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

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 Gutavo 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 de posição de cada ponto entre as imagens de um par. Além dessa informação puramente geométrica, imagens contém 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 uma 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”, ou RNSP's) a problemas de visão computacional, neste trabalho objetivamos investigar a efetividade da aplicação de RNSP’s ao problema de estimar mapas de profunidades. 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 traditionally applied to this problem, *stereo correspondence,* is one of the most intensively studied topics in the field.

Stereo correspondence systems estimate depths from *binocular* features – more specifically, point positioning differences between a stereogram's two 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 *Markov Random Fields* (MRF's) machine-learning systems, Ashutosh Saxena has 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 relative to Saxena's MRF-based approach, and also 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: *estereópsis, desvio de paralaxe* e *pistas monoculares* [MIC05].

Humanos parecem ser extremamente bons 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”, ou 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, assim como as imagens e mapas de profundidades usados no treinamento das redes, e também os resultados experimentais obtidos nos testes, estão disponíveis para download na Internet, a partir do endereço 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. Já no Capítulo 3 detalhamos a estrutura e funcionamento das redes neurais sem peso, com ênfase no modelo VG-RAM, usado pela MAE; também descrevemos as arquiteturas estudadas no trabalho.

No Capítulo 4 apresentamos a metodologia empregada no trabalho, os experimentos efetuados e resultados obtidos. Começamos apresentando a base de dados disponibilizada por Saxena; suas características; 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. Em seguida, descrevemos os experimentos realizados para avaliar os parâmetros ótimos de configuração para cada arquitetura neural, além dos testes adicionais de validação. Passamos então à análise e comentário dos resultados obtidos, comparando também com os alcançados por Saxena.

No Capítulo 5 fazemos a discussão do 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. Em seguida passamos à análise crítica deste trabalho, suas contribuições e deficiências. Finalmente, no Capítulo 6 concluímos com um resumo do trabalho, seus resultados e direções para trabalhos futuros, e no Capítulo 7 relacionamos a bibliografia utilizada.

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.

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.

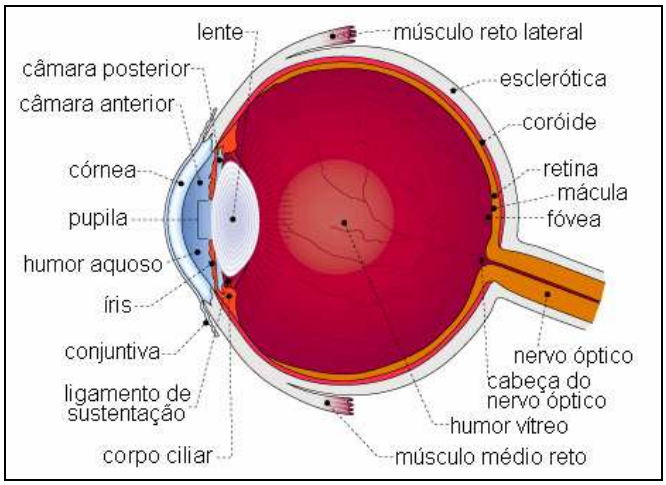


Figura 2-1 – Anatomia do olho humano. Corte medial horizontal do olho direito visto de cima. Figura retirada de http://www.escolavesper.com.br/olho_humano.htm

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 2mm a 8mm, 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 uma 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.

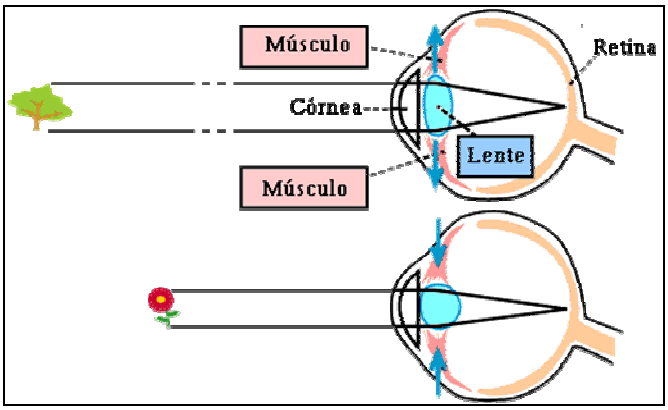


Figura 2-2 – Figura mostrando a acomodação visual do cristalino.

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.

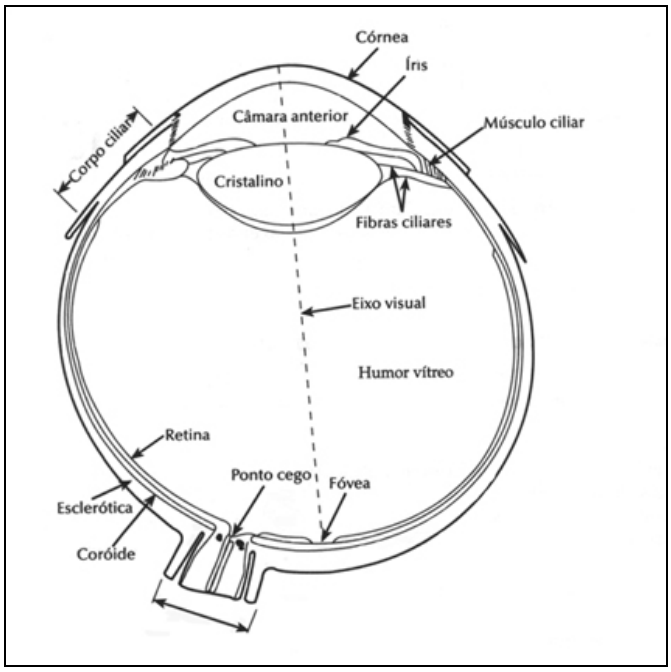


Figura 2-3 – Eixo visual. Figura retirada de http://www.on.br/glossario/alfabeto/o/olho_humano.html

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.

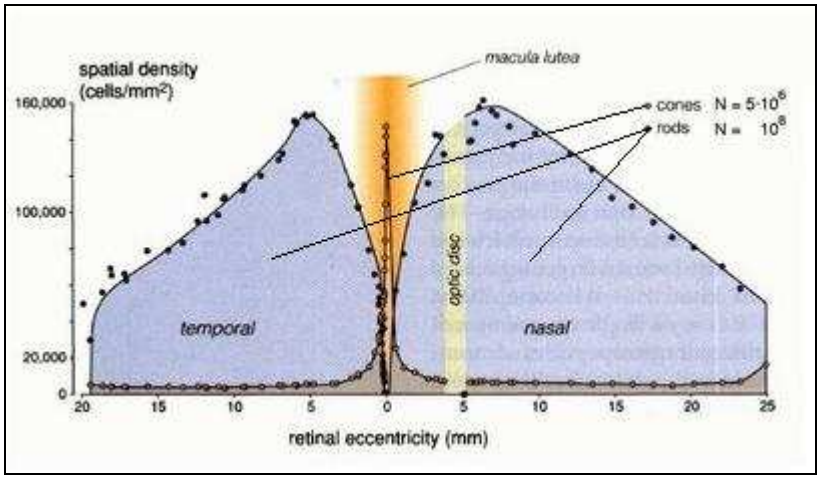


Figura 2-4 – Distribuição de cones e bastonetes (*rods*) na retina. A *macula lutea* (região da fóvea e vizinhanças) possui alta densidade de cones, enquanto que os bastonetes se concentram na periferia. O ponto cego fica na região do disco óptico (*optic disc*). Figura retirada de http://www.brainworks.uni-freiburg.de/group/wac

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.

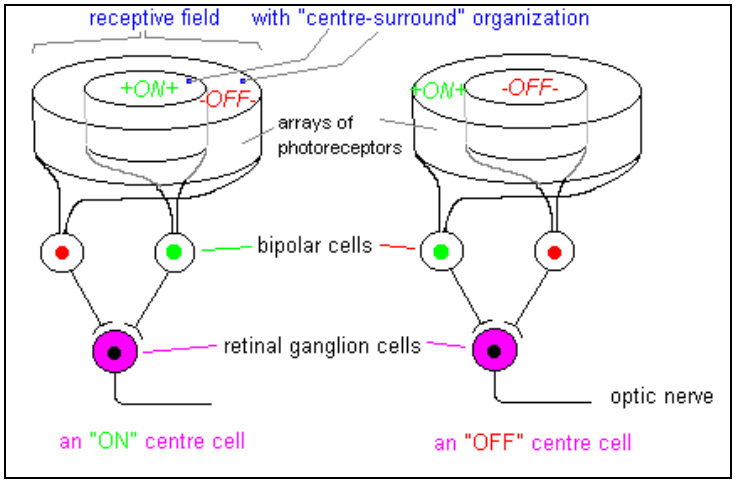


Figura 2-5 – Campo receptivo das células ganglionares. Figura retirada de http://www.cf.ac.uk/biosi/staff/jacob/teaching/sensory/vision.html

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 2-8), 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.

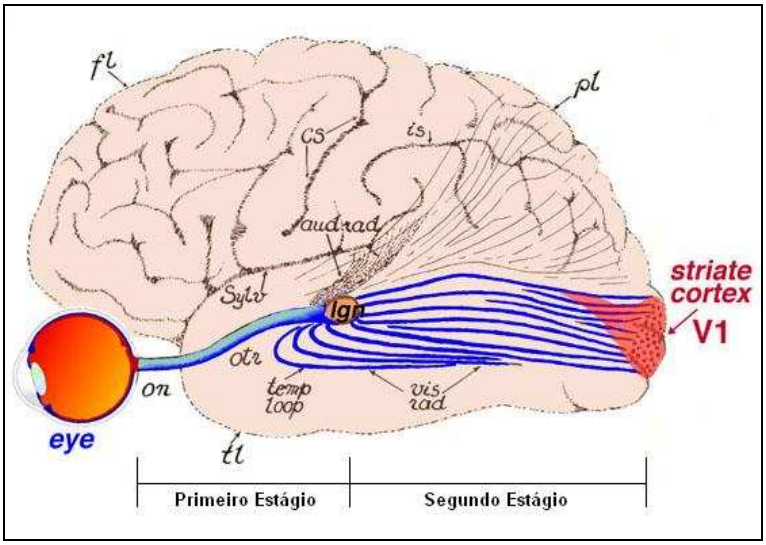


Figura 2-6 – Fluxo das informações visuais. Figura retirada de http://webvision.med.utah.edu/VisualCortex.html e alterada com iserção dos estágios.

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.

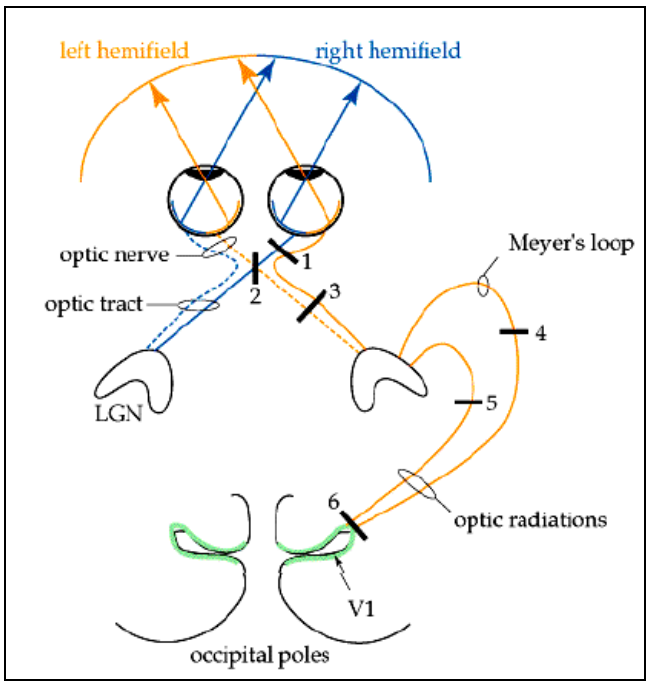


Figura 2-7 – Campo visual. 1) Nervo óptico, 2) Quiasma óptico, 3) Trato óptico. Figura retirada de http://thalamus.wustl.edu/course/basvis.html

O nervo óptico de cada olho projeta-se para o quiasma óptico que é 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.

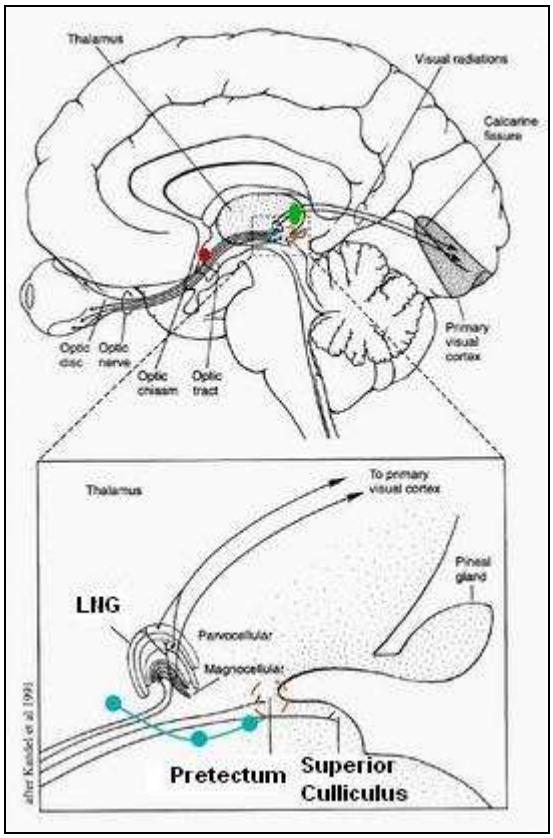


Figura 2-8 – Projeções da retina no mesencéfalo. Figura retirada de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/

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.

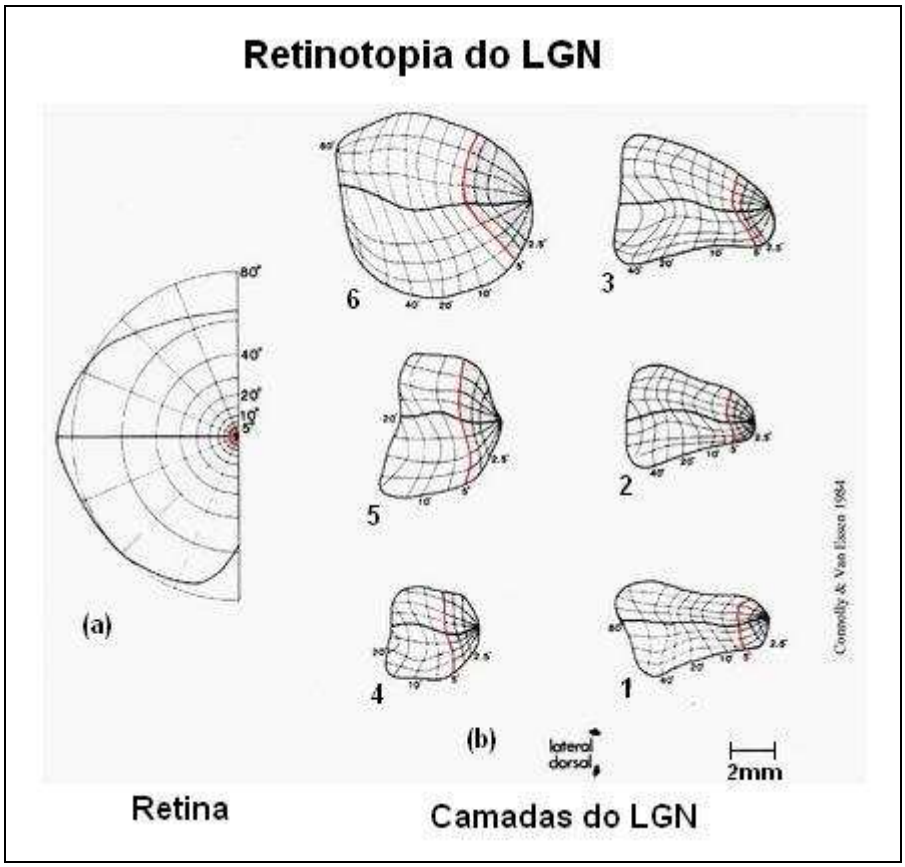


Figura 2-9 – Retinotopia do LGN. (a) Mapeamento da retina. (b) Mapeamento da retina nas camadas 1 à 6 do LGN. Figura retirada e adaptada de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/.

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) contém células relativamente grande 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) contém 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.

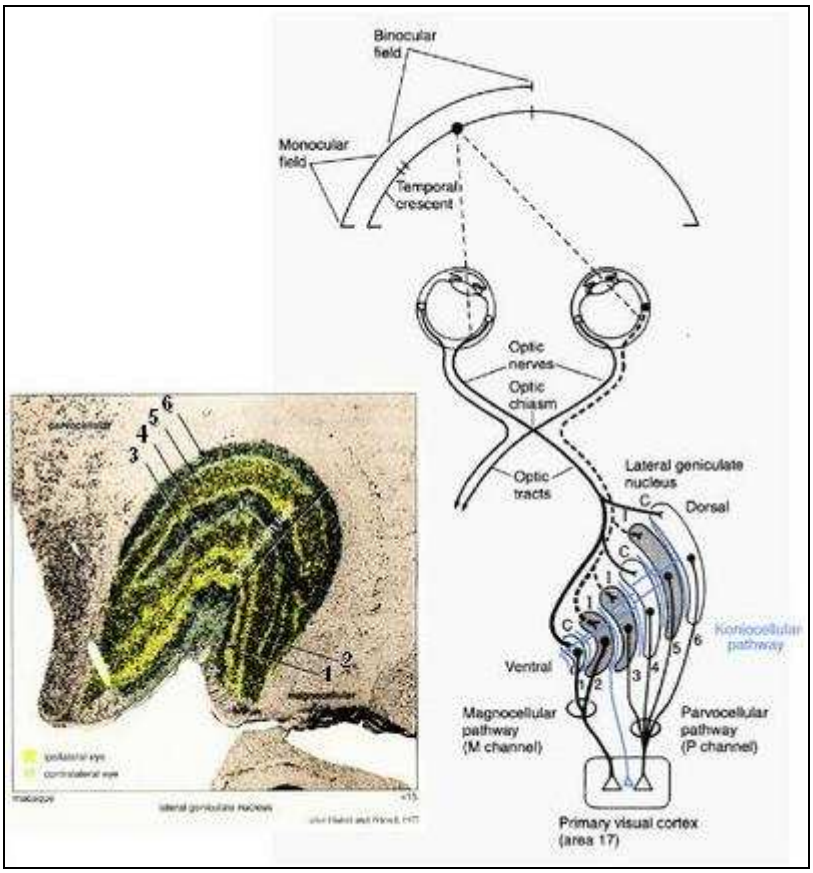


Figura 2-10 – LGN. Figura retirada de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/

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.

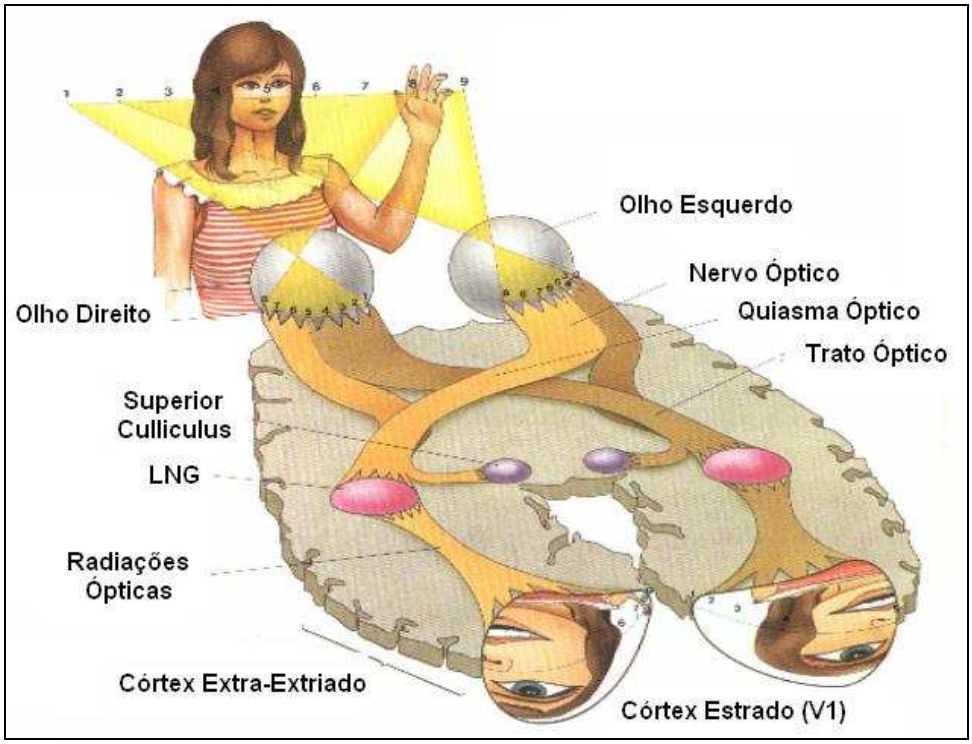


Figura 2-11 – Exemplo ilustrando o fluxo de informações visuais da retina ao córtex estriado. Figura retirada de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/

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.

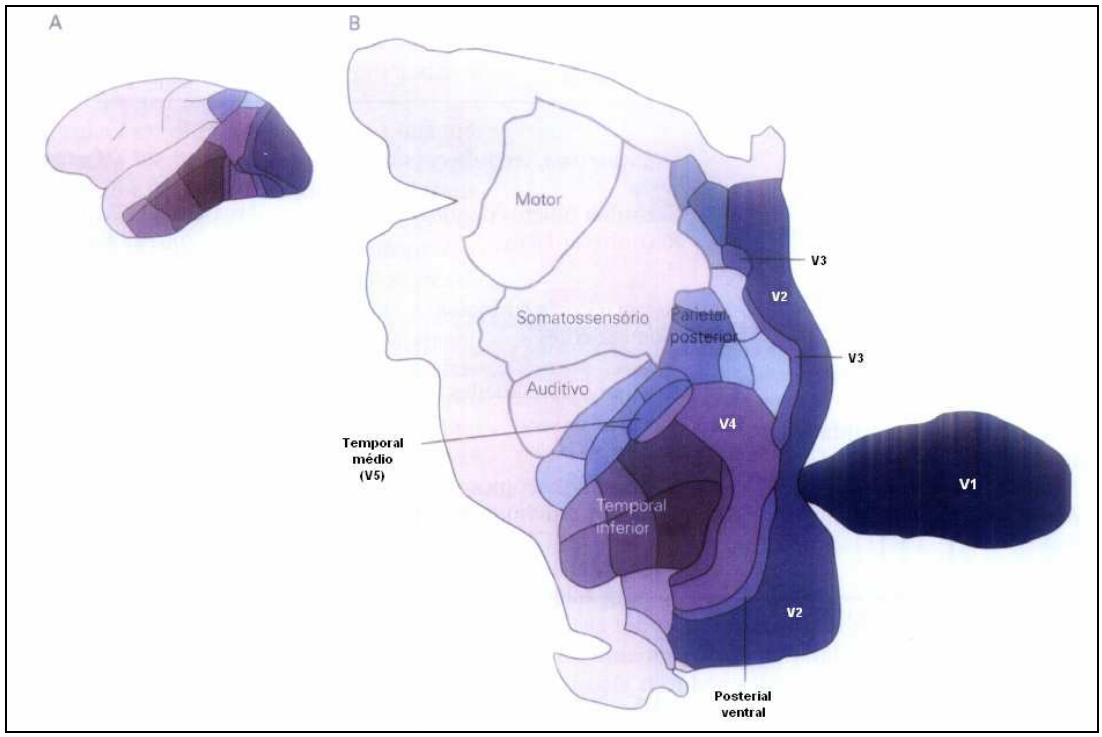


Figura 2-12 – Áreas corticais. Figura retirada de [KAN00].

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 2mm 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 sub camadas: 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.

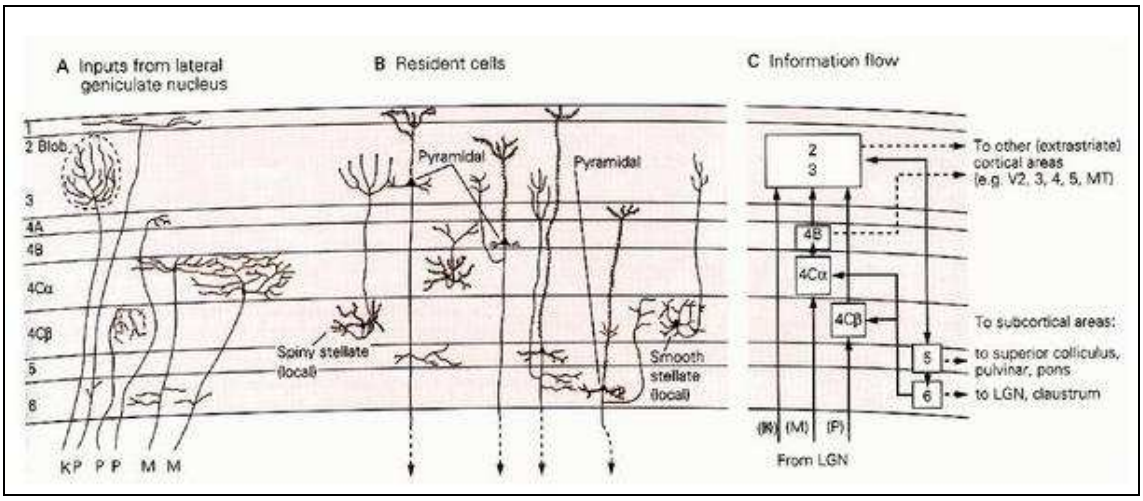


Figura 2-13 – Organização de V1. A) Os axônios dos neurônios P e M do LNG terminam na camada 4; B) Células de V1; C) Concepção do fluxo de informação em V1. Figura retirada de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/

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.

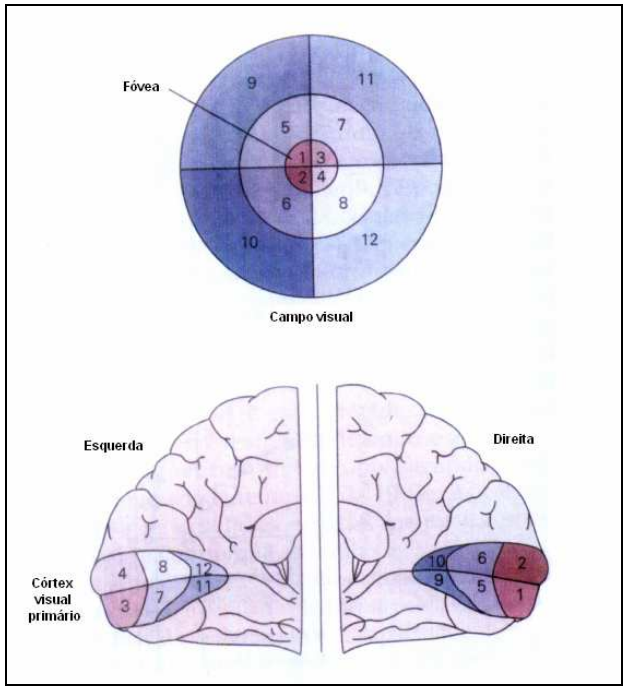


Figura 2-14 – Campo visual representado no córtex visual primário humano. Figura retirada de [KAN00].

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 à estímulos alongados do que à 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 à estímulos na forma de barras com uma orientação específica. Uma célula simples que responde melhor a um estímulo vertical não responderá bem à um estímulo horizontal ou oblíquo, e vice-versa.

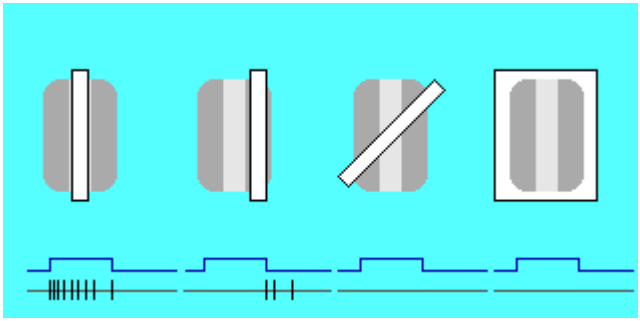


Figura 2-15 – Resposta de uma célula simples em função da projeção de um estímulo em forma de barra. Figura retirada de [MATa]

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.

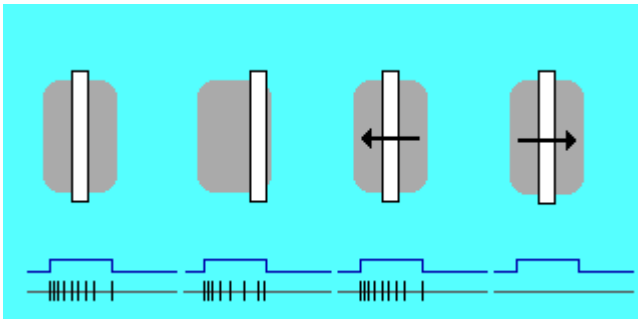


Figura 2-16 – Resposta de uma célula complexa em função da projeção de um estímulo em forma de barra. Figura retirada de [MATa]

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 2mm 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 à 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.

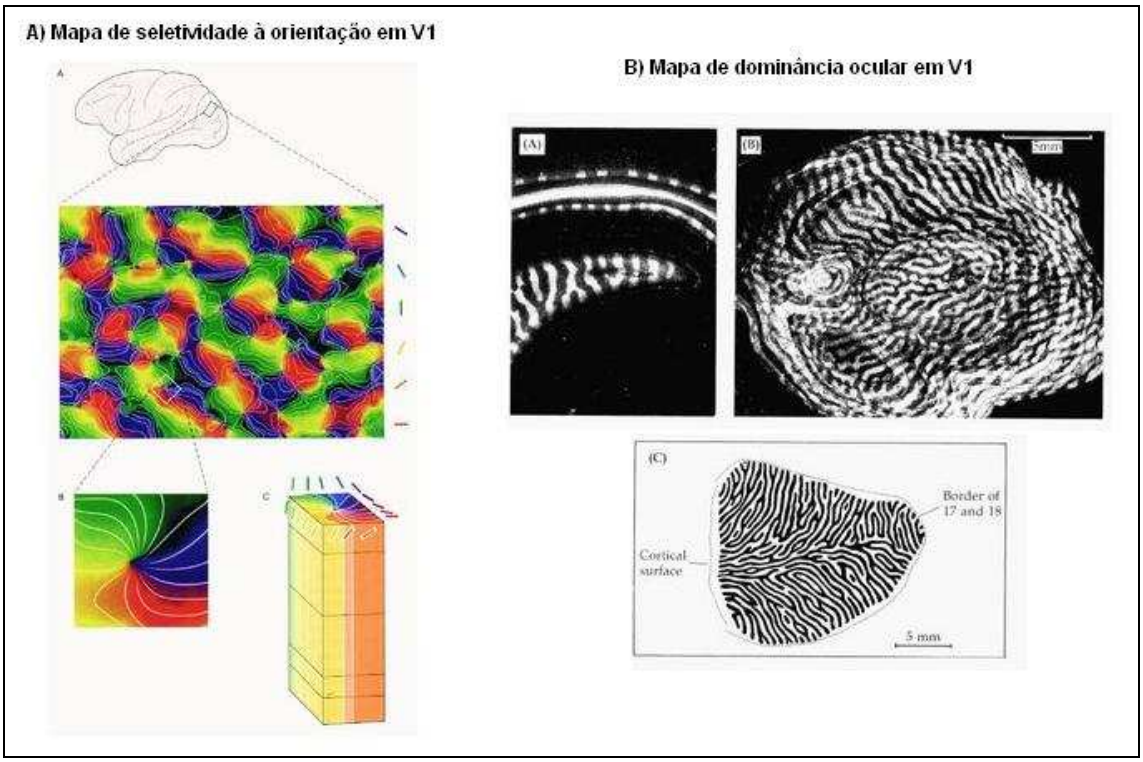


Figura 2-17 – A seletividade à orientação (a) e a dominância ocular (b), variam ao longo da superfície de V1, porém não se alteram numa mesma coluna. Figuras retiradas de http://www.brainworks.uni-freiburg.de/group/wachtler/VisualSystem/.

Um grupo de colunas que respondem à 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.

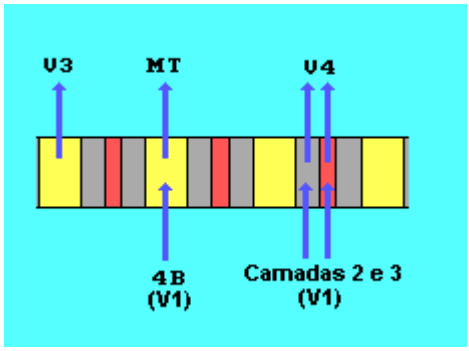


Figura 2-18 – Esquemático de V2. Figura retirada de [MATa]

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.

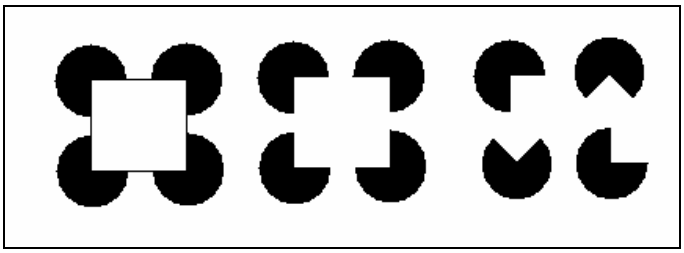


Figura 2-19 – Contorno ilusório. É possível "visualizar" um quadrado branco na figura do meio, apesar de não existir um quadrado desenhado explicitamente.

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 no 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.

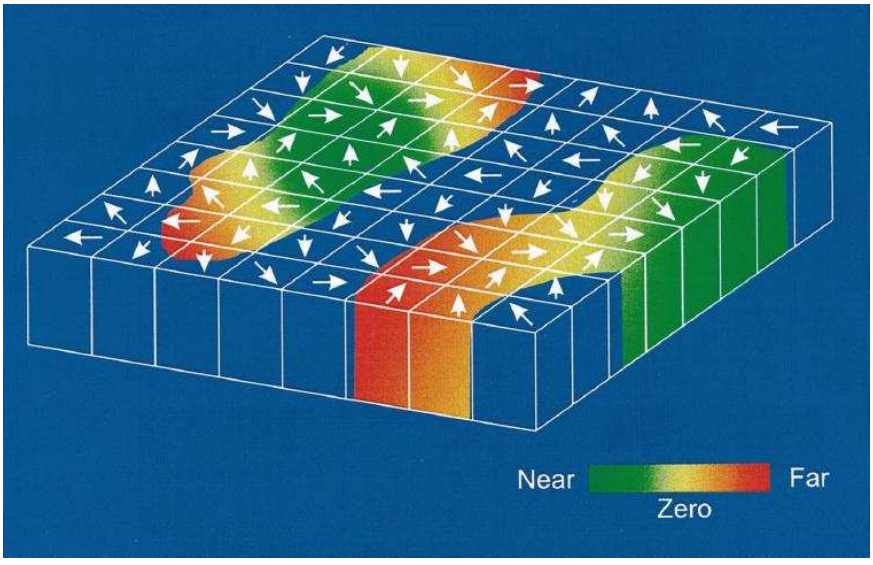


Figura 2-20 – Modelo esquemático da arquitetura funcional de MT. Figura retirada de [DEA99].

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. Um 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.

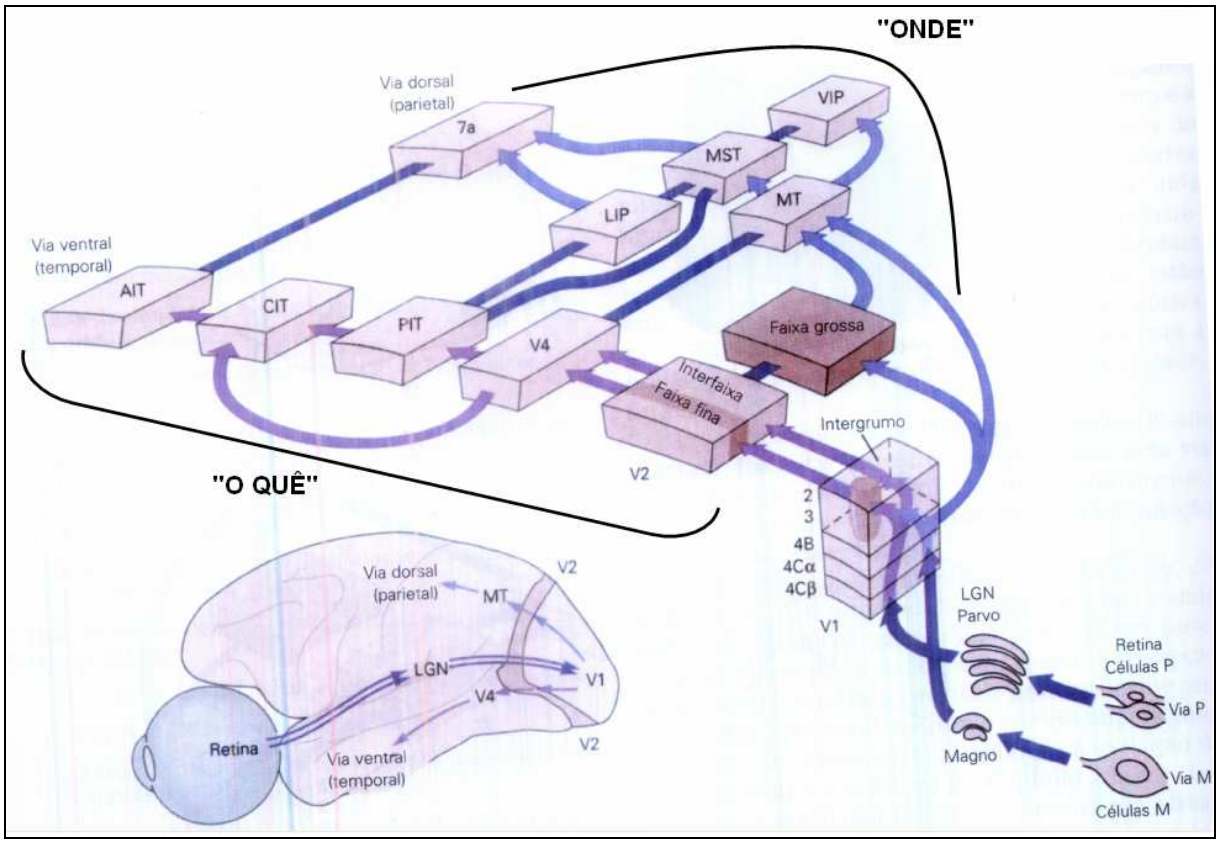


Figura 2-21 – As vias paralelas M e P projetam-se para o córtex visual passando pelo LGN. Figura retirada de [KAN00] e alterada com a inserção .

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*).

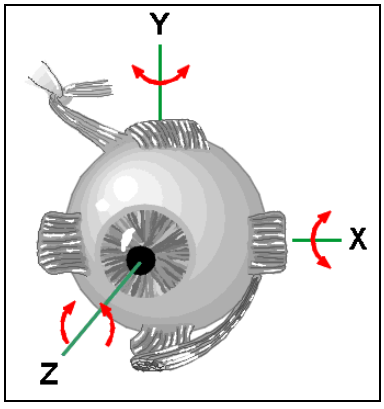


Figura 2-22 – Movimentos oculares. Figura retirada de http://www.auto.ucl.ac.be/EYELAB/Welcome.html com alterações para inclusão dos eixos X, Y e Z.

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].

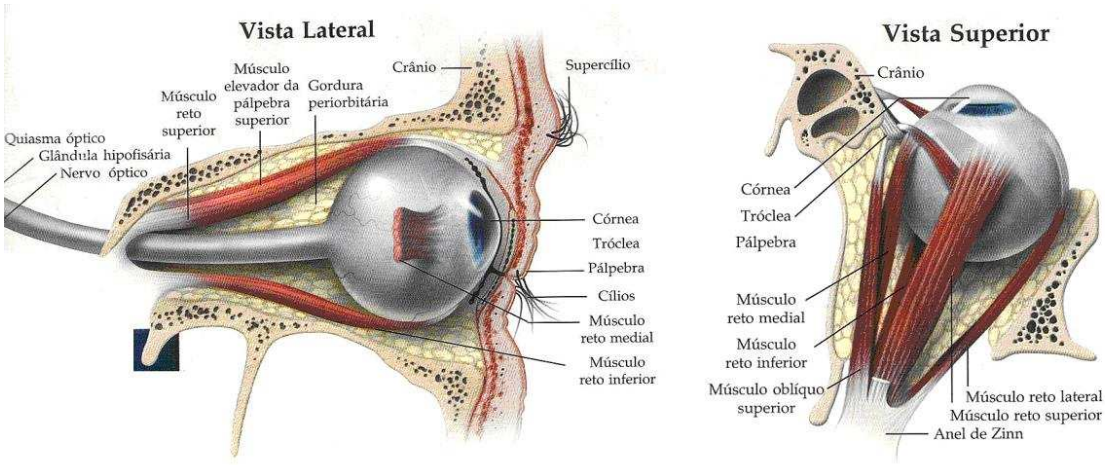


Figura 2-23 – Músculos oculares. A) Vista Lateral; B) Vista Superior. Figuras retiradas de http://www20.brinkster.com/tonho/olho/olhohumano.html

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.

2.1.6. 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 sequencias 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, interposição, oclusão, frequência espacial (tamanho do objeto em relação ao campo de visão), luzes e sombreamento, etc. Gradientes de textura, que capturam a distribuição da direção das bordas em uma imagem, também ajudam a estimar profundidades. Por exemplo, a textura de muitos objetos varia de acordo com a sua distância do observador; da mesma forma, em um chão coberto por ladrilhos, as linhas paralelas formadas pelas juntas parecem divergir com a distância, os ladrilhos mais distantes aparentando divergências aparentes maiores do que os mais próximos. Outra pista monocular utilizável nesse contexto é o efeito de enevoamento, causado pela dispersão da luz causada pela atmosfera terrestre.

2.2. CLASSIFICADOR MRF DE SAXENA

Em [SAX08], Saxena implementa uma rede de *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.

Para cada seção são calculadas um número de *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 informacões: variações de textura, gradientes de textura, e cor. Para capturar as variações de textura, os filtros de Laws são aplicados ao canal de intensidade da imagem. A difração atmosférica da luz (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 orientados são aplicados ao canal de intensidade.

Figura 2-24 – Filtros utilizados para calcular as variações e gradientes de textura. Os primeiros nove são os filtros de Laws 3x3, usados para calcular médias locais, detectar bordas manchas. Os últimos seis são detectores de bordas orientados, espaçados em intervalos de 30º. Figura extraída de [SAX08].

2.2.1. Características absolutas e relativas

Para cada seção *i* da imagem *I(x, y),* a saída dos 17 filtros (9 filtros de Laws, 6 detectores de bordas e 2 canais de cores) F_n, $n = \{1, ..., 17\}$ é usada para calcular um vetor inicial $E_i(n) = \sum_{(x,y)\epsilon patch(i)} |I * F_n|^k$ de dimensão 34, onde k Є {1, 2} nos dá respectivamente a soma absoluta e a soma quadrática da saída de cada filtro. Essa informação local ainda é insuficiente para estimar adequadamente a profundidade da seção; para capturar propriedades mais gerais da imagem, o mesmo vetor é calculado em três escalas espaciais distintas: a mesma da seção original, e dois níveis de distaciamento, 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.

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 $|I * Fn|$, resultando em um vetor de 170 características y_{is}. 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, $y_{ijs} = y_{is} - y_{js}$.

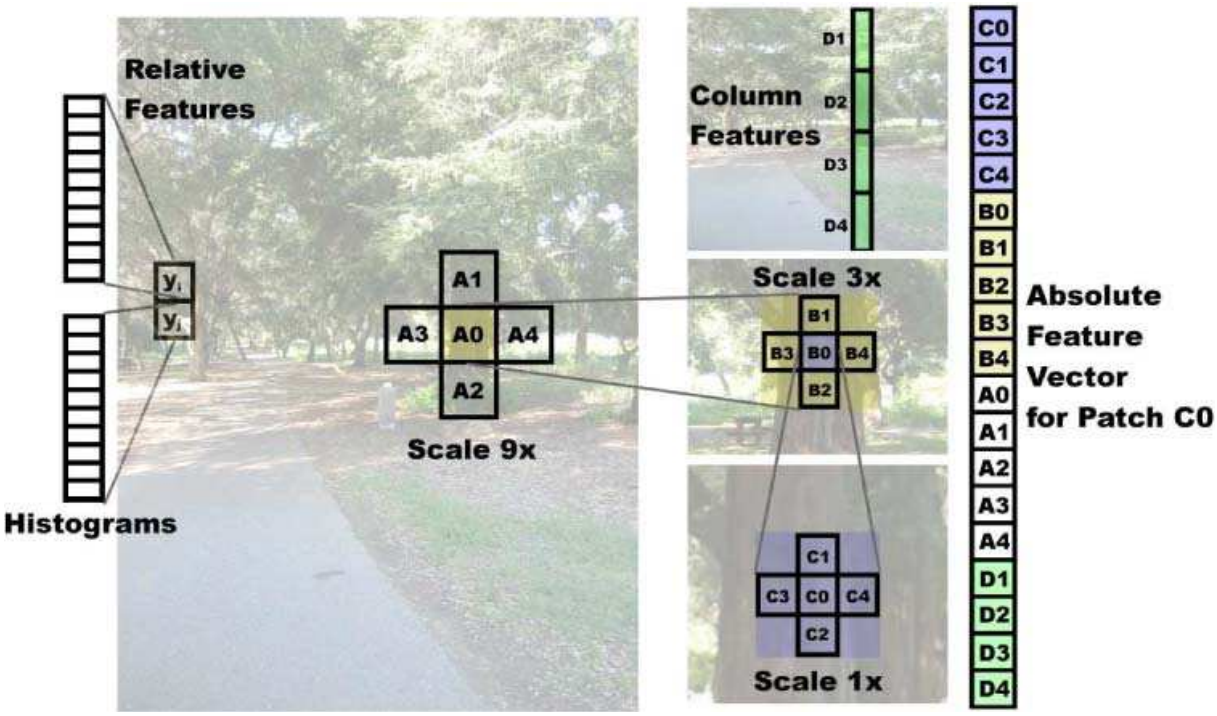


Figura 2-25 – O vetor de características absolutas de uma seção inclui as características dos seus vizinhos imediatos, e também os mais distantes (em escalas espaciais maiores). As características relativas de cada seção usam histogramas das saídas dos filtros. Figura extraída de [SAX08].

2.2.2. Modelo probabilístico

Saxena 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 basica, 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 vetor de parâmetros θ 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 probabibiliades: gaussiana ou laplaciana, sendo que esta última apresentou os melhores resultados.

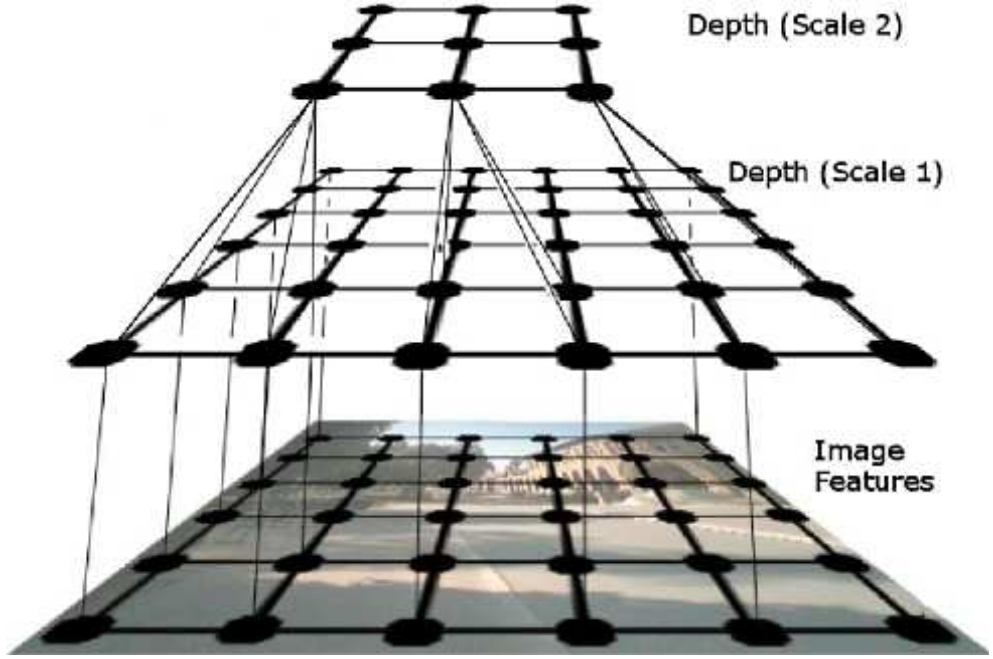


Figura 2-26 – Modelo MRF para modelar relações entre características da imagem e profundidades, entre profundidades na mesma escala, e profundidades em diferentes escalas (apenas 2 de 3 escalas, e um subconjunto das arestas, são representadas). Figura extraida de [SAX08].

3. REDES NEURAIS SEM PESO NA ESTIMATIVA DE PROFUNDIDADE

Neste capítulo detalhamos a estrutura e funcionamento das redes neurais sem peso, com ênfase no modelo VG-RAM, usado pela MAE; também descrevemos as arquiteturas estudadas no trabalho.

3.1. REDES NEURAIS SEM PESO

Redes neurais sem peso (RNSP) são baseadas em *Random Access Memories* (RAM) e não armazenam conhecimento em suas conexões, mas em memórias do tipo RAM dentro dos nós da rede, ou neurônios [ALE66]. Esses neurônios operam com valores de entrada binários e usam RAMs como tabelas-verdade: as sinapses de cada neurônio coletam um vetor de bits da entrada da rede que é usado como o endereço da RAM, e o valor armazenado neste endereço é a saída do neurônio. O treinamento pode ser feito em um único passo e consiste basicamente em armazenar a saída desejada no endereço associado com o vetor de entrada do neurônio.

Apesar da sua notável simplicidade, RNSP são muito efetivas como ferramentas de reconhecimento de padrões, oferecendo treinamento e teste rápidos, e fácil implementação [ALE98]. No entanto, se a entrada da rede for muito grande, o tamanho da memória dos neurônios da RNSP torna-se proibitivo, dado que tem de ser igual a 2^n, onde n é o tamanho da entrada. As RNSP do tipo *Virtual Generalizing RAM* (VG-RAM [LUD99]) são redes neurais baseadas em RAM que somente requerem capacidade de memória para armazenar os dados relacionados ao conjunto de treinamento.

Os neurônios VG-RAM armazenam os pares entrada-saída observados durante o treinamento, em vez de apenas a saída. Na fase de teste, as memórias dos neurônios VG-RAM são pesquisadas mediante a comparação entre a entrada apresentada à rede e todas as entradas nos pares entrada-saída aprendidos. A saída de cada neurônio VG-RAM é determinada pela saída do par cuja entrada é a mais próxima da entrada apresentada – a função de distância adotada pelos neurônios VG-RAM é a distância de *hamming*, i.e., o número de *bits* diferentes entre

dois vetores de *bits* de igual tamanho. Se existir mais do que um par na mesma distância mínima da entrada apresentada, a saída do neurônio é escolhida aleatoriamente entre esses pares.

A Figura 3-1 ilustra a tabela-verdade de um neurônio VG-RAM com três sinapses (*X1*, *X2* e *X3*). Esta tabela-verdade contém três entradas (pares entrada-saída) que foram armazenadas durante a fase de treinamento (*entry* #1, *entry* #2 e *entry* #3). Durante a fase de teste, quando um vetor de entrada é apresentado à rede, o algoritmo de teste VG-RAM calcula a distância entre este vetor de entrada e cada entrada dos pares entrada-saída armazenados na tabela-verdade. No exemplo da Figura 3-1, a distância de *hamming* entre o vetor de entrada (*input*) e a entrada #1 é dois, porque ambos os *bits X2* e *X3* não são semelhantes aos *bits X2* e *X3* do vetor de entrada. A distância da entrada #2 é um, porque *X1* é o único *bit* diferente. A distância da entrada #3 é três, como o leitor pode facilmente verificar. Portanto, para este vetor de entrada, o algoritmo avalia a saída do neurônio, *Y*, como "*category* 2", pois é o valor de saída armazenado na entrada #2.

Lookup table	**X_1**	**X_2**	**X_3**	**Y**
entry #1	1	1	0	category 1
entry #2	0	0	1	category 2
entry #3	0	1	0	category 3
	↑	↑	↑	↓
input	1	0	1	**category 2**

Figura 3-1: Tabela-verdade de um neurônio da RNSP VG-RAM

3.2. ARQUITETURAS DE RNSP PARA RECONHECIMENTO DE PROFUNDIDADES

As próximas seções descrevem as arquiteturas de RNSP VG-RAM propostas neste trabalho para predizer mapas de profundidades a patir de imagens monoculares.

3.2.1. Arquitetura 1

Nossa primeira arquitetura é formada por *z* camadas bidimensionais de neurônios N_o com $m \times n$ neurônios, ligados a uma entrada comum Φ, gerada a partir de uma imagem *I,* ambas de dimensão $u \times v$. Através das camadas, cada neurônio n_{ij}

observa uma região de tamanho *(u/m)×(v/n)* da entrada; neurônios ocupando a mesma posição em camadas diferentes observam a mesma região, enquanto neurônios em posições diferentes observam regiões distintas (não necessariamente disjuntas). A saída da rede é calculada a partir da saída das camadas, pela aplicação do algoritmo *Winner-Take-All* (WTA). A geração da entrada a partir da imagem original e o funcionamento do algoritmo WTA serão discutidos mais adiante.

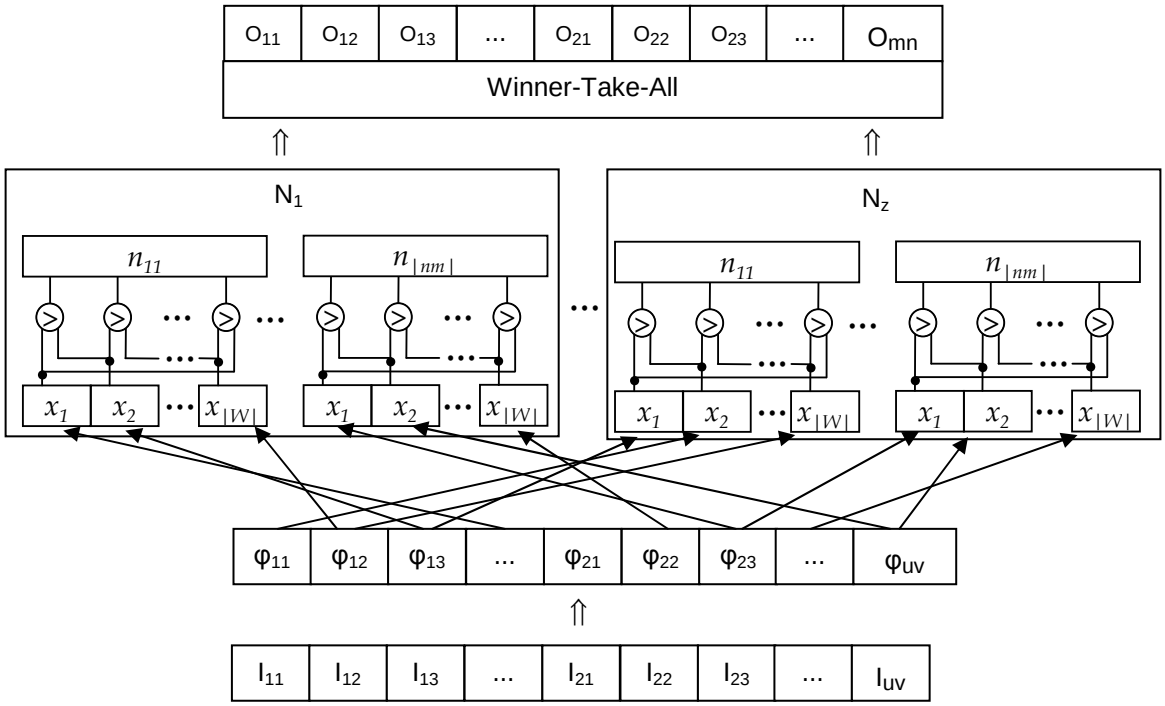


Figura 3-2: Primeira arquitetura neural para o reconhecimento de imagens

Cada neurônio n_{ij} possui um conjunto de sinapses, $W = \{w_1, ..., w_{|W|}\}$. Estas sinapses são conectadas à entrada bidimensional da rede, Φ, composta de $u \times v$ elementos de entrada, $\varphi_{k,l}$, segundo um padrão de interconexão $\Omega_{i,j}(W)$. Este padrão de interconexão sináptica segue uma distribuição randômica bidimensional com PDF (*Probabilistic Density Function*) Normal centrada no pixel $\varphi_{\mu k,\mu l}$, onde $\mu k = i.u/m$ e $\mu l = j.v/n$. Isto é, as coordenadas k e l dos elementos de Φ aos quais $n_{i,j}$ se conecta via W seguem as PDFs:

$$\varphi_{\mu_k,\sigma^2}(k) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-\mu_k)^2}{2\sigma^2}}$$

$$\varphi_{\mu_l,\sigma^2}(l) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(l-\mu_l)^2}{2\sigma^2}}$$

onde σ é um parâmetro de nossa arquitetura neural. A Figura 3-3 mostra de forma gráfica o padrão de interconexão Ω de três neurônios. Este padrão de interconexão sináptica imita aquele observado em muitas classes de neurônio biológico. Em nosso sistema de reconhecimento de imagens, ele é criado juntamente com a rede neural e não muda após a sua criação.

Sinapses de RNSP coletam apenas um bit (0 ou 1) da entrada. Para permitir seu uso com entradas que podem assumir valores não binários, usamos *minchinton cells* [MIT98]. Em nossa arquitetura neural para o reconhecimento de imagens, cada sinapse, w_t, forma uma *minchinton cell* com a próxima, w_{t+1} ($w_{|W|}$ forma uma *minchinton cell* com w_1). Cada uma destas *minchinton cells* retorna 1 se a sinapse w_t está conectada a um elemento da entrada Φ, φ_{ij}, cujo valor, x_t, seja maior que o valor do elemento φ_{kl} ao qual a sinapse w_{t+1} está conectada; caso contrário, retorna zero.

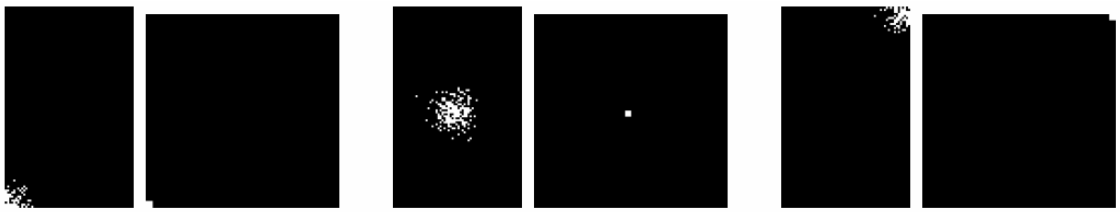


Figura 3-3: O padrão de interconexão sináptica Ω. (a) A imagem da esquerda mostra a entrada Φ: na cor branca, os elementos $\varphi_{i,j}$ da entrada Φ que estão conectados ao neurônio $n_{0,0}$ de N via $w_1, \ldots, w_{|W|}$; a imagem da esquerda mostra a camada bidimensional de neurônios N: na cor branca, o neurônio $n_{0,0}$ de N. (b) Esquerda: em branco, os elementos de $\varphi_{i,j}$ de Φ conectados a $n_{m/2,n/2}$; direita: em branco, o neurônio $n_{m/2,n/2}$ de N. (c) Esquerda: em branco, os elementos de Φ conectados a $n_{m,n}$; direita: em branco, o neurônio $n_{m,n}$.

Para treinar nossa rede neural, inicialmente aplicamos um filtro gaussiano à imagem de entrada: esse filtro "suavisa" a imagem, removendo componentes de alta frequência, e tornando os valores de entrada dos neurônios mais generalizáveis. A imagem filtrada é levada à entrada da rede e, então, cada neurônio é treinado para, dada a região da imagem por ele observada, apresentar como saída o valor de profundidade associado àquela região.

Para usar a rede neural para prever um mapa de profundidades, novamente a imagem precisa ser filtrada e levada à entrada da rede. Cada um dos neurônios apresenta, então, como saída, o valor de profundidade que acredita estar associado à região observada. O resultado é uma matriz tridimensinal de profundidades *P,* de

dimensão *m*×*n*×*z,* que é então combinada em uma saída *O* de acordo com o algoritmo abaixo:

1. Montar uma saída inicial *O(0)* tal que *O(0)[x, y]* seja o valor encontrado com mais frequência no cojunto *P[x, y, 1..z]* (se dois ou mais valores "empatarem", um deles é escolhido aleatoriamente);
2. Montar uma saída intermediária O(t) tal que *O(t)[x, y]* seja escolhido entre os valores do conjunto *P[x, y, 1..z],* como o valor mais "próximo" (em diferenças absolutas) da média dos oito vizinhos de *O(t-1)[x, y];*
3. Repetir o passo (2) até a saída convergir (isto é, *O(t) = O(t-1)*) ou um número máximo de iterações *M* ser atingido.

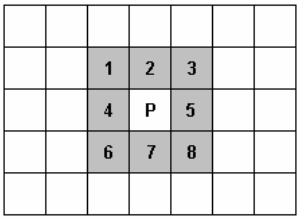


Figura 3-4 – Os oito pontos vizinhos de um ponto P específico. Figura retirada de [OLI05].

3.2.2. Arquitetura 2

Quando observamos os mapas de profundidades de muitas imagens, percebemos que elas tendem a obedecer um padrão de distribuicão em espectro, com as regiões inferiores da imagem tendendo a valores menores, e as regiões superiores, a valores maiores. Intuitivamente, objetos que estejam “no chão” (e o próprio “chão”) tendem a estar mais próximos do observador do que aqueles que se encontram “no céu” (e o próprio “céu”). Isso nos leva a conjecturar que diferentes seções horizontais da imagem apresentam distribuições de profundidades distintas; associando um único neurônio a cada uma dessas seções, obteríamos em princípio uma arquitetura de predição de profundidades mais poderosa do que a anterior, onde o conhecimento de cada neurônio fica restrito a uma região retangular da imagem, e não pode ser aplicado em outras regiões na mesma altura.

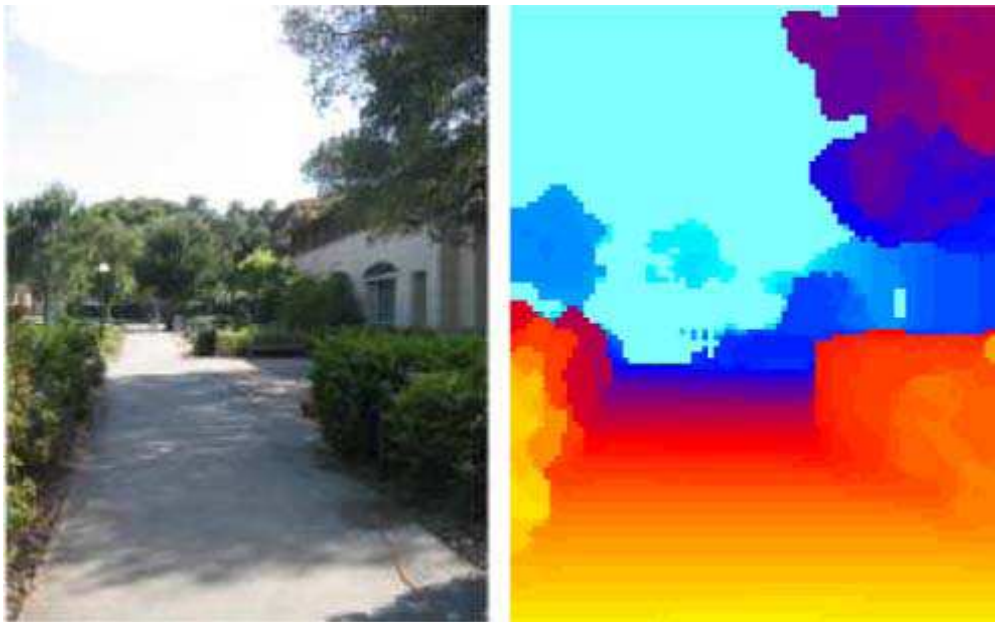


Figura 3-5 – Imagem monocular e mapa de profundidades correspondente. Cores mais “quentes” indicam regiões próximas do observador. Com poucas exceções (veja o canto superior direito) regiões inferiores da imagem estão mais próximas do que as superiores. Figura retirada de [SAX08].

Para investigar essa conjectura, modificamos a Arquitetura 1 de forma que as camadas neurais N_o contenham apenas uma coluna de neurônios cada; a largura da entrada Φ também foi reduzida para 5σ (suficiente para garantir que 99% das sinapses geradas pela gaussiana caiam “dentro” da camada de entrada), e um

elemento de deslizamento foi adicionado entre ela e a imagem, permitindo que diferentes “colunas” da imagem sejam submetidas à rede.

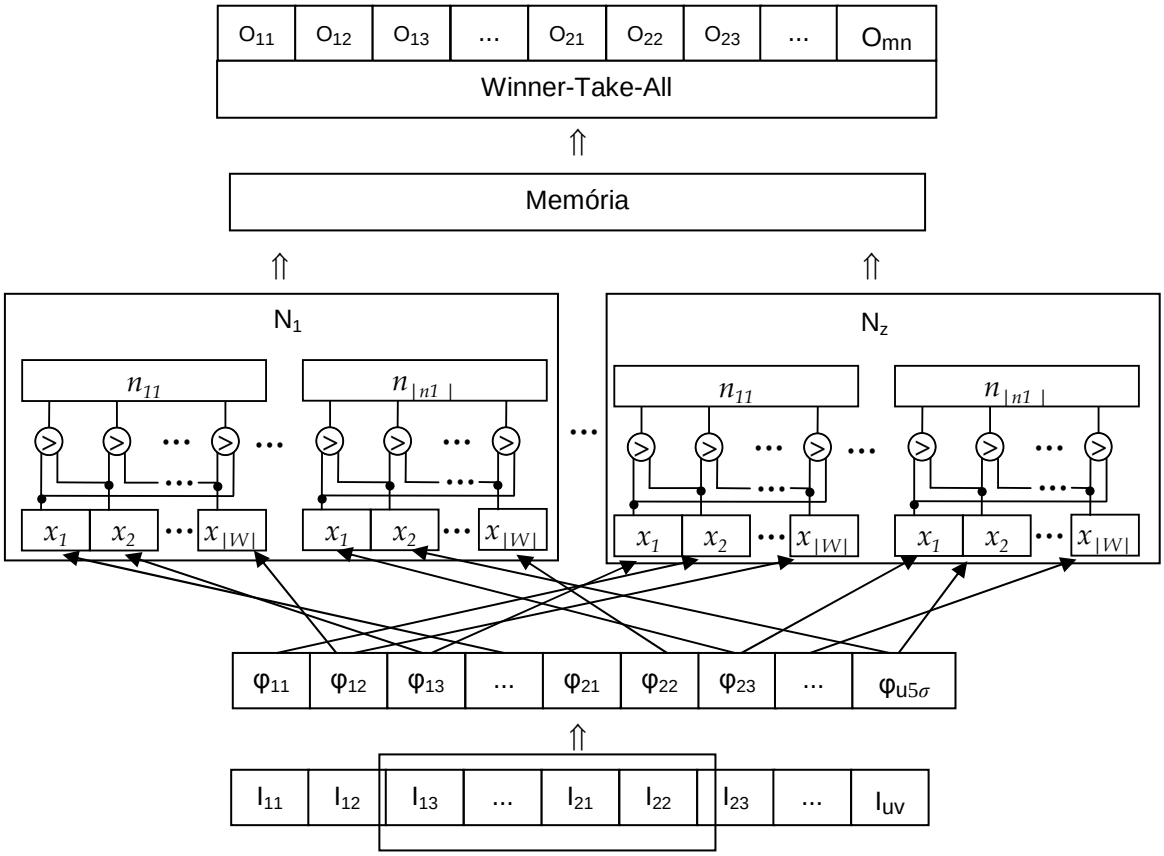


Figura 3-6: Segunda arquitetura neural para o reconhecimento de imagens.

Essa rede neural é treinada em *m* turnos; a cada turno, o elemento de deslizamento (inicialmente centrado na coluna de largura *(v / m)* mais à esquerda da imagem) é deslocado *(v / m)* pixels para a direita, e cada neurônio é treinado para, dada a região da imagem por ele observada, apresentar como saída o valor de profundidade associado àquela região. Dessa forma, cada neurônio aprende as profundidades de *m* regiões dispostas horizontalmente ao longo de uma certa altura da imagem.

Para usar a rede neural para prever um mapa de profundidades, novamente é preciso executar *m* turnos, nos quais diferentes regiões da imagem são expostas aos neurônios. As saídas apresentadas pelos neurônios a cada turno são armazenadas em um elemento de memória, reponsável por construir a matriz de profundidades *P* que será passada ao algoritmo WTA ao término do processo, para calcular a reposta final.

3.2.3. Arquitetura 3

Mesmo agrupando um conjunto de pontos em posições específicas, a identificação de uma seção de imagem pode ser extremamente ambígua, se

consideramos apenas valores no espaço de cores. Isso é indesejável, pois se duas ou mais seções forem mapeadas para um mesmo padrão de entrada, um valor de profundidade comum será associado a todas elas como resultado – mesmo que isso não seja verdade nos dados de treinamento. Para reduzir o risco de colisão, gostaríamos de extrair a maior quantidade possível de informação particular da imagem, produzindo assim “assinaturas” únicas para cada seção.

Voltando ao trabalho de Saxena, tomamos emprestada sua idéia de dividir a imagem em componentes de intensidade e cor, decompondo-a nos três canais do espaço *Hue, Saturation, Value* (HSV); substituímos a Gaussiana pelo primeiro filtro de Laws, e o aplicamos aos três canais. Adicionalmente, usamos os seis detectores de bordas sobre o canal de *Hue* (intensidade) para gerar uma saída representativa dos gradientes de textura da imagem.

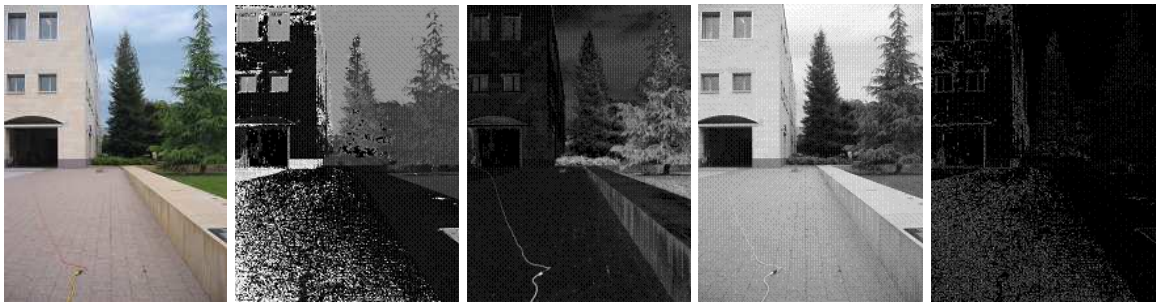


Figura 3-7: Imagem original, canais HSV, e saída do detector de bordas.

Partindo da arquitetura 2, modificamos a entrada para cada neurônio receber as saídas desses quatro filtros para as suas seções da imagem; o restante do funcionamento da rede permanceu inalterado.

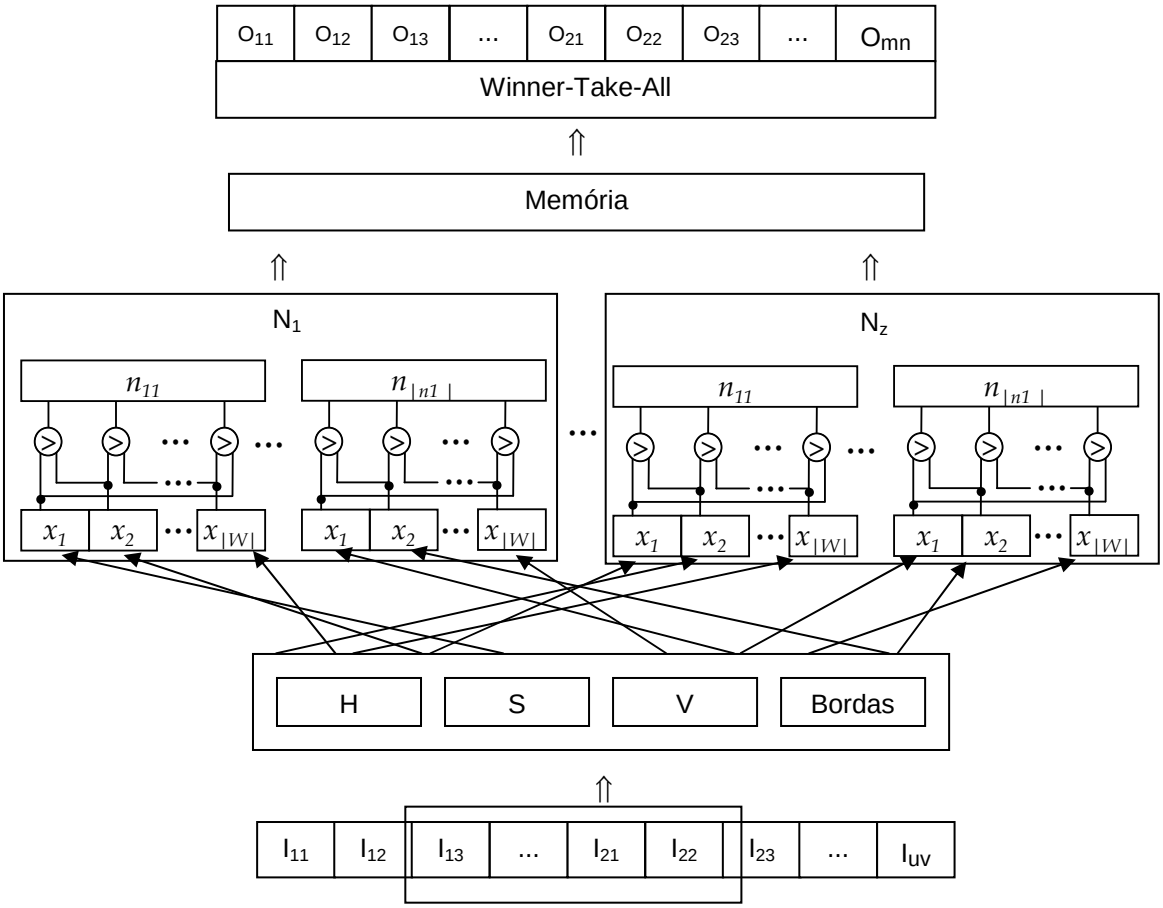


Figura 3-8: Terceira arquitetura neural para o reconhecimento de imagens. Os índices das entradas foram omitidos por simplicidade.

4. METODOLOGIA

Neste capítulo apresentamos a metodologia empregada no trabalho, os experimentos efetuados e resultados obtidos. Começamos apresentando a base de dados disponibilizada por Saxena, suas características, 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. Em seguida, descrevemos os experimentos realizados para avaliar os parâmetros ótimos de configuração para cada arquitetura neural, além dos testes adicionais de validação. Passamos então à análise e comentário dos resultados obtidos, comparando também com os alcançados por Saxena.

4.1. BASE DE DADOS

A base de dados usada nos experimentos foi construída a partir da utilizada em [SAX08]. Segundo o artigo, a base original era composta de 425 casos compostos de uma imagem e um mapa de profundidades (aferido com o auxílio de um *scanner* a laser), com resolucão de 1704×2272 para as imagens, e 86×107 para os mapas de profundidades. As cenas encontradas nas imagens podem ser informalmente classificadas, de acordo com seus elementos dominantes, em três categorias:

- Espaços interiores;
- Paisagens urbanas;
- Paisagens naturais.

Ao tentar obter essa base de dados no *website* do pesquisador (www.cs.cornell.edu/~asaxena/learningdepth/data.html), deparamos com vários contratempos:

1. Os casos estavam separados em dois conjuntos intitulados *Dataset 2* e *Dataset 3,* e encontravam-se misturados com outros 105 casos que não haviam sido usados nos experimentos do artigo (vários dos quais contendo apenas a imagem, sem o mapa de profundidades), sem que houvesse um indicativo claro de quais casos haviam sido descartados;
2. Todas imagens de entrada estavam rotacionadas 90º para a esquerda;
3. 100 mapas de profundidades do *Dataset 2* estavam na resolução 72×58, ao invés de 86×107.

Tendo em vista esses problemas, procedemos à organização da base de dados. Sabíamos que os casos descartados haviam sido postos de lado por problemas de alinhamento entre as imagens e os mapas de profundidades causados por pessoas ou objetos em movimento. Descartando as imagens onde esse tipo de situação era aparente, e também aquelas para as quais não havia um mapa de profundidades associado, conseguimos reconstruir a base de dados original. Em seguida efetuamos a rotação de todas as imagens em 90º para a direita, e redimensionamos os mapas de profundidade fora do padrão para a resolução 86×107. Para a conveniência de trabalhos futuros, os dados selecionados e tratados estão disponíveis para download em (www.lcad.inf.ufes.br/wiki/index.php/Diver).

4.2. MÉTRICAS

Em [SAX08] é utilizada a métrica de Erro Absoluto Médio (*Mean Absolute Error,* ou MAE), aplicada sobre o logaritmo das profundidades, com o objetivo de enfatizar erros multiplicativos sobre erros aditivos. Para permitir uma comparação quantitativa com os resultados desse artigo, adotamos a mesma métrica. Para permitir uma compreensão mais intuitiva da magnitude dos erros, também calculamos a MAE para as profundidades em escala linear. Os valores de MAE e log MAE para cada teste são calculados a partir dos mapas de profundidades previstos pelo sistema e os mapas de referência da base de dados.

4.3. EXPERIMENTOS

Os experimentos foram divididos em duas fases:

1. Sessões de testes para avaliar os parâmetros de configuração ótimos para cada arquitetura;
2. Um teste final de validação empregando os valores ótimos para cada arquitetura.

Dos 425 casos disponíveis na base de dados, 141 foram postos de lado para o teste final de validação; dos demais 284, 190 foram selecionados para treinamento durante as sessões de ajuste de parâmetros, e os restantes 94 utilizados para avaliar a capacidade de predição da rede após cada treinamento.

Todas as arquiteturas têm os mesmos parâmetros de configuração: *z,* o número de camadas neurais; σ*,* que determina o grau de dispersão das sinapses de cada neurônio; e o número de sinapses por neurônio. Para cada arquitetura, configuramos *z* para o número mais alto permitido pelos recursos de memória da plataforma de testes (estações de trabalho Intel de 32 bits, com 2GB de memória), e em seguida executamos 16 testes para ajuste de parâmetros, variando σ de 5 a 40, e o número de sinapses de 32 a 256, em potências de dois (isto é, foram usados os valores 5, 10, 20 e 40 para σ*,* e 32, 64, 128 e 256 para o número de sinapses).

4.3.1. Ajustes da Arquitetura 1

Os resultados dos testes de ajuste de parâmetros para a Arquitetura 1 encontram-se representados abaixo:

σ	Sinapses	MAE	log MAE
5	32	8,122	0,213
5	64	8,231	0,215
5	128	8,518	0,219
5	256	8,967	0,227
10	32	8,109	0,212
10	64	8,191	0,212
10	128	8,463	0,215
10	256	8,944	0,222

σ	Sinapses	MAE	log MAE
20	32	8,150	0,211
20	64	8,171	0,210
20	128	8,477	0,213
20	256	9,008	0,220
40	32	8,186	0,211
40	64	8,231	0,211
40	128	8,544	0,213
40	256	9,309	0,222

Tabela 4-1 – Testes de ajustes de parâmetros para a Arquitetura 1, z = 3

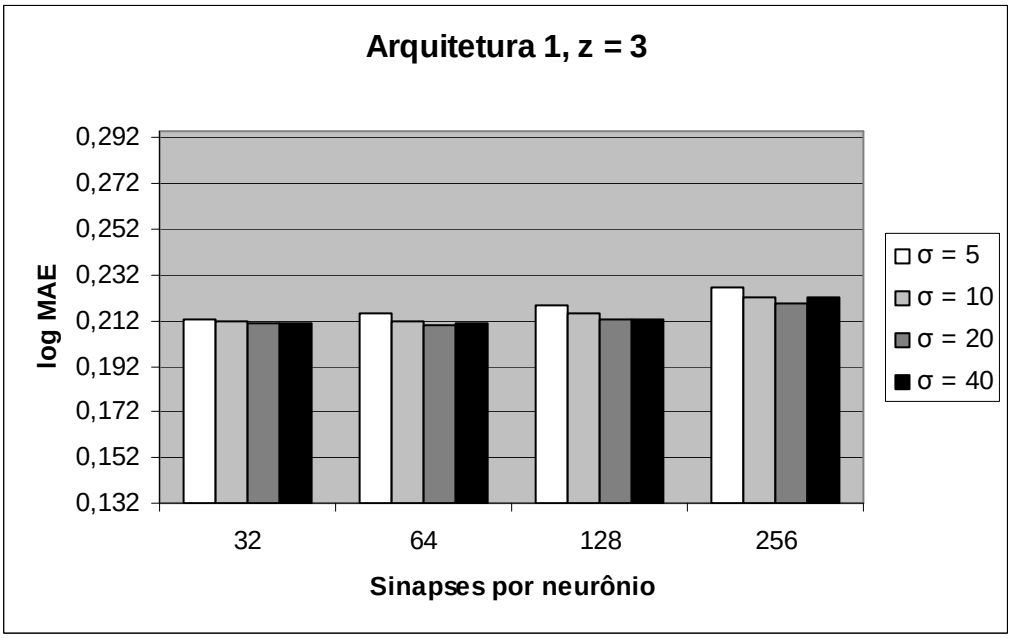


Figura 4-1 – Testes de ajustes de parâmetros para a Arquitetura 1. Barras de mesma cor referem-se a testes feitos com um mesmo valor para *σ*, o fator de dispersão das sinapses (veja legenda na imagem). Eixo X: número de sinapses por neurônio. Eixo Y: margem de erro da arquitetura, em escala log Mean Absolute Error (MAE); valores menores indicam melhor desempenho.

Observando os resultados, vemos que o desempenho da Arquitetura 1 tende a melhorar em função inversa do número de sinapses e direta do valor de σ – pelo menos, até $\sigma = 20$, a partir de onde o desempenho da rede também começa a piorar com valores maiores. Dado que essa é a arquitetura com pior capacidade de generalização, esses resultados não são surpresa: adicionar mais sinapses (i.e. informação) apenas aumenta o nível de ruído, mesmo extender o campos de vicão dos neurônios oferece uma vantagem limitada. Baseados nessas observações, concluímos que a configuração de 32 sinapses e $\sigma = 20$ é ótima para essa arquitetura.

4.3.2. Ajustes da Arquitetura 2

Os resultados dos testes de ajuste de parâmetros para a Arquitetura 2 encontram-se representados abaixo:

σ	Sinapses	MAE	log MAE
5	32	7,946	0,201
5	64	7,823	0,198
5	128	7,677	0,195
5	256	7,538	0,193
10	32	7,927	0,201
10	64	7.864	0,200
10	128	7,745	0,198
10	256	7,652	0,197

σ	Sinapses	MAE	log MAE
20	32	7,935	0,200
20	64	7,774	0,199
20	128	7,749	0,200
20	256	7,762	0,200
40	32	7,930	0,200
40	64	7,746	0,198
40	128	7,648	0,197
40	256	7,741	0,199

Tabela 4-2 – Testes de ajustes de parâmetros para a Arquitetura 2, z = 10

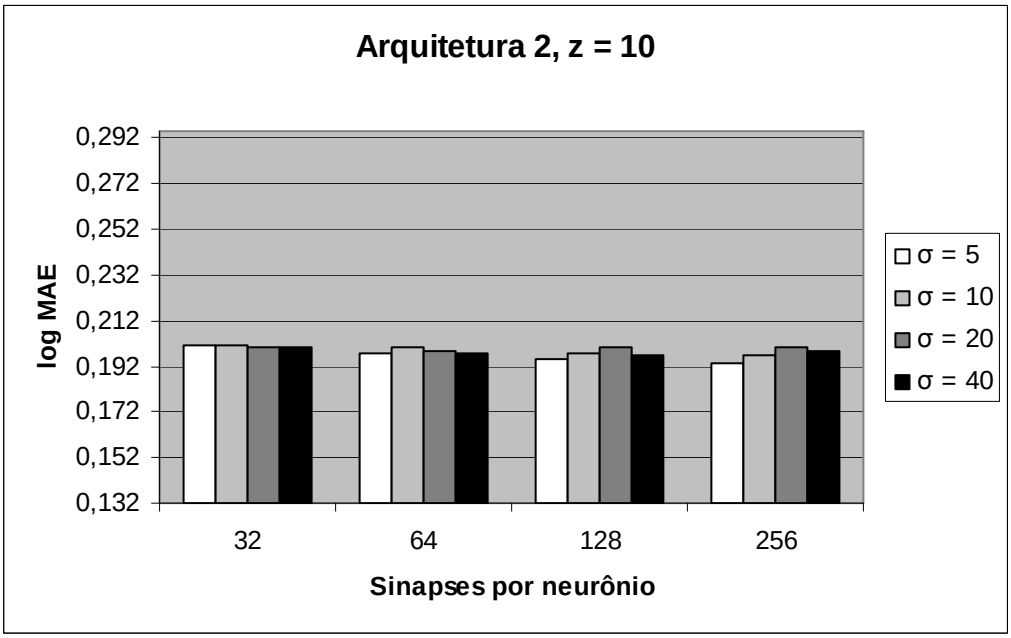


Figura 4-2 – Testes de ajustes de parâmetros para a Arquitetura 2. Barras de mesma cor referem-se a testes feitos com um mesmo valor para *σ*, o fator de dispersão das sinapses (veja legenda na imagem). Eixo X: número de sinapses por neurônio. Eixo Y: margem de erro da arquitetura, em escala log Mean Absolute Error (MAE) ; valores menores indicam melhor desempenho.

Observando os resultados, vemos que a Arquitetura 2, projetada para uma maior capacidade de generalização dos dados de treinamento, se beneficia do aumento do número de sinapses (entretanto a dispersão das sinapses com o aumento do valor de *σ* reduz seu desempenho). Além disso, mesmo nesta fase de testes, é evidente a melhora quantitativa em relação à arquitetura original. Baseados

nessas observações, concluímos que a configuração de 256 sinapses e σ = 5 é ótima para essa arquitetura.

4.3.3. Ajustes da Arquitetura 3

σ	Sinapses	MAE	log MAE
5	32	7,504	0,193
5	64	7,224	0,190
5	128	7,130	0,188
5	256	7,045	0,186
10	32	7,522	0,193
10	64	7,213	0,190
10	128	7,115	0,189
10	256	7,009	0,187

σ	Sinapses	MAE	log MAE
20	32	7,270	0,190
20	64	7,195	0,190
20	128	7,021	0,188
20	256	6,899	0,186
40	32	7,443	0,192
40	64	7,159	0,188
40	128	6,943	0,185
40	256	6,687	0,180

Tabela 4-3 – Testes de ajustes de parâmetros para a Arquitetura 3, z = 10

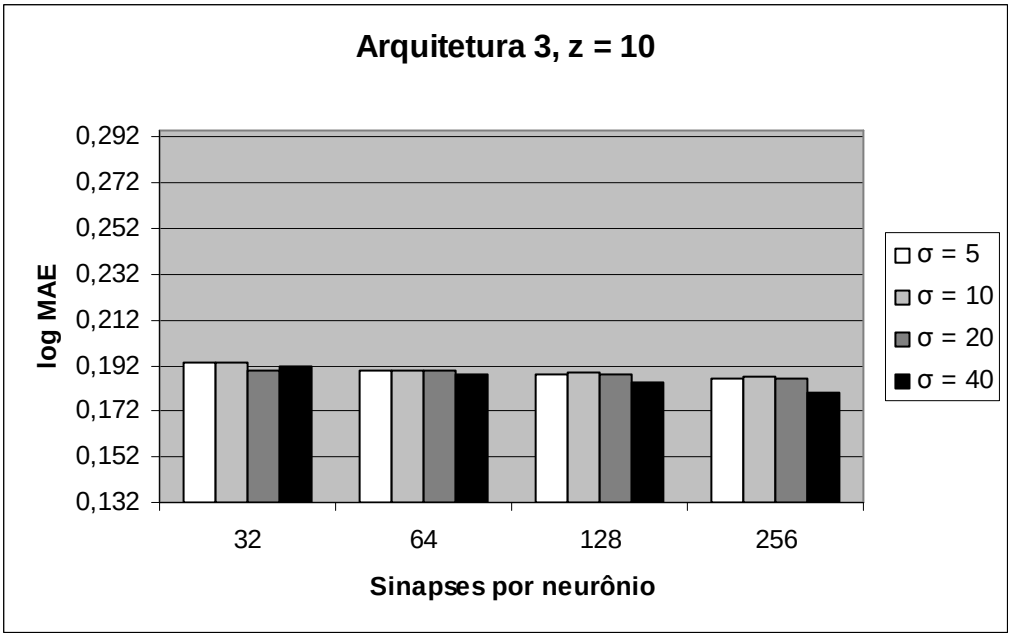


Figura 4-3 – Testes de ajustes de parâmetros para a Arquitetura 3. Barras de mesma cor referem-se a testes feitos com um mesmo valor para σ, o fator de dispersão das sinapses (veja legenda na imagem). Eixo X: número de sinapses por neurônio. Eixo Y: margem de erro da arquitetura, em escala log Mean Absolute Error (MAE) ; valores menores indicam melhor desempenho.

Observando os resultados, vemos que a Arquitetura 3 é a que melhor aproveita os dados de treinamento: seu desempenho melhora em função direta tanto de σ quanto do número de sinapses, sem indicação de reverter essa tendência

no espaço de testes que executamos. Baseados nessas observações, concluímos que a configuração de 256 sinapses e $\sigma = 40$ é ótima para essa arquitetura.

4.3.4. Testes de validação

Concluídos os testes de ajuste de parâmetros, procedemos aos testes finais de validação, cujos resultados encontram-se resumidos na tabela abaixo, junto com os resultados obtidos por Saxena:

Sistema	log MAE
RNSP Arquitetura 1	0,218
RNSP Arquitetura 2	0,198
RNSP Arquitetura 3	0,172
MRF Baseline	0,295
MRF Gaussian (S1,S2,S3, H1,H2, no neighbors)	0,162
MRF Gaussian (S1, H1,H2)	0,171
MRF Gaussian (S1,S2, H1,H2)	0,155
MRF Gaussian (S1,S2,S3, H1,H2)	0,144
MRF Gaussian (S1,S2,S3, C, H1)	0,139
MRF Gaussian (S1,S2,S3, C, H1,H2)	0,133
MRF Laplacian	0,132

Tabela 4-4 – Testes finais de validação das arquiteturas neurais, comparados com os resultados obtidos por Saxena

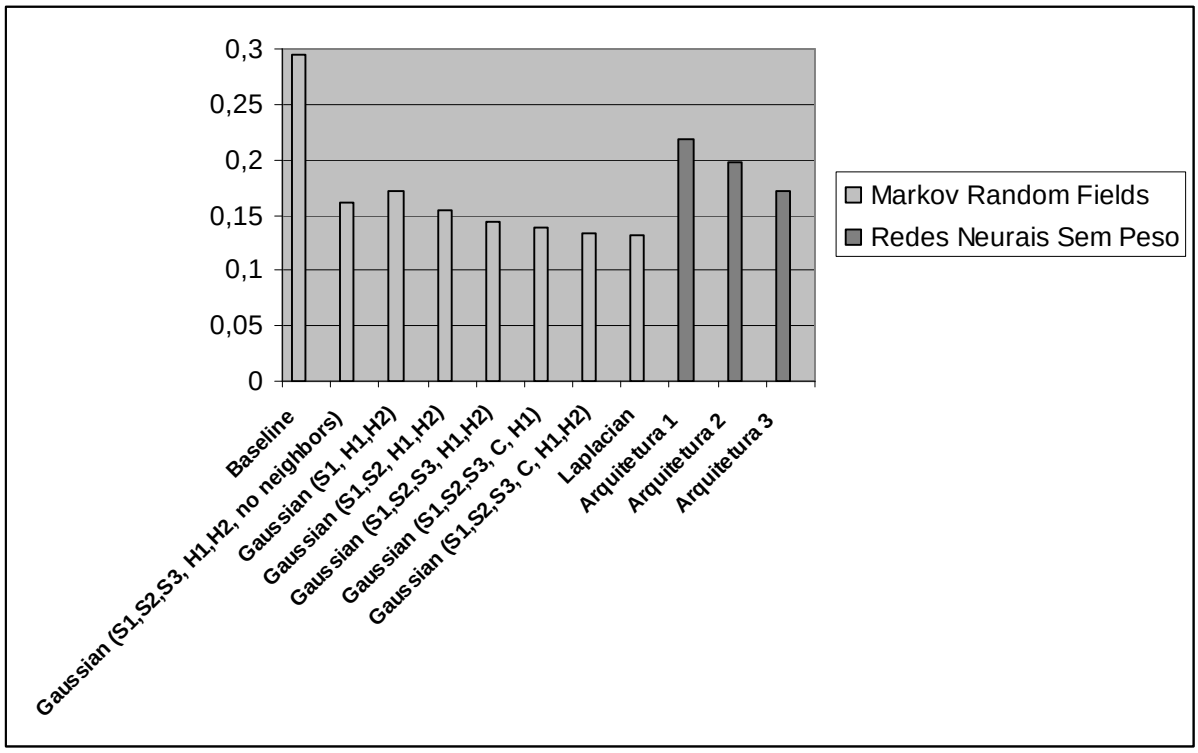
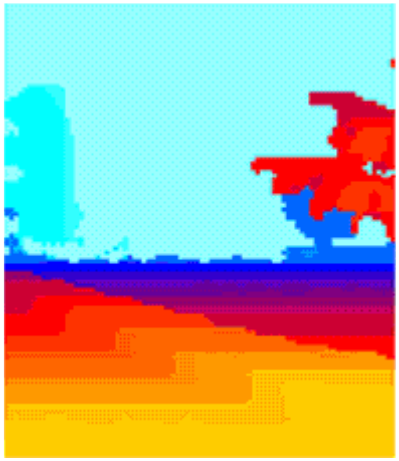
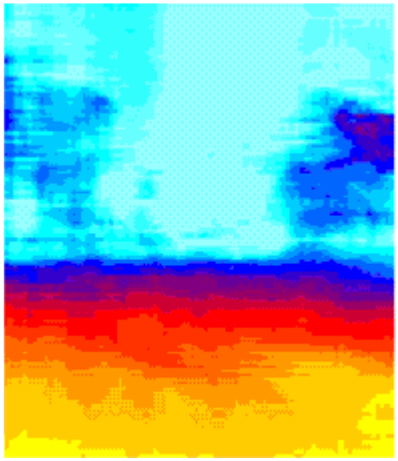


Figura 4-4 – Testes finais de validação das arquiteturas neurais, comparados com os resultados obtidos por Saxena. Eixo X: sistema de estimativa utilizado. Eixo Y: margem de erro, em escala log Mean Absolute Error (MAE).

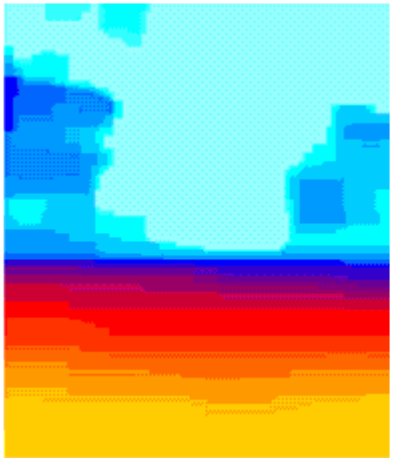
(a) Imagem



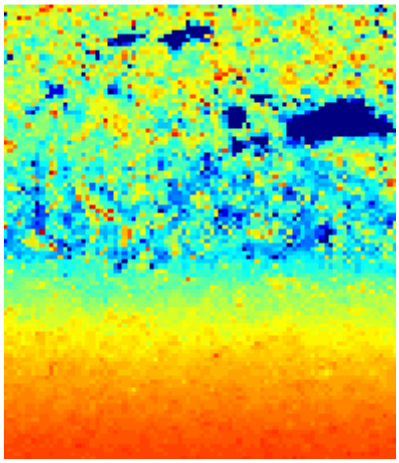
(b) Referência



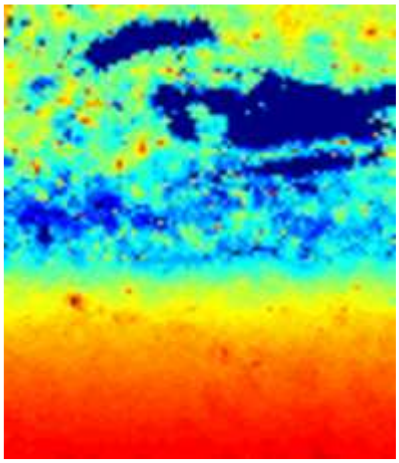
(c) MRF Gaussiana



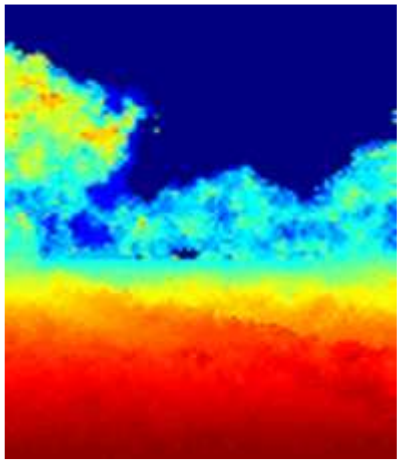
(d) MRF Laplaciana



(e) Arquitetura 1



(f) Arquitetura 2



(g) Arquitetura 3

Figura 4-5 – Comparação entre os resultados obtidos por Saxena e as arquiteturas neurais para uma imagem de teste. (a) Imagem de teste original. (b) Mapa de profundidades de referência. (c) Estimativa gerada pelo sistema MRF gaussiano. (d) Estimativa gerada pelo sistema MRF laplaciano. (e) Estimativa gerada pela arquitetura RNSP 1. (f) Estimativa gerada pela arquitetura RNSP 2. (g) Estimativa gerada pela arquitetura RNSP 3.

Observando os resultados finais, notamos que as arquiteturas neurais conseguem estimar profundidades para áreas próximas no nível do solo com boa precisão, assim como (especialmente a Arquitetura 3) reconhecer os contornos de espaços vazios nas cenas. Entretanto, em comparacão com os resultados obtidos por Saxena, elas demonstram grande dificuldade para reconhecer estruturas verticais, especialmente as mais próximas, e isso compromete seu desempenho.

Não obstante as arquiteturas neurais estudadas não tenham sido capazes de equiparar-se aos sistemas desenvolvidos por Saxena, seus resultados são promissores, e em princípio confirmam a capacidade das RNSP's de estimar profundidades a partir de imagens monoculares. É evidente, entretanto, que arquiteturas mais sofisticadas precisam ser desenvolvidas antes que elas possam rivalizar com sistemas baseados em MRF's.

Entre os mecanismos omitidos nas nossas arquiteturas (e presentes nos sistemas de Saxena) que poderiam contribuir para um melhor desempenho, podemos citar a visualização das imagens de entrada em várias escalas, permitido que os neurônios tenham uma perspectiva mais "global" dos dados de entrada. Opcionalmente, a imagem de entrada poderia ser redimensionada para uma escala com melhor relação sinal/ruído do que a usada atualmente.

Nos sistemas de Saxena, parte do vetor de entrada é dedicado a representar relacionamentos verticais entre elementos da cena. Nas Arquiteturas 2 e 3, entretanto, os neurônios varrem a imagem apenas horizontalmente; portanto, têm pouca oportunidade de aprender padrões relcaionados a estruturas verticais. Modificações quem permitam aos neurônios varrer a imagem verticalmente – talvez apenas duas ou três seções acima e abaixo – poderiam ser a resposta a essa limitação.

Finalmente, é preciso notar que mesmo como está, a Arquitetura 3 não alcançou o ápice de suas capacidades; os testes de ajuste de parâmetros, restritos aos limites de memória da plataforma 32 bits, não chageram a alcançar seu ponto de saturação do desempenho em função do número / dispersão de sinapses. É possível mesmo que, simplesmente portanto essa arquitetura para uma plataforma de 64 bits onde se possa expandir ainda mais o número de conexões por neurônio, uma margem de erro compatível com os resultados de Saxena seja alcançada.

5. DISCUSSÃO

Neste capítulo fazemos a discussão do 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. Em seguida passamos à análise crítica deste trabalho, suas contribuições e deficiências.

5.1. TRABALHOS CORRELATOS

A *framework* MAE, utlizada como plataforma para implementação das redes neurais, é descrita em [OLI05, KOM02]. A inspiração para o uso de uma rede neural sem peso em uma tarefa de reconhecimento visual, assim como a base para a Arquitetura 1, vieram de [ALB08], enquanto a estratégia de usar múltiplas camadas em conjunto com o algoritmo *Winner-Take-All* veio de [OLI05]. As idéias para estender as entradas da Arquitetura 3, assim como o próprio escopo do problema, foram extraídos de [SAX08]. Somos levados a crer que esta é a primeira vez que RNSP's são usadas para estimar profundidades em imagens monoculares, visto que não fomos capazes de encontrar quaisquer referências ao tema em outros trabalhos acadêmicos.

Tradicionalmente, o problema de determinar o mapa de profundidades de uma cena é abordado através do método de *stereo correspondence,* onde as profundidades são estimadas a partir das diferenças de perspectiva observadas em um par de imagens estéreo [SCH02]. Técnicas que utilizam imagens monoculares geralmente se baseiam em sequências de imagens, tais como *optical flow* [BAR94], *structure from motion* [FOR03] e *depth from defocus* [DAS95].

Fora da linha de pesquisa seguida por Saxena, métodos para estimar profundidades a partir de uma única imagem monocular apóiam seu funcionamento em condições bastante específicas. Nagai em [NAG02] estimou profundidades para imagens monoculares de objetos previamente determinados como mãos e faces. Métodos tais como *shape from shading* [ZHA99] e *shape from texture* [MAL97] geralmente assumem a presenca de cores e/ou texturas uniformes, e portanto

tendem a um desempenho ruim em ambientes complexos, como cenas ao ar livre. Hertzmann e Seitz em [HER05] reconstruíram modelos 3D de alta qualidade a partir de várias imagens, mas seu método requer que objetos “assistentes” de dimensões conhecidas estejam presentes na imagem perto do objeto-alvo.

5.2. ANÁLISE CRÍTICA

Certamente a crítica mais grave a este trabalho é o desempenho insatisfatório do sistema, comparando nossos resultados com o de Saxena. Entretanto, esse fato não invalida nossos esforços, visto que não obstante a qualidade inferior, nossos resultados demonstram que RNSP’s são de fato capazes de estimar mapas de profundidades a partir de imagens monoculares. Visto que a decomposição da entrada em elementos de textura, cor e bordas – um processamento que de fato ocorre no cérebro humano – na Arquitetura 3 melhorou o desempenho do sistema, acreditamos que essa falha se deva à ausência, na nossa implementação, de outros mecanismos neurais responsáveis por detectar e refinar estimativas de profundidades. Outra possibilidade é que o cérebro de fato não tenha a mesma capacidade demonstrada pelo sistema de Saxena em estimar profundidades para imagens monoculares – mas podemos descartá-la observando que, se fosse esse o caso, não seríamos capazes de reconhecer cenas em fotografias com tanta clareza.

Uma séria limitação de ordem prática sofrida pelo sistema são seus grandes requisitos computacionais. Mesmo após diversas otimizações para reduzir o consumo de memória e simplificar tarefas de processamento, se forem usadas 256 ou mais sinapses por neurônio em qualquer das três arquiteturas, não é possível efetuar testes de escala relevante (e.g. com mais de 100 casos de treinamento) em computadores com menos de 2GB de memória. Mesmo nesse caso, recuperar um conjunto de estimativas da rede pode levar até mais de 10 minutos, o que certamente inviabiliza o uso da implementação atual em tarefas de tempo-real.

6. CONCLUSÃO

Neste capítulo, apresentamos um sumário do trabalho de pesquisa desenvolvido e, em seguida, discutimos os resultados e conclusões obtidas com o mesmo. Por fim, sugerimos novas linhas de trabalho que podem ser desenvolvidas a partir deste.

6.1. SUMÁRIO

Neste trabalho buscamos investigar – dados os resultados das pesquisas efetuadas por Ashutosh Saxena com sistemas de aprendizado de máquina baseados em *Markov Random Fields* – se haveria arquiteturas de Redes Neurais Sem Peso capazes de estimar mapas de profundidades a partir de imagens monoculares estáticas. Com isso pretendemos desenvolver um sistema com maior plausibilidade biológica para fins de pesquisa, mas que também servisse de base para a criação de soluções de visão artificial para ambientes de produção.

Desenvolvemos três arquiteturas diferentes sobre a plataforma MAE, e também uma aplicação visual para melhor controlá-las e manipular os dados de teste. Em seguida, executamos baterias de testes para avaliar os melhores parâmetros de cada arquitetura, e então executamos testes finais de validação, que nos permitiram comparar nossos resultados com os de Saxena.

6.2. RESULTADOS E CONCLUSÕES

As estimativas de profundidades efetuadas pelo sistema foram comparadas com os mapas de referência produzidos com o auxílio de um *scanner* a laser através de duas métricas: *Mean Absolute Error* (MAE, uma métrica de cálulo e interpretação mais fáceis) e log MAE (utilizada por Saxena, permitiu comparar nossos resultados com os dele). Os resultados obtidos pelas arquiteturas mais sofisticadas (2 e 3) demonstram que o uso de RNSP’s nesse contexto é em princípio viável, e em alguns casos equiparam-se aos melhores rsultados obtidos por Saxena; mas na média ficam muito abaixo do nível de qualidade obtidos por seus sistemas baseados em

MRF. Além do desempenho final insatisfatório, o sistema apresentou requisitos de memória bastante elevados, assim como tempos de processamento que invibializariam a sua aplicação em um contexto de tempo-real.

6.3. TRABALHOS FUTUROS

Embora inferiores aos obtidos por Saxena, os resultados obtidos neste trabalho são promissores, e motivam o desenvolvimento de novos trabalhos para explorar com mais profundidade o uso de RNSP's na estimativa de mapas de profundidades.

Visto que a decomposição da entrada em elementos de textura, cor e bordas – um processamento que de fato ocorre no cérebro humano – na Arquitetura 3 melhorou o desempenho do sistema, conjecturamos que a adição de outros mecanismos neurais responsáveis por fazer e refinar estimativas de profundidade poderiam incrementa-lo ainda mais. Em particular, mecanismos que permitissem à rede "enxergar" a entrada em várias escalas, extensões das entradas de cada neurônio para incluir informações de seus vizinhos, e a implementação de movimentos de varredura vertical sobre a imagem de etrada, poderiam ampliar a capacidade de generalização do sistema.

No sistema de Saxena, as saídas dos vários filtros são agrupados precisamente em função da seção da imagem usada como entrada; em outras palavras, cada vetor de entrada apresenta múltiplas perspectivas da mesma região. Entretanto, na Arquitetura 3, não podemos garantir que cada neurônio observa os mesmos pontos em cada uma de suas entradas. A possibilidade de especificar esse alinhamento, e seus efeitos sobre a capacidade de aprendizado do sistema, ainda devem ser avaliados.

Por fim, pesquisas em novos algoritmos de busca e arquiteturas concorrentes (como a utilização de C+CUDA na implementação dos neurônios) poderiam ser empregadas para reduzir os tempos da fase de estimativa de profundidades. O *port* do sistema para uma plataforma de 64 bits também permitiria o uso de memórias maiores, abrindo assim a possibilidade de estudar configurações com números maiores de camadas neurais e/ou sinapses por neurônio.

7. REFERÊNCIAS BIBLIOGRÁFICAS

[ALB08] DE SOUZA, Alberto F.; BADUE, Claudine; PEDRONI, Felipe; DIAS, Stiven S.; DE SOUZA, Hallysson O.; DE SOUZA, Soterio F. **VG-RAM Weightless Neural Networks for Face Recognition.** Artificial Neural Networks - ICANN 2008, Volume 5163/2008, pp 951-960, 2008.

[ALE09] ALEKSANDER, I.; DE GREGORIO, M.; FRANÇA, F.M.G.; LIMA, P.M.V.; MORTON, H. **A Brief Introduction to Weightless Neural Systems.** ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 Abril 2009.

[BAR94] BARRON, J. L.; FLEET, D. J.; BEAUCHEMIN, S. S. **Performance of optical flow techniques.** International Journal of Computer Vision, 12, 43–77, 1994.

[DAS95] DAS, S.; AHUJA, N. **Performance analysis of stereo, vergence, and focus as depth cues for active vision.** IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(12), 1213–1219, 1995.

[DEA99] DeANGELIS, Gregory C.; NEWSOME, Willian T. **Organization of Disparity-selective Neurons in Macaque Area MT.** The Journal of Neuronscience 19(4): 1398-1415, 1999.

[DOD03] DODGE, R. **Five Types of Eye Movement in the Horizontal Meridian Plane of the Field of Regard.** Am. J. Physiol. 8: 307-329, 1903.

[FOR03] FORSYTH, D. A.; PONCE, J. **Computer Vision: A Modern Approach.** New York: Prentice Hall, 2003.

[GEG97] GEGENFURTNER, Karl R. **Functional Properties of Neurons in Macaque Area V3**. The Journal of Neurophysiol. 77: 1906-1923, 1997.

[HER05] HERTZMANN, A.; SEITZ, S. M. **Example-based photometric stereo: Shape reconstruction with general, varying brdfs.** IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1254–1264, 2005.

[HEY84] VON der Heydt R.; PETHERHANS E.; BAUMGARTNER G. **Illusory Contours and Cortical Neuron Responses**. Science, 224:1260-1262, 1984.

[HUB62] HUBEL, D. H.; WEISEL, T. N. **Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex**. J. Physiol. 160: 106-154, 1962.

[KAN00] KANDEL, Eric R.; SCHWARTZ, James H.; JESSELL Thomas M. **Principles of Neural Science**. 4th Ed. Prentice-Hall International, Inc., 2000.

[KOM02] KOMATI, Karin Satie; DE SOUZA, Alberto Ferreira. **Vergence Control in a Binocular Vision System using Weightless Neural Networks**. In: Proceedings of the 4th International Symposium on Robotics and Automation. Los Alamitos: IEEE, 2002.

[LOO01] LOOMIS, J. M. **Looking down is looking up.** Nature News and Views, 414:155–156, 2001.

[MAL97] MALIK, J.; ROSENHOLTZ, R. **Computing local surface orientation and shape from texture for curved surfaces.** International Journal of Computer Vision, 23(2), 149–168, 1997.

[MATa] MATHER, George. **The Visual Cortex**. [on line] Disponível em http://www.lifesci.sussex.ac.uk/home/George_Mather/Linked%20Pages/Physiol/Cortex.html, último acesso 01/08/2009.

[MIC05] MICHELS, Jeff; SAXENA, Ashutosh; NG, Andrew Y. **High Speed Obstacle Avoidance using Monocular Vision and Reinforcement Learning.** 22nd Int'l Conf on Machine Learning (ICML), 2005.

[MOV85] MOVSHON, J. A.; ADELSON, E. H.; GIZZI, M. S.; NEWSOME, W. T. **The Analysis of Moving Visual Patterns**. In C. Chagas, R. Gattass, C. Gross (eds.), Pattern Recognition Mechanisms. New York, Springer, 117-151, 1985.

[NAG02] NAGAI, T.; NARUSE, T.; IKEHARA, M.; KUREMATSU, A. **Hmm-based surface reconstruction from single images.** IEEE international conference on image processing (ICIP), 2002.

[OLI05] OLIVEIRA, H. **Uma Modelagem Computacional de Áreas Corticais do Sistema Visual Humano Associadas à Percepção de Profundidade**. Dissertação de Mestrado. Programa de Pós-Graduação em Informática, UFES, 2005.

[SAX05] SAXENA, Ashutosh; CHUNG, Sung H.; NG, Andrew Y. **Learning Depth from Single Monocular Images.** Neural Information Processing Systems (NIPS) 18, 2005.

[SAX08] SAXENA, Ashutosh; CHUNG, Sung H.; NG, Andrew Y. **3-D Depth Reconstruction from a Single Still Image.** International Journal of Computer Vision (IJCV), vol. 76, no. 1, pp 53-69, Jan 2008.

[SCH02] SCHARSTEIN, D.; SZELISKI, R. **A taxonomy and evaluation of dense two-frame stereo correspondence algorithms.** IJCV, 47:7–42, 2002.

[ZEK73] ZEKI, S. M. **Colour Coding of the Rhesus Monkey Prestriate Cortex**. Brain Research, 422-427, 1973.

[ZHA99] ZHANG, R.; TSAI, P. S.; CRYER, J. E.; SHAH, M. **Shape from shading: a survey.** IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(8), 690–706, 1999.