Reconhecimento Visual Atencional

Sergio Exel Gonçalves

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO EM ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Prof. Luiz Adauto F. C. Pessoa, Ph.D.

Prof. Luís Alfredo V. de Carvalho, D.Sc.

Prof. Antonio A. F. de Oliveira, D.Sc.

Prof. Antonio Carlos G. Thomé, Ph.D.

Prof. Luciano da Fontoura Costa, Ph.D.

Prof. Paulo Cezar P. Carvalho, Ph.D.

Rio de Janeiro, RJ - Brasil
MARÇO, 1999

GONÇALVES, SERGIO EXEL
Reconhecimento Visual Atencional [Rio de Janeiro] 1999
VII, 127 pp., 29.7 cm, (COPPE/UFRJ, D. Sc., Engenharia de Sistemas e Computação, 1999)
Tese – Universidade Federal do Rio de Janeiro, COPPE
1 – Reconhecimento de Imagens
2 – Representação Multi-escalas
3 – Simulação de Sistemas Atencionais
I. COPPE/UFRJ II. Título (série)

às nossas crianças

Agradecimentos

Ao Luís Alfredo V. Carvalho, orientador inicial nesta jornada, sempre amigo e incentivador. Tê-lo encontrado abriu um novo horizonte na minha busca por estes assuntos fascinantes.

Ao meu lúcido orientador de tese, Luiz Pessoa, por sua direção competente, incentivo e paciência ao longo de toda esta jornada.

Ao meu amigo e competente colega Alexandre Roque, por sua amizade, e cuja ajuda foi essencial para conseguir fazer a infinidade de figuras desta tese e para resolver muitos outros problemas.

À Ana Paula Leitão, por sua amizade, carinho e inestimável ajuda.

A todos meus colegas de laboratório, que juntos formam um ambiente solidário e agradável de trabalho, e especialmente a Márcia Cerioli pela ajuda no LaTeX, Emmanuel Pereira e Vanusa Calegario companheiros das madrugadas e fins de semana de labuta.

Ao pessoal da secretaria, especialmente Cláudia e Solange, sempre atentas e eficientes.

Aos funcionários do laboratório, Carlos Godar e Fred pela ajuda nas horas certas.

À incansável Dona Gersina, por sua presença amiga.

Aos professores Cláudio Esperança e Inês pela ajuda nas horas difíceis da nossa rede.

Ao professor Adilson Xavier, grande incentivador e amigo.

A Cinira e Augusto, por sua amizade.

Aos professores, colegas e amigos do Departamento de Engenharia Mecânica, especialmente aos meus amigos Ricardo Naveiro e José Luis que tanto me ajudaram, incentivaram e aliviaram a carga de aulas.

A minhas ajudantes, Maria da Penha e D. Ivone, cujo trabalho e dedicação são imprescindíveis.

A Angela Vianna, companheira nas horas mais difíceis desta jornada.

Especialmente, e com muito carinho, a Alice Genofre, por incontáveis razões.

E, finalmente, à minha amiga, companheira e amada, que faz tudo isso valer a pena.

Resumo da tese submetida à COPPE como parte dos requisitos necessários para obtenção do grau de Doutor em Ciências (D. Sc.)

Reconhecimento Visual Atencional

Sergio Exel Gonçalves
março, 1999

Orientador: Luiz Adauto F. C. Pessoa
Luís Alfredo V. de Carvalho

Programa: Engenharia de Sistemas e Computação

A visão é um processo ativo, onde informações importantes para o comportamento são seletivamente adquiridas. Nesta tese propomos um modelo de reconhecimento visual atencional no qual uma *fóvea* simulada com alta resolução é dirigida para regiões de *maior interesse* por um processo de atenção seletiva. A região de interesse da imagem bruta é inicialmente representada por um mapa space-variant de respostas das células complexas, ou de contornos orientados. Esta representação é processada por um *Sistema de Decisão* que tenta reconhecer a imagem com as informações parciais adquiridas. Caso a informação disponível seja insuficiente para garantir o reconhecimento, um *Módulo Atencional* entra em ação para determinar a próxima região de interesse de onde extrair informações. A próxima foveação é baseada em informações da imagem (*bottom-up*) e do conjunto de modelos armazenados no sistema (*top-down*) combinadas de acordo com uma *estratégia atencional* adotada pelo sistema. O processo de foveação continua até que a imagem seja reconhecida ou seja atingido um número limite de foveações. Diversas estratégias atencionais foram investigadas, aplicando-se o modelo ao reconhecimento de faces e outros objetos em extensas simulações, que permitem avaliar qualitativamente e quantitativamente os resultados de cada estratégia atencional. Os resultados mostram ser factível reconhecer faces e outros objetos com margens de erros comparáveis ou menores que outros modelos tradicionais, porém a um custo computacional bem mais baixo, utilizando a representação *space-variant* associada a uma *estratégia atencional* adequada.

Abstract of Thesis presented to COPPE/UFRJ as partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

Attentional Visual Recognition

Sergio Exel Gonçalves
março, 1999

Thesis Supervisor: Luiz Adauto F. C. Pessoa
Luís Alfredo V. de Carvalho

Department: Computing and Systems Engineering

Vision is an active process where behaviorally important information is selectively gathered. We present a model of visual recognition in which the high-resolution fovea is deployed to *interesting* regions by selective attention processes. The raw image is initially represented by a space-variant complex-cell, or oriented edge, map. This transformed input is processed by a *decision system* which attempts to recognize the image given *partial* information. In the more typical case where the available information is insufficient to support recognition, an *attentive foveation system* is engaged which is responsible for determining the subsequent image region of interest. The next foveation is based on both bottom-up information from the image and top-down information from the set of stored models, combined according to the attentional strategy in execution. The foveation process continues until the object is recognized or a maximum number of foveations is reached. Several *attentional strategies* were investigated as the model is applied to the task of recognition of faces and other objects. Both qualitative and quantitative evaluations of the atentional strategies are provided. Results show the feasibility of recognition of faces and objects with an error rate comparable or lower to other traditional models, but with much reduced computational costs, a consequence of the space-variant representation used associated with efficient attentional strategy.

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução: o reconhecimento visual humano e a *visão computacional*

A visão humana é um processo ativo, onde um dos componentes mais importantes é um mecanismo de atenção que ajuda a decidir de onde extrair informações. Em uma cena, por exemplo, onde procurar por um determinado objeto, ou, ao visar um objeto, de onde extrair mais informações. A orientação da visão proporcionada pela atenção é essencial, dada a configuração da retina com resolução variável e seu mapeamento para o cortex visual nos mamíferos (Schwartz, 1977 [56]). Enquanto no centro da retina, isto é, na fóvea, forma-se uma "imagem de alta resolução", a periferia somente obtém informação grosseira. Assim a atenção, acoplada a um mecanismo de movimentos oculares rápidos (movimentos sacádicos), compensa o decréscimo de resolução na periferia, e formam uma base poderosa para o comportamento orientado visualmente.

Uma tarefa central desempenhada pela visão é o reconhecimento visual. Alguns autores, como Fermuller e Aloimonos (Fermuller e Aloimonos, 1995 [19]) chegam a argumentar que qualquer problema em visão ou percepção em geral pode ser colocado como um problema de reconhecimento, incluindo nesta categoria até mesmo as tarefas visuais utilizadas para solução dos problemas de navegação (guiar o movimento de um agente dotado de visão) e de manipulação (coordenação manipulador/visão).

No sistema visual humano, o reconhecimento depende críticamente da habilidade de executar os movimentos oculares rápidos, chamados de sacadas, que orientam a fóvea para regiões de interesse em uma cena. Uma questão central é, portanto, determinar o que sejam elementos ou padrões visuais de interesse para onde voltar a atenção. Estes não devem ser vistos como elementos estáticos, pré-definidos, mas sim como dinâmicos e dependentes do contexto. Enquanto em uma situação a cor de um animal pode determinar se ele é amistoso ou nocivo, em outra a forma pode ser o fator decisivo. Assim o reconhecimento visual deve empregar informações correntes, atualizadas a cada momento, para decidir o que é um elemento importante. Para o reconhecimento visual, elementos importantes são aqueles que ajudam a diminuir a ambiguidade do objeto de interesse atual. O melhor dos elementos é aquele capaz de proporcionar a maior discriminação (ou menor ambiguidade) para o processo de

reconhecimento.

Diferenciando-se das tendências mais tradicionais em visão computacional, uma abordagem que vem se desenvolvendo mais recentemente, conhecida como "visão ativa" (" active vision", ou "purposive vision"), procura incorporar o dinamismo da visão humana abordando a visão no contexto das tarefas que um organismo deve efetuar. Enquanto a meta da linha mais tradicional da visão computacional tem proposto " a descrição do mundo tri-dimensional em termos de superfícies e objetos presentes e suas propriedades físicas e relações espaciais" (Black, M.J. in Black et al., 1995 [6]), os defensores da "visão ativa" argumentam que não existe "visão de propósito geral", pois toda atividade visual se dá em um determinado contexto e com determinado objetivo. A proposta desta tendência é "o desenvolvimento de habilidades visuais rápidas ligadas a comportamentos específicos, com acesso à cena visual diretamente, sem representações intervenientes" [6].

Segundo esta perspectiva, o problema do reconhecimento visual em sistemas artificiais se coloca como um processo dinâmico, no qual a imagem do objeto visado não se conserva inalterada durante o processo. Esta imagem pode mudar de várias maneiras e por diversos motivos. Além da possibilidade do deslocamento próprio do objeto, o próprio sistema que processa o reconhecimento pode se deslocar ativamente em relação ao objeto, buscando novas vistas deste, ou, no caso do uso de um dispositivo de aquisação de imagem com resolução variável, como é o caso da retina na visão humana, pode ser produzida uma sequência de imagens correspodentes a diferentes pontos para onde é dirigida a região de maior resolução do dispositivo.

1.1 A proposta deste trabalho

Neste trabalho, introduzimos um modelo de reconhecimento visual atencional com resolução variável, de aplicação geral, isto é, não restrito a um determinado tipo de objeto. O modelo segue a filosofia da visão ativa, e incorpora três aspectos principais: (1) o *reconhecimento incremental* (e.g., como sugerido por Aguilar e Ross [2]), onde informações parciais são fornecidas aos dispositivos de reconhecimento, e novas informações obtidas quando necessário; (2) a adoção de uma *representação em multiplas escalas e "space-variant"* (isto é, com a resolução variando ao longo da representação, decrescendo do centro para a periferia), inspirada na organização do sistema visual dos primatas [7, 58, 56, 57]; e (3) a utilização de *mecanismos atencionais* capazes de orientar o processo de aquisição de informações (como indicado em 1) de forma a guiar o movimento da fóvea para obter informações da imagem em múltiplas escalas (como em 2).

A união destes conceitos permite simular um sistema sacádico que muda seguidamente o ponto de fixação para onde é dirigida a região de maior resolução, processando dinâmicamente as informações assim obtidas, de modo semelhante ao sistema visual humano. Nosso objetivo, no entanto, não é replicar os roteiros de sacadas humanos, mas sim investigar como uma "fóvea" de alta resolução pode efetivamente se deslocar ao longo de uma imagem para focalizar partes relevantes e ignorar as partes menos relevantes. Em nossa investigação, avaliamos diversas *estratégias atencionais* para escolha das regiões de onde extrair informações e os roteiros de sacadas associados a estas estratégias. Resultados preliminares desta nossa

abordagem encontram-se publicados [17, 18, 43, 42, 34, 44]

A Figura 1.1 ilustra a organização macroscópica do modelo de reconhecimento visual atencional. Uma imagem apresentada ao sistema é inicialmente processada de forma a explicitar os contornos presentes. Tal representação intermediária é inspirada no comportamento das células complexas do sistema visual dos primatas. Aqui, a formalização destes mecanismos segue o desenvolvimento original de Grossberg e colaboradores, a qual tem sido empregada em vasta gama de modelos de percepção visual (Grossberg e Mingolla, 1985 [24], Grossberg e Pessoa, 1998 [25], Pessoa et al. 1995 [45]). Este tipo de representação tem sido usada nos últimos anos em um grande número de problemas de visão computacional, incluindo o reconhecimento de objetos (Mel, B., 1997 [37], Rao e Ballard, 1995 [49], Brunelli e Poggio, 1993 [8]), com resultados bastante satisfatórios. Em nosso modelo, a representação de contornos é realizada em múltiplas escalas espaciais, e combinada em uma estrutura *space-variant* simulando uma retina, com alta resolução na "fóvea" e resoluções menores em regiões mais distantes do centro.

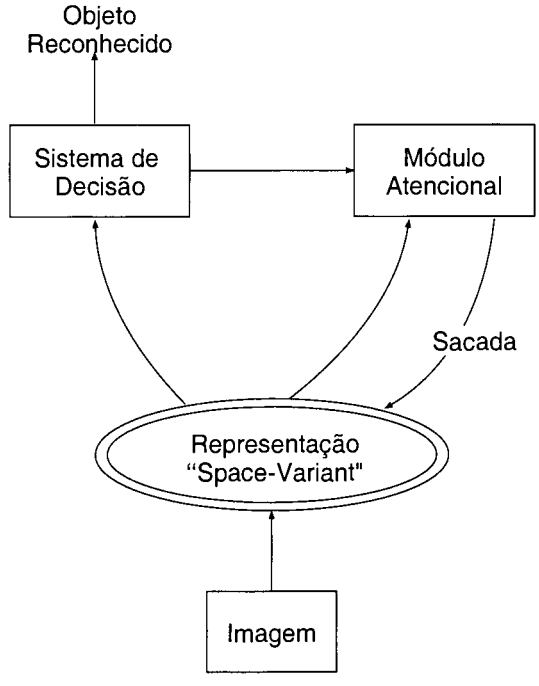

Figura 1.1: **Modelo de Reconhecimento Atencional:** Uma representação *space-variant* centrada em um ponto de fixação inicial é extraída da imagem apresentada para reconhecimento. O Sistema de Decisão compara esta representação com os modelos armazenados no sistema. Caso não seja conseguido o reconhecimento, o Módulo Atencional determina um novo ponto de fixação, isto é, o centro de uma nova representação *space-variant* extraída da imagem. O ciclo se repete até que seja conseguida confiança suficiente no reconhecimento ou um número máximo de "sacadas" seja atingido.

A representação *space-variant* está diretamente associada a um sistema atencional, que é responsável pela determinação de regiões "importantes" na imagem, das quais serão extraídas novas informações para auxiliar no processo de reconhecimento. Diversos sistemas tem empregado estratégias atencionais para determinar

regiões importantes da imagem, e assim, guiar os processos visuais associados. A estratégia mais comum, e também a mais imediata, seria talvez usar a magnitude das respostas de filtros aplicados a imagem para determinar regiões de interesse (Alpaydin, 1996 [3]; Yeshurun e Schwartz, 1989 [68], Rybak, 1998 [54], Aonishi, 1994 [4]). Nesta concepção, regiões importantes são aquelas onde há descontinuidade brusca de luminância, ou bordas. Recentemente, outras estratégias baseadas na imagem (*bottom-up*) tem sido desenvolvidas, tais como operadores de simetria [50], ou detectores de máximos de "curvatura Gaussiana" [30]. Ao mesmo tempo, estratégias que procuram não se limitar aos níveis de luminância da imagem tem proposto usar diferenças e semelhanças dos modelos armazenados para guiar a busca de informações (Aguilar e Ross, 1993 [2]). Na versão mais pura, tais sistemas ignoram inteiramente a imagem de entrada, e, portanto, são denomidados *top-down*. No presente trabalho, argumentamos que, tanto as estratégias *bottom-up* quanto *top-down* são insuficientes por si só. Em uma situação de reconhecimento geral, ou seja, onde uma classe de objetos em particular (e.g., faces) não é assumida, estratégias atencionais *híbridas* são mais efetivas. Em suma, é importante levar em conta tanto a distribuição de luminância da imagem de entrada, como a estrutura dos objetos armazenados na memória. Análises qualitativas, bem como quantitativas, ilustrarão como diversas estratégias atencionais contribuem para a tarefa de reconhecimento, e sugerem como elas podem ser combinadas para um resultado mais efetivo.

No sistema aqui apresentado, informações extraídas de regiões da imagem escolhidas pelo módulo atencional vão sendo acumuladas e usadas para comparação com modelos previamente armazenados, em um processo incremental. Neste processo, novas regiões da imagem são escolhidas e mais informações são extraídas destas enquanto o objeto apresentado não for reconhecido. O processo pára quando as informações já obtidas forem suficientes para definir, com uma certa confiabilidade, que a imagem apresentada ao sistema é uma instância de um dos modelos memorizados (objeto reconhecido), ou quando uma certo número máximo de sacadas foi efetivado sem conseguir a confiabilidade desejada (objeto não reconhecido). A comparação entre as informações extraídas da imagem e os modelos armazenados é baseada em uma medida de correlação entre o vetor que contem as informações já extraídas da imagem a ser reconhecida e os vetores que representam os modelos armazenados. Apenas os componentes dos modelos armazenados correspondentes às informações já extraídas da imagem são utilizados para esta comparação. Os modelos armazenados são vetores sintetizados a partir de um conjunto de imagens de treinamento (ver Pessoa e Leitão, 1999 [44] e Leitão e Pessoa, 1999 [34]). Em resumo, o comportamento do sistema realiza um ciclo onde informações são extraídas de regiões importantes (indicadas pelo sistema atencional), até que o reconhecimento seja possível.

O problema central investigado nesta tese foi o de como selecionar regiões para onde a "fóvea" deve ser dirigida. Esta seleção determina o chamado Roteiro de Sacadas, ou seja, a seqüência de pontos da imagem onde o modelo vai centrar uma representação *space-variant* a fim de obter informações para o processo de reconhecimento. Conforme explorado, tanto do ponto de vista quantitativo quanto qualitativo, esta seqüência influencia decisivamente no desempenho do modelo. Uma seleção adequada pode conduzir rapidamente à discriminação da categoria correta, levando ao reconhecimento com um mínimo de processamento da imagem, pois

só será necessário processar em alta resolução as regiões correspondentes à "fovea" enquanto as regiões periféricas serão processadas em baixa resolução, com grande economia computacional. Por outro lado, uma seleção ineficaz não contribui para a determinação do objeto sendo investigado, pois a informação adquirida não é informativa.

Como mencionado acima, a grande maioria dos modelos atencionais se baseia em estratégias que fazem uso da distribuição de luminância da imagem (*bottom-up*). Ao mesmo tempo, algumas estratégias *top-down* tem sido propostas. No entanto, não se encontra na literatura uma avaliação detalhada de tais estratégias, tanto do ponto de vista qualitativo quanto quantitativo. Desta forma, não é possível avaliar como tais estratégias atencionais de fato contribuem para o processo de reconhecimento, em relação a outras estratégias existentes. A presente tese visa preencher esta lacuna e ao mesmo tempo propor novas estratégias atencionais que, ao serem combinadas, levem a uma performance melhor do sistema.

Em particular, como será mostrado, a utilização de bases de dados com estruturas complementares evidencia as limitações das estratégias *bottom-up* e *top-down* quando usadas isoladamente. Seguindo esta idéia, uma das bases de dados utilizadas foi escolhida por ser muito "densa" no sentido de que os objetos armazenados são bastante semelhantes (e.g. faces), enquanto outra base foi escolhida por ser mais "esparsa", ou seja, os objetos são relativamente distintos entre si (e.g., objetos em 3-D em geral, como um carro e uma xícara de chá).

Em resumo, esta tese apresenta um novo sistema de reconhecimento visual atencional. Enquanto a estrutura geral do sistema é nova, deve ser salientado que diversos dos seus componentes se assemelham a outros propostos na literatura. Por exemplo, a representação de contornos através de filtros inspirados na biologia é amplamente utilizada em sistemas de visão artificial [21, 49, 37, 8]. Ao mesmo tempo, representações *space-variant*, após serem introduzidas nos modelos biológicos por Schwartz (1977 [56] e 1980 [57]), tem sido amplamente utilizados em sistemas de visão ativa [49, 68, 58, 30, 54].

O Sistema de Decisão utiliza uma comparação baseada em uma medida de correlação inspirada na rede neural Fuzzy ART [12], mas outros dispositivos podem ser utilizados, tanto redes neurais (e.g., backpropagation), como métodos tradicionais (e.g., correlação normalizada). Finalmente, algumas das estratégias atencionais exploradas são encontradas na literatura [58, 68, 54, 30, 21]. Neste contexto é importante salientear as principais contribuições desta tese de forma mais explicita:

- Do ponto de vista mais geral, apresentamos a proposta de um sistema de reconhecimento de objetos, com a integração de diversos componentes comumente utilizados em sistemas de visão computacional, mas integrados de forma original (como exibido na Figura 1.1). O sistema que mais se assemelha ao nosso é o de Aguilar e Ross 1993 [2]. No entanto, este sistema não faz uso de uma representação *space-variant*, mas mantem a resolução constante. Aqui utilizamos plenamente uma estrutura *space-variant*, pois a razão essencial de uma estratégia atencional guiando os movimentos "sacádicos" do sensor na imagem é justamente fazer uso otimizado desta estrutura.

- A avaliação qualitativa e quantitativa de estratégias atencionais *bottom-up* e *top-down*, juntamente com a proposta de novas estratégias atencionais

híbridas, as quais efetivamente direcionam a seqüência de aquisições de informação na imagem. A avaliação do desempenho destas estratégias em bases de dados utilizadas por outras abordagens permitiu uma análise comparativa da eficiência do nosso modelo de reconhecimento atencional e dos custos computacionais.

- A utilização de uma estratégia atencional operando numa representação *space-variant*. Tal característica é muito importante pois faz uso de informações em multiplas resoluções para determinar as regiões mais importantes na imagem. Em certos casos, quando os objetos armazenados são mais semelhantes entre si, a informação de alta freqüência espacial pode ser crítica para eliminar a ambigüidade do reconhecimento, enquanto a informação de baixa freqüência pode ser irrelevante. Por outro lado, quando os objetos armazenados são suficientemente diferentes, informações de alta freqüência podem se tornar inúteis (e.g., ruído), enquanto que as informações de baixa freqüência podem ser fundamentais. Outros autores também fazem uso de uma representação *space-variant* num contexto atencional [21, 30, 68].

- A utilização de uma estratégia incremental, onde regiões inteiras são adicionadas ao vetor de entrada a cada fixação. Em contraste, outros sistemas, como o de Aguilar e Ross 1993 [2] adicionam apenas um "pixel" ao vetor de entrada a cada fixação. Por outro lado, o processamento incremental permite que o processo termine muito antes de processar toda a imagem de entrada, pois apenas o mínimo de informação necessário para a tarefa em curso é utilizado, ignorando a grande quantidade de informação redundante característico do processamento de imagens.

- A extensão deste modelo de reconhecimento de objetos para a exploração e reconhecimento de faces em cenas, ou seja, imagens onde diversos objetos (no caso, faces) estão presentes. O sistema para tratamento de cenas inicialmente detecta regiões onde há alta probabilidade de haver uma face. Estas regiões tornam-se então candidatas à investigação detalhada da informação, tarefa que é executada pelo sistema de reconhecimento mostrado na Figura 1.1. Desta forma, o sistema para reconhecimento de um únuco objeto é integrado em um sistema mais genérico de maneira eficaz. Apesar dos resultados apresentados serem preliminares, eles ilustram a possibilidade de estender a presente proposta para a interpretação de cenas.

- Indica alguns caminhos de investigação futura, principalmente no sentido de aperfeiçoar partes componentes do modelo. O Sistema de Decisão, por exemplo, pode ser modificado para "esquecer" informações extraídas no início do processo, quando os mecanismos atencionais ainda não produziram seu melhor efeito. O Módulo Atencional pode incorporar meios de associar de forma mais elaborada as estratégias atencionais para formar as estratégias híbridas. Podem-se também incorporar dispositivos para conseguir invariância das representações usadas no processo de reconhecimento.

A seguir descrevemos a estrutura dos capítulos desta tese. Serão estudados no Capítulo 2 alguns aspectos importantes do problema do reconhecimento, bem como

algumas das principais abordagens e estratégias de reconhecimento descritas na literatura recente. No Capítulo 3 serão descritas detalhadamente cada uma das partes do Modelo de Reconhecimento Atencional incluindo a representação *space-variant*, o Sistema de Decisão e o Módulo Atencional, bem como o funcionamento integrado destes componentes. Um estudo mais aprofundado do funcionamento do Módulo Atencional e das diversas estratégias atencionais é apresentado no Capítulo 4 onde é feita uma analise *qualitativa*. No Capítulo 5 mostramos as simulações feitas com o Modelo de Reconhecimento Atencional para reconhecimento de faces e para objetos em geral, utilizando vários conjuntos de parâmetros e estratégias atencionais, de modo a obter uma avaliação *quantitativa* destas estratégias. O Capítulo 6 apresenta uma discussão deste novo modelo de reconhecimento atencional, incluindo uma avaliação da representação *space-variant* utilizada, uma análise de custos e uma discussão de seu funcionamento e desempenho comparado a outras abordagens, para as diferentes estratégias atencionais. Apresentamos também alguns resultados preliminares da extensão deste modelo atencional para o reconhecimento de imagens em uma cena contendo vários objetos. As conclusões, contribuições deste trabalho, e sugestões para pesquisas futuras estão no Capítulo 7.

Capítulo 2

O Problema do Reconhecimento Visual

O reconhecimento de estímulos provenientes do meio ambiente é uma atividade de suma importância para todos os animais, senão para todos os organismos vivos. Os animais precisam poder oferecer uma resposta adequada aos etsímulos que detectam, no sentido de produzirem um comportamento eficiente para a sobrevivência indivídual e da espécie. Um animal, no entanto, só está equipado para detectar um número reduzido de estímulos entre a enorme diversidade que o ambiente pode produzir, e ele só poderá continuar vivo se puder detectar e reagir aos estímulos que tem significado para a sua sobrevivência. O significado de um estímulo poderia então ser definido funcionalmente, em um nível animal, como o comportamento adequado à sobrevivência em face daquele estímulo. Reconhecer, portanto, seria poder atribuir a um estímulo um significado em termos das atividades básicas necessárias à vida, como alimentar-se, lutar, fugir, reproduzir, etc., e assim poder reagir com o comportamento adequado. Nos animais como os mamíferos, que possuem um sistema visual bem desenvolvido, é através da visão que a maioria dos estímulos importantes podem ser detectados, e por isso o reconhecimento visual ganha uma dimensão fundamental nestes organismos. Dada a relação que se pode estabelecer entre reconhecer e atribuir significado, tanto o reconhecimento de estímulos ou padrões em geral como o reconhecimento visual em particular são fundamentais nos processos cognitivos humanos.

Considerado central em todas os problemas ou tarefas visuais [19] humanas ou artificiais, o reconhecimento visual pode ser compreendido como uma atividade cognitiva que envolve a comparação de um estímulo ou imagem proveniente do mundo exterior com alguma forma de conhecimento que já está presente no sistema ou organismo que faz o reconhecimento. Em termos mais precisos, o reconhecimento compreende: 1) determinação da categoria a que pertence um objeto, o que produz uma *classificação*, 2) avaliação do quanto ele se encaixa em uma categoria, e 3) discriminação entre membros de uma categoria, que é a *identificação* (Liu et al., 1995 [35]). Estas tarefas podem ser referidas como a procura da resposta à pergunta *"o que?"*. Há ainda o problema da busca de um objeto em uma cena, que pode ser relacionado à pergunta *"onde?"*. Alguns autores consideram que o comportamento visual em geral pode ser decomposto em dois comportamentos básicos representados por estas duas perguntas, *"o que?"* e *"onde?"* (Ungerleider e Mishkin, 1982 [64]).

Colocado o problema deste modo, surgem de imediato três questões: como o sistema obtém e armazena o conhecimento, como adquire dados da imagem a ser reconhecida e como faz a comparação com o conhecimento armazenado a fim de obter a associação (ou não) a um objeto ou conjunto de objetos (classe) conhecidos. Estas três questões estão diretamente relacionadas com o modo como o conhecimento ou os objetos são *representados* internamente pelo sistema. Admitindo-se que é possível encontrar uma representação adequada para os objetos conhecidos, estes podem formar uma coleção de modelos armazenados na memória do sistema, e que podem ser comparados com uma representação extraída da imagem a ser reconhecida [1].

As abordagens destas questões diferenciam-se entre as correntes mais tradicionais em visão computacional e a tendência da visão ativa. As abordagens tradicionais, seguindo o paradigma proposto por Marr [36] enfatizaram representações capazes de descrever o mundo tri-dimensional em termos de superfícies, objetos presentes e suas propriedades físicas e relações espaciais [6], uma espécie de *reconstrução* do mundo tri-dimensional. Os defensores da visão ativa, em contraste, enfatizam a tarefa visual a ser executada, e destacam o dinamismo do comportamento visual nos organismos vivos. Segundo esta perspectiva, seriam possíveis outros tipos de representação, não restritos à reconstrução. A modelagem da visão humana baseada em uma retina com resolução variável acoplada a um mecanismo de foveação [2] dá origem aos modelos *atencionais* de reconhecimento visual

Neste capítulo faremos uma revisão dos principais problemas relacionados com as questões apontadas acima e das soluções propostas pelas abordagens tradicionais e pela perspectiva da visão ativa, destacando os modelos atencionais.

2.1 Principais questões sobre o processo de reconhecimento

2.1.1 Representação, reconstrução e dimensionalidade

O problema da representação desempenha um papel central no reconhecimento sendo uma questão importante a distinção entre representação e reconstrução. Edelman e Weinshall [16] enfatizam que, apesar da representação com reconstrução estar de acordo com a tradição de Aristóteles, Hume e Berkeley, e ter sido adotada pela principal corrente da visão computacional nos últimos 15 anos, esta pode não ser a melhor escolha, e apontam para isso três razões. Primeiro, uma representação como esta não é fácil de ser obtida em condições reais sem restrições. Segundo, mesmo que todas as informações necessárias para resolver um dado problema estejam presentes neste tipo de representação, estas podem não estar codificadas da melhor forma, além do que, seria necessário um "homúnculo" para conseguir o reconhecimento, isto é, o problema foi apenas deslocado para dentro do indivíduo ou sistema que

[1] Alguns autores, como Edelman e Weinshall [16] discutem a possibilidade de reconhecimento sem modelos, porém esta possibilidade está restrita a alguns tipos de objetos, ou a processos de classificação e não de identificação.

[2] Mecanismo de foveação é o mecanismo que faz a mudança do ponto da imagem para onde está dirigido o centro da fóvea, que é a região de maior resolução da retina, compreendendo um ângulo sólido de cerca de 3 graus.

deve fazer o reconhecimento, pois nada foi acrescentado no sentido da interpretação da cena [48]. Terceiro, evidencias psicofísicas indicam que a performance humana no reconhecimento não é consistente com a hipótese reconstrucionista. Por outro lado, Locke, no seu *Ensaio sobre o entendimento humano*, sugere que uma idéia representa uma coisa no mundo se ela é naturalmente e previsivelmente evocada por esta coisa, e não necessariamente se a idéia se assemelha à coisa em algum sentido, como os aristotélicos propõem (Edelman e Weinshall, 1994).

Os defensores do paradigma da visão ativa tem proposto outros tipos de representação, capazes de evidenciar certas características presentes na imagem que são úteis para uma determinada tarefa visual. Representações adequadas podem ser indispensáveis para conseguir efetuar alguns passos essenciais para o processo de reconhecimento, como diminuir a dimensionalidade do problema, segmentar a imagem e conseguir algum nível de invariância em relação às transformações que a imagem de um objeto pode sofrer quando visto em condições diferentes. A seguir analisaremos a primeira destas três metas, e na seção seguinte veremos o problema da segmentação e da invariância.

Redução de dimensionalidade

A redução de dimensionalidade é inerente ao processo de reconhecimento. Se reconhecer uma imagem de dimensão N^2 for colocado como o problema de associar a esta imagem um nome, por exemplo, de um conjunto com m nomes de objetos, estaremos diante da tarefa de baixar a dimensionalidade N^2 para m. Esta diminuição pode ser bastante grande já na obtenção de uma representação. Esta, nos estágios iniciais do processamento da visão humana, depende de fatos da neuroanatomia: a informação que sai da retina, por exemplo, pertence a um espaço com cerca de um milhão de dimensões, simplesmente porque o nervo ótico tem este número de fibras. O processamento subsequente da visão humana deve envolver uma importante redução de dimensionalidade, embora a natureza das representações presumidamente de baixa dimensionalidade que suportam o processamento de formas na visão humana não seja conhecida (ver Edelman, 1995 [14]). Se não houvesse redução da dimensionalidade, e o reconhecimento fosse feito pela comparação direta das imagens, ele seria uma busca em um espaço de altíssima dimensão [15], extremamente difícil e custoso. Muitos trabalhos tem tentado encontrar representações que sejam ao mesmo tempo de baixa dimensionalidade e contenham as informações mais importantes para o reconhecimento. Turk e Pentland (M. Turk e A. Pentland, 1991 [62]) propuseram computar os componentes principais de um conjunto de faces e representar cada imagem armazenada e cada nova imagem apresentada por uma combinação linear dos componentes principais do conjunto, que são chamadoas de *eigenfaces*. Deste modo a dimensão da representação é o número de componentes principais adotado para representar o conjunto de faces, e a comparação pode ser feita neste espaço de representações.

Existem três abordagens principais para a representação de formas em visão computacional, conhecidas como pictorial, estrutural e baseada em características. A primeira, pictorial, é a usada por exemplo no modelo de Ullman (Ullman, 1989 [63]), talvez a mais próxima do paradigma reconstrucionista, e visa a construção de uma representação que possa ser comparada com um modelo em 3D depois de um alinhamento adequado. A segunda, estrutural, é a proposta por Biederman [27] em seu

modelo de reconhecimento por componentes, e procura representar partes do objeto e suas relações espaciais. Nesta abordagem há uma diminuição de dimensionalidade porque partes do objeto identificadas na imagem como importantes, tais como cantos, faces ou interseções, são representadas simbolicamente em uma representação da estrutura do objeto. A terceira baseia-se em características extraídas da imagem através de operadores adequados como filtros ou campos receptores, de modo que o estímulo visual é representado por um ponto num espaço de características.

Na abordagem proposta por Rao e Ballard (Rao e Ballard, 1996 [49]), por exemplo, um grande número de filtros é aplicado a um mesmo ponto da imagem, gerando um vetor que contém uma medida multidimensional de propriedades deste ponto. Utilizando esta operação em um número de pontos muito menor que o número de pixels, a representação assim extraída tem dimensão muito menor que a imagem. A representação em multiplas escalas que tem sido utilizada mais recentemente nos modelos de reconhecimento em visão ativa também pertence a este último tipo. Em uma versão mais simples, a representação é construída como uma *pirâmide* de resoluções. Uma *pirâmide* de resoluções é uma estrutura em vários niveis, na qual o nível mais alto, ou "nível zero", contém a imagem em sua resolução original. Nos outros níveis, esta mesma imagem aparece com resolução cada vez mais baixa, de acordo com uma razão de sub-amostragem, que indica qual a diminuição de resolução entre um nível e o seguinte (ver Burt e Adelson, 1983 [10]). Existem muitos outros modelos mais complexos que se assemelham mais de perto a uma retina, como por exemplo a representação *log-polar*, na qual a imagem original é convertida em uma representação onde a resolução varia continuamente, decaindo em direção à periferia. Estes modelos utilizam um *"conformal mapping"*, isto é, uma função complexa que mapeia pontos da imagem em pontos da representação, conservando algumas propriedades importantes, como por exemplo os ângulos entre linhas que se cruzam (para uma revisão ver Bolduc e Levine, 1998 [7]). Alguns destes modelos são implementados em *hardware* [55, 58], proporcionando uma aquisição de imagem rápida e econômica. Estes dispositivos permitem uma redução de dimensionalidade bastante grande, levando a representações com dimensão de cerca de 5% da imagem original [7] ou menores. Este tipo de representação é chamada de "space-variant" porque a resolução com que a imagem é reproduzida na representação varia ao longo desta. Em nosso modelo de reconhecimento, usamos uma simulação simplificada de uma retina, na qual a resolução varia descontinuamente a partir de um disco central de mais alta resolução correspondente à *fóvea*, decaindo em patamares em direção à periferia. Esta representação space-variant, apesar de ser uma simulação apenas aproximada de uma retina, conserva as vantagens proporcionadas pelas múltiplas escalas, ao mesmo tempo que é extremamente simples de implementar. Esta representação será detalhada no Capítulo 3 e seu desempenho avaliado e discutido no Capítulo 6.

2.1.2 Pré-processamento, segmentação e invariância

Uma outra questão diz respeito ao tipo de características que devem ser extraídas da imagem para formar a representação, e isto envolve diretamente a estratégia de pré-processamento adotada. Diversos tipos de pré-processanento podem ser usados para extrair de uma imagem o máximo de informações úteis para a tarefa visual

a ser executada. A questão aqui é saber quais são as propriedades desejadas na representação que podem ser obtidas através do pré-processamento.

Nos modelos que procuram explicitamente a semelhança com o sistema visual humano, são adotados filtros que imitam campos receptores das células da retina e de outros níveis do sistema visual, como por exemplos os campos receptores sensíveis a bordas ou contrastes orientados que imitam as células complexas, capazes de construir uma representação que explicita os contornos existentes na imagem [24, 25, 45]. Alguns destes filtros podem ser agrupados por conexão com unidades de camadas mais elevadas para formar sub-sistemas sensíveis a certas características mais ou menos específicas. Este procedimento pode ser útil para conseguir a *segmentação* da imagem, isto é, a separação desta em regiões significativas, como por exemplo separar um objeto do fundo. A segmentação é muito desejável por ser uma grande ajuda para o processo de reconhecimento mas pode ser considerado um problema tão difícil quanto este, não sendo trivial distinguir as duas coisas. Existem evidências de que algum grau de segmentação é conseguido nos níveis mais baixos do sistema visual humano através da detecção em paralelo de descontinuidades na distribuição de padrões ao longo da imagem [59]. Estas descontinuidades podem ser detectadas ainda durante o processo de filtragem ou pré-processamento da imagem para construir uma representação. O sistema visual humano apresenta uma impressionante habilidade para isso, mesmo sem utilizar a distinção figura-fundo facilitada pela visão estereoscópica.

Os métodos clássicos de segmentação de imagens em visão computacional são: a classificação de pixels, a separação ou conjunção de regiões e a relaxação [5, 53, 41, 69, 26]. Vários modelos de visão computacional, porém, procuram conseguir a segmentação da imagem durante o pré-processamento por outros métodos com diferentes graus de sofisticação. O modelo de Hummel e Biederman [27], por exemplo, agrupa unidades sensíveis a terminações de linhas na primeira camada de uma rede neural por meio de unidades da segunda camada que se tornam então sensíveis a diversos tipos de vértices. Agrupamentos de unidades da segunda camada por sincronização e conexões com uma terceira camada serão sensíveis a formas geométricas elementares chamadas de *"geons"*. A atividade destas unidades pode ser então considerada uma representação destes *geons*. Existem vários tipos possíveis de agrupamentos de elementos da imagem. Um modelo interessante é o de Burbeck e Pizer (Burbeck e Pizer, 1995 [9]), que propõem a detecção de regiões primitivas na imagem, isto é, regiões que pertencem ou configuram um objeto individual pela identificação de *núcleos* ("cores"). Estes núcleos são construídos pela ativação simultânea de campos ou filtros e sensíveis a contrastes de orientações opostas com maior ou menor afastamento. As repostas serão mais fortes ou os *núcleos* mais ativos quanto mais fortes os contrastes e quanto maiores as distâncias entre dois filtros opostos. Esta estratégia procura resolver o problema de isolar a imagem de um objeto individual do contínuo de mudanças de luz que constitui a imagem. Reisfeld, Wolfson e Yeshurun [50] propõem um operador que detecta simetrias baseado apenas nas informações da imagem, isto é, num procedimento *bottom-up*, independente do contexto, procurando também conseguir uma segmentação da imagem.

Invariância

Outra propriedade altamente desejável na representação é a constância ou invariância em relação às modificações que a imagem de um objeto pode sofrer devido a vários fatores como ruído, rotação ou mudanças de pose em relação ao observador, mudanças de distância, de iluminação, etc. A aparência de um objeto tridimensional, isto é, o padrão formado por sua projeção na retina de um olho ou no plano de um filme em uma câmera, depende da posição em que é visto pelo observador. Apesar disso este fato parece ser de importância secundária no caso do reconhecimento feito pelo sistema visual humano, pois este apresenta uma impressionante habilidade de reconhecer um objeto familiar quando visto de uma posição não familiar ou em condições de iluminação diferentes, ou ainda em imagens contendo ruído ou oclusões. Este fenômeno tem sido chamado de constância de forma, por analogia a outras constâncias perceptuais. Encontrar uma interpretação constante de objetos tri-dimensionais diante de condições de visualização (pontos de vista) diferentes tem sido uma das metas da visão computacional. O reconhecimento pode assim ser interpretado como a determinação da constância de forma (ou generalização perceptual), ou seja, identificar estímulos clara e manifestamente diferentes como vistas de um mesmo objeto (Edelman e Weinshall, 1994 [16]).

Muitos modelos de reconhecimento fazem a comparação da representação da imagem a ser reconhecida com modelos previamente adquiridos, também por um pré-processamento que cria uma representação. Esta não é a única abordagem possível, pois existem propostas de *reconhecimento sem modelo*, embora somente aplicáveis em condições muito restritas. No caso mais comum, da comparação com modelos, se fosse possível construir um operador que, quando aplicado a qualquer vista de um dado objeto, produzisse sempre a mesma representação, esta poderia ser usada como modelo, e seria chamada de representação invariante. O reconhecimento seria feito simplesmente aplicando o operador a uma imagem e verificando se a transformação reproduz o modelo. O problema com esta abordagem é que são muito restritos os casos em que isso é possível [16]. O modelo de Hummel e Biederman [27] procura conseguir o reconhecimento através da constância de forma, porque ele tenta construir uma mesma representação estrutural quando atua sobre a imagem de qualquer vista do mesmo objeto.

Uma outra abordagem é procurar uma normalização de pose, de modo a identificar qual a posição em que o objeto está sendo visto e determinar então qual a transformação necessária para obter a mesma posição do modelo. O trabalho de Rybak et al., 1998 [54] apresenta um interessante modelo na linha da visão ativa, onde uma normalização de pose é conseguida durante a extração da representação da imagem.

O uso de uma representação com vários níveis de resolução pode ser bastante interessante para conseguir uma certa invariância, ou robustez, em um sistema de reconhecimento que seja resistente a deformações em relação aos modelos armazenados (ver Fukushima, 1994 [20] e 1982 [22]). Entretanto, esta possibilidade nos remete a um dilema: o reconhecimento envolve a solução do problema de conseguir que imagens diferentes de um mesmo objeto sejam associadas à representação (ou modelo) deste objeto previamente armazenado no sistema. Por maior que seja o número de imagens (instâncias) de um mesmo objeto usadas para construir um (ou vários) modelos deste objeto, sempre haverá, em princípio (quando não é possível encontrar uma representação invariante, como na maioria dos casos), uma diferença

entre a instância particular do objeto apresentada ao sistema para reconhecimento e o modelo armazenado. Mas esta diferença aparece com menor intensidade nos niveis de menor resolução, onde as imagens são mais *"borradas"*. Por outro lado, o reconhecimento só é possível se houver suficiente diferença entre a representação do objeto e os modelos dos outros objetos. Ora, nos níveis de menor resolução, a diferença entre o modelo correto e os outros modelos também é menor, dificultando a discriminação do modelo correto. Por outro lado, nos níveis de maior resolução o problema se inverte, pois é maior a discriminação entre os modelos, mas também é maior a diferença entre a representeção do objeto e o modelo correto.

2.2 Modelos tradicionais de reconhecimento

Tradicionalmente os modelos de reconhecimento procuram resolver o problema utilizando um dos três tipos de representação descritos na seção anterior, ou seja, representações pictoriais (na linha da reconstrução), ou estruturais (como no reconhecimento através de componentes estruturais extraídos da imagem), ou baseadas em características. Nestes modelos, as imagens são dados estáticos, isto é, não mudam durante o processo de reconhecimento, a não ser por causa do movimento próprio dos objetos a serem reconhecidos, movimento este que é um dado do problema e não um recurso do próprio processo para obter mais dados. Além disso toda a imagem deve ser processada, e não há seleção de regiões mais importantes da imagem simulando um processo atencional.

No modelo de Ullman [63], por exemplo, o reconhecimento se baseia em um esforço para alinhar corretamente a representação do objeto a ser reconhecido com o modelo, necessitando para isso de uma representação rica o suficiente para que a forma do objeto possa ser recuperada a partir da imagem. Esta abordagem está de acordo com a linha da reconstrução, onde se procura reconstruir um objeto em 3D antes de processar o alinhamento. As tentativas de economizar processamento se dirigem para restringir o espaço de transformações onde buscar o alinhamento ou restringir o processo de reconstrução a apenas alguns *pontos de ancoragem* [16]. Já no modelo de Hummel e Biederman [27], comentado anteriormente, procura-se uma reconstrução da estrutura do objeto, através da identificação dos *"geons"*. Aqui tambem há o processamento de toda a imagem.

Um dos mais importantes entre os modelos tradicionais utiliza o terceiro tipo de representação: é o proposto por Turk e Pentland [62], já comentado, que usa uma representação de baixa dimensionalidade construída a partir da extração dos componentes principais do conjunto imagens que servem de modelos. Apesar da simplicidade da representação usada para comparação, é necessário processar toda a imagem de entrada para que ela possa então ser representada como uma combinação linear dos componentes principais. Neste modelo, o conjunto de imagens de treinamento, isto é, aquelas que vão formar o conjunto de modelos "conhecidos" pelo sistema, é processado como um conjunto de vetores de um espaço de alta dimensionalidade (a dimensão deste espaço é o número de pixels das imagens). Deste conjunto de vetores extrai-se um conjunto de "componentes principais", isto é, um conjunto de vetores que vai formar uma *base* para o espaço vetorial de alta dimensão. Assim, cada imagem pode ser representada por uma combinação linear dos vetores

desta base. Turk e Pentland usaram este método em um conjunto de imagens de faces, e os vetores desta base, os componentes principais, podem ser interpretados também como faces, e por isso foram chamados de"eigenfaces". O número de vetores desta base, ou número de "eigenfaces", pode ser muito menor que a dimensão das imagens, pois as componentes principais representam justamente aquelas "direções" no espaço de alta dimensionalidade nas quais ocorrem as maiores variações entre os membros do conjunto de treinamento. Deste modo, usando apenas um número restrito de componentes principais se consegue descrever as características mais importantes do conjunto original. Esta descrição é aproximada [3], mas, para fins de reconhecimento, não é preciso ter a descrição exata, mas apenas a descrição aproximada que contenha o número de dimensões mínimo necessário à discriminação entre as imagens. A imagem a ser reconhecida é projetada no subespaço das "eigenfaces" e representada como uma combinação linear destas, ou seja, um vetor que tem a dimensão do número de "eigenfaces" adotado para representar o conjunto de treinamento. O reconhecimento se faz por comparação deste vetor com os vetores que representam os modelos no subespaço das "eigenfaces".

Esta abordagem procura, na verdade, conseguir uma redução de dimensionalidade do problema de reconhecimento, e isto se revela possível porque as imagens, ao mesmo tempo que veiculam uma enorme quantidade de informações, trazem também muita informação redundante ou inútil para uma tarefa visual como o reconhecimento. Desta forma, considerar um certo número de componentes principais equivale a ignorar as dimensões que veiculam informações irrelevantes.

2.3 Modelos Atencionais

A possibilidade de baixar a dimensionalidade do problema de reconhecimento é, como vimos, uma conseqüencia do fato de que as imagens veiculam grande quantidade de informações redundantes ou inúteis para esta tarefa visual. Na verdade, os sistemas visuais dos mamíferos utilizam também os mecanismos da atenção visual para diminuir a dimensionalidade da tarefa. Nas seções seguintes, faremos uma exposição resumida dos mecanismos neurobiológicos da atenção e mostraremos como estas características tem sido modeladas em alguns sistemas artificiais.

2.3.1 A atenção visual

O fenômeno da atenção pode ser estudado do ponto de vista da sua intensidade, isto é, dos distintos graus de clareza com que o sujeito percebe o mundo ou parte dele, ou do ponto de vista da seletividade, isto é, de que (e como) objetos se tornam relevantes em detrimento de outros. A atenção seletiva, segundo William James, ("The Principles of Psychology" [29]), "... é a tomada de posse pela mente, de forma clara e vívida, de um dos que parecem ser vários possíveis e simultâneos objetos ou caminhos do pensamento. Focalização e concentração da consciência é da sua essência. A atenção implica afastar-se de algumas coisas para lidar efetivamente com outras". Ainda de acordo com William James, a atenção seletiva pode

[3]Note-se que, num caso extremo, se fossem usadas tantas componentes quanto o número de dimensões do conjunto original, esta descrição seria exata.

ser sensorial ou intelectual; imediata ou derivada; passiva ou ativa. No nosso caso, o que nos interessa é a atenção visual seletiva sensorial, seja passiva imediata, seja ativa por seleção de uma localização espacial (Rizzolatti et al., 1994 [52]). A orientação passiva da atenção ocorre nos casos em que um estímulo atrai a atenção do indivíduo por suas propriedades intrínsecas ou pelo modo como é apresentado. A orientação ativa provém do próprio sujeito e é caracterizada por um esforço para aumentar a clareza de um dado estímulo externo. Alguns experimentos mostraram que estímulos periféricos apresentados abruptamente causam orientação passiva da atenção, enquanto estímulos apresentados centralmente e que devem ser interpretados para influenciar a atenção causam orientação ativa da atenção [52]). Diversos experimentos, tanto psicológicos como neurofisiológicos, mostram que a atenção é largamente independente do ponto de fixação, podendo ser alocada voluntariamente em qualquer região do campo visual e ainda em qualquer profundidade (Gawryszewski et al., 1987 [23]). Pode também se deslocar involuntariamente para qualquer região e profundidade do campo visual [23]. Por outro lado, há também evidências experimentais de que a atenção pode se encontrar em dois estados: concentrada em um ponto ou pequena região do espaço ou difusa, isto é, não focalizada em nenhum ponto em particular (Posner e Cohen, 1984 [46]). Este segundo estado parece corresponder a situação em que o sistema oculomotor está pronto para que uma sacada (movimento ocular rápido) possa ser *programada.* A atenção focalizada em um ponto corresponderia a uma sacada programada e pronta para ser *executada*, segundo a *"Teoria pré-motora da atenção"* de Rizzolatti e colaboradores [51].

Outras evidências experimentais indicam que os estímulos visuais competem pela atenção [52]. Além disso, esta competição é modulada por centros corticais que podem dirigir a atenção para uma determinada região, facilitando assim a detecção de estímulos nesta região e dificultando a detecção de estímulos em outra região. O sujeito pode dirigir voluntariamente a atenção para um determinado ponto, e, se aparecer um estímulo suficientemente intenso em outro ponto, a atenção pode se deslocar para o novo ponto e uma sacada para este outro ponto pode ser realizada ou não. O padrão de inibições e facilitações entre as diversas regiões do campo visual e sua relação com a atenção mostra que há uma competição entre processos *"bottom-up"*, que seriam devidos às características dos estímulos visuais, e processos *"top-down"*, voluntários, provenientes de centros corticais [52]. Os mecanismos *"bottom-up"* são chamados também de processos pré-atencionais, que operam através de sistemas inatos, com alto grau de paralelismo, rapidez, e independentes de esforço consciente. Os mecanismos *"top-down"* são baseados na memória ou de natureza cognitiva, e são os processos dependentes do contexto, ou da tarefa, e requerem atenção seletiva, operando de forma sequencial, mais lenta e com esforço consciente [59, 30].

Como vemos, os processos atencionais (no caso da visão) podem ser considerados mecanismos de alocação de recursos de processamento a uma dada região da cena visual, permitindo o processamento desta região com maior eficiência, enquanto outras regiões são deixadas em segundo plano ou ignoradas. Nos animais que tem um sistema visual com uma retina de resolução variável, máxima na região central, ou fóvea, e decaindo em direção à periferia do campo visual, a associação deste sistema visual com os mecanismos atencionais proporciona um sistema altamente eficiente. Este sistema se completa com um "sistema sacádico", ou seja, um mecanismo que

produz e controla os movimentos rápidos dos globos oculares chamados de "sacadas". Estes movimentos posicionam os olhos de modo que a imagem da região de maior interesse na cena visual se forme na região de maior resolução, ou fóvea, e por isso são também chamados de movimentos de foveação. Estes movimentos são muito rápidos, podem ocorrer com intervalos de cerca de 300 milisegundos, e durante seu transcurso o movimento da imagem na retina não é percebido. No nosso dia-a-dia, parece natural "prestar atenção" à região da imagem que está projetada sobre a fóvea, ou seja, ao centro da visão. Entretanto, para que isso aconteça, é preciso que a região de interesse seja detectada anteriormante em um outro ponto, que pode estar muito distante da fóvea, numa região vista com menor resolução. Uma vez detectada esta nova região de interesse, pode haver um movimento sacádico. A "Teoria Pré-Motora da Atenção", segundo Rizzolatti e colaboradores [51], argumenta que dirigir a atenção para um ponto equivale a programar uma sacada para este ponto e esta sacada pode então ser realizada ou inibida voluntáriamente.

Os sistemas visuais organizados desta forma podem então realizar o processamento detalhado apenas das regiões de maior interesse na cena visual, evitando processar regiões com informação redundante ou inútil. Para isso articulam os três sistemas apresentados acima: o sistema atencional, que focaliza uma região de interesse, a retina com resolução variável, que extrai informação detalhada desta região, e o sistema sacádico, que desloca os olhos de modo que a região de interesse seja projetada sobre o centro de maior resolução. Desta forma, as regiões de menor interesse serão processadas apenas com baixa resolução, sem ocupar demasiadamente os recursos de processamento que são necessariamente limitados. Nestes termos, o maior problema a ser resolvido é o de como escolher adequadamente quais as regiões que merecem ser processadas em detalhe, ou seja, para onde dirigir a fóvea de forma a extrair informações detalhadas de uma região, em detrimento de outras regiões da cena visual consideradas de menor interesse.

Diversas tentativas tem sido feitas de modelar esta organização do sistema visual para executar as tarefas visuais de reconhecimento e busca de objetos em uma cena visual. Um problema em aberto, porém, é o de como selecionar as regiões de interesse na cena visual, modelando os mecanismos da atenção. Alguma luz tem sido lançada sobre os mecanismos atencionais bottom-up, ou pré-atencionais [60, 31], entretanto, o controle top-down da atenção está relacionado com processos cognitivos de alto nível, que são pouco compreendidos e fracamente formalizados [54]. Neste contexto, a modelagem de estratégias atencionais deve ser mais investigada, pois indica como a informação pode ser extraída incrementalmente evitando o processamento de toda a imagem, o que é extremamente custoso e, além disso, desnecessário, dado o caráter redundante das imagens.

2.3.2 Revisão dos principais modelos de reconhecimento atencional

Os vários modelos de reconhecimento atencional encontrados na literatura distinguem-se pelo modo como propõem soluções para diversos problemas do processo de reconhecimento esquematizados acima. Alguns trabalhos se concentram mais na investigação das representações adequadas, outros enfatizam o caráter in-

cremental do reconhecimento, e outros ainda procuram "reconstruir" o objeto a partir de várias sacadas realizadas com um sensor space-variant, e então classificar ou reconhecer a imagem. Uma ordenação destes modelos ou uma classificação destas abordagens está além do escopo deste trabalho, e nos limitaremos a expor alguns dos principais modelos de reconhecimanto atencional importantes para situar o nosso trabalho entre as diversas linhas de investigação na perspectiva do reconhecimento atencional (ver tabelas 2.1 e 2.2).

Autores	representação	incremental	estratégia atencional	tarefa	processo
Ullman [63]	P	Não	–	R	alinhamento
Hummel [27]	E	Não	–	R	reconstrução
Turk [62]	C	Não	–	R	PCA
Lawrence [33]	C	Não	–	R	SOM + CN

Tabela 2.1: **Modelos tradicionais.** C = vetor de características, E = estrutural, P = pictorial; R = reconhecimento; PCA = análise de componentes principais; SOM = "self organizing map", CN = rede convolucional.

Autores	representação	incremental	estratégia atencional	tarefa	processo
Rao [49]	C	Sim	*top-down*	R, B	correlação
Tistarelli [58]	SV	Sim	fixa	R	PCA
Aguilar [2]	MR	Sim	híbrida	R	FuzzyART
Rybak [54]	SV	Sim	*bottom-up*	R	o que?/onde?
Fukushima [21]	MR	Sim	híbrida	C	reconstrução
Jansen [30]	SV	Sim	híbrida	R	o que?/onde?
Este modelo	SV	Sim	híbrida	R	correlação

Tabela 2.2: **Modelos Atencionais.** C = vetor de características,MR = multiresoluções, SV = *space-variant*; B = busca, C = classificação, R = reconhecimento; PCA = análise de componentes principais; SOM = "self organizing map"; CN = rede convolucional.

Visão ativa baseada em representações icônicas

O trabalho de Rao e Ballard (1995 [49]) apresenta um modelo de visão ativa capaz de realizar tarefas de busca e reconhecimento em um conjunto de imagens de objetos representados por um conjunto de vetores de características. O pré-processamento utiliza filtros Gaussianos com 9 orientações e 5 escalas, gerando uma representação constituída por um vetor de 45 componentes para cada região próxima de um ponto na imagem de um objeto. Os objetos são representados pelos vetores correspondentes a pontos escolhidos usando o centróide do objeto e outros pontos

a certas distâncias e orientações a partir do centróide ou outra estratégia baseada em uma medida de saliência de pontos na imagem do objeto. O reconhecimento é feito por comparação dos vetores extraídos dos pontos da imagem com o conjunto de vetores que constituem a representação de cada objeto através de uma métrica baseada no produto escalar normalizado. A busca, ou localização de objetos, é feita por um processo atencional que utiliza uma *"imagem de distâncias"* construída para cada modelo por comparação com os vetores extraídos de todos os pontos da imagem.

O ponto central investigado neste trabalho é a capacidade desta representação produzir um vetor de características de dimensionalidade elevada para cada um dos pontos da imagem em que é extraído, de tal modo que apenas um número reduzido de pontos da imagem de um objeto pode servir para representá-lo de maneira robusta. O resultado líquido é uma representação de dimensionalidade mais baixa que a imagem, capaz de manter certo nível de invariância à rotações no plano da imagem.

Este modelo de reconhecimento foi aplicado aos objetos da base de dados da Universidade de Colúmbia [4] que também foi utilizada por nós em muitas simulações, principalmente nas simulações para avaliação quantitativa que vamos mostrar no Capítulo 5.

Reconhecimento ativo em um *espaço de faces*

No trabalho de Tistarelli (Tistarelli, 1995 [58]) é utilizado um sensor *space-variant* capaz de adquirir imagens com resolução variável, conseguindo vistas com apenas 8, 8% da extensão da imagem original. Este dispositivo é aplicado em um modelo de reconhecimento que usa a análise de componentes principais (PCA) proposta por Turk e Pentland para o reconhecimento de faces. Após a aquisição de 3 vistas de 7 sujeitos, são extraídos os componentes principais deste conjunto de faces, e novas vistas destes sujeitos, com variações de expressão e pose, são apresentadas como imagens de teste. O modelo reconhece as novas vistas com bastante sucesso, sendo investigado o aspecto incremental deste reconhecimento: uma nova vista dos sujeitos de teste só é utilizada se for necessário. Outro ponto investigado é a estratégia de captação das vistas, que pode ser interpretada como uma estratégia atencional, onde pontos préviamente considerados importantes serão eleitos para centro das representações *space-variants* extraídas. Foi observado que o maior número de acertos no reconhecimento era obtido quando os pontos escolhidos para centro das vistas eram os pontos de uma região central nas faces.

Reconhecimento de faces usando extração incremental de características

O trabalho de Aguilar e Ross (Aguilar e Ross, 1993 [2]) usa uma adaptação da rede ART (de *" Adaptive Resonance Theory"*) proposta por Carpenter e Grossberg [12, 11] para realizar o reconhecimento através de uma comparação incremental de características extraídas da imagem com modelos armazenados no sistema. As representações utilizadas neste sistema são vetores de características obtidos por um pré-processamento com filtros de Gabor de 4 orientações. Foram utilizadas duas escalas ou resoluções para simular a variação de resolução de uma retina. Assim, cada modelo memorizado consiste em um vetor contendo os valores computados

[4] Uma parte desta base de dados, que foi utilizada inicialmente por Murase e Nayar [38] está mostrada no Capítulo 5.

pelos 4 filtros em todos os pontos da imagem para representar à resolução mais alta, e para um número de pontos 16 vezes menor para representar a resolução mais baixa. Durante o reconhecimento, o "matching" é computado inicialmente usando apenas a resolução mais baixa, comparando a representação obtida da imagem com os modelos ativos representados por unidades da última camada da rede ART. Um limiar é estabelecido para identificar unidades cuja ativação é suficientemente alta para torná-las candidatas ao reconhecimento, e um outro limiar define se o reconhecimento foi conseguido. Caso não haja reconhecimento com as informações disponíveis, novas características são extraídas da imagem, desta vez usando a resolução mais alta, o que corresponde a uma visada em um novo ponto da imagem. O novo ponto a ser visado é escolhido por um processo de análise para determinar qual o ponto da imagem tem maior poder de discriminação entre os modelos candidatos, sendo por isso um processo baseado principalmente nos modelos, mas que leva em consideração as ativações das categorias, e estas são resultantes do processo de comparação das características extraídas da imagem com as características dos modelos. Este sistema foi testado para o reconhecimento de uma face em um conjunto de 12 exemplares, a partir de imagens com ruído, conseguindo o reconhecimento usando as 64 características extraídas com resolução baixa e mais 15 regiões de um total de 1024 com resolução alta.

Neste modelo, não é usada uma representação propriamente *space-variant*, mas dois valores de resolução estendidos a toda a imagem, e as características extraídas a cada sacada correspondem aos 4 filtros aplicados a apenas um pixel da imagem em resolução alta. É utilizada uma única estratégia atencional, baseada nos modelos, mas usando a ativação das categorias correspondentes para ponderar a contribuição de cada uma no cálculo do ponto de maior poder de discriminação. Este cálculo é feito para todos os pontos da representação de baixa resolução.

Este trabalho é o que mais se aproxima da nossa abordagem, principalmente pelo caráter incremental. Em nosso modelo de reconhecimento, porém, utilizamos uma representação realmente *space-variant*, e testamos várias estratégias atencionais diferentes, comparando-as inclusive com a estratégia proposta por Aguilar e Ross.

Reconstrução atencional

Um certo número de modelos de reconhecimento procura utilizar estratégias atencionais associadas a uma representação *space-variant* para encontrar na cena visual regiões ou características importantes de algum tipo de objeto, e assim fazer uma espécie de reconstrução que permita o reconhecimento ou classificação. Os objetos que mais se prestam a esse tipo de abordagem são faces, pois as características importantes destes objetos são bem definidas.

O trabalho de Rybak et al., 1998 [54] apresenta um interessante modelo na linha da visão ativa, com representação space-variant, onde, durante o pré-processamento são extraídas características da imagem em diferentes níveis de resolução, e as características extraídas em resolução mais baixa (por isso abrangendo uma região maior) são usadas para estabelecer um *contexto* para as características extraídas em alta resolução. Este *contexto* é usado para encontrar uma direção principal que define a posição do objeto. A representação é então construída em um referencial *do objeto*, tornando possível a comparação com o modelo, cuja direção de referência é conhecida. Apesar de reconhecer a importância dos mecanismos *top-down* de con-

trole atencional, este modelo se baseia apenas em características da imagem para a escolha dos pontos importantes para formar a representação. A invariância é buscada através da determinação da pose do objeto encontrada a partir da determinação da direção principal e portanto da posição do referencial do objeto. Este trabalho segue o "paradigma comportamental", que considera que a representação interna de um objeto no sistema visual humano é uma seqüência de registros sensoriais e motores, ou seja, uma cadeia alternada de características do objeto e traços dos movimentos oculares necessários para levar à próxima característica [39]. Este paradigma procura modelar o comportamento de duas importantes vias do sistema visual humano, que são: a que leva ao córtex parietal, envolvida no processamento da informação espacial (conhecida como *"where pathway"*), e a via que leva ao córtex infero-temporal, envolvida na representação das características dos objetos (conhecida como *"what pathway"* [64]). Nesta perspectiva, isto é, procurando modelar comportamentos *"where?"* e *"what?"*, também se situa o trabalho de Rao e Ballard, referido acima.

Outros trabalhos, como o de Fukushima e Hashimoto, 1997 [21], ou o de Jansen, 1996 [30], procuram utilizar tanto a informação *bottom-up* quanto a informação *top-down* em uma estratégia atencional associada a uma representação *space-variant*, para localizar características importantes de faces em uma cena, e assim determinar se o objeto presente é ou não uma face. No caso do trabalho de Fukushima, as regiões de onde foram extraídas informações com mais alta resolução, se confirmada a expectativa de que ali deveria se encontrar determinada característica, são representadas em alta resolução em um mapa. Este, ao longo do processo, vai sendo preenchido por estas informações de modo a reconstruir uma imagem conhecida, caso ela esteja presente em alguma posição na cena.

Capítulo 3

Descrição do Modelo

O Modelo de Reconhecimento Atencional será apresentado em detelhes neste capítulo. A Figura 3.1 mostra um diagrama do modelo, onde aparecem seus principais componentes.

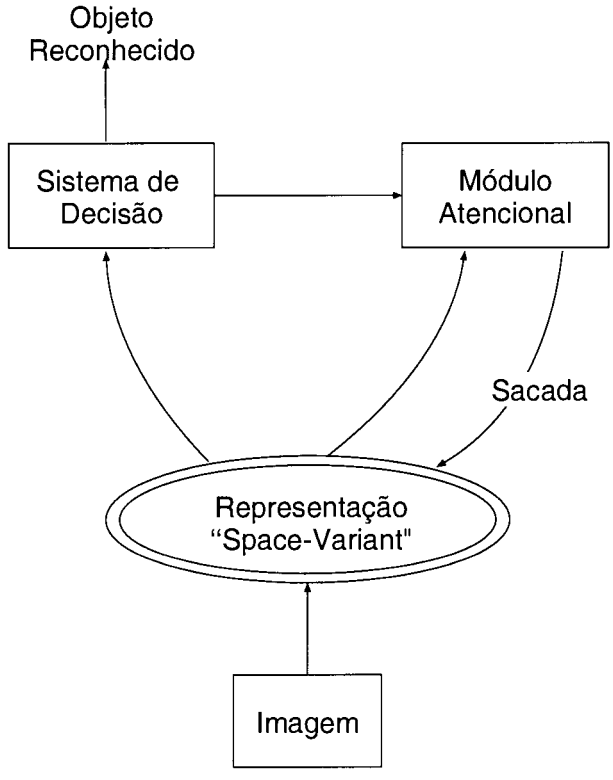

Figura 3.1: **Modelo de Reconhecimento Atencional:** Uma representação *space-variant* centrada em um ponto de fixação inicial é extraída da imagem apresentada para reconhecimento. O Sistema de Decisão compara esta representação com as representações *space-variants* dos modelos armazenados no sistema. Caso não seja conseguido o reconhecimento, o Módulo Atencional determina um novo ponto de fixação, isto é, o centro de uma nova representação *space-variant* extraída da imagem. O ciclo se repete até que seja conseguida confiança suficiente no reconhecimento ou um número máximo de "sacadas" seja atingido.

Neste modelo, o processo de reconhecimento é incremental e se baseia em uma

representação *"space-variant"* da imagem a ser reconhecida. Esta representação simula um mapa de resolução variável das respostas das células complexas do sistema visual humano, correspondente a uma sacada para um determinado ponto da imagem, para onde se dirige o centro da *"fóvea"* simulada. A obtenção desta representação será detalhada mais adiante, nas Seções 3.1 e 3.2. A informação extraída da imagem em uma sacada, através desta representação, é processada pelo *Sistema de Decisão* para tentar o reconhecimento. Esta informação adquirida em uma sacada é sempre parcial, pois apenas a pequena região central da representação, correspondente à *"fóvea"*, tem resolução máxima; o resto da representação tem outros níveis menores de resolução, diminuindo em direção às bordas. O *Sistema de Decisão* tentará processar o reconhecimento com esta informação parcial, por comparação com as respresentações dos modelos armazenados no sistema. No caso mais típico, no qual a informação inicial disponível é insuficiente para o reconhecimento, o *Módulo Atencional* passa a trabalhar para determinar outra região de interesse na imagem, orientando uma sacada para um novo ponto onde será localizado o centro da *"fóvea"* e extraída uma nova representação *space-variant* da imagem. Este *Sistema de Decisão* incremental está detalhado mais adiante, na Seção 3.3. Para determinar a orientação da próxima *sacada*, o *Módulo Atencional* pode utilizar vários tipos de informações, tais como informações da imagem (*bottom-up*) ou informações dos modelos armazenados no Sitema de Decisão (*top-down*), segundo diversas estratégias a serem detalhadas no Capítulo 4. Uma vez determinado o novo ponto de fixação, e extraída uma nova representação como a descrita acima, o processo de decisão recomeça. Este ciclo (decisão - sacada - decisão) se repete até que um critério de decisão seja alcançado, e o objeto reconhecido, ou se esgote um número máximo de sacadas e não ocorra uma decisão. Na Seção 3.4 descrevemos resumidamente o *Módulo Atencional*; entretanto, dada a importância deste módulo para a investigação das estratégias atencionais, ele será abordado com mais profundidade no Capítulo 4. O modelo tem diversos parâmetros de operação que afetam seu desempenho, e algumas medidas devem ser definidas para permitir avaliá-lo; na Seção 3.5 são apresentados os parâmetros do modelo e definidas algumas medidas de desempenho.

3.1 Representação *space-variant* discreta

A representação *space-variant* utilizada aqui é construída a partir de uma "pirâmide" de resoluções sucessivamente menores, cujo *"nível zero"* é formado pela imagem com tamanho original, e os outros níveis (1, 2 e 3) são obtidos por sub-amostragens (ver Burt e Adelson, 1983 [10] e Adelson et al., 1984 [1]. A razão de sub-amostragem é a razão entre a largura (ou altura) da imagem em um nível e a largura (ou altura) do nível seguinte. Deste modo, se uma pirâmide tiver, por exemplo, 4 níveis e razão de sub-amostragem 2, e o nível zero tiver 64×64 pixels, o nível 1 terá 32×32 pixels, o nível 2 terá 16×16 pixels e o nível 3, 8×8 pixels. O modo de fazer esta sub-amostragem depende de como for feito o pré-processamento da imagem, e será detalhado na seção seguinte. Uma representação *space-variant* discreta completa poderá então ser construída utilizando uma região de cada nível de resolução (ver Figuras 3.2 e 3.3). A *fóvea* será formada por um círculo de diâmetro f extraído do nível zero da pirâmide, e as outras regiões por anéis de diâmetros crescentes

extraídos dos outros níveis da pirâmide. Em toda a fóvea e em toda a extensão de cada anel a resolução é uniforme, e por isso esta representação é chamada de *space-variant discreta*, por contraste com as representações em que a resolução decai continuamente em direção à periferia. A razão de crescimento do diâmetro dos anéis é um parâmetro do sistema, mas foi adotado na maioria das simulações o valor 3. Isto significa que cada anel abrange uma área da imagem original correspondente a um diâmetro 3 vezes o do anel anterior. A Figura 3.2 mostra uma visualização dos 4 níveis da pirâmide e das regiões de cada um deles utilizadas na formação da representação.

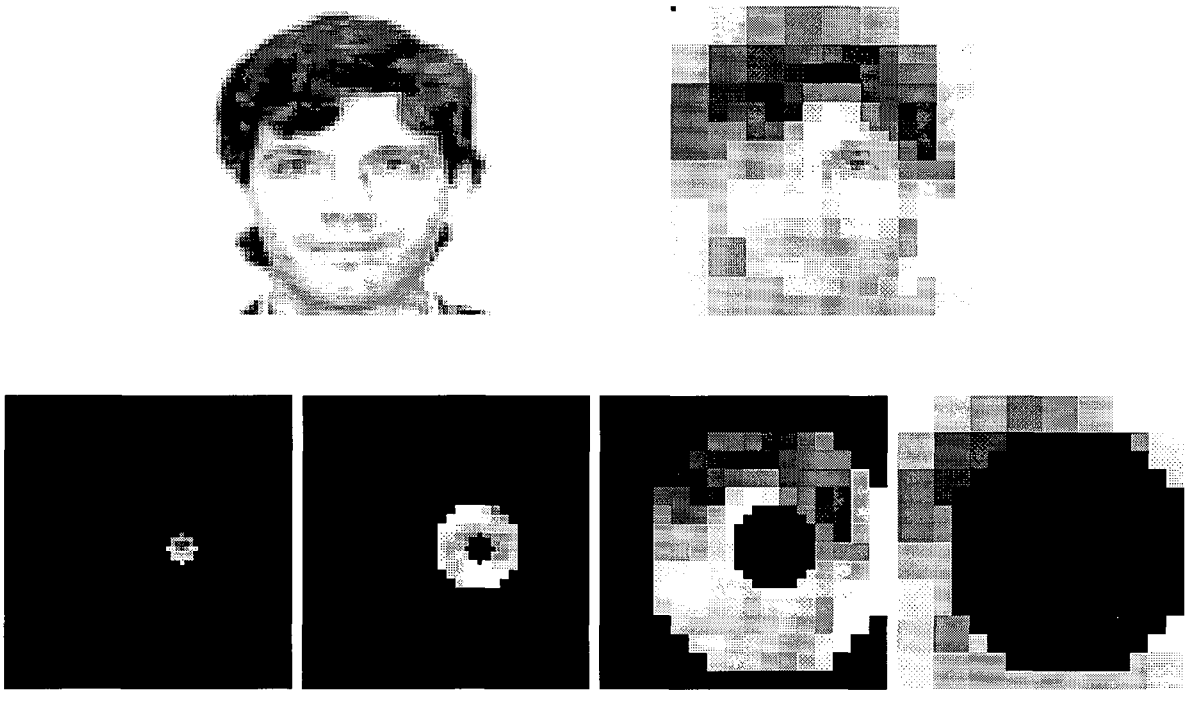

Figura 3.2: **Visualização de uma representação *space-variant* discreta, simulando uma retina:** Acima, esquerda: imagem original. Acima, direita: representação formada com as 4 regiões de resoluções diferentes simulando uma imagem obtida por uma "retina" com decaimento discreto da resolução. Embaixo: regiões de cada nível da pirâmide utilizadas para construir a representação *space-variant* discreta. Note que a extensão de cada nível é menor que a do nível precedente de acordo com a razão de sub-amostragem, como explicado no texto, porém aqui eles aparecem no tamanho correspondente à região que abrangem na imagem original, para visualização.

É importante notar que a Figura 3.2 é apenas uma visualização da representação, e nela os diâmetros dos anéis são mostrados com os tamanhos correspondentes às regiões que abrangem na imagem original. Na realidade, como cada nível da pirâmide é menor que o precedente, os anéis são progressivamente menores do que o tamanho mostrado na Figura 3.2 como ilustração. A Figura 3.3 apresenta os mesmos anéis em seu tamanho verdadeiro. Nela podemos ter uma idéia da grande redução de dimensionalidade obtida por esta representação (cerca de $2,7\%$ do número de pixels de uma imagem com 92×112).

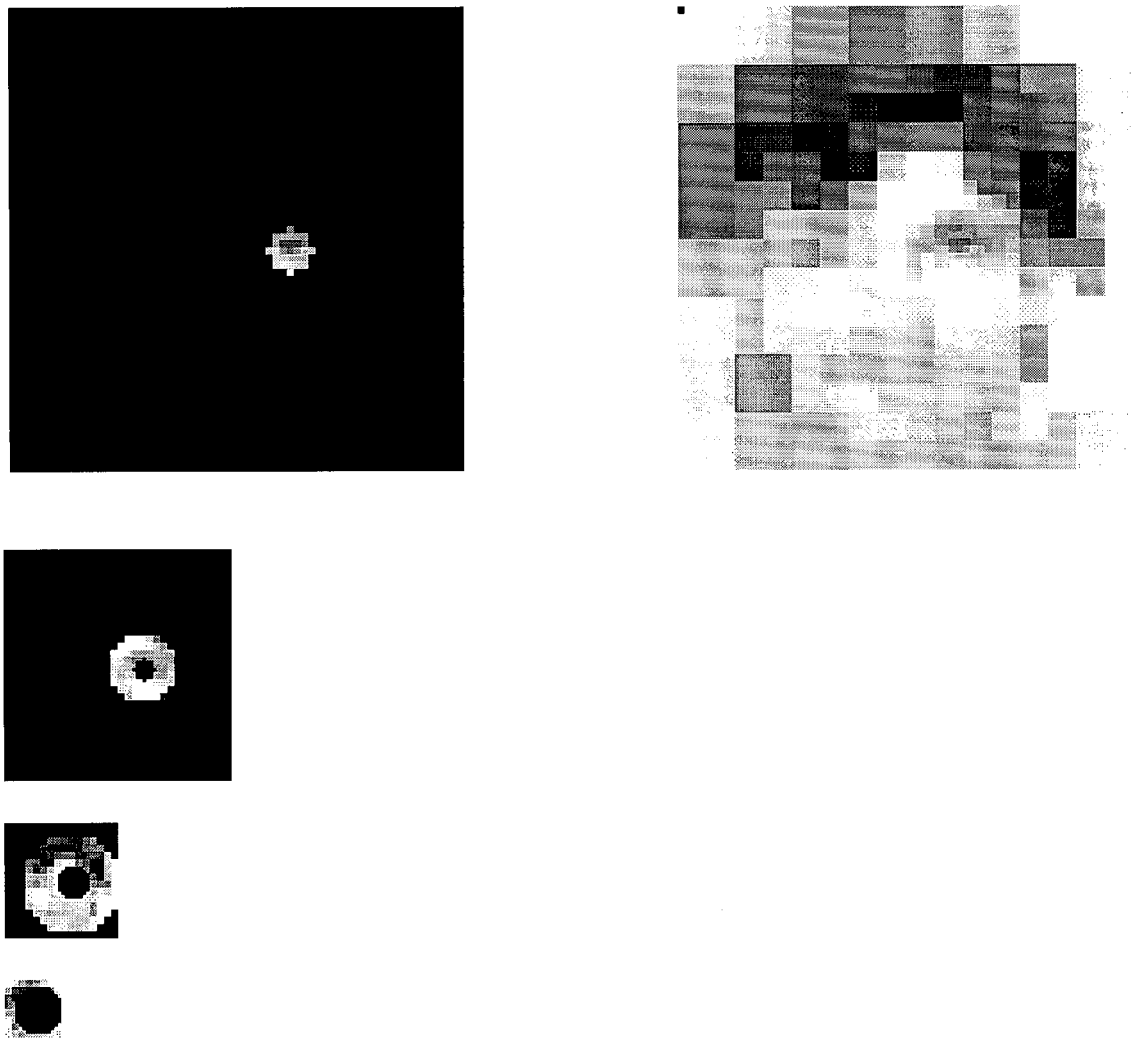

Figura 3.3: **Representação** *space-variant* **discreta:** Esquerda: regiões de cada nível da pirâmide utilizadas para construir a representação *space-variant*. Observe que a extensão de cada nível da pirâmide é menor que a do nível precedente de acordo com a razão de sub-amostragem, como explicado no texto. Direita (acima): visualização da representação formada com as 4 regiões de resoluções diferentes, em uma sacada para o olho esquerdo. Os anéis dos níveis inferiores ao nível zero foram ampliados para permitir a visualização do conjunto.

3.2 Pré-processamento

Muitos modelos de reconhecimento usam uma filtragem inicial da imagem com a finalidade de extrair elementos importantes como bordas, ângulos e linhas de contorno. O sistema visual humano possui células sensíveis a contrastes orientados, as células complexas, que são ativadas pela presença de uma linha de contraste em seu campo receptor. Células complexas são capazes de detectar bordas sem serem (praticamente) afetadas pela quantidade total de iluminação presente na cena, pois o que é percebido é a diferença relativa de luminosidade nos dois lados de uma borda e não seu valor absoluto. Por isso as representações assim obtidas são praticamente invariantes com os valores da iluminação total da imagem. Em nosso modelo

de reconhecimento, simulamos este processo usando filtros que detectam contrastes orientados, construídos com base em outros filtros que simulam as células simples (ver Grossberg e Mingolla, 1985 [24], Grossberg e Pessoa, 1998 [25] e Pessoa et al. 1995 [45]). Estas células detectam contrastes orientados com uma determinada direção (isto é, direção claro-escuro ou direção escuro-claro), dependendo da posição relativa de suas regiões excitatórias e inibitórias. Os filtros que simulam as células simples podem ser construídos a partir de funções gaussianas alongadas, ilustrados na Figura 3.4, pela diferença de duas gaussianas alongadas com uma certa orientação, que tem seus centros deslocados um em relação ao outro de um "off-set" proporcional à largura da gaussiana, como ilustrado na Figura 3.5. Descrevemos a seguir a construção dos filtros e o processo de obtenção de um mapa de respostas das células complexas através da filtragem de uma imagem. Mais adiante, nesta seção, veremos como foram obtidas as representações efetivamente utilizadas no modelo de reconhecimento atencional.

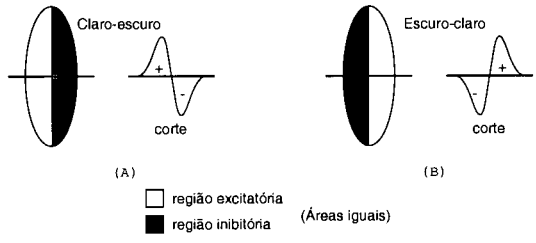

Figura 3.4: **Filtros simulando células simples, orientados na vertical:** O branco representa a região excitatória e o preto representa a região inibitória. Estes filtros respondem com maior intensidade nas regiões da imagem onde há um contraste de mesma orientação (vertical, no exemplo) e direção (claro-escuro ou escuro-claro) Estes filtros podem ser construídos pela diferença de duas gaussianas alongadas, como explicado no texto.

Uma função gaussiana alongada é definida pela equação [28]:

$$G_{\sigma_x\sigma_y}(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y} \exp\left(\frac{-1}{2}\left(\frac{(x-t_x)^2}{\sigma_x^2} + \frac{(y-t_y)^2}{\sigma_y^2}\right)\right) \tag{3.1}$$

onde $G_{\sigma_x\sigma_y}$ é a função gaussiana com largura determinada por σ_x na direção x e por σ_y na direção y. As coordenadas do centro da gaussiana são t_x e t_y. Assim, se $t_y = 0$, $t_x = -\sigma_x$, e $\sigma_x < \sigma_y$, temos uma gaussiana alongada na vertical G_v^{on}, centrada no ponto $(-\sigma_x, 0)$. Subtraindo desta uma outra gaussiana alongada na vertical G_v^{off}, centrada em $(\sigma_x, 0)$, teremos o filtro ilustrado na Figura 3.4 à esquerda, isto é, um filtro sensível a contraste de orientação vertical e direção de contraste claro-escuro (da esquerda para a direita). Este filtro é definido pela equação:

$$s_v^{ce} = G_v^{on} - G_v^{off}.$$

Um filtro de mesma orientação, mas com direção de contraste oposta, isto é, escuro-claro (como à direita na Figura 3.4), será obtido simplesmente trocando os sinais das gaussianas:

$$s_v^{ec} = -G_v^{on} + G_v^{off}.$$

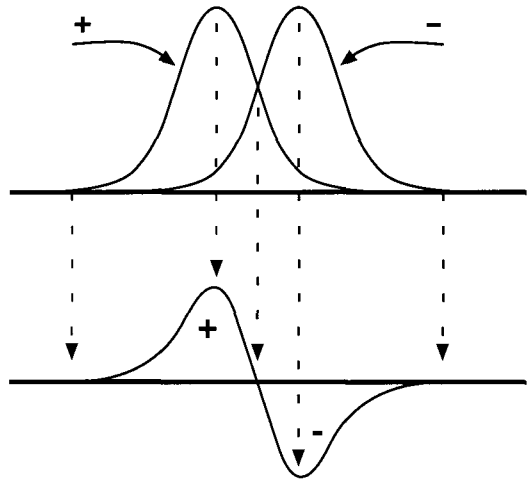

Figura 3.5: **Construção dos filtros ilustrados na Figura 3.4:** Estes filtros podem ser construídos pela diferença de duas gaussianas alongadas de mesma orientação, G^{on} e G^{off} que tem seus centros deslocados um em relação ao outro de um "off-set" proporcional à largura da gaussiana.

A relação entre σ_y e σ_x no exemplo acima, que regula o aspecto alongado na vertical da gaussiana, é chamada de "razão de aspecto", e constitui um parâmetro do pré-processamento (*asp*).

Outros filtros como estes, mas com qualquer orientação k no plano, podem ser conseguidos simplesmente por uma rotação de um ângulo k da função especificada em 3.1. As funções G_v^{on} e G_v^{off} são então substituídas por G_k^{on} e G_k^{off}, produzindo assim os filtros simples s_k^{ce} e s_k^{ec}, que são filtros de orientação k e direção de contraste *ce* e *ec*.

Normalmente um mapa que simula as respostas das células simples do sistema visual pode ser obtido fazendo as convoluções dos filtros que simulam as células simples com a imagem. Entretanto, para que a comparação entre as imagens filtradas seja eficiente, é preciso normalizar os valores das repostas dos filtros. Na realidade, porém, uma vez que o sistema é incremental e por isso utiliza a cada passo apenas informações parciais da imagem, esta normalização precisa ser feita localmente. Para isso, em vez de fazer a convolução dos filtros s_k^{ce} e s_k^{ec} com a imagem, fazemos separadamente a convolução das gaussianas G_k^{on} e G_k^{off} com a imagem, obtendo as respostas I_c e I_e. A seguir calculamos a resposta das células simples fazendo a diferença das convoluções e normalizando este resultado dividindo-o pela soma, como dado pelas equações

$$I_c = G_k^{on} * I, \tag{3.2}$$

$$I_e = G_k^{off} * I, \tag{3.3}$$

$$S_k^{ce} = \frac{I_c - I_e}{A + I_c + I_e}$$

e

$$S_k^{ec} = \frac{I_e - I_c}{A + I_e + I_c},$$

onde A é uma constante, k representa a orientação do filtro no plano da imagem, e *ce* e *ec* representam as direções do contraste, isto é, a posição relativa dos lados claro e escuro. Assim, para cada orientação k, existem duas funções, S_k^{ce} e S_k^{ec} que são as respostas normalizadas das células simples, detectando, respectivamente, contrastes claro-escuro e escuro-claro na orientação k. Entretanto, cada uma destas funções pode ter regiões negativas correspondentes aos contrastes de direção contrária que não devem ser detectados pelo filtro. Por isso estes valores negativos devem ser eliminados, gerando-se então as respostas positivas das células simples P_k^{ce} e P_k^{ec} segundo as expressões:

$$P_k^{ce} = max(S_k^{ce}, 0)$$

e

$$P_k^{ec} = max(S_k^{ec}, 0).$$

Uma filtragem simulando a resposta de uma célula complexa sensível a um contraste de qualquer direção na orientação k pode ser obtida somando as respostas positivas dos filtros que simulam as células simples, como ilustrado na Figura 3.6. Assim, um mapa C_k de respostas das células complexas na orientação k é obtido simplesmente como em 3.4:

$$C_k = P_k^{ce} + P_k^{ec}. \tag{3.4}$$

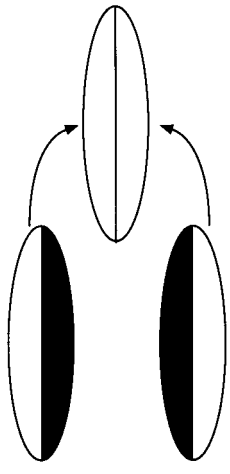

Figura 3.6: **Filtro simulando uma célula complexa:** as respostas positivas de duas células simples de mesma orientação e direções de contraste contrárias somadas formam um filtro que responde a um contraste orientado, seja claro-escuro ou escuro-claro.

Para detectar contornos em qualquer prientação, precisamos de respostas das células complexas em todas as orientações. Na realidade, porém, uma boa aproximação pode ser conseguida utilizando apenas as respostas para quatro orientações separadas de 45 graus, como ilustrado na Figura 3.7. A Figura 3.8 mostra um mapa de respostas das células complexas detectando contrastes em todas as orientações, obtido pela soma das filtragens simulando as respostas das células complexas em 4 orientações.

Construção da pirâmide e da representação *space-variant* discreta

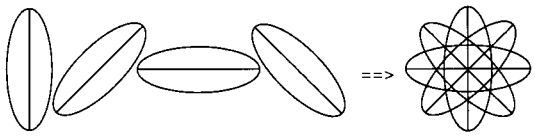

Figura 3.7: **Detecção de contrastes em todas as orientações:** Uma boa aproximação é conseguida somando as respostas para apenas quatro orientações separadas de 45 graus.

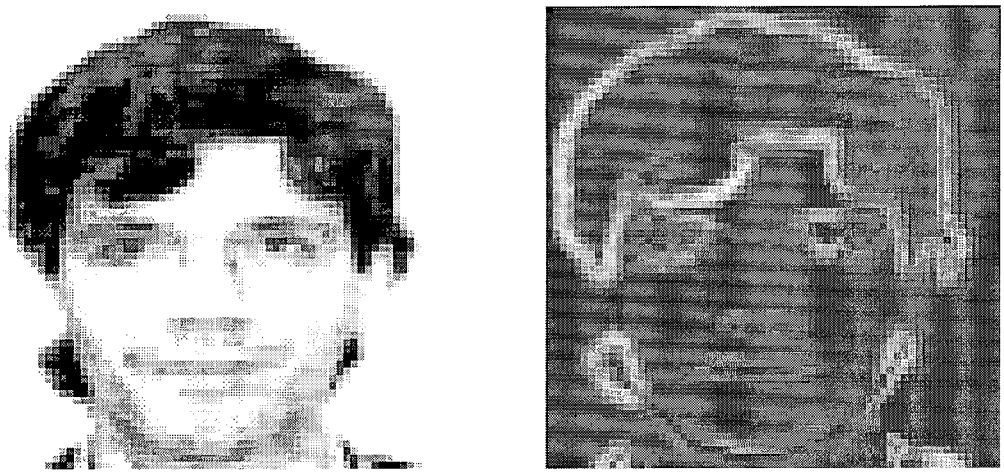

Figura 3.8: **Contraste orientado:** Esquerda: imagem original. Direita: mapa de respostas das células complexas, detectando contrastes em todas as orientações, obtido por filtragem conforme explicado no texto.

Como vimos na seção anterior, em nosso modelo de reconhecimento uma representação *space-variant* discreta é construída baseada em uma pirâmide de 4 níveis de resolução. Esta pirâmide é utilizada para representar os modelos armazenados no sistema. Vamos primeiro mostrar como esta prâmide é obtida, e depois como, durante o processo de reconhecimento, é extraída uma representação *space-variant* para um determinado ponto de fixação.

O nível zero da pirâmide é simplesmente o mapa das respostas das células complexas obtido como explicado acima e ilustrado na Figura 3.8. O tamanho dos filtros, definido pelos valores de σ_x e σ_y nas funções gaussianas G_k^{on} e G_k^{off} em 3.2 e 3.3 são os menores para este nível da pirâmide. Um valor típico usado foi $\sigma_x = 2,0$ e $\sigma_y = 6,0$. Nos níveis de menor resolução, o tamanho dos filtros é maior, já que estamos simulando o crescimento dos campos receptores das células complexas, de modo que a cada nível os valores de σ_x e σ_y são multiplicados por um parâmetro do sistema, a razão de crescimento dos filtros (*cf*). Deste modo os filtros terão uma abrangência maior, na imagem, quanto menor for a resolução. Por este mesmo motivo, não é necessário aplica-los a todas as posições da imagem ao fazer a filtragem nos níveis de menor resolução. Ao contrário, o centro dos filtros podem ser aplicados só a pontos de uma malha que tem, em cada nível, por exemplo, o dobro do espaçamento do nível anterior. A Figura 3.9 mostra, à esquerda, os espaçamentos da malha de localização dos centros dos filtros para o processamento de cada nível da pirâmide. No centro estão ilustrados os tamanhos dos filtros utilizados a cada nível, e à direita o resultado da filtragem. No nível zero, são usados filtros pequenos

para todos os pontos (pixels) da imagem. Como a malha dobra de espaçamento a cada nível, o número de pontos diminui 4 vezes a cada nível, resultando assim num processo de sub-amostragem que produz os 4 níveis da pirâmide. Se, por um lado, a cada nível a área do filtro aumenta, por outro é menor o número de pontos a serem efetivamente filtrados. Por este processo são produzidas as pirâmides de representação dos modelos armazenados no sistema. Observe que, para um conjunto de malhas com os espaçamentos dobrando a cada nível (como ilustrado na Figura 3.9), o total de pontos utilizados nas filtragens, somando todos os níveis, é de $1,33$ vêzes o número de pontos (pixels) da imagem original.

É importante salientar que a representação da imagem que está sendo apresentada, só será extraída à medida que for necessário, isto é, cada vez que o Módulo Atencional determinar uma nova sacada, enquanto não houver informação suficiente para que o Sistema de Decisão encerre o processo. Desta forma, a cada ponto de fixação escolhido pelo sistema, a filtragem será feita apenas nas regiões que vão formar os anéis. Para isso, os filtros serão centrados apenas nos pontos da malha de cada nível localizados nos anéis utilizados para a formação da representação *space-variant* centrada no ponto de fixação atual, como ilustrado na figura 3.10. Somente os pontos pertencentes aos anéis que vão formar a representação são utilizados (à esquerda na figura). Os círculos delimitam as regiões efetivamente utilizadas. Os filtros (centro da figura) serão aplicados nos pontos da malha no interior destas regiões, e somente estas (à direita na figura) vão participar da representação *space-variant* extraída para este ponto de fixação.

3.3 Sistema de Decisão

O Sistema de Decisão é responsável por determinar se a imagem apresentada é uma instância de um dos modelos armazenados. Sua operação básica é determinar o grau de similaridade entre a imagem e cada modelo. Quando uma alta medida de similaridade excede um certo *critério de decisão*, a imagem é reconhecida como sendo o objeto cujo modelo é mais semelhante, e o processo termina. Caso contrário, o Módulo Atencional entra em ação para escolher uma nova região da imagem a ser considerada. No nosso modelo de reconhecimento, o sistema de Decisão a cada momento contém somente *informações parciais* da imagem (ver Aguilar e Ross, 1993 [2]). Para operar deste modo, o Sistema de Decisão deve permitir o uso de vetores de características parciais que vão sendo estendidos incrementalmente.

O Sistema de Decisão pode utilizar diversos métodos para determinar a similaridade entre o vetor de características extraído da imagem e os modelos armazenados. Em nosso modelo de reconhecimento empregamos uma medida de correlação inspirada na rede neural FuzzyArt, desenvolvida por Carpenter e Grossberg [12, 11]. O Sistema de Decisão consiste em uma entrada, F1, que contém o vetor V de características extraídas da imagem apresentado ao sistema, e uma saída, F2, contendo as unidades de categorização (ver Figura 3.11). Cada unidade em F2 representa o modelo de um objeto, e a cada momento o valor y_m da sua ativação reflete a confiança do sistema de que o padrão considerado seja um membro da categoria respectiva. Formalmente, o valor da ativação das unidades de saída refletem uma medida do tipo correlação normalizada da semelhança entre o vetor V de carac-

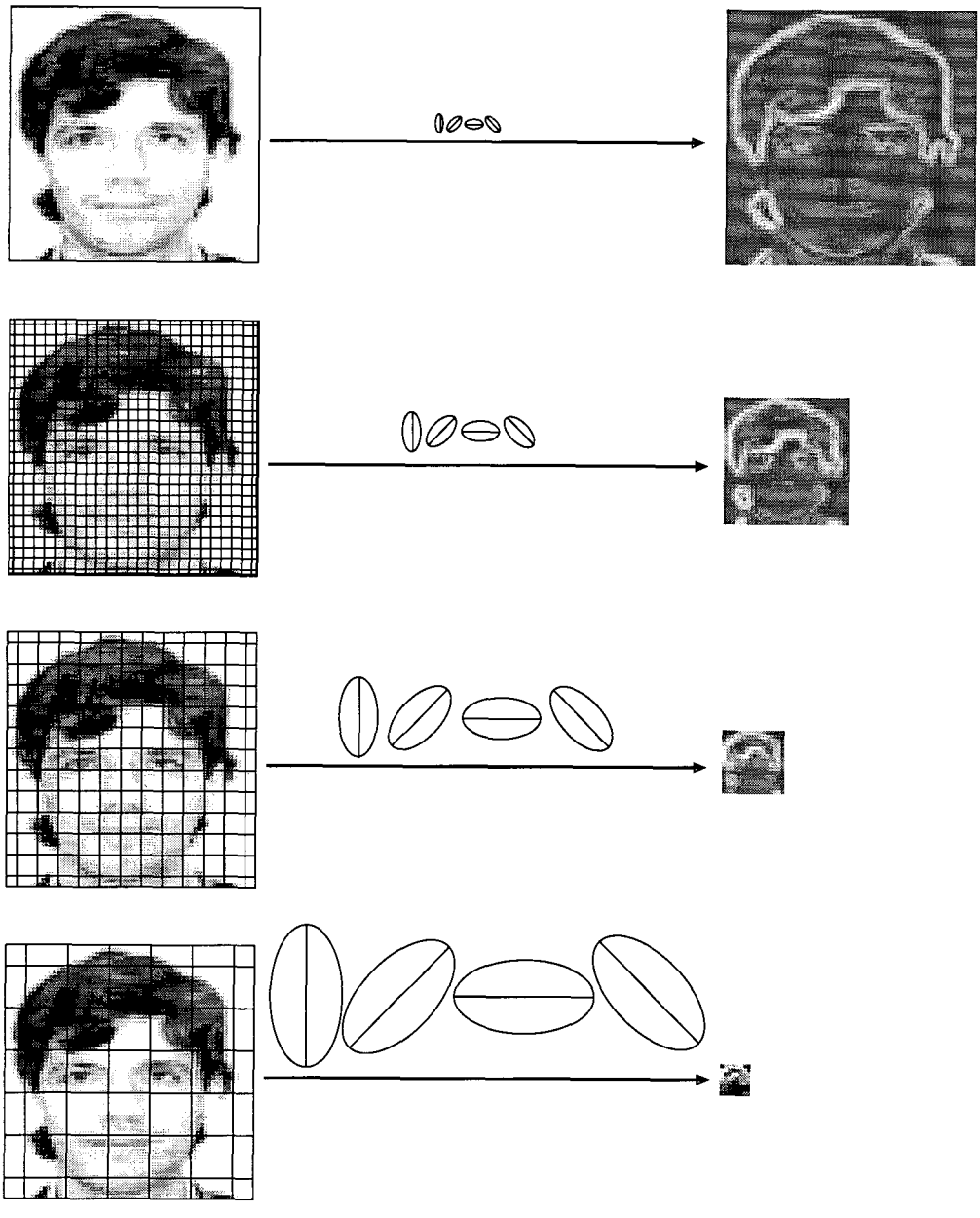

Figura 3.9: **Pré-processamento e construção da pirâmide:** Esquerda: ilustração da malha de localização dos centros dos filtros para o processamento de cada nível da pirâmide, com os espaçamentos crescendo a cada nível. Centro: ilustração dos tamanhos dos filtros utilizados a cada nível. Direita: resultado da filtragem. O número de pontos diminui 4 vêzes a cada nível.

terísticas extraído da imagem de entrada e cada um dos modelos armazenados. Os modelos armazenados são representados por vetores W_m, onde m é o índice da categoria representada, e seus componentes $W_{m,i}$ são os valores dos pixels dos 4 níveis da pirâmide de resoluções que representa o modelo. Assim, formalmente, os valores dos componentes de W_m são dados por:

$$W_{m,i} = C_{m,i},$$

onde $C_{m,i}$ é o valor da resposta dos filtros que simulam as células complexas para a

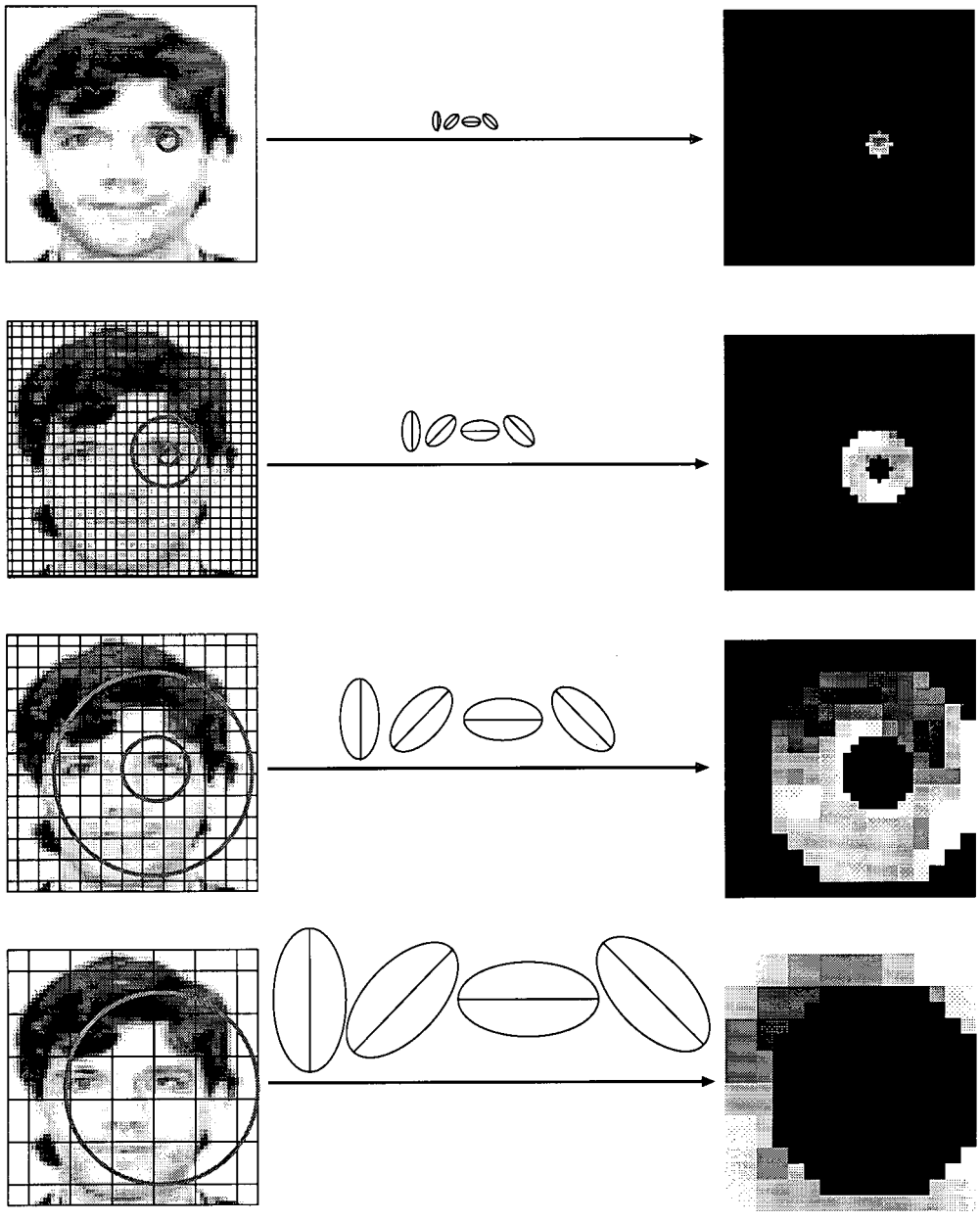

Figura 3.10: **Extração da representação** *space-variant* **da imagem a ser reconhecida:** Esquerda: regiões da malha de localizações dos centros dos filtros a cada nível da pirâmide. Somente os pontos pertencentes aos anéis (delimitados por círculos) que vão formar a representação são utilizados. Centro: ilustração dos tamanhos dos filtros a cada nível. Direita: anéis filtrados a cada nível, ampliados para visualização.

posição i do modelo m.

Como o modelo é incremental, o padrão de entrada não é apresentado ao Sistema de Decisão todo de uma vez, mas sim apenas um certo número de componentes é acrescentado a cada ciclo (ou sacada, no modelo de reconhecimento atencional), aqueles componentes da representação *space-variant* centrada no ponto de fixação atual extraída da imagem. De outro lado, apenas uma parte dos vetores W_m entra

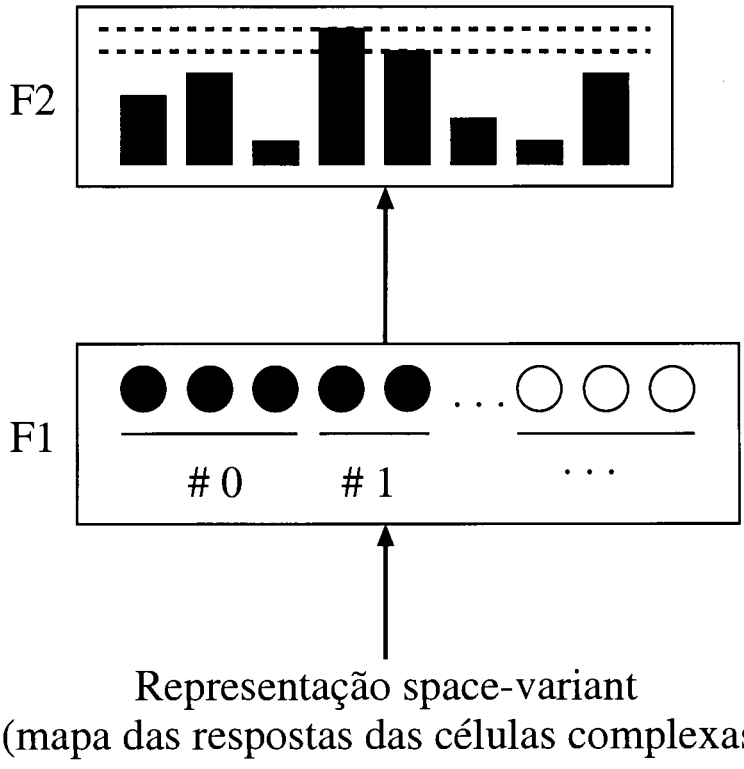

Figura 3.11: **Sistema de Decisão:** A cada sacada, novos dados são acrescentados, aumentando o número de componentes do vetor de entrada em F1. As ativações das unidades em F2 medem o grau de correlação entre as categorias que estas unidades representam e o a parcela do vetor de entrada já extraído da imagem de entrada. Uma decisão pode ser tomada com base na diferença de ativações entre a categoria mais ativa e a segunda mais ativa, assinalada pelas linhas tracejadas horizontais.

em ação a cada momento, especificamente os componentes da representação que correspondem às posições já extraídas da imagem até aquele momento. Esta parcela ativa dos vetores W_m vai aumentando seu número de componentes u a cada ciclo da mesma forma que aumenta o número de componentes ativos na entrada F1. Na Figura 3.11 a parcela ativa do vetor V em F1 está representada pelos pontos pretos, os três primeiros exemplificando os componentes adicionados no ciclo inicial (# 0), os dois seguintes os componentes adicionados no ciclo 1, e os demais pontos (brancos) exemplificam componentes ainda não adicionados. A comparação é então feita entre a parcela de componentes ativos dos vetores W_m que representam cada categoria, e a parcela ativa do vetor V de características extraída da imagem até aquele momento.

A ativação de cada unidade de categorização em F2 mede a correlação entre W_m e V num dado momento, e é dada pela expressão:

$$y_m = \frac{\sum_{i=0}^{u}(W_{m,i} \wedge V_i)}{\sum_{i=0}^{u} W_{m,i} + cp}, \tag{3.5}$$

onde y_m é a ativação da categoria correspondente ao modelo m e $W_{m,i}$ é o componente i do vetor que representa o modelo m. V_i é o componente i do vetor que representa o padrão de entrada na entrada F1, $W_{m,i} \wedge V_i$ é o mínimo "fuzzy" entre

$W_{m,i}$ e V_i, dado por $W_{m,i} \wedge V_i = min(W_{m,i}, V_i)$, cp é um valor muito pequeno para evitar a divisão por zero, e u é o índice do último componente acrescentado ao vetor de entrada (este processo de decisão é inspirado em Carpenter et al., 1991 [12] e Aguilar e Ross, 1993 [2]).

À medida que vão sendo acrescentados dados do padrão de entrada, o vetor de entrada vai se tornando mais extenso, e a comparação vai sendo feita também com uma porção maior dos vetores que representam os modelos. Cada pixel da representação *space-variant* extraída da imagem de entrada fornece um componente do vetor de entrada V. O número de componentes comparados é dado por u, que representa o número total de pixels comparados até um dado momento. A cada sacada, somente os pixels correspondentes à fóvea e ao anel do nível 1 da pirâmide são acrescentados ao vetor de entrada, e um número igual de pixels correspondentes nos modelos é também acrescentada aos vetores W_m, representantes dos modelos.

Os vetores que representam os modelos são obtidos a partir das pirâmides construídas como explicado na seção anterior. Os modelos são construídos a partir de imagens pertencentes a um conjunto de treinamento. Em algumas simulações os modelos foram construídos utilizando uma imagem simples de cada objeto para formar cada modelo ou, em outras simulações, o modelo foi sintetizado, para cada objeto, usando a média de várias imagens do conjunto de treinamento deste objeto. Podem ser utilizadas muitos outros métodos de construir modelos, e uma investigação da eficiência destes métodos pode ser encontrada em Pessoa e Leitão, 1999 [44] e Leitão e Pessoa, 1999 [34].

Canais ON e OFF:

Como a medida de correlação utilizada pelo Sistema de Decisão é baseada no mínimo "fuzzy" entre dois vetores, o sistema dará a mesma resposta para qualquer valor de $V_i > W_{m,i}$, isto é, não haverá discriminação entre diferentes valores de V_i. A capacidade de discriminação deste processo pode ser melhorada através da adoção de um "código complementar" [12], que consiste no uso de dois canais de informações, o "canal ON" e o "canal OFF". Desta forma a comparação passa a ser feita entre os vetores W_m^{ON} e V^{ON} e também entre os vetores W_m^{OFF} e V^{OFF} ao invés da simples comparação entre W_m e V. Para isso as informações devem ser normalizadas para o intervalo $[0, z_{max}]$, onde z_{max} é a amplitude dos canais ON e OFF, e constitui um dos parâmetros do sistema que deve ser escolhido próximo ao valor máximo das respostas dos filtros que simulam as células complexas. O canal ON para um vetor X tem seus componentes calculados por

$$X_i^{ON} = min(X_i, z_{max}),$$

e o canal OFF para o mesmo vetor X tem seus componentes calculados por

$$X_i^{OFF} = z_{max} - X_i^{ON}.$$

Por este processo, quando $V_i^{ON} > W_{m,i}^{ON}$, teremos $V_i^{OFF} < W_{m,i}^{OFF}$, havendo portanto discriminação entre diferentes valores de V_i.

Deste modo, o Sistema de Decisão terá seu vetor de entrada em F1 com o dobro do número de componentes, pois receberá ambos os canais ON e OFF. Para os vetores que representam os modelos também será calculado o canal OFF, de modo que a comparação, na realidade, se dará entre estes vetores de comprimento $2u$. O

uso do canal OFF permite discriminar diferenças que não poderiam ser distinguidas sem o uso dos dois canais. O desempenho melhor com o uso dos dois canais foi comprovado também por várias simulações.

3.4 Módulo Atencional: Mapa de Saliência e estratégias de orientação das sacadas

O Módulo Atencional é responsável por escolher o ponto de fixação [1] da próxima sacada, baseado nas informações presentes no sistema até um dado instante. Esta escolha é feita com base num "Mapa de Saliência" calculado a cada sacada, conforme explicaremos adiante. O próximo ponto de fixação será determinado pela posição onde ocorre o valor máximo do Mapa de Saliência atual.

Construção do Mapa de Saliência

O Mapa de Saliência é também uma estrutura com resolução variável. Como na representação da imagem, seus valores são calculados em alta resolução para a região foveal central (isto é, um disco em torno do ponto de fixação), e em resoluções menores para três anéis em torno da fóvea. Este mapa é construído a partir de uma *função de saliência* calculada para os diversos níveis da pirâmide utilizada nas representações processadas pelo sitema. Assim, em uma pirâmide de quatro níveis, o nível 0 da função de saliência é calculada para uma região circular correspondente à *fóvea*, o nível 1 para um anel circular correspondente à região adjacente à *fóvea*, e assim por diante, até o nível 3, que é o de menor resolução (ver Figura 3.12) [2].

O Mapa de Saliência será inicialmente idêntico à função de saliência; com o decorrer das sacadas algumas modificações serão introduzidas visando impedir que o sistema volte a dirigir sacadas para os mesmos pontos já visitados. O Mapa de Saliência guardará então a memória dos pontos já visitados, passando a ser diferente da função de saliência nas regiões próximas aos pontos de fixação das sacadas já executadas. A função de saliência é, no caso geral, uma função do mapa das respostas das células complexas que forma a representação da imagem de entrada, das ativações na saída F2 do Sistema de Decisão, e das representações dos modelos armazenados no sistema, calculada apenas nas posições dos anéis mostrados na Figura 3.12:

$$S_{n,p} = F(C_{n,p}, M_{n,p}, y)\,,$$

onde $S_{n,p}$ é o valor do Mapa de Saliência no nível n, na posição p deste nível, $C_{n,p}$ é o valor da resposta dos filtros que simulam as células complexas extraída da imagem de entrada no nível n, na posição p; $M_{n,p}$ representa os valores dos pixels das representações dos modelos no nível n, na posição p armazenado na memória; y

[1] Lembrar que "ponto de fixação" é um ponto da imagem de entrada para onde será dirigido o centro da "fóvea," constituindo assim o ponto de destino do "movimento sacádico" simulado e centro da representação *space-variant* que será extraída da imagem.

[2] Por simplicidade, a região circular correspondente à *fóvea* será referida também como *anel*. Assim, daqui em diante, *anel* significará a região de um determinado nível da pirâmide utilizada para simular a representação de uma *retina*. Serão regiões delimitadas por dois círculos concêntricos, sendo que o círculo interno do anel no nível de maior resolução utilizado terá raio nulo, formando uma região circular cheia.

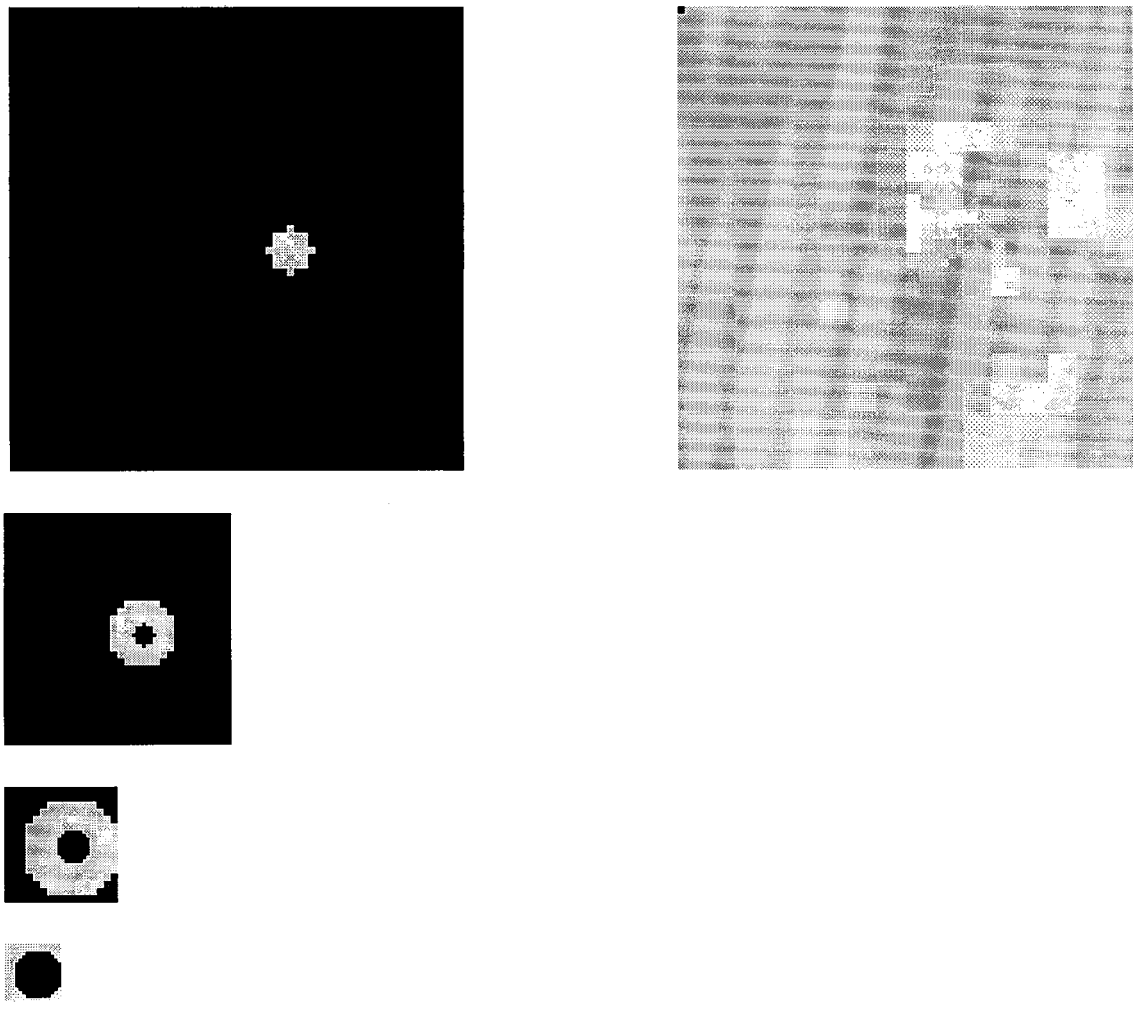

Figura 3.12: **Mapa de Saliência:** Esquerda: regiões de cada nível da pirâmide para as quais é calculada a Função de Saliência. Direita: Mapa de Saliência formado com as 4 regiões de resoluções diferentes, ampliadas para visualização.

representa os valores das ativações das categorias na saída F2 do Sistema de Decisão, e $F()$ é a função de saliência, que exprime o critério de saliência correspondente à estratégia atencional que está sendo utilizada. Cada estratégia atencional combinará de forma diferente as informações dos modelos, do Sistema de Decisão e da imagem, sendo definida por uma expressão diferente de $F()$.

Podemos visualizar melhor a construção do Mapa de Saliência usando um exemplo de estratégia estritamente *bottom-up* (ver Figura 3.13). Neste exemplo, o critério de escolha privilegia os pontos onde é maior o valor das respostas das células complexas, isto é, onde os pixels da representação da imagem apresentam maior valor. Neste caso a função de saliência não utiliza informações provenientes dos modelos. A função de saliência neste caso é a mais simples, e poderia ser escrita como:

$$F_{n,p} = C_{n,p},$$

Como o Mapa de Saliência calculada na sacada inicial é identico à função de saliência, pode ser escrito:

$$S_{n,p} = F_{n,p} = C_{n,p},$$

Esta composição do Mapa de Saliência está ilustrada na Figura 3.13, onde o Mapa de Saliência é construído utilizando apenas informações da imagem porque neste exemplo a estratégia é *bottom-up*, e nem os modelos nem o Sistema de Decisão interferem. A imagem e os modelos são mostrados na Figura 3.13 como simulações de imagens *space-variants* para permitir a visualização, entretanto os mapas de respostas dos filtros que simulam as células complexas é que são realmente utilizados.

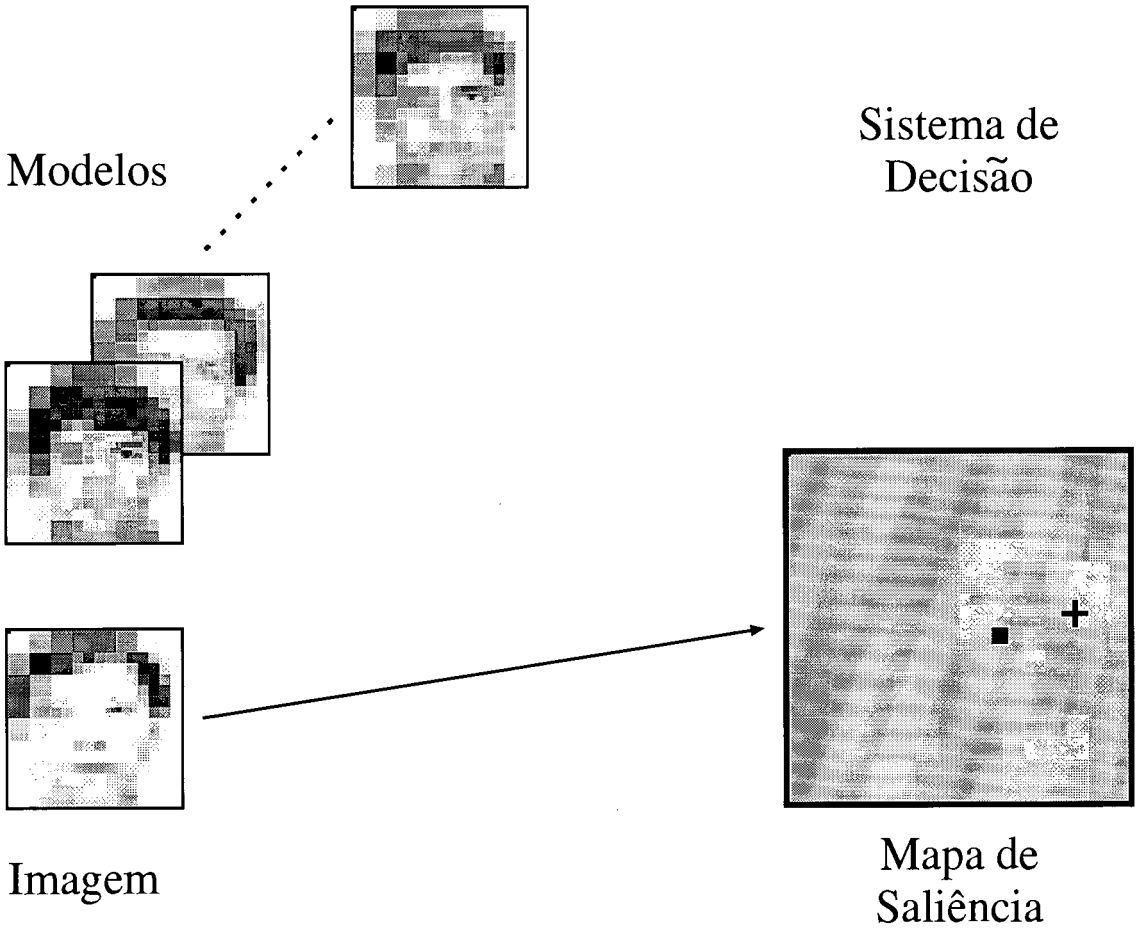

Figura 3.13: **Construção do Mapa de Saliência.** Note que o Mapa de S aliência é construído levando em conta as respostas das células complexas extraídas da imagem (ou dos modelos armazenados), e não as imagens simples, como mostrado apenas para visualização. Como neste exemplo somente são utilizadas informações da imagem (estratégia *bottom-up*) nem os modelos nem o Sistema de Decisão participam da construção do Mapa de Saliência.

Estratégias atencionais

No exemplo acima a estratégia atencional baseava-se inteiramente na imagem. Outras estratégias podem ser implementadas construíndo o Mapa de Saliência usando uma função de saliência que leve em conta os modelos armazenados e também informações provenientes do Sistema de Decisão, como por exemplo o nível de ativação das categorias correspondentes a cada modelo. No Capítulo 4 faremos uma análise detalhada de várias estratégias atencionais, incluindo como construir o Mapa de Saliência para cada uma delas.

3.5 Parâmetros do Modelo de Reconhecimento Atencional e indicadores do processo de reconhecimento

Como vimos, o Sistema de Decisão calcula o valor das ativações y_m das categorias na saída F2 por comparação do vetor de entrada com os vetores representativos dos modelos armazenados. Na verdade o que nos interessa para o processo de reconhecimento não é tanto o valor das ativações mas sim a proporção entre elas. Assim, se uma categoria tem uma ativação muito maior que as outras, este fato reflete a alta probabilidade de que o objeto apresentado pertença a esta categoria. Para tornar as diferenças de ativação mais visíveis, Por causa disso é conveniente o uso de um processo de " amplificação" das ativações através da expressão:

$$A_m = \frac{y_m^q}{\sum_i y_i^q},$$

onde A_m é a ativação amplificada da categoria m, q é um parâmetro que modula esta amplificação e i é o índice dos modelos, variando de 1 a N, onde N é o número total de modelos. Desta forma as diferenças entre as ativações serão realçadas sempre que q for maior que 1, permitindo uma melhor discriminação da categoria mais ativa. Este tipo de amplificação é usada em Aguilar e Ross, 1993 [2].

Discriminação e Critério de Decisão

O indicador mais importante do processo de reconhecimento é a discriminação entre a categoria mais ativa m_1 e a segunda mais ativa m_2. Esta discriminação D é definida pela expressão:

$$D = \frac{A_{m_1}}{A_{m_2}}.$$

A partir deste indicador, torna-se possível definir um *Critério de Decisão cd* que indique o fim do processo de reconhecimento. Se, por exemplo, a discriminação D alcançada em uma determinada sacada for $1,3$, isso significa que a ativação amplificada da categoria mais ativa é 30% maior que a segunda mais ativa, e portanto a confiança do sistema em que o objeto apresentado pertence a esta categoria é de 30%. O reconhecimento será dado por concluído assim que, em qualquer sacada, $D \geq cd$.

As sacadas iniciais, dada a pequena quantidade de informações que trazem, tendem a produzir resultados instáveis e muitas vêzes espúrios. Em muitas situações é conveniente estabelecer um mínimo de sacadas antes de permitir que o sistema decida. Para isso é definido um parâmetro (*rec*) que estabelece este mínimo, de modo que o critério de decisão só será aplicado após este mínimo de sacadas.

Discriminação da Categoria Correta

Um outro indicador muito útil é a *discriminação da categoria correta*, D_c. Quando em uma simulação se sabe qual é a categoria correta m_c e se quer saber se o sistema é capaz de encontrá-la, é conveniente plotar a evolução deste indicador. Ele

é definido por:

$$D_c = \frac{A_{m_c}}{A_{m_2}}, \tag{3.6}$$

onde A_{m_c} é a ativação amplificada da categoria correta. Note que $D_c > 1$ se a categoria correta for a mais ativa; $D_c = 1$ se a categoria correta for a segunda mais ativa, e $D_c < 1$ se a categoria correta for menos ativa que a segunda mais ativa.

Parâmetros usados nas simulações:

- Tamanho (ou largura) dos filtros alongados no nível zero;
- Razão de aspecto dos filtros;
- Razão de crescimento dos filtros a cada nível;
- Diâmetro da *"fóvea"*;
- Diâmetro da região excluída no Mapa de Saliência, explicado no Capítulo 4;
- Razão de crescimento dos anéis;
- Nível de amplificação das ativações;
- Limite entre as altas e baixas ativações, explicado no Capítulo 4;
- Critério de Decisão;
- Mínimo de sacadas;
- Amplitude dos canais ON e OFF;

Capítulo 4

O Módulo Atencional

Como vimos, o "Módulo Atencional" tem a função de determinar novas áreas de interesse para o processo de reconhecimento no objeto ou imagem dada, encontrando um novo ponto de fixação para a próxima sacada, baseado nos dados disponíveis no sistema até um determinado momento. Mas o que significa uma área de interesse?

Alguns pesquisadores tem empregado áreas fixas, como olhos, nariz, boca etc. em faces (ver Tistarelli, 1995 [58] e Wiskott et al., 1997 [65]). Outros tem sugerido que áreas importantes são aquelas que tem alta energia nas respostas a operadores sensíveis a descontinuidades de luminância (Alpaydin, 1996 [3]), o que corresponde a uma estratégia atencional que chamamos de *estratégia baseada na imagem* (ou "bottom-up"), posto que só considera a imagem dada. Uma terceira alternativa é enfatizar regiões nas quais os modelos diferem mais, uma *estratégia baseada nos modelos* (ou "top-down"), que só considera o conjuntó de modelos armazenados no sistema, ignorando a imagem dada. Estas duas últimas estratégias, quando aplicadas isoladamente, não utilizam todas as informações disponíveis no sistema. Assim, uma abordagem estritamente baseada na imagem, por exemplo, não se beneficia das informações armazenadas nos modelos, e vice-versa. Uma possibilidade que vamos explorar é a combinação destas estratégias individuais de processamento, formando uma *estratégia híbrida*. Na verdade podem existir várias estratégias híbridas diferentes, combinando de formas diferentes as informações da imagem e da memória.

Este capítulo se dedica a uma análise detalhada das estratégias atencionais para determinação do próximo ponto de fixação. Cada uma destas estratégias decorre de motivações diferentes, e os critérios para determinação do melhor ponto de fixação, o comportamento do Modelo, sua eficiência e aplicabilidade serão também diferentes. A análise qualitativa apresentada aqui procura explicar as motivações e o comportamento do Modelo nas diferentes estratégias e será ilustrada com algumas simulações. Uma análise quantitativa de todas as estratégias investigadas aqui, baseada em simulações extensas será apresentada no Capítulo 5, onde se procurará apresentar mais elementos para avaliar a eficiência e aplicabilidade destas estratégias.

O primeiro tipo de estratégia (áreas fixas) aplica-se melhor aos sistemas dedicados a uma determinada classe de objetos, na qual os detalhes ou regiões importantes para o reconhecimento já são préviamente conhecidas. O problema de guiar as sacadas, neste caso, se configura como um problema de busca destas regiões [65, 58], ou poderá ser resolvido, eventualmente, por um pré-processamento que normalize a posição das regiões escolhidas nos modelos armazenados e na imagem de entra-

da. Feito isso, o roteiro de sacadas será sempre o mesmo para qualquer imagem de entrada. Num sistema dedicado ao reconhecimento de faces, por exemplo, a posição dos olhos, boca, nariz, etc. (ver Figura 4.1), podem estar normalizados pelo pré-processamento e o roteiro de sacadas visitará sucessivamente estas regiões.

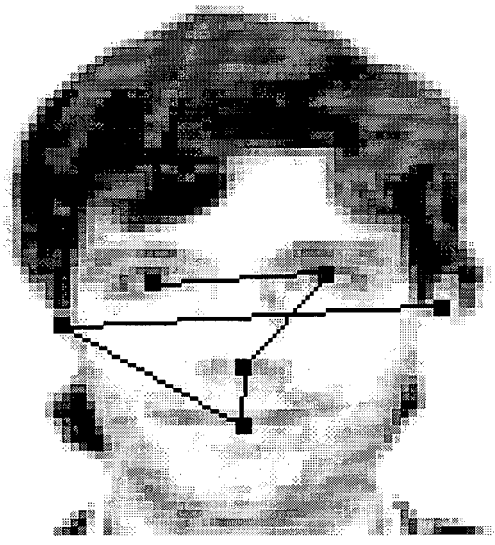

Figura 4.1: **Roteiro fixo:** regiões previamente escolhidas como importantes são alvos de sacadas. No exemplo acima, foram escolhidos os olhos, o nariz, a boca e as duas orelhas.

Nossa investigação voltou-se para o reconhecimento de objetos não restritos a um determinado tipo, e limitamos nosso estudo aos três últimos grupos de estratégias (já citadas: estratégias baseadas na imagem, estratégias baseadas nos modelos e estratégias híbridas), que serão analisadas cuidadosamente nas seções seguintes.

Uma estratégia atencional cria um *Mapa de Saliência*, construído como mostrado na Seção 3.4 do Capítulo 3, (ver Figura 4.2) onde a ativação S_p do mapa na posição p expressa quão relevante é esta posição na imagem dada. Em particular, a maior ativação no mapa determina o centro do próximo ponto de fixação. O Mapa de Saliência não é, contudo, uma estrutura estática, calculada uma vez e associada com um roteiro fixo de sacadas. Ao contrário, é uma estrutura dinâmica, que é atualizada após cada sacada. Deste modo, ele efetivamente leva em consideração o conjunto de pontos visitados, ou a *história sacádica*, pois a localização da fóvea no tempo t dependerá diretamente da localização desta no tempo $t-1$ e indiretamente de todas as localizações prévias.

Uma outra propriedade importante do Mapa de Saliência é que ele é também uma estrutura com resolução variável como a representação da imagem (ver detalhes desta representação no Capítulo 3). Os valores de S_p são calculados em alta resolução para a região foveal central (isto é, um disco em torno do ponto de fixação), e em resoluções menores para três anéis em torno da fóvea. Como será discutido em detalhe nesta seção, em geral o mapa S_p será calculado levando em conta o mapa de respostas das células complexas da imagem dada, o conjunto de modelos armazenados e o conjunto de ativações da saída do *Sistema de Decisão*, que determina a hipótese do sistema sobre qual a unidade de saída que melhor codifica a imagem de entrada num dado momento (Figura 4.2).

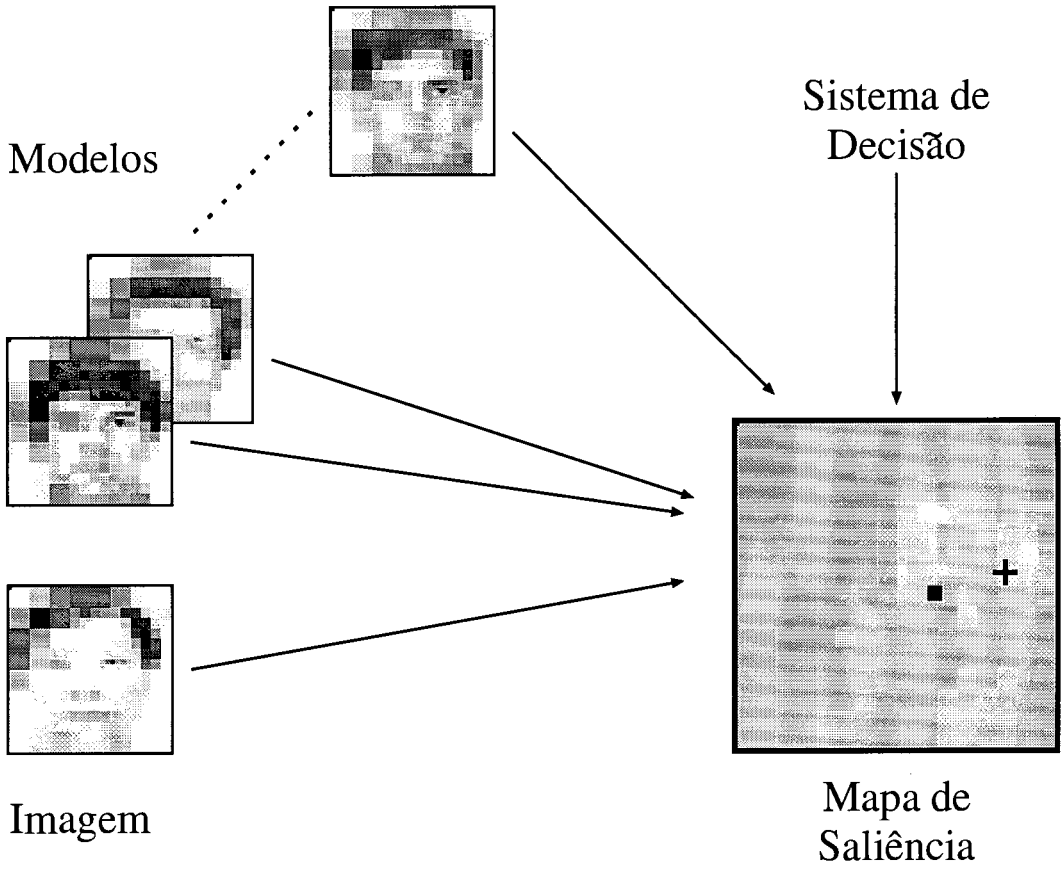

Figura 4.2: **Mapa de Saliência.** O elemento central do sistema atencional é o Mapa de Saliência, que em geral é computado levando em conta o mapa de respostas das células complexas da imagem apresentada, o conjunto de modelos armazenados, e sinais provenientes do Sistema de Decisão. O Mapa de Saliência é uma estrutura dinâmica que é atualizada após cada sacada. Note que o mapa é construído baseado no ponto de fixação atual (ponto preto) e determina o próximo ponto de fixação (cruz) correspondendo à maior ativação no mapa. A saliência é obtida dos mapas multi-escalares das repostas das células complexas e não das imagens originais como mostrado apenas para ilustração. (Estas imagens são parte da "MIT Eigenfaces database", de Turk e Pentaland, 1991 [62].)

4.1 Busca do próximo ponto

Uma vez construído o Mapa de Saliência para um determinado ponto de fixação, o próximo ponto pode ser determinado achando-se a posição do pixel de maior valor. Como o Mapa de Saliência é calculado para os diferentes níveis de resolução (diferentes níveis da pirâmide), o seu máximo será encontrado em um determinado nível (ver Figura 4.3). Cada nível da pirâmide tem, para o processo de reconhecimento em curso, um significado diferente, pelo menos segundo dois aspectos.

Primeiro, os níveis de menor resolução contribuem para o Mapa de Saliência com anéis mais distantes do ponto de fixação atual, portanto se a maior saliência for encontrada em um destes níveis, resultará em uma sacada para um ponto mais distante. Isto pode ser vantajoso, posto que pode facilitar uma abrangência maior

da imagem, e a experiência mostrou que isto leva a resultados melhores do que um traçado de sacadas restrito a uma vizinhança limitada.

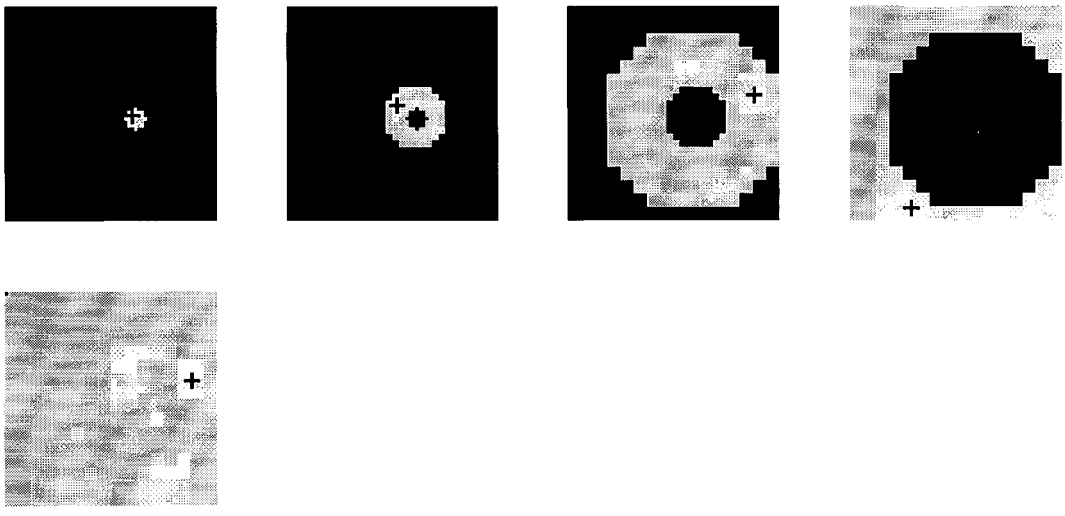

Figura 4.3: **Busca do próximo ponto de fixação.** Acima: Em cada anel de cada nível da pirâmide (da esquerda para a direita: níveis 0, 1, 2 e 3) há um ponto de máxima saliência (cruz). O máximo destes máximos será a posição da próxima sacada, e pode pode ocorrer em qualquer nível (no exemplo, ocorreu no nível 2). Nos níveis de mais baixa resolução corresponderão a pontos mais afastados do ponto de fixação atual. Este ponto será convertido para uma posição no nível zero da pirâmide, no centro da região representada pelo pixel do nível em que ocorreu este máximo. Embaixo: visualização do Mapa de Saliência, com a posição do ponto onde ocorreu o máximo. Este será o próximo ponto de fixação.

Segundo, como vimos no Capítulo 2, o reconhecimento envolve a solução do problema de conseguir que imagens diferentes de um mesmo objeto sejam associadas à representação (ou modelo) deste objeto previamente armazenado no sistema que faz o reconhecimento. Por maior que seja o número de imagens (instâncias) de um mesmo objeto usadas para construir um (ou vários) modelos do objeto, sempre haverá, em princípio, uma diferença entre a instância particular do objeto apresentada ao sistema para reconhecimento e o modelo armazenado. Mas esta diferença aparece com menor intensidade nos níveis mais baixos da pirâmide, onde as imagens são mais "borradas". Por outro lado, o reconhecimento só é possível se houver suficiente diferença entre a representação do objeto e os modelos dos outros objetos.

Trata-se, portanto, de encontrar uma solução de compromisso adequada entre duas tendências opostas: 1) Representações de alta resolução acentuam as diferenças em relação aos modelos dos outros objetos, facilitando a discriminação, mas tem o custo de aumentar a diferença em relação ao modelo correto; 2) Por outro lado, menores resoluções diminuem a diferença entre a representação do objeto e seu modelo, mas tem o custo de também diminuir a diferença em relação aos outros modelos, dificultando a discriminação.

Este problema não parece ser tão importante em um sistema que utilize toda a imagem para o reconhecimento, mas ele se agrava num modelo incremental como o apresentado aqui, onde as informações são parciais e acumuladas gradativamente. Em princípio, o potencial de discriminação varia de região para região ao longo da

imagem, e uma boa escolha do próximo ponto será aquela que encontra uma região na qual a representação do objeto seja mais semelhante ao modelo correto e mais diferente dos outros modelos. Assim, os diversos níveis de resolução terão papéis diferentes nesta escolha.

Um fenômeno que pode ocorrer, dependendo da estratégia e dos parâmetros usados, é o "desbalanceamento" entre os diferentes níveis do Mapa de Saliência, isto é, os valores calculados podem ser sistematicamente maiores em um dos níveis, por exemplo o nível de maior resolução. Neste caso as sacadas serão sempre direcionadas para pontos muito próximos, e o sistema pode não conseguir uma abrangência adequada, computando somente informações provenientes de uma área restrita. Muitas simulações mostraram que quando isto acontece o desempenho do sistema não é bom, e os valores de discriminação alcançados são menores do que quando o Roteiro de Sacadas percorre uma área maior (ver Figura 4.4).

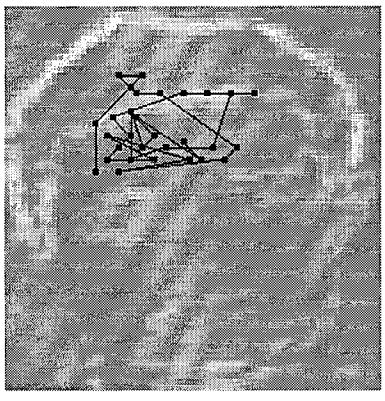

Figura 4.4: **Roteiro de Sacadas restrito:** O predomínio do nível de maior resolução no Mapa de Saliência produz sacadas muito próximas umas das outras, abrangendo uma área restrita da imagem.

De um modo geral, esta tendência ao desbalanceamento aponta para maiores valores de saliência nos níveis de maior resolução, pois nestes níveis as imagens são menos borradas e as diferenças entre os modelos tendem a ser maiores. Uma normalização durante o processo de filtragem (pre-processamento) tende a amenizar este efeito. Podem-se também normalizar os diversos níveis, com o mesmo fim, durante a construção do Mapa de Saliência. Uma solução simplificada adotada foi enfatizar os valores do nível de menor resolução a cada 4 sacadas, dobrando seus valores, o que resultou num Roteiro de Sacadas mais bem distribuído.

Finalmente, o Modelo emprega um dispositivo para evitar sacadas para pontos já visitados ou muito próximos destes. Isto é necessário por três razões: primeiro, porque as informações decorrentes de uma fixação em um ponto já visitado já foram extraídas e incorporadas ao sistema, e não contribuem com informações novas. Segundo, porque a volta a um ponto já visitado poderia redundar num processo cíclico (em "loop"), isto é, o roteiro de sacadas percorreria novamente todos os pontos já visitados. Embora se encontre na literatura que o sistema visual humano tende a roteiros de sacadas cíclicos (Noton e Stark, 1971 [39]), não é nosso objetivo neste trabalho investigar este fenômeno. Em terceiro lugar, uma informação redundante deste tipo poderia causar um desequilíbrio no processo de reconhecimento, por

contribuir indevidamente para ativar a categoria favorecida pelas informações provenientes do ponto repetido. Por razões análogas, um ponto muito próximo a um ponto já escolhido, digamos, um ponto dentro da região abrangida pela "fóvea", também não deve servir como alvo para uma outra sacada.

O modo adotado para impedir este comportamento cíclico, na nossa implementação, foi inibir permanentemente no mapa de saliência pequenas regiões em torno de pontos de fixação já visitados. O tamanho (diâmetro) da região inibida é um parâmetro do sistema, e pode ser ajustado facilmente. Este dispositivo pode ser visto como uma simulação (simplificada) do fenômeno de "inibição de retorno" [47] presente no sistema visual humano [1]. Esta inibição evita que qualquer sacada futura venha a incidir sobre um ponto já selecionado ou muito próximo.

Este conceito de evitar sacadas para uma região próxima a um ponto de fixação já visitado traz uma vantagem adicional ao ser implementado: é possível eliminar do Mapa de Saliência a região correspondente à "fovea", isto é, a região circular extraída do nível de maior resolução (nível 0, o mais extenso) da pirâmide. Isto é possível porque se a maior saliência for encontrada neste nível, o próximo ponto de fixação será na região abrangida pela fóvea, e trará informações muito semelhantes às já obtidas na sacada corrente. É claro que isto só poderia ocorrer se a região inibida for menor que a fóvea, o que nem sempre acontece. Mas se houver de fato um ponto de grande significado numa região não inibida da fóvea, espera-se que ele apareça também no nível seguinte (nível 1) da pirâmide. O Mapa de Saliência será construído, então, utilizando apenas os níveis 1, 2 e 3 da pirâmide (ver Figura 4.5). A busca do máximo do Mapa de Saliência será, portanto, feita apenas na extensão das regiões destes três níveis onde foi calculada a "função de saliência". Se a função de saliência fosse calculada para toda a extensão destes três níveis, isto corresponderia a 31,8% da extensão total do nível 0 (no caso de uma pirâmide de 4 níveis com fator de escala de 2 entre cada nível). Há, portanto, uma economia de processamento, que será melhor analisada no Capítulo 6, dedicado à avaliação do Modelo.

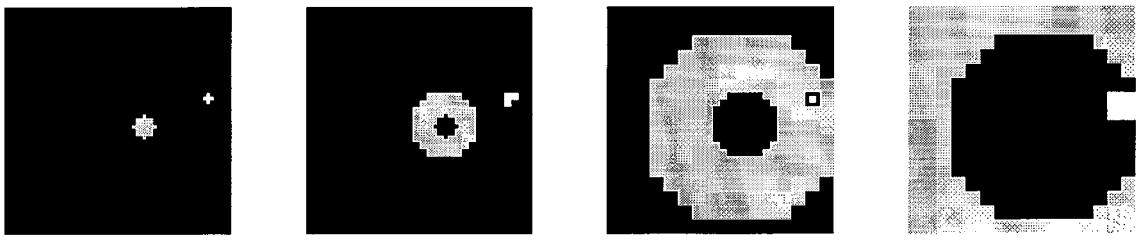

Figura 4.5: **Região excluída:** Os pontos já visitados ficam marcados no Mapa de Saliência, impedindo que as regiões de qualquer nível (da esquerda para a direita, níveis 0, 1, 2 e 3) correspondentes a este ponto sejam alvo de novas sacadas. Na figura o ponto correspondente a uma sacada anterior está marcado nos três níveis. Este ponto está localizado no anel do nível 2 da sacada atual (quadrado preto), mas é marcado também nos outros níveis, para impedir que este ponto seja eleito novamente por efeito de um valor máximo de saliência encontrado no nível 2 ou em qualquer outro nível. A região a ser excluída correspondente ao ponto de fixação atual, no centro da fóvea, ainda não foi marcado.

[1] nos sistemas biológicos esta inibição não é permanente, mas decai com o tempo.

4.2 Estratégias Baseadas na Imagem ("Bottom-Up")

Várias influências podem operar na orientação das sacadas no sistema visual humano. Como vimos no Capítulo 2, os mecanismos baseadaos na imagem, ou *bottom-up* operam através de sistemas inatos, com alto grau de paralelismo, rapidez, e independentes de esforço consciente1985 [59, 30]. Nestes, o tempo de latência não depende da complexidade da imagem, sugerindo um processamento paralelo no cérebro. Estes mecanismos tem a finalidade de encontrar certos padrões ou mudanças de textura na cena, e são menos dependentes de processos cognitivos e intencionais. Uma estratégia de escolha do ponto de fixação que dependa únicamente de características da imagem ("bottom-up") pode ser interpretada como uma forma de modelar este tipo de mecanismo inato.

No Modelo apresentado aqui, a forma mais direta de simular este tipo de estratégia é tomar a própria representação da imagem de entrada como Mapa de Saliência. De acordo com o tipo de pré-processamento que estamos utilizando, isto é, a soma das respostas aos filtros orientados simulando as células complexas do sistema visual humano, isto significa que os pontos de maior saliência serão os de maior contraste orientado na imagem de entrada (ver Figura 4.6). Isto equivale a dizer que a atenção seria atraída para os pontos de maior contraste na imagem.

Figura 4.6: **Contraste orientado:** Esquerda: imagem original. Direita: resposta dos filtros orientados simulando as células complexas do sistema visual humano, correspondente ao nível de maior resolução.

Neste caso, a função de saliência será idêntica à representação da imagem de entrada:

$$S_{n,p} = C_{n,p}\,, \tag{4.1}$$

onde $S_{n,p}$ é o valor do Mapa de Saliência na posição $p \in anel_n$, $n = 1, 2, 3$ e $C_{n,p}$ é a resposta silmulada das células complexas para a imagem I nesta posição e nível [2]

[2]Conforme definição no Capítulo 3, $anel_n$ é a região do nível n da pirâmide que participa da representação para um determinado ponto de fixação.

(ver Figura 4.7). Passamos a analisar a seguir as vantagens e desvantagens deste tipo de estratégia.

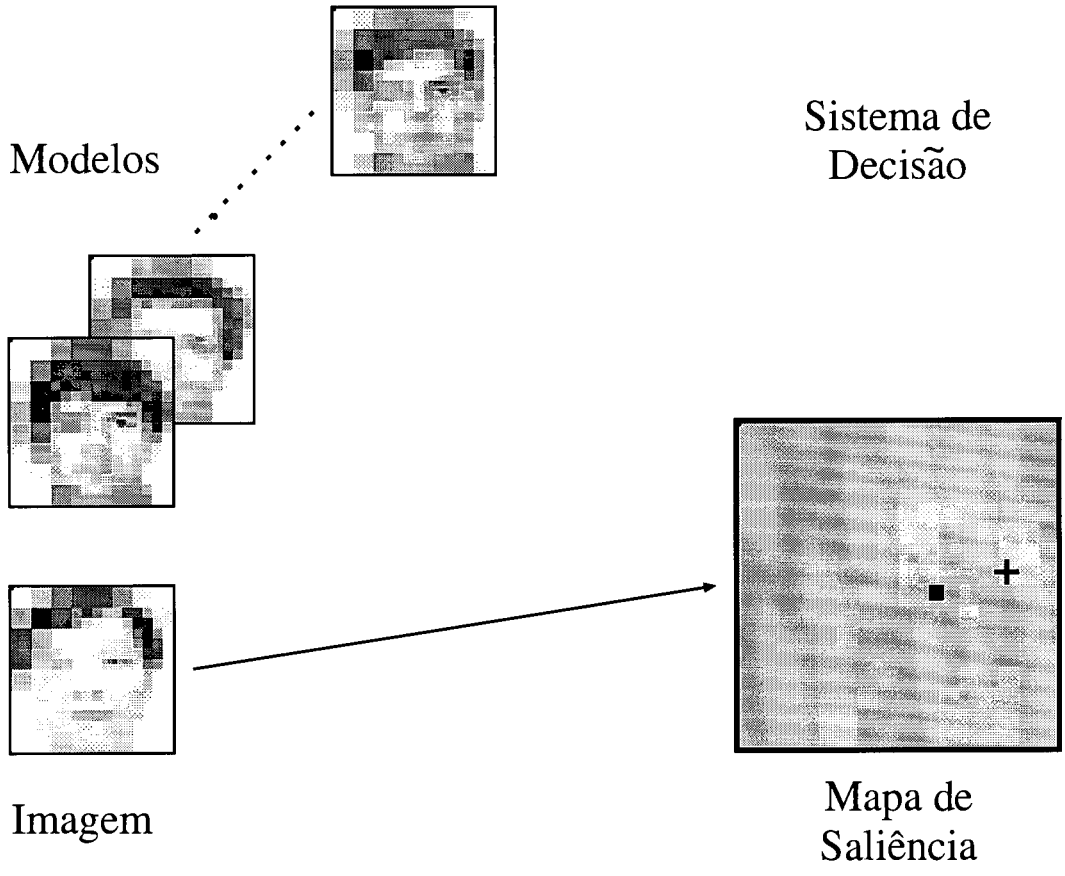

Figura 4.7: **Construção do Mapa de Saliência para estratégia baseada na imagem:** Somente a imagem participa da formação do mapa. Note que o mapa é construído baseado no ponto de fixação atual (ponto preto) e determina o próximo ponto de fixação (cruz) correspondendo à maior ativação no mapa. A saliência é obtida dos mapas multi-escalares das repostas das células complexas e não das imagens originais como mostrado apenas para ilustração.

A principal vantagem é que as sacadas são atraídas para regiões onde há contraste na imagem, e portanto mais informação do que em regiões de luminosidade uniforme. Em faces, por exemplo, as regiões de contorno, onde há contraste com o fundo, ou os olhos, onde há detalhes expressos por mudanças de tons de cinza, ou as regiões de contraste entre o cabelo e a testa, etc., serão mais salientes. Outra vantagem é que, quando um objeto aparece contra um fundo relativamente neutro, as sacadas serão atraídas para regiões do objeto, e não do fundo. Este efeito é particularmente importante quando se trata do reconhecimento de um objeto pertencente a um conjunto relativamente disperso, isto é, quando há diferenças significativas entre os objetos do conjunto, porque neste caso os contornos que aparecerão nas representações que formarão os modelos armazenados serão também bastante diferentes, contribuindo para diferenciar os objetos. Este é o caso do conjunto que chamamos de "base de dados de Colúmbia" formada por imagens obtidas na base de dados da Universidade de Colúmbia, utilizada inicialmente por Murase e Nayar [38] (ver

Figura 5.1 no Capítulo 5).

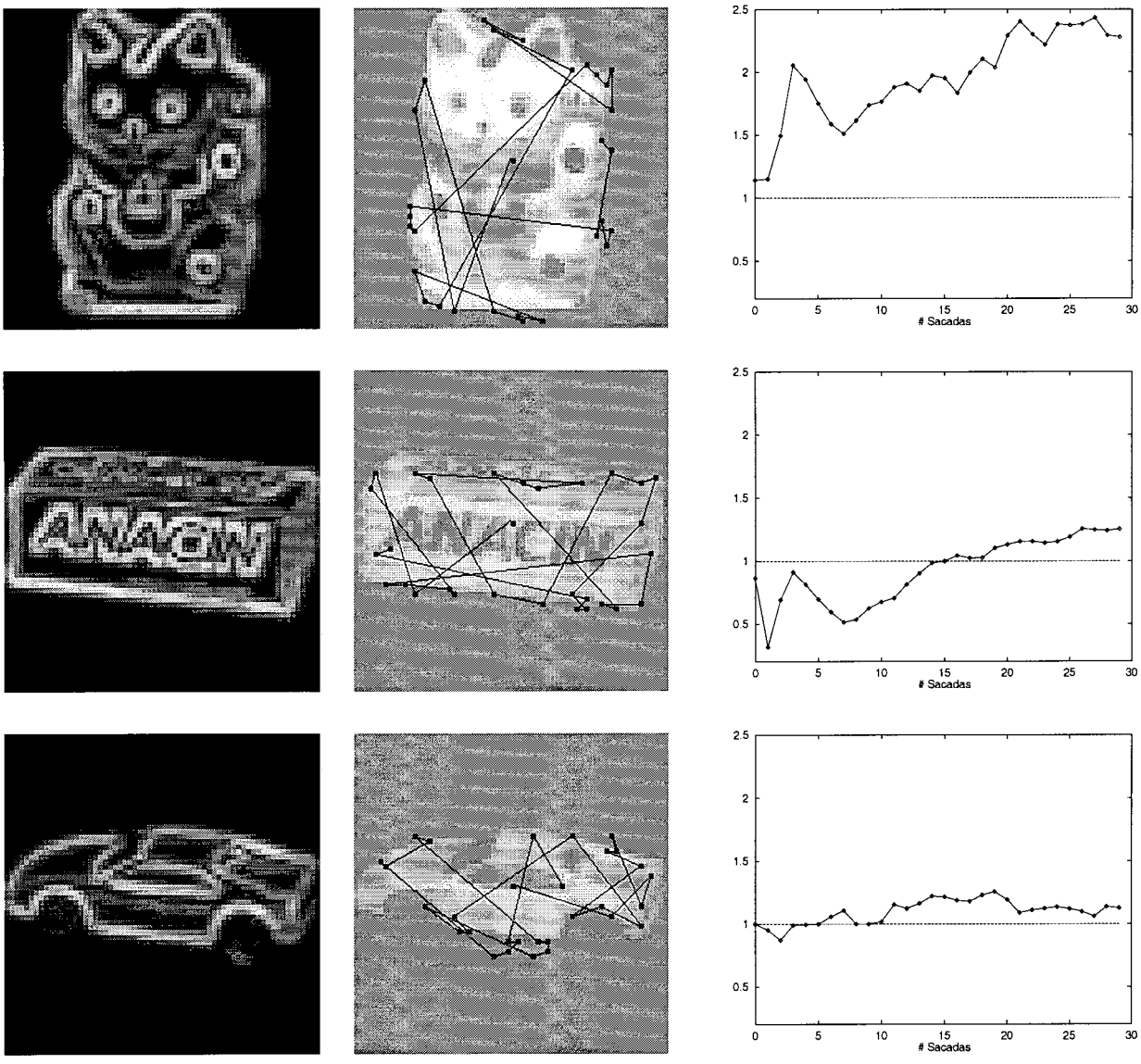

Figura 4.8: **Estratégias baseadas na imagem.** Esquerda: roteiro de sacadas para três objetos apresentados para reconhecimento. Direita: gráficos de discriminação da categoria correta (razão entre a ativação da categoria correta e a segunda maior ativação) em função das sacadas (os valores acima de 1.0 significam reconhecimento correto).

A Figura 4.8 mostra os traçados obtidos quando são usadas só informações da imagem para guiar as sacadas. Os objetos apresentados para reconhecimento estão em uma posição rotacionada de 10 graus em relação a seus modelos (pertencentes a um conjunto de 8 modelos). Entre os outros modelos há objetos com diferentes níveis de semelhança com as imagens apresentadas. É possível observar que as sacadas se mantém na região ocupada pelo objeto, ou bem próximas dele (lembrar que, neste exemplo, nos níveis mais baixos da pirâmide, onde a resolução é muito baixa, um ponto da borda do objeto será traduzido, no nível 0, por um ponto no centro de uma região 8×8, ficando aparentemente fora da borda). Na Figura 4.8 os gráficos de discriminação da categoria correta (razão entre a ativação da categoria correta e

a segunda maior ativação) em função do número de sacadas (Figura 4.8, direita) [3] mostram, para os dois primeiros objetos, que após um período transiente inicial, a discriminação cresce, alcançando seus maiores valores depois de 25 sacadas. Note que a hipótese do sistema baseada na primeira sacada está incorreta para o segundo objeto, e somente após 15 sacadas o modelo correto é identificado. Os resultados são melhores no primeiro caso (alto), de um objeto mais discriminável nesta base de dados. No entanto, efetuar sacadas ao longo dos contornos do objeto, como determinado por esta estratégia pode ser um método fraco de extrair incrementalmente informações. A Figura 4.8 (parte de baixo) mostra que para o caso do carro de brinquedo, a discriminação da categoria correta permanece mais ou menos constante enquanto o sistema efetua as sacadas (embora o sistema encontre corretamente a categoria). Em outras palavras, há um *ganho* muito pequeno ao extrair informações ao longo dos contornos mais fortes do objeto. Note que esta base de dados contém alguns objetos bastante semelhantes a este, tornando-o menos discriminável, pelo menos segundo esta estratégia.

As desvantagens desta estratégia se mostram mais evidentes em uma base de dados mais densa, por exemplo com imagens como faces, onde todos os objetos tem aproximadamente o mesmo contorno. Neste caso, as regiões de contorno não ajudam a discriminar entre os objetos, pois muitos modelos compartilham os contornos que produzem respostas mais intensas dos filtros orientados, e o desempenho com esta etratégia isolada não apresenta resultados muito bons. Simulações mais extensas (apresentadas no Capítulo 5) com a base de dados de Colúmbia e com a base de dados do Olivetti Research Laboratory (ORL) [4], de faces, mostram que a estratégia baseada na imagem dá melhores resultados comparativamente em uma base de dados dispersa (Colúmbia) do que em uma base de dados de faces (mais densa) como a do ORL. Como o mapa das respostas das células complexas tende a apresentar valores relativamente altos em todo o contorno dos objetos, o sistema tende a produzir um Roteiro de Sacadas relativamente abrangente. Esta abrangência é benéfica para o reconhecimento, por levar à extração de informações de várias regiões distintas da imagem. Mas ela por si só não contempla nenhum critério para encontrar pontos relevantes para a discriminação entre os modelos, no caso de objetos com contornos parecidos. No outro caso (base de Colúmbia), com objetos de contornos diferentes, o contorno por si só já é um elemento de discriminação, o que promove resultados satisfatórios.

[3] Lembrar que, conforme definido no Cap. 3, o valor da discriminação é uma medida da confiança do sistema na hipótese corrente. Note que um valor da discriminação da categoria correta (ver equação 3.6) maior que 1 mostra que a unidade mais ativa corresponde à categoria correta, enquanto que um valor igual ou menor que 1 indica que a hipótese do sistema para a categoria correspondente ao objeto apresentado está incorreta.

[4] A base de dados de faces ORL pode ser obtida em http://www.camorl.co.uk/facedatabase.html.

4.3 Estratégias Baseadas nos Modelos ("Top-Down")

Os mecanismos de orientação das sacadas que levam em conta o conhecimento préviamente acumulado, nos sitemas visuais biológicos, são chamados, talvez mais adequadamente que os outros, de " atencionais". Eles pressupõem mecanismos eventualmente decorrentes da tarefa visual atual (ver Yarbus 1967 [66]). A busca do melhor ponto de fixação pode, por exemplo, ser guiada por algum tipo de espectativa derivada da necessidade de confirmar ou desconfirmar a existência de algum aspecto importante (para a tarefa em curso) no objeto analisado. Neste caso o comportamento atencional pode ser considerado um comportamento cognitivo de pleno direito. Por outro lado, os comportamentos atencionais não são, provavelmente, decorrentes de um único mecanismo, mas de uma variedade de esquemas de alocação dos recursos disponíveis (ver Jansen 1996 [30]).

Segundo Jansen [30], a modelagem do comportamento de controle atencional voluntério encontra uma crescente demanda no domínio técnico. Em nosso Modelo, tentamos simular os mecanismos de alocação de sacadas levando em conta o "conhecimento" prévio do sistema, representado aqui pelas categorias (ou modelos) armazenados.

A idéia básica aqui é encontrar regiões em que haja maior variação entre as representações dos modelos, como nas estrategias propostas no trabalho de Aguilar e Ross, 1993 [2]. Informações extraídas de regiões onde os modelos mais se assemelham, pouco contribuiriam para discriminar a categoria correta. Num caso extremo, pode-se pensar que um fundo uniforme e não informativo é exatamente o mesmo para todos os modelos, e uma sacada para posições deste fundo não contribuiriam em nada para o reconhecimento. Por outro lado, sacadas para posições onde os modelos são mais diferentes tem um grande potencial de contribuir para o processo de reconhecimento.

Numa estratégia estritamente baseada nos modelos ("top-down"), as sacadas serão guiadas exclusivamente pelas informações armazenadas no sistema (ver Figura 4.9). Neste caso, as informações provenientes da imagem estão servindo apenas para calcular as ativações das categorias no Sistema de Decisão, e não produzem nenhuma interferência na orientação das sacadas.

A variação entre as representações dos modelos em um determinado ponto pode ser medida de várias maneiras (incluindo variância ou desvio padrão). Por simplicidade, estamos chamando de "variação", para cada modelo, o valor absoluto da diferença entre o valor do modelo em um ponto e um valor de referência neste ponto. O Mapa de Saliência será calculado neste caso somando as diferenças entre as representações dos modelos e algum valor de referência. Este pode ser, por exemplo, a média entre os modelos para cada ponto.

$$S_{n,p} = \sum_m |M_{m,n,p} - \overline{M}_{n,p}|,$$

onde $S_{n,p}$ é o valor do Mapa de Saliência na posição $p \in anel_n$, $n = 1, 2, 3$; $M_{m,n,p}$ é a resposta silmulada das células complexas para o modelo M_m nesta posição; e $\overline{M}_{n,p}$ é a média dos valores dos pixels de todos os modelos na mesma posição e nível.

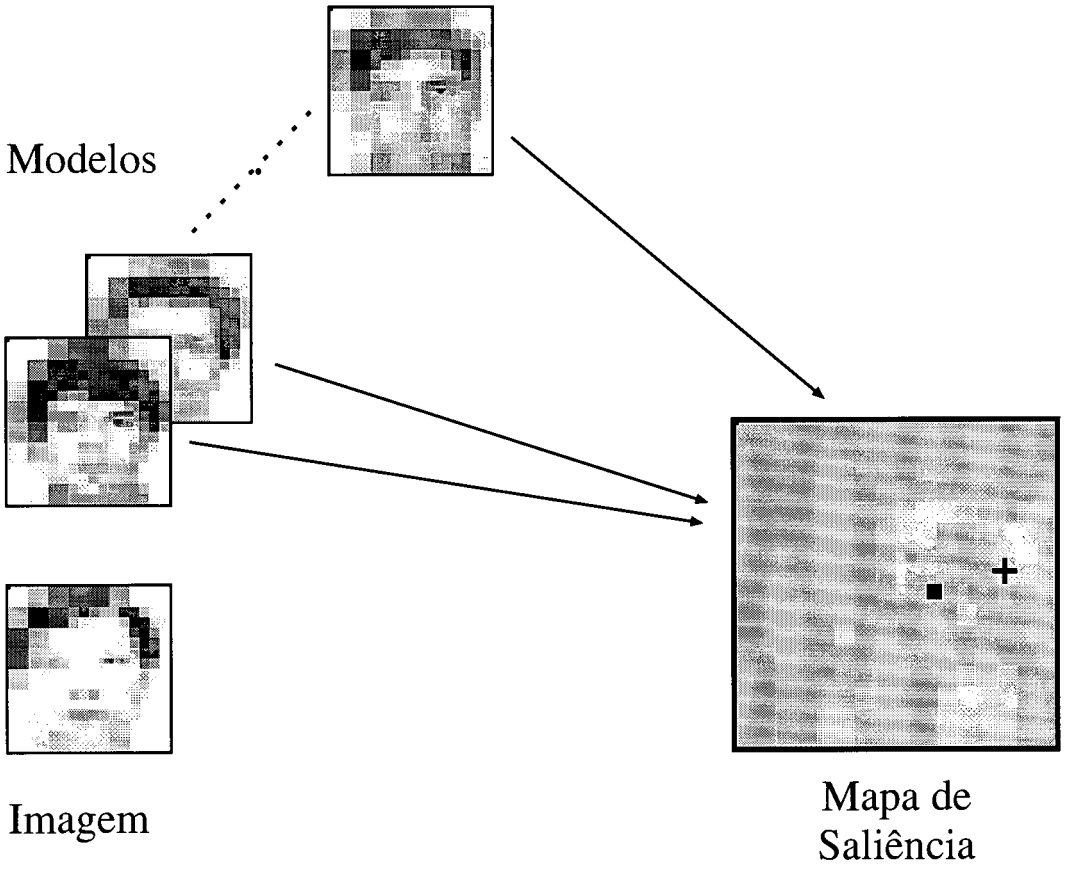

Figura 4.9: **Construção do Mapa de Saliência para estratégia estritamente baseada nos modelos ("top-down"):** Somente os modelos participam da construção do mapa. Note que o mapa é construído baseado no ponto de fixação atual (ponto preto) e determina o próximo ponto de fixação (cruz) correspondendo à maior ativação no mapa. A saliência é obtida dos mapas multi-escalares das repostas das células complexas e não das imagens originais como mostrado apenas para ilustração.

A Figura 4.10 mostra o Roteiro de Sacadas gerado por uma estratégia baseada na memoria para uma imagem de face. Neste caso, os pontos visitados são os que apresentam maior variação nas respostas orientadas para o conjunto de 12 faces armazenadas (todas na mesma escala e em vista frontal). O valor da discriminação oscila inicialmente, depois cresce rapidamente e torna-se assintótico depois de 20 a 25 sacadas. O objeto foi reconhecido corretamente e com boa discriminação neste caso, no qual houve um bom desempenho do sistema.

Deve-se notar que, numa estratégia estritamente baseada na memória, o Roteiro de Sacadas é fixo e exatamente o mesmo para qualquer imagem apresentada, porque os pontos a serem visitados são determinados somente considerando a base de dados dos modelos armazenados. Enquanto esta estratégia pode se revelar razoável para uma base de dados onde os objetos são muito semelhantes (faces por exemplo), ela pode ser desastrosa para casos mais heterogêneos. A Figura 4.11 ilustra esta situação. Para os três objetos mostrados, o mesmo Roteiro de Sacadas é seguido. Dado que os modelos armazenados representam 20 objetos bastante diferentes, para

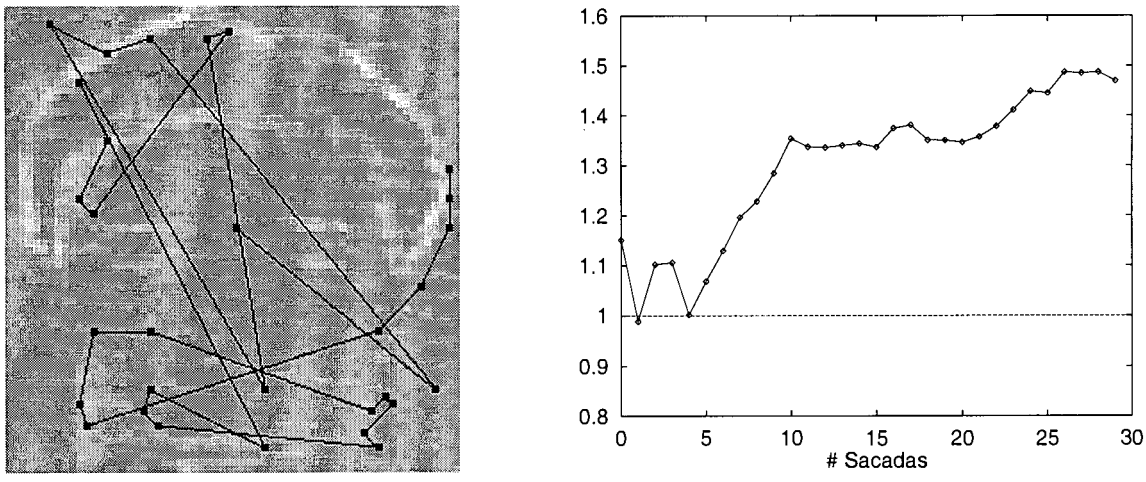

Figura 4.10: **Estratégia baseada nos modelos.** Esquerda: Roteiro de Sacadas; direita: evolução do valor da discriminação da categoria correta.

qualquer imagem apresentada o Roteiro de Sacadas incluirá muitas posições no fundo não informativo. Para a caixa de Anacin, o reconhecimento foi possível, apesar de fraco do ponto de vista incremental, pois a extração de mais informações a partir de 10 sacadas não melhorou a discriminação, enquanto para os outros dois objetos o desempenho foi muito fraco, e foram selecionadas categorias incorretas.

Pudemos observar a partir de muitas simulações, que o valor, para o processo de reconhecimento, das informações extraídas por esta estratégia não é muito consistente, isto é, varia de uma base de dados para outra. Há ainda o caso de um fundo não uniforme; porém, como aqui não contemplamos o problema de busca ou localização de um objeto em uma cena, a investigação restringe-se ao caso em que nas imagens apresentadas, assim como nos modelos armazenados no sistema, os objetos estão centrados e numa escala correspondente ao tamanho da imagem (ver, no Capítulo 5, as figuras que mostram as bases de dados utilizadas). Assim, o fundo não é diferenciado funcionalmente, isto é, nada no Modelo o distingue efetivamente do objeto.

De qualquer modo, em um modelo de reconhecimento ideal, pode-se esperar que as sacadas devam ser mais bem aproveitadas se dirigidas para regiões onde a imagem atual seja mais informativa. "Mais informativa", para ser coerente com o pré-processamento utilizado, seria aquela região onde há algum contraste orientado. Por outro lado, a alternativa baseada na variabilidade de todos os modelos, tem por trás a idéia de que as informações extraídas de uma região de grande variabilidade servirá para ativar mais a categoria correta.

Há, porém, uma observação interessante que pode explicar porque, em muitos casos, esta etratégia não leva a desempenhos muito bons: na maioria das simulações, é comum um número restrito de categorias se tornar mais ativado. Estas seriam as categorias que parecem se assemelhar mais à imagem de entrada em um determinado momento do processo, constituindo as hipóteses mais prováveis naquele momento, e o sistema tem que decidir entre estas qual a correta. Neste caso, a utilização da variabilidade de *todas as categorias* não ajuda muito, pois pode atrair as sacadas para regiões onde as categorias mais ativas (mais prováveis), não necessariamente se distinguem mais.

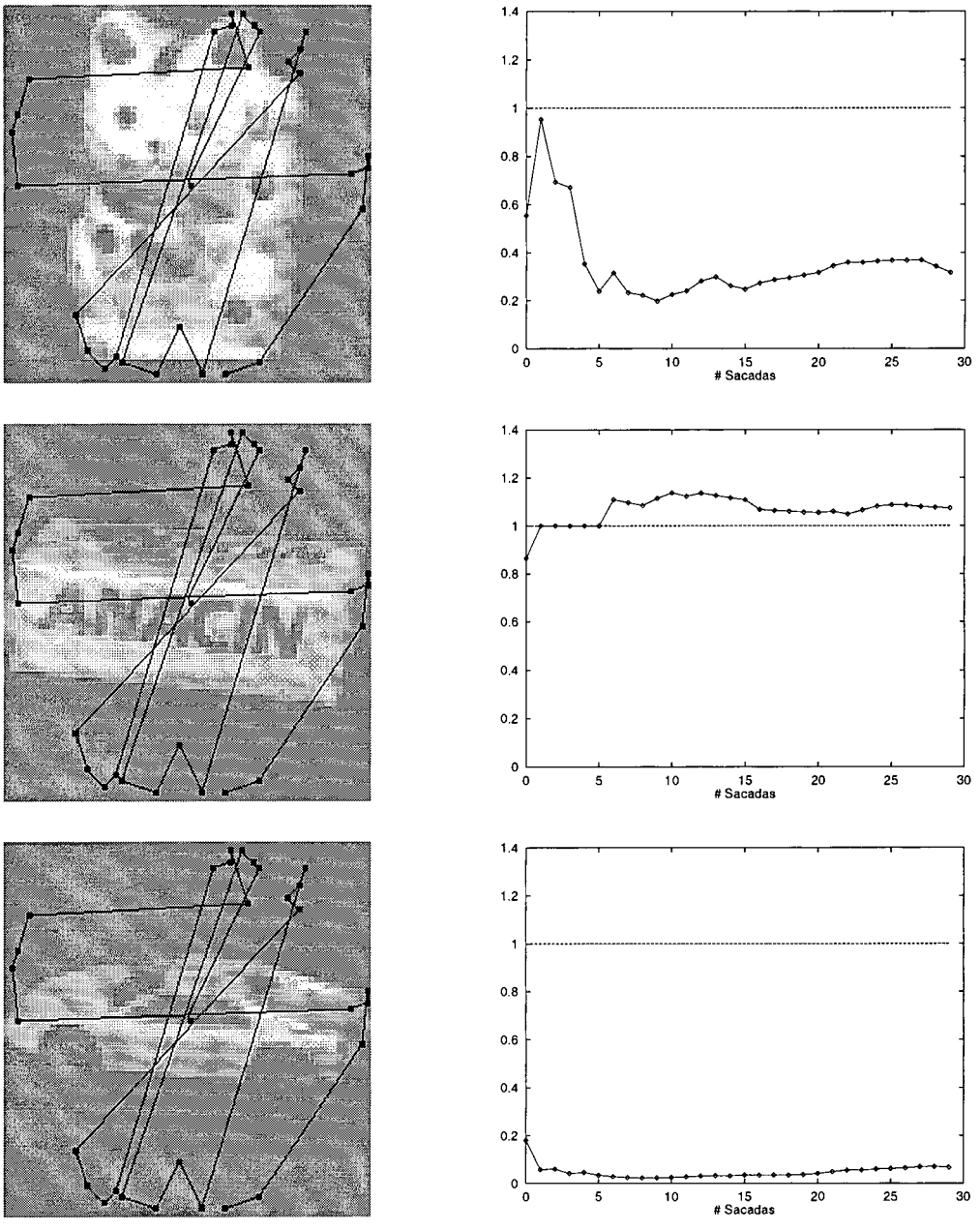

Figura 4.11: **Estratégia estritamente baseada nos modelos.** Esquerda: roteiro de sacadas para três objetos apresentados para reconhecimento (notem-se os roteiros idênticos). Direita: gráficos de discriminação da categoria correta em função das sacadas (os valores acima de 1.0 significam reconhecimento correto).

Estas considerações sugerem duas outras alternativas, onde se procura enfatizar as informações provenientes das categorias mais ativas ao construir o Mapa de Saliência. Para isso, é necessário utilizar os valores das ativações das categorias, fornecidas pelo Sistema de Decisão, como analisaremos nas seções seguintes. Note-se que estas outras alternativas devem ser consideradas híbridas, pois são usadas, ainda que de forma indireta, informações da imagem de entrada na construção do Mapa de Saliência.

4.4 Estratégias Híbridas

Enquanto uma estratégia puramente baseada nos modelos desconsidera inteiramente a imagem apresentada para guiar as sacadas, uma estratégia baseada na imagem não utiliza a informação em potencial contida no conjunto de modelos armazenados. Ambas poderão considerar localizações na imagem que não são muito adequadas para o reconhecimento. Por exemplo, a estratégia baseada nos modelos poderá considerar posições no fundo não informativo, enquanto a estratégia baseada na imagem poderá considerar contornos fortes que são compartilhados por todos os modelos armazenados (e por isso menos informativos para o processo de diminuir a ambiguidade). Por considerar tanto a imagem apresentada quanto as informações armazenadas, uma abordagem *híbrida* é potencialmente mais efetiva.

Uma estratégia híbrida poderá considerar informações provenientes da imagem diretamente para a constução do Mapa de Saliência usando, por exemplo, os valores do mapa das células complexas como na estratégia baseada na imagem, já descrita, ou poderá considerar a influência da imagem através da ativações causadas por esta na saída do Sistema de Decisão. Neste último caso esta ativação pode servir para modular a influência de cada modelo na construção do Mapa de Saliência. Este tipo de estratégia pode ser chamado de "estratégia híbrida indireta" por considerar a imagem apenas através da sua influência nas ativações dos modelos. No caso anterior, as informações provenientes da imagem contribuem por si próprias para a formação do Mapa de Saliência, dando origem às chamadas "estratégias híbridas diretas".

A seguir vamos analisar algumas das estratégias híbridas investigadas, começando pelas estratégias híbridas indiretas.

4.4.1 Estratégia híbrida indireta: uso de pesos para ponderar a importância de cada categoria

As ativações podem ser usadas como pesos para ponderar a contribuição de cada categoria, de modo que a variação entre as categorias mais ativas [5] seja considerada mais importante. Neste caso, todas as categorias contribuiriam para o cálculo da variabilidade, porém com pesos diferentes. Esta foi a estratégia adotada por Aguilar e Ross (1993) [2].

Aqui, as categorias mais ativas em um determinado momento são consideradas as mais prováveis, isto é, assume-se que há maior probabilidade de que a categoria correta fique mais ativa que as outras, pois ela representa a hipótese corrente do sistema para a categoria correta.Por isso, os pontos nos quais a variação destas categorias é maior seriam os pontos mais informativos, os que mais contribuiriam para dirimir dúvidas e discriminar a categoria correta.

O Mapa de Saliência (ver Figura 4.12), neste caso, será calculado pela soma ponderada das diferenças entre o valor da representação de cada modelo e o valor da representação com maior atividade no momento, para cada ponto, e o fator de

[5] Como vimos no Capítulo 3, a ativação de uma categoria é calculada pelo Sistema de Decisão como a atividade de uma unidade da saída F2.

ponderação é o valor da ativação de cada categoria no Sistema de Decisão:

$$S_{n,p} = \sum_m y_m |M_{m,n,p} - M_{v,n,p}|,$$

onde $S_{n,p}$ é o valor do Mapa de Saliência na posição p do anel de nível n $n = 1, 2, 3$, y_m é o valor da ativação da categoria m na saída F2 do Sistema de Decisão, $M_{m,n,p}$ é a resposta silmulada das células complexas para o modelo M_m nesta posição, e $M_{v,n,p}$ é a resposta silmulada das células complexas para o modelo mais ativo, na mesma posição.

Observe que a variação entre as categorias está sendo calculada usando-se a mais ativa como referência, e esta variação é modulada pela atividade de cada categoria. Isto faz com que as diferenças em relação às categorias pouco ativas, ou pouco prováveis, tenham menor influência na localização do próximo ponto de fixação, enquanto que as mais ativas terão maior influência. Deste modo, o Mapa de Saliência assim construído (ver Figura 4.12) dará destaque àquelas posições onde a categoria mais provável no momento, isto é, a mais ativa, apresenta maior diferença em relação às outras categorias também prováveis, que representam neste momento os modelos que parecem mais semelhantes ao modelo representado pela categoria mais ativa. Este procedimento procura, portanto, resolver a ambigüidade entre o modelo mais ativo e os que mais se parecem com ele, segundo os dados extraídos pelo sistema até aquele momento.

A escolha do modelo mais ativo como referência tem o efeito adicional de enfatizar mais os pontos onde este modelo é mais "especial", isto é, onde ele tem algo de característico em relação aos outros modelos. Isto se dá porque a soma das diferenças de um conjunto de valores em relação a um valor que se afasta da mediana do conjunto é maior do que em relaçao a um valor próximo desta. Se interpretamos que o valor de um pixel em um modelo afastado da mediana dos valores dos pixels de mesma posição nos outros modelos como uma característica especial deste modelo, esta posição no modelo mais ativo estará especial deste modelo, esta posição no modelo mais ativo estará sendo enfatizada pelo Mapa de Saliência assim construído.

Em comparação com a anterior, estritamente baseada nos modelos, esta estratégia mostra um desempenho sensivelmente melhor, com valores da discriminação bem mais altos e crescentes, exceto no caso do carro de brinquedo, onde a discriminação, embora mais elevada que nas outras estratégias, não apresentou o crescimento necessário para o reconhecimento. Os gráficos da Figura 4.13 (à direita) mostram a discriminação alcançada pela categoria correta (os valores acima de 1.0 significam reconhecimento correto) para três exemplos de objetos da base de dados de Colúmbia. Pode-se observar que nos casos em que o desempenho é melhor (primeiro e segundo objetos), os Roteiros de Sacadas (à esquerda) são mais coerentes com os objetos apresentados, apesar de ainda se encontrarem muitos pontos de fixação no fundo. Note que apesar de estar sendo usada uma estratégia que depende mais dos modelos, os Roteiros de Sacadas não são os mesmos para diferentes simulações porque a computação do Mapa de Saliência leva em conta as unidades mais ativas, e portanto indiretamente a imagem apresentada.

As desvantagens deste método são análogas às da estratégia anterior: a contribuição de todas as categorias, ainda que ponderadas, para a formação do Mapa de Saliência tende a mascarar o efeito de categorias que representam modelos que po-

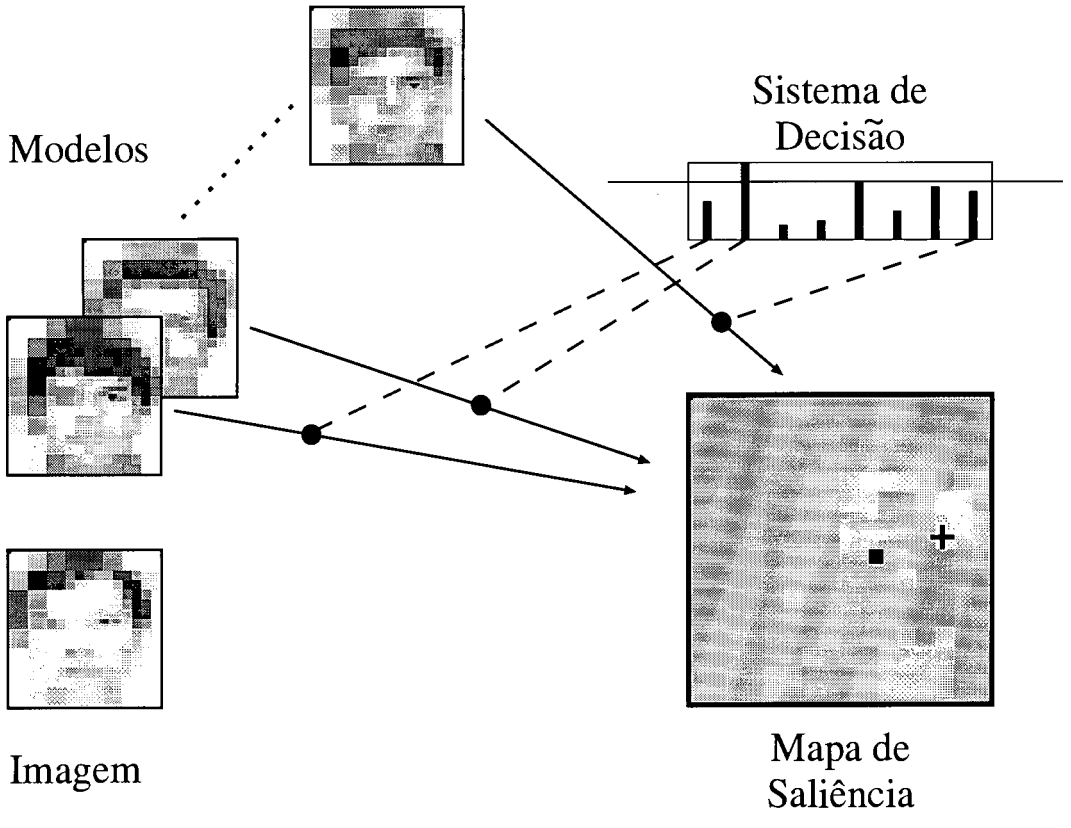

Figura 4.12: **Construção do Mapa de Saliência para estratégia híbrida indireta:** A influência de cada modelo é ponderada pela ativação atual no Sistema de decisão. Note que o mapa é construído baseado no ponto de fixação atual (ponto preto) e determina o próximo ponto de fixação (cruz) correspondendo à maior ativação no mapa.

dem ser mais parecidos com o correto. O Mapa de Saliência pode estar enfatizando uma região onde há grande variação entre os modelos em geral e o modelo mais ativo, mas onde não há grande diferença entre a o modelo correto e os mais semelhantes à imagem. O problema principal é que o sistema pode estar adotando uma hipótese errada, o que é frequente no início das sacadas, enfatizando uma categoria incorreta, enquanto que a correta pode estar com uma ativação baixa, influenciando pouco no Mapa de Saliência. Neste caso estará sendo buscada a região que serve para discriminar mais este modelo (como explicado acima) mas ele pode não ser o correto, e só por acaso estas regiões vão favorecer a ativação do modelo correto. Com um número maior de sacadas, quando mais informações já foram extraídas da imagem, este efeito indesejado tende a diminuir. Uma maneira mais "neutra", isto é, mais independente da hipótese corrente sobre qual é o modelo com mais probabilidade de ser o correto, de procurar pontos onde há maiores diferenças entre os modelos seria usar como referência a média dos valores dos pixels dos modelos em uma determinada posição. Vemos também que as sacadas podem ainda ser atraídas para regiões sem relação com a imagem apresentada. O desempenho melhor desta estratégia em relação à estratégia estritamente baseada nos modelos mostra que estas desvantagens foram bastante amenizadas aqui.

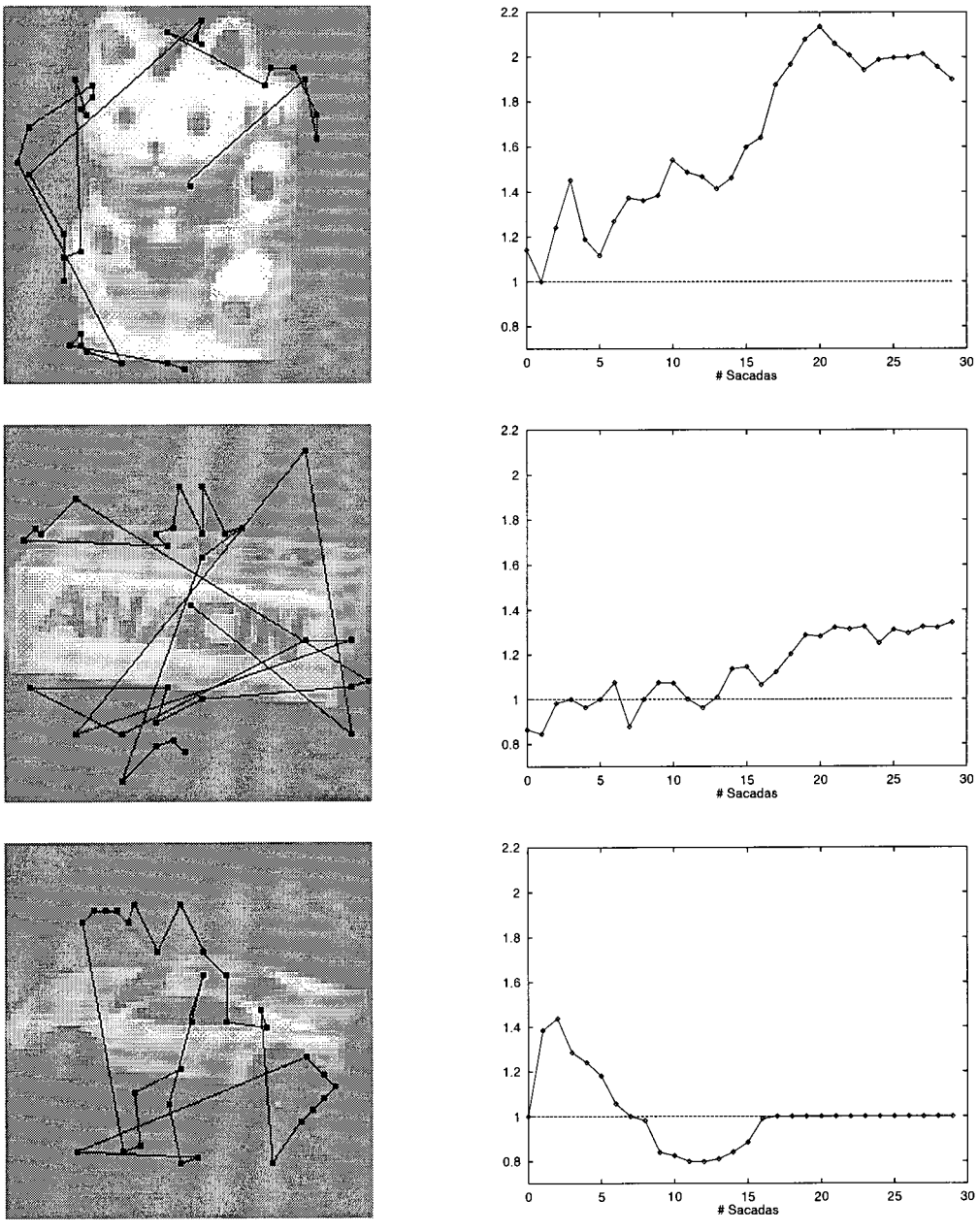

Figura 4.13: **Estratégia híbrida indireta, usando ativações como pesos para ponderar as variações entre as categorias mais ativas.** Esquerda: roteiro de sacadas para três objetos apresentados para reconhecimento. Direita: gráficos de discriminação em função das sacadas (os valores acima de 1.0 significam reconhecimento correto).

4.4.2 Estratégia híbrida indireta: Variação entre as categorias mais prováveis

Nesta estratégia, ao invés de considerar todos os modelos armazenados para a construção do Mapa de Saliência, somente aqueles mais ativos no Sistema de Decisão serão empregados (ver Figura 4.14). As unidades mais ativas indicam as categorias que tem maior probabilidade de ser a categoria à qual pertence a imagem apresentada, de acordo com as informações disponíveis no sistema em um dado momento.

O sistema utiliza, assim, aquelas categorias que apresentaram maior semelhança com a imagem apresentada, até aquele momento, para encontrar o próximo ponto de fixação. A idéia aqui é que, na comparação com os modelos armazenados no sistema, a imagem apresentada em geral é mais parecida com um número restrito de modelos, e as categorias correspondentes a estes modelos tendem a se tornar mais ativas, e também as que tem maior probabilidade de ser a correta. O sistema tem que decidir entre estas categorias, que tendem a ser as mais ambíguas em sua semelhança com a imagem apresentada, e discriminar qual a correta. O Mapa de Saliência será, nesta estratégia, construído de forma a enfatizar as regiões onde estas categorias mais ambíguas apresentam maior variação.

Evidentemente é necessário estabelecer um critério para calcular o limiar que divide o conjunto de modelos em um subconjunto das categorias mais ativas (H), e um subconjunto das menos ativas (L). Este limiar pode, por exemplo, ser o mesmo utilizado para definir o Critério de Decisão, que, como vimos, é o valor da discriminação necessária para considerar um objeto como reconhecido. Assim, se o Critério de Decisão for 1.30, teremos que a ativação da categoria vencedora é 30% maior que a da segunda mais ativa, e estaríamos, neste caso, considerando que não há ambiguidade entre as duas mais ativas. Enquanto o Critério de Decisão não é atingido, podemos interpretar este fato como incerteza do sistema entre as categorias mais ativas, sendo estas as que tem ativação maior que aproximadamente 77% (1/1.30) da mais ativa no momento (neste exemplo). Na maior parte das simulações foi utilizado o limiar de 80% da categoria mais ativa. Caso não exista nenhuma categoria com ativação acima deste limiar em relação à mais ativa, são consideradas as duas categorias mais ativas para formar o conjunto H.

Com este critério o calculo do Mapa de Saliência será dado pela expressão:

$$S_{n,p} = V^H_{n,p} = \sum_{m \in H} |M_{m,n,p} - \overline{M}^H_{n,p}|,$$

onde $S_{n,p}$ é o valor do Mapa de Saliência na posição p do anel do nível n, $n = 1, 2, 3$; $V^H_{n,p}$ é a variação das categorias pertencentes ao conjunto H das categorias de alta ativação na posição p do nível n; $M_{m,n,p}$ é a resposta silmulada das células complexas para o modelo m na posição p do nível n; e $\overline{M}^H_{n,p}$ é a média dos valores dos pixels de todos os modelos do conjunto H na posição p do nível n.

A Figura 4.15 mostra o comportamento do sistema para as mesmas imagens investigadas antes. Em cada simulação, os modelos associados às unidades com ativações maiores que 80% da mais ativa foram usados para construir o Mapa de Saliência. Para estas imagens, a discriminação cresce, chegando a altos níveis no final do processo. Note que os Roteiros de Sacadas não são os mesmos para diferentes simulações devido à influência da imagem apresentada na construção do Mapa de Saliência.

Esta estratégia tem se revelado uma das melhores quando testada tanto em uma base de dados densa, como a base de faces do ORL, como em uma base mais dispersa como a base de Colúmbia, como se pode ver pelos resultados quantitativos das simulações apresentadas no Capítulo 5. Aqui as categorias mais ativas são levadas em consideração de uma maneira mais "neutra" do que a empregada na estratégia anterior, o que tende a atenuar o efeito nocivo de uma alta ativação em uma categoria incorreta associada a uma baixa ativação da categoria correta.

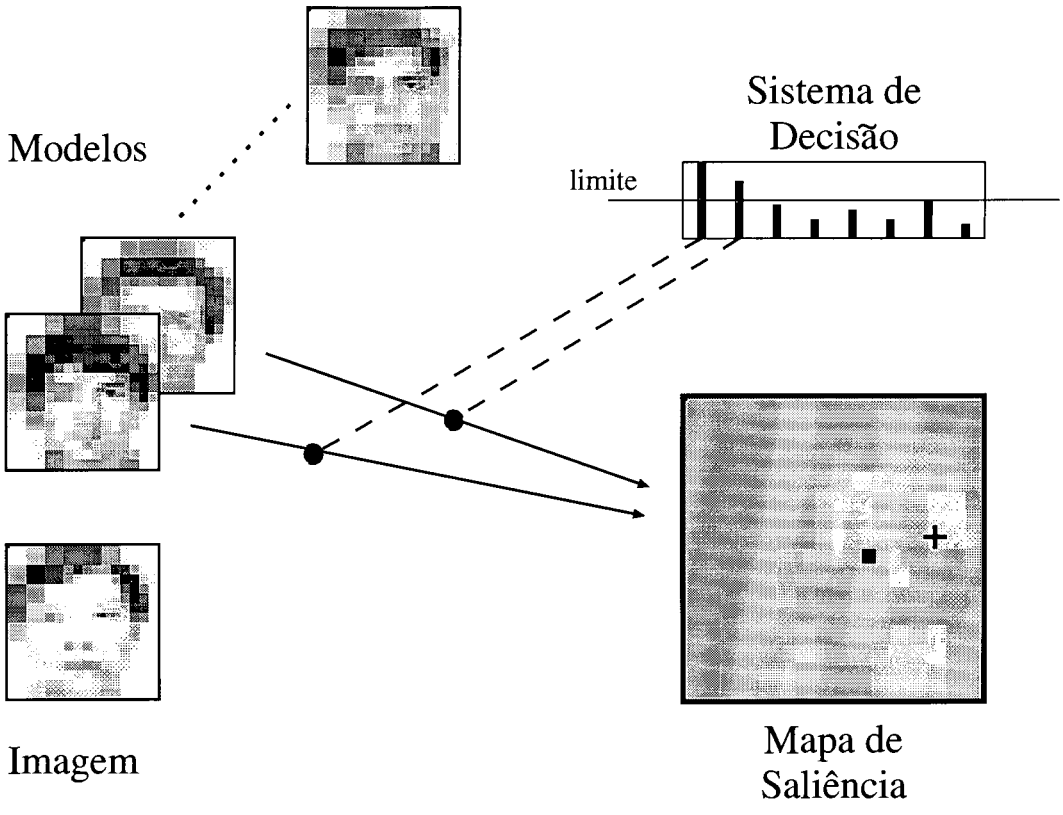

Figura 4.14: **Construção do Mapa de Saliência para estratégia híbrida indireta:** Só as categorias mais ativas participam da constução do mapa. Note que o mapa é construído baseado no ponto de fixação atual (ponto preto) e determina o próximo ponto de fixação (cruz) correspondendo à maior ativação no mapa.

O maior problema detectado em relação a esta estratégia parece depender da escolha do ponto de fixação inicial, e só se manifesta claramente em alguns casos dependendo do conjunto de parâmetros adotado. Como se trata de um Modelo incremental, e a posição da fixação inicial pode ser arbitrária, o valor das ativações das categorias depende muito deste ponto no inicio do processo. Assim, as categorias mais ativas que vão determinar o Mapa de Saliência podem não incluir a correta. Este efeito depende também da extensão das informações extraídas neste início. Quando a quantidade de informações extraídas a cada sacada é pequena, o ponto de fixação inicial pode ativar mais um conjunto de categorias que não contém a correta, fazendo com que esta inicialmente não participe da construção do Mapa de Saliência. Com o decorrer do processo, quando uma massa maior de informações já foi extraída da imagem apresentada, esta tendência pode se reverter. Este efeito parece ter maior importância em uma base de dados mais dispersa onde uma escolha inicial ruim pode levar a muitas sacadas para o fundo ou regiões onde as informações não contribuem para baixar as ativações das categorias incorretas que estão mais ativas. Quando isso acontece, há um periodo inicial de incerteza, no qual uma categoria incorreta pode ter maior atividade, e a discriminação da categoria correta oscila durante um certo número de sacadas, até que se firme uma tendência. Este efeito é, entretanto, difícil de distinguir da oscilação causada por uma outra razão: o fato de que o vetor que serve de dado de entrada para o Sistema de Decisão tem

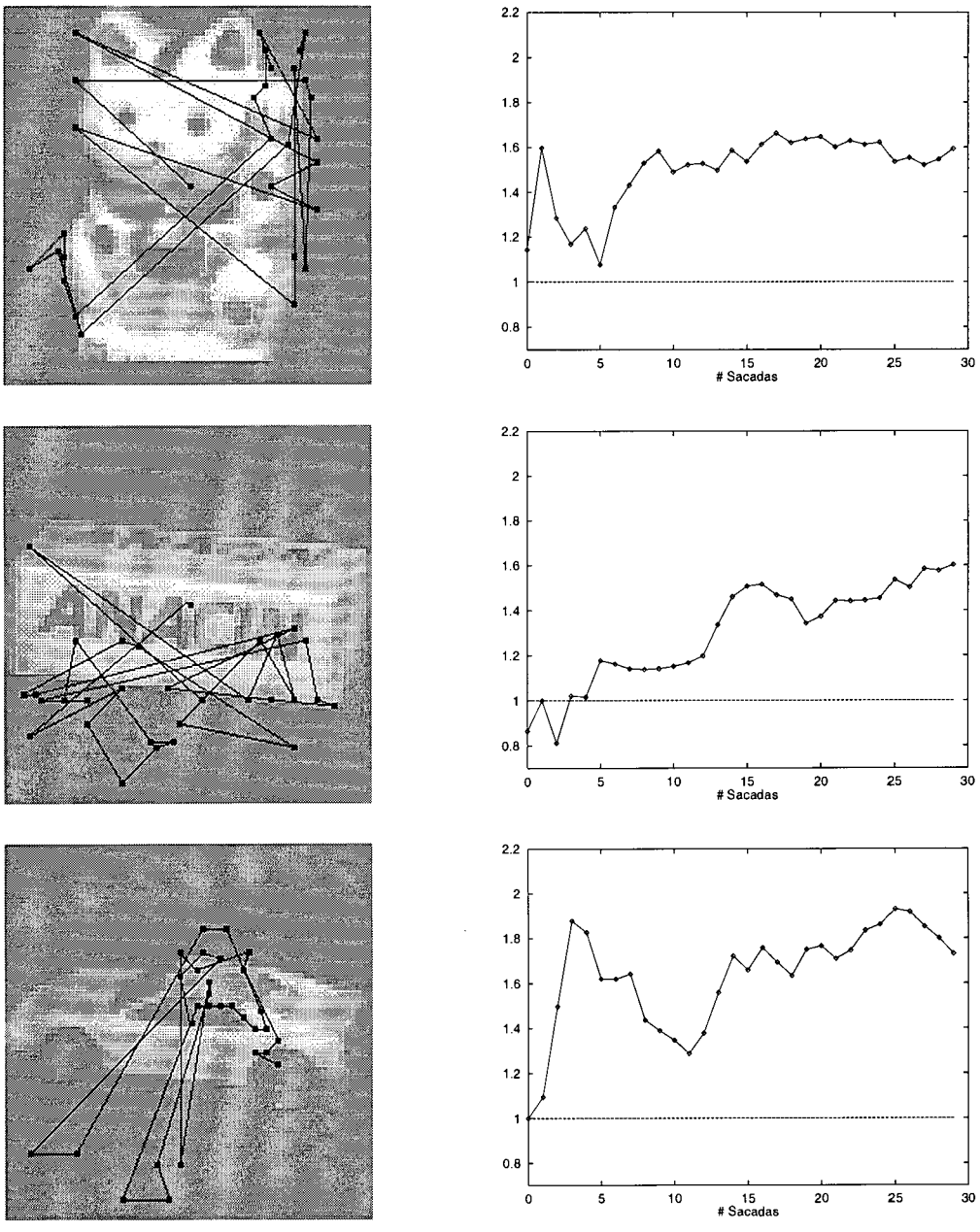

Figura 4.15: **Estratégia híbrida indireta, usando a variação entre as categorias mais ativas**. Esquerda: roteiro de sacadas para três objetos apresentados para reconhecimento. Direita: gráficos de discriminação em função das sacadas (os valores acima de 1.0 significam reconhecimento correto).

um número de componentes pequeno comparado como número de componentes que é acrescentado a cada sacada. Este fato por si só facilita uma oscilação inicial da discriminação, pois os valores das ativações das categorias podem mudar muito de uma sacada para outra. Com o decorrer do processo, o número de componentes do vetor de entrada do Sistema de Decisão passa a ser grande em relação ao acréscimo em cada sacada, e grandes modificações são cada vez menos prováveis.

Em outras palavras, quando o Roteiro de Sacadas depende da atividade das categorias ativadas inicialmente, um erro inicial (baixa ativação da categoria correta) pode demorar a ser corrigido, e esta demora pode ser causada por duas razões

idependentes mas difíceis de serem ditinguidas em seus efeitos.

Observamos aqui ainda muitas sacadas para o fundo das imagens, devido a valores altos de ativação de modelos incorretos.

4.4.3 Estratégia híbrida direta: Desconfirmação das categorias menos ativas

Consideremos que a meta do sistema atencional é determinar a região mais informativa da imagem para a tarefa de reconhecimento. Em princípio, uma localização na imagem com uma boa correlação com o modelo correto será útil pois ela poderá aumentar a ativação da categoria correta. Entretanto, se esta localização tem uma alta correlação com todos, ou com muitos modelos, ela ajudará pouco a determinar o modelo mais adequado. Uma estratégia potencialmente útil é considerar regiões da imagem que possam *desconfirmar* categorias incorretas. Uma localização com uma correlação baixa com os modelos *incorretos* tenderá a diminuir as ativações das unidades associadas a estes modelos. Entretanto, desde que, por definição, nós não conhecemos a categoria correta, isso não pode ser implementado diretamente. Mas, a cada instante, podemos empregar os valores das ativações na saída do Sistema de Decisão como uma indicação de quão provávelmente uma dada unidade representa uma catagoria correta. Em um dado momento, uma baixa ativação indica que a correlação entre modelo e imagem é baixa, e provavelmente o modelo correspondente é incorreto. De outro lado, uma alta ativação expressa a confiança do sistema de que o modelo é potencialmente o correto – quanto maior a ativação, maior a confiança. Separamos então as unidades em dois conjuntos, um das altas ativações (H), e outro das baixas ativações (L). Nossa estratégia será tentar *desconfirmar* as unidades em L. Em outras palavras, selecionar regiões da imagem que dão baixas correlações com as categorias em L, e, fazendo isso, aumentar a diferença entre as ativações das unidades em H e L.

O Mapa de Saliência (ver Figura 4.16), para esta estratégia, será calculado usando um termo de desconfirmação dado pela expressão:

$$S_{n,p} = D^L_{n,p} = \sum_{m \in L} |C_{n,p} - M^L_{m,n,p}|,$$

onde $S_{n,p}$ é o valor do Mapa de Saliência na posição p no anel do nível n, $n = 1, 2, 3$; $D^L_{n,p}$ é o termo de desconfirmação das categoras com baixa ativação para a posição p do nível n; $C_{n,p}$ é a resposta silmulada das células complexas, ou mapa de contornos, para a imagem apresentada, na posição p do nível n; e $M^L_{m,n,p}$ é o valor da representação do modelo $m \in L$ na posição p do nível n. Com isso, as posições p nas quais os modelos pouco ativados e a imagem mais diferem contribuirão mais para o Mapa de Saliência, atraindo as sacadas para estas posições. Na média, estas posições, quando consideradas, tenderão a diminuir mais ainda as ativações das categorias em L. Note que se o termo de desconfirmação fosse computado usando o conjunto completo de modelos ($L \cup H$), isso tenderia a enfatizar localizações onde a imagem e a categoria correta (que não é conhecida) mais diferem, uma propriedade claramente indesejável. A Construção do Mapa de Saliência para esta estratégia é ilustrada na Figura 4.16. Um exemplo de simulação (Figura 4.17) mostra que esta estratégia isolada não fornece bons resultados; entretanto outras simulações (ver

Capítulo 5) mostram que esta estratégia pode ser interessante em associação com outras.

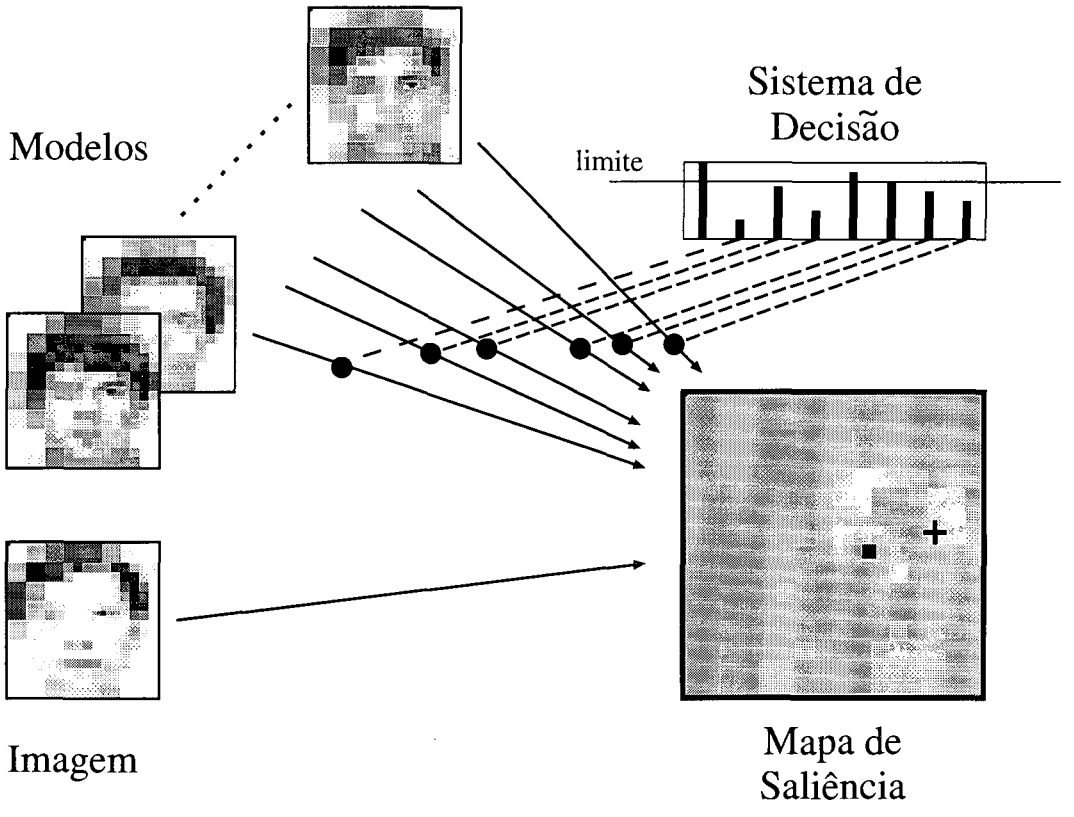

Figura 4.16: **Construção do Mapa de Saliência para estratégia híbrida:** Somente as categorias com baixa ativação e a imagem participam da construção do mapa. Note que o mapa é construído baseado no ponto de fixação atual (ponto preto) e determina o próximo ponto de fixação (cruz) correspondendo à maior ativação no mapa.

A estratégia da variação das categorias mais ativas utilizava um subconjunto dos modelos, aquela parcela que se encontrava mais ativa em cada momento, procurando as localizações onde estas categorias mais diferiam. A estratégia de desconfirmação utiliza o subconjunto complementar, isto é, todos os outros modelos que foram considerados menos ativos. A experiência mostrou que, na maior parte das simulações com as bases de dados de que dispomos, durante o processo de reconhecimento apenas um grupo pequeno de categorias fica mais ativo, isto é, poderia ser incluído no conjunto H das mais ativas (quando os parâmetros do sistema estão bem ajustados). Isso tem acontecido mesmo em bases de dados mais densas como as de faces. Estas seriam as categorias mais semelhantes à imagem apresentada, as categorias que se mostram ambíguas para o reconhecimento, e as outras categorias formam um grupo maior. Este fato tem duas consequências importantes: 1) conseguir resolver a ambiguidade entre os modelos mais parecidos com a imagem apresentada mostrou-se uma tarefa muito importante para o reconhecimento, e por isso a estratégia da variação entre as mais ativas (estratégia anterior) mostra tão bons resultados. 2) Graças ao fato de o conjunto L ser com frequência bem maior que o conjunto H, a estratégia da desconffirmação consegue contribuir bastante para superar os erros

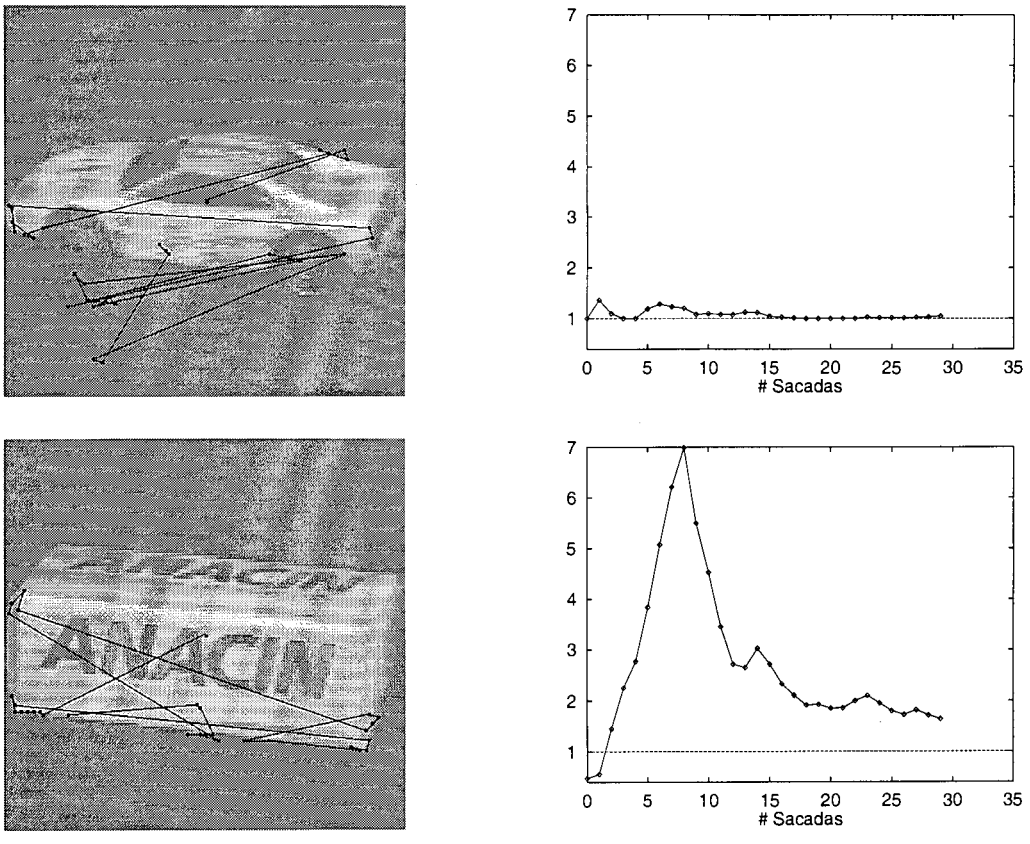

Figura 4.17: **Estratégia híbrida da desconfirmação das categorias menos ativas:** Esquerda: roteiro de sacadas para dois objetos apresentados para reconhecimento. Direita: gráficos de discriminação da categoria correta em função das sacadas (os valores acima de 1.0 significam reconhecimento correto).

iniciais, quando a categoria correta fica inicialmente pouco ativada. Esta virtude se deve a que, sendo o conjunto L razoavelmente extenso, a soma das diferenças entre a imagem e os modelos pode ser alta mesmo nos pontos onde o modelo correto difere pouco da imagem, no caso da categoria correta estar no conjunto L. Assim, se houver um erro inicial, uma escolha de localização baseada nesta estratégia tende a contribuir para que as categorias incorretas diminuam sua ativação, deixando a correta sobressair.

4.4.4 Estratégia híbrida plena: combinação de estratégias

Como vimos, cada uma das estratégias mostradas anteriormente apresenta vantagens e desvantagens que aparecem de modo diferenciado de acordo com algumas características das bases de dados nas quais são aplicadas. Algumas estratégias mostraram resultados melhores, como a da variação das categorias mais ativas, porém o Roteiro de Sacadas ainda visita muitas regiões do fundo, o que é evitado na estratégia baseada na imagem. O que precisamos investigar aqui é o ganho potencial que se pode obter associando estas estratégias, isto é, verificar se as qualidades de cada estratégia podem se associar, levando a melhores resultados. Isso pode ser feito de muitas maneiras, sendo talvez a mais simples a investigada aqui: construir o Mapa de Saliência (ver Figura 4.18) como uma soma ponderada das funções de

saliência correspondentes a cada estratégia, conforme a expressão abaixo:

$$S_{n,p} = \alpha \frac{D^L_{n,p}}{card(L)} + \beta \frac{V^H_{n,p}}{card(H)} + \gamma C_{n,p}, \tag{4.2}$$

onde $D^L_{n,p}$ é o termo de desconfirmação das categorias de baixa ativação; $card(L)$ é o número de categorias $\in L$; $V^H_{n,p}$ é o termo correspondente à variação das categorias de alta ativação; $card(H)$ é o número de categorias $\in H$; $C_{n,p}$ é o termo correspondente à estratégia baseada na imagem (isto é, o mapa das células complexas); e α, β e γ são constantes que podem ser ajustadas para balancear a participação de cada termo. Aqui a divisão pelo número de categorias em cada um dos conjuntos L e H é necessária para balancear os termos correspondentes, de modo que a grandeza deles não fique dependente da extensão destes conjuntos, que varia durante o processo.

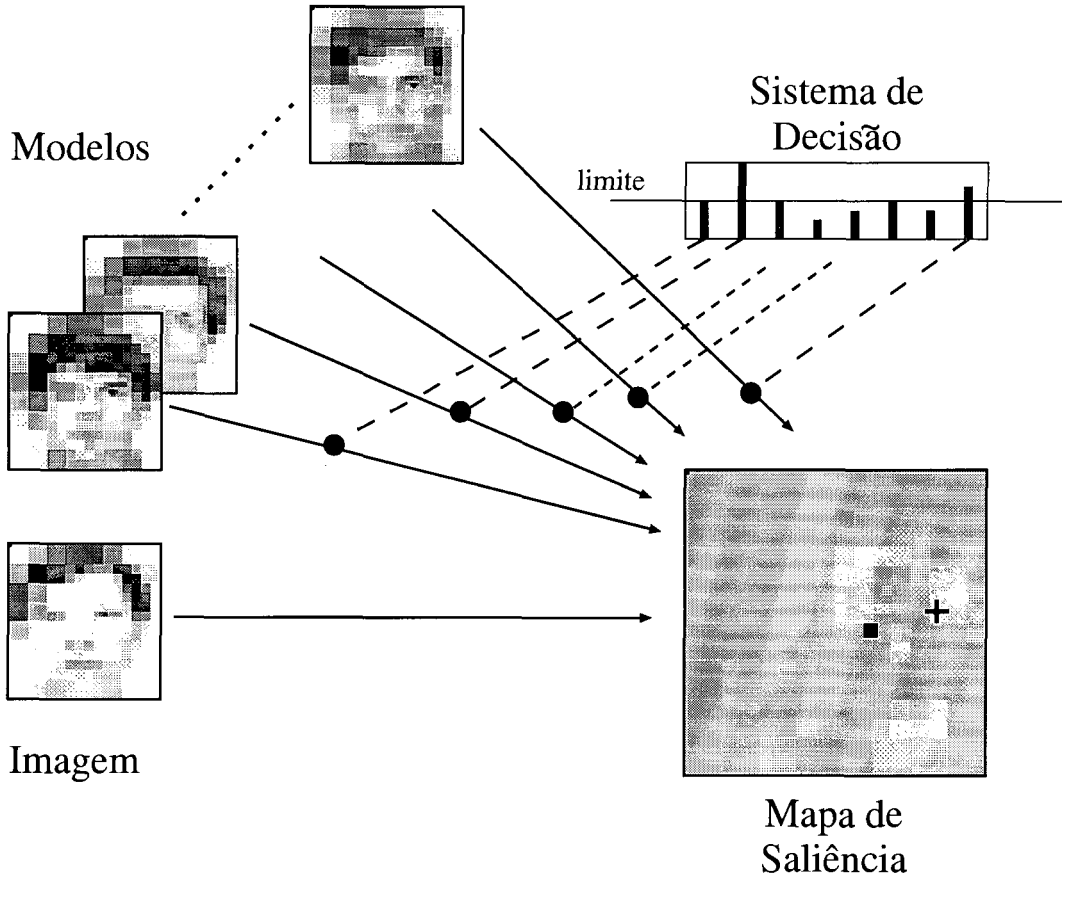

Figura 4.18: **Construção do Mapa de Saliência para estratégia híbrida plena:** Tanto os modelos como a imagem contribuem para a construção do mapa. Note que o mapa é construído baseado no ponto de fixação atual (ponto preto) e determina o próximo ponto de fixação (cruz) correspondendo à maior ativação no mapa.

A Figura4.19 mostra o comportamento do sistema com o Mapa de Saliência dado pela Equação 4.2. Nos três casos a sequência de sacadas provou ser capaz de extrair com sucesso informações para o reconhecimento.

As maiores dificuldades desta estratégia estão relacionadas ao ajuste das constantes α, β e γ da equação 4.2. Este ajuste foi feito de forma empírica, pela observação

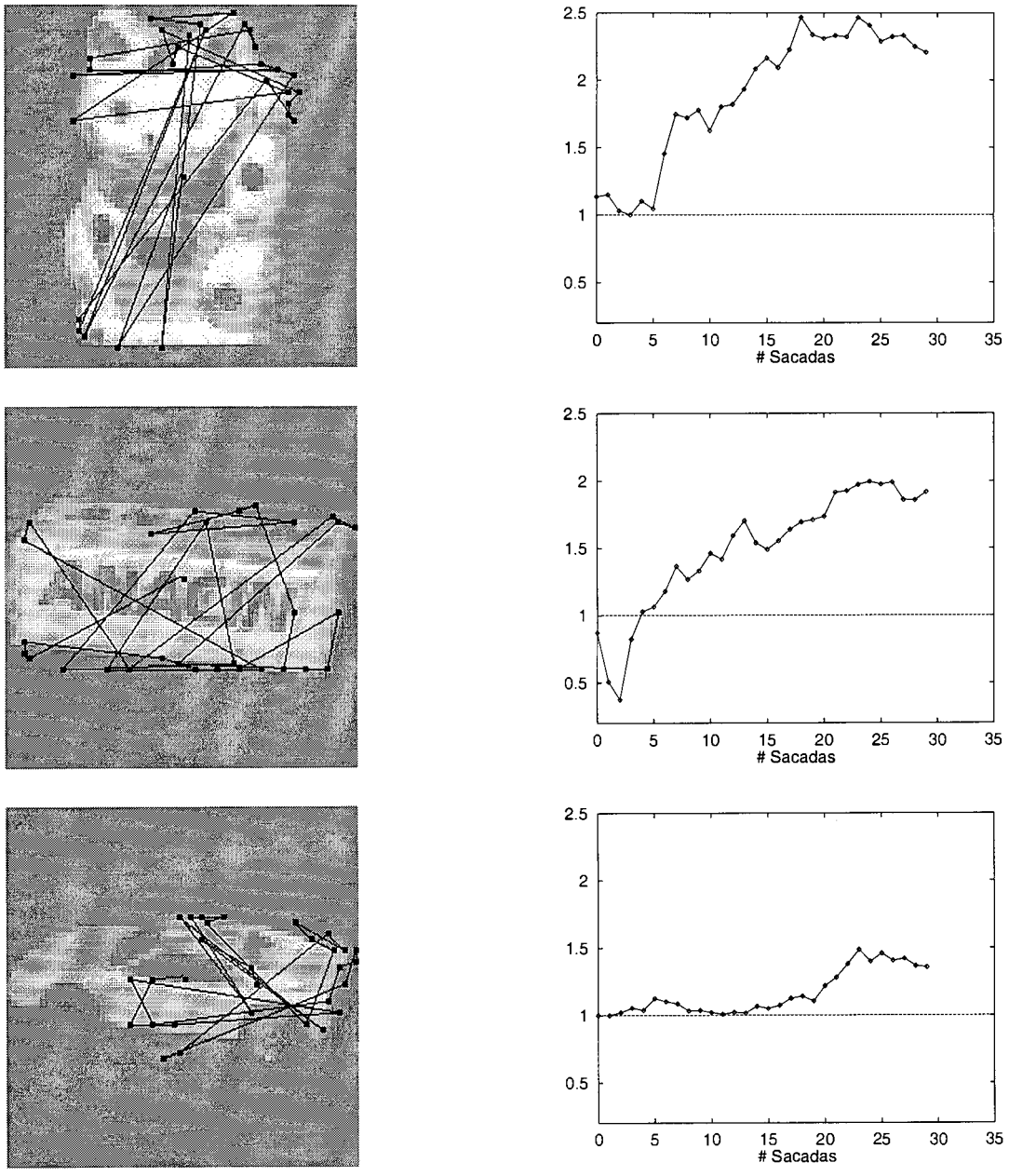

Figura 4.19: **Estratégia atencional híbrida,** dada pela Equação 4.2, usando coeficientes $\alpha = 2$, $\beta = 4$ e $\gamma = 1$. Esquerda: Roteiro de Sacadas; direita: evolução do valor da discriminação.

do comportamento e grandeza de cada uma das parcelas durante as simulações e da eficiência obtida em um certo número de testes. Foi possível identificar algumas tendências, como por exemplo a de que os valores da parcela correspondente ao mapa das células complexas proveniente da imagem tende a ter valores altos, próximos de 1.0, por resultado do processo de filtragem. Já as outras parcelas tendem a ter valores menores. Por este motivo, parece necessário utilizar constantes α e β maiores que γ para evitar que a última parcela sature o Mapa de Saliência, tornando-o insensível às outras contribuições. Um ajuste mal feito, além de fazer com que a influência de alguma estratégia seja desprezada, pode também ter o efeito de desequilibrar as contribuições dos diversos níveis do Mapa de Saliência.

Capítulo 5

Simulações

Neste Capítulo mostraremos as bases de dados utilizadas e as simulações feitas com a finalidade de obter uma avaliação quantitativa do comportamento do Modelo de Reconhecimento Atencional para as diversas estratégias atencionais, analisadas no Capítulo 4. Apresentaremos também uma análise dos resultados quantitativos destas simulações. Um discussão mais aprofundada do comportamento do modelo está no capítulo 6.

Todas as simulações feitas neste trabalho utilizaram imagens de três bases de dados:
1) Base de dados da Universidade de Colúmbia [38];
2) Base dados de faces do Olivetti Research Laboratory [33];
3) Base de dados Eigenfaces, do MIT, de Turk e Pentland [62].

5.1 Reconhecimento de Objetos: base de Colúmbia

5.1.1 Base de dados

A Base de dados da Universidade de Colúmbia, utilizada inicialmente por Murase e Nayar [38], e explorada no trabalho de Rao e Ballard [49] é formada por 20 objetos diferentes, cada um aparecendo em 72 poses rotacionadas de 5 graus em relação à anterior. Estas imagens originalmente tem 256×256 pixels, mas foram reduzidas para 92×92 para serem utilizadas em nossas simulações. A Figura 5.1 mostra os 20 objetos em suas poses com rotação zero.

A Figura 5.2 mostra três dos objetos com as rotações de -35 a $+35$ graus utilizadas nas simulações. A imagem com rotação zero foi usada para extrair a representação do modelo de cada objeto.

É intressante observar os diferentes graus de semelhança entre os 20 objetos, e também a distorção maior ou menor produzida pela rotação. Estas imagens foram usadas nas simulações iniciais para estudo do comportamento qualitativo do Modelo de Reconhecimento Atencional e também nas simulações para a investigação quantitativa.

Figura 5.1: **Base de dados da Universidade de Colúmbia:** Cada um dos 20 objetos mostrados é apresentado em 72 posições rotacionadas de 5 graus em torno de um eixo vertical. Aqui são mostradas as poses com rotação zero.

5.1.2 Simulações

Foram feitas muitas simulações com diferentes condições de pré-processamento, construção de modelos e parâmetros de processamento. Cada uma destas condições tem um tipo de influência no comportamento do modelo, mas aqui são mostrados os resultados de um conjunto de simulações em condições típicas, que descrevemos abaixo.

Pré-processamento

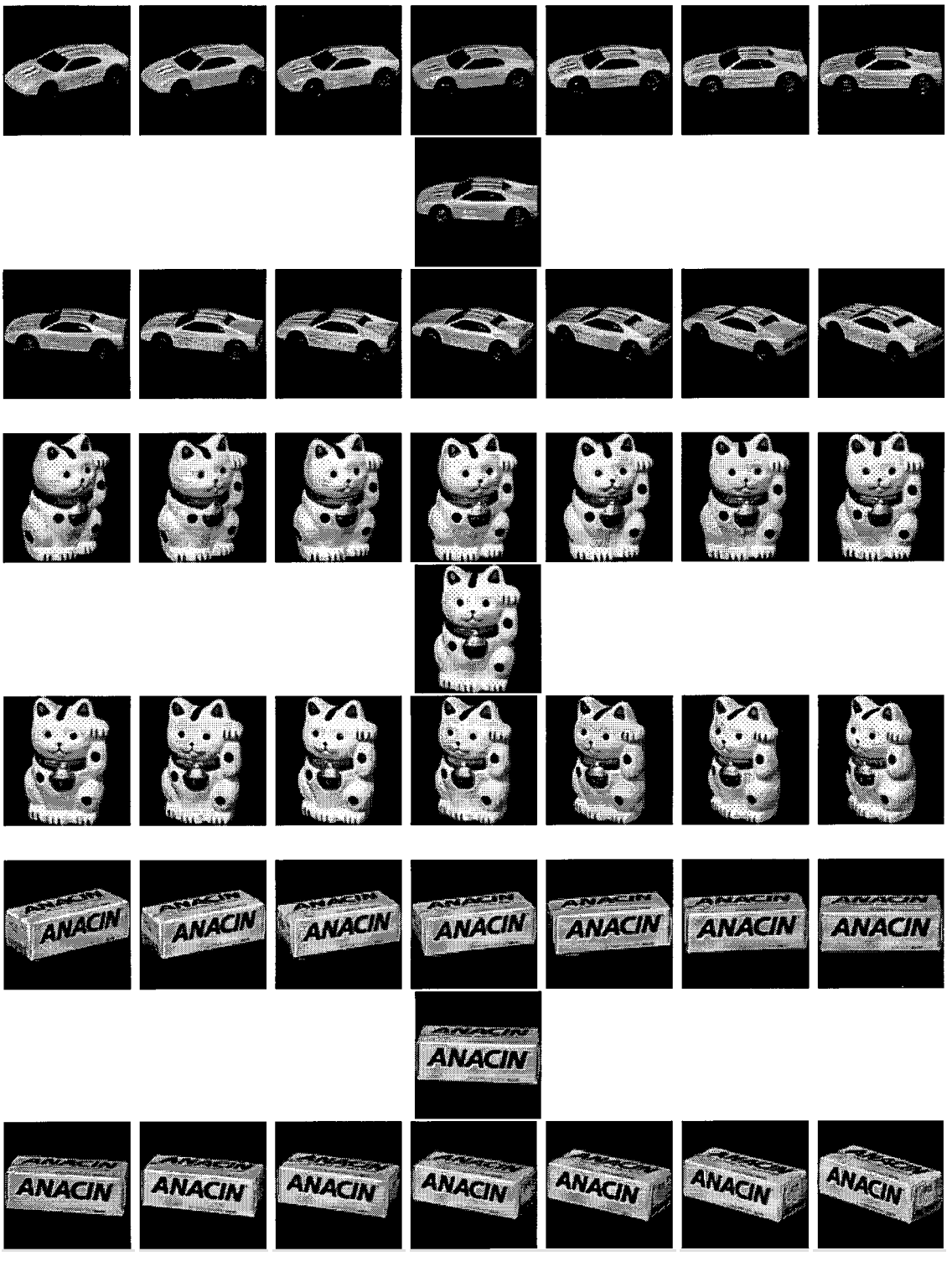

Figura 5.2: **Poses com rotações de -35 a $+35$ graus de 3 dos 20 objetos:** A pose central de cada objeto, com rotação zero, foi usada para extrair a representação do modelo do objeto; as outras poses, rotacionadas com intervalos de 5 graus, formam o conjunto de imagens de teste.

Todas as imagens foram pré-processadas utilizando filtros sensíveis a contrastes orientados, normalizados localmente, simulando as respostas das células complexas do sistema visual humano, conforme explicado no Capítulo 3. Os parâmetros de pré-

processamento foram (ver definições no Capítulo 3): número de níveis da pirâmide = 4; tamanho do filtro do nível zero: $\sigma = 0,5$; razão de aspecto dos filtros: $asp = 3,0$; razão de crescimento dos filtros $cf = 2,0$; razão de sub-amostragem = 2.

Construção dos modelos e imagens de teste:

Os 20 modelos foram construídos utilizando as imagens dos objetos com ângulo de rotação zero. As 280 imagens restantes, isto é, as imagens com rotações variando de 5 em 5 graus, de -35 a $+35$ graus, formaram o conjunto das imagens de teste.

Parâmetros das simulações:

Diâmetro da fóvea: $f = 3$ pixels; diâmetro da região excluída no Mapa de Saliência: $rexc = 3$; razão de crescimento dos anéis: $r = 3$; nível de amplificação das ativações: $q = 4$; limite entre as altas e baixas ativações: $limite = 0,8$; critério de Decisão: $dc = 1,3$ ou 30 sacadas; mínimo de sacadas: $rec = 12$; amplitude dos canais ON e OFF: $zmax = 4,0$.

A Tabela 5.1 mostra os resultados destas simulações.

	estratégias							
ponto inicial	BU	TD	Vy	V^H	D^L	$H061$	$H260$	$H261$
21,21	21,4	22,8	17,1	17,5	23,9	15,7	16,7	18,5
21,63	21,4	23,2	17,1	16,4	23,5	18,5	17,1	18,2
46,46	20,7	21,0	18,9	17,8	24,2	17,5	15,7	15,7
63,21	20,0	25,0	19,6	18,2	23,9	19,6	14,2	15,7
63,63	21,7	23,5	20,3	16,0	26,4	15,7	13,2	16,4
média	**21,04**	**23,10**	**18,60**	**17,18**	**24,38**	**17,40**	**15,38**	**16,90**
indice	1,37	1,50	1,21	1,12	1,59	1,13	1,00	1,10
	pixels comparados no caso mais custoso (%)							
	3,12	3,56	3,36	3,34	3,38	3,20	3,24	3,16

Tabela 5.1: **Simulações com a base de imagens de Colúmbia.** Percentagens de erros de reconhecimento em 280 imagens de teste utilizando diferentes estratégias atencionais. Para cada imagem foram feitas simulações com a sacada inicial em 5 pontos diferentes. A tabela mostra as percentagens de erros e sua média para as cinco posições iniciais, para cada estratégia atencional. O sistema toma uma decisão apontando a categoria mais ativa caso seja atingido o critério de decisão ou o máximo de 30 sacadas. A linha *índice* mostra a razão entre a média de erros para uma estratégia e a menor média de erros, explicitando a hierarquia de eficiência das estratégias. Na última linha estão as percentagens dos pixels totais da imagem que foram utilizados para comparação. (BU = botom-up, TD = top-down, Vy = variação usando ativação como pesos, V^H = variação das categorias mais ativas, D^L = desconfirmação das menos ativas, $H260$ = híbrida com $S = 2D^L + 6V^H + 0BU$, $H061$ = híbrida com $S = 0D^L + 6V^H + 1BU$, e $H261$ = híbrida com $S = 2D^L + 6V^H + 1BU$).

5.2 Reconhecimento de Faces: base do ORL

5.2.1 Base do ORL

A base de dados de faces do "Olivetti Research Laboratory", mostrada na Figura 5.3 (pode ser obtida em http://www.camorl.co.uk/facedatabase.html) é constituída por imagens de faces em posição frontal de 40 pessoas, com 10 poses para cada pessoa, com variações da posição da cabeça e da expressão. A diferença entre as 10 poses de uma dada pessoa varia, de modo que para algumas as poses são bem diferentes, enquanto que outras variam muito pouco. A base de dados ORL é amplamente utilizada para avaliação de modelos de reconhecimento de faces. Por exemplo, Lawrence et al. [33] relatam resultados quantitativos de alguns modelos de reconhecimento utilizando esta base de dados, o que a torna interessante para que se possa comparar resultados (estas comparações serão apresentadas no Capítulo 6).

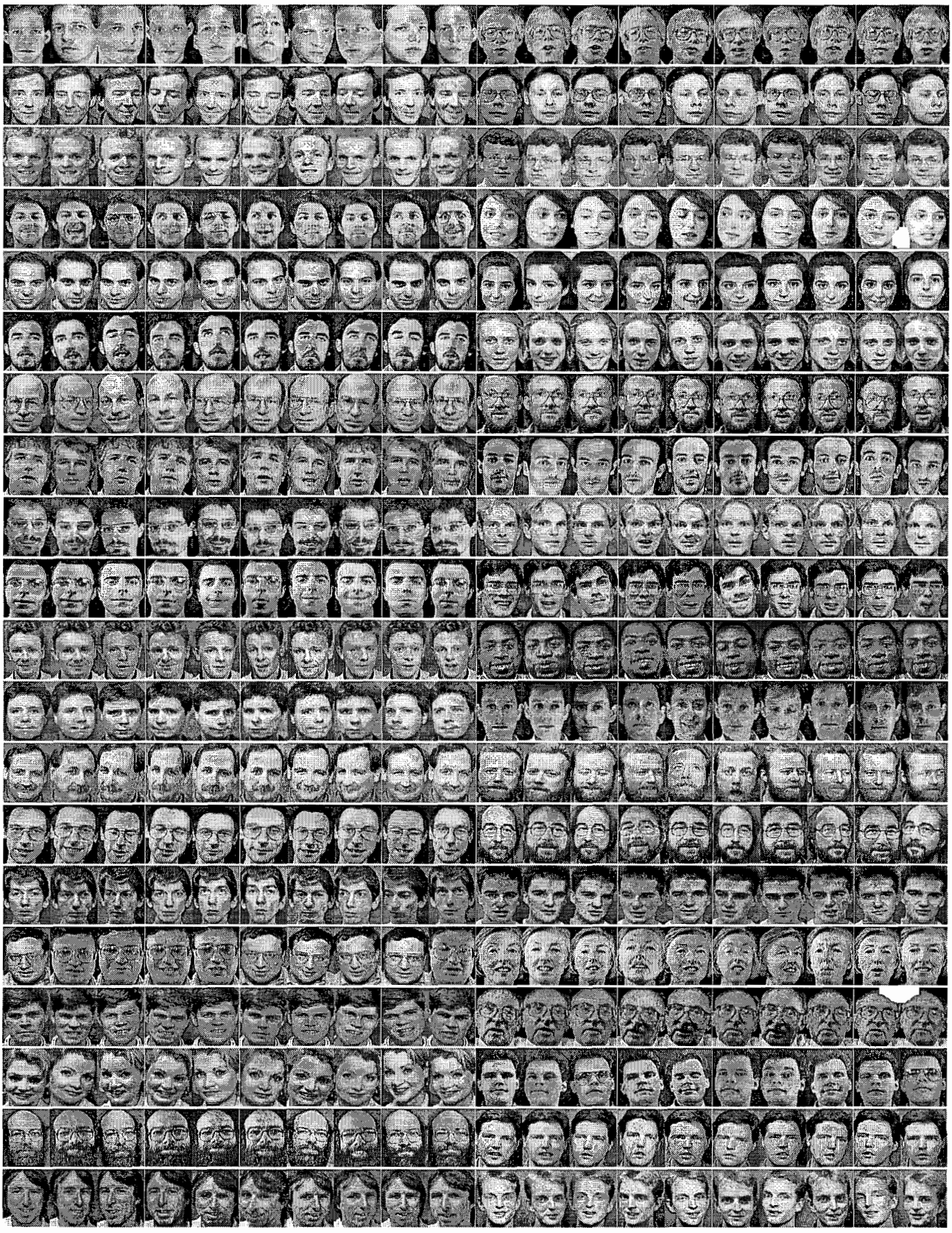

Figura 5.3: **Base de dados de faces do "Olivetti Research Laboratory"**

5.2.2 Simulações

Foram feitas muitas simulações com diferentes condições de pré-processamento, constução de modelos e parâmetros de processamento. Cada uma destas condições tem um tipo de influência no comportamento do sistema, mas aqui são mostrados os resultados de um conjunto de simulações em condições típicas, que descrevemos abaixo.

Pré-processamento:

Todas as imagens foram pré-processadas utilizando filtros sensíveis a contrastes orientados, normalizados localmente, simulando as respostas das células complexas do sistema visual humano, conforme explicado no capítulo 3. Os parâmetros de pré-processamento foram (ver definições no capítulo 3): número de níveis da pirâmide = 4; tamanho do filtro do nível zero: $\sigma = 2,0$; razão de aspecto dos filtros: $asp = 3,0$; razão de crescimento dos filtros: $cf = 1,41$; razão de sub-amostragem = 2.

Construção dos modelos e imagens de teste:

Os modelos foram construídos utilizando a média das representações das 5 primeiras poses de cada pessoa. O conjunto de teste foi formado com as outras cinco imagens de cada pessoa que não foram usadas para construir os modelos [34].

Parâmetros das simulações:

Diâmetro da fóvea: f = 5 pixels; diâmetro da região excluída no Mapa de Saliência: $rexcl = 3$; razão de crescimento dos anéis: $r = 3$; nível de amplificação das ativações: $q = 4$; limite entre as altas e baixas ativações: $limite = 0,8$; critério de Decisão: $dc = 1,3$ ou 30 sacadas; mínimo de sacadas: $rec = 12$; amplitude dos canais ON e OFF: $zmax = 1,0$.

A Tabela 5.2 mostra os resultados destas simulações.

5.3 Base do MIT

A base de imagens de faces do MIT [62] foi usada para os testes qualitativos do modelo de reconhecimento e também para as as simulações da extensão deste modelo de reconhecimento para exploração e reconhecimento de faces em cenas, descrito no Capítulo 6. Mostramos aqui, na Figura 5.4, as imagens desta base de dados que serviram para estes testes.

5.4 Avaliação dos resultados das simulações

A Tabela 5.3 resume os resultados das simulações com as bases de dados de Colúmbia e do ORL. A análise destes resultados permite tirar algumas conclusões a respeito do desempenho do Modelo de Reconhecimento Atencional, que vamos expor aqui. Uma discussão mais geral e comparação com outros modelos será apresentada no Capítulo 6.

Na Tabela 5.3 a relação entre a percentagem de erros em uma estratégia e a menor percentagem de erros para a mesma base de dados está indicada na linhas "índice".

ponto inicial	estrategias							
	BU	TD	Vy	V^H	D^L	$H061$	$H260$	$H261$
21,21	44,5	24,5	20,5	18,0	33,0	20,0	19,5	24,5
21,63	43,5	23,0	19,0	14,5	34,5	20,0	20,0	28,0
46,46	45,5	24,0	20,0	17,0	34,0	20,5	18,5	27,0
63,21	43,0	23,0	25,0	20,5	34,5	19,0	16,5	27,0
63,63	45,0	24,0	21,0	17,0	33,0	20,5	20,0	26,5
média	**44,30**	**23,70**	**21,10**	**17,40**	**33,80**	**20,00**	**19,00**	**26,60**
índice	2,54	1,36	1,21	1,0	1,94	1,15	1,09	1,51
	pixels usados para comparação (%)							
21,21	8,84	9,73	8,62	7,90	8,16	7,61	7,00	7,17
21,63	8,88	9,67	8,50	8,11	8,59	7,80	7,24	7,40
46,46	8,96	9,70	8,72	8,13	8,54	7,79	7,11	7,43
63,21	8,89	9,71	8,59	8,05	8,31	7,55	7,08	7,26
63,63	8,92	9,69	8,60	8,24	8,49	7,73	7,25	7,45
média	8,90	9,70	8,61	8,09	8,42	7,70	7,14	7,34

Tabela 5.2: **Simulações com a base de faces do ORL:** Percentagens de erros de reconhecimento em 200 imagens de teste utilizando diferentes estratégias atencionais. Para cada imagem foram feitas simulações com a sacada inicial em 5 pontos diferentes. A tabela mostra as percentagens de erros e sua média para as cinco posições iniciais, para cada estratégia atencional. O sistema toma uma decisão apontando a categoria mais ativa caso seja atingido o critério de decisão ou o máximo de 30 sacadas. A linha *índice* mostra a razão entre a média de erros para uma estratégia e a menor média de erros, explicitando a hierarquia de eficiência das estratégias. Na parte de baixo da tabela estão as percentagens dos pixels totais da imagem que foram utilizados para comparação; as médias aparecem na última linha. O maior número de pontos pré-processados foi de 10,7 % da imagem original. (BU = botom-up, TD = top-down, Vy = variação usando ativações como pesos, V^H = variação das categorias mais ativas, D^L = desconfirmação das menos ativas, $H260$ = híbrida com $S = 2D^L + 6V^H + 0BU$, $H061$ = híbrida com $S = 0D^L + 6V^H + 1BU$, e $H261$ = híbrida com $S = 2D^L + 6V^H + 1BU$).

Assim a menor percentagem de erros tem índice 1,0, mostrando a hierarquia das estratégias para cada base. Não há sentido em comparar as percentagens de erros entre as bases de dados, pois os ângulos de rotação dos objetos que entraram no conjunto de imagens de teste nas simulações com a base de Colúmbia foram escolhidos de modo a tornar a tarefa de reconhecimento difícil o suficiente para evidenciar as contribuições trazidas pelos métodos explorados. Caso os ângulos de rotação fossem escolhidos numa faixa muito menor que -35 a $+35$ graus, o número de erros seria tão baixo que não se poderia distingüir o efeito das estratégias. Note, no entanto, que em geral a variação nas imagens (ver Figura 5.2) está longe de ser trivial, e desta forma a porção da base de dados utilizada oferece um desafio para métodos de reconhecimento. Assim, como a grandeza da percentagem de erros depende da

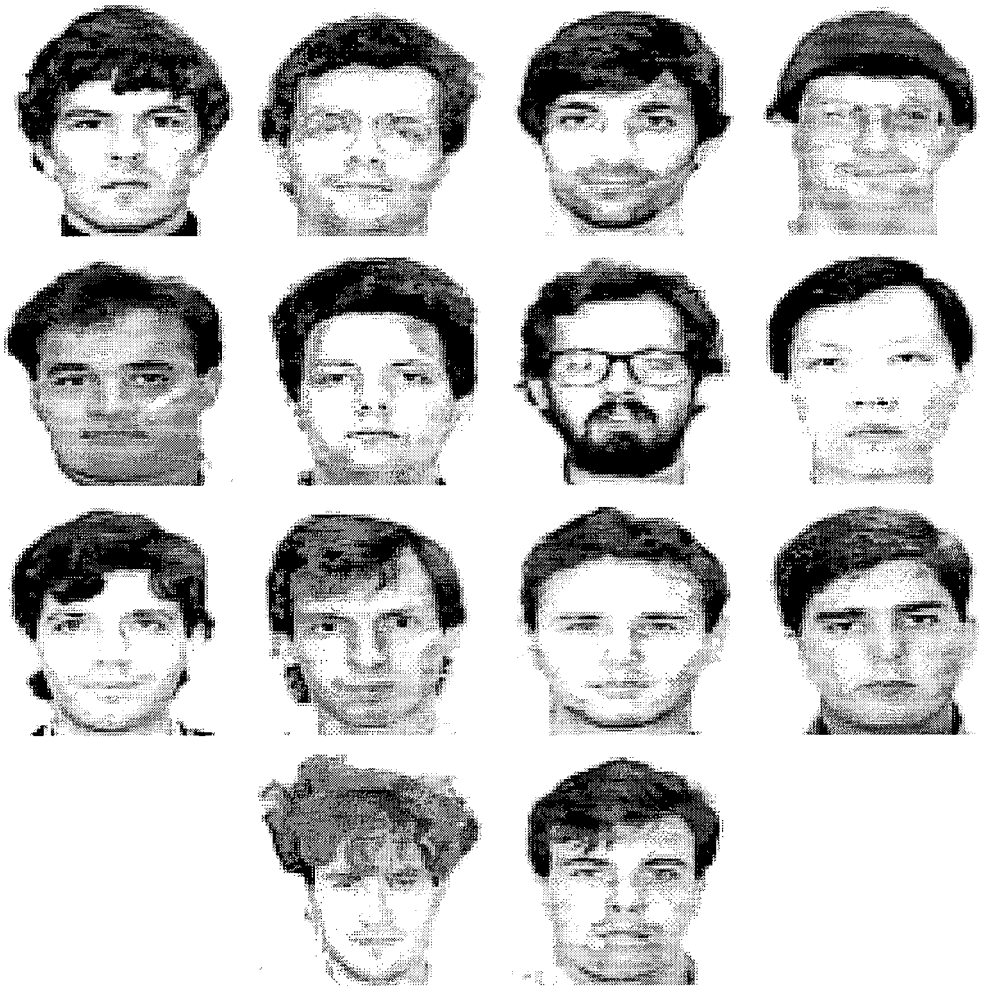

Figura 5.4: **Base de dados de faces do MIT:** as 12 primeiras imagens foram utilizadas como modelos, e as duas últimas como "estranhos" na simulação de reconhecimento.

escolha da faixa de ângulos de rotação dos objetos, não faz sentido comparar com os resultados das imagens de faces. Já o índice pode ser comparado, pois indica quanto uma estratégia é melhor que a mais vantajosa em uma base de dados. Assim, por exemplo, a estratégia TD é claramente muito melhor na base de dados do ORL do que na base de Colúmbia, embora tenha percentagens de erros um pouco maiores na base do ORL, posto que seu índice é 1,36 nesta base contra 1,50 na outra base. A seguir passamos a analisar cada uma das estratégias com base nos valores obtidos nas simulações.

A estratégia estritamente *bottom-up* isolada (BU) produz um índice de erros muito maiores relativamente (2,54) na base ORL do que na base de Colúmbia (1,37). Como já analisado no Capítulo 4 isso se deve a que na base mais dispersa (Colúmbia) as posições do contorno de um objeto são mais discriminativas do que os contornos das faces da base do ORL. No caso das faces, é pouco vantajoso utilizar os contornos para localizar os pontos de onde extrair informações, já que estes são muito semelhantes em toda a base, e esta estratégia dá os piores resultados, 44,30% de erros, certamente alta demais para qualquer método de reconhecimento. Na base de Colúmbia, no entanto, a estratégia *bottom-up* não é uma das piores. É intressante

	estratégias							
	BU	TD	Vy	V^H	D^L	$H061$	$H260$	$H261$
Colúmbia	21,04	23,10	18,60	17,18	24,38	17,40	15,38	16,90
índice	1,37	1,50	1,21	1,12	1,59	1,13	1,00	1,10
ORL	44,30	23,70	21,10	17,40	33,80	20,00	19,00	26,60
índice	2,54	1,36	1,21	1,0	1,94	1,15	1,09	1,51
índice total	1,87	1,37	1,15	1,01	1,69	1,09	1,00	1,24

Tabela 5.3: **Resumo das simulações:** Resumo das tabelas 5.1 e 5.2: percentagens de erros nas simulações com as bases de dados de Colúmbia e do ORL, para cada estratégia atencional. As linhas *índice* indicam a razão entre a percentagem de erros de cada estratégia e a menor percentagem de erros para a *mesma* base de dados (não há sentido em comparar as percentagens de erros *entre* as bases de dados, como explicado no texto). A linha *índice total* foi calculada usando a soma dos índices das duas bases e associando o menor valor a $1,0$. BU = botom-up, TD = top-down, Vy = variação usando ativações como pesos, V^H = variação das categorias mais ativas, D^L = desconfirmação das menos ativas, $H260$ = híbrida com $S = 2D^L + 6V^H + 0BU$, $H061$ = híbrida com $S = 0D^L + 6V^H + 1BU$, e $H261$ = híbrida com $S = 2D^L + 6V^H + 1BU$.

comparar estes resultados com os da estratégia *top-down*. A estratégia estritamente baseada nos modelos (TD) apresenta um comportamento inverso da estratégia BU, isto é, dá resultados muito ruins na base de Colúmbia (índice de 1,50), quase os piores, e resultados um pouco melhores na base do ORL (índice de 1,36).

O resultado mais importante foi o bom desempenho da estratégia híbrida indireta da variação das categorias mais ativas (V^H) quando empregada isoladamente. Nas duas bases o resultado foi bom, obtendo índice 1,0 na base do ORL (o melhor) e índice 1,12 na base de Colúmbia. Este resultado ilustra que a atividade de discriminar entre os modelos mais parecidos com a imagem é muito importante no processo de reconhecimento, e uma estratégia que procura fazer isso produz bons resultados. É interessante notar que esta atividade *é híbrida*, quer dizer, ela só pode ser levada a efeito utilizando informações tanto de baixo nível, proveniente da imagem, quanto cognitivas, de alto nível, provenientes do conhecimento armazenado na memória. Quanto à outra estratégia híbrida indireta (Vy) seus resultados são intermediários entre a estratégia TD e a estratégia V^H. Este comportamento sugere que a utilização de todos os modelos na atividade de dissolver as ambigüidades, ainda que utilizando pesos para modular a influência de cada modelo de acordo com sua probabilidade, acaba por atrapalhar a discriminação que a estratégia V^H consegue.

A estratégia híbrida de desconfirmação das categorias menos ativas, D^L, dá resultados bem fracos nas duas bases (índices 1,59 e 1,94), mas este fato não é de grande importância, já que ela não foi concebida para atuar isoladamente, mas como um complemento à estratégia V^H. Tanto que esta associação dá os melhores resultados na base de Colúmbia (estratégia $H260$), e também ótimos resultados na base do ORL, levando ao melhor resultado global. Este comportamento ilustra a

importância de associar estratégias atencionais distintas, de modo a conseguir resultados melhores do que os obtidos por cada uma delas separadamente. Assim também a associação da estratégia BU com as estratégias V^H e D^L produz (estratégia $H261$) uma percentagem de erros menor que cada uma delas separadamente na base da Colúmbia.

Como uma conclusão parcial, estes resultados indicam alguns pontos importantes: primeiro, a atividade de reconhecimento é intrinsecamente uma atividade híbrida, no sentido de que necessita tanto de mecanismos *bottom-up* quanto *top-down*. Segundo, os mecanismos de baixo nível, pré-atencionais, de processamento em paralelo, e por isso mais rápidos, são mais eficientes para a discriminação entre imagens com níveis de semelhança menores (exemplificados pela base de Colúmbia), enquanto que as estratégias mais dependentes de informações de alto nível são mais eficientes para discriminar objetos muito parecidos, onde os mecanismos *bottom-up* não conseguem dissolver as ambigüidades. Em terceiro, finalmente, estas simulações mostram que é factível e potencialmente profícuo associar estratégias atencionais visando melhores desempenhos nas tarefas visuais de reconhecimento.

Capítulo 6

Avaliação e Discussão

Um modelo de reconhecimento visual atencional como o apresentado neste trabalho tem seu desempenho baseado principalmente em duas caracateristicas: a utilização de uma representação *space-variant* que se assemelha a uma retina como a dos primatas, com alta resolução na região central ou fóvea, e um mecanismo atencional, capaz de escolher dinâmicamente os pontos de fixação nas regiões de maior interesse para a tarefa de reconhecimento. Grande economia de processamento, e portanto rapidez, aliada a uma grande eficiência em tarefas visuais como o reconhecimento, deriva da união destes dois dispositivos. Para proceder a uma avaliação do modelo investigado aqui, vamos analisar separadamente estas duas características importantes. Inicialmente vamos analisar o dispositivo de representação em resolução variável e suas consequências do ponto de vista qualitativo e quantitativo para o desempenho do modelo. Este dispositivo de representação é responsável por um grande ganho em termos de custos de processamento do Modelo de Reconhecimento Atencional, que serão analisados a seguir, e comparados aos de modelos não incrementais. Discutiremos então o Modelo de Reconhecimento Atencional como um todo, as diversas estratégias atencionais, e sua relação com outras abordagens encontradas na literatura recente. Finalmente apresentaremos a extensão desta abordagem para o reconhecimento de faces em cenas contendo vários objetos, mostrando alguns resultados iniciais.

6.1 Representação *space-variant*

A organização do sistema visual dos primatas atende ao compromisso de prover simultâneamente um grande campo visual, uma alta resolução na região de maior interesse, e uma saída de informações que permite um processamento rápido, por ser de dimensionalidademuito mais baixa do que teria se a resolução fosse uniforme. Estas qualidades são altamente desejáveis em sistemas artificiais, como na visão robótica, com agentes autonomos e processamento em tempo real [7].

Enquanto o conhecimento a respeito da atenção visual e do sistema sacádico é ainda fragmentado, a tarefa de modelar o controle comportamental de sistemas no domínio técnico se torna mais e mais urgente [30]. No nosso modelo atencional, utilizamos uma representação *space-variant* que procura atender ao compromisso apontado acima, de modo a associar um grande campo "visual" com alta acuidade

central, mantendo uma baixa dimensionalidade. Esta representação é organizada como uma "pirâmide" de resoluções, típicamente em 4 níveis, como já descrito no Capítulo 3. Embora não seja intenção deste trabalho fazer uma simulação biológica do mapeamento retina–cortex, é interessante simular a *funcionalidade* deste mapeamento, isto é, adotar uma resolução variável e investigar a sua associação com um sistema atencional. Em todo caso, uma comparação com uma retina biológica é útil como ilustração. O gráfico da Figura 6.1 mostra o crescimento linear com a excentricidade (ângulo em relação ao centro da fóvea) do campo receptor das células ganglionares do sistema parvo celular da retina em primatas, em comparação com o crescimento discreto da resolução na representação empregada no nosso modelo. Enquanto o crescimento do campo receptor biológico é linear, a representação utilizada aqui apresenta 4 patamares de resolução constante. A excentricidade da borda de cada anel, ou seu diâmetro, é regulada por um parâmetro do sistema, e o gráfico mostra a posição dos patamares de resolução para uma configuração típica adotada na maioria das simulações, na qual cada anel tem diâmetro 3 vêzes o diâmetro do anel anterior. Admitindo-se que o ângulo sólido "visto" pela "fóvea" simulada seja o mesmo que o ângulo visto por uma fóvea biológica, o crescimento da resolução com a excentricidade na representação *space-variant* acompanha de forma descontínua o crescimento dos campos receptores das células ganglionares da retina. A resolução da representação *space-variant* cai à metade a cada anel, o que corresponde a dobrar a distância entre as regiões da imagem representadas por cada pixel na representação.

6.1.1 Análise qualitativa

A representação *space-variant* discreta oferece uma solução potencial para o dilema (já comentado no Capítulo 2) associado ao problema do reconhecimento: nos níveis de resolução mais baixa a diferença entre um modelo armazenado e uma nova instância do mesmo objeto quando apresentado ao sistema é menor, o que será uma vantagem. Entretanto, nestes níveis a diferença entre os modelos também é menor dificultando a discriminação. Já num nível de maior resolução o problema se inverte, sendo mais fácil discriminar entre os modelos, porém o objeto apresentado também difere mais da instância (ou instâncias) utilizadas como modelo. Fukushima [20, 22] argumenta que a não uniformidade da resoluçã da representação desempenha um papel muito importante na construção de um modelo robusto de reconhecimento, pois as imagens mais "borradas" nos níveis de menor resolução são essenciais para um reconhecimento resistente a deformações. Uma vantagem desta representação *space-variant* discreta é que as qualidades de cada nível de resolução podem ser balanceadas facilmente pela utilização de vários níveis ao fazer a comparação com os modelos. Em nossas simulações utilizamos típicamente a "fóvea" e o primeiro anel para a comparação, com bons resultados. Outras opções foram também testadas, como considerar como "fóvea" o nível 1 da pirâmide e um anel do nível 2, ou iniciar o processo utilizando apenas a representação do nível 3 (menor resolução). Neste contexto, é interessante uma investigação mais detalhada no sentido de encontrar

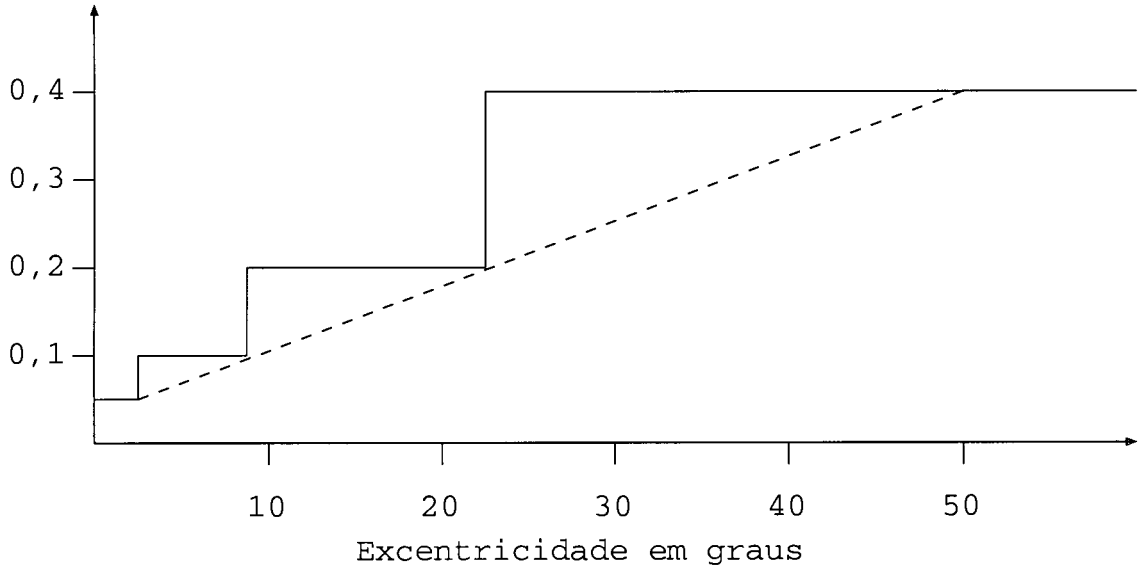

Figura 6.1: **Comparação entre o crescimento descontínuo dos níveis de resolução da representação *space-variant* usada no modelo (linha sólida) e o crescimento dos campos receptores das células ganglionares do sistema parvocelular (linha tracejada) da retina em primatas.** Assumindo-se que o ângulo sólido "visto" pela "fovea" simulada seja o mesmo que o ângulo visto por uma fóvea biológica, o crescimento com a excentricidade (eixo horizontal) da resolução da representação *space-variant* acompanha de forma descontínua o crescimento dos campos receptores das células ganglionares da retina (eixo vertical, em graus) com a excentricidade (dados extraídos de Bolduc e Levine, 1998 [7]).

um modo de otimizar a utilização destes níveis de resolução [1].

6.1.2 Análise quantitativa

Num caso típico, o nível de maior resolução (nível 0) corresponde à região da "fóvea", com um total de 19 pixels (diâmetro de 5 pixels). A periferia da "fóvea" é composta por três aneis com resolução decrescente e com diâmetros crescentes, obtidos cada um de um dos níveis da pirâmide de resoluções (ver Figura 6.2). Assim, apesar de cada anel corresponder a uma região da imagem original com diâmetro maior que o precedente, seu diâmetro ou sua área, em numero de pixels, deve ser computada no seu próprio nível da pirâmide. Como vimos no Capítulo 3, cada nível da pirâmide tem uma área em número de pixels correpondente a 1/4 da área do nível precedente. A Tabela 6.1 mostra os diâmetros e áreas em número de pixels desta representação em cada nível, bem como as percentagens em relação a uma imagem típica de 92×112 pixels onde foi usada uma fóvea de diâmetro 5 pixels. A razão de crescimento do diâmetro dos aneis é um parâmetro do sistema, mas foi adotado na maioria das simulações o valor 3. Isto significa que cada anel abrange uma área da imagem original com 3 vezes o diâmetro do anel anterior. Os valores para o anel do nível 3

[1] Convém lembrar que este problema não é independente do tamanho dos filtros utilizados no pré-processamento.

foram calculados conservativamente como se ele tivesse um diâmetro de 135 pixels, porém na verdade ele será cortado pelas bordas da imagem, e por isso sua área efetiva será bem menor do que o calculado. A Figura 6.2 ilustra esta configuração.

Nível	No.total de pixels	%	diâmetro dos anéis	área em pixels	% da área total
0	10304	100	5	19	0,18
1	2576	25	15	32	0,31
2	644	6,25	45	89	0,86
3	154	1,49	(135)	136	1,32
total					2,67

Tabela 6.1: **Representação** *space-variant* **em 4 níveis:** Em um caso típico usado nas simulações com imagens de 92 × 112 pixels, com diâmetro da fóvea de 5 pixels. A terceira coluna mostra os tamanhos totais de cada nível em percentagens do número de pixels da imagem original. Os diâmetros dos anéis (quarta coluna) estão em pixels da imagem original. Na quinta coluna aparecem as áreas dos anéis em número total de pixels (em seus próprios níveis). A última coluna mostra as áreas dos anéis em percentagens de pixels da imagem original. O tamanho total em pixels desta representação é somente de $2,67\%$ da imagem original.

Como vemos, esta representação proporciona uma grande redução de dimensionalidade em relação a uma imagem de resolução uniforme, pois utiliza apenas $2,67\%$ do número de pixels da imagem total de tamanho 92 × 112. Muitos outros tipos de representação com multi-resolução tem sido propostas, destacando-se as representações de forma log-polar, nas quais o decréscimo de resolução se dá de forma contínua em direção à periferia. No sensor desenvolvido por Sandini e colaboradores (ver Sandini, G. e Tistarelli, M. 1991 [55] e Debusschere et al. 1989 [13]), por exemplo, a representação é de cerca de $8,8\%$ da imagem original. Outros modelos apresentam reduções para cerca de $5,0\%$, ou menos [7].

6.2 Custo computacional

Para avaliar o custo computacional deste modelo de reconhecimento em comparação com outras abordagens com este tipo de representação topográfica, vamos separar este custo em três parcelas: a primeira representa o custo de *comparação* entre as representações extraídas da imagem e as representações dos modelos; a segunda parcela representa o custo de *pré-processamento*, isto é, o custo de filtragem da imagem nas regiões indicadas pelo Módulo Atencional; a terceira parcela representa o custo de processamento para *escolha do próximo ponto de fixação* a cada sacada. As duas primeiras parcelas são comuns a todos os modelos de reconhecimento, enquanto que a última é típica dos modelos atencionais. Além disso, num modelo como o nosso, todos estes custos ocorrem de forma incremental, pois todo o processamento ocorre apenas quando e se for necessário.

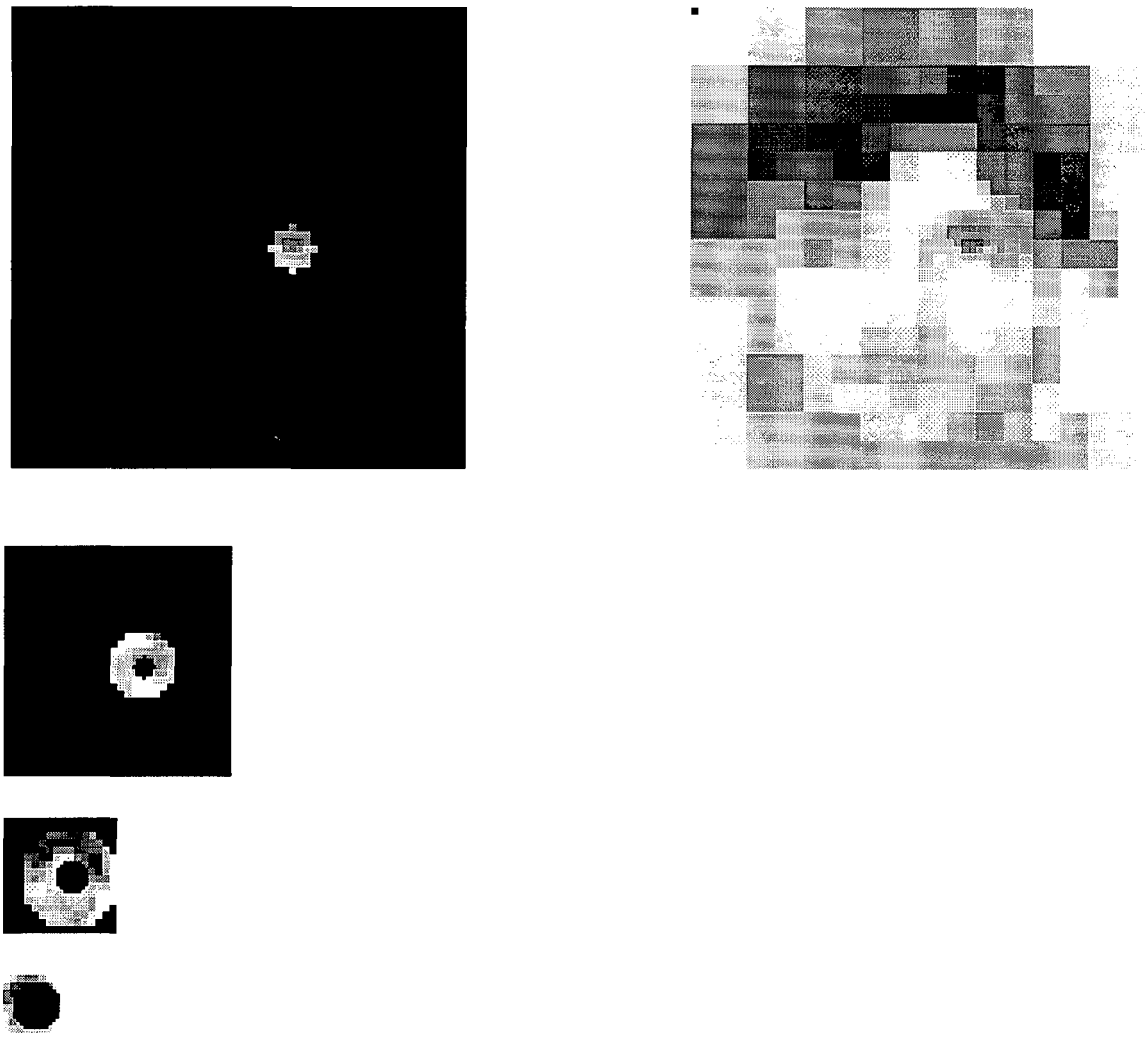

Figura 6.2: **Representação** *space-variant*: Esquerda: regiões de cada nível da pirâmide utilizadas para construir a representação *space-variant*. Observe que a extensão total de cada nível da pirâmide é menor que a do nível precedente de acordo com a razão de sub-amostragem, como explicado no texto. Direita (acima): visualização da representação formada com as 4 regiões de resoluções diferentes, em uma sacada para o olho esquerdo. Os anéis dos níveis inferiores ao nível zero foram ampliados para permitir uma visualização do conjunto.

Para um cálculo do custo da primeira parcela indicada acima, vamos assumir que o custo de comparação da representação extraída da imagem apresentada com as representações dos modelos armazenados, por pixel comparado, é constante para todas as abordagens, e representaremos este custo por C. Assim, em uma abordagem não incremental, onde *toda* a extensão da representação é comparada, este custo seria

$$Custo_c = CN^2,$$

onde $Custo_c$ é o custo total de comparação com os modelos, e N^2 é o tamanho da representação da imagem, em pixels.

No nosso modelo de reconhecimento, o custo de comparação $Custo_c(s)$ é uma função do número de sacadas s necessário para atingir um critério de decisão. Con-

vencionamos aqui que $s = 0$ corresponde à sacada inicial, na qual já há processamento. Portanto o custo é dado por:

$$Custo_c(s) = (s+1)CN^2c, \tag{6.1}$$

onde c é a fração do número de pixels total da imagem que é utilizada para comparação a cada sacada, e s é o número de sacadas. Um valor típico para a fração c corresponde à soma das áreas da fóvea e do anel do nível 1, o que dá $0,49\%$ da área total em uma imagem 92×112 imagem (ver Tabela 6.1). Neste caso $c = 0,0049$ e a expressão 6.1 fica:

$$Custo_c(s) = 0,0049(s+1)CN^2.$$

Custo de pré-processamento

Na maioria dos modelos de reconhecimento há um custo de pré-processamento da imagem, geralmente uma filtragem que pode ser uma convolução com uma função de duas variáveis cujo custo é bastante elevado para cada pixel da imagem pré-processada. Vamos assumir aqui que este custo seja de apenas 10 vezes o custo de comparação de um pixel. No caso de um modelo de reconhecimento que utiliza toda a imagem, o custo de pré-processamento seria:

$$Custo_p = 10CN^2,$$

onde $Custo_p$ é o custo total de pré-processamento.

No caso do modelo incremental, este custo depende de quanto da imagem é pré-processado a cada sacada:

$$Custo_p(s) = 10(s+1)CN^2p,$$

onde p é a fração da imagem total pré-processada a cada sacada (Lembre que $s = 0$ corresponde à sacada inicial).

Em nosso modelo, o valor de p depende de duas coisas: a fração da imagem que vai ser pré processada com a finalidade de comparação com os modelos e a fração que vai servir para construir o Mapa de Saliência, a cada sacada. Típicamente utilizamos para a comparação a fóvea e o primeiro anel (nível 1), e o Mapa de Saliência é construído utilizando apenas os anéis a partir do nível 1, sem a fóvea. Assim, a cada sacada será necessário, na realidade, pré-processar a fóvea e os anéis de todos os níveis da pirâmide. Entretanto, uma vez que um ponto de qualquer nível da pirâmide da representação foi obtido (pré-processado), ele não mais precisará sê-lo, ficando armazenado para ser usado sempre que este ponto participar de um anel em uma sacada futura qualquer. Deste modo, os níveis da pirâmide vão gradativamente sendo pré-processados e "completados" à medida que as sacadas vão acontecendo, e os níveis menores (menor resolução, portanto menos extensos) serão "saturados" primeiro, isto é, totalmente pré-processados depois de algumas sacadas. Por isso o valor de p varia com o número de sacadas. Assim, após as primeiras sacadas o nível de menor resolução cujo anel é mais extenso estará todo pré-processado, não mais participando mais da fração p daí em diante. O mesmo tende a acontecer gradativamente com os outros níveis de modo que a cada sacada não será preciso pré-processar toda a extensão dos 4 níveis da representação. A tendência geral é de que com a sucessão de sacadas apenas uma pequena parte de cada anel, que ainda

não foi obtida nas sacadas anteriores seja pré-processada. Além disso, quando ocorre uma sacada para um ponto próximo à borda da imagem, os anéis são cortados por esta borda, o que também diminui a extensão a ser pré-processada.

Apesar do custo de pré-processamento depender intimamente do Roteiro de Sacadas particular seguido no reconhecimento de uma imagem, é possível estabelecer um limite máximo teórico para este custo, considerando que todas as sacadas serão dadas para regiões que ainda não foram pré-processadas em nenhum dos níveis da pirâmide, com exceção do nível 3, que podemos assumir como sendo "saturado" nas duas primeiras sacadas. Esta saturação deve se dar porque, no exemplo da imagem de 92×112 pixels, o nível 3 da pirâmide de resoluções tem um total de 154 pixels, e a representação *space-variant* utilizada tem até 136 pixels no anel correspondente ao nível 3 (ver Tabela 6.1). A partir daí a área a ser pré-processada crescerá com a soma das áreas dos anéis dos níveis 0, 1 e 2, que somam $(1,34\%)$ a cada sacada, o que faz $p = 0,0134$ até a saturação do nível 2, que deve se dar com cerca de 8 sacadas. Deste ponto em diante $p = 0,0049$, correspondendo às áreas da fóvea e do anel 1, até um número muito elevado de sacadas (66) necessário para saturar o nível 1.

Para um cálculo aproximado e simples do custo máximo teórico, vamos assumir a partir daqui que a curva que representa a fração da área que é pré-processada em função do número de sacadas é uma reta como a mostrada no gráfico da Figura 6.3. Esta reta, descrita pela expressão 6.2, tem inclinação correspondente ao crescimento da área pré-processada somente pelo acréscimo a cada sacada das áreas da fóvea e do anel do nível 1 $(0,49\%$ da imagem total), e tendo ordenada na sacada 8 correspondente às áreas totais do nível 3 e do nível 2, mais a fração das áreas do níveis 0 e 1 pré-processadas até aí. Isto acontece porque até a oitava sacada os níveis 2 e 3 da pirâmide podem ser considerados totalmente pré-processados, não influindo mais nas sacadas subseqüentes. Esta expressão é apenas uma simplificação de uma curva composta por três retas de inclinações diferentes, a primeira correspondendo às sacadas 0 e 1, a segunda correspondendo às sacadas de 2 até 7 e a terceira daí em diante, a única representada. As mudanças de inclinação devem-se à saturação, isto é, pré-processamento total dos níveis 3 e 2 (nesta ordem) da pirâmide. Esta saturação pode ser estimada comparando-se o número de pixels em cada anel da representação *space-variant* com o número total de pixels em cada nível da pirâmide, mostrados na Tabela 6.1. Note que esta expressão não contempla a forma verdadeira deste crescimento apenas antes da sacada de número 8, onde a área pré-processada é menor:

$$p = 0,0820 + 0,0049(s + 1). \tag{6.2}$$

O gráfico da Figura 6.3 mostra uma comparação do crescimento teórico da área pré-processada com as sacadas, para uma imagem com 92×112 pixels, com o crescimento real da àrea pré-processada efetivamente na simulação mais custosa das descritas no Capítulo 5. Observe que a curva de limite máximo teórico tem valores próximos ao dobro dos da curva obtida nas simulações com a base do ORL. A linha horizontal representa a área total da imagem, que tem que ser pré-processada em um modelo não incremental.

Assim, levando em conta esta curva de crescimento da área pré-processada, o

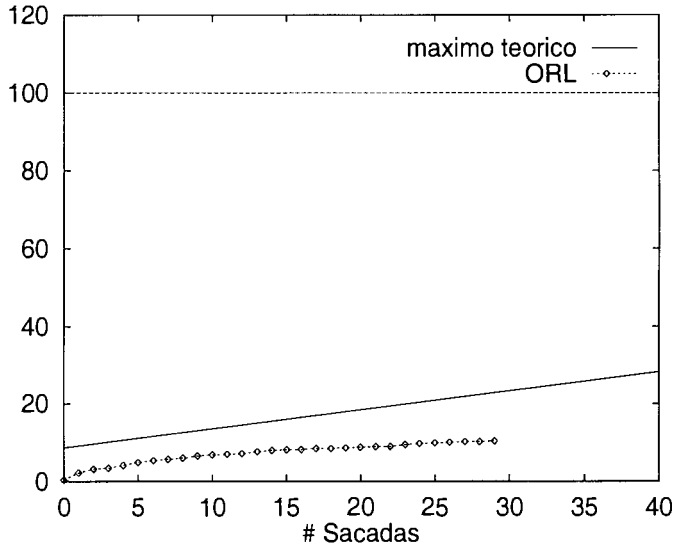

Figura 6.3: **Crescimento da área pré-processada:** percentagem em relação à área total de uma uma imagem com 92×112 pixels (linha horizontal), máximo teórico e área efetivamente pré-processada na simulação mais custosa com a base do ORL.

custo de pré-processamento no modelo incremental deve ser:

$$Custo_p(s) = 0,820CN^2 + 0,049(s+1)CN^2,$$

Custo de escolha do próximo ponto de fixação

O modelo incremental tem ainda um custo extra, que é devido ao processamento do Mapa de Saliência a cada sacada. Vamos assumir que este custo é equivalente ao custo de processamento para comparação com os modelos, por pixel:

$$Custo_m(s) = CN^2ms,$$

onde $Custo_m(s)$ é o custo de processamento do Mapa de Saliência em função do número de sacadas, e m é a fração da área total da imagem correspondente ao número de pixels do Mapa de Saliência. O valor de m corresponde à soma das áreas dos anéis 1, 2 e 3, pois como vimos no Capítulo 4, não é necessário calcular o Mapa de Saliência para o nível zero. Esta fração dá $0,0249$ (considerando o anel 1 como um disco inteiro, isto é, sem descontar a porção relativa à fóvea). Assim o custo de processamento do Mapa de Saliência fica:

$$Custo_m(s) = 0,0249CN^2s.$$

Lembre que o Mapa de Saliência não é processado para a sacada 0, pois o ponto inicial é arbitrário.

Agora podemos escrever as expressões do custo total:

Custo total do reconhecimento utilizando *toda* a imagem:

$$Custo_t = Custo_p + Custo_c = 11CN^2.$$

e

Custo total do reconhecimento incremental:

$$Custo_t(s) = Custo_p(s) + Custo_c(s) + Custo_m(s),$$

ou:

$$Custo_t(s) = CN^2(0,874 + 0,0788s).$$

O gráfico da Figura 6.4 mostra uma comparação entre o custo teórico máximo do reconhecimento no nosso modelo incremental, e o custo do reconhecimento se toda a imagem fosse utilizada para o reconhecimento.

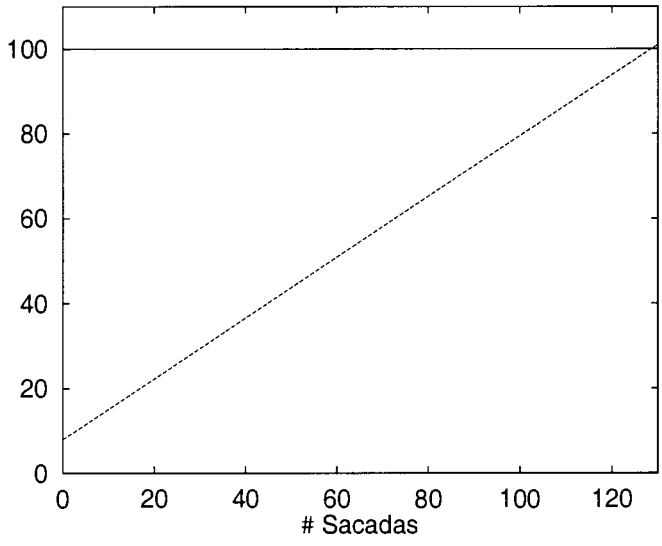

Figura 6.4: **Comparação do custo computacional:** custo do processo incremental em percentagem do custo de reconhecimento utilizando toda a imagem. No caso típico das imagens de 92 × 112 pixels, seriam necessárias 128 sacadas para que o custo máximo teórico do processo de reconhecimento no modelo atencional igualasse o custo de reconhecimento utilizando toda a imagem

O critério adotado para finalizar o processo de reconhecimento nas simulações com a base do ORL foi chegar a uma discriminação maior ou igual a 1,3 ou atingir 30 sacadas. Assim, o máximo de custo atingido nestas simulações corresponde ao custo de 30 sacadas. O máximo teórico para 30 sacadas é de 28,7% do custo de processar toda a imagem. Entretanto, o maior valor *efetivamente* atingido nas simulações corresponde a um pré-processamento de aproximadamente a metade do máximo teórico, como mostra o gráfico da Figura 6.3. Considerando a simulação mais custosa, o total máximo de custo efetivamente atingido em 30 sacadas foi de cerca de 17,2% do processamento não incremental.

Para uma comparação mais detalhada com algumas outras abordagens, é preciso levar em conta que aqui o processo de comparação com os modelos é relativamente simples, enquanto que numa abordagem como por exemplo a descrita por Lawrence et al. [33], que obtém um dos melhores resultados relatados para reconhecimento na base do ORL, a comparação envolve um processo bem mais sofisticado e custoso. Naquela abordagem, a comparação envolve o processamento de um número de pixels

de representações intermediárias em uma "rede convolucional" muito superior ao número de pixels da imagem. Na abordagem já clássica de Turk e Pentland [62], o processo de comparação é também bastante simples, sendo o treinamento do sistema a tarefa mais custosa, mas que não onera o processo de reconhecimento pois pode ser feito préviamente. Neste caso a diferença de custo em relação ao modelo descrito aqui recai exclusivamente na tarefa de pré-processamento. O trabalho de Turk e Pentland não menciona um procedimento de filtragem, mas se houver o pré-processamento da imagem para que o sistema não fique muito sensível a variações de iluminação, este terá que ser feito em toda a extensão da imagem.

Em resumo, a representação *space-variant* associada a uma estratégia atencional adequada demonstrou conseguir reconhecer imagens de faces em uma base de dados de referência (ORL) com uma taxa de erro de 17,4% (ver Tabela 5.2 no Capítulo 5), comparável ou mesmo inferior ao conseguido pelo método de análise de componentes principais (PCA) com o de Turk e Pentland [62], que apresenta taxas de erro de 26% e 10,5% [2]. Estes resultados foram conseguidos com um custo total de 17,2% do processamento de toda a imagem, muito inferior, portanto, ao custo do método de análise de componentes principais, caso fosse utilizado o mesmo tipo de filtragem como pré-processamento.

6.3 Discussão

6.3.1 Representação *space-variant* discreta

A representação *space-variant* comumente adotada na literatura apresenta resolução continuamente decrescente a partir do centro da fóvea (excetuando-se a singularidade no centro) [57, 7]. No entanto, uma representação *space-variant* "discreta" traz inúmeros benefícios para o sistema aqui apresentado de reconhecimento atencional, principalmente no que diz respeito ao armazenamento dos modelos no sistema. Como explicado no Capítulo 3, com a representação em pirâmide de resoluções é possível armazenar os modelos com uma extensão total de 1,33 vêzes o tamanho da imagem original (assumindo-se uma pirâmide em 4 níveis, cada nível com 1/4 do número de pixels do nível precedente). Este valor é aceitável, já que é próximo do custo de armazenar as próprias imagens originais. Além disso o custo de pré-processamento associado aos modelos ocorre uma única vez em "off-line". Uma vez que a imagem filtrada é armazenada na forma de pirâmide, pode-se facilmente computar a representação *space-variant* discreta associada a qualquer ponto de fixação, bastando para isso utilizar apenas as regiões correspondentes a cada anel de seus respectivos níveis (como mostrado na Figura 6.2). Assim, durante o processo de reconhecimento, à medida que o Módulo Atencional for escolhendo os pontos da imagem de onde extrair informações, uma representação *space-variant* centrada no ponto escolhido pode ser extraída, e comparada com as representações *space-variants* dos modelos centradas nos mesmos pontos. Note-se que a imagem de entrada *não* é totalmente pré-processada. Ao contrário, apenas as regiões necessárias são filtradas, procedimento este que está de acordo com a filosofia incremental e atencional

[2]Esta variação se deve ao número de imagens utilizadas para construção do modelo de cada classe (ver tabela 9, em Lawrence et al., 1997 [33])

adotada.

No caso de modelos adquiridos através de representações *space-variants* com decaimento contínuo da resolução (como no trabalho de Tistarelli, 1995 [58]), para que estas representações estivessem disponíveis para comparação, teriam que ser armazenadas imagens de cada modelo para todas as posições possíveis do ponto de fixação. Mesmo que não o fizéssemos para todos os pontos, considerando por exemplo somente posições com um espaçamento de 2, 4 ou 8 pixels, ainda assim teríamos que armazenar cerca de 51 vêzes, 13 vêzes ou 3 vêzes (respectivamente) mais informações, ainda assim com a desvantagem de ter apenas representações aproximadas.

6.3.2 Estratégias atencionais

Grande número de modelos encontrados na literatura propõe um sistema atencional acoplado a uma representação *space-variant* para o reconhecimento de imagens. Este tipo de representação, como vimos é extremamente econômica, mas exige um mecanismo de foveação, baseado em alguma estratégia atencional. Alguns modelos utilizam uma estratégia atencional somente baseada em características extraídas da imagem [54, 50, 68, 67, 4], outros associam a estas as informações dependentes da tarefa, ou seja, informações de natureza cognitiva ou da memória ("top-down") [21, 20, 30, 40, 49]. Outros ainda procuram modelar os mecanismos da atenção humana no sentido de simular algumas de suas capacidades, como por exemplo o controle do tamanho da região realçada pela atenção, ou a invariância a translações e mudanças de escala [61, 40]. Todos estes autores, no entanto, concordam que é fundamental para as tarefas visuais, e em particular para o processo de reconhecimento, o uso de uma estratégia atencional que utilize tanto as informações da imagem ("botom-up") quanto as informações da memória ("top-down"). Por outro lado, é reconhecido que a seleção da seqüência de pontos de fixação quando um sistema "vê" uma imagem pela primeira vez se relaciona com problemas muito complicados e fundamentais da busca visual. Enquanto as propriedades quantitativas da seleção pré-atencional de pontos de fixação são mais conhecidas, o controle top-down da atenção está relacionado com os processos cognitivos de alto nível, que são pouco compreendidos e fracamente formalizados [54]. Em nossa investigação pudemos verificar alguma hipóteses de funcionamento de um sistema atencional que utilize tanto o controle *top-down* da atenção quanto o processamento *bottom-up*.

Os mecanismos atencionais baseados na imagem nos sistemas biológicos operam através de sistemas inatos, com alto grau de paralelismo, rapidez, e independentes de esforço consciente. Já os mecanismos baseados na memória, ou de natureza cognitiva, são os processos dependentes do contexto, ou da tarefa, e requerem atenção seletiva, operando de forma sequencial, mais lenta e com esforço consciente [59, 30].

Mecanismos baseados na imagem, ou *bottom-up*

Reisfeld (Reisfeld et al., 1995 [50]), reconhecendo que o comportamento atencional humano é altamente dependente da tarefa, argumenta que este comportamento se baseia inicialmente (crianças de até 4 anos) em características da imagem, sendo portanto guiado por mecanismos inatos, não dependentes do contexto, já que so-

mente estes mecanismos estariam disponíveis precocemente. Só gradualmente, à medida que são aprendidas mais informações a respeito do ambiente, aparecem os processos de mais alto nível, dependentes do contexto ou da tarefa. Ele propõe que sejam desenvolvidos algoritmos robustos e eficientes de baixo nível para serem usados em visão computacional, posto que estes seriam elementos básicos de construção dos mecanismos atencionais artificiais. Estes autores enfatizam também uma outra razão para a importância dos mecanismos "bottom-up": uma das mais importantes características dos sistemas atencionais é a detecção de sinais *não esperados*. Estes sinais evidentemente precisam ser detectados na imagem, mesmo em pontos não importantes para a tarefa em curso e portanto é preciso haver a possibilidade de que características extraídas da imagem sejam capazes de atuar no sistema atencional rivalizando com a orientação exclusivamente baseada na memória ou na tarefa.

O primeiro passo em práticamente todos os sistemas de visão artificial é algum tipo de detecção de contornos. Este passo, presente em geral nos primeiros estágios dos sistemas visuais biológicos, leva a um simples mecanismo de escolha das regiões de interesse na imagem, que seria a dectecção de concentração de arestas ou contornos [67]. Esta determinação pode ser feita com o uso de diversos operadores baseados na variação da intensidade dos tons de cinza. Em nosso trabalho utilizamos um detetor de contraste para determinar contornos orientados (descrito no Capítulo 3). Este tipo de operador produz respostas mais intensas em bordas onde há grande diferença de tons de cinza e também tende a produzir respostas mais fortes onde uma borda muda bruscamente de direção, pois ali responderão filtros orientados em mais de uma direção. Outros detectores mais complexos podem ser empregados: detectores de fim de linha como em Aonishi e Fukushima, 1994 [4], ou o utilizado em Yeshurun e Schwartz, 1989 [68], que detecta mudança de orientação do contorno. Há ainda o interessante operador proposto por Reisfeld, Wolfson e Yeshurun [50], que localiza simetrias, isto é, cria um mapa de simetria baseado em uma prévia detecção de bordas.

Muitos modelos que utilizam representações *space-variant* associadas a um mecanismo atencional procuram fazer algum tipo de normalização na pose do objeto a ser reconhecido, detectando pontos de interesse na imagem que vão servir como pontos de ancoragem para auxiliar a normalização da pose como em Ribak et al. [54], ou simplesmente para classificar o objeto apresentado, como em Aonishi e Fukushima, 1994 [4]. Como já vimos (Capítulo 4), este tipo de modelo se aplica bem ao reconhecimento de um tipo específico de objetos, quando se pode de antemão determinar que caracteríticas do objeto podem servir para a normalização de pose, ou para a classificação. Nosso modelo não é específico para um tipo de objetos, e não emprega nenhuma hipótese *à priori* para detectar nenhum tipo específico de característica no objeto apresentado. O sistema atencional estudado aqui, portanto, procura regiões de interesse com a única finalidade de extrair dali informações para comparação direta com os modelos, num processo de reconhecimento (identificação) e não de classificação. Não se pode, portanto, comparar diretamente a eficiência desta estratégia atencional em nosso modelo com a eficiência nestes outros modelos, dado que a finalidade com a qual é empregada é diferente. Uma conclusão em comum, porém, pode ser tirada: é a de que a orientação da atenção em um sistema de multi-resolução deve ser uma associação entre os sinais *bottom-up* e *top-down*, ou seja, é preciso preservar a capacidade do sistema de detectar informações não espera-

das e regiões de fortes contornos na imagem, durante um comportamento atencional dirigido pela tarefa.

Os resultados das simulações mostradas nos Capítulos 4 e 5 confirmam isto quando apresentam resultados relativamente fracos da estratégia baseada na imagem (BU) quando utilizada isoladamente (ver a Tabela 5.3, que resume os resultados das tabelas 5.1 e 5.2). Observe que nas linhas *índice* o valor 1,0 indica a melhor estratégia para aquela base de dados. A estratégia estritamente "bottom-up" isolada (BU) produz um indice de erros muito maior relativamente (2,54) na base ORL do que na base de Colúmbia (1,37). Como 'ja analisado no Capítulo 4, isso se deve a que na base mais dispersa (Colúmbia) as posições do contorno de um objeto são mais discriminativas do que os contornos das faces da base do ORL.

Estratégia estritamente baseada nos modelos, ou *top-down*

Por outro lado, como vimos no Capítulo 4, os Roteiros de Sacadas obtidos em uma estratégia estritamente "top-down" são idênticos para todos os objetos apresentados (ver Figura 4.11), refletindo apenas a influência dos modelos (que são os mesmos). Os valores das taxas de erros nas duas bases que vemos na Tabela 5.3 são bastante fracos. Simulações extensas precisam ser feitas usando a associação da estratégia estritamente "top-down" com as outras estratégias para se poder estabelecer seu valor relativo.

Estes resultados confirmam o conceito de que, apesar de ser um comportamento visual altamente orientado pela tarefa, o reconhecimento não pode prescindir das informações provenientes dos níveis mais baixos para dirigir a atenção. A investigação das estratégias híbridas veio corroborar a expectativa de melhores resultados quando se associam veio corroborar a expectativa de melhores resultados quando se associam as duas fontes de informação. Nas duas bases de dados testadas, as estratégias híbridas indiretas (colunas Vy e V^H na Tabela 5.3) mostraram resultados melhores do que as estratégias estritamente "top-down" ou "bottom-up" (colunas TD e BU na mesma tabela).

Estratégias híbridas indiretas

O resultado mais importante de toda a investigação foi o de que a estratégia híbrida indireta da variação das categorias mais ativas utilizada isoladamente (coluna V^H, Tabela 5.3) mostra resultados muito bons nas duas bases de dados, sendo a melhor na base do ORL. Isso mostra que conseguir resolver a ambiguidade entre os modelos mais parecidos com a imagem apresentada é a tarefa mais importante durante o processo de reconhecimento, e por isso esta estratégia mostra tão bons resultados.

As desvantagens desta estratégia híbrida parecem relacionadas à possibilidade de um erro inicial incluir a categoria correta no rol das menos ativas, fazendo com que ela não influencie a orientação das sacadas. Quando o Roteiro de Sacadas depende da atividade das categorias ativadas inicialmente, um erro inicial (baixa ativação da categoria correta) pode demorar a ser corrigido. Este problema sugere uma linha de investigação futura, que seria modelar um sistema capaz de "esquecer"

os dados extraídos nas primeiras sacadas, já que estas informações iniciais podem não contribuir para discriminar a categoria correta, e podem, além disso, retardar o crescimento da ativação da mesma ou das mais semelhantes à imagem apresentada.

Uma comparação com outros modelos de reconhecimento usando a mesma base de dados do ORL apresentada na Tabela 6.2 mostra que os resultados desta estratégia rivalizam com os obtidos pela aplicação da analise de componentes principais (PCA) ("Eigenfaces" [62], ver Capítulo 2). A comparação pode ser feita com os resultados da PCA quando usadas 5 imagens da mesma pessoa para representar uma classe. Neste caso o conjunto de treinamento contém 200 imagens (40×5) e o conjunto de teste outras 200 imagens. Estes dois conjuntos são disjuntos. No nosso Modelo de Reconhecimento Atencional, cada classe é representada por um modelo obtido calculando-se a média entre as 5 representações de cada pessoa utilizadas no treinamento. Os testes foram feitos só com as imagens que não foram utilizadas no treinamento. No caso da PCA, o modelo de cada classe foi construído de duas maneiras: utilizando a média das 5 representações em "eigenfaces" das imagens de uma mesma pessoa no conjunto de treinamento (o número mínimo de "eigenfaces" utilizado foi 40), ou utilizando uma representação em "eigenfaces" para cada imagem utlizada no treinamento, formando 5 modelos para cada classe. No primeiro caso nossos resultados foram melhores ($17,4\%$ de erros). No segundo caso os resultados da PCA foram melhores. Para uma comparação mais completa fazem-se necessárias novas simulações com o nosso sistema de reconhecimento, utilizando a segunda forma de construir os modelos.

A abordagem proposta por Lawrence et al. (Lawrence et al., 1997 [33]) apresenta os melhores resultados na base do ORL, associando análise de componentes principais a uma rede neural "convolucional" (CN). Esta rede tem um número de conexões muito grande, e demanda um treinamento bastante custoso. Esta rede foi testada também em associação com uma rede neural "SOM" ("self-organizing map" introduzida por Kohonen [32]), com melhores resultados que na associação com a PCA. O custo de classificação em ambos os casos é proporcional ao número de conexões da rede CN que é algumas ordens de grandeza superior ao número de pixels da imagem.

Apesar de não apresentar melhores resultados em termos de número de erros em todos os casos, o nosso modelo de reconhecimento apresenta um custo muito baixo, necessitando processar somente $10,7\%$ da imagem no pior caso.

A outra estratégia híbrida indireta, a que usa as ativações das categorias para ponderar sua influência no Mapa de Saliência, mostrou resultados razoáveis nas duas bases de dados (coluna Vy na Tabela 5.3), mas não tão bons quanto a que usa *apenas* as categorias mais ativas. Esta estratégia, no entanto, não pode ser associada às outras com facilidade, porque os pesos usados para ponderar a influênciade cada modelo dificultam uma normalização do peso desta estratégia em relação às outras. Nos outros casos estudados, esta normalização pode ser feita simplesmente dividindo-se a soma das influências de cada modelo pelo número de modelos. Por este motivo investigamos apenas as estratégias híbridas formadas pela associação das outras estratégias.

Modelos de reconhecimento	erros (%)	custo de pré-proc. (%)	custo total de processamento (%)
Modelo de Rec. Atencional	17,4	10,7	17,2
Eigenfaces - media/classe	26,0	–	100
Eigenfaces - 1/imagem	10,5	–	100
PCA+CN	7,5	100	muito maior
SOM+CN	3,8	100	muito maior

Tabela 6.2: **Comparação de desempenho do Modelo de Reconhecimento Atencional com outros modelos não-atencionais (Eigenfaces, de Turk e Pentland [62] e PCA+CN e SOM+CN de Lawrence et al. [33]).** Testes realizados com a base de dados do ORL. Utilizamos em nosso sistema a média de cinco imagens para construção do modelo de cada categoria. Os resultados do "Eigenfaces" são para 40 "eigenfaces" (em um conjunto de 200 imagens de treinamento) e os modelos construídos usando a média de todas as imagens do conjunto de treinamento ou usando separadamente toda as imagens de cada pessoa (deste conjunto) para formar vários modelos da mesma classe. Os resultados de PCA+CN e SOM+CN utilizam uma "rede convolucional" (CN), cuja complexidade é proporcional ao número de conexões desta rede, que é cerca de duas ordens de grandeza maior que o número de pixels da imagem, aqui também totalmente processada. No Modelo de Reconhecimento Atencional, no caso mais custoso apenas $10,7\%$ da imagem foi processada. O maior custo total efetivamente encontrado nas simulações foi $17,2\%$ do custo de reconhecimento usando toda a imagem. Em todos os casos, os conjuntos de treinamento e de teste são disjuntos e contém 200 imagens cada um (40 pessoas, 5 imagens por pessoa). Os dados dos outros modelos são de Lawrence et al [33].

Estratégia híbrida direta

A estratégia de Desconfirmação das categorias menos ativas, pela maneira como foi concebida, não se presta muito bem a ser utilizada isoladamente, e isto foi confirmado pelos resultados fracos nas duas bases de dados (coluna D^L na Tabela 5.3). Entretento sua associação com a estratégia da variação das categorias mais ativas provou ser uma das melhores, sendo a melhor na base de Colúmbia (coluna $H260$ na Tabela 5.3). Graças ao fato de o conjunto das categorias menos ativas (L) ser com frequência bem maior que o conjunto das mais ativas (H), esta estratégia consegue contribuir bastante para superar os erros iniciais, quando a categoria correta fica inicialmente pouco ativada. Esta virtude se deve a que, sendo o conjunto L razoavelmente extenso, a soma das diferenças entre a imagem e os modelos pode ser alta mesmo nos pontos onde o modelo correto difere pouco da imagem, no caso da categoria correta estar no conjunto L. Assim, se houver um erro inicial, uma escolha de localização baseada nesta estratégia tende a contribuir para que as categorias incorretas diminuam sua ativação, deixando a correta sobressair. Caso o conjunto L contenha poucas categorias, as diferenças entre a imagem e o modelo correto poderão dominar, atraindo as sacadas para regiões onde estas diferenças serão enfatizadas, prejudicando o reconhecimento. Tambem pudemos observar que a asssociação desta estratégia com a que leva em conta as categorias mais ativas é

mais profícua na base mais dispersa.

Estratégias híbridas plenas

Outro resultado importante em relação às estratégias híbridas é que é possível obter resultados melhores associando adequadamente estratégias com resultados mais fracos individualmente, como foi tanto o caso acima como o caso da coluna H261 da Tabela 5.3 para a base de Colúmbia, que representa a associação da estratégia "bottom-up" com a estratégia da variação das categorias mais ativas e com a estratégia da desconfirmação. Muitas estratégias híbridas podem ser construídas, combinando de formas diferentes as informações dos modelos e das imagens. Uma questão importante a ser investigada é o modo de ajustar os coeficientes (α, β e γ, na Equação 4.2, Capítulo 4) que definem a contribuição de cada estratégia. Uma análise mais aprofundada do ajuste destas constantes deve ser feita, o que sugere uma linha de investigação posterior no sentido de dotar o sistema de meios automáticos de ajuste, ou a aplicação de algum tipo de algorítmo, como por exemplo um algorítmo genético, para otimizar estes coeficientes.

6.3.3 Outras questões

Dependência da informação inicial

A primeira sacada extrai informações parciais que já vão servir para a construção do Mapa de Saliência, e portanto para dirigir as próximas sacadas. Como a massa de informações da imagem presente no sistema após esta primeira sacada é muito pequena, o comportamento do sistema vai depender muito do ponto (ou pontos) onde se iniciou o processo, pois, em pricípio, o potencial discriminativo de cada região varia ao longo da imagem. Se a primeira sacada for para uma região mais informativa, com maior poder de discriminação, o comportamento do sistema tende a ser melhor que se começar em uma região pouco informativa. Por isso, em nossas simulações, procuramos variar o ponto inicial para obter uma média de comportamento que não trouxesse uma distorção devida à posição da sacada inicial. Esta influência está ilustrada na Figura 6.5.

Rejeição de uma imagem estranha

A Figura 6.6 mostra a evolução da discriminação da categoria correta no processo de reconhecimento da imagem de um estranho ao conjunto de modelos. A discriminação cresce no início, quando a informação inicial é insuficiente, indicando alta probabilidade de uma categoria incorreta. Com mais sacadas o sistema rejeita corretamente a imagem, indicando baixa probabilidade de todos os modelos, e portanto baixa discriminação, nunca atingindo o critério de decisão.

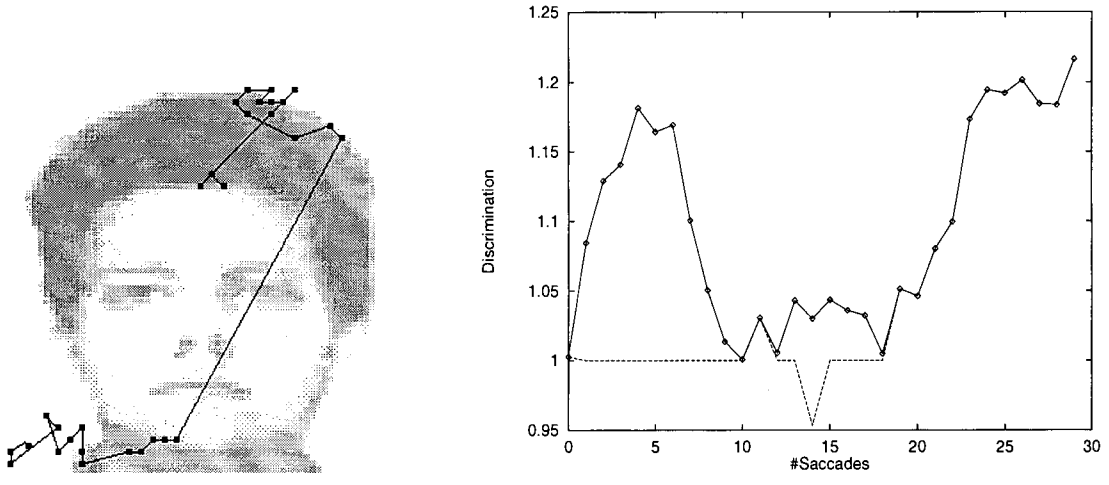

Figura 6.5: **Dependencia da informação inicial:** Uma sacada para uma região pouco informativa pode levar o sistema a uma hipótese inicial incorreta. Este fato é mostrado no gráfico nas regiões em que aparece a linha pontilhada que corresponde aos valores da discriminação da categoria correta. A outra linha (sólida) mostra os valores da discriminação da categoria mais ativa. Depois de 19 sacadas o sistema consegue informação suficiente para reconhecer corretamente a face, o que é evidenciado pelo fato de que a discriminação da categoria correta é a discriminação da categoria mais ativa, assume valores maiores do que 1, e apenas uma linha aparece no gráfico.

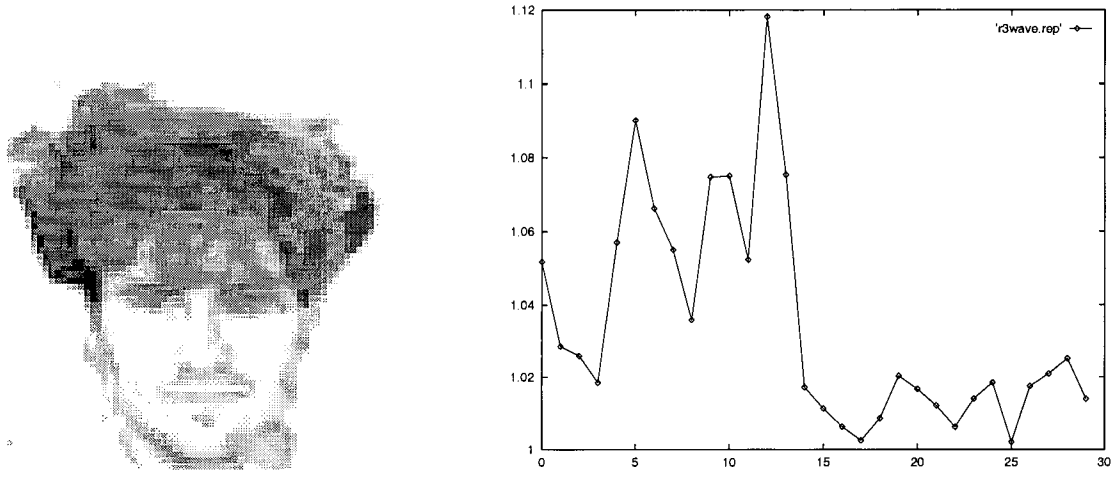

Figura 6.6: **Rejeição de uma imagem estranha ao conjunto de modelos:** Esquerda: imagem estranha. Direita: variação da discriminação com as sacadas; o crescimento inicial da discriminação indica uma hipótese incorreta, seguida de um decaimento correto indicando rejeição.

"Inibição de retorno"

Os modelos de visão computacional que utilizam mecanismos atencionais tem sempre que lidar com o problema de inibir, ou impedir, a foveação repetida do mesmo

ponto ou pontos muito próximos. Nos sistemas biológicos é conhecida a presença da inibição de retorno com esta função. Na maioria dos modelos descritos na literatura esta inibição é implementada através de uma função que decresce com o tempo (ou número de sacadas) [4, 20, 30]), ou propõem uma inibição permanente ou outros tipos de função, por exemplo enfatizando regiões mais distantes do atual ponto de fixação [68]. Em nosso modelo esta inibição foi feita simplesmente impedindo sacadas para regiões dentro de um pequeno raio (parâmetro do sistema) em torno de pontos já visitados. Este dispositivo se revelou indispensável para que o sistema não oscilasse simplesmente entre dois pontos mais salientes. Yeshurun e Shwartz (1989 [68]) propõem que esta inibição tenha um componente aleatório.

Atrator de sacadas

Observamos também em diversas simulações um outro tipo de comportamento muito interessante: aparentemente, para cada conjunto de parâmetros do sistema, existem " atratores" que atuam em uma determinada imagem, fazendo com que o Roteiro de Sacadas, não importa onde ele se inicie, mais cedo ou mais tarde acabe sempre por percorrer um determinado caminho. O significado deste comportamento ainda não está totalmente claro, mas sugere que este modelo de reconhecimento, apesar de um pouco sensível às condições iniciais, isto é, a qual o ponto da imagem escolhido para a sacada inicial, é suficientemente robusto para encontrar os melhores pontos de fixação mesmo que o ponto inicial seja desfavorável. Uma investigação sugerida é se este comportamento pode estar modelando algum comportamento encontrado em sistemas biológicos.

Invariância

Neste trabalho, o modelo de reconhecimento não inclui nenhum mecanismo para conseguir invariância de forma, além da que pode ser conseguida pela representação *space-variant*. Invariância a diferentes condições de iluminação e ruído é conseguida através da utilização de filtros no pré-processamento. Os resultados mostraram que apesar disso podem ser conseguidas altas taxas de sucesso nas bases de dados investigadas.

Este modelo de reconhecimento não é de modo nenhum incompatível com a utilização de métodos de pré-processamento que procurem conseguir invariância de pose ou de escala. Este problema não foi investigado aqui porque a meta era pesquisar o componente atencional, mas este modelo pode servir de referência para modificações futuras que incorporem algum mecanismo de normalização da posição do objeto a ser reconhecido em relação a translação ou pose ou escala. Uma investigação preliminar foi feita construindo um sistema que busca um objeto (face) em uma cena, procura extrair uma imagem deste objeto alinhada com os modelos armazenados e por fim efetuar incrementalmente o reconhecimento [43].

6.4 Exploração e reconhecimento de faces em cenas

Uma importante tarefa visual consiste em *exploração de cenas*, durante a qual toda uma cena é explorada seqüencialmente. Neste processo, objetos familiares são reconhecidos — enquanto ao mesmo tempo, objetos desconhecidos são rotulados como tais — por uma fóvea de alta resolução que se desloca ao longo da cena. Este comportamento deve ser distingüido da busca visual, na qual um dado objeto ou característica é ativamente procurada na cena. Embora durante a exploração da cena, informações *top-down*, tais como expectativas de faces humanas possam estar em jogo, nenhum realce explícito de características de baixo nível (como arestas horizontais, ou regiões verdes) deve ser assumido. Não obstante, o deslocamento seqüencial da fóvea ao longo da cena não é randômico. Uma questão chave é, então, estabelecer os fatores que determinam o roteiro de exploração em cenas. Nosso objetivo, ao adaptar nosso modelo de reconhecimento atencional para exploração de cenas, não foi replicar os roteiros humanos de exploração, mas sim investigar como uma fóvea com alta resolução pode efetivamente se deslocar através de uma cena para reconhecer objetos de interesse e ignorar outros menos relevantes. Para um exemplo concreto, aplicamos nosso modelo à tarefa de reconhecer faces em cenas, embora ele não tenha sido projetado especialmente para este domínio. Este exemplo ilustra os resultados preliminares de uma pesquisa em andamento [43, 42].

6.4.1 Descrição do modelo de exploração de cenas

A Figura 6.7 mostra o esquema macroscópico do sistema. A cena é inicialmente representada como uma pirâmide de resoluções construída a partir do mapa de respostas das células complexas, onde estão disponíveis informações sobre contornos orientados. Todo o processo de exploração de cenas é baseado nesta estrutura em pirâmide (ver Figura 6.8). O sistema inclui dois sub-sistemas principais: um *sistema de exploração* e um *sistema de decisão* que interagem para determinar como regiões da cena devem ser investigadas. A operação do sistema alterna entre um *modo de exploração* e um *modo de reconhecimento*. No modo de exploração, são determinadas regiões de interesse. No modo de reconhecimento, uma região de interesse é cuidadosamente examinada enquanto procura processar o reconhecimento nesta região particular da cena. Inicialmente, toda a cena é usada para gerar um *mapa de interesse*, que determina as regiões de interesse a serem investigadas mais cuidadosamente.

Uma vez que o reconhecimento ocorra para um determinado objeto, o sistema muda para o modo de exploração. O processo global de exploração-reconhecimento interage até que não haja mais nenhuma área com alta ativação no mapa de interesse. Adiante os componentes do sistema são especificados em detalhe.

Regiões de interesse são determinadas por um *mapa de interesse* computado com base tanto em informações *bottom-up* quanto *top-down*. O componente *bottom-up* consiste no mapa de respostas das células complexas gerados pelos filtros orientados. O componente *top-down* é baseado no realce de formas *esperadas* (esboços de objetos). Uma classe de formas que comumente ocorre em cenas são faces. No exem-

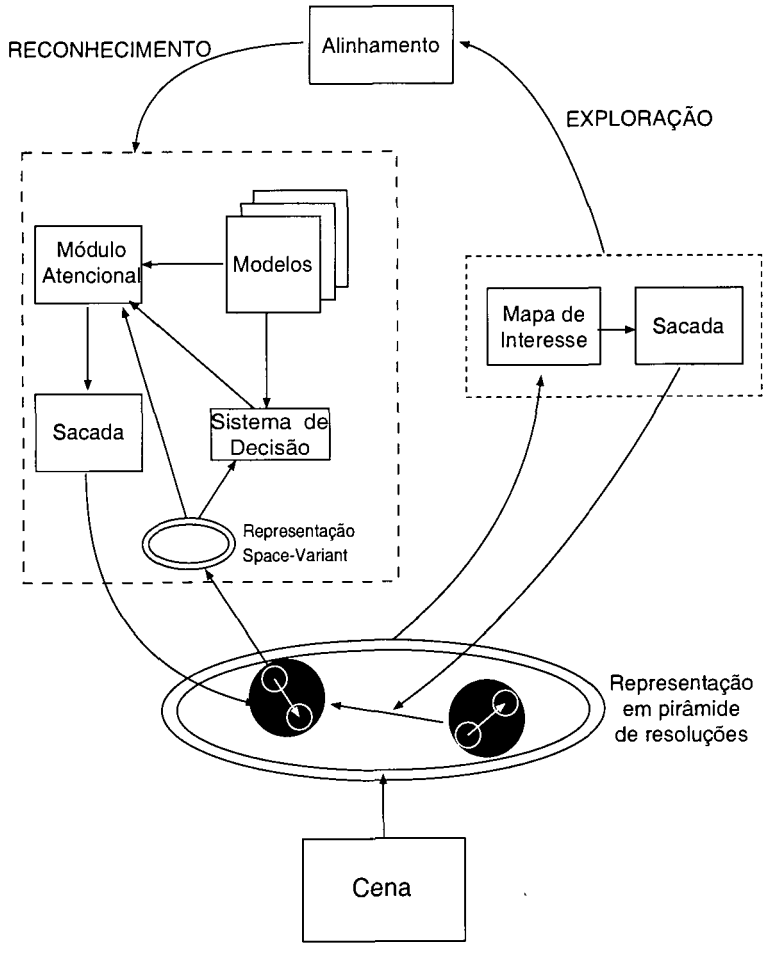

Figura 6.7: **Modelo de exploração atencional de cenas.** Uma representação multi-escalas de um mapa de respostas de células complexas é empregado pelos sistemas de *exploração* e de *reconhecimento*. Os discos pretos grandes representam regiões de interesse especificadas pelo *mapa de interesse*. As setas longas que as conectam são grandes *sacadas de exploração*. As pequenas áreas delimitadas por círculos brancos representam informações parciais extraídas pelas foveações de reconhecimento, de modo a reconhecer incrementalmente objetos de interesse. As setas curtas brancas são pequenas *sacadas de reconhecimento*.

plo mostrado aqui, o componente *top-down* consiste no realce de faces. Um mapa de respostas das células complexas para uma face típica determina uma máscara prototípica que é usada então para procurar na cena por esboços de formas semelhantes. Aqui tiramos partido da estrutura em pirâmide para, em vez de procurar na resolução mais alta (a resolução original da cena), procurar em uma resolução mais baixa (penúltimo nível na pirâmide da Figura 6.8).

Sacadas de exploração

O mapa de interesse determina regiões a serem investigadas. Quanto mais alta for a ativação no mapa de interesse, maior a precedência conferida ao exame de uma dada região. Depois que o reconhecimento é tentado em uma região, uma nova região da cena é visitada. Esta deve necessariamente estar posicionada à alguma distância dos locais investigados anteriormente, devido a uma inibição de longo termo das

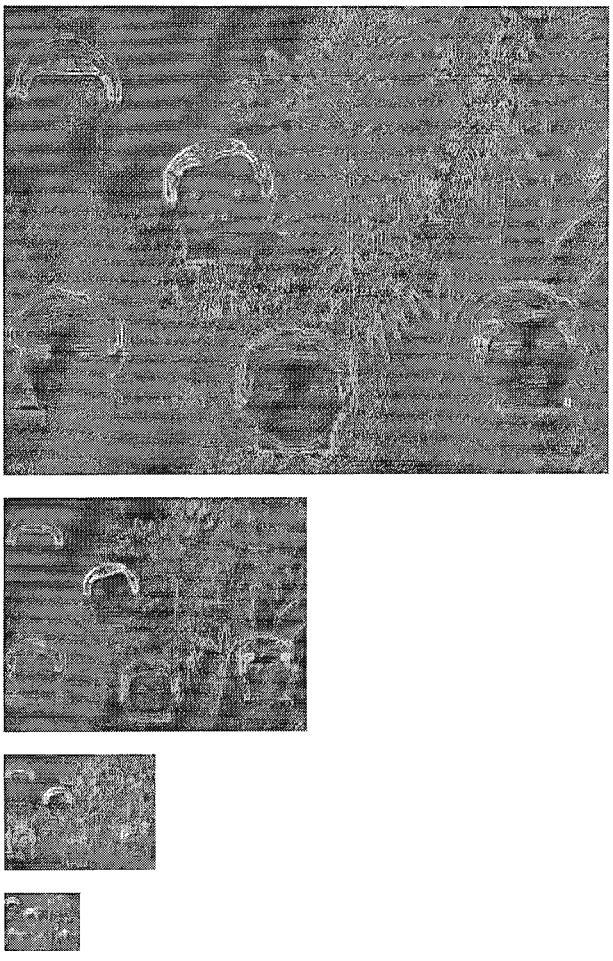

Figura 6.8: **Representação da cena em pirâmide de resoluções.** As 4 imagens mostram a representação em pirâmide de resoluções das respostas das células complexas. Note-se que, apesar de mostrados os 4 níveis desta representação, as respostas das células complexas não precisam ser computadas para toda a cena em todos os níveis. Sua extração pode ser determinada pela ativação do mapa de interesse do sistema de exploração e pelo mapa de saliência do sistema de reconhecimento.

regiões próximas a um local já visitado. Deste modo, longas sacadas de exploração são geradas, e outras regiões da cena são visitadas.

6.4.2 Sistema de reconhecimento

Alinhamento dos modelos com a cena

Uma vez que as regiões de interesse são determinadas, o processo de reconhecimento pode ser iniciado. A cena e os modelos tem diferentes tamanhos e origens, definindo um *espaço da cena* e um *espaço do modelo.* Como a representação adotada aqui é topográfica (respostas orientadas em localizações espaciais dadas), é necessário alinhar os modelos com a cena. Mas como se pode alinhar cena e modelo sem resolver o próprio problema colocado de reconhecer um objeto em uma posição de interesse na cena?

Alinhamento por deslisamento de todos os modelos na região da cena indicada pelo ponto de interesse corrente é impraticável, dado o alto custo computacional.

Nossa sugestão é fazer uso da estrutura em pirâmide de resoluções adotada, de modo a reduzir em muito as computações. O deslizamento é feito então no mais baixo nível de resolução da pirâmide. Além disso, apenas uma pequena janela em torno do ponto de interesse é usada para determinar a correlação entre cena e modelo. Em nossas simulações obtivemos resultados confiáveis com uma janela de 3×3 pixels.

Qual é o resultado do alinhamento? O Alinhamento produz uma lista de posições em *modelos candidatos* (modelos armazenados). Estas posições permitirão encontrar na cena o ponto que deve coincidir com a origem do espaço de cada modelo candidato. Dada esta origem, pode ser extraído da cena a região que vai servir de entrada para o processo de reconhecimento. A lista é então visitada pela ordem dos valores de correlação enquanto o sistema tenta reconhecer o objeto. Em resumo, o processo de alinhamento procura gerar rapidamente uma lista de modelos candidatos *potenciais*, para então disparar um processo mais custoso de reconhecimento.

Reconhecimento

O reconhecimento do objeto presente na cena, uma vez feito o alinhamento, será processado pelo sistema de reconhecimento atencional descrito nesta tese, ou seja, usando informações parciais extraídas incrementalmente da região eleita na cena. Neste processo, serão executadas as pequenas sacadas de reconhecimento, enquanto o sistema compara informações parciais extraídas da região alinhada na cena com os modelos.

6.4.3 Simulações

Doze imagens de faces em um fundo branco uniforme foram inicialmente armazenadas no sistema [3]. Para testar o sistema, as imagens originais foram coladas em posições arbitrárias sobre um fundo com uma imagem de textura rica. Uma simulação típica é mostrada na Figura 6.9. A imagem original é inicialmente transformada em uma pirâmide de resoluções, e o nível 2 (o penúltimo) é representado pela resposta das células complexas. Esta representação é então usada para determinar o mapa de interesse, que especifica as regiões que devem ser examinadas. Nesta versão foi usado o realce de faces para determinar as ativações do mapa de interesse. Este realce é dado pela correlação entre uma máscara representando uma face (mapa de respostas de células complexas) e a imagem da cena. O máximo local das medidas de correlação especifica então o centro das regiões de interesse.

Na simulação mostrada, 8 regiões candidatas foram obtidas (marcadas com uma cruz ou um círculo, na Figura 6.9). As 8 regiões candidatas foram então visitadas de acordo com o valor da ativação no mapa de interesse. No processo, 4 faces foram reconhecidas corretamente (círculo) e uma não foi reconhecida (cruz). Três das faces foram reconhecidas já na sacada inicial, enquanto a outra disparou duas pequenas *sacadas de reconhecimento* (face à esquerda, embaixo). Note também que uma das faces foi reconhecida mesmo sendo o ponto de interresse marcado um pouco acima da face (ultima à direita). Neste caso, as informações provenientes de região em torno do ponto de interesse foram suficientes provenientes de região em torno do ponto

[3] As imagens usadas pertencem à base de dados "MIT Eigenfaces database" [62], mostrada no Capítulo 5.

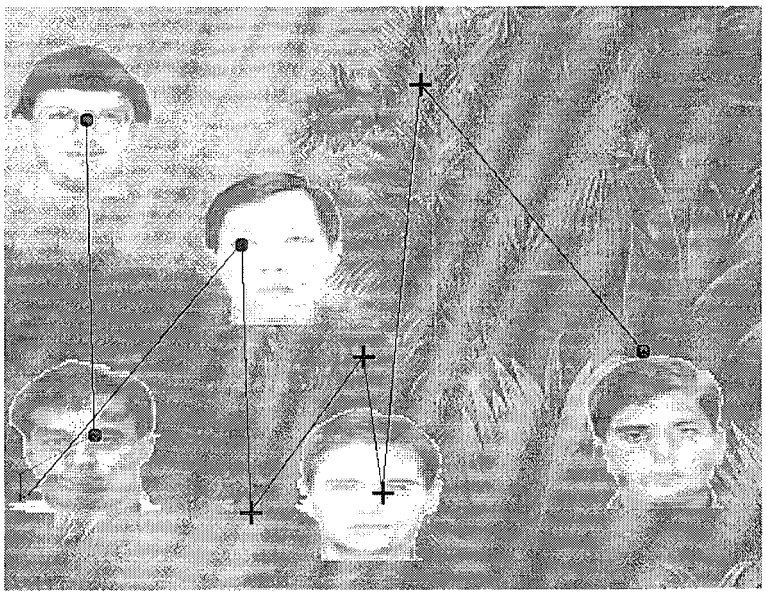

Figura 6.9: **Exploração e reconhecimento:** De um conjunto de 8 regiões candidatas, 4 faces foram reconhecidas corretamente (marcadas com círculos), uma perdida (cruz), e três corretamente rejeitadas.

de interesse foram suficientes para levar ao reconhecimento com os mesmos critérios de confiança usados nas outras faces [4]. O restante dos pontos do fundo visitados foram rejeitados de pronto, pois nenhuma pequena sacada de reconhecimento foi disparada. Esperamos que, no futuro, métodos mais robustos de realce inicial e uma computação mais elaborada do mapa de interesse possa eliminar estes pontos indesejáveis do fundo. Entretanto, em geral, sempre haverá falsos positivos a serem verificados, se não se deseja perder objetos importantes.

[4]Para a face à esquerda embaixo, as sacadas para o fundo também extraem informações da face.

Capítulo 7

Conclusões

Nesta tese, foi apresentado um novo modelo de reconhecimento visual atencional de aplicação geral, isto é, não restrita a um determinado tipo de objetos, e investigado seu desempenho em diferentes condições. Este modelo segue a filosofia da visão ativa, e incorpora três aspectos principais: (1) o *reconhecimento incremental* (como proposto por Aguilar e Ross [2]), onde informações parciais são fornecidas aos dispositivos de reconhecimento, e novas informações obtidas quando necessário; (2) a adoção de uma *representação em múltiplas escalas e "space-variant"*, inspirada na organização do sistema visual dos primatas [58, 7]; e (3) a utilização de *mecanismos atencionais* capazes de orientar o processo de aquisição de informações (como indicado em 1) de forma a guiar o movimento da fóvea para obter informações da imagem em múltiplas escalas (como em 2). Embora diversas partes do sistema sejam baseadas em modelos propostos na literatura, a estrutura deste sistema e a integração de seus componentes, assim como o modo de utilizar estes conceitos é original, conforme será explicitado na seção seguinte onde aparecem as principais contribuições deste trabalho e os resultados da investigação realizada. Na última seção apresentamos as propostas de pesquisas futuras.

7.1 Contribuições e resultados

Proposta de um novo sistema de reconhecimento de objetos

Foi proposto e testado um novo sistema de reconhecimento de objetos, construído pela integração de diversos componentes comumente utilizados em sistemas de visão computacional, mas integrados de forma original (como exibido na Figura 7.1). Em contraste com o sistema que mais se assemelha a este, o proposto por Aguilar e Ross [2], aqui utilizamos plenamente uma estrutura *space-variant* para representar os dados extraídos da imagem. Este tipo de representação se acopla ao uso de um sistema atencional para guiar a extração de dados e assim realizar um reconhecimento incremental.

As extensas simulações realizadas comprovam que a concepção incremental deste sistema permite conseguir o reconhecimento com um mínimo de aquisição de informações extraídas da imagem apresentada. Comparado com o modelo tradicional de reconhecimento por Análise de Componentes Principais (PCA) de Turk e Pentland [62, 33], que processa *toda* a imagem, o nosso modelo, aplicado à base de faces

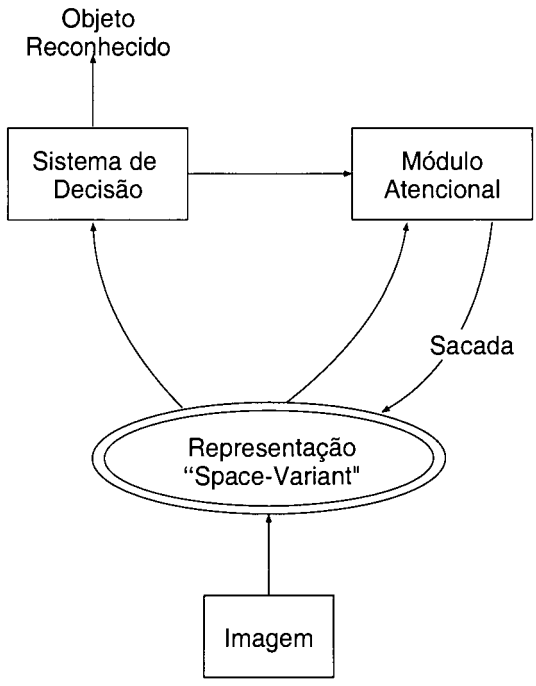

Figura 7.1: **Modelo de Reconhecimento Atencional:** Uma representação *space-variant* centrada em um ponto de fixação inicial é extraída da imagem apresentada para reconhecimento. O Sistema de Decisão compara esta representação com outros modelos armazenados no sistema. Caso não seja conseguido o reconhecimento, o Módulo Atencional determina um novo ponto de fixação, isto é, o centro de uma nova representação *space-variant* a ser extraída da imagem. O ciclo se repete até que seja conseguida confiança suficiente no reconhecimento ou um número máximo de "sacadas" seja atingido.

do ORL conseguiu resultados 12% melhores (taxa de erros de 17,4% contra 26,0%, quando usada a média de 5 imagens para formar o modelo de uma catagoria), extraindo apenas 10,7% de dados das imagens, no caso mais custoso (ver tabela 6.2). Quando são usadas as representações de 5 imagens da mesma pessoa para formar 5 modelos da mesma classe, o reconhecimento por "Eigenfaces" consegue uma taxa de erros menor (10,5%), mas o nosso processo de reconhecimento atencional não foi ainda testado com este método de construção de modelos. Note que a construção de modelos utilizada no reconhecimento por "Eigenfaces" utiliza 40 "eigenfaces" (em um conjunto de 200 imagens de treinamento). Por outro lado, os resultados de PCA+CN e SOM+CN de Lawrence e colaboradores são os melhores mostrados na literatura até aqui, para esta base de dados. Neste último modelo, a imagem é toda processada, e a complexidade do reconhecimento é proporcional ao número de conexões da "rede convolucional" (CN), que é cerca de duas ordens de grandeza maior que o número de pixels da imagem.

Avaliação qualitativa e quantitativa de estratégias atencionais *bottom-up*, *top-down* **e híbridas**

O problema central investigado nesta tese foi o de como selecionar regiões para onde a "fóvea", ou o centro de um sensor *space-variant* deve ser dirigido de modo a

coletar informações incrementalmente. Esta seleção determina o chamado Roteiro de Sacadas, ou seja, a seqüência de regiões da imagem de onde o modelo vai extrair informações. Conforme se verificou, tanto do ponto de vista quantitativo quanto qualitativo, esta seqüência influencia decisivamente no desempenho do modelo. Uma seleção adequada pode conduzir rápidamente à discriminação da categoria correta, levando ao reconhecimento com um mínimo de processamento da imagem. Em nossa investigação, verificamos algumas hipóteses de funcionamento de um sistema atencional que utiliza tanto o controle *top-down* da atenção quanto o processamento *bottom-up* e determinamos quais as estratégias capazes de levar a melhores resultados no processo de reconhecimento.

Além disso, esta tese contribui para preencher a lacuna deixada na literatura pela ausência de avaliações detalhadas, tanto do ponto de vista quantitativo quanto qualitativo, de estratégias atencionais em comparação umas com as outras e com outros métodos de reconhecimento. Em particular, como se verificou, a utilização de bases de dados com estruturas complementares (esparsas e densas) evidenciou as limitações das estratégias *bottom-up* e *top-down* quando usadas isoladamente.

A comparação entre os desempenhos das estratégias atencionais aparece nítidamente na Tabela 5.3. A comparacao da estratégia *bottom-up* com a estratégia *top-down* nas duas bases de dados confirma a tese de que o reconhecimento deve ser um processo híbrido, como advogam diversos autores [50, 67]. Ambas as estratégias, quando usadas isoladamente, dão resultados mais fracos que quando combinadas de alguma forma. Além disso, ficou clara a diferença de comportamento destas estratégias de acordo com a densidade das bases de dados. A estratégia *top-down* é melhor que a *bottom-up* para a base mais densa (ORL) e a posição se inverte na base mais esparsa (Colúmbia). Este comportamento pode ser entendido, na medida em que as informações *bottom-up* são menos importantes em um universo onde todos os objetos tem aproximadamente a mesma forma, pois nos contornos das imagens que são compartilhados por quase todos os objetos, não há informação rica. Por outro lado, se uma das mais importantes características dos sistemas atencionais é a detecção de sinais *não esperados*, e estes sinais precisam ser detectados na imagem, é preciso haver a possibilidade de que características extraídas da imagem sejam capazes de atuar no sistema atencional rivalizando com a orientação exclusivamente baseada na memória ou na tarefa. Este fato ganha importância na base mais esparsa (Colúmbia) como vemos pelos resultados na tabela 5.3.

O resultado mais importante de toda a investigação foi o de que a estratégia híbrida indireta da variação das categorias mais ativas (coluna V^H, tabela 5.3) mostra resultados muito bons nas duas bases de dados, sendo a melhor na base do ORL. Isso mostrou que conseguir resolver a ambigüidade entre os modelos mais parecidos com a imagem apresentada é a tarefa mais importante durante o processo de reconhecimento, e por isso esta estratégia mostra tão bons resultados. É interessante notar que esta atividade *é híbrida*, isto é, ela só pode ser levada a efeito utilizando informações tanto de baixo nível, proveniente da imagem, quanto cognitivas, de alto nível, provenientes do conhecimento armazenado na memória.

Outra conclusão importante é a de que é factível e potencialmente profícuo associar estratégias atencionais visando melhores desempenhos nas tarefas visuais de reconhecimento. Isto aparece nítidamente na tabela 5.3 nas colunas $H260$ e $H261$ para a base de Colúmbia, onde o resultado das associações foi melhor do que o de

cada estratégia isolada.

Utilização de uma estratégia atencional operando numa representação *space-variant*

A utilização da representação *space-variant* da forma como foi construída, isto é, com variação descontínua de resolução, se mostrou bastante econômica, pois utiliza apenas $2,67\%$ do número de pixels da imagem original (de 112×92), e ao mesmo tempo eficiente, pois é de implementação bastante simples, e se ajusta perfeitamente ao objetivo de permitir uma comparação com os modelos armazenados. Estes podem ser representados também de forma bastante econômica, na forma de pirâmides, ocupando uma extensão total de $1,33$ vêzes o tamanho da imagem original para cada modelo. Destas pirâmides se pode extrair então as representações correspondentes a sacadas para qualquer ponto da imagem, bastando para isso utilizar apenas as regiões correspondentes a cada anel de seus respectivos níveis. Além disso, esta representação torna disponíveis, ao mesmo tempo, informações de alta freqüência espacial, que pode ser crítica para eliminar a ambigüidade do reconhecimento, e informações de baixa freqüência, que podem ser fundamentais face a ruído, por exemplo, ou pequenas diferenças de pose e escala, quando os objetos armazenados são bastante distintos, pois aí as informações de alta freqüência podem se tornar inúteis.

Por outro lado, a construção de um Mapa de Saliência utilizando esta estrutura também se mostrou bastante eficiente, permitindo varrer um amplo "campo visual" combinando as virtudes dos diferentes níveis de resolução.

Custo computacional

O custo computacional do modelo de reconhecimento proposto nesta tese, em uma aplicação típica como no caso da base de dados do ORL, mostrou-se extremamente baixo, em comparação com os outros modelos sobre os quais existem dados de comparação. Este baixo custo deve-se principalmente ao caráter incremental do modelo, o que permite utilizar somente o mínimo de informação necessária para o reconhecimento; deve-se também a uma estratégia atencional adequada, para que as informações extraídas tenham qualidade suficiente para permitir o reconhecimento. Assim, o maior custo do processo, que é o de filtragem da imagem, incidirá apenas sobre uma área muito pequena em relação à área total da imagem. O gráfico da Figura 6.3 ilustra o pequeno pré-processamento necessário em uma simulação típica com 30 sacadas.

O gráfico da Figura 6.4 mostra a evolução do custo máximo teórico total do reconhecimento, num caso típico. Note-se que o custo máximo teórico em 30 sacadas é de $28,7\%$, enquanto que o custo efetivamente encontrado na simulação mais custosa foi de $17,2\%$ do custo do reconhecimento não incremental.

7.2 Pesquisas futuras

O sistema de reconhecimento aqui proposto, apesar de ser relativamente simples, permite aperfeiçoamentos em quase todas as suas partes. Em relaçõe às estratégias atencionais, que foram o centro de nossa investigação, o desdobramento mais imediato seria investigar métodos de melhorar a combinação das estratégias, através do ajuste dos coeficientes que modulam a participação de cada uma. Neste sentido abre-se um leque de propostas, desde usar um algorítmo genético para ajustar os coeficientes, até investigar meios automáticos de ajuste dinâmico dos mesmos. Outra linha seria tentar combinações mais sofisticadas como produto de estratégias, etc.

Quanto ao Sistema de Decisão, uma proposta imediata seria modelar um sistema capaz de "esquecer" os dados extraídos nas primeiras sacadas, já que estas informações iniciais podem não contribuir para discriminar a categoria correta, e podem, além disso, retardar o crescimento da ativação desta ou das mais semelhantes à imagem apresentada.

Em relação ao pré-processamento, como vimos, os níveis de resolução tem influência no grau de invariância da representação. Devem ser pesquisados meios de quantificar esta influência, visando otimizar o tamanho dos filtros utilizados e também o peso de cada nível de resolução, tanto no processo de comparação com modelos quanto na construção do Mapa de Saliência. Outra investigação que diz respeito ao pré-processamento é tentar, no caso de determinados objetos como faces, restringir a região das sacadas à região do rosto, através de alguma segmentação que evite características contingenciais como a forma do cabelo ou a roupa.

A exploração de cenas pode ser melhorada com outros processos de encontrar as regiões de interesse, com a pesquisa de outros métodos de alinhamento e também incorporando qualquer aperfeiçoamento no sitema de reconhecimento como os propostos acima.

Finalmente, um dos temas mais importantes no reconhecimento é o problema da invariância. Como vimos, o modelo aqui apresentado não incorpora nenhum recurso adicional para conseguir invariância nas representações além do conseguido nas filtragens. Este modelo, porém, não é incompatível com processos que procurem encontrar representações invariantes a translações, pose ou escala. As propostas preliminares no sentido de conseguir alinhamento entre modelos e imagem (Seção 6.4 [43, 42]), e outros mecanismos podem ser acoplados visando esta invariância. Neste sentido, parece promissor tentar tirar partido do modelo incremental de modo a conseguir mais facilmente a invariância, talvez modelando um sistema que dirija as sacadas através de um mapeamento que compense as deformações de pose em relação aos modelos, utilizando um referencial centrado no objeto como o proposto por Rybak e colaboradores [54].

Referências Bibliográficas

[1] E.H. Adelson, C.H. Andersen, J.R. Bergen, P.J. Burt, e J.M. Ogden. Pyramid methods in image processing. *RCA Engineer*, (29-6):33–41, 1984.

[2] M. Aguilar e W. Ross. Incremental ART: A neural network system for recognition by incremental feature extraction. *Proc. WCNN-93*, 1993.

[3] E. Alpaydin. Selective attention for handwritten digit recognition. *In: Advances in Neural Information Processing Systems (NIPS'95), MIT Press*, 1996.

[4] T. Aonishi e K. Fukushima. Eye movement model based on non-uniformity of the retina and feature extraction in the cortex. *Proceedings of International Conference on Neural Information Processing*, (3):1521–1526, 1994.

[5] D.H. Ballard e C.M. Brown. *Computer Vision.* Prentice-Hall, Englewood Cliffs, New Jersey, 1982.

[6] M.J. Black, J. Aloimonos, C.M. Brown, I. Horswill, J. Malik, G. Sandini, e M.J. Tarr. Action, representation, and purpose: re-evaluating the foundations of computational vision. *Proc. International Congress of Artificial Intelligence*, 1995.

[7] M. Bolduc e M. Levine. A review of biologicaly motivated space-variant data reduction models for robotic vision. *Computer Vision and Image Understanding*, (69):170–184, 1998.

[8] R. Brunelli e T. Poggio. Face recognition: Features versus templates. *IEEE PAMI*, páginas 1042–1052, 1993.

[9] C. Burbeck e S. Pizer. Object representation by cores: Identifying and representing primitive spatial regions. *Vision Research*, (13):1917–1930, 1995.

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

[11] G. Carpenter e Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*, (37):54–115, 1987.

[12] G. Carpenter, S. Grossberg, e D. Rosen. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, (4):759–771, 1991.

[13] I. DEBUSSCHERE, E. BRONCKAERS, E C. CLAEYS. A 2d retinal ccd sensor for fast 2d shape recognition and tracking. *Proceedings of the 5th International Solid-State and Transducers Conference, Montreux, Switzerland*, 1989.

[14] S. EDELMAN. Representation of similarity in 3d object discrimination. *Neural Computation*, (7):407–422, 1995.

[15] S. EDELMAN, D. REISFELD, E Y. YESHURUN. Learning to recognize faces from examples. *Proceedings of the 2th European Comference on Computer Visio*, páginas 787–791, 1992.

[16] S. EDELMAN E D. WEINSHALL. *Computational approaches to shape constancy. In: Perceptual Constancies.* V. Walsh and J. Kulikowski, Cambridge U. Press, 1994.

[17] S. EXEL E L. PESSOA. Attentive Visual Recognition . Em *Proceedings of the 14th International Conference on Pattern Recognition*, 1998.

[18] S. EXEL E L. PESSOA. Space-variant representation for active object recognition . Em *Proceedings of the International Symposium on Computer Graphics, Image Processing and Vision, SIBGRAPI'98, Rio de Janeiro, Brasil*, 1998.

[19] C. FERMULLER E Y. ALOIMONOS. Recognizing 3-d motion. *Proceedings of the 13th International Conference on Artificial Intteligence, Chambèry, France*, páginas 1624–1638, 1993.

[20] K. FUKUSHIMA. Neural networks for seletive looking. *Proceedings of International Conference on Neural Information Processing*, (3):1367– 1372, 1994.

[21] K. FUKUSHIMA E H. HASHIMOTO. Recognition and segmentation of components of a face by a multi-resolution neural network. *Lecture Notes in Computer Science - Artificial Neural Networks, 7th International Conference , preceedings*, (1327):931–937, 1997.

[22] K. FUKUSHIMA E S. MIYAKE. Neocognitron: a new algorithm for pattern reconition tolerant of deformation and shifts in position. *Pattern Recognition*, (15):455–469, 1982.

[23] L.G. GAWRYSZEWSKI, L. RIGGIO, G. RIZZOLATTI, E C. UMILTÁ. Movenents of attention in three spatial dimensionas and the meaning of "neutral" cues. *Neurofisiologia*, (1a):19–29, 1987.

[24] S. GROSSBERG E E. MINGOLLA. Neural dynamics of form and perception: Boundary completion, illusory figures and neon color spreading. *Psychological Review*, (92):173–211, 1985.

[25] S. GROSSBERG E L. PESSOA. Texture segregation, surface representation, and figure-ground separation. *Vision Research*, (38):2657–2684, 1998.

[26] R.M. HARALICK E L.G. SHAPIRO. Image segmentation techniques. *Computer Vision, Graphics, and Image Processing*, (5):100–132, 1985.

[27] J.E. HUMMEL E I. BIEDERMAN. Dynamic binding in a neural network for shape recognition. *Psychological Review*, (99):480–517, 1992.

[28] A.K. JAIN. *Fundamentals of Digital Image Processing.* Prentice Hall, Englewood Cliffs, 1989.

[29] W. JAMES. *The Principles of Psychology Vol. I.* Dover Publications, New York, 1950.

[30] H. JANSEN. *Saccadic Camera Control for Scene Recognition on an Autonomous Vehicle, in: Visual Attention and Cognition.* Elsevier Science B. V., 1996.

[31] B. JULESZ E J.R. BERGEN. Textons, the fundamental elements in preattentive vision and the perception of textures. *Bell System Technical Journal*, (62):1619–1644, 1983.

[32] T. KOHONEN. The self-organizing map. *Proceedings of the IEEE*, (78):1464–1480, 1990.

[33] S. LAWRENCE, C.L. GILES, A.C. TSOI, E A. BACK. Face recognition: A convolutional neural network approach. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks and Pattern Recognition*, (8):98–113, 1997.

[34] A. LEITÃO E L. PESSOA. Pyramid Representation for Face Recognition . Em *Proceedings of the III Workshop on Cybernetic Vision, Campinas, Brasil*, fevereiro 1999.

[35] Z. LIU, D.C. KNILL, E D. KERSTEIN. Object classification for human and ideal observers. *Vision Research*, (35):549–168, 1995.

[36] D. MARR. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H.Freeman, San Francisco, 1982.

[37] B. MEL. SEEMORE: Combining color, shape,and texture histogramming in a neural-inspired approach to visual object recognition. *Neural Computation*, (9):777–804, 1997.

[38] H. MURASE E S.K. NAYAR. Visual learning and recognition of 3d objects from appearence. *International Journal of Computer Vision*, (14):5–24, 1995.

[39] D. NOTON E L. STARK. Scanpaths in saccadic eyemovements while viewing and recognizing patterns. *Vision Research*, (11):929–942, 1971.

[40] B.A. OLSHAUSEN, C.H. ANDERSON, E D.C. VANESSEN. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, (13(11)):4700–4719, 1993.

[41] T. PAVLIDIS. *Structural Pattern Recognition.* Springer-Verlag, New York, 1977.

[42] L. PESSOA, S. EXEL, A. ROQUE, E A. LEITÃO. Active Scene Exploration Vision System. Em *Proceedings of the International Conference on Neural Network and Brain*, 1998.

[43] L. PESSOA, S. EXEL, A. ROQUE, E A. LEITÃO. Atentive Vision Recognition for Scene Exploration . Em *Proceedings of the International Conference on Neural Information Processing, ICONIP'98, Japão*, 1998.

[44] L. PESSOA E A. P. LEITÃO. Complex cell prototype representation for face recognition. *IEEE Transactions on Neural Networks (submetido)*, 1999.

[45] L. PESSOA, E. MINGOLLA, E H. NEUMANN. A contrast-and luminance-driven multiscale network model of brightness perception. *Vision Research*, (35):2201–2223, 1995.

[46] M.I. POSNER E Y. COHEN. *Components of visual orienting. In: Attention and Performance, Vol. 10.* Bouma, H. Bouwhuis, D., Erlbaum, Hillsdale, NJ, 1984.

[47] M.I. POSNER, C.R.R. SNYDER, E J. DAVIDSON. Attention and the detection of signals. *Journal of Experimental Psychology: General*, (109):160–174, 1980.

[48] Z. PYLYSHYN. What the mind's eye tells to the mind's brain: a critique of mental imagery. *Psychological Bulletin*, (80):1–24, 1973.

[49] R. RAO E D. BALLARD. An active vision architecture based on iconic representations. *Artificial Intelligence*, (78):461–505, 1995.

[50] D. REISFELD, H. WOLFSON, E Y. YESHURUN. Context free attentional operators: The generalized symmetry transform. *Journal of Computer Vision*, (14):119–130, 1995.

[51] G. RIZZOLATTI, I. DASCOLA, E C. UMILTÁ. Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, (1a):31–40, 1987.

[52] G. RIZZOLATTI, L. RIGGIO, SHELIGA B.M, , E C. UMILTÁ. *Space and Seletive Attention, in: Attention and Performance XV.* Moscovitch, M. (eds.), Erlbaum, Hillsdale, NJ., 1994.

[53] A. ROSENFELD E A.C. KAK. *Digital Picture Processing.* Academic Press, New York, 1982.

[54] I.A. RYBAK, V.I. GUSAKOVA, L.N. PODLADCHIKOVA, E N.A. SHEVTSOVA. A model of attention-guided visual perception and recognition. *Vision Research, Special Issue: Models of Recognition*, 1998.

[55] G. SANDINI E M. TISTARELLI. Vision and space-variant sensing. *In: H. Wechsler (ed), Neural Networks for Perception: Human and Machine Perception*, 1991.

[56] E. SCHWARTZ. Spatial mapping in the primate sensory projection: analytic structure and relevance to perception. *Biological Cybernetics*, (25):181–194, 1977.

[57] E. SCHWARTZ. Computational anatomy and functional architecture of striate cortex: a spatial-mapping approach to perceptual coding. *Vision Research*, (20):645–670, 1980.

[58] M. TISTARELLI. Active/space-variant object recognition. *Image and Vision Computing*, (13):215–226, 1995.

[59] A. TREISMAN. Preattentive processing in vision. *Computer Graphics and Image Processing*, (31):156–177, 1985.

[60] A. TREISMAN E G. GELADE. A feature integration theory of attention. *Cognitive Psychology*, (12):97–113, 1980.

[61] J.K. TSOTSOS, S.M. CULHANE, E K.Y.W. WAI. Modeling visual attention via selective tuning. *Artificial Intteligence*, (78):507–545, 1995.

[62] M. TURK E A. PENTLAND. Face recognition using eigenfaces. *Journal of Cognitive Neuroscience*, (3):71–86, 1991.

[63] S. ULLMAN. Aligning pictorial descriptions: An approach to object recognition. *Cognition*, (32):193–245, 1989.

[64] L.G. UNGERLEIDER E M. MISHKIN. *Two cortical visual systems. In: Analysis of visual behavior.* Ingle, D.j. and Goodale, M.A. and Mansfield, R.J.W. (Eds.), MIT Press, Cambridge, MA, 1982.

[65] L. WISCOTT, J.M. FELLOUS, N. KRUGER, E C. MALSBURG. Face recognition by elastic bunch graph matching. *Proceedins of the 7th Int. Conference on Computer Analysis of Images and Patterns, Kiel, Germany*, 1997.

[66] A. YARBUS. *Eye Movements and Vision.* Plenum Press, New York, 1967.

[67] Y. YESHURUN. *Attention Mechanisms in Computer Vision. In: Artificial Vision: Image Description, Recognition and Communications.* V. Cantoni and S. Levialdi, 1995.

[68] Y. YESHURUN E E. SCHWARTZ. Shape description with a space-variant sensor: Algorithms for scanpath, fusion, and convergence over multiple scans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (11):1217–1222, 1989.

[69] S.W. ZUCKER. Region growing: Childhood and adolescence. *Computer Graphics and Image Processing*, (5):382–399, 1976.