GERALDO AUGUSTO MASSAHUD RODRIGUES DOS SANTOS

ALINHAMENTO TEMPORAL DE SEQÜÊNCIAS DE VÍDEO ADQUIRIDAS POR CÂMERAS PERSPECTIVAS E CATADIÓPTRICAS

Belo Horizonte
04 de agosto de 2006

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

ALINHAMENTO TEMPORAL DE SEQÜÊNCIAS DE VÍDEO ADQUIRIDAS POR CÂMERAS PERSPECTIVAS E CATADIÓPTRICAS

Proposta de dissertação apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

GERALDO AUGUSTO MASSAHUD RODRIGUES DOS SANTOS

Belo Horizonte
04 de agosto de 2006

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Alinhamento Temporal de Seqüências de Vídeo Adquiridas por Câmeras Perspectivas e Catadióptricas

GERALDO AUGUSTO MASSAHUD RODRIGUES DOS SANTOS

Proposta de dissertação defendida e aprovada pela banca examinadora constituída por:

Ph. D. Mario Montenegro Campos – Orientador
Universidade Federal de Minas Gerais

Ph. D. Rodrigo Lima Carceroni – Co-orientador
Universidade Federal de Minas Gerais

Ph. D. Arnaldo de Albuquerque Araújo
Universidade Federal de Minas Gerais

Ph. D. Flávio Luis Cardeal Pádua
Centro Federal de Educação Tecnológica de Minas Gerais

Belo Horizonte, 04 de agosto de 2006

Resumo

Este trabalho trata o problema de estimar o alinhamento temporal entre 2 seqüências de vídeo não sincronizadas da mesma cena 3D, capturadas de pontos de vista distintos por uma câmera perspectiva e uma câmera para-catadióptrica.

Apresentamos um método que é uma extensão de uma técnica já existente para alinhar temporalmente múltiplas seqüências de vídeo adquiridas por câmeras perspectivas.

Resultados experimentais com seqüências reais mostram que nosso método pode alinhar os vídeos mesmo quando possuem grandes desalinhamentos. Finalmente, resultados experimentais com seqüências sintéticas mostram como nosso método se comporta na presença de erros no sistema.

Abstract

This work adresses the problem of estimating the temporal alignment between 2 unsynchronized video sequences of the same dynamic 3D scene, captured from distinct viewpoints by a perspective camera and a catadioptric camera.

We present a method that is an extension of an existing technique for temporal aligning multiple video sequences acquired by perspective cameras.

Experimental results with real world sequences show that our method can accurately align the videos even when they have large misalignments. Finally, experimental results with synthetic sequences show how our method behaves in the presence of errors in the system.

Para meus pais e minha irmã, que sempre me apoiaram durante toda minha vida.

Agradecimentos

Devo agradecimentos a várias pessoas por finalmente ter chegado à conclusão deste trabalho.

Obrigado ao meu co-orientador, Rodrigo Lima Carceroni, por dar todo o suporte e dividir as idéias necessárias para a conclusão deste trabalho.

Obrigado ao meu amigo Flávio Cardeal Pádua, por trabalhar junto no desenvolvimento de todo este trabalho e ajudar a resolver os problemas encontrados.

Muito obrigado ao meu orientador Mario Montenegro Campos, por acreditar em mim e ajudar à conclusão desta jornada.

Obrigado aos meus familiares, por todo o apoio e compreensão. Principalmente meus pais, minha irmã e meu cunhado, que nunca me deixaram desistir.

Obrigado a todos os meus amigos, meus colegas do VERLab e todas as pessoas que participaram na conclusão deste trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Vários trabalhos em visão computacional utilizam seqüências de vídeos, e grande parte dos trabalhos que utilizam mais de uma seqüência de vídeo necessitam que estas seqüências estejam alinhadas temporalmente.

O alinhamento temporal entre duas seqüências de vídeo permite que, para uma dada coordenada temporal ou número de quadro t em uma seqüência de referência, seja possível determinar a correspondente coordenada temporal t' na outra seqüência. Alguns exemplos de aplicações que utilizam seqüências de vídeo alinhadas temporalmente são reconstrução tridimensional, captura de movimento, rastreamento entre várias câmeras, fusão sensorial e efeitos especiais cinematográficos.

As seqüências de vídeo podem ser alinhadas no momento da gravação ou a posteriori, através de vídeos gravados, mas não alinhados. Para alinhar temporalmente seqüências de vídeo no momento da gravação normalmente utiliza-se hardware especializado, conseguindo-se erros de alinhamento extremamente baixos. Este método é muito utilizado na indústria cinematográfica, porém possui um custo elevado, principalmente quando um número grande de câmeras precisa ser alinhado.

Para seqüências gravadas mas não alinhadas temporalmente, como por exemplo jogos de futebol ou gravações das câmeras de seguranças de empresas, o alinhamento pode ser feito manualmente ou automaticamente, através de métodos de visão computacional. O alinhamento manual é trabalhoso e sua acurácia varia de indivíduo para

indivíduo, além disso este método não permite alinhamento sub-quadro. Existem vários métodos para alinhamento temporal automático de seqüências de vídeo (Pádua et al., 2004; Pádua, 2005; Caspi e Irani, 2000, 2001; Caspi et al., 2002; Lee et al., 2000; Wolf e Zomet, 2002a,b; Rao et al., 2003; Stein, 1998). Entre estes métodos, está o método de alinhamento de seqüências obtidas a partir de câmeras perspectivas, proposto por Pádua et al.(Pádua et al., 2004; Pádua, 2005), o qual foi estendido neste trabalho, permitindo sua utilização na determinação do alinhamento temporal entre câmeras perspectivas e para-catadióptricas.

Sistemas catadióptricos são sistemas visuais normalmente formados por uma câmera presa a um espelho convexo, o que permite um amplo campo de visão. São utilizados por exemplo na navegação de robôs e em vigilância. A geometria destes sistemas, em especial os de projeção central, já foi bastante estudada pela comunidade científica, sendo bem conhecida e de fácil acesso (Svoboda e Pajdla, 2002; Micusik et al., 2002). O sistema formado por espelhos parabolóides com câmeras ortográficas possui inclusive uma definição própria, para-catadióptricos.

Uma das motivações da utilização de sistemas catadióptricos no método de alinhamento é a possibilidade de uma câmera catadióptrica compartilhar a área de visão de todas as outras câmeras da montagem. Essa câmera poderia então ser utilizada como referência temporal de todas as câmeras.

A definição do problema tratado neste trabalho é a seguinte:

Alinhar temporalmente uma seqüência de vídeo gravada com uma câmera perspectiva com outra seqüencia gravada com uma câmera para-catadióptrica, ambas estáticas de um mesmo evento temporal e com taxas de quadros constantes, dado como entrada a geometria epipolar das seqüências e trajetórias de objetos encontrados em ambas as seqüências.

Em outras palavras, com a geometria epipolar de duas seqüências de vídeo de um mesmo evento e as trajetórias dos objetos existentes nessas seqüências, ambas gravadas

com uma câmera perspectiva e outra para-catadióptrica, de taxas de quadro constantes, é possível alinhar temporalmente estas seqüências.

A solução do problema de alinhamento temporal desenvolvida trata dos seguintes casos:

Taxa de quadros desconhecidas As taxas de quadros das seqüências são desconhecidas, podendo ser diferentes, porém devem ser constantes.

Deslocamento temporal arbitrário O deslocamento temporal entre as seqüências é desconhecido e pode ser grande.

Movimento desconhecido O movimento dos objetos no espaço é desconhecido, não sendo necessariamente planar.

Falhas no rastreamento Podem existir falhas no rastreamento dos objetos nas seqüências.

A idéia básica da abordagem é a definição de uma reta N-dimensional que captura as relações temporais entre N seqüências de vídeo. A propriedade fundamental desta reta é que a estimativa de pontos sobre a reta pode ser feita sem o conhecimento prévio da reta. Assim o problema de se estimar o alinhamento temporal entre N seqüências é reduzido para o problema de se estimar uma única reta a partir de um conjunto de pontos gerados em $\Re^N$. Neste trabalho, o número de seqüências trabalhadas será 2.

O resultado do alinhamento é uma equação linear $t' = \alpha t + \beta$ onde t é o tempo de uma seqüência e t' é o tempo deste evento na outra seqüência. Os parâmetros α e β são a razão entre as taxas de quadros das seqüências e o deslocamento temporal entre as mesmas, respectivamente. Esta representação funciona para quaisquer seqüências de vídeo com taxas de quadros fixas, que são as encontradas normalmente.

O método desenvolvido utiliza o cruzamento das trajetórias dos objetos rastreados em uma seqüência com as linhas epipolares geradas pelos objetos rastreados na outra seqüência. Neste trabalho será mostrado que ele é válido não apenas entre câmeras

Figura 1.1: Exemplo de imagem obtida por uma câmera catadióptrica, note que as quatro paredes que formam o laboratório são capturadas.

perspectivas, mas entre qualquer conjunto de câmeras onde seja possível obter uma geometria epipolar.

Este trabalho trata em especial do caso de câmeras perspectivas alinhadas com câmeras para-catadióptricas, que são câmeras catadióptricas montadas com um espelho parabólico e uma câmera ortográfica.

1.1 Organização da dissertação

Esta seção apresenta uma descrição sucinta de cada parte desta dissertação.

No Capítulo 1, é dada uma introdução do problema que será tratado, com alguns exemplos de aplicações reais do mesmo.

O Capítulo 2 é uma revisão de trabalhos já desenvolvidos na área de alinhamento temporal de seqüências de vídeo.

O Capítulo 3 apresenta fundamentos teóricos necessários para entender a metodologia utilizada para se resolver o problema. Como sistemas catadióptricos e a geometria

epipolar das câmeras.

A técnica de alinhamento temporal utilizada e sua extensão para câmeras para-catadióptricas e perspectivas se encontra no Capítulo 4.

Os experimentos feitos, com imagens reais e simulação, são apresentados e discutidos no Capítulo 5.

O Capítulo 6 conclui o trabalho, discutindo os resultados como um todo, problemas encontrados, e propostas de trabalhos futuros.

Por fim o Apêndice A mostra como pode ser feita a extensão do método de refinamento para câmeras para-catadióptricas e perspectivas.

Capítulo 2

Trabalhos relacionados

Os principais trabalhos sobre alinhamento espaço-temporal podem ser divididos em duas categorias: métodos baseados em características e métodos diretos. Os métodos baseados em características (Pádua et al., 2004; Pádua, 2005; Caspi et al., 2002; Rao et al., 2003; Wolf e Zomet, 2002a,b; Lee et al., 2000; Stein, 1998) extraem informações das trajetórias dos objetos rastreados, e os métodos diretos (Caspi e Irani, 2000, 2001) extraem as informações das intensidades dos pixels das imagens. Os métodos diretos normalmente alinham apenas seqüências com aparência similar, enquanto os métodos baseados em características conseguem alinhar seqüências em situações mais desafiadoras, tais como aquelas nas quais as câmeras possuem diferentes ampliações (zoom), sensibilidades espectrais distintas, e lentes com grande campo de visão.

Este trabalho utiliza um método baseado em características desenvolvido anteriormente (Pádua et al., 2004; Pádua, 2005), porém agora estendido para aplicações onde não somente câmeras perspectivas, mas também câmeras para-catadióptricas são utilizadas. Como nenhum método específico para câmeras catadióptricas foi encontrado, será feita a comparação dos métodos para câmeras perspectivas.

O método utilizado reduz a computação do alinhamento temporal a uma regressão linear, sendo uma solução robusta, que funciona mesmo na presença de outliers. Os outros métodos baseados em características existentes (Caspi et al., 2002; Rao et al., 2003; Wolf e Zomet, 2002a,b; Lee et al., 2000; Stein, 1998) procuram em todo espaço

de possíveis alinhamentos temporais. Estes métodos necessitam, portanto, de assumir várias restrições sobre os dados, como por exemplo o número de seqüências de vídeo ser restrito a duas; o desalinhamento temporal ser inteiro; a taxa de quadros das câmeras ser conhecida; e o desalinhamento temporal estar dentro de um pequeno intervalo especificado pelo usuário. O método utilizado neste trabalho alinha N seqüências de vídeo em um único passo, encontra a solução mesmo com um grande desalinhamento temporal e sem a necessidade de conhecer previamente a razão as taxas de quadros das câmeras. Além disso, o método faz a correspondência entre posições instantâneas em uma seqüência com todo o espaço de trajetórias das outras seqüências, assim a qualidade do alinhamento é invariante à magnitude da diferença temporal inicial entre as seqüências.

O método de Pádua et al. (Pádua, 2005; Pádua et al., 2004) é mais relacionado com o desenvolvido por Caspi et al. (Caspi et al., 2002), no qual a geometria epipolar e o desalinhamento temporal são recuperados da imagem da trajetória de um único ponto da cena, que é visível nas duas seqüências, e depois são refinados utilizando mais pontos. Para isso, eles assumem taxas de quadro conhecidas e resolvem um problema de otimização não linear, sendo necessária uma boa estimativa inicial do desalinhamento temporal e da geometria epipolar. Esse método ainda assume que os objetos são rastreados sem interrupções durante toda a seqüência, o que pode ser difícil de se conseguir em vídeos reais. Diferente disso, no método de Pádua et al. as trajetórias rastreadas não necessitam de continuidade durante toda a cena, isto é, a solução requer a habilidade de se rastrear pontos na cena somente ao longo de dois quadros consecutivos na mesma seqüência. Além disso, não é necessário estabelecer correspondências dos objetos rastreados entre as seqüências .

Os outros métodos baseados em características encontrados na literatura (Rao et al., 2003; Wolf e Zomet, 2002a,b; Lee et al., 2000; Stein, 1998) utilizam restrições posição-para-posição (*position-to-position*). Através das correspondências de posições instantâneas de pontos rastreados nas duas cenas, eles verificam se algum dos possíveis de-

salinhamentos temporais são consistentes, ou seja, quando aplicados nas seqüências, fazem com que elas se relacionem por uma única transformação rígida. Todos esses métodos precisam fazer a procura em todo o espaço de possível desalinhamento temporal, o que torna impraticável o cálculo de desalinhamentos temporais grandes. Além disso, nenhum deles consegue resolver desalinhamentos temporais com precisão sub-quadro, e apenas um (Wolf e Zomet, 2002b) tolera outliers, assumindo que as câmeras são ortográficas. Por outro lado, o método de Pádua et al. não é sensível ao tamanho do desalinhamento temporal, tolera grande quantidade de *outliers* e faz alinhamento sub-frame.

Finalmente, pode-se citar o método proposto por Caspi e Irani (Caspi e Irani, 2000), que é um método direto. Nele, as seqüências de vídeo são tratadas como dois volumes, e o alinhamento é encontrado por meio da solução de sistemas lineares que modelam a transformação de um volume em outro, levando-se em consideração as intensidades dos pixels e seus gradientes espaço-temporais. Na medida que este método modela as transformações espaciais entre seqüências como homografias, ele não é apropriado para alinhar seqüências com significativas descontinuidades de profundidade (cenas não planas).

Capítulo 3

Fundamentos teóricos

3.1 Sistemas catadióptricos

Sistemas catadióptricos são sistemas de visão formados por uma câmera anexada a um espelho convexo (Nayar, 1997). Este tipo de montagem aumenta consideravelmente o campo de visão, permitindo a visualização de 180^o em torno do eixo da câmera. A Figura 1.1 é um exemplo de uma imagem obtida por uma câmera catadióptrica formada por um espelho parabólico e uma câmera ortográfica, denominada câmera para-catadióptrica. Como pode ser notado, quanto mais próximo do centro do espelho, menor é a resolução obtida. O mesmo ocorre para quando se chega perto das bordas da imagem.

O interesse deste trabalho é em sistemas catadióptricos de projeção central, que se caracterizam por possuir um único ponto de projeção. A propriedade mais importante dos sistemas de projeção central para este trabalho é possuir geometria epipolar, necessária para o método de alinhamento.

Existem vários tipos de montagens para sistemas catadióptricos de projeção central (Svoboda e Pajdla, 2002; Nayar, 1997), as mais utilizadas são espelhos hiperbólicos com câmeras perspectivas e espelhos parabólicos com câmeras ortográficas. Na montagem do sistema hiperbólico, a câmera perspectiva deve ter seu centro de projeção coincidente com o segundo foco do hiperbolóide do espelho, pois todos os raios que vão em direção

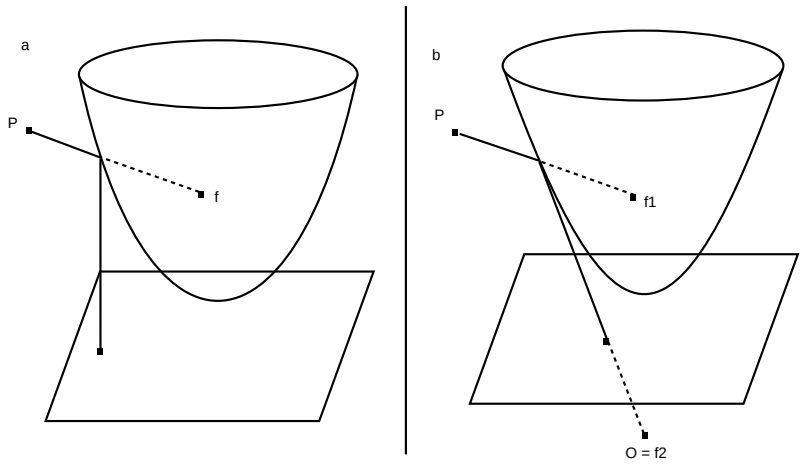


Figura 3.1: Exemplos de câmeras catadióptricas de projeção central. a) espelho parabólico com câmera ortográfica. b) espelho hiperbólico com câmera perspectiva. Note que, na montagem com espelho hiperbólico, o centro de projeção da câmera perspectiva deve coincidir com o segundo foco do espelho.

ao primeiro foco, são projetados para o segundo foco, que é o centro de projeção do sistema. No sistema para-catadióptrico, os raios que vão em direção ao foco são projetados perpendicular ao eixo de simetria do espelho, como mostra a Figura 3.1, portanto a câmera utilizada deve ser ortográfica, e o centro de projeção é o foco do espelho.

3.2 Geometria epipolar

3.2.1 Modelo de câmera

Uma câmera perspectiva é modelada por vários parâmetros, esses parâmetros são conhecidos como parâmetros intrínsecos e parâmetros extrínsecos da câmera. Os parâmetros extrínsecos definem a orientação e localização da câmera com respeito a um sistema coordenado conhecido do mundo, e os parâmetros intrínsecos são os parâmetros necessários para associar coordenadas de pixels de um ponto na imagem com as

coordenadas correspondentes no sistema coordenado da câmera.

Uma escolha típica de parâmetros extrínsecos para se descrever a posição da câmera no mundo são um vetor de translação T_w e uma matriz de rotação R_w. A translação descreve as posições relativas das origens do sistema coordenado do mundo e do sistema coordenado da câmera, enquanto a rotação alinha os eixos correspondentes dos dois sistemas.

Os parâmetros intrínsecos para uma câmera *pinhole* capturam a projeção perspectiva, a transformação entre coordenadas da câmera e coordenadas de pixel e a distorção geométrica introduzida pela óptica. O parâmetro que modela a projeção perspectiva é a distância focal. A transformação de coordenadas de câmera para coordenadas de pixel é modelada pelo centro de projeção em coordenadas de pixel e pelo tamanho efetivo de um pixel nas direções horizontal e vertical. Por fim a distorção geométrica é capturada pela distorção radial da câmera.

Ignorando a distorção radial, os parâmetros intrínsecos e extrínsecos podem ser colocados em forma matricial:

$$M = \begin{bmatrix} f/s_x & 0 & o_x \\ 0 & f/s_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3.1}$$

, onde f é a distância focal, (o_x, o_y) são as coordenadas do centro de projeção em pixels, e (s_x, s_y) são o tamanho efetivo de um pixel na horizontal e vertical.]

A projeção de um ponto do mundo p em um pixel na imagem da câmera $\overline{p}$ é portanto modelada pela equação:

$$\overline{p} = M(R_w p + T_w) \tag{3.2}$$

3.2.2 Restrição epipolar

A geometria epipolar associa pontos entre duas câmeras. Os pontos nos sistemas coordenados das câmeras são associados por uma matriz E, chamada Matriz Essencial,

já os pixels são associados através de outra matriz F, chamada Matriz Fundamental. A matriz essencial captura os parâmetros extrínsecos das duas câmeras, e a matriz fundamental captura os parâmetros extrínsecos e intrínsecos.

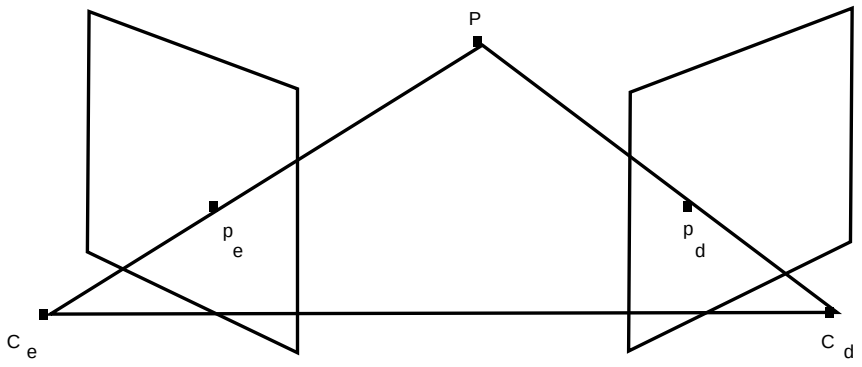


Figura 3.2: Geometria epipolar entre duas câmeras perspectivas.

As duas câmeras perspectivas da figura 3.2 tem seus sistemas coordenados relacionados por uma matriz de rotação R e uma matriz de translação $T = C_d - C_e$. Os pontos P, C_e e C_d formam um plano no espaço, os pontos p_e e p_d pertencem a este plano e podem ser relacionados entre si pela equação:

$$p_e = Rp_d + T, \tag{3.3}$$

$$p_e - T = Rp_d. \tag{3.4}$$

T, p_e e $p_e - T$ podem ser considerados pontos, e estão no mesmo plano formado por C_e, P e C_d. A equação deste plano pode ser escrita utilizando-se a condição de coplanaridade de T, p_e e $p_e - T$:

$$(p_e - T)^T T \times p_e. \tag{3.5}$$

Substituindo 3.4, se obtém

$$(R^T p_d)^T T \times p_e. \tag{3.6}$$

Um produto vetorial pode ser reescrito como uma multiplicação por uma matriz

$$T \times p_e = S p_e, \tag{3.7}$$

$$S = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix}. \tag{3.8}$$

Usando este fato, a equação 3.6 se transforma em

$$p_d^T E p_e = 0, \tag{3.9}$$

$$E = RS. \tag{3.10}$$

A matriz E é chamada de matriz essencial, e contém a ligação entre a restrição epipolar e os parâmetros extrínsecos do sistema. Para obter a equação em relação aos pixels $\overline{p_e}$ e $\overline{p_d}$ das câmeras, tem-se que

$$p_d = M_d^{-1} \overline{p_d} \tag{3.11}$$

$$p_e = M_e^{-1} \overline{p_e}. \tag{3.12}$$

, onde M_d e M_e são as matrizes de parâmetros intrínsecos da câmera direita e esquerda, respectivamente. Substituindo 3.11 e 3.12 em 3.2.2 se obtém a restrição epipolar à partir dos pixels:

$$\overline{p_e}^T M_e^{-1} E M_d^{-1} \overline{p_d} = 0, \tag{3.13}$$

$$\overline{p_e}^T F \overline{p_d} = 0, \tag{3.14}$$

F é chamada de matriz fundamental, e relaciona diretamente os pixels das imagens.

Capítulo 4

Metodologia

4.1 Sistema de alinhamento temporal

O sistema de alinhamento temporal utilizado é uma modificação do sistema apresentado por Pádua et al. (Pádua et al., 2004; Pádua, 2005). Este sistema é constituído por algumas etapas, que vão desde a obtenção das seqüências, rastreamento de objetos móveis presentes nas seqüências, até a identificação da linha temporal em si. Nesta seção será apresentada uma visão geral das etapas deste sistema.

A Figura 4.1 é um diagrama do sistema, com suas cinco etapas principais: 1) obtenção das seqüências de vídeo; 2) obtenção da geometria epipolar; 3) rastreamento; 4) geração do mapa de votos; e 5) obtenção da reta temporal. No fluxo do diagrama existem ações e dados, sendo as ações representadas por losangos, e os dados por retângulos. Os dados produzidos em uma etapa superior são utilizados como entrada para as ações das etapas inferiores, e o último dado obtido é a linha temporal que descreve o desalinhamento temporal das seqüências.

A seguir será apresentada uma breve descrição das ações executadas em cada etapa, serão identificados também os dados de entrada para estas ações e quais dados são gerados por elas.

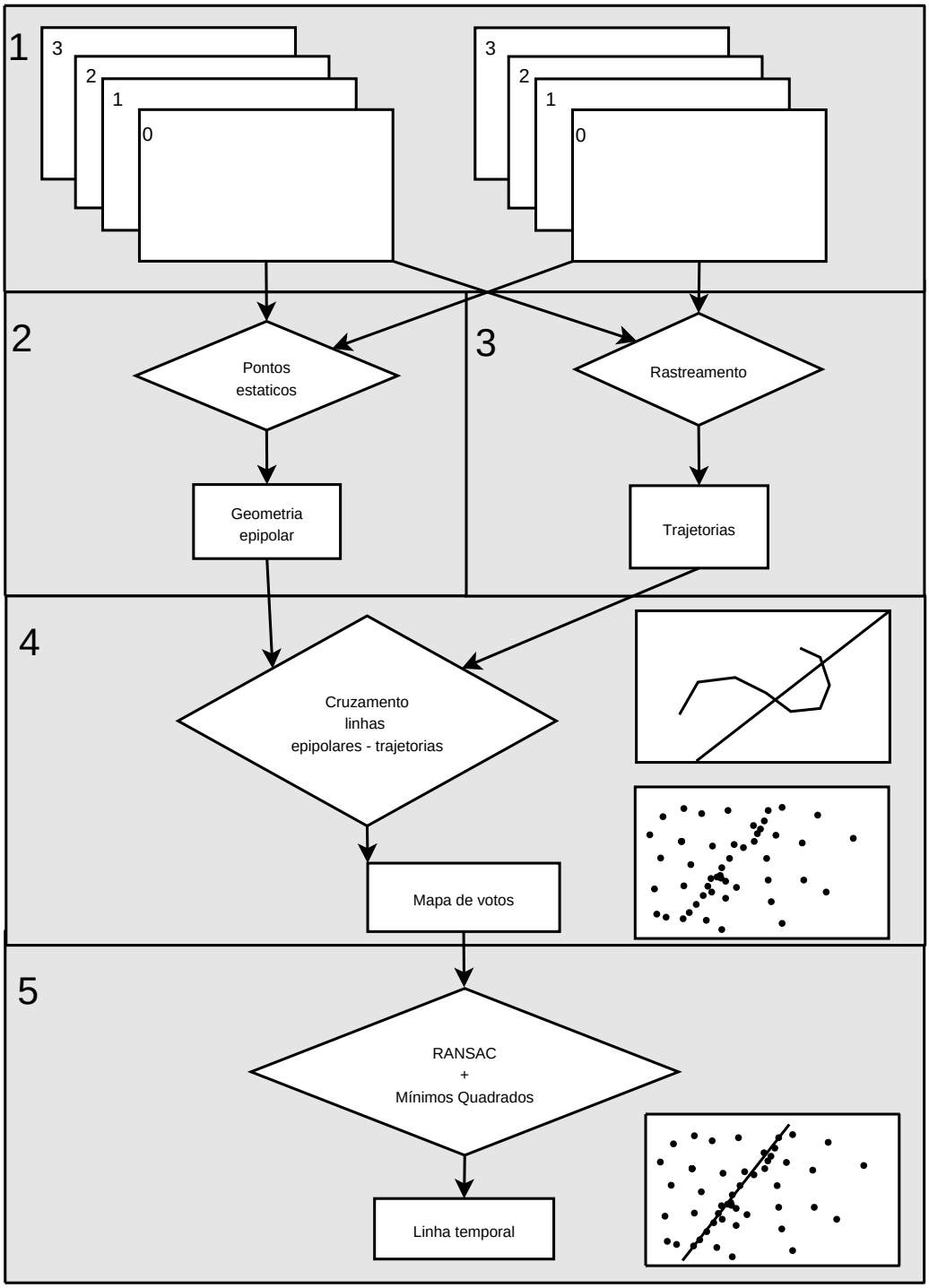


Figura 4.1: Visão geral do sistema de alinhamento temporal.

1. **Obtenção das seqüencias de vídeo**

   A primeira etapa do sistema é obter as seqüências de vídeo. Algumas restrições são necessárias na obtenção de seqüências para o método. As seqüências devem ser obtidas por sistemas ópticos onde seja possível extrair a geometria epipolar, pois o método utiliza a geometria epipolar para encontrar o desalinhamento. As câmeras devem ter taxas de quadro fixas, pois a linha temporal encontrada não captura variação de taxas de quadro em uma seqüência. As seqüências também devem possuir interseções espacial e temporal.

   Ao final desta etapa, as duas seqüências de vídeo são obtidas.

2. **Geometria epipolar entre as seqüencias**

   Após obter as seqüências de vídeo, deve-se extrair a geometria epipolar das mesmas. No diagrama da Figura 4.1 a geometria epipolar é obtida através da extração de pontos estáticos correspondentes nas seqüências. Com um sistema de equações montadas a partir destes pontos é possível obter a geometria epipolar capturada na matriz fundamental. Qualquer outro método que obtenha a geometria epipolar pode ser utilizado, como calibração das câmeras, desde que no final seja possível obter a geometria epipolar do sistema, e dela as linhas epipolares.

   Ao final desta etapa, a geometria epipolar das seqüências é calculada.

3. **Rastreamento**

   Cada seqüência de vídeo é submetida a um rastreador, que encontrará trajetórias dos objetos na seqüência. Nenhum rastreador em específico precisa ser utilizado, e não é necessário associar objetos rastreados em uma seqüência com objetos rastreados na outra seqüência. Também não é necessário que os objetos sejam rastreados durante toda a seqüência, é permitido ao rastreador perder objetos e também adicionar novos objetos no meio da seqüência.

Esta etapa gera as trajetórias dos objetos presentes das seqüências, cada ponto rastreado tem associado uma posição x e y na imagem e o número do quadro que pertence.

4. **Geração do mapa de votos**

A geração do mapa de votos pode ser considerada a principal etapa do método. Os dados necessários para esta etapa são a geometria epipolar das seqüências e as trajetórias rastreadas. O mapa de votos é um espaço euclidiano cuja dimensão é uma função do número de câmeras utilizadas. Os eixos detes espaço contém as coordenadas temporais (números dos quadros) das seqüências de vídeo. Cada ponto neste espaço é um voto para o verdadeiro alinhamento temporal entre as seqüências.

Para se obter os votos, para cada ponto rastreado na primeira seqüência, utiliza-se a geometria epipolar para gerar uma linha epipolar na segunda seqüência - no caso da Figura 4.1 esta linha é uma reta. Esta linha epipolar é então cruzada com todas as trajetórias rastreadas na segunda seqüência.

Cada cruzamento de linha epipolar com trajetória irá gerar um voto no mapa de votos. Como cada ponto rastreado possui uma coordenada temporal, a coordenada temporal do voto na seqüência 1 é a coordenada temporal do ponto utilizado para gerar a linha epipolar, e a coordenada temporal do voto na seqüência 2 é a coordenada temporal calculada na trajetória a partir do ponto de cruzamento. Para calcular esta coordenada temporal utiliza-se os dois pontos pertencentes aos extremos do segmento de reta cortado pela linha epipolar, considera-se que a velocidade do movimento neste segmento é constante, assim a coordenada temporal do cruzamento é a coordenada temporal do primeiro ponto do segmento somado à fração de tempo calculada pela posição em que a linha epipolar cortou o segmento da trajetória.

Após gerar uma linha epipolar para cada ponto da trajetória da seqüência 1 e

calcular seus votos, o resultado desta etapa será o mapa de votos.

5. **Obtenção da reta temporal**

   O mapa de votos possui votos de todos os cruzamentos das linhas epipolares de uma câmera com as trajetórias da outra câmera, portanto possui muitos votos que não são uma associação correta entre os quadros das seqüências, estes votos são considerados pontos espúrios (*outliers*).

   Espera-se que os *outliers* estejam distribuídos pelo mapa de forma aleatória, enquanto todos os pontos que associam corretamente os quadros (*inliers*) estão próximos/formam uma mesma reta no mapa.

   È necessário, portanto, um algoritmo que consiga extrair a melhor reta do mapa, mesmo na presença de muitos *outliers*. O algoritmo utilizado é o Random Sample Consensus (Fischler e Bolles, 1981) - RANSAC. O que o RANSAC faz é obter o maior conjunto de pontos do sistema que instancie o modelo proposto, este modelo seria uma reta.

   Após obter este melhor conjunto de pontos, ou conjunto consenso, aplica-se uma regressão linear neste conjunto, por exemplo através de mínimos quadrados, a reta obtida dessa regressão é a linha temporal.

   Ao final desta etapa, obtém-se o alinhamento temporal entre as seqüências, capturado na linha temporal.

4.2 Geometria epipolar para-catadióptrica - perspectiva

Como as câmeras para-catadióptricas e perspectivas são de projeção central, é possível definir a geometria epipolar entre elas. Uma diferença nesta geometria é o fato da restrição epipolar ser entre o espelho e a câmera perspectiva, e não diretamente com os pontos da imagem da câmera ortográfica.

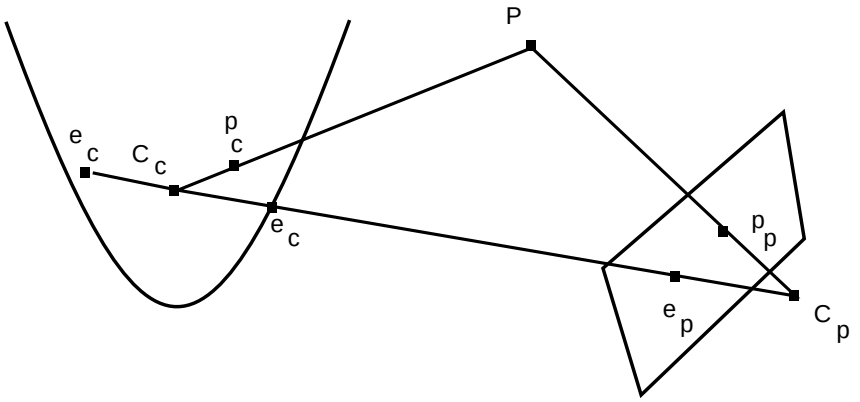


Figura 4.2: Geometria epipolar do sistema para-catadióptrico - perspectivo

A Figura 4.2 ilustra a geometria de nosso sistema. Analogamente à geometria epipolar de duas câmeras perspectivas, tem-se o plano formado por C_c, P e C_p, os pontos na superfície do espelho p_c e na câmera p_p, pode-se então escrever a seguinte restrição epipolar:

$$p_c^T M_c^{-T} E M_p^{-1} p_p = 0, \tag{4.1}$$

onde M_c e M_p são as matrizes de parâmetros intrínsecos do espelho e da câmera perspectiva, respectivamente. E E é a matriz essencial.

A matriz fundamental do sistema é portanto definida da mesma forma da seção anterior.

$$p_c^T F p_p = 0 \tag{4.2}$$

$$F = M_c^{-T} E M_p^{-1}. \tag{4.3}$$

Mas mesmo com a geometria epipolar da equação 4.11, o ponto p_c está definido

na superfície do espelho, e os dados obtidos da câmera para-catadióptrica são pixels na câmera ortográfica. Portanto, é necessário transformar um ponto da superfície da imagem da câmera ortográfica para a superfície do espelho, pois só assim é possível utilizar a restrição epipolar.

O primeiro passo é descobrir como transformar um ponto p_o em coordenadas de câmera da câmera ortográfica em um ponto no espelho, note que p_o está no em coordenadas da câmera, o ponto correspondente a p_o em coordenadas de pixel é $\overline{p_o}$, a obtenção de p_o a partir de $\overline{p_o}$ será tratada posteriormente.

Considerando que o sistema coordenado da câmera ortográfica seja o mesmo do espelho, exceto por uma translação em z, p_o é definido como

$$p_o = \begin{bmatrix} x_c \\ y_c \\ 1 \end{bmatrix}. \tag{4.4}$$

O espelho parabólico com o sistema coordenado centrado no foco tem a equação

$$z = \frac{x^2 + y^2 - b^2}{2b}, \tag{4.5}$$

onde b é um parâmetro do espelho definido como duas vezes a distância do vértice ao foco. Portanto, com as coordenadas x e y do ponto p_o é possível obter a coordenada z.

A partir das Equações 4.4 e 4.5, p_c é definido como

$$p_c = \begin{bmatrix} x_c \\ y_c \\ \frac{x_c^2 + y_c^2 + b^2}{2b} \end{bmatrix}. \tag{4.6}$$

Obtendo p_c a partir de p_o basta encontrar p_o à partir do pixel $\overline{p_o}$ na imagem da câmera ortográfica. É necessário portanto transformar o sistema coordenado da imagem para o sistema coordenado da câmera, como mostra a Figura 4.3. Para isto basta

uma transformação afim que corrija a imagem do espelho para um círculo, e translade o centro do sistema coordenado para o centro deste círculo, a transformação inversa a esta pode ser considerada a matriz de parâmetros intrínsecos da câmera ortográfica, M_o, de onde

$$p_o = M_o^{-1}\overline{p_o}. \tag{4.7}$$

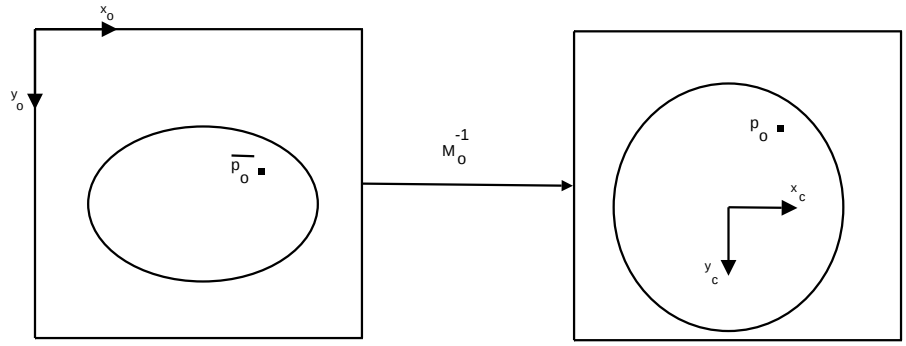


Figura 4.3: Transformação na imagem da câmera ortográfica, colocando o sistema coordenado no centro da imagem do espelho e fazendo com que esta imagem tenha o formato de um círculo.

Unindo as Equações 4.11, 4.4, 4.6, e 4.7, tem-se a definição completa da geometria epipolar do sistema, a partir das imagens capturadas:

$$p_c^T F p_p = 0 \tag{4.8}$$

$$p_c = \begin{bmatrix} x_c \\ y_c \\ \frac{x_c^2+y_c^2-b^2}{2b} \end{bmatrix} \tag{4.9}$$

$$\begin{bmatrix} x_c \\ y_c \\ 1 \end{bmatrix} = M_o^{-1}\overline{p_o} \tag{4.10}$$

$$\begin{bmatrix} x_c & y_c & \frac{x_c^2+y_c^2}{2b} \end{bmatrix} F \begin{bmatrix} x_p \\ y_p \\ 1 \end{bmatrix}. = 0 \tag{4.11}$$

Para resolver este sistema, é necessário encontrar F e b. Micusik et al. (Micusik et al., 2002) resolvem um sistema parecido transformando a restrição epipolar em um Problema de Autovalor Polinomial. A transformação da Equação 4.11 em um Problema de Autovalor Polinomial gera a seguinte equação:

$$\left(D_1 - bD_2 - b^2 D_3\right) f = 0, \tag{4.12}$$

onde b e f são os parâmetros que devem ser estimados e

$$\begin{aligned} f &= \begin{bmatrix} f_{11} & f_{12} & f_{13} & \dots & f_{33} \end{bmatrix}^T \\ D_1 &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & (x_c^2+y_c^2)x_p & (x_c^2+y_c^2)y_p & (x_c^2+y_c^2) \end{bmatrix} \\ D_2 &= \begin{bmatrix} 2x_cx_p & 2x_cy_p & 2x_c & 2y_cx_p & 2y_cy_p & 2y_c & 0 & 0 & 0 \end{bmatrix} \\ D_3 &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -x_p & -y_p & -1 \end{bmatrix} \end{aligned}$$

, onde $f_{11}, f_{12}, \dots$ são os valores da matriz fundamental nas coordenadas $(1,1), (1,2), \dots$.

Para resolver este problema existem algoritmos eficientes já implementados, como a função polyeig do MATLAB. As matrizes D_i devem ser quadradas para a solução, portanto são necessárias 9 correspondências de pontos entre as imagens para solucionar o problema.

A solução retornada pela função polyeig é um vetor de 18 b's e uma matriz 9x18, onde cada coluna corresponde a uma solução de f, escolhendo apenas soluções reais positivas para b (normalmente existem de de 1 a 3) obtém-se uma matriz F respectiva para cada b. A solução com o menor erro é escolhida.

Micusik et al. (2002) utiliza os ângulos entre os raios e o plano epipolar para o

cálculo do erro, através da seguinte fórmula:

$$\epsilon(p_p, p_c, F) = A/s - \sqrt{\frac{A^2}{4-B}},$$
$$A = p_p^T F^T F p_p + p_c^T F F^T p_c, B = (p_c^T F p_p)^2.$$

4.3 Geração dos votos

Com a geometria epipolar resolvida, resta encontrar as equações das linhas epipolares nas imagens das câmeras, para que se possa cruzar as linhas com as trajetórias, gerando os votos no espaço de votos. Dois casos são possíveis, gerar as linhas epipolares na câmera perspectiva a partir de pontos da câmera para-catadióptrica ou gerar as linhas epipolares na imagem da câmera para-catadióptrica a partir de pontos das câmeras perspectivas. Primeiramente será visto o caso mais simples de gerar as linhas epipolares na câmera perspectiva.

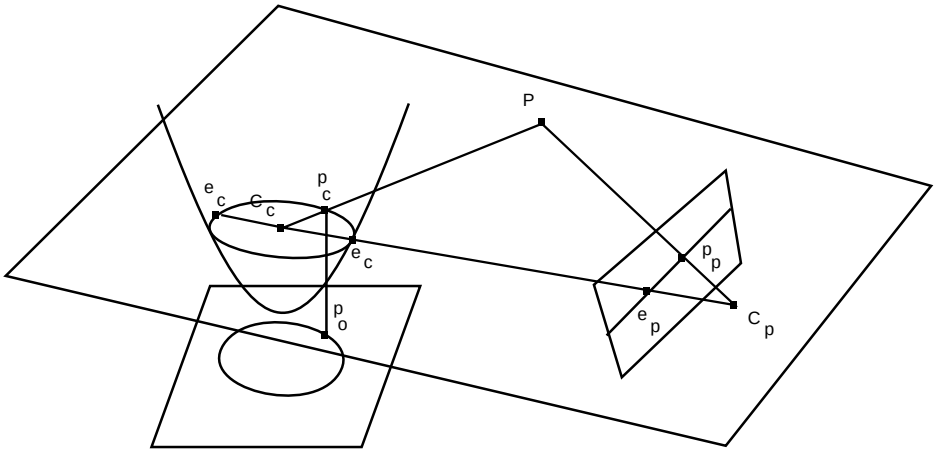


Figura 4.4: Geometria epipolar do sistema para-catadióptrico - perspectivo

Como mostra a Figura 4.4, a interseção do plano epipolar com o plano de imagem da câmera perspectiva é uma reta. A equação desta reta é obtida diretamente da

restrição epipolar (Eq 4.11).

$$l = p_c^T F, \tag{4.13}$$

onde $p_c = \begin{bmatrix} x_c \\ y_c \\ \frac{x_c^2+y_c^2+b^2}{2b} \end{bmatrix}$. Os pontos da câmera para-catadióptricas viram retas na câmera perspectiva, os votos são gerados portanto através do cruzamento de retas com as trajetórias

Para pontos da câmera perspectiva gerarem as linhas epipolares na câmera para-catadióptrica, a geometria do sistema é um pouco diferente. Neste caso, primeiro é necessário verificar o resultado do cruzamento entre o plano epipolar com a superfície do espelho catadióptrico, um esboço pode ser visto na Figura 4.4. Esta interseção é uma cônica no espaço, mais precisamente uma elipse ou uma parábola (Svoboda e Pajdla, 2002).

A forma matricial da projeção ortográfica desta cônica no plano xy é a equação

$$p_o^T Q(p_p) p_o = 0, \tag{4.14}$$

onde $Q(p_p)$ é a forma matricial da cônica nas coordenadas da câmera ortográfica e

$$p_o = \begin{bmatrix} x_c \\ y_c \\ 1 \end{bmatrix} \tag{4.15}$$

(Svoboda e Pajdla, 2002). Svoboda e Pajdla mostra que a equação da cônica é

$$sx^2 + 2bpx + sy^2 + 2bqy - sb^2 = 0, \tag{4.16}$$

onde $px + qy + sz = 0$ é a equação do plano epipolar, $n = [\ p \quad q \quad s\ ]^T$ é a normal do plano. Svoboda e Pajdla também mostra que a normal do plano é obtida a partir da matriz fundamental, multiplicando o ponto correspondente à cônica epipolar na outra

câmera, então

$$n = \begin{bmatrix} p \\ q \\ s \end{bmatrix} = Fp_p. \tag{4.17}$$

Transformando a Equação 4.16 para a forma matricial, obtém-se a matriz $Q(p_p)$ da Equação 4.14, que é

$$Q(p_p) = \begin{bmatrix} s & 0 & bp \\ 0 & s & bq \\ bp & bq & -b^2 s \end{bmatrix}. \tag{4.18}$$

A Equação 4.14 é a equação da cônica projetada no plano xy do sistema coordenado do espelho, mas a cônica de interesse é no sistema coordenado da imagem do espelho.

A transformação de $\overline{p_o}$ em p_o é dada pela Equação 4.7, substituindo a equação 4.7 em 4.14 se obtém a equação

$$\overline{p_o}^T M_o^{-T} Q(p_p) M_o^{-1} \overline{p_o} = 0, \tag{4.19}$$

que é a equação da cônica gerada por p_p na imagem da câmera ortográfica, portanto a matriz da equação da cônica na imagem é

$$A(p_p) = M_o^{-T} Q(p_p) M_o^{-1}. \tag{4.20}$$

Para encontrar os votos utilizando pontos na câmera perspectiva e trajetórias na câmera para-catadióptrica é necessário encontrar o cruzamento de cônicas $\overline{p_p}^T A(p_p) \overline{p_o} = 0$ com os segmentos de retas das trajetórias.

Cada voto possui duas coordenadas temporais, a primeira é a coordenada temporal do ponto que gerou a linha epipolar, a segunda é a coordenada temporal da interseção encontrada, sendo essa coordenada temporal não necessariamente um inteiro. Assumindo que em cada segmento da trajetória na imagem o objeto rastreado se move com

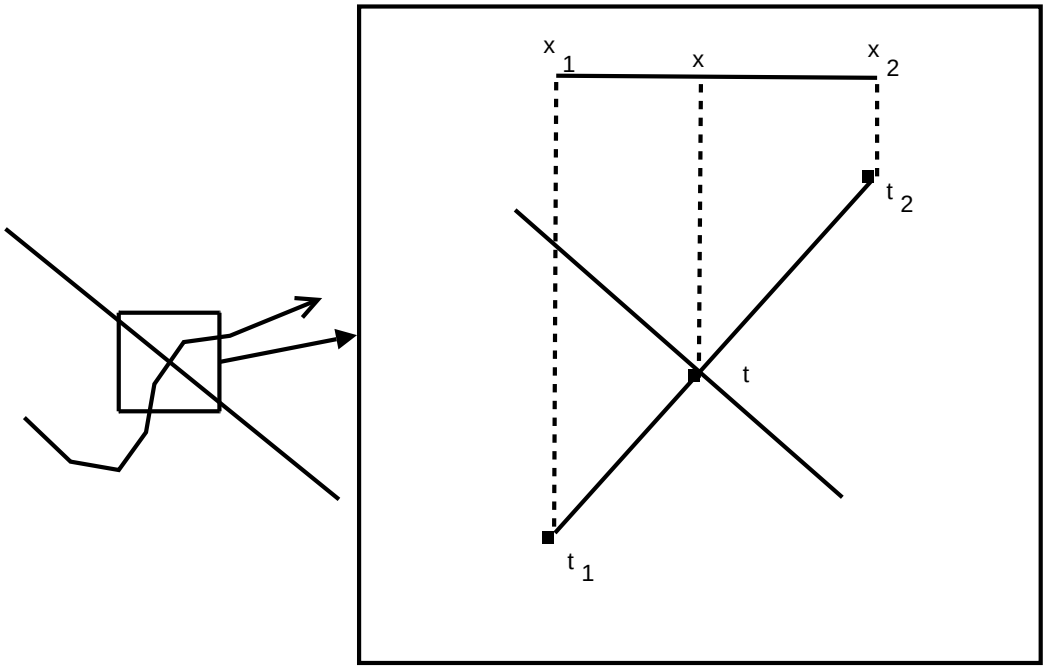


Figura 4.5: Relação geométrica utilizada para calcular o tempo de cada voto. Assumindo que o objeto se moveu com velocidade constante entre os instantes t_1 e t_2, a coordenada temporal do cruzamento da trajetória com a linha epipolar, t, é obtida através da relação $\frac{x-x_1}{x_2-x_1} = \frac{t-t_1}{t_2-t_1}$.

velocidade constante, a coordenada temporal do ponto de interseção pode ser calculado através de relações geométricas, como mostra a Figura 4.5.

4.4 Obtenção da linha temporal

O último passo do método é encontrar a linha temporal que descreve o desalinhamento temporal entre as duas seqüências de vídeo. Para isso utiliza-se o mapa de votos gerado na etapa anterior.

Os votos encontrados pelos métodos da seção anterior são colocados em um mapa de votos, em que cada coordenada é o tempo de uma das câmeras. A Figura 4.6 mostra

um exemplo de mapa de votos obtido.

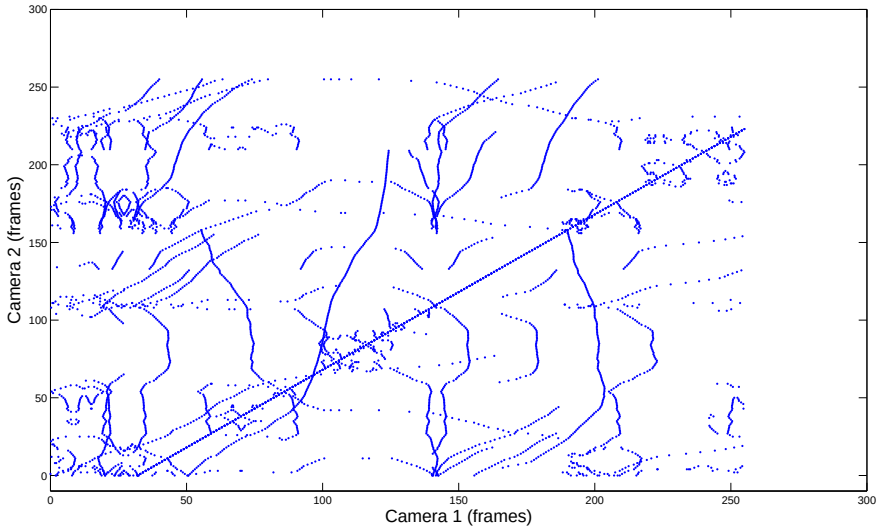


Figura 4.6: Exemplo de um mapa de votos obtido.

Através do mapa de votos, encontrar o desalinhamento temporal se resume a encontrar a melhor reta do mapa. Para esta tarefa o algoritmo RANSAC (Fischler e Bolles, 1981) é utilizado. O RANSAC é um algoritmo que procura a melhor ocorrência de um modelo nos dados, mesmo com a presença de muitos *outliers*.

Para encontrar a melhor reta, o RANSAC primeiro escolhe dois pontos aleatoriamente dos dados e gera uma reta que passa por estes dois pontos. Com a reta definida, ele passa por todos pontos dos dados, procurando os pontos que estão próximos da reta até uma distância máxima pré-definida, os pontos que estão dentro desta distância máxima são armazenados em um conjunto, chamado consenso.

O RANSAC repete o passo anterior um número finito de vezes, guardando no final de cada iteração o maior consenso obtido. Quando terminar todas as iterações ele terá uma certa probabilidade do consenso guardado ser o melhor conjunto de pontos que descreve uma reta nos dados.

O número de vezes que o RANSAC deve iterar nos dados é definido pela equação

$$n = \left\lceil \frac{log(1-p)}{log(1-r^2)} \right\rceil, \tag{4.21}$$

onde p é a probabilidade requerida de que pelo menos uma das iterações do RANSAC tenha o melhor modelo de reta, e r é a probabilidade de um ponto aleatório escolhido pertencer à melhor reta.

Encontrado o maior consenso após todas as iterações do RANSAC, basta executar um algoritmo de regressão linear nos pontos do consenso para obter a linha temporal. A Figura 4.7 mostra uma linha temporal encontrada no mapa de votos da Figura 4.6.

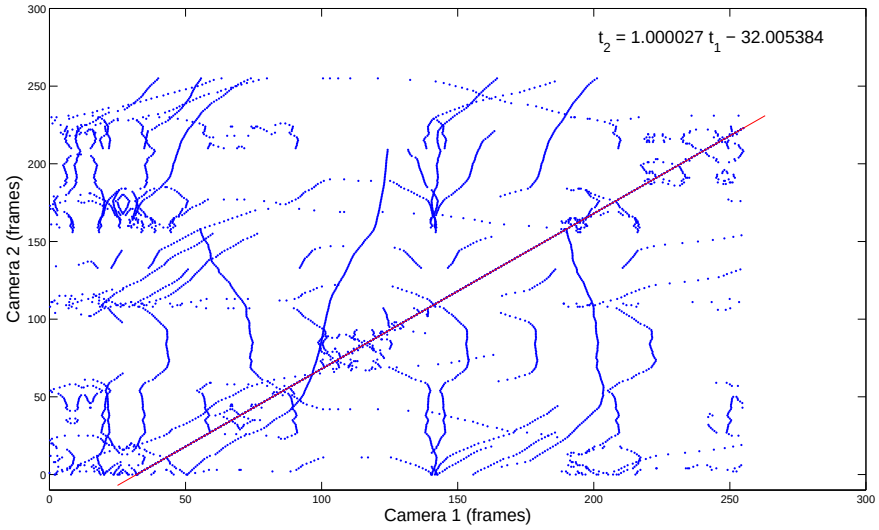


Figura 4.7: Exemplo de linha temporal encontrada.

Capítulo 5

Experimentos

5.1 Vídeo

Utilizando uma câmera para-catadióptrica e uma câmera perspectiva gravou-se dois vídeos de uma cena contendo duas pessoas caminhando no interior de um laboratório. Na cena as duas pessoas utilizam cones coloridos na cabeça. O rastreador WSL(Jepson et al., 2003) foi utilizado para rastrear esses cones.

A seqüência da câmera para-catadióptrica possui 768 quadros, e a seqüência da câmera perspectiva possui 899 quadros. A resolução da imagem da câmera perspectiva é 320x240 pontos, e da câmera para-catadióptrica 640x480. A taxa de quadros de ambas é 30 quadros por segundo. A reta temporal calculada manualmente é

$$t_c = t_p - 44, \tag{5.1}$$

com uma incerteza de 5 quadros.

Na câmera perspectiva o rastreador perdeu os objetos a partir do quadro 598, e começou o rastreamento no quadro 55, portanto as trajetórias da câmera perspectiva são apenas de 543 quadros. As Figuras 5.1 e 5.2 mostram as trajetórias sobrepostas às imagens das seqüências catadióptrica e perspectiva, respectivamente.

A matriz M_o foi calculada transformando um retângulo que contém a imagem do

espelho em um quadrado e centrando o sistema coordenado no centro deste quadrado. A partir de M_o é possível calcular F e b montando o Problema de Autovalor Polinomial.

Foram escolhidos 9 pontos correspondentes nas duas imagens para resolver o Problema de Autovalor Polinomial. O resultado da solução encontrada pode ser visto nas Figuras 5.3 e 5.4, com alguns pontos e as linhas epipolares correspondentes. A Figura 5.5 mostra o plano de imagem expandido, para que seja possível observar o formato das cônicas.

A partir da geometria epipolar e das trajetórias é possível montar os mapas de votos. Foram gerados os dois mapas de votos: de cruzamentos de retas epipolares com trajetórias e de cruzamento de cônicas epipolares com trajetórias. O resultado do alinhamento de cada mapa de votos será mostrado nas subseções seguintes.

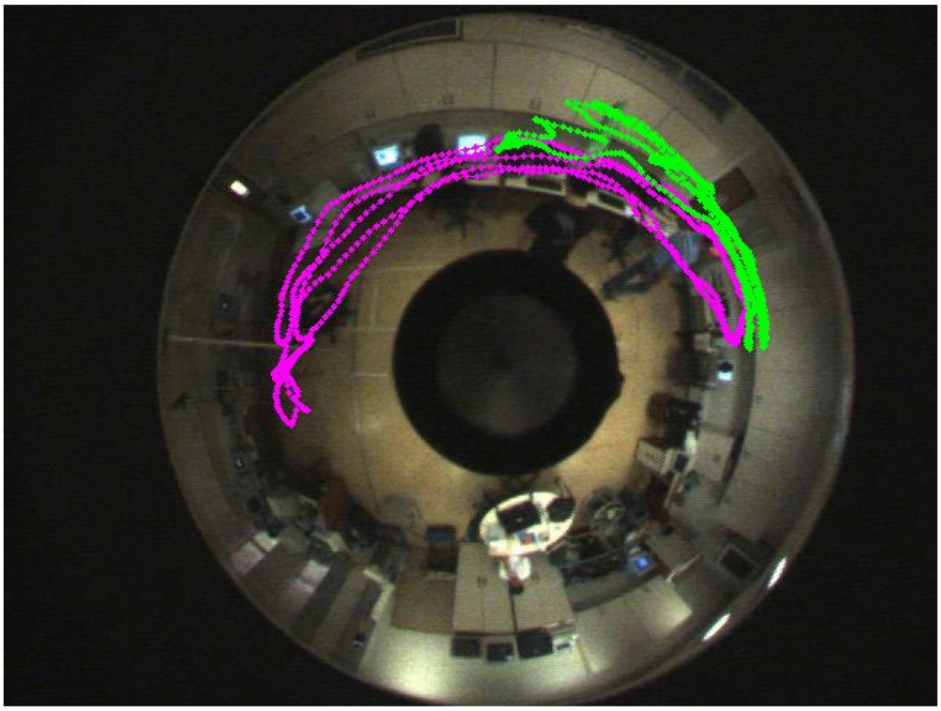


Figura 5.1: Imagem da câmera para-catadióptrica com as trajetórias encontradas.

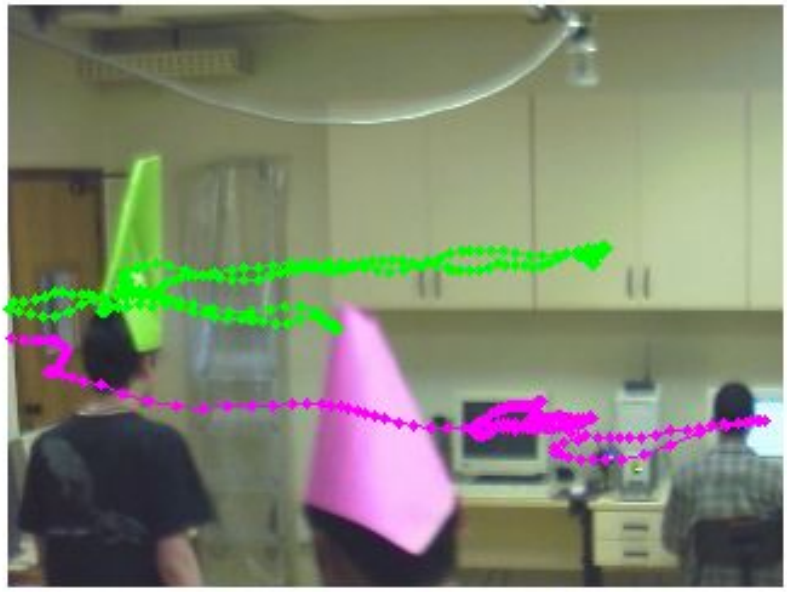


Figura 5.2: Imagem da câmera perspectiva com as trajetórias encontradas.

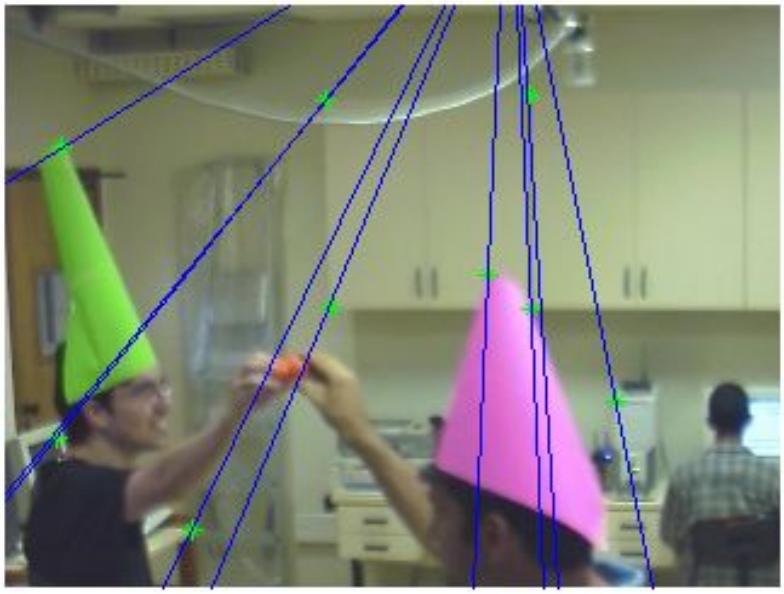


Figura 5.3: Calibração da câmera perspectiva.

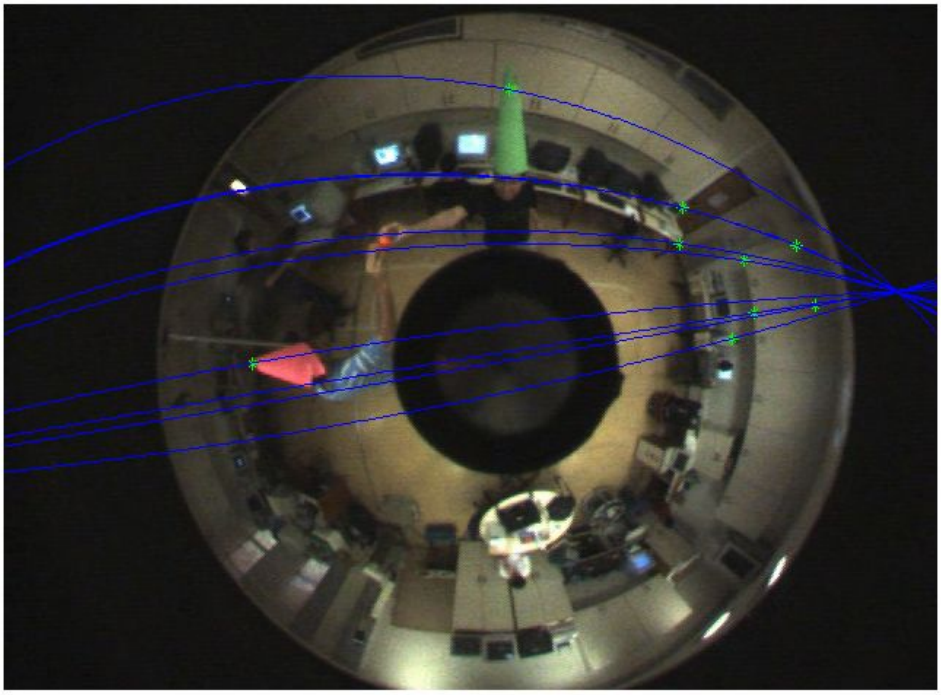


Figura 5.4: Calibração da câmera para-catadióptrica.

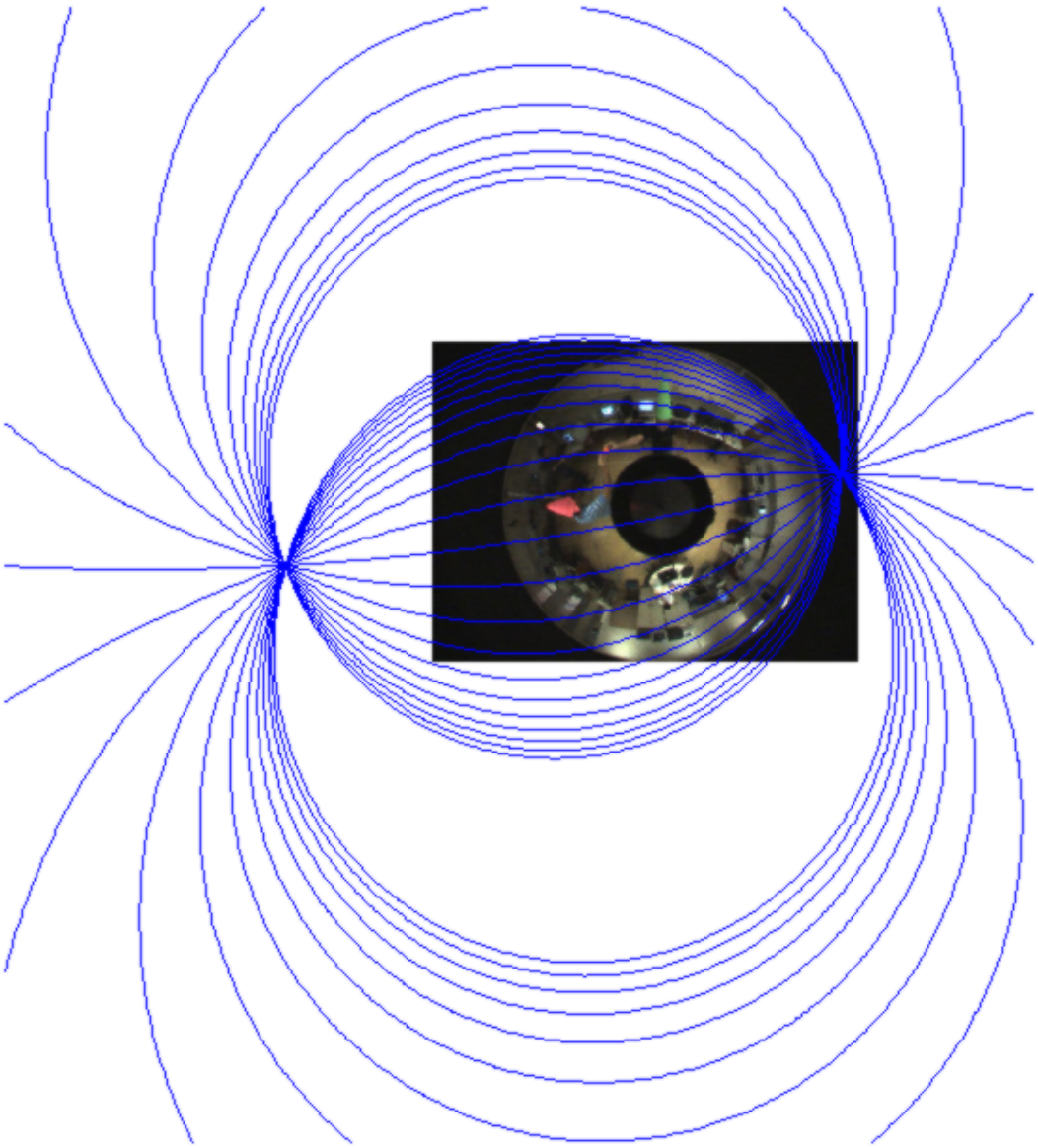


Figura 5.5: Cônicas epipolares geradas à partir de pontos igualmente espaçados na câmera perspectiva.

5.1.1 Retas epipolares x trajetórias

O primeiro mapa de votos gerado é o de retas epipolares cruzando com trajetórias na câmera perspectiva. O mapa de votos gerado é mostrado na Figura 5.6.

A reta encontrada tem a equação

$$t_c = 0.9801 t_p - 34.6409 \tag{5.2}$$

e o erro médio entre a reta da Equação 5.1 e a reta encontrada, levando-se em consideração os 768 quadros da câmera para-catadióptrica é de 3.85 quadros. A Figura 5.7 mostra algumas correspondências de quadros feitas a partir de 5.2.

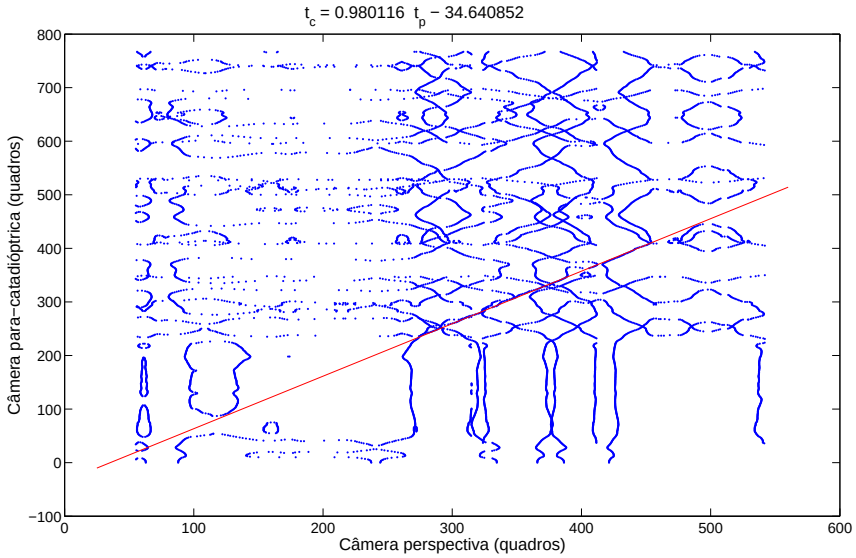


Figura 5.6: Mapa de votos gerado a partir das retas epipolares na câmera perspectiva.

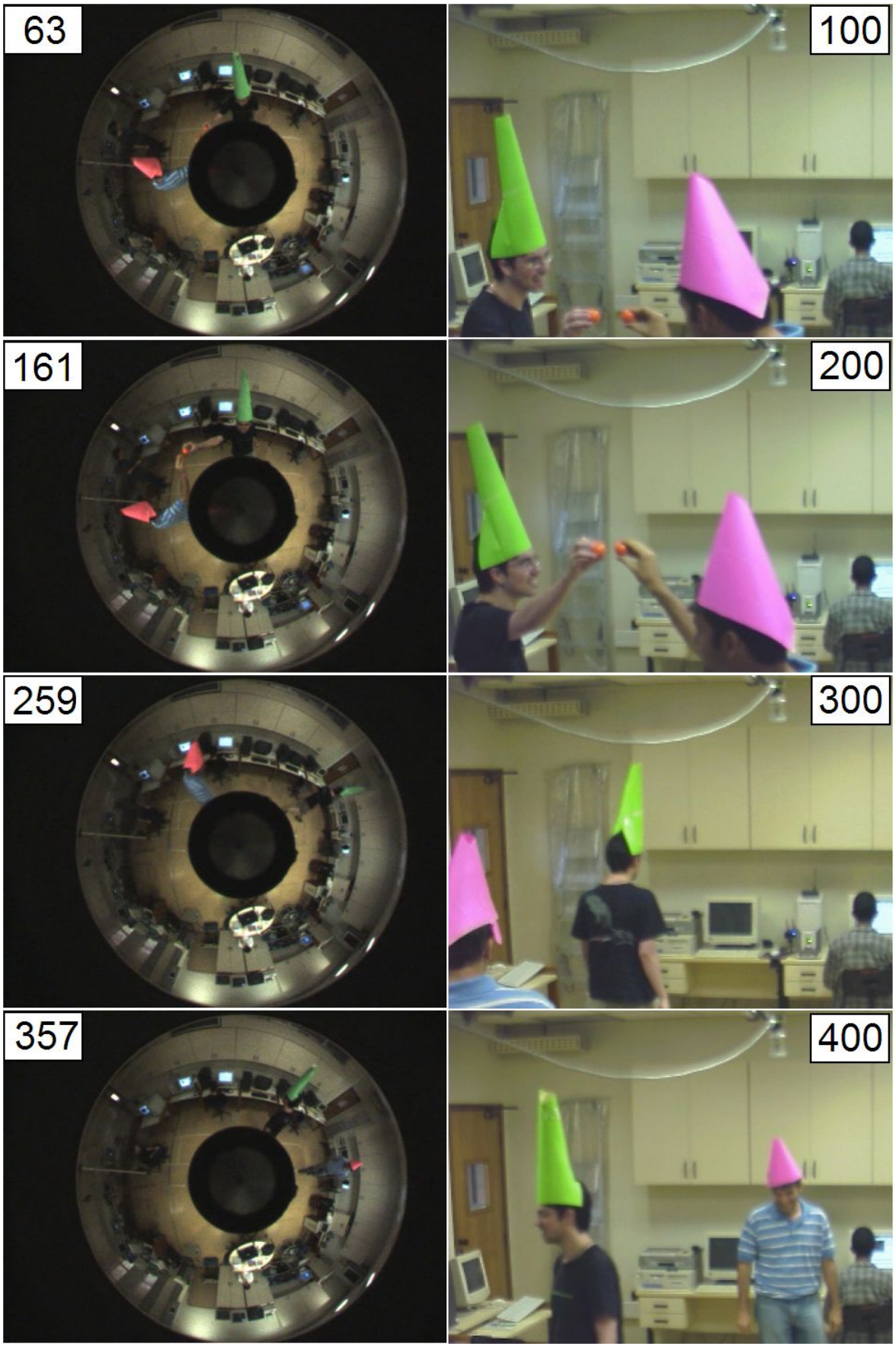


Figura 5.7: Correspondências de quadros à partir da Equação 5.2.

5.1.2 Cônicas epipolares x trajetórias

O segundo mapa de votos gerado é do cruzamento de cônicas epipolares com as trajetórias na imagem da câmera para-catadióptrica. A reta obtida tem a equação

$$t_c = 0.980447 t_p - 35.686815 \tag{5.3}$$

e o mapa de votos gerado pode ser visto na Figura 5.8.

O erro médio entre a reta calculada manualmente (Eq. 5.1) e a reta encontrada é de 3.83 quadros. O que demonstra que o método funciona também na imagem da câmera para-catadióptrica.

A Figura 5.9 mostra algumas correspondências de quadros feitas a partir de 5.3.

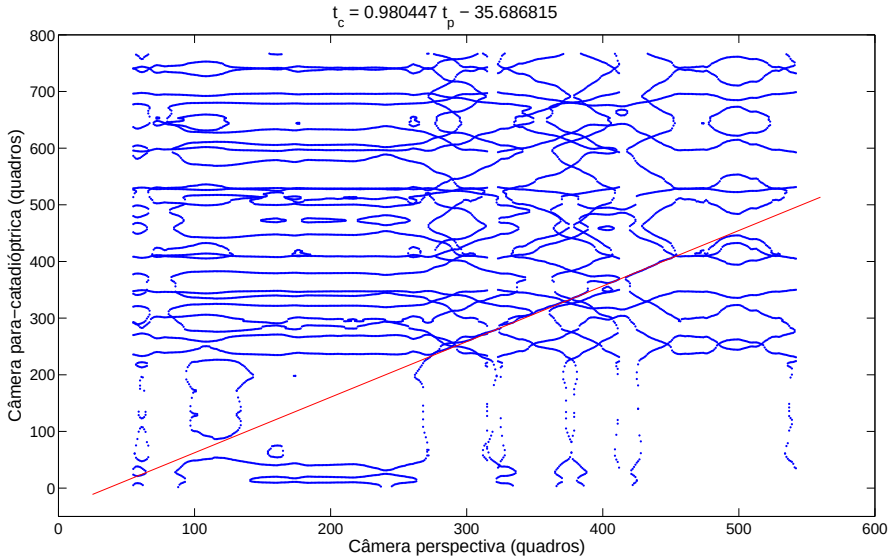


Figura 5.8: Mapa de votos gerado a partir das cônicas epipolares na câmera para-catadióptrica.

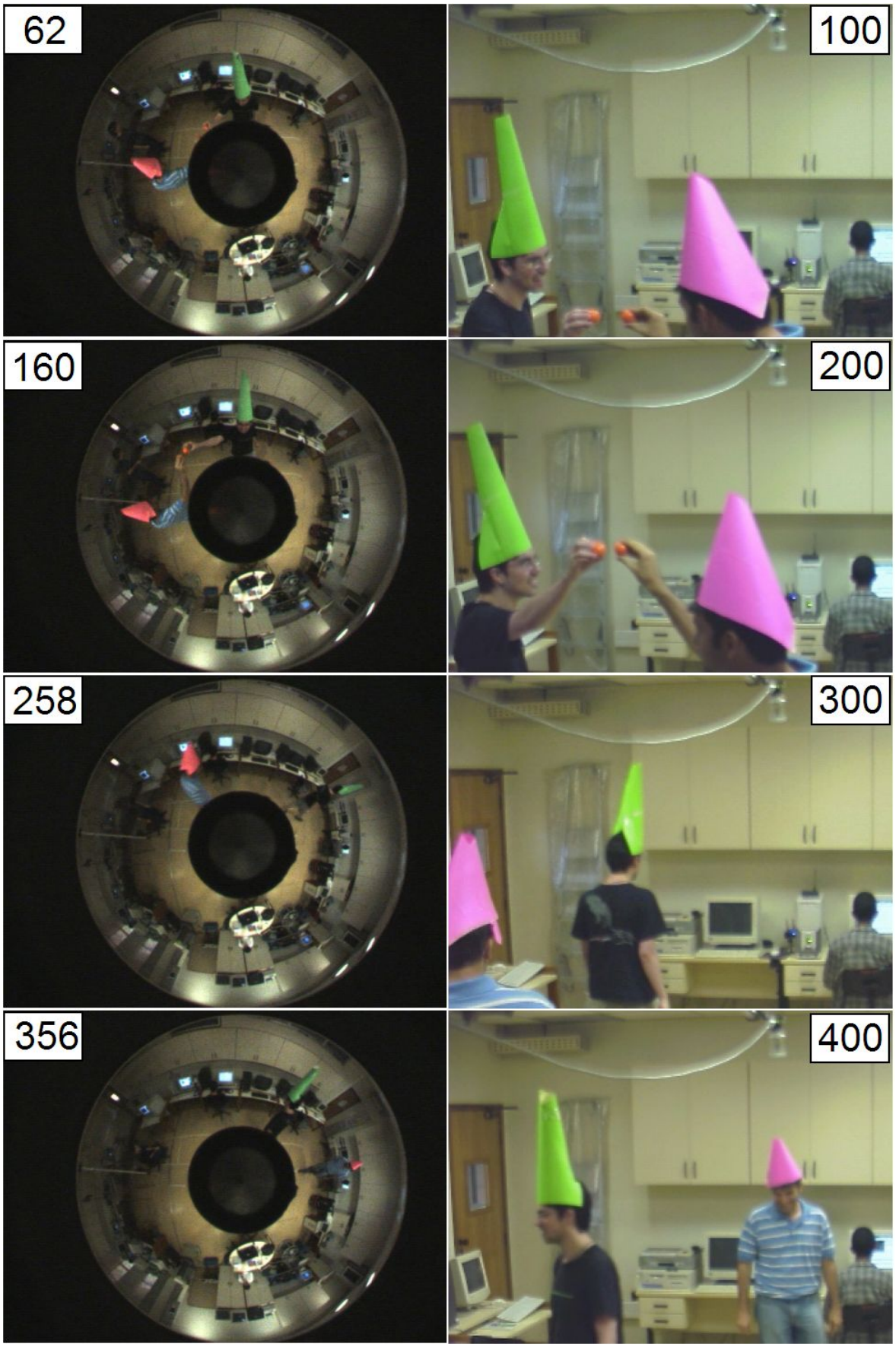


Figura 5.9: Correspondências de quadros à partir da Equação 5.3.

5.1.3 Discussão

Este experimento teve como objetivo principal demonstrar que o método funciona, utilizando seqüências de vídeo reais. Um erro não desprezível no alinhamento era esperado, pois o rastreador utilizado não conseguiu sempre garantir que a posição dos objetos rastreados fosse em seu centro de massa. Porém o resultado obtido foi um erro médio de menos de 4 quadros, e a diferença visual do resultado obtido no alinhamento é pequena.

Para se obter as retas o RANSAC foi executado 100 vezes para cada mapa de votos, e a melhor das 100 retas temporais foi escolhida. O mapa de votos gerado projetando cônicas epipolares na imagem da câmera para catadióptrica obteve um menor número de retas próximas do alinhamento real do que o mapa que foi gerado projetando retas epipolares na imagem da câmera perspectiva.

Uma consideração importante baseia-se no fato de que a imagem catadióptrica possui um volume maior de informação por pixel do que a imagem perspectiva, pois com aproximadamente o mesmo número de pixels a imagem catadióptrica captura informação de tudo à sua volta, inclusive toda a informação que a câmera perspectiva capta. Isso se traduz como uma menor precisão nas posições e trajetórias dos objetos rastreados do que a precisão possível com uma câmera perspectiva.

Segmentos de trajetória na câmera para-catadióptrica são considerados como retas. Se o movimento do objeto na cena é retilíneo, a trajetória dele na imagem é uma curva. O fato de que objetos que se movimentam em direção radial na imagem do espelho deveriam ter suas velocidade consideradas é importante, pois isso modela objetos acelerados no mundo. A Figura 5.10 demonstra ambos os casos, pode-se considerar uma linha de pontos da malha como a trajetória de um objeto se movendo com em linha reta e velocidade constante na cena. Fica claro que as trajetórias viram curvas, e o espaçamento entre os pontos é maior no centro, mostrando que deve ser considerada uma aceleração no segmento de trajetória da imagem para modelar objetos que se movimentam com velocidade constante na cena.

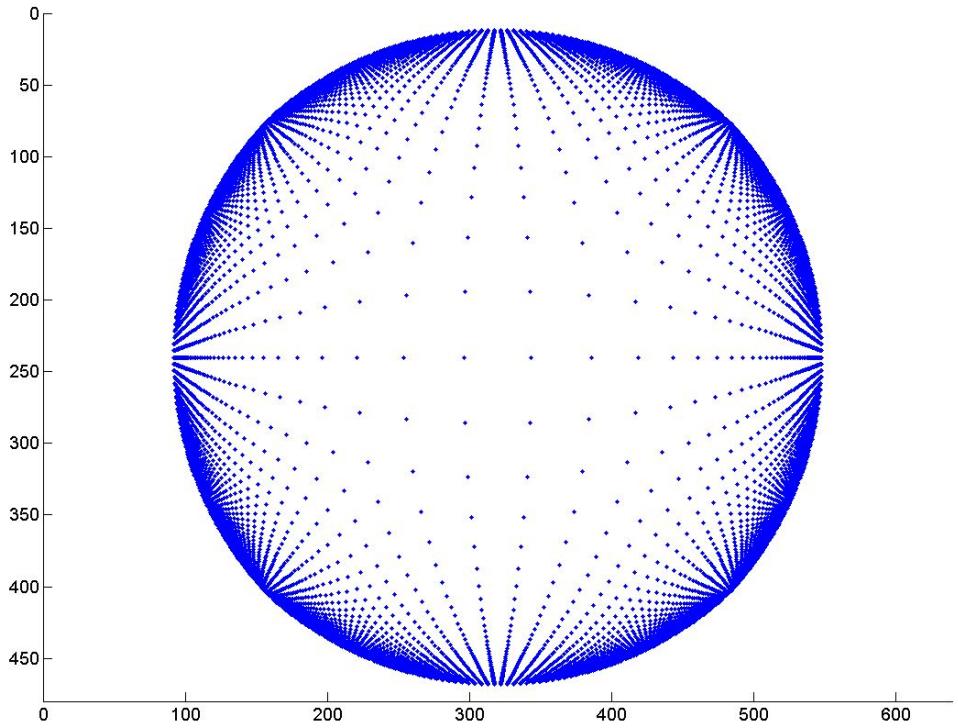


Figura 5.10: Malha de pontos imageada por uma câmera para-catadióptrica. Os pontos na cena são todos coplanares, em um plano perpendicular ao eixo do espelho, e estão igualmente espaçados.

Como este trabalho assume que os segmentos de trajetórias são retas, e a velocidade no segmento é constante, um segmento de trajetória na câmera perspectiva corresponde a um objeto que movimentou-se em linha reta na cena, e a uma velocidade constante. Já na câmera catadióptrica um segmento de trajetória corresponde a um objeto se movimentando em curva na cena, e acelerado. Na realidade os objetos se movimentam com uma mistura de curvas, retas, e velocidades variadas, e como os segmentos de trajetórias nas imagens são pequenos, ambos os modelos de movimentação de objetos na cena são aceitáveis.

5.2 Simulador

O experimento com o vídeo tem como finalidade mostrar que o método é aplicável utilizando seqüências reais. Porém, apenas este experimento não permite a obtenção de informações sobre o comportamento do método quando se varia algum parâmetro do sistema, como erros do rastreador ou número de pontos. Para se entender melhor este comportamento será utilizado um simulador. O simulador utilizado gera trajetórias tridimensionais e projeta os pontos destas trajetórias em modelos de câmeras.

O gerador de trajetórias é o mesmo utilizado por Pádua (2005), para cada objeto da cena ele sorteia uma variação angular não muito brusca no movimento e uma variação na velocidade de movimento. Objetos nascem e morrem durante a seqüência, mas sempre é garantido um número de objetos simultâneos.

Com as trajetórias 3D e os modelos das câmeras, são geradas as trajetórias nas imagens das câmeras, através da projeção dos pontos das trajetórias 3D. A matriz fundamental obtida não possui erros, pois os pontos correspondentes utilizados são conseguidos através de projeções de pontos do mundo.

Após as trajetórias serem projetadas nas câmeras, é possível inserir ruídos controlados nas trajetórias, simulando rastreadores reais. Para os experimentos com o simulador neste trabalho, serão variados os erros inseridos no rastreamento e o número de objetos simultâneos na cena. O número de objetos varia entre $1, 2, 4, 8, 16$ e 32 e os erros de rastreamento de 0 a 10 pixels.

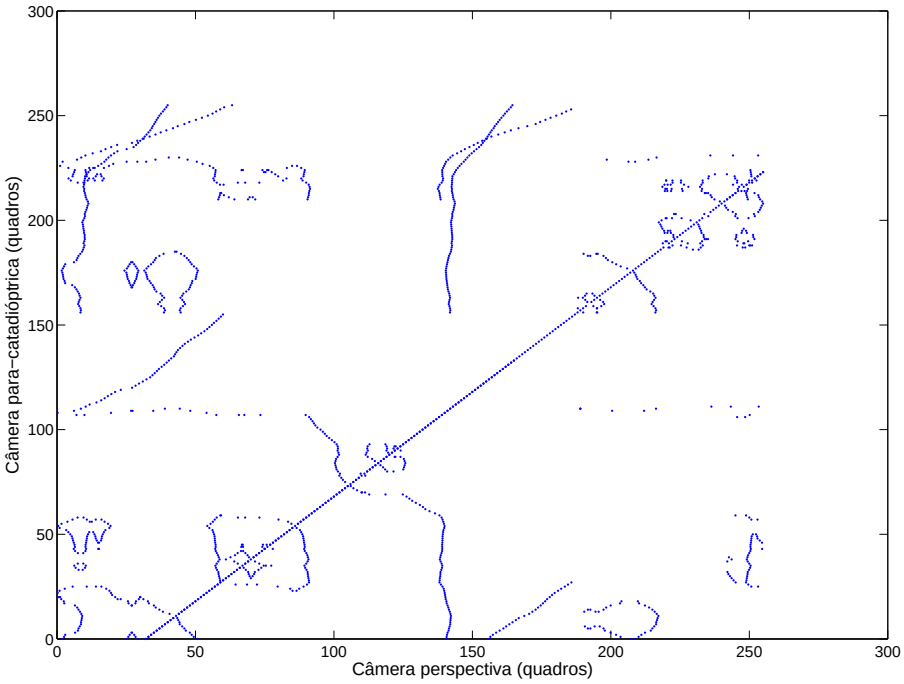
A avaliação da precisão do método é feita da seguinte forma: primeiro gera-se um grande número de retas temporais a partir de cada mapa de votos. Após isso, verifica-se a porcentagem de retas geradas que desvia de no máximo 5 quadros em média da reta temporal real, foram escolhidos 5 quadros por ser a incerteza do cálculo manual do alinhamento. Esta porcentagem de acertos é a medida utilizada para avaliar as execuções. A meta que se gostaria de atingir é 95% das retas geradas desviarem no máximo 5 quadros do alinhamento temporal real.

A primeira observação feita durante a execução dos experimentos é o grande au-

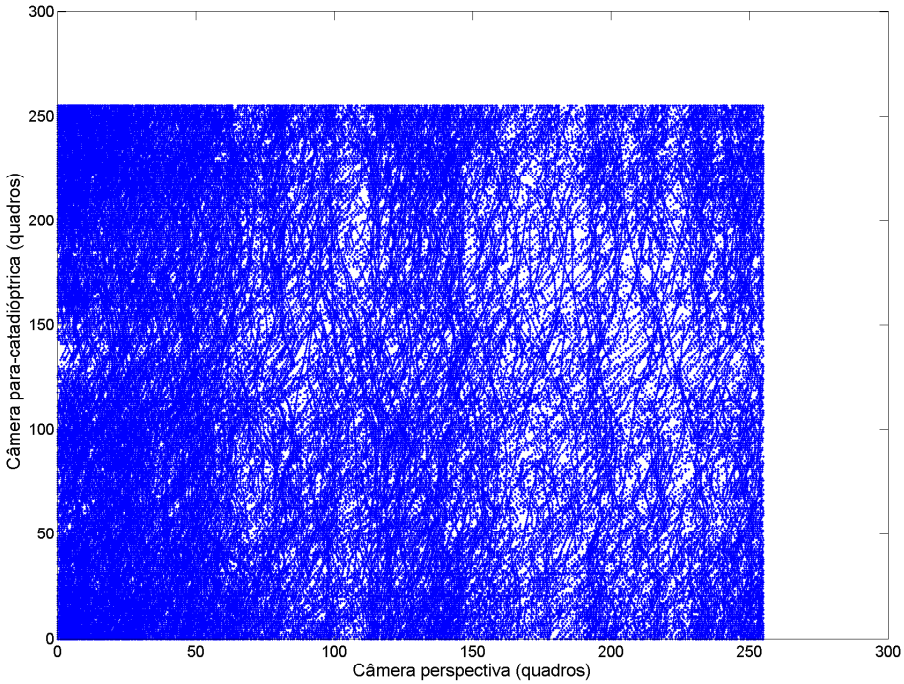
mento do número de votos com o aumento do número de objetos simultâneos rastreados. Com 16 objetos simultâneos o mapa de votos possui 58 mil votos, e com 32 objetos este número passa para 206 mil, cobrindo quase completamente o mapa. A Figuras 5.11 são exemplos de alguns dos mapas de votos obtidos. A Tabela 5.1 exibe o número de votos gerados de acordo com o número objetos simultâneos na cena.

O próximo passo foi verificar o que ocorre quando se aumenta o número de objetos simultâneos. A Figura 5.12 mostra o comportamento do aumento do número de objetos simultâneos para vários erros do rastreador. É perceptível que à medida que o erro aumenta, ter mais objetos simultâneos na cena diminui a porcentagem de acerto do método, até que finalmente se atinge um um erro grande o suficiente que impossibilita a utilização do método, como por exemplo o erro médio de 10 pixels na Figura 5.2.1. Em alguns casos o acerto de mapas com menos objetos é menor do que de mapas com mais objetos, isso ocorre pela formação de novas retas no mapa com a inserção de ruido. A explicação para os mapas com mais votos terem menor porcentagem de acerto também é devida à formação de retas concorrentes à reta de alinhamento verdadeira. E com mais votos no mapa, a chance destas retas surgirem é maior.

A última observação feita foi sobre o comportamento do método quando se aumenta o ruído. Na Figura 5.14 estão alguns gráficos com o resultado. Como era o esperado, à medida que o ruído aumenta, a taxa de acerto diminui. Algumas anomalias foram encontradas, principalmente para erros de 4 e 6 pixels, isso se deve à forma como o ruido transformou a trajetória, fazendo com que a reta temporal se espalhasse mais no mapa de votos (Figura 5.13), como a inserção de ruído é aleatória, situações como estas podem ocorrer. Para comprovar que isto foi gerado apenas por uma combinação não oportuna de erros no rastreador, novamente foi inserido um ruido aleatório de 6 pixels nas trajetórias de 1 objeto, e com estas novas trajetórias com ruídos de 6 pixels a taxa de acerto passou de 0% para 93%.



(a) 2 objetos

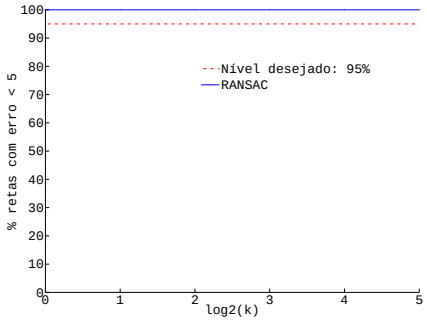


(b) 32 objetos

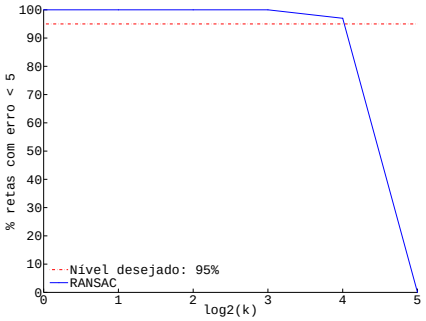
Figura 5.11: Dois mapas de votos obtidos

Objetos	1	2	4	8	16	32
Votos	188	1610	4740	17034	58792	206173

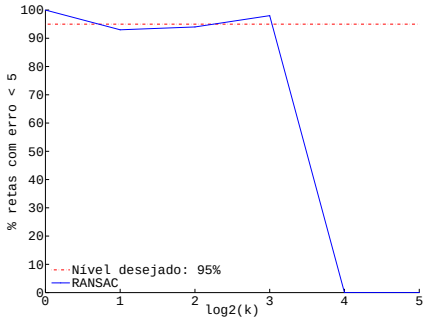
Tabela 5.1: Número de votos obtidos de acordo com número de objetos rastreados simultaneamente.



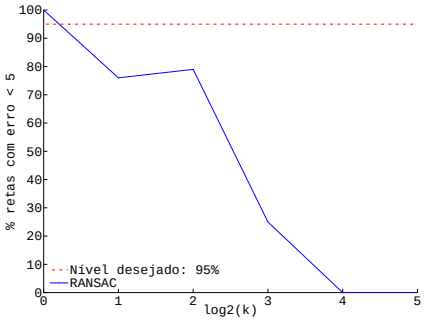
(a) Erro do rastreador 0 pixels



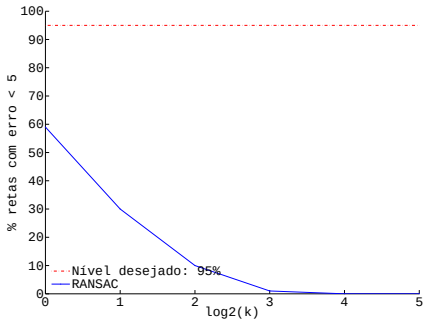
(b) Erro do rastreador 1 pixel



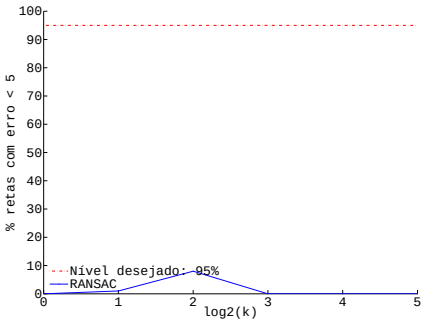
(c) Erro do rastreador 3 pixels



(d) Erro do rastreador 5 pixels



(e) Erro do rastreador 8 pixels



(f) Erro do rastreador 10 pixels

Figura 5.12: Impacto do aumento do número de objetos simultâneos para vários erros de rastreamento. k é o número de objetos rastreados simultânemante.

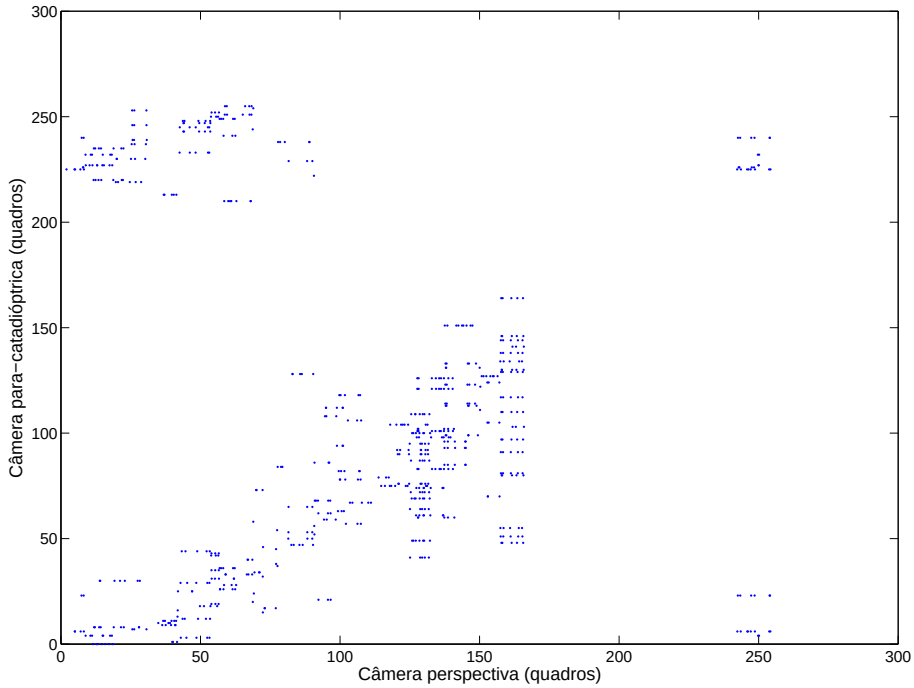
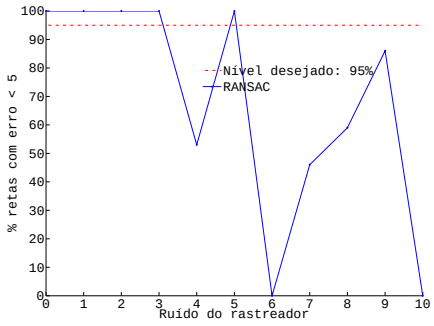
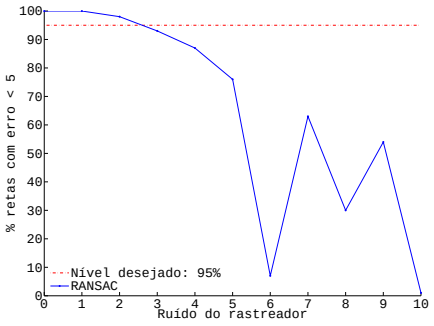


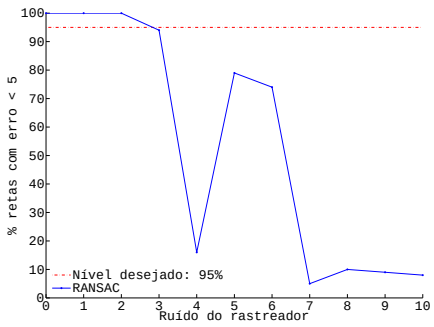
Figura 5.13: Mapa de votos para erro de 6 pixels e 1 objeto



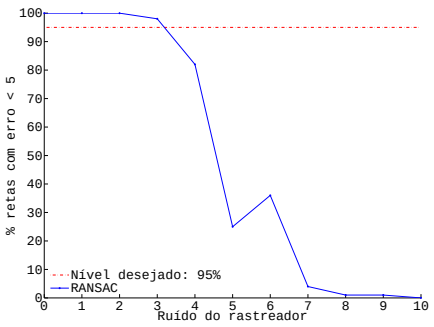
(a) 1 objeto



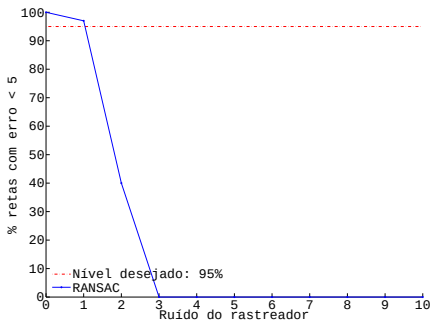
(b) 2 objetos



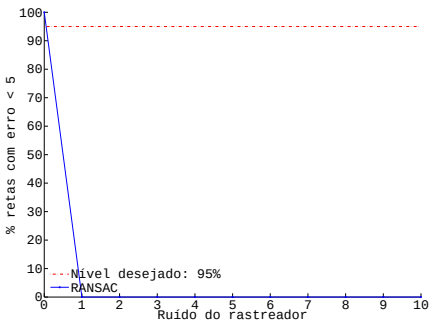
(c) 4 objetos



(d) 8 objetos



(e) 16 objetos



(f) 32 objetos

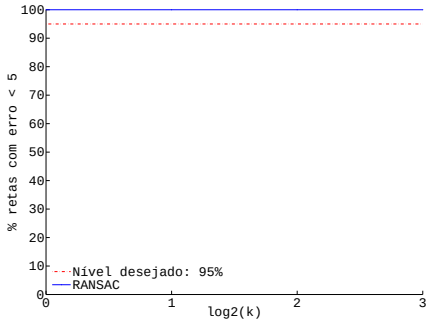
Figura 5.14: Impacto do aumento do erro do rastreador.

5.2.1 Comparação com cônicas epipolares

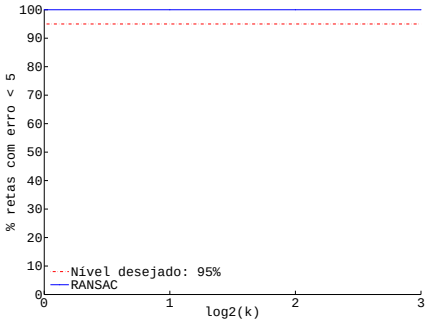
Todos os dados do simulador até o momento foram gerados a partir de retas epipolares na câmera perspectiva, nesta seção será realizada uma comparação do resultado das retas epipolares na imagem perspectiva com o resultado das cônicas epipolares na imagem para-catadióptrica.

As Figuras 5.15 e 5.16 exibem o comportamento do método quando se varia o número de objetos rastreados simultaneamente e o erro do rastreador. Percebe-se que os resultados são inferiores aos gerados na câmera perspectiva (Figura 5.12 e 5.14). Isto já era esperado, pois como foi colocado anteriormente, as trajetórias na câmera catadióptrica mapeiam objetos que se movem em curva e acelerados no mundo real, e nas trajetórias simuladas todos os objetos andam em linhas retas e velocidade constante a cada nova posição. Além disso, 1 pixel de erro na imagem catadióptrica equivale a uma área maior da cena do que na câmera perspectiva, pois a imagem da câmera catadióptrica tem aproximadamente o mesmo número de pixels ativos da imagem da câmera perspectiva, e a área imageada pela câmera perspectiva ocupa apenas uma parte da imagem da câmera catadióptrica.

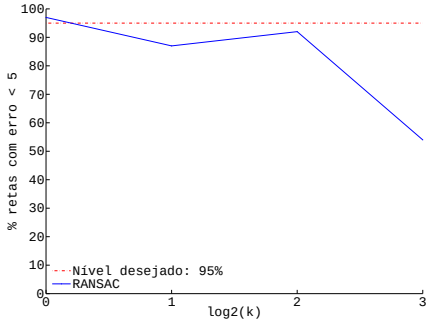
Apesar disso ainda é possível encontrar o desalinhamento temporal através dos votos gerados na câmera para-catadióptrica, mas o rastreador deve ser mais preciso do que quando se gera os votos na câmera perspectiva.



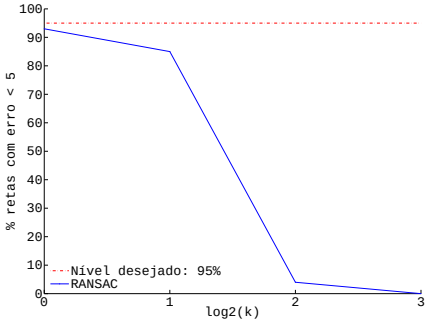
(a) Erro do rastreador 0



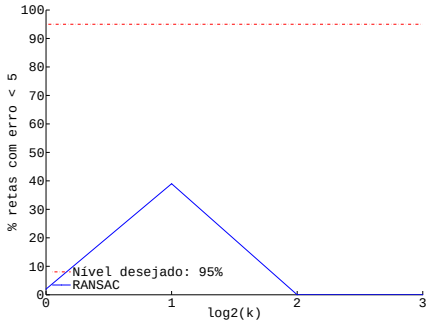
(b) Erro do rastreador 1



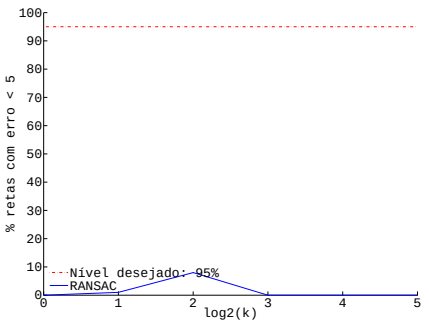
(c) Erro do rastreador 3



(d) Erro do rastreador 5

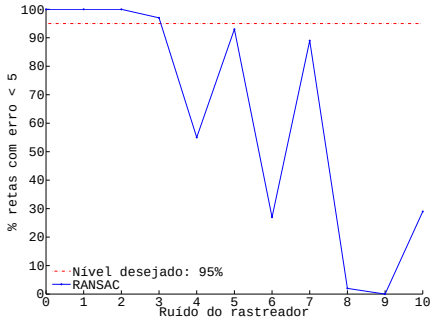


(e) Erro do rastreador 8

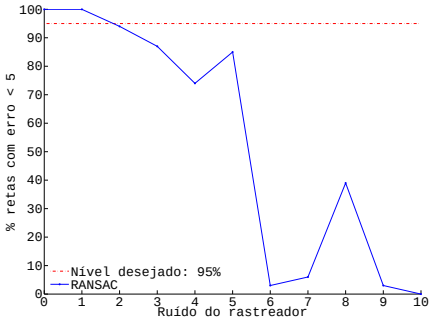


(f) Erro do rastreador 10

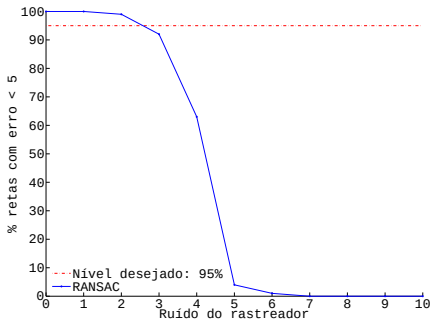
Figura 5.15: Impacto do aumento do número de objetos simultâneos para vários erros de rastreamento, gerado a partir das cônicas epipolares.



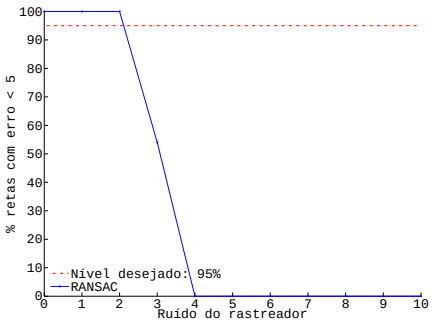
(a) 1 objeto



(b) 2 objetos



(c) 4 objetos



(d) 8 objetos

Figura 5.16: Impacto do aumento do erro do rastreador com votos gerados a partir das cônicas epipolares.

Capítulo 6

Conclusão

Este trabalho demonstrou que o método de alinhamento temporal permite alinhar temporalmente seqüências de vídeo gravadas com câmeras perspectivas e catadióptricas. Com este alinhamento é possível alinhar temporalmente vários vídeos de câmeras perspectivas que não possuem interseção entre os campos de visão. Basta adicionar uma câmera catadióptrica e alinhar todas as câmeras perspectivas com ela.

Uma vantagem do método de alinhamento é a de não precisar da correspondência dos objetos rastreados entre as duas câmeras, na realidade estas podem ser obtidas durante sua execução. Se nos votos do mapa de votos estiver também armazenado quais objetos rastreados geraram o voto, basta verificar os votos pertencentes ao conjunto consenso do RANSAC para se encontrar as correspondências de objetos entre as câmeras.

Foi mostrado que os votos gerados a partir de retas epipolares na câmera perspectiva levam a resultados melhores do método do que quando os votos são gerados a partir de cônicas epipolares na imagem catadióptrica. Isso se da principalmente devido ao fato da imagem catadióptrica agregar mais volume da cena por pixel do que a imagem da câmera perspectiva, esta imprecisão se agrava se as trajetórias dos objetos imageados estiverem muito próximas do centro ou da borda da imagem do espelho, que são as áreas onde existe maior acumulo de informações por pixel.

A modelagem de segmentos trajetórias na câmera catadióptrica foi definida da

mesma forma que na câmera perspectiva, como retas com velocidade constante. Porém esta modelagem é mapeada na cena como objetos se movimentando em curvas e com aceleração. Como o tempo entre os quadros de uma cena é muito curto e objetos reais não andam apenas em linha reta e com a mesma velocidade, essa modelagem também é válida, porém isto pode ter afetado os experimentos do simulador, já que todos os objetos do simulador possuem segmentos de trajetórias em linha reta e velocidade constante.

O método funciona com votos gerados por cônicas epipolares na imagem catadióptrica, isso significa que o método também funciona para duas câmeras catadióptricas, pois ambas são de projeção central, possuem geometria epipolar e suas linhas epipolares são cônicas. De fato, é possível alinhar temporalmente qualquer câmera que consiga gerar cônicas epipolares na imagem catadióptrica, basta gerar os mapas de votos à partir do cruzamento de cônicas epipolares com trajetórias, na câmera para-catadióptrica.

No método de alinhamento original de Pádua et al. a reta temporal obtida pelo RANSAC não era na verdade a reta temporal definitiva. Após se obter a reta temporal pelo RANSAC, um refinamento da reta e da matriz fundamental era efetuado. Este refinamento é um processo de otimização linear que aumenta consideravelmente a acurácia do método. Este trabalho se concentrou em mostrar como estender o método de alinhamento temporal para câmeras para-catadióptricas e perspectivas, não aplicando o refinamento nos experimentos executados, porém o refinamento também pode ser estendido para câmeras para-catadióptricas e perspectivas. A dedução do refinamento para câmeras para-catadióptricas e perspectivas se encontra no Apêndice A.

A acurácia do método depende da qualidade do rastreador utilizado, pois como foi visto, algumas configurações de erros no rastreamento podem levar a mapas de votos que não geram retas temporais próximas da solução real.

Outra problema do método é o ângulo de cruzamento entre as trajetórias e as linhas epipolares. Se este ângulo for pequeno, a acurácia do método será deteriorada, pois a linha epipolar ira cruzar com vários pixels do segmento de trajetória.

6.1 Direções futuras

A principal modificação que deveria ser implementada no método para melhorar sua porcentagem de acerto é a etapa de refinamento. Acredito que o refinamento melhoraria consideravelmente as retas epipolares obtidas pelo RANSAC se a matriz fundamental encontrada fosse refinada. Talvez a implementação do refinamento do parâmetro b também possa ser implementada, utilizando técnicas de otimização não linear.

Outra modificação interessante é a modelagem dos segmentos de trajetórias na câmera catadióptrica. Modelá-los para que os objetos tenham movimento retilíneo uniforme na cena para se comparar o resultado obtido com a modelagem feita neste trabalho.

Modificações para aumento da velocidade de execução também devem ser feitas. Atualmente o RANSAC compara todos os votos existentes com o modelo de reta gerado, 100 execuções completas do RANSAC em um CPU de 2Ghz e 2GB de memória demoram 5 horas para 200.000 votos. Uma subdivisão espacial dos votos, como uma *quadtree*, aceleraria consideravelmente a execução, pois grande parte dos votos seriam descartados ao verificar a interferência da *quadtree* com o modelo gerado com o RANSAC.

Apêndice A

Refinamento

A etapa de refinamento do método de Pádua et al. (2004); Pádua (2005) refina a matriz fundamental e os parâmetros da equação do alinhamento temporal simultaneamente através de sistemas lineares. Neste apêndice será mostrado como estender o refinamento entre câmeras perspectivas para câmeras para-catadióptricas e perspectivas. Uma explicação completa do processo de refinamento pode ser encontrada em Pádua (2005).

Seja p_{ci} a projeção de um ponto da cena no plano de imagem do sistema coordenado da câmera no quadro t_i. A projeção instantânea deste mesmo ponto da cena na câmera perspectiva, no quadro t_j pode ser parametrizada como:

$$p_{pj} = p_{pa} + (t_j - a)\frac{p_{pb} - p_{pa}}{t_b - t_a} \tag{A.1}$$

onde p_{pa} e p_{pb} são os extremos de um segmento linear que contém a posição p_{pj} e t_a e t_b seus respectivos quadros. Como p_{ci} e p_{pj} são pontos correspondentes, tem-se que:

$$p_{ci}^T F p_{pj} = 0. \tag{A.2}$$

Combinando as Equações A.1 e A.2 obtém-se

$$p_{ci}^T F \left\{ p_{pj} + (t_i - a)\frac{p_{pb} - p_{pa}}{t_b - t_a} \right\} = 0, \tag{A.3}$$

que pode ser escrita de uma forma mais compacta:

$$p_{ci}^T \{ F k t_j + F m \} = 0, \tag{A.4}$$

onde

$$k = \frac{p_{pb} - p_{pa}}{t_b - t_a}. \tag{A.5}$$

$$m = p_{pa} - t_a k. \tag{A.6}$$

Considerando que $t_j = \alpha t_i + \beta$, onde $\alpha = \hat{\alpha} + \Delta\alpha$ e $\beta = \hat{\beta} + \Delta\beta$, onde α e β são os parâmetros estimados após o refinamento, $\hat{\alpha}$ e $\hat{\beta}$ são a estimativa corrente e $\Delta\alpha$, $\Delta\beta$ são os termos do refinamento. Similarmente escrevendo $F = \hat{F} + \Delta F$ e substituindo na Equação A.6, ignorando os termos de segunda ordem, obtemos a seguinte restrição linear em $\Delta\alpha$, $\Delta\beta$ e ΔF:

$$p_{ci} \left\{ t_i \hat{F} k \Delta\alpha + \hat{F} k \Delta\beta + \Delta F h \right\} = -p_{ci}^T \hat{F} h, \tag{A.7}$$

onde

$$h = (t_i \hat{\alpha} + \hat{\beta}) k + m. \tag{A.8}$$

A Equação A.7 é a restrição linear do refinamento se não refinarmos o parâmetro b da câmera para-catadióptrica. Ela pode ser reescrita como um produto de dois vetores, um vetor linha de 11 elementos que contém apenas coeficientes e um vetor coluna de 11 elementos que contém as 9 incógnitas de ΔF seguidas de $\Delta\alpha$ e $\Delta\beta$. Restrições lineares nesta forma, geradas por interseções temporalmente consistentes ($t_a < \alpha t_i + \beta < t_b$) entre linhas epipolares e trajetórias geram um sistema linear

sobre-determinado $A_{n\times 11}x_{11\times 1} = b_{1\times 11}$.

Tratando o parâmetro b do espelho como$b = \hat{b} + \Delta b$, o ponto p_{ci} se transforma em $\hat{p_{ci}} + \Delta p_{ci}$, e a Equação A.7 é reescrita como:

$$(\hat{p_{ci}} + \Delta p_{ci})^T \left\{ t_i \hat{F} k \Delta\alpha + \hat{F} k \Delta\beta + \Delta F h \right\} = -(\hat{p_{ci}} + \Delta p_{ci})^T \hat{F} h, \tag{A.9}$$

onde

$$\hat{p_{ci}} = \begin{bmatrix} x_{ci}^2 \\ y_{ci}^2 \\ \frac{x_{ci}^2 + y_{ci}^2}{2\hat{b}} - \frac{\hat{b}}{2} \end{bmatrix}, \tag{A.10}$$

$$\Delta p_{ci} = \begin{bmatrix} 0 \\ 0 \\ \frac{\Delta b}{\hat{b}^2} - \frac{\Delta b}{2} \end{bmatrix}. \tag{A.11}$$

Desenvolvendo a Equação A.9 tem-se:

$$\hat{p}_{ci} \left\{ t_i \hat{F} k \Delta\alpha + \hat{F} k \Delta\beta + \Delta F h \right\} + \Delta p_{ci} \hat{F} h = -\hat{p}_{ci} \hat{F} h, \tag{A.12}$$

que é a restrição do refinamento considerando-se $b = \hat{b} + \Delta b$, e o sistema de refinamento pode ser montado da mesma forma que na Equação A.7, neste caso o sistema terá 12 incógnitas (as 11 anteriores adicionadas de Δb), sendo da forma $A_{n\times 12}x_{12\times 1} = b_{1\times 12}$

Referências Bibliográficas

Caspi, Y. e Irani, M. (2000). A step towards sequence-to-sequence alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 682–689, Hilton Head Island, South Carolina. IEEE.

Caspi, Y. e Irani, M. (2001). Alignment of non-overlapping sequences. In *Proc. IEEE International Conference on Computer Vision*.

Caspi, Y.; Simakov, D. e Irani, M. (2002). Feature-based sequence-to-sequence matching. In *VAMODS (Vision and Model ling of Dynamic Scenes) workshop with ECCV*.

Fischler, B. e Bolles, R. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

Jepson, A.; Fleet, D. e El-Maraghi, T. (2003). Robust on-line appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1296–1311.

Lee, L.; Romano, R. e Stein, G. (2000). Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Transactions on Pattern Analysis and Machine Intel ligence (PAMI)*, 22:758–767.

Micusik, B.; e Pajdla, T. (2002). Para-catadioptric camera auto-calibration from epipolar geometry. *International Journal of Computer Vision*, 49(1):23–37.

Nayar, S. K. (1997). Catadioptric omnidirectional camera. In *Proc. of IEEE Computer Vision and Pattern Recognition Conference*.

Pádua, F. L. C. (2005). *Alinhamento Espaço-Temp oral de Sequências de Vídeo Capturadas a Partir de Múltiplos Pontos de Vista*. PhD thesis, Universidade Federal de Minas Gerais.

Pádua, F. L. C.; Carceroni, R. L.; dos Santos, G. A. M. R. e Kutulakos, K. N. (2004). Sequence-to-sequence alignment. In *Proc. of IEEE Computer Vision and Pattern Recognition Conference*.

Rao, C.; Gritai, A.; Shah, M. e Syeda-Mahmood, T. (2003). View-invariant alignment and matching of video sequence. In *Proc. of IEEE International Conference on Computer Vision*.

Stein, G. (1998). Tracking from multiple view points: Self-calibration of space and time. In *DARPA Image Understanding Workshop*, pp. 521–527.

Svoboda, T. e Pajdla, T. (2002). Epipolar geometry for central catadioptric cameras. *International Journal of Computer Vision*, 49(1):23–37.

Wolf, L. e Zomet, A. (2002a). Correspondence-free synchronization and reconstruction in a non-rigid scene. In *Workshop on Vision and Model ling of Dynamic Scenes*.

Wolf, L. e Zomet, A. (2002b). Sequence-to-sequence self calibration. In *Proc. of European Conference on Computer Vision*.