UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

WILSON PIRES GAVIÃO NETO

Sumarização de Vídeos de Histeroscopias Diagnósticas

Tese apresentada como requisito parcial
para a obtenção do grau de
Doutor em Ciência da Computação

Prof. Dr. Jacob Scharcanski
Orientador

Porto Alegre, Agosto de 2009

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Gavião Neto, Wilson Pires

Sumarização de Vídeos de Histeroscopias Diagnósticas / Wilson Pires Gavião Neto. – Porto Alegre: PPGC da UFRGS, 2009.

104 f.: il.

Tese (doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2009. Orientador: Jacob Scharcanski.

1. Sumarização de vídeos. 2. Indexação de vídeos. 3. Browsing de vídeos. 4. Histeroscopias. 5. Vídeo médico. I. Scharcanski, Jacob. II. Título.

AGRADECIMENTOS

Agradeço as oportunidades que me foram oferecidas no desenvolvimento deste trabalho. Neste sentido, começo agradecendo meu orientador pela responsabilidade assumida na condução deste trabalho e por importantes ensinamentos, sobretudo no sentido de sempre buscar o máximo de qualidade no desenvolvimento de um trabalho. Agradeço a Universidade Federal do Rio Grande do Sul, em particular ao Instituto de Informática e todos os seus membros com os quais eu interagi durante estes anos. Também, sou especialmente grato a Universidade da Carolina do Norte (UNC) e ao Departamento de Ciência da Computação onde parte do programa de doutorado foi realizado. Agradeço ao Professor Marc Pollefeys pela oportunidade de desenvolver parte do trabalho junto a um qualificado grupo de pesquisa. Neste grupo, inúmeras discussões foram importantes para a compreensão de idéias centrais deste trabalho, bem como para apontar caminhos na direção de possíveis soluções. Em especial, as discussões com Jan-Michael Frahm foram bastante proveitosas, bem como a oportunidade oferecida pelo Professor Steve Marron para debater o problema junto a sua classe de doutorandos no departamento de estatística da UNC. Agradeço também ao Professor Fernando Sabino do departamento de estatística da UFRGS, que desenvolvia seu trabalho de doutorado em estatística na UNC, pelo suporte técnico prestado e amizade durante o tempo que passei na Carolina do Norte. A valiosa oportunidade de me dedicar exclusivamente a este trabalho durante um bom tempo deve-se ao suporte financeiro provido pelo governo brasileiro através do CNPq (no Brasil) e CAPES (no exterior).

No campo da medicina, sou especialmente grato ao Dr. Paulo Cará, e sua assistente Inês, por todo suporte prestado na condução de experimentos documentados neste trabalho. Agradeço também ao Professor João Sabino Cunha Filho e toda sua equipe do Departamento de Ginecologia e Obstetrícia do Hospital de Clínicas de Porto Alegre.

Quando trata-se de oportunidades, certamente que sempre vou agradecer a meus pais, Sérgio Gavião e Ivone Krüger, sobretudo pela oportunidade de escolher uma profissão. Finalmente, muito especialmente agradeço a quem mais de perto participou, apoiou e incondicionalmente seguiu os "distantes destinos" visitados por este trabalho, *Obrigado Karen!*.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS 6

LISTA DE FIGURAS 7

LISTA DE TABELAS 13

RESUMO 14

ABSTRACT 15

1 INTRODUÇÃO E CONTRIBUIÇÕES 16

2 VíDEOS DE HISTEROSCOPIAS 19
2.1 O Exame de Histeroscopia 19
2.2 Características de Vídeos de Histeroscopias 20
2.2.1 Informação Espacial 20
2.2.2 Informação Temporal 21
2.2.3 Informação Geométrica 22

3 FUNDAMENTOS: ESTRUTURA E INDEXAÇÃO DO CONTEÚDO DE VíDEOS DIGITAIS 23
3.1 A Estrutura de um Vídeo 23
3.2 Indexação de Vídeos Digitais 24
3.3 Sumarização de Vídeos Digitais 26
3.3.1 Extração de Feições 26
3.3.2 Segmentação Temporal do Vídeo 26
3.3.3 Seleção de Quadros-chave 27
3.3.4 Aplicações 28
3.4 Conclusões 28

4 FUNDAMENTOS: MOVIMENTO DE CÂMERA A PARTIR DE IMAGENS 30
4.1 Conceitos e Abordagens Comumente Adotadas 30
4.1.1 Extração e Rastreamento de Feições 30
4.1.2 Geometria de Cenas Estáticas 3-D 31
4.1.3 Movimento de Câmera a partir de Pontos Correspondentes em Duas Imagens 32
4.1.4 Configurações Degeneradas de Cena 36
4.1.5 Correção de Distorções de Lente e Calibração da Câmera 38
4.2 Seleção Robusta do Modelo de Movimento de Câmera 39
4.2.1 O Algoritmo QDEGSAC 42

4.2.2 Saída do Algoritmo QDEGSAC 46
4.2.3 Custo Computacional do Algoritmo QDEGSAC 46

5 TRABALHOS RELACIONADOS E ABORDAGENS ALTERNATIVAS DESENVOLVIDAS NESTE TRABALHO 47
5.1 Análise de Movimento na Sumarização de Vídeos 47
5.2 Sumarização de Vídeos de Histeroscopias 51
5.2.1 Método Baseado na Decomposição de Valor Singular (SVD - *Singular Value Decomposition*) 51
5.2.2 Método Baseado em um Modelo para Valores de Distâncias entre Quadros 56
5.2.3 Discussão 60
5.3 Conclusões 62

6 SUMARIZAÇÃO DE VíDEOS DE HISTEROSCOPIAS DIAGNÓSTICAS BASEADA EM MOVIMENTOS DE CÂMERA 64
6.1 Visão Geral do Método 64
6.2 Correspondências entre Quadros do Vídeo 65
6.3 Validação Geométrica das Correspondências 66
6.3.1 Restrição Epipolar 66
6.3.2 Cenas Degeneradas e Quase-Degeneradas 66
6.4 Representação Hierárquica para Vídeos de Histeroscopias 68
6.4.1 Pontos Consistentes 69
6.4.2 Pontos Persistentes e Sobreposição de Conteúdo 70
6.4.3 Árvores de Segmentos de Vídeo 70
6.5 Seleção de Quadros-Chave 73

7 EXPERIMENTOS 75
7.1 Experimentos com Dados Sintéticos 75
7.2 Experimentos com Imagens Reais 83
7.2.1 Avaliação do Algoritmo QDEGSAC em Seqüências Reais 84
7.2.2 Resultados em Sumarização de Vídeos 90
7.3 Discussão 94

8 CONCLUSÕES 97

REFERÊNCIAS 98

LISTA DE ABREVIATURAS E SIGLAS

SVD	*Singular Value Decomposition*
KLT	Kanade-Lucas-Tomasi
RANSAC	*Random Sample Consensus*
SFM	*Structure-From-Motion*
QDEGSAC	RANSAC para Dados Quase Degenerados
u.d.f.	Unidades de Distância Focal

LISTA DE FIGURAS

Figura 2.1: Anatomia do útero. 19

Figura 2.2: Imagens capturadas durante as fases de uma histeroscopia diagnóstica: (a) Panorâmica da cavidade uterina; (b) Abertura trompa esquerda; (c) Abertura trompa direita; (d) Fundo do Útero. 20

Figura 2.3: Alguns quadros retirados de segmentos de vídeo com pouco importância para propósitos de diagnósticos/prognósticos. Estes quadros são caracterizados por apresentar regiões obstruídas por muco e efeitos luminosos indesejáveis. 21

Figura 2.4: Endoscópios com campo de visão oblíquo permitem aos especialistas mudarem a direção de visão simplesmente rotacionando o cilindro do endoscópio em torno de seu eixo. Idealmente, esta é uma das configurações críticas de cena que devem ser detectadas para se evitar resultados espúrios em termos de restrições de movimento 3D de câmera. 22

Figura 3.1: Estrutura de um Vídeo 24

Figura 3.2: Diagrama de uma metodologia clássica de análise de vídeos digitais. . 25

Figura 3.3: Diagrama do sistema de sumarização de vídeos de ecocardiografia proposto por Ebadollahi et al (EBADOLLAHI; CHANG; WU, 2001) 29

Figura 4.1: Projeções $\mathbf{x}$ e $\mathbf{x}'$ de um ponto no espaço $\mathbf{X}$ recaem sobre o mesmo plano, conhecido com plano epipolar. Esta é uma importante propriedade amplamente explorada na busca por correspondências que irão dar suporte a estimativas de movimento de câmera $\mathbf{E}$ entre $\mathbf{C}$ e $\mathbf{C}'$. . 34

Figura 4.2: Projeções ao longo de raios que passam por um ponto comum $\mathbf{C}$ (o centro de projeção) definem um mapeamento entre planos. Se um sistema de coordenadas é definido em cada plano e pontos são representados em coordenadas homogêneas, então este mapeamento pode ser expresso como uma transformação projetiva $\mathbf{H}$, onde $\mathbf{x}' = \mathbf{H}\mathbf{x}$. . 35

Figura 4.3: Relação entre uma geometria epipolar $\mathbf{E}$ e uma homografia $\mathbf{H}$ induzida pelo plano π. Um ponto $\mathbf{x}$ mapeado pela homografia recai sobre sua linha epipolar correspondente $l' = \mathbf{E}\mathbf{x}$. 36

Figura 4.4: Configurações degeneradas de cena onde uma transformação projetiva $\mathbf{H}$ explica a relação entre as imagens, onde $\mathbf{x}' = \mathbf{H}\mathbf{x}$. (a) Cena Planar. (b) Cena em que a câmera executa apenas rotações em torno de seu eixo ótico. 37

Figura 4.5: Visão geral do algoritmo QDEGSAC. 44

Figura 5.1: Gráfico mostrando a norma $\|\psi\|$ dos quadros como barras verticais. O eixo horizontal representa cada quadro i na ordem temporal do vídeo, e o eixo vertical representa $\|\psi_i\|$. A linha tracejada representa τ. Barras em cinza representam os segmentos de vídeo discartados, enquanto que as barras pretas representam os segmentos selecionados. A menor barra preta dentro de cada segmento selecionado mostra o quadro-chave correspondente. As setas indicam estes quadros- chave. 54

Figura 5.2: Ilustração do processo de fusão de segmentos de vídeo. Os diagramas acima mostram parte de uma seqüência temporal de valores da norma $\|\psi\|$ (barras verticais) para os quadros de um vídeo histeroscópico particular. Os segmentos horizontais de reta indicam os segmentos relevantes do vídeo: (a) abaixo, quadros chave representando os segmentos antes do processo de fusão; (b) quadros chave obtidos de segmentos que sofreram o processo de fusão (cada segmento é representado por um quadro chave). 55

Figura 5.3: Gráficos mostrando os segmentos selecionados e o limiar δ^l em cada nível de sumarização. O eixo horizontal representa cada quadro X_i na seqüência temporal do vídeo e o eixo vertical representa as distâncias entre os quadros adjacentes $D(H(X_i), H(X_{i+1}))$. A linha contínua representa o limiar δ^l em cada nível de sumarização l: (a) nível 1; (b) nível 2 e (c) nível 3. A linha irregularmente tracejada, acima do limiar, indica a localização temporal dos segmentos relevantes selecionados pelo método e, a linha acima, representa a localização temporal dos segmentos selecionados pelos especialistas. 59

Figura 5.4: Diagrama mostrando características visuais e distâncias de quadros para um segmento de vídeo de histeroscopia diagnóstica. O eixo horizontal representa cada quadro i na ordem temporal do vídeo, e o eixo vertical representa distâncias entre quadros adjacentes (barras em cinza). A curva sobreposta ao gráfico representa uma versão suavizada dos valores de distâncias. Abaixo, mostra-se uma escala visual onde cada quadro é representado por uma fatia vertical de 5 *pixels* extraída do seu centro. Padrões visuais mais estáveis (isto é, segmentos relevantes) são verificados para as regiões de vale da curva. 61

Figura 5.5: Quadros de número 445, 550 e 595 extraídos da seqüência temporal mostrada na figura 5.4. O quadro 550, proveniente de uma região de vale, corresponde a fase de inspeção da abertura da trompa direita (seção 2.1). 62

Figura 5.6: Gráficos mostrando distâncias (barras verticais) entre quadros para um segmento de vídeo de histeroscopia diagnóstica em particular. O eixo horizontal representa cada quadro i na ordem temporal do vídeo. (a) Distâncias entre quadros adjacentes $n = 1$; (b) Distância média $\widehat{D_i}$ de um quadro X_i em relação a todos os quadros X_j que compõe o segmento de vídeo, $n > 1$. Alguns picos e vales tornam-se mais evidentes para valores maiores de n. 63

Figura 6.1: Quantidade de inliers computados entre os quadros (a) e (b) para dois modelos de movimento: uma matriz essencial $\mathbf{E}$ e uma homografia $\mathbf{H}$. (c-d) Inliers sobrepostos sobre o segundo quadro para $\mathbf{H}$ (c) e $\mathbf{E}$ (d) ($\tau = 2$ pixels). (e) Comparação entre os modelos $\mathbf{E}$ e $\mathbf{H}$ considerando diferentes níveis τ de erro. O modelo $\mathbf{E}$ é um modelo de movimento menos restritivo que $\mathbf{H}$, conseqüentemente $\mathbf{E}$ pode explicar pontos correspondentes que $\mathbf{H}$ não pode e, por esta razão, geralmente entrega uma quantidade maior de inliers. 67

Figura 6.2: Notação associada aos quadros considerados no processo de validação geométrica de pontos correspondentes. Um ponto $\mathbf{X}_i$ no espaço é projetado sobre os quadros I^j, $I^{j+\Delta}$, $I^{j+2\Delta}$ e $I^{j+k\Delta}$ respectivamente como $\mathbf{x}_i^j$, $\mathbf{x}_i^{j+\Delta}$, $\mathbf{x}_i^{j+2\Delta}$ e $\mathbf{x}_i^{j+k\Delta}$, estabelecendo pontos correspondentes como $\mathbf{x}_i^j \leftrightarrow \mathbf{x}_i^{j+\Delta}$ e $\mathbf{x}_i^{j+\Delta} \leftrightarrow \mathbf{x}_i^{j+2\Delta}$. 69

Figura 6.3: A sobreposição dos campos de visão determinará os quadros que serão agrupados primeiro. Os quadros centrais $I^{j-\Delta}$ e I^j constituirão o segmento de vídeo $\delta^{j-\Delta \mapsto j}$ se a sobreposição de conteúdo entre eles $\theta_{j-\Delta}^j$ é maior que as sobreposições de conteúdo $\theta_{j-2\Delta}^{j-\Delta}$ e $\theta_j^{j+\Delta}$, as quais foram computadas com relação a seus quadros vizinhos $I^{j-2\Delta}$ e $I^{j+\Delta}$. 71

Figura 6.4: O processo iterativo constrói segmentos de vídeo (i) agrupando quadros e formando novos segmentos, (ii) agregando quadros a segmentos de vídeo já existentes ou (iii) agrupando segmentos de vídeo vizinhos em segmentos maiores. 72

Figura 6.5: Uma árvore de segmento de vídeo particular na qual a tarefa de *browsing* de seus quadros é representada como uma linha horizontal (setas em cinza). Enquanto está linha desloca-se através dos níveis da árvore, sumários de vídeos mais, ou menos, compactos são gerados em termos de quantidade de quadros-chave. As intersecções $\times$ determinam uma hierarquia de sub-segmentos de vídeo (sub-árvores), os quais constituirão o segmento de vídeo final no topo da árvore. . . . 73

Figura 7.1: Quantidade de inliers computados e a quantidade média de trials requerida pelo algoritmo QDEGSAC para computar inliers como uma função do ruído, onde $|\mathbf{t}| = 5$ u.d.f e $\alpha = 5$ graus (configuração genérica de cena). 76

Figura 7.2: Quantidade de inliers computados e a quantidade média de trials requerida pelo algoritmo QDEGSAC para computar inliers como uma função do ruído, onde $|\mathbf{t}| = 5$ u.d.f e $\alpha = 0$ graus (configuração genérica de cena). 77

Figura 7.3: Desempenho do algoritmo QDEGSAC para cenas planares (degeneradas) como função do ruído, onde $(|\mathbf{t}|, \alpha) = (5, 5)$. (Acima) Quantidade de inliers computados e quantidade média de trials do algoritmo QDEGSAC para computá-los. (Abaixo) Proporção de inliers degenerados computados para $\tau \in \{1, 2, 3\}$. 77

Figura 7.4: Desempenho do algoritmo QDEGSAC para cenas degeneradas (rotação pura) como função do ruído, onde $(|\mathbf{t}|, \alpha) = (0, 5)$. (Acima) Quantidade de inliers computados e quantidade média de trials do algoritmo QDEGSAC para computá-los. (Abaixo) Proporção de inliers degenerados computados para $\tau \in \{1, 2, 3\}$. 78

Figura 7.5: Em termos da relação correta de movimento T_{Right}, erros médios são computados para os inliers $\{in\}^{\tau}_{QDEGSAC}$ entregues pela algoritmo QDEGSAC ($\tau \in \{1, 2, 3\}$) como uma função de níveis de ruído. (a) À direita, erro computado para cenas não-degeneradas. À esquerda, erro computado para cenas planares (degeneradas), $(|\mathbf{t}|, \alpha) = (5, 5)$. (b-c) Erro computado sobre os inliers $\{in\}^{\tau}_{QDEGSAC}$ entregues pelo QDEGSAC contra o erro computado sobre o conjunto de inliers reais $\{in\}^{\tau}_{Right}$ (*ground truth*) para cenas não-degeneradas (b) e cenas degeneradas (c). 79

Figura 7.6: Desempenho do algoritmo QDEGSAC como um função da quantidade de outliers reais no contexto de cenas genéricas (não-degeneradas), onde $|\mathbf{t}| = 5$ u.d.f, $\alpha = 5$ graus e $\sigma = 0.1$ pixels. À esquerda, quantidade de inliers reais perdidos. À direita, quantidade média requerida de trials pelo algoritmo QDEGSAC. 80

Figura 7.7: Quantidade de inliers reais perdidos e a média de trials requerida pelo algoritmo QDEGSAC como função da quantidade de outliers reais presentes nos dados, onde $|\mathbf{t}| = 5$ u.d.f., $\alpha = 0$ graus (configuração genérica de cena) e $\sigma = 0.1$ pixels. 81

Figura 7.8: Desempenho do algoritmo QDEGSAC como função da quantidade de outliers reais para cenas degeneradas, onde $|\mathbf{t}| = 0$ u.d.f., $\alpha = 5$ graus e $\sigma = 0.1$ pixels. Acima, quantidade de inliers reais perdidos e a média de trials requerida pelo algoritmo QDEGSAC. Abaixo, proporção de inliers degenerados para $\tau \in \{1, 2, 3\}$. 81

Figura 7.9: Erro médio calculado sobre os conjuntos de inliers $\{in\}^{\tau}_{QDEGSAC}$ entregues pelo algoritmo QDEGSAC ($\tau \in \{1, 2, 3\}$) e o conjunto de inliers reais $\{in\}_{Right}$ (*ground truth*) produzido pela relação T_{Right}, como uma função da quantidade de outliers reais presentes nos dados de entrada. À esquerda, erro calculado para cenas genéricas com $(|\mathbf{t}|, \alpha) = (5, 5)$. À direita, erro calculado sobre cenas degeneradas, incluído o erro computado sobre o conjunto de todos os pontos correspondentes (inliers e outliers reais). 82

Figura 7.10: Comparação do algoritmo QDEGSAC contra uma simples abordagem RANSAC no contexto de cenas degeneradas, considerando diferentes quantidades de outliers reais, onde $(|\mathbf{t}|, \alpha) = (0, 5)$ (apenas rotação de câmera). O erro simétrico de transferência é calculado sobre os inliers entregues pelo QDEGSAC $\{in\}^{\tau}_{QDEGSAC}$ e RANSAC $\{in\}^{\tau}_{RANSAC}$ para $\tau \in \{1, 2, 3\}$. 82

Figura 7.11: **Acima**: quadros 1,7 e 13 da seqüência *Checkerboard*. **Meio**: movimento 2D dos pontos correspondentes computados pelo rastreador KLT, do quadro corrente para o próximo ("→"). **Abaixo**: Quadros com a distorção de lente removida. (a) Quadro 1. (b) Quadro 7. (c) Quadro 13. (d) Movimento 2D do quadro 1 para o quadro 7. (e) Movimento 2D do quadro 7 para o quadro 13. (f) Movimento 2D do quadro 1 para o quadro 13. 85

Figura 7.12: Da esquerda para direita, resultados do quadro 1 para 7, 7 para 13 e 1 para 13 da seqüência *checkerboard*. **(a-c)** Quantidade de inliers (eixo y) quando emprega-se 6, 7 e 8 restrições (eixo x). (a) $\tau = 1$. (b) $\tau = 2$. (c) $\tau = 3$. Inliers adicionais computados pelo algoritmo QDEGSAC aparecem empilhados (cinza). Linha horizontal pontilhada representa a quantidade total de pontos correspondentes entregue pelo rastreador KLT. **(d)** Histogramas do erro epipolar residual para os inliers computados pelo algoritmo QDEGSAC com $\tau = 3$. . 87

Figura 7.13: **Acima**: quadros 1, 10 e 20 da seqüência *Tubal Orifice*. **Abaixo**: movimento 2D dos pontos correspondentes computados pelo rastreador KLT, do quadro corrente para o próximo ("→"). (a) Quadro 1. (b) Quadro 10. (c) Quadro 20. (d) Movimento 2D do quadro 1 para o quadro 10. (e) Movimento 2D do quadro 10 para o quadro 20. (f) Movimento 2D do quadro 1 para o quadro 20. 88

Figura 7.14: Da esquerda para direita, resultados do quadro 1 para 10, 10 para 20 e 1 para 20 da seqüência *Tubal Orifice*. **(a-c)** Quantidade de inliers (eixo y) quando emprega-se 6, 7 e 8 restrições (eixo x). (a) $\tau = 1$. (b) $\tau = 2$. (c) $\tau = 3$. Inliers adicionais computados pelo algoritmo QDEGSAC aparecem empilhados (cinza). Linha horizontal pontilhada representa a quantidade total de pontos correspondentes entregue pelo rastreador KLT. **(d)** Histogramas do erro epipolar residual para os inliers computados pelo algoritmo QDEGSAC com $\tau = 3$. . 89

Figura 7.15: **Acima**: quadros 1, 5 e 10 da seqüência *Fundus*. **Abaixo**: movimento 2D dos pontos correspondentes computados pelo rastreador KLT, do quadro corrente para o próximo ("→"). (a) Quadro 1. (b) Quadro 5. (c) Quadro 10. (d) Movimento 2D do quadro 1 para o quadro 5. (e) Movimento 2D do quadro 5 para o quadro 10. (f) Movimento 2D do quadro 1 para o quadro 10. 90

Figura 7.16: Da esquerda para direita, resultados do quadro 1 para 5, 5 para 10 e 1 para 10 da seqüência *Fundus*. **(a-c)** Quantidade de inliers (eixo y) quando emprega-se 6, 7 e 8 restrições (eixo x). (a) $\tau = 1$. (b) $\tau = 2$. (c) $\tau = 3$. Inliers adicionais computados pelo algoritmo QDEGSAC aparecem empilhados (cinza). Linha horizontal pontilhada representa a quantidade total de pontos correspondentes entregue pelo rastreador KLT. **(d)** Histogramas do erro epipolar residual para os inliers computados pelo algoritmo QDEGSAC com $\tau = 3$. 91

Figura 7.17: Quadros cuja ordem temporal é a mediana dentro de segmentos (importantes) selecionados com auxílio de especialistas. (a) *hyst1*. (b) *hyst2*. (c) *hyst3*. (d) *hyst4*. 93

Figura 7.18: Árvores de segmentos de vídeo computadas através dos vídeos. Eixo x representa os quadros na seqüência temporal do vídeo. Linhas horizontais em vermelho representam os segmentos de vídeo (importantes) selecionados com o auxílio de especialistas. (a) *hyst1*. (b) *hyst2*. (c) *hyst3*. (d) *hyst4*. Observa-se que segmentos importantes geralmente aparecem associados a árvores de altura destacada. . . . 94
Figura 7.19: Árvores de segmentos de vídeo e quadros-chave associados, ambos computados sobre uma seqüência extraída do vídeo *hyst1*. **(Acima)** 9 árvores de segmentos de vídeo como um função da seqüência temporal dos quadros. **(Meio)** Seqüência de vídeo amostrada em intervalos regulares de 25 quadros. **(Abaixo)** Os 9 quadros-chave computados para cada uma das árvores de segmento. Segmentos cuja duração é menor que $1/3$ de segundo foram descartados como irrelevantes. . . 95

LISTA DE TABELAS

Tabela 7.1: Parâmetros estimados para a câmera histeroscópica utilizada na aquisição dos vídeos analisados nos experimentos 75
Tabela 7.2: Resultados para a seqüência *Checkerboard* (média sobre 100 execuções do algoritmo QDEGSAC) 86
Tabela 7.3: Resultados para a seqüência *Tubal Orifice* (média sobre 100 execuções do algoritmo QDEGSAC) 86
Tabela 7.4: Resultados para a seqüência *Fundus* (média sobre 100 execuções do algoritmo QDEGSAC) 88
Tabela 7.5: Valores dos parâmetros empregados nos experimentos com os vídeos. 91
Tabela 7.6: Descrição/caracterização dos vídeos utilizados nos experimentos. . . 92
Tabela 7.7: Sumário de resultados e comparação do método proposto (M1) contra o método proposto em (SCHARCANSKI; GAVIãO, 2006) (M2) . . . 94

RESUMO

Dada uma biblioteca com milhares de vídeos de histeroscopias diagnósticas, sobre a qual deseja-se realizar consultas como *"retornar imagens contendo miomas submucosos"* ou *"recuperar imagens cujo diagnóstico é pólipo endometrial"*. Este é o contexto deste trabalho. Vídeos de histeroscopias diagnósticas são usados para avaliar a aparência do útero e são importantes não só para propósitos de diagnóstico de doenças mas também em estudos científicos em áreas da medicina, como reprodução humana e estudos sobre fertilidade. Estes vídeos contêm uma grande quantidade de informação, porém somente um número reduzido de quadros são úteis para propósitos de diagnósticos e/ou prognósticos. Esta tese apresenta um método para identificar automaticamente a informação relevante em vídeos de histeroscopias diagnósticas, criando um sumário do vídeo. Propõe-se uma representação hierárquica do conteúdo destes vídeos que é baseada no rastreamento de pontos geometricamente consistentes através da seqüência dos quadros. Demonstra-se que esta representação é uma maneira útil de organizar o conteúdo de vídeos de histeroscopias diagnósticas, permitindo que especialistas possam realizar atividades de *browsing* de uma forma rápida e sem introduzir informações espúrias no sumário do vídeo. Os experimentos indicam que o método proposto produz sumários compactos (com taxas de redução de dados em torno de 97.5%) sem descartar informações clinicamente relevantes.

Palavras-chave: Sumarização de vídeos, indexação de vídeos, browsing de vídeos, histeroscopias, vídeo médico.

Content-Based Summarization of Diagnostic Hysteroscopy Videos

ABSTRACT

Given a library containing thousands of diagnostic hysteroscopy videos, which are only indexed according to a patient ID and the exam date. Usually, users browse through this library in order to obtain answers to queries like *retrieve images of submucosal myomas* or *recover images whose diagnosis is endometrial polyp*. This is the context of this work. Specialists have been used diagnostic hysteroscopy videos to inspect the uterus appearance, once the images are important for diagnosis purposes as well as in medical research fields like human reproduction. These videos contain lots of information, but only a reduced number of frames are actually useful for diagnosis/prognosis purposes. This thesis proposes a technique to identify clinically relevant information in diagnostic hysteroscopy videos, creating a rich video summary. We propose a hierarchical representation based on a robust tracking of image points through the frame sequence. We demonstrate this representation is a helpful way to organize the hysteroscopy video content, allowing specialists to perform fast browsing without introducing spurious information in the video summary. The experimental results indicate that the method produces compact video summaries (data-rate reduction around 97.5%) without discarding clinically relevant information.

Keywords: Video Summarization, Video Indexing, Video Browsing, Hysteroscopy, Medical Video.

1 INTRODUÇÃO E CONTRIBUIÇÕES

Dada uma biblioteca com mais de 10.000 vídeos de histeroscopias diagnósticas indexadas apenas por um código de paciente e a data em que os exames foram realizados. Sobre esta base de vídeos realiza-se consultas do tipo: *"Buscar imagens que apresentam miomas submucosos"*. Um especialista inicia então uma atividade de *browsing* (pesquisar/navegar) sobre os quadros dos vídeos para resolver a consulta. Este é o contexto sobre o qual este trabalho se desenvolve.

A histeroscopia diagnóstica é um exame clínico no qual um especialista utiliza um instrumento telescópico pequeno, chamado histeroscópio, para capturar imagens do útero. O histeroscópio transmite imagens do canal uterino para um monitor, permitindo que o especialista guie o instrumento dentro da cavidade uterina. O resultado deste procedimento é um vídeo. Vídeos de histeroscopias são importantes para a avaliação das condições de saúde das pacientes, pois possibilitam a visualização da aparência do útero.

Na prática hospitalar/clínica vários vídeos de histeroscopias diagnósticas são produzidos por dia. Usualmente estes vídeos são gravados por completo devido a dinâmica do procedimento. Tais vídeos apresentam uma duração média de 3 minutos (aproximadamente 4500 quadros a uma taxa de 25 quadros por segundo) e, de acordo com especilistas, é importante guardar algumas imagens junto aos prontuários das pacientes.

Assim, vídeos de histeroscopias contêm uma grande quantidade de informações, porém somente um número reduzido de quadros são úteis para propósitos de diagnóstico/prognóstico. Especialistas realizam uma busca manual e seqüencial no conteúdo visual do vídeo para selecionar imagens relevantes. Um método capaz de apontar quadros e segmentos de video relevantes dentro de um vídeo de histeroscopia diagnóstica auxiliaria os médicos na atividade de *browsing* dos quadros do vídeo. Deste modo, o tempo necessário para procurar imagens relevantes e fazer a descrição do conteúdo do vídeo seria menor.

A parte da literatura que trata da análise e sumarização de vídeos apresenta poucos trabalhos no contexto de vídeos médicos. De uma forma geral, a literatura provê apenas direções que podem ser seguidas na construção de um método apropriado para o contexto deste trabalho, não oferecendo estudos que possam ser tomados de maneira comparativa quando trata-se da análise digital do conteúdo de vídeos de histeroscopias. Neste sentido destacam-se algumas diferenças (contribuições) deste trabalho em relação a literatura:

1. Um método de sumarização projetado de acordo com uma análise sobre as características e o conteúdo de vídeos de histeroscopias;

2. Uma representação eficiente para vídeos de histeroscopias, de maneira que especialistas possam realizar uma tarefa de *browsing* rápida sobre o conteúdo visual sem desconsiderar regiões clinicamente importantes dos vídeos;

3. Em geral, métodos de sumarização de vídeos propostos na literatura estão focados na representação do conteúdo dinâmico dos vídeos. Basicamente, estes métodos exploram a estrutura de vídeos comerciais editados, sendo que o reconhecimento de transições entre unidades desta estrutura, como cenas e *shots*[1], estão no centro das técnicas propostas. Contudo, videos de histeroscopias são gravados como uma seqüencia ininterrupta de images, onde a câmera é manualmente guiada através da cavidade uterina. Assim, este trabalho apresenta como contribuição um primeiro passo em direção ao que poderia ser entendido como um *shot* dentro de um vídeo de histeroscopia.

4. Usualmente, rastreamento de pontos em seqüencias de imagens não é utilizado por técnicas de sumarização de vídeos devido a conhecidas dificuldades em casar pontos correspondentes entre imagens que apresentam movimentos rápidos, o que freqüentemente aparece na prática. Felizmente, no contexto de vídeos de histeroscopias, movimento rápido de câmera é uma evidência de que o conteúdo das imagens não atraiu a atenção do especialista. Por esta razão, as limitações no processo de rastreamento de pontos apareceria em segmentos de video que carregam um conteúdo de pouca relevância clínica, o que direcionaria para poucas perdas de representatividade no sumário de vídeo resultante. Sendo assim, este trabalho apresenta uma maneira de explorar o rastreamento de pontos em seqüências de imagens para propósitos de sumarização e estruturação do conteúdo de vídeos que permitem tal rastreabilidade. Com base nos experimentos realizados, a metodologia proposta permite computar boas estimativas em termos de alterações de conteúdo visual ao longo de uma seqüência de quadros de vídeo. A partir de medidas simples, como a persistência de pontos ao longo das imagens, é possível obter respostas razoáveis para questões clássicas no campo de pesquisa que trata da representação de vídeos com base em conteúdo visual, tal como *"qual a quantidade de sobreposição de conteúdo que há entre os quadros dados?"*.

5. Neste trabalho discute-se o problema de sumarizar longas seqüências de imagens exploratórias, onde a câmera é manualmente guiada (i. e. direção e dimensão de movimentos pouco previsíveis) através de regiões igualmente importantes. A maioria das técnicas propostas na literatura detectam unidades semânticas nos vídeos, como *shots* ou mesmo seqüências com padrões similares de movimento de câmera, e utilizam regras predeterminadas para selecionar quadros-chave dentro de tais unidades. A metodologia proposta neste trabalho pode estimar a sobreposição de conteúdo visual entre quadros vizinhos e, por esta razão, pode produzir um sumário de vídeo que é adaptativamente construído seguindo um equilíbrio entre redundância e descarte de informações na seleção de quadros-chave.

O conteúdo desta tese está organizado em sete partes. O capítulo 2 introduz o procedimento histeroscópico e apresenta as características do vídeo produzido em termos de conteúdo espacial e temporal, além das configurações geométricas típicas de cenas histeroscópicas. Tendo em vista que esta tese trata sobre sumarização de vídeos baseada em conteúdo, apresenta-se no capítulo 3 conceitos e abordagens clássicas dentro desta área de estudo. O capítulo 4 apresenta conceitos e idéias que fundamentam a abordagem proposta neste trabalho. No capítulo 5 discute-se a literatura para propósitos de estruturação

[1]Um *shot* pode ser entendido como uma tomada de uma cena. A seção 3.1 apresenta detalhes sobre a estrutura de um vídeo.

do conteúdo de vídeos de histeroscopias diagnósticas. A metodologia proposta é apresentada no capítulo 6. Os experimentos realizados, uma análise crítica do método proposto e resultados são descritos no capítulo 7. Por fim, o capítulo 8 conclui este trabalho e resume as principais contribuições.

2 VÍDEOS DE HISTEROSCOPIAS

Vídeos de histeroscopia diagnóstica são usados para avaliar a aparência do útero e são de grande importância para propósitos de diagnósticos/prognósticos. Nas próximas seções apresenta-se uma introdução sobre o procedimento de histeroscopia bem como uma análise sobre as características do vídeo histeroscópico. Esta analise é de grande importância na definição da abordagem proposta neste trabalho.

2.1 O Exame de Histeroscopia

A histeroscopia é um procedimento cirúrgico no qual um ginecologista utiliza um instrumento telescópico pequeno, chamado histeroscópio, para diagnosticar e tratar problemas do útero. O histeroscópio transmite imagens do canal uterino (Figura 2.1) para um monitor, permitindo que o ginecologista guie o instrumento dentro da cavidade endometrial. Há dois tipos de histeroscopia: diagnóstica e operativa. A histeroscopia diagnóstica é um procedimento realizado para examinar o útero e verificar a existência de sinais de anormalidades. A histeroscopia operativa é realizada com o objetivo de tratar problemas diagnosticados previamente. A abordagem proposta neste trabalho está direcionada para a histeroscopia diagnóstica.

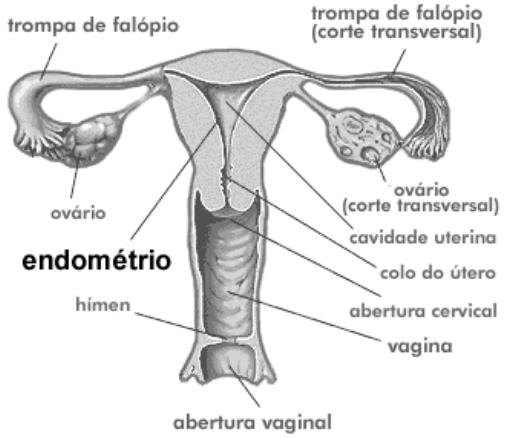
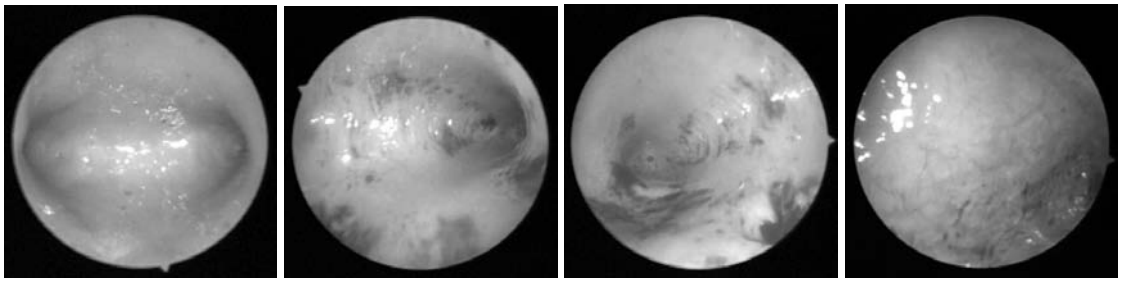


Figura 2.1: Anatomia do útero.

Em geral, uma histeroscopia diagnóstica típica constitui-se de quatro fases. Em cada fase objetivos específicos são alcançados:

Cavidade Uterina: quando a abertura cervical (Figura 2.1) é ultrapassada, a cavidade



(a) (b) (c) (d)

Figura 2.2: Imagens capturadas durante as fases de uma histeroscopia diagnóstica: (a) Panorâmica da cavidade uterina; (b) Abertura trompa esquerda; (c) Abertura trompa direita; (d) Fundo do Útero.

uterina é examinada panoramicamente. A Figura 2.2(a) exemplifica uma imagem capturada durante esta fase. Na seqüência o exame procede com a identificação dos orifícios das trompas;

Orifício trompa esquerda (ou direita): nesta fase examina-se a abertura da trompa esquerda. A Figura 2.2(b) mostra uma imagem capturada durante esta fase;

Orifício trompa direita (ou esquerda): nesta fase examina-se a abertura da trompa direita. A Figura 2.2(c) exemplifica uma imagem capturada durante esta fase;

Fundo do Útero: o especialista aproxima a microcâmera da parede do útero para examinar as características do endométrio. A Figura 2.2(d) exemplifica uma imagem capturada durante esta fase.

Um vídeo de histeroscopia tem uma duração aproximada de 3 minutos. Usualmente estes vídeos são gravados por completo devido à dinâmica do procedimento. Assim, considerando uma taxa de amostragem de 25 quadros por segundo, um vídeo de histeroscopia contém por volta de 4.500 quadros gravados de forma ininterrupta. Médicos salientam que estes vídeos contêm uma grande quantidade de informações, porém somente um número reduzido de quadros são úteis para propósitos de diagnóstico/prognóstico. Na prática, é importante guardar algumas imagens junto aos prontuários das pacientes. Para isso os médicos realizam uma busca manual no conteúdo visual do vídeo para selecionar imagens relevantes.

2.2 Características de Vídeos de Histeroscopias

A informação contida em uma cena de vídeo pode ser descrita através de três componentes fundamentais (IRANI et al., 1997): informação espacial (aparência), informação temporal e informação geométrica da cena (estrutura 3D da cena e movimento de câmera). Deste modo descreve-se então o vídeo de histeroscopia:

2.2.1 Informação Espacial

A intensidade dos pixels em imagens de histeroscopias é determinada pela orientação da extremidade do cilindro histeroscópico, pois este contém uma fonte de luz. Por esta razão, reflexos especulares aparecem com relativa freqüência, como observado nas imagens

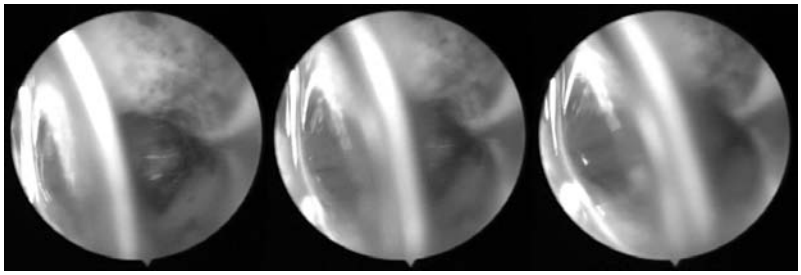


Figura 2.3: Alguns quadros retirados de segmentos de vídeo com pouco importância para propósitos de diagnósticos/prognósticos. Estes quadros são caracterizados por apresentar regiões obstruídas por muco e efeitos luminosos indesejáveis.

da Figura 2.2. Além disso, devido ao amplo campo de visão da lente, as imagens apresentam uma distorção nítida. Deste modo, se um modelo de câmera linear é assumido, uma fonte de erro será introduzida em processos cujo objetivo é, por exemplo, computar estimativas de movimento de câmera.

Clinicamente, imagens de histeroscopias são espacialmente classificadas como relevantes e irrelevantes. Imagens relevantes são caracterizadas por um campo de visão desobstruído em relação a parede uterina, como mostrado nas imagens da Figura 2.2. Por outro lado, imagens irrelevantes não apresentam detalhes do útero claramente. Basicamente, tais imagens são geradas em dois contextos. Primeiro, quando o especialista move a câmera procurando por regiões de interesse, ele tende a movê-la de uma maneira bem mais rápida em comparação a quando ele está de fato observando detalhes de seu interesse. Por esta razão, a qualidade das imagens é degradada pelos efeitos provocados pelo movimento rápido da câmera, tal como borramento de detalhes. Segundo, a secreção de muco juntamente com o gás insuflado produz bolhas, as quais podem aparecer repentinamente no campo de visão obstruindo informações importantes. A Figura 2.3 exemplifica como estas imagens são degradas.

2.2.2 Informação Temporal

A maioria das técnicas para indexação de vídeos com base em conteúdo propostas na literatura estão direcionadas para vídeos comerciais. Estas técnicas exploram a maneira na qual os vídeos são editados e métodos para reconhecer transições entre *shots* constituem-se na idéia central destas abordagens (LEW, 2001). Diferentemente de vídeos comerciais, uma histeroscopia diagnóstica é gravada como uma seqüência ininterrupta de imagens na qual a câmera é manualmente guiada através da cavidade uterina. Usualmente, especialistas movem a câmera lentamente quando estão examinando importante regiões do útero. Neste sentido, uma questão poderia ser feita: *Por que o especialista não simplesmente pressiona um botão de captura quando imagens relevantes estão sendo observadas?* A resposta não é tão óbvia. Na prática, o procedimento histeroscópico é baseado na inspeção de diferentes regiões da parede do útero. Por esta razão a câmera é reposicionada freqüentemente, e não raramente artefatos degradam os quadros (ou parte deles), como ilustrado na Figura 2.3. Estes artefatos indesejáveis aparecem e desaparecem do campo de visão repentinamente durante o exame, como explicado na seção anterior. Assim, mesmo que uma captura seletiva de quadros fosse realizada, tais artefatos ainda afetariam a qualidade do material gravado. Por esta razão é desejável que os vídeos sejam gravados completamente.

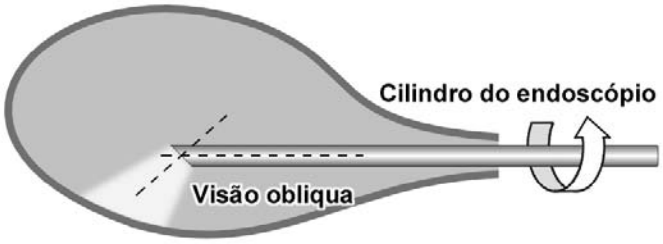


Figura 2.4: Endoscópios com campo de visão oblíquo permitem aos especialistas mudarem a direção de visão simplesmente rotacionando o cilindro do endoscópio em torno de seu eixo. Idealmente, esta é uma das configurações críticas de cena que devem ser detectadas para se evitar resultados espúrios em termos de restrições de movimento 3D de câmera.

2.2.3 Informação Geométrica

Conforme discutido na seção 2.1, pode-se distinguir quatro fases distintas em um vídeo de histeroscopia, cada qual com seus respectivos objetivos (HAMOU, 1991). Contudo, em termos de análise de movimento de câmera, identifica-se configurações de cenas distintas associadas com as fases mencionadas. Basicamente, o vídeo de histeroscopia é obtido em termos de três contextos geométricos de cena:

1) Um *exame panorâmico*, envolvendo translação e rotação de câmera dentro de um ambiente 3D (a cavidade uterina) com variações claras de profundidade, como mostrado nas imagens da Figura 2.2(a-c).

2) Uma configuração de *cena aproximadamente planar* que ocorre quando o especialista aproxima a parede do útero para examinar características do endométrio. Este contexto é típico da fase de inspeção do *Fundo do Útero*. Figura 2.2(d) mostra um quadro característico desta fase.

3) Endoscópios com campo de visão oblíquo são amplamente utilizados em histeroscopias devido ao fato que a direção de visão pode ser alterada simplesmente rotacionando-se o cilindro endoscópico em torno de seu eixo, como ilustrado na Figura 2.4. Na prática, este recurso é freqüentemente utilizado pelos especialistas e, como resultado, uma configuração de cena na qual a câmera é apenas rotacionada (sem translação) pode aparecer. Matematicamente este movimento pode ser modelado por duas rotações sucessivas, como proposto em (YAMAGUCHI et al., 2004).

A presença de determinadas configurações de cena é um importante aspecto que motiva a abordagem proposta neste trabalho, uma vez que vídeos de histeroscopias são adquiridos por uma câmera que é manualmente (livremente) guiada, onde o movimento induzido pela câmera nas imagens, e conseqüentemente a configuração da cena, são difíceis de serem preditos. Tal importância deve-se a necessidade de detecção de algumas configurações de cena devido a limitações de modelos paramétricos utilizados para estimar movimento de câmera (TORR; FITZGIBBON; ZISSERMAN, 1999).

3 FUNDAMENTOS: ESTRUTURA E INDEXAÇÃO DO CONTEÚDO DE VÍDEOS DIGITAIS

Os avanços tecnológicos nos últimos anos têm provido maior capacidade de processamento e armazenamento de dados na indústria digital. Assim, a manipulação de grandes quantidades de dados torna-se cada vez mais rápida. Aliado a isso, o crescimento da internet, em termos de trafegabilidade de informações e número de usuários, tem impulsionado a evolução da tecnologia multimidia. Neste contexto estão os vídeos digitais, os quais requerem novas tecnologias para prover métodos de manipulação de dados visuais com um baixo custo de armazenamento e uma metodologia de indexação que permita um acesso eficiente aos dados armazenados. Baseado nisso, a abstração/sumarização de vídeos digitais torna-se um tópico de estudo importante. Idealmente, técnicas de sumarização de vídeos devem ser capazes de eliminar redundâncias e detectar os eventos de maior importância dentro da seqüência de imagens, de maneira que um usuário não necessite verificar o vídeo inteiro para obter informações a respeito destes eventos.

Neste capítulo apresenta-se uma introdução sobre sumarização de vídeos digitais bem como uma visão geral sobre os principais desafios encontrados no desenvolvimento de técnicas destinadas a este fim.

3.1 A Estrutura de um Vídeo

Ao contrário de uma simples imagem, os vídeos representam sua informação através de múltiplos planos de comunicação. Isto inclui os efeitos de edição que revelam a maneira pela qual os quadros são ligados. Neste contexto, as mudanças de cor, textura, forma e movimento ao longo da seqüência dos quadros são importantes para a identificação da estrutura do vídeo. Além disso, cada tipo de vídeo (por exemplo: comerciais, notícias, filmes, esportes ou vídeos médicos) tem características peculiares que devem ser levadas em consideração em processos automáticos de extração, indexação e acesso ao seu conteúdo.

Basicamente, a estrutura de um vídeo é composta por três unidades principais (Figura 3.1):

- Os *quadros* são a unidade básica de informação dentro de um vídeo. Normalmente são amostrados a uma taxa de 25 ou 30 quadros por segundo;

- Os *shots* são conjuntos de quadros dispostos temporalmente entre dois efeitos de transição, como cortes ou outro efeito de edição. Os *shots* caracterizam-se por uma continuidade perceptual, formando os segmentos elementares na estrutura de um vídeo;

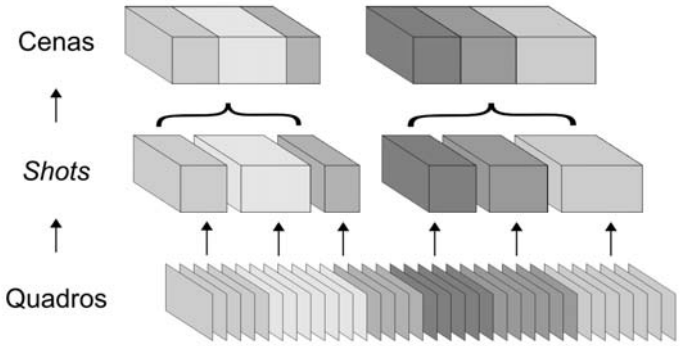


Figura 3.1: Estrutura de um Vídeo

- As *cenas* são formadas por coleções de *shots* que mostram uma seqüência temporal de eventos ocorridos em um local físico. Cenas podem ser classificadas como estáticas ou dinâmicas. Por exemplo, cenas de diálogos em filmes são classificadas como estáticas.

3.2 Indexação de Vídeos Digitais

Os avanços na digitalização, armazenamento e tecnologias de comunicação de dados têm criado grandes quantidades de vídeos digitais. Contudo, a interação com dados multimídia requer mais do que simples conexões com dados bancários ou envio de dados via internet para um usuário consumidor. A análise do conteúdo de vídeos deveria ser análoga à análise de documentos textuais, onde um procedimento estrutural decompõe o documento em parágrafos, sentenças e palavras, antes da construção de um índice. Para prover um acesso rápido e confiável a dados de vídeos, deveria-se segmentar este documento visual em tomadas (*shots*) e cenas, com base na extração de quadros-chave, para compor uma tabela de conteúdo. Isso não é uma tarefa fácil. Ferramentas destinadas a descrever, organizar e gerenciar dados visuais ainda apresentam grandes limitações. O ponto central discutido na literatura científica diz respeito à indexação e estruturação do conteúdo deste tipo de mídia, sendo que o maior desafio está em estabelecer uma relação entre informações de baixo nível, como cor e vetores de movimento, e o conteúdo das imagens, como *"pôr do sol"*, *"corrida de carros"*, *"montanhas"*, etc.

Um esquema típico de análise do conteúdo de vídeos digitais, seguido por muitos pesquisadores, envolve quatro processos principais (DIMITROVA et al., 2002):

1. *Extração de feições*
2. *Análise da estrutura ou segmentação temporal do vídeo*
3. *Sumarização*
4. *Indexação*.

A Figura 3.2 ilustra a relação existente entre estes processos. Cada etapa desta análise apresenta suas próprias dificuldades.

A *extração de feições* é um processo crítico no contexto da análise de vídeos digitais. A eficiência de um esquema de indexação de vídeo, por exemplo, depende diretamente

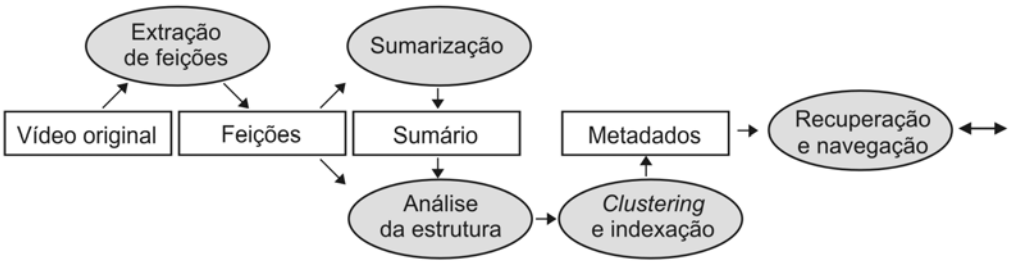


Figura 3.2: Diagrama de uma metodologia clássica de análise de vídeos digitais.

dos atributos utilizados na representação do seu conteúdo. As dificuldades iniciam-se com um problema clássico de recuperação de informações visuais (SMEULDERS et al., 2000): mapear feições visuais como cor, textura, forma e movimento em conceitos semânticos, tais como pessoas, corrida de carros e cenas. Estratégias que vêm sendo adotadas para contornar este tipo de problema incluem o uso de outras informações contidas em um vídeo digital, tais como textos sobrepostos nas imagens e áudio (LI; ZHANG; TRETTER, 2001; DIMITROVA et al., 2002).

O próximo passo em uma abordagem clássica de análise do conteúdo de vídeos é a *segmentação temporal do vídeo*. Este processo consiste em extrair a estrutura da informação temporal contida na seqüência de quadros do vídeo. Isso envolve detectar fronteiras temporais, como cortes, e identificar segmentos importantes do vídeo, tais como cenas e *shots*. Vários métodos para a segmentação temporal de vídeos são encontrados na literatura (DIMITROVA et al., 2002; NGO; PONG; ZHANG, 2001; CERNEKOVA; PITAS; NIKOU, 2006).

A *sumarização de vídeos* é o processo de criação e apresentação de uma versão resumida da informação visual contida na estrutura do vídeo (DIMITROVA et al., 2002). Este processo é similar a extração de palavras chave ou sumários em processamento de textos. Sendo assim, o procedimento para vídeos tem como objetivo a extração de um subconjunto dos dados do vídeo original, tais como quadros chave e cenas. A sumarização é um processo especialmente importante, tendo em vista a grande quantidade de dados contida mesmo em vídeos de pouca duração. O resultado do processo de sumarização forma uma base não só para a representação do conteúdo, mas também para a indexação e recuperação de vídeos. Neste ponto é importante destacar o papel dos quadros chave no processo de sumarização. Quadros chave são imagens estáticas, extraídas do vídeo original, que melhor representam o conteúdo do vídeo de uma maneira abstrata. Isto justifica-se pelo fato que nem todos quadros dentro de uma seqüência são igualmente descritivos, sendo que o maior desafio é determinar automaticamente os quadros que são mais representativos. A literatura apresenta muitos trabalhos neste sentido (CERNEKOVA; PITAS; NIKOU, 2006; HANJALIC; ZHANG, 1999; NGO; PONG; ZHANG, 2001; GONG; LIU, 2000; DEMENTHON; KOBLA; DOERMANN, 1998; LI; ZHANG; TRETTER, 2001), porém construir uma metodologia robusta para a extração de quadros chave em vídeos de propósito geral ainda constitui-se em uma tarefa desafiadora.

O processo de *indexação de vídeos* baseia-se em metadados, que são os atributos estruturais e de conteúdo extraídos nos processos de extração de feições, segmentação temporal e sumarização do vídeo. Com base nestes atributos, pode-se então construir índices e tabelas de conteúdo utilizando-se, por exemplo, técnicas de *clustering*. Estas técnicas classificam os segmentos de vídeo em diferentes categorias visuais, que podem

estar mapeadas em uma estrutura indexada (HANJALIC; ZHANG, 1999; GONG; LIU, 2000). Assim como em sistemas de banco de dados tradicionais, esquemas e ferramentas tornam-se necessários para o uso dos índices e metadados em procedimentos de consulta e navegação em bancos de vídeos. A literatura apresenta várias propostas neste sentido (DIMITROVA et al., 2002), porém métodos robustos e eficientes para manipular grandes conjuntos de dados ainda são necessários.

3.3 Sumarização de Vídeos Digitais

Sumarizar um vídeo é um processo que envolve a criação e apresentação de uma versão concisa da informação visual (isto é, o conteúdo) contida na estrutura do vídeo (DIMITROVA et al., 2002). Segundo Li et al (LI; ZHANG; TRETTER, 2001), o sumário de um vídeo é uma seqüência de imagens estáticas ou dinâmicas representando o seu conteúdo, de maneira que eventos chave sejam providos de forma rápida e concisa enquanto a mensagem do vídeo, como um todo, é bem preservada. A literatura apresenta dois tipos fundamentais de sumarização de vídeos (HANJALIC; ZHANG, 1999; LI; ZHANG; TRETTER, 2001): (a) métodos que extraem uma coleção de imagens estáticas e (b) técnicas que geram seqüências de imagens dinâmicas como resultado da sumarização.

Especial atenção é dada à extração automática de quadros-chave neste trabalho. Este tópico tem sido discutido em muitas pesquisas sobre análise de vídeos nos últimos anos (DEMENTHON; KOBLA; DOERMANN, 1998; HANJALIC; ZHANG, 1999; SAHOURIA; ZAKHOR, 1999; GONG; LIU, 2000; LI; ZHANG; TRETTER, 2001; CHANG; MACIEJEWSKI; BALAKRISHNAN, 1999; NGO; PONG; ZHANG, 2001; CERNEKOVA; PITAS; NIKOU, 2006). O desafio é determinar automaticamente os quadros que são mais representativos. A maioria dos métodos propostos na literatura caracterizam-se por algumas etapas bem definidas no processo de sumarização, a qual leva a escolha dos quadros-chave: a extração de feições, a segmentação do vídeo (detecção de suas unidades temporais) e um critério que define os quadros-chave, como introduzido na seção anterior.

3.3.1 Extração de Feições

A extração de feições é a primeira etapa no processo de sumarização, e consiste na construção de uma representação da informação contida nos quadros do vídeo, tendo em vista que processar a informação de cada *pixel* contido nos quadros pode exigir um esforço computacional excessivo. Em geral, a extração de feições baseia-se em técnicas de análise e indexação de imagens estáticas, as quais, na maioria das aplicações, buscam representar o conteúdo visual através de vetores de características de baixo nível, como cor, textura e forma (SMEULDERS et al., 2000; DEL BIMBO, 1999; LEW, 2001). Além disso, a informação de movimento também é explorada na construção de representações de baixo nível para vídeos. Inúmeras técnicas fazem uso de vetores de movimento disponíveis em formatos como *mpeg*. Limitações destas representações dizem respeito a dificuldade em modelar a percepção visual humana através de medições de baixo nível (SMEULDERS et al., 2000), como histogramas de cor, vetores de movimento e outras.

3.3.2 Segmentação Temporal do Vídeo

A detecção automática das unidades de um vídeo (*video parsing*) consiste em extrair a estrutura temporal da informação contida no vídeo (DIMITROVA et al., 2002; DEL BIMBO, 1999; HANJALIC, 2002). A maioria dos trabalhos focam na segmentação dos *shots* através da identificação de suas fronteiras temporais (cortes e efeitos de

transições), ou por meio de algoritmos de *clustering*, que agrupam quadros com características semelhantes. Métodos fundamentados na estrutura de *shots* do vídeo (NGO; PONG; ZHANG, 2001; LI; ZHANG; TRETTER, 2001) propõe a extração de pelo menos um quadro-chave por *shot* (LI; ZHANG; TRETTER, 2001), ou mesmo montam um sumário dinâmico (isto é, um segmento de vídeo) composto por *shots* que contêm os quadros-chave apontados pelo método (HANJALIC; ZHANG, 1999).

3.3.3 Seleção de Quadros-chave

Um dos principais aspectos em técnicas de sumarização de vídeos é o critério de seleção dos quadros-chave. Avanços na seleção de quadros-chave utilizam teoria de grafos (NGO; PONG; ZHANG, 2005; CHANG; SULL; LEE, 1999), simplificação de curvas (DEMENTHON; KOBLA; DOERMANN, 1998), *clustering* (HANJALIC; ZHANG, 1999) e decomposição em valor singular (GONG; LIU, 2000; CHANG; MACIEJEWSKI; BALAKRISHNAN, 1999; SAHOURIA; ZAKHOR, 1999). A idéia central consiste em selecionar um subconjunto de pontos representativos (associados a quadros em um espaço de feições) dentro de uma determinada distância, ou pontos/quadros que capturam mudanças significativas de conteúdo dentro de cada *shot*.

Chang et al. (CHANG; SULL; LEE, 1999) interpreta um *shot* como um grafo de proximidade onde cada quadro é um vértice neste grafo. Procura-se encontrar o conjunto de vértices que minimiza a distância total entre os vértices e seus pontos vizinhos. Dementhon (DEMENTHON; KOBLA; DOERMANN, 1998) trata a seleção de quadros-chave como um problema de simplificação de curvas. Um *shot* é visto como uma trajetória, ou uma curva composta por pontos em um espaço de feições de alta dimensionalidade. A partir disso o método procura detectar junções (quebras) nesta curva, as quais identificam os quadros-chaves. Deste modo é possivel obter uma visão hierárquica dos quadros-chave recursivamente, identificando junções da curva enquanto esta é simplificada para níveis de menor detalhe. Um problema potencial desta abordagem diz respeito a dificuldade em avaliar a aplicabilidade dos quadros-chave obtidos, tendo em vista a ausência de um estudo que comprove que o conjunto de quadros-chave extraídos de fato captura instantes importantes do vídeo (HANJALIC; ZHANG, 1999).

Inúmeros métodos propostos na literatura representam quadros como pontos em um espaço de feições, como mencionado acima. Assim, quadros similares tendem a formar agrupamentos (*clusters*) neste espaço de feições, traduzindo o processo de sumarização em técnicas de *clustering*, onde os quadros chave são selecionados de acordo com os centros dos *clusters* computados (GONG; LIU, 2000; CHANG; MACIEJEWSKI; BALAKRISHNAN, 1999; NGO; PONG; ZHANG, 2001; HANJALIC; ZHANG, 1999; SAHOURIA; ZAKHOR, 1999). Um bom exemplo deste tipo de abordagem é proposto por Hanjalic et al. (HANJALIC; ZHANG, 1999), onde um algoritmo de validação é empregado para selecionar a quantidade ideal de *clusters* dentro dos *shots*. Os *clusters* resultantes são ótimos em termos da medida de distância empregada. Contudo, isto é feito com relativo esforço computacional, pois os quadros são agrupados em n *clusters*, e para cada possibilidade de agrupamento (isto é, para cada n) uma análise de validação $\rho(n)$, baseada na distância de centros de *cluster* e na distância *intra-cluster*, é empregada (Equação 3.1). Neste caso, ξ_i representa a dispersão dos pontos do cluster i e μ_{ij} representa a distância entre os centróides dos *clusters* i e j. Deste modo, quanto menor o valor de $\rho(n)$, maior é a chance da opção de n *clusters* ser a melhor configuração para o vídeo analisado. Os quadros que possuem suas projeções no espaço de feições mais próximas dos centróides dos *clusters* são selecionados como quadros-chave.

$$\rho(n) = \frac{1}{n}\sum_{i=1}^{n} max_{1\leq j\leq n \wedge i\neq j}\left(\frac{\xi_i + \xi_j}{\mu_{ij}}\right), \qquad n \geq 2 \tag{3.1}$$

A fim de obter um espaço de feições mais conciso e representativo para os quadros, técnicas de redução de dimensionalidade também são empregadas em uma etapa anterior a execução de um algoritmo de *clustering*. Com isto busca-se preservar a estrutura do espaço de feições (eixos linearmente independentes), ao mesmo tempo que menores diferenças (como o ruído) são removidas da representação dos quadros (eixos linearmente dependentes). Gong e Liu (GONG; LIU, 2000) representam a informação dos *shots* em uma matriz **A**, onde cada coluna representa um quadro através de um vetor de feições. Esta matriz é decomposta por uma transformada chamada decomposição em valor singular (SVD) (JACKSON, 1991), $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. A partir da SVD deriva-se um espaço de feições refinado de menor dimensão. Como resultado, o algoritmo de *clustering* pode ser aplicado de maneira mais eficiente no novo espaço gerado.

3.3.4 Aplicações

Aplicações para técnicas de sumarização de vídeos digitais são muitas. Por exemplo, na literatura pode-se encontrar vários trabalhos referentes a análise de vídeos esportivos (SAHOURIA; ZAKHOR, 1999; LI; SEZAN, 2002; CHANG, 2002). Li et al. (LI; SEZAN, 2002) publicou um trabalho propondo algoritmos para a detecção automática de segmentos de vídeos em transmissões esportivas. Neste contexto, eventos importantes são definidos de acordo com cada esporte, tal como a bola em jogo no *football* americano e a rebatida com o taco no *baseball*.

Na área médica os trabalhos são mais recentes, conforme (EBADOLLAHI; CHANG; WU, 2001). Ebadollahi et al. propõe uma técnica para sumarização de vídeos de ecocardiografia. O método detecta e reconhece os diferentes segmentos do vídeo, extrai os quadros clinicamente importantes e gera representações concisas do conteúdo do vídeo. Deste modo, muitas atividades são facilitadas. Na medicina remota, por exemplo, sumários clínicos destes vídeos podem ser enviados via internet, mostrando somente informações importantes para especialistas remotos. No caso de diagnósticos auxiliados por computador, o sistema pode descobrir casos médicos que apresentam atributos espaço-temporais similares, revelando, assim, informações clínicas importantes. A Figura 3.3 mostra uma visão geral deste sistema.

3.4 Conclusões

Neste capítulo foram apresentados conceitos que fundamentam a sumarização de vídeos digitais. Além disso, discutiu-se o contexto atual desta área de pesquisa. De modo geral, a sumarização automática de vídeos digitais enfrenta problemas desafiadores que ainda não foram devidamente resolvidos. Um dos tópicos mais estudados objetiva a definição clara de uma ligação entre feições de baixo nível (cor, textura e movimento) e o conteúdo visual. Ainda sim, em domínios específicos, o problema pode ser simplificado em razão das necessidades das aplicações. Neste contexto encontram-se os vídeos médicos. Segundo Chang (CHANG, 2002), a análise de vídeos médicos é uma área de pesquisa promissora, onde as tarefas são bem definidas, tendo em vista que os médicos procuram analisar imagens e vídeos em função de diagnósticos e prognósticos. A semântica dos dados é definida por uma perspectiva clínica que, basicamente, diferencia

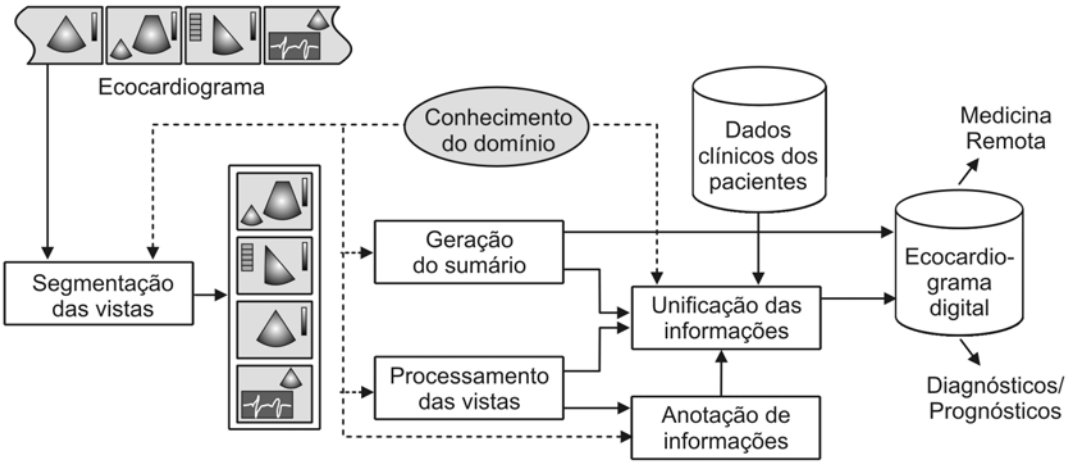


Figura 3.3: Diagrama do sistema de sumarização de vídeos de ecocardiografia proposto por Ebadollahi et al (EBADOLLAHI; CHANG; WU, 2001)

as situações normais das anormais, e classifica estes dados em categorias específicas. Por fim, observa-se que as técnicas propostas na literatura estão focadas em detecção de *shots*, cortes e cenas. Vídeos de histeroscopias não apresentam tal estrutura, e conseqüentemente devem ser abordados de maneira diferente.

4 FUNDAMENTOS: MOVIMENTO DE CÂMERA A PARTIR DE IMAGENS

Neste capítulo apresenta-se conceitos preliminares que fundamentam a abordagem proposta no capítulo 6. Além de conceitos geométricos envolvidos na representação de cenas estáticas 3-D, discute-se o emprego de abordagens utilizadas no processo de estimar movimento de câmera, uma vez que tais abordagens apresentam limitações e nem sempre são apropriadas em determinados contextos. Maiores detalhes sobre os conceitos apresentados a seguir são encontrados em (MA et al., 2003; HARTLEY; ZISSERMAN, 2000; TORR; FITZGIBBON; ZISSERMAN, 1999).

4.1 Conceitos e Abordagens Comumente Adotadas

Computar a estrutura de uma cena em 3 dimensões, juntamente com a posição da câmera no espaço, a partir de imagens tomadas de diferentes posições da própria cena é uma área de pesquisa conhecida como *structure-from-motion* (SFM). Detectar pontos potencialmente adequados para propósitos de rastreamento, estabelecer efetivamente a correspondência entre estes pontos através das imagens e validá-los geometricamente de acordo com restrições de movimentos rígidos são etapas típicas de abordagens propostas no contexto de SFM. Nesta seção apresenta-se detalhes sobre estas etapas, as quais são empregadas na metodologia proposta nesta tese.

4.1.1 Extração e Rastreamento de Feições

Um primeiro passo para computar a geometria de uma cena a partir de uma seqüência de imagens de vídeo consiste em estabelecer correspondências entre seus quadros. Usualmente, pontos nos quadros são detectados e rastreados/casados através de características de sua vizinhança, uma vez que nem toda região de uma imagem é igualmente apropriada para propósitos de rastreamento. Idealmente, estes pontos deveriam ser projeções do mesmo ponto no espaço, como ilustrado para o caso de duas imagens na Fig. 4.1, onde $\mathbf{x}$ e $\mathbf{x}'$ são projeções de $\mathbf{X}$. Contudo, este não é um problema de fácil resolução, uma vez que não se pode esperar que um ponto em uma imagem aparecerá com as mesmas características (isto é, as mesmas coordenadas de imagem e intensidade de pixels) nas imagens subseqüentes. Métodos usuais de rastreamento freqüentemente entregam ambigüidades e falsas correspondências (MA et al., 2003), e por esta razão os próximos passos em um processo de estimar a geometria de uma cena devem ser robustos para contornar as limitações de técnicas de rastreamento.

A extração de feições consiste em detectar pontos "interessantes"em cada quadro do vídeo. Essencialmente busca-se detectar pontos com boas propriedades de rastreabilidade, os quais podem ser detectados usando-se o detector de cantos de Harris (HARRIS; STEPHENS, 1988). Basicamente, o operador de Harris detecta cantos sobre uma pequena janela espacial onde há uma grande variação na direção do gradiente da imagem. Na prática, antes de cantos propriamente ditos, métodos que detectam cantos identificam pontos de uma maneira geral, uma vez que estes pontos são caracterizados pela presença de grandes diferenças de intensidade em sua vizinhança.

Formalmente, em cada imagem de uma seqüência, um conjunto do pontos $L = \{\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_M\}$ é extraído, onde $\mathbf{x}_i = (x_i, y_i)^\mathsf{T}$. Dois conjuntos L e L' detectados em duas imagens sucessivas podem ser casados usando-se o rastreador de pontos de Kanade-Lucas-Tomasi (KLT) (LUCAS; KANADE, 1981; SHI; TOMASI, 1994). Dadas duas imagens I e I' separadas por pontos de vista próximos e um ponto $\mathbf{x}_i$ em I, o algoritmo KLT iterativamente procura pela localização de $\mathbf{x'}_i$ em I' minimizando a diferença de intensidades entre janelas de tamanho fixo, W_i e W'_i, centradas em $\mathbf{x}_i$ e $\mathbf{x'}_i$ respectivamente. A versão mais simples do algoritmo KLT é baseada em um modelo de movimento de translação local, onde o deslocamento $d = (d_x, d_y)$ de cada ponto é estimado pela minimização de

$$\sum_{(x,y)\in W} [I'(x + d_x, y + d_y) - I(x, y)]^2. \tag{4.1}$$

Dimensões usuais para W são 5×5, 7×7 ou 9×9 pixels. Contudo, deve-se levar em consideração um equilíbrio na escolha destas dimensões: W deve ser tão grande quanto possível para evitar os efeitos provocados pelo ruído e deslocamentos mais amplos, mas também deve ser tão pequena quanto possível para aproximar deformações locais entre os quadros, uma vez que emprega-se um modelo simplificado de translação. Deste modo, para tratar deslocamentos maiores que as dimensões de W, e ao mesmo tempo manter o tamanho de W pequeno, uma representação piramidal/multi-escala do algoritmo KLT pode ser empregado (BOUGUET, 2000a). Neste processo as imagens I e I' são suavizadas e reamostradas sucessivas vezes. Então, o algoritmo KLT é aplicado primeiro em escalas mais baixas (menos detalhes), provendo uma estimativa de deslocamento d para imagens em escalas mais altas (mais detalhes). Este processo é sucessivamente aplicado a partir de imagens em níveis menos detalhados até imagens em níveis com mais detalhes, sendo que uma estimativa final para d é produzida de acordo com os critérios de equilíbrio mencionado acima.

O rastreador de pontos KLT produz um conjunto de pontos potencialmente correspondentes $\{\mathbf{x}_i \leftrightarrow \mathbf{x'}_i\}$ entre pares de imagens. Cada correspondência é computada independentemente das outras, sem considerar uma consistência global de movimento que envolveria todas as correspondências computadas para duas imagens. Isso é feito pela validação destes pontos de acordo com um modelo global de movimento, que é, neste trabalho, um modelo de movimentos rígidos induzidos pelo movimento de câmera. Assim, o processo de rastreamento de feições discutido nesta seção provê o conjunto inicial de pontos potencialmente correspondentes, os quais serão refinados em termos de um modelo de movimento de câmera, como descrito na seção 4.1.3.

4.1.2 Geometria de Cenas Estáticas 3-D

O processo de formação de imagens pode ser definido como uma projeção do espaço 3-D sobre o plano 2-D da imagem. Seguindo um modelo simples de camera *pinhole*, as

coordenadas de um ponto em 3-D no espaço $\mathbf{X} = (X, Y, Z, 1)^\mathsf{T}$ e sua correspondente projeção sobre o plano da imagem $\mathbf{x} = (x, y, 1)^\mathsf{T}$, ambos representados em coordenadas homogêneas, estão relacionados pela equação projetiva (MA et al., 2003)

$$\lambda \mathbf{x} = \mathbf{P}\mathbf{X} \tag{4.2}$$

onde λ é um fator de escala desconhecido (que é proporcional a profundidade de $\mathbf{X}$ relativa a câmera) e $\mathbf{P}$ é uma matriz 3×4 de projeção de câmera, que pode ser fatorada como:

$$\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}], \tag{4.3}$$

onde:

$$\mathbf{K} = \begin{bmatrix} f & s & x_0 \\ 0 & f & y_0 \\ 0 & 0 & 1 \end{bmatrix}$$

A matriz de calibração de câmera $\mathbf{K}$ mapeia coordenadas métricas em coordenadas de imagem (pixels). $\mathbf{K}$ contém os parâmetros internos (ou intrínsecos) de câmera, onde f representa a distância focal da câmera; $\mathbf{c} = [x_0, y_0]^\mathsf{T}$ é o ponto principal, que representa as coordenadas da imagem onde ocorre a intersecção do eixo ótico e o plano da imagem; s é referido como o fator de inclinação e diz respeito a formatos não retangulares de pixels no sensor CCD da câmera (s é bastante próximo de zero para a maioria das câmeras). A matriz 3×4 de parâmetros externos $[\mathbf{R}|\mathbf{t}]$ representa a orientação e a posição da câmera. $\mathbf{R}$ é uma matriz de rotação e $\mathbf{t}$ é um vetor de translação.

4.1.3 Movimento de Câmera a partir de Pontos Correspondentes em Duas Imagens

Considera-se a situação onde duas imagens são tomadas com uma câmera que move-se em relação a uma cena estática. Neste contexto, relações de movimento de câmera podem ser estabelecidas com base em pontos correspondentes detectados nas imagens. Pontos correspondentes $\mathbf{x} \leftrightarrow \mathbf{x}'$ são ilustrados nas Figuras 4.1 e 4.4. Esta seção dedica-se ao estudo das relações geométricas entre dois pontos de vista de uma cena. No centro da discussão está a *geometria epipolar*. Para um modelo de câmera *pinhole*, apresentado na seção anterior, se duas imagens da mesma cena são tomadas de diferentes posições, pontos correspondentes nas imagens satisfazem uma simples restrição geométrica conhecida como *restrição epipolar*. Além disso, dependendo de características espaciais da cena e do movimento realizado pela câmera, também é possível estabelecer uma relação entre pontos correspondentes através de uma transformação projetiva conhecida como *homografia*. Deste modo, geometria epipolar e homografia constituem dois modelos distintos de *movimento de câmera*, os quais podem explicar a relação de movimento que há entre duas imagens. A seguir apresenta-se detalhes sobre *geometria epipolar*, *homografias* e a relação que há entre estes conceitos.

4.1.3.1 Restrição Epipolar

Segundo um modelo de câmera *pinhole*, cada imagem está associada a uma matriz de projeção de câmera $\mathbf{P}$, onde um ponto 3-D do espaço $\mathbf{X}$ é projetado no plano da primeira imagem como $\mathbf{x} = \mathbf{P}\mathbf{X}$ e como $\mathbf{x}' = \mathbf{P}'\mathbf{X}$ na segunda imagem. Uma vez que $\mathbf{x}$ e $\mathbf{x}'$ são imagens do mesmo ponto do espaço 3-D, eles são entendidos como pontos correspondentes $\mathbf{x} \leftrightarrow \mathbf{x}'$. Assumindo que a câmera está calibrada, isto é $\mathbf{K}$ é a matriz

identidade, os pontos correspondentes $\mathbf{x}$ e $\mathbf{x}'$ satisfazem a restrição epipolar (LONGUET-HIGGINS, 1981)

$$\mathbf{x'}^{\mathsf{T}}\mathbf{E}\mathbf{x} = \mathbf{0}, \tag{4.4}$$

onde $\mathbf{E}$ é uma matriz 3×3 que carrega a posição relativa (rotação $\mathbf{R}$ e translação $\mathbf{t}$) entre as duas câmeras $\mathbf{C}$ e $\mathbf{C}'$, como ilustrado na Figura 4.1. $\mathbf{E}$ é conhecida como *matriz essencial* e, dada uma quantidade suficiente de pontos correspondentes, as entradas de $\mathbf{E}$ podem ser estimadas a partir de um conjunto de equações epipolares, como explicado ainda dentro desta seção.

Na maioria das situações práticas a matriz de calibração $\mathbf{K}$ não é uma matriz identidade e, conseqüentemente, ela deve ser estimada. A partir de uma boa aproximação de $\mathbf{K}$ é possível recuperar $\mathbf{E}$ com base em pontos correspondentes. Para isso deve-se remover os efeitos de $\mathbf{K}$ como segue

$$\mathbf{x'}^{\mathsf{T}}\mathbf{K}^{-\mathsf{T}}\mathbf{E}\mathbf{K}^{-1}\mathbf{x} = \mathbf{0}. \tag{4.5}$$

Situações nas quais não é possível recuperar $\mathbf{K}$ não são raras. Felizmente, a restrição epipolar é também válida para câmeras não calibradas e pode ser derivada diretamente da Equação 4.5 como

$$\hat{\mathbf{x}}'^{\mathsf{T}}\mathbf{F}\hat{\mathbf{x}} = \mathbf{0}. \tag{4.6}$$

onde $\hat{\mathbf{x}} = \mathbf{K}^{-1}\mathbf{x}$, significando que $\mathbf{x}$ não é mais dado em coordenadas métricas, mas em coordenadas de imagem (pixels). $\mathbf{F}$ é amplamente conhecida como *matriz fundamental*, onde $\mathbf{F} = \mathbf{K}^{-\mathsf{T}}\mathbf{E}\mathbf{K}^{-1}$ (MA et al., 2003; HARTLEY; ZISSERMAN, 2000).

Do ponto de vista geométrico, a restrição epipolar é uma condição de coplanaridade, como ilustrado na Figura 4.1. O vetor conectando os dois centros óticos $\overrightarrow{\mathbf{CC}'}$ e os vetores conectando estes centros óticos ao ponto $\mathbf{X}$ no espaço, $\overrightarrow{\mathbf{CX}}$ e $\overrightarrow{\mathbf{C}'\mathbf{X}}$, claramente formam um plano, conhecido como *plano epipolar*. Por esta razão, a restrição epipolar pode ser expressa como um produto escalar triplo

$$\overrightarrow{\mathbf{CC}'}.(\overrightarrow{\mathbf{CX}} \times \overrightarrow{\mathbf{C}'\mathbf{X}}) = 0, \tag{4.7}$$

o qual pode ser escrito como a Equação 4.4 (MIKHAIL; BETHEL; MCGLONE, 2001).

Conforme a Figura 4.1, o ponto $\mathbf{X}$ e suas projeções $\mathbf{x}$ e $\mathbf{x}'$ estão sobre o plano epipolar. Esta é uma importante propriedade que é explorada no processo de validação de pontos correspondentes, os quais irão dar suporte a estimativas de $\mathbf{E}$ (movimento de câmera). Para cada ponto $\mathbf{x}$ na primeira imagem, haverá uma *linha epipolar* correspondente l' na segunda imagem. A linha epipolar l' é determinada pela intersecção do plano epipolar com o plano que define a segunda imagem, como ilustrado na Figura 4.1. Assim, há um mapeamento projetivo $\{\mathbf{x} \mapsto l' \mid \mathbf{x}' \subset l'\}$ representado pela matriz essencial $\mathbf{E}$, onde

$$l' = \mathbf{E}\mathbf{x} \quad \text{e} \quad l = \mathbf{E}^{\mathsf{T}}\mathbf{x}'. \tag{4.8}$$

Os pontos e e e' são conhecidos como *epipólos* e são determinados pela intersecção da linha que conecta os centros de câmera (linha base) com os planos das imagens.

Dada uma quantidade de pontos correspondentes $\mathbf{x}_i \leftrightarrow \mathbf{x'}_i$, a restrição epipolar (Eq. 4.4) é válida para qualquer par destes pontos em uma cena estática e rígida. Assim, cada correspondência pode constituir uma restrição em $\mathbf{E}$. Uma vez que $\mathbf{E}$ é uma matriz 3×3 definida em função de um fator de escala arbitrário, tem-se $3 \times 3 - 1$ incógnitas. Por esta razão, 8 pares de pontos correspondentes são suficientes para computar as entradas de $\mathbf{E}$ linearmente. Esta é a essência do *algoritmo de oito pontos* (LONGUET-HIGGINS, 1981;

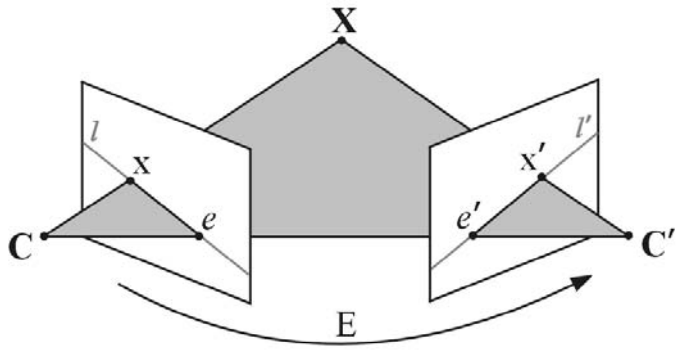


Figura 4.1: Projeções $\mathbf{x}$ e $\mathbf{x}'$ de um ponto no espaço $\mathbf{X}$ recaem sobre o mesmo plano, conhecido com plano epipolar. Esta é uma importante propriedade amplamente explorada na busca por correspondências que irão dar suporte a estimativas de movimento de câmera $\mathbf{E}$ entre $\mathbf{C}$ e $\mathbf{C}'$.

HARTLEY, 1997). Basicamente, a Equação 4.4 é reescrita em termos das coordenadas conhecidas de $\mathbf{x} = [x\ y\ 1]^\mathsf{T}$ e $\mathbf{x}' = [x'\ y'\ 1]^\mathsf{T}$:

$$[\ xx'\ yx'\ x'\ xy'\ yy'\ y'\ x\ y\ 1\ ]\ \mathbf{E}^s = 0 \tag{4.9}$$

onde $\mathbf{E}^s = [E_{11}\ E_{12}\ E_{13}\ E_{21}\ E_{22}\ E_{23}\ E_{31}\ E_{32}\ E_{33}]^\mathsf{T}$ é um vetor obtido pelo empilhamento das entradas da matriz $\mathbf{E}$. Assim, a partir de um conjunto de 8 pares de pontos correspondentes em duas imagens, é possível estruturar suas coordenadas em uma matriz $\mathbf{A}$ e, na ausência de ruído, o vetor $\mathbf{E}^s$ irá satisfazer

$$\begin{bmatrix} x_1x_1' & y_1x_1' & x_1' & x_1y_1' & y_1y_1' & y_1' & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ x_8x_8' & y_8x_8' & x_8' & x_8y_8' & y_8y_8' & y_8' & x_8 & y_8 & 1 \end{bmatrix} \mathbf{E}^s = \mathbf{A}\mathbf{E}^s = 0. \tag{4.10}$$

Este sistema de equações é facilmente resolvido usando-se decomposição em valor singular (SVD) (GOLUB; LOAN, 1989). Uma vez que a matriz $\mathbf{A}$ tenha sido formada a partir de 8 equações linearmente independentes, a solução $\mathbf{E}^s$ do sistema é a última coluna da matriz $\mathbf{V}$, onde $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ é a SVD de $\mathbf{A}$.

4.1.3.2 Homografias

Transformações projetivas são úteis para explicar relações de movimento de câmera sob certas configurações de cenas que aparecem na prática. Uma transformação projetiva planar $\mathbf{H}$ (também conhecida como colineação ou homografia) é um mapeamento linear representado por uma matriz 3×3 não-singular:

$$\mathbf{x}' = \mathbf{H}\mathbf{x}. \tag{4.11}$$

onde $\mathbf{x}$ e $\mathbf{x}'$ são pontos 2D representados em coordenadas homogêneas. A Figura 4.2 ilustra este mapeamento para o caso de uma transformação projetiva em perspectiva. Transformações projetivas em perspectiva são de especial interesse quando as imagens são capturadas segundo um modelo de câmera pinhole, apresentado na seção 4.1.2. Este tipo de transformação caracteriza-se por um ponto central de projeção $\mathbf{C}$ (centro ótico

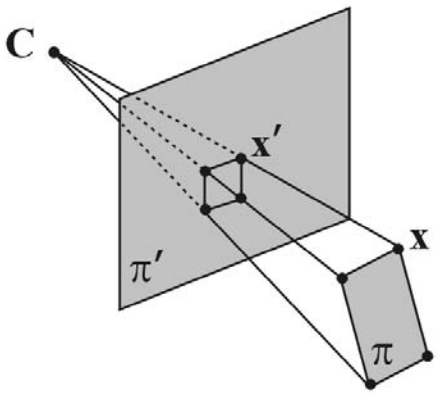


Figura 4.2: Projeções ao longo de raios que passam por um ponto comum $\mathbf{C}$ (o centro de projeção) definem um mapeamento entre planos. Se um sistema de coordenadas é definido em cada plano e pontos são representados em coordenadas homogêneas, então este mapeamento pode ser expresso como uma transformação projetiva $\mathbf{H}$, onde $\mathbf{x}' = \mathbf{H}\mathbf{x}$.

de uma câmera), onde os raios projetivos convergem para este ponto. Raios projetivos através de $\mathbf{C}$ definem um mapeamento de um plano para outro, conforme ilustrado na Figura 4.2 para os planos π e π', onde $\mathbf{x}' = \mathbf{H}\mathbf{x}$.

O processo de formação de uma imagem ocorre pela projeção de uma cena sobre o plano da imagem. Este processo está ilustrado na Figura 4.2, onde o plano π' representa o plano da imagem, o plano π representa um objeto genérico da cena e o ponto $\mathbf{C}$ representa a posição da câmera (o centro ótico) em relação a cena. Sendo assim, a Figura 4.4 ilustra duas situações onde o movimento de câmera pode ser explicado através de uma transformação projetiva $\mathbf{H}$ entre planos de imagens associados a dois pontos de vista.

Na Figura 4.4(a) observa-se o caso de uma cena planar, o que simplifica para uma transformação projetiva planar a relação existente entre pontos correspondentes nas imagens ($\mathbf{x}$ e $\mathbf{x}'$). A razão para isso é que a relação entre o espaço 3D (representado por uma superfície planar π) e o plano da imagem também pode ser explicada por uma transformação projetiva planar $\mathbf{H}$, conforme a Equação 4.11. Sendo assim, uma transformação projetiva perspectiva $\mathbf{H}_1$ explica a relação entre $\mathbf{x}$ (plano da imagem 1) e $\mathbf{X}$ (plano π) como $\mathbf{x} = \mathbf{H}_1\mathbf{X}$, da mesma forma que uma transformação projetiva perspectiva $\mathbf{H}_2$ explica a relação entre $\mathbf{x}'$ (plano da imagem 2) e $\mathbf{X}$ (plano π) como $\mathbf{x}' = \mathbf{H}_2\mathbf{X}$. Deste modo, a relação $\mathbf{H}$ entre os pontos das imagens $\mathbf{x}$ e $\mathbf{x}'$ pode ser explicada como uma composição de duas transformações projetivas perspectivas, onde $\mathbf{x}' = \mathbf{H}_2\mathbf{H}_1^{-1}\mathbf{x} = \mathbf{H}\mathbf{x}$ (HARTLEY; ZISSERMAN, 2000). Este mapeamento de $\mathbf{x}$ para $\mathbf{x}'$ é comumente conhecido na literatura como uma homografia induzida pelo plano π.

Na Figura 4.4(b) observa-se o caso em que a câmera executa apenas movimentos de rotação em torno de seu eixo ótico, não havendo translação. Nesta configuração de cena a relação entre os pontos $\mathbf{x}$ e $\mathbf{x}'$ resume-se a um mapeamento direto entre os planos das imagens (onde $\mathbf{x}' = \mathbf{H}\mathbf{x}$) na forma de uma transformação projetiva perspectiva $\mathbf{H}$, conforme discutido no início desta seção.

4.1.3.3 Homografias Compatíveis com Geometria Epipolar

Supondo a situação mostrada na Figura 4.4(a), onde pontos $\mathbf{X}_i$ tomados no espaço

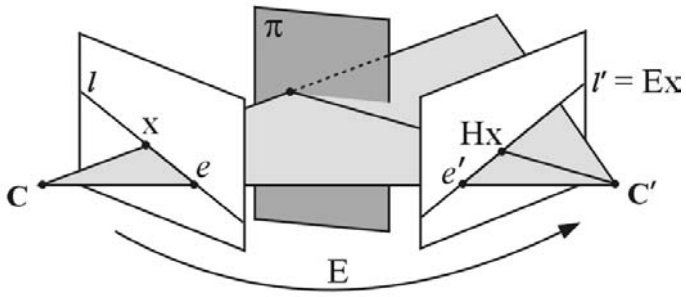


Figura 4.3: Relação entre uma geometria epipolar $\mathbf{E}$ e uma homografia $\mathbf{H}$ induzida pelo plano π. Um ponto $\mathbf{x}$ mapeado pela homografia recai sobre sua linha epipolar correspondente $l' = \mathbf{E}\mathbf{x}$.

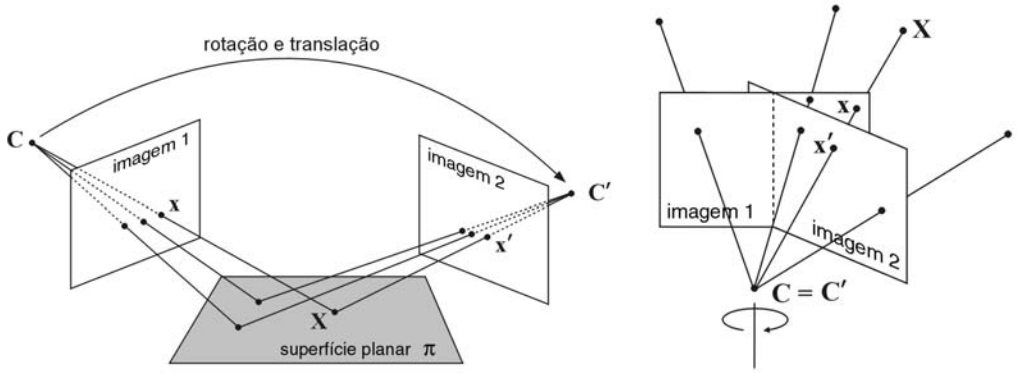
são coplanares. Deste modo, as correspondências $\mathbf{x}_i \leftrightarrow \mathbf{x'}_i$ geradas nas imagens definem uma homografia $\mathbf{H}$, que é induzida pelo plano π, onde $\mathbf{x'}_i = \mathbf{H}\mathbf{x}_i$, conforme discutido na seção 4.1.3.2. Estas correspondências também obedecem a restrição epipolar, isto é $\mathbf{x'}_i^{\mathsf{T}}\mathbf{E}\mathbf{x}_i = 0$. A Figura 4.3 ilustra a relação de uma homografia $\mathbf{H}$ que é dita consistente com a geometria epipolar $\mathbf{E}$, onde qualquer ponto $\mathbf{x}$ mapeado pela homografia recai sobre sua linha epipolar correspondente $l' = \mathbf{E}\mathbf{x}$ (HARTLEY; ZISSERMAN, 2000; MA et al., 2003).

Contudo, estimar uma geometria epipolar $\mathbf{E}$ a partir de correspondências $\mathbf{x}_i \leftrightarrow \mathbf{x'}_i$, que são projeções de pontos $\mathbf{X}_i$ espacialmente coplanares, resulta em uma família de possíveis matrizes essenciais $\mathbf{E}$ como solução do sistema de equações 4.10, onde a matriz de equações $\mathbf{A}$, derivada do conjunto de correspondências, deve ter um posto de no máximo 6 (HARTLEY; ZISSERMAN, 2000). Sendo assim, os dados $\mathbf{x}_i \leftrightarrow \mathbf{x'}_i$ não provêem restrições suficientes para se computar $\mathbf{E}$ como uma solução única derivada do sistema de equações 4.10, caracterizando uma situação conhecida como uma configuração degenerada de cena.

4.1.4 Configurações Degeneradas de Cena

A tarefa de rastrear pontos consistentes com o movimento de câmera em seqüências de imagens é importante no contexto deste trabalho, conforme discute-se na seção 6.1. Na literatura pode-se encontrar alguns métodos com este propósito (BEARDSLEY; ZISSERMAN; MURRAY, 1997; FITZGIBBON; ZISSERMAN, 1998; NISTéR, 2000). Um componente importante destes métodos é o uso da geometria epipolar para estabelecer pontos geometricamente correspondentes entre pares de imagens. Usualmente, computa-se um conjunto inicial de potenciais correspondências através de algoritmos como o rastreador KLT (seção 4.1.1), sendo que logo após estas correspondências são validadas geometricamente de acordo com uma geometria epipolar $\mathbf{E}$ estimada (seção 4.1.3.1). Contudo, há algumas configurações de cena na qual esta validação geométrica falha. Por esta razão, estimar a matriz essencial $\mathbf{E}$ a partir de pontos correspondentes requer algumas suposições sobre o movimento de câmera e a estrutura da cena em si. Configurações de cenas em desacordo com estas suposições são conhecidas como *configurações degeneradas de cena*.

Dois tipos de configurações degeneradas de cena aparecem mais freqüentemente na prática: *degeneração de movimento* (Figura 4.4(b)), onde a câmera executa apenas ro-



(a) (b)

Figura 4.4: Configurações degeneradas de cena onde uma transformação projetiva $\mathbf{H}$ explica a relação entre as imagens, onde $\mathbf{x}' = \mathbf{H}\mathbf{x}$. (a) Cena Planar. (b) Cena em que a câmera executa apenas rotações em torno de seu eixo ótico.

tações em torno de seu eixo ótico (não há translação) e *degeneração estrutural* (Figura 4.4(a)), onde a estrutura espacial da cena observada é planar ou, de uma forma aproximada, a variação de profundidade dentro da cena é pequena quando comparada a distância entre a câmera e a própria cena (TORR; FITZGIBBON; ZISSERMAN, 1999), como ocorre com imagens aéreas por exemplo.

Em ambos casos de degenerações a geometria epipolar não pode ser determinada (TORR; FITZGIBBON; ZISSERMAN, 1999). Este fato verifica-se claramente em casos de degenerações de movimento (Figura 4.4(b)), uma vez que não há translação de câmera e, conseqüentemente, os centros de câmera $\mathbf{C}$ e $\mathbf{C}'$ coincidem no espaço. Por outro lado, no caso de degeneração estrutural, um conjunto de pontos correspondentes degenerados, os quais são projeções de pontos espacialmente coplanares, não provêem restrições suficientes para se computar unicamente $\mathbf{E}$ (Equação 4.10), havendo assim uma família de relações de movimento que podem explicar as correspondências igualmente bem (HARTLEY; ZISSERMAN, 2000). Contudo, observa-se que para ambos os casos de degenerações uma homografia $\mathbf{H}$ pode ser empregada, em lugar de uma matriz essencial $\mathbf{E}$, para explicar a relação de movimento de câmera entre os pontos correspondentes $\mathbf{x} \leftrightarrow \mathbf{x}'$ nas imagens, uma vez que estes pontos estão em conformidade com as configurações de cenas descritas na seção 4.1.3.2. Deste modo, a estratégia parece ser utilizar um modelo de movimento de câmera baseado em homografias $\mathbf{H}$, em lugar de um modelo epipolar $\mathbf{E}$, sempre que configurações degeneradas de cena são detectadas (TORR, 1997; CHUM; WERNER; MATAS, 2005; POLLEFEYS; VERBIEST; VAN GOOL, 2002).

Na ausência de ruído, a tarefa de detectar degenerações estruturais e de movimento não seria problemática. No processo de estimar $\mathbf{E}$ a partir de pontos correspondentes, degenerações estruturais e de movimento podem ser matematicamente expressas em termos do posto da matriz $\mathbf{A}$, conforme a Equação 4.10. Em situações livres de ruído a matriz $\mathbf{A}$ deve ter posto 8 para prover uma solução única, e posto 6 quando $\mathbf{A}$ é derivada de correspondências degeneradas (HARTLEY; ZISSERMAN, 2000). Contudo, partindo de dados reais (ruidosos), o problema torna-se bastante complicado, pois as restrições não providas por equações degeneradas podem ser determinadas pelo ruído.

Os efeitos de assumir configurações não-degeneradas em situações degeneradas incluem a perda de pontos que podem estar sendo consistentemente rastreados, além de

permitir a inclusão de pontos inconsistentes devido ao sobre-ajuste do modelo geométrico estimado, como detalhado em (TORR; FITZGIBBON; ZISSERMAN, 1999). No contexto deste trabalho este fato reduziria a consistência do método proposto como um todo, uma vez que a abordagem proposta fundamenta-se no comportamento de pontos rastreados ao longo de uma seqüência de imagens.

Computar o movimento de câmera em configurações de cena que envolvem combinações de translações e rotações bem como configurações degeneradas decorrentes da estrutura espacial da cena são discutidos com profundidade em (VIéVILLE; LINGRAND, 1999; LUONG; FAUGERAS, 1996; MAYBANK, 1992).

4.1.5 Correção de Distorções de Lente e Calibração da Câmera

Normalmente, o campo de visão de endoscópios é estreito e lentes de amplo-ângulo de visão são empregadas para contornar esta limitação. Contudo, lentes de amplo-ângulo causam distorção radial nas imagens capturadas através destas lentes. Essa distorção faz com que pontos em uma imagem sejam deslocados em direções radiais a partir do centro de distorção, o qual normalmente coincide com o centro da imagem. Deste modo, a distorção é menor em regiões próximas ao centro da imagem e cresce em direção às bordas da imagem. Neste contexto, um modelo linear (pinhole) de câmera (Equação. 4.3) não pode ser diretamente aplicado para, por exemplo, propósitos de estimar movimento de câmera, uma vez que a imagem $\mathbf{x} = (x, y)$ de um ponto $\mathbf{X}$ no espaço é projetada além da posição determinada por uma projeção perspectiva, como aquela induzida por uma câmera pinhole.

Abordagens que tratam com movimentos de câmera usualmente apresentam um processo preliminar para estimar a distorção de lente a fim de permitir que passos subseqüentes sejam desenvolvidos sob o contexto de um modelo linear de câmera. Em termos de lentes endoscópicas, a relação entre a posição distorcida de um ponto (x_d, y_d) na imagem e sua posição corrigida (x, y) pode ser modelada como um polinômio (SHAHIDI et al., 2002)

$$x = x_d(1 + k_1 r^2 + k_2 r^4 + ...)$$
$$y = y_d(1 + k_1 r^2 + k_2 r^4 + ...)$$

onde $r^2 = x_d^2 + y_d^2$ e k_i são coeficientes de distorção. Vários autores relatam uma precisão abaixo de 0.5 pixels considerando até dois coeficientes para distorção radial (k_1 and k_2). Na prática, estes coeficientes podem ser estimados usando-se métodos bem estabelecidos de calibração de câmera (ZHANG, 1999; TSAI, 1987; HEIKKILä; SILVéN, 1997), os quais estão disponíveis como pacotes de software. Além disso, o propósito de procedimentos de calibração de câmera é estimar os parâmetros internos de câmera (isto é a matriz $\mathbf{K}$), permitindo a remoção de seus efeitos de, por exemplo, pontos correspondentes $\mathbf{x} \leftrightarrow \mathbf{x}'$ em uma seqüência de imagens. Assim, a partir do procedimento de calibração de câmera estima-se $\mathbf{K}$ e os parâmetros de distorção de lente, deixando pontos correspondentes com coordenadas corrigidas para servirem de entrada para algoritmos lineares que estimam o movimento de câmera, como por exemplo, o algoritmo de oito pontos (Equação 4.10).

Neste trabalho utiliza-se o pacote (BOUGUET, 2000b) de calibração de câmera que é baseado no trabalho de Heikkilä e Silvén (HEIKKILä; SILVéN, 1997). Resumidamente, imagens tomadas de diferentes posições de um objeto plano, com um padrão quadriculado (como um tabuleiro de xadrez) de dimensões conhecidas, é utilizado para estimar os

parâmetros internos e coeficientes de distorção. Em cada imagem, os cantos internos do padrão quadriculado são detectados automaticamente e, uma vez que suas reais posições são conhecidas, as entradas da matriz de câmera $\mathbf{P}$ podem ser estimadas como incógnitas na Equação 4.2.

4.2 Seleção Robusta do Modelo de Movimento de Câmera

Dado um conjunto com M pontos correspondentes $\{\mathbf{x}_i \leftrightarrow \mathbf{x'}_i\}_{i=1}^{M}$ entre duas imagens, estimar uma relação de movimento de câmera T entre estas imagens é um problema amplamente estudado em visão computacional. Geralmente imagens reais contém ruído e assim não é possível tratar com correspondências perfeitas entre as imagens. Além disso, métodos que estabelecem potenciais correspondências podem entregar falsas combinações de pontos devido a ambigüidades na descrição das características dos pontos. O algoritmo RANSAC (*RANdom SAmple Consensus*) (FISCHLER; BOLLES, 1981) é uma metodologia comumente empregada para tratar a presença de combinações incorretas. O RANSAC pode estimar uma relação de movimento T, que melhor se ajusta aos dados, e conseqüentemente computar as correspondências que são consistentes com T. Pontos correspondentes que dão suporte a relação de movimento estimada T são denominados de *inliers*, o restante dos pontos são *outliers*.

Na seqüência adota-se a seguinte notação para representar os conjuntos de dados trabalhados na abordagem RANSAC:

- $\{p\}$: por simplicidade adota-se $\{p\}$ para representar o conjunto de pontos correspondentes $\{\mathbf{x}_i \leftrightarrow \mathbf{x'}_i\}_{i=1}^{M}$ pré-estabelecidos entre duas imagens I e I', onde M é o total de correspondências detectadas. Além de ruído, correspondências incorretamente estabelecidas podem fazer parte de $\{p\}$;

- $\{in\}$: refere-se ao subconjunto de $\{p\}$ que dá suporte a uma determinada relação de movimento estimada T (conjunto de inliers segundo T);

- $\{out\}$: refere-se ao subconjunto de $\{p\}$ que não dá suporte a uma determinada relação de movimento estimada T (conjunto de outliers segundo T).

O algoritmo RANSAC é simples e pode ser descrito nos seguintes passos:

1. Seleciona-se aleatoriamente m pares de pontos correspondentes a partir de um conjunto de potenciais correspondências $\{p\}$ e computa-se uma relação candidata T_c com base nessa amostra de m correspondências. Usualmente, m é a quantidade mínima de elementos necessários para computar a relação T.

2. Aplica-se T_c a $\{p\}$ e classifica-se cada potencial correspondência usando-se um limiar de erro. Tem-se então o conjunto de inliers $\{in_c\}$ e o conjunto de correspondências classificadas como outliers $\{out_c\}$ segundo T_c.

3. A melhor relação candidata T_c é aquela que gera o maior conjunto (consenso) de inliers.

4. Os passos de 1 a 2 repetem-se até uma quantidade suficiente de amostragens ter sido avaliada, ou até que uma desejada probabilidade ρ de que uma boa relação candidata tenha sido computada.

Geralmente é desnecessário, e algumas vezes infactível, testar todas as combinações de m amostras. Deste modo, a quantidade de iterações (*trials*) S de um processo RANSAC padrão pode ser determinada a partir de resultados teóricos. Usualmente, S é um valor escolhido suficientemente alto para assegurar com uma probabilidade ρ que pelo menos uma das amostragens aleatórias de m pontos está livre de combinações incorretas de pontos. Quando isso ocorre, a relação T resultante é provavelmente útil. Sendo assim, ρ pode ser entendido como a probabilidade do algoritmo produzir um resultado útil. Valores usuais para ρ estão na faixa de 0.95 - 0.99.

Seja w a probabilidade de escolher uma correspondência corretamente estabelecida (potencial inlier) a partir de $\{p\}$. Por exemplo, dado que a relação de movimento correta T é conhecida, teria-se

$$w = \frac{|\{in\}|}{|\{p\}|}, \tag{4.12}$$

onde $|\cdot|$ denota a quantidade de elementos do conjunto. Assumindo que as m correspondências necessárias para estimar T são selecionadas independentemente a partir de $\{p\}$, w^m é a probabilidade de que todas estas m amostras sejam correspondências corretamente estabelecidas (potenciais inliers) e, conseqüentemente, $1 - w^m$ é a probabilidade de que pelo menos uma das m correspondências seja uma correspondência incorretamente estabelecida (um potencial outlier), o que implicaria em uma estimativa incorreta de T a partir deste conjunto de m correspondências. Deste modo, para S amostragens aleatórias de m correspondências, $(1 - w^m)^S$ é a probabilidade do algoritmo nunca selecionar um conjunto de m correspondências onde todas elas são corretamente estabelecidas. Nestes termos, $(1 - w^m)^S$ deve ser igual a $1 - \rho$, ou seja, $(1 - w^m)^S = 1 - \rho$, de onde pode-se isolar a quantidade S de iterações de um processo RANSAC como

$$S = \frac{\log(1 - \rho)}{\log(1 - w^m)}. \tag{4.13}$$

Na prática w não é conhecido a priori. Nestes casos o algoritmo é inicializado com uma estimativa de pior caso para w, sendo que esta estimativa pode ser atualizada cada vez que um conjunto maior de inliers é encontrado. Por exemplo, se a suposição de pior caso é $w = 0,5$ e um conjunto de inliers com 70% dos dados é encontrado em alguma iteração do algoritmo RANSAC, então uma estimativa atualizada é $w = 0,7$. Deste modo, dada uma probabilidade desejada $\rho = 0.99$ de que pelo menos uma das amostragens aleatórias de m pontos está livre de combinações incorretas de pontos, a quantidade S de iterações do algoritmo RANSAC pode ser adaptativamente computada conforme a Equação 4.13, onde w é atualizado, conforme a Equação 4.12, para cada iteração que aponte uma quantidade de inliers maior do que a maior estimativa corrente. Assim, o algoritmo termina logo que S amostragens aleatórias tenham sido executadas, o que pode ocorrer quando uma atualização de w resulte em um valor de S menor do que a quantidade de amostragens executadas até então. Este processo adaptativo para computar S é sumarizado no algoritmo 1.

A quantidade de iterações S cresce com a taxa de correspondências incorretamente estabelecidas em $\{p\}$, porém é fácil observar que esta taxa mantém-se baixa para valores razoáveis de ρ, w, e m.

A quantidade mínima $m = \lceil \frac{n}{r} \rceil$ de elementos necessários para computar uma relação T depende do número r de restrições provida por cada elemento e do número n de parâmetros da relação que deseja-se estimar. Por exemplo, dados pelo menos $m = 8$ pontos

Algorithm 1 Algoritmo adaptativo para determinar o número de amostragens S em um processo RANSAC

$S \Leftarrow \infty$
amostragens $\Leftarrow 0$
while $S >$ amostragens **do**
  Selecionar amostras e contar o número de inliers
  Atualizar w conforme Eq. 4.12
  Atualizar S a partir de w conforme Eq. 4.13 com $\rho = 0.99$
  amostragens $\Leftarrow$ amostragens + 1
end while

correspondentes, é possível computar linearmente uma solução para uma matriz essencial $\mathbf{E}$ (assumindo um parâmetro de escala arbitrário), uma vez que cada par de pontos correspondentes gera uma equação/restrição linear ($r = 1$) nas entradas de $\mathbf{E}$ ($n = 8$ parâmetros), como mostrado na Equação 4.10 e discutido na seção correspondente.

No passo 2 do algoritmo RANSAC menciona-se um limiar de erro que irá determinar se um par de pontos correspondentes $\mathbf{x}_i \leftrightarrow \mathbf{x'}_i$ será um inlier ou um outlier. Em termos de uma matriz essencial, esta classificação pode ser determinada com base em um limiar τ de erro sobre uma métrica como a *distância epipolar simétrica* (HARTLEY; ZISSERMAN, 2000). Esta métrica mede o quão próximo o par de pontos correspondentes $\mathbf{x}_i \leftrightarrow \mathbf{x'}_i$ satisfaz a restrição epipolar (Eq. 4.4). A distância epipolar simétrica é baseada nas Equações 4.8 e considera a distância d (em pixels) de um ponto $\mathbf{x'}$ a sua linha epipolar projetada $l' = \mathbf{E}\mathbf{x}$, como ilustrado na Fig. 4.1. Deste modo, dada uma estimativa de $\mathbf{E}^s$, um par de pontos correspondentes $\mathbf{x}_i \leftrightarrow \mathbf{x'}_i$ é classificado com inlier se

$$d(\mathbf{x'}_i, \mathbf{E}\mathbf{x}_i) + d(\mathbf{x}_i, \mathbf{E}^\mathsf{T}\mathbf{x'}_i) < \tau. \tag{4.14}$$

Além dos problemas envolvendo ruído e a presença de correspondências incorretas nos dados de entrada, se somente correspondências degeneradas estão presentes em $\{p\}$ não é possível computar uma relação correta T. Como discutido na seção 4.1.4, correspondências degeneradas significam que não há restrições suficientes para se computar uma relação/solução única. Em resumo, um método deveria ser capaz de detectar configurações degeneradas de cenas a partir de dados ruidosos e ainda tratar a presença de correspondências incorretas nos dados de entrada. No sentido de computar uma matriz essencial ou matriz fundamental (Equações. 4.4 e 4.6 respectivamente), algumas abordagens tem sido propostas (TORR, 1997; CHUM; WERNER; MATAS, 2005; FRAHM; POLLEFEYS, 2006). Basicamente, dado um conjunto de pontos correspondentes, estas abordagens propõem uma seleção automática entre dois modelos que melhor explicam a relação de movimento de câmera: uma homografia $\mathbf{H}$ para cenas degeneradas e uma matriz essencial $\mathbf{E}$ para situações em que configurações de cenas não-degeneradas são detectadas.

Neste trabalho adota-se uma abordagem proposta por Frahm e Pollefeys (FRAHM; POLLEFEYS, 2006) conhecida como QDEGSAC. Este método destaca-se pelo fato de não requerer qualquer informação sobre os tipos de degenerações que ocorrem nos dados de entrada. Assim, o desempenho do algoritmo QDEGSAC é avaliado na detecção automática de configurações degeneradas de cenas de histeroscopias e na seleção do modelo de movimento de câmera apropriado para pares de quadros em longas seqüências de imagens histeroscópicas. Por esta razão, detalha-se a seguir o método QDEGSAC em termos do algoritmo de oito pontos, e explica-se seu comportamento no contexto da seleção de

um modelo apropriado de movimento de câmera. Este é um importante passo no processo de rastreamento de pontos geometricamente consistentes ao longo de quadros de um vídeo e, conseqüentemente, em estimar alterações dentro do campo de visão devido a movimentos de câmera.

4.2.1 O Algoritmo QDEGSAC

O QDEGSAC foi originalmente motivado pelas limitações do RANSAC em estimar a relação correta quando os dados são quase-degenerados, o que significa que a maioria dos pontos correspondentes não provêem restrições suficientes para se computar a relação unicamente (dados degenerados), sendo que somente uma pequena fração dos pontos provêem as restrições restantes. Cenas predominantemente planares são exemplos práticos deste tipo de situação, onde a maioria dos pontos correspondentes detectados nas imagens são projeções de pontos espacialmente coplanares (parte degenerada da cena, ou parte degenerada dos dados), sendo que uma pequena parte das correspondências é oriunda de projeções de pontos dispostos em posições genéricas no espaço (pontos não coplanares), o que determina a parte não-degenerada dos dados/cena. Para dados quase-degenerados a relação pode sempre ser definida unicamente, porém o algoritmo RANSAC apresenta uma baixa probabilidade de computar a relação correta neste caso, como discutido em (FRAHM; POLLEFEYS, 2006) e (CHUM; WERNER; MATAS, 2005).

O algoritmo QDEGSAC emprega a metodologia RANSAC para computar uma solução e explora o número de restrições providas pela matriz de dados de entrada $\mathbf{A}$. O QDEGSAC pode ser aplicado em vários problemas em visão computacional, contudo neste trabalho foca-se em estimativas lineares da matriz essencial. Sendo assim, a matriz $\mathbf{A}$ é dada em termos do algoritmo de oito pontos, isto é, conforme a Equação 4.10. Logo, a relação $\mathbf{E}^s$ é determinada como o espaço-nulo da matriz de dados $\mathbf{A}$. Como mencionado na seção 4.1.4, $\mathbf{A}$ deveria ter um posto $r_{\mathbf{A}}$ de 8 para se obter uma solução não trivial para a Equação 4.10. Na ausência de ruído, $r_{\mathbf{A}} = 8$ significa que os dados proveram 8 restrições linearmente independentes e, conseqüentemente, as entradas do vetor $\mathbf{E}^s$ podem ser computadas unicamente (com um parâmetro arbitrário de escala) como o espaço-nulo 1-dimensional de $\mathbf{A}$ (HARTLEY; ZISSERMAN, 2000). Neste caso os dados são denominados aqui como não-degenerados. Por outro lado, se $r_{\mathbf{A}} < 8$, uma quantidade menor de restrições independentes é provida pelos dados e a solução $\mathbf{E}^s$ torna-se ambigüa, caracterizando-se assim uma configuração de cena degenerada. Degenerações de movimento e degenerações estruturais aparecem freqüentemente em vídeos de histeroscopias, como discutido nas seções 2.2.3 e 4.1.4. Ambas são caracterizadas por $r_{\mathbf{A}} = 6$ e, assim, a relação $\mathbf{E}$ degenera em uma homografia $\mathbf{H}$ (TORR; FITZGIBBON; ZISSERMAN, 1999).

Com base na discussão acima, o posto da matriz de dados $\mathbf{A}$ pode ser usado para detectar dados degenerados em casos de ausência de ruído. Na prática contudo, computar o posto de $\mathbf{A}$ pode resultar em valores incorretos, uma vez que este cálculo é sensível à presença de ruído nos dados de entrada. Os distúrbios causados pelo ruído resultam em pequenos valores singulares, sendo assim ainda parece ser possível estimar o posto de $\mathbf{A}$ utilizando-se um limiar apropriado sobre os valores singulares. Contudo, se além de inliers degenerados, os dados contêm outliers, é possível que estes últimos aumentem o posto $r_{\mathbf{A}}$ da matriz de dados de maneira que $r_{\mathbf{A}}$ seja igual ao valor de posto esperado para uma configuração de cena não-degenerada. Por esta razão, a ambigüidade não pode ser detectada somente pela análise dos valores singulares da matriz de dados $\mathbf{A}$.

Ainda assim, o algoritmo QDEGSAC pode ser interpretado como uma maneira robusta de medir o posto $r_{\mathbf{A}}$ da matriz de dados $\mathbf{A}$. Para dados degenerados o algoritmo sele-

ciona a relação apropriada para representar os dados. Se os dados são quase-degenerados, o algoritmo busca por inliers adicionais dentro do conjunto inicial de outliers, os quais são computados em um processo RANSAC inicial, como explicado nas seções seguintes. Isso é feito com o objetivo de identificar o maior valor possível de posto $r_{\mathbf{A}}$ que seja capaz de explicar adequadamente os dados.

O algoritmo QDEGSAC consiste em três fases conforme ilustrado na Figura 4.5:

1. O *primeiro RANSAC* estima a relação assumindo que os dados são não-degenerados (isto é, assumindo que $r_{\mathbf{A}} = 8$). A partir deste primeiro passo, os dados são classificados em inliers $\{in_8\}$ e outliers $\{out_8\}$.

2. Na seqüência o posto da matriz de dados é estimado robustamente a partir dos inliers $\{in_8\}$ computados no primeiro RANSAC. Esta fase é denominada de *seleção do modelo* na Figura 4.5, e determina o mais baixo valor de posto possível para a parte degenerada dos dados. Isso é feito mesmo no caso de possíveis correspondências incorretamente estabelecidas determinarem as restrições restantes para um modelo com $r_{\mathbf{A}} = 8$.

3. Por fim, a fase de *complementação do modelo* inspeciona os outliers, computados nas fases anteriores, tentando encontrar inliers adicionais para prover o maior valor de posto $r_{\mathbf{A}}$ possível para a matrix $\mathbf{A}$. Deste modo, se os dados são quase-degenerados, busca-se no conjunto inicial de outliers por inliers adicionais não-degenerados que possam prover as restrições restantes para se computar os 8 graus de liberdade de uma relação $\mathbf{E}^s$.

Nas seções seguintes são apresentados detalhes de cada uma das fases mencionadas acima.

4.2.1.1 Primeiro RANSAC

Nos moldes do algoritmo de oito pontos, um processo RANSAC é executado para computar-se uma relação T com 8 graus de liberdade. Este processo é denominado aqui como RANSAC(8). Deste modo são tomadas amostras aleatórias de $m = 8$ pontos correspondentes para formar matrizes candidatas $\mathbf{A}$ como na Equação 4.10. O processo RANSAC(8) produz uma relação $T_{RANSAC,8}$ que espera-se que empregue 8 restrições, classificando o conjunto de potenciais correspondências em inliers $\{in_8\}$ e outliers $\{out_8\}$.

Para dados não-degenerados, o processo RANSAC(8) entregará a relação correta T com uma probabilidade que depende do limiar de confiança ρ previamente definido. No caso de haver somente amostras de dados degenerados, o posto $r_{\mathbf{A}}$ deveria ser menor que 8. Este é o resultado esperado quando um número insuficiente de restrições é provido por uma dada amostra de 8 pontos correspondentes na forma da matriz $\mathbf{A}$ (Equação 4.10). No caso de uma amostra conter dados quase-degenerados, o posto $r_{\mathbf{A}}$ deveria ser igual a 8. Para uma amostra contendo dados degenerados e outliers, o posto $r_{\mathbf{A}}$ deveria ser também igual a 8, uma vez que os outliers provêem as restrições que complementam aquelas restrições providas pelos dados degenerados. É importante notar que amostras que contêm somente dados degenerados, ou dados degenerados juntamente com outliers, entregam uma alta quantidade de inliers, como discutido (FRAHM; POLLEFEYS, 2006). Deste modo, é possível que o processo RANSAC(8) entregue uma relação $T_{RANSAC,8}$ e pare sem alcançar a probabilidade de que uma relação correta tenha sido computada, conforme o critério apresentado na Equação 4.13. Por esta razão, os esperados 8 graus de liberdade

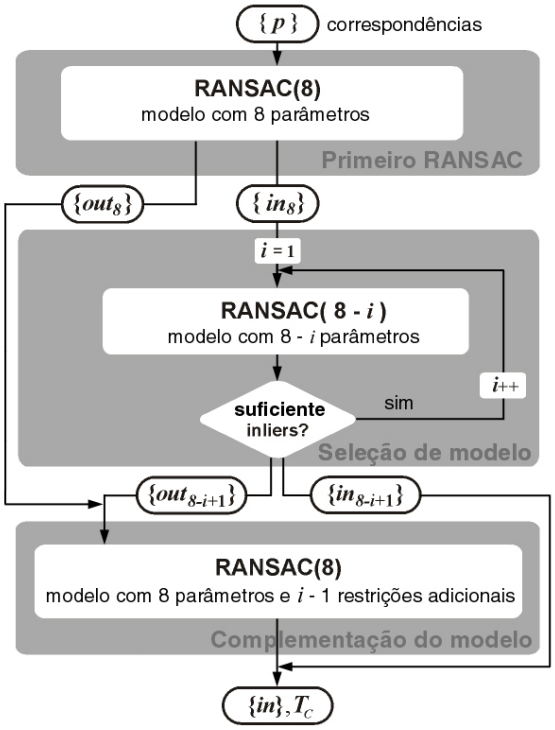


Figura 4.5: Visão geral do algoritmo QDEGSAC.

da relação $T_{RANSAC,8}$ podem não ser estabelecidos corretamente pelas restrições providas pelo conjunto de inliers computados, uma vez que podem haver restrições que resultaram do ruído presente em amostras degeneradas ou a partir de outliers. Assim, o objetivo dos próximos passos do algoritmo QDEGSAC é detectar os casos onde $r_{\mathbf{A}} < 8$, sendo que isso é feito através de testes realizados sobre o conjunto de inliers $\{in_8\}$ e outliers $\{out_8\}$ entregues pelo processo RANSAC(8).

4.2.1.2 Seleção de Modelo

Para determinar o posto $r_{\mathbf{A}}$, uma série de RANSACs é executada para estimar relações $T_{RANSAC,dim}$ com um menor número de restrições, onde $dim < 8$. O processo inicia-se com o RANSAC(7), diminuindo em 1 a quantidade de restrições empregadas inicialmente no processo RANSAC(8). RANSAC(7) determina a relação $T_{RANSAC,7}$ empregando uma aproximação de posto 8-1 da matriz $\mathbf{A}$ em cada amostra de pontos. Neste sentido, RANSAC(7) também usa uma quantidade menor de elementos/pontos $\tilde{m} = \lceil \frac{n=7}{r=1} \rceil$ em cada amostra. A entrada para o processo RANSAC(7) é o conjunto de inliers $\{in_8\}$ computado no processo RANSAC(8). RANSAC(7) testa para os inliers em $\{in_8\}$ se eles são também inliers para todas as relações do espaço-nulo da matriz de dados $\mathbf{A}$, que é composta agora por 7 equações. A avaliação para cada correspondência em $\{in_8\}$ emprega o mesmo critério do RANSAC(8) para verificar se ela é um inlier para todas as relações no espaço-nulo de $\mathbf{A}$. Na verdade, mostra-se em (FRAHM; POLLEFEYS, 2006) que inliers podem ser determinados observando-se somente os erros computados para as relações base do espaço-nulo da matriz de dados $\mathbf{A}$. Deste modo, se uma relação $T_{RANSAC,7}$, com menos parâmetros, recebe um suporte suficientemente grande dentro de $\{in_8\}$, conclui-se que os dados não proveram uma quantidade suficiente de restrições para determinar os 8 graus de liberdade da relação $T_{RANSAC,8}$. O suporte é computado como a razão $\frac{|\{in_7\}|}{|\{in_8\}|}$, ex-

pressando a proporção de inliers entregue pela relação $T_{RANSAC,7}$ e aqueles entregue pela relação original $T_{RANSAC,8}$. Por outro lado, se a relação $T_{RANSAC,7}$ não tem um suporte suficiente dentro do conjunto de inliers $\{in_8\}$, conclui-se que os dados $\{in_8\}$ proveram as 8 restrições para $T_{RANSAC,8}$.

O processo de reduzir a quantidade de restrições, que é explorado para computar a relação $T_{RANSAC,7}$ como explicado acima, é continuado até a relação $T_{RANSAC,8-i}$ não apresentar um suporte suficiente no conjunto de inliers originais $\{in_8\}$. Assim, o processo pára quando

$$\frac{|\{in_{8-i}\}|}{|\{in_8\}|} < \gamma, \tag{4.15}$$

onde $\{in_{8-i}\}$ representa o conjunto de inliers computados pelo processo RANSAC$(8-i)$. Deste modo, "suporte suficiente"é determinado por um limiar γ. Mostra-se experimentalmente em (FRAHM; POLLEFEYS, 2006) que γ não é um limiar crítico, e pode ser seguramente definido dentro de uma faixa de 50%-80% sem um impacto significante.

Se uma relação que não pode representar os inliers $\{in_8\}$ é encontrada, o modelo apropriado para os inliers $\{in_8\}$ foi computado na iteração anterior $i-1$ (veja esquema na Figura 4.5). Este modelo tem então $8-i+1$ graus de liberdade e a dimensão do espaço-nulo de $\mathbf{A}$ é $i-1$. A sucessiva redução da quantidade de restrições usada para computar a relação determina quantos restrições são entregues pelos inliers $\{in_8\}$. Neste ponto o processo de seleção do modelo é terminado e o modelo selecionado consiste em uma familia de relações que satisfazem as restrições providas pelos inliers $\{in_8\}$.

4.2.1.3 Complementação do Modelo

A fase de seleção do modelo pode entregar uma relação T_{Deg} que foi computada a partir de dados degenerados, o que significa que os dados não proveram os 8 graus de liberdade necessários para computar $\mathbf{E}$ como uma solução única. É demonstrado em (FRAHM; POLLEFEYS, 2006) que isso pode ocorrer mesmo para casos em que os dados são quase-degenerados, pois o algoritmo RANSAC tem uma baixa probabilidade de incluir a parte não-degenerada dos dados no processo de estimar uma relação. Neste caso, ainda é possível estimar a relação correta $\mathbf{E}$ com 8 parâmetros, sendo que para isso pode-se explorar os dados que não foram cobertos pelo processo de seleção de modelo que entregou T_{Deg} como solução.

Se o processo de seleção de modelo entrega uma relação degenerada T_{Deg}, os dados $\{p\}$ são classificados em inliers degenerados $\{in_{Deg}\}$ e outliers $\{out\}$. Há ainda a matriz de dados $\mathbf{A}_{Deg}$, a partir da qual a solução T_{Deg} foi computada (conforme a Equação 4.10). É importante observar que quando uma relação degenerada é entregue pelo processo de seleção de modelo, a matriz $\mathbf{A}_{Deg}$ é composta por uma quantidade $\tilde{m}$ de linhas/equações inferior a 8. Por exemplo, se o processo de redução da quantidade de restrições atingiu a dimensão 6, uma relação $T_{RANSAC,6}$ foi computada e a matriz $\mathbf{A}_{Deg}$ consta de 6 linhas.

O processo de complementação do modelo utiliza então uma aproximação de posto $8-i+1$ de $\mathbf{A}_{Deg}$, estendendo-a com matrizes provenientes de amostras de $\lceil \frac{i-1}{r} \rceil$ pontos tomados do conjunto $\{out\}$. Por exemplo, seguindo o exemplo acima onde uma relação $T_{RANSAC,6}$ foi computada, tem-se que a matriz $\mathbf{A}_{Deg}$ (com 6 linhas) é aproximada via SVD para posto 6. Então, o processo de complementação do modelo estende $\mathbf{A}_{Deg}$ incluindo duas novas linhas (2 + 6 = 8) provenientes do conjunto de outliers $\{out\}$. Isso ocorre no contexto de uma abordagem RANSAC, onde os outliers são testados para serem consistentes com $\mathbf{A}_{Deg}$ e, se de fato houverem outliers consistentes com $\mathbf{A}_{Deg}$, as restrições não empregadas no processo que gerou T_{Deg} são providas e a relação resultante T_C

constará de efetivos 8 parâmetros, como necessário para uma matriz essencial $\mathbf{E}$. Caso contrário, a hipótese de dados degenerados é aceita e, se $dim = 6$, a relação apropriada para representar os dados seria uma homografia $\mathbf{H}$.

4.2.2 Saída do Algoritmo QDEGSAC

No sentido de avaliar o desempenho do abordagem QDEGSAC, é importante observar a saída produzida por este método, uma vez que dados quase-degenerados não são raros em contextos práticos. Para uma cena, o QDEGSAC geralmente determina dois conjuntos disjuntos de inliers: *inliers degenerados* são aqueles inliers que estão em condições degeneradas na cena, por exemplo, pontos que são espacialmente coplanares dentro da cena. Estes inliers não determinam todos os graus de liberdade de uma relação estimada em contextos mais genéricos de cenas, como explicado acima em termos de uma matriz essencial $\mathbf{E}$. Por outro lado, inliers que não estão em condições degeneradas de cenas são denominados de *inliers não-degenerados*. Usualmente, refere-se a estes pontos como estando em posições genéricas na cena, isto é, não são pontos coplanares ou aproximadamente coplanares.

É importante notar que se os dados são quase-degenerados, o processo de *seleção de modelo* será responsável por entregar os *inliers degenerados*. Após, o processo de *complementação do modelo* investigará entre os outliers providos nos processos anteriores, *primeiro RANSAC* e *seleção de modelo*, a existência de *inliers adicionais* que sejam condizentes com o conjunto de *inliers degenerados* já computado pelo processo de *seleção de modelo*. Estes inliers adicionais conterão então a parte não-degenerada dos dados de entrada.

4.2.3 Custo Computacional do Algoritmo QDEGSAC

Com base no critério de parada do algoritmo RANSAC (Equação 4.13), observa-se que o processo RANSAC mais caro computacionalmente dentro da abordagem QDEGSAC é o RANSAC($8 - i$), o qual necessitará uma quantidade significativa de tentativas (*trials*) para provar que os dados não dão suporte para uma relação estimada a partir de somente $8 - i$ restrições. Deste modo, para reduzir um esforço computacional desnecessário em casos de cenas degeneradas, o processo de redução do número de restrições é interrompido quando a dimensão 5 é alcançada, uma vez que uma homografia $\mathbf{H}$ é uma relação de movimento apropriada para explicar dados de cenas degeneradas que aparecem na prática (NISTéR, 2000; TORR; FITZGIBBON; ZISSERMAN, 1999).

5 TRABALHOS RELACIONADOS E ABORDAGENS ALTERNATIVAS DESENVOLVIDAS NESTE TRABALHO

A maioria das técnicas de sumarização de vídeos digitais descritas na literatura propõe métodos para identificar quadros-chave que são representativos do conteúdo de segmentos distintos do vídeo. Usualmente este objetivo é alcançado pela redução de quadros redundantes na sequência do vídeo (GONG; LIU, 2000; HANJALIC; ZHANG, 1999; SAHOURIA; ZAKHOR, 1999; LI; ZHANG; TRETTER, 2001; NGO; PONG; ZHANG, 2001). Contudo, no caso de vídeos de histeroscopia diagnóstica a redundância visual entre quadros provê informações úteis. Conforme discutido na seção 2.2.2, ao fazer a histeroscopia, o especialista captura grande quantidade de imagens irrelevantes tentando encontrar as regiões de interesse (imagens úteis). Assim, quando estas regiões são encontradas, mais tempo é dispendido pelo especialista observando o conteúdo de tais imagens, produzindo um segmento de vídeo redundante em termos de informação visual. Deste modo, abordagens que quantificam diferenças visuais entre quadros de um vídeo, bem como métodos que quantificam/qualificam movimentos de câmera, são de especial interesse no contexto deste trabalho.

Neste capítulo discute-se a literatura de sumarização de vídeos com vistas a vídeos de histeroscopia diagnóstica. Na seção 5.1 discutem-se abordagens para a análise de movimento em vídeos como forma de quantificar/qualificar o conteúdo visual. Enquanto que a seção 5.2 apresenta técnicas desenvolvidas especificamente para a estruturação automática do conteúdo de vídeos histeroscópicos, as quais são fundamentadas na análise da redundância visual entre quadros vizinhos do vídeo. Estas técnicas foram desenvolvidas e publicadas como resultado deste trabalho.

5.1 Análise de Movimento na Sumarização de Vídeos

Análise de movimento é amplamente utilizada em tarefas que envolvem processamento de vídeo, contudo não é trivial representar o conteúdo visual convenientemente em termos de feições de baixo nível como cor e movimento (LEW, 2001; CHANG, 2002; NGO; PONG; ZHANG, 2001; DEL BIMBO, 1999). Por esta razão um grande número de abordagens estão sendo propostas dentro deste contexto de pesquisa (DUAN et al., 2006; NGO; PONG; ZHANG, 2003; ZHU et al., 2005; LIU; ZHANG; QI, 2003; VASCONCELOS; LIPPMAN, 2000; PIRIOU; BOUTHEMY; YAO, 2006; YOU et al., 2007; HO et al., 2006; MA et al., 2005). Como discutido na seção 2.2.1, o movimento de câmera é uma feição apropriada para ser explorada no contexto de sumarização de vídeos de histeroscopias. A abordagem apresentada neste trabalho visa quantificar movimentos de câmera, e por esta razão apresenta-se na seqüência uma revisão de técnicas orientadas

neste sentido.

Uma vez que o movimento de câmera é um indicador da intenção do autor do vídeo e freqüentemente expressa alterações em termos de conteúdo visual, métodos que caracterizam movimentos de câmera qualitativamente têm sido propostos na literatura não-paramétrica de indexação de vídeos digitais. Duan *et al.* (DUAN et al., 2006) propõe um espaço de feições onde o algoritmo *mean-shift* é utilizado para reconhecer padrões de movimento de câmera, como *panning*, *tilting* e *zooming*. Ngo *et al.* (NGO; PONG; ZHANG, 2003) defende classificar movimentos de câmera através da análise temporal de padrões extraídos de fatias espaciais de cada quadro do vídeo. Estes padrões delimitam sub-*shots*, os quais são posteriormente agrupados de acordo com similaridades cromáticas e proximidade temporal. Zhu *et al.* (ZHU et al., 2005) também propõe classificar movimentos de câmera qualitativamente. Neste trabalho histogramas são computados a partir de vetores de movimento e movimentos típicos de câmera, como *panning* e *zooming*, são associados com formatos distintos de histogramas. Contudo, a combinação de diferentes movimentos de câmera aparece freqüentemente em vídeos adquiridos com uma câmera manualmente controlada, tal como ocorre em vídeos de histeroscopias. Conseqüentemente, poderia ser bastante complicado caracterizar padrões de movimento de câmera em termos de *panning*, *tilting*, *zooming* e rotações, como proposto pelas abordagens mencionadas acima.

Muitos métodos também quantificam a intensidade de movimentos como forma de medir o conteúdo de vídeos digitais. Liu *et al.* (LIU; ZHANG; QI, 2003) propõe selecionar quadros-chave em um vídeo quantificando o movimento dominante, o qual é estimado a partir da magnitude e consistência angular de vetores de movimento. Wolf (WOLF, 1996) também utiliza uma métrica de movimento a partir da magnitude de vetores de movimento. Esta métrica é analisada em função do tempo e quadros-chave são selecionados de acordo com mínimos locais desta função. Ma *et al.* (MA et al., 2005) discute aspectos psicológicos da percepção humana e propõe uma metodologia de sumarização baseada em modelos de atração da atenção do usuário. Do ponto de vista de análise de movimento, os autores sugerem que a alta intensidade de movimentos normalmente atrai mais a atenção humana. Por esta razão a coerência espacial e a magnitude de vetores de movimento são utilizados para a construção de modelos de atenção baseados em movimentos. Uma desvantagem característica destas abordagens é a dependência de um crítico ajuste de limiares, os quais são geralmente determinados a partir de experimentos controlados (LIU; ZHANG; QI, 2003) ou a partir de uma suposta experiência por parte dos usuários (MA et al., 2005; WOLF, 1996).

No contexto de representação do conteúdo de vídeos, muitos métodos iniciam distinguindo o movimento de câmera (algumas vezes assumido como o movimento dominante da cena) do movimento independente de objetos (entendido como o resíduo do movimento dominante). Neste sentido, modelos afins 2-D de movimento são amplamente utilizados para explicar o movimento 2-D nas imagens que é induzido pelo movimento 3-D da câmera (BOUTHEMY; GELGON; GANANSIA, 1999; YOU et al., 2007; HO et al., 2006; MA et al., 2005; PIRIOU; BOUTHEMY; YAO, 2006; TAN et al., 2000; PEYRARD; BOUTHEMY, 2005). Em geral, os autores argumentam que um modelo afim pode lidar com grande parte das cenas em vídeos de propósito geral e, mesmo quando ele não pode, resultados satisfatórios são obtidos para propósitos de representação de movimento, mantendo um equilíbrio entre complexidade do modelo assumido e níveis aceitáveis de qualidade dos resultados. Contudo, uma vez que um modelo de movimento de câmera afim é adotado, algumas suposições são feitas a respeito da cena (LONGUET-

HIGGINS; PRAZDNY, 1980; MA et al., 2003), sendo que erros grosseiros devem ser tratados no ajuste do modelo aos dados quando estas suposições não forem satisfeitas. Por exemplo, se os movimentos da câmera são apenas translações, a variação de profundidade dentro da cena deve ser pequena quando comparada a distância entre a câmera e a cena em si. Por outro lado, se a variação de profundidade 3-D na cena é significante, a câmera deve executar apenas rotações em torno de seu eixo ótico. Deste modo, considerando a natureza de cenas histeroscópicas (seção 2.2.3), as restrições de cenas mencionadas acima não podem ser satisfeitas e por isso não considera-se apropriado o emprego de modelos afins de movimento de câmera unicamente para explicar cenas típicas de um vídeo de histeroscopia diagnóstica.

Para contornar limitações de modelos de movimento afim, uma solução imediata seria incorporar modelos 3-D de movimento de câmera, como restrições de movimentos rígidos entre múltiplos pontos de vista da cena (HARTLEY; ZISSERMAN, 2000), como forma de tratar configurações de cenas mais genéricas. Contudo, esta idéia tem atraído pouca atenção da comunidade científica que pesquisa sobre indexação de vídeos. Algumas técnicas existentes (IRANI; ANANDAN, 1998; SAWHNEY; AYER, 1996) estimam o movimento 2-D em imagens fundamentando-se na abordagem *plane-plus-parallax* (SAWHNEY, 1994), a qual assume a existência de uma superfície planar dentro da cena, ou que a variação de profundidade em uma região da cena é pequena. Abordagens que adotam restrições 3-D de movimento de câmera explicitamente para indexação de vídeos são recentes (ROTHGANGER et al., 2007; WAIZENEGGER; FELDMANN; SCHREER, 2008). Isso poderia ser justificado sob três aspectos básicos:

1. A análise qualitativa de movimento em vídeos, como a detecção de eventos de câmera, é em geral efetiva para propósitos de caracterização do conteúdo de vídeos e, conseqüentemente, estimativas precisas de parâmetros não se fazem necessárias (DUAN et al., 2006; LIU; ZHANG; QI, 2003; NGO; PONG; ZHANG, 2003), como discutido acima.

2. Usualmente, pontos correspondentes são necessários através da seqüência de quadros para estimar relações 3-D de movimento entre quadros (NISTéR, 2000; FITZGIBBON; ZISSERMAN, 1998; BEARDSLEY; ZISSERMAN; MURRAY, 1997). Contudo, estabelecer estas correspondências constitui-se ainda hoje em uma tarefa desafiadora, sendo que métodos usuais de rastreamento de pontos apresentam limitações sobretudo em casos de movimentos rápidos de câmera e freqüentemente entregam ambigüidades e falsas correspondências (MA et al., 2003).

3. Como discutido em áreas de pesquisa relatas (TRON; VIDAL, 2007; HARTLEY; ZISSERMAN, 2000; MA et al., 2003), modelos de câmera 3-D são mais complexos e também requerem algumas suposições sobre as configurações da cena no processo de estimar o movimento de câmera confiavelmente. Por exemplo, é bem estabelecido que dois pontos de vista de uma mesma cena rígida estão relacionados pela restrição epipolar (LONGUET-HIGGINS, 1981) e, uma das dificuldades para computar a geometria epipolar é que o movimento de câmera não pode ser estimado confiavelmente sob certas configurações de cena (TORR; FITZGIBBON; ZISSERMAN, 1999). Estes casos são conhecidos como configurações degeneradas de cena. Na prática, dois casos de degeneração surgem mais freqüentemente: *degeneração de movimento*, onde a câmera executa apenas rotações em torno de seu eixo ótico (isto é, não há translação de câmera entre os pontos de vista considerados da cena), e *degeneração estrutural*, onde a configuração 3-D da cena observada

é planar ou, de uma forma aproximada, a profundidade dentro da cena é pequena quando comparada a distância entre a câmera e a própria cena.

Apesar das dificuldades em estimar movimentos de câmera com base em modelos 3-D, pode-se concluir que ambos modelos 2-D e 3-D se complementam: modelos 2-D não são apropriados para estimar movimentos em configurações de cenas 3-D, as quais apresentam variações de profundidade juntamente com translações de câmera, o que pode ser tratado por modelos 3-D de movimento. Por outro lado, em configurações de cenas 2-D (*layout* da cena aproximadamente planar ou câmera executando apenas rotações), modelos de câmera 2-D mostram-se apropriados e modelos 3-D degeneram a não entregam soluções seguras. Como discutido na seção 2.2.3, ambas configurações de cena aparecem em vídeos de histeroscopias. Por esta razão, entende-se a tarefa de estimar movimentos de câmera em vídeos histeroscópicos como um problema de seleção de modelo, e por isso emprega-se técnicas recentemente propostas (FRAHM; POLLEFEYS, 2006) para selecionar o modelo de movimento apropriado através da seqüência de quadros para então estimar/quantificar o movimento de câmera em tais vídeos.

Seguindo a literatura de reconstrução 3-D de cenas a partir de múltiplos pontos de vista (NISTéR, 2000; FITZGIBBON; ZISSERMAN, 1998; BEARDSLEY; ZISSERMAN; MURRAY, 1997), o próximo passo seria computar a trajetória 3-D da câmera ao mesmo tempo em que quantifica-se movimentos de translação e rotação como uma função do tempo (velocidade de câmera), onde segmentos relevantes do vídeo de histeroscopia seriam selecionados de acordo com os mínimos locais desta função, uma vez que objetiva-se detectar movimentos lentos de câmera. Contudo, estimar pequenas translações e rotações de câmera a partir do movimento detectado nas imagens constitui-se em uma difícil tarefa (STEWéNIUS; ENGELS; NISTéR, 2007). Estimativas efetivas de movimento 3-D de câmera, evitando configurações degeneradas de cena, requerem uma certa disparidade entre os pontos de vista (quadros do vídeo), o que significa que em contextos de pequenas translações não é possível recuperar a trajetória 3-D da câmera confiavelmente a partir de correspondências pré-estabelecidas através dos quadros (HARTLEY; ZISSERMAN, 2000).

A fim de contornar problemas de configurações degeneradas de cenas em uma seqüência de imagens, alguns trabalhos propõe selecionar quadros-chave cuja disparidade é suficientemente grande para dar suporte a estimativas de movimento não-degeneradas (NISTéR, 2000; THORMAEHLEN; BROSZIO; WEISSENFELD, 2004; REPKO; POLLEFEYS, 2005). Contudo, a tarefa de rastrear pontos através de quadros largamente separados é difícil, uma vez que estes pontos são eventualmente perdidos ou movem-se para fora do campo de visão à medida que a seqüência de imagens decorre. Assim, a quantidade de pontos correspondentes entre quadros pode ser insuficiente para permitir uma estimativa confiável do movimento de câmera. Por esta razão, quadros-chave são selecionados de acordo com um equilíbrio entre a perda de correspondências ao longo dos quadros e a necessidade de um disparidade suficientemente grande entre estes quadros (NISTéR, 2000). Em vídeos de histeroscopias pode-se observar longas seqüências de quadros dentro de configurações degeneradas de cena, por esta razão retardar o processo de estimar a relação de movimento até configurar-se uma disparidade adequada entre os quadros pode ocasionar uma perda considerável de pontos correspondentes e uma conseqüente estimativa de relação de movimento não confiável. Além disso, considerar somente pares de quadros não-degenerados no processo de construção de sumários histeroscópicos poderia prejudicar a representatividade do conteúdo resultante no sumário de vídeo computado,

uma vez que quadros que constituem cenas degeneradas também são importantes clinicamente, como destacado na seção 2.2.3.

Desconsiderando as dificuldades com configurações degeneradas de cena, ainda é bastante difícil tratar o ruído presente nos pontos correspondentes previamente computados através dos quadros. Especialmente em casos de pequenos movimentos de câmera, o ruído pode ter maior influência que o próprio movimento 2-D detectado nas imagens no processo de estimar relações 3-D de movimento de câmera. Por esta razão, soluções provenientes do estado da arte são iterativas e dependem de otimizações não-lineares de restrições geométricas, o que é suscetível a problemas de mínimos locais e por isso não são confiáveis (SIM; HARTLEY, 2006).

5.2 Sumarização de Vídeos de Histeroscopias

Histeroscopias produzem uma quantidade grande de quadros, sendo que somente poucos destes quadros são verdadeiramente relevantes do ponto de vista de diagnósticos e prognósticos. Baseado nisso, a sumarização de vídeos de histeroscopias é usada para identificar segmentos de vídeo importantes e extrair quadros-chave destes. Nesta seção são apresentadas algumas abordagens propostas para a sumarização destes vídeos. Basicamente, o conteúdo apresentado a seguir foi proposto dentro do contexto desta tese e está publicado nas seguintes referências: (SCHARCANSKI; GAVIãO, 2006; GAVIãO et al., 2007; SCHARCANSKI; GAVIãO; CUNHA-FILHO, 2005; GAVIãO; SCHARCANSKI, 2005; CUNHA-FILHO et al., 2004). Na seção 5.2.1 apresenta-se uma metodologia que explora propriedades da decomposição de valor singular para sumarizar vídeos e, na seção 5.2.2, apresenta-se um método fundamentado na proposição de um modelo que caracteriza distâncias entre quadros como sendo distâncias estáticas ou dinâmicas.

5.2.1 Método Baseado na Decomposição de Valor Singular (SVD - *Singular Value Decomposition*)

A fim de caracterizar segmentos e quadros de vídeos de histeroscopias diagnósticas de acordo com uma medida de redundância/relevância (veja seção 2.2), apresenta-se nesta seção uma abordagem baseada na decomposição de valor singular (SVD) (GOLUB; LOAN, 1989). A SVD é uma técnica usada para a redução de dimensionalidade (JACKSON, 1991) de espaços de feições, enquanto preserva a estrutura essencial dos dados originais (GONG; LIU, 2000). Neste espaço reduzido, relações entre conjuntos de dados podem ser reveladas. A utilização da SVD para sumarizar e indexar vídeos é discutida na literatura (GONG; LIU, 2000; LEE; HAYES, 2004), porém, neste trabalho, um enfoque diferente é apresentado tendo em vista as características dos vídeos histeroscópicos. Na próxima seção são detalhadas as propriedades da SVD que são utilizadas na caracterização de segmentos e quadros de vídeos de acordo com sua dinamicidade. Na seção 5.2.1.2 apresenta-se um método de sumarização de vídeos de histeroscopia baseado nestas propriedades.

5.2.1.1 Propriedades da SVD

Considerando uma matriz $\mathbf{A}$ cuja i-ésima coluna representa um histograma de cor H_i, computado a partir de um quadro i. Esta matriz $\mathbf{A}$ é uma representação temporal da informação cromática do vídeo. Se $\mathbf{A}$ tem dimensões $m \times n$, onde m é o número de bins do histograma e n é a quantidade de quadros do vídeo, sendo $m \geq n$ a SVD de $\mathbf{A}$ é definida como:

$$\mathbf{A} = \mathbf{UDV}^T \tag{5.1}$$

Onde $\mathbf{U} = [u_{ij}]$ é uma matriz ortonormal $m \times n$ cujas colunas são chamadas de vetores singulares a esquerda; $\mathbf{D} = diag(\sigma_1, \sigma_2, \ldots, \sigma_n)$ é uma matriz diagonal $n \times n$ cujos elementos são chamados de valores singulares não negativos, os quais aparecem ordenados de forma decrescente; e $\mathbf{V} = [v_{ij}]$ é uma matriz ortonormal $n \times n$ cujas colunas são chamadas de vetores singulares a direita.

Se $posto(\mathbf{A}) = q$ a SVD pode ser interpretada como o mapeamento de um espaço m-dimensional, caracterizado pelos histogramas de cor dos quadros, em um espaço de feições refinado (reduzido) de q dimensões, as quais são linearmente independentes. Cada vetor coluna A_i de $\mathbf{A}$, representando os histogramas de cor do quadro i, é mapeado para um vetor coluna q-dimensional $\psi_i = [v_{i1}\ v_{i2}\ \cdots\ v_{iq}]^T$ da matriz $\mathbf{V}^T$.

A distância de ψ_i até a origem neste espaço reduzido (isto é, a norma euclidiana de ψ_i) é dada por (GONG; LIU, 2000):

$$\|\psi_i\| = \sqrt{\sum_{j=1}^{posto(\mathbf{A})} v_{ij}^2} \tag{5.2}$$

Considerando que $\mathbf{V}$ é ortonormal, se $posto(\mathbf{A}) = n$ então $\|\psi_i\| = 1$, onde $i = 1, 2, \ldots, n$ (GONG; LIU, 2000).

Neste trabalho explora-se basicamente duas propriedades da SVD (LEE; HAYES, 2004; GONG; LIU, 2000):

Propriedade 1: Sendo $\mathbf{B} = [B_1, B_2, ..., B_n]$ uma matriz $m \times n$ onde $m \geq n$ e $posto(\mathbf{B}) = n$, isto é, as colunas de $\mathbf{B}$ são linearmente independentes. Sendo B_0 um vetor coluna m-dimensional que é linearmente independente em relação aos vetores B_i em $\mathbf{B}$, e sendo $\mathbf{B}_D$ uma matriz $m \times d$ obtida através da replicação do vetor coluna B_0 d vezes. Assim,

$$\mathbf{B}_D = [B_0, B_0, ..., B_0] \quad \text{e} \quad posto(\mathbf{B}_D) = 1. \tag{5.3}$$

Sendo $\mathbf{A}'$ uma matriz $m \times (n + d)$ que é constituida pelos vetores coluna de $\mathbf{B}$ e $\mathbf{B}_D$ arranjados em qualquer ordem, como por exemplo:

$$\mathbf{A}' = [B_1 ... \overbrace{B_0^1 ... B_0^d} ... B_n]. \tag{5.4}$$

Calculando-se a SVD de $\mathbf{A}'$ (Eq. 5.1) e considerando que $\mathbf{V}'^T = [\psi_1 ... \overbrace{\phi_1 ... \phi_d} ... \psi_n]$ é a matriz de vetores singulares a direita obtida. Pode ser mostrado que a norma euclidiana $\|\phi_j\|^2 = 1/d$, onde $j = 1, 2, \ldots, d$ e d é o número de vezes que B_0 é replicado em $\mathbf{A}'$ (veja prova em (LEE; HAYES, 2004)).

A conclusão acima implica que o vetor coluna A_i, que é linearmente independente em $\mathbf{A}$, é projetado pela SVD em um vetor ψ_i, cuja distância em relação a origem do espaço de feições reduzido é 1, de acordo com a Equação 5.2. Quando A_i tem cópias $A_i^{(j)}$ (i.e. A_i é redundante, ou linearmente dependente), a distância de seu vetor projetado ϕ_j em relação a origem do novo espaço diminui. Quanto mais cópias de A_i há (isto é, quanto mais redundante ele é), menor será sua distância da origem (GONG; LIU, 2000).

Propriedade 2: Se $\mathbf{A}$ é uma matriz $m \times n$ de posto k com uma SVD definida pela Equação 5.1, então a distância euclidiana entre quaisquer dois vetores coluna de $\mathbf{A}$ é igual a distância euclidiana entre os vetores coluna correspondentes de $\mathbf{V}^T$ ponderados

pelos correspondentes valores singulares (LEE; HAYES, 2004). Conseqüentemente, a distância euclidiana entre pares de histogramas H_i e H_j (representando os quadros i e j em **A**) (GONG; LIU, 2000) é:

$$D(\psi_i, \psi_j) = \sqrt{\sum_{l=1}^{k} \sigma_l (v_{il} - v_{jl})^2}, \tag{5.5}$$

onde ψ_i e ψ_j representam H_i e H_j no espaço de feições obtido pela SVD, e σ_l são os valores singulares.

5.2.1.2 Explorando a SVD para Sumarizar Vídeos de Histeroscopias Diagnósticas

Objetiva-se estimar se um segmento de vídeo é estático ou dinâmico (redundante ou não) com base nas propriedades deste segmento no espaço da SVD. Especificamente, utiliza-se a *propriedade 1* mencionada na seção 5.2.1.1. Com base na equação 5.2, computa-se a norma Euclidiana de cada vetor de feição (representando um quadro do vídeo) ψ_i no espaço $\mathbf{V}^T$ definido pela SVD. Deste modo, vetores representando quadros provenientes de um segmento de video redundante são projetados próximos da origem, enquanto vetores de quadros oriundos de segmentos menos redundantes são projetados em posições mais afastadas da origem no espaço da SVD. Assim, valores $\|\psi_i\|$ menores sugerem segmentos de vídeo redundantes (isto é, estáticos), e valores maiores indicam segmentos menos estáticos. Conseqüentemente, $\|\psi_i\|$ é usado como uma medida de redundância, ou estaticidade, para cada quadro i de um vídeo histeroscópico.

Neste ponto é necessário definir um limiar τ para discriminar segmentos estáticos, caracterizados por valores menores de $\|\psi_i\|$, e segmentos não estáticos. É importante notar que determinar um valor para τ não é uma tarefa simples, desde que a decisão entre o "estático" e o "dinâmico" tende a ser subjetiva. Com base nisso, propõe-se um limiar τ adaptativo.

Considerando $P(\|\psi\|)$ como uma estimativa da função densidade de probabilidade de $\|\psi\|$, dados os valores de $\|\psi_i\|$, $i = 1, ..., n$. O limiar τ é escolhido como sendo a moda de $P(\|\psi\|)$ para um dado vídeo de histeroscopia diagnóstica. A fim de obter uma estimativa de moda mais precisa modela-se $P(\|\psi\|)$ através de uma mistura de kernels gaussianos (BOWMAN; AZZALINI, 1997).

Assim, um quadro i é considerado redundante, isto é, um quadro oriundo de um segmento de vídeo estático se:

$$\|\psi_i\| \leq \tau \tag{5.6}$$

Caso contrário, este quadro é considerado não redundante e, consequentemente, não relevante. Todos os quadros temporalmente adjacentes que satisfazem a equação 5.6 formam os segmentos de vídeo classificados como relevantes para a construção do sumário do vídeo.

Com base na equação 5.6, diferentes grupos de quadros adjacentes são tomados como segmentos relevantes S_k, $k = 1, ..., M$, onde M é a quantidade de segmentos relevantes para um vídeo. Estes segmentos constituem o sumário inicial do vídeo. Cada segmento tem um quadro-chave KS_k associado. Seguindo a hipótese inicial de associar redundância com relevância, um quadro-chave KS_k é o quadro $i \in S_k$ com o menor valor $\|\psi_i\|$, isto é, adota-se o quadro mais redundante dentro do segmento relevante S_k como quadro-chave. Neste ponto salienta-se que o nível mais alto no processo de sumarização é definido

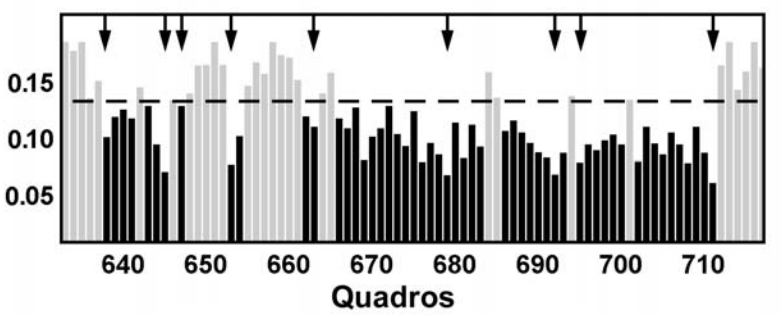


Figura 5.1: Gráfico mostrando a norma $\|\psi\|$ dos quadros como barras verticais. O eixo horizontal representa cada quadro i na ordem temporal do vídeo, e o eixo vertical representa $\|\psi_i\|$. A linha tracejada representa τ. Barras em cinza representam os segmentos de vídeo discartados, enquanto que as barras pretas representam os segmentos selecionados. A menor barra preta dentro de cada segmento selecionado mostra o quadro-chave correspondente. As setas indicam estes quadros- chave.

pela escolha manual dos quadros-chave feita pelos especialistas, tomando como base os quadros-chave providos por esta abordagem.

A Figura 5.1 mostra uma série de valores $\|\psi_i\|$, representando cada quadro na seqüência do vídeo, para um vídeo de histeroscopia em particular. No eixo horizontal estão representados os quadros i na seqüência temporal do vídeo e a linha tracejada representa o limiar τ. As barras consecutivas em cinza representam os segmentos de vídeo discartados de acordo com τ, sendo que as barras consecutivas em preto representam os segmentos relevantes. Nestes últimos, a menor barra (isto é, o menor valor de $\|\psi_i\|$ dentro do segmento) indica o quadro-chave do respectivo segmento. As setas acima indicam os quadros-chave em cada segmento.

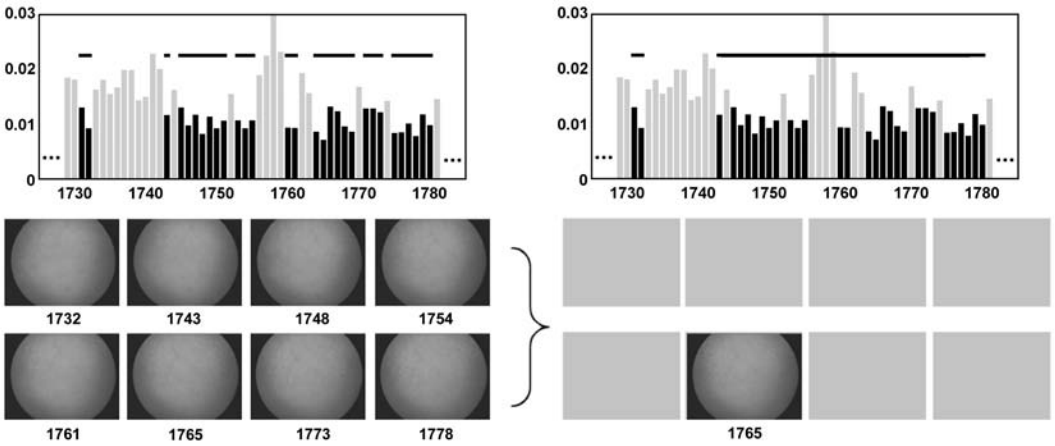
Os quadros-chave extraídos de acordo com o limiar τ podem ser muito similares (isto é, redundantes). Isto pode ser tratado como uma desvantagem deste método. Basicamente, alguns segmentos de vídeo selecionados como relevantes, S_k e S_{k+1}, estão localizados temporalmente próximos na seqüência do video, gerando uma super segmentação do vídeo. Lembrando que quadros temporalmente próximos tendem a ser similares, sobretudo quando são extraídos de segmentos estáticos. Os três primeiros segmentos relevantes à esquerda na Figura 5.1 ilustra este problema. Para contornar este problema apresenta-se uma etapa de pós-processamento na próxima seção.

5.2.1.3 Pós-processamento

Para reduzir a super segmentação do vídeo e a redundância de quadros-chave propõe-se uma estratégia de fusão de segmentos. Dois segmentos relevantes e adjacentes S_k e S_{k+1} são fundidos se as seguintes condições são satisfeitas:

- Considerando $\overline{\|\psi_{S_k}\|}$ e $\overline{\|\psi_{S_k+1}\|}$ como o valor médio de $\|\psi\|$ para os segmentos relevantes S_k e S_{k+1}, respectivamente. Sendo $\overline{\|\psi_{S_\times}\|}$ o valor médio de $\|\psi\|$ para um segmento de vídeo não relevante que está localizado temporalmente entre S_k e S_{k+1}. Deste modo, a primeria condição para fundir S_k e S_{k+1} é:

$$\frac{\overline{\|\psi_{S_k}\|} + \overline{\|\psi_{S_k+1}\|} + \overline{\|\psi_{S_\times}\|}}{3} \leq \tau, \tag{5.7}$$



(a) (b)

Figura 5.2: Ilustração do processo de fusão de segmentos de vídeo. Os diagramas acima mostram parte de uma seqüência temporal de valores da norma $\|\psi\|$ (barras verticais) para os quadros de um vídeo histeroscópico particular. Os segmentos horizontais de reta indicam os segmentos relevantes do vídeo: (a) abaixo, quadros chave representando os segmentos antes do processo de fusão; (b) quadros chave obtidos de segmentos que sofreram o processo de fusão (cada segmento é representado por um quadro chave).

onde τ é a moda de $P(\|\psi\|)$, como mencionado anteriormente. De acordo com esta condição, o segmento resultante da fusão de S_k,$S_\times$ e S_{k+1} ainda pode ser consiredado estático;

- A fim de minimizar a chance de fusão de segmentos de vídeo que são visualmente distintos, propõe-se uma condição baseada nos vetores de feições dos quadros que constituem o início e o fim de um segmento relevante. Computa-se a distância $D(\psi_i, \psi_{i+1})$ para todos quadros adjacentes dentro dos segmentos relevantes S_k (Eq. 5.5). Denota-se $P_C(D(\psi_i, \psi_{i+1}))$ como a probabilidade acumulada dos valores de $D(\psi_i, \psi_{i+1})$. Sendo ψ^k_{Last} e ψ^{k+1}_{First} o último e o primeiro vetor de feições dos quadros dos segmentos S_k e S_{k+1}, respectivamente, no espaço da SVD. Assim, para um dado video de histeroscopia diagnóstica, os segmentos S_k e S_{k+1} são fundidos se:

$$P_C(D(\psi^k_{Last}, \psi^{k+1}_{First})) \leq \nu \tag{5.8}$$

Em todos os experimentos realizados utiliza-se $\nu = 0.99$. Deste modo, os segmentos S_k e S_{k+1} são fundidos se a distância entre o último quadro ψ^k_{Last} de S_k e o primeiro quadro ψ^{k+1}_{First} de S_{k+1} resultar em um valor que caracterize uma similaridade nos mesmos padrões de similaridade verificados dentro de segmentos relevantes.

Com base na discussão acima, dois segmentos relevantes de vídeo S_k e S_{k+1} são fundidos se as equações 5.7 e 5.8 são satisfeitas. A Figura 5.2 ilustra este processo de fusão para um segmento de vídeo em particular.

5.2.2 Método Baseado em um Modelo para Valores de Distâncias entre Quadros

Seguindo a mesma proposição fundamental, isto é, associando redundância com relevância para identificar segmentos de quadros clinicamente importantes em vídeos de histeroscopia diagnóstica, apresenta-se nesta seção uma abordagem baseada na modelagem das distância entre quadros. O objetivo é classificar valores menores de distâncias entre quadros como sendo distâncias estáticas, e valores maiores como sendo distâncias dinâmicas. Neste contexto, distâncias estáticas caracterizam os segmentos de vídeo redundantes (relevantes) e distâncias dinâmicas caracterizam os segmentos não redundantes (não relevantes).

A distância Jeffrey (RUBNER et al., 2001) entre histogramas de cor H_i, provenientes de quadros temporalmente adjacentes X_i e X_{i+1}, é utilizada como métrica $D(H_i, H_{i+1})$:[1]

$$D(H(X_i), H(X_{i+1})) = \sum_j H_j(X_i) \log \frac{H_j(X_i)}{\overline{H}(j)} \quad + \\ + \quad H_j(X_{i+1}) \log \frac{H_j(X_{i+1})}{\overline{H}(j)} \quad , \quad i \in [1, N-1] \tag{5.9}$$

onde $H_j(X_i)$ and $H_j(X_{i+1})$ são os valores do bin j para os quadros sucessivos X_i e X_{i+1}; $\overline{H}(j) = [H_j(X_i) + H_j(X_{i+1})]/2$ é a histograma médio; e N é a quantidade de quadros do vídeo.

Pequenos valores de $D(H(X_i), H(X_{i+1}))$ correspondem a pequenas diferenças entre histogramas de quadros adjacentes (isto é, quadros similares). Consequentemente, $D(H(X_i), H(X_{i+1}))$ provê uma medida de redundância para o quadro X_i. Segmentos de vídeo estáticos são caracterizados por pequenos valores de diferenças inter-quadros $D(H(X_i), H(X_{i+1}))$, e são discriminados de quadros não estáticos (segmentos dinâmicos) por uma limiar adaptativo δ.

Seja $P(d)$ a probabilidade de ocorrência dos valores de distância d, dados todos os valores $D(H(X_i), H(X_{i+1}))$, $i = 1, ..., N$. Seja h_0 a hipótese que um dado valor d caracterize um quadro redundante, e h_1 seja a hipótese que d caracterize um quadro não redundante. A probabilidade de d dado que h_1 ocorre, denotada por $P(d|h_1)$, torna-se maior na medida que os valores de d também são maiores. Consequentemente, a probabilidade de d dado que h_0 ocorre, isto é, $P(d|h_0)$, torna-se menor conforme os valores de d aumentam. Logo, a probabilidade acumulada $P_C(d)$ é adotada como um modelo de probabilidade a priori para $P(d|h_1)$:

$$P(d|h_1) \equiv P_C(d) = \sum_{\gamma=0}^{d} P(\gamma) \tag{5.10}$$

onde $d \in \{D(H(X_i), H(X_{i+1}))\}$. Consequentemente tem-se:

$$P(d|h_0) = 1 - P(d|h_1) = 1 - \sum_{\gamma=0}^{d} P(\gamma) \tag{5.11}$$

O limiar δ é o valor d que torna $P(d|h_0) = P(d|h_1)$, minimizando o erro de confirmar h_0 quando h_1 é verdadeira, e vice-versa. Assim,

[1] Conceitualmente, histogramas são distribuições de probabilidades empíricas que deveriam ser comparadas por uma métrica adequada para distribuições de probabilidade (por exemplo, a distância Jeffrey).

$$P(\delta|h_0) = P(\delta|h_1) \tag{5.12}$$

$$\sum_{\gamma=0}^{\delta} P(\gamma) = 1 - \sum_{\gamma=0}^{\delta} P(\gamma) \tag{5.13}$$

e,

$$\sum_{\gamma=0}^{\delta} P(\gamma) = \frac{1}{2} \tag{5.14}$$

Com base na discussão acima, conclui-se que uma estimativa razoável é $\delta = \textit{median}\{d\}$. O limiar δ é escolhido como a mediana da distância dos histogramas para um dado vídeo de histeroscopia diagnóstica. Assim, um quadro X_i é redundante, isto é, provém de um segmento de vídeo estático (relevante), se:

$$D(H(X_i), H(X_{i+1})) \leq \delta \tag{5.15}$$

Quadros adjacentes que satisfazem a equação 5.15 constituem os segmentos relevantes do vídeo.

5.2.2.1 Sumarização Hierárquica

Com base na metadologia apresentada na seção anterior desenvolveu-se uma técnica hierarquica para selecionar segmentos de vídeo relevantes. Níveis de sumarização mais baixos contém um sumário mais detalhado (isto é, o sumário é menos compacto que em níveis mais altos de sumarizaçao). Em cada nível de sumarização l, um sumário inicial é constituido pelos segmentos relevantes do vídeo S_j, os quais podem ser navegados e acessados com base em seus quadros-chave $\overline{X}_j^k$. O conjunto final de quadros-chave é resultado da escolha manual das imagens (entre os quadros-chave do sumário) por parte dos especialistas. Estes quadros são, então, utilizados para descrever o vídeo junto ao prontuário da paciente.

No nível de sumarização mais baixo $l = 1$, os segmentos relevantes S_j^1 contém todos os quadros $X_i \in S_j^1$ cuja distância entre quadros adjacentes satisfaz a Equação 5.15. Níveis de sumarização mais altos $l > 1$ são obtidos selecionando-se diferentes valores δ^l recursivamente. Um dado quadro X_i é considerado redundante em um nível de sumarização l se:

$$D(H(X_i), H(X_{i+1})) \leq \delta^l. \tag{5.16}$$

Dado que $X_i \in S_j^l$, define-se os quadros provenientes de segmentos relevantes de vídeo S_j^l no nível l como sendo $URS^l = \bigcup_j S_j^l$ (isto é, $X_i^l = \{X_i \in URS^l\}$, onde $i = 1, ..., N^l$, e N^l é a quantidade de quadros pertencentes a segmentos relevantes de vídeo no nível l). Considerando que d^l denota os valores de distâncias $D(H(X_i^l), H(X_{i+1}^l))$, define-se o limiar adaptativo no nível $l+1$ como $\delta^{l+1} = \textit{median}\{d^l\}$. Este limiar é utilizado para definir os segmentos relevantes do vídeo neste nível S_j^{l+1}.

O procedimento hierárquico de sumarização descrito acima é aplicado recursivamente, produzindo diferentes níveis de sumarização l. O processo de recursão continua enquanto a seguinte razão probabilística é satisfeita:

$$\frac{\overline{P}(d^l|h_0)}{1 - \overline{P}(d^l|h_0)} > 1, \tag{5.17}$$

onde,

$$\overline{P}(d^l|h_0) \equiv E[P(d^l|h_0)] = \sum_{\gamma=\min\{d^l\}}^{\max\{d^l\}} P(\gamma)P(\gamma|h_0), \tag{5.18}$$

e $P(\gamma|h_0)$ é o modelo de probabilidade a priori para $P(d|h_0)$ definido de acordo com a Equação 5.11 ($d^l = \{d^l_{i=1,\ldots,N^l}\}$). Deste modo, $\overline{P}(d^l|h_0)$ pode ser interpretado como a probabilidade média de que quadros de segmentos relevantes no nível l de sumarização são redundantes (isto é, quadros oriundos de segmentos estáticos); e $1 - \overline{P}(d^l|h_0)$ pode ser interpretado como a probabilidade média destes quadros serem não redundantes (isto é, quadros oriundos de segmentos dinâmicos).

Como o processo de recursão direciona-se para níveis mais altos de sumarização, o valor de δ^l aumenta e os segmentos tendem a conter mais quadros não redundantes. Na próxima seção apresenta-se uma etapa de seleção dos segmentos mais representativos entre os segmentos S^l_j já definidos.

5.2.2.2 Seleção de Segmentos de Vídeo Relevantes em cada Nível de Sumarização

Na prática, segmentos S^l_j contém quadros redundantes e não redundantes. Por esta razão, segmentos relevantes de vídeo S^l_j, em níveis mais altos de sumarização, tendem a apresentar maiores proporções de quadros não redundantes, quando comparados com segmentos em níveis mais baixos de sumarização. Deste modo, a probabilidade média de que quadros pertencentes a S^l_j sejam redundantes *a priori* provê uma medida de redundância para um segmento de vídeo:

$$\overline{P}(d^l_j|h_0) \equiv E[P(d^l_j|h_0)] = \sum_{\gamma=\min\{d^l_j\}}^{\max\{d^l_j\}} P(\gamma)P^{'}(\gamma|h_0), \tag{5.19}$$

onde $d^l_j = \{d \in S^l_j\}$, $P^{'}(\gamma = d|h_0) \equiv 1 - \sum_{\gamma=\min\{d^l\}}^{d} P(\gamma)$ e $d^l = \{d^l_{i=1,\ldots,N^l}\}$.

Neste ponto é importante notar que $P^{'}(\gamma|h_0)$ é calculado com base nos segmentos selecionados S^l_j, e esta discussão somente aplica-se a níveis de sumarização maiores que 1, isto é, $l > 1$.

No nível de sumarização l, o comprimento dos segmentos variam, e estão na faixa $[\min\{L^l\}, \max\{L^l\}]$, onde $\{L^l\}$ denota o conjunto dos valores de comprimentos de segmentos. Geralmente, segmentos curtos são irrelevantes do ponto de vista prático, pois são associados a um tempo de observação curto de uma região do útero gerado não intencionalmente. Deste modo, os segmentos de interesse são primariamente redundantes e não tão curtos. Nesta abordagem, atribui-se relevância somente a segmentos de vídeo S^l_j que tem comprimento $L^l_j \geq \overline{L}^l$, onde $\overline{L}^l \equiv E\{L^l\}$ (para $l > 1$, como mencionado).

Finalmente, os segmentos relevantes de vídeo SR^l_j são selecionados como os segmentos S^l_j que satisfazem a seguinte razão de probabilidades:

$$\frac{\overline{P}(d^l_j|h_0)}{1 - \overline{P}(d^l_j|h_0)} > 1, \tag{5.20}$$

e apresenta comprimento $L^l_j \geq \overline{L}^l$.

A Figura 5.3 ilustra distâncias entre quadros adjacentes (Equação 5.9) para um vídeo em particular. O eixo horizontal representa cada quadro X_i na seqüência temporal do vídeo e o eixo vertical representa as distâncias entre os quadros adjacentes $D(H(X_i), H(X_{i+1}))$.

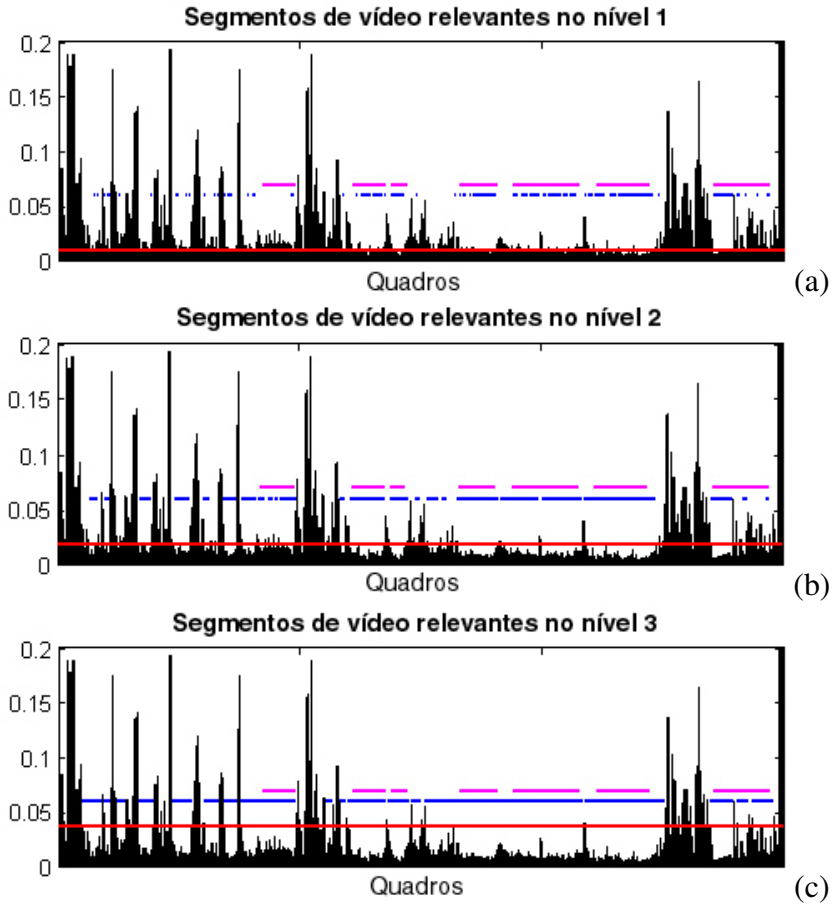


Figura 5.3: Gráficos mostrando os segmentos selecionados e o limiar δ^l em cada nível de sumarização. O eixo horizontal representa cada quadro X_i na seqüência temporal do vídeo e o eixo vertical representa as distâncias entre os quadros adjacentes $D(H(X_i), H(X_{i+1}))$. A linha contínua representa o limiar δ^l em cada nível de sumarização l: (a) nível 1; (b) nível 2 e (c) nível 3. A linha irregularmente tracejada, acima do limiar, indica a localização temporal dos segmentos relevantes selecionados pelo método e, a linha acima, representa a localização temporal dos segmentos selecionados pelos especialistas.

As Figuras 5.3(a-c) mostram δ^l e a localização dos segmentos relevantes em cada nível de sumarização l. A linha contínua (na parte de baixo de cada gráfico) representa o limiar δ^l e a linha irregularmente tracejada, acima do limiar, indica a localização temporal dos segmentos relevantes selecionados pelo método. A localização temporal dos segmentos selecionados pelos especialistas também é ilustrada acima dos segmentos selecionados pelo método.

5.2.2.3 Seleção de Quadros-chave

Define-se um quadro-chave $\overline{X}_j^l$ como o quadro X_i mais redundante dentro de uma vizinhança temporal W^l de um segmento relevante SR_j^l (isto é, o quadro com a menor distância $D(H(X_i), H(X_{i+1}))$ dentro de W^l). A dimensão da vizinhança temporal W^l assume valores no intervalo $[\min\{L^l\}, \lambda^l]$, sendo que o comprimento λ^l é definido adaptativamente para cada nível de sumarização l:

$$\lambda^l = \textit{median}\{L^l\}. \tag{5.21}$$

Pelo menos um quadro-chave $\overline{X}_j^l$ é extraído de cada segmento de vídeo relevante SR_j^l. Se um segmento relevante SR_j^l tem comprimento $L_j^l \leq \lambda^l$, um quadro-chave $\overline{X}_j^l$ é selecionado dentro da vizinhança temporal $W^l = L_j^l$. Por outro lado, o quadro-chave é escolhido como o quadro mais redundante em cada posição não sobreposta da vizinhança $W^l = \lambda^l$ ao longo de SR_j^l. Deste modo, segmentos SR_j^l longos têm mais quadros-chave para melhor representar seu conteúdo visual, objetivando maior praticidade na atividade de *browsing* do vídeo.

5.2.3 Discussão

Os métodos apresentados nas seções 5.2.1 e 5.2.2 apresentam uma limitação comumente encontrada em abordagens destinadas à sumarização de vídeos. A similaridade entre quadros é um aspecto essencial no processo de sumarização, uma vez que a redundância de informações deve ser detectada e eliminada tanto quanto possível. Contudo, ao utilizar-se uma métrica de similaridade entre quadros, faz-se necessário uma estratégia para definir um conjunto de quadros com os quais um dado quadro i deve ser comparado em termos de similaridade. Em muitos casos, a comparação de um quadro i com todos os demais quadros do vídeo é desnecessária e impraticável por questões computacionais. Por exemplo, o método apresentado na seção 5.2.1 estabelece uma métrica de comparação que envolve um espaço de feições derivado de todos os quadros do vídeo, sendo que no contexto de vídeos histeroscopias há pouco sentido em comparações por similaridade entre quadros oriundos de fases distintas do exame. Por outro lado, o método apresentado na seção 5.2.2 fundamenta-se em distâncias computadas apenas entre pares de quadros temporalmente adjacentes (quadros i e $i+1$). Assim, uma técnica capaz de estabelecer uma vizinhança temporal para um dado quadro i já seria de grande valor. Idealmente, os limites da vizinhança de um quadro i deveria ser definido de maneira a desconsiderar quadros cujo conteúdo pouco tem em comum com i. Na sequência desta seção discutem-se alguns pontos que podem ser explorados na construção de uma abordagem capaz de associar um conjunto de quadros vizinhos a cada quadro i do vídeo para propósitos de comparação por similaridade. Com isso, o problema é melhor compreendido e a solução apresentada no próximo capítulo é justificada.

Conforme discutido no início deste capítulo, o movimento de câmera realizado pelo especialista ao capturar imagens importantes em uma histeroscopia diagnóstica tende a ser mais lento quando comparado a outros momentos do exame, produzindo segmentos de vídeo com quadros redundantes em termos de informação visual. Deste modo, ao computar-se distâncias entre quadros vizinhos no vídeo produzido, é possível verificar-se uma caracterização da atividade da câmera na seqüência temporal dos valores de distância. Por exemplo, a Figura 5.4 mostra uma seqüência temporal de distâncias (barras verticais em cinza) entre quadros adjacentes de um vídeo de histeroscopia em particular. A curva sobreposta ao gráfico representa uma versão suavizada dos valores de distâncias. Observa-se nesta curva que vales podem caracterizar uma região de interesse, pois refletem uma redução na atividade da câmera e uma consequente similaridade entre quadros temporalmente próximos, subentendendo-se que o especialista está focalizando detalhes que chamaram sua atenção. Isto é visualmente verificado através da escala mostrada abaixo do gráfico de barras, onde cada quadro está representado espacialmente por uma fatia vertical de 5 *pixels* extraída do centro de cada quadro. Verifica-se claramente nesta

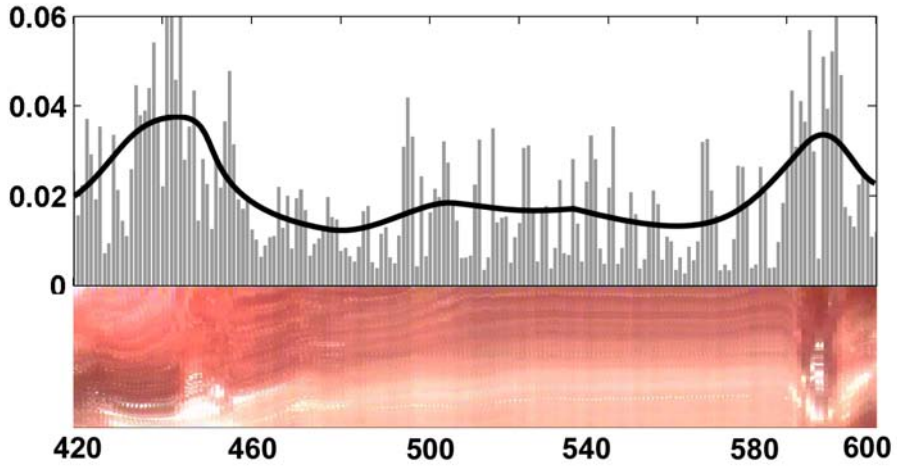


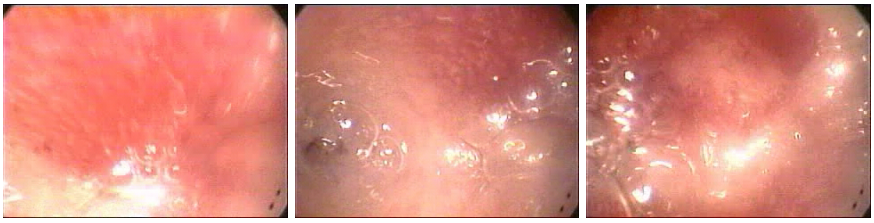
Figura 5.4: Diagrama mostrando características visuais e distâncias de quadros para um segmento de vídeo de histeroscopia diagnóstica. O eixo horizontal representa cada quadro i na ordem temporal do vídeo, e o eixo vertical representa distâncias entre quadros adjacentes (barras em cinza). A curva sobreposta ao gráfico representa uma versão suavizada dos valores de distâncias. Abaixo, mostra-se uma escala visual onde cada quadro é representado por uma fatia vertical de 5 *pixels* extraída do seu centro. Padrões visuais mais estáveis (isto é, segmentos relevantes) são verificados para as regiões de vale da curva.

escala que a variação do padrão visual está correlacionado, de alguma forma, com o padrão da curva. Por exemplo, a região central do gráfico está caracterizada por um vale. A Figura 5.5 mostra os quadros 445, 550 e 595 da seqüência temporal mostrada na Figura 5.4. Observa-se que a região de vale contém exatamente imagens que correspondem a fase de inspeção do orifício da trompa direita, mencionada como uma das fases importantes do exame de histeroscopia diagnóstica na seção 2.1. Os demais quadros caracterizam imagens menos importantes segundo sua informação visual, sobretudo o quadro 445, onde verifica-se um borramento de detalhes decorrente de movimentos mais rápidos da câmera, o que foi determinado pelo especialista no momento do exame.

De acordo com as evidências apresentadas acima, verifica-se a possibilidade de construção de um processo adaptativo de seleção de quadros-chave, uma vez que os vales podem determinar regiões importantes do vídeo e consequentemente responder a desafiadora pergunta: *Onde estão e quantos são os quadros-chave que representam adequadamente o conteúdo de um vídeo de histeroscopia diagnóstica?* A resposta poderia ser encontrada nos vales da curva computada sobre as distâncias entre quadros, onde a parte mais profunda (região menos dinâmica do segmento de vídeo) de cada vale definiria o quadro-chave. Além disso, seria razoável associar a relevância de um quadro-chave de acordo com a forma e/ou comprimento temporal de seu respectivo vale.

Com base no exposto acima, aponta-se alguns pontos fundamentais em uma metodologia destinada a caracterizar vales e picos em uma seqüência de valores de distâncias entre quadros de um vídeo:

- Tomando como base apenas distâncias entre quadros adjacentes, torna-se complicado determinar a relevância de cada vale. Isto verifica-se na Figura 5.6(a), onde as barras verticais em preto denotam distâncias entre quadros adjacentes em um segmento relevante, para um vídeo em particular. Observa-se a existência de inúmeros



445 550 595

Figura 5.5: Quadros de número 445, 550 e 595 extraídos da seqüência temporal mostrada na figura 5.4. O quadro 550, proveniente de uma região de vale, corresponde a fase de inspeção da abertura da trompa direita (seção 2.1).

pequenos vales, o que resultaria em uma grande quantidade de quadros chaves de acordo com a estratégia mencionada acima;

- Uma estratégia interessante seria analisar a distância de um quadro X_i em relação aos seus n quadros temporalmente mais próximos. Sendo $D(X_i, X_j)$ a distância entre os quadros X_i e X_j, onde $|i - j| \leq n$. Sendo $\widehat{D_i}$ a distância média de X_i em relação a todos os quadros X_j definida como:

$$\widehat{D_i} = \frac{\sum_{j=1}^{n} D(X_i, X_j)}{n}, \qquad i \neq j. \tag{5.22}$$

  A Figura 5.6(b) mostra valores $\widehat{D_i}$, como barras verticais em preto, para um segmento de vídeo histeroscópico particular. Neste caso $n > 1$, sendo que para o exemplo da Figura 5.6(a) $n = 1$. As Figuras 5.6(a) e (b) representam o mesmo segmento de vídeo. Experimentalmente, verifica-se que conforme o valor de n aumenta, alguns vales e picos tornam-se mais evidentes e persistem quando emprega-se diferentes valores de n.

Basicamente, estruturar um vídeo de histeroscopia consiste em definir os limites de segmentos de vídeo cujos quadros apresentam um conteúdo visual correlacionado e clinicamente relevante. Estimar um valor adequado para n ao longo do vídeo traduz a essência deste problema, pois caracteriza-se com isso a informação contida em uma região do vídeo, possibilitando que quadros-chave sejam selecionados com base na redundância/relevância da informação contida em cada quadro do vídeo.

5.3 Conclusões

Neste capítulo apresentou-se uma análise da literatura de sumarização de vídeos para propósitos de sumarização de vídeos de histeroscopias diagnósticas. Exceto por resultados derivados deste trabalho, conclui-se que não há trabalhos apropriadamente projetados para gerar sumários de vídeos de histeroscopias diagnósticas. Há, contudo, inúmeros trabalhos que provêem direções a serem seguidas na construção de uma abordagem para tais vídeos, conforme discutido na seção 5.1.

Além disso, apresentou-se duas abordagens que exploram a redundância de feições dos quadros para extrair segmentos e quadros relevantes em vídeos histeroscópicos. Na seção 5.2.3 discutiu-se um importante aspecto/limitação destes métodos: a definição de

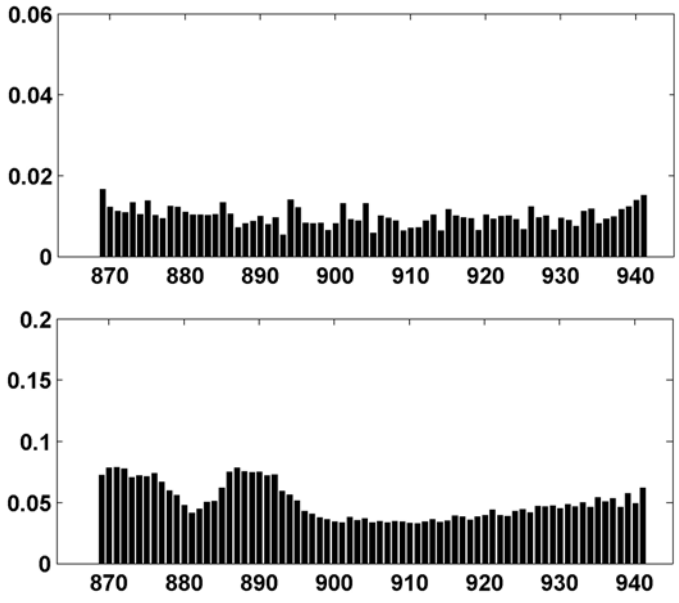


Figura 5.6: Gráficos mostrando distâncias (barras verticais) entre quadros para um segmento de vídeo de histeroscopia diagnóstica em particular. O eixo horizontal representa cada quadro i na ordem temporal do vídeo. (a) Distâncias entre quadros adjacentes $n = 1$; (b) Distância média $\widehat{D_i}$ de um quadro X_i em relação a todos os quadros X_j que compõe o segmento de vídeo, $n > 1$. Alguns picos e vales tornam-se mais evidentes para valores maiores de n.

um tamanho flexível de vizinhança temporal (quantidade de quadros) para cada quadro i do vídeo, a qual utiliza-se para propósitos de quantificação da redundância visual associada a cada quadro. Assim, a utilização de um valor não flexível para o tamanho de vizinhança faz de etapas de pós-processamento e seleção de quadros-chave passos críticos nos métodos de sumarização apresentados nas seções 5.2.1 e 5.2.2 (GAVIãO et al., 2007; SCHARCANSKI; GAVIãO, 2006). Com o objetivo de contornar estas limitações, apresenta-se nos próximos capítulos uma abordagem capaz de identificar segmentos de vídeos de histeroscopias com base em estimativas de sobreposição de conteúdo entre quadros vizinhos do vídeo, o que permite uma análise da vizinhança de um quadro adaptada a variações temporais do conteúdo do vídeo.

6 SUMARIZAÇÃO DE VÍDEOS DE HISTEROSCOPIAS DIAGNÓSTICAS BASEADA EM MOVIMENTOS DE CÂMERA

Neste capítulo apresenta-se um método capaz de organizar o conteúdo visual de vídeos de histeroscopias, permitindo que especialistas possam realizar uma navegação (*browsing*) rápida através do conteúdo dos vídeos. O método explora o rastreamento de pontos através dos quadros como forma de quantificar alterações no campo de visão. A idéia central é que se um conjunto de pontos é observado através de uma seqüência de quadros, estes quadros apresentam uma quantidade de sobreposição de conteúdo visual e podem ser agrupados em algum nível de similaridade visual.

A seção 6.1 apresenta uma visão geral do método proposto. Na seção 6.2 discute-se o método empregado para computar o conjunto inicial de pontos correspondentes através da seqüência de quadros, bem como a metodologia empregada na remoção da distorção provocada pela lente do equipamento ótico. Na seção 6.3 apresenta-se a estratégia de validação geométrica das correspondências computadas no passo anterior. A seção 6.4 apresenta o método proposto para representar vídeos de histeroscopias em termos de seu conteúdo visual. Por fim, a seção 6.5 apresenta o critério adotado para a escolha de quadros-chave, os quais constituirão o sumário do vídeo que guiará o especialista na atividade de *browsing* dos quadros do vídeo.

6.1 Visão Geral do Método

Como discutido na seção 2.2.2, seqüências de quadros relevantes estão associadas com movimentos lentos de câmera em vídeos de histeroscopias diagnósticas (SCHARCANSKI; GAVIãO, 2006; GAVIãO et al., 2007; SCHARCANSKI; GAVIãO; CUNHA-FILHO, 2005; GAVIãO; SCHARCANSKI, 2005; CUNHA-FILHO et al., 2004). Sendo assim, propõe-se reconhecer/sumarizar segmentos relevantes nestes vídeos através da análise do movimento 2-D induzido nas imagens pelo movimento 3-D da câmera no espaço. Neste sentido, assume-se a cavidade uterina como um ambiente rígido e adota-se uma abordagem conhecida como *Structure-From-Motion* (SFM) (MA et al., 2003) para rastrear e classificar pontos através dos quadros como sendo consistentes com um modelo paramétrico de movimento de câmera. Movimentos não-rígidos e independentes são assumidos como associados a quadros corrompidos por fatores biológicos indesejáveis, como discutido na seção 2.2.1. Assim, para pares de quadros na seqüência do vídeo, pontos são rastreados e geometricamente validados de acordo com um modelo de movimento de câmera. Pontos classificados como válidos são referidos como inliers e, uma vez que eles tenham sido computados, simplesmente mede-se alterações no campo de visão pela persistência dos inliers através da seqüência de quadros.

A fim de estruturar vídeos de histeroscopias em termos de alterações no campo de visão, um processo iterativo agrupa quadros vizinhos de acordo com a quantidade de inliers que eles tem em comum. Quadros vizinhos que apresentam uma alta quantidade de inliers são agrupados primeiro. De fato, estes grupos formam segmentos de video, os quais são agrupados recursivamente seguindo o mesmo critério, isto é, segmentos de vídeo vizinhos que apresentam maiores quantidades de inliers em comum são agrupados primeiro. O processo iterativo pára quando não há mais segmentos vizinhos a serem agrupados, o que acontece quando não há mais inliers em comum entre os segmentos formados, ou quando a quantidade de inliers observada em segmentos de vídeo vizinhos decresce abaixo de um limiar específico. Além disso, para cada um dos segmentos de vídeo formados associa-se um quadro-chave. Assim, ao final do processo iterativo, um conjunto de segmentos de vídeo é gerado e, para cada um deles, um quadro-chave é selecionado, formando deste modo o sumário do vídeo.

Cada segmento de vídeo gerado pode ser representado por uma estrutura de árvore binária, a qual guarda a informação sobre os pares de segmentos de vídeo que foram sendo agrupados recursivamente até a constituição de cada segmento resultante. Esta representação é uma maneira útil de organizar o conteúdo de vídeos de histeroscopias diagnósticas, uma vez que ela permite aos especialistas gerar um sumário do vídeo com mais, ou menos, detalhes/quadros-chave sem introduzir quadros espúrios no sumário gerado. Isto pode ser realizado simplesmente percorrendo-se a árvore do segmento de vídeo através de seus níveis: a partir do topo em direção a nodos folha para aumentar a quantidade de quadros-chave, e a partir de nodos folha em direção ao topo para gerar um sumário mais compacto do vídeo, como ilustrado na Figura 6.5.

6.2 Correspondências entre Quadros do Vídeo

O ponto de partida para a abordagem proposta é a detecção e rastreamento de pontos através de cada par de quadros consecutivos, I^j e I^{j+1}, na seqüência do vídeo. Para este propósito emprega-se o rastreador de pontos KLT (seção 4.1.1), uma vez que este método tem mostrado resultados satisfatórios em cenas endoscópicas (RAI; MERRITT; HIGGINS, 2006; WU; SUN; CHANG, 2007).

O algoritmo KLT entrega um conjunto de potenciais pontos correspondentes $\{\mathbf{x}_i^j \rightarrow \mathbf{x}_i^{j+1}\}$ para cada par de quadros consecutivos I^j e I^{j+1} do vídeo. Por simplicidade de notação abrevia-se o conjunto de potenciais pontos correspondentes entre os quadros I^j e I^{j+1} do vídeo como $\{p^j\}$. Deste modo, sendo N a quantidade de quadros no vídeo, tem-se que $j = 1 \cdots N-1$. Além disso, sendo M a quantidade de pontos correspondentes rastreados do quadro I^j para I^{j+1}, tem-se $i = 1 \cdots M$.

A fim de manter-se uma quantidade constante de M pontos para cada quadro, se k pontos são perdidos em um quadro I^j (por exemplo, saíram do campo de visão), k novos pontos são detectados e rastreados a partir de I^j, como implementado em (BIRCHFIELD, 2006).

O próximo passo consta de uma validação geométrica das correspondências computadas para pares de quadros I^j e $I^{j+\Delta}$ ($\Delta > 1$), onde assume-se um projeção linear do espaço 3-D para o plano 2-D das imagens. Por esta razão, se faz necessário a remoção das distorções não lineares causadas pela lente endoscópica. Os coeficientes de distorção da lente, juntamente com os parâmetros internos de câmera $\mathbf{K}$, são estimados no processo de calibração da câmera, como apresentado na seção 4.1.5. Assim, por simplicidade,

refere-se a $\{\mathbf{x}_i^j\}_{i=1}^M$ como o conjunto de pontos detectados no quadro I^j, cujas coordenadas já sofreram o processo de remoção de distorção de acordo com a matriz de calibração estimada $\mathbf{K}$ e os coeficientes de lente também estimados.

6.3 Validação Geométrica das Correspondências

Nesta seção apresenta-se a abordagem para integrar o processo de rastreamento de pontos e a validação geométrica dos mesmos segundo um modelo rígido de movimento de câmera. Isso se faz necessário pois os pontos entregues pelo algoritmo KLT são rastreados independentemente e de acordo com um modelo (simplificado) de movimento de translação, o qual não é apropriado para explicar movimentos de câmera em cenas mais genéricas. Além disso, assume-se a cavidade uterina como um ambiente rígido no qual movimentos não-rígidos e independentes são atribuídos a características biológicas (seção 2.2.1), as quais obstruem detalhes do útero e fazem com que os quadros tornem-se inapropriados para propósitos de diagnósticos, como mostrado na Figura 2.3.

6.3.1 Restrição Epipolar

Neste trabalho, objetiva-se quantificar movimentos de câmera pela análise do movimento 2-D detectado nas imagens. Deste modo, o processo de validação geométrica consiste em selecionar pontos correspondentes de $\{p^j\}$ que sejam consistentes com um modelo de movimento de câmera. Essencialmente utiliza-se a restrição epipolar (Eq. 4.4) sobre pares de pontos correspondentes $\{\mathbf{x}_i^j \rightarrow \mathbf{x}_i^{j+\Delta}\}$ em $\{p^j\}$. Devido ao ruído, pontos correspondentes consistentes (denominados de inliers) são selecionados em termos da distância epipolar simétrica, conforme a Equação 4.14. Assim, dada uma estimativa de $\mathbf{E}$ entre os quadros I^j e $I^{j+\Delta}$, um par de pontos correspondentes $\mathbf{x}_i^j \rightarrow \mathbf{x}_i^{j+\Delta}$ é considerado como um inlier se

$$d(\mathbf{x}_i^{j+\Delta}, \mathbf{E}\mathbf{x}_i^j) + d(\mathbf{x}_i^j, \mathbf{E}^\mathsf{T}\mathbf{x}_i^{j+\Delta}) < \tau, \tag{6.1}$$

onde $d(\cdot)$ representa a distância ortogonal entre um ponto e uma reta e τ é um limiar de erro.

6.3.2 Cenas Degeneradas e Quase-Degeneradas

Vídeos de histeroscopias diagnósticas são adquiridos com uma câmera manualmente guiada por um operador e com isso configurações distintas de cenas podem ser observadas em algumas fases típicas do exame histeroscópico, como introduzido na seção 2.2.3. Infelizmente, algumas configurações de cena são críticas no processo de estimar uma relação de movimento de câmera entre dois quadros de um vídeo, o que faz necessário uma etapa preliminar de detecção de cenas degeneradas, como discutido na seção 4.1.4. Por exemplo, no contexto das fases típicas do exame de histeroscopia, a inspeção do *fundo do útero* caracteriza cenas que são degeneradas estruturalmente, onde o *layout* tridimensional da cena é aproximadamente planar (veja seção 4.1.4). Neste caso, uma homografia $\mathbf{H}$ poderia ser empregada como uma relação de movimento de câmera entre os quadros, sendo suficiente para determinar um conjunto de inliers consistente com o movimento de câmera. Por outro lado, a fase do *exame panorâmico* tipicamente caracteriza uma configuração de cena genérica, onde uma matriz essencial $\mathbf{E}$ poderia ser estimada e utilizada para computar o conjunto de inliers. Em resumo, *degenerações de movimento* e *degenerações estruturais* são tipos comuns de degenerações de cenas em vídeos de histeroscopias.

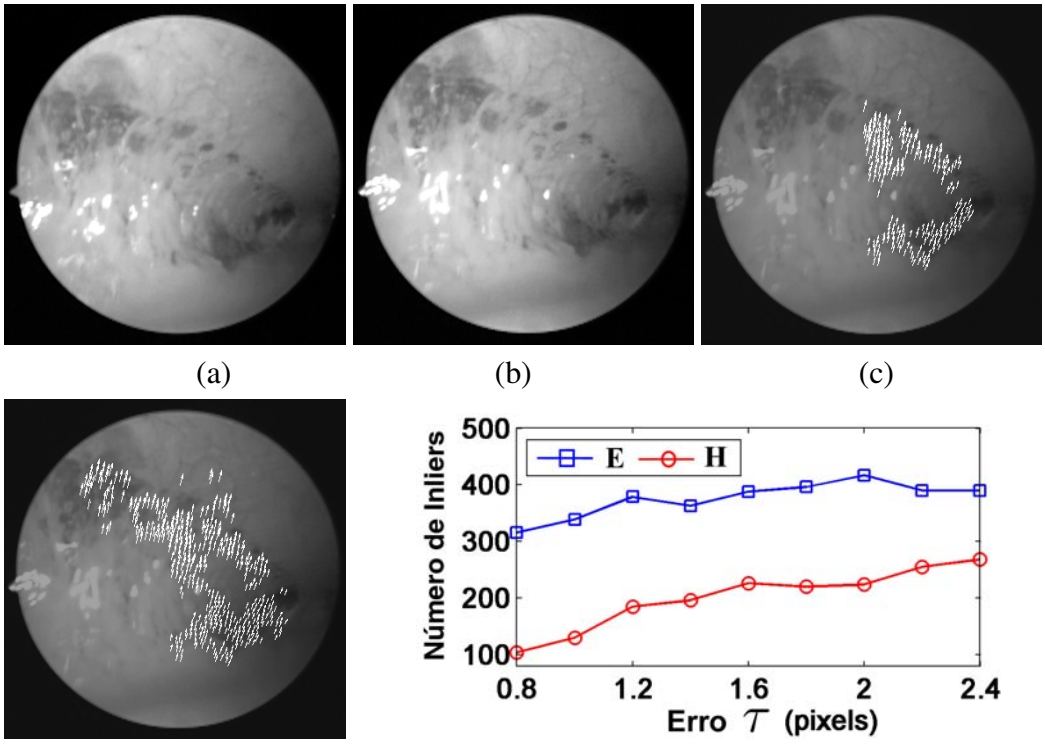


Figura 6.1: Quantidade de inliers computados entre os quadros (a) e (b) para dois modelos de movimento: uma matriz essencial $\mathbf{E}$ e uma homografia $\mathbf{H}$. (c-d) Inliers sobrepostos sobre o segundo quadro para $\mathbf{H}$ (c) e $\mathbf{E}$ (d) ($\tau = 2$ pixels). (e) Comparação entre os modelos $\mathbf{E}$ e $\mathbf{H}$ considerando diferentes níveis τ de erro. O modelo $\mathbf{E}$ é um modelo de movimento menos restritivo que $\mathbf{H}$, conseqüentemente $\mathbf{E}$ pode explicar pontos correspondentes que $\mathbf{H}$ não pode e, por esta razão, geralmente entrega uma quantidade maior de inliers.

Deste modo, configurações de cenas necessitam ser detectadas a fim de contornar limitações dos modelos utilizados para estimar movimentos de câmera. Caso contrário, em contextos de cenas degeneradas por exemplo, métodos que estimam $\mathbf{E}$, como o algoritmo de oito pontos, podem incluir outliers como suporte da solução, produzindo resultados de baixa confiabilidade (seção 4.1.4).

Além de um certo nível de acurácia, objetiva-se também estimar o modelo apropriado de câmera para computar tantos inliers quanto possível, uma vez que a abordagem proposta fundamenta-se no consenso de inliers. Como discutido na seção 5.1, muitas abordagens na literatura de recuperação de vídeos propõem o emprego de um modelo (simplificado) afim de movimento (que é um caso particular de um modelo de homografia $\mathbf{H}$) para explicar o movimento de câmera em vídeos de modo geral. Contudo, em algumas cenas histeroscópicas típicas, este modelo restringiria a quantidade de inliers, mantendo somente aqueles que são consistentes com um parte da cena. Essa situação é demonstrada na Figura 6.1 para um par de quadros histeroscópicos típico, onde compara-se a quantidade de inliers provida pelos modelos $\mathbf{H}$ e $\mathbf{E}$ para diferentes níveis de erros τ. Ambos modelos foram estimados segundo a metodologia RANSAC, como descrito na seção 4.2.

A fim de tratar cenas degeneradas e suas conseqüentes implicações, adota-se o algoritmo QDEGSAC como descrito na seção 4.2. O QDEGSAC pode detectar configurações de cenas degeneradas automaticamente e escolher um potencial modelo de movimento

de câmera entre dois quadros. Além disso, tal abordagem é capaz de tratar correspondências incorretas presentes no conjunto de potenciais correspondências $\{p^j\}$, bem como encontrar inliers a partir de pontos correspondentes cujas coordenadas nas imagens estão contaminadas por ruído. Como saída, o algoritmo QDEGSAC entrega um conjunto de inliers $\{in\}$ para cada par de quadros considerados na seqüência do vídeo. Somente pontos rastreados como inliers serão considerados nas próximas fases da abordagem proposta nesta tese, as quais são detalhadas nas próximas seções.

6.4 Representação Hierárquica para Vídeos de Histeroscopias

Nesta seção apresenta-se uma representação hierárquica para que especialistas possam realizar uma busca por conteúdo em vídeos de histeroscopias de uma maneira não seqüencial. Basicamente, explora-se o comportamento de pontos geometricamente rastreados através do vídeo. Estes pontos são consistentes com um modelo de movimento de câmera (seção 6.4.1), logo, a persistência destes pontos através dos quadros pode ser utilizada para estimar alterações no campo de visão que são decorrentes de movimentos da câmera. Uma seqüência de quadros que apresenta grandes quantidades destes pontos rastreados através de si revela um campo de visão mais estável (seção 6.4.2). Deste modo, propõe-se estruturar um vídeo de histeroscopia em segmentos de vídeo, os quais são formados em um processo de agrupamento de quadros vizinhos. Quadros são agregados a segmentos de vídeo com base na quantidade de pontos que persistem através deles (seção 6.4.3). Pontos correspondentes geometricamente validados são denominados de inliers $\{in\}$ e são computados pelo algoritmo QDEGSAC, que é aplicado sobre pares de quadros separados por intervalos regulares de Δ quadros, como ilustrado na Figura 6.2.

Antes de formalizar as idéias que estão por trás da representação proposta, alguns pontos são justificados:

- Utilizando-se o rastreador de pontos KLT, pontos correspondentes $\{p^j\}_{j=1}^{N-1}$ são computados para cada par de quadros consecutivos I^j e I^{j+1} através do vídeo. Computar pontos correspondentes para todos os quadros vizinhos do vídeo é uma prática adotada para estabelecer correspondências entre quadros cuja distância temporal é de $\Delta > 1$ quadros, uma vez que o casamento automático de feições é uma tarefa difícil e melhores resultados são alcançados quando o movimento de câmera entre quadros é relativamente pequeno. Contudo, a fim de reduzir um esforço computacional desnecessário no contexto da abordagem proposta nesta tese, não computa-se inliers para quadros consecutivos I^j e I^{j+1} através do vídeo. Utiliza-se uma amostragem de $\Delta > 1$ para validar geometricamente (como inliers) os pontos correspondentes entregues pelo algoritmo KLT. Isso não deve causar problemas significativos, uma vez que segmentos de vídeos de histeroscopias relevantes são geralmente adquiridos com movimentos lentos de câmera, o que produz uma quantidade excessiva de quadros (pontos de vistas redundantes de uma mesma região do útero) mesmo para propósitos clínicos. Por outro lado, pontos que estão sendo rastreados são eventualmente perdidos, ou saem do campo de visão, visto que o especialista movimenta a câmera no sentido de observar outras regiões do útero. Por esta razão, a quantidade de pontos correspondentes previamente estabelecidos entre pontos de vista muito separados pode não ser suficiente para estimar confiavelmente uma relação de movimento como uma matriz essencial $\mathbf{E}$ ou uma homografia $\mathbf{H}$. Sendo assim, opta-se por um valor conservador (pequeno) para Δ, uma vez que o

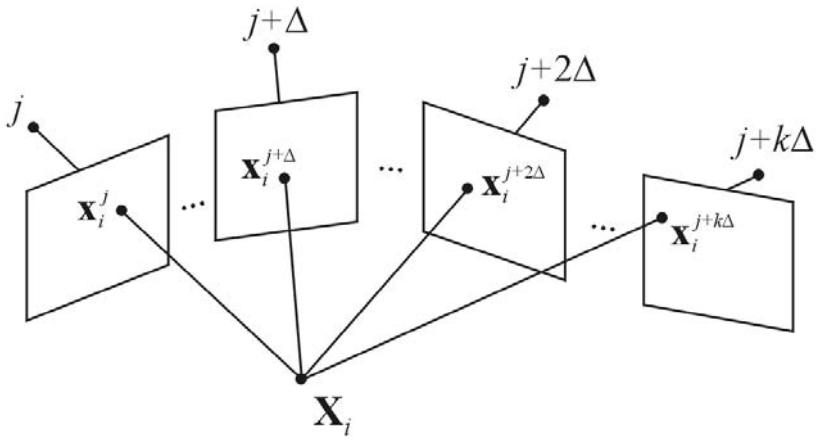


Figura 6.2: Notação associada aos quadros considerados no processo de validação geométrica de pontos correspondentes. Um ponto $\mathbf{X}_i$ no espaço é projetado sobre os quadros I^j, $I^{j+\Delta}$, $I^{j+2\Delta}$ e $I^{j+k\Delta}$ respectivamente como $\mathbf{x}_i^j$, $\mathbf{x}_i^{j+\Delta}$, $\mathbf{x}_i^{j+2\Delta}$ e $\mathbf{x}_i^{j+k\Delta}$, estabelecendo pontos correspondentes como $\mathbf{x}_i^j \leftrightarrow \mathbf{x}_i^{j+\Delta}$ e $\mathbf{x}_i^{j+\Delta} \leftrightarrow \mathbf{x}_i^{j+2\Delta}$.

interesse é construir uma representação de vídeo que inicie quantificando pequenas alterações de pontos de vista (entre quadros temporalmente próximos), e ao mesmo tempo evitando-se computar relações entre quadros consecutivos, I^j e I^{j+1}, e potencialmente redundantes.

- Dois quadros I^j e $I^{j+\Delta}$ são considerados irrelevantes, e uma relação entre eles não é computada via QDEGSAC, se o quantidade de pontos correspondentes M entre eles está abaixo de um limiar ζ, cujo valor também é escolhido de uma maneira conservadora (isto é, $\zeta = 50$ pontos).

De acordo com os experimentos realizados, o método proposto não é sensível aos parâmetros fixos Δ e ζ, os quais podem ser escolhidos dentro de uma faixa de valores razoável sem um impacto significativo nos resultados.

6.4.1 Pontos Consistentes

Um ponto $\mathbf{x}_i^j$ detectado no quadro I^j é considerado como um *ponto consistente* se as correspondências $\mathbf{x}_i^{j-\Delta} \leftrightarrow \mathbf{x}_i^j$ e $\mathbf{x}_i^j \leftrightarrow \mathbf{x}_i^{j+\Delta}$ são validadas como inliers pelo algoritmo QDEGSAC, onde $\mathbf{x}_i^{j-\Delta}$, $\mathbf{x}_i^j$ e $\mathbf{x}_i^{j+\Delta}$ são projeções do ponto $\mathbf{X}_i$ no espaço sobre os quadros $I^{j-\Delta}$ (vizinho da esquerda), I^j e $I^{j+\Delta}$ (vizinho da direita) respectivamente. O conjunto de pontos consistentes computados para um quadro I^j é representado como $\{\mathbf{x}_{con}^j\}$.

Formalmente, seja $\{in\}_j^{j+\Delta}$ o conjunto de inliers computados pelo algoritmo QDEGSAC entre os quadros I^j e $I^{j+\Delta}$. Da mesma forma que $\{in\}_{j-\Delta}^{j}$ representa o conjunto de inliers computados entre os quadros $I^{j-\Delta}$ e I^j. Um ponto $\mathbf{x}_i^j$ é um ponto consistente associado ao quadro I^j, isto é, $\mathbf{x}_i^j \in \{\mathbf{x}_{con}^j\}$ se

$$(\mathbf{x}_i^{j-\Delta} \leftrightarrow \mathbf{x}_i^j) \in \{in\}_{j-\Delta}^{j} \quad \text{e} \quad (\mathbf{x}_i^j \leftrightarrow \mathbf{x}_i^{j+\Delta}) \in \{in\}_j^{j+\Delta} \tag{6.2}$$

6.4.2 Pontos Persistentes e Sobreposição de Conteúdo

Cada quadro I^j é agora representado pelo seu conjunto de pontos consistentes $\{\mathbf{x}_{con}^j\}$. Assim, dados dois quadros I^j e $I^{j+k\Delta}$ como ilustrado na Figura 6.2, um ponto $\mathbf{x}_i^j$ é considerado como um *ponto persistente* de I^j a $I^{j+k\Delta}$ se

$$\mathbf{x}_i^{j+t\Delta} \in \{\mathbf{x}_{con}^{j+t\Delta}\} \quad \text{para} \quad t = 0 \ldots k, \tag{6.3}$$

o que significa que $\mathbf{x}_i^j$ e seus pontos correspondentes no intervalo de quadros de $I^{j+\Delta}$ a $I^{j+k\Delta}$ devem ser pontos consistentes.

Uma vez definida a idéia de persistência de pontos consistentes, apresenta-se o noção de sobreposição de conteúdo entre quadros. Dados dois quadros I^j e $I^{j+k\Delta}$, como ilustrado na Figura 6.2, e seus respectivos conjuntos de pontos consistentes $\{\mathbf{x}_{con}^j\}$ e $\{\mathbf{x}_{con}^{j+k\Delta}\}$, define-se a sobreposição de conteúdo $\theta_j^{j+k\Delta}$ entre os quadros I^j e $I^{j+k\Delta}$ como a quantidade de pontos persistentes de I^j a $I^{j+k\Delta}$.

Seja $per(\mathbf{x}_i^j, k)$ uma função booleana que verifica a persistência do ponto $\mathbf{x}_i^j$ no intervalo de quadros de I^j a $I^{j+k\Delta}$:

$$per(\mathbf{x}_i^j, k) = \begin{cases} 1, & \text{se } \mathbf{x}_i^{j+t\Delta} \in \{\mathbf{x}_{con}^{j+t\Delta}\} \quad \text{para} \quad t = 0 \ldots k \\ 0, & \text{caso contrário.} \end{cases} \tag{6.4}$$

Assim, dados dois quadros quaisquer I^j e $I^{j+k\Delta}$, defini-se a sobreposição de conteúdo $\theta_j^{j+k\Delta}$ entre estes quadros como

$$\theta_j^{j+k\Delta} = \sum_{i=1}^{M} per(\mathbf{x}_i^j, k), \tag{6.5}$$

onde M é a quantidade de pontos correspondentes estabelecidos pelo algoritmo KLT entre os quadros I^j e I^{j+1}.

Utiliza-se então a sobreposição de conteúdo como métrica de similaridade entre quadros. Por exemplo, dados três quadros $I^{j-\Delta}$, I^j e $I^{j+\Delta}$, a sobreposição de conteúdo $\theta_{j-\Delta}^j$ entre os quadros $I^{j-\Delta}$ e I^j é maior que a sobreposição $\theta_j^{j+\Delta}$ entre I^j e $I^{j+\Delta}$ se

$$\theta_{j-\Delta}^j \; > \; \theta_j^{j+\Delta}, \tag{6.6}$$

o que significa que os quadros $I^{j-\Delta}$ e I^j contêm mais pontos consistentes rastreados através de si do que os quadros I^j e $I^{j+\Delta}$.

6.4.3 Árvores de Segmentos de Vídeo

Um processo iterativo é empregado para agrupar quadros em segmentos de vídeo. Um segmento de vídeo é notado como $\delta^{a \mapsto b}$, onde a representa o primeiro quadro do segmento e b representa o último quadro do segmento. Novos segmentos são formados com base no teste de sobreposição de conteúdo da expressão 6.6, onde quadros que apresentam maior sobreposição de conteúdo são agrupados primeiro.

Dados quatro quadros Δ-espaçados na seqüência temporal do vídeo $I^{j-2\Delta}$, $I^{j-\Delta}$, I^j e $I^{j+\Delta}$, e seus respectivos conjuntos de pontos consistentes $\{\mathbf{x}_{con}^{j-2\Delta}\}$, $\{\mathbf{x}_{con}^{j-\Delta}\}$, $\{\mathbf{x}_{con}^j\}$ e $\{\mathbf{x}_{con}^{j+\Delta}\}$. Os dois quadros centrais $I^{j-\Delta}$ e I^j serão agrupados em um novo segmento

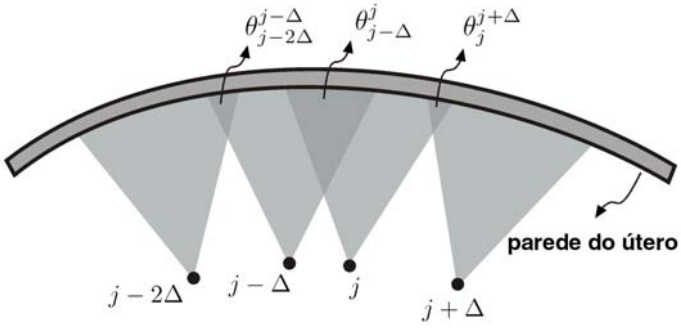


Figura 6.3: A sobreposição dos campos de visão determinará os quadros que serão agrupados primeiro. Os quadros centrais $I^{j-\Delta}$ e I^{j} constituirão o segmento de vídeo $\delta^{j-\Delta \mapsto j}$ se a sobreposição de conteúdo entre eles $\theta^{j}_{j-\Delta}$ é maior que as sobreposições de conteúdo $\theta^{j-\Delta}_{j-2\Delta}$ e $\theta^{j+\Delta}_{j}$, as quais foram computadas com relação a seus quadros vizinhos $I^{j-2\Delta}$ e $I^{j+\Delta}$.

de vídeo $\delta^{j-\Delta \mapsto j}$ se a sobreposição $\theta^{j}_{j-\Delta}$ entre os quadros $I^{j-\Delta}$ e I^{j} é maior que as sobreposições $\theta^{j-\Delta}_{j-2\Delta}$ e $\theta^{j+\Delta}_{j}$ que envolvem os quadros $I^{j-2\Delta}$ e $I^{j+\Delta}$, ou seja, se

$$\theta^{j-\Delta}_{j-2\Delta} \; < \; \theta^{j}_{j-\Delta} \; > \; \theta^{j+\Delta}_{j} \tag{6.7}$$

A Figura 6.3 ilustra esta idéia em termos de quatro pontos de vista e as sobreposições de campo de visão que podem ocorrer entre eles. Neste sentido é razoável assumir que a sobreposição de campo de visão é proporcional a métrica de sobreposição de conteúdo expressa na Equação 6.5, uma vez que quanto maior é a região da cena comum a dois quadros, maior a chance destes quadros apresentarem uma grande quantidade de pontos consistentes rastreados através de si.

Da mesma forma que quadros são representados individualmente pelo seu conjunto de pontos consistentes, segmentos de vídeo também são representados por pontos consistentes que persistem do início ao fim do segmento. Por exemplo, um segmento $\delta^{j-\Delta \mapsto j}$ possui um conjunto de pontos consistentes associado, notado como $\{\mathbf{x}^{j-\Delta \mapsto j}_{con}\}$, do qual só fazem parte os pontos consistentes que persistem do quadro $I^{j-\Delta}$ ao quadro I^{j}, conforme definido na Expressão 6.3.

Dada uma seqüência de quadros I^{a}, I^{b}, I^{c} e I^{d} tomada de um vídeo. Durante o processo iterativo de formação de segmentos, um segmento de vídeo $\delta^{b \mapsto c}$ agregará quadros, e a quantidade de pontos persistentes através de $\delta^{b \mapsto c}$ tenderá a decrescer, uma vez que pontos persistentes são eventualmente perdidos ou movem-se para fora do campo de visão devido ao movimento de câmera. Neste sentido, um segmento de vídeo $\delta^{b \mapsto c}$ é considerado estável e não agregará mais quadros quando a sobreposição de conteúdo entre ele e seus quadros vizinhos, I^{a} e I^{d}, cai abaixo de um limiar ζ, ou seja, quando

$$\theta^{a \mapsto b}_{a} \; < \; \zeta \quad \text{e} \quad \theta^{d}_{a \mapsto b} \; < \; \zeta, \tag{6.8}$$

onde $\theta^{a \mapsto b}_{a}$ representa a sobreposição de conteúdo entre o quadro I^{a} e o segmento de vídeo $\delta^{b \mapsto c}$ (Eq. 6.5), enquanto $\theta^{d}_{a \mapsto b}$ representa a sobreposição de conteúdo entre o segmento $\delta^{b \mapsto c}$ e seu quadro vizinho I^{d}. Note que I^{a} e I^{d} são os quadros considerados para serem agregados ao segmento $\delta^{b \mapsto c}$ pelo processo iterativo.

Em cada iteração sobre a seqüência de quadros, o processo iterativo analisa quatro quadros na seqüência temporal por vez I^{a}, I^{b}, I^{c} e I^{d}, formando um novo segmento

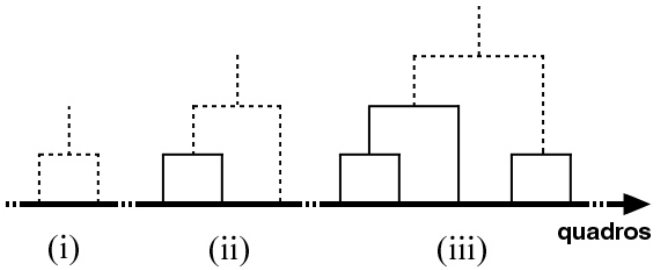


Figura 6.4: O processo iterativo constrói segmentos de vídeo (i) agrupando quadros e formando novos segmentos, (ii) agregando quadros a segmentos de vídeo já existentes ou (iii) agrupando segmentos de vídeo vizinhos em segmentos maiores.

de vídeo $\delta^{b \mapsto c}$ sempre que a expressão 6.7 torna-se verdadeira (isto é, sempre que θ_b^c é maior que θ_c^d e θ_a^b). Cada novo segmento será representado por um conjunto de pontos consistentes $\{\mathbf{x}_{con}^{b \mapsto c}\}$ que persiste do quadro I^b ao quadro I^c. Assim, $\{\mathbf{x}_{con}^{b \mapsto c}\}$ substitui os dois conjuntos de pontos consistentes, $\{\mathbf{x}_{con}^{b}\}$ e $\{\mathbf{x}_{con}^{c}\}$, na representação do vídeo como um seqüência de conjuntos de pontos consistentes.

É importante notar que o processo de formar segmentos de vídeo procede de três maneiras: (i) agrupando quadros em novos segmentos de vídeo, (ii) agregando quadros a segmentos de vídeo já existentes ou (iii) agrupando segmentos de vídeo vizinhos em segmentos maiores, conforme ilustrado na Figura 6.4. O processo iterativo terá chegado ao fim quando os segmentos de vídeo tornarem-se estáveis, isto é, quando a sobreposição de conteúdo entre cada segmento de vídeo e seus vizinhos, sejam simples quadros ou outros segmentos de vídeo, decrescer abaixo de um limiar ζ, como representado na Expressão 6.8.

Para cada segmento de vídeo formado associa-se um quadro-chave, como descrito na seção 6.5. Ao final do processo, um conjunto de segmentos de vídeo terá sido constituído e, para cada um destes segmentos, um quadro-chave será definido, formando o sumário do vídeo que servirá para guiar o especialista através do conteúdo do vídeo. Uma vez que um especialista tenha escolhido um quadro-chave $I_{kf}^{j \mapsto j+k\Delta}$, o conteúdo do segmento de vídeo correspondente $\delta^{j \mapsto j+k\Delta}$ pode ser acessado. Neste sentido, propõe-se explorar a estrutura hierárquica deixada pelo processo de constituição de cada segmento de vídeo.

Cada segmento de vídeo constituído no final do processo pode ser representado por uma árvore binária, que mantém a informação sobre quais pares de segmentos de vídeo foram agrupados até a constituição final de tal segmento. A Figura 6.4(iii) ilustra uma árvore binária em particular, onde o passo final para constituir o segmento de vídeo é representado por linhas tracejadas, e os segmentos de vídeo agrupados são representados em linha cheia. Deste modo, denomina-se uma árvore binária associada a um segmento de vídeo de *árvore de segmento de vídeo*.

A representação do vídeo em termos de árvores de segmentos de vídeo é uma maneira útil de organizar o conteúdo de vídeos de histeroscopias para propósitos de *browsing* do vídeo, uma vez que permite-se a especialistas gerar um sumário de vídeo com mais, ou menos, detalhes/quadros-chave sem introduzir quadros espúrios no sumário do vídeo. Isso pode ser alcançado navegando-se através dos níveis de árvores de segmentos de vídeo: de níveis superiores para níveis inferiores no sentido de aumentar a quantidade de quadros-chave, e de níveis mais baixos para níveis mais altos para gerar sumários mais compactos do vídeo. Esta idéia é ilustrada na Figura 6.5, que mostra uma árvore

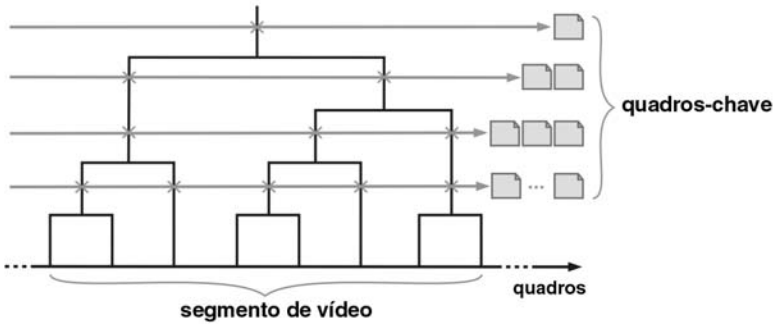


Figura 6.5: Uma árvore de segmento de vídeo particular na qual a tarefa de *browsing* de seus quadros é representada como uma linha horizontal (setas em cinza). Enquanto está linha desloca-se através dos níveis da árvore, sumários de vídeos mais, ou menos, compactos são gerados em termos de quantidade de quadros-chave. As intersecções $\times$ determinam uma hierarquia de sub-segmentos de vídeo (sub-árvores), os quais constituirão o segmento de vídeo final no topo da árvore.

de segmento de vídeo em particular, na qual a tarefa de navegar através do conteúdo do segmento de vídeo pode ser representada como uma linha horizontal imaginária (setas em cinza), onde sumários de vídeo com maior ou menor redundância/quadros-chave são gerados pelo deslocamento vertical desta linha. Em cada nível da árvore, sub-árvores (sub-segmentos) são determinadas pela intersecção desta linha com a estrutura da árvore. Assim como cada segmento de vídeo gerado no final do processo, um sub-segmento também tem associado a ele um quadro-chave, o qual constitui o conjunto de quadros-chave para o nível da árvore em questão, conforme ilustrado na Figura 6.5. Deste modo, uma vez que um especialista tenha selecionado um quadro-chave no topo de uma árvore de segmento de vídeo, ele pode navegar sobre o conteúdo do segmento de vídeo correspondente sem introduzir quadros que estariam fora do contexto do quadro-chave inicialmente selecionado.

Uma vez que uma árvore de segmento de vídeo delimita uma seqüência de quadros de conteúdo correlacionado, esta seqüência de vídeo delimitada pode ser entendida como um *shot*, que é uma unidade de vídeo amplamente explorada dentro da literatura de indexação de vídeos para representar o conteúdo de vídeos comerciais/editados (capítulo 3). Além disso, uma importante vantagem da representação em árvore é que permite-se uma atividade de *browsing* guiado do conteúdo de segmentos de vídeos. Uma vez que o processo de formação dos segmentos inicia agrupando quadros com um alto grau de sobreposição de conteúdo, quadros com menor grau de sobreposição de conteúdo são agrupados em níveis mais altos da árvore. Sendo assim, enquanto atravessa-se a árvore a partir do topo, quadros redundantes são progressivamente inseridos no sumário do vídeo. Isso ocorre, contudo, seguindo a estrutura da árvore, que por sua vez faz com que os sumários gerados tendam a conter quadros tão distintos quanto possível, o que é uma característica desejável em aplicações cujo o propósito é facilitar a navegação sobre o conteúdo de vídeos (isto é, evitando mostrar informações redundantes).

6.5 Seleção de Quadros-Chave

Dado um segmento de vídeo $\delta^{a \mapsto b}$, ou um sub-segmento, e seu conjunto de pontos

consistentes $\{\mathbf{x}_{con}^{a \mapsto b}\}$, o quadro com a maior sobreposição de conteúdo sobre os quadros do segmento de vídeo é selecionado como o quadro-chave $I_{kf}^{a \mapsto b}$. No sentido de quantificar este critério de seleção de quadros-chave, considera-se a *duração de um ponto consistente*, que é a quantidade de quadros consecutivos em que um ponto é rastreado como consistente. Cada quadro em um segmento de vídeo $\delta^{a \mapsto b}$ possui um conjunto de pontos consistentes associado $\{\mathbf{x}_{con}^{a}\}, \{\mathbf{x}_{con}^{a+\Delta}\}, \ldots, \{\mathbf{x}_{con}^{b-\Delta}\}, \{\mathbf{x}_{con}^{b}\}$, como definido na seção 6.4.1. Para cada um desses conjuntos computa-se a *duração média* (quantidade de quadros) de seus pontos consistentes, sendo que o quadro com a maior duração média dentro do segmento é selecionado como o quadro-chave $I_{kf}^{a \mapsto b}$.

Seja $con(\mathbf{x}_i^j)$ uma função booleana que verifica se $\mathbf{x}_i^j$ é um ponto consistente associado ao quadro I^j:

$$con(\mathbf{x}_i^j) = \begin{cases} 1, & \text{se } \mathbf{x}_i^j \in \{\mathbf{x}_{con}^j\} \\ 0, & \text{caso contrário.} \end{cases} \tag{6.9}$$

Dentro de um segmento de vídeo $\delta^{a \mapsto b}$, pode-se definir a *duração média* dos pontos $\{\mathbf{x}_i^j\}_{i=1}^M$ associados ao quadro I^j (onde $a \leq j \leq b$) como

$$\widehat{dur}(I^j) = \frac{\sum_{i=1}^{M} \sum_{t=a}^{b} con(\mathbf{x}_i^t)}{M}, \tag{6.10}$$

onde M é a quantidade de pontos rastreados pelo rastreador KLT em cada quadro. Assim, o quadro-chave $I_{kf}^{a \mapsto b}$ associado ao segmento de vídeo $\delta^{a \mapsto b}$ será o quadro I^j onde

$$I_{kf}^{a \mapsto b} = \arg \max_{j=a \ldots b} \widehat{dur}(I^j) \tag{6.11}$$

De fato, o valor de duração média associado a um quadro I^j em um segmento $\delta^{a \mapsto b}$ será determinado pela duração dos pontos consistentes em $\{\mathbf{x}_{con}^j\}$ que não estão dentro do conjunto $\{\mathbf{x}_{con}^{a \mapsto b}\}$. Isso deve-se ao fato de que os pontos consistentes em $\{\mathbf{x}_{con}^{a \mapsto b}\}$ aparecem em todos os quadros do segmento de vídeo $\delta^{a \mapsto b}$, conseqüentemente sua duração é constante para todos os quadros de $\delta^{a \mapsto b}$. Dessa forma, diferenças de valores de duração média ocorrem em termos de pontos consistentes que não estão presentes ao longo de todo o segmento de vídeo.

7 EXPERIMENTOS

Neste capítulo apresenta-se os experimentos realizados com dados sintéticos e vídeos reais. Testes foram conduzidos em 4 vídeos de histeroscopias diagnósticas pré-interpretados. A escolha destes vídeos é justificada na seção 7.2.2. Basicamente, avaliou-se o desempenho do método proposto sobre resultados do rastreamento de pontos consistentes através do vídeo e em termos do sumário de vídeo produzido. A Tabela 7.1 mostra os parâmetros da câmera histeroscópica utilizada na aquisição dos vídeos.

7.1 Experimentos com Dados Sintéticos

Inicialmente testa-se o comportamento do algoritmo QDEGSAC em dados sintéticos, uma vez que há poucos relatos na literatura sobre o desempenho de tal algoritmo em diferentes condições de ruído e quantidades de correspondências incorretamente estabelecidas (outliers, segundo uma relação corretamente estimada). Um conjunto de 100 pontos em 3D é gerado aleatoriamente com uma variação de profundidade de 10-50 unidades de distância focal (u.d.f). A rotação entre quadros é de $\alpha \in \{0, 5\}$ graus em torno de um eixo de rotação aleatório, sendo que a translação entre quadros é de $|\mathbf{t}| \in \{0, 5\}$ u.d.f com uma direção aleatória de translação. Os pontos de vista são obtidos por projeção perspectiva com imagens de dimensões de 512×512 pixels. Ruído gaussiano de média zero e desvio padrão de $\sigma = [0, 1]$ pixels é adicionado às coordenadas de pontos correspondentes em ambos quadros. As Figuras mostram medidas que resultam de uma média de 100 trials do algoritmo QDEGSAC.

A partir de dados sintéticos é possível avaliar os resultados entregues pelo algoritmo QDEGSAC, uma vez que a relação de movimento correta é conhecida e, conseqüentemente, os conjuntos reais de inliers e outliers (*ground truth*) também são conhecidos. Além disso, no contexto de cenas degeneradas, pode-se avaliar a qualidade dos resultados em termos de inliers degenerados, tendo em vista que a indesejável detecção de inliers não-degenerados ocorre por limitações do algoritmo em tratar o ruído e a presença de cor-

Tabela 7.1: Parâmetros estimados para a câmera histeroscópica utilizada na aquisição dos vídeos analisados nos experimentos

Distância Focal f	284.10 $\pm$ 0.75
Ponto principal $\mathbf{c} = (x_0, y_0)$	(159.14,194.11) $\pm$ (0.79,0.77)
Dimensão das Imagens	(373,373)
Distorção da lente (k_1,k_2)	(-0.440,0.182) $\pm$ (0.0046,0.0080)

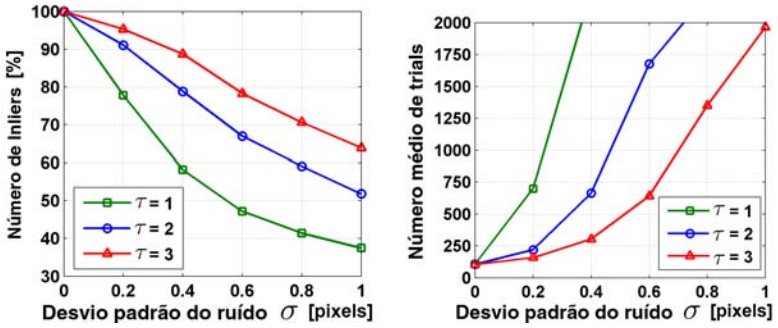


Figura 7.1: Quantidade de inliers computados e a quantidade média de trials requerida pelo algoritmo QDEGSAC para computar inliers como uma função do ruído, onde $|\mathbf{t}| = 5$ u.d.f e $\alpha = 5$ graus (configuração genérica de cena).

respondências incorretamente estabelecidas. A seguinte notação é utilizada na avaliação dos resultados:

- T_{Right} representa a relação correta de movimento de câmera entre os quadros (pontos de vista) considerados;
- $\{in\}_{Right}$ representa o *conjunto de inliers reais* segundo a relação T_{Right};
- $\{out\}_{Right}$ representa o *conjunto de outliers reais* presentes nos dados de entrada, ou seja, representa as correspondências que não dão suporte a relação correta de movimento T_{Right};
- $\{in\}^{\tau}_{QDEGSAC}$ representa o conjunto de inliers entregue pelo algoritmo QDEGSAC, onde τ é o limiar de erro (em pixels) utilizado para aceitar pontos correspondentes como inliers, segundo a relação de movimento estimada $T_{QDEGSAC}$.

As Figuras 7.1 e 7.2 mostram o desempenho do algoritmo QDEGSAC para computar uma relação com $(|\mathbf{t}|, \alpha) = (5, 5)$ e $(|\mathbf{t}|, \alpha) = (5, 0)$ respectivamente, como uma função do nível de ruído σ. A medição de desempenho se dá em termos da porcentagem de inliers computados. Assim, dado um limiar de erro de τ pixels para aceitar um par de pontos correspondentes como um inlier, computa-se a quantidade de inliers entregue pelo algoritmo QDEGSAC. Uma vez que os dados são contaminados apenas por ruído, a quantidade de inliers entregue pelo QDEGSAC deveria ser tão alta quanto possível. Resultados são mostrados para $\tau \in \{1, 2, 3\}$.

As Figuras 7.3 e 7.4 mostram o desempenho do QDEGSAC no contexto de configurações degeneradas de cenas para $(|\mathbf{t}|, \alpha) = (5, 5)$ sobre pontos coplanares e $(|\mathbf{t}|, \alpha) = (0, 5)$ (rotação pura). Estas figuras também mostram a proporção de inliers degenerados, a qual deveria ser tão alta quanto possível, uma vez que ambas configurações de cena são degeneradas e, conseqüentemente, a detecção de inliers não-degenerados ocorre devido ao ruído.

Dada a relação correta de movimento de câmera T_{Right} entre os pontos de vista, calcula-se o erro médio para os inliers $\{in\}^{\tau}_{QDEGSAC}$ entregues pelo algoritmo QDEGSAC, onde $\tau \in \{1, 2, 3\}$. Para pontos de vista degenerados, emprega-se uma homografia $\mathbf{H}$ como o modelo de movimento correto. Neste caso, utiliza-se o erro simétrico de transferência, que mede o quão próximo um par de pontos $\mathbf{x}_i \rightarrow \mathbf{x'}_i$ satisfaz a relação $\mathbf{H}$ estimada (HARTLEY; ZISSERMAN, 2000). O erro de transferência é a distância Euclidiana,

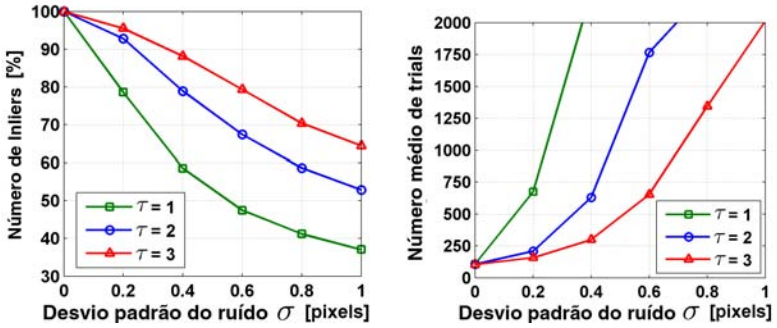


Figura 7.2: Quantidade de inliers computados e a quantidade média de trials requerida pelo algoritmo QDEGSAC para computar inliers como uma função do ruído, onde $|\mathbf{t}| = 5$ u.d.f e $\alpha = 0$ graus (configuração genérica de cena).

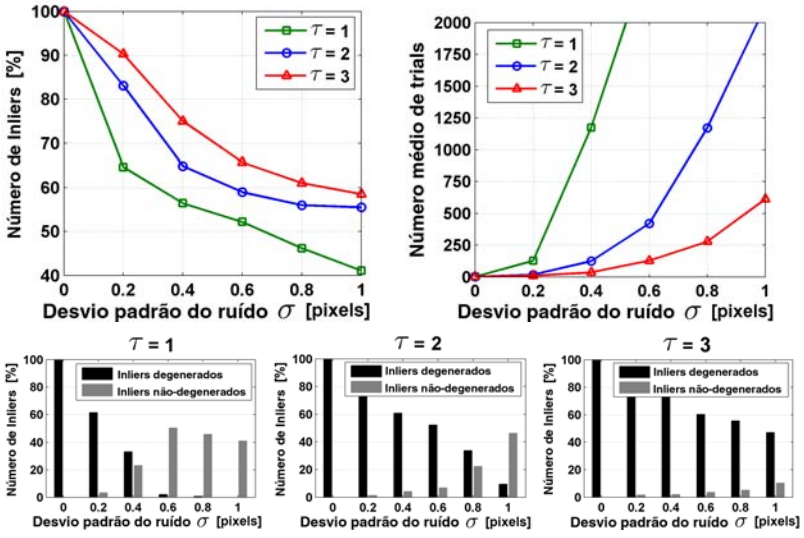


Figura 7.3: Desempenho do algoritmo QDEGSAC para cenas planares (degeneradas) como função do ruído, onde $(|\mathbf{t}|, \alpha) = (5, 5)$. (Acima) Quantidade de inliers computados e quantidade média de trials do algoritmo QDEGSAC para computá-los. (Abaixo) Proporção de inliers degenerados computados para $\tau \in \{1, 2, 3\}$.

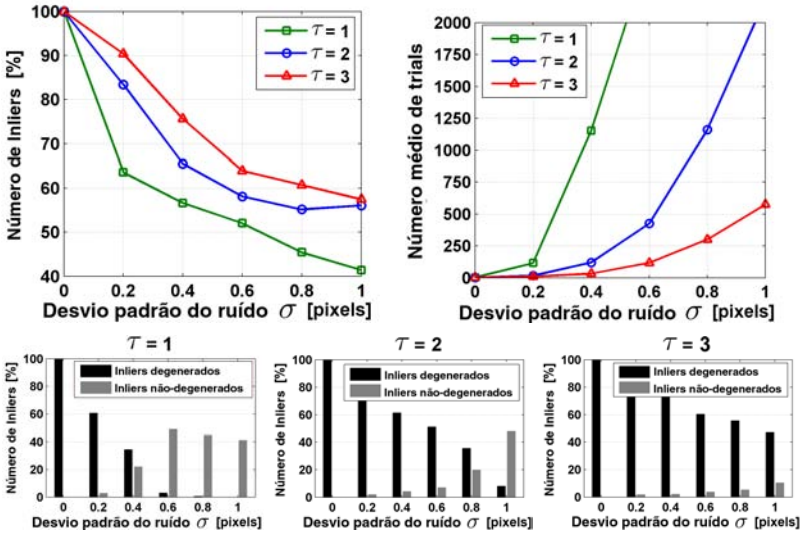


Figura 7.4: Desempenho do algoritmo QDEGSAC para cenas degeneradas (rotação pura) como função do ruído, onde $(|\mathbf{t}|, \alpha) = (0, 5)$. (Acima) Quantidade de inliers computados e quantidade média de trials do algoritmo QDEGSAC para computá-los. (Abaixo) Proporção de inliers degenerados computados para $\tau \in \{1, 2, 3\}$.

no plano da imagem, entre os pontos medidos e a projeção dos pontos correspondentes em ambas imagens:

$$d(\mathbf{x}_i, \mathbf{H}^{-1}\mathbf{x'}_i)^2 + d(\mathbf{x'}_i, \mathbf{H}\mathbf{x}_i)^2. \tag{7.1}$$

Para configurações de cenas não-degeneradas emprega-se o erro epipolar simétrico, como definido na Equação 4.14. A Figura 7.5 mostra o erro para ambas configurações de cenas, degeneradas e não-degeneradas, como uma função do nível de ruído. Cenas planares e apresentando apenas rotação de câmera apresentaram praticamente os mesmos valores de erro.

A Figura 7.5 também compara o erro computado sobre $\{in\}^{\tau}_{QDEGSAC}$ contra o erro *ground truth*, o qual é computado sobre o conjunto de inliers $\{in\}^{\tau}_{Right}$ entregue pela relação T_{Right}. Este conjunto é constituído por pontos correspondentes que produziram um erro menor que τ pixels em termos da relação correta de movimento T_{Right}. Deve-se notar que os dados são corrompidos apenas por ruído neste estágio, isto é, os dados não contém pontos correspondentes incorretamente estabelecidos. Com base nos gráficos/testes apresentados, algumas observações podem ser feitas:

1. O desempenho do algoritmo QDEGSAC deteriora com a quantidade de ruído.

2. Como esperado, o esforço computacional aumenta de acordo com a quantidade de ruído, uma vez que o ruído faz a proporção de inliers decrescer e, de acordo com a Equação 4.13, a quantidade necessária de trials S do RANSAC irá crescer.

3. A quantidade necessária de trials do QDEGSAC em configurações degeneradas de cena é significativamente menor que a quantidade de trials em casos não-degenerados. Em cenas não-degeneradas, o algoritmo QDEGSAC necessita de uma quantidade significativa de trials para provar que os dados não dão suporte para uma relação de movimento que emprega apenas 7 restrições. Por outro lado, quando o QDEGSAC atinge a dimensão 6 em casos de cenas degeneradas, o processo de redução

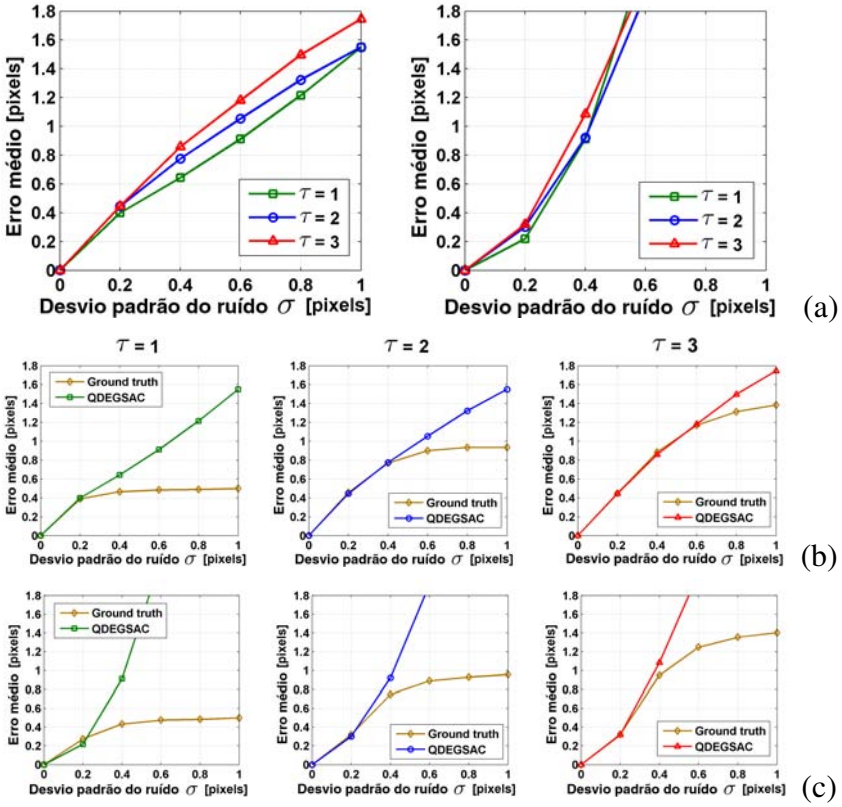


Figura 7.5: Em termos da relação correta de movimento T_{Right}, erros médios são computados para os inliers $\{in\}^{\tau}_{QDEGSAC}$ entregues pela algoritmo QDEGSAC ($\tau \in \{1, 2, 3\}$) como uma função de níveis de ruído. (a) À direita, erro computado para cenas não-degeneradas. À esquerda, erro computado para cenas planares (degeneradas), $(|\mathbf{t}|, \alpha) = (5, 5)$. (b-c) Erro computado sobre os inliers $\{in\}^{\tau}_{QDEGSAC}$ entregues pelo QDEGSAC contra o erro computado sobre o conjunto de inliers reais $\{in\}^{\tau}_{Right}$ (*ground truth*) para cenas não-degeneradas (b) e cenas degeneradas (c).

de dimensão é interrompido. Isso evita a execução de trials desnecessárias dentro do processo RANSAC(5) (como discutido na seção 4.2.3), que pode ser tão caro computacionalmente quanto o processo RANSAC(7) em cenas não-degeneradas.

4. O melhores resultados são alcançados quando emprega-se o QDEGSAC com $\tau = 3$. Esta configuração resulta no maior erro, contudo obtém-se o menor esforço computacional bem como uma esperada quantidade maior de inliers.

Deve-se observar que objetiva-se rastrear tantos inliers quanto possível através da seqüência de imagens. Conseqüentemente, o foco está em evitar a perda destes inliers por razões de ruído ou devido a presença de correspondências incorretas no conjunto inicial de pontos correspondentes provido pelo rastreador KLT. Sendo assim, uma vez que o conjunto de inliers reais $\{in\}_{Right}$ é conhecido, avalia-se a seguir como a presença de correspondências incorretas afeta o desempenho do algoritmo QDEGSAC em termos da perda destes inliers reais. Neste contexto, especial atenção é dada para valores maiores de τ, uma vez que, quanto maior o valor de τ, maior são as chances de classificar correspondências incorretas como inliers. Lembrando (seção 4.1.4) que correspondências incorretas podem levar a uma alta taxa de perda de inliers reais, desde que uma relação de

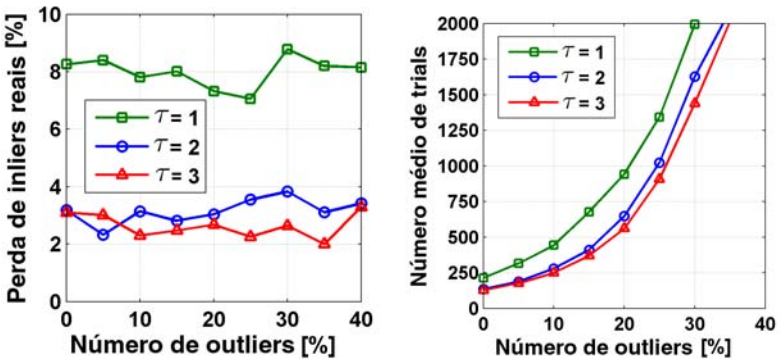


Figura 7.6: Desempenho do algoritmo QDEGSAC como um função da quantidade de outliers reais no contexto de cenas genéricas (não-degeneradas), onde $|\mathbf{t}| = 5$ u.d.f, $\alpha = 5$ graus e $\sigma = 0.1$ pixels. À esquerda, quantidade de inliers reais perdidos. À direita, quantidade média requerida de trials pelo algoritmo QDEGSAC.

movimento T_{out}, cujo suporte está em correspondências incorretas (outliers reais), pode interromper o processo de rastreamento de inliers reais, fazendo com que estes sejam considerados como outliers em termos de T_{out} e sejam conseqüentemente descartados.

As Figuras 7.6 e 7.7 mostram a quantidade de inliers reais perdidos como uma função da quantidade de outliers reais presente nos dados de entrada, para $(|\mathbf{t}|, \alpha) = (5, 5)$ e $(|\mathbf{t}|, \alpha) = (5, 0)$ respectivamente. Outliers reais são introduzidos no domínio da imagem, e ruído gaussiano com $\sigma = 0.1$ pixels é adicionado as coordenadas dos pontos correspondentes em ambas imagens.

A Figura 7.8 mostra a quantidade de inliers reais perdidos no contexto de cenas degeneradas para $(|\mathbf{t}|, \alpha) = (0, 5)$ (rotação pura de câmera). A Figura 7.8 também mostra a proporção de inliers corretamente detectados como degenerados. Uma vez que não há translação de câmera, a geometria epipolar não pode ser definida e, conseqüentemente, a quantidade de inliers degenerados entregue pelo QDEGSAC deveria ser tão alta quanto o total de inliers reais (*ground truth*) presentes nos dados de entrada. Assim, inliers não-degenerados são detectados incorretamente como inliers pela relação de movimento estimada, fato que decorre da influência negativa causada por ruído e/ou outliers reais presentes nos dados de entrada.

Com o objetivo de avaliar o quão diferentes são os inliers $\{in\}^{\tau}_{QDEGSAC}$ entregues pelo algoritmo QDEGSAC e os inliers reais $\{in\}_{Right}$ entregues pela relação correta de movimento T_{Right}, computa-se o erro sobre as correspondências classificadas como inliers pelo QDEGSAC utilizando-se a relação *ground truth* T_{Right}. Para cenas não-degeneradas (onde T_{Right} é uma matriz essencial $\mathbf{E}$) utiliza-se o erro epipolar simétrico, como definido na Equação 4.14, e para cenas degeneradas (onde T_{Right} é uma homografia $\mathbf{H}$) emprega-se o erro de transferência como definido na Equação 7.1. A Figura 7.9 mostra o erro para ambas configurações de cenas, degeneradas e não-degeneradas. A Figura 7.9 também compara o erro calculado sobre $\{in\}^{\tau}_{QDEGSAC}$ contra o erro computado sobre o conjunto de inliers reais $\{in\}_{Right}$ (erro *ground truth*). Além disso, com o objetivo de prover uma idéia sobre os outliers reais presentes nos dados sintéticos de entrada, calcula-se o erro sobre todos os pontos correspondentes, incluindo inliers reais e outliers reais.

A fim de verificar experimentalmente a necessidade de tratar degenerações de cena, a Figura 7.10 mostra uma comparação entre a abordagem QDEGSAC, que trata con-

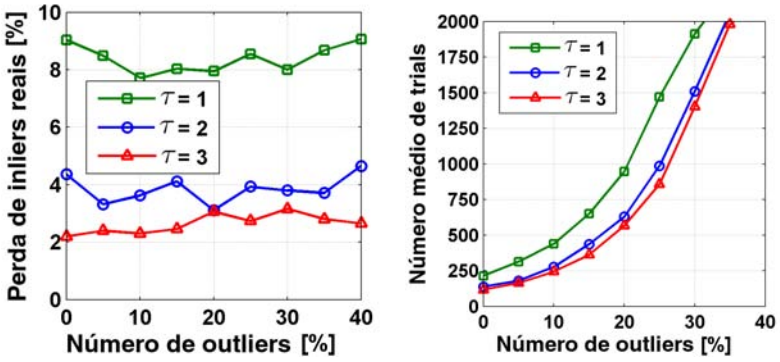


Figura 7.7: Quantidade de inliers reais perdidos e a média de trials requerida pelo algoritmo QDEGSAC como função da quantidade de outliers reais presentes nos dados, onde $|\mathbf{t}| = 5$ u.d.f., $\alpha = 0$ graus (configuração genérica de cena) e $\sigma = 0.1$ pixels.

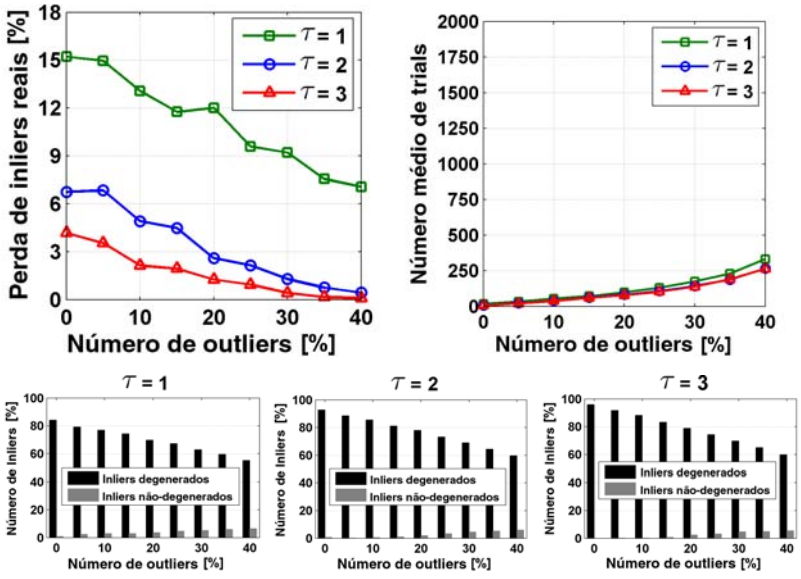


Figura 7.8: Desempenho do algoritmo QDEGSAC como função da quantidade de outliers reais para cenas degeneradas, onde $|\mathbf{t}| = 0$ u.d.f., $\alpha = 5$ graus e $\sigma = 0.1$ pixels. Acima, quantidade de inliers reais perdidos e a média de trials requerida pelo algoritmo QDEGSAC. Abaixo, proporção de inliers degenerados para $\tau \in \{1, 2, 3\}$.

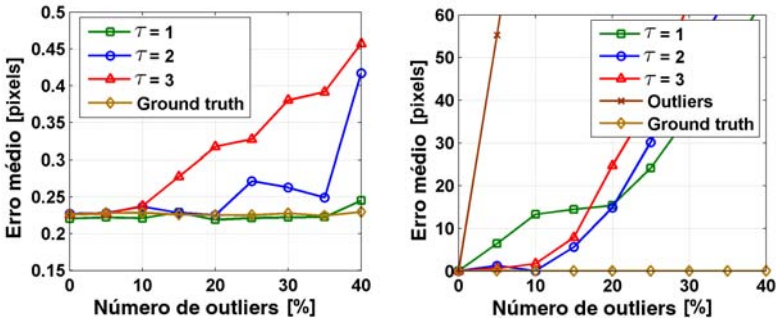


Figura 7.9: Erro médio calculado sobre os conjuntos de inliers $\{in\}^{\tau}_{QDEGSAC}$ entregues pelo algoritmo QDEGSAC ($\tau \in \{1, 2, 3\}$) e o conjunto de inliers reais $\{in\}_{Right}$ (*ground truth*) produzido pela relação T_{Right}, como uma função da quantidade de outliers reais presentes nos dados de entrada. À esquerda, erro calculado para cenas genéricas com $(|\mathbf{t}|, \alpha) = (5, 5)$. À direita, erro calculado sobre cenas degeneradas, incluído o erro computado sobre o conjunto de todos os pontos correspondentes (inliers e outliers reais).

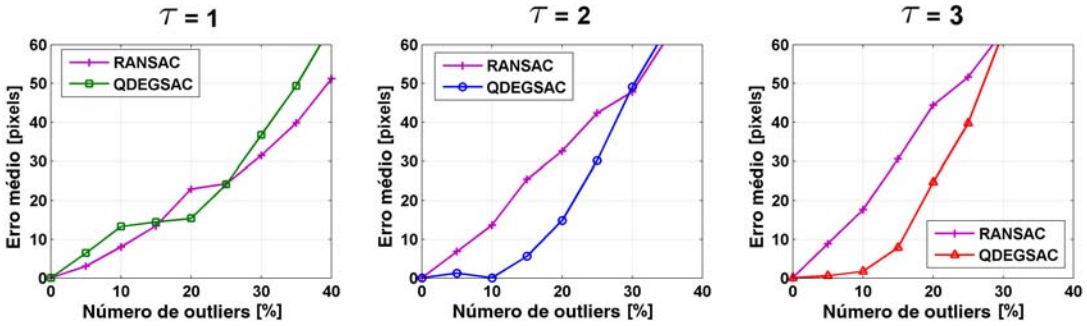


Figura 7.10: Comparação do algoritmo QDEGSAC contra uma simples abordagem RANSAC no contexto de cenas degeneradas, considerando diferentes quantidades de outliers reais, onde $(|\mathbf{t}|, \alpha) = (0, 5)$ (apenas rotação de câmera). O erro simétrico de transferência é calculado sobre os inliers entregues pelo QDEGSAC $\{in\}^{\tau}_{QDEGSAC}$ e RANSAC $\{in\}^{\tau}_{RANSAC}$ para $\tau \in \{1, 2, 3\}$.

figurações degeneradas de cena, e uma simples abordagem RANSAC que estima uma relação de movimento de câmera sem considerar contextos degenerados. Dada a relação correta de movimento T_{Right} (uma homografia $\mathbf{H}$), calcula-se o erro simétrico de transferência sobre o conjunto de inliers $\{in\}^{\tau}_{QDEGSAC}$ entregue pelo QDEGSAC e RANSAC $\{in\}^{\tau}_{RANSAC}$, onde $\tau \in \{1, 2, 3\}$. A Figura 7.10 mostra o erro para cenas degeneradas, com $(|\mathbf{t}|, \alpha) = (0, 5)$, como uma função da quantidade de outliers reais presentes nos dados de entrada.

Considerando os efeitos causados pela presença de outliers reais nos dados de entrada, algumas observações podem ser feitas com base nos experimentos realizados:

1. O esforço computacional cresce de acordo com a quantidade de outliers reais presente nos dados, uma vez que o algoritmo RANSAC necessitará de um número maior de trials para alcançar um consenso dentro do conjunto de pontos correspondentes provido.

2. Em termos de perdas de inliers reais, melhores resultados são alcançados empregando-se o QDEGSAC com $\tau = 3$. Contudo, nota-se claramente a existência de um equilíbrio na escolha do limiar de erro τ: ele deveria ser suficientemente alto para evitar que inliers reais sejam perdidos devido a efeitos de ruído e, ao mesmo tempo, deveria ser tão baixo quanto possível para evitar classificar outliers reais como inliers. Neste sentido, para os experimentos realizados neste trabalho, $\tau = 3$ proveu os melhores resultados. Por exemplo, em configurações degeneradas de cena, $\tau = 1$ apontou incorretamente a presença de inliers não-degenerados desde baixos níveis de contaminação por outliers reais, enquanto $\tau = 3$ mostrou-se menos sensível a este problema, conforme mostrado na Figura 7.8.

3. Qualitativamente, o desempenho do algoritmo QDEGSAC deteriora-se a partir da presença de 10% de outliers reais nos dados de entrada. Neste sentido, observa-se que o erro computado para cenas degeneradas é significativamente maior que o erro computado em contextos não-degenerados (Figura 7.9). Uma vez que outliers reais equivocadamente dão suporte a relação de movimento estimada, eles são entendidos pelo algoritmo QDEGSAC como inliers não-degenerados. Deste modo, o erro de transferência, estimado em termos de uma homografia $\mathbf{H}$, será alto, dado que $\mathbf{H}$ não pode explicar os falsos inliers não-degenerados.

4. A abordagem QDEGSAC é menos suscetível a presença de outliers reais que uma simples abordagem RANSAC que ignora cenas degeneradas, conforme mostrado na Figura 7.10 para $\tau \in \{2, 3\}$. Nota-se, contudo, que ambos apresentam um desempenho similar para cenas genéricas (não-degeneradas), uma vez que, ao detectar o contexto não-degenerado, o algoritmo QDEGSAC apresentará resultados provenientes de um simples processo RANSAC, conforme descrito na seção 4.2.1.

7.2 Experimentos com Imagens Reais

Nesta seção apresenta-se experimentos com imagens de histeroscopias reais. Avalia-se o desempenho do algoritmo QDEGSAC em condições reais de ruído e presença de outliers reais bem como o desempenho da abordagem proposta no sentido de apontar informações relevantes em vídeos de histeroscopias diagnósticas.

7.2.1 Avaliação do Algoritmo QDEGSAC em Seqüências Reais

O desempenho do algoritmo QDEGSAC é avaliado sobre pares de quadros extraídos de três seqüências: *Checkerboard* (Fig 7.11), *Tubal Orifice* (Fig 7.13) e *Fundus* (Fig 7.15). Estas seqüências são capturadas com uma câmera manualmente guiada e, por esta razão, informações precisas a respeito da geometria da cena (*ground truth*) não são conhecidas. Ainda sim, é possível estabelecer algumas características para estas seqüências:

Checkerboard é uma seqüência de imagens utilizada no processo de calibração da câmera histeroscópica. O propósito é avaliar os potenciais efeitos causados pelo processo de remoção de distorção da lente bem como avaliar o desempenho do algoritmo QDEGSAC em condições reais e no contexto de uma configuração de cena sabidamente degenerada (cena planar).

Tubal Orifice é uma seqüência histeroscópica típica da fase de *exame panorâmico* (seção 2.2.3), onde configuram-se cenas genéricas com variações claras de profundidade.

Fundus é uma seqüência cujo layout da cena é aproximadamente planar (degenerado). Esta seqüência foi adquirida no contexto da fase de exame do *fundo do útero*, cujas características são discutidas na seção 2.2.3.

Devido a dificuldades em estabelecer a geometria correta para cenas de histeroscopias, testa-se o algoritmo QDEGSAC em termos da quantidade de pontos consistentes rastreados com sucesso e o erro epipolar correspondente. Embora esta seja uma maneira indireta de medir a qualidade dos resultados, é possível obter uma boa indicação da qualidade da relação estimada quando outliers reais estão presentes nos dados de entrada.

O algoritmo QDEGSAC foi executado 100 vezes sobre os potenciais pontos correspondentes entregues pelo rastreador KLT. Mede-se a quantidade média de inliers computada para relações que empregam 6, 7 e 8 restrições, considerando contextos de erro máximo permitido de $\tau \in \{1, 2, 3\}$ (pixels). Para relações (degeneradas) que empregam 6 e 7 restrições mede-se também a quantidade de inliers adicionais encontrados pelo algoritmo QDEGSAC (seção 4.2.2). Lembra-se que configurações degeneradas de cena caracterizam-se por uma grande quantidade de inliers entregues por uma relação que emprega apenas 6 restrições (dimensão 6). Além disso, histogramas que mostram o erro epipolar residual são computados sobre as 100 execuções. As Figuras 7.12, 7.14 e 7.16 mostram um avaliação qualitativa a respeito dos inliers computados para as seqüências *Checkerboard*, *Tubal Orifice* e *Fundus*, respectivamente. As Tabelas 7.2, 7.3 e 7.4 sumarizam os resultados para as seqüências *Checkerboard*, *Tubal Orifice* e *Fundus* respectivamente.

Com base nos testes realizados sobre as seqüências de imagens mencionadas acima, destaca-se os seguintes pontos sobre o desempenho do algoritmo QDEGSAC:

1. O erro cresce com a distância temporal entre os quadros.

2. Embora o QDEGSAC com $\tau = 3$ entregue uma quantidade não tão baixa de inliers na dimensão 6 para a seqüência *Tubal Orifice*, esta configuração apresenta o melhor desempenho em termos de quantidades de inliers e trials.

3. O QDEGSAC com $\tau = 2$ apresenta os resultados mais coerentes em termos da detecção de configurações de cena: para cenas planares (degeneradas) ou quase planares, como nas seqüências *Checkerboard* e *Fundus*, computou-se uma quantidade

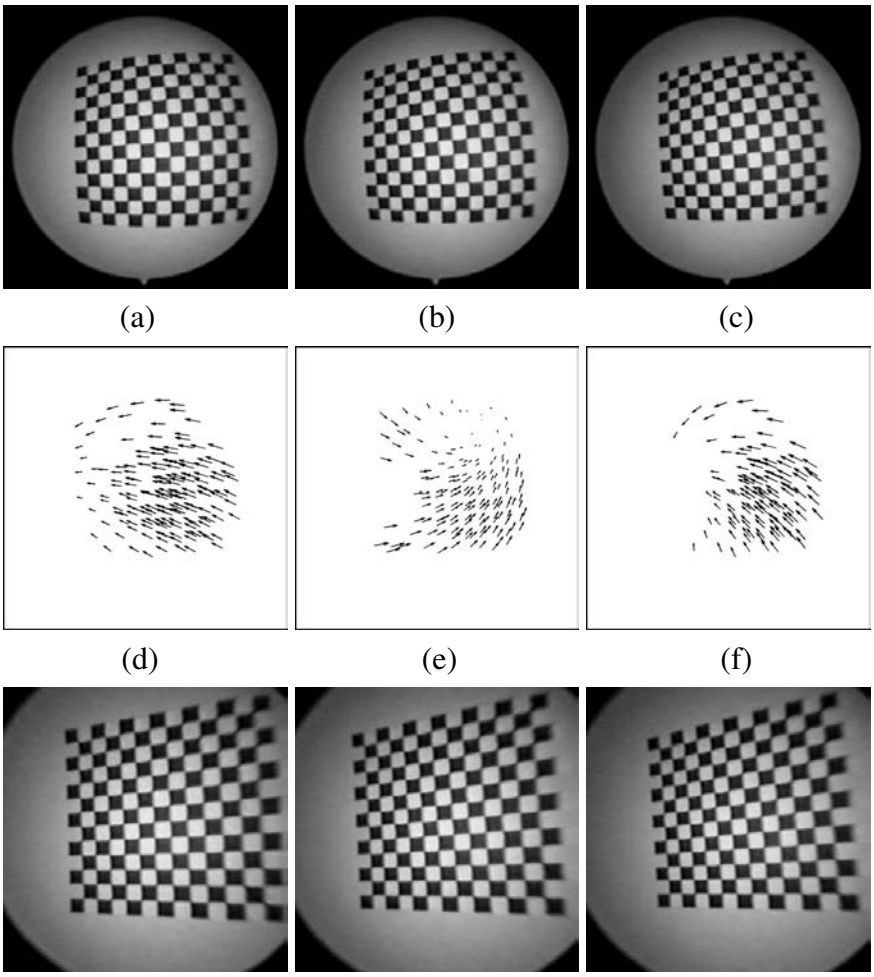


Figura 7.11: **Acima**: quadros 1,7 e 13 da seqüência *Checkerboard*. **Meio**: movimento 2D dos pontos correspondentes computados pelo rastreador KLT, do quadro corrente para o próximo ("→"). **Abaixo**: Quadros com a distorção de lente removida. (a) Quadro 1. (b) Quadro 7. (c) Quadro 13. (d) Movimento 2D do quadro 1 para o quadro 7. (e) Movimento 2D do quadro 7 para o quadro 13. (f) Movimento 2D do quadro 1 para o quadro 13.

Tabela 7.2: Resultados para a seqüência *Checkerboard* (média sobre 100 execuções do algoritmo QDEGSAC)

Quadros	**1-7**	**7-13**	**1-13**
Total de pontos computados (KLT)	155	163	111
Total de inliers para $\tau = 1$	151.98	160.00	104.16
Total de inliers para $\tau = 2$	154.17	162.2	110.05
Total de inliers para $\tau = 3$	154.50	162.52	110.42
Erro residual (pixels) para $\tau = 1$ (σ)	0.21	0.21	0.23
Erro residual (pixels) para $\tau = 2$ (σ)	0.33	0.35	0.37
Erro residual (pixels) para $\tau = 3$ (σ)	0.43	0.42	0.51
Quantidade de trials requeridas para $\tau = 1$	26.06	23.38	62.20
Quantidade de trials requeridas para $\tau = 2$	9.59	8.57	13.93
Quantidade de trials requeridas para $\tau = 3$	6.44	6.64	8.80

Tabela 7.3: Resultados para a seqüência *Tubal Orifice* (média sobre 100 execuções do algoritmo QDEGSAC)

Quadros	**1-10**	**10-20**	**1-20**
Total de pontos computados (KLT)	547	469	426
Total de inliers para $\tau = 1$	359	333.10	266.31
Total de inliers para $\tau = 2$	480.02	397.08	355.44
Total de inliers para $\tau = 3$	507.01	428.67	389.23
Erro residual (pixels) para $\tau = 1$ (σ)	0.26	0.25	0.28
Erro residual (pixels) para $\tau = 2$ (σ)	0.42	0.44	0.45
Erro residual (pixels) para $\tau = 3$ (σ)	0.63	0.64	0.64
Quantidade de trials requeridas para $\tau = 1$	786.77	694.27	2129.10
Quantidade de trials requeridas para $\tau = 2$	192.63	206.42	295.80
Quantidade de trials requeridas para $\tau = 3$	68.28	98.39	231.45

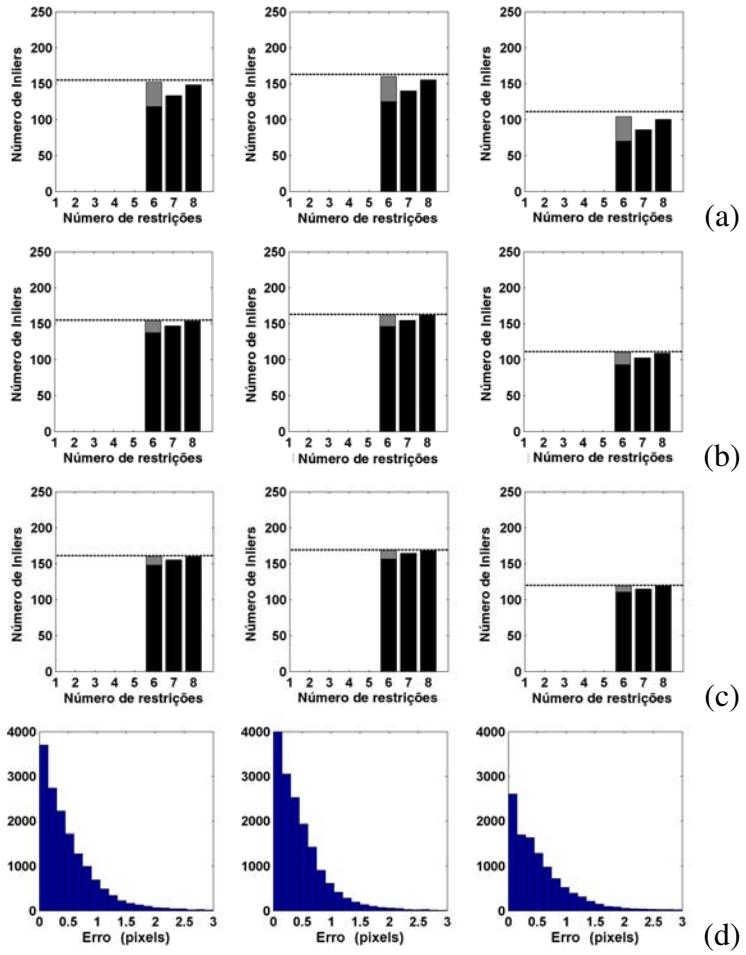


Figura 7.12: Da esquerda para direita, resultados do quadro 1 para 7, 7 para 13 e 1 para 13 da seqüência *checkerboard*. **(a-c)** Quantidade de inliers (eixo y) quando emprega-se 6, 7 e 8 restrições (eixo x). (a) $\tau = 1$. (b) $\tau = 2$. (c) $\tau = 3$. Inliers adicionais computados pelo algoritmo QDEGSAC aparecem empilhados (cinza). Linha horizontal pontilhada representa a quantidade total de pontos correspondentes entregue pelo rastreador KLT. **(d)** Histogramas do erro epipolar residual para os inliers computados pelo algoritmo QDEGSAC com $\tau = 3$.

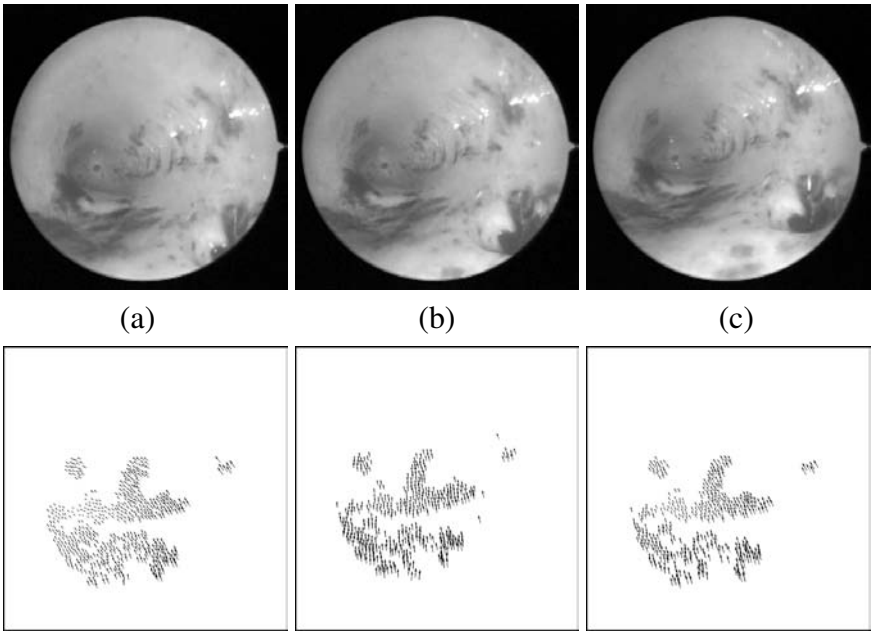


Figura 7.13: **Acima**: quadros 1, 10 e 20 da seqüência *Tubal Orifice*. **Abaixo**: movimento 2D dos pontos correspondentes computados pelo rastreador KLT, do quadro corrente para o próximo ("→"). (a) Quadro 1. (b) Quadro 10. (c) Quadro 20. (d) Movimento 2D do quadro 1 para o quadro 10. (e) Movimento 2D do quadro 10 para o quadro 20. (f) Movimento 2D do quadro 1 para o quadro 20.

Tabela 7.4: Resultados para a seqüência *Fundus* (média sobre 100 execuções do algoritmo QDEGSAC)

Quadros	**1-5**	**5-10**	**1-10**
Total de pontos computados (KLT)	380	317	200
Total de inliers para $\tau = 1$	274.28	234.36	143.72
Total de inliers para $\tau = 2$	370.66	308.10	182.85
Total de inliers para $\tau = 3$	376.52	314.65	194.55
Erro residual (pixels) para $\tau = 1$ (σ)	0.25	0.26	0.27
Erro residual (pixels) para $\tau = 2$ (σ)	0.40	0.39	0.47
Erro residual (pixels) para $\tau = 3$ (σ)	0.50	0.46	0.58
Quantidade de trials requeridas para $\tau = 1$	240.90	314.87	675.98
Quantidade de trials requeridas para $\tau = 2$	26.35	37.42	63.72
Quantidade de trials requeridas para $\tau = 3$	14.00	18.37	25.22

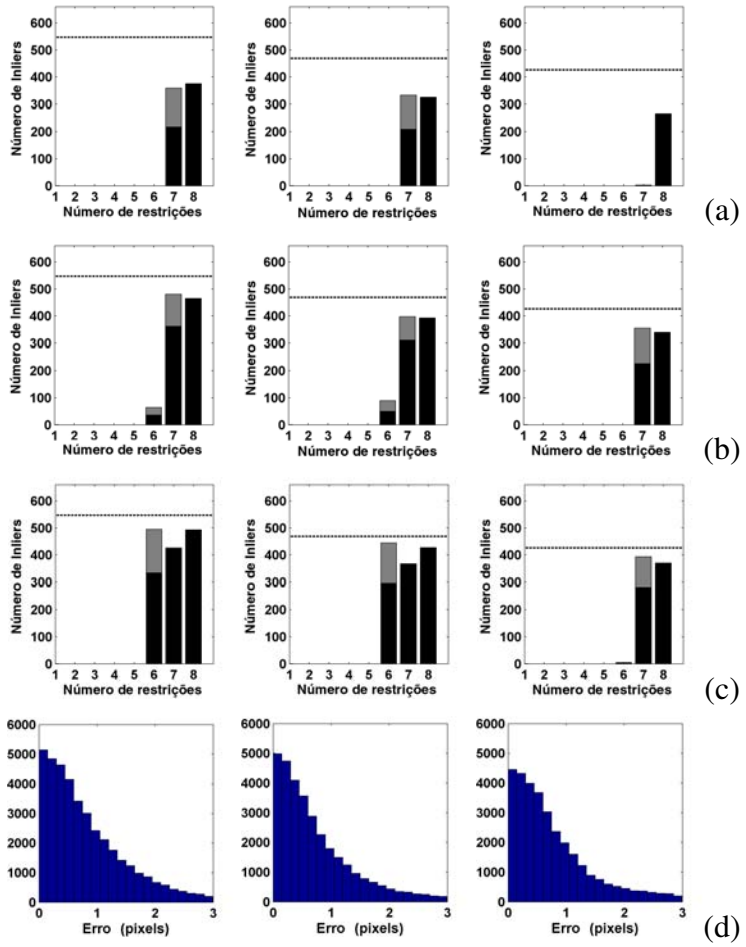


Figura 7.14: Da esquerda para direita, resultados do quadro 1 para 10, 10 para 20 e 1 para 20 da seqüência *Tubal Orifice*. **(a-c)** Quantidade de inliers (eixo y) quando emprega-se 6, 7 e 8 restrições (eixo x). (a) $\tau = 1$. (b) $\tau = 2$. (c) $\tau = 3$. Inliers adicionais computados pelo algoritmo QDEGSAC aparecem empilhados (cinza). Linha horizontal pontilhada representa a quantidade total de pontos correspondentes entregue pelo rastreador KLT. **(d)** Histogramas do erro epipolar residual para os inliers computados pelo algoritmo QDEGSAC com $\tau = 3$.

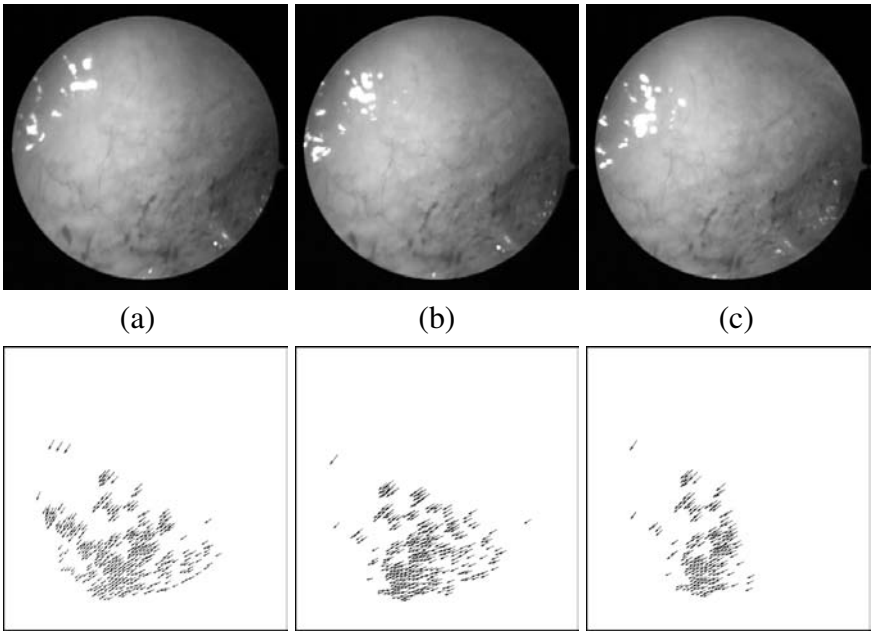


Figura 7.15: **Acima**: quadros 1, 5 e 10 da seqüência *Fundus*. **Abaixo**: movimento 2D dos pontos correspondentes computados pelo rastreador KLT, do quadro corrente para o próximo ("→"). (a) Quadro 1. (b) Quadro 5. (c) Quadro 10. (d) Movimento 2D do quadro 1 para o quadro 5. (e) Movimento 2D do quadro 5 para o quadro 10. (f) Movimento 2D do quadro 1 para o quadro 10.

alta de inliers na dimensão 6, enquanto que para a seqüência *Tubal Orifice* (não-planar) verificou-se uma baixa quantidade de inliers na dimensão 6. Isso verifica-se mesmo para quadros temporalmente próximos (1-10 e 10-20) dentro da seqüência *Tubal Orifice*.

7.2.2 Resultados em Sumarização de Vídeos

Nesta seção avalia-se o desempenho da abordagem proposta na sumarização de vídeos. Os experimentos foram conduzidos em 4 vídeos de histeroscopias diagnósticas pré-interpretados: *hyst1*, *hyst2*, *hyst3* e *hyst4*. Estes vídeos foram selecionados a partir de uma biblioteca que conta com mais de 10.000 exames. A Tabela 7.6 apresenta uma breve descrição sobre os vídeos utilizados nos experimentos, justificando sua escolha em termos de características que constituem potenciais dificuldades para a abordagem proposta.

Usualmente, os vídeos são gravados no formato DVD, armazenando quadros de forma entreleçada. Assim, com o objetivo de minimizar dificuldades no processo de rastreamento de pontos, emprega-se o método *bob deinterlacer* disponível no software VirtualDub (www.virtualdub.org) para obter versões dos vídeos sem entrelaçamento de quadros. O processo de desentrelaçamento divide cada quadro em duas imagens, conseqüentemente tem-se os vídeo originais em uma taxa de amostragem dobrada nos experimentos deste trabalho. Os resultados apresentados a seguir foram computados com base nos valores de parâmetros mostrados na Tabela 7.5.

A Figura 7.17 mostra os quadros medianos dos segmentos de vídeos manualmente selecionados com o auxílio de especialistas (segmentos importantes). Neste ponto deve-se observar a estratégia adotada para que especialistas definissem os limites de cada seg-

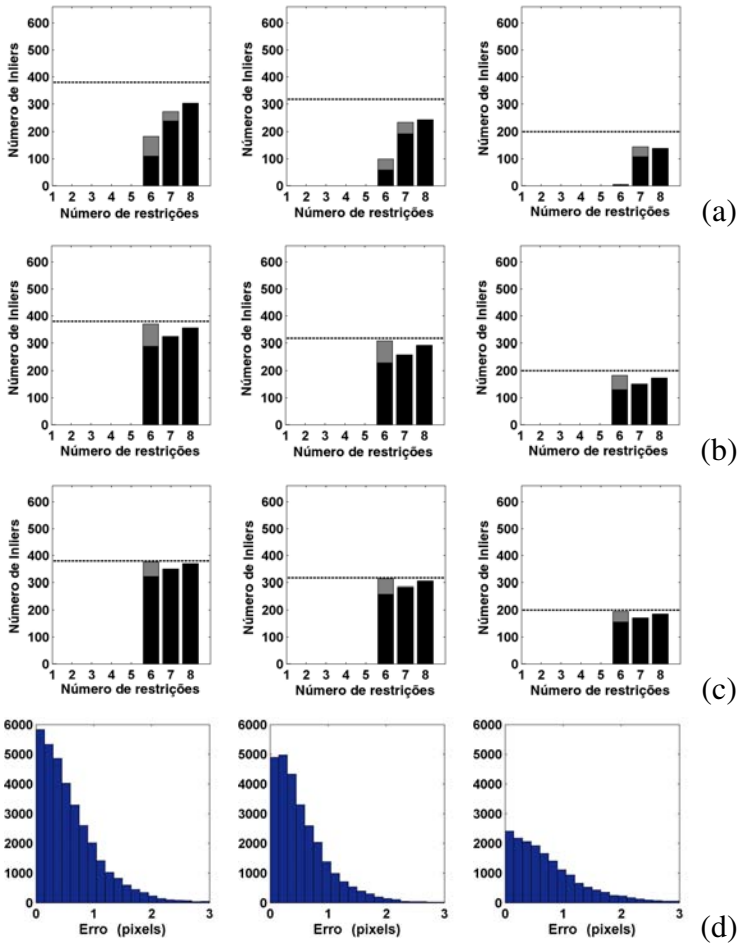


Figura 7.16: Da esquerda para direita, resultados do quadro 1 para 5, 5 para 10 e 1 para 10 da seqüência *Fundus*. **(a-c)** Quantidade de inliers (eixo y) quando emprega-se 6, 7 e 8 restrições (eixo x). (a) $\tau = 1$. (b) $\tau = 2$. (c) $\tau = 3$. Inliers adicionais computados pelo algoritmo QDEGSAC aparecem empilhados (cinza). Linha horizontal pontilhada representa a quantidade total de pontos correspondentes entregue pelo rastreador KLT. **(d)** Histogramas do erro epipolar residual para os inliers computados pelo algoritmo QDEGSAC com $\tau = 3$.

Tabela 7.5: Valores dos parâmetros empregados nos experimentos com os vídeos.

Parâmetro	Unidade	Descrição
$\tau = 3$	pixels	algoritmo QDEGSAC (Equação 4.14)
$\gamma = 60\%$	inliers	algoritmo QDEGSAC (Equação 4.15)
$\Delta = 5$	quadros	amostragem do vídeo (Seção 6.4)
$W = 19$	pixels	rastreador KLT (Seção 4.1.1)
$\zeta = 50$	pontos	quantidade mínima de pontos correspondentes (Seção 6.4)

Tabela 7.6: Descrição/caracterização dos vídeos utilizados nos experimentos.

Vídeos	Descrição	Total de quadros	Total de segmentos de vídeo selecionados por especialistas
hyst1	Histeroscopia tradicional com fases bem definidas e achados não-normais	3013	9
hyst2	Exame de pequena duração com quadros importantes degradados por inúmeras regiões com reflexo especular	2080	5
hyst3	Várias seqüências de quadros interrompidas repentinamente por efeitos indesejáveis ou movimentos rápidos de câmera	5500	5
hyst4	Imagens com pouco textura, representando dificuldades em termos de rastreamento e manutenção da persistência dos pontos	7500	12

mento de vídeo julgado como relevante: é esperado que pelo menos um quadro de cada segmento de vídeo selecionado esteja presente no sumário do vídeo. Deste modo, se há uma seqüência de quadros cobrindo diferentes propósitos no exame, os especialistas são orientados a fracionar a seqüência em tantas partes quantos são os propósitos identificados por eles na seqüência de quadros.

A Figura 7.18 mostra as árvores de segmentos de vídeo computadas para os quatro vídeos, revelando a localização de segmentos potencialmente importantes dentro de cada vídeo. Lembra-se que quadros que compartilham mais pontos consistentes são agrupados primeiro, sendo que o processo de agrupamento chega ao fim quando a quantidade de pontos compartilhados entre quadros vizinhos (ou segmentos de vídeo vizinhos) cai abaixo de um limiar ϱ. Nestes experimentos o valor de ϱ adotado foi de 10 pontos consistentes. Deste modo, árvores mais altas resultam de seqüências de quadros caracterizadas por movimentos lentos de câmera, sendo que, quanto maior é a altura da árvore, maior é a quantidade de quadros no segmento de vídeo associado.

A Figura 7.19 mostra em detalhe 9 árvores de segmentos de vídeo e os quadros-chave associados para uma seqüência de 700 quadros proveniente do vídeo *hyst1*. Como pós-processamento, descartou-se árvores cujo segmento de vídeo associado não ultrapassou a duração de $1/3$ de segundo. Observa-se que a seqüência inicia com uma inspeção panorâmica, então o operador prossegue através de quadros não-importantes e alcança uma região do útero na qual despende a maior parte do tempo observando/diagnosticando potenciais problemas.

A Tabela 7.7 mostra um sumário dos resultados, além de uma comparação com o método proposto em (SCHARCANSKI; GAVIãO, 2006), onde a métrica de *Precision* (comumente utilizada para avaliar abordagens em recuperação de informações) é definida

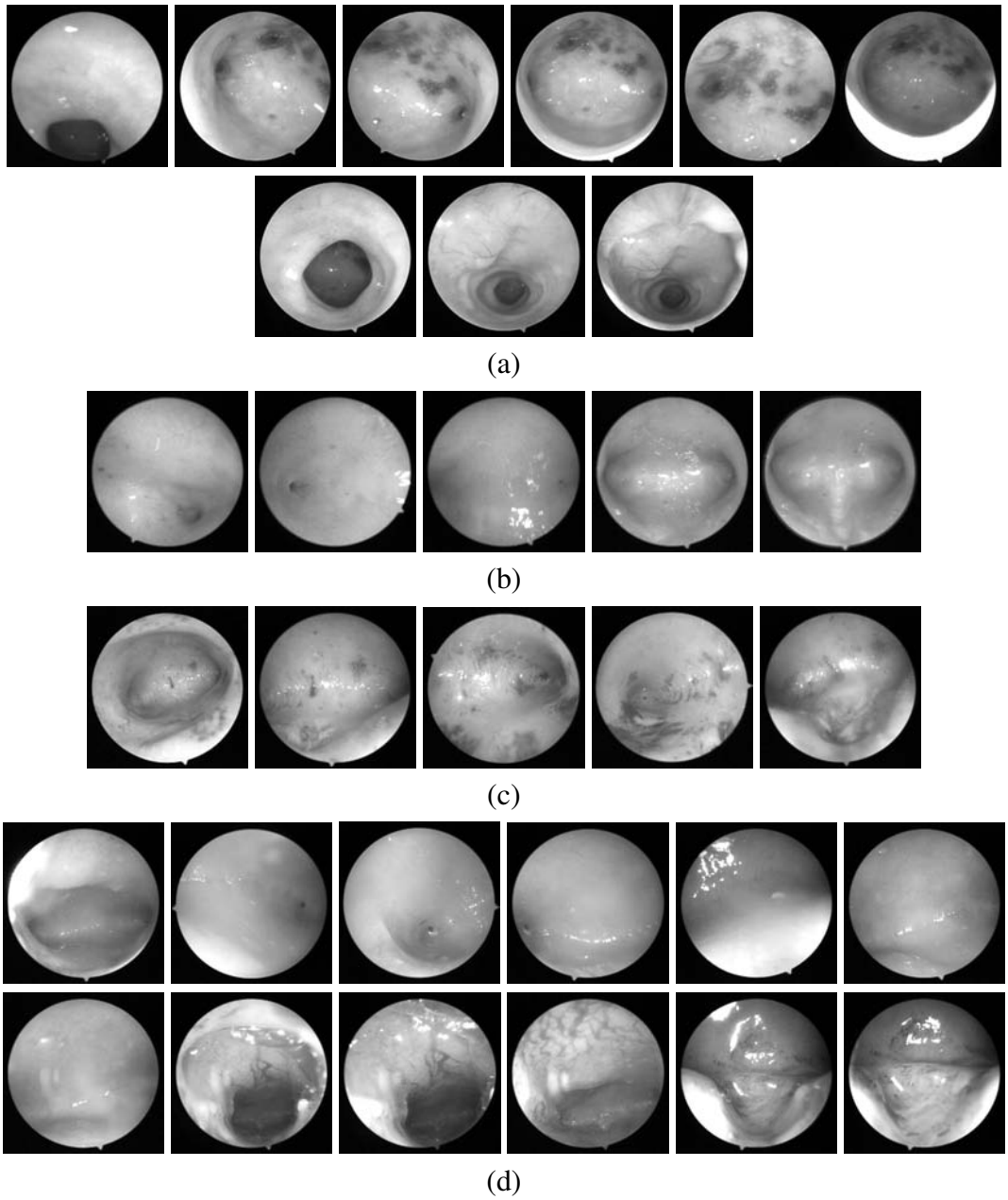


Figura 7.17: Quadros cuja ordem temporal é a mediana dentro de segmentos (importantes) selecionados com auxílio de especialistas. (a) *hyst1*. (b) *hyst2*. (c) *hyst3*. (d) *hyst4*.

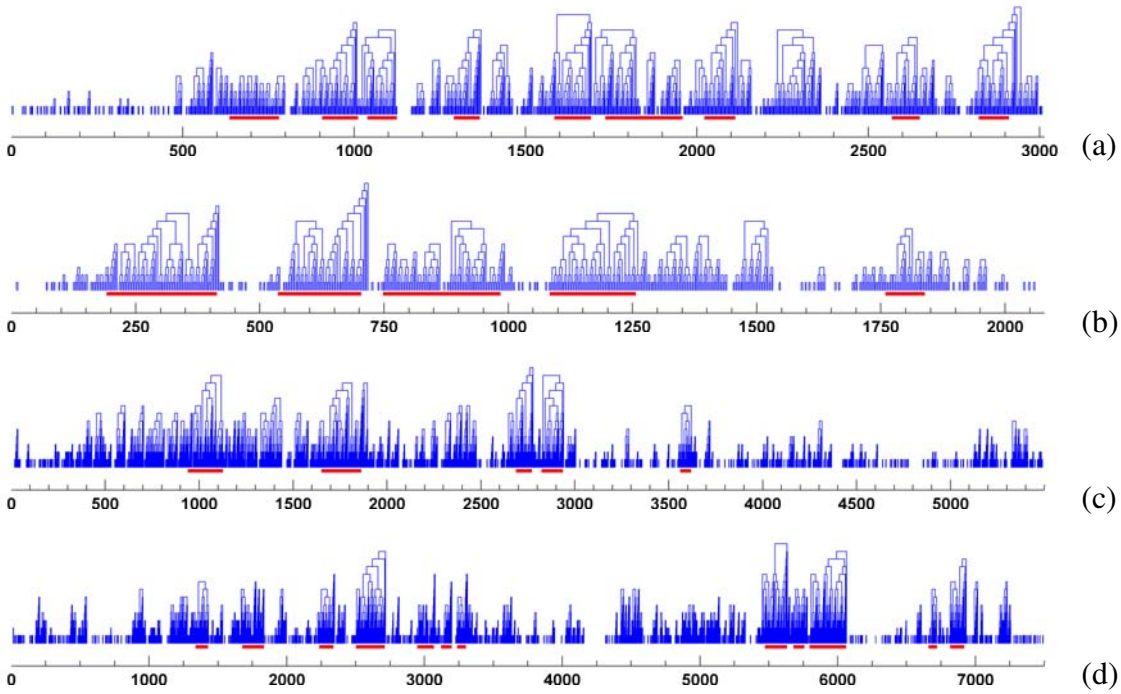


Figura 7.18: Árvores de segmentos de vídeo computadas através dos vídeos. Eixo x representa os quadros na seqüência temporal do vídeo. Linhas horizontais em vermelho representam os segmentos de vídeo (importantes) selecionados com o auxílio de especialistas. (a) *hyst1*. (b) *hyst2*. (c) *hyst3*. (d) *hyst4*. Observa-se que segmentos importantes geralmente aparecem associados a árvores de altura destacada.

como:

$$Precision = \frac{\#quadros\ relevantes}{\#quadros\ relevantes + \#quadros\ irrelevantes} \tag{7.2}$$

Para os vídeos testados o método não apontou falso negativos, ou seja, retornou pelo menos um quadro para cada segmento de vídeo delimitado manualmente com o auxílio de especialistas.

7.3 Discussão

De modo geral, os experimentos indicam que a abordagem proposta pode produzir

Tabela 7.7: Sumário de resultados e comparação do método proposto (M1) contra o método proposto em (SCHARCANSKI; GAVIãO, 2006) (M2)

Vídeos	*hyst1*		*hyst2*		*hyst3*		*hyst4*	
Métodos	M1	M2	M1	M2	M1	M2	M1	M2
Redução de dados (%)	97.5	96.1	97.6	94.6	97.2	95.8	97.5	93.9
Total de quadros-chave	76	117	51	111	154	226	191	456
Precision	0,98	0,97	1	0.91	0,92	0.84	0,90	0.82

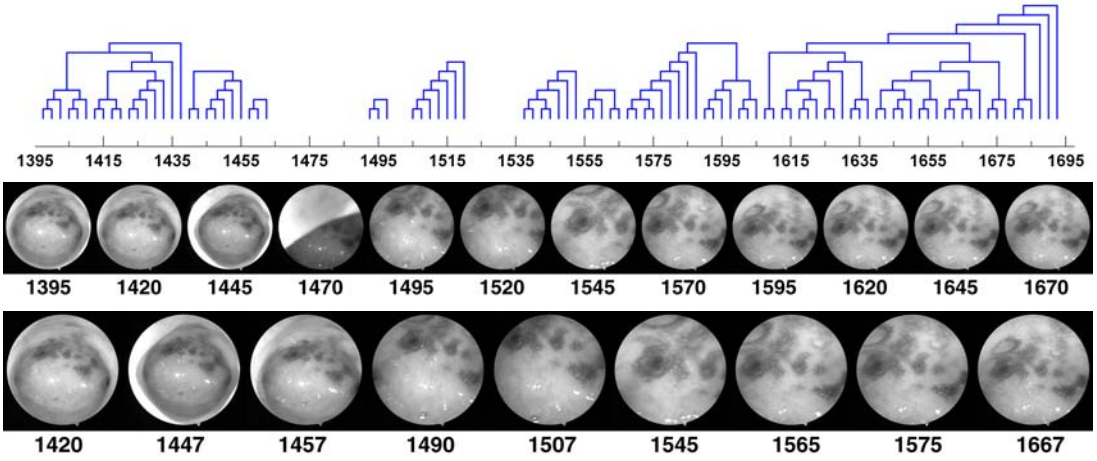


Figura 7.19: Árvores de segmentos de vídeo e quadros-chave associados, ambos computados sobre uma seqüência extraída do vídeo *hyst1*. **(Acima)** 9 árvores de segmentos de vídeo como um função da seqüência temporal dos quadros. **(Meio)** Seqüência de vídeo amostrada em intervalos regulares de 25 quadros. **(Abaixo)** Os 9 quadros-chave computados para cada uma das árvores de segmento. Segmentos cuja duração é menor que $1/3$ de segundo foram descartados como irrelevantes.

sumários de vídeos muito compactos, incluindo quadros de cada segmento selecionado como importante na ótica de especialistas. Contudo, alguns pontos podem ser aprofundados e possivelmente refinados:

- Idealmente, a abordagem proposta deveria considerar somente pontos que movem-se para fora do campo de visão, uma vez que a idéia central é medir mudanças visuais resultantes do movimento de câmera. Neste sentido, deve-se notar que oclusões de pontos não são tratadas, conseqüentemente pontos que são perdidos devido a, por exemplo, reflexos especulares, afetarão negativamente a métrica de sobreposição de conteúdo proposta (uma vez que a perda destes pontos não foi necessariamente causada pelo movimento de câmera). O implementação do rastreador de pontos KLT (BIRCHFIELD, 2006) empregada nos experimentos oferece uma interessante classificação de pontos que foram perdidos. A proximidade dos limites da imagem é levada em conta como razão para a perda de pontos. Assim, resultados melhores poderiam ser alcançados considerando apenas pontos que escapam do campo de visão no processo de rastreamento.

- Em alguns vídeos verifica-se a presença de segmentos de vídeos irrelevantes do ponto de vista clínico, porém com baixa atividade da câmera por alguns segundos, fato este que determinaria a existência de falsos positivos no sumário de vídeo gerado. De acordo com ginecologistas experientes, isto ocorre devido a inexperiência de alguns especialistas no procedimento de histeroscopia, ou devido a existência de obstáculos indesejáveis que eventualmente são difíceis de contornar, tal como bolhas/muco (seção 2.2.1) que aderem à lente histeroscópica. A abordagem proposta nesta tese não trata este tipo de situação adequadamente, pois sempre que houver uma boa quantidade de pontos consistentes com um modelo rígido de movimento, o método irá sugerir o segmento de vídeo associado a estes pontos como sendo relevante. Este tipo de situação degrada a precisão (Equação 7.2) dos sumários de

vídeo entregues pela abordagem proposta.

- Do ponto de vista de esforço computacional, observou-se que o processo de rastreamento de pontos pelo algoritmo KLT consome a maior parte do tempo. São necessárias horas para processar um vídeo de histeroscopia por inteiro, considerando um processador Intel Core 2 Duo com 2 núcleos de 1.66GHz e 2GB de RAM. A abordagem proposta não é pretendida para produzir resultados em tempo real, contudo o desempenho do processo de detecção e rastreamento de pontos poderia ser dramaticamente melhorado empregando-se o rastreador KLT projetado para executar em hardware gráfico, conforme proposto em (SINHA et al., 2006).

- Um importante aspecto que está sob investigação é a capacidade do algoritmo KLT em rastrear pontos através de segmentos importantes de vídeos de histeroscopias diagnósticas, dadas conhecidas dificuldades envolvidas na tarefa de estabelecer correspondências entre quadros, sobretudo entre aqueles que são tomados de pontos de vista separados por distâncias maiores. Considerando taxas de amostragens convencionais, os experimentos indicaram que um especialista não move a câmera suficientemente rápido para fazer com que o algoritmo KLT perca pontos devido a movimentos rápidos de câmera, pelo menos em segmentos de vídeos que são considerados relevantes.

8 CONCLUSÕES

Nesta tese apresentou-se uma abordagem para extrair eventos importantes em vídeos de histeroscopias diagnósticas. Pontos são detectados e rastreados através dos quadros, de maneira que sejam consistentes com um modelo rígido de movimento de câmera. Dessa forma, um processo de agrupamento de quadros organiza o vídeo de acordo com a quantidade de pontos/informações que há em comum entre quadros vizinhos. Além disso, uma representação hierárquica, orientada de acordo com o conteúdo destes vídeos, é proposta com o objetivo de facilitar a tarefa de (*browsing*) navegar/selecionar quadros importantes, permitindo que especialistas possam navegar através do conteúdo destes vídeos sendo guiados por um sumário visual que é gerado automaticamente.

Fazem parte deste trabalho também as abordagens apresentadas na seção 5.2. Embora estas abordagens tenham apresentado resultados satisfatórios, discutiu-se na seção 5.2.3 uma importante limitação destas técnicas que motivou a proposição principal desta tese. Sumarizar o conteúdo visual de um vídeo em termos de seus quadros, requer, de alguma forma, a comparação de quadros com sua vizinhança (quadros vizinhos), sendo que uma parte importante do problema está em definir os limites de tal vizinhança para cada quadro considerado no vídeo. Neste sentido as técnicas discutidas na seção 5.2 apresentam suas limitações, pois utilizam um valor fixo de tamanho de vizinhança (quantidade de quadros) para avaliar a similaridade de um dado quadro com relação ao conteúdo que está temporalmente próximo. Sendo assim, a essência do método proposto nesta tese está em fazer uma análise de vizinhança adaptada a variações temporais do conteúdo do vídeo.

Apesar do custo computacional elevado e das dificuldades em rastrear pontos geometricamente, os experimentos indicam sumários surpreedentemente compactos e que são gerados sem descartar informações potencialmente importantes na forma de *falsos negativos*. Melhores resultados são alcançados, não só em termos da precisão do sumário gerado (menor taxa de *falsos positivos*) mas também em termos de uma quantidade significativamente menor de quadros-chave apontados pelo método. Trabalhos futuros estão no sentido de gerar representações de quadros-chave mais compactas, na forma de mosaicos para contextos de cenas degeneradas.

REFERÊNCIAS

BEARDSLEY, P. A.; ZISSERMAN, A.; MURRAY, D. W. Sequential updating of projective and affine structure from motion. **Int. Journal of Computer Vision**, [S.l.], v.23, n.3, p.235–259, 1997.

BIRCHFIELD, S. **KLT**: an implementation of the kanade-lucas-tomasi feature tracker. Disponível em: <http://vision.stanford.edu/birch/klt/>. Acesso em: Set. 2006.

BOUGUET, J.-Y. **Pyramidal Implementation of the Lucas Kanade Feature Tracker**. [S.l.]: OpenCV Documentation, Intel Corporation, 2000.

BOUGUET, J.-Y. **Camera Calibration Toolbox for Matlab**. Disponível em: <http://www.vision.caltech.edu/bouguetj/calib_doc/>. Acesso em: Nov. 2006.

BOUTHEMY, P.; GELGON, M.; GANANSIA, F. A Unified Approach to Shot Change Detection and Camera Motion Characterization. **IEEE Trans. on Circuits Systems for Video Technology**, [S.l.], v.9, n.7, p.1030–1044, October 1999.

BOWMAN, A. W.; AZZALINI, A. **Applied Smoothing Techniques for Data Analysis**. [S.l.]: Oxford University Press, 1997.

CERNEKOVA, Z.; PITAS, I.; NIKOU, C. Information Theory-Based Shot Cut/Fade Detection and Video Summarization. **IEEE Trans. Circuits Syst. Video Technol.**, [S.l.], v.16, n.1, p.82–91, Jan 2006.

CHANG, C.-Y.; MACIEJEWSKI, A. A.; BALAKRISHNAN, V. Eigendecomposition-based analysis of video images. In: SPIE 11TH INT. SYMPOSIUM ON ELECTRONIC IMAGING: STORAGE AND RETRIEVAL FOR IMAGE AND VIDEO DATABASES VII, 1999, San Jose, CA. **Proceedings...** [S.l.: s.n.], 1999. v.3656, p.186–195.

CHANG, H.; SULL, S.; LEE, S. Efficient video indexing scheme for content-based retrieval. **IEEE Trans. Circuits Syst. Video Technol.**, [S.l.], v.9, n.8, p.1269–1279, Dec. 1999.

CHANG, S. The holy grail of content-based media analysis. **IEEE Multimedia**, [S.l.], v.9, n.2, p.6–10, April-June 2002.

CHUM, O.; WERNER, T.; MATAS, J. Two-view geometry estimation unaffected by a dominant plane. In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 2005. **Proceedings...** [S.l.: s.n.], 2005. p.772–779.

CUNHA-FILHO, J. S.; SCHARCANSKI, J.; GAVIAO, W.; PASSOS, P. E. Digital hysteroscopy: a new diagnostic method for the mid-secretory endometrium. In: ANNUAL MEETING OF THE AMERICAN SOCIETY FOR REPRODUCTIVE MEDICINE, 60., 2004, Philadelphia, USA. **Anais...** ASRM, 2004.

DEL BIMBO, A. **Visual Information Retrieval**. San Francisco, USA: Morgan Kaufmann, 1999.

DEMENTHON, D.; KOBLA, V.; DOERMANN, D. Video Summarization by curve simplification. In: ACM INT. MULTIMEDIA CONF., 6., 1998, Bristol, U.K. **Proceedings...** [S.l.: s.n.], 1998.

DIMITROVA, N.; ZHANG, H.; SHAHRARAY, B.; SEZAN, I.; HUANG, T.; ZAKHOR, A. Applications of video content analysis and retrieval. **IEEE Multimedia**, [S.l.], v.9, n.3, p.42–55, July-Sept. 2002.

DUAN, L.-Y.; JIN, J. S.; TIAN, Q.; XU, C.-S. Nonparametric Motion Characterization for Robust Classification of Camera Motion Patterns. **IEEE Trans. on Multimedia**, [S.l.], v.8, n.2, p.323–340, April 2006.

EBADOLLAHI, S.; CHANG, S.; WU, H. Echocardiogram video summarization. In: SPIE MEDICAL IMAGING, 2001, Bellingham, Wash. **Proceedings...** [S.l.: s.n.], 2001.

FISCHLER, M.; BOLLES, R. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. **Communications of the ACM**, [S.l.], v.24, p.381–385, 1981.

FITZGIBBON, A. W.; ZISSERMAN, A. Automatic camera recovery for closed or open image sequences. In: EUROPEAN CONF. ON COMPUTER VISION, 1998. **Proceedings...** [S.l.: s.n.], 1998. p.311–326.

FRAHM, J.-M.; POLLEFEYS, M. RANSAC for (Quasi-)Degenerate data (QDEGSAC). In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 2006, New York, USA. **Proceedings...** [S.l.: s.n.], 2006. p.453–460.

GAVIãO, W.; SCHARCANSKI, J. Content-Based Diagnostic Hysteroscopy Summaries for Video Browsing. In: SIBGRAPI, 2005, Natal, Brasil. **Proceedings...** IEEE, 2005.

GAVIãO, W.; SCHARCANSKI, J.; PASSOS, E. P.; CUNHA-FILHO, J. S. Evaluating the Mid-Secretory Endometrium Appearance Using Hysteroscopic Digital Video Summarization. **Image and Vision Computing**, [S.l.], v.25, n.1, p.70–77, January 2007.

GOLUB, C.; LOAN, C. **Matrix computations**. 2nd.ed. Baltimore: John Hopkins University Press, 1989.

GONG, Y.; LIU, X. Video summarization using singular value decomposition. In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 2000, Hilton Head, South Carolina. **Proceedings...** [S.l.: s.n.], 2000.

HAMOU, J. **Hysteroscopy and Microcolpohysteroscopy, Text and Atlas**. USA: Appleton and Lange, 1991.

HANJALIC, A. Shot-boundary detection: unraveled and resolved? **IEEE Trans. Circuits Syst. Video Technol.**, [S.l.], v.12, n.2, p.90–105, Feb. 2002.

HANJALIC, A.; ZHANG, H. An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis. **IEEE Trans. on Circuits Systems for Video Technology**, [S.l.], v.9, n.8, p.1280–1289, Dec 1999.

HARRIS, C.; STEPHENS, M. A Combined Edge and Corner Detector. In: ALVEY VISION CONFERENCE, 4., 1988. **Proceedings...** [S.l.: s.n.], 1988. p.147–151.

HARTLEY, R. In Defense of the Eight-Point Algorithm. **IEEE Trans. on Pattern Analysis and Machine Intelligence**, [S.l.], v.19, n.6, p.580–593, October 1997.

HARTLEY, R.; ZISSERMAN, A. **Multiple View Geometry in Computer Vision**. [S.l.]: Cambridge University Press, 2000.

HEIKKILä, J.; SILVéN, O. A Four-step Camera Calibration Procedure with Implicit Image Correction. In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 1997. **Proceedings...** [S.l.: s.n.], 1997.

HO, Y.-H.; LIN, C.-W.; CHEN, J.-F.; LIAO, H.-Y. M. Fast Coarse-to-Fine Video Retrieval Using Shot-Level Spatio-Temporal Statistics. **IEEE Trans. on Circuits Systems for Video Technology**, [S.l.], v.16, n.5, p.642–648, May 2006.

IRANI, M.; ANANDAN, P. Video indexing based on mosaic representations. **Proceedings of the IEEE**, [S.l.], v.86, n.5, p.905–921, May 1998.

IRANI, M.; SAWHNEY, H. S.; KUMAR, R.; ANANDAN, P. Interactive Content-Based Video Indexing and Browsing. In: IEEE WORKSHOP ON MULTIMEDIA SIGNAL PROCESSING, 1997. **Proceedings...** [S.l.: s.n.], 1997. p.313–318.

JACKSON, J. E. **A User's Guide to Principal Components**. New York, USA: John Wiley and Sons, 1991.

LEE, S.; HAYES, M. H. Video summarization and retrieval using singular value decomposition. **IEEE Signal Processing Letters**, [S.l.], v.11, n.11, p.862–886, Nov 2004.

LEW, M. S. (Ed.). **Principles of Visual Information Retrieval**. London, UK: Springer-Verlag, 2001.

LI, B.; SEZAN, M. Event detection and sumarization in american football broadcast video. In: IS&T/SPIE CONF. STORAGE AND RETRIEVAL FOR MEDIA DATABASES, 2002, Bellingham, Wash. **Proceedings...** [S.l.: s.n.], 2002. p.202–215.

LI, Y.; ZHANG, T.; TRETTER, D. **An Overview of Video Abstraction Techniques**. [S.l.]: Hewlett-Packard Company, 2001. (HPL-2001-191).

LIU, T.; ZHANG, H.; QI, F. A Novel Video Key-Frame-Extraction Algorithm Based on Perceived Motion Energy Model. **IEEE Trans. on Circuits Systems for Video Technology**, [S.l.], v.13, n.10, p.1006–1013, Aug 2003.

LONGUET-HIGGINS, H. C. A computer algorithm for reconstructing a scene from two projections. **Nature**, [S.l.], v.293, p.133–135, September 1981.

LONGUET-HIGGINS, H. C.; PRAZDNY, K. The Interpretation of a Moving Retinal Image. In: ROYAL SOC. LONDON, 1980. **Proceedings...** [S.l.: s.n.], 1980. p.385–397. (B, v.208).

LUCAS, B. D.; KANADE, T. An Iterative Image Registration Technique with an Application to Stereo Vision. In: INT. JOINT CONF. ARTIFICIAL INTELLIGENCE, 1981. **Proceedings...** [S.l.: s.n.], 1981. p.674–679.

LUONG, Q. T.; FAUGERAS, O. D. The fundamental matrix: theory, algorithms, and stability analysis. **Int. Journal of Computer Vision**, [S.l.], v.17, n.1, p.43–76, 1996.

MA, Y.-F.; HUA, X.-S.; LU, L.; ZHANG, H.-J. A Generic Framework of User Attention Model and Its Application in Video Summarization. **IEEE Trans. on Multimedia**, [S.l.], v.7, n.5, p.907–919, October 2005.

MA, Y.; SOATTO, S.; KOSECKA, J.; SASTRY, S. S. **An Invitation to 3-D Vision**: from images to geometric models. [S.l.]: Springer Verlag, 2003.

MAYBANK, S. **Theory of Reconstruction from Image Motion**. NJ, USA: Springer-Verlag, 1992.

MIKHAIL, E. M.; BETHEL, J. S.; MCGLONE, J. C. **Introduction to Modern Photogrammetry**. [S.l.]: John Wiley and Sons, 2001.

NGO, C.-W.; PONG, T.-C.; ZHANG, H.-J. Recent Advances in Content-Based Video Analysis. **Int. Journal of Image and Graphics**, [S.l.], v.1, n.3, p.445–468, 2001.

NGO, C.-W.; PONG, T.-C.; ZHANG, H.-J. Motion Analysis and Segmentation Through Spatio-Temporal Slices Processing. **IEEE Trans. on Image Processing**, [S.l.], v.12, n.3, p.341–355, March 2003.

NGO, C.-W.; PONG, Y.-F. M.; ZHANG, H.-J. Video Summarization and Scene Detection by Graph Modeling. **IEEE Trans. on Circuits Systems for Video Technology**, [S.l.], v.15, n.2, p.296–305, Feb 2005.

NISTéR, D. Reconstruction from Uncalibrated Sequences with a Hierarchy of Trifocal Tensors. In: EUROPEAN CONF. ON COMPUTER VISION, 2000. **Proceedings...** [S.l.: s.n.], 2000. p.649–663.

PEYRARD, N.; BOUTHEMY, P. Motion-Based Selection of Relevant Video Segments for Video Summarization. **Multimedia Tools and Applications**, [S.l.], n.26, p.259–276, 2005.

PIRIOU, G.; BOUTHEMY, P.; YAO, J.-F. Recognition of Dynamic Video Contents With Global Probabilistic Models of Visual Motion. **IEEE Trans. on Image Processing**, [S.l.], v.15, n.11, p.3418–3431, Nov 2006.

POLLEFEYS, M.; VERBIEST, F.; VAN GOOL, L. Surviving Dominant Planes in Uncalibrated Structure and Motion Recovery. In: EUROPEAN CONF. ON COMPUTER VISION, 2002. **Proceedings...** [S.l.: s.n.], 2002. v.1, p.837–851.

RAI, L.; MERRITT, S. A.; HIGGINS, W. E. Real-time Image-based Guidance Method for Lung-Cancer Assessment. In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 2006. **Proceedings...** [S.l.: s.n.], 2006.

REPKO, J.; POLLEFEYS, M. 3D Models from Extended Uncalibrated Video Sequences: addressing key-frame selection and projective drift. In: INT. CONF. ON 3-D DIGITAL IMAGING AND MODELING, 2005. **Proceedings...** [S.l.: s.n.], 2005.

ROTHGANGER, F.; LAZEBNIK, S.; SCHMID, C.; PONCE, J. Segmenting, Modeling, and Matching Video Clips Containing Multiple Moving Objects. **IEEE Trans. on Pattern Analysis and Machine Intelligence**, [S.l.], v.29, n.3, p.477–491, March 2007.

RUBNER, Y.; PUZICHA, J.; TOMASI, C.; BUHMANN, J. M. Empirical evaluation of dissimilarity measures for color and texture. **Computer Vision and Image Understanding**, [S.l.], v.84, n.1, p.25–43, Oct 2001.

SAHOURIA, E.; ZAKHOR, A. Content Analysis of Video Using Principal Components. **IEEE Trans. on Circuits Systems for Video Technology**, [S.l.], v.9, n.8, p.1290–1298, Dec 1999.

SAWHNEY, H. S. 3D geometry from planar parallax. In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 1994, Seattle, USA. **Proceedings...** [S.l.: s.n.], 1994. p.929–934.

SAWHNEY, H. S.; AYER, S. Compact representations of videos through dominant and multiple motion estimation. **IEEE Trans. on Pattern Analysis and Machine Intelligence**, [S.l.], v.18, n.8, p.814–830, August 1996.

SCHARCANSKI, J.; GAVIãO, W. Hierarchical Summarization of Diagnostic Hysteroscopy Videos. In: IEEE INT. CONF. ON IMAGE PROCESSING, 2006, Atlanta, EUA. **Proceedings...** [S.l.: s.n.], 2006.

SCHARCANSKI, J.; GAVIãO, W.; CUNHA-FILHO, J. S. Dyagnostic Hysteroscopy Summarization and Browsing. In: ANNUAL INT. CONF. OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY, 2005, Shanghai, China. **Proceedings...** [S.l.: s.n.], 2005.

SHAHIDI, R.; BAX, M.; MAURER, C.; JOHNSON, J.; WILKINSON, E.; WANG, B.; WEST J.AND CITARD, M.; MANWARING, K.; KHADEM, R. Implementation, Calibration and Accuracy Testing of an Image-Enhance Endoscopy System. **IEEE Trans. on Medical Imaging**, [S.l.], v.21, n.12, p.1524–1535, Dec 2002.

SHI, J.; TOMASI, C. Good Features to Track. In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 1994. **Proceedings...** [S.l.: s.n.], 1994.

SIM, K.; HARTLEY, R. Recovering Camera Motion Using L infinity Minimization. In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 2006. **Proceedings...** [S.l.: s.n.], 2006. p.1230–1237.

SINHA, S.; FRAHM, J.-M.; POLLEFEYS, M.; GENC, Y. GPU-Based Video Feature Tracking and Matching. In: WORKSHOP ON EDGE COMPUTING USING NEW COMMODITY ARCHITECTURES, 2006. **Proceedings...** [S.l.: s.n.], 2006.

SMEULDERS, A. W. M.; WORRING, M.; SANTINI, S.; GUPTA, A.; JAIN, R. Content-Based Image Retrieval at the End of the Early Years. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v.22, n.12, p.1349–1380, Dec 2000.

STEWéNIUS, H.; ENGELS, C.; NISTéR, D. An Efficient Minimal Solution for Infinitesimal Camera Motion. In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 2007. **Proceedings...** [S.l.: s.n.], 2007. p.1–8.

TAN, Y.-P.; SAUR, D. D.; KULKARNI, S. R.; RAMADGE, P. J. Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation. **IEEE Trans. on Circuits Systems for Video Technology**, [S.l.], v.10, n.1, p.133–146, February 2000.

THORMAEHLEN, T.; BROSZIO, H.; WEISSENFELD, A. Keyframe Selection for Camera Motion and Structure Estimation from Multiple Views. In: EUROPEAN CONF. ON COMPUTER VISION, 2004. **Proceedings...** [S.l.: s.n.], 2004. p.523–535.

TORR, P. H. S. An Assessment of Information Criteria for Motion Model Selection. In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 1997. **Proceedings...** [S.l.: s.n.], 1997.

TORR, P. H. S.; FITZGIBBON, A.; ZISSERMAN, A. The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences. **Int. Journal of Computer Vision**, [S.l.], v.32, n.1, p.27–44, August 1999.

TRON, R.; VIDAL, R. A Benchmark for the Comparison of 3-D Motion Segmentation Algorithms. In: IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 2007, Minneapolis,USA. **Proceedings...** [S.l.: s.n.], 2007.

TSAI, R. Y. A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology using off-the-shelf TV cameras and Lenses. **IEEE Journal of Robotics and Automation**, [S.l.], v.3, n.4, p.323–344, 1987.

VASCONCELOS, N.; LIPPMAN, A. Statistical Models of Video Structure for Content Analysis and Characterization. **IEEE Trans. on Image Processing**, [S.l.], v.9, n.1, p.3–19, 2000.

VIéVILLE, T.; LINGRAND, D. Using Specific Displacements to Analyze Motion without Calibration. **Int. Journal of Computer Vision**, [S.l.], v.31, n.1, p.5–29, 1999.

WAIZENEGGER, W.; FELDMANN, I.; SCHREER, O. Semantic annotation and retrieval of unedited video based on extraction of 3D camera motion. In: INT. WORKSHOP ON CONTENT-BASED MULTIMEDIA INDEXING, 2008, London, UK. **Proceedings...** [S.l.: s.n.], 2008. p.265–271.

WOLF, W. Key frame selection by motion analysis. In: IEEE INT. CONF. ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 1996, Atlanta, USA. **Proceedings...** [S.l.: s.n.], 1996. p.1228–1231.

WU, C.-H.; SUN, Y.-N.; CHANG, C.-C. Three-Dimensional Modeling From Endoscopic Video Using Geometric Constraints Via Feature Positioning. **IEEE Trans. on Biomedical Engineering**, [S.l.], v.54, n.7, p.1199–1211, Jul 2007.

YAMAGUCHI, T.; NAKAMOTO, M.; SATO, Y.; KONISHI, K.; HASHIZUME, M.; SUGANO, N.; YOSHIKAWA, H.; TAMURA, S. Development of a camera model and calibration procedure for oblique-viewing endoscopes. **Computer Aided Surgery**, [S.l.], v.9, n.5, p.203–214, 2004.

YOU, J.; LIU, G.; SUN, L.; LI, H. A Multiple Visual Models Based Perceptive Analysis Framework for Multilevel Video Summarization. **IEEE Trans. on Circuits Systems for Video Technology**, [S.l.], v.17, n.3, p.273–285, March 2007.

ZHANG, Z. Flexible camera calibration by viewing a plane from unknown orientations. In: IEEE INT. CONF. ON COMPUTER VISION, 1999. **Proceedings...** [S.l.: s.n.], 1999. v.1, p.666–673.

ZHU, X.; ELMAGARMID, A.; XUE, X.; WU, L.; CATLIN, A. Insight Video: toward hierarchical video content organization for efficient browsing, summarization and retrieval. **IEEE Trans. on Multimedia**, [S.l.], v.7, n.4, p.648–666, August 2005.