Universidade Federal de Pernambuco
Centro de Informática

Graduação em Engenharia da Computação

Reconhecimento de Objetos por Regiões

Thyago Neves Porpino

Trabalho de Graduação

Recife
29 de dezembro de 2011

Universidade Federal de Pernambuco
Centro de Informática

Thyago Neves Porpino

Reconhecimento de Objetos por Regiões

Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.

Orientador: *Prof. Dr. Tsang Ing Ren*

Recife
29 de dezembro de 2011

Eu dedico este trabalho à José Roberto Lins Porpino, meu pai.

Agradecimentos

Agradeço à Deus, e à todas as pessoas que contribuiram para minha formação acadêmica e profissional.

If you seek mastery of the sword, seek first sincerity of the heart, for the former is but a relection of the latter.

—IWAKURA YOSHINORI

Resumo

O reconhecimento de objetos é um dos problemas em aberto na área de visão computacional. Através da combinação de abordagens bottom-up (características como brilho, cor e textura) e top-down (informações sobre categorias de objetos e exemplares) os pesquisadores estão buscando aumentar o índice de reconhecimento destes sistemas.

Entre as abordagens mais promissoras em reconhecimento de objetos, está o uso de regiões como unidades de segmentação. O uso de regiões é vantajoso, pois as mesmas codificam informações sobre forma e dimensão de um objeto, além de representarem os domínios para o cálculo das características da imagem, sem serem afetadas pela desordem fora da região.

Neste trabalho, é dada uma introdução ao tema do reconhecimento de objetos usando regiões. As técnicas usadas em [1] para criação de um framework de reconhecimento de objetos do estado da arte são investigadas. E uma implementação desse framework foi desenvolvida, e mostra-se aqui os resultados observados e possíveis melhoras futuras.

Palavras-chave: Reconhecimento de objetos, regiões, segmentação, processamento de imagem.

Abstract

Object recognition is one of the open problems in computer vision. Through the combination of bottom-up (brightness, color and texture cues) and top-down (object categories and exemplars) approaches, researchers are trying to improve the recognition rate of such systems

Between the more promissing approaches in object recognition, the use of regions as segmentation units deserve special mention. Their use is an advantage beccause they can codify form and scale information on a natural way, and they represent the domains where the feature extraction will take place, without being affected by cluster outside the region.

In this document, a brief introduction to object recognition using regions is given, and the techniques used in [1] for creating a state of the art object recognition framework are examined. An implementation this framework was developed and its results and future improvements will be discussed.

Keywords: Object Recognition, regions, segmentation, image processing.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Neste capítulo será dada uma introdução à área de Reconhecimento de Objetos. Para melhor entendimento deste assunto, será apresentada um pouco da motivação por trás da pesquisa na área de Visão Computacional, e parte da sua história. Após isso, serão apresentados o objetivo do presente trabalho e uma descrição de como ele está estruturado.

1.1 Motivação e Contextualização

Todos os dias, os seres humanos reconhecem uma multidão de objetos familiares e novos. Eles fazem isso sem esforço, embora esses objetos possam variar um pouco em sua forma, cor, textura, etc. Os objetos são reconhecidos de vários pontos de observação diferentes (pela frente, lado, ou por trás), em vários locais diferentes, e com vários tamanhos diferentes. Objetos podem até ser reconhecidos quando eles estão parcialmente obstruídos da visão.

Embora possa ser óbvio que as pessoas são capazes de reconhecer objetos em uma enorme variedade de condições, isto não é verdade para as máquinas. Mesmo depois de décadas de pesquisa, as máquinas ainda estão longe de possuir uma visão tão sofisticada quanto a visão humana.

O objetivo da área de visão computacional é desenvolver algoritmos e representações que irão permitir que uma máquina analise informações visuais de forma autônoma. Como tal, o reconhecimento de objetos é um problema fundamental de visão, e busca responder a simples pergunta: O que está nessa imagem, e onde? Reconhecimento continua sendo um problema desafiador, em parte devido às grandes variações exibidas por imagens do mundo real. Oclusões parciais, mudanças no ponto de vista, iluminação variável, planos de fundo desordenados, etc. Todos esses problemas tornam necessário o desenvolvimento de modelos cada vez mais robustos de reconhecimento.

As aplicações que estão surgindo na área de visão, e principalmente na área de reconhecimento de objetos, são muitas. Desde reconhecimento de faces em redes sociais à reconhecimento automático de tipo sanguíneo através de processamento de imagem, as aplicações são inúmeras e só tendem a aumentar no futuro.

Nesse contexto, o uso de regiões no reconhecimento de objetos, pode ser visto de duas formas. Primeiramente, o seu uso é uma maneira de analisar o contexto (características globais) da imagem ao invés de somente características locais, e segundo, regiões parecem ser uma representação mais fiel de como o ser humano analisa uma imagem.

1.2 Histórico

A pesquisa em visão computacional pode ser traçada desde o final dos anos 50, quando trabalhos iniciais foram feitos em domínios simples. O mundo era modelado como composto por blocos definidos pelas coordenadas de seus vértices e a especificação da informação de borda. O bloco representava áreas de brilho uniforme na imagem e as bordas dos blocos estavam localizadas em áreas de intensa descontinuidade. Essa pesquisa foi realizada por Larry Roberts, o pai da visão computacional, que desenvolveu um sistema sequencial onde a primeira etapa envolvia a segmentação de uma imagem, e que produzia um conjunto de linhas formando um desenho. Essa informação era então passada para a etapa de interpretação, que buscava analisar o desenho em termos de modelos de câmera e protótipos 3-D.

Baseando-se no sistema proposto por Roberts, pesquisadores perceberam que a desvantagem neste tipo de sistema sequencial era que a qualidade dos resultados na etapa de interpretação era fortemente dependente da habilidade de segmentar a imagem original. Efeitos naturais como sombras e reflexão especular, que são tratados com facilidade pelo sistema visual humano, representavam grandes problemas para esses sistemas de visão iniciais. Logo, em resposta à isto, o final dos anos 60 e na década de 70 viram o desenvolvimento de sistemas que integravam as etapas de segmentação e interpretação. Exemplos clássicos como o sistema de Falk, reconheciam objetos por correspôndencias parciais à modelos e então verificava a hipótese, estabelecendo um conjunto mais completo de correspondência de características com o candidato proposto. Esta época também viu a rápida melhora na aquisição de imagens com o range imaging e a melhora na qualidade dos equipamentos de stereo sensing.

Em 1978, Marr propôs uma abordagem mais rigorosa para o desenvolvimento de sistemas de visão computacionais. As teorias, que ficaram conhecidas como o paradigma de Marr, envolviam o desenvolvimento de níveis computacionais, algorítmicos e de implementação para o sistema. O efeito desse paradigma foi encorajar mais pesquisadores à se concentrar no desenvolvimento de soluções para problemas menores e mais bem definidos em visão de baixo nível, como detecção de borda, crescimento e segmentação de regiões, e em processos de alto nível como reconhecimento de formas e raciocínio.

Uma nova direção para visão computacional, que emergiu na metado dos anos 80 foi a active vision. Nela, a percepção visual é tratada como um processo ativo porque o sistema de visão se adapta constantemente ao ambiente em mudança e uma variedade de requisitos de tarefas como exploração e busca por informação. Esse sistema necessita de cooperação ativa entre diferentes módulos do sistema de visão, e/ou da cooperação entre diferentes sensores, que também é chamado de sensor fusion . O controle de sensores é necessário para atingir de forma ativa a atenção seletiva. Teoria da decisão é o framework por trás a integração das informações e o controle dos diferentes sensores. Já foi verificado que os problemas que não são adequados para um observador passivo, podem ser resolvidos por um observador ativo [2].

1.3 Classificação de Sistemas de Visão Computacional

Um sistema de visão é tipicamente composto pelos seguinte módulos:

1. **Captura**: O processo de gera uma imagem visual.
2. **Pré-processamento**: Tratando de técnicas como redução de ruído e acentuação de características.
3. **Segmentação**: O processo de particionar a imagem em regiões de interesse.
4. **Descrição**: Tratando do cálculo de características (tamanho, forma, etc) adequadas para diferenciar um objeto do outro.
5. **Reconhecimento**: O processo que identifica objetos, e.g. caneca, chapéu. etc.

Neste trabalho, foca-se nos módulos de segmentação, descrição e reconhecimento, embora a tarefa de captura seja muito importante. Sistemas de reconhecimento de objetos normalmente consistem apenas das últimas duas etapas: descrição e reconhecimento.

O reconhecimento de objetos têm dominado a atenção de vários pesquisadores em visão computacional, já que ele é um passo necessário para o desenvolvimento de sistemas de visão computacionais completos. Sistemas de reconhecimento de objetos atribuem uma definição de alto nível de um objeto baseando-se nos dados que são representados. Desta interpretação, o sistema pode tomar decisões acerca do ambiente à sua volta.

Pesquisadores de visão computacional desenvolveram dois paradigmas para o desenvolvimento de sistemas de reconhecimento de objetos.

1. **Abordagem Bottom-up**: Essa é uma abordagem muito boa de maneira geral, que se baseia somenta dos dados que são disponibilizados pelos sensores e busca fazer nenhuma suposição a priori. Este esquema é exemplificado pela teoria de Marr e a abordagem caixa-preta. Redes neurais normalmente pertencem à esta abordagem. Nesse caso, a semântica interna da rede não é tão importante, e dados de treinamento são usados para atingir um bom mapeamente de padrões.
2. **Abordagem Top-down**: Essa abordagem assume a presença de um objeto particular ou de uma família de objetos e então utiliza algoritmos para localizar o objeto na cena. Um clássico exemplo dessa abordagem é o reconhecimento de objetos baseados em modelos.

1.4 Reconhecimento de Objetos por Regiões

O reconhecimento de objetos é uma área com intensa pesquisa nos dias atuais, devido às suas inúmeras aplicações (e.g. reconhecer faces em imagens). A estratégia dominante na detecção de objetos em uma cena é o multi-scale scanning, onde uma janela de tamanho e forma fixas varre a imagem, e o seu conteúdo é usado como entrada para um classificador, que informa se um dado objeto C (e.g. face, carro, etc) está presente na janela. Porém, essa abordagem possui algo de profundamente insatisfatória.

Primeiramente, classificar uma janela como contendo um objeto (e.g. garrafa) não é a mesma coisa que extrair os pixels que formam esse mesmo objeto do plano de fundo, e para isso ser feito, é necessário um pós-processamento. Em segundo lugar, a natureza força bruta do processamento na classificação por janelas não é particularmente atraente, e, em último lugar, a classificação por janelamento difere significativamente de como os seres humanos reconhecem um objeto numa imagem, pois estes o fazem através de análise de pistas de contexto e de identificação de regiões mais relevantes. Uma alternativa para o janelamento é o que o movimento Gestalt no ínicio do séc. 20 chamava de "organização de percepções". Segundo essa teoria, a visão humana possui vários níveis de abstração, e os níveis inferior e intermediário são responsáveis pela criação de entidades que serão manipuladas nos níveis mais altos da visão, nos quais o reconhecimento de objetos opera. Essas entidades, podem ser divididas em: pontos, curvas e regiões. Na última década, os pesquisadores deram mais atenção aos pontos como entidades úteis ao reconhecimento de objetos.

Entre as abordagens menos utilizadas, se encontra o reconhecimento de objetos por regiões. Essa abordagem foi pouca pesquisada no passado, pois as técnicas para identificar contornos e regiões não estavam maduras o suficiente. Porém, com a evolução das mesmas, o reconhecimento de objetos por regiões se torna uma alternativa atraente às outras abordagens usadas na área. Entre as suas vantagens estão: (1) regiões codificam forma e dimensão de um objeto naturalmente; (2) elas especificam os domínios em que se devem calcular várias características, sem serem afetadas pela desordem presentes fora da região [1].

1.5 Objetivo

O objetivo deste trabalho é replicar o framework de reconhecimento de objetos desenvolvido em [1], e comparar os resultados da implementação obtida, com os da técnica do artigo original na base ETHZ shape. No final do trabalho será possível avaliar as vantagens da técnica utilizada, assim como seus problemas, e possivelmente propor trabalhos futuros em cima desta técnica.

1.6 Estrutura do Trabalho

O restante deste trabalho está organizado da seguinte forma: no capítulo 2, toda a teoria e explicação do framework de reconhecimento utilizado é feita, entrando em detalhes nas diversas técnicas utilizadas, indo desde a etapa de extração de regiões até a etapa de verificação. No

capítulo 3 temos uma descrição dos experimentos realizados, além dos resultados obtidos. E por fim, no capítulo 4 temos a conclusão do presente trabalho, com um resumo do que foi feito, análise das vantagens e desvantagens da técnica empregada e possíveis trabalhos futuros.

CAPÍTULO 2

Framework

2.1 Obtenção de Regiões

Neste sessão será dada uma dada uma introdução à algumas técnicas de segmentação de imagens, mais especificamente, segmentação baseada em regiões. As técnicas aqui apresentadas são usadas para extrair as regiões de uma imagem.

2.1.1 Visão Geral

Várias tarefas em visão podem se beneficiar da redução de complexidade atingida ao se transformar uma imagem com milhões de pixels em um conjunto com centenas ou milhares de superpixels (segmentos). Nesse contexto, regiões são segmentos que são usados como entidades de alto nível para o processamento de uma imagem.

A segmentação de imagens é um processo fundamental em várias imagens, vídeos e aplicações de visão computacional. Ela decompõe uma imagem em vários componentes constituintes, que idealmente correspondem à objetos do mundo real. Segmentação de imagens baseada em regiões é uma abordagem popular ao problema de segmentação, na qual uma imagem é particionada em regiões conectadas através do agrupamento de pixels vizinhos que possuem características semelhantes. Características de interesse normalmente incluem cor, textura, forma, etc.

Neste sessão, será apresentado um algoritmo que produz uma segmentação hierárquica a partir da saída de qualquer detector de contornos. O primeiro passo desse algoritmo é uma variante da watershed transform, a Oriented Watershed Transform (OWT) [3], que produz um conjunto inicial de regiões a partir da saída de um detector de contornos. Depois disso, um Ultrametric Contour Map (UCM) [4] é construído a partir das fronteiras dessas regiões iniciais.

Essa sequência de operações (owt-ucm) pode ser vista como um método genérico para se obter uma árvore hierárquica de regiões a partir de uma imagem, dado seus contornos. Os contornos codificados na segmentação hierárquica resultante, possuem pesos representados por números reais, indicando a sua probabilidade de ser uma fronteira de verdade. Para um dado threshold, a saída desse framework é um conjunto de contornos fechados, que podem ser tratados como uma segmentação ou como um detector de fronteiras para o propósito de avaliação de performance.

O detector de contornos escolhido para fornecer a entrada do owt-ucm é o gPb (Global Probability of Boundary) [5], pois testes [3] mostraram que ele produz resultados melhores do que outros detectores conhecidos (e.g. Canny). Essa sequência de técnicas usadas para obter uma árvore hierárquica de regiões, será chamada no restante desse trabalho de *gPb-owt-ucm*.

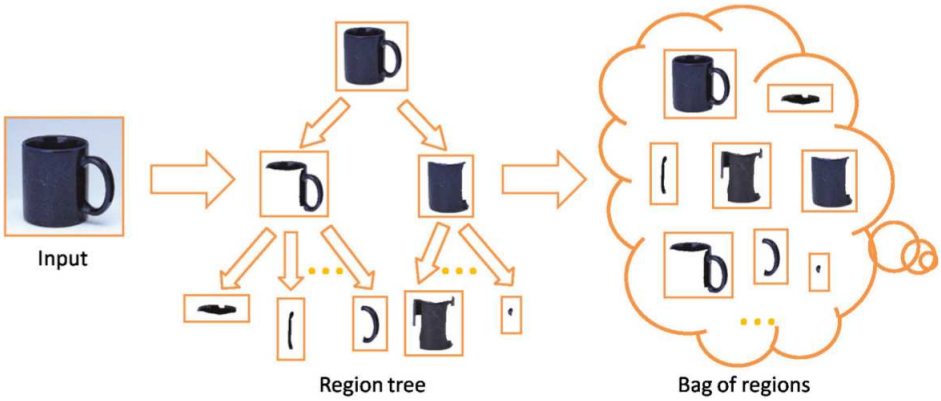

Figura 2.1 Bolsa de regiões, gerada após a executação do gPb-owt-ucm. O gPb-owt-ucm retorna como saída um UCM da imagem, que deve ser processado em um passo posterior para extrair as regiões.

A saída do gPb-owt-ucm é processada para que se destrua a estrutura de árvore, e se crie uma bolsa de regiões, como mostrado na Figura 2.1.

2.1.2 Detecção de Contornos

Neste trabalho, o problema classicamente chamado de detecção de bordas é distinguido da detecção de contornos. Um contorno é uma divisão na imagem que representa uma mudança na posse do pixel, de um objeto (ou superfície) para outro(a). Em contraste, uma borda costuma ser definida como uma mudança abrupta em alguma característica de baixo nível da imagem, como brilho ou cor. Logo, a detecção de bordas é uma técnica de baixo nível que é normalmente aplicada visando atingir o objetivo de detectar contornos.

Figura 2.2 Detecção de Borda com o Canny. À esquerda, a imagem original, e à direita, a saída do detector de borda Canny.

Existe uma extensa literatura em detecção de contornos. Para os propósitos deste trabalho, considera-se duas abordagens principais para esta tarefa. A primeira família de métodos tem por objetivo quantificar a presença de fronteiras em uma dada localização na imagem, usando medidas locais. Abordagens locais antigas, como o detector Canny [6], modela as bordas como discontinuidades abruptas no brilho. Uma descrição mais rica pode ser obtida ao considerar a resposta da imagem em relação à um conjunto de filtros de diferentes escalas e orientações. Um exemplo, é a abordagem Oriented Energy [7] [8], na qual o banco de filtros é composto por duplas de quadratura de filtros simétricos pares e ímpares. Abordagens mais recentes também levam em consideração informações de cor e textura, e fazem uso de técnicas de aprendizado para a agregação de características [9] [10].

À despeito desse progresso, sem algum mecanismo para forçar o fechamento dos contornos, uma segmentação construída a partir de contornos detectados localmente irá com frequência unir erroneamente partes da imagem devido aos furos no contorno delimitador, o que resulta em subsegmentação.

Uma segunda família de métodos, se baseia na combinação de informações globais da imagem em um processo de integração. Teoria de grafos espectrais[4] tem sido comumente usada para esse propósito, particularmente, o critério de Corte Normalizado [11] [12]. Neste framework, dada uma matriz de afinidade W cujas entradas codificam a similaridade entre pixels, define-se $D_{ii} = \sum_j W_{ij}$ e resolve-se o sistema linear para os autovetores:

$$(D - W)v = \lambda Dv \tag{2.1}$$

Tradicionalmente, depois desse passo, o agrupamento é aplicado para se obter a segmentação em regiões. Essa abordagem normalmente quebra regiões uniformes onde os autovetores têm gradientes suaves. Uma solução é reponderar a matriz de afinidade [13]; outros já propuseram formulações alternativas de particionamento de grafos [14] [15] [16]. Embora métodos de detecção de contornos baseados em particionamento espectral terem se prestado bem ao regime high precision / low recall, a sua performance é geralmente pobre no regime high recall / low precision [14] [16].

2.1.2.1 Global Probability of Boundary

O Global Probability of Boundary (gPb) [5], é uma evolução da técnica de detecção de contornos desenvolvida em [10]. A técnica original, cuja saída $Pb_{\sigma}(x, y, \theta)$ prediz a probabilidade de cada pixel da imagem, pertencer à uma fronteira, ao medir a diferença entre vários canais de características, nas duas metades de um disco σ centrado em (x, y) e dividido por um diâmetro com ângulo θ.

O detector gPb combina gradientes multidimensionais de brilho, cor, e textura, com um sinal espectral orientado dessas características. Em particular, a combinação ponderada das características locais é definida como:

$$mPb(x, y, \theta) = \sum_s \alpha_{i,s} G_{i,\sigma(s)}(x, y, \theta) \tag{2.2}$$

onde s indexa as escalas, i indexa os canais de características (brilho, cor, textura), e $G_{i,\sigma(s)}(x, y, \theta)$

mede a diferença entre duas metades do disco de raio $\sigma(s)$ centrado em (x,y) e dividido por um diâmetro no ângulo θ no canal i.

Do *mPb*, define-se uma matriz de afinidade W entre pixels, usando a pista do contorno de intervenção [17]. Escrevendo $D_{ii} = \sum_j W_{ij}$, calcula-se os autovetores $v_1, \ldots, v_n$ do sistema $(D-W)v = \lambda Dv$ correspondendo aos n menores autovalores. Tratando cada autovetor v_k como uma imagem, convolui-se com derivadas Gaussianas direcionais para obter sinais de contorno orientados $sPb_{v_k}(x,y,\theta)$, e o combina em

$$sPb(x,y,\theta) = \sum_{k=1}^{n} \frac{1}{\sqrt{\lambda_k}} \cdot sPb_{v_k}(x,y,\theta) \tag{2.3}$$

O detector gPb final, é um somatório ponderado de sinais locais e espectrais, que é subsequentemente redimensionado usando uma sigmóide:

$$gPb(x,y,\theta) = \sum_s \sum_i \beta_{i,s} G_{i,\sigma(s)}(x,y,\theta) + \gamma \cdot sPb(x,y,\theta) \tag{2.4}$$

Figura 2.3 gPb. À esquerda, a imagem original, e à direita, uma saída visualizável do gPb. Percebe-se que a saída do gPb ainda contêm curvas abertas, o que precisa ser consertado pela Oriented Watershed Transform para que as regiões sejam recuperadas.

2.1.3 Oriented Watershed Transform

Usando o sinal de contorno, primeiramente é construída uma partição mais detalhada para a hierarquia, uma sobresegmentação cujas regiões determinam o maior nível de detalhe considerado. Isto é feito ao se calcular $E(x,y) = \max_\theta E(x,y,\theta)$, a resposta máxima do detector de contorno sobre as orientações. O mínimo regional de $E(x,y)$ é usado como localização base para segmentos homogêneos e aplica-se a Watershed Transform usada em morfologia matemática [18] [19] na superfície topográfica definida por $E(x,y)$. A catchment basin (bacia) do mínimo, chamada de $\mathcal{P}_0$, fornece as regiões da partição mais detalhada e os arcos watershed correspondentes, $\mathcal{K}_0$,as possíveis localizações das fronteiras.

Depois disso, a intensidade das fronteiras, dada por $E(x,y,\theta)$ são transferidas para as localizações $\mathcal{K}_0$. Para esse propósito, os arcos watershed são aproximados para segmentos de linha, e pondera-se cada ponto em $\mathcal{K}_0$ pelo valor de $E(x,y,\theta)$ nesse ponto, na direção θ dada pela orientação do segmento linha correspondente, como detalhado na Figura 2.4. Este procedimento, chamado de Oriented Watershed Transform (OWT), reforça a consistência entre a força das fronteiras de $\mathcal{K}_0$ e do sinal subjacente $E(x,y,\theta)$ e remove artefatos deixados pela watershed transform padrão[1].

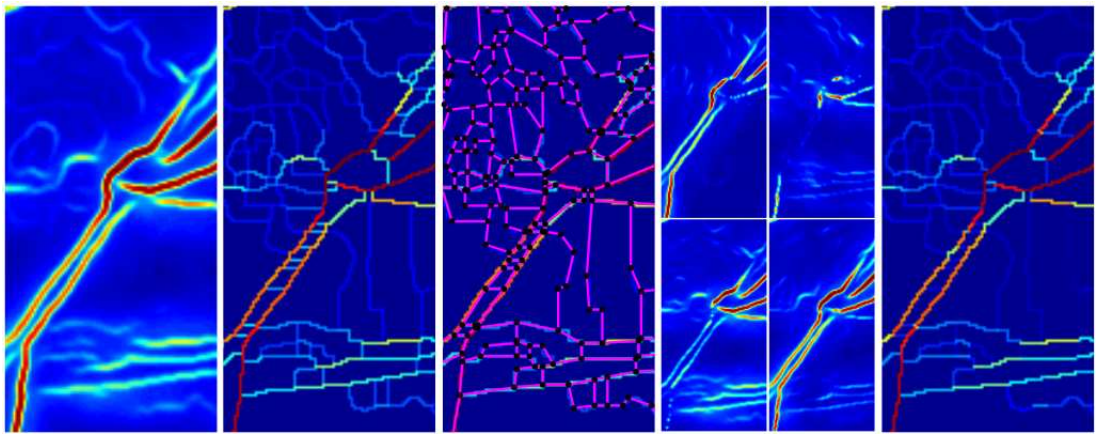

Figura 2.4 Oriented Watershed Transform. (imagem de [3]) Esquerda: Sinal de entrada de fronteira $E(x,y) = \max_\theta E(x,y,\theta)$. **Meio Esquerda:** arcos watershed calculados de $E(x,y)$. Note que regions finas dão origem à artefatos. **Meio:** Arcos watershed com uma subdivisão a partir de segmentos de linha aproximadamente retos. **Meio Direita:** Força orientada de fronteira $E(x,y,\theta)$ para quatro orientações θ. Na prática 8 orientações são usadas. **Direita:** Arcos watershed reponderados de acordo com E na orientação do seu segmento de linha associado.

2.1.4 Ultrametric Contour Maps

Contornos possuem a vantagem de ser relativamente simples representar incerteza na presença de um contorno subjacente de verdade, ou seja, associando-o à uma variável aleatória binária. Porém, não é tão óbvio assim como representar incerteza sobre uma segmentação. Uma das possibilidades, que é usado nesse trabalho, é o Ultrametric Contour Map [4], que define uma dualidade entre contornos fechados, ponderados, que não se intersectam, e uma hierarquia de regiões. Essa mudança de representação de uma segmentação única para uma conjunto aninhado de segmentações se mostra bastante poderosa [3].

A hierarquia é construída através de um algoritmo guloso baseado em grafos que é responsável pela combinação das regiões. Um grafo inicial é definido, onde os nós são as regiões em $\mathcal{P}_0$, as arestas ligam regiões adjacentes e são ponderadas por uma medida de similaridade entre regiões. O algoritmo segue ordenando as arestas por similaridade e iterativamente combinando as regiões mais similares. Este processo produz uma árvore de regiões, onde as folhas são elementos de $\mathcal{P}_0$, a raiz é todo o domínio da imagem, e as regiões são ordenadas por uma relação de inclusão.

A similaridade entre duas regiões adjacentes, é definida como a força média do seu contorno em comum em $\mathcal{K}_0$, inicializado pela OWT. A árvore de regiões construída através deste procedimento possui a estrutura de uma hierarquia indexada e pode ser descrita por um dendrograma, onde a altura de cada segmento é o valor da similaridade na qual ele surgi pela primeira vez e a distância entre duas regiões é a altura entre o menor segmento na hierarquia que as contêm. Além disso, toda a hierarquia pode ser representada por um Ultrametric Contour Map (UCM), uma imagem com valores reais obtida ao se ponderar cada borda entre duas regiões por sua escala de transparência.

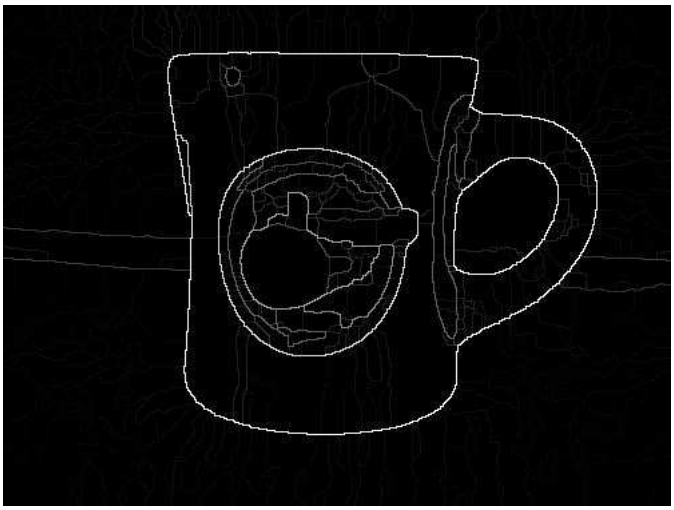

Figura 2.5 UCM da imagem, ou seja, uma representação de uma árvore hierárquica de regiões, na qual pode ser aplicado um threshold *k* para se obter um conjunto de contornos fechados.

A Figura 2.5 demonstra um exemplo desse método. A UCM é uma imagem de contornos ponderados, que por sua própria construção, possui a propriedade de produzir um conjunto de curvas fechadas para qualquer limiar. Da mesma forma, ela é uma representação conviniente da árvore de regiões, já que a segmentação na escala k pode ser facilmente recuperada ao se aplicar o limiar k à UCM. Como a noção de escala usada, é a força média do contorno, os valores da UCM refletem o contraste entre regiões vizinhas.

2.1.5 Extração de Regiões

Ao se ter o UCM da imagem, pode-se escolher em que nível de detalhe quer-se chegar, ao escolher qual threshold será aplicado, como mostrado à esquerda da Figura 2.6. A abordagem seguida nesse trabalho é varrer todos os possíveis thresholds, escolhendo para isso um passo apropriado (para acelerar o processamento), obtendo todas as segmentações que disto resultarem. Para cada segmentação, deve-se extrair todas as regiões como mostrado à direita da Figura 2.6 e excluir as regiões que se repetem. Cada região é armazenada como uma máscara (Figura 2.7), e para obter a bolsa de regiões de uma imagem, basta combinar todas as regiões que restaram de todas as segmentações analisadas.

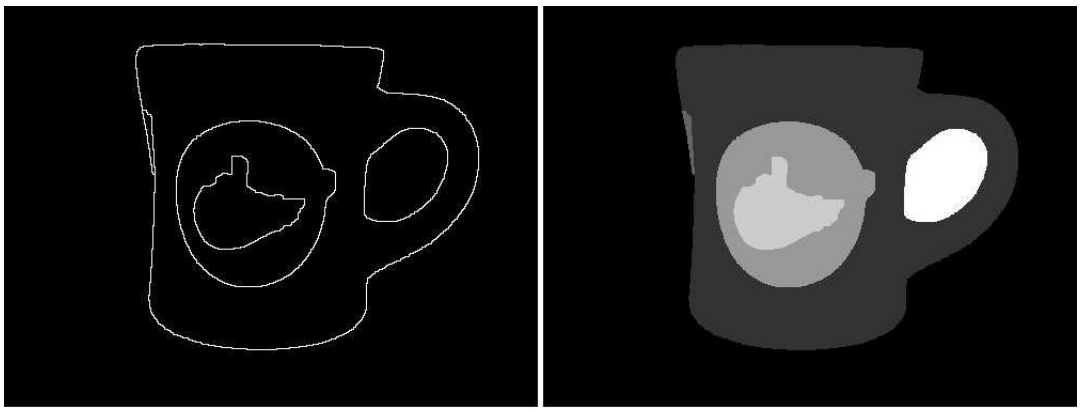

Figura 2.6 Thresholded UCM. Ao se aplicar um limiar k à UCM, obtêm-se um conjunto de curvas fechadas, que por sua vez podem ser usadas para obter regiões no nível k.

Figura 2.7 Máscara que representa uma região. À esquerda, uma máscara que é usada para armazenar uma das regiões obtidas da imagem, e à direita, a aplicação da máscara na imagem original, obtendo a região verdadeira da imagem.

2.2 Extração de Características

Em reconhecimento de objetos, uma miríade de características são usadas para ajudar no processo de reconhecimento. Neste trabalho, cada região é descrita subdividindo uniformemente sua bounding box em uma grade n por n. No artigo analisado foi usado n = 4. Cada célula codifica informação apenas dentro da região. Diferentes características das regiões são capturadas dentro das células, e cada tipo de característica é codificada concatenando-se os sinais das células em um histograma. Neste trabalho, os seguintes atributos foram analisados:

- Formato de contorno, dado pelo histograma de respostas orientadas do detector de contorno gPb [22]
- Formato de borda, onde a orientação é dada pelo gradiente local da imagem (calculado através da convolução com um filtro [-1 0 1] sobre os eixos x e y). Este processo captura informação de alta frequência (e.g. textura), enquanto o gPb é projetado para suprimí-la.

- Cor, representada pelo L*, histogramas *a* e *b* no espaço de cor CIELAB.
- Textura, descrita por histogramas texton.

Essa representação de regiões possui várias propriedades interessantes. Primeiramente, a natureza da invariância de dimensão dos descritores de regiões, permite que regiões sejam comparadas independentemente de seus tamanhos relativos. Em segundo lugar, "poluição"de plano de fundo afeta os descritores de regiões de forma amena, se comparado com os descritores de ponto de interesse. E em terceiro, e último lugar, esses descritores de regiões herdam idéias de representações de imagem populares e recentes, como o GIST [20], HOG [21] e SIFT [22].

2.3 Aprendizado Discriminativo de Pesos

Nem todas as regiões possuem igual significância para discriminar um objeto de outro. Por exemplo, regiões circulares são mais importantes do que partes uniformes para distinguir uma bicicleta de uma caneca. Neste trabalho, o framework desenvolvido em [23] é adaptado para o aprendizado dos pesos das regiões. Dado um exemplar $\mathcal{L}$ contendo uma instância de objeto e uma consulta $\mathcal{I}$, denota-se $f_i^{\mathcal{L}}$, $i = 1,2,\ldots,M$ e $f_j^{\mathcal{I}}$, $j = 1,2,\ldots,N$ características de suas bolsas de regiões.

A distância de $\mathcal{L}$ para $\mathcal{I}$ é definida como:

$$\mathcal{D}(\mathcal{L} \to \mathcal{I}) = \sum_{i=1}^{M} w_i^{\mathcal{L}} d_i^{\mathcal{L}\mathcal{I}} = \langle w^{\mathcal{L}}, d^{\mathcal{L}\mathcal{I}} \rangle, \tag{2.5}$$

onde $w_i^{\mathcal{L}}$ é o peso da característica $f_i^{\mathcal{L}}$, e

$$d_i^{\mathcal{L}\mathcal{I}} = \min d(f_i^{\mathcal{L}}, f_j^{\mathcal{I}}) \tag{2.6}$$

é a distância elementar entre $f_i^{\mathcal{L}}$ e o atributo mais próximo em $\mathcal{I}$. Note que a distância entre o exemplar e a consulta é assimétrica, i.e. $\mathcal{D}(\mathcal{L} \to \mathcal{I}) \neq \mathcal{D}(\mathcal{I} \to \mathcal{L})$.

Na etapa de aprendizado dos pesos, supondo que $\mathcal{L}$ é um objeto da categoria $\mathcal{C}$, acha-se um par de $\mathcal{I}$ e $\mathcal{K}$ para que $\mathcal{I}$ seja um objeto da mesma categoria $\mathcal{C}$ e $\mathcal{K}$ seja um objeto de uma categoria diferente. O algoritmo de aprendizado reforça a seguinte condição:

$$\mathcal{D}(\mathcal{L} \to \mathcal{K}) > \mathcal{D}(\mathcal{L} \to \mathcal{I}) \tag{2.7}$$

$$\Longrightarrow \langle w^{\mathcal{L}}, d^{\mathcal{L}\mathcal{K}} \rangle > \langle w^{\mathcal{L}}, d^{\mathcal{L}\mathcal{I}} \rangle \tag{2.8}$$

$$\Longrightarrow \langle w^{\mathcal{L}}, x^{\mathcal{L}\mathcal{I}\mathcal{K}} \rangle > 0, \tag{2.9}$$

onde $x^{\mathcal{L}\mathcal{I}\mathcal{K}} = d^{\mathcal{L}\mathcal{K}} - x^{\mathcal{L}\mathcal{I}}$. Supondo T destes pares sejam construídos para $\mathcal{L}$ do conjunto de treinamento, logo $x_1, x_2, \ldots, x_T$ (superscripts omitidos para clareza). A otimização é formulada da seguinte forma:

$$\min_{w,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{T} \xi_i \tag{2.10}$$

$$s.t. : w^T x_i \geq 1 - \xi_i, \xi_i \geq 0, \forall i = 1, 2, \ldots, T \tag{2.11}$$

$$w \succeq 0. \tag{2.12}$$

Quando se integra múltiplas características para uma única região, aprende-se um peso para cada característica.

Assim como em [23], a probabilidade da consulta $\mathcal{I}$ ser da mesma categoria do exemplar $\mathcal{L}$ foi modelada como uma função logística:

$$p(\mathcal{L}, \mathcal{I}) = \frac{1}{1 + exp[-\alpha_{\mathcal{L}} \mathcal{D}(\mathcal{L} \rightarrow \mathcal{I}) - \beta_{\mathcal{L}}]} \tag{2.13}$$

onde $\alpha_{\mathcal{L}}$ e $\beta_{\mathcal{L}}$ são parâmetros aprendidos no treinamento.

2.4 Algoritmos de Detecção e Segmentação

O framework unificado de reconhecimento de objetos contêm três componentes: votação, verificação e segmentação. Para uma dada imagem de consulta, a etapa de votação gera uma hipótese inicial sobre as posições, escalas e suporte dos objetos baseando-se em correspondência de regiões. Essas hippóteses são então refinadas através de um classificador de verificação e um segmentador de restrição, respectivamente, para obter a detecção final e os resultados de segmentação. A Figura 2.8 mostra o pipeline do algoritmo de reconhecimento utilizado, usando um logo de maçã (retirado de [1]). A imagem de consulta é comparada com cada examplar do logo da maçã no conjunto de treinamento, cujas bounding boxes e máscaras de suporte são fornecidas como entrada. Todos os pesos das regiões são determinadas como na Sessão 2.3

2.4.1 Votação

O objetivo aqui, é dado uma imagem de consulta e uma categoria de objetos, gerar uma hipótese sobre as bounding boxes e suportes (parciais) dos objetos dessa categoria na imagem. Para atingir isto, um esquema de votação de Hough é usado, baseando-se na transformação entre regiões correspondentes, assim como nos objetos associados nos exemplares.

Especificamente, dado o exemplar $\mathcal{L}$, sua ground truth bounding box $B^{\mathcal{L}}$ e máscara de suporte $M^{\mathcal{L}}$, a região $R^{\mathcal{L}}$ em $\mathcal{L}$ é comparada com outra região $R^{\mathcal{I}}$ na consulta $\mathcal{I}$. E então, o voto para a bounding box $\hat{\mathrm{B}}$ do objeto em $\mathcal{I}$ é caracterizada por:

$$\theta_{\hat{B}} = \tau(\theta_{B^{\mathcal{L}}} | \theta_{R^{\mathcal{L}}}, \theta_{R^{\mathcal{I}}}) \tag{2.14}$$

onde $\theta = [x, y, s_x, s_y]$ caracteriza as coordenadas centrais $[x, y]$ de uma região ou bounding box, e τ é uma função transformação pré-definida cujos parâmetros são derivados das regiões correspondentes $\theta_{R^{\mathcal{L}}}$ e $\theta_{R^{\mathcal{I}}}$.

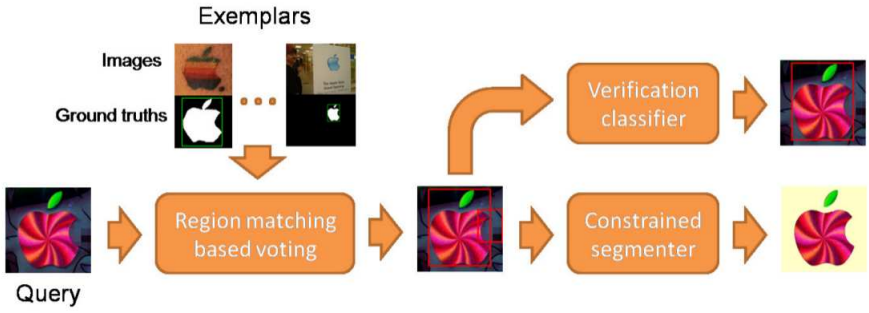

Figura 2.8 Pipeline de reconhecimento. Esse algoritmo de reconhecimento de objetos consiste de três etapas. Dada uma imagem de consulta, a etapa de votação gera hipóteses iniciais sobre as posições dos objetos, suas escalas e sua forma, baseadass em regiões correspondentes dos exemplares. Estas hipóteses são usadas como entrada para os próximas etapas e são refinadas através de um classificador de verificação e um segmentador de restrição, respectivamente, para obter os resultados de detecção e segmentação.

Uma pontuação de votação é atribuída a cada box através da combinação de múltiplos termos:

$$S_{vot}(\hat{B}) = \bar{w}_{R^{\mathcal{L}}} \cdot g(d_{R^{\mathcal{L}}}, d_{R^{\mathcal{I}}}) \cdot h(R^{\mathcal{L}}, R^{\mathcal{I}}) \tag{2.15}$$

onde $\bar{w}_{R^{\mathcal{L}}}$ é o peso aprendido de $R^{\mathcal{L}}$ depois da normalização, $g(d_{R^{\mathcal{L}}}, d_{R^{\mathcal{I}}})$ caracteriza a similaridade entre os descritores $d_{R^{\mathcal{L}}}$, $d_{R^{\mathcal{I}}}$, e $h(R^{\mathcal{L}}, R^{\mathcal{I}})$ penaliza diferenças na forma da região entre as imagens.

No geral, τ na eq. 2.14 pode ser qualquer função transformação. Nos experimentos realizados no trabalho referência, o modelo de transformação é restringido para permitir apenas tradução e redimensionamento tanto no eixo-x quanto no eixo-y. Logo, na direção x:

$$x^{\hat{B}} = x^{R^{\mathcal{I}}} + \left(x^{B^{\mathcal{L}}} - x^{R^{\mathcal{I}}}\right) \cdot s_x^{R^{\mathcal{I}}} / s_x^{R^{\mathcal{L}}} \tag{2.16}$$

$$s_x^{\hat{B}} = s_x^{B^{\mathcal{L}}} \cdot s_x^{R^{\mathcal{I}}} / s_x^{R^{\mathcal{L}}} \tag{2.17}$$

e as mesmas equações se aplicam na direção y.

Eqs. 2.15, 2.16 e 2.17 sumarizam a votação de bounding boxes entre um par de regiões correspondentes. Uma rejeção prematura é aplicada na box votada, seja se a sua pontuação de votação for muito baixa ou se a box for (parcialmente) fora da imagem. Para todas as regiões correspondentes entre a consulta $\mathcal{I}$ e todos os exemplares de uma categoria, gera-se um conjunto de bounding boxes para objetos daquela categoria em $\mathcal{I}$ e para cada par de regiões. Finalmente, agrupa-se essas bounding boxes usando o algoritmo mean-shift [6] no espaço de características θ_B. O mean-shift é escolhido ao invés de outros métodos pois ele permite adaptação da configuração de banda para clusters diferentes. Logo, duas bounding boxes grandes estão mais sujeitas à serem combinadas do que duas bounding boxes pequenas se elas diferem na mesma medida no espaço de características.

Uma grande vantagem desse algoritmo de votação baseado em correspondência de regiões é que ele é capaz de recuperar o suporte total de um objeto, se possuir apenas uma fração desse objeto seja correspondente. Ele fornece não só a posição, mas também uma estimativa confiável de dimensão das bounding boxes.

2.4.2 Verificação

Um classificador de verificação é aplicado à cada hipótese de bounding box da votação. No geral, qualquer modelo de objeto, e.g. [24] e [25], podem ser aplicados em cada hipótese. Porém, para se aproveitar totalmente do uso da representação por regiões, o método de [23] é seguido, usando os pesos das regiões derivadas da Seção 2.3.

A pontuação de verificação de um bounding box $\hat{B}$ no que diz respeito à categoria $\mathcal{C}$ é definida como a média das probabilidades de $\hat{B}$ para todos os exemplares da categoria $\mathcal{C}$:

$$S_{ver}(\hat{B}) = \frac{1}{N}\sum_{i=1}^{N} p\left(\mathcal{L}_{\mathcal{C}_i}, \hat{B}\right) \tag{2.18}$$

onde $\mathcal{L}_{\mathcal{C}_1}$, $\mathcal{L}_{\mathcal{C}_2}$, ..., $\mathcal{L}_{\mathcal{C}_N}$ são exemplares da categoria $\mathcal{C}$, e $p\left(\mathcal{L}_{\mathcal{C}_i}, \hat{B}\right)$ são computadas usando a eq. 2.13. A pontuação geral de detecção $S_{det}(\hat{B})$ de $\hat{B}$ para a categoria $\mathcal{C}$ é uma combinação das pontuações de votação $S_{vot}(\hat{B})$ e da pontuação de verificação $S_{ver}(\hat{B})$, por exemplo, o produto das duas:

$$S_{det}(\hat{B}) = S_{vot}(\hat{B}) \cdot S_{ver}(\hat{B}) \tag{2.19}$$

2.4.3 Segmentação

Considera-se aqui, como segmentação, a tarefa de extrair precisamente o objeto da imagem. Esta tarefa foi tratada no passado, por técnicas como OBJCUT [26]. No framework aqui apresentado, a árvore de regiões é o resultado de processamento bottom-up; conhecimento top-down, derivado do examplar correspondente é usado para marcar algumas das folhas da árvore de regiões como definitavemente pertencentes ao objeto, e algumas outras como pertencentes ao plano de fundo da imagem. Esses labels são propagados para o resto das folhas usando o método de [23], conseguindo assim, o benefício tanto do processamento de baixo pra cima quanto de cima pra baixo.

Mais precisamente, deixe $\mathcal{L}$, $M^{\mathcal{L}}$ e $B^{\mathcal{L}}$ ser o exemplar, e sua ground truth support mask e bounding box, respectivamente. Logo, para a região $R^{\mathcal{L}}$ em $\mathcal{L}$ e uma das suas regiões correspondentes $R^{\mathcal{I}}$ na imagem de consulta $\mathcal{I}$, computa-se $\mathcal{T}(M^{\mathcal{L}})$, a transformação da ground truth mask $M^{\mathcal{L}}$ em $\mathcal{I} \cdot \mathcal{T}(M^{\mathcal{L}})$ provê uma suposição top-down inicial para a localização, dimensão e forma do objeto em $\mathcal{I}$. O seu complemento fornece a hipótese top-down para o plano de fundo. Já que, é indesejável que a segmentação seja feita, somente a partir dessas hipóteses top-down, é utilizada uma zona de pixels "don't know"em uma vizinhança fixa da fronteira da máscara transformada do exemplar, e considera-se como pré-requisitos para o objeto e plano de fundo apenas pixels com uma distância maior que uma dada distância Euclidiana da fronteira da ground truth mask projetada $\mathcal{T}(M^{\mathcal{L}})$. Já que tem-se a restrição que toda a região correspon-

dente $R_{\mathcal{I}}$ deve ser uma parte do objeto, une-se isso com a máscara do objeto para produzir uma "constrained mask".

Assim, um segmento $\mathcal{M}$ é construído na consulta ao usar tanto uma exemplar mask e as informações de baixo nível da imagem de consulta, como ilustrado na Figura 2.4.3. Como um teste inicial de rejeição, computa-se o overlap entre $\mathcal{M}$ e a máscara transformada $\mathcal{T}(M^{\mathcal{L}})$, e discarta-a se a pontuação for muito baixa.

Também é atribuído uma pontuação $S_{seg}(\mathcal{M})$ para $\mathcal{M}$ baseado nas regiões correspondentes $R^{\mathcal{L}}$ e $R^{\mathcal{I}}$:

$$S_{seg}(\mathcal{M}) = \bar{w}_{R^{\mathcal{L}}} \cdot g(d_{R^{\mathcal{L}}}, d_{R^{\mathcal{I}}}) \tag{2.20}$$

onde $\bar{w}_{R^{\mathcal{L}}}$ e $g(d_{R^{\mathcal{L}}}, d_{R^{\mathcal{I}}})$ são definidos na Seção 2.4.1. Logo, defini-se o confidence map de $\mathcal{I}$ para $\mathcal{L}$ baseado em $R^{\mathcal{L}}$ como a resposta máxima de cada região em $\mathcal{I}$.

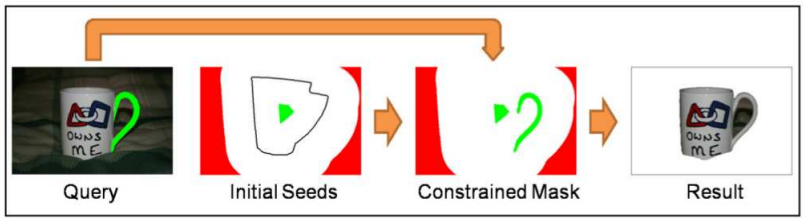

Figura 2.9 Etapa de segmentação. As hipóteses iniciais (verde para objeto e vermelho para plano de fundo) são derivadas da transformação da máscara do exemplar. A máscara restringida é a combinação das hipóteses e a parte correspondente (alça da caneca nesse caso).

CAPÍTULO 3

Experimentos

3.1 Descrição da Base

Para avaliar o framework proposto, foi utilizado um subconjunto da base ETHZ shape, base de dados criada para o teste de algoritmos de detecção de classes de objetos.

Esta base contêm imagens de cinco classes (applelogos, bottles, giraffes, mugs e swans) com formas variadas coletadas do Flickr e Google Images. Os maiores desafios que ela apresenta são a desordem, variância de forma intra-classe e mudanças de dimensão. Além disso, a maioria dos objetos não apresenta oclusão e são apresentados do mesmo ângulo (lateral).

Classe	Número de Imagens	Número de Instâncias
applelogos	40	44
bottles	48	55
giraffes	87	91
mugs	48	66
swans	32	33
Total	255	289

Tabela 3.1 Distribuição de imagens na Base ETHZ Shape original

Devido ao tempo de execução e consumo excessivo de memória de algumas técnicas usadas, teve-se que selecionar manualmente apenas algumas imagens da base de dados (àquelas de tamanho menor, que consumiriam menos memória em seu processamento).

Classe	Número de Imagens	Número de Instâncias
applelogos	22	22
bottles	26	26
giraffes	47	47
mugs	21	21
swans	14	14
Total	130	130

Tabela 3.2 Distribuição no subconjunto da base ETHZ Shape usado nesse trabalho.

3.2 Metodologia

Inicialmente, contruiu-se árvores de regiões (UCM) de cada imagem da base, gerando-se na média aproximadamente 80 regiões por imagem. Já que cor e textura não são muito úteis nessa base de dados, usou-se apenas a forma de contorno baseada no gPb como descritor das regiões. O conjunto de dados foi dividido, metade para treinamento e metade para teste (em cada categoria).

Na etapa de aprendizado de pesos, contruiu-se imagens exemplares e seus pares de similaridade e dissimilaridade da seguinte forma: as bounding boxes de objetos no conjunto de treinamento são tomadas como exemplares. Para cada exemplar, instâncias similares são as bounding boxes contendo objetos da mesma categoria do exemplar, e instâncias dissimilares são aquelas que contêm objetos de categorias diferentes.

Na etapa de votação, foram escolhidas as seguintes funções para a eq. 2.15:

$$g(d_{R^{\mathcal{L}}}, d_{R^{\mathcal{J}}}) = \max\{0.1 - \sigma \cdot \mathcal{X}^2(d_{R^{\mathcal{L}}}, d_{R^{\mathcal{J}}})\} \tag{3.1}$$

$$h(R^{\mathcal{L}}, R^{\mathcal{J}}) = 1\left[\alpha \leq Asp(R^{\mathcal{L}})/Asp(R^{\mathcal{J}}) \leq 1/\alpha\right] \tag{3.2}$$

onde $\mathcal{X}^2(\cdot)$ especifica a distância chi-square, e $Asp(R)$ é o aspect ratio do bounding box de R. A última equação reforça consistência de aspect ratio entre regiões correspondentes. Neste experimento, $\sigma = 2$ e $\alpha = 0.6$ foram usados.

Para avaliar a performance da técnica, ela foi executada quatro vezes para cada categoria, e calculou-se a média da taxa de reconhecimento obtida para cada uma delas.

3.3 Resultados

Os resultados se mostraram satisfatórios, mesmo que abaixo dos resultados obtidos pelo artigo original. Na tabela 3.3, pode-se ver às taxas de acerto obtidas pelo artigo original, em cada uma das classes da base ETHZ shape. E na tabela 3.4, pode-se ver os resultados obtidos neste trabalho. A discrepância nos resultados pode estar relacionada ao menor conjunto de treinamento usado, além de um possível tuning feito na etapa de extração de regiões.

Classe	Apenas Votação	Apenas Verificação	Combinação
applelogos	87.2	85.4	90.6
bottles	93.0	93.2	94.8
giraffes	79.4	73.6	79.8
mugs	72.6	81.4	83.2
swans	82.2	80.8	86.8
Média	82.9	82.9	87.1

Tabela 3.3 Média da taxa de reconhecimento em cada uma das classes da base ETHZ shape, usando somente a etapa de votação, somente a de verificação, ou usando uma combinação (produto) das duas.

Classe	Apenas Votação	Apenas Verificação	Combinação
applelogos	75.5	72.8	78.3
bottles	85.1	84.6	86.2
giraffes	78.8	74.2	79.4
mugs	70.4	78.7	80.8
swans	77.5	74.1	79.9
Média	77.46	76.88	80.9

Tabela 3.4 Média da taxa de reconhecimento obtida neste trabalho, para comparação com a performance do artigo original.

CAPÍTULO 4

Conclusão

Para melhorar a taxa de reconhecimento a nova abordagem de combinar características locais e globais têm sido bastante popular. A escolha de regiões neste contexto, têm se mostrado uma vantagem em relação à outros tipos de unidades de segmentação, como pontos, pois regiões têm a capacidade de codificar formas de uma maneira natural.

Baseando-se nessa abordagem, este trabalho buscou replicar o framework de reconhecimento de objetos por regiões desenvolvido em [1], que busca a detecção e segmentação de um determinado objeto de uma imagem de consulta.

Para avaliar os resultados desta técnica, alguns experimentos foram realizados em cima de um subconjunto da base ETHZ shape, resultando numa taxa de reconhecimento, abaixo do artigo original, porém ainda satisfatória.

Entre os pontos de destaque da técnica utilizada estão a alta taxa de acerto alcançada e a não existência de muitos parâmetros para tuning dos algoritmos, sendo uma técnica quase que automática. Entre os problemas, está o uso do gPB como detector de contornos, o que torna muito custosa a etapa de extração de regiões, encarecendo o valor dessa técnica para reconhecimento de objetos em tempo real.

Para trabalhos futuros, seria interessante investigar melhor o possível uso de outros detectores de contorno para acelerar o processo de extração de regiões, além de testar outras métricas de similaridade entre regiões, buscando maiores taxas de reconhecimento.

Referências Bibliográficas

[1] Chunhui Gu, Joseph J. Lim, Pablo Arbelaez, and Jitendra Malik. Recognition using regions. In *CVPR* [1], pages 1030–1037.

[2] M. Bennamoun and G. J. Mamic. *Object recognition: fundamentals and case studies*. Springer-Verlag New York, Inc., New York, NY, USA, 2002.

[3] Pablo Arbelaez, Michael Maire, Charless C. Fowlkes, and Jitendra Malik. From contours to regions: An empirical evaluation. In *CVPR* [3], pages 2294–2301.

[4] P. Arbelaez. Boundary extraction in natural images using ultrametric contour maps. In *Proceedings POCV*, 2006.

[5] Michael Maire, Pablo Arbelaez, Charless Fowlkes, and Jitendra Malik. Using contours to detect and localize junctions in natural images. In *CVPR* [5].

[6] J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8:679–698, June 1986.

[7] M. C. Morrone and R. A. Owens. Feature detection from local energy. *Pattern Recogn. Lett.*, 6:303–313, December 1987.

[8] Pietro Perona and Jitendra Malik. Detecting and localizing edges composed of steps, peaks and roofs. In *In Proc. 3rd Intl. Conf. Computer Vision*, pages 52–57, 1991.

[9] Piotr DollÃ¡r. Supervised learning of edges and object boundaries. In *In CVPR*, pages 1964–1971, 2006.

[10] David R. Martin, Charless C. Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26:530–549, May 2004.

[11] Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi. Contour and texture analysis for image segmentation. *Int. J. Comput. Vision*, 43:7–27, June 2001.

[12] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 1997.

[13] Gary L. Miller and David Tolliver. Graph partitioning by spectral rounding: Applications in image segmentation and clustering. In *In CVPR*, pages 1053–1060, 2006.

[14] Charless Fowlkes and Jitendra Malik. How much does globalization help segmentation. Technical report, 2004.

[15] Song Wang, Toshiro Kubota, Jeffrey Mark Siskind, and Jun Wang. Salient closed boundary extraction with ratio contour. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:546–561, April 2005.

[16] Stella X. Yu. Segmentation induced by scale invariance. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 444–451, Washington, DC, USA, 2005. IEEE Computer Society.

[17] Charless Fowlkes, David Martin, and Jitendra Malik. Learning affinity functions for image segmentation: combining patch-based and gradient-based approaches. In *In Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, pages 54–61, 2003.

[18] S. Beucher and F. Meyer. *Mathematical morphology in image processing*. 1992.

[19] Laurent Najman and Michel Schmitt. Geodesic saliency of watershed contours and hierarchical segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(12):1163–1173, 1996.

[20] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[21] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005.

[22] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.

[23] A Frome, Y Singer, and J Malik. Image retrieval and classification using local distance functions. *In Practice*, 19(4):417, 2007.

[24] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[25] Subhransu Maji, Alexander C. Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient.

[26] M. Pawan Kumar, Philip H. S. Torr, and A. Zisserman. Obj cut. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 18–25, Washington, DC, USA, 2005. IEEE Computer Society.

[27] M. Pawan Kumar, Philip H. S. Torr, and A. Zisserman. Obj cut. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 18–25, Washington, DC, USA, 2005. IEEE Computer Society.