3
Fundamentos Teóricos

Neste capítulo são abordados fundamentos teóricos importantes para o entendimento do método de segmentação proposto no capítulo 4, como o conceito de imagens digitais, e alguns outros conceitos de segmentação de imagens. Além disso são apresentadas a teoria básica do modelo de *level sets* e uma breve introdução sobre algoritmos genéticos.

3.1
Imagens Digitais

Uma imagem digital pode ser definida por uma função bi-dimensional, $f(x, y)$, onde x e y são coordenadas espaciais, e a amplitude f representa a intensidade da imagem para as coordenadas x e y. Imagens monocromáticas seguem exatamente este modelo, enquanto imagens coloridas são formadas por uma combinação de imagens 2D. Por exemplo, no formato RGB, existem 3 componentes 2D que representam cada qual uma cor primária diferente (vermelho, azul e verde) que ao serem compostas geram a cor em cada pixel.

Imagens podem ser analógicas, isto é, ser contínuas no que diz respeito aos eixos x e y, e também na amplitude f. Em visão computacional se trabalha normalmente com imagens digitais, e portanto se faz necessária a conversão de um tipo para outro. Como em toda conversão análogo-digital (A/D) existe perda de informação na amostragem e quantização que é intrínseca à imagem digital mas, ao fim do processo, é obtida uma representação matricial da imagem contendo apenas números reais, com os quais se pode realizar diversas operações matemáticas que permitem extrair informação útil das imagens.

Neste trabalho especificamente, foram utilizadas imagens digitalizadas no tomógrafo, de forma que a definição espacial delas é estabelecida pelo técnico que o opera. Também é possível considerar cada exame como uma pilha regular de imagens, com espaçamento conhecido entre elas, de forma que cada exame pode ser representado por uma função $f(x, y, z)$, onde x, y e z são as coordenadas espaciais e f é a intensidade do voxel. Vale lembrar que o voxel tem a forma de um paralelepípedo já que a escala de x e y é diferente da escala em z. Esta pilha de imagens nos dá informação volumétrica como

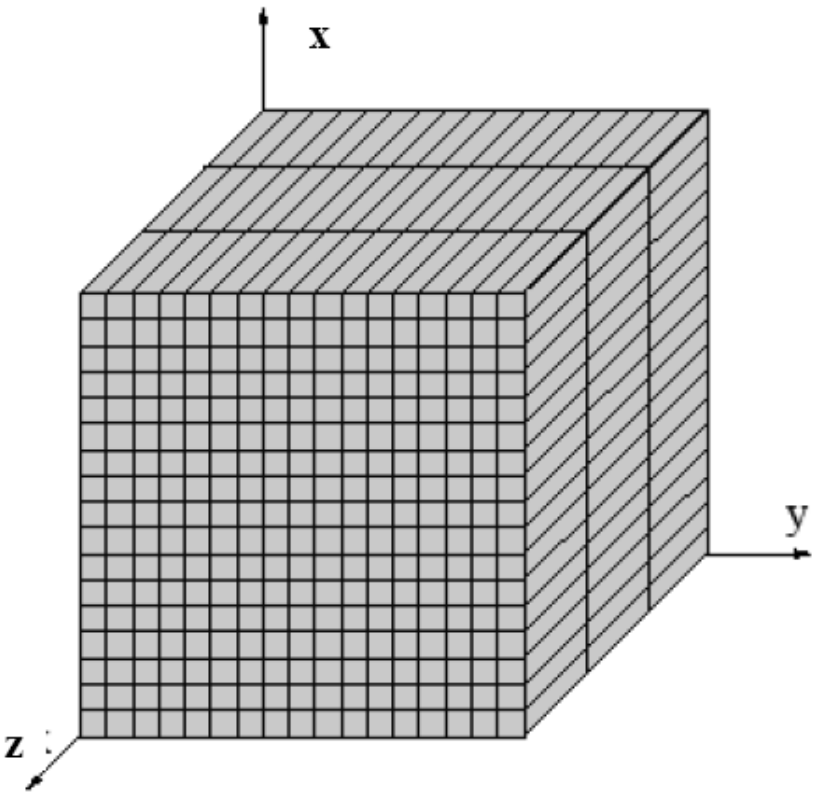

Figura 3.1: Imagem tridimensional.

mostrado na figura 3.1.

3.2 Segmentação de Imagens

A segmentação é um processo que nos permite distinguir em uma imagem, os objetos que a compõem. É geralmente o passo mais delicado em processamento de imagens e por isso um tema intensamente pesquisado ainda atualmente. Nesta dissertação ela constitui o principal objetivo: a segmentação do fígado e de suas principais estruturas internas.

Este é um processo extremamente dependente da aplicação abordada e, portanto, o conhecimento prévio sobre a área de aplicação normalmente se faz necessário. Neste trabalho, a medicina nos fornece o conhecimento das estruturas anatômicas do fígado e suas características, que foi utilizado no desenvolvimento de heurísticas usadas pelo método de segmentação, e que podem ser implementadas através da combinação de diversas técnicas de processamento de imagens.

A visão humana baseia-se em certas características observáveis dos objetos em uma cena, tal como forma e cor, para formar agrupamentos dentro de um contexto estabelecido pela imagem, ou seja, decompor a imagem

em elementos significantes. Métodos de segmentação procuram emular esse processo cognitivo humano para identificar objetos relevantes na imagem e podem ser divididos basicamente em dois diferentes grupos (Masutani06): orientados a dados, e orientados a modelo.

3.2.1 Segmentação Orientada a Dados

Técnicas de segmentação orientadas a dados (Fujimoto02, Kim00) tentam emular a capacidade humana de identificar objetos usando alguma informação de similaridade presente nos dados de imagem, detectando e classificando automaticamente objetos e feições nas imagens. Muitos deles utilizam técnicas conhecidas como crescimento de regiões e limiarizações, combinadas com algum conhecimento a priori sobre o objeto a ser segmentado.

Segmentação por Contorno

Este modelo de segmentação baseia-se no fato de que regiões semanticamente diferentes geralmente formam grupos com diferentes níveis de intensidade que acabam por gerar nas suas fronteiras uma variação abrupta de intensidade. Várias razões fazem com que o estudo dessas regiões de grande variação se tornem interessantes. Geralmente fronteiras de objetos apresentam variação de intensidade, pois de outra forma não seriam visíveis na imagem. Além disso padrões de repetição de fronteiras podem ser utilizados em heurísticas de identificação, indicando diferentes texturas de objetos diferentes.

Quando as bordas são detectáveis em todo o contorno do objeto, esta é uma abordagem suficiente. Entretanto, o fígado, caso de estudo desta dissertação geralmente não possui a borda claramente definida em toda sua extensão, principalmente devido às estruturas anatômicas que fazem fronteira com ele, que tem densidade próxima a sua, e, portanto, níveis de cinza próximos (veja o capítulo 2). Isto faz com que não haja contraste suficiente entre estes diferentes órgãos para indicar com clareza o contorno do fígado utilizando os métodos de detecção de contorno tradicionais.

Segmentação por Região

Regiões são partes da imagem que compartilham características comuns como textura e intensidade semelhantes, sendo formadas por pixels conectados entre si. O processo de segmentação por regiões visa identificar objetos e estruturas em uma imagem digital, através da busca de características que definem as diferentes regiões.

O modelo de segmentação por regiões procura identificar estas regiões seguindo algum critério de homogeneidade. Desta forma agloremados de pixels ou voxels, que contém características em comum e são vizinhos são rotulados como pertencendo a um mesmo dado semântico.

Crescimento de Regiões por Histerese

O método de agrupamento por crescimento de regiões é um procedimento que agrupa pixels ou sub-regiões em regiões maiores. Cada região é investigada e os pixels ou sub-regiões que estão ao seu redor e que formam uma região ainda homogênea, segundo algum critério de homogeneidade, são fundidos à região investigada.

O procedimento pode ser descrito da seguinte forma:

1. Um ponto da imagem pertencente à região que se deseja segmentar é selecionado, recebendo o nome de semente.

2. Agregam-se os pixels vizinhos que são similares a ele segundo algum critério pré-definido (intensidade, cor, textura, etc.).

3. Para cada pixel agregado repete-se o passo 2 até que mais nenhum pixel satisfaça o critério de similaridade.

Uma forma bastante simples de agrupar pixels é calculando intervalos de intensidade que caracterizam de algum modo determinado conceito ou estrutura da imagem. Esta abordagem é conhecida como crescimento de regiões por histerese, e a definição das sementes e do conjunto de vizinhos passíveis de serem agregados às sementes é feita de forma simples. As sementes são definidas por um intervalo de níveis de cinza, mais restrito, que define os pixels que com certeza pertencem a um determinado conceito. Um segundo intervalo, mais abrangente, determina um grupo de pixels maior, que contém pixels que só serão agregados ao conceito, se forem similares às sementes, e estiverem na sua vizinhança.

Desta forma o procedimento de crescimento de regiões por histerese pode ser definido em 5 passos:

1. Segmentam-se as sementes como os grupos de pixels adjacentes, cujas intensidades pertencem ao intervalo de intensidade $[I_L, I_H]$. Estes pixels pertencerão ao objeto que se deseja segmentar.

2. Segmenta-se um conjunto de candidatos ao objeto que se deseja segmentar, com menor grau de certeza, como os pixels pertencentes ao intervalo

de intensidade $[I_L - \Delta_L, I_H + \Delta_H]$. Considera-se que pixels com valor de intensidade fora deste intervalo não pertencem ao objeto a ser segmentado.

3. Inicialmente estima-se o objeto a ser segmentado como sendo igual aos pixels do conjunto de sementes.

4. Adicionam-se ao objeto estimado os pixels do conjunto de candidatos adjacentes em 3 dimensões.

5. Repete-se o passo anterior até que objeto estimado pare de crescer, isto é, mantenha seu volume inalterado entre uma iteração e outra.

O objeto resultante é formado então pela união dos pixels cuja intensidade é característica do objeto que se deseja segmentar (sementes), com os pixels cuja intensidade ainda pode indicar o conceito desejado, e que encontram-se na vizinhança do conjunto de sementes. Este modelo é especialmente atraente para conceitos onde há dispersão ou quando o valor de intensidade dos pixels do conceito varia de acordo com a sua posição do espaço, como no caso de vasos sanguíneos em imagens de TC.

3.2.2 Segmentação Orientada a Modelo

Técnicas de segmentação orientadas a modelo (Lamecker05, Soler01) utilizam modelos pré-definidos para segmentar o objeto desejado nas imagens disponíveis. Neste tipo de técnica é definido um modelo descrevendo, por exemplo, o orgão a ser segmentado, em termos das características do objeto como a posição espacial, textura e relação espacial com outros objetos. Um vez desenhadas estas características, o algoritmo procura nas imagens instâncias que correspondem ao modelo dado.

Existem diversas categorias de métodos orientados a modelo como por exemplo os probabilísticos, os geométricos, e os deformáveis. Esta última vem sendo largamente utilizada em processamento de imagens médicas, e dentre os modelos deformáveis se destaca o *level sets* (Osher03, Sethian99).

3.3 Level Sets

Como dito anteriormente, a segmentação de imagens médicas é um passo importante em várias aplicações, como visualização, análise quantitativa de órgãos e lesões, e auxílio em diagnóstico. Diversos métodos foram propostos para segmentação de órgãos. Métodos de baixo nível, como clusterização,

crescimento de regiões e detecção de bordas, geralmente precisam de pré e/ou pós processamento e ainda assim não garantem o resultado preciso, se órgãos adjacentes compartilham valores de intensidade de pixel.

Neste contexto surgiram os modelos deformáveis, que representam explicitamente o contorno e a forma do objeto segmentado. Eles combinam diversas propriedades desejáveis como suavidade e conectividade, que muitas vezes conseguem reprimir erros de segmentação, e principalmente permitem que se incorpore ao modelo conhecimento sobre o objeto de interesse.

Existem basicamente dois tipos de modelos deformáveis: os paramétricos, e os geométricos (Hamarnerh06).

Os modelos deformáveis paramétricos, dentre os quais há grande destaque para os contornos ativos e os balões, possuem duas grandes limitações. A primeira é conseqüência do fato destes modelos acompanharem a evolução de um número pré-definido de pontos. O que ocorre é que se o contorno inicial difere muito de tamanho e forma do contorno final, então é necessário reparametrizar o modelo dinamicamente, de forma a garantir um número suficiente de pontos que permita que a evolução do modelo transcorra adequadamente e que o contorno final seja próximo do desejado. A segunda limitação é a grande dificuldade existente ao se dividir ou unir pedaços do modelo em evolução. Esta dificuldade é causada pelo fato de uma nova parametrização ser necessária sempre que uma mudança topológica ocorre, o que requer funções sofisticadas para definição correta do contorno resultante. Esta segunda limitação é bastante prejudicial em segmentação de imagens médicas pois freqüentemente se observam alterações topológicas em um órgão visto tomo a tomo.

Os modelos deformáveis geométricos, também chamados de *level sets*, promoveram uma elegante solução para os problemas observados nos modelos paramétricos. Dentre as vantagens observadas na formulação implícita dos modelos geométricos em relação aos paramétricos, pode-se destacar a não necessidade de parametrização do contorno, a flexibilidade topológica, estabilidade numérica, e extensão natural do método de 2D para N dimensões.

Nesta seção serão apresentados conceitos matemáticos de superfícies implícitas relevantes para o entendimento do método de *level sets*, e então o método de *level sets* em maiores detalhes.

3.3.1 Superfícies Implícitas

Quando se deseja segmentar algum objeto em uma cena, uma abordagem comum é encontrar o contorno do objeto em questão (vide seção 3.2.2). Em

imagens médicas este tipo de abordagem pode ser utilizada quando se deseja segmentar um determinado órgão, onde os contornos em cada tomo compõem uma superfície tridimensional que descreve o órgão segmentado.

Tanto a superfície 3D, como cada curva 2D podem ser descritas analiticamente de forma explícita ou implícita. Na forma explícita uma função com o mesmo número de variáveis necessárias para representar o objeto segmentado define explicitamente todos os pontos da curva. Por exemplo, um volume qualquer tridimensional é representado explicitamente por uma função definida em coordenadas x, y, z. Esta tarefa pode ser bastante complexa no caso de imagens médicas, uma vez que o objeto segmentado pode compor curvas bastante complexas, dificultando a tarefa de definir a função que o representa.

Existe também a possibilidade de se definir uma curva de forma implícita, onde a função que descreve o objeto segmentado pode ser embutida em uma outra de maior dimensão, ou seja, com mais uma variável, de modo a ser caracterizada como uma isosuperfície da função final. Na forma implícita a função embutida permanece perfeitamente representada, isto é, todos os seus pontos são definidos.

Este tipo de abordagem pode passar a idéia de desperdício, já que todas as outras isosuperfícies disponíveis a princípio não têm muita utilidade. Entretanto este tipo de representação oferece algumas vantagens importantes. A principal delas é a facilidade de tratamento no caso de divisão ou agrupamento de curvas, como explicado logo adiante. No caso de segmentação em imagens médicas isto tem grande utilidade, uma vez que a perseguição de estruturas anatômicas pelos tomos precisa prever a possibilidade do objeto se dividir, ou de objetos se agruparem em determinado tomo.

A descrição da divisão de uma determinada curva em outras, ou o agrupamento de várias curvas em uma única é bastante complexa. Se feita de forma explícita, geralmente é necessário descrever os objetos como um conjunto de funções independentes. Se for considerado que comumente esta subdivisão pode continuar acontecendo subseqüentemente, a descrição e acompanhamento das curvas tornam-se quase impraticáveis.

A forma implícita oferece uma grande vantagem neste tipo de situação. Como neste caso a curva que está sendo modificada é de dimensão maior do que a que de fato se quer encontrar, a divisão em outras curvas, bem como a união de curvas em uma só, é natural. É possível associar esta idéia ao caso das curvas de nível, onde existe a clara possibilidade de se descrever duas curvas, a princípio independentes, em uma única função, de ordem maior, que as descreve implicitamente.

Há uma outra vantagem importante no uso deste tipo de representação:

a definição da posição relativa de um ponto em relação à curva. Considere uma função $\phi(\overrightarrow{x})$ uma função implícita onde $\overrightarrow{x}$ é um vetor de coordenadas qualquer. Designamos por conveniência a isosuperfície $\phi(\overrightarrow{x}) = 0$ como a superfície de interesse, chamada aqui de interface, e limitamos que esta interface deve ser necessariamente fechada. Deste modo a verificação da posição relativa de um determinado ponto em relação ao contorno pode ser realizada pela simples inferência do sinal que esta posição retorna em $\phi(\overrightarrow{x})$. A mesma operação quando usamos a representação explícita da curva é bastante mais complexa, necessitando geralmente de uma função auxiliar. Uma operação bastante utilizada para este fim traça uma reta entre o ponto $\overrightarrow{x}$ e um ponto muito distante, e verifica se o número de vezes que ela intercepta a curva é ímpar, se sim o ponto encontra-se no interior da curva, caso contrário, encontra-se no exterior.

Funções implícitas também possibilitam grande facilidade em operações de lógica booleana e de construção de novos objetos a partir da aglomeração de outros existentes. Se ϕ_1 e ϕ_2 são duas funções implícitas diferentes, então $\phi(\overrightarrow{x}) = min(\phi_1(\overrightarrow{x}), \phi_2(\overrightarrow{x}))$ é a função implícita que representa a união das regiões internas de ϕ_1 e ϕ_2. Da mesma forma, $\phi(\overrightarrow{x}) = max(\phi_1(\overrightarrow{x}), \phi_2(\overrightarrow{x}))$ é a função implícita que representa a interseção das regiões internas de ϕ_1 e ϕ_2. O complemento de ϕ_1 é definido como $\phi(\overrightarrow{x}) = -\phi_1(\overrightarrow{x})$. Também $\phi(\overrightarrow{x}) = max(\phi_1(\overrightarrow{x}), -\phi_2(\overrightarrow{x}))$ representa a região obtida subtraindo a região interna de ϕ_2 da região interna de ϕ_1.

Quando é necessário discretizar curvas, como no nosso caso, é desejável que a curva a ser discretizada tenha algumas características, para fins de estabilidade. Um subgrupo especialmente interessante de funções implícitas com este perfil é o das funções de distância sinalizadas. Estas funções possuem características como suavidade, módulo do gradiente igual a 1, e sinais opostos na parte interna e na parte externa à interface.

Seja $\partial\Omega$ o contorno de uma função implícita $\phi(\overrightarrow{x})$, Ω^- a região interna ao contorno e Ω^+ a região externa ao contorno. Uma função de distância $d(\overrightarrow{x})$ é assim definida:

$$d(\overrightarrow{x}) = min(|\overrightarrow{x} - \overrightarrow{x}_i|) \text{ para todo } \overrightarrow{x}_i \in \partial\Omega \tag{3-1}$$

Isto implica que $d(\overrightarrow{x}) = 0$ no contorno, onde $\overrightarrow{x} \in \partial\Omega$, e que fora do contorno, onde $\overrightarrow{x} \ni \partial\Omega$, a função retorna a distância enclidiana mínima entre $\overrightarrow{x}$ e $\partial\Omega$.

Um função de distância ϕ é dita sinalizada quando $|\phi(\overrightarrow{x})| = d(\overrightarrow{x})$ para todo $\overrightarrow{x}$. Esta função deve ter $|\phi(\overrightarrow{x})| = d(\overrightarrow{x}) = 0$ para todo $\overrightarrow{x} \in \partial\Omega$, $|\phi(\overrightarrow{x})| = -d(\overrightarrow{x})$ para todo $\overrightarrow{x} \in \Omega^-$, e $|\phi(\overrightarrow{x})| = d(\overrightarrow{x})$ para todo $\overrightarrow{x} \in \Omega^+$. Funções

de distância sinalizada são monotônicas na direção normal ao contorno, e podem ser diferenciadas quando $\overrightarrow{x} \in \partial\Omega$ com uma confiança significativamente alta, no que diz respeito à ausência de problemas relacionados à instabilidade numérica.

3.3.2 Métodos de Level Sets

Os métodos de *level sets* (Osher03, Sethian99) são uma ferramenta poderosa para acompanhamento e definição de superfícies, e adicionam dinâmica a curvas implicitamente representadas. Suas aplicações vão desde modelagem de fenômenos físicos e geração de modelos complexos de computação gráfica, até a segmentação de imagens, discutida neste trabalho.

A idéia principal consiste em embutir um modelo deformável em um espaço de dimensão $d+1$, para segmentar iterativamente um objeto definido num espaço de dimensão d. Para tanto geralmente define-se uma curva inicial, implicitamente representada, que é iterativamente recalculada e atualizada levando-se em conta deslocamentos definidos por equações diferenciais parciais, que a deformam em direção aos contornos do objeto que se deseja segmentar.

Para fins de estabilidade numérica, a função ϕ deve ser suave e bem comportada (Osher03), e pelos motivos descritos na seção 3.3.1 foi utilizada neste trabalho a função de distância sinalizada.

Seja então $\phi(\overrightarrow{x}, t)$ uma função, chamada de função de *level sets*, que representa implicitamente a superfície em evolução, ou interface, $\phi(\overrightarrow{x}, t) = 0$ em qualquer tempo t da evolução. Neste caso a interface pode ser interpretada como a curva de nível zero da função ϕ.

A evolução da interface em *level sets* segue uma dinâmica Euleriana relacionada à mecânica dos fluidos, em contraposição à dinâmica Lagrangiana dos modelos deformáveis paramétricos (Malladi03). A formulação básica de *level sets* define a evolução da interface através de uma equação diferential que caracteriza um deslocamento espacial da função ϕ dado um deslocamento de tempo Δt regido por um campo vetorial de velocidade $\overrightarrow{V}$. Tal formulação pode ser explicitamente definida de forma simples:

$$\frac{d\overrightarrow{x}}{dt} = \overrightarrow{V}(\overrightarrow{x}) \tag{3-2}$$

A resolução do deslocamento da interface consistiria basicamente em resolver a equação 3-2 para todo ponto pertencente à interface. Entretanto, pelos mesmos motivos descritos para justificar a representação implícita da interface, a forma implícita é mais adequada também para descrição da evolução da interface. Deste modo a equação pode ser reescrita da seguinte

maneira:

$$\frac{\partial \phi}{\partial t} + \overrightarrow{V} \cdot \nabla \phi = 0 \tag{3-3}$$

Considerando ∇ o operador de gradiente e $\overrightarrow{V} = \{u, v, w\}$ o campo vetorial de velocidade, temos:

$$\overrightarrow{V} \cdot \nabla \phi = u\frac{\partial \phi}{\partial x} + v\frac{\partial \phi}{\partial y} + w\frac{\partial \phi}{\partial z}$$

A equação diferencial parcial (EDP) 3-3 define o deslocamento da interface onde $\phi(\overrightarrow{x}) = 0$, que é conduzida segundo um campo vetorial $\overrightarrow{V}$. Este é o chamado termo de advecção em um método *level sets*, e incorpora a informação da influência de um campo externo na evolução da interface.

Existem outros fatores que podem influenciar no resultado final de segmentação, e nesse sentido a EDP de *level sets* pode ser enriquecida. Para obtermos contornos mais suaves, uma informação importante é a curvatura média.

O deslocamento na interface devido à curvatura média é definido de forma a ocorrer na direção normal à interface e com uma velocidade proporcional à sua curvatura. Matematicamente temos $\overrightarrow{V} = -b\kappa\overrightarrow{N}$ onde $b > 0$ é uma constante e κ é a curvatura média, calculada como o divergente da normal à superfície (3-4). Desta forma se $b < 0$, a interface se move na direção da concavidade, neste caso círculos se constraem até formarem um único ponto e desaparecem. Se $b > 0$, a interface propaga na direção da convexidade, e círculos se expandem.

$$\kappa = \nabla \cdot \vec{n} \tag{3-4}$$

Este comportamento é bastante interessante e esta componente é capaz de evitar vazamentos na segmentação, comuns em métodos como crescimento de regiões, em nome de uma curvatura média desejada pré-determinada.

O movimento causado pelo campo vetorial induzido pela curvatura média da curva pode ser expresso de forma análoga à equação 3-3. Desta forma teríamos a EDP que descreve o descolamento devido à curvatura média da interface:

$$\frac{\partial \phi}{\partial t} - b\kappa\overrightarrow{N} \cdot \nabla \phi = 0 \tag{3-5}$$

É possível entretanto simplificar esta equação já que por definição o movimento devido à curvatura média só possui componente na direção normal à interface. Desta forma teríamos:

$$\overrightarrow{N} \cdot \nabla \phi = \frac{\nabla \phi}{|\nabla \phi|} \cdot \nabla \phi = \frac{|\nabla \phi|^2}{|\nabla \phi|} = |\nabla \phi|$$

E a equação 3-5 se resume a:

$$\frac{\partial \phi}{\partial t} - b\kappa|\nabla\phi| = 0 \tag{3-6}$$

A EDP 3-6 define o deslocamento da interface devido à curvatura média da interface e compõe o chamado termo de curvatura média em *level sets*.

Outro termo também definido na modelagem clássica de *level sets*, é relacionado a um campo de velocidade internamente gerado, cujo movimento ocorre também na direção normal à interface. Neste caso o campo de velocidade pode ser definido simplesmente como $\overrightarrow{V} = a\overrightarrow{N}$, onde a é uma constante. Levando em conta as mesmas considerações adotadas para o campo de velocidade na direção normal gerado pela curvatura média, podemos obter a EDP relativa a este deslocamento de forma análoga à equação 3-6:

$$\frac{\partial \phi}{\partial t} + a|\nabla\phi| = 0 \tag{3-7}$$

Neste caso é possível observar que se $a > 0$ a interface se move na direção normal, e se $a < 0$ a interface se move na direção contrária à normal.

A equação 3-7 descreve o chamado termo de propagação do *level sets*, e basicamente define o componente responsável pelo deslocamento normal à interface, análogo a uma força de inflar/esvaziar a interface.

Os três termos apresentados: advecção, curvatura média e propagação, compõem a equação geral de *level sets*. Esta EDP, representada pela equação 3-8, descreve a evolução de uma interface no tempo, que define características para a curva final, modelando-a através dos parâmetros da equação.

$$\frac{\partial \phi}{\partial t} + \overrightarrow{V} \cdot \nabla\phi - b\kappa|\nabla\phi| + a|\nabla\phi| = 0 \tag{3-8}$$

O resultado final é obtido de acordo com algum critério de parada na evolução da curva. Um critério bastante comum consiste em um valor mínimo de variação da curva, onde a evolução pára se a diferença entre a superfície gerada em um passo e a superfície gerada no passo posterior, for menor que este valor.

3.4 Algoritmos Genéticos

3.4.1 Fundamentos

Algoritmos Genéticos (AG) são uma ferramenta computacional de busca para encontrar soluções aproximadas em problems de otimização. Eles são baseados na teoria da evolução das espécies de Charles Darwin (Darwin03). O princípio geral da teoria da evolução de Darwin é que características individuais

são transmitidas de pais para filhos através das gerações, e os indivíduos mais adaptados ao ambiente têm maiores chances de sobreviver e gerar descendentes, passando suas características particulares para gerações futuras.

Os AG se baseiam na forma como os seres vivos evoluem para atingirem soluções ótimas para um problema. Para tal, passam por uma série de etapas inspiradas na seleção natural e na reprodução sexuada dos seres vivos.

Em termos computacionais um indivíduo representa uma solução potencial para um dado problema, e suas características importantes em relação ao problema são chamadas genes. Nesta dissertação, como a aplicação consiste em encontrar parâmetros ótimos de segmentação, os genes são um conjunto de parâmetros do método proposto de segmentação.

Uma população é um conjunto de indivíduos de uma determinada geração, e os indivíduos de uma população são avaliados de acordo com a sua capacidade de resolver o problema proposto. Esta capacidade é determinada através de uma função de aptidão, que indica numericamente quão bom um indivíduo é como solução do problema (Michalewicz96).

AG propõem um processo evolucionário para encontrar soluções que maximizem (ou minimizem) uma função de aptidão. Esta busca é realizada iterativamente através das gerações de indivíduos. Em cada geração os indivíduos menos aptos são descartados, e novos indivíduos são gerados pela reprodução dos mais aptos, através de operadores genéticos.

Assim como na natureza, onde os seres mais adaptados ao meio transmitem seus genes para a próxima geração, em AG acontece algo semelhante: os indivíduos que mais se aproximam da solução ideal têm maiores chances de serem escolhidos para passarem suas características para a próxima população.

Ao final do processo evolucionário o indivíduo mais apto representa a solução aproximada do problema de otimização. A figura 3.2 ilustra o processo.

3.4.2 Representação

Um questão importante em AG é a forma como se representa um indivíduo. Esta representação deve expressar objetivamente a qualidade da solução do problema, e ao mesmo tempo estar descrita de forma a ser manipulável por um processo computacional.

Existem diversas formas diferentes de representar ou codificar um determinado problema. Algumas delas são, entretanto, mais freqüentes: representação binária, em números reais e por permutação de símbolos. Na representação binária cada indivíduo tem suas características representadas por 0 ou 1, onde cada conjunto de genes (0 ou 1) com um significado para a solução

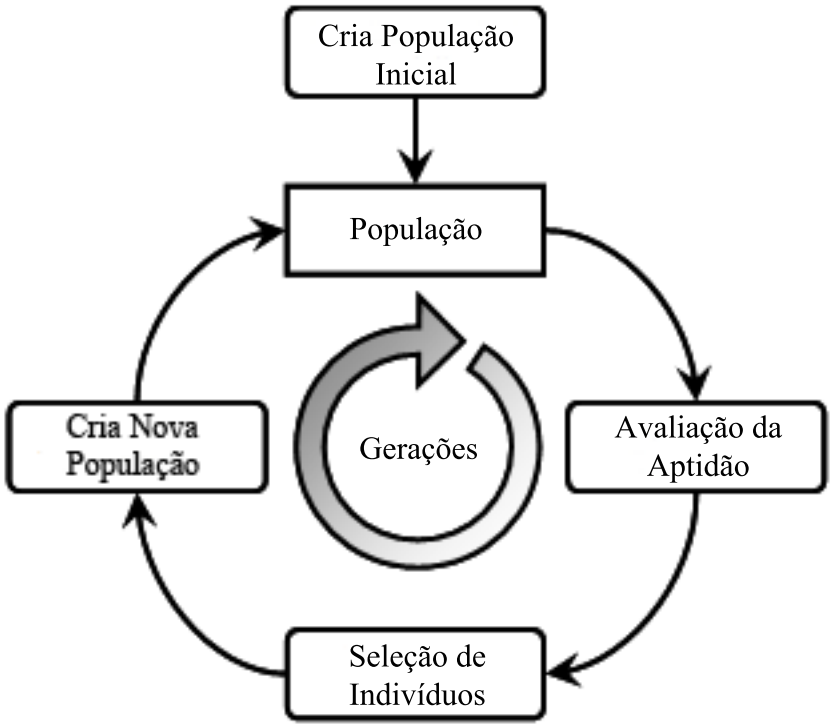

Figura 3.2: Processo evolucionário.

é chamado de cromossomo. Na represetação por números reais os genes podem assumir qualquer valor dentro do conjunto de números reais, e este tipo de representação é bastante utilizado, devido à fácil associação com o problema real. A representação por permutação de símbolos é utilizada em problemas onde a ordem é um dado relevante.

3.4.3 Avaliação e Seleção

Na natureza indivíduos mais aptos sobrevivem e passam suas características genéticas para seus descendentes. Em AG ocorre o mesmo, com a diferença que neles a avaliação de aptidão de um indivíduo deve ser explicitamente definida, e o valor numérico gerado nesta avaliação é utilizado então para a seleção dos mais aptos.

Todos os indivíduos são avaliados verificando-se o quão próxima a solução por eles representada está da solução ideal. Através dessa verificação é gerado um valor real que quantifica este conceito de proximidade, chamado de aptidão. A forma como este valor é calculado para um determinado indivíduo é absolutamente dependente do problema para o qual se busca a solução.

A seleção é um processo baseado nos valores de aptidão que escolhe indivíduos para reprodução, isto é, para gerarem novos indivíduos com parte

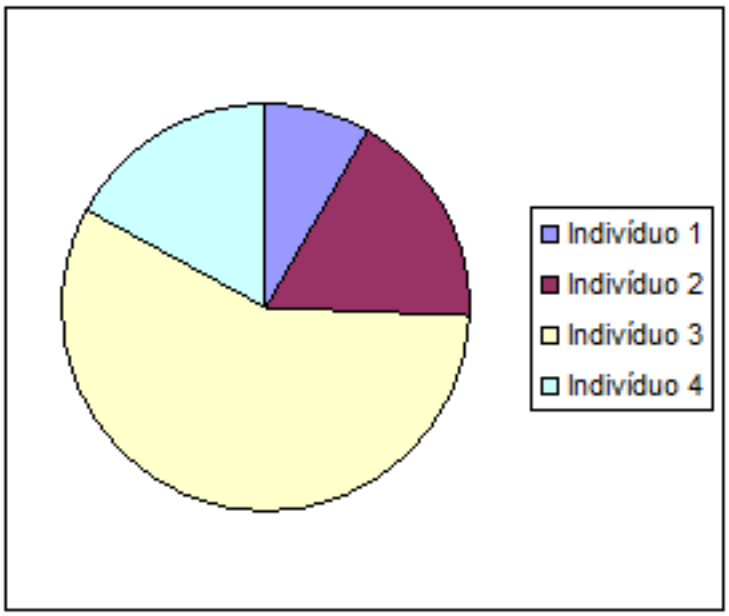

Figura 3.3: Seleção por roleta.

do seu material genético. Este mecanismo ocorre de forma que as soluções com maiores valores atribuídos pela função de avaliação, possam ter maiores chances de reproduzir e criarem novos indivíduos.

Desta forma espera-se que um indivíduo entre os mais aptos contenha material genético importante na busca da solução ótima, e que a combinação do material genético dos indivíduos mais aptos leve a soluções cada vez mais próximas da solução ideal.

Existem diferentes formas de fazer seleção de indivíduos, priorizando ou amenizando a participação de certos grupos de indivíduos no processo evolucionário. A mais comum de todas é a seleção por roleta. Nela, a probabilidade de um indivíduo ser selecionado é proporcional ao valor de sua aptidão, e este cálculo é representado pela equação 3-9, onde N é tamanho da população, f é a aptidão, e p a probabilidade de seleção:

$$p_i = \frac{f_i}{\sum_{j=1}^{N} f_j} \tag{3-9}$$

Uma representação gráfica para o entendimento deste tipo de seleção é dada na figura 3.3. Cada fatia do disco representa a probabilidade de um indivíduo ser escolhido para reprodução, dada uma população com quatro indivíduos.

A seleção por roleta tem a seu favor a sua simplicidade de implementação. Apresenta algumas desvantagens como uma situação em que um indivíduo possua uma probabilidade de escolha muito maior que os demais, isto pode

fazer com que o algoritmo chegue ao fim muito rapidamente sem ter alcançado uma solução satisfatória.

Uma vez selecionados os indivíduos mais aptos, uma nova geração é criada, através de um processo de reprodução que utiliza operadores genéticos. Esta nova geração é composta geralmente dos indivíduos mais aptos da geração anterior, seus descendentes e indivíduos gerados aleatoriamente,.

3.4.4 Operadores Genéticos

Um operador genético representa uma regra para a geração de novos indivíduos. Os operadores genéticos clássicos são o *crossover* e a mutação.

O *crossover*, também conhecido como reprodução ou cruzamento, se baseia na reprodução sexuada dos seres vivos e consiste na mistura de genes de dois indivíduos escolhidos na etapa de seleção, gerando novos indivíduos que herdam características dos indivíduos originais. A idéia geral é que a aptidão de um indivíduo é função de suas características, e a troca de genes bons pode gerar indivíduos ainda mais aptos, dependendo dos genes herdados de seus pais. Embora indivíduos menos aptos possam ser gerados neste processo, eles tendem a ser eliminados durante o processo de seleção.

Os indivíduos escolhidos na etapa de seleção podem ou não participar do processo de reprodução. Isto é determinado pela chamada taxa de *crossover*, um valor entre 0 e 1 que indica a probabilidade de se aplicar o operador de cruzamento sobre um indivíduo. Os indivíduos selecionados que não participam da reprodução são repetidos para geração seguinte, os que participam tem seus descendentes na próxima geração.

Existem diversas formas de se aplicar o operador de *crossover* sendo as duas mais comuns a que utiliza um ponto de corte (*one-point crossover*) e a que utiliza dois pontos de corte (*two-point-crossover*). Outra abordagem bastante comum, também utilizada neste trabalho, é o chamado *crossover* aritmético, onde é realizada uma combinação linear entre dois indivíduos, sendo os coeficientes da combinação escolhidos aleatoriamente.

No *crossover* com um ponto de corte um ponto nos cromossomos de dois indivíduos I1 e I2 é escolhido aleatoriamente, separando-os em uma ala esquerda e uma ala direita. A ala esquerda do cromossomo de I1 se junta com a ala direita do cromossomo de I2 e a ala direita do cromossomo de I1 se junta com a ala esquerda do cromossomo de I1, gerando dos novos indivíduos com características de seus genitores. Este procedimento é ilustrado na figura 3.4.

No *crossover* com dois pontos de corte dois pontos nos cromossomos de dois indivíduos I1 e I2 são escolhidos aleatoriamente, separando-os assim em

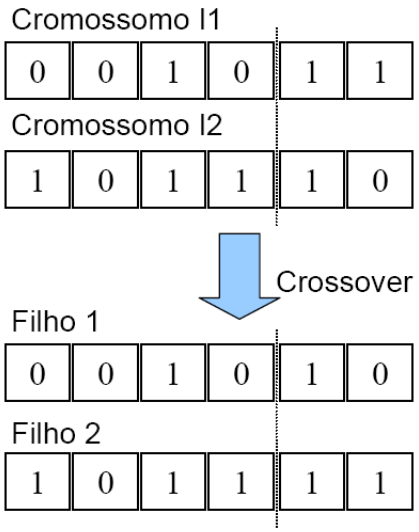

Figura 3.4: *Crossover* de um ponto.

uma ala esquerda, uma parte central e uma ala direita. Neste método somente as partes centrais são trocadas, ou seja, a parte central de I1 se junta com a ala esquerda e direita de I2 e a parte central de I2 de junta com a ala esquerda e direita de I1. Este procedimento é ilustrado na figura 3.5.

A mutação muda os valores dos genes de modo aleatório, respeitando os intervalos de busca estabelecidos para cada gene. Esta operação desempenha um papel importante no processo evolucionário porque introduz um componente aleatório na busca da solução ótima, servindo como uma ferramenta para evitar convergência para mínimos locais, já que pode direcionar a evolução para regiões ainda não visitadas do espaço de busca.

A figura 3.6 ilustra o procedimento de mutação na codificação binária, entretanto este tipo de operação também pode ser realizada com números reais, como neste trabalho, onde o valor final do gene sorteado é calculado aleatoriamente dentro de um intervalo pré-definido.

Estes dois tipos de operadores genéticos tem comportamentos complementares. De um lado o *crossover* tende a indicar o caminho para a solução local ótima, contribuindo para a convergência da população durante o processo evolucionário. Por outro lado, a mutação implementa mudanças bruscas nos genes da população, e com isso permite que regiões do espaço de busca ainda não visitadas, e que podem incluir a solução ótima, sejam contempladas no processo evolucionário.

Outros operadores genéticos podem ser encontrados na literatura (Michalewicz96). A maioria deles são variantes de *crossover* e mutação, adaptados para tipos específicos de problemas.

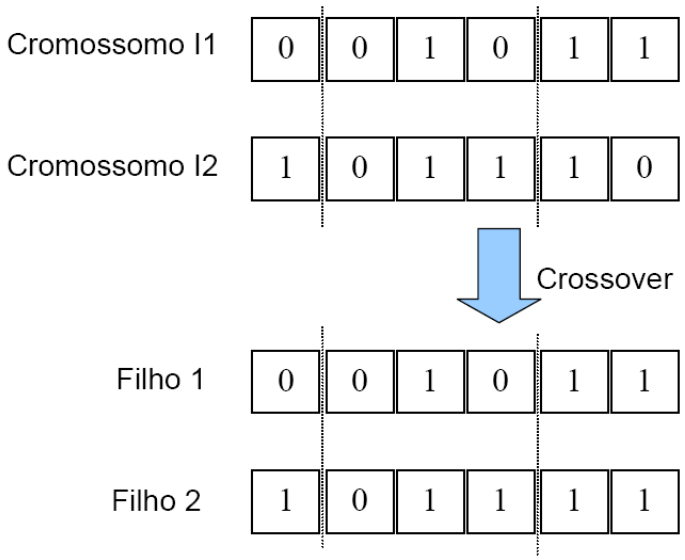

Figura 3.5: *Crossover* de dois pontos.

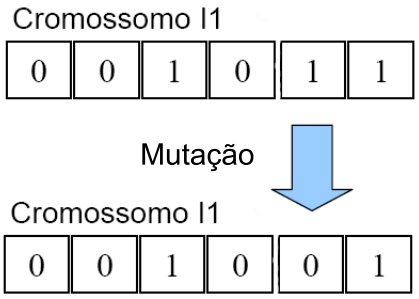

Figura 3.6: Mutação.