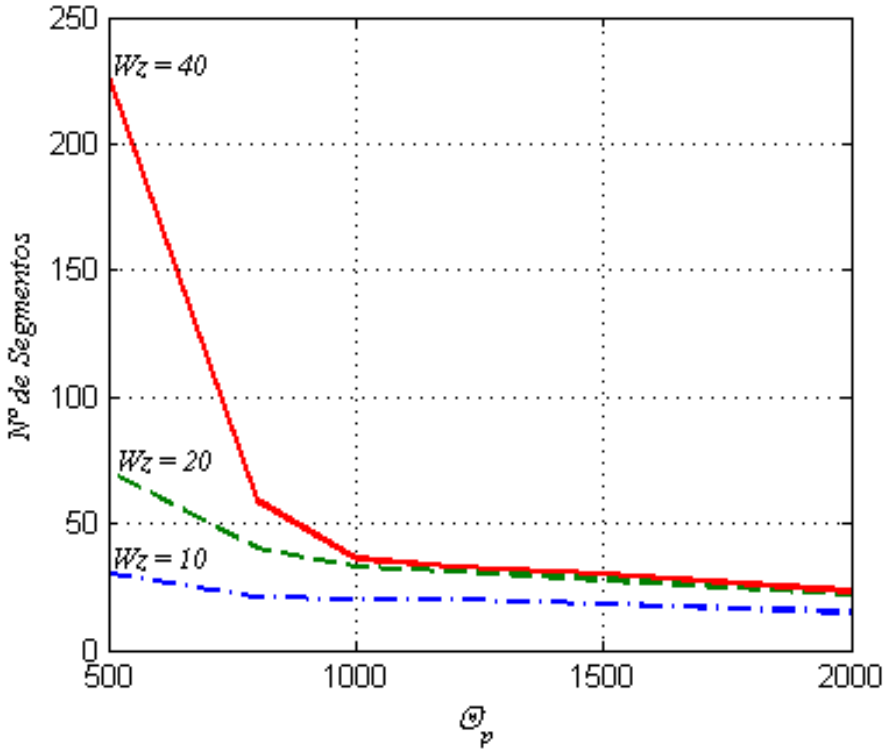


Figura 5.26: Gráfico do comportamento da rede LEGION baseado em variações de W_z e θ_p.

Conforme descrito anteriormente, neste proposta também consideramos a possibilidade de modulações *bottom-up* e *top-down* para o desenvolvimento da atenção visual. Atributos visuais primitivos, ou seja, cores, intensidades, etc, definem a informação *bottom-up* do modelo. Por outro lado, informações sobre objetos previamente conhecidos (modulações *top-down*) são responsáveis por guiar o processo de seleção. Assim, de forma a aplicar este modelo para a seleção de objetos salientes em um cena, a rede MLP deve ser treinada com um conjunto de objetos representando os alvos desejados.

O conjunto de objetos utilizados para o treinamento da MLP é composto por imagens binárias extraídas a partir da cena. Para isto, uma cena contendo o objeto alvo é apresentada à rede LEGION, que retornará os objetos segmentados. Estes objetos são manualmente rotulados e utilizados para a definição dos pesos da MLP. A Figura 5.28 apresenta um conjunto de objetos que podem ser reconhecidos pela MLP, e assim modular o processo de seleção do objeto.

Após o processo de treinamento, a MLP está apta a reconhecer o conjunto de segmentos (objetos). O valor de saída da MLP indica se o objeto está entre àqueles

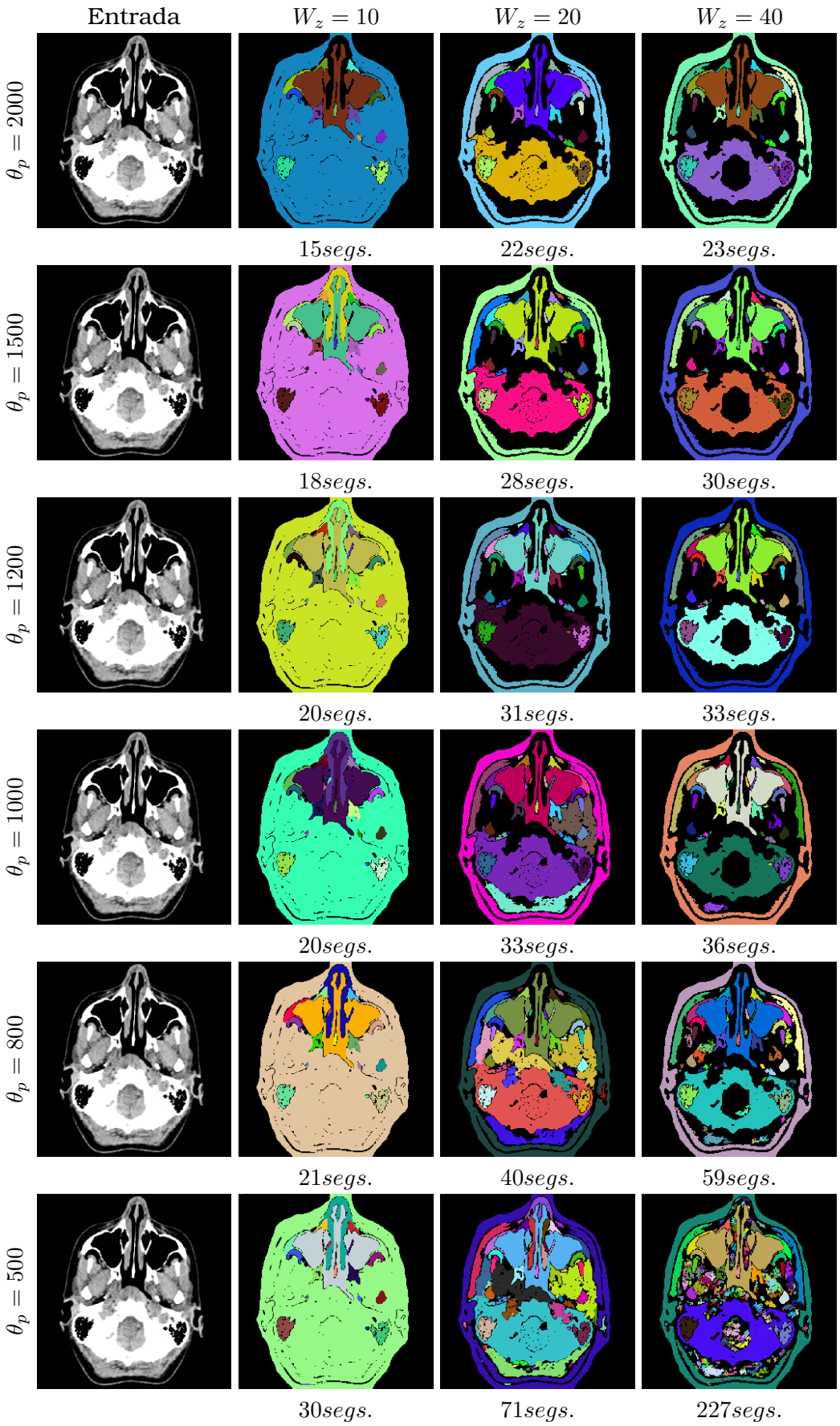


Figura 5.27: Segmentação LEGION com variações dos parâmetros W_z e θ_p. A coluna Entrada apresenta uma MRI de 250x250 *pixels*. Os valores dos parâmetros W_z e θ_p estão descritos nas colunas e linhas, respectivamente. O número de segmentos gerados é mostrado abaixo de cada simulação.

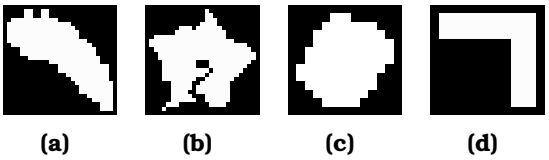


Figura 5.28: Exemplos de objetos para o treinamento do módulo de reconhecimento.

memorizados pelo sistema de reconhecimento ou não. Caso o objeto seja reconhecido pela MLP, o valor de saída da rede é utilizado para configurar o atributo referente ao valor de reconhecimento R (Figura 5.29), de todos os neurônios dentro de cada segmento reconhecido. Inicialmente $R = 0$ para todos os neurônios. No final do processo, todos os neurônios relacionados aos objetos que deverão receber atenção estarão associados a um valor de reconhecimento $R = 1$, o que permitirá a modulação *top-down*, baseada em objetos, do processo atencional.

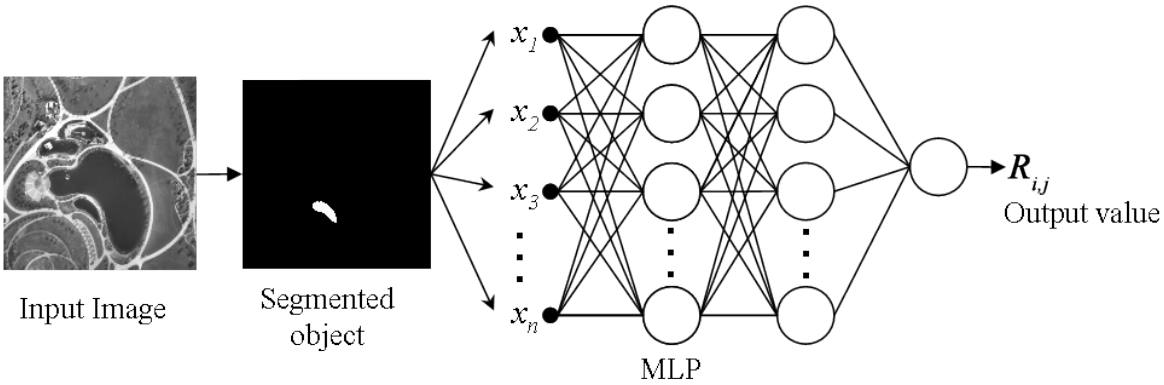


Figura 5.29: Segmentation and recognition value.

O mapa SOM utilizado nesta proposta é baseado diretamente nas descrições apresentadas na Seção 5.2.1, de forma que os padrões utilizados durante a fase de treinamento do mapa são extraídos da cena, descrito por $p = [r,\ g,\ b,\ I,\ RG,\ BY,\ O_0,\ O_{45},\ O_{90},\ O_{135},\ x, y,\ R]$, exceto R, obtido pela saída da MLP. Em uma topologia retangular, o mapa SOM é a partir da Equação (5.23).

Conforme mencionado inicialmente nesta seção, a descrição do modelo, com exceção da rede LEGION, é baseada na descrição apresentada na Seção 5.2.1. Entretanto, é importante notar que, para identificar o objeto saliente, consideramos o termo v_i da Equação 5.24. Apresentaremos a seguir as simulações realizadas.

Simulações Computacionais

Com o objetivo de analisar o comportamento do modelo proposto considerando imagens reais, as simulações foram realizadas considerando os canais de cores r, g e b , o contraste de intensidades I, a diferença em cores RG e BY, orientações O_0, O_{45}, O_{90} e O_{135}, localização em um plano bidimensional x e y, e o valor de reconhecimento R. Os valores de parâmetros utilizados encontram-se

descritos em cada simulação, uma vez que tratam de domínios distintos, ajustes foram realizados de acordo com cada cena.

Inicialmente, a Figura 5.30 (Entrada), apresenta uma cena contendo, aparentemente, três objetos salientes (o carro "vermelho", a placa de sinalização "azul" e a faixa da estrada) contrastantes com os demais segmentos de fundo. Nesta mesma figura, nas colunas à direita, são apresentados a segmentação gerada pela rede LEGION, o mapa SOM, o mapa de atributo-saliência e o local de maior saliência. Tratando-se de uma imagem com grande diversidade de informações, demonstramos, em cada linha, variações do parâmetro W_j com o objetivo de realizar o enviesamento *top-down* de características desejadas e também de testar a robustez do modelo, quanto sua precisão na localização da características desejadas. Nesta simulação, os parâmetros utilizados foram: rede LEGION ($\theta_p = 800$ e $W_z = 20$), gerando um total de 12 segmentos, SOM ($\alpha_k = 0.5$ e $\sigma = 50$) e MAS ($W_Y = 3$). Como pode-se observar na Figura 5.30 (LMS), de acordo com o enviesamento dos atributos RG, BY e O_θ, as três regiões comentadas inicialmente apresentaram maior valor de saliência, recebendo o foco da atenção.

Outro fator importante apresentado na Simulação 1 foi a utilização de valores individuais para o parâmetro Wj, de modo a salientar objetos específicos. Entretanto, a utilização simultânea para outros valores de j não altera o comportamento do modelo, de modo que o potencial de saliência dos neurônios do mapa de atributo-saliência pode ser utilizado para a seleção visual dos objetos salientes, de acordo com os valores de v_i. Para demonstrar esta informação, na Figura 5.31 (a), (b), (c), são apresentados os gráficos obtidos a partir dos valores de v_i dos mapas de atributo-saliência da simulação 1, referente às características RG, BY, O_θ, respectivamente. Em (d) é apresentado o gráfico dos valores de v_i considerando $W_{5..10} = 1.0$, que considera todas as regiões salientes. Nas próximas seções abordaremos esta questão com maiores detalhes.

Para uma comparação qualitativa, a cena de entrada apresentada na Figura 5.30 foi submetida ao modelo de saliência de Itti et al. (1998). Os resultados são apresentados na Figura 5.32. A partir da observação das características *bottom-up*, podemos concluir que os resultados obtidos por ambos os modelos foram bastante semelhantes.

Baseado na cena e na diversidade de objetos salientes existentes, o ajuste do parâmetro W_j (Equação 5.25) permite associar pesos à características desejadas. Na simulação seguinte, apresentada na Figura 5.33, consideramos como entrada uma imagem MRI (250x206 *pixels*) com diagnóstico de câncer. O objetivo desta simulação é observar que a atenção visual também pode ser guiada por características específicas, possibilitando, neste caso, o enviesamento atencional para possíveis lesões. O objeto, previamente conhecido, é apresentado na Figura 5.28 (c). De acordo com a Figura 5.33, a partir da imagem de entrada são gerados o

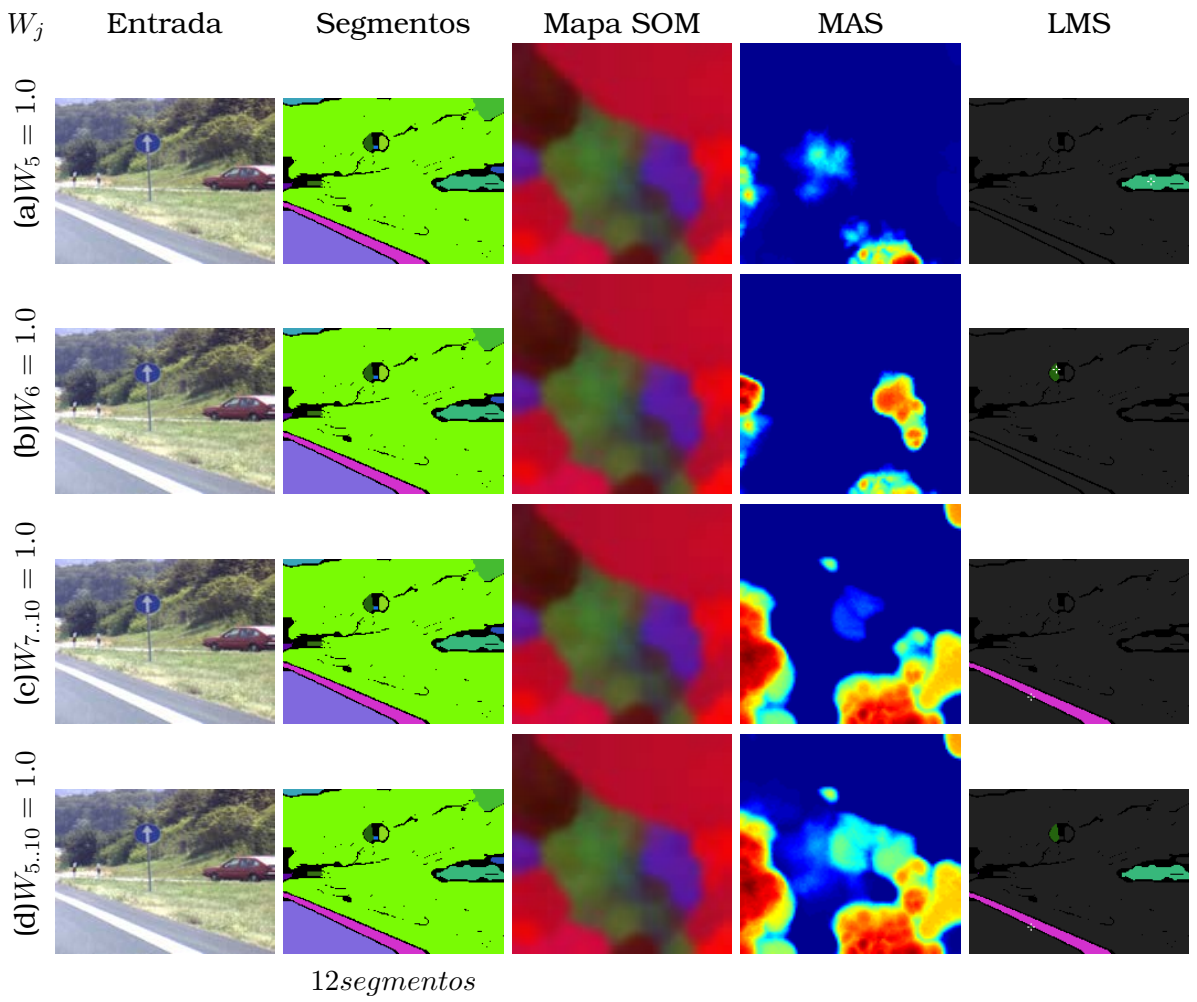
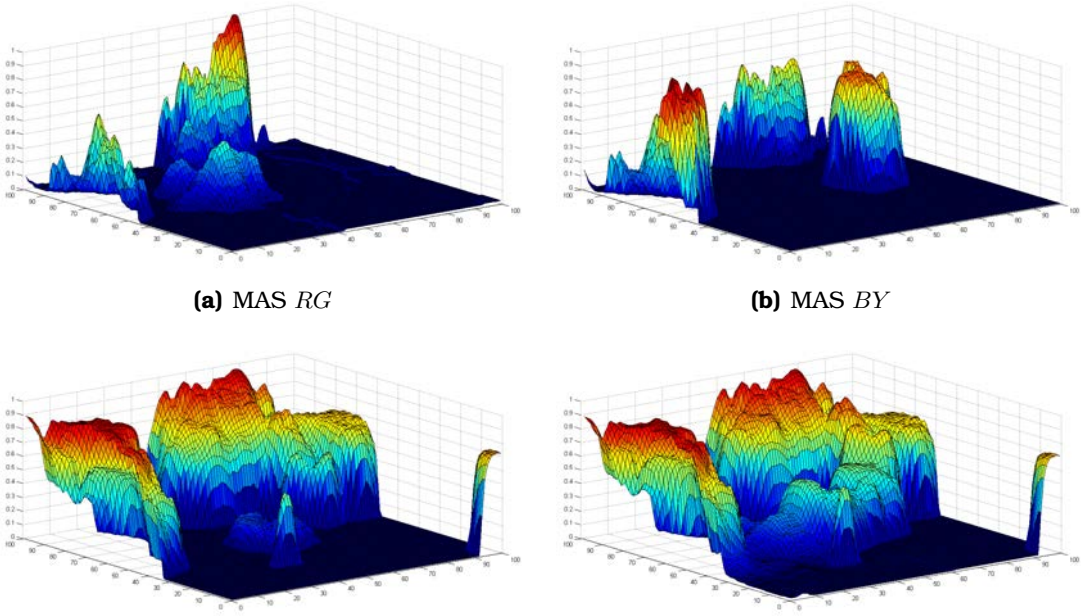


Figura 5.30: Modelo de atenção *top-down* e *bottom-up* II. Simulação 1. Modulações do parâmetro W_j para o enviesamento *top-down* de características desejadas. O valor do parâmetro W_j encontra-se na primeira coluna. Para todas as simulações, foi utilizado $W_j = 0.0$ para todo j não informado. Figura da base de imagens disponibilizada publicamente por Itti (200x150 *pixels*).

mapa SOM e os segmentos (total de vinte). Todos os segmentos geradas são então submetidos à rede de reconhecimento (MLP) e, neurônios que compõe um objeto reconhecido são configurados com $R = 1$. Nesta simulação consideramos $W_{13} = 1.0$ e $W_j = 0$ para $j = [1..12]$, caracterizando o modelo como puramente *top-down*. De maneira descritiva, quando um neurônio com valor de reconhecimento $R = 1$ pulsar na LEGION, este será representado pelo neurônio m_s do SOM que apresente maior valor de similaridade (Equação 5.23). Consequentemente, o neurônio v_i do mapa de atributo-saliência associado ao neurônio m_s do som será estimulado conforme a Equação 5.24 e, devido aos termos de acoplamento excitatórios e inibitórios desta equação, os neurônios com valor de $R = 1$ mantêm-se com altos valores de v_i (MAS). Finalmente, a localização x e y referente ao neurônio v_i é utilizada para indicar o local de maior saliência (LMS).



(a) MAS RG **(b)** MAS BY

(c) MAS O_θ **(d)** MAS RG, BY e O_θ

Figura 5.31: Gráficos dos MAS referente à Simulação 1.

5.3 Competição por Atenção Visual Baseada em Objetos

De acordo com os conceitos apresentados na Seção 4.3, evidências comportamentais e neurofisiológicas demonstram que mecanismos de seleção visual baseado em objetos podem ser empregado no sistema visual dos primatas (O'Craven et al., 1999; Roelfsema et al., 1998; Desimone and Duncan, 1995), propondo que um mecanismo pré-atentivo realize, inicialmente, a segmentação do campo visual em um conjunto de objetos que possam servir de alvo para a atenção visual (O'Craven et al., 1999).

Embora os modelos apresentados anteriormente neste capítulo apresentem

(a) Modelo de Itti et al. (1998) **(b)** Modelo proposto

Figura 5.32: Comparação qualitativa de seleção de objetos em cenas reais. (a) Seleção visual das localizações salientes a partir do modelo proposto por Itti et al. (1998) e (b) Resultado do modelo apresentado nesta seção.

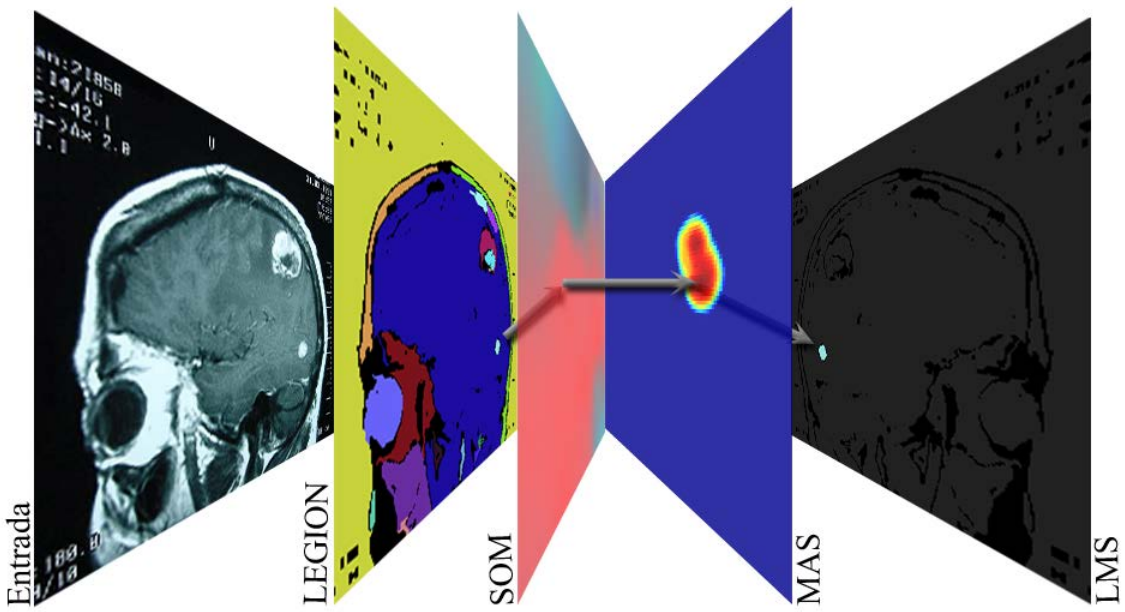


Figura 5.33: Modelo de atenção *top-down* e *bottom-up* II. Simulação 2.

em suas propostas mecanismos de seleção visual baseados em objetos, de forma que a seleção visual ocorra sequencialmente entre os objetos, a competição pela atenção visual tem sido desenvolvida baseada não em objetos, mas sim na atividade correlacionada entre neurônios que compõem os objetos. Desta forma, como principal objetivo considerado nesta seção, pretendemos desenvolver um mecanismo para competição pela atenção visual baseado estritamente em objetos, denominado por mapa de objeto-saliente (MOS) e, consequentemente, possibilitar a redução do esforço computacional para o desenvolvimento da seleção visual.

5.3.1 Atenção Visual Baseada em Objetos

Propomos nesta pesquisa, como alternativa à utilização da rede MLP para o reconhecimento de objetos, o uso de um modelo de classificação de objetos baseado na combinação de classificadores de baixo e alto nível, proposto por Silva and Zhao (2012). Segundos os autores, a classificação de baixo nível é implementada a partir de técnicas tradicionais de classificação, como por exemplo, uma MLP. Por outro lado, a classificação de alto nível explora propriedades topológicas complexas, correspondentes à rede gerada a partir dos padrões de entrada, o que proporciona maior robustez para a classificação em relação a variações nos padrões de reconhecimento.

Sendo assim, este modelo foi composto pelos seguintes módulos: a extração de características visuais primitivas, responsável pela extração dos mapas de conspicuidades contendo as características visuais primitivas da cena, o modelo LEGION, para a segmentação da imagem, o classificador de alto nível baseado em padrões topológicos complexos de rede (*Network-Based High Level Classification* - HLC) (Silva and Zhao, 2012), e finalmente, a geração do mapa de objeto-saliente, que

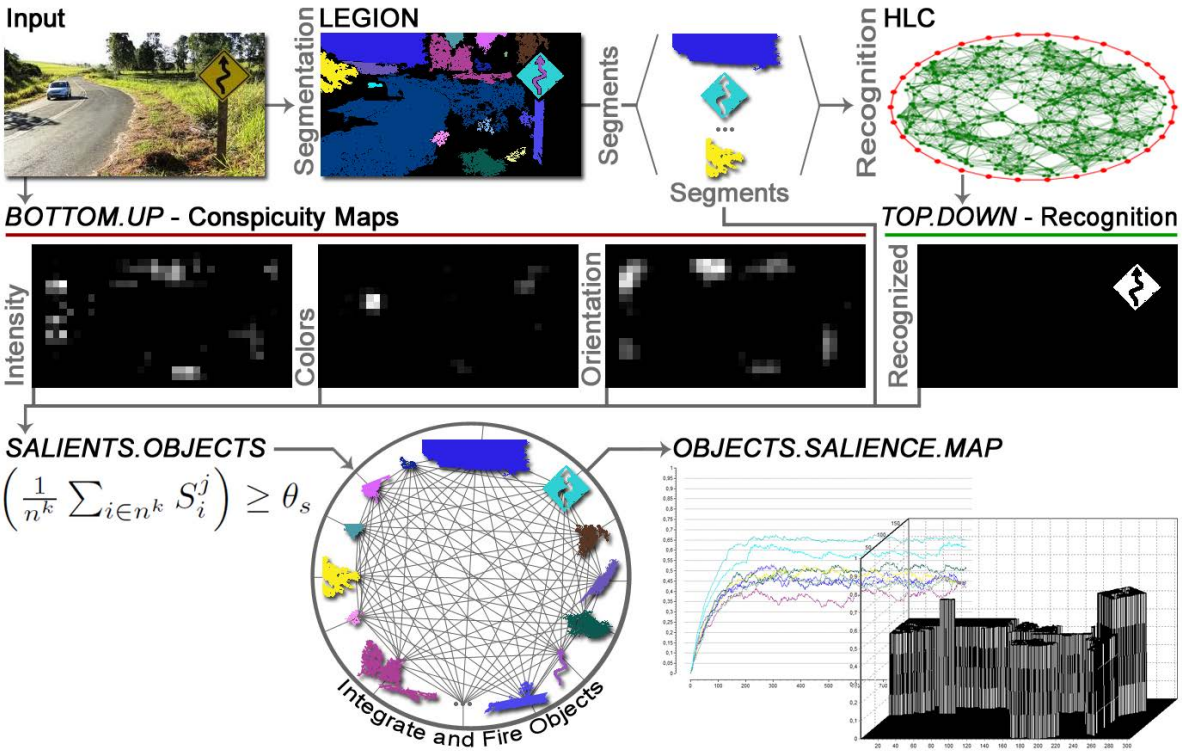


Figura 5.34: Diagrama do modelo de seleção baseada em objetos.

torna possível a seleção visual dos objetos salientes. Os resultados deste trabalho encontram-se submetidos para publicação em (Benicasa et al., 2013). A seguir apresentaremos a descrição do modelo e simulações realizadas a partir de cenas sintéticas e reais.

Descrição do Modelo

De acordo com o diagrama do modelo proposto (Figura 5.34), inicialmente a cena é apresentada, paralelamente, aos módulos responsáveis pela extração de características visuais primitivas e segmentação da cena. Para a implementação destes módulos, os modelos descritos nas Seções 3.1 e 5.2.2 são utilizados, respectivamente. As saídas obtidas são então utilizadas como alimentação para o modelo de reconhecimento de objetos e geração do mapa de objeto-saliente, apresentados a seguir.

Diferente das propostas apresentadas, consideramos aqui a extração de características primitivas baseada no processo de geração do mapa de saliência proposto por Itti et al. (1998). Entretanto, utilizaremos somente os mapas de conspicuidades de contraste de intensidades, cores e orientações, com o objetivo de manter informações primitivas em mapas individuais.

Como extensão das propostas apresentadas nas Seções 5.1.3 e 5.2.1, utilizaremos informações *bottom-up* e *top-down* para o desenvolvimento da atenção visual. Como já mencionado, características visuais primitivas, ou seja, contraste de intensidades, cores e orientações, definem o sinal *bottom-up*. Por outro lado,

informações sobre objetos previamente memorizados (top-down) também são responsáveis por guiar o processo de seleção visual. Nesta proposta, com o objetivo de melhorar a precisão na classificação dos segmentos gerados pela rede LEGION, propusemos a substituição da MLP pela rede HLC, recentemente proposto por Silva and Zhao (2012). Os processos de treinamento e classificação dos segmentos LEGION mantêm-se baseados na proposta apresentada na Seção 5.2.2, de modo que, quando um segmento sincronizado pulsa na rede LEGION, este é diretamente apresentado para à HLC. O valor de saída da HLC indica se o objeto está entre àqueles memorizados pelo sistema de reconhecimento. Caso o objeto seja reconhecido pela HLC, o valor de saída da rede é utilizado para configurar o atributo referente ao valor de reconhecimento R de todos os neurônios dentro do segmento reconhecido. Inicialmente, $R = 0$ para todos os neurônios. No final do processo, todos os neurônios relacionados aos objetos que deverão receber atenção estarão associados a um valor de reconhecimento $R = 1$, o que permitirá a modulação atencional *top-down* baseada em objetos.

Devido aos valores de R gerados pela rede de reconhecimento, verificamos a possibilidade de segmentos representando objetos desconhecidos também apresentarem valores de reconhecimento ($0 < R < 1$). Com o objetivo de evitar que tais objetos possam influenciar o processo de modulação atencional *top-down*, adotamos o limiar θ_R. Assim, o valor de reconhecimento R é definido como:

$$R = \begin{cases} 1, & \text{se } R \geq \theta_R \\ 0, & \text{caso contrário} \end{cases} \tag{5.28}$$

Outro fator considerado importante neste trabalho é a seleção prévia de objetos salientes. Para cada *pixel* da imagem de entrada foram extraídos os seguintes descritores: contraste de intensidades I, diferença espacial em cores RG e BY, orientações O_θ com $\theta \in \{0°, 45°, 90°, 135°\}$, localizações espaciais x e y dos *pixels*, e o valor de reconhecimento R. Tendo $l_k = [I, C, O, L, R]$ (*Intensidade*, *Cor*, *Orientação*, *Localização* e *Reconhecimento*) como um padrão pertencente a um segmento ativo na rede LEGION e, uma vez que o processo de segmentação esteja concluído e os valores de saliência de todos os *pixels* pertencentes à imagem de entrada encontrem-se devidamente calculados, o cálculo da média de saliência de cada característica k dos segmentos j pode ser definido como:

$$S_k^j = \frac{1}{n^j} \sum_{i \in n^j} l_{k_i}^j, \tag{5.29}$$

onde n^j representa o número de neurônios do segmento j e $l_{k_i}^j$ é o valor do mapa de saliência do neurônio i pertencente à característica k do segmento j. É importante notar que cada objeto continua sendo representado por k características, preservando a informação do segmento. Com a definição da saliência média dos segmentos

S_k^j, torna-se possível o desenvolvimento do mecanismo que permitirá a competição pela atenção visual baseada em objetos, diferente dos modelos anteriores, onde v_i representa o potencial de saliência de cada neurônio. Discutiremos sobre esta questão ainda nesta seção.

Para (D., 1988; Treisman and Gormican, 1988; Wolfe and Horowitz, 2004), outra característica importante que pode guiar o desenvolvimento da atenção visual é o tamanho do objeto que, neste trabalho, é representado por n^j, ou seja, o tamanho do objeto é incorporado ao valor de saliência S_k^j de acordo com o número de neurônios presentes no segmento j. Portanto, o vetor de características foi redefinido e normalizado, de modo que $l_k = [I, C, O, L, R, T]$, onde T representa o tamanho do segmento. A normalização dos valores de l_k para o intervalo entre $[0, 1]$ é definida como:

$$l'_k = \frac{l_k - min_{l_k}}{max_{l_k} - min_{l_k}}, \tag{5.30}$$

onde max_{l_k} and min_{l_k} representam o máximo e o mínimo valor da característica k, respectivamente. Neste caso, os valores das médias das saliências (Equação 5.29) devem ser atualizadas.

O valor de saliência de um objeto é considerado um dos principais, senão o principal, componente deste trabalho, pois somente a partir destes valores torna-se possível a seleção de objetos. De acordo com os trabalhos apresentados no Capítulo 4, os valores de saliência de um objeto são baseados em características *bottom-up*, *top-down*, ou ainda, em ambas. Desta forma, a ausência de características salientes, pode excluir, automaticamente, regiões ou objetos da cena. Este comportamento é denominado biologicamente por busca assimétrica que, de acordo com Treisman and Gormican (1988) e Bruce and Tsotsos (2009), a assimetria atribuída à presença ou ausência de uma determinada característica, deva ser considerada como uma importante informação para o desenvolvimento da atenção visual. Consideremos as imagens apresentadas na Figura 5.35 (a), (b) e (c), e seus respectivos mapas de conspicuidades de intensidades, cores e orientações, e o mapa de saliência proposto por Itti et al. (1998). Como pode ser observado na Figura 5.35 (a), o objeto representado pelo sinal de "mais", apresenta alto valor de saliência em relação aos demais objetos, sendo destacado no mapa de saliência. Na Figura 5.35 (b), embora o objeto representando o sinal de "menos", de acordo com uma inspeção visual humana, seja o objeto mais saliente, este não é destacado no mapa de saliência, devido aos baixos valores de saliência apresentados nos mapas I, C e O. A mesma situação é apresentada na Figura 5.35 (c), onde um único objeto, representado pela cor "laranja", não apresenta valores de saliência suficientes para ser destacado no mapa de saliência. Entretanto, analisando a atenção visual baseada no contraste de características, e não em níveis de saliência, consideramos que a ausência de uma determinada característica possa, também, ser utilizada para o desenvolvimento da

atenção visual. Deste modo, propomos neste trabalho a normalização dos valores de saliência dos objetos, para valores de contrastes reais de saliência, descrito como segue:

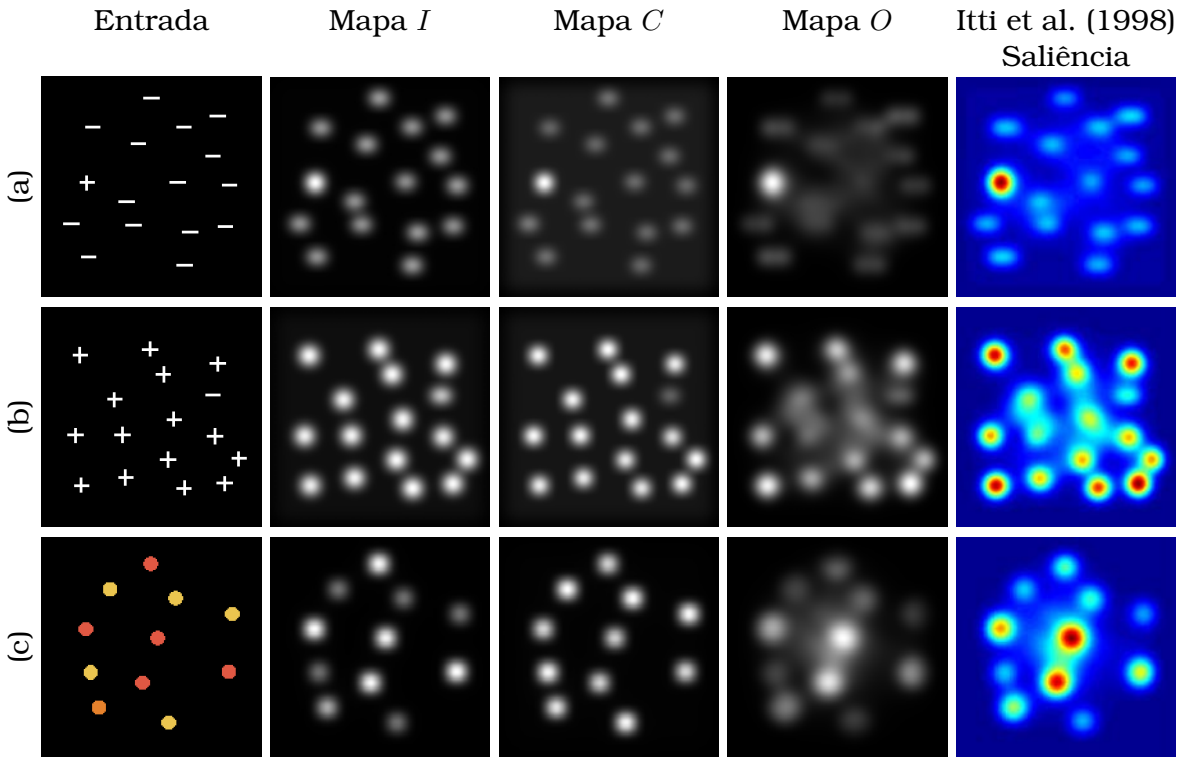


Figura 5.35: Análise da Saliência de Objetos Imagens do *benchmark* disponibilizado publicamente por Bruce and Tsotsos (2009).

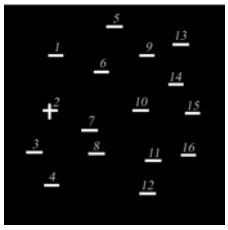
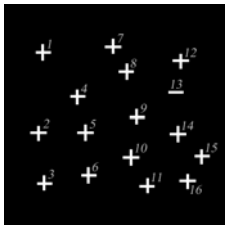
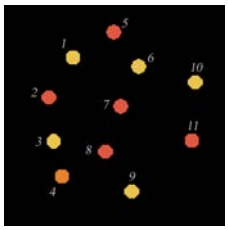
$$\mathcal{N}_s(S_k^j) = \frac{1}{n^s} \sum_{i \in n^s} |D(S_k^j) - D(S_k^i)|, \tag{5.31}$$

$$D(S_k) = \frac{1}{n^s} \sum_{i \in n^s} |S_k - S_k^i|, \tag{5.32}$$

onde n^s representa o número de segmentos gerados pela rede LEGION e $D(S_k)$ é a média da similaridade entre o valor de saliência da característica k de um objeto j em relação aos demais objetos. De modo geral, de acordo com a Equação 5.32, objetos com valores de saliência próximos à média, deverão apresentar baixos valores de contraste, enquanto que objetos distantes, serão definidos com altos valores de contraste. De acordo com os valores apresentados na Tabela 5.1, considerando as imagens apresentadas na Figura 5.35 (a) e (b), o objeto representado pelo sinal de "mais" ((a) objeto 2), e o objeto representado pelo sinal de "menos" ((b) objeto 13), apresentam altos valores de contraste em relação à característica I (veja coluna $D(S_1^j)$ da Tabela 5.1). Entretanto, baseando-se ainda nos valores obtidos pela Equação 5.32, o objeto de cor "laranja" ((c) objeto 4), apresenta o menor valor de contraste possível, uma vez que é definido com valores de saliência médios no mapa de conspicuidades de intensidades (Figura 5.35 (c) Mapa I). Neste caso, a Equação 5.31

é responsável por definir a média da similaridade entre o valor de contraste $D(S_k^j)$, em relação aos demais objetos. Como pode ser observado na Tabela 5.1, os valores de contrastes reais de saliência são apresentados na coluna $\mathcal{N}_s(S_1^j)$ e normalizados, em seguida, para valores entre $[0, 1]$. Concluímos que a normalização dos valores de saliência dos objetos pode ser um importante mecanismo para a identificação de objetos contrastantes na cena. Na seção seguinte serão apresentadas simulações baseadas nesta hipótese.

Tabela 5.1: Geração de valores de contrastes reais de saliência.

Figura 5.35	Objeto j	Saliência I	$D(S_1^j)$	$\mathcal{N}_s(S_1^j)$	Valores Normalizados entre $[0, 1]$
(a) 	1	139	10.75	6.92	0.00
	2	**252**	**104.63**	**86.95**	**1.00**
	3	135	12.38	6.66	0.00
	4	136	11.63	6.56	0.00
	5	149	14.50	8.09	0.02
	6	147	13.00	6.97	0.01
	7	141	11.00	6.75	0.00
	8	138	10.88	6.83	0.00
	9	137	11.13	6.69	0.00
	10	135	12.38	6.66	0.00
	11	145	12.00	6.56	0.00
	12	145	12.00	6.56	0.00
	13	139	10.75	6.92	0.00
	14	137	11.13	6.69	0.00
	15	147	13.00	6.97	0.01
	16	136	11.63	6.56	0.00
(b) 	1	249	11.00	5.08	0.00
	2	238	11.13	5.09	0.00
	3	235	13.00	5.86	0.02
	4	241	10.00	5.36	0.01
	5	238	11.13	5.09	0.00
	6	243	9.63	5.61	0.02
	7	274	32.63	20.11	0.44
	8	246	10.00	5.36	0.01
	9	242	9.75	5.52	0.01
	10	251	12.50	5.61	0.02
	11	247	10.25	5.27	0.01
	12	249	11.00	5.08	0.00
	13	**187**	**54.38**	**39.14**	**1.00**
	14	230	16.75	8.20	0.09
	15	243	9.63	5.61	0.02
	16	249	11.00	5.08	0.00
(c) 	1	111	67.82	3.06	0.12
	2	252	75.18	5.23	0.58
	3	106	70.82	2.83	0.07
	4	**167**	**62.73**	**7.22**	**1.00**
	5	240	69.36	2.49	0.00
	6	109	68.55	2.68	0.04
	7	241	69.64	2.51	0.01
	8	250	73.73	4.16	0.35
	9	108	69.18	2.50	0.00
	10	110	68.09	2.88	0.08
	11	251	74.36	4.56	0.44

Considerando a normalização dos valores das médias de saliência de cada segmento S^j concorrente à atenção visual, podem existir objetos com baixos valores de contraste ou, como tem sido referenciado, saliência. Como solução, propusemos a utilização de um limiar de valor de saliência do segmento, denominado por θ_s, de modo que, segmentos com valor total de saliência menor que θ_s não poderão participar da competição pela atenção visual. A seleção prévia de objetos é definida

por:

$$P_{priorselection}(S^j) = \begin{cases} 1, & \text{se } \left(\frac{1}{n^k}\sum_{k=1}^{6} S_k^j\right) \geq \theta_s \\ 0, & \text{caso contrário} \end{cases}, \tag{5.33}$$

onde n^k é o número de características responsáveis por guiar a atenção visual consideradas neste trabalho.

De modo diferente dos modelos que tratam neurônios como unidades individuais para a representação da saliência da cena (Quiles et al., 2011; Benicasa et al., 2012, 012b), ou seja, um neurônio por *pixel*, consideramos nesta proposta a geração de um mapa de objeto-saliente a partir da atividade entre neurônios que representam objetos inteiros, o que significa a representatividade de um neurônio por objeto.

Assim, o mapa de objeto-saliente é definido como uma rede neural composta por objetos com dois tipos de conexões: conexões excitatórias e conexões inibitórias. As conexões excitatórias formam um mecanismo cooperativo responsável por sincronizar grupos de objetos que representam padrões similares (objetos com características similares). Por outro lado, as conexões inibitórias têm como objetivo inibir objetos de fundo da cena, permitindo que objetos mais salientes sejam selecionadas. O MOS foi constituído pelo mesmo número de segmentos (objetos) gerados pela rede LEGION. Sua dinâmica foi definida da seguinte forma: considera-se uma imagem de entrada, sincronizada e segmentada. Quando um objeto pulsar na LEGION, seu sinal é apresentado para todos os outros objetos selecionados, sendo atualizados como segue:

$$\dot{v}_j = -v_j + E_j - W_Y Y_j, \qquad j = 1, \ldots, n^j \tag{5.34}$$

onde n^j é o número de objetos do MOS. Considerando que o MOS seja composto pelo mesmo número de objetos pulsantes na rede LEGION, pode-se concluir que cada objeto da LEGION possui seu respectivo objeto no MOS. A variável v_j representa o potencial de saliência do objeto j e W_Y é o peso de inibição a partir do termo de acoplamento inibitório Y_j.

Consideramos S_k^j sendo um objeto referente a um segmento ativo na rede LEGION, e k seu respectivo índice de características. Os termos de acoplamento excitatório E_j e inibitório Y_j são definidos pela seguinte equação:

$$E_j = Y_j = exp^{-d(S^j, S^s)}, \qquad s = 1, .., n^j, \tag{5.35}$$

de forma que E_j será atualizado, se e somente se, o valor de E_j, em outro instante $t+1$, for superior t, ou seja, o termo E_j conterá o valor máximo excitatório do objeto j e S^s representa cada objeto pulsante. A medida de similaridade entre as

características do objeto S^j e os demais objetos é definida por:

$$d(S^j, S^s) = \| S^j - S^s \| = \sqrt{\sum_{k=1}^{6} W_k (S_k^j - S_k^s)^2}, \quad k \in \{I, C, O, L, R, T\} \tag{5.36}$$

onde W_k define o peso associado a cada característica k. É importante notar que, com o parâmetro W_k é possível realizar o ajuste dos pesos, de acordo com as características que se deseja salientar da cena. Assim, se $W_k = 0$ para todas as informações primitivas da imagem de entrada, o modelo proposto torna-se estritamente *top-down* e, *bottom-up*, caso $W_k = 0$ para todas as informações relacionadas ao reconhecimento de objetos.

As conexões inibitórias são determinadas com base no contraste entre atributos. Desta forma, se dois objetos são alimentados por atributos semelhantes, ou seja, o contraste entre eles é pequeno ou zero, o termo Y_j da Equação 5.34 aproxima zero e, devido a função exponencial negativa da Equação 5.35, o peso de acoplamento inibitório assume um alto valor. Por outro lado, quando os sinais de alimentação de tais objetos são definidos por atributos distintos, o peso da conexão inibitória entre eles é pequeno ou mesmo zero. Assim, objetos com características semelhantes são mutuamente inibidos, devido a competição gerada pelas conexões inibitórias. Podemos concluir que um objeto que apresenta um alto contraste com os demais não é inibido e permanece oscilando, salientando o objeto sob o foco de atenção. A Figura 5.36 apresenta, de forma ilustrativa, o processo descrito para a geração do mapa de objeto-saliente, onde uma arquitetura em camadas apresenta a imagem de entrada, a segmentação LEGION e o MOS, gerado a partir da competição entre objetos.

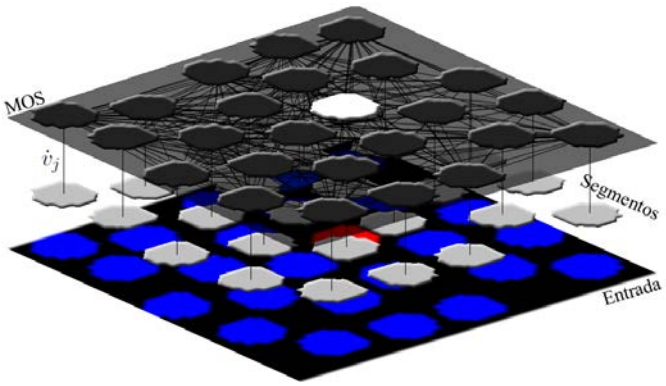


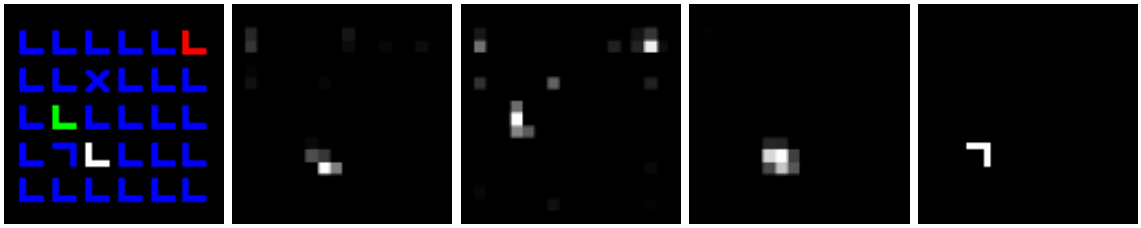
Figura 5.36: Mapa de Objeto-Saliente gerado a partir da competição entre objetos.

Para demonstrar o comportamento do modelo proposto, a Figura 5.37 apresenta uma simulação baseada em um ambiente sintético, composto por todas as características considerados neste trabalho. Neste caso, $W_k = 1.0$ para todos os valores de k, ou seja, nenhum enviesamento *top-down* foi realizado. Na Figura 5.37 (b), (c), (d) e (e) são apresentados os mapas de contraste de intensidades, cores,

orientações e reconhecimento, respectivamente. Informação sobre a localização e tamanho dos objetos podem ser observadas a partir da imagem de entrada, apresentada em (a). Em (f) é mostrado o resultado da segmentação LEGION. Com o objetivo de apresentar a evolução do potencial de saliência v_j de cada objeto, em (g) é apresentado o gráfico do potencial saliência v_j sob o tempo t. É importante notar que, a unidade de tempo é incrementada de acordo com a dinâmica que define a geração do MOS, ou seja, $t = t + 1$ quando um objeto S^j é apresentado aos demais objetos S^s. Sendo assim, faz-se necessário $t = n^j$ para que a medida de similaridade (Equação 5.36) ocorra, ao menos uma vez, entre todos os objetos pulsantes. De acordo com o gráfico apresentado na Figura 5.37 (g), o comportamento do modelo torna-se estável quando $t \approx 150$. Entretanto, os parâmetros W_k e θ_s exercem um importante papel para o tempo de seleção visual dos objetos. O MOS é apresentado em (h). Para fins de esclarecimentos, como pode ser observado no gráfico (g), o objeto que apresenta maior valor de saliência é o "X". A diferença considerável na amplitude deste objeto ocorre devido a normalização da característica tamanho (T), uma vez que todos os demais objetos possuam o mesmo número de neurônios, X é definido com $T = 1$ e os demais com $T = 0$.

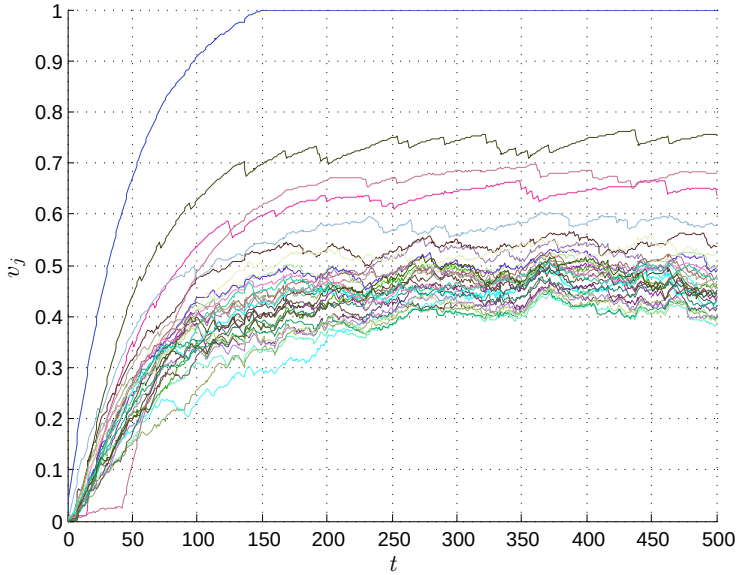
Na Figura 5.38 é apresentada a análise das variações do parâmetro W_k para a cena utilizada na Figura 5.37. Nas Figuras 5.38 (a), (b), (c), (d), (e) e (f), foram utilizados valores de W_k para o enviesamento individual das características: intensidade, cor, orientação, localização, reconhecimento e tamanho, respectivamente. A característica L (Localização) encontra-se presente em todos os objetos, de modo que seu contraste é associado à proximidade entre os objetos. Conforme apresentado em (d), e devido ao posicionamento dos objetos na forma de um quadrado, os maiores valores de saliência estão localizados nas extremidades da imagem. Em (g) são apresentados os objetos salientes considerando os valores de enviesamento informados e, em (h), todas as características são consideradas.

De acordo com os valores de W_k apresentados na Figura 5.38, na Figura 5.39 são apresentados os gráficos dos potenciais de saliência. O objetivo é apresentar a sensibilidade do modelo quanto às variações do parâmetro W_k e o tempo necessário para a seleção dos objetos e estabilidade do modelo proposto. Considerando inicialmente os gráficos gerados a partir de características individuais, apresentados na Figura 5.39 (a), (b), (c), (e) e (f), pode-se observar a rápida identificação do objetos salientes, onde $0 < t < 50$. Em (d), devido à característica localização não apresentar altos níveis de contraste entre os objetos, os valores de v_j oscilam entre ≈ 0.3 e 0.5. Entretanto, como pode ser visto na Figura 5.39 (g), através do enviesamento *top-down* do parâmetro W_k, foi possível obter-se objetos salientes oscilando em faixas distintas, permitindo a seleção visual baseada nos valores de v_j. É importante notar que, conforme apresentado em (g), a ordem de saliência dos objetos encontra-se definida a partir de $t > 50$. Em (h), todas as características são consideradas

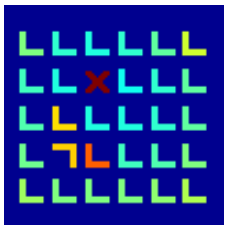


(a) Entrada (b) I (c) C (d) $O\theta$ (e) R

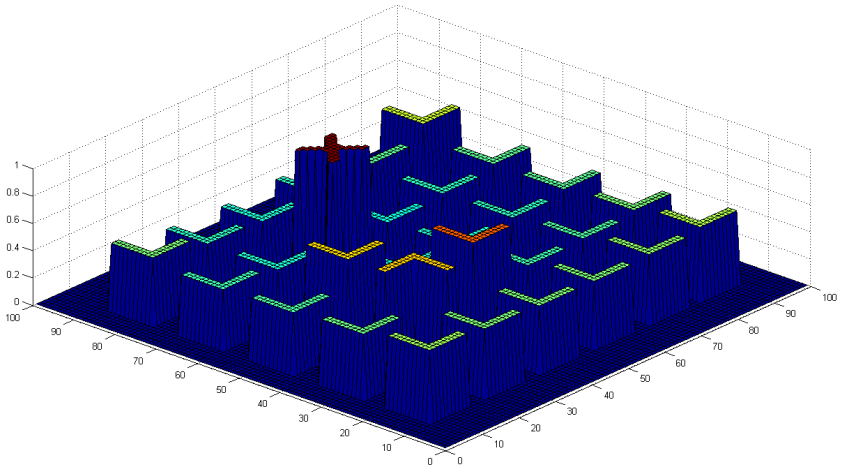
(f) Segmentos



(g) Gráfico v_j

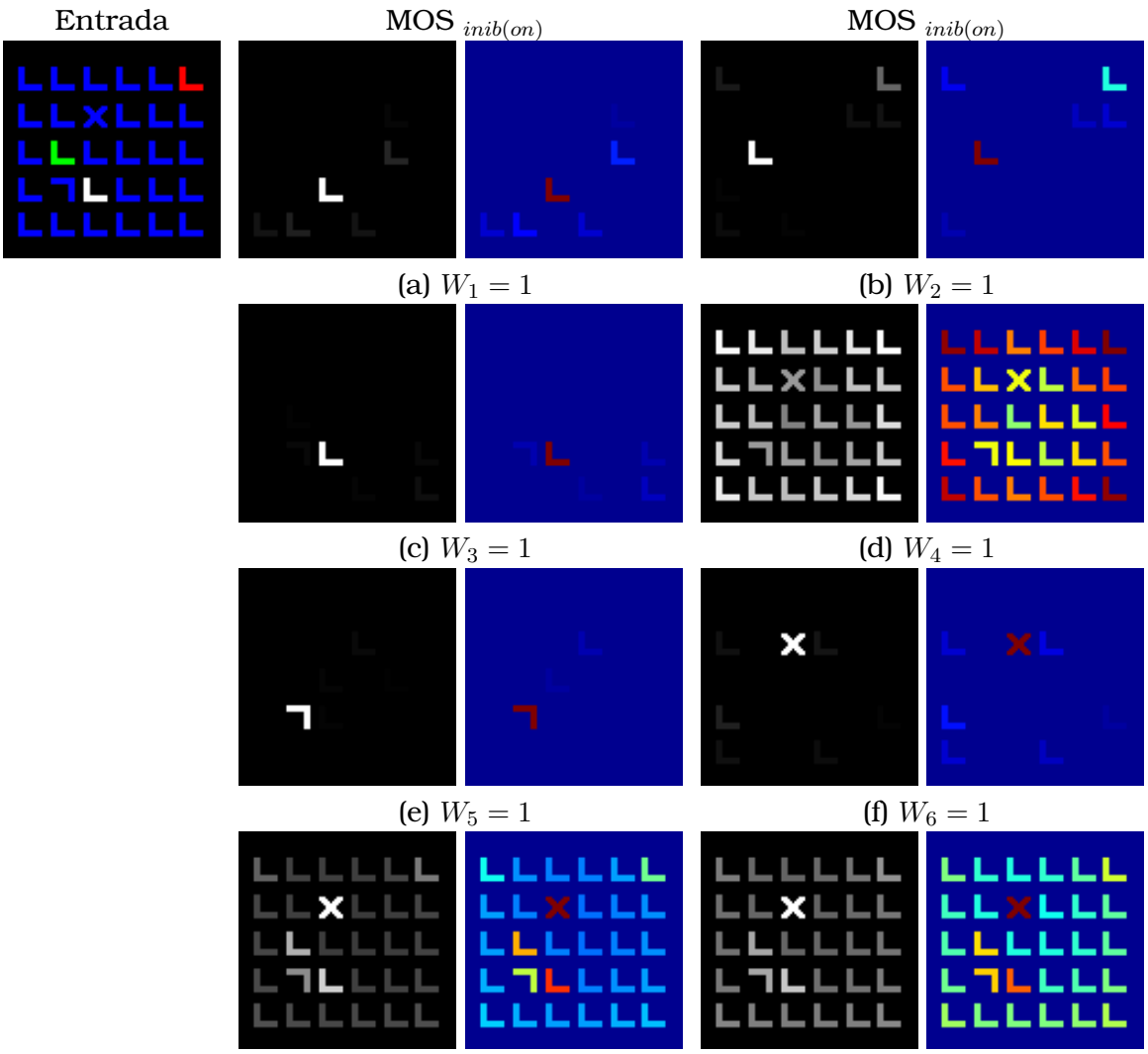


(h) MOS $_{inib(on)}$



(i) MOS $_{inib(on)}3D$

Figura 5.37: Comportamento do modelo. (a) Imagem de entrada. Mapas de conspicuidades: (b) Intensidades, (c) Cores, (d) Orientações, (e) Mapa de reconhecimento de objetos, (f) Segmentação LEGION, (g) Gráfico dos potenciais de saliência, (h) e (i) Mapa de objeto-saliente. Os valores de parâmetros utilizados foram: rede LEGION $\theta_p = 1200$ e $W_z = 20$, gerando um total de 30 segmentos e para a geração do MOS, $W_Y = 1$, $\theta_r = 0.5$, $\theta_s = 0$ e $W_k = 1$ para todos os valores de k.



(a) $W_1 = 1$ (b) $W_2 = 1$

(c) $W_3 = 1$ (d) $W_4 = 1$

(e) $W_5 = 1$ (f) $W_6 = 1$

(g) $W_1 = 1$, $W_2 = 1$, $W_3 = 1$, $W_4 = 0.1$, $W_5 = 0.4$ e $W_6 = 0.4$ (h) $W_k = 1$ para $\forall k$

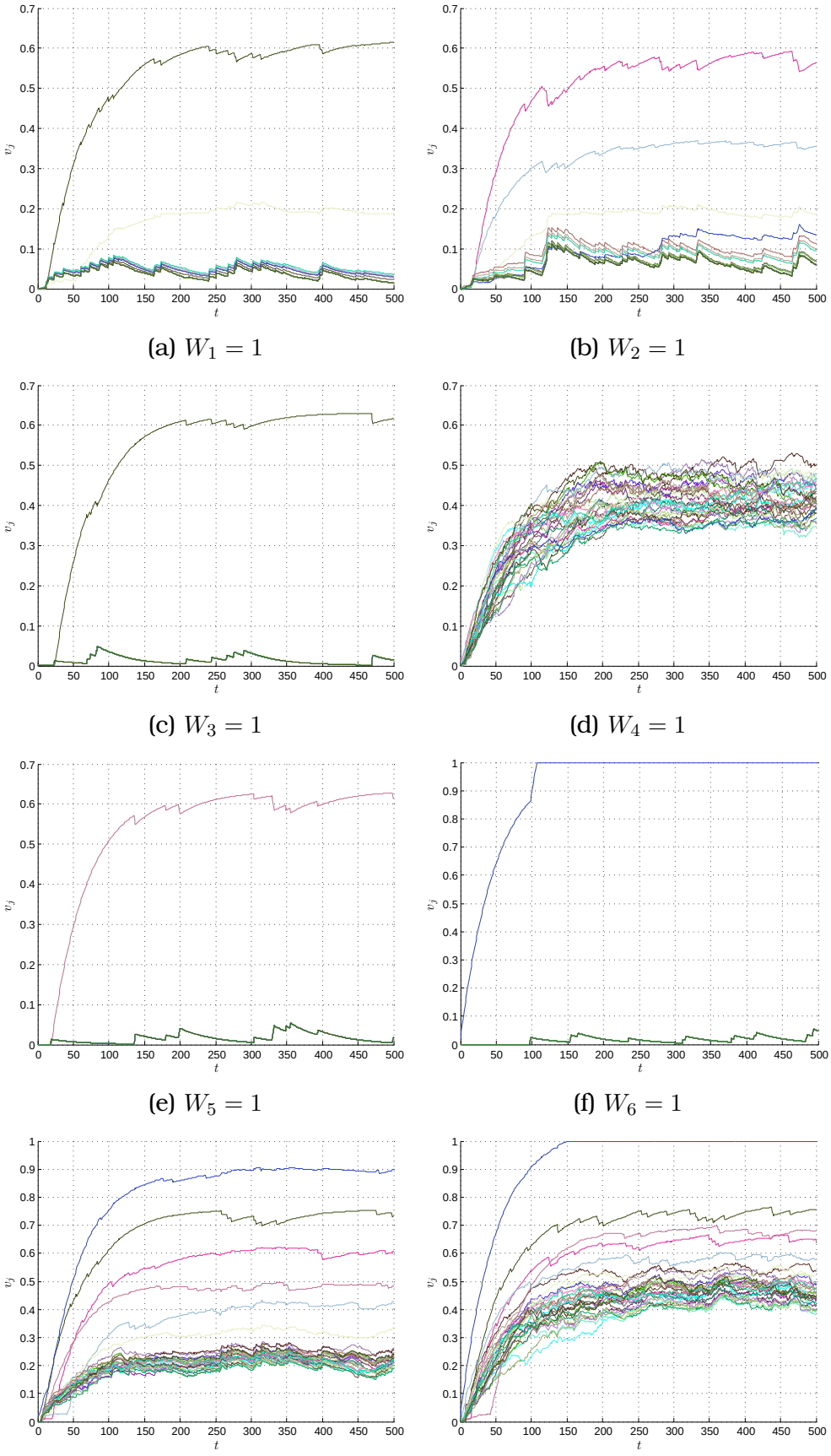
Figura 5.38: Influência do enviesamento *top-down* de características específicas.

igualmente. Neste caso, devido aos valores próximos de v_j, a estabilidade do modelo somente ocorre a partir de $t > 150$.

De maneira geral, assumimos que o objeto saliente é definido como sendo àquele que apresente um maior contraste com os demais objetos presentes na cena, denominados por objeto saliente e objetos de fundo, respectivamente. Esta suposição recebe suporte direto de experimentos biológicos que têm demonstrado que o contraste dos objetos que compõem uma determinada cena é mais importante que o nível absoluto de cada um dos atributos visuais em tarefas de inspeção visual (Wolfe and Horowitz, 2004; Yantis, 2000).

Simulações Computacionais

Simulações computacionais são inicialmente apresentadas para demonstrar o comportamento do modelo proposto considerando características isoladas. Imagens reais e sintéticas, com altos valores de contraste, relacionados às características



(a) $W_1 = 1$

(b) $W_2 = 1$

(c) $W_3 = 1$

(d) $W_4 = 1$

(e) $W_5 = 1$

(f) $W_6 = 1$

$W_1 = 1$, $W_2 = 1$, $W_3 = 1$,
$W_4 = 0.1$, $W_5 = 0.4$ e $W_6 = 0.4$,

$W_k = 1$ para $\forall k$

Figura 5.39: Tempo e estabilidade do modelo de acordo com variações de W_k.

consideradas neste trabalho foram selecionadas. A seguir, enfatizamos o comportamento do modelo a partir de cenas contendo objetos com as seguintes características predominantes: reconhecimento de objetos, objeto com tamanho contrastante, objeto pertencente a uma determinada classe, porém com posicionamento distante de seu grupo, ou seja, com contraste de localização, objeto com contraste de intensidade, cor e orientação. Para a conclusão das simulações foram utilizadas imagens contendo objetos salientes relacionados a mais de uma característica, com o objetivo de avaliar a eficácia do modelo proposto.

De forma a explorar a rede HLC, utilizada neste trabalho para o reconhecimento de objetos, a primeira simulação a ser apresentada desenvolve a atenção visual baseada nos valores de reconhecimento de dígitos manuscritos, apresentados em Silva and Zhao (2012). De acordo com os experimentos apresentados pelos autores, a rede HLC foi aplicada a um base de dados composta por 60.000 padrões de treinamento e 10.000 padrões de testes. Nos resultados apresentados, a rede HLC, proposta a partir da composição de classificadores de baixo e alto nível, apresentou taxas de classificações superiores, comparados a outros modelos de classificação. Na Figura 5.40 são apresentadas as classes reais dos dígitos manuscritos considerados nesta simulação.

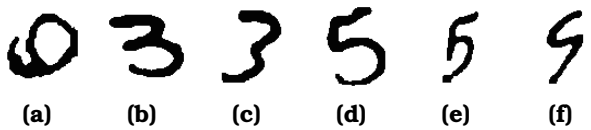
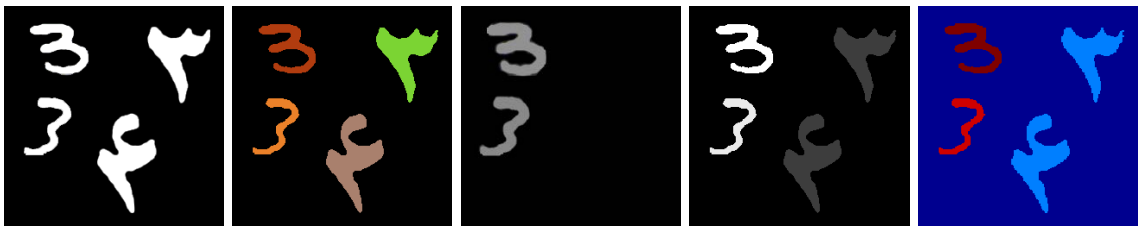


Figura 5.40: Classificação real. (a) Classe 0 (b) Classe 3 (c) Classe 3 (d) Classe 5 (e) Classe 5 (f) Classe 9. Baseado nos experimentos apresentados em (Silva and Zhao, 2012).

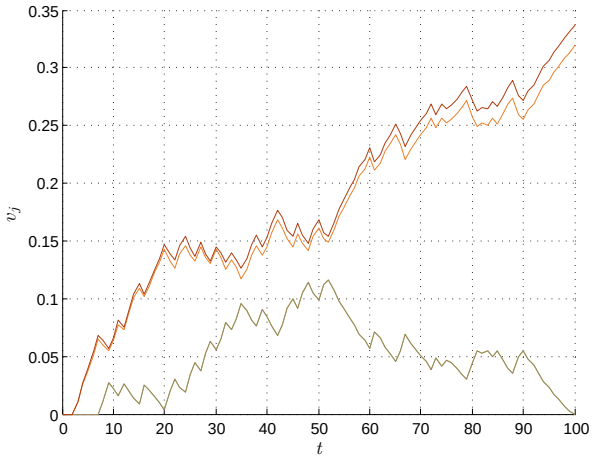
De acordo com os experimentos apresentados em (Silva and Zhao, 2012), quando somente um classificador tradicional (baixo nível) é utilizado ($\lambda = 0$), padrões que sejam facilmente classificados (Figura 5.40 (b) e (d)), classificadores de baixo nível retornam resultados satisfatórios. Entretanto, esta afirmação não é válida para padrões considerados difíceis de classificar (Figura 5.40 (a), (c), (e) e (f)). Como exemplo, os padrões apresentados na Figura 5.40(b) e (d), pertencentes as classes reais 3 e 5, respectivamente, foram classificados com valores 0.74 para a classe 3 e 0.79 para a classe 5, o que é correto, porém, a Figura 5.40(c) (classe real 3) foi classificada com o valor máximo de 0.44 para a classe 7, implicando em uma classificação incorreta. O mesmo ocorre nas Figuras 5.40 (a), (e) e (f). Considerando o valor do parâmetro $\lambda = 0.2$, o padrão apresentado em (c) foi corretamente classificado com valor 0.54 para a classe 3. Isto ocorre devido a HLC examinar não somente a similaridade entre o padrão de entrada e os padrões treinados, mas também pela análise das propriedades topológicas da rede correspondente a cada dígito (para maiores detalhes veja (Silva and Zhao, 2012)). Conforme descrito na seção anterior,

a saída gerada pela rede HLC indica se o objeto está entre àqueles memorizados pelo sistema de reconhecimento, em seguida, o valor de saída da rede é utilizado para configurar o atributo referente ao valor de reconhecimento R.

Na Figura 5.41 é apresentado o comportamento do modelo de atenção em relação a classe 3, considerando $\lambda = 0.2$. Na Figura 5.41 (b) são apresentados os segmentos LEGION. Em (c) são apresentados os valores de reconhecimento de todos os segmentos, sendo que, o valor de saída da HLC para o objeto "3" (superior esquerdo) foi $R = 0.57$ e, para o segundo objeto "3" (inferior esquerdo), o valor de classificação foi $R = 0.54$. Nesta simulação definimos $\theta_R = 0.5$. Demais segmentos tiveram seus valores de R definidos como 0 (zero). Na Figura 5.41 (d) é apresentado o mapa de objeto-saliente que, para fins de visualização, é apresentado através de variações de intensidades e cores. Embora os valores dos potenciais de saliência v_j dos objetos "3" sejam bastante próximos, no gráfico, apresentado em (e), pode ser visto o comportamento individual de cada objeto em relação ao tempo t. Devido aos valores idênticos de R, referente aos dois objetos menos salientes, as representações gráficas dos valores de v_j são apresentadas sobrepostas. Em simulações seguintes apresentaremos o comportamento do modelo a partir do reconhecimento de objetos em imagens reais.



(a) Entrada (b) Segmentos (c) Mapa R (d) MOS $_{inib(on)}$

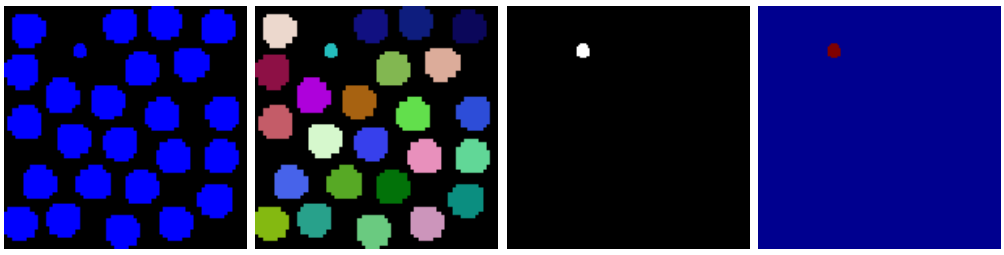


(e) Gráfico v_j

Figura 5.41: Modelo baseado em objetos. Simulação 1 - Saliência de objetos reconhecidos. Valores de parâmetros utilizados: rede LEGION $\theta_p = 1200$ e $W_z = 20$, gerando um total de 4 segmentos e para a geração do MOS, $W_Y = 1.2$, $\theta_r = 0.5$, $\theta_s = 0$ e $W_5 = 1$. Para os demais valores de k, $W_k = 0$. Imagem 255x255 *pixels*.

De acordo com D. (1988), Treisman and Gormican (1988) e Wolfe and Ho-

rowitz (2004), a característica relacionada ao tamanho que define um alvo pode, eficientemente, guiar a atenção visual. Conforme apresentado na Seção 4.3, o primeiro trabalho a considerar a característica tamanho para guiar a atenção, baseada na correlação oscilatória, foi proposto por Wang (1999), de forma que, em uma cena contendo diversos objetos (segmentos), a rede seleciona o maior objeto encontrado, enquanto que os demais são inibidos, devido a um mecanismo de inibição responsável por armazenar o tamanho do maior segmento encontrado durante a oscilação da rede que, consequentemente, inibe a ativação de segmentos menores. Outros trabalhos têm sido desenvolvidos considerando a característica tamanho para guiar a atenção, como por exemplo em (Quiles et al., 2011), contudo, considerando que a atenção visual, baseada no tamanho do objeto, deva ser direcionada para os objetos maiores em uma escala de grandeza. Entretanto, consideramos neste trabalho que o tamanho de um objeto também pode estar relacionado, por exemplo, a um objeto menor entre vários outros objetos maiores e, neste caso, a atenção também deve ser direcionada para objetos menores, ou seja, objetos contrastantes. Este comportamento também se estende para as demais características, como por exemplo, um objeto desconhecido meio a objetos conhecidos, etc. Na Figura 5.42 é apresentado o comportamento do modelo proposto a partir de uma cena sintética composta por um objeto pequeno meio a objetos de maiores. O objeto saliente é apresentado na Figura 5.42(c).

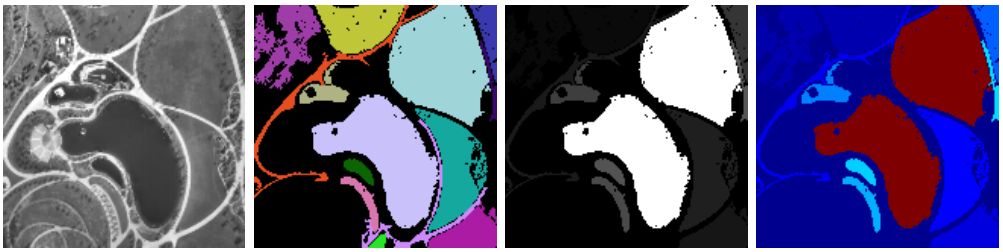


(a) Entrada (b) Segmentos (d) MOS $_{inib(on)}$

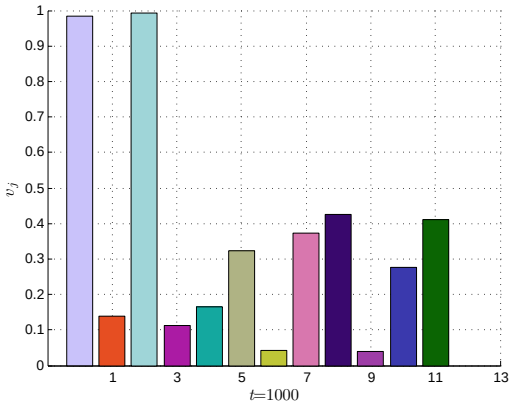
Figura 5.42: Modelo baseado em objetos. Simulação 2 - Saliência de objetos menores. Valores de parâmetros utilizados: rede LEGION $\theta_p = 1200$ e $W_z = 20$, gerando um total de 25 segmentos e para a geração do MOS, $W_Y = 1.2$, $\theta_r = 0.5$, $\theta_s = 0$ e $W_6 = 1$. Para os demais valores de k, $W_k = 0$. Imagem 150x150 *pixels*.

Ainda analisando o comportamento do modelo em relação ao tamanho dos objetos, na simulação apresentada na Figura 5.43, diferente da simulação anterior, devido ao grande número de objetos menores de tamanhos similares, o modelo guia a atenção visual para os objetos maiores. Entretanto, como pode ser visto nas Figuras 5.43 (c), (d) e (d), após os dois primeiros objetos salientes, representados em (d) pelas colunas 0 e 2, a atenção é direcionada para objetos menores, representados pelas colunas 8, 11, 7, etc. Isto ocorre devido ao contraste entre o tamanho de cada objeto em relação ao tamanho médio dos objetos.

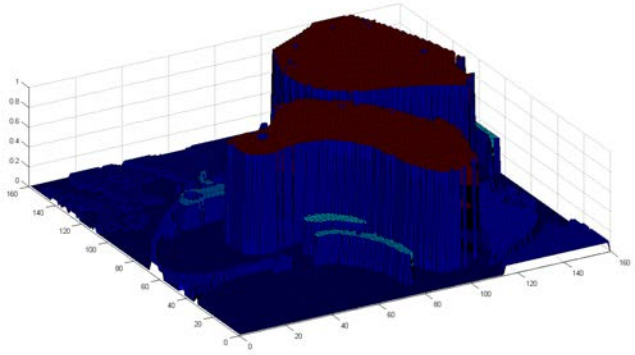
Na Simulação 4, apresentada na Figura 5.44, pode ser observado o comportamento do modelo a partir de variações do parâmetro θ_s. A cena utilizada



(a) Entrada (b) Segmentos (c) MOS $_{inib(on)}$



(d) Gráfico MOS v_j com $t = 1000$



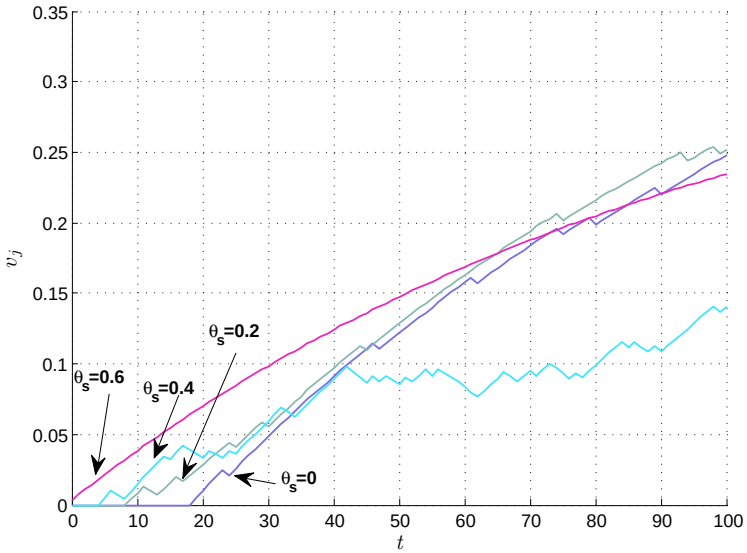
(e) Gráfico MOS v_j 3D

Figura 5.43: Modelo baseado em objetos. Simulação 3 - Saliência de objetos maiores. Valores de parâmetros utilizados: rede LEGION $\theta_p = 1300$ e $W_z = 20$, gerando um total de 14 segmentos e para a geração do MOS, $W_Y = 1.3$, $\theta_r = 0.5$, $\theta_s = 0$ e $W_6 = 1$. Para os demais valores de k, $W_k = 0$. Imagem aérea 160x160 *pixels*, citada inicialmente em (Wang and Terman, 1997).

apresenta segmentos com altos valores de saliência relacionados à característica de intensidade. De acordo com os estudos realizados sobre os parâmetros θ_p e W_z da rede LEGION (Seção 5.2.2), nesta simulação, com o objetivo de gerar um maior número de segmentos, definimos $\theta_p = 600$ e $W_z = 45$. Na Figura 5.44 (b) é apresentada a segmentação LEGION, gerando um total de 21 segmentos. O principal objetivo para o uso do parâmetro θ_s é de evitar a competição entre objetos com baixos valores de saliência e, consequentemente, aumentar a taxa de seleção dos objetos salientes da cena. Como pode ser visto no mapa I, apresentado em (c), o maior valor de saliência da cena está relacionado ao segmento representando a "lua". Neste caso, analisamos o tempo t em que a atenção é direcionada para este segmento, a partir de diferentes valores de θ_s, sendo realizadas quatro simulações. A primeira simulação, representada pela cor "azul" na Figura 5.44 (g), foi utilizado o valor de $\theta_s = 0$, de modo que todos os segmentos participaram da competição pela atenção. Na segunda simulação, representada em (g) pela cor "verde", com $\theta_s = 0.2$, o número de segmentos participantes foi reduzido para 3. A terceira simulação, representada pela cor "ciano", com $\theta_s = 0.4$, somente 2 segmentos foram aptos a competir. Finalmente, a quarta simulação ("rosa"), com $\theta_s = 0.6$, somente 1 segmento superou o valores do limiar. De maneira geral, os tempos necessários para a seleção visual do objeto "lua" foram: $t \approx 18$ para $\theta_s = 0$, $t \approx 8$ para $\theta_s = 0.2$, $t \approx 4$ para $\theta_s = 0.4$ e $t \approx 1$ para

$\theta_s = 0.6$. Desta forma, concluímos que o parâmetro θ_s exerce uma importante tarefa para o mecanismo de atenção visual proposto neste trabalho, principalmente no que refere-se ao tempo necessário para a seleção dos objetos salientes.

(a) Entrada (b) Segmentos (c) Mapa I

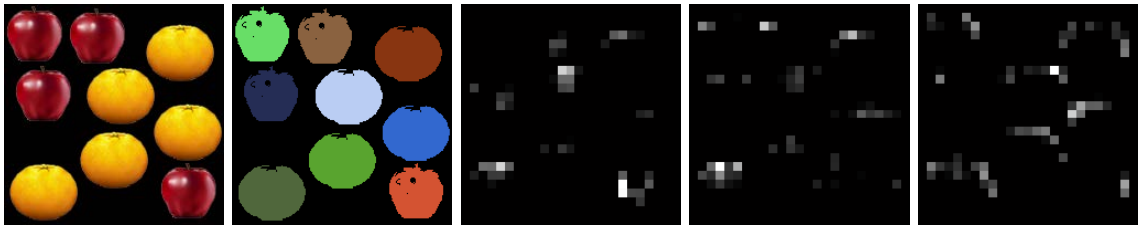


(d) Gráfico MOS v_{lua}

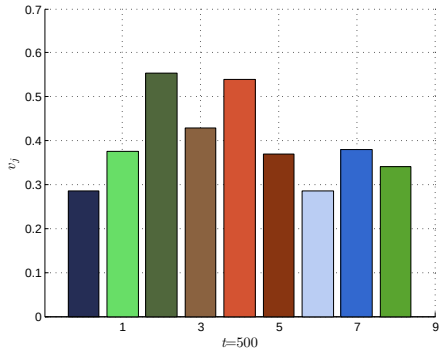
Figura 5.44: Modelo baseado em objetos. Simulação 4 - Variações do parâmetro θ_s. Valores de parâmetros utilizados: rede LEGION $\theta_p = 600$ e $W_z = 45$, gerando um total de 21 segmentos e para a geração do MOS, $W_Y = 1.3$, $\theta_r = 0.5$, $\theta_s = 0$ e $W_1 = 1$. Para os demais valores de k, $W_k = 0$. Imagem 120x202 *pixels*.

Considerado neste trabalho como um atributo primitivo da imagem, a localização espacial do objeto na cena também é considerada como uma importante informação para o desenvolvimento da atenção. A Figura 5.45 (a) apresenta uma cena onde, o contraste apresentado pelos mapas de conspicuidades, relacionadas à cor, intensidade ou orientação (Figuras 5.45 (c), (d) e (e)), não são suficientes para selecionar a "maça", localizada na região inferior direita da cena, como o objeto mais saliente. Assim, propomos neste trabalho a utilização da informação sobre a localização espacial dos objetos como uma característica para o desenvolvimento da atenção visual. Na Figura 5.45 (f) são apresentados os valores do MOS e, em (g), o gráfico dos potenciais de saliência no tempo t. Entretanto, como pode ser observado em (f) e (g), com o uso das características cor, intensidade, orientação e localização, a atenção é direcionada para os objetos mais distantes, portanto, não evidenciando a saliência do objeto considerado alvo. Devido à natureza conjuntiva

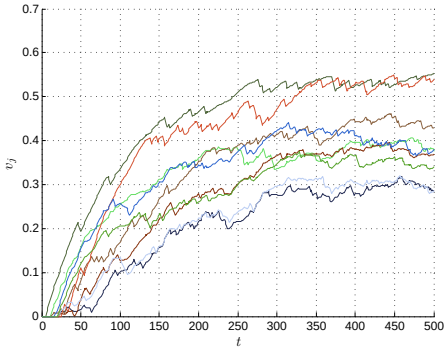
desta busca, outras características devem ser consideradas, de forma a distinguir os diferentes agrupamentos presentes na cena. Desta forma, o tamanho do objeto foi utilizado em conjunção com as demais características, possibilitando a seleção do objeto alvo. As Figuras 5.45 (h) e (i) apresentam os resultados obtidos. De maneira geral, a atenção *bottom-up* foi modulada automaticamente por um estado atencional *top-down*, ou seja, o tamanho do objeto "maça" é relevante para guiar atenção quando esta característica estiver cognitivamente associada às demais. É importante notar que, nesta simulação, nenhum ajuste no enviesamento *top-down* através do parâmetro W_k foi realizado, sendo utilizados os valores 0 ou 1, conforme descritos na Figura 5.45.



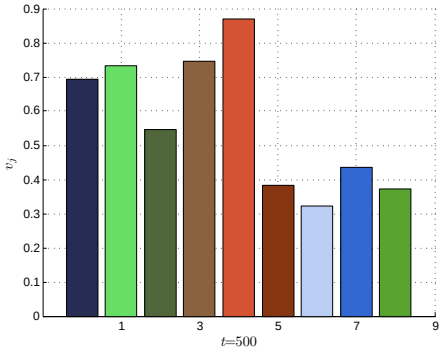
(a) Entrada (b) Segmentos (c) Mapa C (d) Mapa I (e) Mapa O_θ



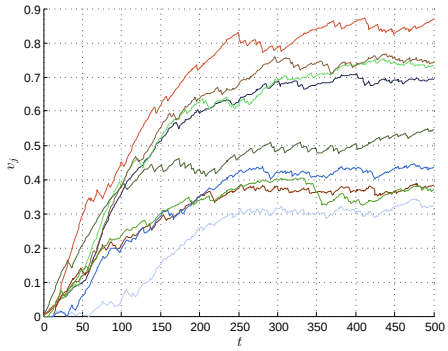
(f) Gráfico MOS



(g) Gráfico v_j



(h) Gráfico MOS

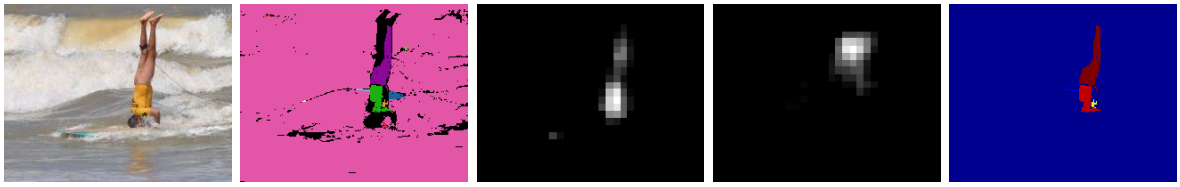


(i) Gráfico v_j

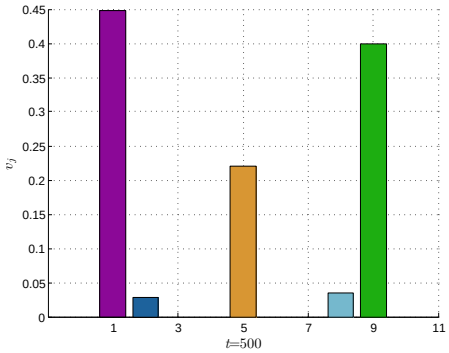
Figura 5.45: Modelo baseado em objetos. Simulação 5 - Busca conjuntiva. Valores de parâmetros utilizados: rede LEGION $\theta_p = 1200$ e $W_z = 20$, gerando um total de 9 segmentos e para a geração do MOS, $W_Y = 1.3$, $\theta_r = 0.5$, $\theta_s = 0$. Para (f) e (g) $W_{1..5} = 1$ e $W_6 = 0$. Para (h) e (i) $W_k = 1$ para $\forall\ k$. Imagem 200x200 *pixels*.

Na Figura 5.46 (a) é apresentada uma cena com alto contraste relacionado às características de cor e orientação. Os mapas de conspicuidades de cores e orientações são apresentados em (c) e (d), respectivamente. A rede LEGION gerou um total

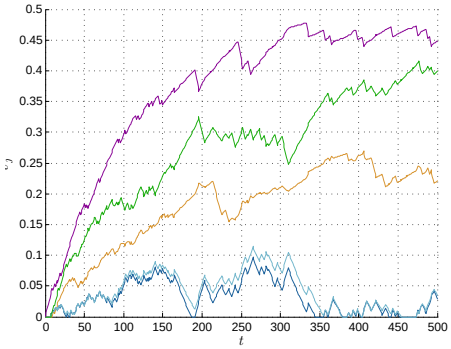
de 11 segmentos, porém, devido ao uso de $\theta_s = 0.01$, somente 5 segmentos estiveram aptos à competição pela atenção visual. O contraste em cores pode ser observado na camiseta do surfista, enquanto que suas pernas contrastam verticalmente com as ondas do mar. Para analisar o comportamento do modelo, nesta simulação, os valores de W_k foram definidos como: $W_2 = 1.0$, $W_3 = 1.0$ e $W_k = 0.0$ para os demais valores de k. A Figura 5.46 (e) apresenta o MOS gerado. De acordo com o comportamento do modelo, apresentado em (f) e (g), a identificação dos objetos saliência foram obtidas rapidamente, com $0 < t < 50$. Finalmente, consideramos a ordem de seleção visual, apresentada em (f), consistente com a inspeção visual humana.



(a) Entrada (b) Segmentos (c) Mapa C (d) Mapa O_θ (e) MOS



(f) Gráfico MOS



(g) Gráfico v_j

Figura 5.46: Modelo baseado em objetos. Simulação 6 - Cor e Orientação. Valores de parâmetros utilizados: rede LEGION $\theta_p = 400$ e $W_z = 10$, gerando um total de 11 segmentos e para a geração do MOS, $W_Y = 1$, $\theta_r = 0.5$, $\theta_s = 0.01$. Para (f) e (g) $W_2 = 1$ e $W_3 = 1$. Para (h) e (i) $W_k = 1$ para $\forall\ k$. Imagem 200x256 *pixels*.

De acordo com o parâmetro W_5, a atenção visual pode ser modulada para objetos de interesse na cena. Nas simulações seguintes, consideramos um motorista dirigindo por uma estrada, onde uma grande quantidade de informações, tanto *bottom-up* quanto *top-down*, podem emergir dinamicamente. Com isto, pretendemos demonstrar que o modelo de atenção visual proposto, juntamente com a rede HLC podem, efetivamente, contribuir para a resolução de problemas reais, especificamente àqueles pertencentes à área de atenção visual e reconhecimento de objetos. Neste contexto, a rede HLC foi treinada a partir de padrões representativos de placas de sinalização, apresentados na Tabela 5.2.

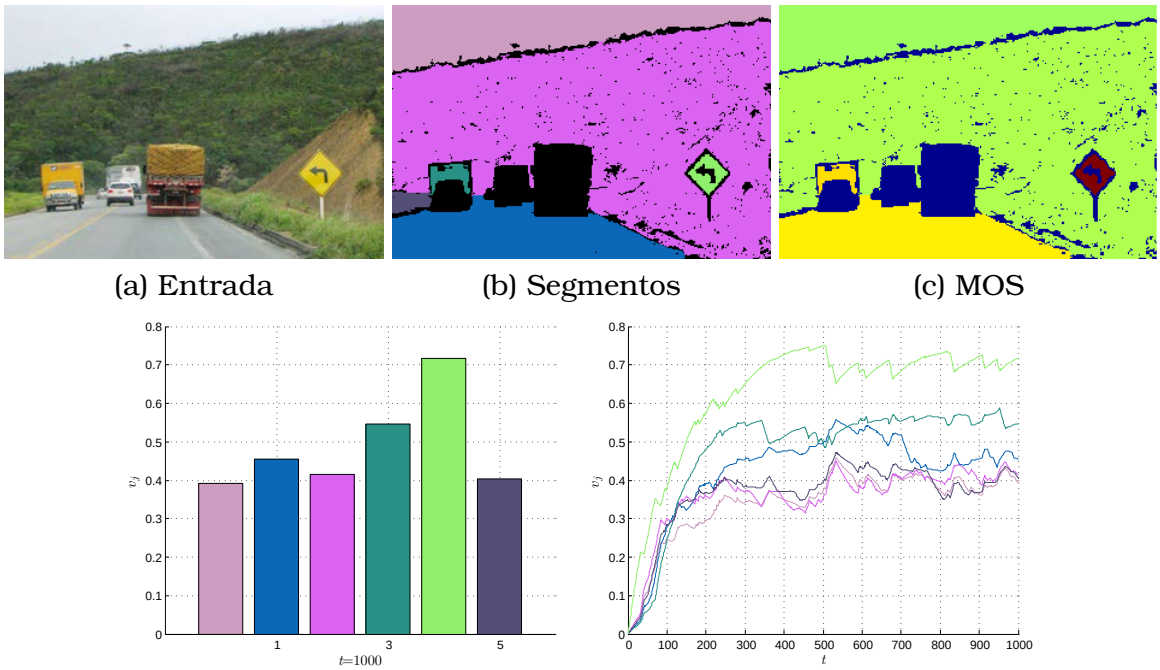
Conforme apresentado na seção anterior, quando um objeto pulsa na rede LEGION, este é apresentado para a HLC, responsável pela classificação do segmento. Na Tabela 5.2 são apresentados todos os segmentos gerados pela rede LEGION que

Tabela 5.2: Classificações geradas pelos classificadores de alto e baixo nível.

Segmentos	Classe Real	Valores Estimados de Classificação (Classes)					Predição Final	
		1	2	3	4	5	Low	HLC
Simulação 7							$\lambda = 0.00$	$\lambda = 0.08$
0	-	0.23 0.21	0.25 0.23	0.21 0.23	0.16 0.18	0.15 0.15	-	-
1	-	0.20 0.19	0.22 0.18	0.24 0.26	0.17 0.20	0.17 0.17	-	-
2	-	0.21 0.20	0.20 0.19	0.16 0.20	0.21 0.18	0.22 0.23	-	-
3	-	0.24 0.21	0.15 0.21	0.14 0.23	0.24 0.21	0.23 0.14	-	-
4	Classe 1	**0.39** 0.25	0.16 0.23	0.15 0.20	0.27 **0.27**	0.03 0.05	**Classe 4**	**Classe 1**
5	-	0.22 0.21	0.24 0.22	0.20 0.20	0.18 0.21	0.16 0.16	-	-
Simulação 8							$\lambda = 0.00$	$\lambda = 0.12$
0	-	0.23 0.21	0.25 0.20	0.22 0.20	0.20 0.20	0.10 0.19	-	-
1	-	0.22 0.22	0.24 0.20	0.18 0.16	0.20 0.24	0.16 0.18	-	-
2	-	0.19 0.20	0.23 0.18	0.21 0.22	0.22 0.21	0.15 0.19	-	-
3	-	0.17 0.19	0.19 0.18	0.22 0.23	0.16 0.15	0.26 0.25	-	-
4	Classe 2	0.25 0.27	**0.37** **0.28**	0.15 0.15	0.20 0.25	0.03 0.05	**Classe 2**	**Classe 2**
5	-	0.20 0.20	0.20 0.20	0.20 0.20	0.20 0.20	0.20 0.20	-	-
6	-	0.20 0.21	0.20 0.18	0.19 0.19	0.21 0.22	0.20 0.20	-	-
7	-	0.21 0.23	0.22 0.20	0.20 0.24	0.20 0.14	0.17 0.19	-	-
8	-	0.21 0.23	0.19 0.21	0.22 0.20	0.20 0.18	0.18 0.18	-	-
9	-	0.19 0.20	0.18 0.21	0.24 0.22	0.17 0.16	0.22 0.21	-	-

deverão ser classificados nas simulações 7 e 8. O limiar de reconhecimento utilizado em ambas as simulações foi definido como $\theta_R = 0.27$. Para fins de esclarecimento sobre a Tabela 5.2, as entradas da coluna "Valores Estimados de Classificação" são compostas por dois valores: o superior, representando o valor de classificação gerado pela rede HLC, e o valor logo abaixo, que apresenta a classificação gerada pelo classificador de baixo nível (*Low*). Os valores apresentados na coluna "Predição Final" indicam os resultados obtidos a partir do classificador de baixo nível ($\lambda = 0$) e da rede HLC, respectivamente. É importante notar que, quando maior o valor de λ, maior será o peso atribuído à saída da rede HLC.

Na simulação 7, apresentada na Figura 5.47, foram utilizados os seguintes valores de parâmetros: $\lambda = 0.08$ e $R_j \geq \theta_R$. Neste caso, somente o segmento 4 (classe real 1) atingiu o limiar de reconhecimento. Como pode ser visto na Tabela 5.2, o segmento 4 foi classificado incorretamente pelo classificador de baixo nível, entretanto, na rede HLC, o segmento 4 foi classificado corretamente, com o valor de 0.39 para a classe 1. Nesta simulação, o parâmetro de enviesamento *top-down* W_k foi considerado, de forma que, a atenção seja direcionada, principalmente, para os objetos reconhecidos. Sendo assim, os valores de W_k foram definidos como: $W_1 = 0.6$, $W_2 = 0.4$, $W_3 = 0.2$, $W_4 = 0$, $W_5 = 1.0$, $W_6 = 0.0$. De acordo com o comportamento do modelo, apresentado na Figura 5.47 (f) e (g), a seleção do objeto mais saliente (placa de sinalização) ocorreu corretamente, com $0 < t < 50$.



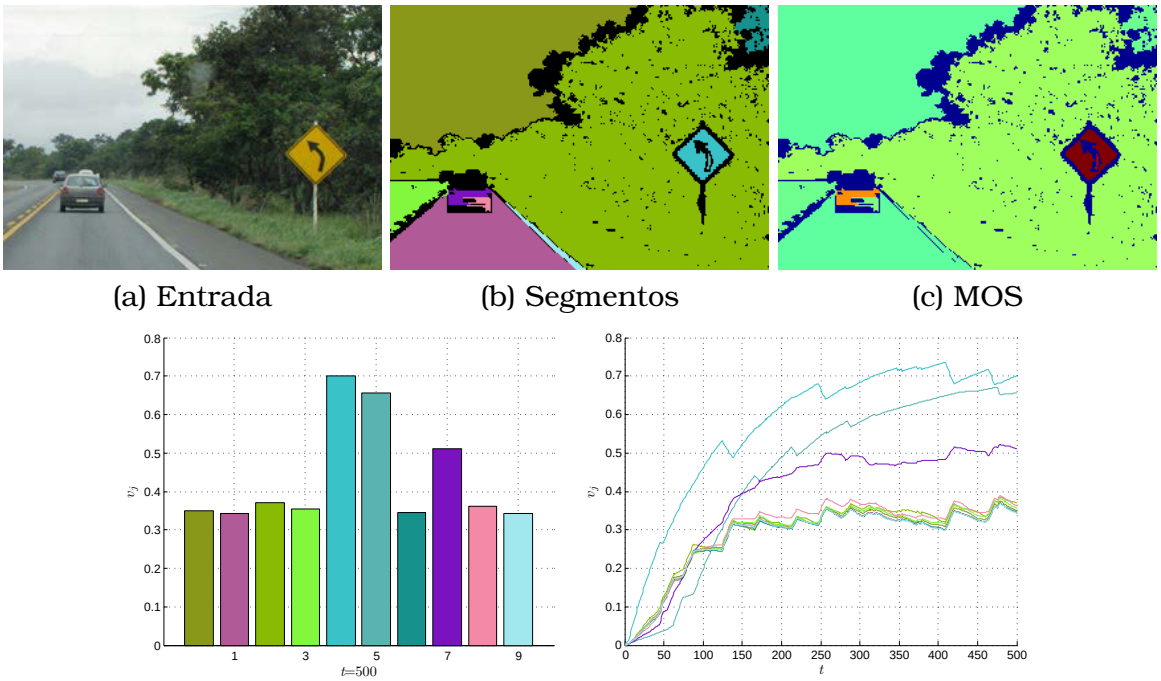
(a) Entrada (b) Segmentos (c) MOS

(d) Gráfico MOS (e) Gráfico v_j

Figura 5.47: Modelo baseado em objetos. Simulação 7 - Reconhecimento de Placas de Sinalização. Valores de parâmetros utilizados: rede LEGION $\theta_p = 600$ e $W_z = 10$, gerando um total de 6 segmentos e para a geração do MOS, $W_Y = 0.8$, $\theta_r = 0.27$, $\theta_s = 0$. Imagem 211x315 *pixels*.

Na simulação 8 (Figura 5.48), o parâmetro λ foi definido com o valor de 0.12.

Assim como na simulação anterior, somente o segmento 4 (classe real 2) foi reconhecido. Neste caso, ambos os classificadores identificaram corretamente a classe real do segmento 4, porém, a rede HLC apresentou um maior valor de reconhecimento. De modo a testar a robustez do modelo proposto, os valores atribuídos ao parâmetro W_k foram os mesmos utilizados na simulação 7. Como pode ser observado na Figura 5.48 (g), o objeto alvo (placa de sinalização) foi o primeiro a ser selecionado pelo modelo, corroborando com os objetos desta proposta.



(a) Entrada (b) Segmentos (c) MOS

(d) Gráfico MOS (e) Gráfico v_j

Figura 5.48: Modelo baseado em objetos. Simulação 8 - Reconhecimento de Placas de Sinalização. Valores de parâmetros utilizados: rede LEGION $\theta_p = 600$ e $W_z = 10$, gerando um total de 10 segmentos e para a geração do MOS, $W_Y = 0.8$, $\theta_r = 0.27$, $\theta_s = 0$. Imagem 200x286 *pixels*.

5.3.2 Enviesamento *Top-Down* e Atenção Visual Baseada em Objetos

Como uma evolução do modelo apresentado na Seção 5.3.1, este modelo tem como principal objeto a proposta de um modelo de atenção visual baseado em objetos com características *bottom-up* e *top-down*. Nos modelos com características *top-down*. propostos anteriormente nesta tese, a atenção *top-down* está relacionada às intenções do observador, podendo ser visto como um processo de modulação, através do mecanismo de seleção visual. Caso o observador esteja à procura de uma forma ou cor específica, modulações *top-down* permitem o enviesamento do processo de busca do segmento com maior propensão a alvo. Assim, objetos que não apresentem características desejadas são definidos com baixos valores de potenciais de saliência. Contudo, estes segmentos mantêm-se ativos durante todo processo de seleção visual, de forma que possam ser selecionados, em uma ordem decrescente

de saliência.

Neste trabalho, considerando cenas onde se conheça previamente quais informações deseja-se encontrar, propomos um mecanismo para que somente segmentos candidatos a alvo estejam ativos e aptos à competição pela atenção visual. Os resultados deste trabalho encontram-se publicados em (Benicasa et al., 013b).

Sendo assim, e tratando-se de uma evolução do modelo apresentado na seção anterior, apresentaremos a seguir a descrição do modelo, a partir das modificações propostas.

Descrição do Modelo

O modelo proposto para a seleção de objetos salientes, conforme diagrama apresentado na Figura 5.49, foi composto pelos seguintes módulos: a extração de características visuais primitivas, responsável pela extração dos mapas de conspicuidades, contendo as características visuais primitivas da cena, o enviesamento *top-down* da segmentação LEGION, que atribui pesos específicos (*top-down*) aos mapas de conspicuidades, o reconhecimento de objetos e, finalmente, a geração do mapa de objeto-saliente (MOS), que torna possível a seleção visual dos objetos salientes.

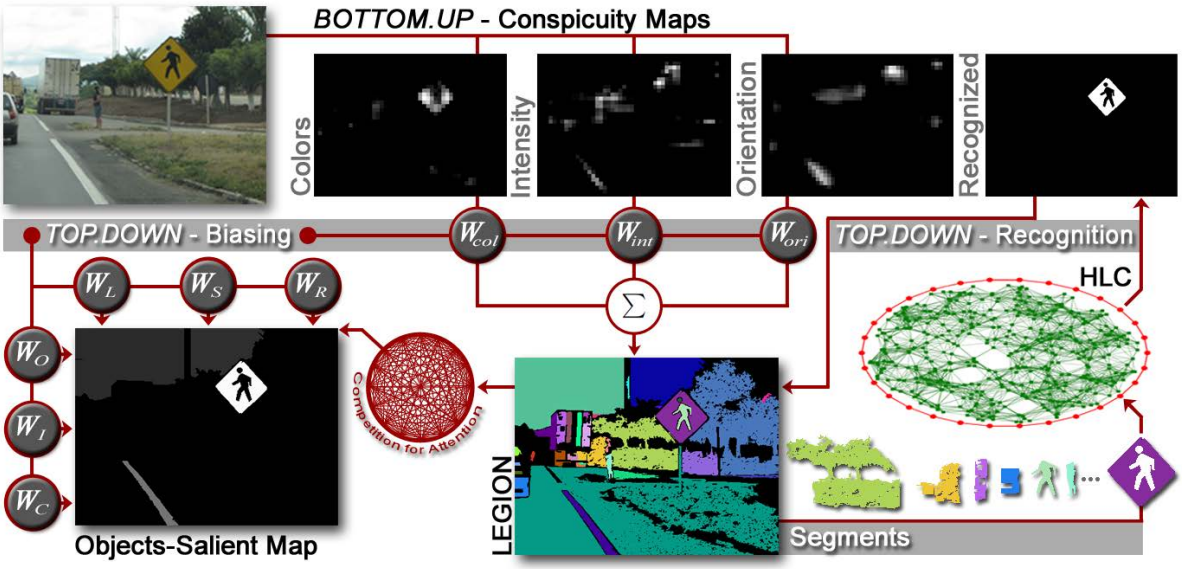


Figura 5.49: Diagrama do modelo de seleção baseada em objetos II.

Inicialmente, o enviesamento *top-down* proposto é definido pela associação de pesos aos mapas de conspicuidades. Desta forma, os valores de saliência dos mapas de conspicuidades são pesadas e combinados em um mapa de saliência S_m, definido como:

$$S_m = \frac{1}{3}\left(W_{int}C_{int} + W_{cor}C_{cor} + W_{ori}C_{ori}\right), \tag{5.37}$$

onde C_{int}, C_{cor} e C_{ori} representam, respectivamente, os mapas de conspicuidades

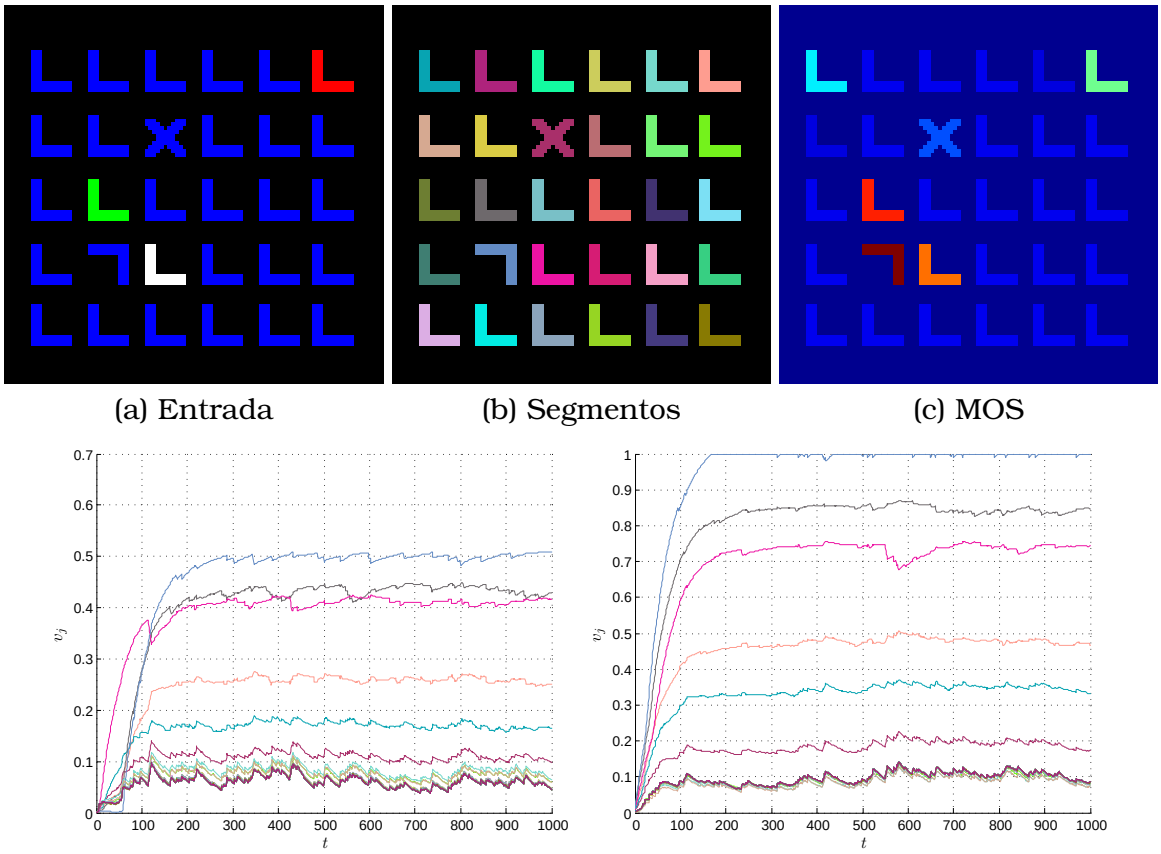
de intensidades, cores, e orientações, e W_{int}, W_{cor} e W_{ori}, determinam os pesos atribuídos a cada mapa de conspicuidade.

De acordo com os conceitos apresentados na Seção 3.2.3, a segmentação LEGION é baseada em potenciais laterais $p_i \geq \theta_p$, o que permite um neurônio i tornar-se líder de um segmento e pulsar. Entretanto, para o enviesamento *top-down* deste trabalho, propomos que, um oscilador i definido como líder, somente poderá pulsar se seu valor de saliência S_{m_i} exceder o limiar de enviesamento, denominado por θ_{bias}. Esta condição garante a segmentação de regiões com valores de saliência desejados, permitindo a redução do número de segmentos gerados e, consequentemente, reduzindo custos computacionais para a segmentação de regiões não salientes. É importante notar que, embora os pesos *top-down* sobre determinados alvos possam ser obtidos automaticamente a partir do próprio alvo (Navalpakkam and Itti, 2005; Frintrop et al., 2010; Borji et al., 2011), consideramos neste trabalho os valores dos parâmetros W_{int}, W_{cor}, W_{ori} e do limiar θ_{bias}, definidos de acordo com o domínio em questão.

Os demais módulos que compõem este modelo são baseados na descrição apresentada na proposta anterior (Seção 5.3.1). Entretanto, consideramos uma pequena modificação na Equação 5.34, responsável pela geração do mapa de objeto-saliente, sendo re-escrita como segue:

$$\dot{v}_j = -v_j + E_j - W_Y Y_j + \sum_{k=1}^{6} W_k S_k^j, \qquad j = 1, \ldots, n^j, \tag{5.38}$$

onde v_j representa o potencial de saliência do objeto j, E_j é o termo de acoplamento excitatório e W_Y é o peso de inibição aplicado ao termo de acoplamento inibitório Y_i. A modificação realizada é apresentada no último termo da equação, representada pelo somatório dos valores de saliência do segmento ativo S_k^j, enviesado pelo peso W_k, associado a cada característica. Embora a competição pela atenção visual ocorra baseada nos valores de enviesamento e medidas de similaridade entre os objetos concorrentes, a inserção do termo apresentado tem como objetivo proporcionar maior sensibilidade ao modelo proposto, em relação ao enviesamento *top-down*, uma vez que os valores dos enviesamentos são imputados diretamente na equação que define v_j. Para demonstrar o comportamento do modelo diante desta modificação, foram realizadas duas simulações utilizando valores de parâmetros idênticos. Entretanto, para a geração do MOS, a primeira simulação é baseada na Equação 5.34, apresentada na Seção 5.3.1 e, na segunda, o MOS é gerado a partir da Equação 5.38. De acordo com os gráficos (d) e (e), apresentados na Figura 5.50, podemos concluir que, o termo proposto tornou o modelo mais estável em relação aos valores de oscilações dos objetos, apresentou maior precisão para a seleção visual, além de aumentar a taxa de seleção dos objetos salientes da cena, com $0 < t < 50$.



(a) Entrada (b) Segmentos (c) MOS

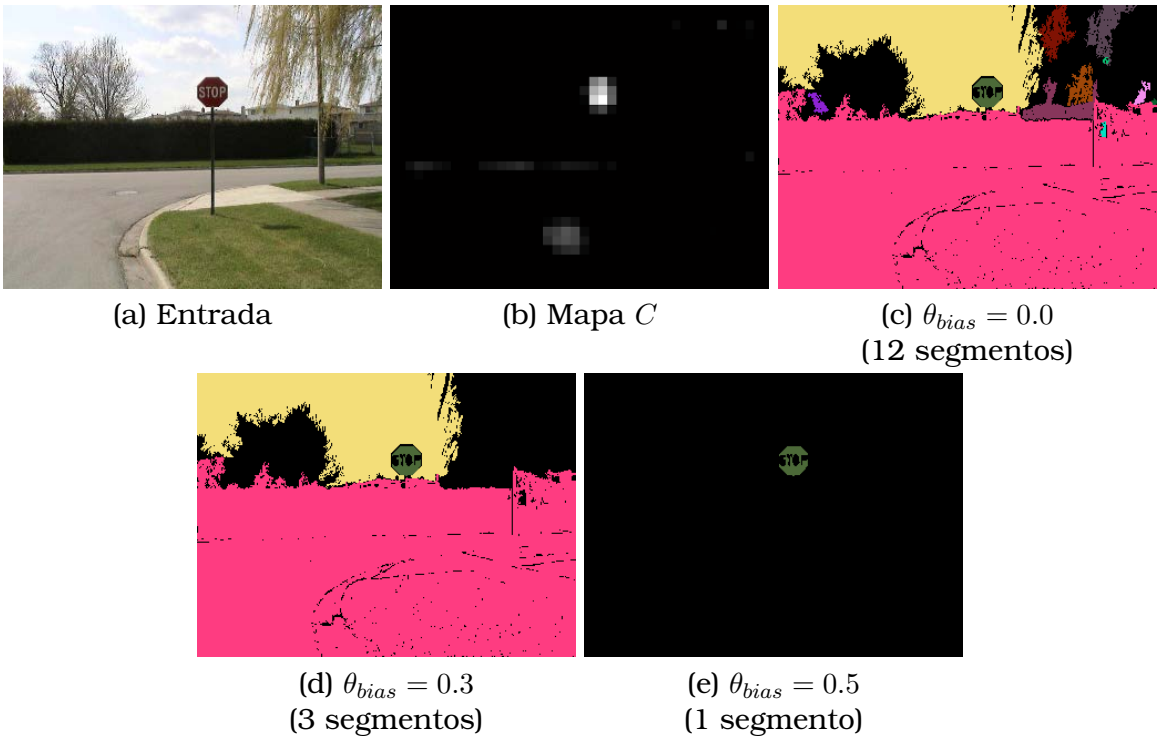
(d) Gráfico v_j (Equação 5.34) (e) Gráfico v_j (Equação 5.38)

Figura 5.50: Comportamento do modelo baseado nas Equações 5.34 e 5.38. Valores de parâmetros utilizados: enviesamento *top-down*, $W_{int} = 1$, $W_{col} = 1$, $W_{ori} = 1$, $\theta_{bias} = 0$, $W_1 = 0.3$, $W_2 = 0.4$, $W_3 = 0$, $W_4 = 0$, $W_5 = 0.5$ e $W_6 = 0.0$; rede LEGION, $\theta_p = 1200$ e $W_z = 20$, gerando um total de 30 segmentos e, para a geração do MOS, $W_Y = 1.0$, $\theta_r = 0.5$, $\theta_s = 0$. Imagem 100x100 *pixels*.

Simulações Computacionais

Um importante tipo de informação *top-down* é o conhecimento prévio sobre o alvo, utilizado para a busca visual (Frintrop et al., 2010). Nas simulações apresentadas nesta seção, com o objetivo de demonstrar o comportamento do modelo proposto em relação ao enviesamento do processo de segmentação, informações sobre o alvo desejado serão utilizadas.

Inicialmente, na Figura 5.51 é apresentada uma cena com alto contraste relacionado à característica cor ("placa de sinalização"). Neste caso, podemos considerar que a busca por placas de sinalização está relacionada, principalmente, à alvos que apresentem alto contraste em cores. Assim, o enviesamento *top-down* foi definido com os seguintes valores de parâmetros: $W_{int} = 0.0$, $W_{col} = 1.0$, $W_{ori} = 0.0$ e $\theta_{bias} = [0, \ldots, 0.5]$. É importante notar que, variações nos valores destes parâmetros podem alterar, significantemente, o número de segmentos a participar da competição pela atenção. Na Figura 5.51 (c), (d) e (e) são apresentados os segmentos obtidos a partir de variações do parâmetro θ_{Bias}.



(a) Entrada (b) Mapa C (c) $\theta_{bias} = 0.0$ (12 segmentos)

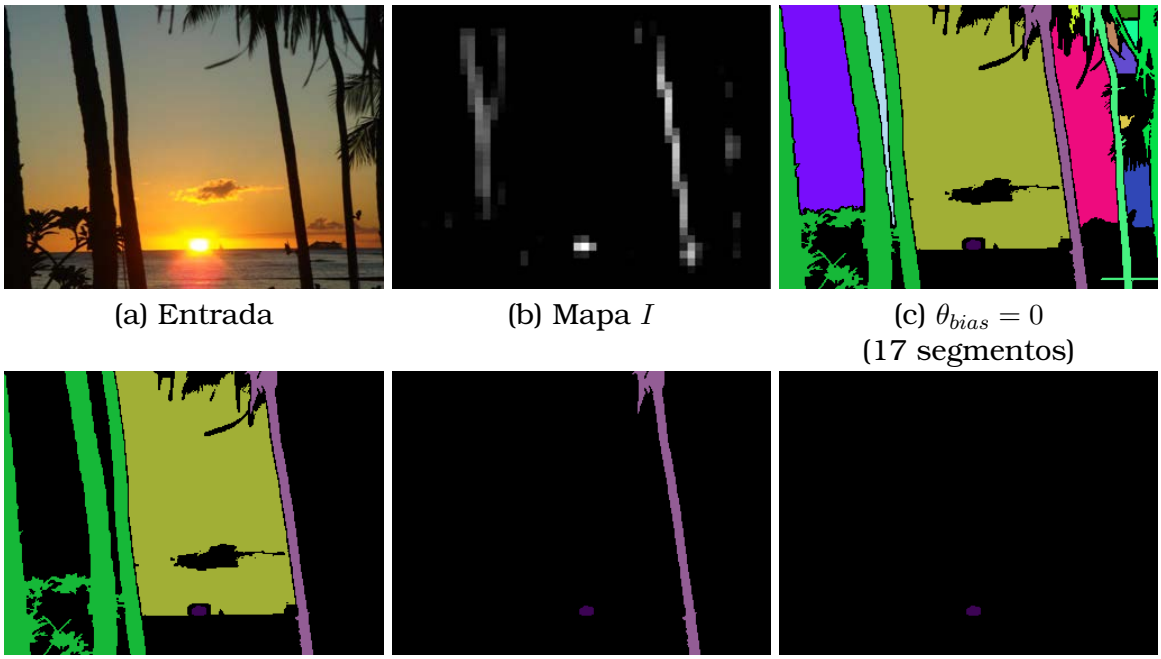
(d) $\theta_{bias} = 0.3$ (3 segmentos) (e) $\theta_{bias} = 0.5$ (1 segmento)

Figura 5.51: Modelo Baseado em Objetos II. Simulação 1 - Enviesamento *top-down* baseado no mapa de conspicuidades de cores. Valores de parâmetros utilizados: enviesamento *top-down*, $W_{int} = 0$, $W_{cor} = 1$, $W_{ori} = 0$, $\theta_{Bias} = [0, \ldots, 0.5]$, e rede LEGION, $\theta_p = 800$ e $W_z = 20$. Imagem 256x342*pixels* do *benchmark* disponibilizado publicamente por Bruce and Tsotsos (2009).

Na Simulação 2, apresentada na Figura 5.52, pode ser observado o uso do enviesamento *top-down*, a partir da característica de intensidades. Neste caso, os valores dos parâmetros de enviesamento foram definidos como: $W_{int} = 1$, $W_{cor} = 0$, $W_{ori} = 0$ e $\theta_{bias} = [0, \ldots, 0.9]$. Conforme apresentado na descrição desta proposta, o principal objetivo da utilização do enviesamento *top-down* para o processo de segmentação, é a exclusão de segmentos com baixos valores de saliência. Conforme apresentado na Figura 5.52 (c), e com a utilização de $\theta_{bias} = 0$, a rede LEGION gerou um total de 17 segmentos, contudo, somente alguns segmentos apresentam valores significativos de saliência. Embora uma solução para esta questão possa ser obtida através do ajuste do limiar de saliência (θ_s), o uso do limiar de enviesamento θ_{bias} permite a redução do custo computacional e também torna possível a maximização da taxa de seleção de objetos.

5.4 Considerações Finais

Neste capítulo foram apresentados os modelos computacionais propostos nesta tese para atenção visual. Foram propostos um total de cinco modelos, contendo características originais e também extensões de modelos anteriores. Dentre as principais características propostas, destacamos inicialmente o MAS, que permitiu a



(a) Entrada (b) Mapa I (c) $\theta_{bias} = 0$ (17 segmentos)

(d) $\theta_{bias} = 0.4$ (4 segmentos) (e) $\theta_{bias} = 0.6$ (2 segmentos) (e) $\theta_{bias} = 0.9$ (1 segmento)

Figura 5.52: Modelo Baseado em Objetos II. Simulação 1 - Enviesamento *top-down* baseado no mapa de conspicuidades de intensidades. Valores de parâmetros utilizados: enviesamento *top-down*, $W_{int} = 1$, $W_{cor} = 0$, $W_{ori} = 0$, $\theta_{Bias} = [0, \ldots, 0.9]$, e rede LEGION, $\theta_p = 800$ e $W_z = 20$. Imagem 253x338*pixels* do *benchmark* disponibilizado publicamente por Judd et al. (2012).

auto-organização de características distintas de um objeto e também, a partir do processo de competição por atenção proposto, a localização de objetos salientes, de modo que o desenvolvimento da atenção visual tenha sido baseado inicialmente em informações *bottom-up*. Mecanismos para a enviesamento da atenção visual a partir de informações *top-down* foram apresentados nos modelos seguintes, considerando sua aplicabilidade tanto para cenas sintéticas quanto para cenas reais. Por fim, foram apresentados os modelos para atenção visual baseados no MOS, o que permitiu a redução do esforço computacional para o desenvolvimento da seleção visual.

No capítulo seguinte serão apresentadas análises qualitativas e quantitativas dos resultados obtidos para um total de 104 simulações, compostas por imagens reais, sintéticas psicofísicas e de domínio específico, comparados aos mapas de fixação, gerados através da atenção visual humana, e também a mapas de saliência, gerados por quatro outros modelos de atenção visual.

CAPÍTULO

6

Análise dos Modelos Propostos

Considerando a evolução natural do modelos apresentados no Capítulo 5, utilizaremos aqui, para fins comparativos, as saídas geradas pelos modelos propostos apresentados nas Seções 5.3.1 e 5.3.2.

De acordo com Judd et al. (2012), diversos modelos de atenção visual têm sido propostos com o objetivo de predizer locais da cena em que humanos direcionam a atenção visual, contudo, para cada novo modelo introduzido, o mecanismo de atenção proposto é avaliado a partir de novas imagens, o que torna difícil a comparação dos resultados gerados. Assim, para a análise qualitativa e quantitativa dos resultados obtidos a partir dos modelos propostos nesta tese, utilizaremos mapas de fixação (FM)[1], gerados através do rastreamento dos movimentos dos olhos de observadores humanos, para uma variedade de imagens, disponibilizadas publicamente por diversos pesquisadores (veja (Judd et al., 2012) para uma listagem abrangente). Dentre estes, devido a sua ampla variedade de imagens reais em diferentes domínios, selecionamos para análise qualitativa e quantitativa, um total de sessenta e uma cenas reais, obtidas dos *benchmarks* propostos em Bruce and Tsotsos (2009) (36 cenas) e Judd et al. (2012) (25 cenas). Analisamos também o comportamento de modelo em relação a uma variedade de resultados clássicos derivados da literatura psicofísica, disponibilizados no *benchmark* de Bruce and Tsotsos (2009) (total de 15 imagens sintéticas). Por fim, apresentamos a análise do modelo proposto, baseado porém em um domínio específico, com o objetivo de demonstrar a robustez do modelo em relação ao ajuste de parâmetros. Neste caso, os parâmetros são ajustados inicialmente e aplicados em um conjunto de cenas propostas para esta tese (total de 28 cenas).

De acordo com as características biológicas desta tese, focamos nossa pes-

[1] do inglês Fixations Maps

quisa em modelos de atenção visual baseados nos processos cognitivos de atenção, conforme apresentado na Seção 4. Entretanto, para fins comparativos, além dos mapas de fixações e do mapa de saliência proposto por Itti et al. (1998), analisamos também alguns dos principais modelos de atenção visual propostos nos últimos anos, de acordo com Judd et al. (2012). Outro fator considerado foi a disponibilidade dos códigos destes modelos por parte dos autores, facilitando a comparação dos resultados obtidos. Sendo assim, a análise apresentada nesta seção baseia-se na comparação do mapa de objeto-saliente (MOS) proposto, em relação ao mapa de fixação humana (FM) e os mapas de saliências gerados a partir dos modelos propostos por Itti et al. (1998)[2], Harel et al. (2006)[3], Achanta et al. (2009)[4] e Cheng et al. (2011)[5], categorizados respectivamente como: modelo cognitivo, modelo baseado em grafos e, para os dois últimos, modelos baseados em análises espectrais.

De maneira elucidativa, consideramos a descrição das principais características dos modelos utilizados para análise nesta seção, de forma que o primeiro, proposto por Itti et al. (1998), encontra-se descrito na Seção 3.1. No segundo, proposto por Harel et al. (2006), foi introduzido um modelo de saliência baseado em grafos (GBVS - *Graph-Based Visual Saliency*). No GBVS, os mapas de características (intensidades, cores e orientações) são extraídos de maneira similar à apresentada em (Itti et al., 1998), onde, para cada mapa de característica, um grafo totalmente conectado é utilizado, de modo que nodos são responsáveis pela representação de todas as localizações. Pesos entre dois nodos são associados proporcionalmente à similaridade em relação às suas características e pesadas por suas distâncias espaciais. Em um processo de normalização, nodos que apresentem altos valores de dissimilaridades em relação à vizinhança são definidos com altos valores de saliência. Os mapas são finalmente normalizados para enfatizar detalhes conspicuosos, e combinados em um único mapa de saliência. O modelo proposto por Achanta et al. (2009) apresenta como principal objetivo a segmentação de regiões salientes, propósito bastante comum aos apresentados nesta tese. O mecanismo de saliência proposto por Achanta et al. (2009) é definido como: $S(x,y) = |I_\mu - I_{\omega_{hc}}|$, onde I_μ representa a média de cores da imagem I, $I_{\omega_{hc}}$ é uma versão Gaussiana (5x5) da imagem original, $|.|$ representa a distância Euclidiana e x, y são as coordenadas dos *pixels*. O último trabalho considerado para análise comparativa, proposto por Cheng et al. (2011), baseia-se inicialmente na segmentação da imagem em regiões (Felzenszwalb and Huttenlocher, 2004) e, em seguida, valores de saliência são associados a cada região. De acordo com Cheng et al. (2011), o valor de saliência de cada região é calculado a partir do contraste de cores da região em relação à distância espacial das demais regiões.

[2] http://www.saliencytoolbox.net/
[3] http://www.klab.caltech.edu/~harel/share/gbvs.php
[4] http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/index.html
[5] http://cg.cs.tsinghua.edu.cn/people/~cmm/Saliency/

6.1 Domínios Heterogêneos

Para o primeiro conjunto de simulações realizadas, o conhecimento sobre as características primitivas do alvo são utilizados para o enviesamento do processo de segmentação. Sendo assim, nas Figuras 6.1 à 6.13 são apresentadas 61 cenas de diferentes domínios, contendo objetos contrastantes em relação às seguintes características: intensidades, cores, orientações, localização e tamanho. O valor de reconhecimento do objeto não foi considerado como informação para o enviesamento da atenção visual, uma vez que as cenas apresentadas neste conjunto de simulações tratam de domínios distintos. Em uma análise qualitativa dos resultados apresentados nas Figuras 6.1 à 6.13, podemos concluir que os resultados apresentados nos mapas de objetos-salientes (MOS) são bastante coerentes com os resultados apresentados nos mapas de fixação de Judd et al. (2012). É importante notar que, a comparação qualitativa deve ocorrer entre o mapa de fixação e cada um dos mapas de saliência apresentados.

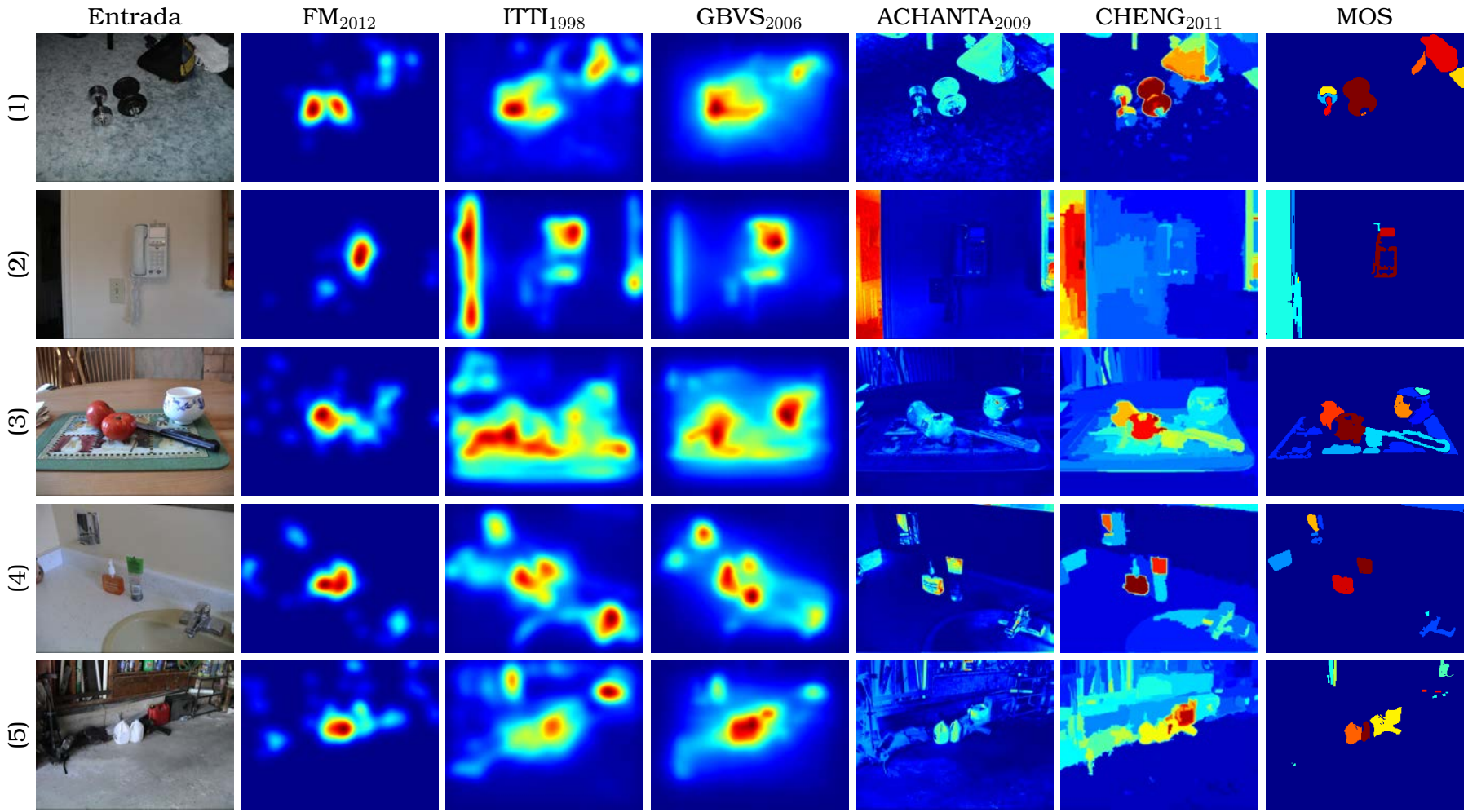


Figura 6.1: Análise qualitativa (1-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

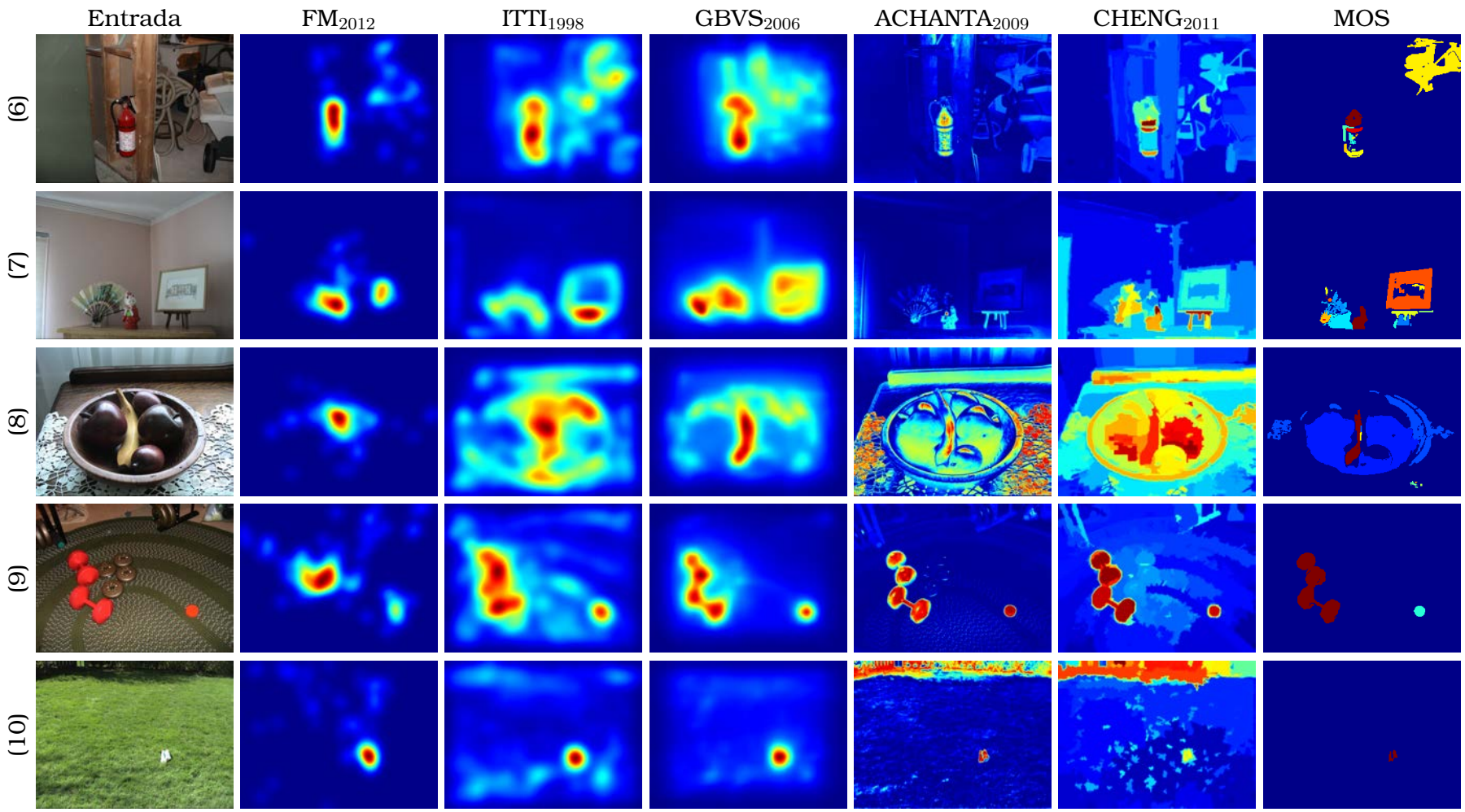


Figura 6.2: Análise qualitativa (2-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

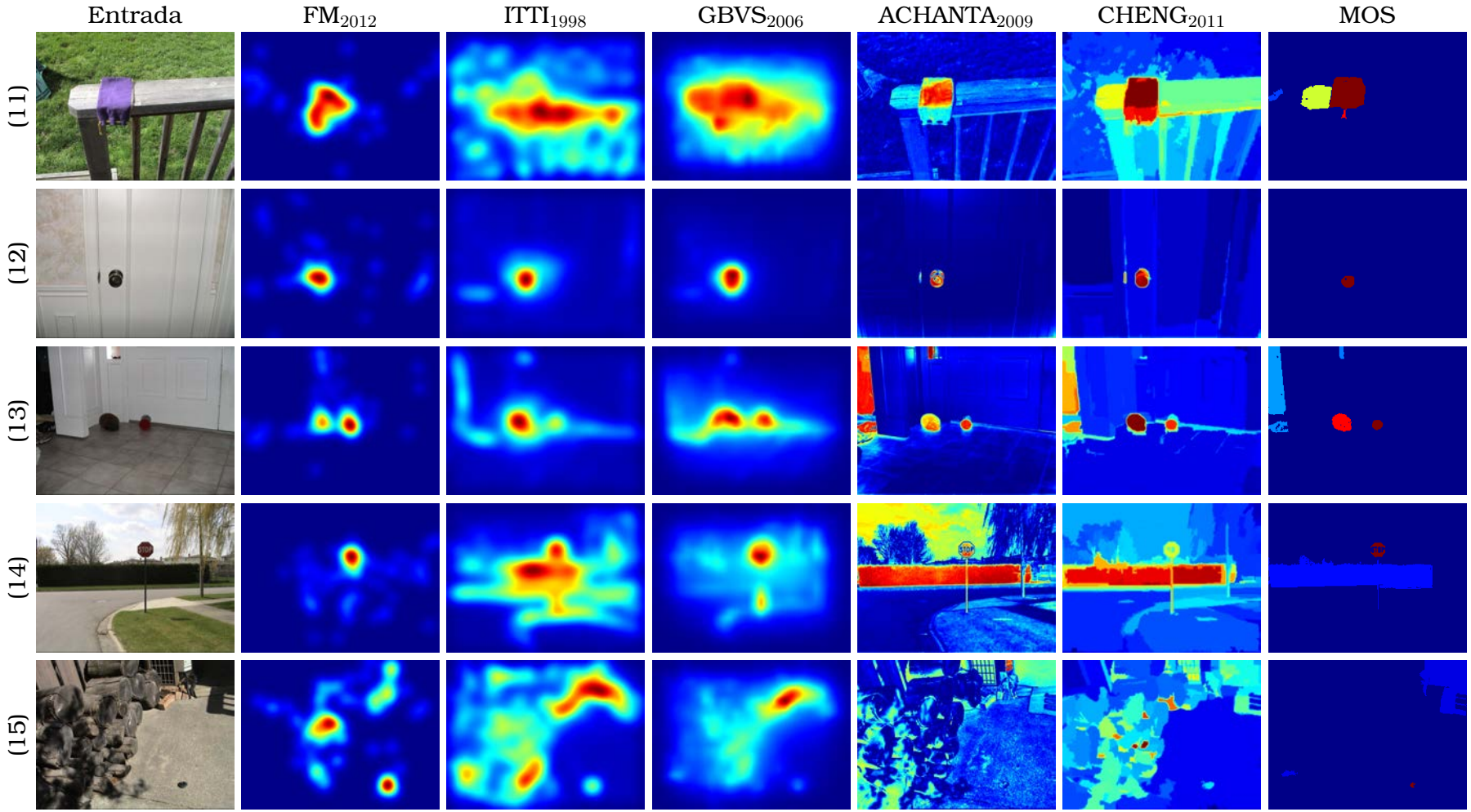


Figura 6.3: Análise qualitativa (3-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

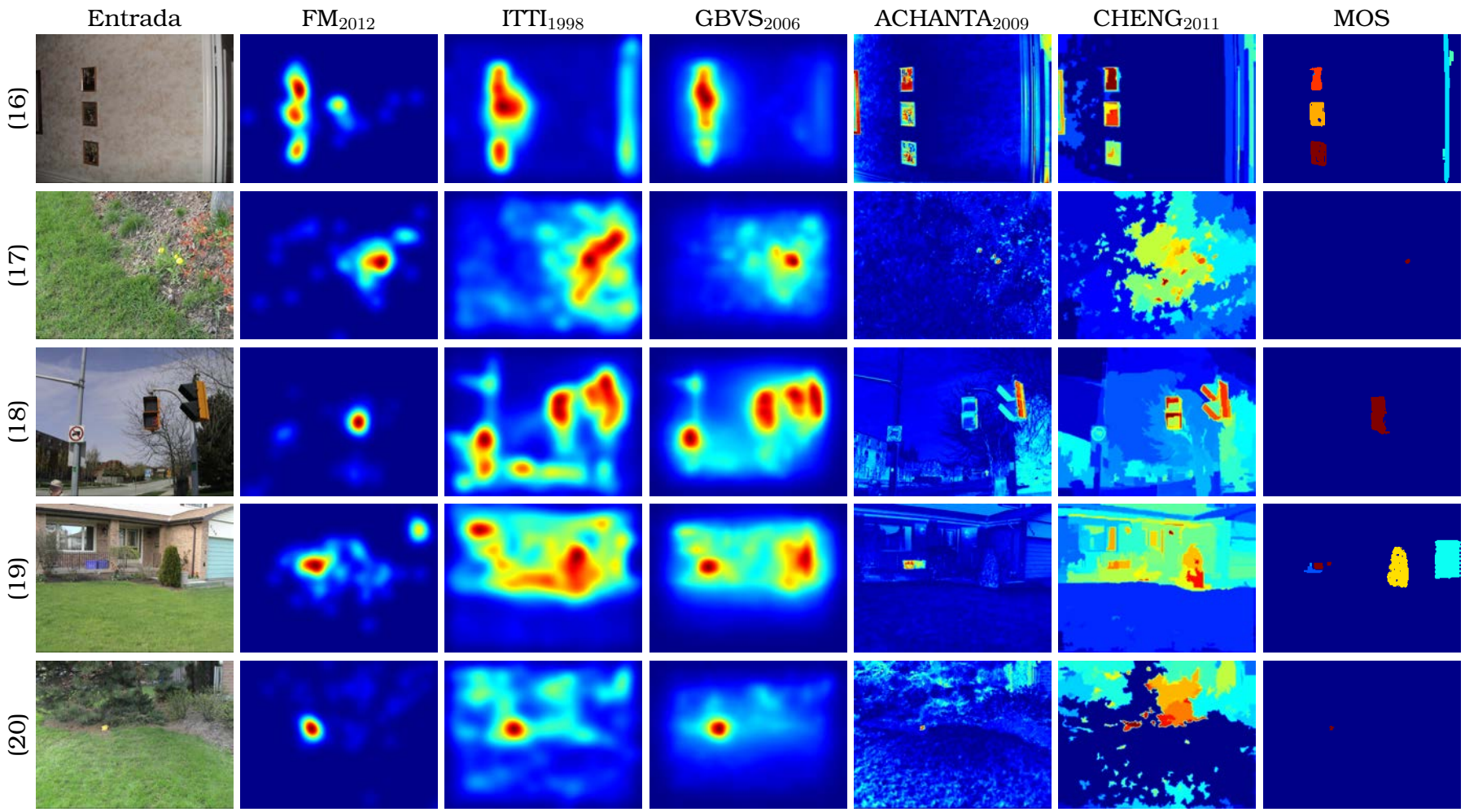


Figura 6.4: Análise qualitativa (4-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

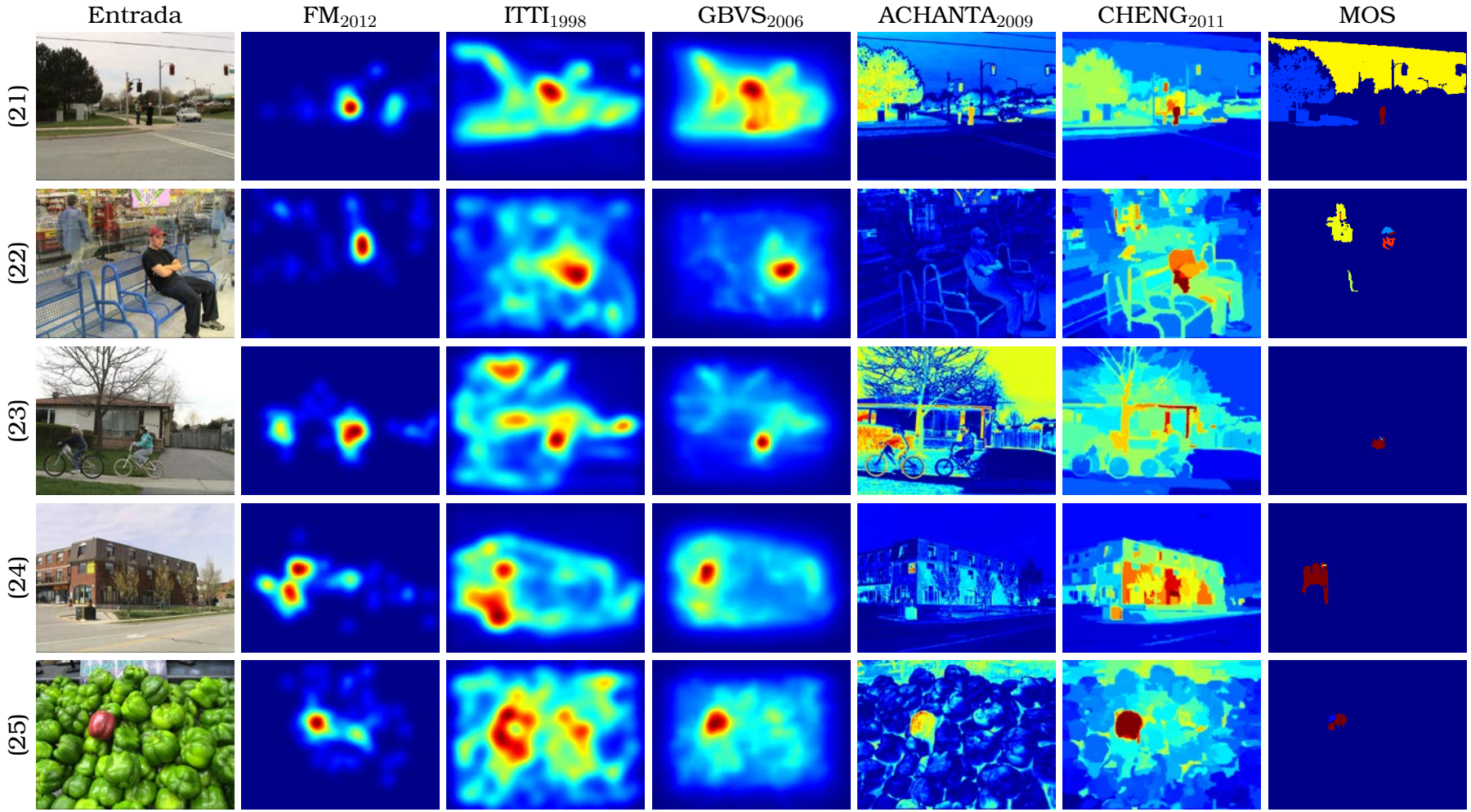


Figura 6.5: Análise qualitativa (5-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

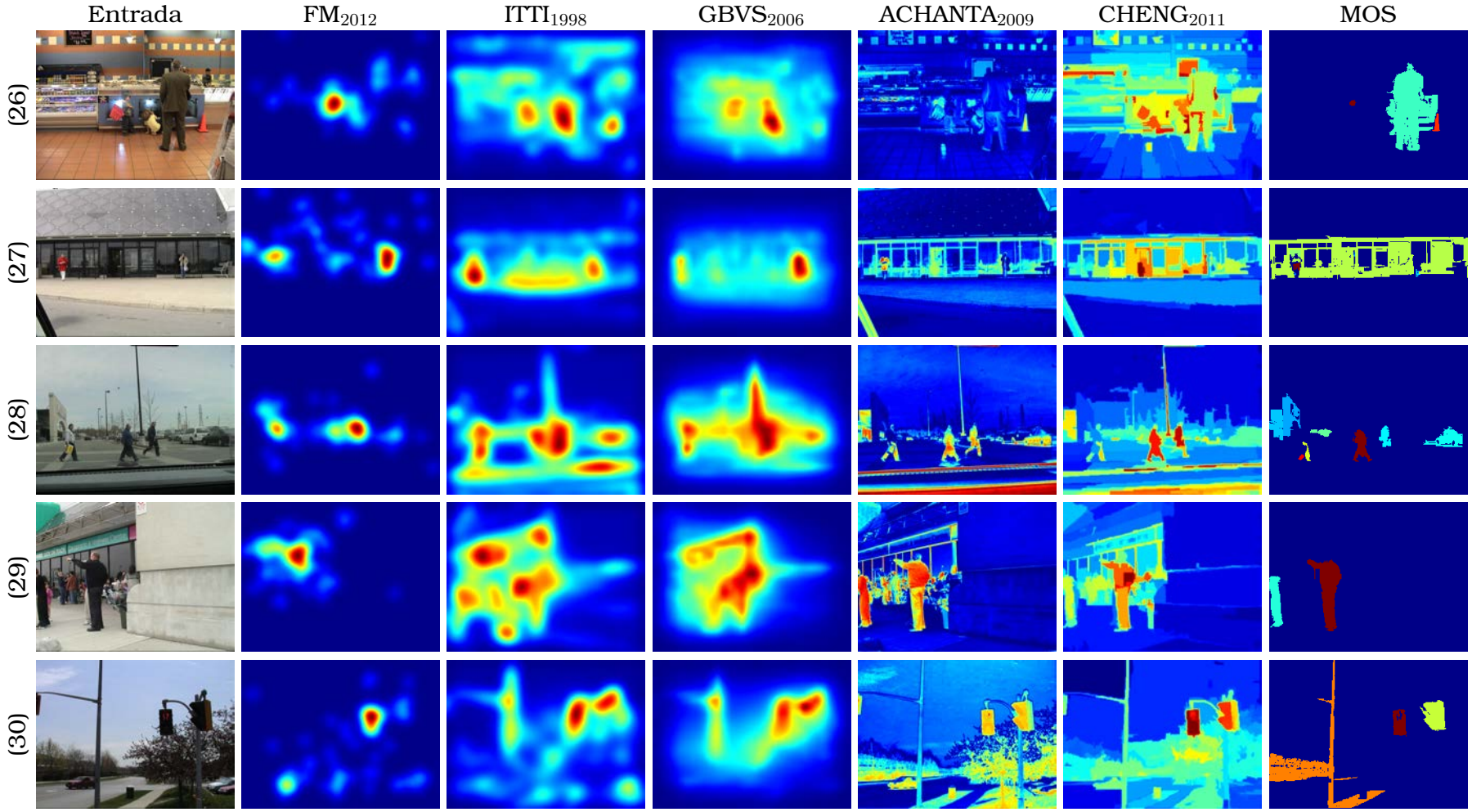


Figura 6.6: Análise qualitativa (6-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

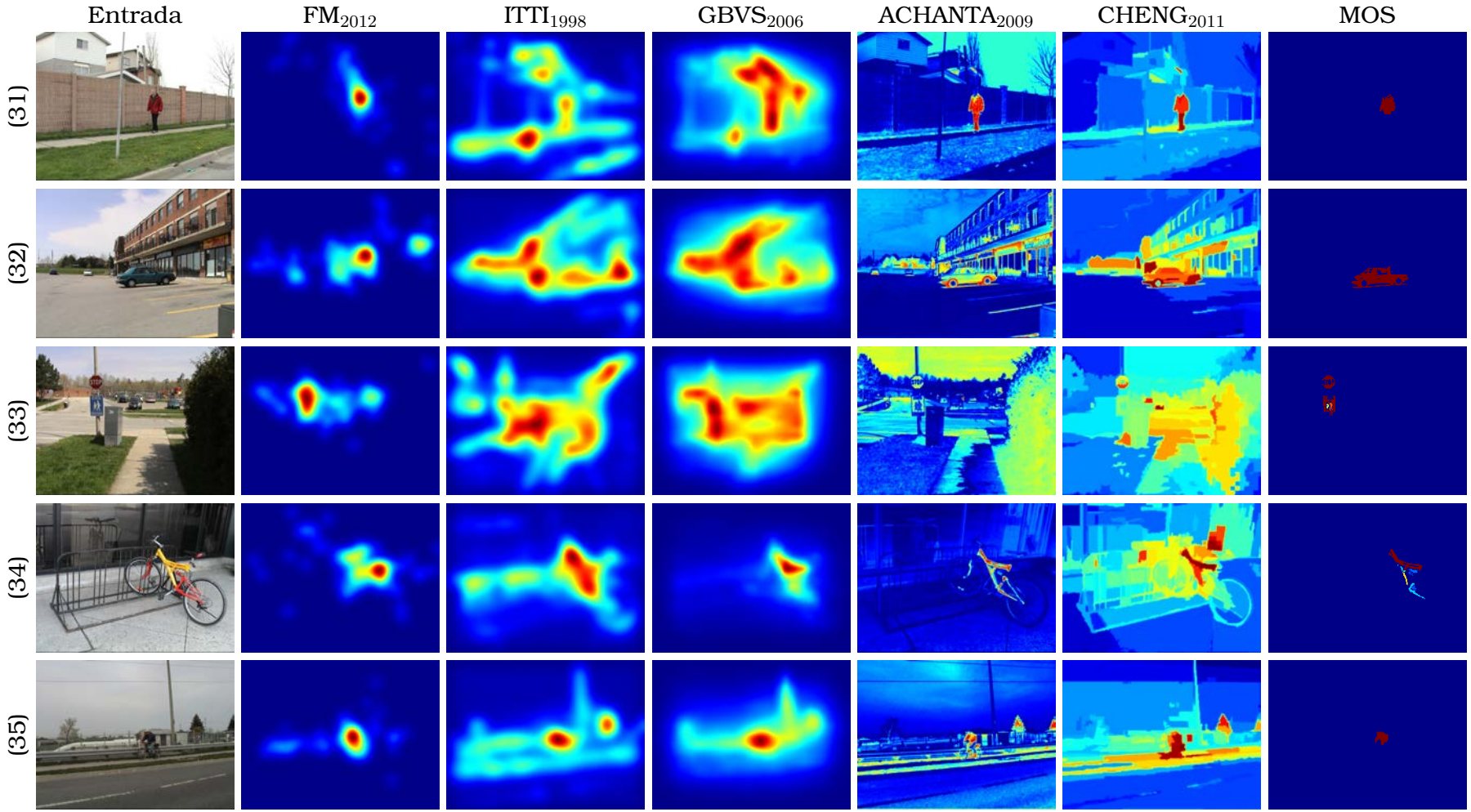


Figura 6.7: Análise qualitativa (7-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

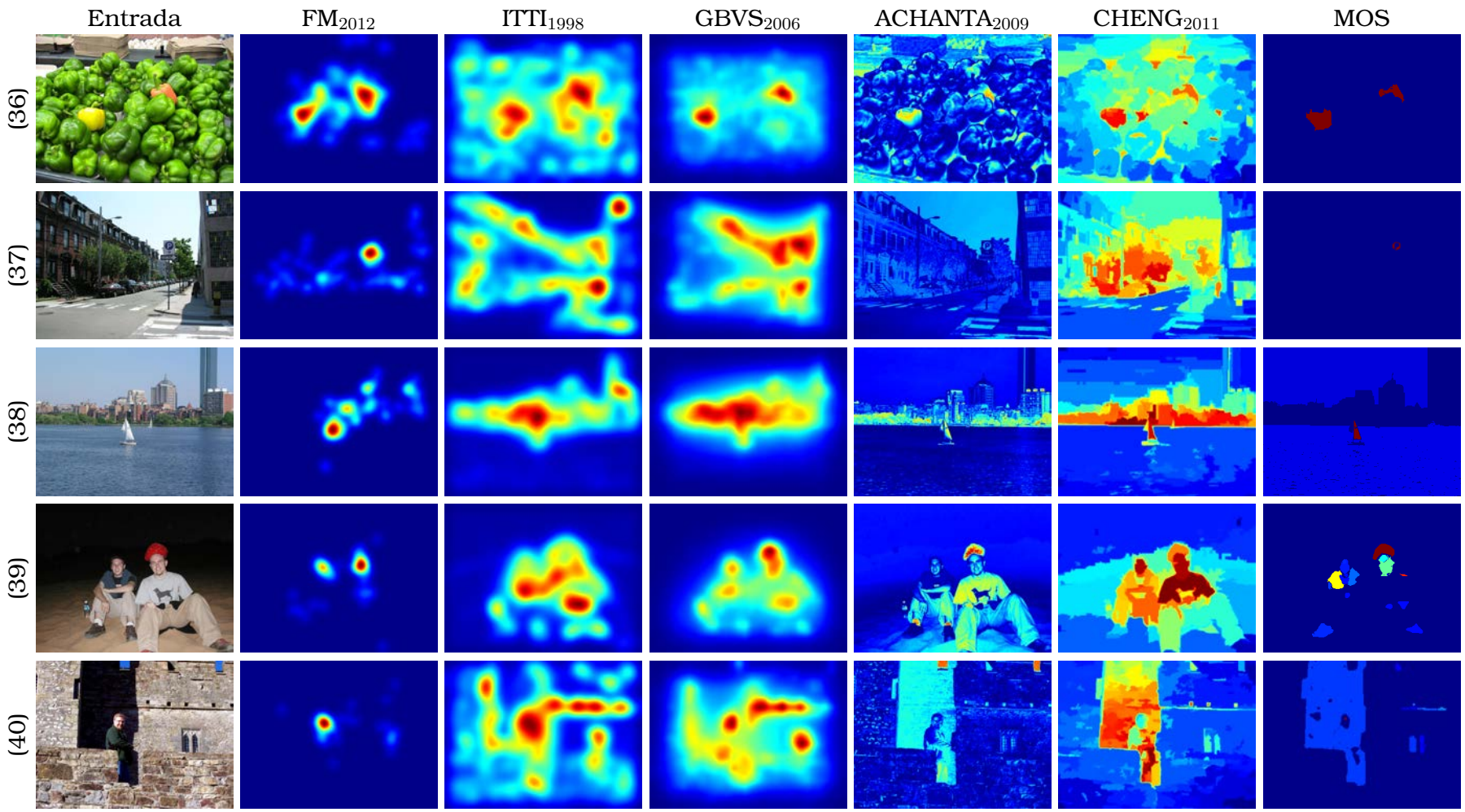


Figura 6.8: Análise qualitativa (8-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

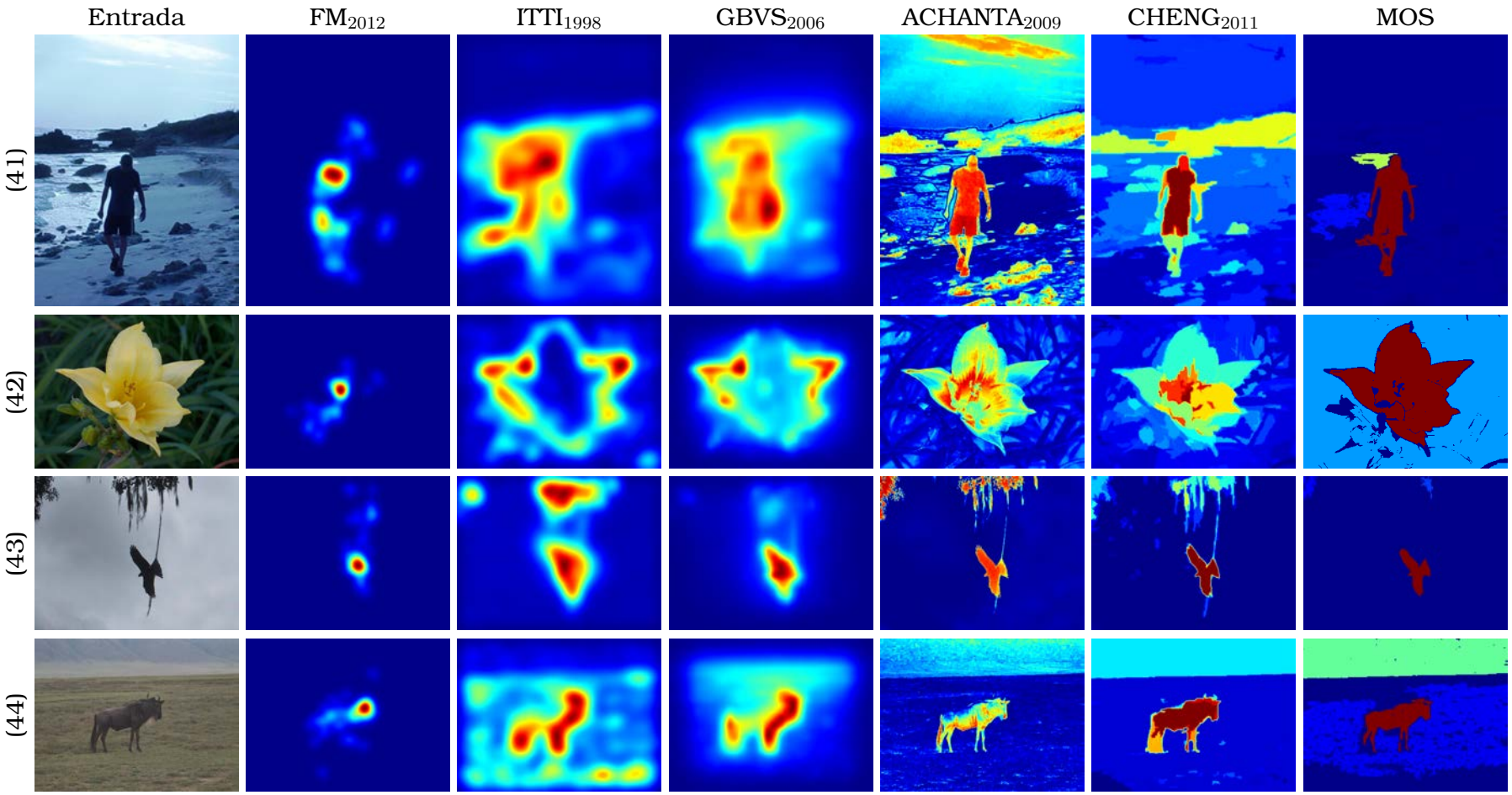


Figura 6.9: Análise qualitativa (9-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

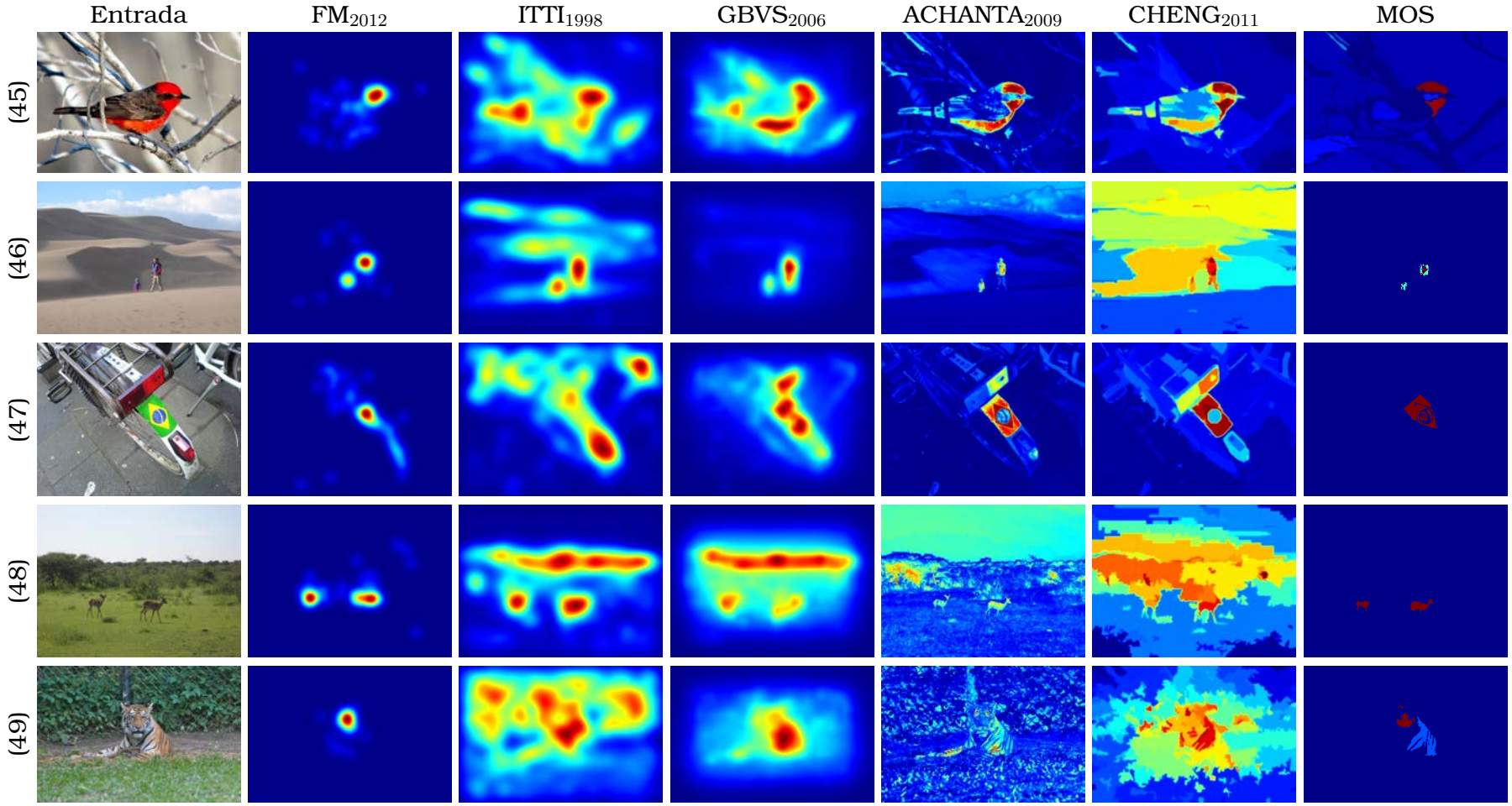


Figura 6.10: Análise qualitativa (10-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

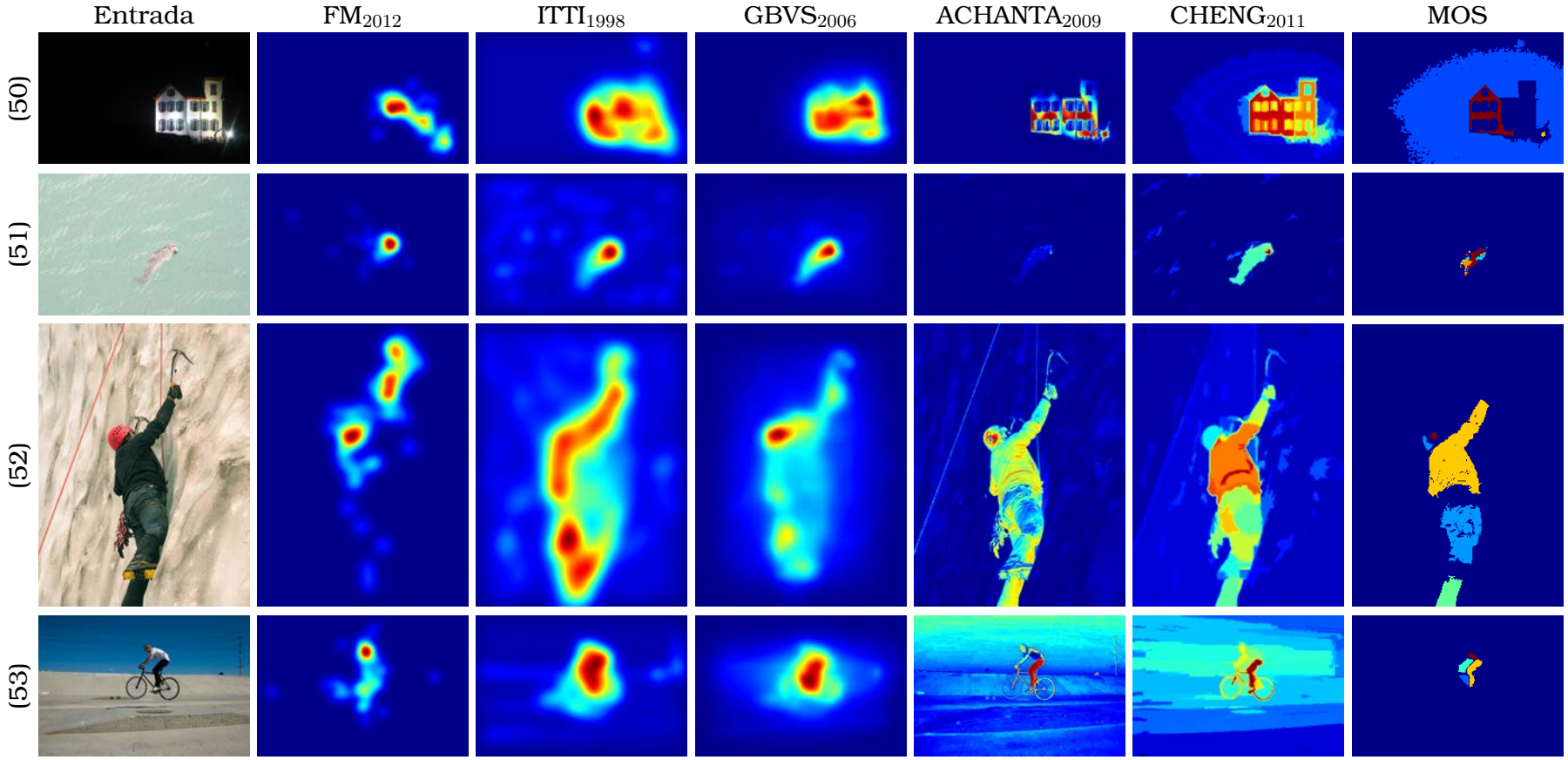


Figura 6.11: Análise qualitativa (11-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

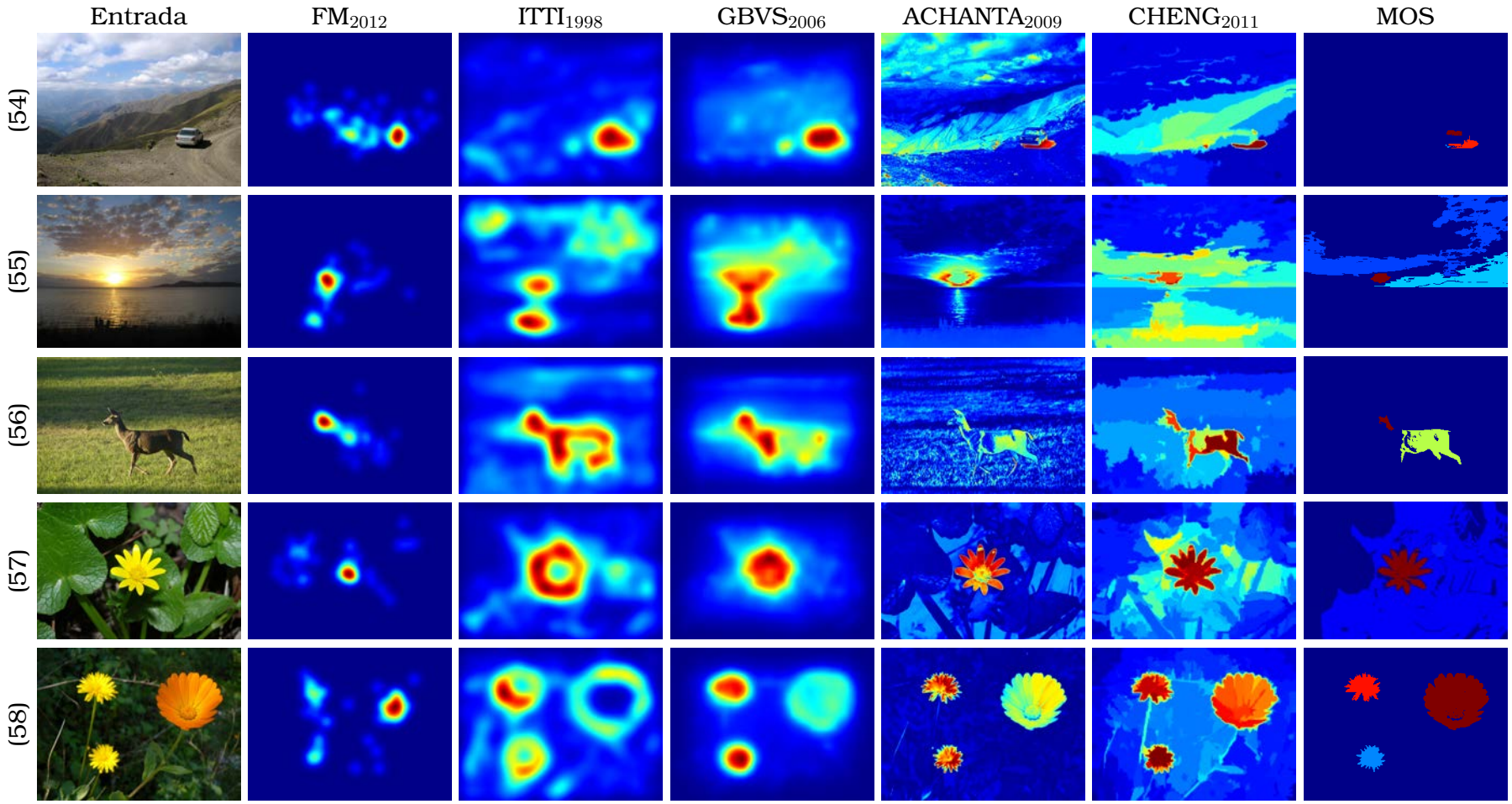


Figura 6.12: Análise qualitativa (12-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

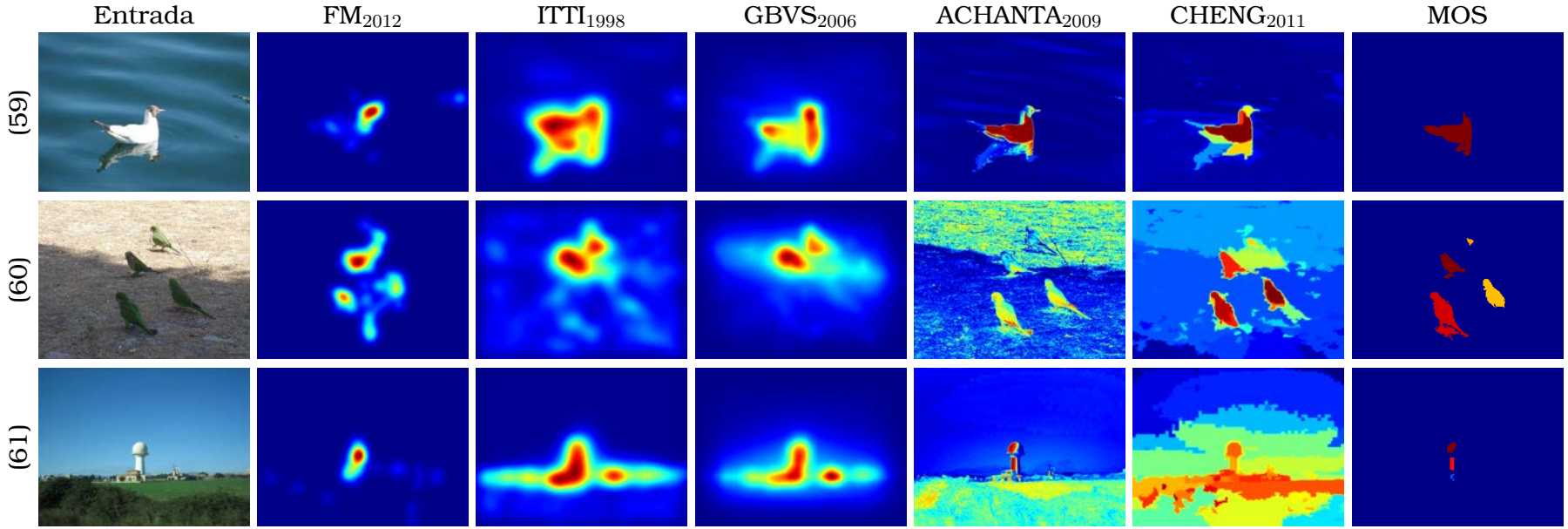


Figura 6.13: Análise qualitativa (13-13) do MOS proposto em cenas reais, comparado com o mapa de fixação humana (FM) de Judd et al. (2012) e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens dos *benchmarks* disponibilizados publicamente por Bruce and Tsotsos (2009) (1-36) e Judd et al. (2012) (37-61).

De acordo com os resultados apresentados no MOS, de uma maneira qualitativa, podemos concluir a eficiência do modelo proposto para a predição de regiões da cena nas quais observadores humanos tendem a fixar. Contudo, considerando a existência dos mapas de fixação, apresentados nas Figuras 6.1 à 6.13, torna-se também possível a análise quantitativa do comportamento de cada modelo, baseada em uma medida de similaridade entre o mapa de saliência ou mapa de objeto-saliente, e o respectivo mapa de fixação. Sendo assim, com o objetivo de realizar uma análise espacial entre os mapas, utilizaremos a medida de similaridade apresentada em (Judd et al., 2012). De acordo com os autores, a medida de similaridade (S) baseia-se inicialmente na redistribuição dos valores dos mapas de saliência a serem comparados, de forma que a soma dos valores das coordenadas i, j seja igual a 1. A similaridade é então calculada através da soma dos valores mínimos dos pontos redistribuídos de cada mapa. A similaridade entre dois mapas P e Q é definida como segue:

$$S = \sum_{i,j} \min(P_{i,j}, Q_{i,j}) \quad onde \quad \sum_{i,j} P_{i,j} = \sum_{i,j} Q_{i,j} = 1, \tag{6.1}$$

de forma que $S = 1$ quando as distribuições forem idênticas e, $S = 0$, caso não exista qualquer coincidência entre os mapas.

A medida de similaridade S utilizada em (Judd et al., 2012) para a comparação qualitativa entre mapas de saliência tem se mostrado bastante eficiente para a classificação de modelos de atenção visual. Entretanto, conforme apresentado nas seções anteriores, e principalmente nas duas últimas seções, propomos mecanismos para a identificação da saliência da cena baseada em objetos. Notamos que, devido ao MOS apresentar a saliência baseada em objetos, a região saliente no mapa limita-se à região ocupada pelo próprio objeto segmentado, o que faz com que a medida de similaridade da vizinhança imediata ao objeto seja nula, quanto comparada a um mapa de fixação, ou a qualquer mapa de saliência, no qual o objeto é representado pela saliência de uma região. Como solução, aplicamos um filtro Gaussiano (7x7 e $\sigma = 3$) ao MOS, previamente ao cálculo de similaridade, de forma a suavizar as bordas dos objetos segmentados. Nas Figuras 6.14, 6.15 e 6.16 são apresentados os valores de similaridades S dos resultados apresentados pelo modelo proposto e demais modelos analisados, de acordo com as cenas apresentadas nas Figuras 6.1 à 6.13.

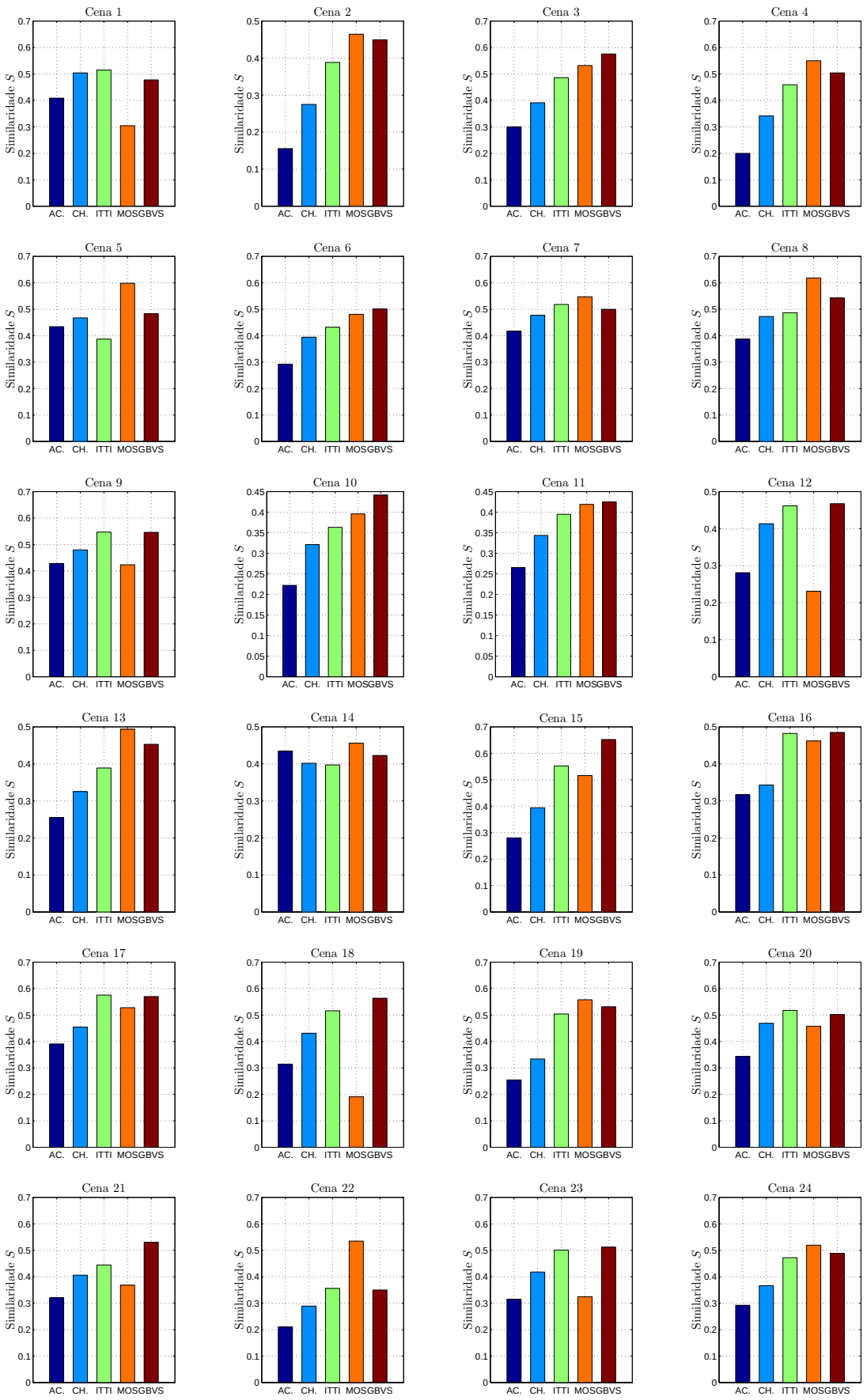


Figura 6.14: Similaridade S (1-3) dos mapas de saliência apresentados nas Figuras 6.1 à 6.13, em relação aos respectivos mapas de fixação (FM) de Judd et al. (2012).

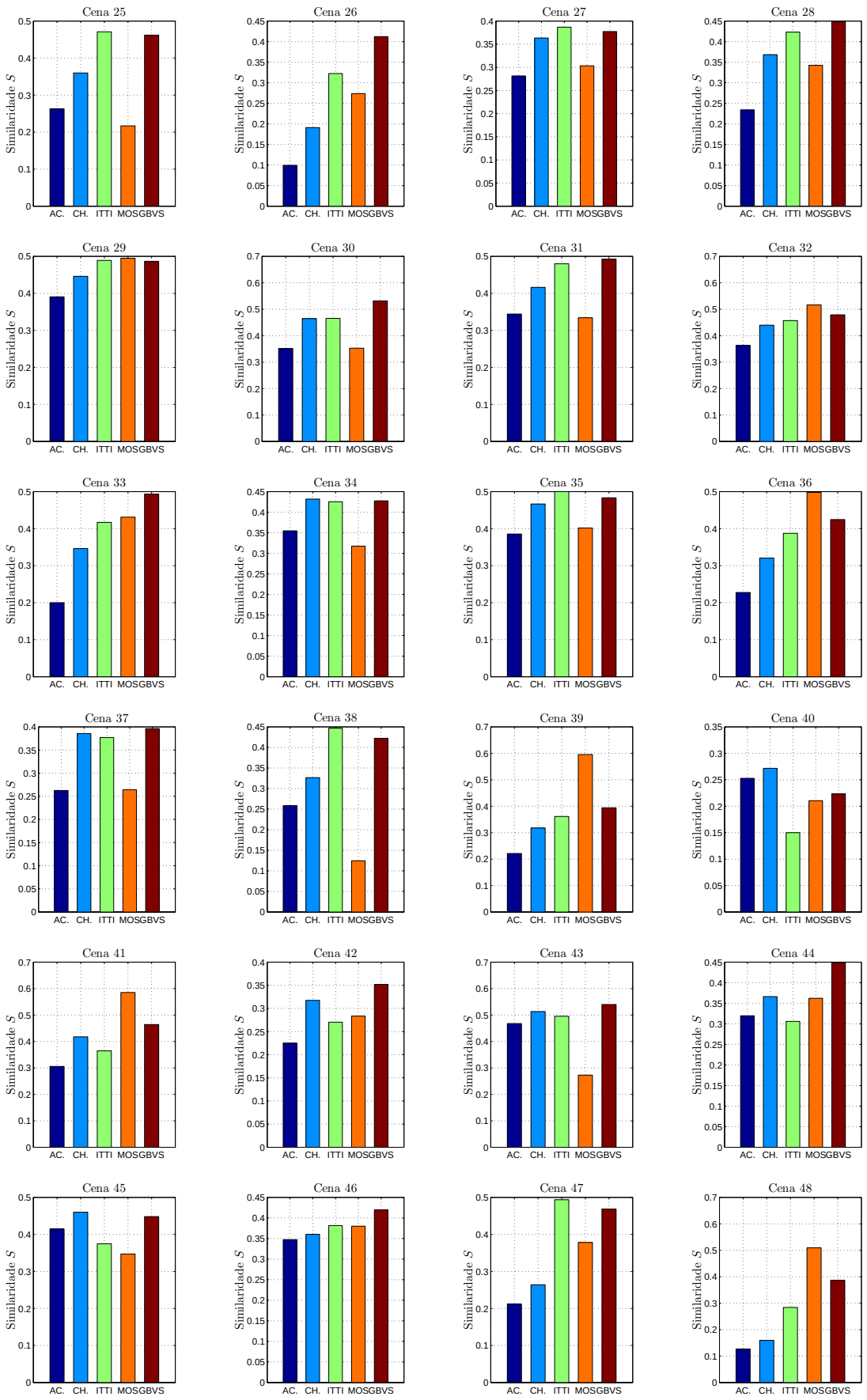


Figura 6.15: Similaridade S (2-3) dos mapas de saliência apresentados nas Figuras 6.1 à 6.13, em relação aos respectivos mapas de fixação (FM) de Judd et al. (2012).

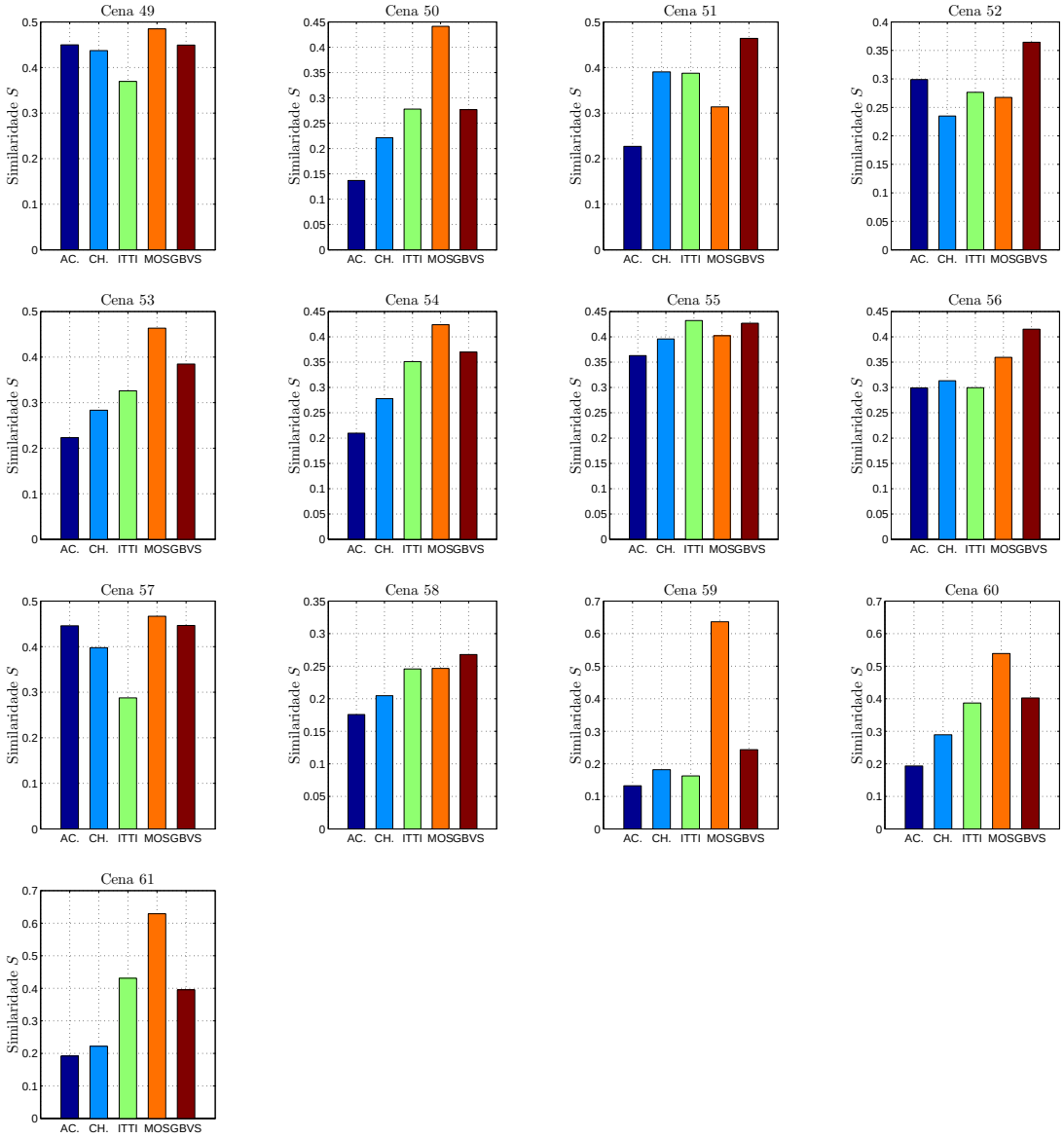


Figura 6.16: Similaridade S (3-3) dos mapas de saliência apresentados nas Figuras 6.1 à 6.13, em relação aos respectivos mapas de fixação (FM) de Judd et al. (2012).

Na Figura 6.17 são apresentadas as médias de similaridades obtidas pelo modelo proposto e demais modelos analisados. De acordo com a análise do comportamento do modelo proposto, apresentado nas Figuras 6.14, 6.15 e 6.16, e na média de similaridade obtida, concluímos também sua eficiência, de forma quantitativa, em relação aos demais modelos analisados, apresentado resultados concludentes para a predição da atenção visual humana. Observamos ainda que, os resultados obtidos por nosso modelo de atenção poderiam ser mais representativos, pois em algumas simulações, como por exemplo, a simulação 38, apresentada na Figura 6.8, embora o local de saliência tenha sido identificado de forma precisa no MOS, o valor de similaridade obtido para esta simulação não foi representativo. Na Figura 6.18 é apresentada uma visão geral do comportamento dos modelos analisados em relação aos valores de similaridades obtidos.

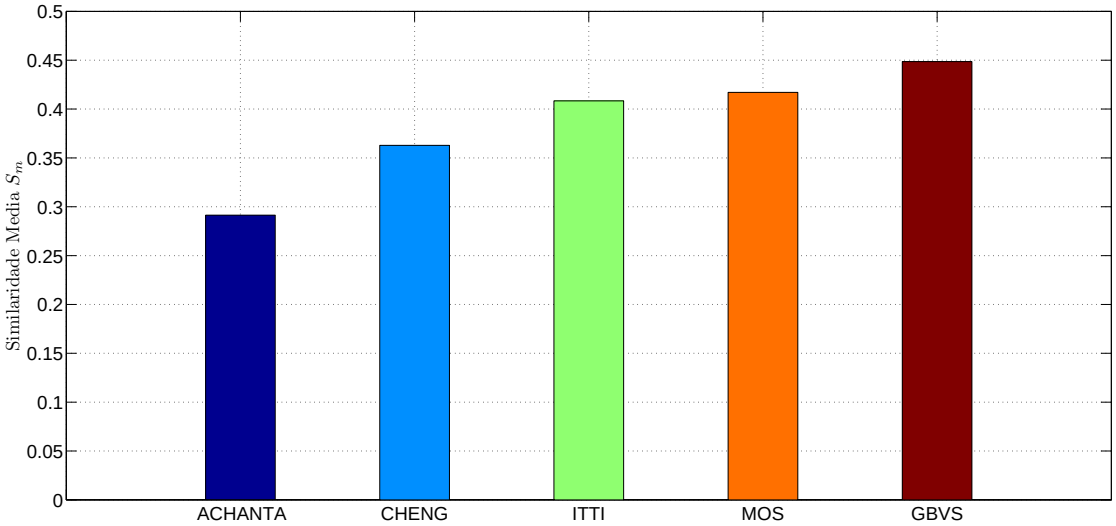


Figura 6.17: Médias de similaridades dos modelos analisados em relação aos mapas de fixações humanas (FM) de Bruce and Tsotsos (2009) e Judd et al. (2012) apresentados nas Figuras 6.1 à 6.13.

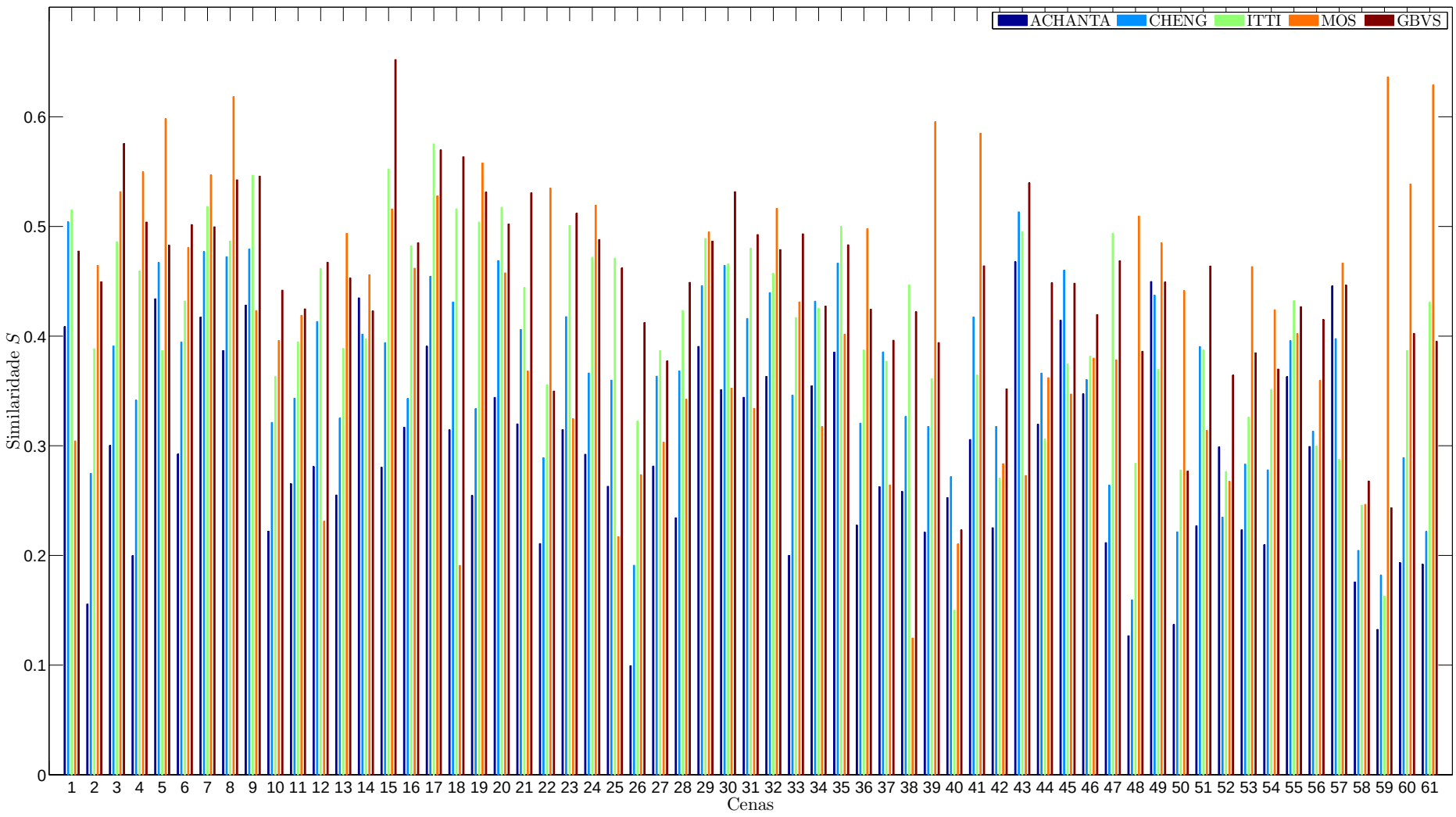


Figura 6.18: Visão geral da similaridade S dos mapas de saliência apresentados nas Figuras 6.1 à 6.13, em relação aos respectivos mapas de fixação (FM) de Judd et al. (2012).

6.2 Domínio Psicofísico

De acordo com Bruce and Tsotsos (2009), modelos computacionais desenvolvidos na área de atenção visual têm sido de grande importância para o entendimento da atenção visual biológica. Conforme apresentado na Seção 2, diversos experimentos foram realizados com o objetivo de demonstrar as reações de observadores humanos e primatas, diante de cenas contendo diferentes valores de saliência, a partir de diferentes estímulos visuais. Sendo assim, no conjunto de simulações apresentados a seguir, consideramos uma variedade de resultados clássicos derivados da literatura psicofísica, disponibilizados publicamente no *benchmark* de Bruce and Tsotsos (2009). O objetivo é demonstrar o comportamento do modelo proposto, e dos demais modelos analisados, diante de condições psicológicas específicas. Para este conjuntos de experimentos não foram disponibilizados os respectivos mapas de fixação. Sendo assim, propomos empiricamente, para fins comparativos, a geração dos mapas de fixação, com base nos resultados experimentais descritos no trabalho de Bruce and Tsotsos (2009).

Inicialmente, as imagens 1 e 2, apresentadas na Figura 6.19, baseiam-se nos experimentos relacionados à teoria da integração de características de Treisman and Gelade (1980), onde, objetos que apresentem características contrastantes com os demais objetos da cena, *pop-out*, recebendo a atenção. Entretanto, na imagem 3, apresentada na mesma figura, devido à presença de distratores heterogêneos, a busca deverá ocorrer de forma conjuntiva, considerando as características de cor e orientação. Na imagem 4, as características primitivas relacionadas à cor, orientação e tamanho guiam, sequencialmente, a atenção visual, porém, entre os objetos "5"*s* existe um "2", que também dependerá de uma busca conjuntiva para tornar-se saliente. Nas imagens 5 à 10 (Figuras 6.20 e 6.21), são apresentados experimentos com variações nos valores de similaridade entre o objeto alvo e os distratores. Neste caso, de acordo com Desimone and Duncan (1995), Wolfe (2004) e Bruce and Tsotsos (2009), quando um objeto apresenta um certo contraste em relação aos demais objetos, este deverá receber a atenção. Em seguida, nas imagens 11 à 15 (Figuras 6.19 e 6.22), consideramos as teorias de Treisman and Gormican (1988) e Rosenholtz et al. (2004), baseadas na assimetria atribuída à presença ou ausência de características saliências para o direcionamento da atenção. Nas imagens 11 e 12, por exemplo, a tarefa de encontrar o objeto "+" entre os "-" é mais fácil do que encontrar o "-" entre os "+", contudo a saliência deverá existir em ambas as buscas. De maneira similar, nas imagens 13 (objeto "vermelho" meio a objetos "rosas"), 14 (objeto "rosa" meio a objetos "vermelhos") e 15 (objeto "rosa" meio a objetos "vermelhos" e fundo com baixo valor de contraste), o objeto contrastante deverá ser salientado, mesmo em ambientes com baixos de valores contraste. O conjunto de imagens de entrada, obtidos do *benchmark* proposto por Bruce and Tsotsos (2009), os respectivos mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al.,

2009) e (Cheng et al., 2011), e o mapa de objeto-saliente proposto neste trabalho, são apresentados nas Figuras 6.19, 6.20, 6.21 e 6.22.

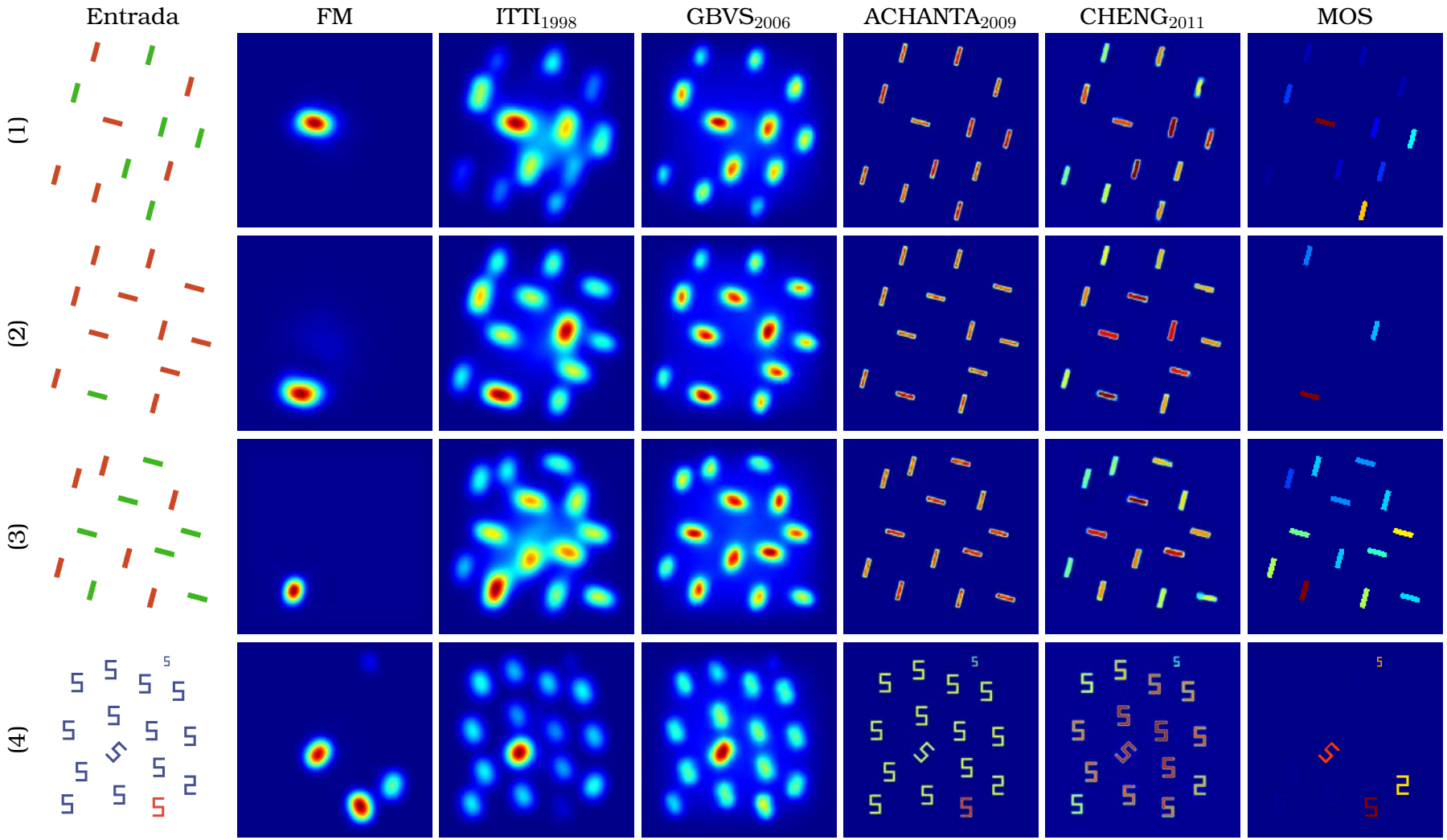


Figura 6.19: Análise qualitativa (1-4) do MOS proposto em cenas sintéticas, comparado com o mapa de fixação gerado empiricamente para fins comparativos e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens do *benchmark* disponibilizado publicamente por Bruce and Tsotsos (2009).

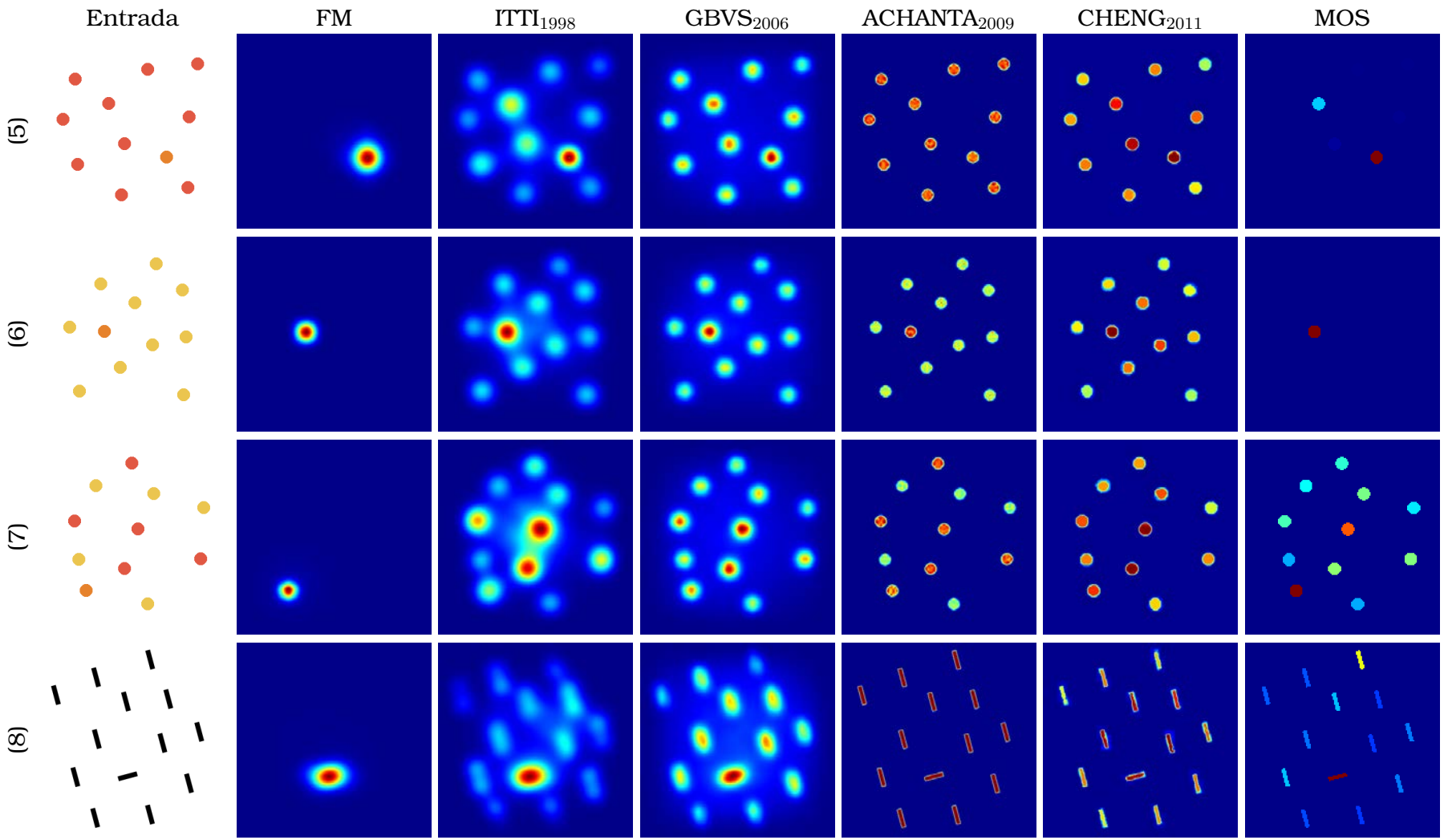


Figura 6.20: Análise qualitativa (2-4) do MOS proposto em cenas sintéticas, comparado com o mapa de fixação gerado empiricamente para fins comparativos e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens do *benchmark* disponibilizado publicamente por Bruce and Tsotsos (2009).

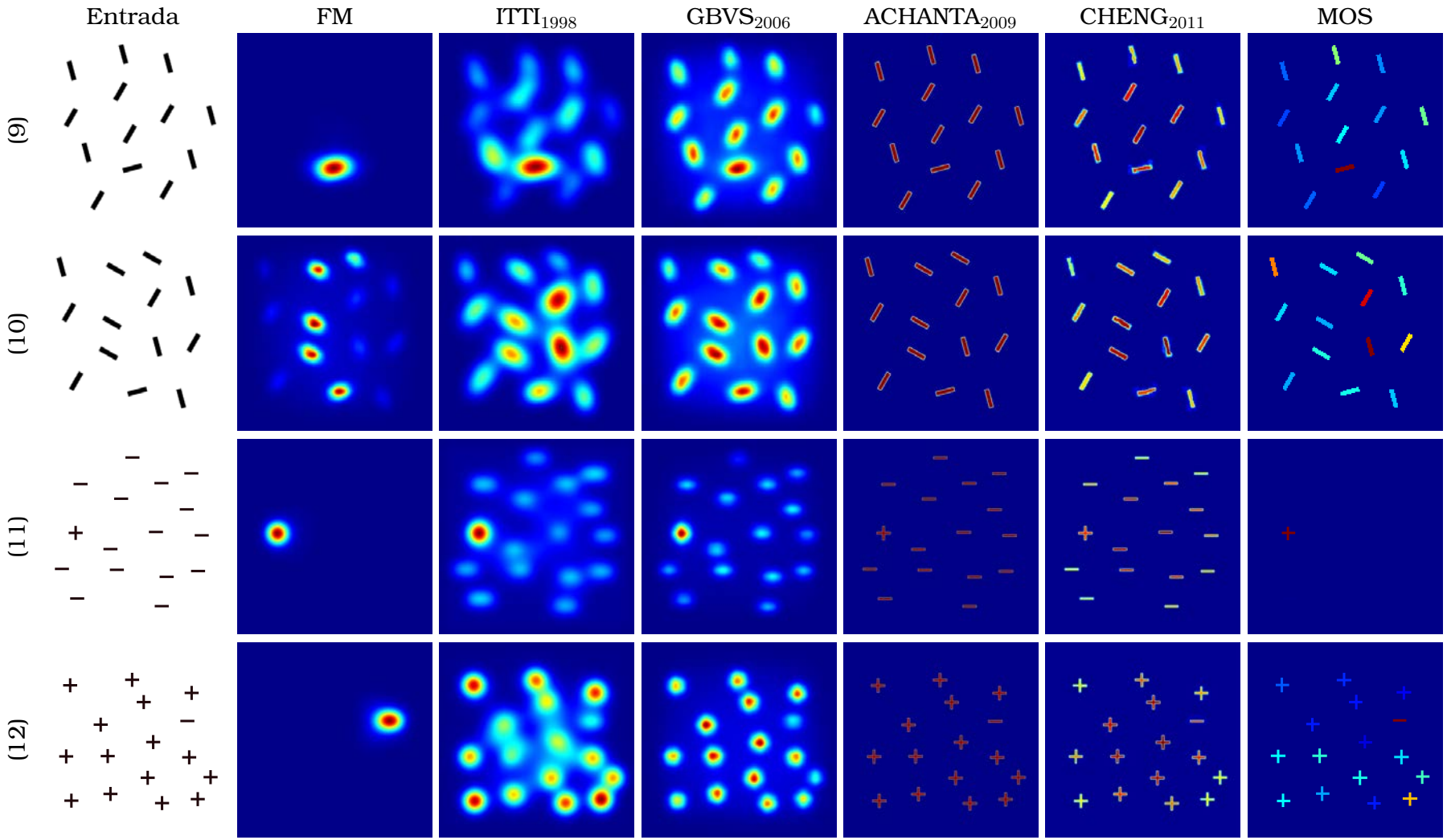


Figura 6.21: Análise qualitativa (3-4) do MOS proposto em cenas sintéticas, comparado com o mapa de fixação gerado empiricamente para fins comparativos e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens do *benchmark* disponibilizado publicamente por Bruce and Tsotsos (2009).

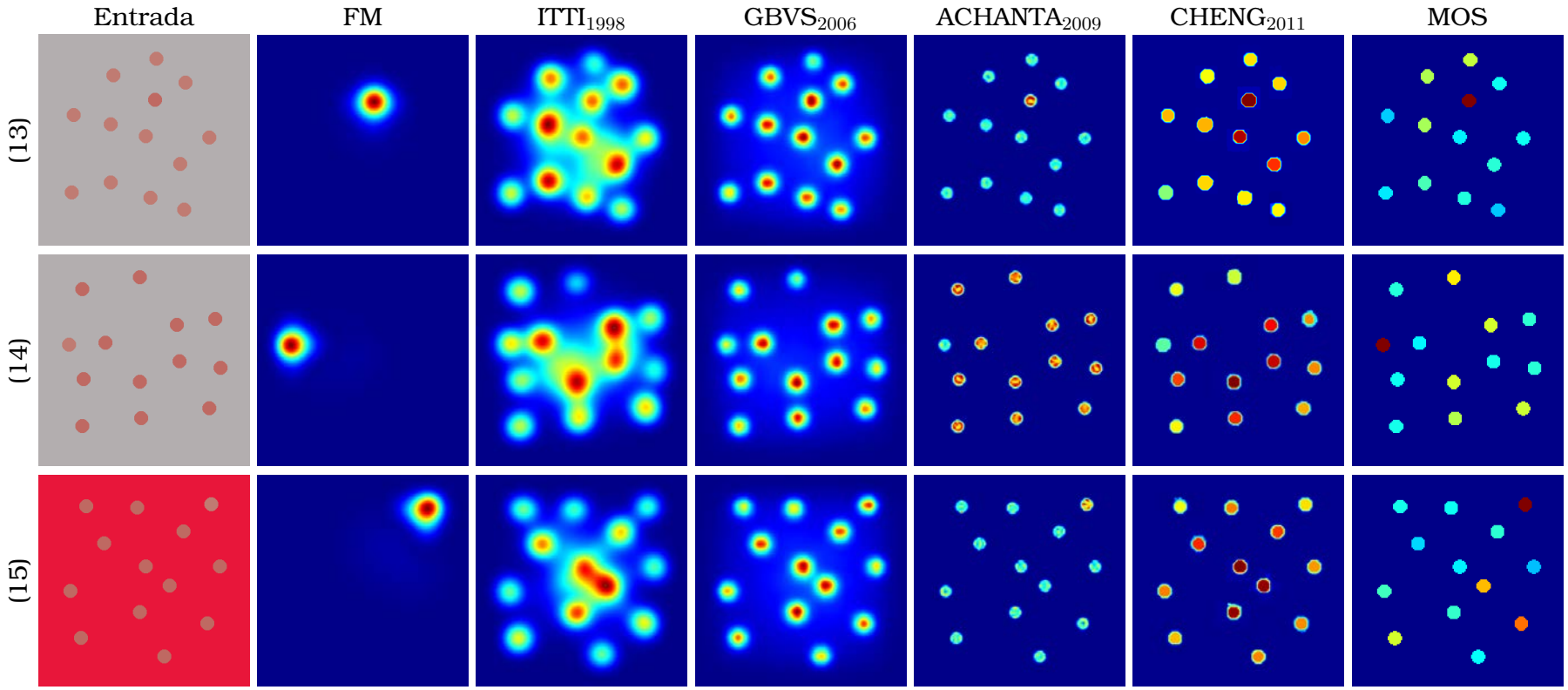


Figura 6.22: Análise qualitativa (4-4) do MOS proposto em cenas sintéticas, comparado com o mapa de fixação gerado empiricamente para fins comparativos e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens do *benchmark* disponibilizado publicamente por Bruce and Tsotsos (2009).

De acordo com a análise qualitativa dos mapas apresentados nas Figuras 6.19 à 6.22, podemos concluir que o comportamento do modelo proposto, para a predição dos locais de maior saliência, foram coerentes com os mapas de fixação apresentados. Os modelos de Achanta et al. (2009) e Cheng et al. (2011), devido à proposta de mecanismos de saliência baseados somente no contraste de cores da imagem, não apresentaram valores significativos de predição. Por outro lado, os modelos de Itti et al. (1998) e Harel et al. (2006), por apresentarem mapas de saliência gerados a partir das características de intensidades, cores e orientações, apresentaram maior precisão na predição dos locais de maior saliência. De maneira quantitativa, o comportamento dos modelos analisados, para este conjunto de simulações, foi bastante uniforme (veja Figura 6.24). Observamos ainda que, devido ao mecanismo de competição por atenção baseado em objetos, proposto em nosso modelo, em imagens contendo distratores homogênos, objetos com características semelhantes foram totalmente inibidos, o que aumentou, significativamente, a taxa de similaridade do modelo proposto. Isto pode ser observado nas imagens 2, 4, 5, 6 e 11 das Figuras 6.19, 6.20 e 6.21, e também nos valores de similaridades apresentados na Figura 6.24. A Figura 6.23 apresenta a média de similaridade obtida por cada modelo.

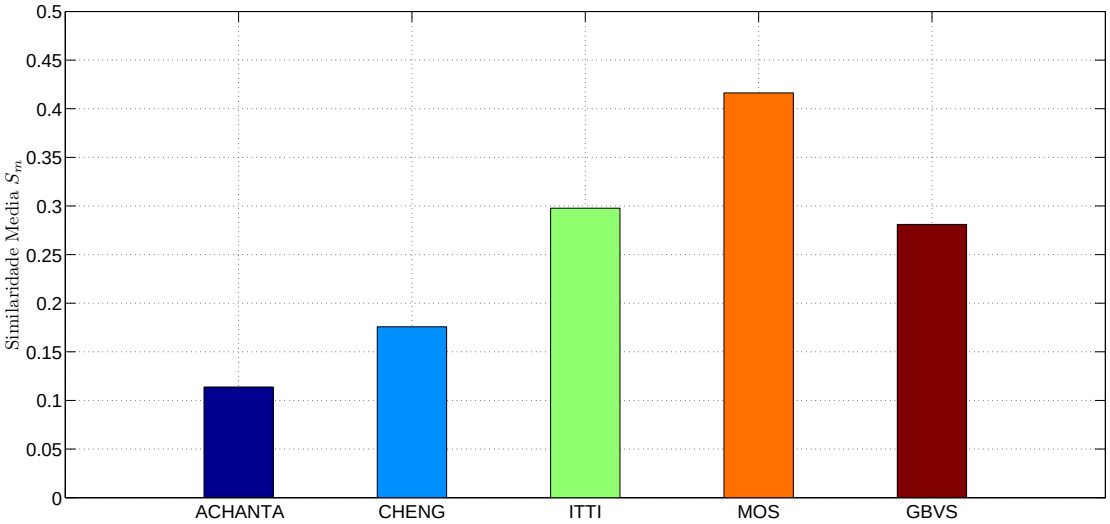


Figura 6.23: Médias de similaridades dos modelos analisados em relação aos mapas de fixações humanas (FM) das Figuras 6.19, 6.20 e 6.21.

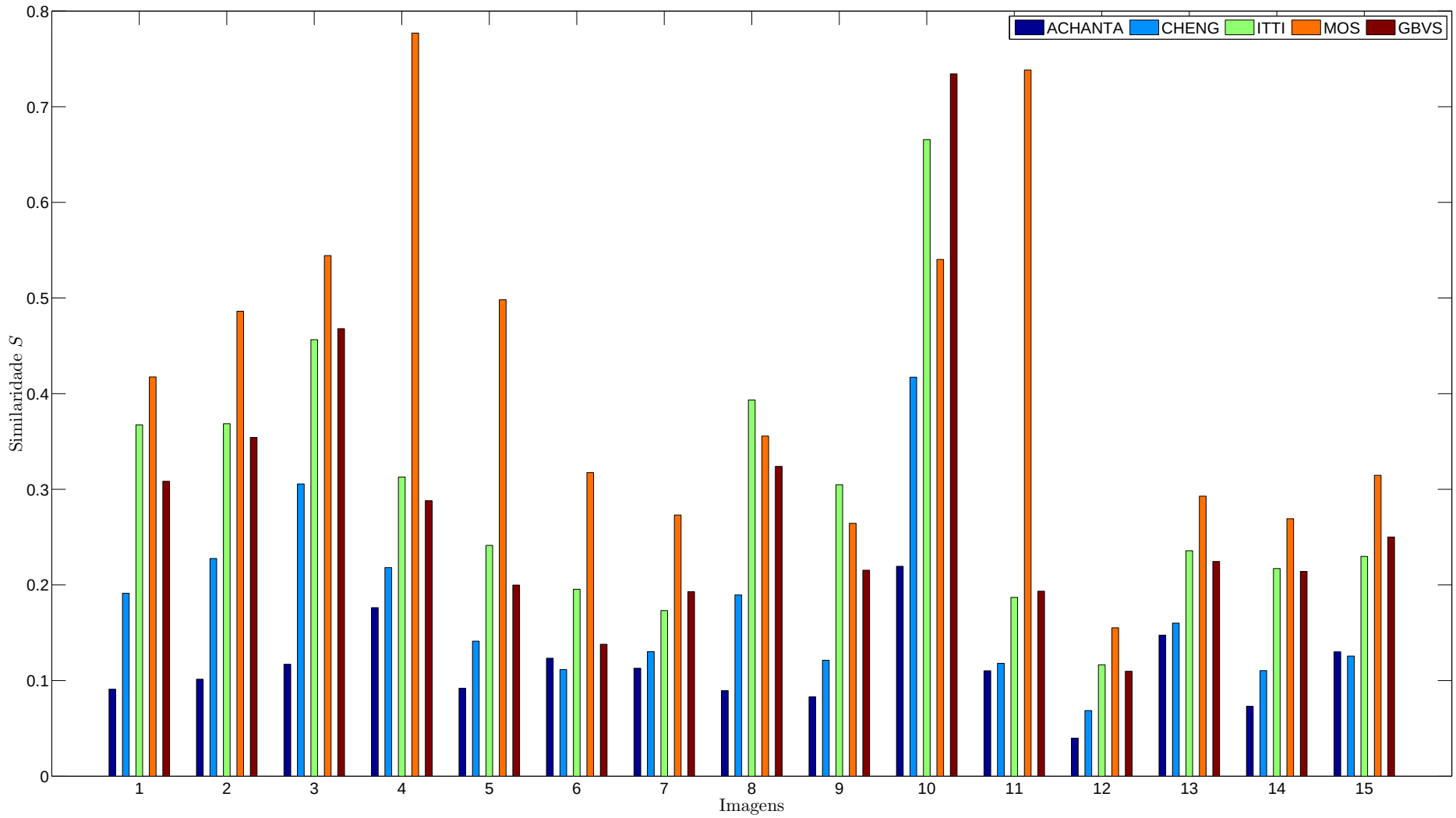


Figura 6.24: Visão geral da similaridade S dos mapas de saliência apresentados nas Figuras 6.19 à 6.22, em relação aos respectivos mapas de fixação (FM).

6.3 Domínio Homogêneo

No conjunto de simulações apresentadas a seguir, considerado como uma extensão das simulações 7 e 8 (Seção 5.3.1), temos como objetivo demonstrar a estabilidade do modelo proposto para um domínio específico de cenas. É importante notar que, devido à similaridade das cenas e do alvo, foi possível definir valores de parâmetros comuns às simulações. Consideramos a configuração inicial dos valores de parâmetros definidos como: enviesamento *top-down* da segmentação: $W_{int} = 0$, $W_{col} = 1$, $W_{ori} = 0$ e $\theta_{bias} = 0.6$; rede LEGION $\theta_p = 600$ e $W_z = 20$; geração do MOS: $W_Y = 1$, $\theta_r = 0.27$, $\theta_s = 0.01$; e enviesamento *top-down* da atenção visual: $W_1 = 0$, $W_2 = 0$, $W_3 = 0$, $W_4 = 0$, $W_5 = 1$, $W_6 = 0$. Outra característica importante apresentada, é a efetividade do enviesamento *top-down* do processo de segmentação, proposto na Seção 5.3.2, utilizado para enviesar o processo de seleção visual de placas de sinalização. Para a análise qualitativa e quantitativa, os mapas de fixação foram gerados empiricamente, baseados na busca específica por placas de sinalização. O conjunto de 28 cenas, mapas de fixação, saliência, e objeto-saliente são apresentados nas Figuras 6.25 à 6.30.

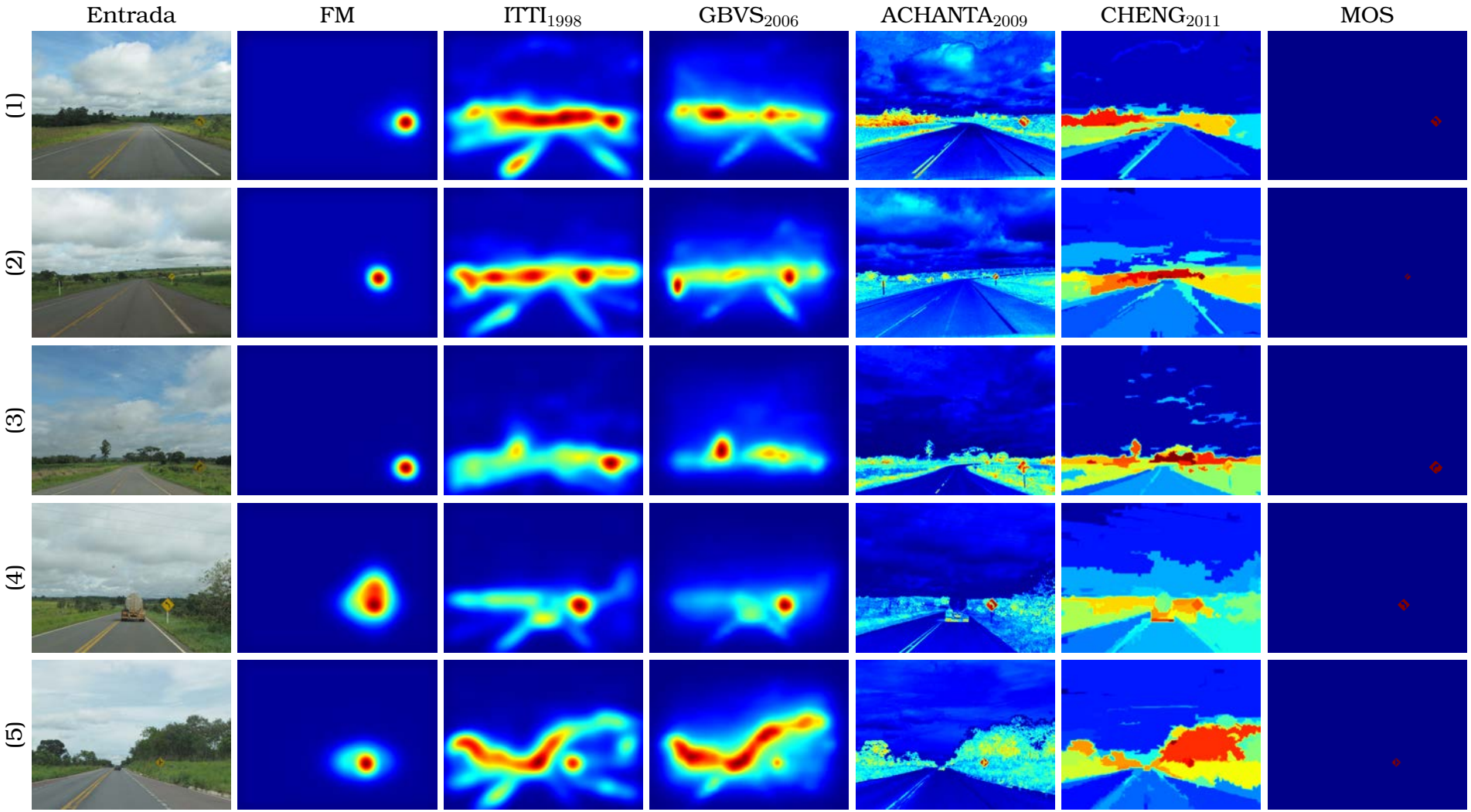


Figura 6.25: Análise qualitativa (1-6) do MOS proposto em cenas contendo placas de sinalização, comparado com o mapa de fixação gerado empiricamente para fins comparativos e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens propostas inicialmente nesta tese.

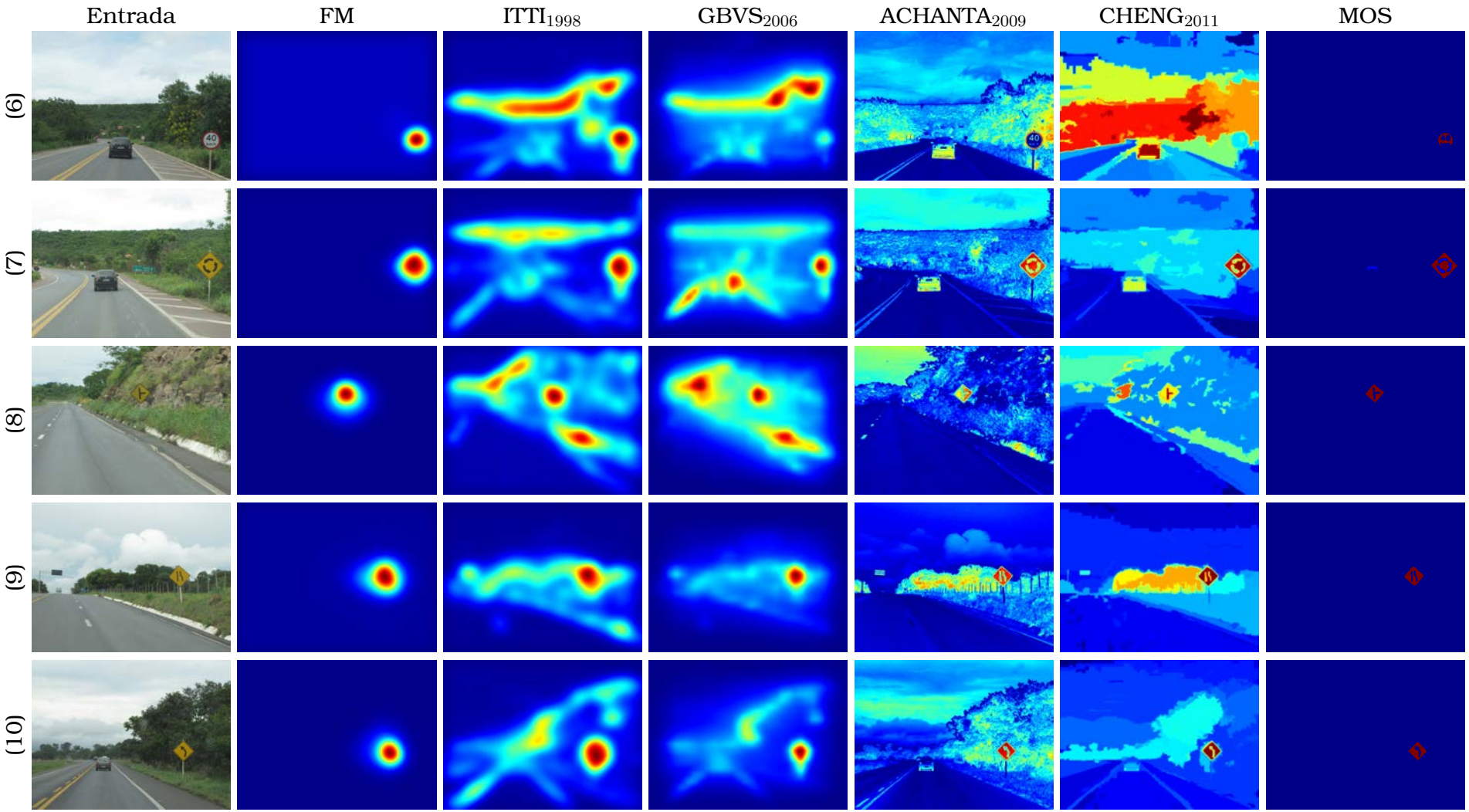


Figura 6.26: Análise qualitativa (2-6) do MOS proposto em cenas contendo placas de sinalização, comparado com o mapa de fixação gerado empiricamente para fins comparativos e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens propostas inicialmente nesta tese.

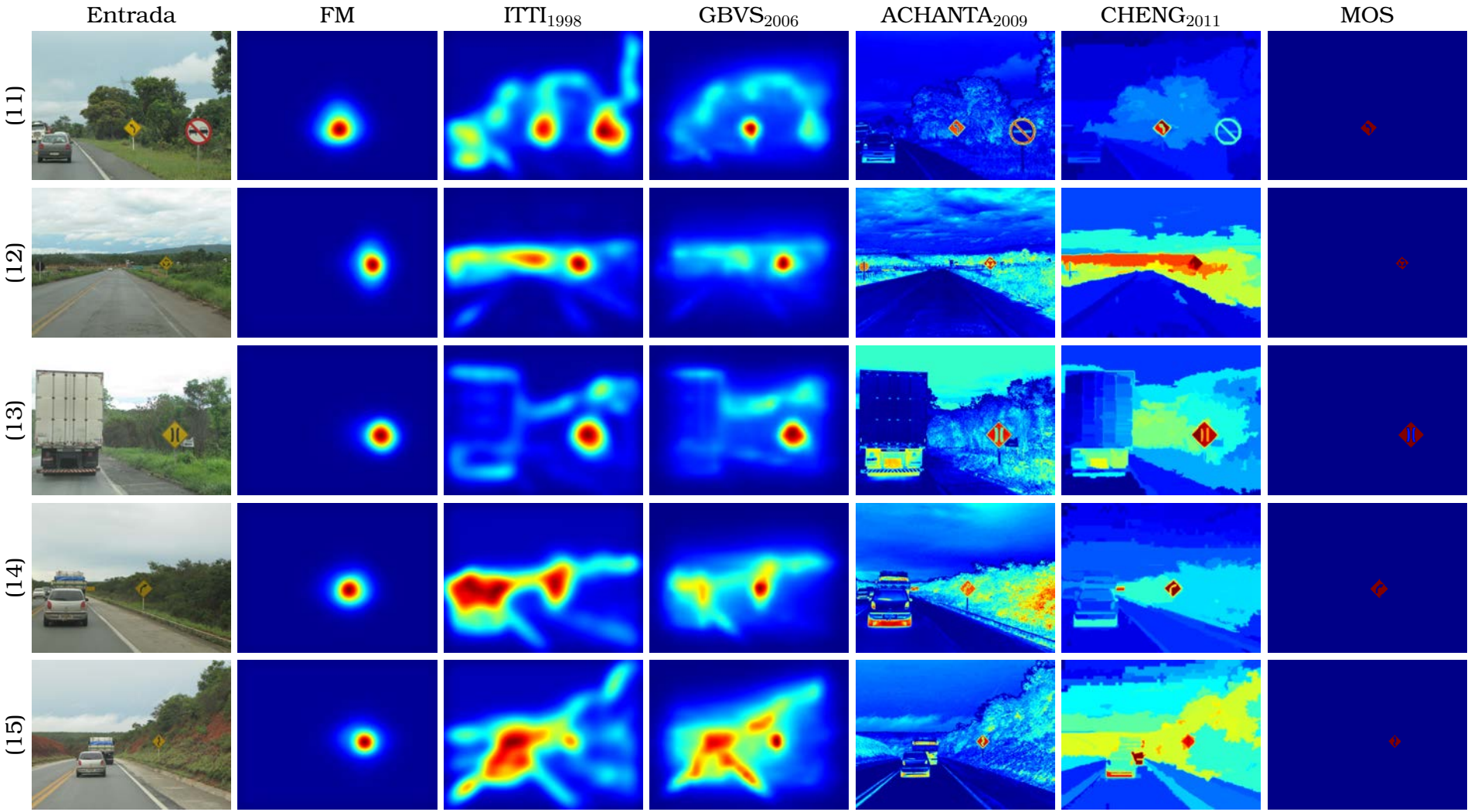


Figura 6.27: Análise qualitativa (3-6) do MOS proposto em cenas contendo placas de sinalização, comparado com o mapa de fixação gerado empiricamente para fins comparativos e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens propostas inicialmente nesta tese.

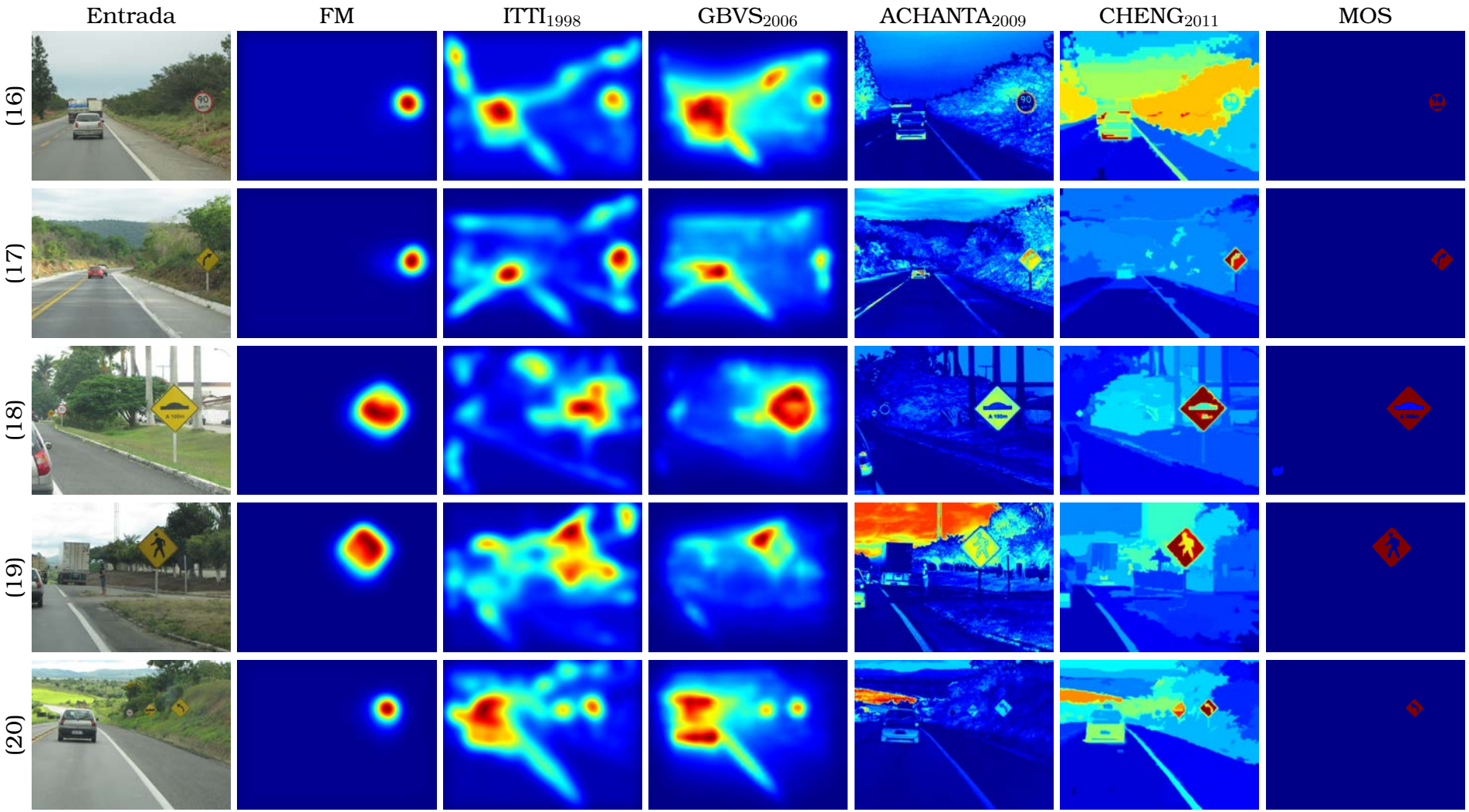


Figura 6.28: Análise qualitativa (4-6) do MOS proposto em cenas contendo placas de sinalização, comparado com o mapa de fixação gerado empiricamente para fins comparativos e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens propostas inicialmente nesta tese.

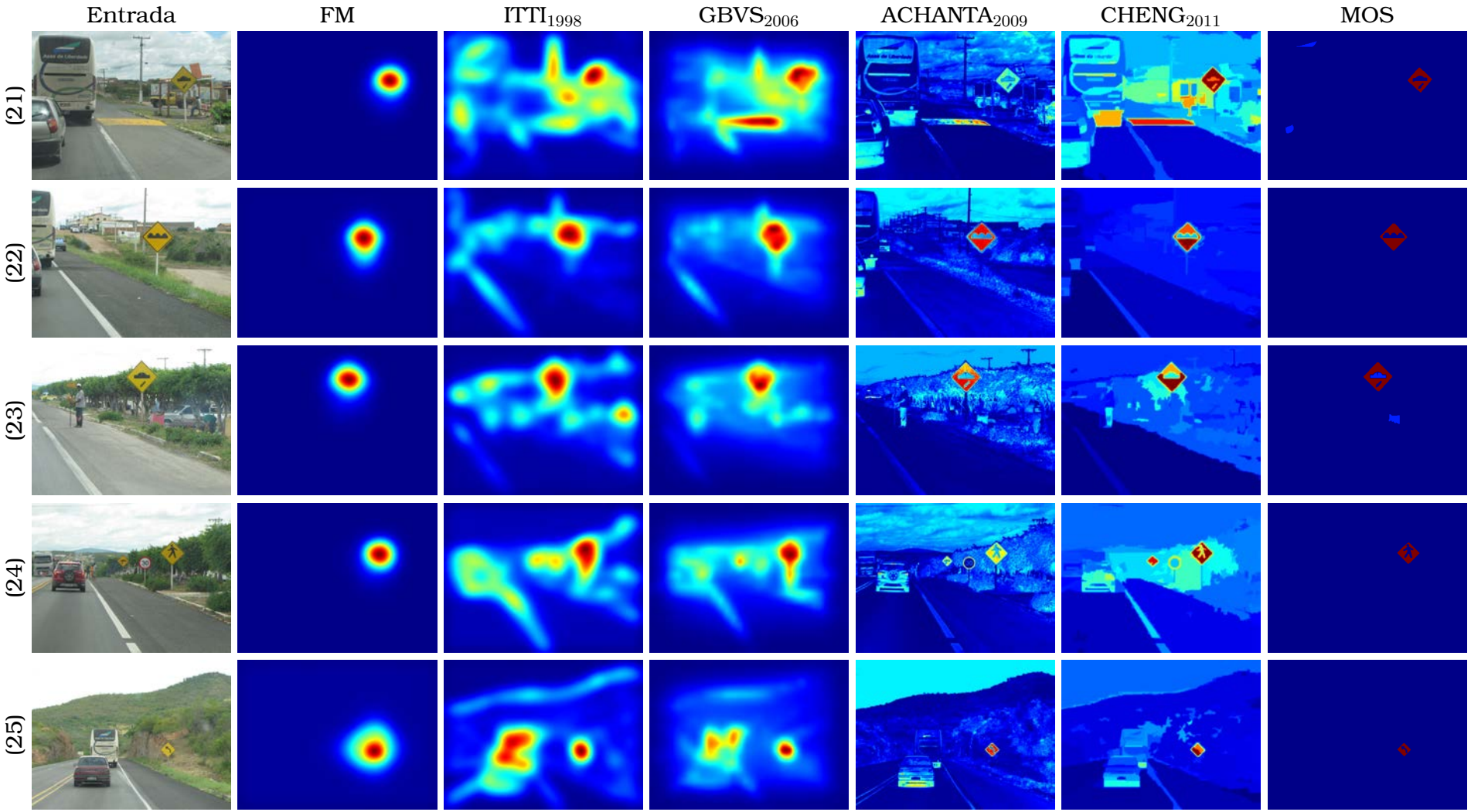


Figura 6.29: Análise qualitativa (5-6) do MOS proposto em cenas contendo placas de sinalização, comparado com o mapa de fixação gerado empiricamente para fins comparativos e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens propostas inicialmente nesta tese.

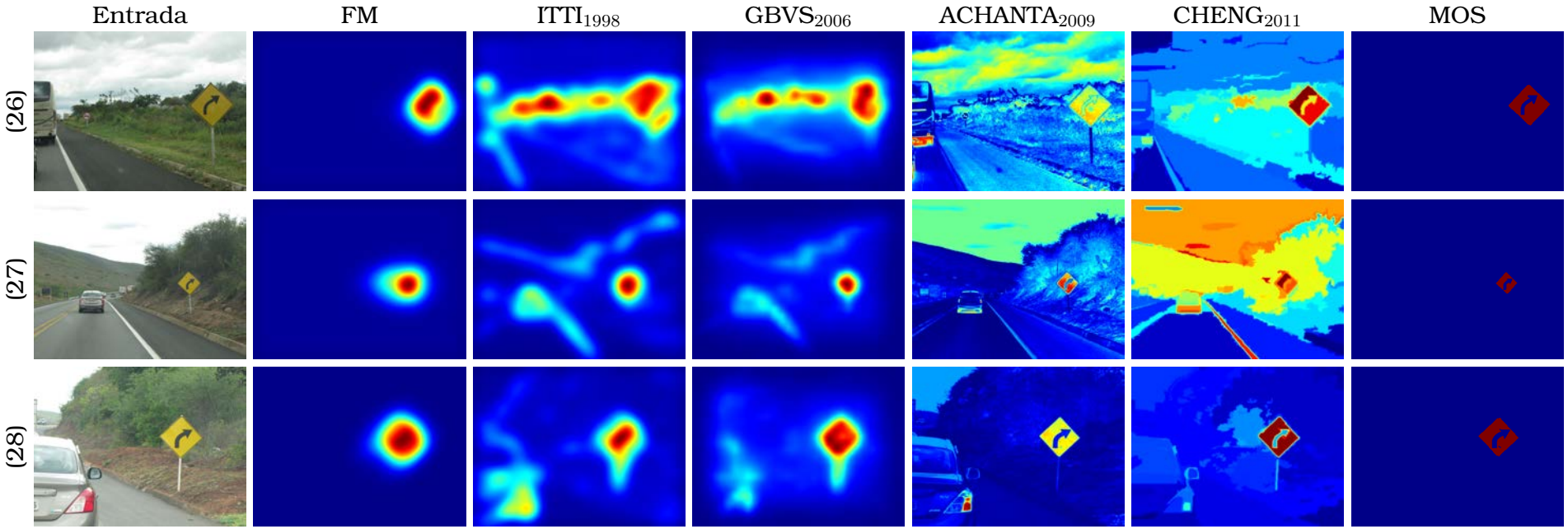


Figura 6.30: Análise qualitativa (6-6) do MOS proposto em cenas contendo placas de sinalização, comparado com o mapa de fixação gerado empiricamente para fins comparativos e também com os mapas de saliência propostos em (Itti et al., 1998), (Harel et al., 2006), (Achanta et al., 2009) e (Cheng et al., 2011), respectivamente apresentados da esquerda para a direita. Imagens propostas inicialmente nesta tese.

Nas simulações apresentadas nas Figuras 6.25 à 6.30, como pode ser observado, os modelos de Achanta et al. (2009) e Cheng et al. (2011) apresentaram melhor predição dos locais de atenção, em relação ao comportamento apresentado nos conjuntos de simulações anteriores, de modo que podemos concluir sua aplicabilidade para domínios nos quais o alvo apresente contraste de cor suficiente para guiar a atenção. Os modelos de Itti et al. (1998) e Harel et al. (2006) também apresentaram bons resultados para este conjunto de simulações. Entretanto, devido a não influência de informações *top-down* para o direcionamento da atenção, diversas outras regiões da imagem são também salientadas, como por exemplo, nas imagem 16, 17 e 20 (Figura 6.28), onde a região contendo o carro foi considerada o local de maior saliência da cena. De acordo com os valores de parâmetros utilizados, o MOS apresentou a saliência somente do objeto de interesse, neste caso, a placa de sinalização. Observamos também a importância do mecanismo de enviesamento *top-down* para o processo de segmentação, onde a definição do parâmetro $W_{cor} = 1$ possibilitou a segmentação e análise somente de regiões com contraste de cores, de forma que demais regiões da imagem tenham sido desconsideradas, otimizando o tempo de identificação e seleção do objeto saliente e, consequentemente, proporcionando a redução do custo computacional. Na Figura 6.31 são apresentadas as médias similaridades, obtidas a partir das simulações apresentadas nas Figuras 6.25 à 6.30. O comportamento individual dos modelos, para cada simulação, pode ser visto na Figura 6.32.

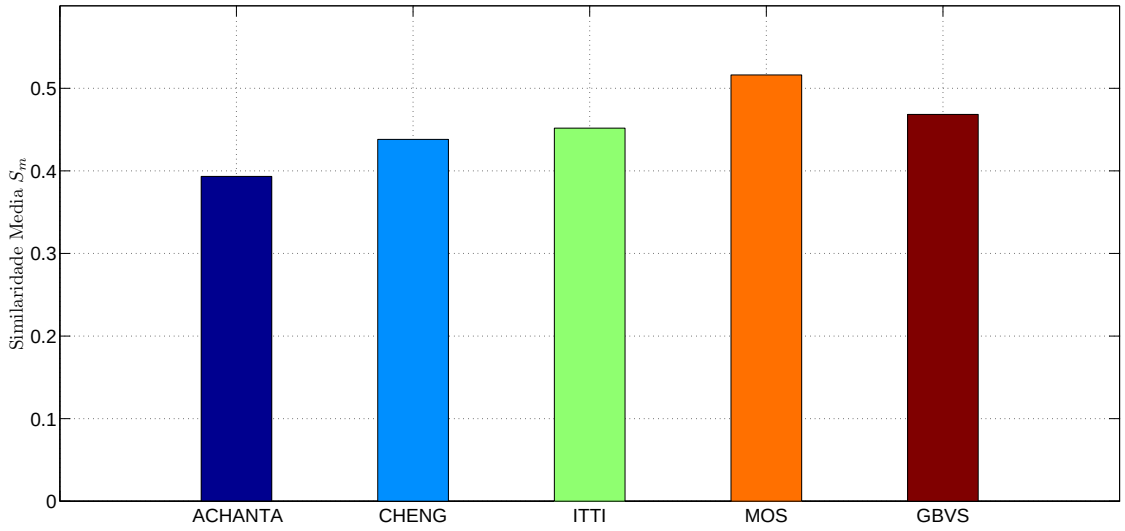


Figura 6.31: Médias de similaridades dos modelos analisados em relação aos mapas de fixações humanas (FM) das Figuras 6.25 à 6.30.

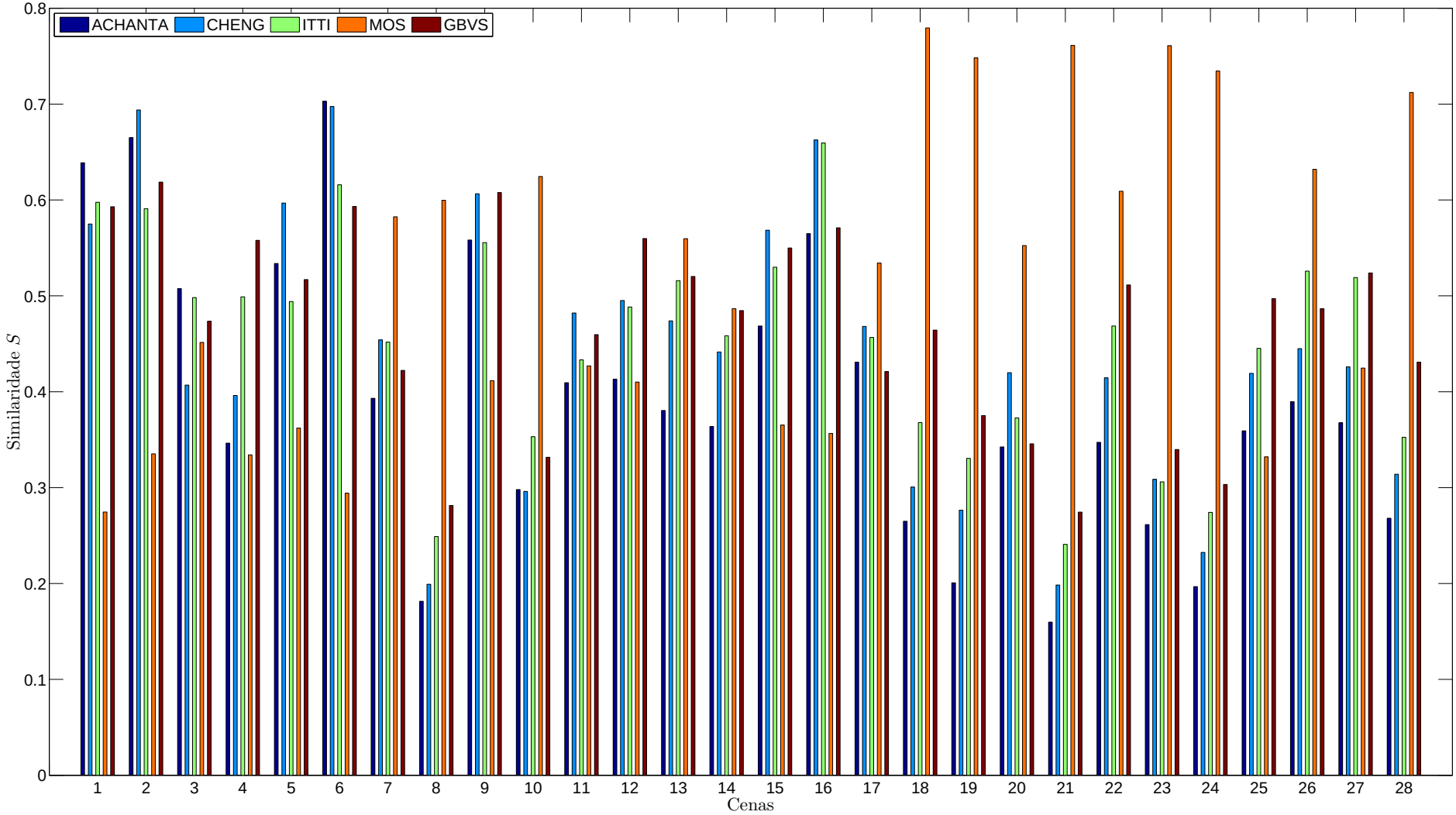


Figura 6.32: Visão geral da similaridade S dos mapas de saliência apresentados nas Figuras 6.25 à 6.30, em relação aos respectivos mapas de fixação (FM).

6.4 Considerações Finais

Neste capítulo, o mapa de objeto-saliente proposto foi comparado com mapas de fixação humana e também com mapas de saliência, gerados a partir de quatro modelos de atenção visual. A medida de similaridade entre os mapas foi utilizada para as análises qualitativas e quantitativas dos resultados obtidos. Conforme apresentado, a capacidade preditiva do modelo proposto pôde ser comparado à importantes modelos de atenção visual propostos recentemente.

No próximo capítulo serão apresentadas as conclusões e principais contribuições desta tese, assim como as perspectivas de trabalhos futuros.

CAPÍTULO

7

Conclusões e Trabalhos Futuros

Nesta tese de doutorado foram introduzidos novos modelos computacionais de atenção visual para a detecção e identificação de objetos salientes em cenas reais e sintéticas. A seleção do objeto saliente foi determinada por dois tipos de influências: *bottom-up* e *top-down*.

De acordo com a evolução dos modelos propostos, destacamos, cronologicamente, as respectivas conclusões, contribuições e desafios encontrados em cada momento do desenvolvimento deste trabalho, conforme descrito a seguir.

Na Seção 5.1.1 foi apresentado um mecanismo de saliência de atributos pulsantes através de um mapa de atributo-saliência gerado a partir de modelos de redes neurais artificiais. Os resultados mostraram que o modelo como um promissor mecanismo para a composição de um sistema de atenção visual. Embora o modelo proposto não tenha sido aplicado à seleção visual de objetos, nas propostas seguintes o potencial de saliência dos neurônios do MAS foi utilizado como principal mecanismo para a identificação da saliência.

A principal contribuição da pesquisa apresentada na Seção 5.1.2 foi a possibilidade da identificar o posicionamento, em um plano 2D, do objeto representado em regiões de maior saliência no mapa de atributo-saliência, ou seja, baseado no potencial de saliência dos neurônios do MAS, foi possível identificar o grupo de osciladores da rede I&F e, consequentemente, a seleção do objeto saliente. A partir das simulações realizadas, foi possível também identificar a limitação no modelo proposto em relação ao conjunto de padrões utilizados durante a fase de treinamento da rede SOM. Concluímos que, com o aumento do número de atributos a ser mapeado bidimensionalmente no SOM, neste caso somente os atributos cor e posição, a utilização de padrões aleatórios durante a fase de treinamento ocasionou alterações significativas sobre os valores de convergência do mapa, vindo a inferir

diretamente na precisão do modelo. Entretanto, este experimento foi elucidativo quanto às propostas apresentadas a seguir.

Sendo o primeiro trabalho publicado decorrente desta tese (Benicasa and Romero, 2010), a pesquisa apresentada na Seção 5.1.3 pode ser considerada como uma síntese das propostas anteriores. Assim, todas as contribuições realizadas até o momento são também atribuídas a esta proposta. Como contribuição particular desta proposta, destacamos o treinamento da rede SOM a partir de padrões baseados na própria cena. Embora este processo deva ocorrer a cada simulação, podemos concluir que o modelo apresentou maior precisão na localização dos objetos salientes.

Na Seção 5.2.1 foram considerados para o direcionamento da atenção visual os seguintes atributos: canais de cores, contraste de intensidades, diferença espacial em cores, orientações, localizações em um plano bidimensional e o fator de reconhecimento do objeto. Como uma das principais contribuições, destacamos inicialmente sua característica *bottom-up* e *top-down*, obtida através da proposta de um mecanismos que tornou possível realizar modulações *top-down* baseadas no espaço e também baseada no objeto, permitindo o direcionamento da atenção tanto para alvos desejados, quanto para demais regiões salientes. Outra contribuição proposta foi a possibilidade do direcionamento da atenção visual para objetos que apresentem padrões cognitivos diferentes dos distratores, mesmo em situações onde a saliência *bottom-up* seja considerada nula. Conforme mencionado inicialmente na Seção 5.2.1, os resultados obtidos encontram-se publicados em Benicasa et al. (2012).

Considerando os modelos propostos descritos, destacamos suas limitações relacionadas às simulações envolvendo imagens reais. Esta limitação está diretamente ligada ao processo de segmentação da cena. Consideramos que, apesar da rede de neurônios I&F apresentar eficiência e simplicidade no ajuste de parâmetros para a segmentação de imagens sintéticas, não obtivemos sucesso na segmentação de imagens reais.

Na Seção 5.2.2 foi introduzido um modelo de atenção visual baseada em objetos, com modulações *bottom-up* e *top-down* aplicadas à imagens reais. Baseado na proposta apresentada na Seção 5.2.1, este modelo mostrou-se apto a selecionar objetos de acordo com suas características visuais primitivas e também a partir de conhecimentos prévios sobre determinados alvos. Além disso, as simulações apresentadas demonstraram a capacidade do modelo LEGION para a segmentação de imagens reais, permitindo a este modelo trabalhar com imagens reais, fator considerado como limitação nos modelos apresentados nas seções anteriores.

Considerando os modelos propostos, o mapa SOM apresentou uma importante função para o desenvolvimento do mecanismo de atenção visual desta tese. Nas Seções 5.1.1 e 5.1.2, o mapa SOM foi gerado baseado em padrões aleatório e, a partir da Seção 5.1.3, foi gerado por padrões extraídos da própria cena,

proporcionando maior precisão na localização dos objetos salientes. Entretanto, sua geração, considerando as fase de treinamento e classificação em tempo real, paralelamente ao processo de segmentação, torna o custo computacional para o processo de atenção bastante alto. Além de inviabilizar a aplicabilidade dos modelos propostos em sistemas de tempo real como, por exemplo, a saliência baseada em objetos a partir de imagens em vídeo. Com isto, tivemos como objetivo nas propostas seguintes, apresentar um mecanismo alternativo ao SOM para o cálculo da saliência de objetos. Outro fator considerado foi relacionado ao módulo de classificação dos segmentos gerados pela rede LEGION. Embora a MLP tenha sido satisfatória para demonstrar o comportamento *top-down* dos modelos apresentados, nas propostas futuras outros modelos foram considerados para a classificação de objetos.

Na Seção 5.3.1, foi apresentado um modelo de atenção visual baseado em objetos, gerado a partir da integração dos mecanismos *top-down* e *bottom-up*. A principal contribuição realizada, em relação aos modelos propostos anteriormente, foi o desenvolvimento do mecanismo de competição por atenção baseado em objetos. Desta forma, após o processo de segmentação, cada segmento foi representado por um único neurônio, responsável por participar da competição pela atenção visual e geração do mapa de objeto-saliente. Assim, com a redução do número de neurônios competidores, tornou-se possível reduzir o custo computacional para o desenvolvimento da seleção visual. Outra contribuição considerada importante foi em relação à normalização dos valores de saliência dos objetos. Neste caso, a presença ou ausência de uma característica, foi utilizada para definir o valor de contraste real do objeto. O modelo proposto também permitiu a modulação da atenção para informações específicas da cena, tornando-o apto a ser utilizado em diferentes domínios. Simulações foram realizadas considerando imagens reais e sintéticas, demonstrando o comportamento do modelo para diferentes condições. Não menos importante, o uso da rede HLC para a identificação de objetos considerados difíceis de classificar, permitiu a geração de valores de saliência *top-down*, possibilitando maior precisão na identificação e saliência dos objetos.

De acordo com as simulações e análises apresentadas na Seção 5.3.1, o número de objetos concorrentes à atenção visual pode influenciar diretamente no comportamento do modelo. Embora tenha sido proposto um mecanismo para a seleção prévia de objetos, onde objetos com valores saliências abaixo de um determinado limiar não tornam-se aptos a participarem da competição pela atenção visual, alguns segmentos ainda poderiam apresentar valores de saliência obtidos a partir da saliência de segmentos vizinhos. Isto pode ocorrer devido à disposição da saliência apresentada nos mapas de conspicuidades utilizados, uma vez que estes mapas destacam a saliência baseadas no espaço, e não em objetos. Desta forma, objetos que não deveriam apresentar valores de saliência, sofrem a influência da vizinhança.

Outra consideração, agora relacionada ao número de segmentos gerados ini-

cialmente, diz repeito ao custo computacional necessário para a segmentação de toda a cena, o que inviabiliza a aplicabilidade do modelo proposto em sistemas de tempo real como, por exemplo, a saliência baseada em objetos a partir de imagens dinâmicas, como por exemplo, em vídeo.

Com base nas considerações apresentadas, tem-se ainda o objetivo de propor um mecanismo de enviesamento *top-down* do processo de segmentação, de modo que a segmentação ocorra somente em regiões enviesadas. Pretendemos com isso, reduzir o número de segmentos e, consequentemente, aumentar a taxa de seleção visual do modelo.

Assim, a pesquisa apresentada na Seção 5.3.2 mostrou-se apta a salientar objetos relacionados às intenções do observador. De maneira específica, através do ajuste de pesos relacionados às características visuais primitivas, foi possível realizar o enviesamento prévio do processo de segmentação. Concluímos que o comportamento do modelo proposto possa ser considerado plausível biologicamente, de forma que, em determinadas situações, objetos com baixos valores de contraste são automaticamente desconsiderados durante uma tarefa de busca visual. Por exemplo, durante a busca por um alvo de cor "vermelha", a característica cor será altamente enviesada, enquanto que as demais serão desconsideradas. Assim, informações irrelevantes da cena, de acordo com uma tarefa de busca específica, podem ser desconsideradas.

Como conclusão final dos modelos propostos, no Capítulo 6 simulações em 104 imagens de domínio heterogêneos, psicofísicos e homogêneos, demonstraram a eficiência dos modelos propostos para a predição dos objetos da cena que observadores humanos tendem a fixar, fornecendo bons resultados em todos os cenários analisados.

7.1 Trabalhos Futuros

O atenção visual é uma área de pesquisa amplamente estuda atualmente. Avanços em pesquisas nesta área poderão proporcionar uma maior compreensão da atenção visual humana e, consequentemente, possibilitar a proposta de novos modelos computacionais.

Por outro lado, existem vários aspectos nos quais os modelos propostos nesta tese podem ser melhorados e estendidos. Destacamos os seguintes:

- Implementação em hardware com arquitetura paralela: de acordo com as características individuais dos módulos que compõem os modelos para atenção visual propostos, sua implementação paralela em hardwares dedicados pode ser considerada uma interessante aplicação para sistemas de tempo real;

- Consideração da característica de movimento: considerando aplicações de tempo real, uma característica importante que pode ser considerada para o direciona-

mento da atenção visual é o enviesamento *top-down* do processo de segmentação baseado no contraste de movimento na cena;

- Mecanismo atencional aplicado ao controle de robôs: uma possível aplicabilidade dos modelos propostos nesta tese é a navegação autônoma de robôs, contudo, para o desenvolvimento da atenção visual demais informações sensoriais também podem ser utilizadas como, por exemplo, dados gerados a partir de escaneamentos em 3D ou ainda informações auditivas presentes no ambiente;

- Aprendizagem de valores de parâmetros: embora tenham sido apresentados ajustes específicos de valores de parâmetros para determinados domínios, em condições heterogêneas, a proposta de mecanismos para o aprendizado automático de parâmetros é uma área interessante a ser pesquisada;

- Aplicações para robótica educacional: diante da capacidade em identificar regiões de maior saliência em uma cena, os modelos para atenção visual propostos podem servir de inspiração para diversas aplicações na área da educação como, por exemplo, o auxílio à aprendizagem de símbolos e formas, de modo que o alvo apresentado pelo usuário seja localizado de acordo com suas características salientes e, em seguida, reconhecido, possibilitando uma interação construtivista.

Referências Bibliográficas

(1999). Human perception and performance. *Journal of Experimental Psychology 25*.

Achanta, R., S. Hemami, F. Estrada, and S. Sï¿½sstrunk (2009). Frequency-tuned Salient Region Detection. In *IEEE CVPR*, pp. 1597 – 1604.

Bacon, W. and H. Egeth (1994). Overriding stimulus-driven attentional capture. *Perception & Psychophysics 55*, 485–496.

Benicasa, A. X., M. G. Quiles, L. Zhao, and R. A. Romero (2012b, oct.). An object-based visual selection model with bottom-up and top-down modulations. In *Neural Networks (SBRN), 2012 Brazilian Symposium on*, Curitiba,PR-Brasil, pp. 238 –243.

Benicasa, A. X., M. G. Quiles, L. Zhao, T. C. Silva, and R. A. Romero (2013). A unified top-down and bottom-up model for object-based visual selection [submetido para publicação].

Benicasa, A. X., M. G. Q. Quiles, L. Zhao, and R. A. Romero (2013b). Top-down biasing and modulation for object-based visual attention. In *The 20th International Conference on Neural Information Processing (ICONIP'2013),*, Daegu,Korea.

Benicasa, A. X. and R. A. F. Romero (2010, oct.). Localization of salient objects in scenes through visual attention. In *Neural Networks (SBRN), 2010 Eleventh Brazilian Symposium on*, São José dos Campos-SP, pp. 103 –108.

Benicasa, A. X., L. Zhao, and R. A. Romero (2012, june). Model of top-down / bottom-up visual attention for location of salient objects in specific domains. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, Brisbane-AU, pp. 1582–1589.

Bonaiuto, J. and L. Itti (2006). Using attention and spatial information for rapid facial recognition in video. *Image and Vision Computing 24(6)*, 557–563.

Borji, A., M. N. Ahmadabadi, and B. N. Araabi (2011). Cost-sensitive learning of top-down modulation for attentional control. *Machine Vision and Applications 22*(1), 61–76.

Borji, A. and L. Itti (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(1), 185–207.

Braithwaite, J. and G. Humphreys (2003). Inhibition and anticipation in visual search: Evidence from effects of color foreknowledge on preview search. *Perception & Psychophysics 65*, 213–237.

Bruce, N. and J. Tsotsos (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision 9*(3), 1–24.

Buia, C. and P. Tiesinga (2006). Attentional modulation of firing rate and synchrony in a model cortical network. *Journal of Computational Neuroscience 20*, 247–264.

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

Campbell, S. R. and D. Wang (1996). Synchronization and desynchronization in a network of locally coupled wilson-cowan oscillators. *IEEE Transactions on Neural Networks 7*, 541–554.

Campbell, S. R., D. L. Wang, and C. Jayaprakash (1999). Synchrony and desynchrony in integrate-and-fire oscillators. *Neural Computation 11*, 1595–1619.

Carota, L., G. Indiveri, and V. Dante (2004). A softwarehardware selective attention system. *Neurocomputing 58-60*, 647–653.

Cheng, H. D., X. H. Jiang, Y. Sun, and J. Wang (2001). Color image segmentation: advances and prospects. *Pattern Recognition 34*, 2259–2281.

Cheng, M., G. Zhang, N. Mitra, X. Huang, and S. Hu (2011). Global contrast based salient region detection. In *IEEE CVPR*, pp. 409–416.

Clark, J. and N. Ferrier (1988). Modal control of an attentive vision system. *Proc. Int. Conf. Computer Vision (ICCV'88)*.

Clark, J. and N. Ferrier (1989). Control of visual attention in mobile robots. In *Robotics and Automation, 1989. Proceedings., 1989 IEEE International Conference on*, pp. 826–831. IEEE.

Cohen, J. D. and D. Servan-Schreiber (1992). The prefrontal cortex and cognitive control. *Psychological Review 99*(1), 45–77.

Connor, C. E., H. E. Egeth, and S. Yantis (2004). Visual attention: Bottom-up versus top-down. *Current biology 14*, 850–852.

Corbetta, M. (1998). Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? *Proceedings of the National Academy of Sciences 95*(3), 831–838.

D., S. (1988). The combination of spatial frequency and orientation is effortlessly perceived. *Percept. Psychophys 43*, 601–603.

Deco, G. and E. T. Rolls (2005). *Neurobiology of Attention*, Chapter 100 - The Role of Short-Term Memory in Visual Attention, pp. 610–617. Elsevier, Oxford.

Desimone, R. and J. Duncan (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience 18*, 193–222.

Egeth, H. E. and S. Yantis (1997). Visual attention: Control, representation, and time course. *Annual Review of Psychology 48*(1), 269–297.

Elazary, L. and L. Itti (2008). Interesting objects are visually salient. *Journal of Vision 8(3)*, 1–15.

Elazary, L. and L. Itti (2010). A bayesian model for efficient visual search and recognition. *Vision Research 50*(14), 1338–1352.

Engel, S., X. Zhang, and B. Wandell (1997). Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature 388*(6637), 68–71.

Feldman, J. (1982). Dynamic connections in neural networks. *Biological Cybernetics 46*, 27–39.

Felzenszwalb, P. F. and D. P. Huttenlocher (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision 59*(2), 167–181.

Findlay, J. M. and I. D. Gilchrist (2001). *Active vision perspective. In Vision & Attention, M. Jenkin and L. R. Harris*, Chapter Chapter 5, pp. 83–103. Eds. Springer Verlag.

Fries, P., J. H. Reynolds, A. E. Rorie, and R. Desimone (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science 291*, 1560–1563.

Frintrop, S. (2006). *VOCUS - A Visual Attention System of Object Detection and Goal-directed Search*. Ph. D. thesis, PhD thesis, Lecture Notes in Artificial Intelligence (LNAI).

Frintrop, S., E. Rome, and H. I. Christensen (2010, January). Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept. 7*(1), 6:1–6:39.

Greenspan, H., S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C. H. Anderson (1994). Overcomplete steerable pyramid filters and rotation invariance. *IEEE Computer Vision and Pattern Recognition*, 222–228.

Harel, J., C. Koch, and P. Perona (2006). Graph-based visual saliency. In *Advances in neural information processing systems*, pp. 545–552.

Haykin, S. (2001). *Redes Neurais - Princípios e Práticas*. Bookman.

Hopfield, J. and A. V. M. Herz (1995). Rapid local synchronization of action potentials: Toward computation with coupled integrate-and-fire oscillator neurons. *Proceedings of the National Academy of Sciences of the USA 92*, 6655–6662.

Hulle, M. M. V. (2000). *Faithful Representations and Topographic Maps: From Distortion- to Information-Based Self-Organization*. New York, NY, USA: John Wiley & Sons, Inc.

Itti, L. (2005). *Neurobiology of Attention*, Chapter 94 - Models of bottom-up attention and saliency, pp. 576–582. Elsevier, Oxford.

Itti, L. and C. Koch (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research 40*, 1489–1506.

Itti, L. and C. Koch (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience 2*, 194–203.

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

Izhikevich, E. (2006). Polychronization: Computation with spikes. *Neural Computation 18*, 245–282.

Izhikevich, E. M. (2004). Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks 15*(5), 1063–1070.

Jermakowicz, W. J. and V. A. Casagrande (2007). Neural networks a century after cajal. *Brain Research Reviews 55*(2), 264–284.

Judd, T., F. Durand, and A. Torralba (2012). A benchmark of computational models of saliency to predict human fixations. *MIT Computer Science and Artificial Intelligence Laboratory Technical Report*.

Kandel, E. R., J. H. Schwartz, and T. M. Jessell (1997). *Fundamentos da neurociência e do comportamento*. Prentice-Hall.

Kazanovich, Y. B. and R. Borisyuk (1999). Dynamics of neural networks with a central element. *Neural Networks 12*(3), 441–454.

Kazanovich, Y. B. and R. Borisyuk (2002). Object selection by an oscillatory neural network. *BioSyst. 67*, 103–111.

Kim, M.-S. and K. Cave (1999). Top-down and bottom-up attentional control: On the nature of interference from a salient distractor. *Perception & Psychophysics 61*, 1009–1023.

Kim, Y. J., M. Grabowecky, K. A. Paller, K. Muthu, and S. Suzuki (2007). Attention induces synchronization-based response gain in steady-state visual evoked potentials. *Nature Neuroscience 10*(1), 117–125.

Koch, C. and S. Ullman (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology 4*, 219–227.

Kohonen, T. (2001). *Self-Organizing Maps* (3th Edition ed.). Information Science. Springer.

Lamy, D., Y. Tsal, and H. Egeth (2003). Does a salient distractor capture attention early in processing? *Psychonomic Bulletin & Review 10*, 621–629.

Lau, J. (2013, May). Philosophy & cognitive science. in http://philosophy.hku.hk/courses/cogsci/ncc.php.

LindBlad, T., K. J. (2005). *Image Processing Using Pulse-Coupled Neural Network*. Secaucus, NJ, USA: Springer.

Liu, X., K. Chen, and D. Wang (2001). Extraction of hydrographic regions from remote sensing images using an oscillator network with weight adaptation. *Geoscience and Remote Sensing, IEEE Transactions on 39*(1), 207–211.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, Volume 2, pp. 1150–1157 vol.2.

Álvaro Corral, C. J. Pérez, A. Díaz-Guilera, and A. Arenas (1995). Synchronization in a lattice model of pulse-coupled oscillators. *Physical Review Letters 75*(20), 3697–3700.

Maass, W. (1997). Networks of spiking neurons: The third generation of neural network models. *Neural Networks 10(9)*, 1659–1671.

McCulloch, W. S. and W. Pitts (1943). A logical calculus of the ideias immanente in nervous activity. *Bulletin of Mathematical Biophysics 5*, 115–133.

Miller, E. (2000). The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience 1*(1), 59–66.

Mirollo, R. E. and S. H. Strogatz (1990). Synchronization of pulse-coupled biological oscillators. *SIAM J. Appl. Math. 50*(6), 1645–1662.

Murthy, A., K. G. Thompson, and J. D. Schall (2001). Dynamic dissociation of visual selection from saccade programming in frontal eye field. *Journal of Neurophysiology 86*(5), 2634–2637.

Navalpakkam, V. and L. Itti (2005). Modeling the influence of task on attention. *Vision research 45*(2), 205–231.

Navalpakkam, V. and L. Itti (2006a, Jun). An integrated model of top-down and bottom-up attention for optimal object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, pp. 2049–2056.

Navalpakkam, V. and L. Itti (2006b). Top–down attention selection is fine grained. *Journal of Vision 6*(11).

Niebur, E. and C. Koch (1994). A model for neuronal implementation of selective visual attention based on temporal correlation among neurons. *Journal of Computational Neuroscience 1*, 141–158.

Nothdurft, H. (2005). *Salience of feature contrast. In Neurobiology of Attention, L. Itti, G. Rees, and J. K. Tsotsos*, pp. 233–239. Eds. Elsevier.

O'Craven, K. M., P. E. Downing, and N. Kanwisher (1999). fmri evidence for objects as the units of attentional selection. *Nature 401*(6753), 584–587.

Ogawa, T. and H. Komatsu (2004). Target selection in area v4 during a multidimensional visual search task. *Journal of Neuroscience 24*(28), 6371–6382.

Oliva, A., A. Torralba, M. S. Castelhano, and J. M. Henderson (2003). Top-down control of visual attention in object detection. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, Volume 1, pp. I–253. IEEE.

Pashler, H. (1998). Introduction. in h. pashler (org.). *Attention. Hove (Reino Unido): Psychology Press.*.

Quiles, M., L. Zhao, and R. Romero (2007). A selection mechanism based on a pulse-coupled neural network. In *The 2007 International Joint Conference on Neural Networks (IEEE-IJCNN2007)*, Orlando-US, pp. 1–6.

Quiles, M. G., R. A. F. Romero, and L. Zhao (2006). A pulse-coupled neural network as a simplified bottom-up visual attention model. In *IEEE Proceedings of the Ninth Brazilian Symposium on Artificial Neural Networks (SBRN'2006)*, pp. 1–6.

Quiles, M. G., D. Wang, L. Zhao, R. A. Romero, and D.-S. Huang (2011). Selecting salient objects in real scenes: An oscillatory correlation model. *Neural Networks 24*(1), 54 – 64.

Quiles, M. G., L. Zhao, F. Breve, and R. A. F. Romero (2009). A network of integrate and fire neurons for visual selection. *Neurocomputing 72*, 2198–2208.

Rayner, K. (1998). Eye movements in reading and information processing. *Psychological Bulletin 85*(3), 618–660.

Riesenhuber, M. and T. Poggio (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience 2*(11), 1019–1025.

Roelfsema, P. R., V. A. F. Lamme, and H. Spekreijse (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature 395*, 376–381.

Rosenholtz, R., A. L. Nagy, and N. R. Bell (2004). The effect of background color on asymmetries in color search. *Journal of Vision 4*(3).

Rossini, J. C. and C. Galera (2006). Atenção visual: estudos comportamentais da seleção baseada no espaço e no objeto. *Estudos de Psicologia 11(1)*, 79–86.

Rutishauser, U., D. Walther, C. Koch, and P. Perona (2004). Is bottom-up attention useful for object recognition? In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, Volume 2, pp. II–37. IEEE.

Sejnowski, T. J. and O. Paulsen (2006). Network oscillations: Emerging computational principles. *The Journal of Neuroscience 26*(6), 1673–1676.

Shareef, N., D. L. Wang, and R. Yagel (1999). Segmentation of medical images using legion. *Medical Imaging, IEEE Transactions on 18*(1), 74–91.

Shic, F. and B. Scassellati (2007). A behavioral analysis of computational models of visual attention. *International Journal of Computer Vision 73*(2), 159–177.

Siagian, C. and L. Itti (2007). Rapid biologically-inspired scene classification using features shared with visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 29*(2), 300–312.

Siklossy, I. (2005). *Mimicking the visual pathway*. Ph. D. thesis, University of Edinburgh - Division of Informatics, Edinburgh.

Silva, T. C. and L. Zhao (2012). Network-based high level data classification. *IEEE Transactions on Neural Networks 23*, 954–970.

Singer, W. and C. M. Gray (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience 18*(1), 555–586. PMID: 7605074.

Terman, D. and D. Wang (1995). Global competition and local cooperation in a network of neural oscillators. *Physica D 81*, 148–176.

Theeuwes, J. (1991). Cross-dimensional perceptual selectivity. *Perception & Psychophysics 50*, 184–193.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics 51*, 599–606.

Treisman, A. (1998). Feature binding, attention and object perception. *353*, 1295–1306.

Treisman, A. and S. Gormican (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychol. Rev. 95*, 15–48.

Treisman, A. M. and G. Gelade (1980). A feature-integration theory of attention. *Cognitive Psychology 12*(1), 97 – 136.

Tsotsos, J. K. (1992). On the relative complexity of active vs. passive visual search. *International Journal of Computer Vision 7*, 127–141.

Tsotsos, J. K. (2011). *A computational perspective on visual attention*. MIT Press, Cambridge.

Tsotsos, J. K., S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo (1995). Modeling visual attention via selective tuning. *Artificial Intelligence 78*, 507–545.

van der Pol, B. (1926). Relaxation oscillations. *Philosophical Magazine 2(11)*, 978–992.

von der Malsburg, C. (1981). The correlation theory of brain function. Technical report, Internal report 81-2: Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany.

von der Malsburg, C. and W. Schneider (1986). A neural cocktail-party processor. *Biological Cybernetics 54*, 29–40.

Walther, D., L. Itti, M. Riesenhuber, T. Poggio, and C. Koch (2002). Attentional selection for object recognition-a gentle way.

Walther, D. and C. Koch (2006). Modeling attention to salient proto-objects. *Neural Networks 19*(9), 1395–1407.

Walther, D., U. Rutishauser, C. Cock, and P. Perona (2005). Selective visual attention enables learning and recognition of multiples objects in cluttered scenes. *Computer Vision and Image Understanding 100*, 41–63.

Wang, D. (1996). Primitive auditory segregation based on oscillatory correlation. *Cognitive Science 20*(3), 409–456.

Wang, D. (1999). Object selection based on oscillatory correlation. *Neural Networks 12*, 579–592.

Wang, D. (2005). The time dimension for scene analysis. *IEEE Transactions on Neural Networks 16*(6), 1401–1426.

Wang, D. and G. J. Brown (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks 10*, 684–697.

Wang, D. and D. Terman (1995). Locally excitatory globally inhibitory oscillator networks. *IEEE Transactions on Neural Networks 6*(1), 283–286.

Wang, D. and D. Terman (1997). Image segmentation based on oscillatory correlation. *Neural Computation 9*, 805–836.

Wang, D., L. X. (2002). Scene analysis by integrating primitive segmentation and associative memory. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics 32(3)*, 254–268.

Wolfe, J. M., H. T. S. (2004). What attributes guide the deployment of visual attention and how do they do it ? *Nature Review Neuroscience 5*, 495–501.

Wolfe, J. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review 1*, 202–238.

Wolfe, J. (2005). Guidance of visual search by preattentive information. *Neurobiology of attention*, 101–104.

Wolfe, J., K. Cave, and S. Franzel (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance 15*(3), 419.

Wolfe, J. M. (2007). Guided search 4.0 - current progress with a model of visual search. *Integrated models of cognitive systems*, 99–119.

Wolfe, J. M. and T. S. Horowitz (2004). What attributes guide the deployment of visual attention and how do they do it ? *Nature Review Neuroscience 5*, 495–501.

Wright, R. D. and L. M. Ward (1998). The control of visual attention. *Visual attention*, 132–186.

Wrigley, S. and G. J. Brown (2004). A computational model of auditory selective attention. *Neural Networks, IEEE Transactions on 15*(5), 1151–1163.

Yantis, S. (2000). *Attention and Performance XVIII*, Volume 18, Chapter Goal-directed and stimulus-driven determinants of attentional control, pp. 73–103. MIT Press, Cambridge.

Yegnanarayana, B. (2005). *Artificial Neural Networks*. Prentice-Hall.

Zuchini, M. H. (2003). *Aplicações de mapas auto-organizáveis em mineração de dados e recuperação de informação*. Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.