Fausto Neri da Silva Vanin

Caracterização de Níveis de Ação em Vídeos Estruturados

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Curitiba
2005

Fausto Neri da Silva Vanin

Caracterização de Níveis de Ação em Vídeos Estruturados

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Área de Concentração: Inteligência Artificial

Orientador: Díbio Leandro Borges

Curitiba
2005

Vanin, Fausto Neri da Silva
Caracterização de Níveis de Ação em Vídeos Estruturados. Curitiba, 2005.

Dissertação - Pontifícia Universidade Católica do Paraná Programa de Pós-Graduação em Informática Aplicada.

1. Inteligência Artificial 2. Reconhecimento de Padrões 3.
I.Pontifícia Universidade Católica do Paraná. Centro de Ciências Exatas e Tecnologia. Programa de Pós-Graduação em Informática Aplicada II - t

À minha família pelo apoio em todos os momentos. Aos colegas Paulo, Islenho e Ederson pelo trabalho em equipe e ao Prof. Díbio pela orientação e amizade.

Agradecimentos

Ao Colega David Menoti por acreditar no meu trabalho, ao colega Carlos Silla pelo apoio técnico. Ao Prof. Celso Kaestner pela presteza nas questões referentes ao curso. Ao Prof. Alceu de Britto Jr. pelas orientações durante todo o curso. Ao suporte técnico e à Secretaria do PPGIA pelo profissionalismo. Ao Prof. Carlos Maziero pela disponibilidade. Aos membros, efetivos e desligados do LUCI^{2}A: Paulo Cavalin, Islenho de Almeida, Ederson Sgarbi, Fernanda Ramos, Cristiane e Willian Ferreira e, especialmente ao Prof. Díbio Borges pela oportunidade.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

SC	Característica de Cor
SM	Característica de Movimento
LC	Limiares de interseção de cores
LM	Limiares de interseção de movimento
λ	Precisão
φ	Recobrimento

Lista de Abreviações

DVD	*Digital Versatile Disc*
MPEG	*Moving Pictures Expert Group*
MEM	*Modelo Escondido de Markov*
KL	*Kullback-Liebler*
HSI	*Matiz, Saturação, Intesidade*
QPS	*Quadros por Segundo*

Resumo

Este trabalho descreve um método de caracterização de vídeos estruturados do tipo filmes. Estes vídeos serão analisados através de características de cor e movimento. São definidas 20 subregiões para cada imagem e de cada subregião calcula-se a Interseção de Histogramas HSI quantizados (Cores) e a Máxima Verossimilhança (Movimento). Foi criada uma base de dados com 20 mil imagens de dois filmes. O método foi aplicado a esta base de dados e seus resultados são avaliados através de Precisão e Revocação.

Abstract

This work describes a structured video characterization method. These videos where analised through color and motion features. Twenty subregions are defined for each image and for each subregion we extract the HSI Histograms Intersection (color feature) and the Maximum Likelihood (motion feature). A 20,000 thousand images base was created with images from 2 movies. The method was aplied to the image base and the obtained results where evaluated over the criteria of Precision-Recall.

Capítulo 1

Introdução

Este trabalho foi desenvolvido utilizando os conceitos de Inteligência Artificial, especificamente o Reconhecimento de Padrões.

O uso de vídeos digitais em ambientes domésticos é cada vez mais comum na vida cotidiana. O advento de tecnologias como as que popularizaram o DVD (*Digital Versatile Disc*) e a grande difusão de vídeos através da *internet* - este por causa do uso crescente de conexões que suportam grandes transferências de dados - são fatores que demonstram a grande inserção destes vídeos no cotidiano. Essas novidades são acompanhadas e/ou originadas por evoluções técnicas como a compactação de vídeo, caracterizada no desenvolvimento de padrões que reduzam o espaço consumido pelos arquivos, sem perda de qualidade. Com tantos fatores - deixando de lado as questões comerciais -, compartilhar ou adquirir vídeos digitalizados tem se tornado uma atividade tão comum à vida das pessoas quanto o compartilhamento de imagens digitais.

Esse uso comum das novidades tecnológicas traz consigo a necessidade de criação de soluções capazes de facilitar o uso e o acesso a estes produtos, ou até mesmo tratar questões que o próprio uso destas novidades pode trazer.

Justamente nas questões relacionadas à manipulação e análise dos vídeos, têm sido desenvolvidas diversas ferramentas computacionais. Entre essas ferramentas estão sistemas de segmentação de vídeo, rotulação [GULER et al., 2003], [FAN and LUO, 2003], detecção de movimentos não convencionais em vídeos de segurança [STAUFFER, 2003], [CHOWDURRY and CHELLAPA, 2003], sumarização de vídeos esportivos [EKIN et al., 2003], [EKIN and TEKALP, 2003], entre outros.

Uma questão que desperta interesse entre todas estas ferramentas computacionais e soluções matemáticas para a análise de vídeo, a caracterização do conteúdo em vídeos estruturados é uma área que possui uma grande diversidade de aplicações e de possibilidades de desenvolvimento.

Um vídeo estruturado, como um filme por exemplo, possui características estruturais nas suas imagens e na variação entre essas imagens que possibilitam o desenvolvimento de técnicas que permitam avaliar e caracterizar os dados contidos na seqüência.

As técnicas de análise e caracterização de vídeo utilizam o vídeo em duas principais formas: 1) trabalhando diretamente com o arquivo de vídeo em formato compactado; e 2) trabalhando com as imagens extraídas do vídeo.

Uma questão interessante que surge é a caracterização de vídeos estruturados de forma automática. Esta questão está relacionada a um método capaz de separar os diferentes níveis de ação contidos neste vídeo.

Este trabalho descreve um método para caracterização dos níveis de ação contidos em vídeos estruturados do tipo filmes, através da combinação de características de cor e movimento.

A estrutura deste documento está como segue: Capítulo de Fundamentação Teórica no Capítulo 2, em seguida, no Capítulo 3 o Estado da Arte. A próxima seção é a descrição do método, e depois no Capítulo 5 os Experimentos Realizados, sendo que por fim as Conclusões.

Capítulo 2

Fundamentação Teórica

Esta seção descreve o detalhes sobre as formas de representação de vídeos e também conceitos relacionados à cinegrafia de vídeos estruturados.

2.1 Representação de vídeo

Um vídeo pode ser representado basicamente de duas formas: compactada e não-compactada.

Os vídeos compactados são formas de representação da seqüência de imagens baseada na eliminação de redundâncias. Uma das principais formas de compactação de vídeo é o formato MPEG (*Moving Pictures Expert Group*). Este formato utiliza os coeficientes resultantes da aplicação da transformada cosseno para representacão de cada imagem, sendo que cada uma das imagens é dividida em uma série de macroblocos [1]. Para cada macrobloco $k(i, j)$, onde i representa um índice em linha e j em coluna, as diferenças espaciais destes macroblocos através da seqüência em uma região de busca são computadas e de acordo com um nível de similaridade, assume-se que um determinado valor se repete ou não nos próximos $k(i + x, j + y)$ macroblocos.

Estas diferenças são computadas em posições relacionais à cada quadro, definindo dois tipos diferentes de quadros para codificação:

- Quadros P: São quadros que são codificados utilizando como base de comparação as informações espaciais apenas dos quadros anteriores.

- Quadros B: Estes quadros, diferentemente dos quadros P utilizam as informações espaciais tanto dos quadros anteriores quanto posteriores para a codificação. É o

[1] Macrobloco: são quadrados, geralmente de dimensões 8X8, formados pelos píxeis de uma imagem

tipo de quadro que proporciona uma maior compressão do arquivo.

- Quadros I: A representação dos quadros I é a que proporciona menor compressão, porque estes quadros fazem a compactação baseado apenas nas informações contidas no próprio quadro.

Desta forma, ao invés dos valores representativos de cada quadro, quando possível, apenas o macrobloco de referência e sua posição são armazenadas, reduzindo desta forma a quantidade de dados representativos e conseqüentemente reduzindo o tamanho do arquivo.

Esta forma de armazenagem incide em uma perda de informação no momento da codificação do arquivo. O grau de perda de informação é dado pela forma de codificação escolhida. A maioria dos vídeos utilizam uma estrutura que possui 1 quadro do tipo I a cada 8 na seqüência.

Os vídeos em formato não compactado são representados de duas formas:

- Existem formas de geração de vídeo digital que não utilizam compactação ou estruturação espacial dos quadros contidos na seqüência. Cada imagem é independente das outras e é apenas acrescentada à seqüência, sem nenhum tipo de compactação, preservando, desta forma, a qualidade original do vídeo.

- Outra forma de representação utiliza todas as imagens do vídeo de forma separada, sem nenhuma forma de agrupamento. O espaço consumido por essas imagens vai ser da mesma ordem do método anterior, com a diferença de que cada imagem será um arquivo separado.

Esta representação incide certamente em um espaço de armazenamento maior. Basta tomarmos como exemplo a armazenagem de um filme completo com cerca de 90 minutos. Este filme, com uma taxa média de 25 quadros por segundo vai ter um total de imagens em torno de 130.000.

2.2 Vídeos Estruturados

Um vídeo estruturado é todo vídeo que possui um tratamento ou desenvolvimento técnico no seu feitio. Geralmente esses vídeos são editados em estúdios especializados que concatenam as diversas tomadas captadas de forma semântica ao objetivo do vídeo.

Para o entendimento deste texto consideremos a partir de agora por **tomada** todo período de imagens captado por uma mesma câmera, até a ocorrência de um corte físico, visual ou técnico. Entenda-se também por **cena** todo conjunto de tomadas que possuam ligação semântica e espacial no vídeo estruturado.

No momento da edição técnica e composição do vídeo completo através das seqüências distintas capturadas pelas câmeras, várias formas de transição e agrupamento de tomadas podem ser utilizadas:

Os cortes bruscos nas tomadas podem ser devidos a ocorrência de um corte físico, que é quando a câmera que estava captando a seqüência encerra a captação. Pode ser técnica quando a ferramenta de edição do vídeo insere um final abrupto a transposição das imagens da seqüência captada para o vídeo final e pode ser visual quando um elemento alheio à seqüência passa repentinamente a tomar parte significativa do quadro da tomada.

Além das transições bruscas, existem também as transições graduais entre tomadas. Geralmente as transições graduais são utilizadas para não ferir a troca entre as "histórias" do vídeo ou para dar idéia de continuidade. As transições graduais podem ser:

- *Fade-in*: quando as imagens finais da tomada gradativamente tendem a um quadro todo preto.
- *Fade-out*: quando as imagens partem gradativamente de um quadro preto para o começo da tomada seguinte.
- Dissolução: é caracterizado pela troca gradual das imagens finais de uma tomada com as imagens iniciais da tomada seguinte.
- Outros: transições baseadas em efeitos visuais como elementos geométricos ou efeitos de animação sintéticos.

Em muitos casos as transições graduais entre tomadas podem ocorrer combinadas, tornando a tarefa dos sistemas de detecção de tomadas complicada devido ao seu aspecto gradual de mudança.

Existem também características cinemáticas presentes na estrutura das imagens dos vídeos. As nuances de cor aplicadas à seqüência estão relacionadas à comumente chamada fotografia. Da mesma forma, a incidência de iluminação artificial em estúdios de gravação.

O posicionamento da câmera em relação aos eventos capturados é controlada, sendo que o objeto de atenção está sendo bem acompanhado, em tomadas de longa, média e curta distância.

Podem também existir efeitos técnicos no vídeo, como repetições (*replays*) e variações sintéticas de cor.

Outro fator bastante utilizado na estrutura de um determinado tipo de vídeo é a duração de cada tomada. Geralmente em seqüências mais introspectivas, como diálogo

por exemplo, as tomadas são mais duradouras e nos casos de maior ação o tamanho é reduzido.

Existem diversos tipos de vídeos estruturados. Entre eles podemos destacar como exemplo os filmes e as transmissões televisivas.

As transmissões por televisão, mesmo as transmitidas ao vivo, possuem uma estrutura definida que compele a demonstração da notícia ou reportagem, auxiliada de imagens, geralmente alternando entre a apresentação e a demonstração dos vídeos.

Na classe dos filmes, encontram-se vídeos como propagandas, curta-metragens e longa-metragens. Cada um destes filmes, devido a sua duração possui características semânticas de dinâmica diferentes. Diferentemente dos vídeos televisivos os filmes possuem uma maior liberdade de apresentação de cenas diferentes no decorrer da seqüência. Essa liberdade de apresentação das imagens está diretamente ligada à mensagem a ser passada pelo vídeo.

Neste caso a associação de características como cor (fotografia) e dinâmica das cenas são associadas para produzir o efeito desejado ao espectador. Geralmente as trocas bruscas de cores, associadas a mudanças rápidas de tomadas, visam um estado de tensão maior no espectador, devido as constantes mudanças.

Da mesma forma a estabilidade das cores e a maior duração das tomadas ou menor movimentação da câmera tencionam dar maior tranqüilidade ou suspense ao espectador. O conjunto destas características cinemáticas não pode ser ignorado no momento de fazer um sistema de análise automática de um vídeo estruturado, haja visto que as características semânticas neles contidas proporcionam um maior entendimento e uma possível caracterização destes vídeos baseado na verificação destas estruturas.

Capítulo 3

Estado da Arte

Entre os diversos estudos desenvolvidos para tratar de questões relacionadas ao processamento de vídeos estruturados podemos destacar alguns.

Em [IYENGAR, 2002] descreve-se uma técnica de caracterização de "trailers" de filmes baseada em a estrutura semântica do filme e em heurísticas relacionadas às formas de edição dos filmes, baseada na relação entre a quantidade de ação e movimentação com a predominância das personagens do filme como centro das atenções na tomada.

O método recai entre os que utilizam uma abordagem não-compactada e toma como características a duração das tomadas e uma medida de ação, utilizando-as para treinar um modelo escondido de Markov (MEM).

O MEM (*Modelo Escondido de Markov*) foi treinado utilizando como característica uma razão entre a energia contida entre as imagens da tomada e o tamanho da mesma.

Foi utilizada um característica visual que é a distância KL (*Kullback-Liebler*) para histogramas RGB normalizados, dada por:

$$d = \sum_{k} log_2 \left(\frac{p_k}{q_k} \right) \tag{3.1}$$

onde p e q representam os histogramas e k representa o conjuntos dos canais de cada histograma.

Da base de dados extraem-se alguns filmes para treinamento e baseado nas características visuais (KL) são geradas duas funções de densidade-probabilidade para representar ação e tomadas de maior atuação das personagens.

Um ponto importante deste método é a combinação de características visuais extraídas das relações das imagens da seqüência e características estruturais diretamente relacionadas ao feitio e à intenção na concepção da obra.

No trabalho [EKIN and TEKALP, 2003] Ahmet Ekin descreve um modelo para detecção probabilística de eventos em tempo real ou quase real em vídeos de televisão utilizando características cinemáticas e classes de tomada em eventos esportivos. Ele é aplicado em dois esportes: futebol e basquetebol. Os eventos de detectados são agrupados com o objetivo de compor um sumário do vídeo.

A detecção de tomadas é um passo importante em um detector geral de eventos. Alguns algoritmos utilizam grandes transições entre histogramas como uma evidência de troca de tomada. Para melhorar o resultado do método é proposta como nova característica a diferença absoluta da taxa de pixeis coloridos do campo entre dois quadros. Estas duas características são combinadas para a detecção de transições graduais e abruptas.

Para a classificação as tomadas foram divididas em quatro tipos: tomadas distantes, médias no campo e *close-up* ou fora do campo. A maioria dos quadros de uma tomada definirão o tipo desta tomada.

Foi utilizado um algoritmo para detecção de *replay* para a verificação do gol no futebol.

O uso de características cinemáticas (como *replays*, por exemplo) tem o objetivo de dar maior robustez na detecção de eventos em esportes diferentes, considerando que características baseadas em objeto são dependentes de cada categoria.

Alan Hanjalic em [HANJALIC et al., 1997] descreve um método de segmentação de filmes utilizando características visuais, objetivando a extração das chamadas Unidades Lógicas da História (ULH), que são composições de tomadas do filme que possuam relação visual. Aplica-se um cálculo de dissimilaridade entre dois quadros k_1 e $k_1 + p_1$ onde p_1 é uma medida de distância entre esses quadros. Caso esse valor de dissimilaridade supere um determinado limiar, todos os *quadros* compreendidos entre k_1 e $k_1 + p_1$ serão considerados parte da mesma ULH.

Cada tomada identificada é representada por um ou mais quadros-chave. Os quadros-chave de uma tomada são agrupados em uma grande imagem chamada **imagem de tomada**.

No trabalho, assume-se que o vídeo está segmentado em tomadas previamente.

Para representar cada quadro-chave utilizada como característica a média das cores no espaço de cores $L * u * v$[1]de cada uma das imagens compostas.

O autor comenta de uma forma de separar os *quadros*-chave de cada ULH através uma medida de granularidade que não está esclarecida no artigo. Este recurso está associado com a possibilidade de navegar pelo conteúdo do vídeo de uma forma dinâmica

[1]Espaço de cor $L * u * v$: este é um espaço de cor onde L=[0, 100]; u=[-134,220]; e v=[-140,122].

apenas utilizando os *quadros*-chave, mas a técnica de agrupamento não está descrita.

O texto descreve que se assume que dois *quadros* k e $k + p_1$ pertencem ao mesmo contexto e então aplica uma medida de similaridade que, se ultrapassar um determiado limiar, será realmente considerado parte de um mesmo contexto. Mas no texto não está descrito qual objeto é utilizado para medida de similaridade.

A imagem de tomada obtida dos *quadros*-chave é dividida em blocos de HxW *pixels*. O significado de H e W não está mencionado. Se considerarmos estes valores como altura (*height*) e largura (*width*), da mesma forma não faz sentido unir os *quadros*-chave se depois eles serão divididos novamente em imagens de mesmo tamanho.

Em [RASHEED and Shah, 2003] descreve-se um sistema de detecção de limite entre cenas utilizando características como o movimento, o tamanho da tomada e propriedades de cor das tomadas de vídeos estruturados do tipo filmes.

O passo inicial do sistema descrito consiste em fazer a detecção das tomadas. Os quadros são representados por um histograma de cores HSV de 16 canais (8 canais para Matiz, 4 para Saturação e 4 canais pra Valor). Considerando f^x o x-ésimo quadro e H_x o seu histograma de cores, C é o conjunto de todos os canais do histograma, a intersecção entre os quadros x e $x + 1$, $D(f^x, f^{x+1})$ é dada por:

$$D(f^x, f^{x+1}) \sum_{b \ \epsilon \ C} min(H_x(b), H_x + 1(b)) \tag{3.2}$$

Uma troca de tomada é detectada quando o valor de D for menor do que um determinado limiar.

A seleção dos quadros-chave é feita da seguinte forma: o quadro do meio é adicionado ao conjunto (ainda vazio) dos quadros-chave. Após, cada quadro da tomada é comparado com todos os quadros. Se a intersecção D registrada para todos os quadros-chave for menor do que um limiar (diferente do limiar anterior), este frame será adicionado ao conjunto dos quadros-chave.

O algoritmo utiliza o tamanho da tomada e o movimento como características para o tipo de tomada. As informações de movimento já fazem parte do próprio código do arquivo MPEG. No momento da decodificação, obtém-se essas informações de movimento. Primeiro, um modelo global afim é estimado através da aplicação do método de mínimos quadrados sobre os vetores de movimento. Segundo, as velocidades dos blocos são reprojetadas. Terceiro, é feita uma comparação entre a velocidade atual e a velocidade reprojetada do bloco. Esta informação será utilizada como característica de movimento para a detecção de cenas.

Para a detecção de cenas é utilizada a informação de cor de cada tomada. Uma medida chamada Coerência anterior de tomada. Para todas as N tomadas identificadas calcula-se:

$$SC_i^j = max(D(f^x, f^y)) \tag{3.3}$$

para cada par de tomadas (i, j).

As características tamanho da tomada e movimento da tomada são utilizadas para prevenir a sobre segmentação nos casos de cenas com muita ação onde o algoritmo pode identificar seqüências de uma mesma cena como cenas separadas.

Uma medida chamada **Dinâmica de Cena** é calculada para relacionar tamanho da tomada e sua movimentação como segue:

$$SD_i = \frac{\sum\limits_{j \ \epsilon \ scene_i} SMC_j}{\sum\limits_{j \ \epsilon \ scene_i} L_i} \tag{3.4}$$

3.1 Discussão

As características cinemáticas são tratadas nos trabalhos como um conhecimento estrutural que permite adequar o sistema para uma melhor abordagem dos problemas em vista dos resultados esperados em cada trabalho. Os tipos específicos de imagens captados pelas câmeras, a duração das tomadas de acordo com o tipo de seqüência, a tentativa de representar o comportamento de um período de *replay*, todas essas características, na área em que cada método se propõe, são fruto de um conhecimento prévio do tipo de vídeo a ser manipulado e demonstram como o trabalho com vídeos estruturados pode ser orientado diretamente pelas características estruturais destes.

A representação dos dados nos arquivos compactados (com os quadros P, B e I) também podem ser utilizada nos métodos de manipulação de vídeos, pois informações referentes às diferenças entre as imagens da seqüência são utilizadas pelo processo de compactação das imagens e geração do vídeo.

Da mesma forma, a representação da imagens da seqüência separadamente, nos diversos formatos disponíveis, permite ressaltar características das imagens que não estão disponíveis no modo compactado. Parte-se sempre de uma melhor representação visual dos dados contidos, sem nenhuma perda de conteúdo (como no vídeo compactado),

assumindo-se o custo computacional necessário para desenvolver tais tarefas.

Dos trabalhos descritos nesta seção serão utilizadas várias técnicas no decorrer do trabalho.

A representação das imagens em formato HSV em histogramas de 16 níveis, e também a utilização de características de movimento (mesmo que através de outra representação) serão aplicadas neste trabalho.

Outras técnicas que não foram utilizadas neste trabalho terão a sua aplicabilidade discutida na parte final do trabalho, pois várias deste técnicas podem ser úteis a um sistema de caracterização de vídeos estruturados.

Capítulo 4

Método Proposto

O método em questão aplica algumas técnicas descritas nas seções anteriores e propõe uma nova abordagem para caracterizar e classificar as imagens contidas em um vídeo estruturado. Como principais características exploradas nesse método estão a diferença de cores em um espaço HSI (*Matiz, Saturação, Intesidade*) e as características de movimentação do vídeo.

Os vídeos estruturados são seqüências que possuem um formato tecnicamente definido e características peculiares como posicionamento de câmera, iluminação e cortes técnicos. Os vídeos possuem uma boa resolução e uma taxa de amostragem em torno de 25 quadros por segundo (qps).

O método implementado possui três partes definidas, a partir das imagens extraídas do vídeo. O segundo passo consiste em tomar essas imagens HSI quantizadas e calcular o histograma, dividindo a imagem em 20 partes iguais. Dessa forma, para cada par de imagens, suas partes serão comparadas considerando uma medida de Interseção de Cores e outra de Interseção de Movimento.

A terceira parte será uma classificação da diferença entre os quadros, baseado nos 20 valores de Cores e Movimento. Dessa forma os pares de imagens serão rotulados de acordo com a quantidade de movimentação existente, considerando quatro níveis para os pares de quadros.

A estrutura completa do método está mostrada na Figura 4.1, e a descrição completa de cada uma das partes mencionadas é feita nas seções seguintes.

4.1 Especificações Técnicas

Os recursos técnicos utilizados no estudo foram:

- Processador Athlon 1.662 MHz, 512 MB de memória RAM, 40 GB de Disco Rígido.

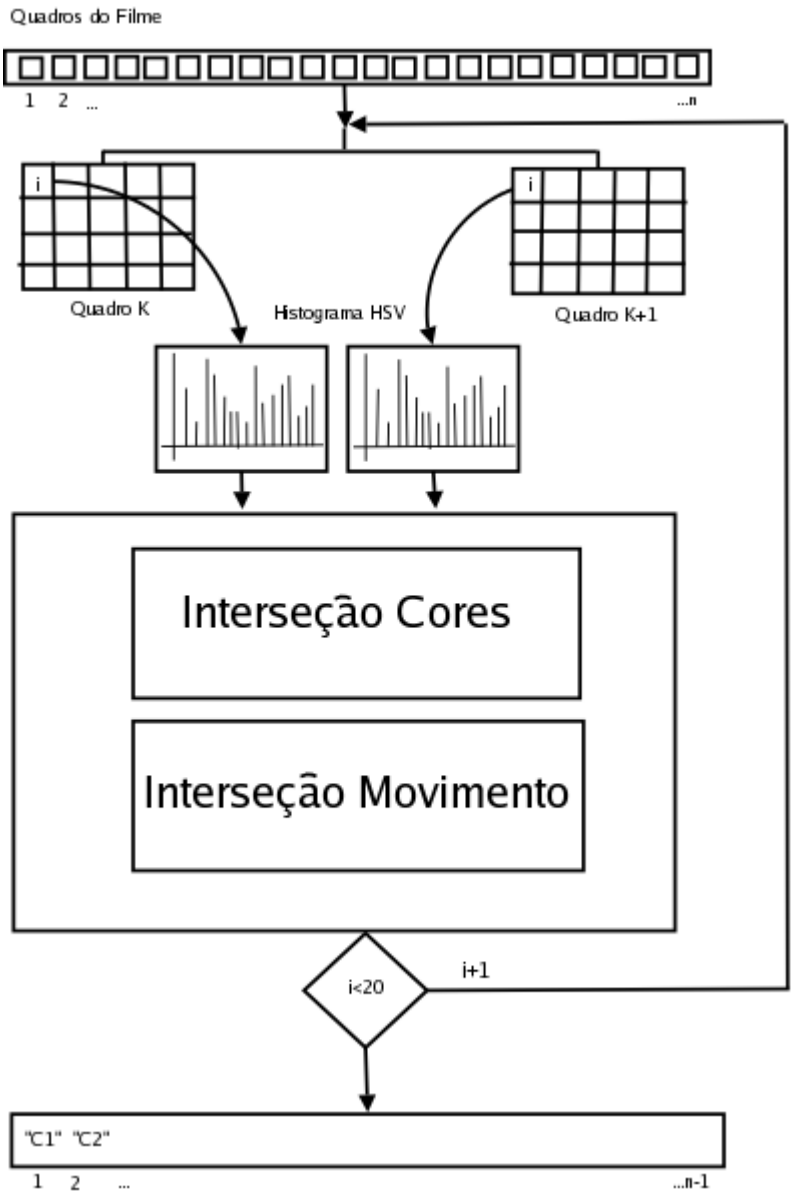


Figura 4.1: Diagrama do sistema

- Processador Athlon 1.466 Mhz, 256 MB de memória RAM, 36 GB de Disco Rígido.
- Máquina bi-processada Intel Pentium 4 2.600 MHz, 1,5 GB de memória RAM e 90 GB de Disco Rígido.

Os recursos foram utilizados para armazenagem e processamento dos filmes.

Para os processos de extração das imagens dos vídeos foi utilizada a ferramenta FFMPEG [FFMPEG, 2003] As operações desenvolvidas pelo sistema foram desenvolvidas em *scripts* feitos para a ferramenta de cálculos matemáticos Octave [OCTAVE, 2003].

4.2 Representação dos Quadros

De um vídeo em formato digital são extraídos os quadros que o compõem. As imagens extraídas de cada seqüencia são originalmente representadas em RGB.

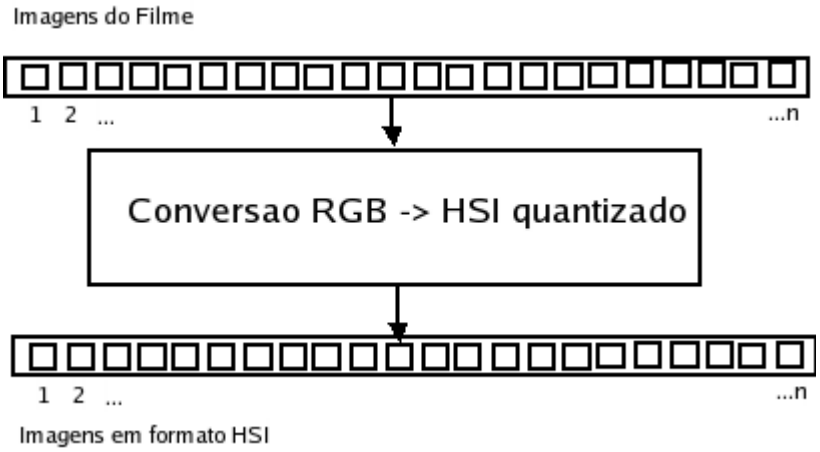


Figura 4.2: Primeira Parte do Sistema

Para obter uma resposta mais precisa referente à luminosidade contida em cada quadro, estes são convertidos do formato RGB para o formato HSI (Figura 4.2), de acordo com a definição:

$$H(i,j) = cos^{-1}\left[\frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right] \tag{4.1}$$

$$S(i,j) = 1 - \frac{3}{R+G+B}[min(R,G,B)] \tag{4.2}$$

$$I(i,j) = \frac{1}{3}(R+G+B) \tag{4.3}$$

Os coeficientes HSI contidos nas imagens serão quantizados para equalizar os valores contidos em cada um dos coeficientes. Dessa forma foram definindos 16 níveis de

representação para os canais de cor: 8 níveis para o canal H, 4 para S e 4 para I.

4.3 Histogramas HSI

Figura 4.3: Exemplo imagem dividida em 20 partes filme 'Cidade de Deus'

Partindo de um grupo de imagens em formato HSI quantizado em 16 níveis, o sistema irá carregar estas imgagens e fazer a sua comparação em tempo de execução. A Figura 4.8 mostra a estrutura desta parte do sistema.

As imagens serão divididas em 20 partes cada uma, e para cada parte será computado um histograma HSI de 16 canais. Dessa forma, as relações entre as imagens serão tomadas para cada par de subpartes. A Figura 4.3 representa uma imagem dividida em 20 partes.

As subpartes de cada par de imagens são tomadas par a par e seus histogramas, baseados nos coeficientes HSI quantizados, são extraídos. Cada histograma é formado através da associação das respostas para cada um dos canais HSI dos segmentos. As Figuras 4.4, 4.4 e 4.6 demonstram histogramas comums dos canais HSI extraídos da subparte número 7 (linha 2 coluna 2) da Figura 4.3, sem quantização. O histograma quantizado desta mesma subparte da imagem de exemplo é demonstrado na imagem 4.7. Desta forma podemos verificiar que a representação de HSI quantizado possibilita agrupar os três histogramas em um só devido a variabilidade dos valores estar em um mesmo conjunto.

Estes histogramas são comparados baseado em duas medidas.

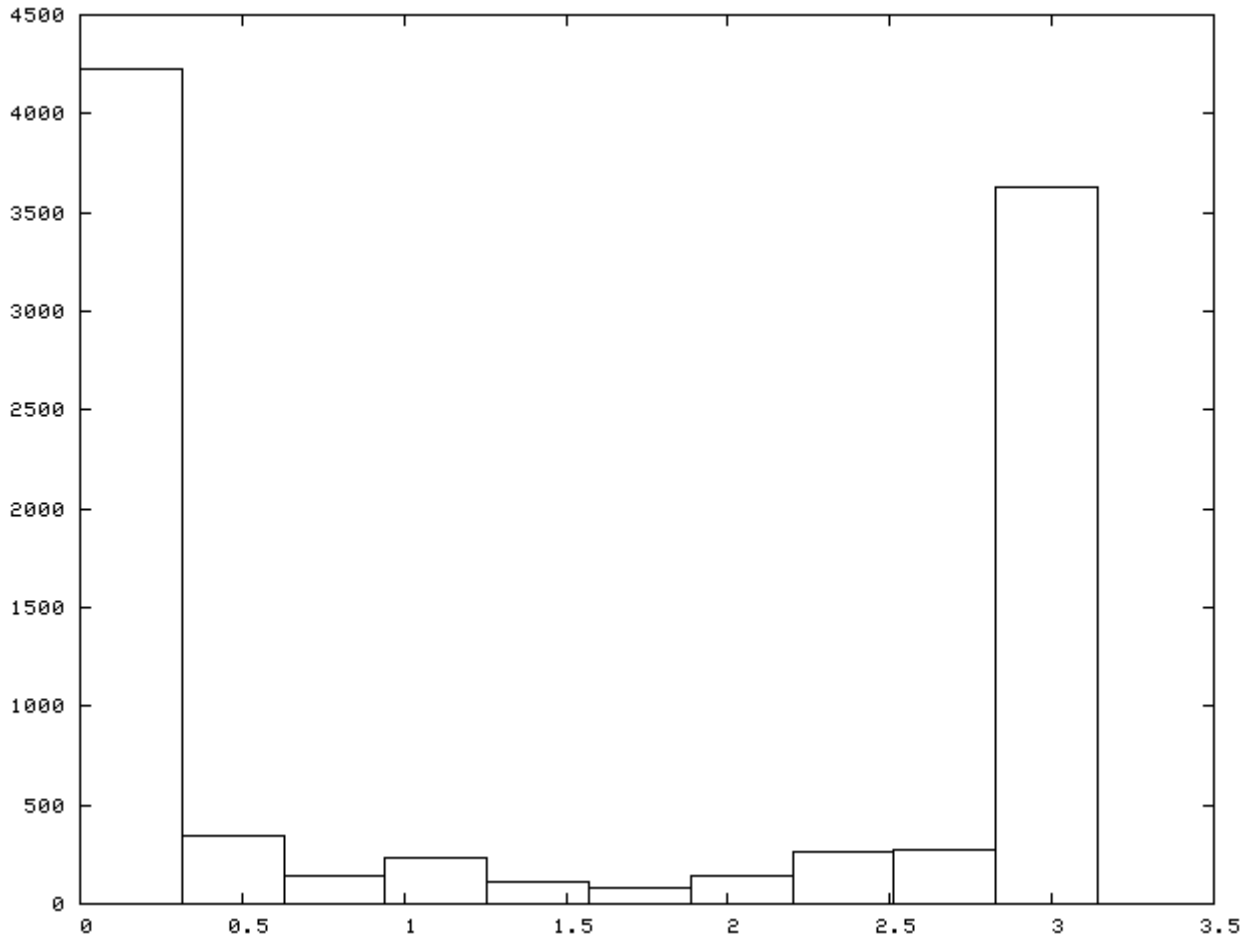


Figura 4.4: Exemplo de Histograma do Canal H extraído da subparte 7 da Figura 4.3

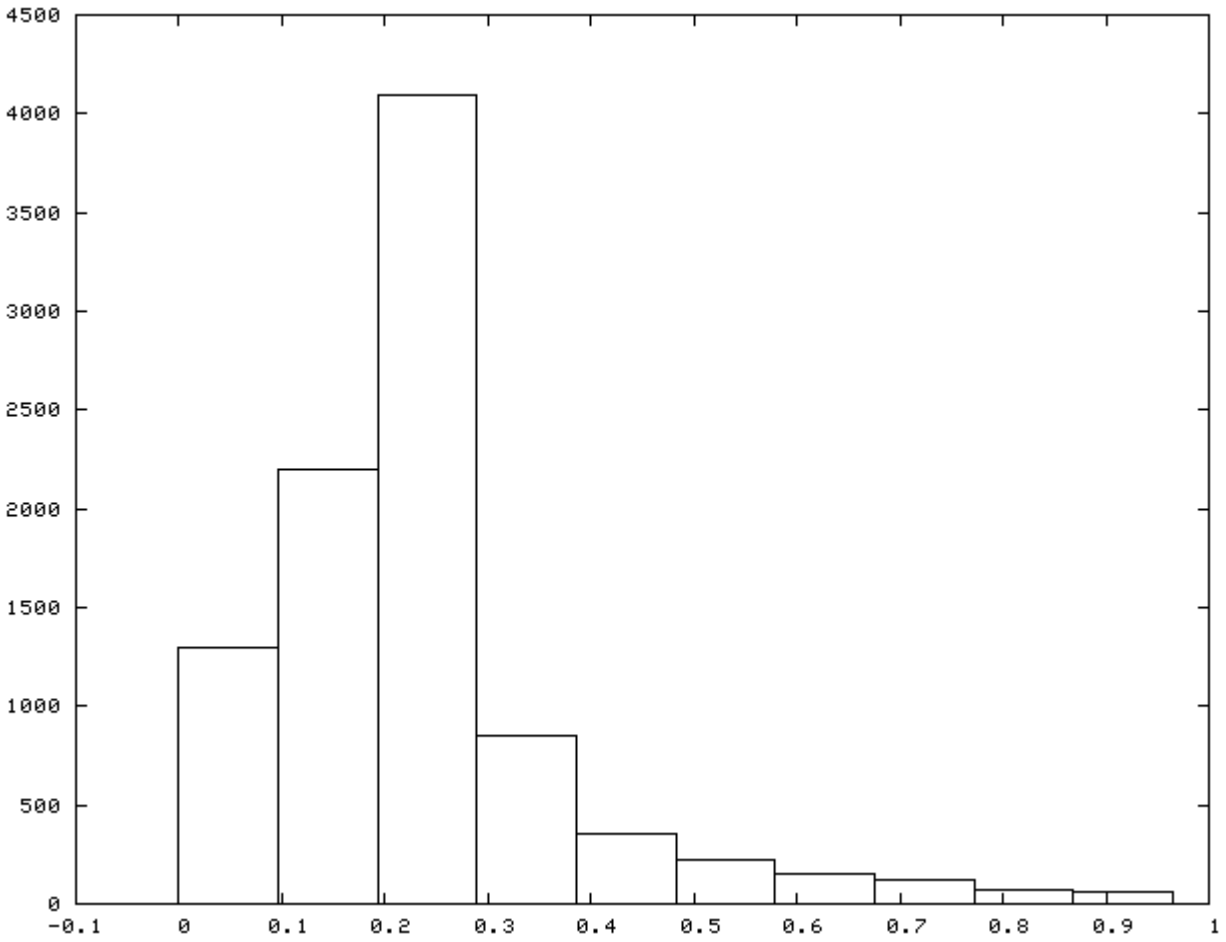


Figura 4.5: Histograma do Canal extraído da subparte 7 da Figura 4.3

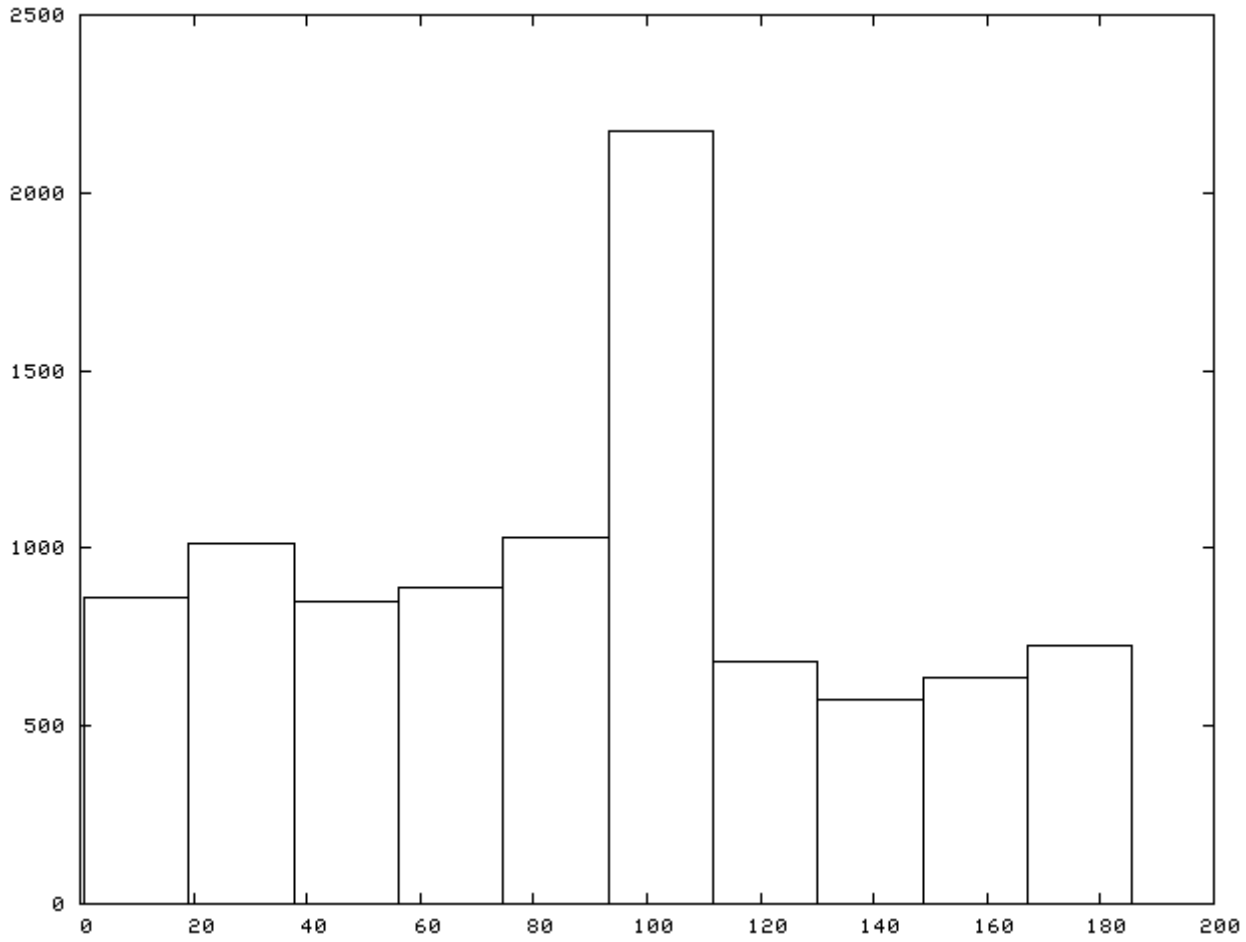


Figura 4.6: Histograma do Canal I extraído da subparte 7 da Figura 4.3

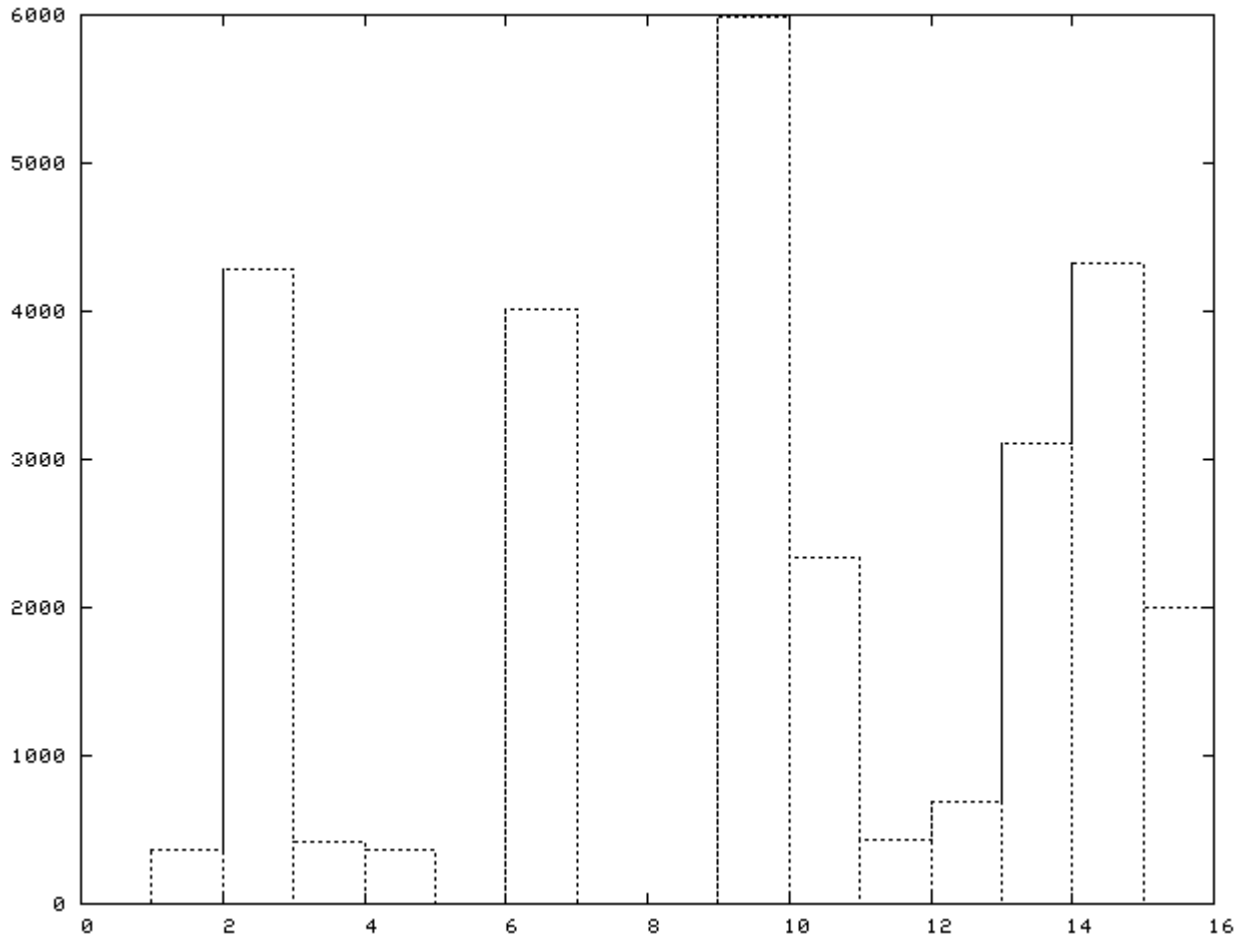


Figura 4.7: Histograma Quantizado dos canais HSI extraído da subparte 7 da Figura 4.3

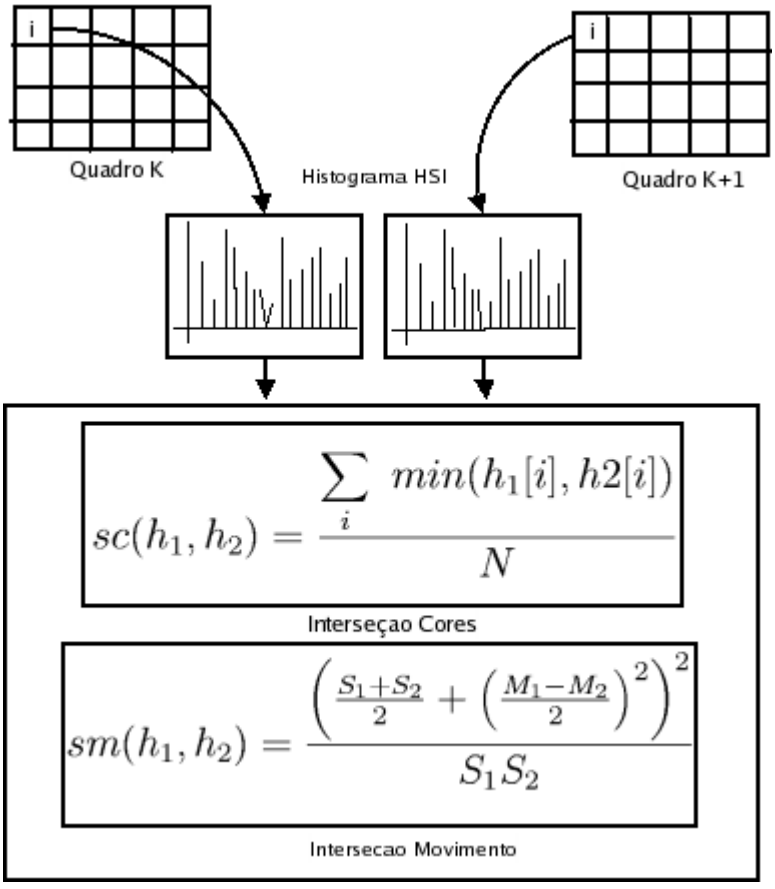


Figura 4.8: Segunda parte do Sistema

- Uma medida de Interseção de Cores, definida por:

$$sc(h_1, h_2) = \frac{\sum_i min(h_1[i], h2[i])}{N} \tag{4.4}$$

onde h_1 e h_2 representam os histogramas da primeira e da segunda imagem respectivamente.

Esta função terá como valor máximo 1; quanto mais próximo do máximo esse valor estiver, mais semelhança têm as imagens entre si.

- A outra medida utilizada é a de Interseção de Movimento, que consiste no cálculo da Máxima Verossimilhança entre os histogramas, dada por:

$$sm(h_1, h_2) = \frac{\left(\frac{S_1+S_2}{2} + \left(\frac{M_1-M_2}{2}\right)^2\right)^2}{S_1 S_2} \tag{4.5}$$

onde M_1 e M_2, S_1 e S_2 representam a média e o desvio-padrão da primeira e da segunda subparte respectivamente.

Esta função tem como valor mínimo 0, que quando ocorre significa que as subpartes não possuem movimentação entre si.

Vamos chamar de $SC(sc1, sc2, sc3, ..., sc20)$ os valores de interseção para cada subparte do par e $SM(sm1, sm2, sm3, ..., sm20)$ os valores de verossimilhança. Estes somatórios terão valores máximo de 20 para cores e mínimo de 0 para movimento.

4.4 Caracterização

Um par de imagens possui um conjunto de 20 valores de intereseção de cores e 20 valores de intereseção de movimento. Cada um desses grupos de 20 coeficientes é somado, gerando dois valores representantes de cores e movimento para cada par de imagens.

Dessa forma, os dados foram analisados levando em consideração os seus valores representativos e a variabilidade possível dentro destes valores. Essa análise deu origem a uma caracterização dos pares de imagens definindo níveis de diferença. Assim, cada par terá um rótulo relacionado a quantidade de diferença entre as imagens.

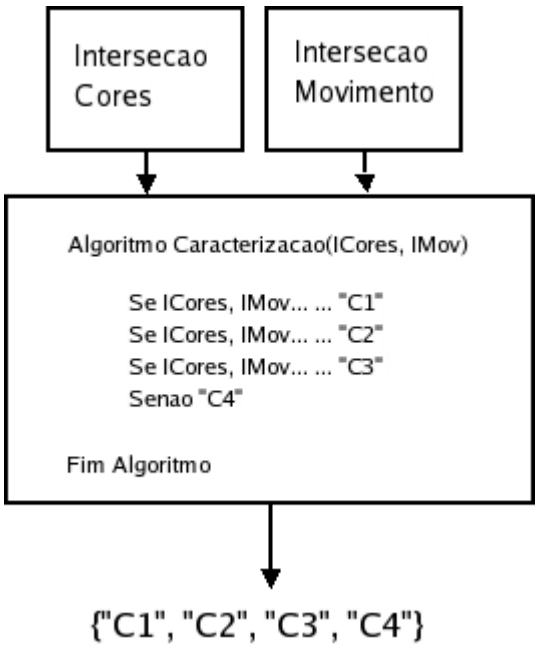


Figura 4.9: Terceira parte do Sistema

A classificação dos pares de tomadas considera quatro valores possíveis estados para os pares de tomadas: C1, C2, C3 e C4 (Figura 4.9). Estes níveis definem a quantidade de ação entre o par e são dados através de SC e SM, indo da menor diferença em C1 para a maior diferença em C4.

Essa rotulação da base gera um vetor de rótulos de tamanho $n - 1$ onde n é a quantidade de imagens contidas na seqüência. Este vetor vai representar toda a seqüência do vídeo estruturado, permitindo agrupar ou separar partes do vídeo de acordo com a quantidade de movimentação.

Baseado na variabilidade dos dados, e feitas análises do comportamento dos dados contidos em diversas seqüências de vídeos foram definidos limiares de separação entre os níveis diferentes de movimentação contidos em um vídeo estruturado.

Os rótulos do tipo $C1$ caracterizam a mínima diferença entre o par de imagem. Suas respostas tendem ao valor mínimo em SM e ao valor máximo em SC.

Os rótulos de $C2$ e $C3$ caracterizam os níveis intermediários de ação contidos em um par de imagens. Eles representam as movimentações ocorridas em uma mesma tomada que contenham ação em nível intermediário.

A representação dos rótulos do tipo $C4$ buscam representar imagens com alto nível de divergência, possivelmente imagens que pertencem a tomadas diferentes, ou que sofrem um tipo de corte brusco, como os definidos na Seção 2.2.

Para separar devidamente os valores de SC e SM foram definidos quatro valores de limiares em de interseção de cores (LC) e quatro para interseção de movimento (LM),

onde $LC = \{LC1, LC2, LC3, LC4\}$ e $LM = \{LM1, LM2, LM3, LM4\}$. Estes valores serão aplicados nos testes de ground-truth, descritos na seção de Testes.

Dado que o valor máximo de SC seja $MAX(SC)$ e o valor mínimo de SM seja $MIN(SM)$, a variação dos limiares para os rótulos é dada da seguinte forma:

- C1: MAX(SC) $\geq$ SC $>$ $LC1$ e MIN(SM) $\leq$ SM $<$ $LM1$
- C2: $LC1 \geq$ SC $>$ $LC2$ e $LM1 \leq$ SM $<$ $LM2$
- C3: $LC2 \geq$ SC $>$ $LC3$ e $LM2 \leq$ SM $<$ $LM3$
- C4: $LC3 \geq$ SC $>$ $LC4$ e $LM4 \leq$ SM $<$ $LM4$

Este vetor passa por um processo de *ground-truth* com o vetor de rótulos identificados previamente, avaliando os resultados através do método qualitativo de Precisão-Recobrimento [GARGI et al., 2000]. Defina-se λ como Precisão e φ com Recobrimento através de:

$$\lambda = \frac{D}{D + DE} \tag{4.6}$$

$$\varphi = \frac{D}{D + AF} \tag{4.7}$$

onde D é a quantidade de rotulações corretas, DE é a quantidade de detecções errôneas (casos onde o sistema não identifica um determinado elemento do grupo) e AF representa os alarmes falsos (casos onde o sistema classifica com um determinado elemento que não condiz com o rótulo previamente definido).

Cada seqüência avaliada terá quatro valores para *Precisão* e quatro para *Recobrimento*, sendo um para cada possível rótulo do par de imagem. Assim pode-se traçar uma curva com o desempenho do método para cada nível de movimentação das imagens, a medida que LC e LM variam.

Capítulo 5

Experimentos

A fase de experimentos foi composta por uma série de passos. A primeira de todas foi a parte de seleção dos filmes para compor a base de dados. Em seguida foram aplicadas as técnicas descritas no Capítulo 4.

Foram escolhidos dois filmes, e para cada um destes filmes foram separados dois conjuntos de testes: um com imagens rotuladas manualmente e outro sem rotulaçãoo. Cada um destes grupos contémm um total de 10.000 imagens. O conjunto de imagens rotuladas foi dividido em duas partes neste trabalho para facilitar a descrição e mostração dos detalhes do trabalho.

Esta Seção começa com uma descrição dos filmes escolhidos para a base de dados e o processo que definiu essa escolha. Em seguida, os processos descritos na Seção Método Proposto são revistos, levando em consideração o seu comportamento aplicado aos filmes escolhidos. Para finalizar, uma análise dos resultados obtidos, assim como comentários e apontamentos.

5.1 Base de Dados

Para o processo de escolha dos filmes, foram analisados vários títulos, levando em consideração a sua relevância para o trabalho. Esta relevância está relacionada com a diversidade das cenas contidas nos filmes, características cinemáticas que possam definir o filme, e à variabilidade da dinâmica das seqüências e movimentação de câmera contidas neste. Para este trabalho foram escolhidos dois filmes.

Os filmes utilizados foram *"Matrix Reloaded"* [WARNER, 2003], *"Cidade de Deus"* [MIRAMAX, 2003]. Cada um destes possui características visuais peculiares como mostrado a figura 5.1.

Cada um destes filmes possui mais de 100 mil quadros.

(a) Quadro de “Matrix Reloaded” 480X208 (b) Quadro de “Cidade de Deus” 592X320

Figura 5.1: Exemplos de Imagens da base de dados.

Uma das etapas dos experimentos realizados foi a análise e rotulação manual do filme, que consistiu na avaliação quadro a quadro de cada grupo de 10.000 imagens dos filmes escolhidos. Cada par avaliado obteve um rótulo entre os C1, C2, C3 ou C4 previamente descritos neste trabalho. Posteriormente esses rótulos foram comparados com os rótulos obtidos pelo sistema. Os resultados obtidos com essa aplicação (*ground-truth*) serão descritos nas próximas seções.

5.1.1 Detalhes técnicos dos filmes

Os vídeos foram obtidos em formato compactado e suas imagens foram extraídas para arquivos do tipo JPG. A Tabela 5.1 descreve os detalhes principais detalhes técnicos de cada um dos filmes, como tipo do arquivo, quantidade de QPS (*Quadros por Segundo*), resolução e duraçãoo dos filmes.

Tabela 5.1: Especificações técnicas dos filmes

Nome	QPS	Tipo Arquivo	Resolução	Duração
“Matrix Reloaded”	25	MPEG-4 XVid	480x208	2:06:44
“Cidade de Deus”	25	MPEG-4 XVid	592x320	2:04:09

5.1.2 “Matrix Reloaded”

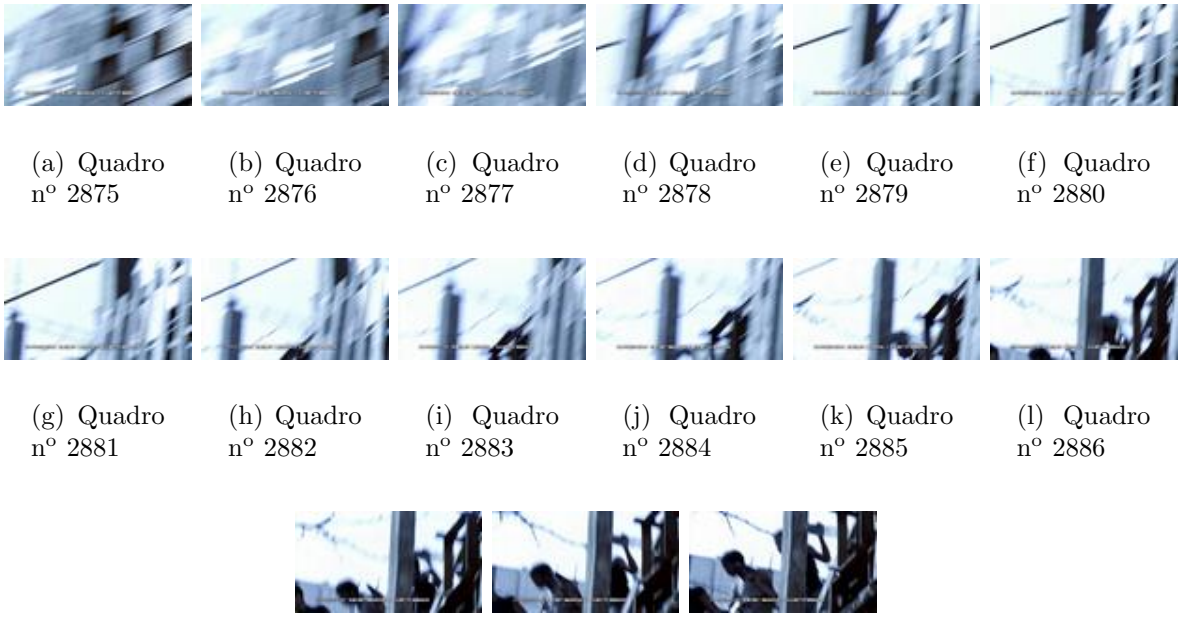
Este filme possui como uma de suas principais características a baixa iluminação. A maior parte das cenas é capturada em ambientes fechados ou à noite, o que explica a pouca quantidade de luz presente na seqüência. Existem também muitas seqüências de ação e diálogo, tornando o filme bem variado em tipos de cena. Esse fator torna impactante o aparecimento repentino de elementos como pessoas, por exemplo. A figura

(a) Quadro nº 4325 (b) Quadro nº 9680 (c) Quadro nº 9744 (d) Quadro nº 13807 (e) Quadro nº 15328

Figura 5.2: Exemplo de seqüência “Matrix Reloaded”

5.1.2 mostra imagens peculiares do filme.

5.1.3 “Cidade de Deus”



(a) Quadro nº 2875 (b) Quadro nº 2876 (c) Quadro nº 2877 (d) Quadro nº 2878 (e) Quadro nº 2879 (f) Quadro nº 2880

(g) Quadro nº 2881 (h) Quadro nº 2882 (i) Quadro nº 2883 (j) Quadro nº 2884 (k) Quadro nº 2885 (l) Quadro nº 2886

(m) Quadro nº 2887 (n) Quadro nº 2888 (o) Quadro nº 2889

Figura 5.3: Exemplo de seqüência “Cidade de Deus”

No caso deste filme, as cores utilizadas foram mais claras do que as aplicadas no filme *“Matrix Reloaded”*. A dinâmica das cenas resulta em tomadas menores e a captação de cenas com câmera na mão proporciona grande movimentação. A figura 5.1.3 mostra um exemplo de uma tomada do filme.

5.2 Conversão RGB para HSI

Em sua totalidade, cada um dos filmes conta com mais de 100 mil quadros. Como os grupos definidos para este trabalhos foram de 20 mil quadros cada, foram escolhidos os 20 mil primeiros quadros de cada filme para o procedimento de testes.

O processo de conversão de RGB para HSI quantizado foi implementado através de scripts para o ambiente Octave e demanda um tempo de em média 40 segundos por imagem. Para um total de 20 mil imagens, como foram definidos os grupos de teste, o tempo de conversão das imagens da base foi em torno de 22 horas.

Cada arquivo salvo pelo método tem um tamanho médio de 4.000 KB, demandando um espaço em disco da ordem de 78 GB por grupo de imagens.

Estas informações foram obtidas com os recursos técnicos descritos na Seção 4.1.

5.3 Comparação de imagens

Para a comparação dos histogramas HSI quantizados, as imagens armazenadas da seqüência são carregadas par a par e o processo de extração de SC e SM é feito.

Para cada par de imagens, o tempo de processamento desta parte do método é de em torno de 2 minutos. Para um total de 20 mil imagens esse tempo de processamento eh da ordem de 500 horas.

Durante essa comparação é salvo um arquivo para cada par de imagens, contendo dos valores de sc e sm (dois vetores de 20 valores cada). Estes arquivos consomem um espaço em disco rígido em torno de 160KB cada. Estes arquivos salvos são a representação das diferencas entre os quadros do filme e os valores serão avaliados nos processos posteriores.

5.4 Extração de Características

Nesta Seção serão descritos os resultados obtidos para os grupos de 5.000 imagens de cada base.

Os arquivos referentes a cada par de imagens são carregados e avaliados segundo as regras definidas na Seção 4.4. A partir destes critérios obtém-se um vetor de tamanho $n - 1$ onde n é a quantidade de imagens contidas em cada seqüência. Este vetor contem os rótulos para todos os pares de quadros das imagens.

Em seguida, esse arquivo é comparado com o arquivo dos pré-rotulados para obter Precisão e Recobrimento para cada um dos rótulos.

De cada grupo obtém-se um vetor de 4.999 rótulos, representando cada um dos pares de imagens da base testados.

5.4.1 Experimentos com o filme "Matrix Reloaded"

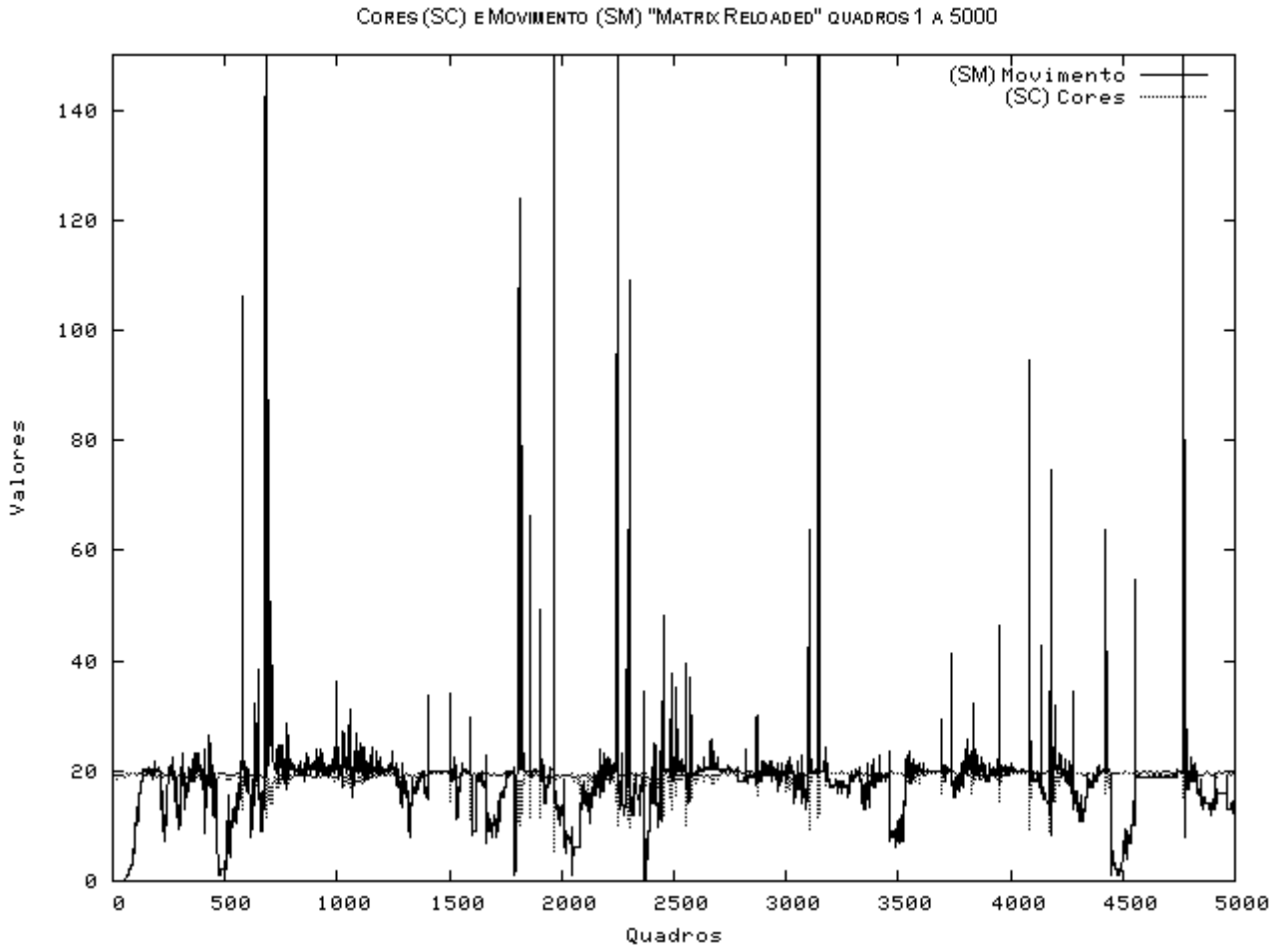


Figura 5.4: Gráfico Cores e Movimento filme "Matrix Reloaded" quadros 1 a 5.000

A figura 5.4 mostra os valores de SC e SM para os primeiros 5.000 quadros do filme. Os vales no vetor de cores representam maior diferença entre os valores de SC das imagens, da mesma forma acontece com os picos do vetor dos valores de movimento.

A figura 5.5 mostra as variações de SC e SM para o segundo grupo de 5.000 imagens.

5.4.2 Experimentos com o filme "Cidade de Deus"

Os valores para SC e SM estão representados na figura 5.6. Da mesma forma que no filme *"Matrix Reloaded"* os vales no vetor de cores representam maior diferença entre os quadros assim como os picos no vetor de movimento.

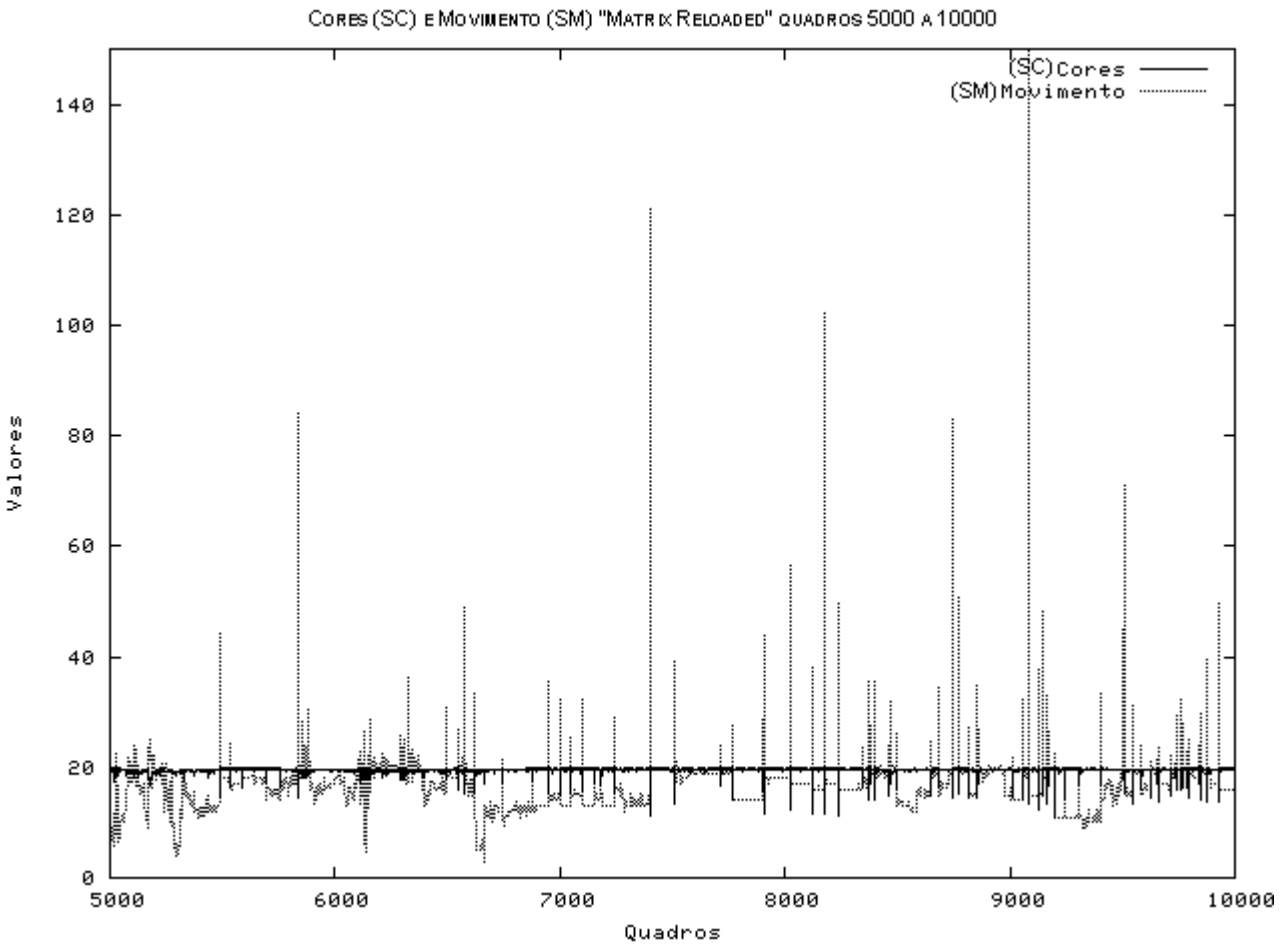


Figura 5.5: Gráfico Cor e Movimento filme “Matrix Reloaded” quadros 5000 a 10000

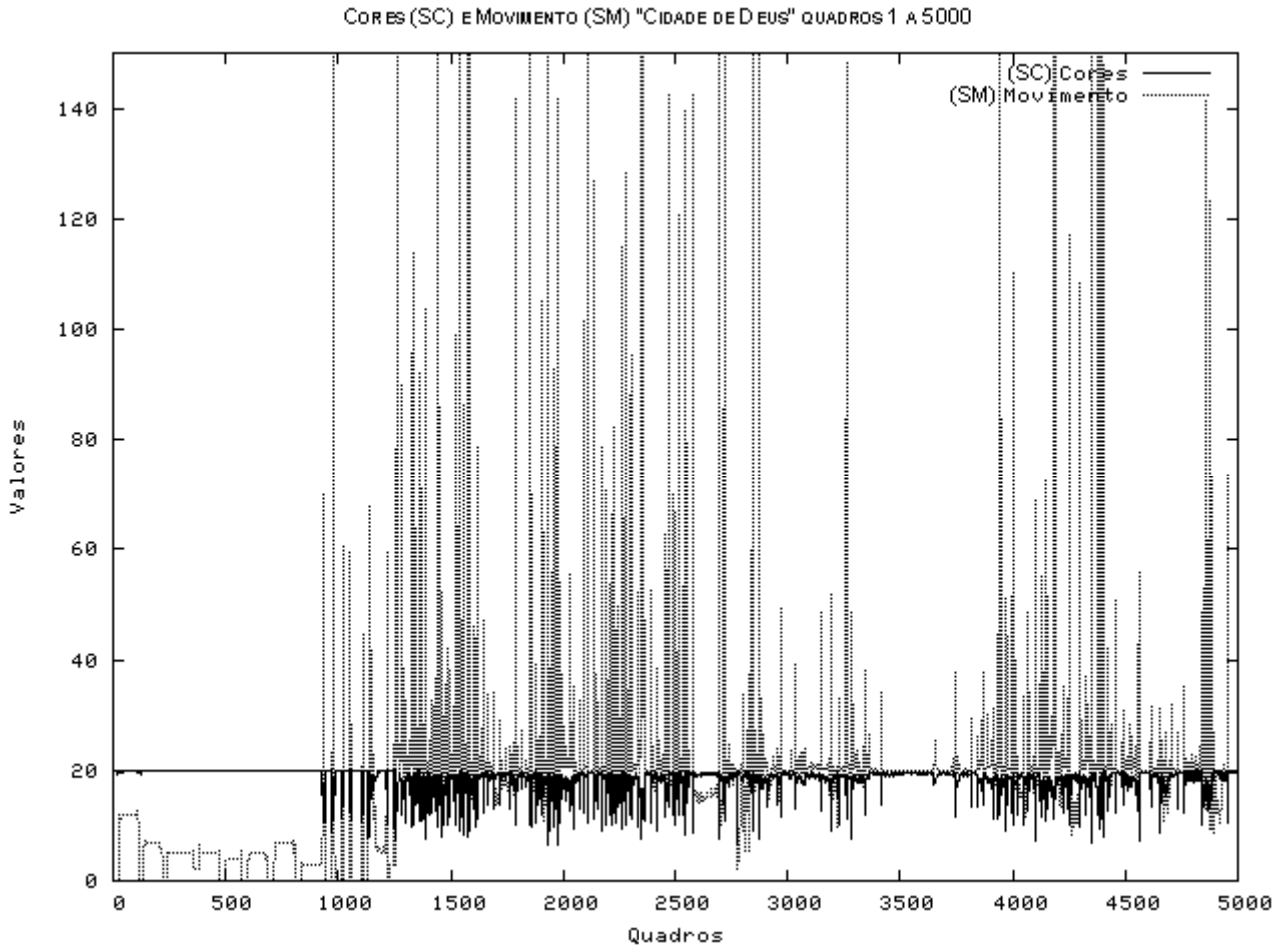


Figura 5.6: Gráfico Cor e Movimento filme “Cidade de Deus” quadros 1 a 5000

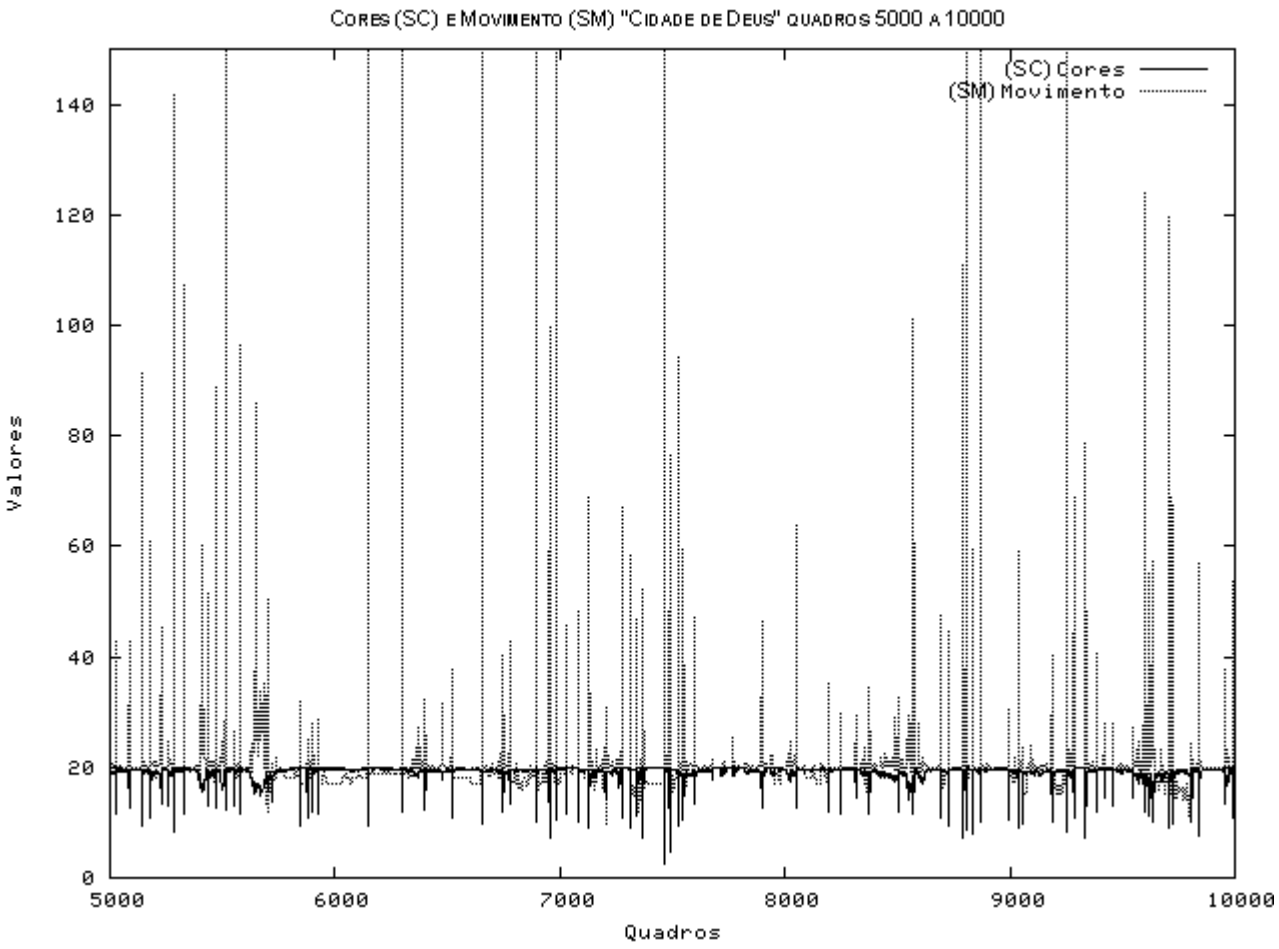


Figura 5.7: Gráfico Cor-Movimento filme “Cidade de Deus” quadros 5000 a 10000

Os valores de SC e SM para o segundo grupo de 5.000 imagens para este filme está mostrado na figura 5.7.

5.5 Rotulação

A aplicação do método sobre a base rotulada de 20 mil imagens, permite obter uma medida de avaliação qualitativa do método através das curvas de Precisão e Recobrimento.

O processo de rotulação manual das tomadas consistiu em uma análise detalhada da variação de cada par de quadros em 10 mil imagens de cada um dos filmes. Essa avaliação é subjetiva e serviu de base aos limiares arbitrários que cercearam a definição das métricas de rotulação da base através do sistema desenvolvido.

Aplicado o processo de *ground-truth* aos resultados obtidos do sistema aos grupos de 5.000 imagens para cada filme, variam-se os limiares de caracterização dos níveis (LC e LM) de ação e dessa forma podemos extrair a curva de Precisão e Recobrimento a medida que estes valores variam.

Como são quatro níveis de diferença entre os pares de imagens, obtém-se uma curva para cada rótulo.

Os valores de LC e LM foram projetados em cinco possíveis valores para a geração das curvas. Estes valores são mostrados nas Tabelas 5.2 e 5.3, respectivamente.

Representativamente, cada grupo de limiares será representado com um número e estes números irão compor o eixo das coordenadas nos gráficos que se seguem. Estes valores foram aplicados para cada um dos limiares $LC = \{LC1, LC2, LC3, LC4\}$ e $LM = \{LM1, LM2, LM3, LM4\}$ utilizando as definições descritas na Seção 4.4.

Tabela 5.2: Variação dos Limiares LC

Grupo	1	2	3	4	5	6	7	8
LC1	17	16	19	19	17	17	17	17
LC2	17	16	19	16	15	16	16	16
LC3	12	12	14	14	12	13	13	13
LC4	12	12	14	12	8	13	13	13

Tabela 5.3: Variação dos Limiares LM

Grupo	1	2	3	4	5	6	7	8
LM1	21	22	23	23	20	20	21	21
LM2	23	24	25	25	23	30	30	35
LM3	30	31	32	32	25	40	35	45
LM4	50	51	52	52	52	60	35	45

Nesta parte, estão os gráficos e tabelas referentes ao testes aplicados aos grupos de limiares previamente descritos. Cada gráfico possui a sua tabela de dados correspondente mostrada.

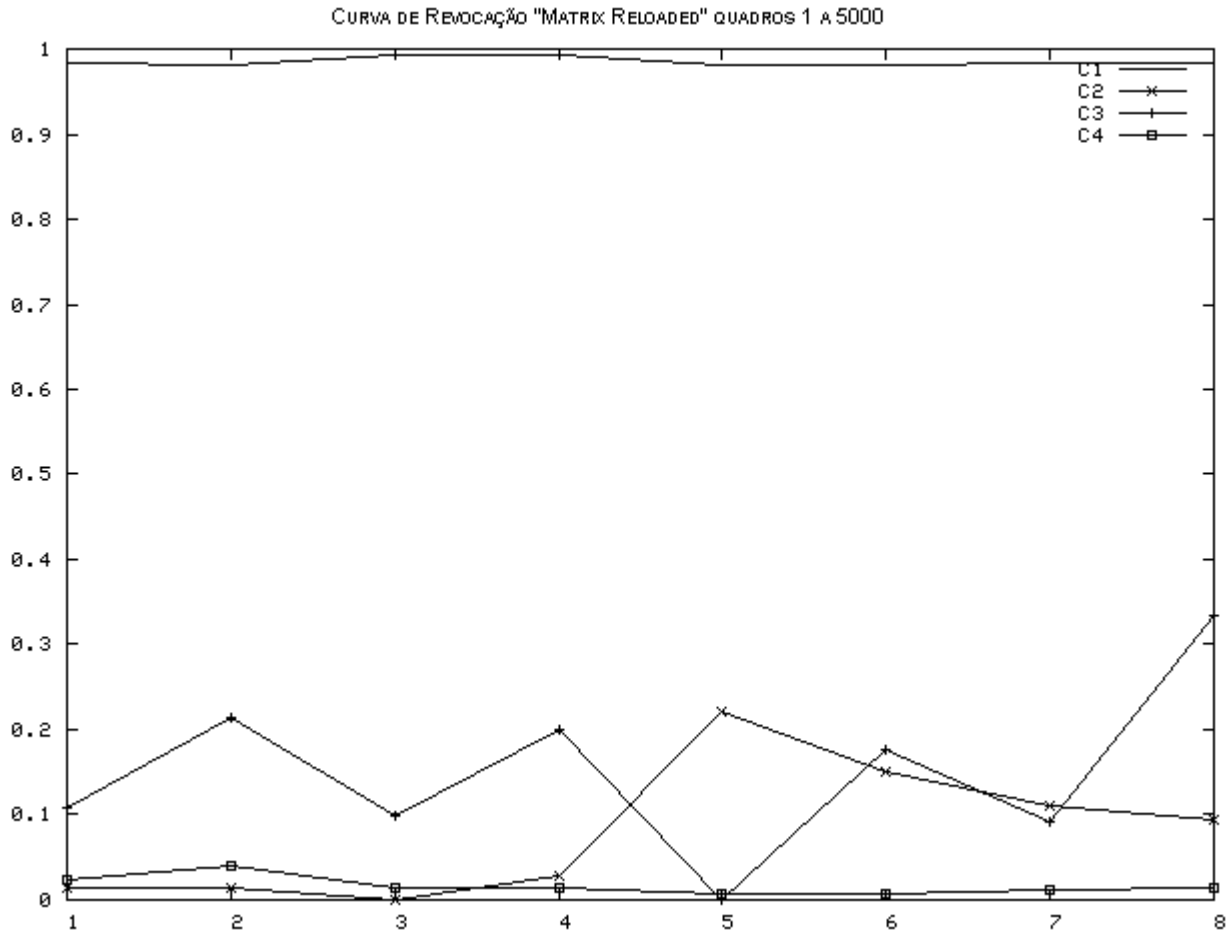


Figura 5.8: Curva Recobrimento "Matrix Reloaded" primeiro grupo de 5.000 imagens

Baseado nos limiares definidos na Tabela 5.2 e 5.3 a figura 5.8 mostra a curva de Recobrimento para o primeiro grupo de 5.000 imagens do filme *"Matrix Reloaded"* e a figura 5.9 mostra a distribuição de Precisão do mesmo. Os dados contidos na figura 5.8 estão mostrados na Tabela 5.4 e os dados referentes ao gráfico de Precisão estão descritos na Tabela 5.5.

O segundo grupo de 5.000 imagens tem suas curvas de Precisão e Recobrimento mostrados nas figuras 5.10 e 5.11 respectivamente. Seus dados estão dispostos nas tabelas 5.6 e 5.7.

As 5.000 primeiras imagens têm a sua curva Recobrimento mostrada na figura 5.12 e a figura 5.13 mostra os valores de Precisão do mesmo período, com os dados mostrados nas Tabelas 5.9 e 5.8. As curvas de Precisão e Recobrimento para o segundo grupo de 5.000 imagens do filmes estão mostrados nas figuras 5.14 e 5.15 e seus dados dispostos

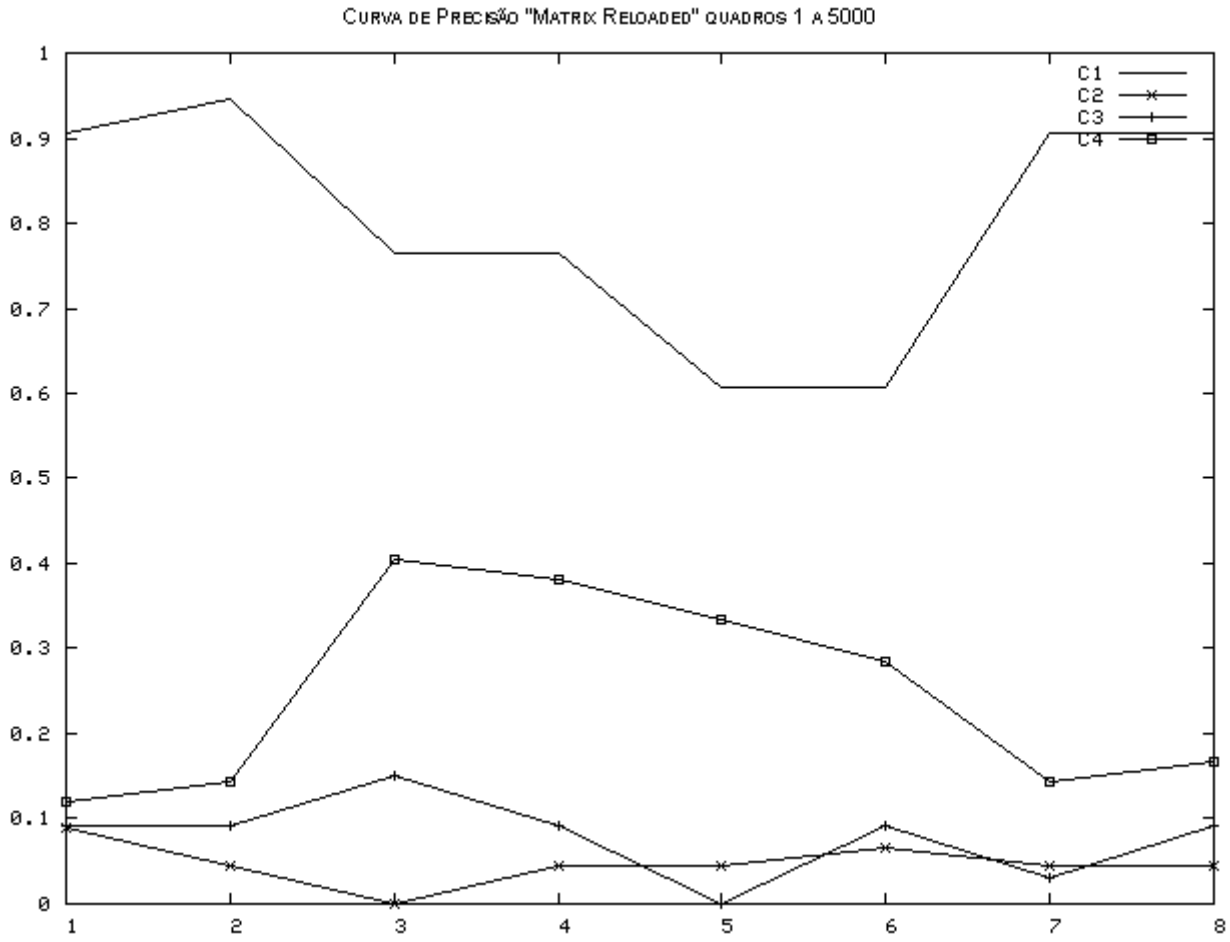


Figura 5.9: Curva Precisão "Matrix Reloaded" primeiro grupo de 5.000 imagens

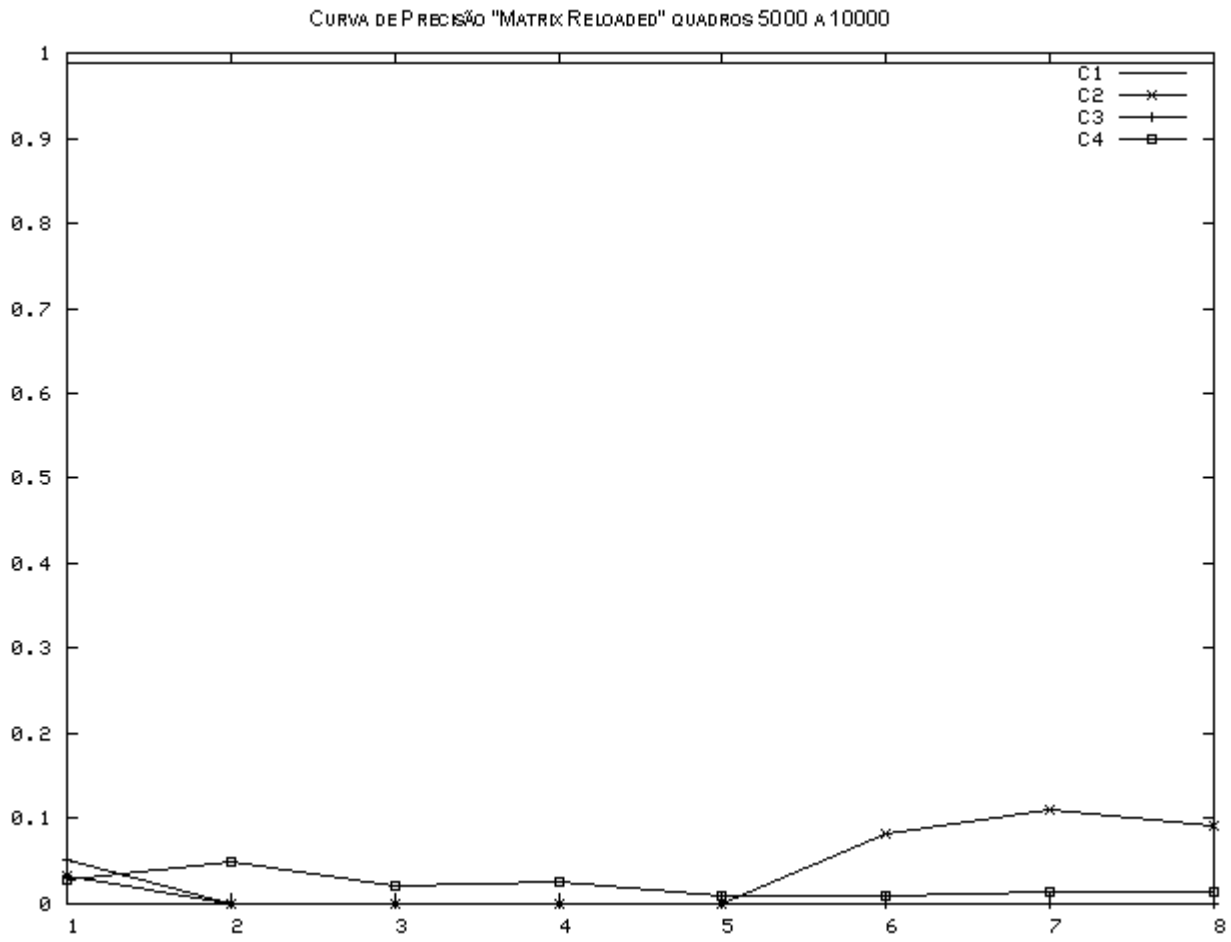


Figura 5.10: Curva Precisão “Matrix Reloaded” segundo grupo de 5000 imagens

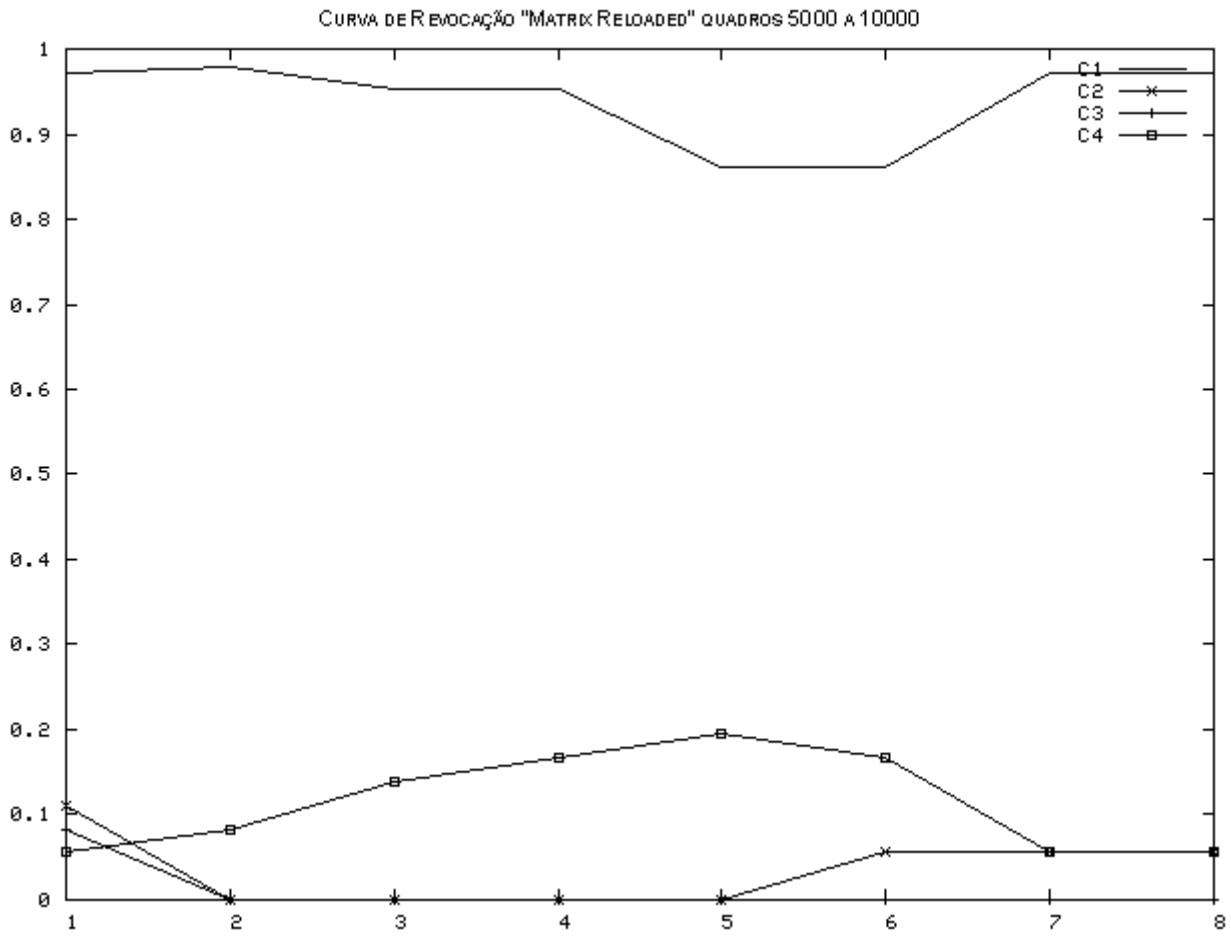


Figura 5.11: Curva Recobrimento “Matrix Reloaded” segundo grupo de 5000 imagens

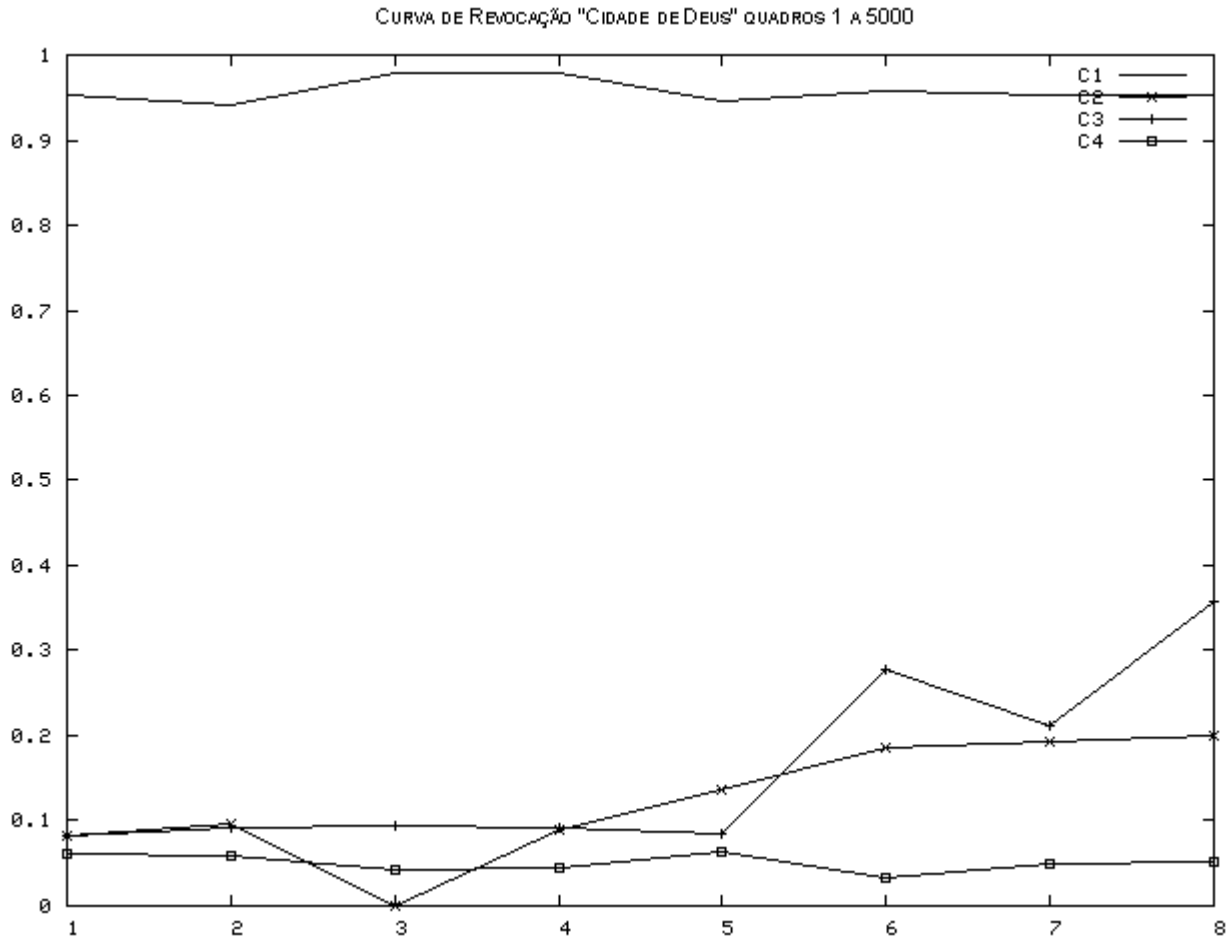


Figura 5.12: Curva Recobrimento “Cidade de Deus” primeiro grupo de 5.000 imagens

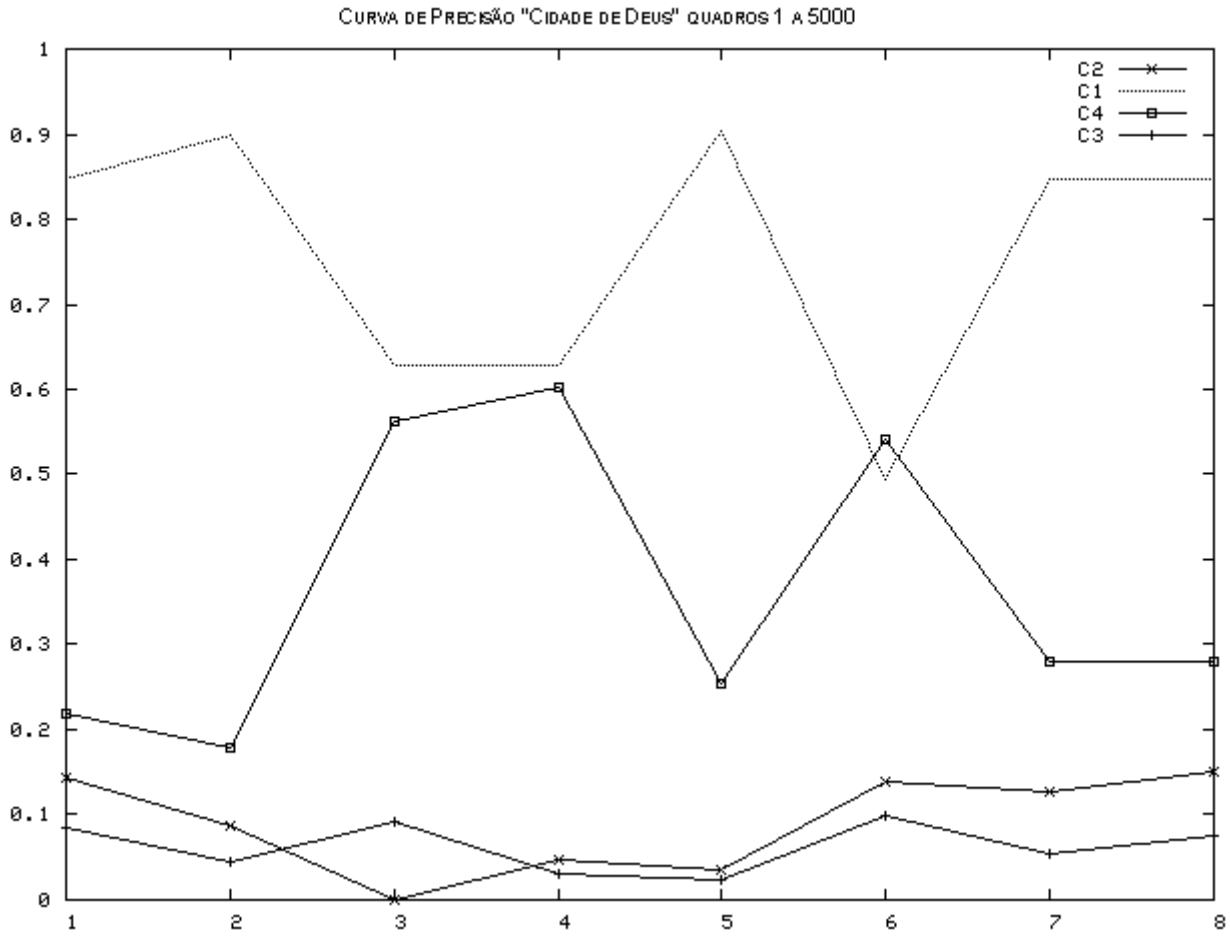


Figura 5.13: Curva Precisão "Cidade de Deus" primeiro grupo de 5.000 imagens

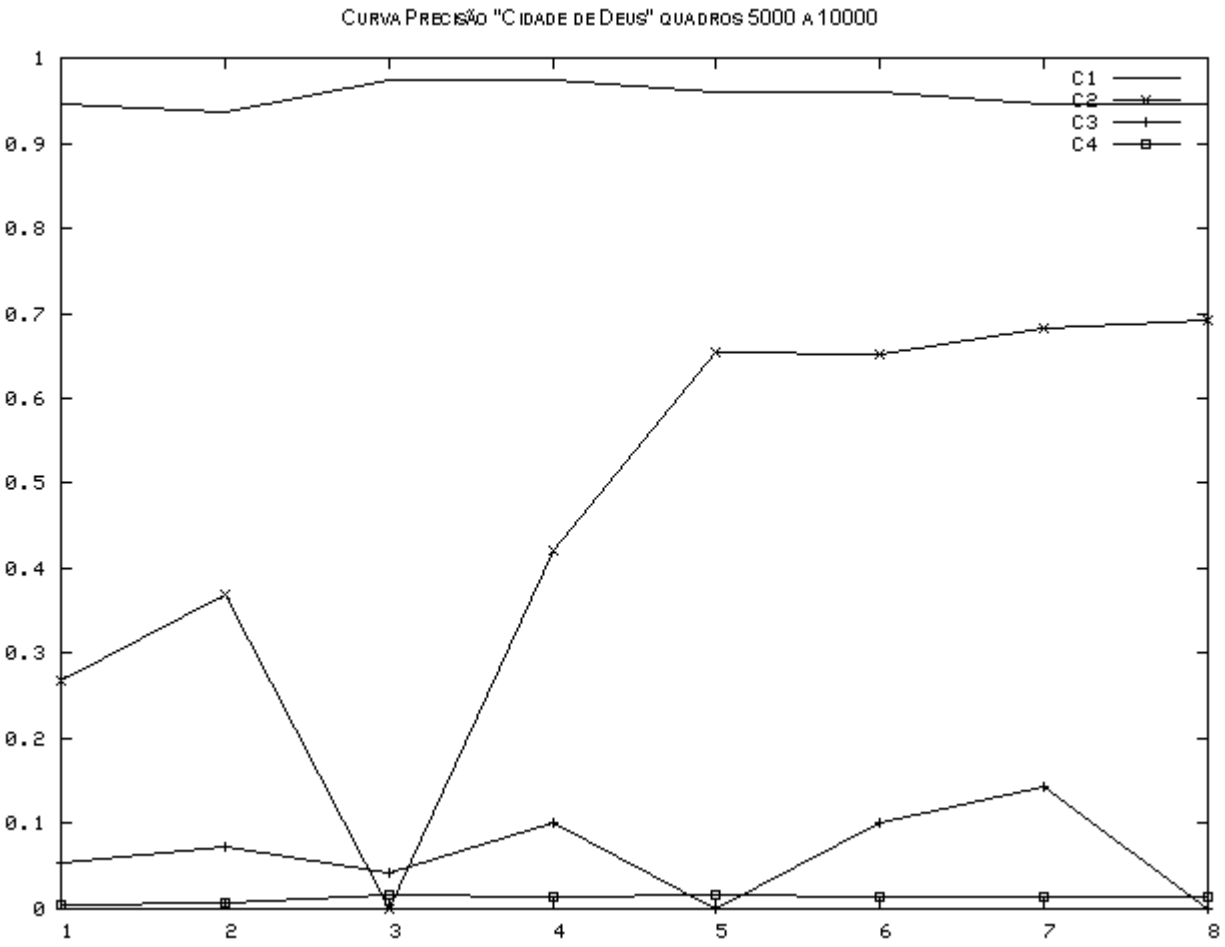


Figura 5.14: Curva de Precisão “Cidade de Deus” segundo grupo de 5.000 imagens

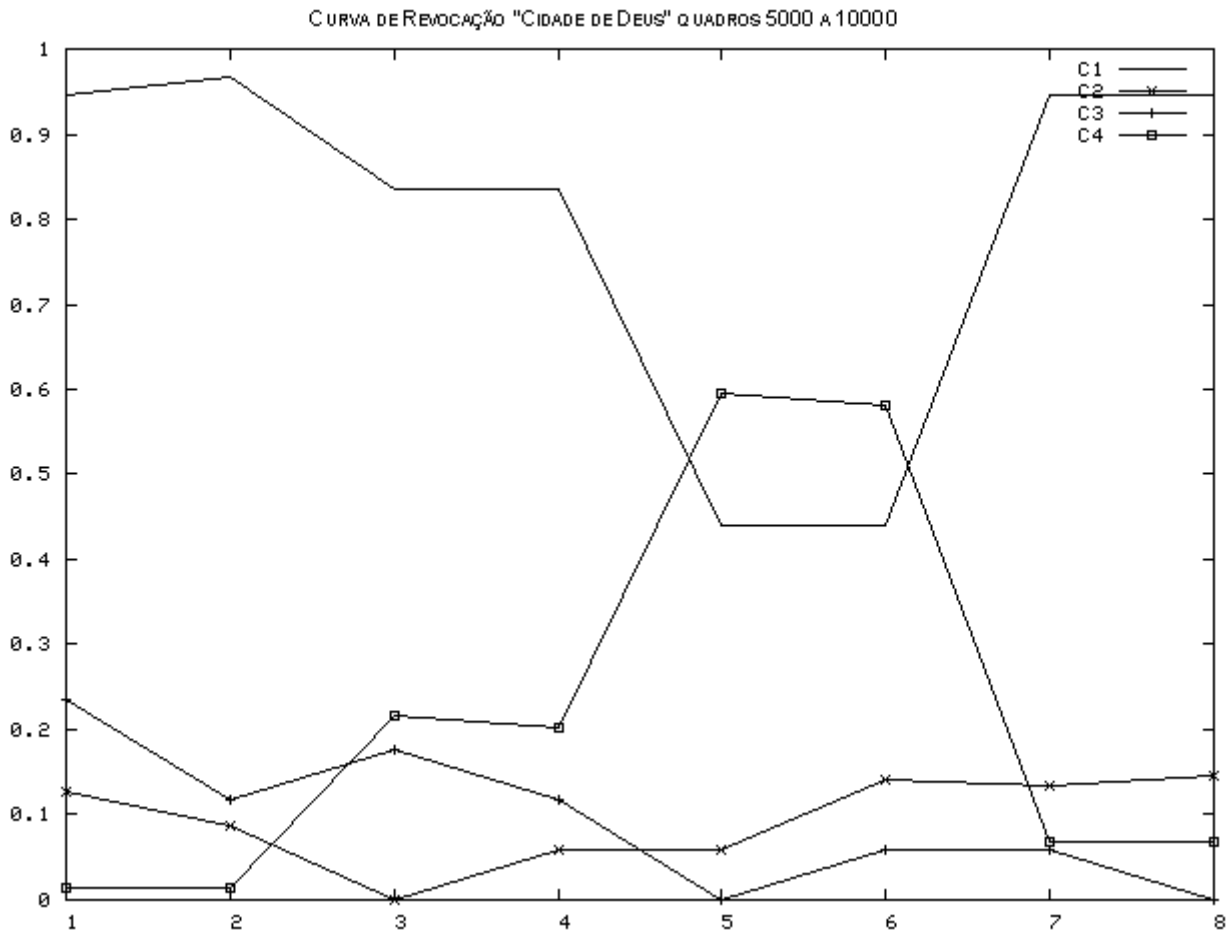


Figura 5.15: Curva de Recobrimento “Cidade de Deus” segundo grupo de 5.000 imagens

Tabela 5.4: Dados de Recobrimento para "Matrix Reloaded" primeiro grupo de 5.000 imagens

Grupos Limiares	C1	C2	C3	C4
1	90,61%	8,89%	9,09%	12,50%
2	94,59%	4,44%	9,09%	15,00%
3	76,47%	0,00%	15,15%	42,50%
4	76,47%	4,44%	9,09%	40,00%
5	60,83%	26,67%	0,00%	15,00%
6	60,83%	31,11%	9,09%	10,00%
7	90,61%	15,56%	3,03%	10,00%
8	90,61%	15,56%	9,09%	12,50%

Tabela 5.5: Dados de Precisão para "Matrix Reloaded" primeiro grupo de 5.000 imagens

Grupos Limiares	C1	C2	C3	C4
1	98,46%	1,45%	10,71%	2,46%
2	98,23%	1,44%	21,43%	4,11%
3	99,28%	0,00%	9,80%	1,45%
4	99,28%	2,38%	20,00%	1,40%
5	98,25%	0,68%	0,00%	2,73%
6	98,25%	0,75%	17,65%	4,60%
7	98,46%	1,75%	9,09%	4,21%
8	98,46%	1,73%	33,33%	5,38%

nas Tabelas 5.10 e 5.11.

O ponto mais estável de detecção verificado pelo processo de *ground-truth* está nos grupos de limiares definidos para os valores $LC = \{17, 16, 13, 13\}$ e $LM = \{21, 35, 45, 45\}$. Estes valores são utilizados como balizadores para a geração dos rótulos para o conjunto das 20 mil imagens não rotuladas da base.

A porcentagem de cada rótulo obtido através do melhor limiar para ambos filmes está mostrada na Tabela 5.12.

Tabela 5.6: Dados de Precisão para "Matrix Reloaded" segundo grupo de 5.000 imagens

Grupos Limiares	C1	C2	C3	C4
1	98,87%	3,33%	5,26%	2,86%
2	98,83%	0,00%	0,00%	5,00%
3	98,95%	0,00%	0,00%	2,22%
4	98,95%	0,00%	0,00%	2,71%
5	98,91%	0,49%	0,00%	3,61%
6	98,91%	0,62%	0,00%	0,00%
7	98,87%	3,33%	0,00%	0,00%
8	98,87%	3,12%	0,00%	0,00%

Tabela 5.7: Dados de Recobrimento para "Matrix Reloaded" segundo grupo de 5.000 imagens

Grupos Limiares	C1	C2	C3	C4
1	97,16%	11,11%	8,33%	5,88%
2	97,99%	0,00%	0,00%	8,82%
3	95,32%	0,00%	0,00%	14,71%
4	95,32%	0,00%	0,00%	17,65%
5	86,24%	16,67%	0,00%	8,82%
6	86,24%	22,22%	0,00%	0,00%
7	97,16%	16,67%	0,00%	0,00%
8	97,16%	16,67%	0,00%	0,00%

Tabela 5.8: Dados de Precisão para "Cidade de Deus" primeiro grupo de 5.000 imagens

Grupos Limiares	C1	C2	C3	C4
1	95,39%	8,20%	8,33%	6,14%
2	94,21%	9,55%	9,68%	5,87%
3	97,84%	0,00%	9,45%	4,21%
4	97,84%	8,70%	9,09%	4,53%
5	95,87%	3,39%	0,00%	6,19%
6	95,87%	4,10%	27,66%	6,22%
7	95,39%	10,59%	21,21%	6,27%
8	95,39%	10,76%	35,71%	6,65%

5.6 Discussão

O método aplicado foi testado para um total de 40 mil imagens retiradas de dois filmes diferentes, sendo que as imagens foram separadas em dois grupos, onde 20 mil imagens foram previamente rotuladas e outras 20 mil foram submetidas ao método sem rotulação.

A partir destes testes, algumas conclusões podem ser tiradas, tanto sobre os resultados obtidos quanto à aplicações do método futuramente.

Tabela 5.9: Dados de Recobrimento para "Cidade de Deus" primeiro grupo de 5.000 imagens

Grupos Limiares	C1	C2	C3	C4
1	84,70%	14,45%	8,46%	22,07%
2	89,78%	8,67%	4,62%	17,93%
3	62,81%	0,00%	9,23%	56,55%
4	62,81%	4,62%	3,08%	60,69%
5	49,46%	40,46%	0,00%	24,83%
6	49,46%	52,60%	10,00%	16,55%
7	84,70%	31,21%	5,38%	17,93%
8	84,70%	33,53%	7,69%	17,93%

Tabela 5.10: Dados de Precisão para “Cidade de Deus” segundo grupo de 5.000 imagens

Grupos Limiares	C1	C2	C3	C4
1	94,58%	26,80%	5,48%	0,56%
2	93,68%	36,84%	7,41%	0,61%
3	97,53%	0,00%	4,29%	1,61%
4	97,53%	39,58%	10,00%	1,50%
5	96,01%	6,85%	0,00%	0,45%
6	96,01%	8,20%	10,00%	0,00%
7	94,58%	38,70%	14,29%	0,00%
8	94,58%	39,11%	0,00%	0,00%

Tabela 5.11: Dados de Recobrimento para “Cidade de Deus” segundo grupo de 5.000 imagens

Grupos Limiares	C1	C2	C3	C4
1	94,66%	12,81%	23,53%	1,37%
2	96,62%	8,75%	11,76%	1,37%
3	83,61%	0,00%	17,65%	21,92%
4	83,61%	5,94%	11,76%	20,55%
5	44,02%	57,19%	0,00%	1,37%
6	44,02%	70,62%	5,88%	0,00%
7	94,66%	31,56%	5,88%	0,00%
8	94,66%	33,12%	0,00%	0,00%

Quanto à implementação do método, devemos considerar que o tempo de processamento é uma condição determinante à aplicabilidade desta técnica em condições práticas.

A utilização da ferramenta *Octave* para o desenvolvimento do método proporcionou implementar e testar diversas técnicas baseado em suas facilidades de utilização de funções matemáticas. Como contrapartida de sua aplicação, o tempo de processamento em relação a códigos compilados é muito grande, e não permite grandes otimizações devido ao *scripts* serem executados através do ambiente.

A ferramenta é muito útil ao desenvolvimento da técnica e na aplicação dos testes matemáticos e, a partir do momento que estes testes são efetivamente validados e avaliados, podem ser implementados em uma linguagem compilada, gerando assim, códigos muito mais rápidos e praticamente aplicáveis.

Tabela 5.12: Porcentagem dos rótulos na base não rotulada

Rótulo	Matrix	“Cidade de Deus”
C1	79,7%	76,1%
C2	11,2%	8,7%
C3	4,4%	2,3%
C4	5,7%	13,9%

A rotulação do grupo de imagens que compuseram o grupo de não rotuladas que foram utilizados servirão como forma de projetar a curva de Precisão-Recobrimento para mais dados e conseqentemente ampliar a variabilidade dos dados utilizados e aumentar a robustez do método para outros filmes.

A medida que se refinam as projeções feitas pelo sistema, novas análises podem ser feitas sobre os valores obtidos nas diferenças. As variações locais para os quadros de cada par da imagem podem ser analisados separadamente, ou determinadas regiões da imagem podem ter peso maior sobre as decisões tomadas pelo sistema de rotulação dos pares de imagens.

A detecção de $C1$ para os grupos de limiares foi a que obteve os melhores resultados. O melhor resultado foi obtido com o segundo grupo de limiares, onde os valores de Precisão e Recobrimento.

Os valores de $C2$ têm uma variabilidade maior devido a quantidade de verificações deste tipo de nível de ação ser menor. O grupo de limiares que teve a melhor resposta para este foi o nmero 8.

O nível $C3$ de ação também sofre com a pouca quantidade de dados contidas nas seqüências. O melhor grupo de limiares para este nível foi o 4, seguido do nmero 8. Possivelmente a superioridade do grupo 4 sobre o grupo 8 seja devido a uma menor separação entre os valores de $LM2$ e $LM3$ no grupo 4 (25 e 32 respectivamente) para os do grupo 8 (35 e 45 respectivamente). Como o valor de $LM3$ neste caso é consideravelmente maior no grupo 4 que no grupo 8, este fator teve impacto na classificação.

Outro fator ponderante ao resultado é a similaridade entre os níveis $C2$ e $C3$. Esse fator influencia muito mais o caráter subjetivo da rotulação da base do que a própria saída do método.

Os resultados de $C4$ sofre com um outro fator que é a pouca quantidade de incidência deste tipo de quadros na seqüência. Desta forma, qualquer detecção imprecisa tem um impacto maior no resultado obtido. Os melhores resultados para $C4$ foram obtidos com o grupo de limiares 5. Estes melhores resultados para $C4$ foram obtidos pelo detrimento de $C3$.

A grande diferença de precisção entre classificar os tipos de cena com mínimo de ação ($C1$) está na grande similaridade das imagens que compõem este tipo de distribuição.

A separação dos níveis superiores de ação ($C2$, $C3$ e $C4$) sofre com alguns fatores:

- A subjetividade presente na rotulação da base resulta em uma dificuldade de identificar os diferentes tipos de ação na seqüência de forma padronizada.

- A proximidade de $C2$, $C3$ e $C4$ revelaram uma sensibilidade maior às variações de

LM

As possíveis soluções para estas questões:

- Fazer uma rotulação mais apurada quanto à separação dos níveis de ação $C2$, $C3$ e $C4$;
- Analisar o comportamento dos dados após a rotulação direcionada para definir outros valores para LC e LM;
- Testar estes novos limiares com diversas combinações em busca de uma melhor representação.

Capítulo 6

Conclusões

Este trabalho descreve um sistema de carcterização e classificação de conteúdo em vídeos estruturados, tendo níveis diferentes de movimentação como forma de representação dos mesmos.

Os objetivos deste trabalho, como descrito no Capítulo **??** são:

- Fazer um estudo das diferentes técnicas de manipulação e análise de vídeos estruturados para identificar formas de representação de vídeos que permitam identificar a ação contida nos mesmos;
- Definir níveis diferentes de ação para caracterizar os vídeos;
- A criação de uma base de dados de vídeos estruturados para o trabalho;
- Aplicar a caracterização dos níveis de ação à base de dados criada.

Desta forma podemos analisar separadamente os objetivos específicos do trabalho e os resultados obtidos.

Através do estudo das técnicas de manipulação de vídeos e imagens decidiu-se pela utilização da representação das imagens através do formato HSV. As características utilizadas para representar os pares de imagens foram a Interseção de Histogramas HSV e a Máxima Verossimilhança. Estas características foram utilizadas com o objetivo de obter uma representação da variação de cor e de movimento contida em cada par de imagens.

Foram definidos quatro níveis diferentes de ação (C1, C2, C3, C4), indo da menor quantidade de ação para a maior. Estes quatro níveis foram criados através da análise da base de dados criada. A definição de níveis de ação para as seqüências de vídeos é uma das principais características do trabalho. Estes rótulos podem servir como balisadores para os processos posteriores de análise de vídeo.

Foi criada uma base de dados a partir de dois filmes. De cada filme foram extraídos cerca de 10.000 quadros. Cada par de quadros das seqüências foi rotulado considerando quatro níveis diferentes de ação. A criação da base de dados rotulados é um passo muito importante do trabalho, pois não existe nenhuma base de vídeos disponíveis para pesquisa. Desta forma a criação desta base é um passo inicial para que se crie uma base, com mais dados, que sirva de parâmetro para as pesquisas em vídeos estruturados.

A aplicação da técnica de rotulação está descrita e analisada detalhadamente no Capítulo 5. As principais questões a serem analisadas pela aplicação do método são:

- O caráter subjetivo da caracterização da base de dados influenciou os resultados obtidos;
- A quantidade de filmes e de quadros existentes na base de dados denota a necessidade de uma quantidade e uma variabilidade de filmes e estilos cinemáticos maior;
- As distribuições de cor e movimento para as seqüências demonstram uma separação correspondente dos picos e vales destes elementos. Este fator ressalta a pertinência destas características para representar os pares de quadros.

De uma forma geral o trabalho foi realizado de acordo com os objetivos definidos previamente e os resultados obtidos pelo método apontam diversos caminhos possíveis na análise de vídeos estruturados.

A pouca quantidade de publicações diretamente relacionadas ao tema torna a quantidade de alternativas grande, e também torna necessárias atividades como a de desenvolvimento da base de dados, e outros trabalhos futuros já citados. Os resultados descritos neste trabalho devem ser aplicados em uma base de dados maior, sofrendo mudanças de ajustes nos seus pesos para geração de novos dados qualitativos e novas avaliações do método.

A aplicação de técnicas de análise de conteúdo em vídeos estruturados pode ser aplicada em atividades práticas como a recuperação de vídeos por conteúdo, aplicável à bibliotecas digitais. Estas técnicas também podem ser aplicadas em sistemas de classificação de conteúdos em reporduutores de vídeos digitais, tanto em computadores, quanto em aparelhos comerciais como tocadores de DVD e aparelhos digitais.

A obtenção de uma representação dos níveis de ação contídos em um vídeo proporciona também uma base para estudos posteriores que permitem analisar e tipificar uma seqüência de vídeo, rotulando o período de imagens como diálogo, paisagem, violência, ação, através da adição de outros elementos estruturais ou específicos de cada um dos tipos de cena.

6.1 Trabalhos Futuros

Para aumentar a robustez do método são necessárias e possíveis algumas atividades que são como passos posteriores no desenvolvimento deste trabalho.

A necessidade de uma base de vídeo rotulada com, ao menos, 5 filmes completos é um fator que proporciona ao estudo uma quantidade de amostras de tomadas muito variadas. Para tanto, deve-se sselecionar mais filmes que possuam cenas com representação adequada de tomadas tanto de ação quanto tomadas mais introspectivas. A geração de tal base (até hoje indisponível) pode servir como base para o desenvolvimento de outros estudos em vídeos estruturados.

A projeção dos limiares testados em uma quantidade maior de combinações, analisando as variações dos valores de Precisão-Recobrimento às mudanças destes parâmetros. Desta forma pode-se obter uma representação ótima das variações de movimentação em vídeo, assim como uma análise do impacto das mudanças de cada mum dos limiares nos resultados obtidos. Essa questão está diretamente relacionada à robustez do método, porque testar uma grande variabilidade de limiares em uma base de dados satisfatoriamente representativa, proporciona uma confiabilidade ao método.

A análise local das diferenças entre as imagens, através dos subpartes resultantes da divisão das imagens em 20 partes, é uma forma diferente de ver os mesmos dados contidos nas imagens.

Essa visão local pode ser implementada de várias formas:

- 1) Definir pesos diferentes para determinadas regiões ou grupos de subpartes baseados na região de atenção da imagem capturada, geralmente no centro. Esta é uma característica cinemática que pode ser explorada.

- 2) Adicionar a informação da posição so subquadro ao valor da diferença. Desta forma podem-se identificar variações em regiões "vizinhas"espacialmente nas imagens. Assim pode-se aumentar a robustez à grandes variações esparsas no espaço, assim como contemplar variações locais.

A caracterização manual de pares com muita ou pouca ação ($C4$ e $C1$ respectivamente) se apresenta mais simples devido a quantidade extrema de movimentação ou diferença de cores presentes nos mesmos. Os casos de rotulação de pares $C2$ e $C3$ possibilitam a ocorrência de mais divergências devido à proximidade que ambos tipos possuem. Este fatores têm seu impacto na saída do sistema a partir do momento que tanto $C2$ quanto $C3$ definem pares com ação, mesmo que em niveis diferentes. Como o objetivo primeiro deste trabalho é o de caracterizar períodos de ação e não ação na seqüência de

vídeo, o tratamento minucioso da separação dos níveis intermediários de ação torma-se um trabalho posterior, a partir do momento em que tipos variados de ação sejam buscados (violência, perseguição, luta, por exemplo).

Das técnicas de outros trabalhos descritos na seção 2 podemos apontar alguns elementos que podem ser aplicados a este trabalho:

- A representação de quadros-chave em [HANJALIC et al., 1997] pode ser aplicada ao trabalho como elemento representativo para agrupar as tomadas semelhantes em processos posteriores a caracterização.

- O acréscimo da informação de tamanho da tomada [IYENGAR, 2002] pode ser utilizado como elemento ponderante na classificação dos níveis de ação.

Referências Bibliográficas

[FFMPEG, 2003] Fast fourier MPEG Software. *http://ffmpeg.sourceforge.net/, 2003.*

[OCTAVE, 2003] John W. Eaton. *University of Wisconsin. Department of Chemical Engineering* Gnu Octave, http://www.octave.org/, 2003.

[ADAMS et al., 2003] ADAMS, B., AMIR, A., DORAI, C., and GHOSAL, S. (2003). Ibm research trec-2002 video retrieval system.

[ARMAN et al., 1994] ARMAN, F., HSU, A., and CHIU, M.-Y. (1994). Image processing on encoded video sequences. In *ACM Multimedia Systems Journal.*

[BARRON et al., 1994] BARRON, J. L., BEAUCHEMIN, S. S., and FLEET, D. J. (1994). On optical flow. In *6th Int. Conf. on Artificial Intelligence and Information-Control Systems of Robots (AIICSR).*

[CHOWDURRY and CHELLAPA, 2003] CHOWDURRY, A. R. and CHELLAPA, R. (2003).

[DUGAD et al., 1998] DUGAD, R., RATAKONDA, K., and AHUJA, N. (1998). Robust video shot change detection. In *IEEE Workshop on Multimedia Signal Processing.*

[EKIN and TEKALP, 2003] EKIN, A. and TEKALP, A. M. (2003). Generic event detection in sports video using cinematic features. In *2nd IEEE Workshop on Event Mining : Detection and Recognition of Events in Video.*

[EKIN et al., 2003] EKIN, A., TEKALP, A. M., and MEHROTRA, R. (2003). Automatic soccer video analysis and summarization. *IEEE Trans. Image Processing.*

[FAN and LUO, 2003] FAN, J. and LUO, H. (2003). Principal video shot: Linkig low-level perception features to semantic video events.

[GARGI et al., 2000] GARGI, U., Kasturi, R., and Strayer, S. H. (2000). Performance characterization of video-shot-change detection methods. *IEEE Trans. Circuits Syst. Video Techn.*, 10(1):1–13.

[GULER et al., 2003] GULER, S., LIANG, W. H., and PUSHEE, I. A. (2003). A video event detection and mining framework.

[HANJALIC et al., 1997] HANJALIC, A., CECCARELLI, M., LAGENDIJK, R. L., and BIEMOND, J. (1997). Automation of systems enabling search on stored video data. In *Storage and Retreval for Image and Video Databases V.*

[HANJALIC et al., 1999] HANJALIC, A., LAGENDIJK, R. L., and BIEMOND, J. (1999). Automatically segmenting movies into logical story units. In *Visual Information and Information Systems.*

[IYENGAR, 2002] IYENGAR, G. R. (2002). *Characterization of Unstructured Video.* PhD thesis, Massachusetts Institure of Technology.

[KOBLA et al., 1996] KOBLA, V., DOERMANN, D., and ROSENFELD, A. (1996). Compressed domain video segmentation.

[LEE et al., 2003] LEE, J. H., LEE, G. G., and KIM, W. Y. (2003). Automatic video summarizing tool using mpeg-7 descriptors for personal video recorder. In *IEEE Transactions on Consumer Eletronics.*

[LI and SEZAM,] LI, B. and SEZAM, M. I.

[LIN et al., 2002] LIN, C.-Y., TSENG, B., and SMITH, J. (2002). Universal mpeg content access using compressed-domain system stream editing techniques. In *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference*, volume 2, pages 73–76.

[MIRAMAX, 2003] International MIRAMAX (2003). Cidade de deus. Brasil.

[PATEL and SETHI,] PATEL, N. and SETHI, I. Compressed video processing for cut detection.

[RASHEED and Shah, 2003] RASHEED, Z. and Shah, M. (2003). Scene detection in hollywood movies and tv shows. In *IEEE International Conference on Coputer Vison and Pattern Recognition (CVPR).*

[RUI et al., 1998] RUI, Y., HUANG, T. S., and MEHROTRA, S. (1998). Exploring video structure beyond the shots. In *ICMCS '98: Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, page 237, Washington, DC, USA. IEEE Computer Society.

[SETHI and PATEL, 1995] SETHI, I. K. and PATEL, N. V. (1995). Statistical approach to scene change detection. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 329–338.

[SIKORA, 2003] SIKORA, T. (2003). *Digital Consumer Eletronics Handbook*. McGraw Hill Book Company.

[STAUFFER, 2003] STAUFFER, C. (2003). Estimating tracking sources and sinks. In *Second IEEE Workshop on Event Mining*.

[VASCONCELOS and LIPPMAN, 2000] VASCONCELOS, N. and LIPPMAN, A. (2000). Feature representations for image retrieval: Beyond the color histogram. In *IEEE International Conference on Multimedia and Expo (II)*, pages 899–902.

[WARNER, 2003] WARNER Bros. (2003). The matrix reloaded. Estados Unidos da America.