4
Segmentação

Este capítulo apresenta primeiramente o algoritmo proposto para a segmentação do áudio em detalhes. Em seguida, são analisadas as inovações apresentadas.

É importante mencionar que as mudanças de cena mencionadas durante este capítulo não se referem apenas a mudanças claras como duas cenas diferentes de um filme, ou a entrada de um intervalo comercial durante um programa de televisão. Com a análise a partir do áudio, podem-se encontrar outros tipos de mudanças significativas, como a entrada do refrão ou de um solo de guitarra em uma música, que são muito úteis para a indexação de vídeos. Durante este capítulo, vamos nos referir a todos esses tipos de mudanças como mudanças de cena. Diferentes tipos de indexações que são encontrados pelo algoritmo vão ser discutidos no capítulo 5, que expõe os seus resultados.

4.1.
Algoritmo proposto

O algoritmo desenvolvido para a segmentação foi inspirado na idéia apresentada em Foote (2000), já mencionado no capítulo 2. A idéia do algoritmo é analisar o quanto de mudança ocorre em determinado instante no áudio, e estabelecer os momentos onde as maiores mudanças acontecem como candidatos a mudanças de cenas. Além disso, são adicionados mais dos fatores. Um deles é a detecção de intervalos de silêncio longos, que também sugerem uma mudança de cena, por serem muito comuns entre dois programas, em um intervalo comercial, ou até entre cenas diferentes de um filme ou seriado de televisão. O outro é o resultado da classificação; mudanças no tipo de áudio, indicadas por esse resultado, são fortes indicativos de mudanças de cenas. O algoritmo foi desenvolvido visando ao propósito de trabalhar em paralelo à análise do vídeo, para que os resultados combinados possam ser usados em busca de uma segmentação mais precisa do que a que é possível individualmente.

Primeiramente, é estabelecida a duração da janela de análise para cada *frame*. Nos resultados apresentados nesta dissertação, a janela trabalhada foi de 1 segundo. Esse tamanho de janela é um parâmetro do algoritmo e portanto pode ser facilmente alterado. A janela de 1 segundo foi escolhida após testes como um tamanho geral para a análise, de modo que o algoritmo trabalhe sem informações prévias do vídeo, a não ser as que foram encontradas na classificação. Em amostras com vídeos de características diferentes, podem eventualmente ser encontrados melhores resultados utilizando outros tamanhos de janelas. Isso será mais detalhado na parte de trabalhos futuros desta dissertação, na seção 6.2.

Os primeiros passos do algoritmo de segmentação são realizados para todos os *frames* do áudio onde seja possível trabalhar com a janela estabelecida. Sendo assim, os frames anteriores a 1 segundo de vídeo, no caso dessa janela, e os frames pertencentes ao último segundo do vídeo não são considerados candidatos.

Para cada *frame*, o corrente é estabelecido como candidato a fronteira de segmentos, e são estabelecidos dois segmentos para comparação. O segmento anterior consiste dos *frames* anteriores ao *frame* que está sendo analisado, com uma duração total correspondente ao tamanho da janela de análise. Já o segmento seguinte consiste do grupo de *frames* iniciado no frame corrente e que possui a mesma duração do fragmento anterior.

Para cada um dos segmentos, calcula-se o somatório dos valores dos fatores de escala de cada sub-banda. O grau de diferença entre os dois segmentos definidos pelo *frame* é então calculado a partir desses somatórios. Para isso, primeiramente é calculada a diferença entre os volumes totais de cada sub-banda para os segmentos. É calculada a diferença entre o volume total dos segmentos em cada sub-banda, e o grau de diferença entre os dois segmentos consiste do somatório dos módulos dessas diferenças. Esse grau de diferença é considerado o grau de novidade do *frame* corrente.

Quando o grau de novidade já foi calculado para todos os *frames*, o algoritmo passa a procurar os *frames* que indicam picos de mudanças. Inicialmente, são classificados como picos de mudanças os *frames* em que o grau de novidade é maior do que o dos seus *frames* adjacentes, tanto o anterior quanto o seguinte. Todos os outros *frames* são então considerados candidatos inválidos e excluídos da análise.

A partir dos graus de novidades dos *frames* considerados picos, é calculada a média do valor dos picos. A partir dessa média, é estabelecido um limite mínimo de grau de novidade para que um *frame* que é pico seja considerado um candidato a mudança de cena. Nos resultados apresentados aqui, o limite foi calculado a partir da soma da média dos picos somada a 60. Esse valor, assim como o tamanho da janela de análise, é um parâmetro, e pode ser alterado de acordo com as características do vídeo e da quantidade de picos que se espera serem encontrados. Isso vai ser mais detalhado na parte de trabalhos futuros desta dissertação, na seção 6.2.

Os candidatos a mudança de cena são, então, estabelecidos de duas maneiras. A primeira consiste nos picos cujo grau de novidade é maior que o limite mínimo. A segunda utiliza-se da classificação do áudio como um indicativo da ocorrência de mudança. Quando dois envelopes consecutivos têm classificações distintas, sabe-se que a mudança do tipo de áudio encontra-se nos 2 segundos de sobreposição entre esses envelopes. Assim, picos que estão contidos na sobreposição de dois envelopes de classificação diferente também são considerados candidatos válidos.

A próxima etapa filtra os candidatos a mudança de cena que vão ser sugeridos pelo algoritmo. Essa etapa baseia-se em outro parâmetro que pode variar na análise, que é o tamanho mínimo de um segmento. Nos testes realizados, esse parâmetro foi estabelecido em 1,5 segundos, porém pode ser alterado de acordo com as necessidades do usuário. A partir desse tamanho mínimo, é estabelecido que não pode haver um candidato a mudança de cena dentro de um dos segmentos. Com isso, são eliminados todos os candidatos que possuem um grau de novidade menor do que a de qualquer outro candidato que esteja a até 1,5 segundos de distância dele no fluxo do áudio. Isso garante que os segmentos encontrados pelo algoritmo vão respeitar o tamanho mínimo estabelecido.

Com isso, temos estabelecidos todos os candidatos selecionados a partir do grau de novidade do áudio. A última etapa do algoritmo de segmentação parte da detecção de intervalos longos de silêncio como indicativo de mudanças de cena, e inclui os candidatos encontrados por esse método na lista de candidatos. Esse método é utilizado também com um filtro final para os candidatos estabelecidos anteriormente, como explicado a seguir.

A detecção de silêncio funciona a partir da mesma técnica apresentada no capítulo 3, porém sem um tamanho de segmento específico. Quando um *frame* é detectado como silêncio, o algoritmo procura por *frames* consecutivos que mantenham o volume em silêncio e assim calcula o tamanho do intervalo de silêncio encontrado.

Como são muito comuns intervalos de silêncio durante a fala, é recomendado que o tamanho mínimo que um segmento de silêncio deve ter para indicar uma mudança de cena seja pelo menos de 2 segundos. Esse valor é mais um parâmetro, e pode ser alterado de acordo com o tipo de vídeo e as necessidades do usuário.

A partir dos intervalos de silêncio encontrados, são realizados dois procedimentos. Primeiramente, os pontos centrais desses intervalos são indicados como candidatos a mudanças de cena. Eles não podem ser encontrados a partir do grau de novidade, pois durante o silêncio o som é bastante estável, porém são fortes indicativos de mudanças de cena quando ocorrem. O ponto central é indicado como representante dessa mudança a partir do áudio por estar exatamente no meio do intervalo, o que deixa um final de silêncio em um segmento e um início no segmento seguinte, ambos com a mesma duração. Porém, as informações da imagem são muito importantes para definir o *frame* exato da fronteira de segmentos nesses casos, pois nem sempre o *frame* central é o indicativo correto. Isso vai ser mais detalhado na parte de trabalhos futuros desta dissertação, na seção 6.2.

A segunda função da detecção de intervalos de silêncio é servir como filtro para as mudanças de cena detectadas anteriormente. Candidatos que estão dentro de um intervalo de silêncio detectado, ou em suas fronteiras, geralmente têm o seu momento incorreto, já que é muito raro que a fronteira de dois segmentos esteja localizada exatamente no início ou no fim de um intervalo de silêncio. Normalmente, há pelo menos um pequeno silêncio antes da troca da cena, ou no seu início.

Com as características da análise por grau de novidade, essas fronteiras tornam-se muitas vezes candidatos fortes para esse tipo de análise, já que a distância entre o silêncio e o não-silêncio é sempre alta. Assim, a inclusão desse filtro evita erros eventuais causados por isso e garante que o candidato a ser indicado pela análise do áudio vai ser central em intervalos de silêncio de tamanho

considerável. Na seção 6.2, é comentada também a hipótese da utilização desse mesmo método para evitar *false hits* durante um período de fala, onde intervalos pequenos na fala muitas vezes causam a detecção de candidatos que não representam mudanças verdadeiras.

4.2. Inovações apresentadas

A grande maioria dos métodos encontrados na literatura que fazem a segmentação do áudio trabalha exclusivamente a partir de sua classificação. Nesses métodos, o único indicativo de uma mudança de cena analisado é a mudança de classificação entre dois segmentos de áudio. Portanto, esses métodos trabalham exclusivamente na procura dessas fronteiras, e deixam de lado a possibilidade de mudanças sem que a classificação dos dois segmentos analisados seja diferente.

O algoritmo desenvolvido para a segmentação nesta dissertação baseia-se na idéia de similaridade momentânea do áudio apresentada no artigo de J. Foote (2000), explicado no capítulo 2. Esse método permite uma análise geral do áudio, não se baseando exclusivamente nos resultados da classificação, os quais são utilizados nesta dissertação como apenas uma parte das informações analisadas. Como esse artigo não trabalha no domínio comprimido, a única idéia adotada dele retirada foi a procura por essa alteração no áudio no domínio de freqüências. A essa idéia, foi acoplada a análise tradicional com a inclusão dos picos encontrados em momentos onde há mudança da classificação do áudio como candidatos a mudanças de cena.

A partir da intenção de trabalhar com a similaridade dos segmentos, foi desenvolvida a idéia de se utilizar a energia de cada sub-banda como um fator de comparação entre os segmentos. Foram experimentados diferentes métodos de análise do grau de novidade. O método de somatório dos módulos das diferenças simples entre as energias de cada sub-banda demonstrou ser o mais eficiente, por ter um grau de detecção de mudanças de cena muito alto, conforme demonstram os resultados apresentados no capítulo 5.

Em seguida, o desenvolvimento do algoritmo focou-se na construção de filtros para que os candidatos encontrados sejam selecionados de modo que apenas as mudanças reais de cena fossem indicadas.

O primeiro filtro desenvolvido, a partir dos picos, é uma idéia comum em diferentes tipos de sistemas, por simplesmente selecionar os candidatos mais fortes, primeiro globalmente, e depois localmente. Tanto esse filtro quanto a idéia de que um segmento deve ter um tamanho mínimo são comuns na literatura, e foram explorados neste trabalho por terem sua efetividade comprovada em inúmeros trabalhos anteriores. Essa idéia serve como base para o segundo filtro, que também trabalha a partir de forças dos candidatos, porém a partir de restrições de tamanho dos segmentos.

A detecção de intervalos de silêncio é um procedimento comum nas análises realizadas a partir apenas da classificação. Nesses métodos, muitas vezes esses intervalos são o primeiro indicativo de um fim de segmento, como em S. Kiranyaz, M. Aubazac e M. Gabbouj (2003), explicado no capítulo 2. Como o método de classificação apresentado nesta dissertação trabalha em envelopes de 4 segundos, muitas vezes intervalos menores não são detectados como silêncio na classificação. Assim, esse método foi incluído como possível indicador de segmentos e como um filtro, de modo a minimizar a quantidade de *false hits* detectados.

Pode-se observar que o método desenvolvido baseou-se em diversas idéias pré-existentes, porém que nunca foram utilizadas de forma conjunta. Essa união de técnicas permite um alto grau de eficiência na detecção de mudanças significativas. Além disso, nenhuma das técnicas estudadas trabalha a partir apenas dos fatores de escala. Com isso, todos os algoritmos tiveram que ser adaptados e re-criados para que pudessem realizar o esperado com uma quantidade significativamente menor de informações.

O algoritmo resultante é relativamente simples, porém seus resultados demonstram todo o seu potencial. O capítulo 5 vai detalhar esses resultados, porém é importante citar que a porcentagem das mudanças de cena detectadas pelo algoritmo é extremamente alta.

Além disso, o algoritmo se apóia em filtros para detalhar sua análise. Isso permite que, com o desenvolvimento de novos tipos de filtro, o grau de precisão da análise realizada seja aumentado. O algoritmo apresentado aqui é uma forma

bem básica, com apenas filtros comuns tendo sidos incorporados na sua análise. Na seção 6.2 desta dissertação, são mencionadas as possibilidades de desenvolvimento de novos filtros e da junção da avaliação do áudio com a do vídeo, que podem permitir a eliminação de *false hits* e uma maior precisão para a definição de fronteiras de segmentos.