UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
MESTRADO EM INFORMÁTICA

HÉLIO PERRONI FILHO

Predição de Mapas de Profundidades A Partir de Imagens Monoculares Por Meio de Redes Neurais Sem Peso

Vitória – ES
2010

HÉLIO PERRONI FILHO

Predição de Mapas de Profundidades A Partir de Imagens Monoculares Por Meio de Redes Neurais Sem Peso

Dissertação apresentada ao Mestrado de Informática do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Mestre em Informática.

Orientador: Prof. Dr. Alberto Ferreira De Souza.

Vitória – ES
2010

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil)

Perroni Filho, Hélio, 1979-
P459p Predição de mapas de profundidade a partir de imagens monoculares por RNSP's / Hélio Perroni Filho. – 2010.
92 f. : il.

Orientador: Alberto Ferreira de Souza.
Dissertação (mestrado) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Visão artificial. 2. Redes neurais (Computação). 3. Aprendizado do computador. 4. Profundidade - Percepção. I. Souza, Alberto Ferreira de. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 004

HÉLIO PERRONI FILHO

Predição de Mapas de Profundidades A Partir de Imagens Monoculares Por Meio de Redes Neurais Sem Peso

Dissertação apresentada ao Mestrado de Informática do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Mestre em Informática.

Aprovada em 27 de Fevereiro de 2010.

COMISSÃO EXAMINADORA

Prof. Dr. Aberto Ferreira De Souza
Universidade Federal do Espírito Santo
Orientador

Profa. Dra. Claudine Santos Badue Gonçalves
Universidade Federal do Espírito Santo

Prof. Dr. Wagner Meira Jr.
Universidade Federal de Minas Gerais

Vitória – ES
2010

Para minha família.

AGRADECIMENTOS

Agradeço aos meus pais, Hélio e Diana, por representarem os pilares dos meus valores pessoais, e modelos nos quais me inspiro para continuar evoluindo enquanto indivíduo.

Agradeço ao professor Alberto Ferreira De Souza, por sua amizade e orientação, além da paciência e incentivo incansáveis, que tornaram possível o desenvolvimento deste trabalho.

Agradeço aos professores Claudine Santos Badue Gonçalves e Wagner Meira Jr. por terem gentilmente aceitado participar de minha avaliação, mesmo que convidados com pouca antecedência.

Aos meus amigos do LCAD André Gustavo Almeida, Avelino Forechi e Felipe Pedroni, pelo apoio inestimável durante o desenvolvimento deste trabalho.

Aos professores e funcionários do Departamento de Informática da UFES, pelo bom trabalho e a boa convivência no decorrer do curso.

Por fim, gostaria de agradecer aos vários amigos de casa, São Paulo, dentre eles Gustavo Saita, Vanderlei Andrade, Marco Oliveira e Thiago Nishio, e à minha querida Juliana Shizue Nishio, pela amizade e incentivo.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Um problema central para a Visão Computacional é o de *depth estimation* (“estimativa de profundidades”) – isto é, derivar, a partir de uma ou mais imagens de uma cena, um *depth map* (“mapa de profundidades”) que determine as distâncias entre o observador e cada ponto das várias superfícies capturadas. Não é surpresa, portanto, que a abordagem de *stereo correspondence* (“correspondência estéreo”), tradicionalmente usada nesse problema, seja um dos tópicos mais intensamente investigados do campo.

Sistemas de correspondência estimam profundidades a partir de características *binoculares* do par estéreo – mais especificamente, a diferença entre as posições de cada ponto em um par de imagens. Além dessa informação puramente geométrica, imagens contêm uma série de características *monoculares* – tais como variações e gradientes de textura, variações de foco, padrões de cores e reflexão, etc. – que podem ser exploradas para derivar estimativas de profundidade. Para isso, entretanto, é preciso acumular certa quantidade de conhecimento *a priori,* uma vez que há uma ambiguidade intrínseca entre as características de uma imagem e variações de profundidade.

Através de suas pesquisas com sistemas de aprendizado de máquina baseados em *Markov Random Fields* (MRF’s), Ashutosh Saxena demonstrou ser possível estimar mapas de profundidades com grande precisão a partir de imagens monoculares estáticas. Sua abordagem, entretanto, carece de plausibilidade biológica, visto que não há correspondência teórica conhecida entre MRF’s e as redes neurais do cérebro humano.

Motivados por sucessos anteriores na aplicação de *Weightless Neural Networks* (“Redes Neurais Sem Peso" – RNSP's) a problemas de visão computacional, neste trabalho objetivamos investigar a efetividade da aplicação de RNSP’s ao problema de estimar mapas de profundidades. Com isso, esperamos alcançar uma melhoria em relação ao sistema baseado em MRF’s de Saxena, além de desenvolver uma arquitetura mais útil para a avaliação de hipóteses sobre o processamento de informações visuais no córtex humano.

ABSTRACT

Depth estimation – taking one or more images from a scene and estimating a *depth map,* which determines distances between the observer and points taken from various object surfaces – is a central problem in computer vision. It's not surprising, then, that the approach of *stereo correspondence,* traditionally applied to this problem, is one of the most intensively studied topics in the field.

Stereo correspondence systems estimate depths from *binocular* features – more specifically, the difference between the positions of each point in a pair of images. Besides this purely geometrical information, images contains many *monocular* features – such as texture variations and gradients, focus, color patterns and reflection – which can be explored to derive depth estimates. For this, however, a certain amount of *a priori* knowledge must be gathered, since there is an inherent ambiguity between an image's characteristics and depth variations.

Through his research on machine-learning systems based on *Markov Random Fields* (MRF's), Ashutosh Saxena proved that it is possible to estimate very accurate depth maps from a single monocular image. His approach, however, lacks biological plausibility, since there is no known theoretical correspondence between MRF's and the human brain's neural networks.

Motivated by past successes in applying *Weightless Neural Networks* (WNN's) to computer vision problems, in this paper we investigate the effectiveness of applying WNN's to the problem of depth map estimation. With this, we hope to achieve performance improvements over Saxena's MRF-based approach, and develop a more useful architecture for evaluating hypotheses about visual information processing in the human cortex.

1. INTRODUÇÃO

Tal como câmeras, os olhos humanos captam imagens do ambiente num formato bidimensional. De posse desta informação, o cérebro – em particular as áreas primária (V1) e temporal medial (MT) do córtex visual, responsáveis pela maior parte do processamento da percepção de profundidade – é capaz de construir uma representação tridimensional do mundo exterior, inferindo a profundidade dos objetos. Basicamente, três categorias de informação são usadas para isso: *estereopsis, desvio de paralaxe* e *pistas monoculares* [MIC05].

Humanos parecem ser extremamente competentes em estimar profundidades a partir de imagens monoculares estáticas [LOO01]. Isso é feito explorando pistas monoculares tais como variações e gradientes de textura, oclusão, objetos de tamanho conhecido, enevoamento, etc. [SAX05]. Ashutosh Saxena, através de sua pesquisa com algoritmos para o aprendizado de máquina baseados em *Markov Random Fields* (MRF's), demonstrou [SAX08] que é possível estimar mapas de profundidade a partir unicamente de informações monoculares; entretanto, seus resultados não possuem plausibilidade biológica (pelo menos não aparente), uma vez que, há nosso ver, não há relação entre MRF's e o funcionamento do córtex visual.

Weightless Neural Networks ("Redes Neurais Sem Peso" – RNSP's) modelam a dinâmica de modulação de entradas encontrada nas árvores dendríticas de neurônios biológicos [ALE09], sendo, portanto, uma alternativa de sistema de aprendizado de máquina mais próxima do domínio biológico do que os MRF's. Sucessos anteriores na sua aplicação a problemas de visão computacional (veja, por exemplo, [ALB08]) indicam a sua viabilidade também como base para desenvolvimento de sistemas de produção, como soluções de visão para robôs.

Neste trabalho, buscou-se examinar algumas alternativas de arquiteturas de RNSP capazes de inferir profundidades a partir de imagens monoculares. Com isso, esperávamos alcançar uma melhoria de desempenho em relação ao sistema baseado em MRF's de Saxena, além de desenvolver uma arquitetura mais útil para a avaliação de hipóteses sobre o processamento de informações visuais no córtex humano. Embora não tenha sido possível superar o nível de precisão obtido por Saxena – um objetivo para pesquisas adicionais – nossos resultados são coerentes com os dele, e permitem inferir importantes características do córtex visual.

Para simplificar a operação e avaliação das arquiteturas neurais, desenvolvemos uma aplicação visual chamada *Diver,* que oferece um número de facilidades para controle das RNSP's, manipulação e visualização de dados, execução de treinamentos e testes, e compilação de resultados experimentais. Diver integra-se à plataforma de pesquisa *MAE* (Máquina Associadora de Eventos), sobre a qual as redes neurais descritas neste trabalho foram implementadas. O código do Diver e das redes neurais, as imagens e mapas de profundidades usados no treinamento das redes, e os resultados experimentais obtidos nos testes estão disponíveis para download na Internet em www.lcad.inf.ufes.br/wiki/index.php/Diver.

O restante deste trabalho está estruturado como segue. No Capítulo 2, discutimos o problema da estimativa de profundidades em uma cena a partir de uma única imagem monocular. Apresentamos os mecanismos da visão humana responsáveis pelo reconhecimento de profundidades, em particular as pistas monoculares exploradas pelo cérebro; relacionamos quais informações são ou não usadas pelas nossas redes neurais; e apresentamos a solução empregada por Saxena. No Capítulo 3, detalhamos a estrutura e o funcionamento das redes neurais sem peso, com ênfase no modelo VG-RAM, usado pela MAE; e descrevemos as arquiteturas estudadas no trabalho.

No Capítulo 4, apresentamos a metodologia empregada no trabalho, os experimentos efetuados e os resultados obtidos. Começamos apresentando as características da base de dados disponibilizada por Saxena; as várias adaptações que precisaram ser realizadas para adequá-la ao cenário de teste original e às nossas necessidades; e as métricas utilizadas para avaliar quantitativamente o desempenho do sistema em relação aos dados de referência.

No Capítulo 5, descrevemos as sessões de validação realizadas para determinar os parâmetros ótimos de configuração para cada arquitetura neural, além das sessões de testes para avaliar o desempenho final. Passamos então à análise e comentário dos resultados obtidos, comparando também com os alcançados por Saxena.

No Capítulo 6, apresentamos uma discussão sobre o trabalho. Em primeiro lugar, relacionamos outros trabalhos que abordaram o problema de estimativa de profundidades em uma cena a partir de uma única imagem monocular, inclusive o trabalho de Saxena [SAX08]. Em seguida, passamos à análise crítica deste trabalho,

suas contribuições e deficiências. Finalmente, no Capítulo 7, concluímos com um resumo do trabalho, seus resultados e direções para trabalhos futuros.

2. ESTIMATIVA DE PROFUNDIDADES A PARTIR DE IMAGENS MONOCULARES

Neste capítulo, discutimos o problema da estimativa de profundidades em uma cena a partir de uma única imagem monocular. Apresentamos os mecanismos da visão humana responsáveis pelo reconhecimento de profundidades, em particular as pistas monoculares exploradas pelo cérebro; relacionamos quais informações são ou não usadas pelas nossas redes neurais; e apresentamos a solução empregada por Saxena [SAX08].

2.1. SISTEMA VISUAL HUMANO

Nesta seção é descrito sumariamente o sistema visual humano. Ela apresenta conceitos e termos que são essenciais para compreender as contribuições deste trabalho. Seu conteúdo foi, fundamentalmente, extraído de [OLI05].

2.1.1. O olho

O globo ocular, com cerca de 25 milímetros de diâmetro, é o responsável pela captação da luz refletida pelos objetos. Anatomicamente, o globo ocular fica alojado em uma cavidade formada por vários ossos chamada órbita e é constituído por três túnicas (camadas): túnica fibrosa externa, túnica intermédia vascular pigmentada e túnica interna nervosa.

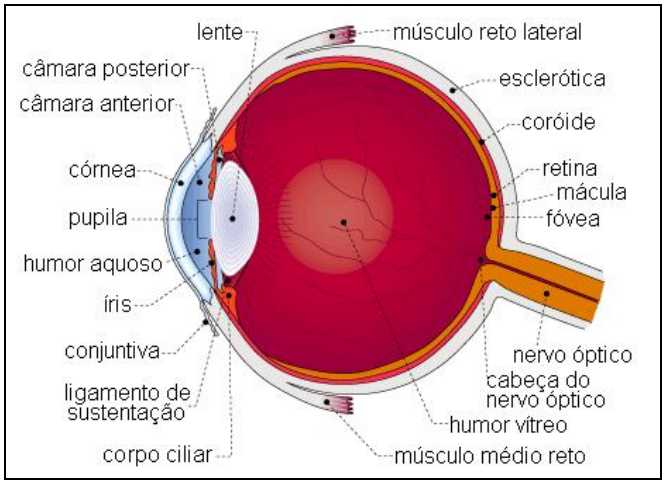


Figura 2-1 – Anatomia do olho humano. Corte medial horizontal do olho direito visto de cima. Figura retirada de http://www.escolavesper.com.br/olho_humano.htm

Na Figura 2-1 estão representadas todas as partes que formam as túnicas fibrosa externa, intermédia vascular pigmentada e a túnica interna nervosa. A túnica fibrosa externa ou esclerótica, também chamada de “branco do olho”, tem uma função protetora. É resistente, de tecido fibroso e elástico, e envolve externamente o olho (globo ocular). A maior parte da esclerótica é opaca e chama-se esclera. A ela estão conectados os músculos extra-oculares que movem os globos oculares, dirigindo-os ao seu objetivo visual. A parte anterior da esclerótica chama-se córnea, que é transparente e atua como uma lente convergente.

A túnica intermédia vascular pigmentada ou úvea compreende a coróide, o corpo ciliar e a íris. A coróide está situada abaixo da esclerótica e é bastante pigmentada para absorver a luz que chega a retina, evitando sua reflexão dentro do olho. Ela é intensamente vascularizada e tem também como função nutrir a retina. A íris é uma estrutura muscular de cor variável (parte circular que dá cor aos olhos), é opaca e tem uma abertura central, chamada pupila, por onde a luz passa. O diâmetro da pupila varia, aproximadamente de 2 mm a 8 mm, de acordo com a intensidade luminosa do ambiente. Em ambientes claros a pupila se estreita, diminuindo a passagem de luz, evitando a saturação das células detectoras de luz da retina. No escuro a pupila se dilata, aumentando a passagem de luz e sua captação pela retina. A luz que passa pela pupila atinge imediatamente o cristalino, uma lente gelatinosa que focaliza os raios luminosos sobre a retina.

O corpo ciliar é uma estrutura formada por musculatura lisa e que envolve o cristalino (a lente do olho). Ele á capaz de mudar a forma do cristalino permitindo

assim ajustar a visão para objetos próximos ou distantes. Este processo é conhecido como acomodação visual. A convergência correta do cristalino faz com que a imagem seja projetada nitidamente na retina. Se a imagem for maior ou menor que a necessária, fica fora de foco. Se o cristalino está ajustado para certa distância de um objeto, e este objeto se aproxima, a imagem perde a nitidez. Para recuperá-la, o corpo ciliar aumenta a convergência do cristalino, acomodando-o, diminuindo a distância focal. Caso o objeto se afaste, ocorre o processo inverso. Este processo está ilustrado na Figura 2-2.

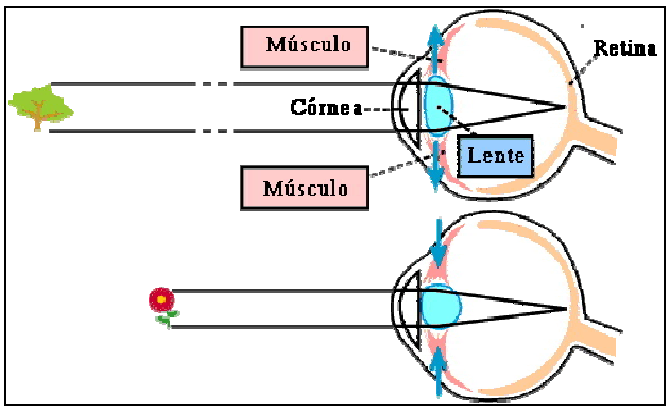


Figura 2-2 – Figura mostrando a acomodação visual do cristalino.

A túnica interna nervosa é a retina. É na retina que se formam as imagens visualizadas. A imagem projetada na retina é invertida, mas isto não causa nenhum problema já que o cérebro se adapta a isto desde o nascimento. Para que a imagem seja projetada na retina, a luz percorre o seguinte caminho: primeiramente a luz atinge a córnea, que conforme visto anteriormente é um tecido transparente, passa pela pupila, que é a abertura situada na íris que regula a intensidade de luz que entra no olho, atravessa o cristalino, que é a lente gelatinosa e que tem a função de focalizar a imagem na retina, atravessa um fluido viscoso chamado humor vítreo, que preenche a região entre o cristalino e a retina, e finalmente atinge a retina.

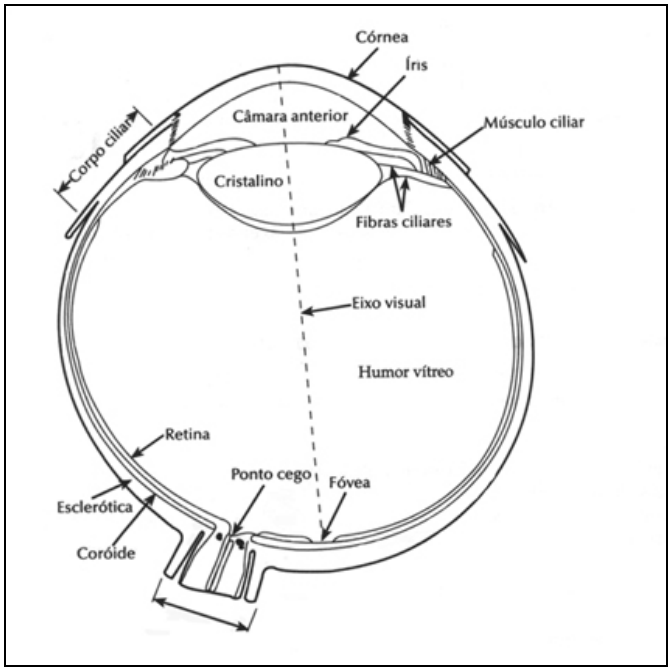


Figura 2-3 – Eixo visual. Figura retirada de http://www.on.br/glossario/alfabeto/o/olho_humano.html

A retina é composta por mais de 100 milhões de células fotossensíveis: cerca de 7 milhões de cones e entre 75 milhões e 150 milhões de bastonetes. Estas células, quando excitadas pela energia luminosa, estimulam as células nervosas adjacentes, gerando um impulso nervoso que é propagado pelo nervo óptico. A imagem fornecida pelos cones é mais nítida e rica em detalhes. Os cones são sensíveis a cores. Há três tipos de cones: um que se excita com luz vermelha, outro com luz verde e outro com luz azul. Os bastonetes não têm poder de resolução visual tão bom nem conseguem detectar cores, mas são mais sensíveis à luz. Em situações de pouca luminosidade a visão passa a depender exclusivamente dos bastonetes.

As imagens dos objetos visualizados diretamente são projetadas normalmente numa região da retina chamada *fovea centralis* ou simplesmente fóvea com cerca de 1,5 mm de diâmetro e que fica na direção da linha (eixo visual) que passa pela córnea, pupila e pelo centro do cristalino (Figura 2-3). Os cones são encontrados na retina central, em um raio de aproximadamente 10 graus a partir da fóvea. Os bastonetes estão localizados principalmente na retina periférica. O local da retina de onde sai o nervo óptico não possui cones nem bastonetes. Este local é chamado de ponto cego porque uma imagem que forme sobre este ponto, não é vista.

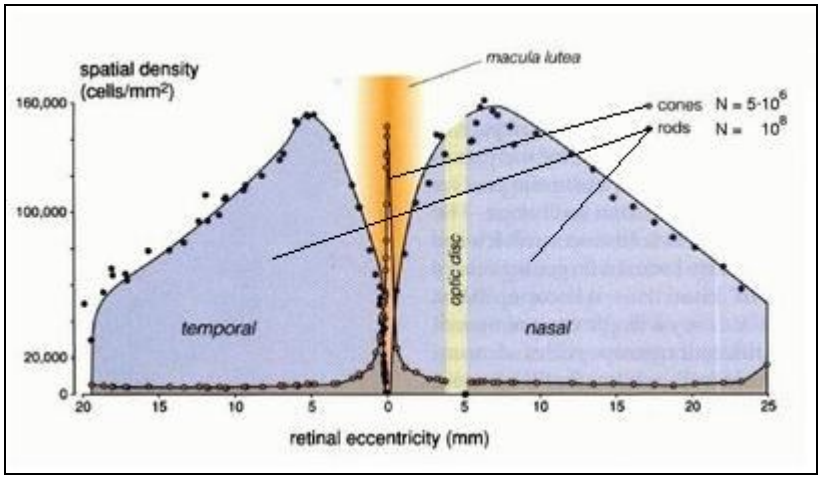


Figura 2-4 – Distribuição de cones e bastonetes (*rods*) na retina. A *macula lutea* (região da fóvea e vizinhanças) possui alta densidade de cones, enquanto que os bastonetes se concentram na periferia. O ponto cego fica na região do disco óptico (*optic disc*). Figura retirada de http://www.brainworks.uni-freiburg.de/group/wac

Os neurônios de saída da retina são as células ganglionares que projetam seus axônios através do nervo óptico, levando a informação visual para o cérebro. Cada célula ganglionar recebe informações de um conjunto de células fotorreceptoras vizinhas em uma área circunscrita na retina que é o seu campo receptivo. Duas características importantes podem ser percebidas nos campos receptivos das células ganglionares. Primeiro, são aproximadamente circulares. Segundo, são divididos em duas partes: um círculo central e um anel periférico.

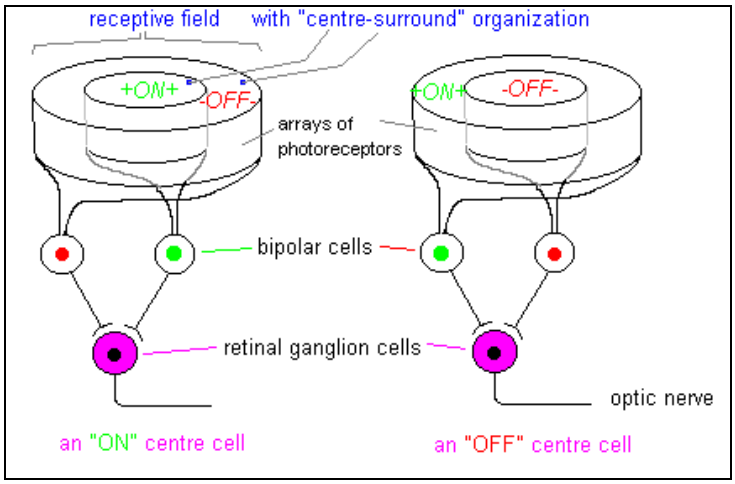


Figura 2-5 – Campo receptivo das células ganglionares. Figura retirada de http://www.cf.ac.uk/biosi/staff/jacob/teaching/sensory/vision.html

As células ganglionares respondem bem a uma iluminação diferencial entre o centro e a periferia dos seus campos receptivos. Assim, é possível identificar dois tipos de células ganglionares: *on center* e *off center.* Células ganglionares com campos receptivos *on center* ficam excitadas quando a luz estimula o centro e ficam inibidas quando a luz estimula o contorno do campo receptivo. Células *off center* funcionam ao contrário, ficam excitadas quando a luz estimula a periferia e ficam inibidas quando a luz estimula o centro do campo receptivo. Os sinais visuais de intensidade luminosa (na verdade, de contraste), depois das transformações feitas na retina, são levados até o cérebro pelo nervo ótico.

2.1.2. Fluxo de informações visuais

Nesta seção será descrito o fluxo de informações visuais em dois estágios: primeiro, a informação visual saindo da retina e indo para o mesencéfalo e tálamo (Figura 2-6), e depois, a informação saindo do tálamo para o córtex visual primário, conforme indicado na Figura 2-6. Para tanto, serão definidos alguns conceitos a seguir.

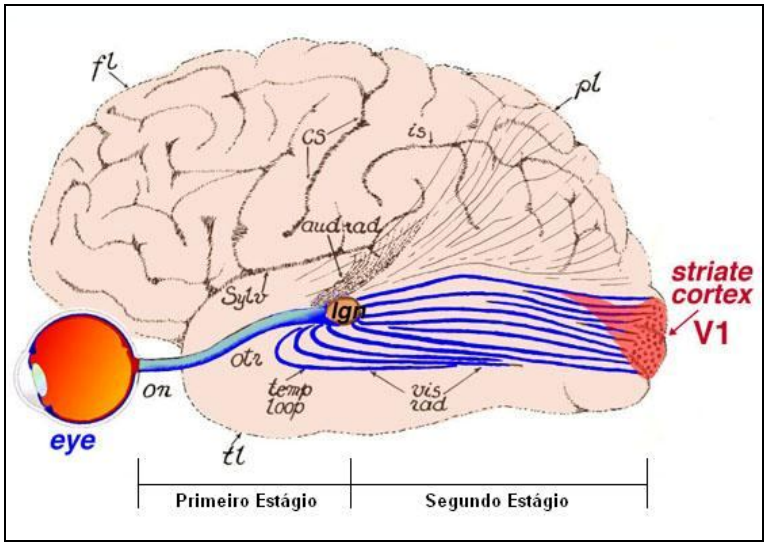


Figura 2-6 – Fluxo das informações visuais. Figura retirada de http://webvision.med.utah.edu/VisualCortex.html e alterada com iserção dos estágios.

A retina pode ser dividida em duas partes: hemirretina nasal e hemirretina temporal, cuja separação é uma linha imaginária que corta o olho de cima a baixo passando pela fóvea. Numa situação em que as fóveas de ambos os olhos estão

fixas num ponto do espaço situado em linha reta com o nariz, é possível dividir o campo visual em *left hemifield* (campo visual esquerdo) à esquerda do ponto fixo no espaço e *right hemifield* (campo visual direito) à direita do ponto fixo no espaço. O *left hemifield* é projetado na hemirretina nasal do olho esquerdo e na hemirretina temporal do olho direito. O *right hemifield* é projetado na hemirretina nasal do olho direito e na hemirretina temporal do olho esquerdo, conforme mostrado na Figura 2-7.

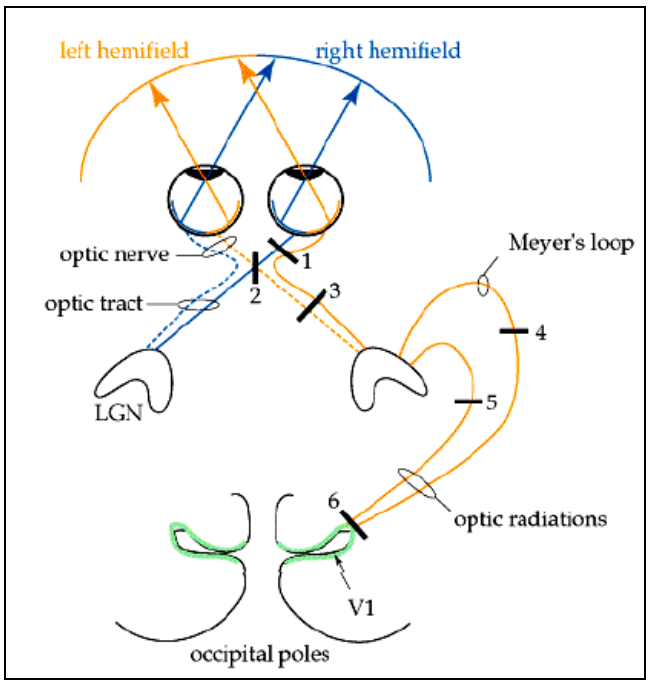


Figura 2-7 – Campo visual. (1) Nervo óptico, (2) Quiasma óptico, (3) Trato óptico. Figura retirada de http://thalamus.wustl.edu/course/basvis.html

O nervo óptico de cada olho projeta-se para o quiasma óptico, região onde é feita a separação das fibras de cada olho, em tratos (feixes de axônios) ópticos, destinadas para um mesmo lado do cérebro. Os tratos ópticos se projetam cada uma para três áreas subcorticais simétricas (que existem nos dois lados de cérebro): *região pretectal* ou *pretectum,* o *superior culliculus* do mesencéfalo e o *lateral geniculate nucleus* (LGN) do tálamo, conforme mostra a Figura 2-8.

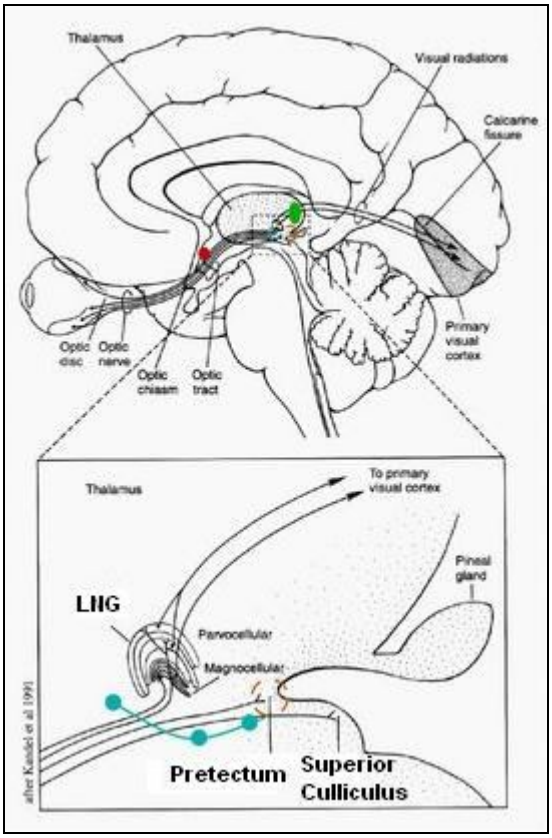


Figura 2-8 – Projeções da retina no mesencéfalo. Figura retirada de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/

A região pretectal do mesencéfalo possui células que se projetam bilateralmente para os neurônios do sistema simpático/parassimpático que controla os reflexos pupilares, contraindo e dilatando a pupila de acordo com a quantidade de luz que incide nos olhos. O *superior culliculus* é uma estrutura de camadas alternantes cinzentas e brancas localizada no teto do mesencéfalo. As células das camadas superficiais projetam-se para uma vasta área do córtex cerebral, formando uma via indireta da retina para o córtex cerebral. As camadas superficiais também recebem sinais provenientes do córtex visual enquanto que as camadas mais profundas recebem projeções de várias outras áreas do córtex ligadas a outros sentidos. O *superior culliculus* possui um mapeamento visual além de responder a estímulos auditivos e somatossensórios.

Células das camadas mais profundas do *superior culliculus* respondem positivamente antes dos movimentos sacádicos dos olhos, no qual os olhos trocam rapidamente de um ponto de fixação para outro numa cena. Estas células formam um mapa de movimento sacádico ordenado com o mapa visual. Para controlar os

movimentos sacádicos, o *superior culliculus* recebe informações não só da retina, mas também do córtex cerebral.

O *lateral geniculate nucleus* (LGN) é o ponto de retransmissão das informações visuais provenientes da retina para o córtex visual. Cerca de 90% dos axônios da retina chegam até o LGN, que possui uma representação retinotópica da metade contralateral (lado oposto) do campo visual.

A razão entre uma área do LGN e uma área correspondente da retina que representa um grau do campo visual é chamada de fator de magnificação daquela área do LGN. A fóvea possui uma representação relativamente maior que a retina periférica no LGN, ou seja, as regiões do LGN que monitoram a fóvea possuem um maior fator de magnificação.

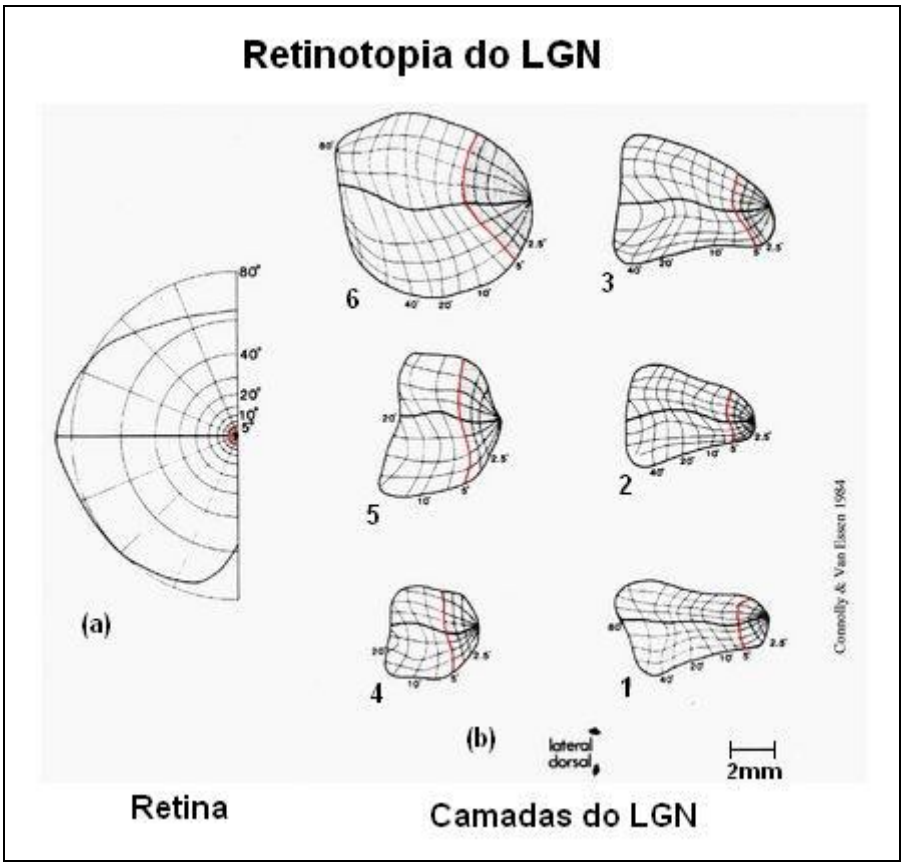


Figura 2-9 – Retinotopia do LGN. (a) Mapeamento da retina. (b) Mapeamento da retina nas camadas 1 a 6 do LGN. Figura retirada e adaptada de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/

Nos primatas, incluindo os humanos, o LGN é formado por seis camadas numeradas de 1 (ventral) a 6 (dorsal). As duas camadas mais ventrais (camadas 1 e 2) contêm células relativamente grandes que recebem conexões das células

ganglionares M da retina e são conhecidas como camadas magnocelulares enquanto que as outras 4 camadas dorsais (camadas 3, 4, 5 e 6) contêm células que recebem conexões das células ganglionares P da retina e são conhecidas como camadas parvocelulares. A Figura 2-9 mostra de forma esquemática o mapeamento da retina nas diversas camadas do LGN. Nela é possível ver como o fator de magnificação varia da fóvea para a periferia no LGN.

Todas as camadas do LGN possuem células com campo receptivo *on center* e *off center,* sendo que cada camada recebe sinais somente de um olho. As fibras da hemirretina nasal contralateral (outro lado) são projetadas nas camadas 1, 4 e 6, enquanto que as fibras da hemirretina temporal ipsilateral (mesmo lado) são projetadas nas camadas 2, 3 e 5 conforme mostrado na Figura 2-10.

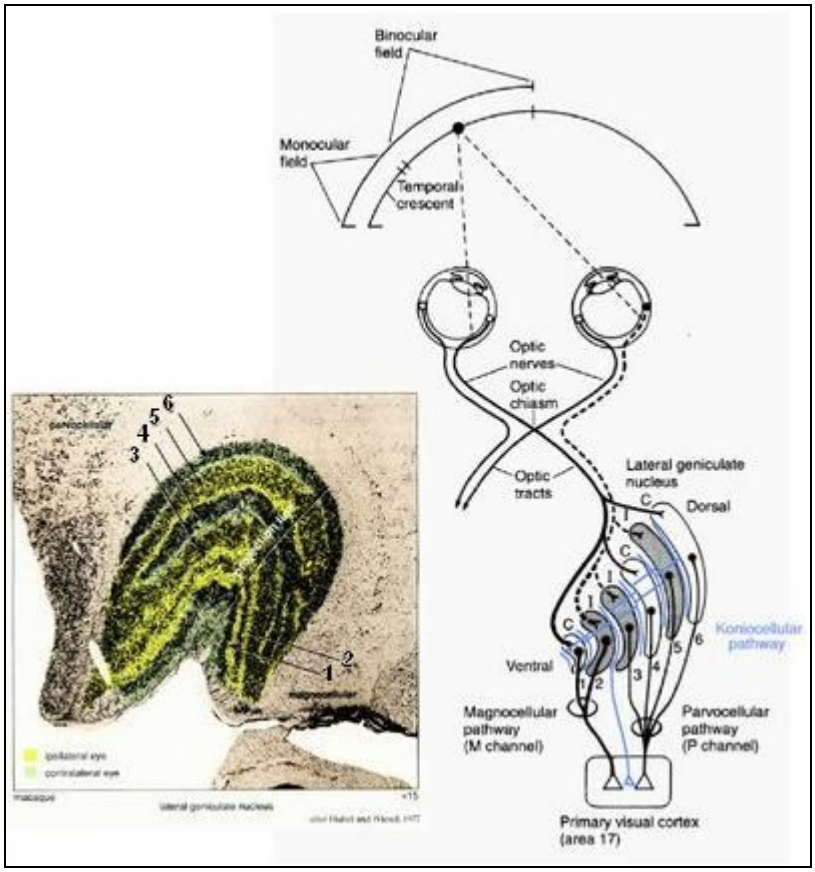


Figura 2-10 – LGN. Figura retirada de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/

As células das camadas magnocelulares e parvocelulares do LGN projetam-se para o córtex visual formando duas vias independentes (vias M e P) que se estendem desde a retina até o córtex visual primário. A via P é essencial para a

visão de cores e sensível a estímulos de alta freqüência espacial e baixa freqüência temporal da imagem na retina. A via M é mais sensível a estímulos de baixa freqüência espacial e alta freqüência temporal.

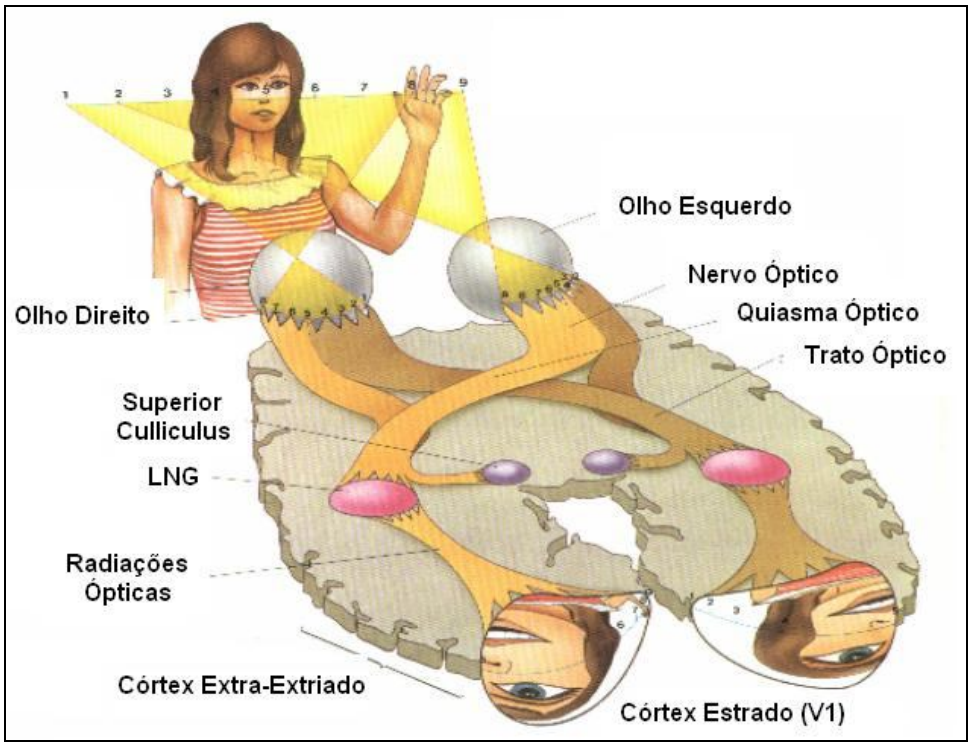


Figura 2-11 – Exemplo ilustrando o fluxo de informações visuais da retina ao córtex estriado. Figura retirada de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/

2.1.3. Organização do córtex visual

A informação visual é processada em diversas áreas corticais, sendo que cada uma delas contribui diferencialmente para o processamento da percepção de movimento, profundidade, forma e cor. Aqui serão descritas brevemente cinco áreas corticais visuais mais diretamente ligadas a este trabalho: V1, V2, V3, V4 e MT (também conhecida como V5). Na Figura 2-12-A é apresentada uma vista lateral de um hemisfério do cérebro de um macaco e na Figura 2-12-B é mostrado este hemisfério estendido de modo a formar um plano, onde são indicadas com uma tonalidade mais escura as áreas corticais visuais e são apontadas as áreas V1, V2, V3, V4 e MT. Como é possível observar na Figura 2-12-A, as áreas corticais visuais ocupam aproximadamente metade do córtex de um macaco.

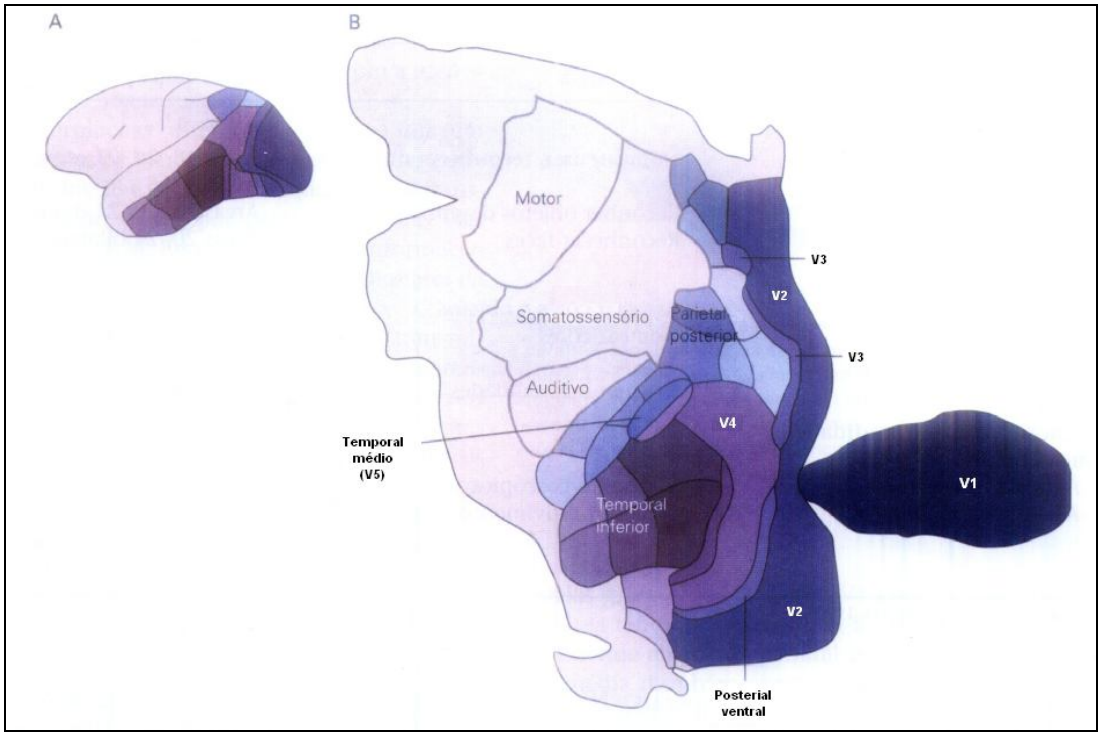


Figura 2-12 – Áreas corticais. Figura retirada de [KAN00]

V1

Quase toda informação visual vinda da retina entra no córtex via a área V1 que, devido a sua aparência estriada, também é conhecida como córtex estriado. As outras áreas são conhecidas como córtex extra-estriado. V1 também é chamada de córtex visual primário e de área 17 de Brodmann [KAN00]. Nos humanos o córtex visual primário possui cerca de 2 mm de espessura e é dividido em 6 camadas numeradas de 1 a 6 (Figura 2-13). A camada 4 é a que recebe a maioria das projeções dos axônios do LGN e pode ser dividida em 4 subcamadas: 4A, 4B, 4Cα e 4Cβ. Os axônios de células das camadas parvocelulares (P) do LGN terminam principalmente na camada 4Cβ com algumas poucas projeções para 4A e 1, enquanto que os axônios de células das camadas magnocelulares (M) terminam principalmente na camada 4Cα.

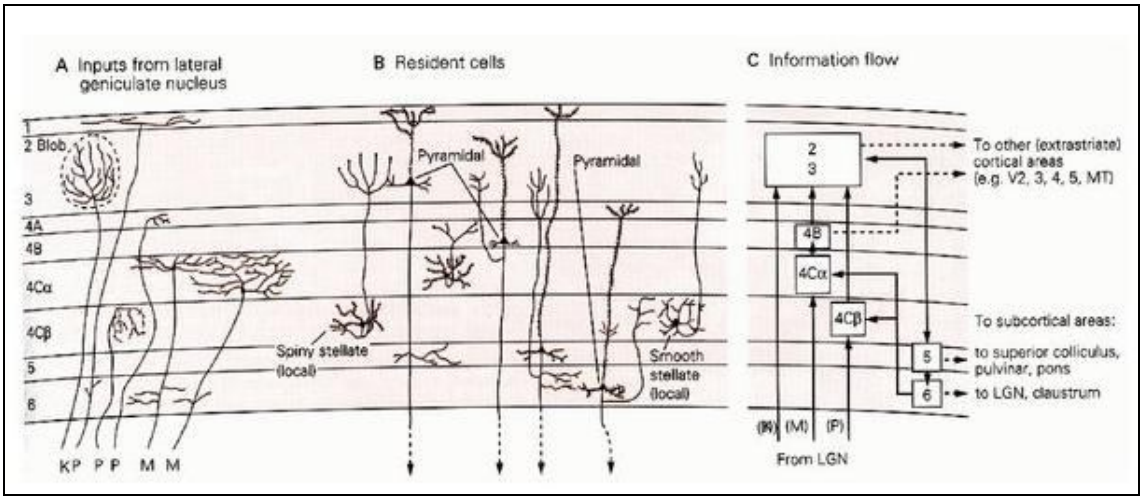


Figura 2-13 – Organização de V1. A) Os axônios dos neurônios P e M do LNG terminam na camada 4; B) Células de V1; C) Concepção do fluxo de informação em V1. Figura retirada de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/

Através de estudos feitos em macacos verificou-se que o córtex estriado, assim como LGN, possui um mapa retinotópico, isto é, áreas do campo visual vizinhas da retina são também vizinhas em V1, do campo visual contralateral [TOT82]. O aspecto mais importante deste mapa é que cerca da metade das projeções da retina sobre o córtex visual primário são provenientes da fóvea e regiões circunvizinhas. Esta área apresenta a maior acuidade visual.

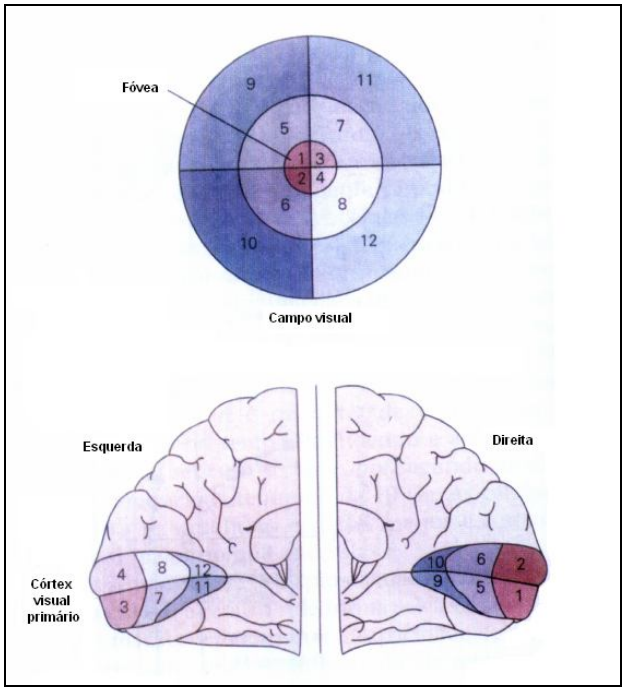


Figura 2-14 – Campo visual representado no córtex visual primário humano. Figura retirada de [KAN00]

O formato do campo receptivo das células de V1 é diferente do formato dos campos receptivos das células da retina e do LGN que são circulares. Em V1, os campos receptivos das células são alongados e, conseqüentemente, respondem melhor a estímulos alongados do que a estímulos pontuais. Hubel e Wiesel [HUB62] classificaram as células de V1 de acordo com a complexidade de sua resposta, dividindo-as em dois grupos chamados simples e complexos.

As células simples também possuem campo receptivo com regiões excitatórias e inibitórias, contudo estas regiões têm seu formato alongado. Estas células respondem melhor a estímulos na forma de barras com uma orientação específica. Uma célula simples que responde melhor a um estímulo vertical não responderá bem a um estímulo horizontal ou oblíquo, e vice-versa.

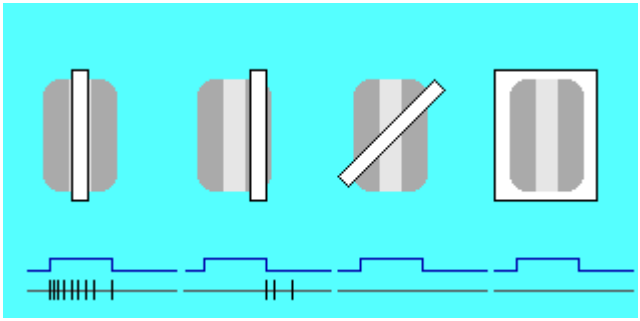


Figura 2-15 – Resposta de uma célula simples em função da projeção de um estímulo em forma de barra. Figura retirada de [MATa]

Na Figura 2-15 é apresentado de forma esquemática o comportamento de uma célula simples quando um estímulo em forma de barra é projetado em seu campo receptivo. Para produzir os resultados mostrados na Figura 2-15 o pesquisador, tipicamente, monitora a tensão no interior da célula através de um microeletrodo (na verdade, uma micropipeta que perfura a parede celular), ao mesmo tempo em que o animal cuja célula está sendo monitorada observa um estímulo. Na Figura 2-15, quatro estímulos na forma de barra são mostrados posicionados sobre uma representação do campo receptivo da célula; os três primeiros, da esquerda para a direita, são barras brancas estreitas e de mesma largura, e o quarto é uma barra branca larga que cobre todo o campo receptivo. O campo receptivo em questão possui orientação vertical, sendo que a parte do mesmo que excita a célula é central e as que inibem ficam nas laterais esquerda e

direita. Abaixo de cada conjunto estímulo-campo receptivo são mostrados dois gráficos. O primeiro, imediatamente abaixo de cada conjunto estímulo-campo receptivo, mostra o momento em que o estímulo está desligado ou ligado (trata-se do comportamento de um sinal elétrico ao longo do tempo que, no nível baixo, indica que o estímulo está inativo e, no nível alto, indica que o estímulo está ativo). O segundo representa o sinal capturado pelo microeletrodo conectado à célula.

Como o primeiro par de gráficos (Figura 2-15) (o mais a esquerda) mostra, a célula simples responde fortemente (emitindo vários pulsos pouco afastados no tempo, que é o modo como as células do córtex sinalizam sua ativação) quando o estímulo é ligado estando corretamente orientado e posicionado sobre a parte central do campo receptivo. A resposta da célula é mais vigorosa imediatamente após o acionamento do estimulo, o que mostra um aspecto temporal da resposta da célula. Na verdade, permanecendo o estímulo por muito tempo (de dezenas de segundos a alguns minutos), a resposta da célula desapareceria totalmente, por um processo conhecido como acomodação. Mas um estímulo constante por muito tempo não ocorre naturalmente, uma vez que movemos os olhos continuamente.

Como o segundo par de gráficos da Figura 2-15 mostra, quando o estímulo é posicionado sobre a parte do campo receptivo que inibe a célula ela não responde no momento em que o estímulo é ligado, embora responda, fracamente, imediatamente após o estímulo ser desligado (também uma evidência do aspecto temporal da resposta da célula). Nos outros casos a célula não responde.

As células complexas são mais numerosas em V1 do que as células simples e, assim como as células simples, respondem bem apenas para um estímulo com uma orientação específica. Porém, diferentemente das células simples, a resposta das células complexas não é seletiva à posição espacial do estímulo, ou seja, não varia com a posição do estímulo dentro do seu campo receptivo. Muitas células complexas são sensíveis ao sentido e direção do movimento do estímulo dentro do seu campo receptivo, respondendo somente quando este estímulo se move numa determinada direção e sentido.

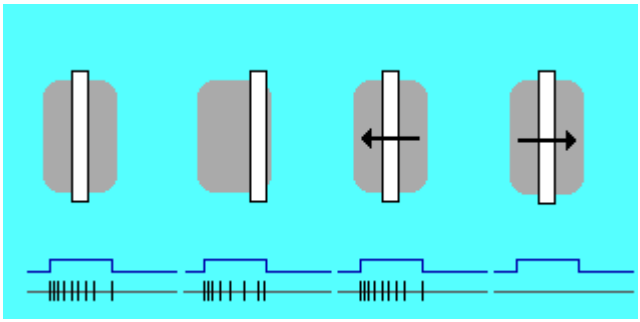


Figura 2-16 – Resposta de uma célula complexa em função da projeção de um estímulo em forma de barra. Figura retirada de [MATa]

Na Figura 2-16 é apresentado o comportamento de uma célula complexa quando um estímulo em forma de barra é projetado no seu campo receptivo. A célula complexa responde independente da posição do estímulo no campo receptivo, diferentemente da célula simples. As células complexas também são sensíveis à movimentação do estímulo dentro de seu campo receptivo. Caso o estímulo se mova no mesmo sentido em que o campo receptivo esteja sintonizado, a célula continua respondendo, caso contrário para de responder ao estímulo.

Hubel e Wiesel foram os primeiros a descobrir que as células de V1 são arranjadas e organizadas de uma forma precisa em relação à sensibilidade à orientação. Ao longo da superfície de V1, a sensibilidade à orientação varia gradualmente, mas permanece constante ao longo dos 2 mm de córtex (de uma coluna do córtex). Hubel e Wiesel também descobriram que a resposta das células varia de acordo com o olho estimulado. Muitas células de V1 respondem de forma aproximadamente equivalente a estímulos provenientes de ambos os olhos, mas a maioria das células de V1 respondem preferencialmente a estímulos provenientes de um determinado olho. Esta característica, chamada de dominância ocular, é organizada no córtex visual primário de uma forma que não varia verticalmente (em colunas), mas alterna ao longo da superfície de V1.

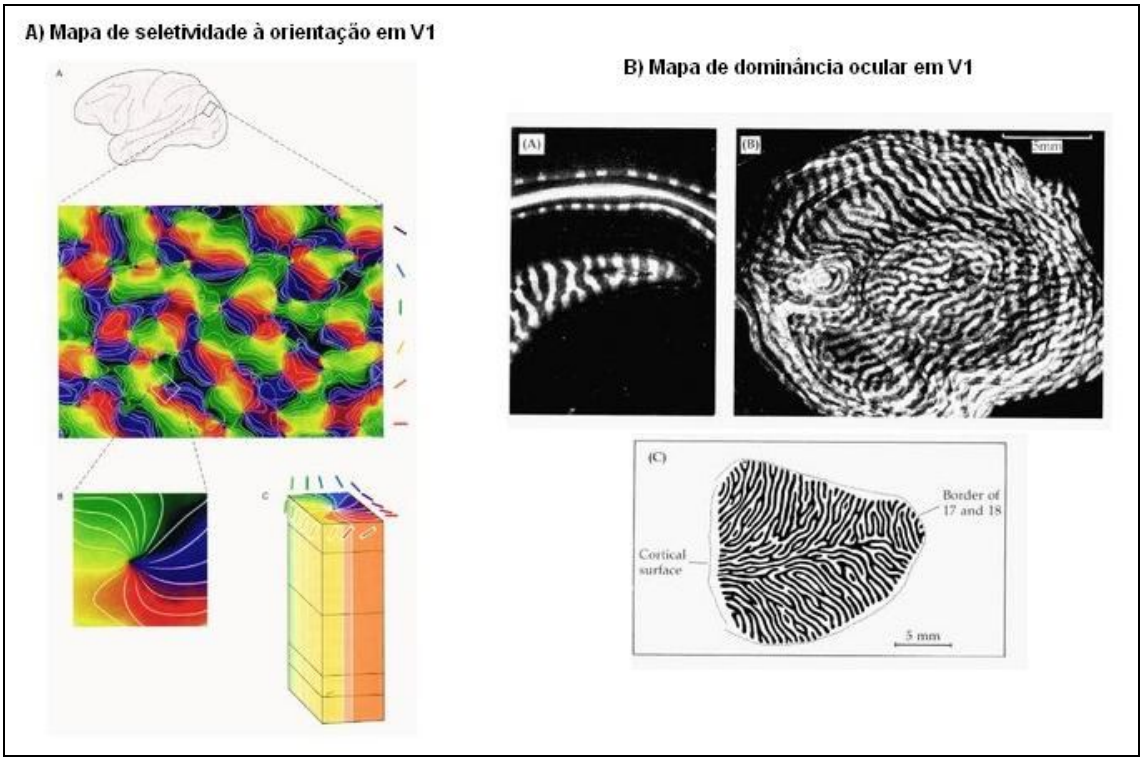


Figura 2-17 – A seletividade à orientação (a) e a dominância ocular (b) variam ao longo da superfície de V1, porém não se alteram numa mesma coluna. Figuras retiradas de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/

Um grupo de colunas que respondem a linhas (estímulos) com todas as orientações numa região particular do campo visual foi denominado de hipercoluna por Hubel e Wiesel. Essas hipercolunas aparecem repetidas regularmente e precisamente sobre a superfície de V1, ocupando cada uma cerca de 1mm^2. Essa organização sugere uma modularização do córtex cerebral, na qual cada módulo de uma área do córtex processaria todas as variantes locais da informação visual tratada naquela área cortical. Assim, se uma determinada área processa orientações do estímulo visual, uma hipercoluna dela codifica todas as orientações da região do campo visual monitorada pela hipercoluna. O mesmo ocorrendo para áreas corticais responsáveis por processar profundidade, movimento, etc.

V2

A área V2 possui uma extensa fronteira com a área V1. Esta área apresenta regiões chamadas de faixas grossas e faixas finas separadas por regiões chamadas de interfaixas (vide Figura 2-18). As faixas finas e interfaixas recebem projeções da via parvocelular que vêm das camadas 2 e 3 da área V1 enquanto que as faixas grossas recebem projeções da via magnocelular que vêm das camadas 4B, 4Cα e

4Cβ. As faixas finas e as interfaixas se projetam para a área V4, enquanto que as faixas grossas se projetam para área temporal média (MT ou V5). Estes caminhos não são totalmente separados conforme descrito, pois existem conexões entre as faixas finas e grossas e também projeções da área V4 de volta para as faixas finas de V2. Também existem conexões das faixas grossas com a área V3.

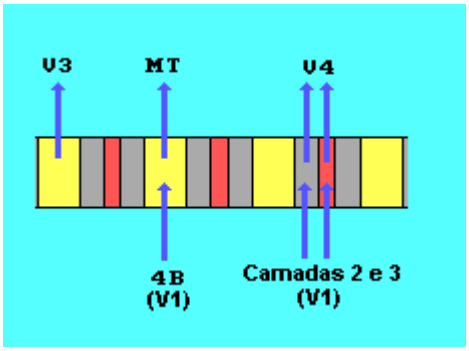


Figura 2-18 – Esquemático de V2. Figura retirada de [MATa]

As células de V2, assim como as células de V1, são sensíveis à orientação, cor e à profundidade dos estímulos, ou seja, estas células continuam a análise iniciada em V1. A resposta das células de V2 para contornos reais e ilusórios foram testados juntamente com as células de V1 em alguns experimentos. Um exemplo de percepção de contornos ilusórios pode ser visto na Figura 2-19. No desenho da esquerda, existe realmente um quadrado desenhado. No desenho do centro, apesar de não existir um quadrado desenhado, é possível facilmente “enxergar” um contorno ilusório de um quadrado, enquanto que no desenho da direita, apesar dos objetos serem os mesmos que no desenho do centro, não é possível “enxergar” o mesmo quadrado.

Muitas células de V2 responderam aos contornos ilusórios exatamente como responderam às bordas, enquanto que poucas células de V1 responderam aos mesmos contornos ilusórios da mesma forma como responderam às bordas [HEY84]. Essas observações sugerem que na área V2 é feito um processamento de contornos num nível acima do processamento que ocorre em V1, constituindo assim uma evidência da análise progressiva que ocorre no córtex visual.

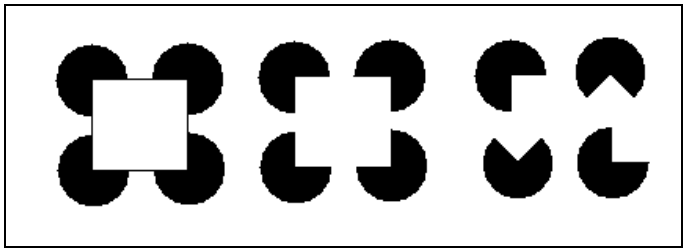


Figura 2-19 – Contorno ilusório. É possível "visualizar" um quadrado branco na figura do meio, apesar de não existir um quadrado desenhado explicitamente.

V3

Pouco se sabe sobre as propriedades funcionais dos neurônios da área extra-estriada V3. Esta área recebe informações das faixas grossas da área V2 e da camada 4B da área V1, e faz projeções tanto para a área temporal média (MT ou V5) quanto para a área V4. Grande parte das células de V3 são seletivas em relação à orientação e direção do estímulo visual, sendo que algumas células são seletivas a cores, o que sugere que em V3 ocorre uma interação entre o processamento de cor e movimento [GEG97].

V4

A área extraestriada V4 foi estudada em profundidade inicialmente por Semir Zeki [ZEK73]. Esta área recebe projeções principalmente das faixas finas e interfaixas de V2, provenientes do caminho parvocelular que vêm do LGN e retina, mas também recebe projeções das áreas V1 e V3. Inicialmente pensava-se que as células de V4 fossem exclusivamente dedicadas ao processamento de cores, porém estudos posteriores mostraram que as células de V4 são sensíveis a combinações de cores e formas. As células de V4 se projetam principalmente para o córtex temporal inferior, onde é feito o reconhecimento de faces e de outras formas complexas.

V5 ou MT

A área V5, também conhecida como área temporal média ou MT, recebe projeções da camada 4B do córtex visual primário (V1) e das faixas grossas de V2. Estas conexões são provenientes do caminho magnocelular que parte das células M da retina, passa pelas camadas magnocelulares do LGN e chega ao córtex MT. Assim como a área V1, MT possui um mapa retinotópico do campo visual

contralateral, porém os campos receptivos das células de MT são bem maiores que os campos receptivos das células de V1.

O processamento de movimento começa de forma rudimentar em V1 atingindo formas bem mais abstratas em MT numa abstração sucessiva, ou seja, em etapas. Anthony Movshon *et al.* [MOV85] testou a hipótese de que o movimento é processado em 2 etapas, registrando a resposta de células em V1 e MT para um padrão de linhas cruzadas em forma de xadrez em movimento. As células de V1 responderam ao movimento dos elementos isolados do padrão, que são as linhas, enquanto que as células em MT responderam ao movimento do padrão em forma de xadrez por completo.

A percepção de profundidade também é processada em MT. Embora células sensíveis à disparidade binocular sejam encontradas em várias áreas corticais, como V1, V2 e V3, as células de MT respondem melhor a estímulos em distâncias específicas do plano de fixação (mais próximos ou mais distantes do ponto de fixação). Esse processamento da disparidade binocular pode ser utilizado tanto para a percepção de profundidade quanto para o controle do movimento de vergência dos olhos.

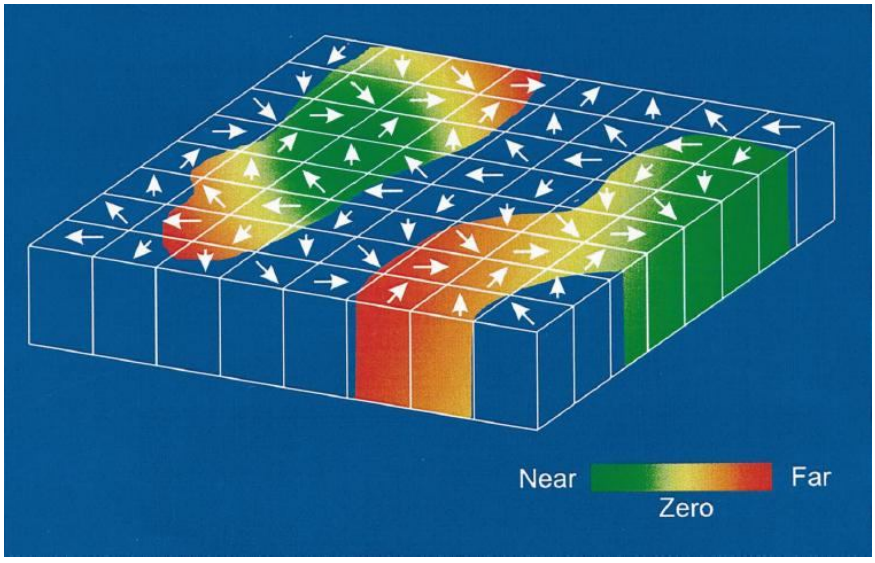


Figura 2-20 – Modelo esquemático da arquitetura funcional de MT. Figura retirada de [DEA99].

A Figura 2-20 apresenta um modelo esquemático da arquitetura funcional de MT que mostra que esta área processa tanto informações de movimento (direção) quanto de profundidade. Na figura, as setas indicam a direção preferencial dos neurônios numa coluna. Estas direções preferenciais variam suavemente através da

superfície de MT. A percepção de disparidade é representada pela faixa colorida que varia de *near* (estimulo afastado do ponto de fixação, mas perto do observador), em verde, até *far* (estimulo afastado do ponto de fixação, mas longe do observador), em vermelho, passando pela disparidade zero (estímulo à mesma distância do observador do que o ponto de fixação), codificada em amarelo. As regiões do modelo que estão em azul são regiões da área MT que aparentemente têm pouca seletividade para disparidade.

2.1.4. Vias paralelas

As informações visuais vindas da retina são conduzidas através de vias paralelas que se iniciam na retina, passam pelo LGN, chegam em V1, e depois continuam até os córtices parietal posterior (via dorsal) e temporal inferior (via ventral), como mostra a Figura 2-21. As células P da retina se projetam para as camadas parvocelulares (camadas 3, 4, 5 e 6) do LGN e seguem para o córtex visual primário, recebendo o nome de via P ou via parvocelular. A partir de V1, esta via se projeta para as faixas finas e interfaixas de V2, que depois seguem para a área V4, formando assim a via ventral que alcança o córtex temporal inferior (Figura 2-21). Os neurônios que fazem parte da via ventral são mais sensíveis em relação ao contorno das imagens, sua orientação e bordas. Outro aspecto importante que é processado nesta via é a percepção de cores. Estas células possuem alta resolução espacial, baixa resolução temporal e alta sensibilidade a cores e bordas, o que proporciona a este sistema a capacidade de analisar “o quê” é visto. Lesões no lobo temporal inferior causam deficiências relacionadas ao reconhecimento de objetos complexos, inclusive o reconhecimento de faces.

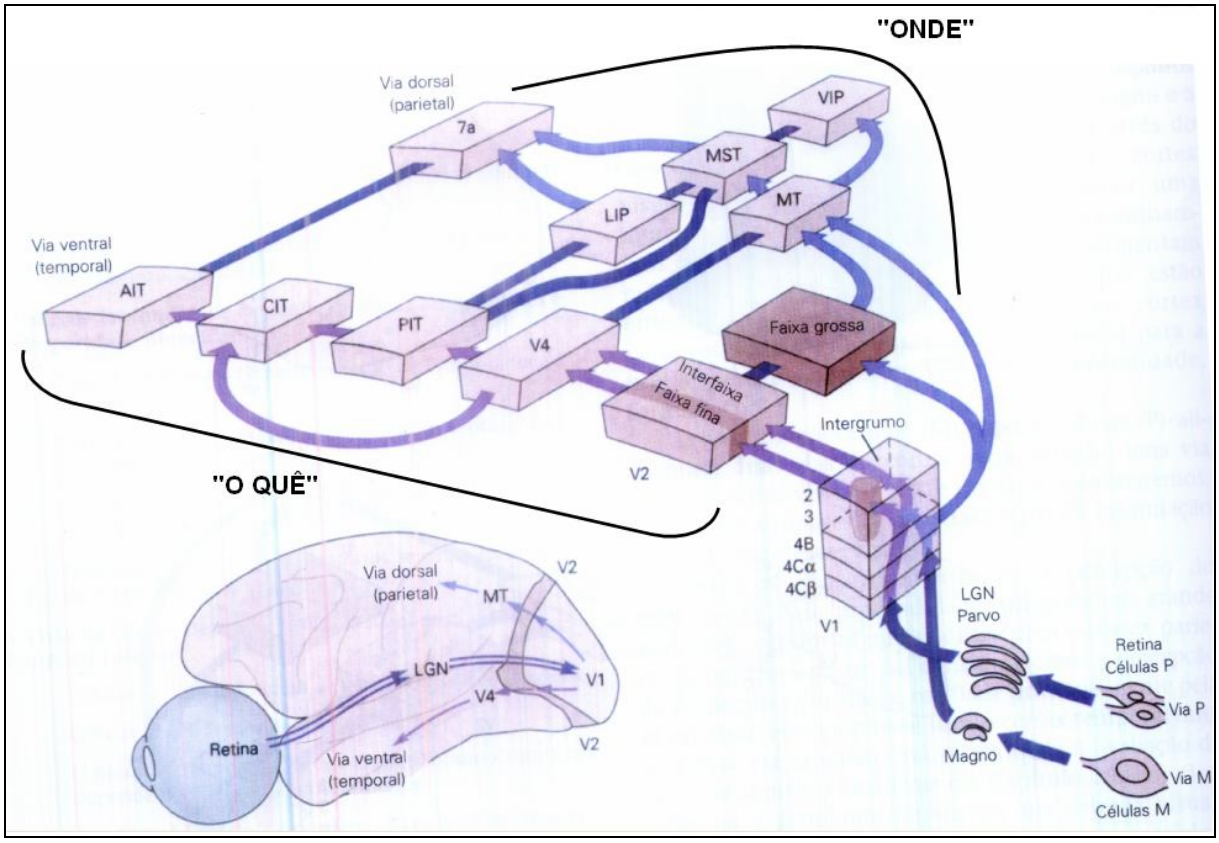


Figura 2-21 – As vias paralelas M e P projetam-se para o córtex visual passando pelo LGN. Figura retirada de [KAN00] e alterada com a inserção dos indicadores "ONDE" e "O QUÊ".

As células M da retina se projetam para as camadas magnocelulares (camadas 1 e 2) do LGN e também seguem para o córtex visual primário, recebendo o nome de via M ou via magnocelular. A via M se estende de V1 até as faixas grossas de V2, que depois se projetam para a área temporal média (MT) formando a via dorsal que se estende até o córtex parietal posterior (Figura 2-21). Conforme visto anteriormente, o MT (também chamado de V5) está relacionado ao processamento do movimento e profundidade. Os neurônios que formam este sistema são poucos sensíveis a cor e objetos parados, diferentemente dos neurônios associados à via ventral, mas possuem alta resolução temporal e sensibilidade a disparidade binocular, o que faz com que este sistema tenha capacidade de analisar "onde" estão os objetos vistos. Lesões na via dorsal causam deficiência na percepção de movimentos e nos movimentos dos olhos dirigidos a alvos em movimento (movimento de perseguição suave).

A via dorsal, responsável pela análise de "o quê" é visualizado, continua até terminar numa região do córtex pré-frontal especializada na memória de trabalho visual espacial enquanto que a via ventral, responsável pela análise de "onde" estão

os objetos visualizados, continua até terminar numa outra região também do córtex pré-frontal especializada na memória de trabalho de cognição. Esta análise mostra que o sistema visual está organizado em vias paralelas bem definidas, com uma organização seqüencial e hierárquica em cada uma delas.

2.1.5. Sistema óculo-motor

O ângulo total de visão humana possui arco de cerca de 200 graus. A melhor definição fica na fóvea, que tem pouco menos de 1 mm de diâmetro e representa cerca de 2 graus de arco no centro do campo de visão (aproximadamente o tamanho da unha do polegar à distância do braço estendido). Cerca de metade de V1 é devotada inteiramente às fóveas e esta concentração de recursos neurais, que vem da retina e persiste nas outras áreas corticais visuais além de V1, resulta em uma percepção visual muito melhor das imagens que estão sobre as fóveas. Por essa razão, a visão é um sistema bastante elaborado para a movimentação rápida e precisa dos olhos.

Os movimentos dos olhos se dão em 3 eixos de rotação (Figura 2-22): vertical (eixo X, movimento para baixo e para cima - *depression* e *elevation*), horizontal (eixo Y, movimento de lado para outro, *abduction* e *adduction*) e torsional (eixo Z, *extorsions* e *intorsions*).

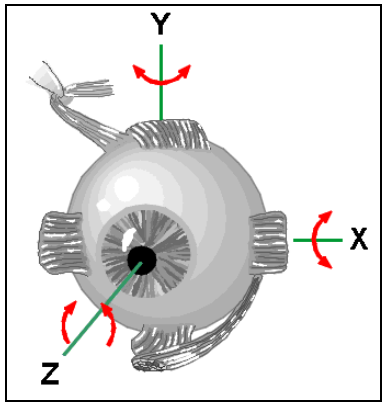


Figura 2-22 – Movimentos oculares. Figura retirada de http://www.auto.ucl.ac.be/EYELAB/Welcome.html com alterações para inclusão dos eixos X, Y e Z.

Cada um dos olhos possui 3 pares de músculos extra-oculares, que operam antagonicamente (Figura 2-23): *Medial Rectus* (*adduction*) e *Lateral Rectus*

(*abduction*); *Superior Rectus* (*elevation*) e *Inferior Rectus* (*depression*); *Superior Oblique* (*extorsion*) e *Inferior Oblique* (*introsion*) [KAN00].

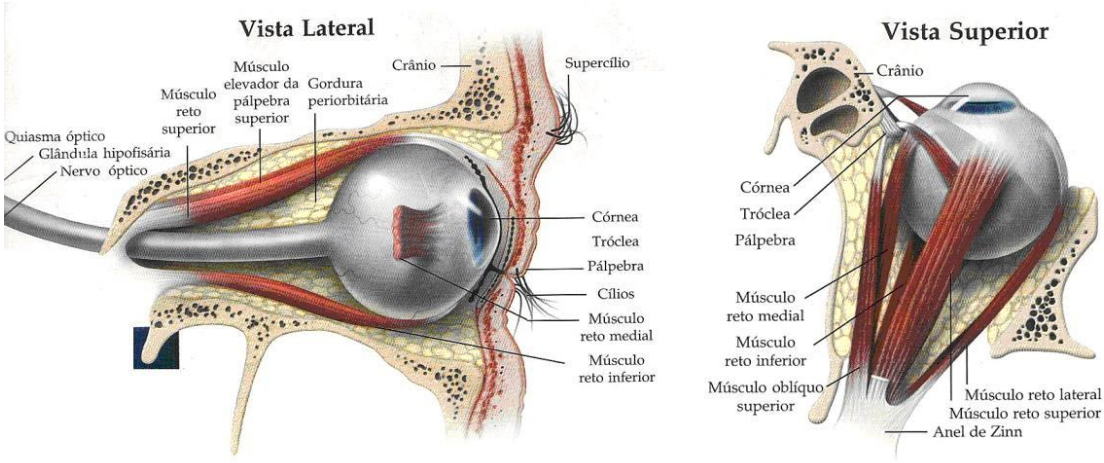


Figura 2-23 – Músculos oculares. A) Vista Lateral; B) Vista Superior. Figuras retiradas de http://www20.brinkster.com/tonho/olho/olhohumano.html

Olhar de forma exploratória em busca de um ponto de interesse requer mover os olhos rapidamente de modo que a imagem dos objetos seja projetada sobre nossas fóveas. Uma vez localizado o ponto de interesse, contudo, precisamos estabilizar sua imagem na retina, mesmo que a cabeça se movimente. O sistema óculo-motor tem, então, duas grandes funções:

1. Posicionar a imagem do ponto de interesse – o alvo – na parte da retina com maior acuidade, a fóvea;
2. Manter a imagem estacionária na fóvea, independente de movimentos do alvo ou da cabeça.

Por volta de 1902, Raymond Dodge descreveu 5 sistemas separados de controle da posição dos olhos [DOD03]. Estes 5 sistemas podem ser divididos em dois grupos, segundo as duas grandes funções descritas do sistema óculo-motor, como mostrado na Tabela 2-1.

Os primeiros 4 movimentos são conjugados, cada olho se move na mesma direção e na mesma quantidade. O último é desconjugado: os olhos se movem em direções diferentes e, muitas vezes, de diferentes quantidades.

Movimento	Função
Movimentos que estabilizam o olho quando a cabeça se move	
Vestíbulo-ocular	Mantém as imagens estáveis na retina durante rápidas rotações da cabeça
Optokinético	Mantém as imagens estáveis na retina durante rotação lenta e contínua da cabeça
Movimentos que mantém a fóvea no alvo	
Sacada	Trás novos pontos de interesse para a fóvea
Perseguição suave	Mantém a imagem de um alvo em movimento na fóvea
Vergência	Ajusta os olhos para que o mesmo ponto seja levado a ambas as fóveas

Tabela 2-1 – Movimentos Oculares.

Existem outros tipos de movimentos que têm como principal característica amplitudes muito pequenas. Estes movimentos são involuntários e ocorrem quando se está observando um objeto fixo. Estes movimentos são chamados de movimentos de *sustaining* e possuem como principal função manter o foco sobre o objeto que está sendo observado, e produzir variações constantes, mesmo que pequenas, da imagem na retina. Sem estas variações a imagem “desapareceria” devido à acomodação dos neurônios do sistema visual.

2.2. PISTAS MONOCULARES

Uma das principais funções do sistema visual é reconstruir uma representação tridimensional do mundo à nossa volta a partir das imagens bidimensionais projetadas na retina. Estudos indicam que esta reconstrução é baseada tanto em pistas estereoscópicas oriundas da disparidade binocular causada pela leve diferença das imagens projetadas nas retinas, quanto em pistas monoculares. Quando os objetos observados estão a distâncias maiores do que cerca de 30 metros, as imagens projetadas nas retinas são praticamente idênticas, eliminando quase que totalmente a disparidade binocular, fazendo com que a percepção de profundidade seja baseada principalmente nas pistas monoculares.

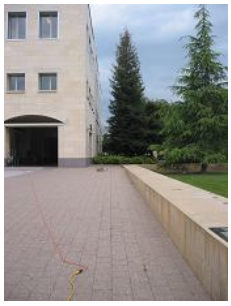
Algumas pistas monoculares se prestam mais à exploração em sequências contínuas de imagens do que em imagens avulsas. O desvio de paralaxe, por exemplo, permite inferir as distâncias relativas entre dois objetos a partir do grau de deslocamento aparente do objeto mais próximo em relação ao mais distante, quando o observador se move paralelamente aos seus planos de profundidade. De forma

similar, a alteração no grau de desfocamento de um objeto à medida que se varia a distância focal do aparato visual pode ser usada para estimar sua distância.

Outras pistas são mais adequadas à estimativa de profundidade em imagens estáticas, tais como diferenças de textura, gradientes de textura, distribuições de cores e bordas. As definições e utilidades dessas pistas são discutidas nas próximas seções.

2.2.1. Filtros e Convolução

Ao considerar as pistas monoculares utilizadas por sistemas biológicos e computacionais para estimar profundidades em uma imagem, é útil ter em mente o conceito de *filtro.* Um filtro é um construto matemático tal que, dado um conjunto de valores de entrada (por exemplo, uma matriz), produz um conjunto de saída que se distingue da entrada por ter mais realçada (ou atenuada) alguma "característica" particular. Exemplos de filtros e as "características" das imagens que eles manipulam são visíveis na Figura 2-24: filtros podem ser usados para detectar os limites entre objetos em uma cena (Detector de Bordas de Sobel), atenuar ou realçar o contraste (Enevoamento Gaussiano, Realce de Contraste), ou mesmo adicionar ruído (Dispersão).



Detector de Bordas de Sobel
⇒

Enevoamento Gaussiano
⇒

Realce de Contraste
⇒

Dispersão
⇒

Figura 2-24 – Exemplos de filtros e seus efeitos sobre uma imagem de entrada. Esquerda: imagem de entrada. Centro: filtro aplicado sobre a imagem. Direita: saída do filtro.

Computacionalmente, filtros podem ser facilmente implementados através do operador Convolução (*). Dada uma matriz I de dimensões $m \times n$ e um *kernel* (matriz) F de dimensões $k \times l$, tais que $k << m$ e $l << n$, definimos o operador convolução:

$$I * F = O \rightarrow O[x,y] = \sum_{i=1}^{k} \sum_{j=1}^{l} I[x+i-1, y+j-1] F[i,j] \quad \forall\, O[x,y] \in O$$

A Figura 2-25 apresenta um esquema simplificado da operação de convolução. Cada célula *(x, y)* de *O* é calculada como o somatório da multiplicação entre os elementos do kernel *F* e uma "submatriz" de *I,* de dimensão igual ao kernel e centrada na célula *(x, y)* de *I.*

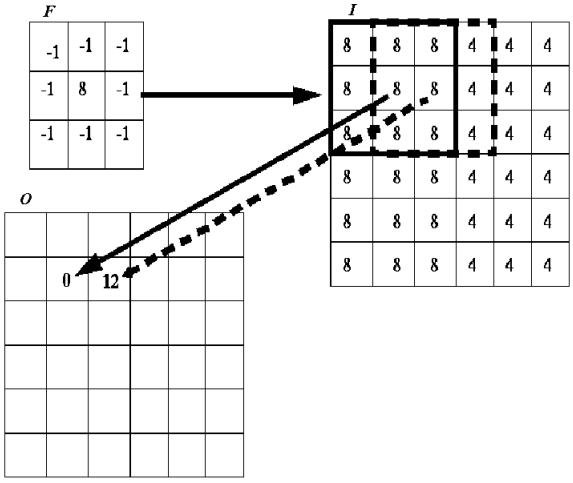


Figura 2-25 – Esquema simplificado da operação de convolução. Cada célula *(x, y)* de *O* é calculada como o somatório da multiplicação entre os elementos do kernel *F* e uma "submatriz" de *I,* de dimensão igual ao kernel e centrada na célula *(x, y)* de *I.* Figura adaptada de http://www.geo.hunter.cuny.edu/~yllik/gis2/lectures/lecture5/lecture5.html

Apesar de aparentar uma relativa simplicidade, a convolução é uma operação poderosa, sendo a principal ferramenta para implementação de filtros visuais. Por exemplo, o Detector de Bordas de Sobel $G(I)$, ilustrado na Figura 2-24 pode ser facilmente implementado como a média geométrica de duas convoluções:

$$G(I) = \sqrt{G_x^2(I) + G_y^2(I)}\text{, onde } G_x(I) = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * I \text{ e } G_y(I) = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} * I$$

2.2.2. Texturas e Gradientes de Textura

Intuitivamente, a *textura* é a característica *visual* de um objeto que sugere uma sensação *táctil* (ex: "macio", "áspero", "úmido", etc.). Por exemplo, na Figura 2-26, a textura do objeto representado (um tronco de árvore) sugere uma sensação áspera, mesmo que não nos seja possível tocá-lo. Mais formalmente, a textura é uma propriedade relacionada a flutuações periódicas de luminosidade na imagem de

uma superfície, que nos permitem interpretá-la como uma estrutura homogênea; texturas podem ser caracterizadas por características tais como regularidade, esparsidade, contraste e direcionalidade [AND02].

Figura 2-26 – A textura de uma imagem é uma informação visual, entretanto sugere uma sensação táctil. Figura extraída de http://catedral2.weblog.com.pt/flora/

A percepção visual da textura de uma superfície varia com a distância do observador: à medida que a superfície se afasta, a textura adquire traços mais finos e suaves, eventualmente tornando-se indistinguível. A esse fenômeno se dá o nome *gradiente de textura* [JOH03]. Por exemplo, na Figura 2-27, os paralelepípedos do calçamento tornam-se cada vez menores e menos definidos à medida que a rua "afasta-se" do observador, formando assim um gradiente de textura que sugere a sensação de profundidade da cena.

Figura 2-27 – Exemplo de gradiente de textura. O efeito de profundidade da cena é criado pelo calçamento da rua, cujos paralelepípedos, ao tornarem-se menores e menos definidos na medida de sua "distância" ilusória do observador, criam um gradiente de textura. Figura extraída de [JOH03]

2.2.3. Bordas e Cores

No sistema visual humano, as imagens captadas pela retina são decompostas em várias dimensões antes de serem processadas: existem circuitos diferentes para captar a luminância (brilho) e as cores das imagens, além de estruturas especializadas em identificar "bordas" (isto é, os limites entre regiões de texturas ou cores distintas). O resultado é uma codificação multidimensional, englobando brilho, cores e bordas.

Em computação, uma das maneiras de representar os pixels de uma imagem é através do espaço de cores *Hue, Saturation, Value* (HSV). Em contraste ao mais comum *Red, Green, Blue* (RGB) – no qual as cores são definidas como a adição de proporções distintas das cores básicas vermelho, verde e azul – o formato HSV descreve a cor de um pixel em termos de *Hue* (a "cor básica" do pixel, relativa ao espectro visível da luz), *Saturation* (a "intensidade" da cor do pixel relativa ao "brilho") e *Value* (o "brilho", ou a "quantidade de luz emitida" pelo pixel). Normalmente *Hue, Saturation* e *Value* são representados por valores numéricos entre 0 e 255 (a faixa de valores possível em um byte) que indicam as coordenadas em um sólido de cor, visível na Figura 2-28:

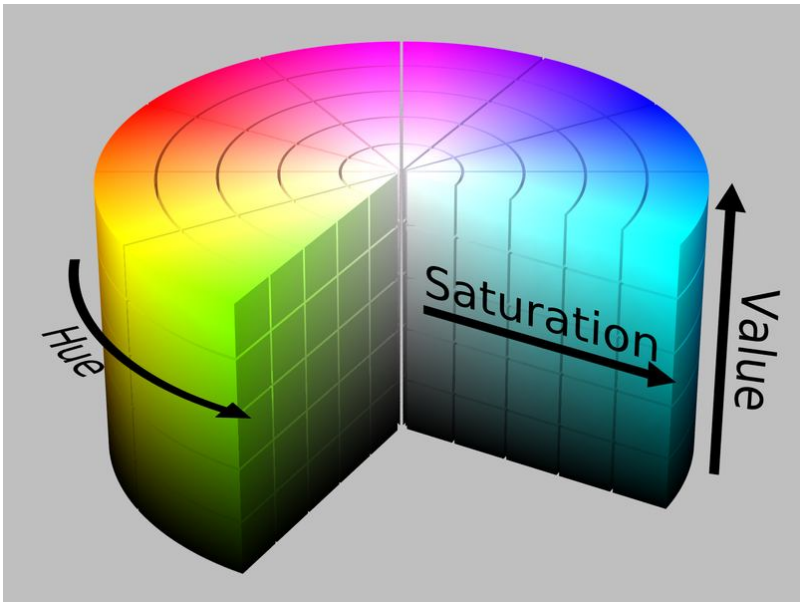


Figura 2-28 – Valores de *Hue, Saturation* e *Value* como coordenadas em um sólido de cor. Figura extraída de http://en.wikipedia.org/wiki/File:HSV_color_solid_cylinder_alpha_lowgamma.png

Ao dividir a informação visual em brilho (*Value*) e cor (*Hue, Saturation*), o formato HSV aproxima-se da abordagem biológica, tornando-o útil para a implementação de sistemas de visão artificial. Adicionalmente, as distribuições de valores normalmente encontradas no canal de *Value* fazem dele uma entrada conveniente para filtros de detecção de bordas. A Figura 2-29 apresenta uma imagem colorida, os mapas de valores para seus canais de *Hue, Saturation* e *Value,* e a saída de um detector de bordas (Filtro de Sobel) aplicado ao canal de *Value.*

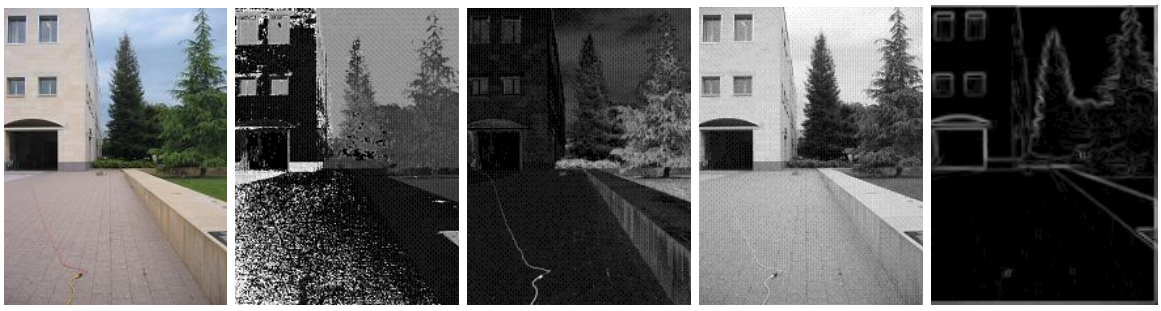


Figura 2-29 – Imagem original, canais HSV e saída de detector de bordas, aplicado ao canal de *Value.*

Bordas e cores proporcionam importantes pistas para a detecção de profundidades. É nos limites entre os objetos da cena (indicados pelas bordas) onde normalmente encontram-se variações bruscas de profundidade. Por outro lado, regiões contíguas preenchidas por uma mesma cor (ou tonalidades muito semelhantes) frequentemente pertencem a um objeto específico, e/ou sugerem uma profundidade constante; adicionalmente, o fenômeno de “névoa seca” ou dispersão cromática (*haze*), ao atenuar (“acinzentar”) as cores de objetos distantes, é uma pista útil para distinguir os limites do espaço de profundidades da imagem.

2.3. CLASSIFICADOR MRF DE SAXENA

Em [SAX08], Saxena implementa um classificador baseado em *Markov Random Fields* (MRF) para aprender o relacionamento entre as pistas monoculares contidas em uma imagem e o mapa de profundidades da cena. Sua abordagem consiste em dividir as imagens em pequenas seções retangulares, e então estimar uma profundidade média para cada seção. A Figura 2-30 ilustra esse conceito: as imagens de entrada (a) são divididas em seções (b) e em seguida diferentes valores

de profundidade (aqui representados por cores, por simplicidade) são atribuídos a cada seção (c).

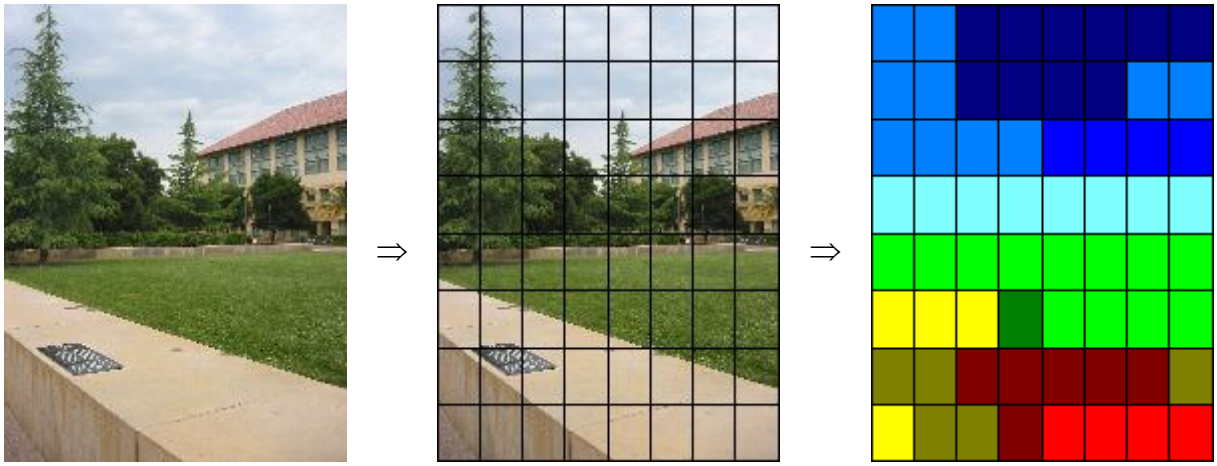
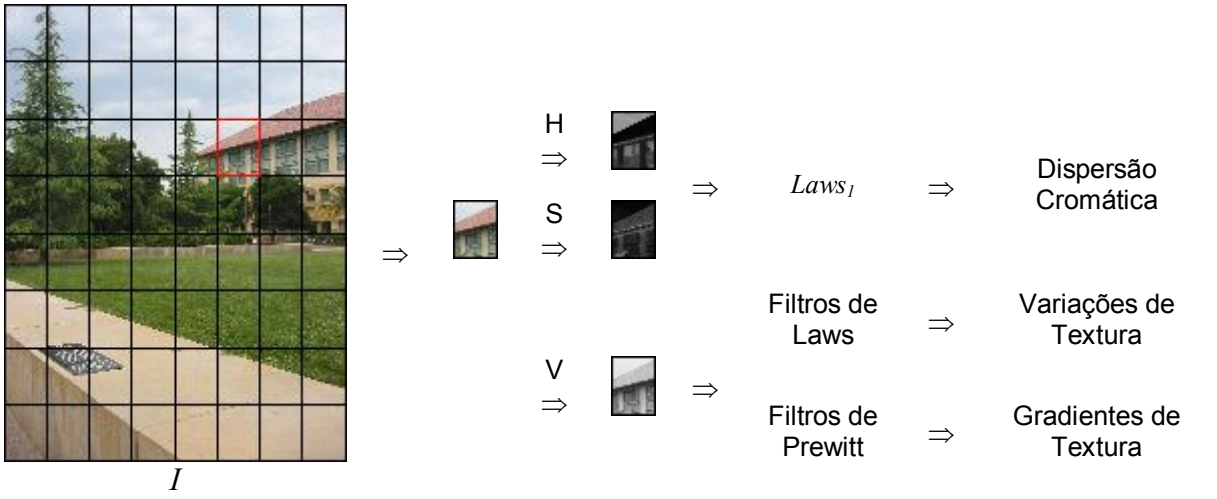


Figura 2-30 – Esquema simplificado da estratégia de estimativa de profundidades empregada por Saxena: as imagens de entrada (a) são divididas em seções (b) e em seguida diferentes valores de profundidade (aqui representados por cores, por simplicidade) são atribuídos a cada seção (c).

Para cada seção, são calculadas várias *características,* que podem ser divididas em dois tipos: *absolutas,* usadas para estimar a profundidade absoluta de uma seção; e *relativas,* usadas para estimar a magnitude da diferença de profundidade entre duas seções. As características extraídas das seções buscam capturar três tipos de informações: variações de textura, gradientes de textura, e cor.

Para capturar as variações de textura, os filtros de Laws [DAV97] são aplicados ao canal de intensidade da imagem. A dispersão cromática (enevoamento) reflete-se nas frequências mais baixas dos canais de cor, e essa informação é capturada aplicando-se um filtro de média local (o primeiro filtro de Laws) aos canais de cor. Finalmente, para calcular uma estimativa dos gradientes de textura que seja robusta a ruídos, seis filtros de bordas (Filtros de Prewitt) orientados são aplicados ao canal de intensidade.

A Figura 2-31 ilustra a decomposição e aplicação de filtros às imagens de entrada. Inicialmente, cada segmento da imagem original é decomposto nos três canais do espaço de cores HSV; em seguida, o filtro $Laws_1$ é aplicado aos canais de cores (H, S), enquanto o canal de intensidade (V) é passado como entrada para os Filtros de Laws e os Detectores de Bordas. Os filtros são representados como *kernels* de convolução.



Laws$_1$ $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$

Filtros de Laws $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 2 & -1 \\ -2 & 4 & -2 \\ -1 & 2 & -1 \end{bmatrix}\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & -1 \end{bmatrix}\begin{bmatrix} 1 & -2 & 1 \\ 0 & 0 & 0 \\ -1 & 2 & -1 \end{bmatrix}\begin{bmatrix} -1 & -2 & -1 \\ 2 & 4 & 2 \\ -1 & -2 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 & -1 \\ -2 & 0 & 2 \\ 1 & 0 & -1 \end{bmatrix}\begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix}$

Detectores de Bordas $\begin{bmatrix} -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 \end{bmatrix}\begin{bmatrix} -1 & 0 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & -1 & -1 & 0 \\ -1 & -1 & -1 & -1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 & -1 \\ 0 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 0 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 0 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & 0 & -1 & -1 & -1 \end{bmatrix}$

Figura 2-31 – Separação das características da imagem nos classificadores MRF de Saxena. Cada segmento da imagem original é primeiramente decomposto nos três canais do espaço de cores HSV; em seguida, o filtro *Laws$_1$* é aplicado aos canais de cores (H, S), enquanto o canal de intensidade (V) é passado como entrada para os Filtros de Laws e os Detectores de Bordas. Os filtros são representados como *kernels* de convolução.

2.3.1. Características absolutas

As características absolutas de uma seção retangular i da imagem I são dadas pela fórmula:

$$E_i = \begin{Bmatrix} E(Laws_1, H_i), E(Laws_1, S_i), \\ E(Laws_1, V_i), \ldots, E(Laws_9, V_i), \\ E(\mathrm{Pr}ewitt_1, V_i), \ldots, E(\mathrm{Pr}ewitt_6, V_i) \end{Bmatrix}$$

onde $Laws_1$,...,$Laws_9$ correspondem aos nove filtros de Laws, $Prewitt_1$,...,$Prewitt_6$ correspondem aos seis filtros de Prewitt, H_i, S_i e V_i correspondem aos canais H, S e V da seção i, e:

$$E(F,I)=\left(\sum_{(x,y)}|I*F|,\sum_{(x,y)}(I*F)^2\right)$$

nos dá respectivamente a soma absoluta e a soma quadrática da saída de cada filtro. O vetor de características absolutas contém a soma absoluta e quadrática da saída de 17 filtros (9 filtros de Laws, 6 filtros de Prewitt e 2 canais de cores), totalizando 34 dimensões.

Infelizmente, a informação local de uma seção é insuficiente para estimar adequadamente a sua profundidade. Para capturar propriedades mais gerais da imagem, o valor de E_i é calculado em três escalas espaciais distintas: a mesma da seção original, e dois níveis de distanciamento, x3 e x9. Além disso, para capturar outras características gerais (como relacionamentos de oclusão), as características das quatro seções vizinhas também são adicionadas ao vetor de cada seção, nas três escalas espaciais. Finalmente, muitas estruturas encontradas ao ar livre, como árvores e construções, demonstram uma orientação vertical na sua estrutura; para representar esse fato, as características da coluna da imagem onde a seção se encontra também são adicionadas ao vetor.

Para cada seção, após incluir suas próprias características e as dos seus quatro vizinhos em três escalas espaciais, mais as características das quatro seções da coluna da imagem, o vetor de características absolutas X é de dimensão $19 . 34 = 646.$ A Figura 2-32 ilustra a montagem do vetor de características absolutas de uma seção.

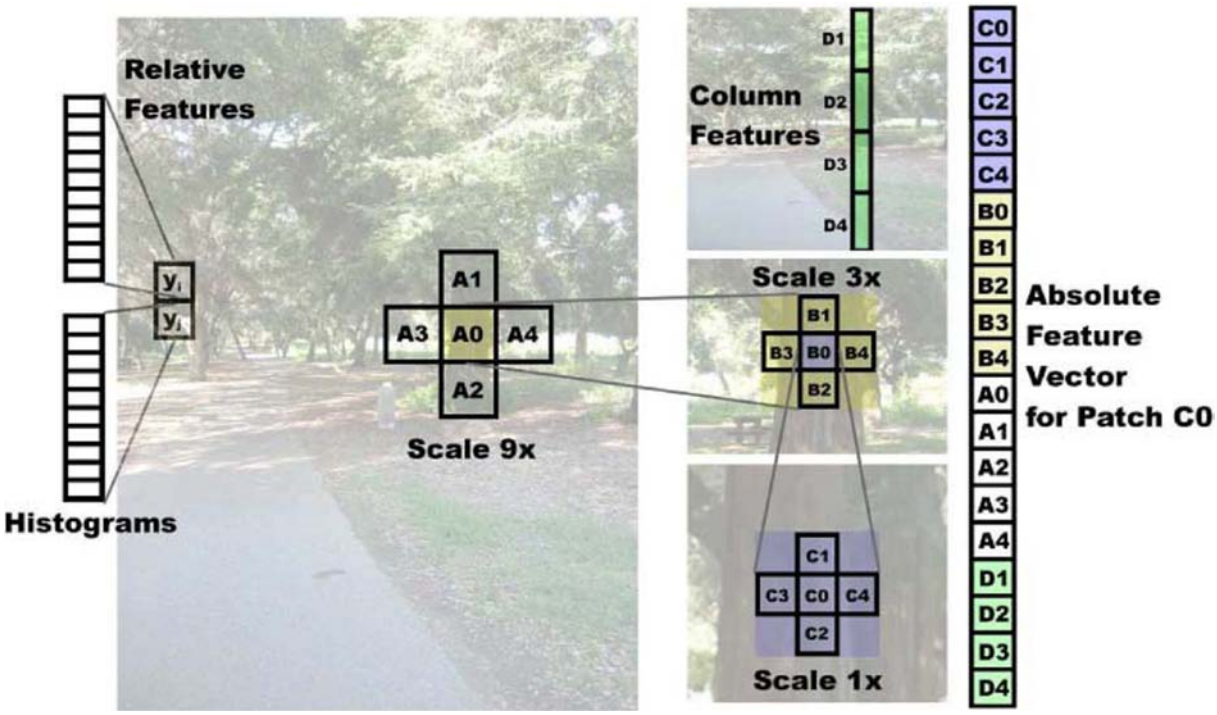


Figura 2-32 – O vetor de características absolutas de uma seção inclui as características dos seus vizinhos imediatos, e também os mais distantes (em escalas espaciais maiores). As características relativas de cada seção usam histogramas das saídas dos filtros. Figura extraída de [SAX08].

2.3.2. Características relativas

Para cada seção i da imagem $I(x, y)$ na escala espacial s, também é calculado um histograma de 10 colunas para cada saída de filtro $|F * I|$, resultando em um vetor de 170 características y_{is}:

$$y_{is} = \begin{Bmatrix} Hist_{10}(Laws_1 * H_{is}), Hist_{10}(Laws_1 * S_{is}), \\ Hist_{10}(Laws_1 * V_{is}), \ldots, Hist_{10}(Laws_9 * V_{is}), \\ Hist_{10}(\mathrm{Pr}ewitt_1 * V_{is}), \ldots, Hist_{10}(\mathrm{Pr}ewitt_1 * V_{is}) \end{Bmatrix}$$

onde $Laws_1,\ldots,Laws_9$ correspondem aos nove filtros de Laws, $Prewitt_1,\ldots,Prewitt_6$ correspondem aos seis filtros de Prewitt, H_{is}, S_{is} e V_{is} correspondem aos canais H, S e V da seção i na escala s, e:

$$Hist_{10}(X) = \{m_1(X), m_2(X), ..., m_{10}(X)\}$$

$$m_i(X) = \sum_{(x,y)} \begin{cases} 1, \, se \; Lo_i(X) \le X(x,y) < Hi_i(X) \\ 0, \, c.c. \end{cases}$$

$$Lo_i(X) = \min(X) + (i-1) \cdot \frac{\max(X) - \min(X)}{10}$$

$$Hi_i(X) = \min(X) + i \cdot \frac{\max(X) - \min(X)}{10}$$

As características relativas entre duas seções i e j na escala s, por sua vez, são calculadas como as diferenças entre seus histogramas, isto é, $y_{ijs} = y_{is} - y_{js}$. A Figura 2-32 ilustra a montagem do vetor de características relativas de uma seção.

2.3.3. Modelo probabilístico

Saxena [SAX08] emprega um modelo hierárquico multi-escalar baseado em *Markov Random Fields* para modelar o relacionamento entre as características de uma seção e as profundidades das suas vizinhas em múltiplas escalas. Seu modelo leva em consideração os seguintes relacionamentos:

1. A profundidade de uma seção depende das suas características; portanto, o relacionamento entre a profundidade e o vetor de características da seção é modelada;
2. A profundidade de uma seção também está relacionada às profundidades de seus vizinhos (seções que recaem sobre um mesmo objeto terão todas profundidades semelhantes);
3. Além das interações com suas vizinhas imediatas, também há ocasionalmente relacionamentos fortes entre uma seção e vizinhas mais distantes (por exemplo, as seções que recaem sobre a fachada de uma construção terão todas profundidades semelhantes). Na escala espacial básica, pode ser difícil reconhecer uma seção como parte de um objeto muito maior; por isso, o modelo também leva e consideração relacionamentos entre profundidades em múltiplas escalas espaciais.

Com esses relacionamentos em mente, Saxena aplica algoritmos de otimização para ajustar um conjunto de parâmetros do seu modelo estatístico, de forma a maximizar a probabilidade $P(d|X)$ (probabilidade de uma profundidade d condicionada a um vetor de características X) de cada seção das imagens de treinamento. Seus sistemas trabalham com uma de duas possíveis distribuições de probabilidades, gaussiana ou laplaciana, modeladas nas fórmulas abaixo:

$$P_G(d \mid X;\theta,\sigma) = \frac{1}{Z_G}\exp\left(-\sum_{i=1}^{M}\frac{\left(d_i(1)-x_i^T\theta_r\right)^2}{2\sigma_{1r}^2}-\sum_{s=1}^{3}\sum_{i=1}^{M}\sum_{j\in N_s(i)}\frac{\left(d_i(s)-d_j(s)\right)^2}{2\sigma_{2rs}^2}\right) \quad (1)$$

$$P_L(d \mid X;\theta,\lambda) = \frac{1}{Z_L}\exp\left(-\sum_{i=1}^{M}\frac{\left|d_i(1)-x_i^T\theta_r\right|}{2\lambda_{1r}}-\sum_{s=1}^{3}\sum_{i=1}^{M}\sum_{j\in N_s(i)}\frac{\left|d_i(s)-d_j(s)\right|}{2\lambda_{2rs}}\right) \quad (2)$$

Na distribuição Gaussiana (1), os parâmetros são θ_r e σ^2_{1r} (modelado tal que $\sigma^2_{1r} = v^T_r x_i$) para as seções na linha r, e σ^2_{2rs} (modelado tal que $\sigma^2_{2rs} = u^T_{rs}|y_{ijs}|$) para a linha r na escala s. Na distribuição Laplaciana (2), os parâmetros são θ_r e λ_{1r} (modelado tal que $\lambda_{1r} = v^T_r x_i$) para a linha r, e λ_{2rs} (modelado tal que $\lambda_{2rs} = u^T_{rs}|y_{ijs}|$) para a linha r na escala s. Nos dois casos, M é o número total de seções retangulares da imagem (na escala x1); Z é a constante de normalização do modelo; x_i é o vetor de características absolutas da seção i; θ_r, σ_{1r} e σ_{2r} são os parâmetros do modelo para as seções da linha r; $N_s(i)$ é a lista dos vizinhos da seção i na escala s; e $d_i(s)$ é o valor da profundidade da seção i na escala s (equivalente a um nível de distanciamento $x3^{(s-1)}$), sujeito à restrição:

$$d_i(s+1) = (1/5)\sum_{j\in N_s(i)\cup\{i\}} d_j(s) \quad \text{para } s < 3.$$