MÉTRICA OBJETIVA PARA AVALIAÇÃO DO CONFORTO NA VISUALIZAÇÃO DE VÍDEOS ESTEREOSCÓPICOS.

Marcelo de Azevedo Miguel

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Eduardo Antônio Barros da Silva

Rio de Janeiro
Junho de 2012

MÉTRICA OBJETIVA PARA AVALIAÇÃO DO CONFORTO NA VISUALIZAÇÃO DE VÍDEOS ESTEREOSCÓPICOS.

Marcelo de Azevedo Miguel

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Cláudio Rosito Jung, Ph.D.

Prof. Eduardo Antônio Barros da Silva, Ph.D.

Prof. José Gabriel Rodriguez Carneiro Gomes, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2012

Miguel, Marcelo de Azevedo

Métrica objetiva para avaliação do conforto na visualização de vídeos estereoscópicos./Marcelo de Azevedo Miguel. – Rio de Janeiro: UFRJ/COPPE, 2012.

X, 38 p.: il.; 29, 7cm.

Orientador: Eduardo Antônio Barros da Silva

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2012.

Referências Bibliográficas: p. 35 – 38.

1. Qualidade. 2. Estereoscopia. 3. 3D. I. Silva, Eduardo Antônio Barros da. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

À minha família, em especial à memória de Levy de Azevedo e Clóvis Geraldino.

Agradecimentos

Gostaria de agradecer a todos que me apoiaram no desenvolvimento desse trabalho. Em especial:

- Ao Eduardo A. B. da Silva pela orientação desde o momento em que decidi iniciar os estudos para o mestrado,
- Aos colegas do DOSRF, da TV Globo, pelo apoio dado para que eu pudesse conciliar os horários de trabalho e de estudo.
- À minha esposa Lilian, pelo carinho e paciência, pelas horas de lazer que deixamos de lado para que eu pudesse concluir esse trabalho.
- À minha família, por me guiar desde o início pelo o caminho de conquistas que me trouxe até aqui.
- Aos professores Cláudio e José Gabriel por aceitarem o convite para participar da banca.
- Aos demais professores que em algum momento fizeram parte de minha formação.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MÉTRICA OBJETIVA PARA AVALIAÇÃO DO CONFORTO NA VISUALIZAÇÃO DE VÍDEOS ESTEREOSCÓPICOS.

Marcelo de Azevedo Miguel

Junho/2012

Orientador: Eduardo Antônio Barros da Silva

Programa: Engenharia Elétrica

Apresenta-se, nesta tese um método automático objetivo para avaliação do conforto na visualização de vídeos estereoscópicos baseado na geometria da visão estéreo, visando auxiliar o ajuste do conjunto de câmeras para captação do conteúdo.

Utilizando-se da técnica SIFT (scale-invariant feature transform), determinam-se pontos homólogos entre as vistas esquerda e direita de cada vídeo. A geometria da captação e da exibição do conteúdo estereoscópico é estudada para a elaboração de um método de classificação desses pontos.

A classificação dos pontos homólogos é estendida para o restante do quadro do vídeo através da segmentação dos quadros, obtendo-se qual a proporção em cada cena de regiões exibidas que causam desconforto ao observador.

O método é validado através de uma base de vídeos estereoscópicos da EPFL[1], e o resultado é comparado com o do método proposto por MITTAL *et al.* [2].

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

METHOD TO EVALUATE THE QUALITY OF 3D EXPERIENCE IN STEREOSCOPIC IMAGES AND VIDEOS.

Marcelo de Azevedo Miguel

June/2012

Advisor: Eduardo Antônio Barros da Silva

Department: Electrical Engineering

In this work, we present a method for automatically assessing the comfort associated with viewing stereoscopic videos based on the stereo vision geometry, to assist the camera stereo rig adjustment.

Usign SIFT (scale-invariant feature transform), it finds homologous points for left and right views of each stereo video. The geometry of both the camera and the stereo vision are studied in order to elaborate a method to classify those points.

The points' classes are extended to the rest of the video frame by the frame segmentation, obtaining the proportion of each scene which causes eye strain to the observer.

The method is validated with a stereo video database made by EPFL[1], and the results are compared to the results of the method proposed by MITTAL *et al.* [2].

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Na última década o desenvolvimento tecnológico da indústria eletrônica permitiu o advento e comercialização de televisores capazes de reproduzir conteúdo estereoscópico a preços próximos aos dos televisores convencionais [4]. Houve também um substancial aumento na quantidade de salas de cinema que dispõem desse tipo de tecnologia.

Uma das maiores preocupações dos produtores de conteúdo é a de permitir aos usuários a melhor experiência tri-dimensional(3D) possível, sem causar incômodos como enjôos, dores de cabeça ou tonturas [5][6][7].

A criação de conteúdo para o cinema e televisão tem se desenvolvido rapidamente, com cada vez mais filmes lançados em 3D. Uma das dificuldades da produção de conteúdo estereoscópico é o longo tempo necessário para realizar os ajustes do arranjo de câmeras para a captação desse conteúdo [7].

Além da dificuldade do ajuste, ainda existe dificuldade na avaliação da qualidade do efeito 3D durante o processo de captação, fazendo com que a pós-produção do vídeo tenha que incluir algumas correções no conteúdo visando evitar efeitos indesejáveis da estereoscopia [7].

A percepção 3D produzida por um vídeo estereoscópico é gerada a partir da exibição de 2 diferentes sinais de vídeo: um para o olho esquerdo e um para o olho direito. As diferenças entre os vídeos exibidos devem simular as diferenças percebidas naturalmente por conta da distância entre os olhos. É importante salientar que a percepção 3D só acontece se ao visualizar as 2 imagens, o observador seja capaz de fundí-las em algo inteligível, ou seja, que a convergência das imagens de um mesmo objeto retrate de maneira fidedigna a visualização real desse objeto [8].

Para que isso aconteça, é necessário que a convergência de pontos comuns em relação aos olhos esteja à frente do observador. Há de se observar que essa condição é dependente da distância do observador à tela de exibição e do tamanho da tela. Portanto, ao realizar a produção do conteúdo, é necessário considerar as possibilidades de exibição do conteúdo para um bom ajuste das câmeras.

Este estudo busca desenvolver uma métrica objetiva para avaliação de conforto de visualização de vídeos estereoscópicos. Ou seja, desenvolver um método com o objetivo de obter a “Quality of Experience"(QoE) de estímulos 3D. O método deve auxiliar no ajuste do afastamento do par de cameras do sistema estereoscópico para que o vídeo gravado seja o mais confortável possível para o observador humano, sem que sejam necessários testes subjetivos. Os resultados são comparados com os dos testes subjetivos reportados em [1] usando a mesma base de dados. Além disso, os resultados da métrica proposta serão comparados com os resultados obtidos em [2].

1.1 Organização da Dissertação

O Capítulo 2 mostrará a geometria da visão estereoscópica, sendo que primeiramente será apresentada a geometria de captação do conteúdo, comparando dois tipos diferentes de arranjo. Em seguida será apresentada a geometria da exibição do conteúdo. Nessa seção serão apresentadas as restrições necessárias para uma percepção confortável do observador.

No Capítulo 3 será desenvolvido o método de análise para a avaliação da qualidade do vídeo estéreo, baseando-se na geometria apresentada no capítulo anterior.

O Capítulo 4 apresentará os testes realizados para validar o método. A base de vídeos utilizada será apresentada, e os resultados serão comparados o resultado obtido em [2].

Enfim, o Capítulo 5 resume a dissertação, apresentando as conclusões obtidas após os testes com a base, sugerindo possíveis aplicações para o método e propondo trabalhos futuros.

Capítulo 2

Geometria da visão Estereoscópica

A geometria da visão estereoscópica pode ser determinada como um mapeamento de um ponto (X,Y,Z) no espaço real, através do conjunto de lentes das duas câmeras para dois pontos ($X_{\mathrm{ccd_L}}$,$Y_{\mathrm{ccd_L}}$), ($X_{\mathrm{ccd_R}}$,$Y_{\mathrm{ccd_R}}$), no sensor de captura de cada uma delas. Esses pontos por sua vez são mapeados para pontos na tela de exibição: ($X_{\mathrm{s_L}}$,$Y_{\mathrm{s_L}}$), ($X_{\mathrm{s_L}}$,$Y_{\mathrm{s_R}}$). O observador, ao visualizar as imagens, reconstrói a cena no espaço tridimensional, em um ponto (X',Y',Z') [3].

2.1 Ajuste do Conjunto de Câmeras

Para a captação do vídeo estereoscópico, são utilizadas duas câmeras tradicionais montadas sobre um suporte (*rig*). Esse suporte permite o ajuste das câmeras de forma a produzir as imagens relativas aos olhos esquerdo e direito. A princípio, tende-se a imaginar que a distância entre os centros das câmeras deve ser de 64 mm, de forma a simular à distância interocular humana [9], porém isso não é válido. Dependendo do conjunto de lentes utilizados, a distância entre as câmeras deve variar para manter a percepção visual livre de problemas, como será exposto adiante.

As câmeras são dispostas no suporte de forma que os centros de ambas as câmeras estejam alinhados em relação ao eixo vertical. Existem dois tipos de arranjo quanto à rotação das câmeras: o arranjo convergente, no qual os eixos das câmeras convergem em um ponto à frente do arranjo, e arranjo paralelo, onde as câmeras são ajustadas com seus eixos ópticos paralelos [3][10].

A figura 2.1 mostra um arranjo convergente. As equações 2.1 a 2.4 representam o mapeamento de um ponto sobre os planos dos sensores das câmeras para esse arranjo, e são desenvolvidas por semelhança de triângulos [3], considerando o modelo *pinhole* de câmera.

$$X_{\mathrm{ccd_L}} = f.tan(\theta_L) \tag{2.1}$$

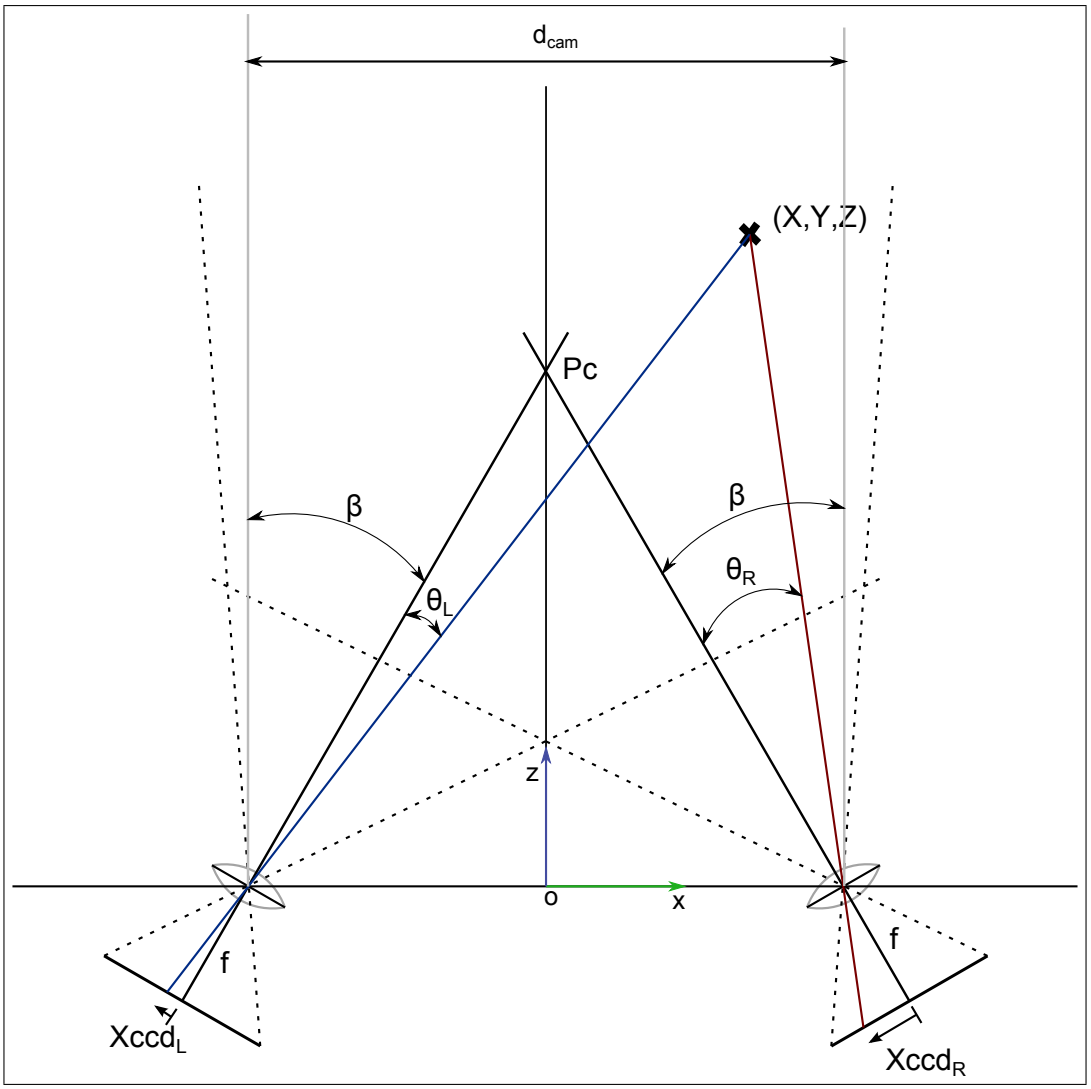

Figura 2.1: Arranjo convergente de câmeras, baseado no diagrama de [3]

$$tan(\theta_L + \beta) = \frac{X + \frac{d_{\text{cam}}}{2}}{Z} \tag{2.2}$$

$$\theta_L = arctan\left(\frac{2X + d_{\text{cam}}}{2Z}\right) - \beta \tag{2.3}$$

$$X_{\text{ccd}_\text{L}} = f.tan\left(arctan\left(\frac{2X + d_{\text{cam}}}{2Z}\right) - \beta\right) \tag{2.4}$$

analogamente, para θ_R:

$$X_{\text{ccd}_\text{R}} = f.tan(\theta_R) \tag{2.5}$$

$$tan(\beta - \theta_R) = \frac{\frac{d_{\text{cam}}}{2} - X}{Z} \tag{2.6}$$

$$\theta_R = -arctan\left(\frac{d_{\text{cam}} - 2X}{2Z}\right) + \beta \tag{2.7}$$

$$X_{\text{ccd}_\text{R}} = -f.tan\left(arctan\left(\frac{d_{\text{cam}} - 2X}{2Z}\right) - \beta\right) \tag{2.8}$$

Já no caso do arranjo de câmeras paralelo, pode ser necessário um deslocamento horizontal h do centro do sensor em relação ao centro óptico da lente, para aumentar a área de intersecção do campo de visão das câmeras. A figura 2.2 exibe o esquema relativo a esse arranjo, e as equações 2.9 e 2.12 mostram os valores relativos às projeções do objeto nos sensores [3].

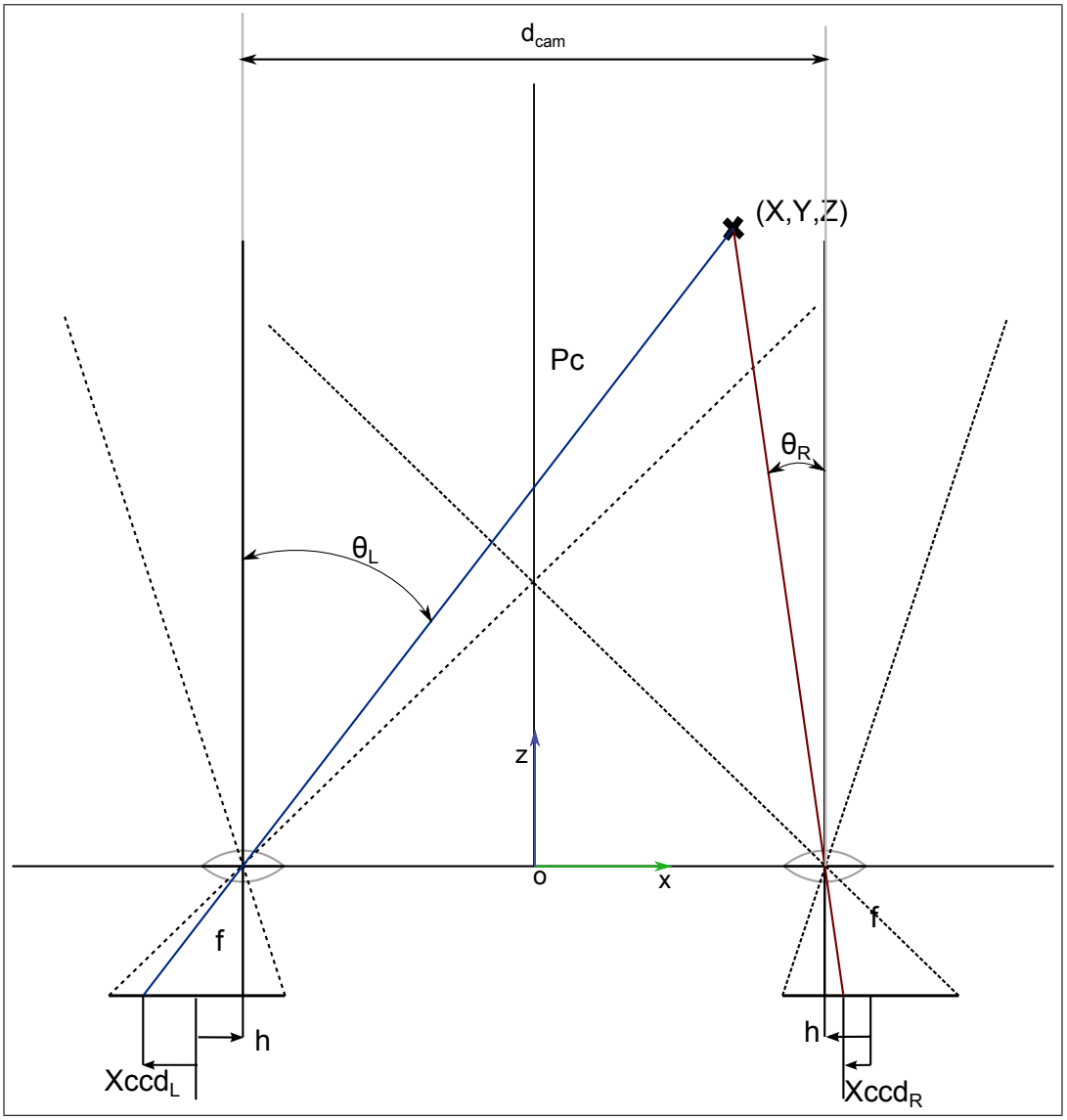

Figura 2.2: Arranjo paralelo de câmeras, baseado no diagrama de [3]

$$\frac{X_{\text{ccd}_\text{L}} + h}{f} = \frac{2X + d_{\text{cam}}}{2Z} \tag{2.9}$$

$$X_{\mathrm{ccd_L}} = f.\frac{2X + d_{\mathrm{cam}}}{2Z} - h \tag{2.10}$$

$$\frac{X_{\mathrm{ccd_R}} - h}{f} = \frac{2X - d_{\mathrm{cam}}}{2Z} \tag{2.11}$$

$$X_{\mathrm{ccd_R}} = f.\frac{2X - d_{\mathrm{cam}}}{2Z} + h \tag{2.12}$$

WOODS *et al.* [3], demonstram em seu artigo que o arranjo paralelo é vantajoso em relação ao convergente, uma vez que o arranjo convergente gera uma curvatura nos planos de profundidade, enquanto no arranjo paralelo, os planos não sofrem com essa distorção.

De fato, a relação entre a profundidade percebida pelo observador e a distância entre os pontos exibidos na tela é linear, conforme será demonstrado na próxima seção. Ao considerar diversos pontos a uma mesma distância Z_0 do arranjo de câmeras, é desejável que, para o observador, todos os pontos sejam percebidos a uma mesma distância.

Para o caso do arranjo convergente, tem-se que a distância entre os pontos exibidos na tela é:

$$P_{\mathrm{conv}} = f\left(tan\left(arctan\left(\frac{2X + d_{\mathrm{cam}}}{2Z}\right) - \beta\right) + tan\left(arctan\left(\frac{d_{\mathrm{cam}} - 2X}{2Z}\right) - \beta\right)\right) \tag{2.13}$$

Para o caso do arranjo paralelo, tem-se:

$$P_{\mathrm{par}} = f.\frac{2X + d_{\mathrm{cam}}}{2Z} - h - \left(f.\frac{2X - d_{\mathrm{cam}}}{2Z} + h\right) \tag{2.14}$$

$$P_{\mathrm{par}} = f.\frac{d_{\mathrm{cam}}}{Z} - 2h \tag{2.15}$$

A equação 2.13 mostra que a distância percebida pelo observador no caso das câmeras em arranjo convergente é dependente da posição do ponto analisado sobre o eixo X. Isso gera a distorção no plano de profundidade. No caso da equação 2.14, que representa o arranjo paralelo, vê-se que a distância entre os pontos projetados, e consequentemente a percepção de profundidade é dependente somente da distância Z entre o ponto e as câmeras, não gerando distorção.

2.2 Geometria da Percepção do Observador.

Para que uma pessoa assistindo a um vídeo estereoscópico tenha a sensação de tridimensionalidade, é necessário que o vídeo satisfaça a uma série de restrições,

conforme será demonstrado a seguir. A percepção de tridimensionalidade ocorre quando imagens deslocadas de um mesmo objeto são exibidas para cada olho [11]. Dependendo da profundidade do objeto na cena, o deslocamento será maior ou menor. Esse deslocamento é denominado disparidade (d_{imagem}), e é representado pela diferença entre a posição do ponto na imagem esquerda P_l e direita P_r, conforme a equação 2.16 [12][13] .

$$d_{\text{imagem}} = P_r - P_l \tag{2.16}$$

Nas duas imagens do mesmo objeto, os pontos que correspondem a uma mesma localização da cena são denominados pontos homólogos, ou pontos equivalentes. A fusão dessas imagens no cérebro gera a sensação de profundidade. As figuras 2.3 e 2.4 exibem exemplos da geometria da visão estéreo.

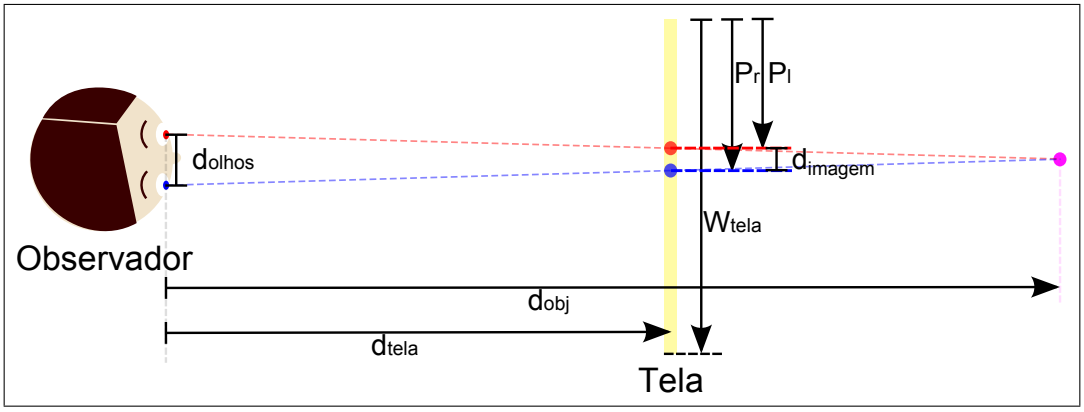

Figura 2.3: Dimensões para o cálculo da geometria da visão estéreo - objeto percebido atrás da tela

Considerando-se que os pontos equivalentes de um mesmo objeto estejam alinhados na vertical, podemos dividir a localização da percepção dos objetos em cena em 3 espaços distintos:

1. Sobre a tela de projeção ou exibição.
2. Atrás da tela de projeção ou exibição.
3. À frente da tela de projeção ou exibição.

Para que os objetos sejam percebidos atrás da tela de exibição, sua imagem direcionada ao olho esquerdo deve estar localizada à esquerda da imagem direcionada ao olho direito, nesse caso, vê-se pela equação 2.16 que a disparidade é positiva (figura 2.3). Já para o caso onde os objetos são percebidos à frente da tela de exibição, a imagem referente ao olho esquerdo está à direita da imagem exibida

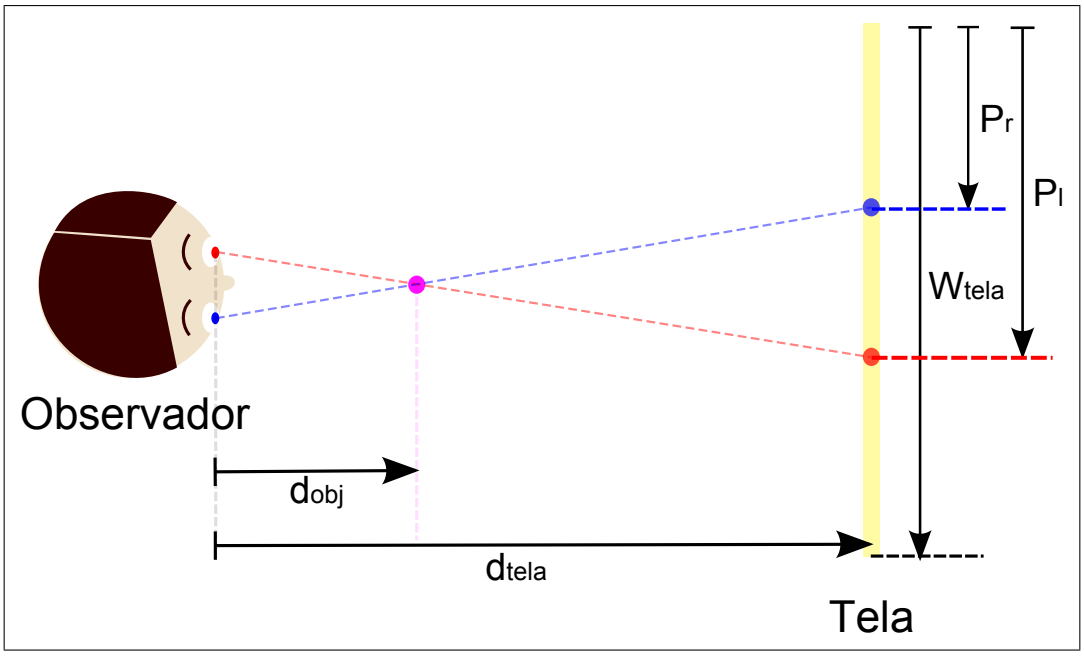

Figura 2.4: Dimensões para o cálculo da geometria da visão estéreo - objeto percebido à frente da tela

para o olho direito, logo a disparidade é negativa (figura 2.4). Quando o objeto é percebido sobre a tela, as imagens são coincidentes e a disparidade é nula [10].

Tendo como referência a figura 2.3, por semelhança de triângulos pode-se mostrar que a distância entre o observador e o objeto é dada pela equação 2.17 [3][13].

$$d_{\text{obj}} = \frac{d_{\text{olhos}}.d_{\text{tela}}}{d_{\text{olhos}} - d_{\text{imagem}}} \tag{2.17}$$

Onde d_{obj} é a distância do observador à imagem percebida do objeto, d_{olhos} é a distância entre os olhos do observador (cerca de 64 mm [9]), e d_{tela} é a distância entre o observador e a tela, conforme o esquema da figura 2.3.

Considerando a exibição de um vídeo estereoscópico com resolução de W_{pixels} pixels de largura, e a disparidade em pixels dada por d_{pixels} podemos verificar a influência do tamanho da tela de exibição (representado aqui pela largura da tela, W_{tela}) na percepção do vídeo.

$$d_{\text{imagem}} = \frac{W_{\text{tela}}.d_{\text{pixels}}}{W_{\text{pixels}}} \tag{2.18}$$

Substituindo a equação 2.18 em 2.17, vem:

$$d_{\text{obj}} = \frac{d_{\text{olhos}}.d_{\text{tela}}}{d_{\text{olhos}} - \frac{W_{\text{tela}}.d_{\text{pixels}}}{W_{\text{pixels}}}} \tag{2.19}$$

Além disso, nos casos em que um vídeo estereoscópico pré-gravado vai ser exibido,

d_{pixels} e W_{pixels} são fixos. Neste caso, a equação 2.19, quando d_{pixels} é positivo, nos mostra que há um limite no tamanho da tela para a percepção de distância do objeto, uma vez que o denominador da expressão tem que ser um número positivo. Se as dimensões da tela extrapolam esse limite, o observador tem uma cena que o cérebro não será capaz de reconstruir, pois os olhos teriam que ficar em posição divergente, conforme a figura 2.5 [14]. A inequação 2.20 expõe esse limite.

$$W_{tela} < \frac{d_{olhos}.W_{pixels}}{d_{pixels}} \tag{2.20}$$

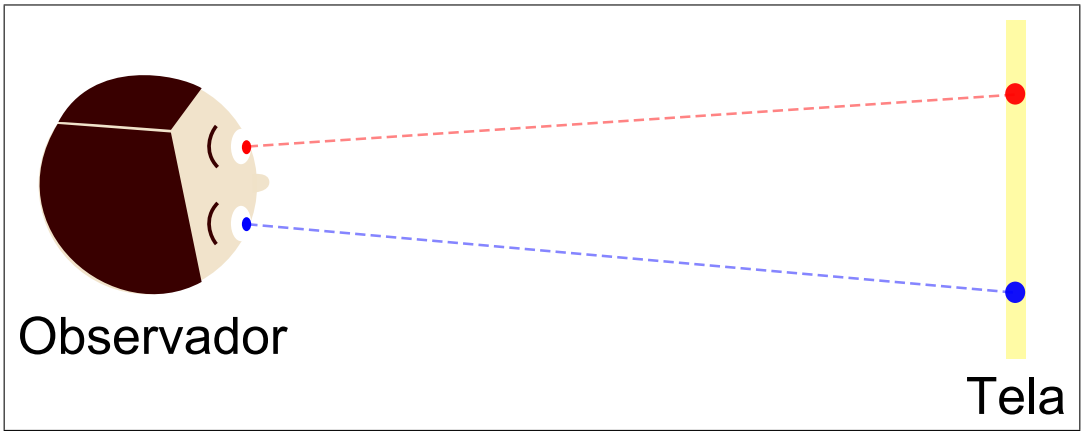

Figura 2.5: Visão Divergente. Se a tela cresce muito, pontos correspondentes podem ficar longe o suficiente para causar visão divergente, o que limita na prática o tamanho da tela de exibição.

Inversamente, podemos pensar na captação do filme. Levando em consideração a maior tela em que o filme será exibido, temos um valor máximo que podemos ter de disparidade no vídeo (ou seja, da distância máxima que podemos ter em pixels entre a imagem do olho esquerdo e do olho direito).

$$d_{pixels} < \frac{d_{olhos}.W_{pixels}}{W_{tela}} \tag{2.21}$$

Existem outras restrições em relação às imagens e como elas compõem a cena percebida pelo observador. Essas restrições estão relacionadas à percepção de profundidade na reconstrução da cena em relação à profundidade da cena real, e ao quanto os olhos conseguem convergir para observar uma imagem específica [15].

Para uma mesma distância de observação, podemos ter objetos com sua forma distorcida por um tamanho de tela inadequado, fazendo-os parecerem mais longos, caso a tela seja menor que o necessário, ou mais curtos, caso a tela seja maior. No primeiro caso, a profundidade da cena parece maior do que se pretendia mostrar, e no segundo, os objetos ficam achatados, como no cenário de uma peça de teatro [16]. A figura 2.6 mostra a variação da razão entre largura e profundidade de um mesmo objeto para tamanhos de tela diferentes.

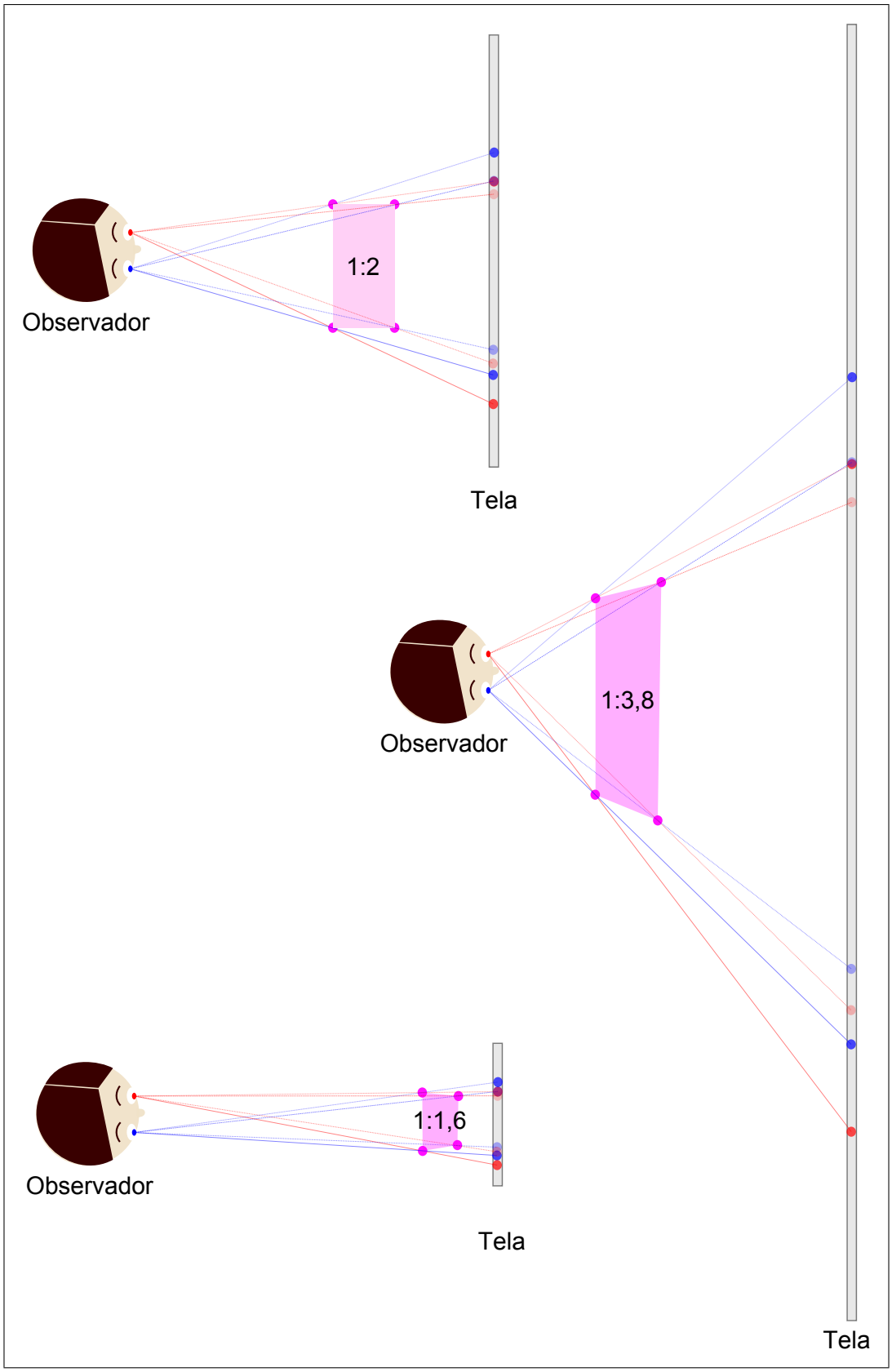

Figura 2.6: Diferença na percepção de profundidade de um objeto para diferentes tamanhos de tela.

Em certos casos há deslocamento vertical entre os pontos equivalentes das duas vistas. Esse tipo de distorção deve ser evitado, mantendo o alinhamento vertical das câmeras durante a captação. Quanto maior o deslocamento, pior é a percepção do observador. Outro fator que influencia negativamente a qualidade é a variância do deslocamento vertical entre diferentes pontos de uma mesma imagem, normalmente causado pela rotação das câmeras, ou ainda por distorções causadas pelas lentes [3].

Existem outros fatores citados por [17] que também influenciam no conforto do observador, dentre eles, a relação entre vergência, que é o quanto os olhos têm de convergir um em direção ao outro para observar os pontos equivalentes, e acomodação, que é o ajuste dos cristalinos para a focalização das imagens. Ao observar uma cena 3D real, as distâncias de vergência e a acomodação são iguais. Essa relação é denominada linha de Donder [18]. Porém nos vídeos estéreo, a acomodação é constante e fixa na tela de exibição e a vergência varia de acordo com o conteúdo da cena.

A definição de um limite de conforto para vergência e acomodação, denominado CVR (*comfortable viewing range*) foi alvo de diversos estudos [13][19][20][21][22], onde se verificou que há desconforto visual quando a distância percebida do objeto (d_{objeto}) está fora da região de foco dos olhos. Essa região é de 0,2 dioptria [21][22], onde dioptria é o inverso da distância focal em metros, assim as distâncias máxima e mínima de percepção do objeto são dadas pelas inequações 2.22 e 2.23, considerando que todas as distâncias são dadas em metros [18][23].

$$d_{obj} > \frac{1}{\frac{1}{d_{tela}} + 0,2m^{-1}} \tag{2.22}$$

$$d_{obj} < \begin{cases} \frac{1}{\frac{1}{d_{tela}} - 0,2m^{-1}} & \text{, se } d_{tela} < 5m \\ \infty & \text{, caso contrário} \end{cases} \tag{2.23}$$

Substituindo d_{obj} nas inequações 2.22 e 2.23 pela expressão da equação 2.19, e colocando em função de d_{pixels}, vêm:

$$d_{pixels} > -\frac{0,2m^{-1}.d_{olhos}.d_{tela}.W_{pixels}}{W_{tela}} \tag{2.24}$$

$$d_{pixels} < \frac{0,2m^{-1}.d_{olhos}.d_{tela}.W_{pixels}}{W_{tela}}, \text{ se } d_{tela} < 5m \tag{2.25}$$

As inequações 2.21, 2.24 e 2.25 estabelecem limites de conforto na observação das imagens. Esses limites serão usados na construção do método de avaliação objetiva proposto.

No capítulo 2 foi estudada a geometria da visão estereoscópica, tendo por objetivo principal obter os limites da visão confortável para o observador. A partir dos

resultados obtidos, no próximo capítulo será elaborado um método objetivo para a avaliação da qualidade de vídeos estéreo através da detecção da existência de regiões que não respeitem os limites encontrados.

Capítulo 3

Métrica Objetiva

A equação 2.19 mostra que a percepção de profundidade da estereoscopia durante a exibição do conteúdo está relacionada a duas grandezas: o tamanho da tela de exibição, e a distância do observador para a tela.

Durante o processo de captação das imagens, nem sempre será possível para a equipe técnica dispor de uma tela de tamanho suficiente para verificar durante a gravação se o produto gravado gera os efeitos de profundidade aos quais ele se propõe. Observar o mesmo vídeo em um monitor de 9 ou 14 polegadas (tamanhos comuns em estúdios de gravação) pode esconder determinados efeitos indesejáveis que aparecerão na exibição nas telas de 40 a 50 polegadas normalmente usadas nas residências, ou ainda nas telas de cinema.

O técnico responsável, conhecendo a geometria da exibição, deve ajustar a estrutura das câmeras e lentes de forma a garantir a inexistência de disparidades proibidas assim como de distorções verticais causadas pelo desalinhamento vertical das câmeras ou ainda da rotação das câmeras em relação ao plano da tela [7].

Em cenas de estúdio, o controle dos parâmetros de profundidade das cenas é relativamente simples, por se tratar de um espaço limitado. Porém em gravações de cenas externas e em eventos ao vivo, o controle desses parâmetros pode ser algo bastante complicado, pois muitas vezes o espaço não é limitado e os objetos da cena podem se mover livremente.

Uma forma de tentar minimizar a presença de conteúdo inadequado para a percepção estereoscópica é gerar um modelo de análise do vídeo, para identificar automaticamente as regiões que não obedeçam às restrições de exibição.

Se para cada pixel de cada objeto da cena captada para o olho direito for possível determinar o seu correspondente na cena captada para o olho esquerdo, usando a geometria da visão estereoscópica, será possível verificar se o vídeo possui ou não áreas ou objetos que gerariam problemas ao observador, dado que tem-se o tamanho da tela e a distância do observador.

Vale ressaltar que não se pode usar na análise quaisquer informações do arranjo

das câmeras, ou ainda das lentes utilizadas, para a solução do problema, visto que o objetivo do algoritmo é exatamente encontrar automaticamente falhas na elaboração desse arranjo.

Uma solução para o problema seria utilizar um método automático para encontrar as correspondências entre pixels dos objetos dos vídeos da esquerda e da direita. Porém, a maior parte dos métodos desenvolvidos utiliza informações relacionadas à geometria das câmeras e das lentes utilizadas, de maneira a reduzir o espaço de busca do problema, usando a geometria epipolar. Sem essa redução o custo computacional destes métodos os tornam proibitivos [24][25][26].

O método desenvolvido nesse estudo tem como objetivo verificar se existem falhas no arranjo de câmeras que gerem desconforto ao observador. Por isso, não devem ser usadas informações sobre o arranjo de câmera como dados de entrada para o método. Como os métodos tradicionais de correspondência estéreo necessitam dessa informação, não se pode utilizá-los diretamente, sendo necessário recorrer a outras formas de abordar o problema.

O método proposto para a elaboração da métrica de qualidade de experiência de vídeos estereoscópicos é baseado na busca de correspondências através da técnica SIFT (*Scale Invariant Feature Transform*). O SIFT irá verificar as semelhanças entre as imagens da vista esquerda e direita, marcando os pontos correspondentes (representados na figura 2.4 pelos pontos vermelho e azul na tela), depois uma técnica de segmentação e busca do vizinho mais próximo irá aproximar a correspondência SIFT, que é esparsa, para uma correspondência estéreo densa.

3.1 Scale Invariant Feature Transform

A técnica SIFT é um "método para extrair descritores distintos e invariantes de imagens que podem ser usados para executar uma correspondência confiável entre diferentes vistas de um mesmo objeto ou cena", conforme escreveu David G. Lowe, no artigo em que se descreve o método [27].

Ao extrair os descritores, as principais características apresentadas são invariância à escala e rotação, e correspondência robusta mesmo em situações que apresentem variação de iluminação, mudança de ponto da câmera, ruído, e variações de perspectiva.

Essa técnica se encaixa bem como método para o início da análise de qualidade. Isto acontece porque ela consegue extrair os descritores das imagens para determinar as correspondências estéreo de modo que as mesmas estejam de acordo com as premissas definidas na formulação do problema descrita anteriormente.

O método é composto por uma sequência de passos para que as operações de mais alta complexidade computacional somente sejam executadas em pontos chaves

aprovados em passos anteriores, estes computacionalmente mais simples.

Detecção de Pontos Chaves

Primeiramente, é realizada uma busca na imagem para detecção de áreas na imagem cujas características permaneçam inalteradas por mudanças de escala. Isso é obtido através de uma função denominada espaço de escala, buscando características estáveis nessas diferentes escalas [28].

A função espaço de escala $L(x, y, \sigma)$ é obtida a partir da convolução da imagem $I(x, y)$ com uma função gaussiana bidimensional $G(x, y, \sigma)$, definida pela equação 3.1:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.1}$$

LOWE [29] propôs um método para detectar os pontos mais estáveis do espaço de escalas através de função diferença de gaussianas convoluída com a imagem (DoG), que é calculada a partir da diferença entre duas escalas separadas por um fator k, conforme a equação 3.2.

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \star I(x, y) \tag{3.2}$$

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{3.3}$$

O método proposto por Lowe consiste em convoluir a imagem original incrementalmente com filtros gaussianos para produzir as imagens separadas por um fator de escala k. Lowe divide o espaço de escala em oitavas, ou seja, em cada oitava o valor de σ dobra. Cada oitava é dividida em um número s de intervalos, de forma que $k = 2^{1/s}$. Com isso, são necessárias $s + 3$ imagens filtradas para gerar, através da subtração das imagens adjacentes, as $s + 2$ imagens da diferença de gaussianas.

A imagem é então reduzida em resolução para metade da resolução original (subamostragem [30]), gerando uma nova oitava, onde o procedimento de geração da DoG é repetido.

A localização dos extremos (máximos e mínimos) é feita através de uma busca local. Cada pixel é comparado com 26 vizinhos (8 vizinhos na própria escala, e os 9 vizinhos das escalas anterior e posterior), e é marcado como ponto chave se for o maior ou o menor do conjunto de vizinhos.

Como o máximo ou mínimo da função $D(x, y, \sigma)$ podem não coincidir em posições exatas de pixel, é feita uma estimação do valor extremo em subpixels, visando a próxima etapa, onde o valor estimado do extremo será utilizado.

Para isso, é utilizada a aproximação de Taylor de segundo grau da função es-

paço de escala $D(x, y, \sigma)$, deslocada de forma que a origem localize-se no ponto de amostragem, ou seja, que $\mathbf{x}$ seja o deslocamento em relação ao ponto de amostragem.

$$D(\mathbf{x}) = D + \left(\frac{\partial D}{\partial \mathbf{x}}\right)^T \mathbf{x} + \frac{1}{2}\mathbf{x}^T \frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x} \tag{3.4}$$

$$\text{onde, } \mathbf{x} = \begin{bmatrix} x \\ y \\ \sigma \end{bmatrix} \text{ e } \frac{\partial D}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial D}{\partial x} \\ \frac{\partial D}{\partial y} \\ \frac{\partial D}{\partial \sigma} \end{bmatrix}$$

A localização em subpixel do extremo ($\hat{\mathbf{x}}$) é determinada pelo extremo da equação 3.4, bastando para isso, derivar a equação 3.4 em relação a $\mathbf{x}$, e igualar o resultado a zero.

$$\frac{\partial D}{\partial \mathbf{x}} + \frac{\partial^2 D}{\partial \mathbf{x}^2}\hat{\mathbf{x}} = 0 \tag{3.5}$$

Temos então que o offset da localização em subpixel do extremo em relação ao ponto $\mathbf{x}$ é:

$$\hat{\mathbf{x}} = -\left(\frac{\partial^2 D}{\partial \mathbf{x}^2}\right)^{-1} \frac{\partial D}{\partial \mathbf{x}} \tag{3.6}$$

As derivadas são calculadas numericamente pela diferença de pixels vizinhos, conforme proposto por [31]. Caso o offset seja maior que 0,5 em quaisquer de suas dimensões, o ponto de amostra é deslocado ao mais próximo do offset e o cálculo é repetido para esse novo ponto.

Eliminação dos Pontos Instáveis

Os pontos com pouco contraste ou que se encontram sobre arestas da imagem não geram bons descritores para encontrarmos seu correspondente. Com isso, esses pontos têm que ser descartados da análise. Assim, se a magnitude de um ponto chave está abaixo de um limite, ele é descartado.

A magnitude do ponto chave é dada pela equação 3.7 abaixo. Normalizando os valores dos pixels para o intevalo [0,1], os valores de $|D(\hat{\mathbf{x}})|$ menores que 0,03 são descartados.

$$D(\hat{\mathbf{x}}) = D + \frac{1}{2}\left(\frac{\partial D}{\partial \mathbf{x}}\right)^T \mathbf{x} \tag{3.7}$$

Ao usar o método da DoG, podem ser detectados pontos chaves sobre arestas da imagem. Isso não é desejável pois a localização de pontos chaves ao longo da borda não é bem determinada. Esses pontos devem ser detectados e eliminados. Para isso, Lowe propõe utilizar a matriz heissiana 2x2 calculada na posição e escala

do ponto-chave.

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{3.8}$$

$$D_{xx} = D(x+1, y, \sigma) - 2D(x, y, \sigma) + D(x-1, y, \sigma)$$

$$D_{yy} = D(x, y+1, \sigma) - 2D(x, y, \sigma) + D(x, y-1, \sigma)$$

$$D_{xy} = \frac{D(x-1, y-1, \sigma) + D(x+1, y-1, \sigma) - D(x+1, y-1, \sigma) - D(x-1, y+1, \sigma)}{4}$$

As magnitudes das curvaturas principais de D são proporcionais aos autovalores de H. Lowe emprega a abordagem de [32], utilizando a razão entre os autovalores para evitar a necessidade de calcular os autovalores explicitamente.

Sendo α o autovalor de maior magnitude e β o de menor, temos que:

$$Tr(H) = D_{xx} + D_{yy} = \alpha + \beta \tag{3.9}$$

$$Det(H) = D_{xx}D_{yy} - D_{xy}^2 = \alpha\beta \tag{3.10}$$

Sendo r a razão entre o maior e o menor autovalor, onde $\alpha = r\beta$, temos que:

$$\frac{Tr(H)^2}{Det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r} \tag{3.11}$$

O valor de $(r+1)^2/r$ cresce conforme a razão entre os autovalores cresce. Assim, eliminam-se os pontos cuja razão esteja acima de um limiar de r. Lowe sugere que faça-se $r = 10$.

Orientação dos Pontos-chave

A cada ponto chave deve ser definida uma orientação baseada nas características dos pixels ao redor do ponto chave. A construção dos descritores invariantes à rotação será feita relativamente à orientação definida nesse ponto.

Para cada imagem $L(x, y, \sigma)$, é calculada a magnitude do gradiente, $m(x, y)$, e sua orientação, $\theta(x, y)$, de acordo com as equações 3.12 e 3.13.

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{3.12}$$

$$\theta(x,y) = arctan\left(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}\right) \tag{3.13}$$

Para cada ponto chave é montado um histograma da orientação de seus pixels vizinhos. O histograma é dividido em 36 regiões de 10° cada, abrangendo assim o intervalo de 0° a 360°.

Cada pixel vizinho é adicionado ao histograma com um peso proporcional à sua magnitude e a uma janela circular gaussiana centrada no ponto chave e de desvio igual a 1,5 vezes o valor da escala do ponto chave.

O pico do histograma é considerado como a orientação daquele ponto chave. Caso existam outros picos, com até 80% do valor do pico máximo, são criados novos pontos chaves com a mesma localização e escala do ponto original, mas a eles são associadas as orientações dos picos respectivos.

Extração das Características

Por fim, uma vez que os pontos chave estão definidos, suas características devem ser extraídas e associadas a ele. Isso é feito a partir da magnitude e orientação do gradiente na região ao redor do ponto chave.

Para que haja a invariância à rotação, a orientação do gradiente ao redor do ponto chave é tomada em relação à orientação do ponto chave calculado anteriormente.

Uma janela quadrada ao redor do ponto chave é dividida em nxn regiões com kxk pixels de tamanho. Uma função gaussiana centrada no ponto chave e de desvio igual à metade do tamanho da janela é multiplicada ao valor da magnitude do gradiente dos pontos pertencentes à janela.

Para cada região da janela é montado um histograma da orientação do gradiente, onde o histograma é dividido em 8 regiões de 45° cada. O peso de cada ponto no histograma é proporcional à magnitude multiplicada pela função gaussiana. No total são gerados $8n^2$ descritores para cada ponto chave. [27] mostra que os valores ideais são $k = 4$ e $n = 4$, resultando em 128 descritores.

Os descritores são invariantes a variações de brilho da imagem, pois os mesmos são calculados a partir da diferença dos pixels, porém para que não sofram influência de variações homogêneas de contraste da cena, os descritores são normalizados de forma que o descritor de maior magnitude tenha sempre valor unitário. Isso torna os descritores robustos a variações lineares de iluminação.

Os efeitos de variações não-lineares de iluminação causam mudanças grandes na magnitude do gradiente, porém causam pouca influência em sua direção [27]. Para mitigar esses efeitos, todos os descritores maiores que 0,2 são igualados a 0,2 e re-normalizados para a escala de 0 a 1.

Uma vez que os descritores dos pontos chaves das imagens esquerda e direita

foram extraídos, eles são comparados pela distância euclidiana de seus descritores, para encontrar os pontos de correspondência entre as imagens. São considerados pontos correspondentes os pontos com a mínima distância euclidiana, desde que esta distância seja menor que um limiar de 0,8, para a eliminação de *outliers*.

3.2 Classificação dos Pares de Pontos

A correspondência através do SIFT é realizada para cada par de quadros correspondentes do vídeo estéreo. Os pares de pontos-chaves, determinados pelo SIFT, são utilizados para a verificação da conformidade na percepção do observador. Porém, em alguns casos, a quantidade de pontos em um quadro é pequena e pouco representativa da cena como um todo. Para melhorar a quantidade de pontos da cena, os quadros vizinhos (anterior e posterior) são comparados com o atual, e em caso de não haver movimento de um determinado ponto-chave do quadro vizinho, o quadro atual passa a possuir também esse ponto chave.

Para determinar se não houve movimento na cena, é calculada a diferença entre os quadros consecutivos. Considerando cada componente de cor (R, G e B) variando de 0 a 255, foi definido que não há movimento quando o módulo da diferença é menor que um limite l, conforme a equação 3.14. O limite l foi determinado experimentalmente igual a 4.

$$\sqrt{(R_k - R_{k+1})^2 + (G_k - G_{k+1})^2 + (B_k - B_{k+1})^2} < l \tag{3.14}$$

Após a listagem dos pares, é verificado se cada par de pontos está localizado em uma região permitida ou proibida em relação à percepção pelo observador.

Conforme foi visto no item anterior, a percepção do vídeo estereoscópico está associada às dimensões da tela de exibição, e à distância do observador à tela. Com isso, podemos dividir a profundidade em três regiões diferentes, conforme ilustrado na figura 3.1:

1. Região divergente (DIV) - é composta pelos pontos onde não é possível recriar a imagem percebida através da convergência dos olhos, ou seja, são classificados como DIV os pontos que não atendam à restrição da equação 2.21.

2. Região de conflito de vergência e acomodação (NCVR) - é composta pelos pontos onde há convergência, ou seja, onde a restrição da equação 2.21 é atendida, porém as equações 2.24 e 2.25 não são atendidas. Essa região gera desconforto quando há movimento em profundidade, fazendo com que a vergência varie rapidamente, forçando os músculos dos olhos [17].

3. Região de conforto (CVR) - é composta pelos pontos em que a convergência ocorre normalmente, ou seja, são classificados como OK os pontos que atendam a todas as restrições anteriores.

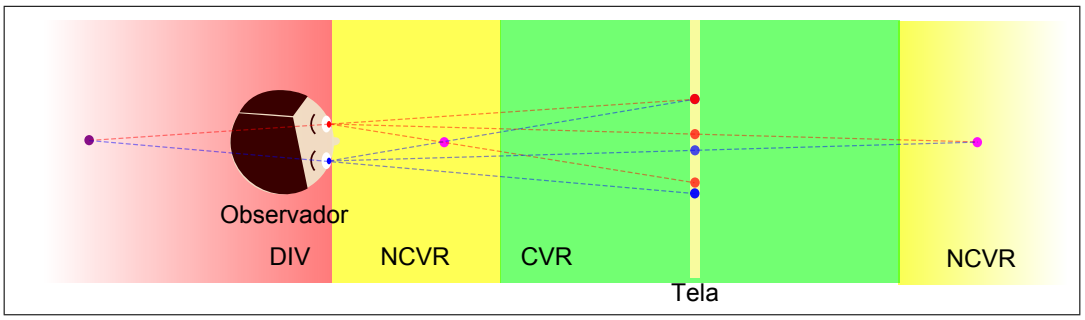

Figura 3.1: Diagrama com as regiões de profundidade

3.3 Segmentação das Imagens

Na última seção, os pontos chave de cada quadro do vídeo foram classificados nas regiões CVR, NCVR e DIV, conforme sua posição. Através da segmentação dos quadros, a região de cada segmento será classificada de acordo com os pontos chave pertencentes a ela.

A imagem da vista esquerda é submetida a uma segmentação, utilizando o software Edison, desenvolvido no *Robust Image Understanding Laboratory* da *Rutgers University*, NJ - EUA, baseado nos artigos [33], [34] e [35], configurado de forma que não haja regiões com menos de 2.000 pixels. Os parâmetros utilizados na segmentação são:

- Synergistic - perform synergistic segmentation = true
- SpatialBandWidth - segmentation spatial radius (integer) = 7
- RangeBandWidth - segmentation feature space radius (float) = 6,5
- MinimumRegionArea - minimum segment area (integer) = 2000
- SpeedUp - algorithm speed up = 1
- GradientWindowRadius - synergistic parameters (integer) = 2
- MixtureParameter - synergistic parameter (float) = 0,3
- EdgeStrengthThreshold- synergistic parameter (float) = 0,3

Para cada ponto-chave, verifica-se em qual segmento ele se encontra. Se o ponto é o único no segmento, todos os pontos do segmento passam a pertencer à mesma região do ponto original. Para o cálculo da métrica proposta, caso haja mais de um ponto no segmento, todos os pontos daquele segmento são distribuídos na proporção dos pontos que ali apareçam, por exemplo: se há um segmento com X pixels, e nesse segmento há L pontos CVR, J pontos NCVR e K pontos DIV, teremos então X*L/(L+J+K) pontos CVR, X*J/(L+J+K) pontos NCVR, e X*K/(L+J+K) pontos DIV, conforme exemplificado na figura 3.2

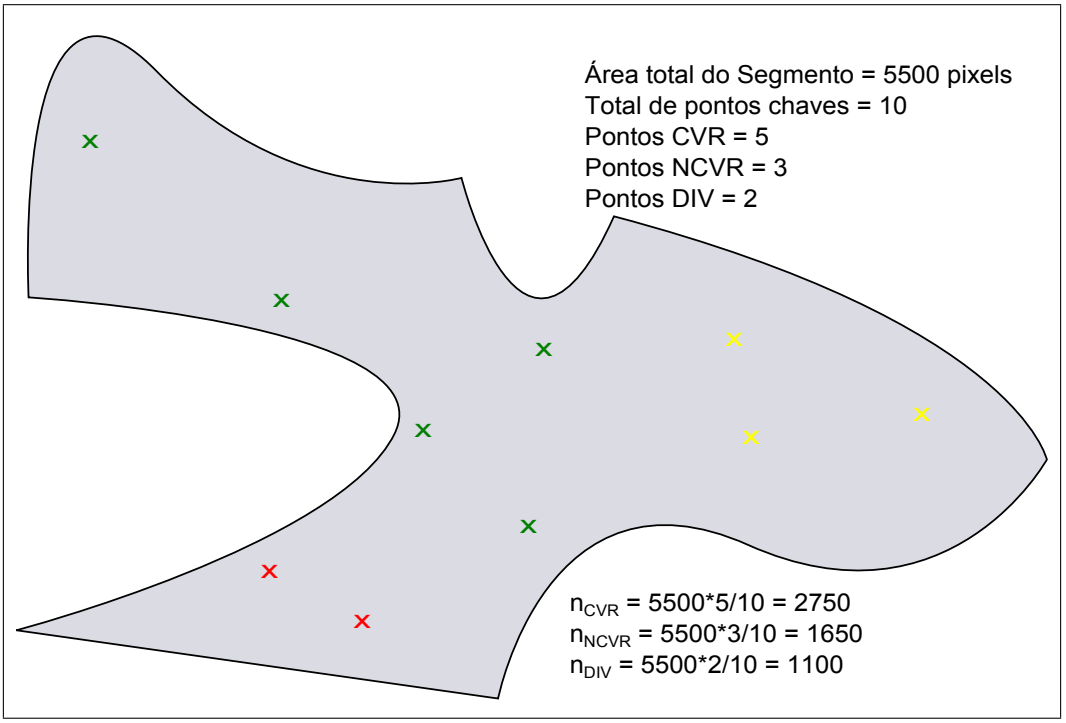

Figura 3.2: Exemplo de classificação de um segmento

Algumas regiões não terão nenhum ponto-chave associado. Nesse caso, o centróide da região é calculado e verifica-se a classificação do ponto-chave mais próximo a ele. A região recebe então a classificação desse vizinho mais próximo. A figura 3.3 exemplifica a classificação das regiões sem ponto chave.

As figuras 3.4 e 3.5 mostram a imagem original e a imagem segmentada com a marcação dos pontos determinados pelo SIFT. Os segmentos que não possuem nenhum ponto originário do SIFT são contabilizados como estado desconhecido (UNK). A figura 3.6 ilustra um quadro de um vídeo marcado com as regiões CVR, NCVR, DIV e UNK. A figura 3.7 ilustra a mesma imagem da figura 3.6, com as regiões UNK classificadas de acordo com o vizinho mais próximo de seu centróide.

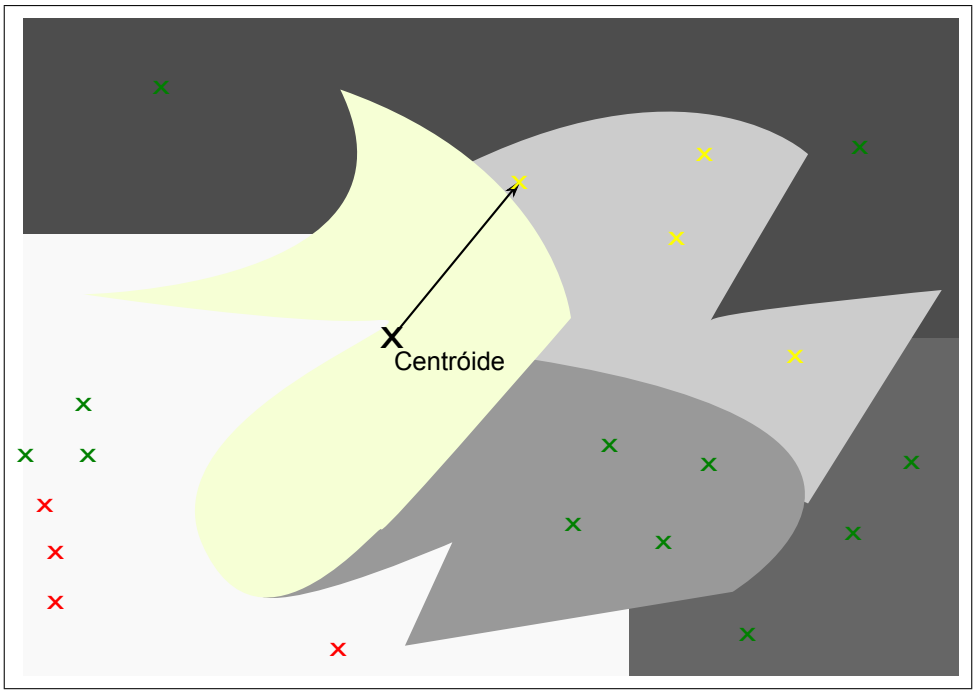

Figura 3.3: Classificação de um segmento sem ponto chave associado

3.4 Cálculo da Nota Objetiva

A métrica para a qualidade objetiva é calculada a partir da quantidade de pontos classificada em cada uma das regiões, re-escalonada para a escala de 0 a 100, de acordo com a equação 3.15. Os pontos classificados como DIV, por serem a falha mais severa[17] recebem pontuação nula. Os pontos classificados como NCVR recebem pontuação k (onde $0 < \mathrm{k} < 1$), e os pontos classificados como CVR recebem pontuação 1. O valor ótimo de k será estimado a partir de um conjunto de treinamento.

$$N_{OBJ} = 100 * \frac{n_{CVR} + k * n_{NCVR}}{n_{DIV} + n_{CVR} + n_{NCVR}} \tag{3.15}$$

No capítulo atual o método para a avaliação da qualidade de vídeos estereoscópicos foi desenvolvido, primeiramente com a busca de pares de pontos correspondentes pelo método SIFT. Em seguida é realizada o reaproveitamento dos pontos chave entre quadros onde não há movimento. Cada quadro é segmentado e cada segmento é classificado de acordo com os pontos chaves a ele pertencente. Por fim, os segmentos que não contém pontos chave são classificados de acordo com o ponto chave vizinho mais próximo de seu centróide. No próximo capítulo o método será validado com a base de vídeos descrita em [1].

Figura 3.4: Amostra de um quadro da vista esquerda de um vídeo estéreo

Figura 3.5: Imagem da figura 3.4 segmentada, com as marcações dos pontos-chave SIFT.

Figura 3.6: Imagem da figura 3.4 com as regiões CVR em verde, as regiões DIV em vermelho, as regiões NCVR em amarelo e as regiões sem classificação em azul.

Figura 3.7: Imagem da figura 3.4 após a classificação das regiões sem ponto chave

Capítulo 4

Resultados e Discussões

Para validação do método, foi utilizada a base de vídeos estereoscópicos da *École Polytechnique Federale de Lausanne* (EPFL), descrita em [1]. A base é formada por 30 vídeos de 10 segundos cada, sendo divididos em 5 diferentes vídeos representando a mesma dentre 6 cenas. Cada cena foi gravada 5 vezes, a primeira vez com as câmeras com uma distância horizontal entre elas de 10 cm, sendo que a cada nova tomada a distância entre as câmeras era aumentada em 10 cm.

A resolução dos vídeos é de 1920 x 1080 pixels, a 25 quadros por segundo. As cenas possuem conteúdo gravado em áreas internas e externas, com variedades de cor, movimentos, texturas e profundidades. Como cada tomada foi gravada separadamente, o conteúdo entre 2 vídeos da mesma cena pode variar ligeiramente, porém mantendo as características gerais citadas acima.

A distância máxima entre as câmeras foi calculada em [1] usando a equação de Bercovitz simplificada [36] (equação 4.1), que associa a distância máxima entre câmeras(b), com as características do conjunto de lentes (f - distância focal, do limite de paralaxe (p) e da geometria da cena (n - distância do objeto mais próximo à câmera e l - distância do objeto mais afastado da câmera), e é válida quando a distância focal é bem menor que a mínima distância.

$$b = \frac{p}{f}.\frac{l.n}{l-n} \tag{4.1}$$

A tabela 4.1 mostra para cada vídeo a distância do objeto da cena mais próximo da câmera (n), a distância do objeto mais distante da câmera (l), e a distância máxima calculada (b) para cada cena, conforme calculado por [1].

Os vídeos foram submetidos pelos autores de [1] para a realização de testes subjetivos de acordo com as normas ITU-R BT.500 [37], ITU-R BT.710 [38] e ITU-R BT.1438 [39] utilizando-se de um monitor estereoscópico de 46" e resolução de 1920 x 1080 pixels, com o observador posicionado a 2 metros da tela. A partir dos testes subjetivos, foram obtidas as notas MOS (*mean opinion score*) relativas

Tabela 4.1: Características das cenas captadas - extraído de [1]

ID	VIDEO	D MIN (n) - m	D MAX (l) - m	D CAMERA (b) - cm
1	sofa	3	6	17
2	bike	10	150	30
6	feet	2	4	11
8	hallway	2	20	6
11	notebook	3	10	12
12	car	8	120	24

à qualidade dos vídeos da base. Para a avaliação, os 30 vídeos foram apresentados aleatoriamente, com intervalos de 5 segundos entre eles, para que o avaliador pudesse anotar sua nota. Ao todo, 20 avaliadores participaram do processo.

Os resultados dos testes subjetivos estão apresentados na tabela 4.2, e a figura 4.1 possui uma amostra do 100° quadro extraída do vídeo designado ao olho direito de cada uma das cenas. Percebe-se na imagem a variedade de conteúdo apresentado na base.

Tabela 4.2: Avaliação subjetiva dos vídeos da base EPFL

DIST	sofa	bike	feet	hallway	notebook	car
10	74.529	86.941	77.529	68.118	83.882	82.176
20	64.059	81.765	56.118	53.412	71.294	87
30	56.706	71.235	54.882	41.706	52.471	78.412
40	43.647	75.529	42.941	19.529	31.765	76
50	35.471	62.294	22.882	13.471	18.118	68.882

Figura 4.1: Amostra do 100° quadro dos vídeos referentes ao olho direito

Cada um dos 30 vídeos foram processados com o software Autopano-sift-c (disponível em wiki.panotools.org), para a extração dos pontos correspondentes através do método SIFT. O software encontra os pontos chave de cada quadro do vídeo a partir da média das componentes de cor de cada pixel. Em seguida, é realizado o processo de reaproveitamento dos pontos chave entre quadros onde não há movimento.

Se o processo de reaproveitamento dos pontos-chave não for executado, a proporção de regiões sem classificação do vídeo, em relação à quantidade total de pontos do vídeo é de em média 41,61%. Após o processamento, a proporção de regiões sem classificação é reduzida para uma média de 11,40% do conteúdo. Os resultados por vídeo estão exibidos na tabela 4.3.

Tabela 4.3: Redução das regiões classificadas como desconhecidas.

Vídeo	Sem processamento temporal	Com processamento temporal
sofa-10	17,26%	2,55%
sofa-20	38,43%	4,14%
sofa-30	57,81%	6,74%
sofa-40	70,4%	11,16%
sofa-50	79,13%	14,71%
bike-10	3,41%	0,22%
bike-20	8,57%	1,01%
bike-30	14,18%	1%
bike-40	18,74%	2,41%
bike-50	20,99%	2,66%
feet-10	15,03%	2,5%
feet-20	40,1%	8,51%
feet-30	42,08%	10,44%
feet-40	50,84%	18,42%
feet-50	58,51%	23,19%
hallway-10	46,73%	3,73%
hallway-20	74,19%	14,49%
hallway-30	79,19%	18,78%
hallway-40	82,04%	26,24%
hallway-50	83,65%	31,97%
notebook-10	34,4%	9,46%
notebook-20	46,92%	20%
notebook-30	60,79%	28,26%
notebook-40	64%	29,48%
notebook-50	67,66%	40,88%
car-10	8,29%	0,76%
car-20	8,65%	1,12%
car-30	14,09%	1,69%
car-40	18,07%	2,25%
car-50	24,1%	3,28%
Média	41,61%	11,4%

A tabela 4.4 mostra a quantidade total de pontos de cada vídeo, e a quantidade de pontos classificados como CVR, NCVR e DIV após o processamento dos vídeos.

As regiões sem classificação são tratadas com o método descrito na seção 3.2. A contagem final de pixels por vídeo classificados como CVR, NCVR e DIV está exposta na tabela 4.5.

Tabela 4.4: Resultados da análise por SIFT

Vídeo	n_{CVR}	n_{NCVR}	n_{DIV}
sofa-10	3141238	26414	1074
sofa-20	1253946	685472	250
sofa-30	354574	733614	23442
sofa-40	156228	340220	183484
sofa-50	43109	222225	188265
bike-10	2738111	9209	0
bike-20	2319287	18915	250
bike-30	1672306	81357	250
bike-40	444535	1075940	2329
bike-50	240742	1019395	3500
feet-10	1221897	50930	0
feet-20	176870	289294	6
feet-30	90665	14456	102828
feet-40	28744	1027	129609
feet-50	7866	2270	97535
hallway-10	542972	211319	0
hallway-20	26296	192503	37187
hallway-30	3521	17125	91045
hallway-40	1429	3614	72864
hallway-50	0	236	53949
notebook-10	554416	1486501	250
notebook-20	13302	1479782	33486
notebook-30	5245	6504	1275513
notebook-40	4579	9021	924991
notebook-50	526	4858	812949
car-10	5061089	33094	1086
car-20	4296028	127348	500
car-30	1200621	2193057	4000
car-40	255038	2758878	9026
car-50	87531	2284941	69808

Os 30 vídeos foram divididos em 25 vídeos de treinamento e 5 vídeos de teste para que pudesse ser calculado o valor ótimo de k. Todas as C_{30}^{25} combinações foram calculadas. A otimização buscou o valor de k que maximizasse a correlação de Pearson entre as notas N_{OBJ} e MOS dos vídeos pertencentes ao conjunto de treinamento. Caso a base de vídeos fosse maior, a quantidade de combinações seria muito maior. Para que seja possível realizar a otimização, deve ser usada validação cruzada do tipo *k-fold*, dividindo os vídeos em subconjuntos. Os subconjuntos são então divididos em treinamento e teste, realizando-se todas as combinações de subconjuntos.

O valor de k determinado em cada uma das combinações foi aplicado ao conjunto de teste, onde foram calculados a correlação de Pearson, a correlação de Spearman, o erro médio absoluto normalizado (NMAE) e o erro médio quadrático normalizado (NMRSE). O valor médio de k é de 0,7225, com desvio de 0,0428. A figura 4.2 mostra o histograma dos valores de k calculados, com intervalo de 0,01 e a figura 4.3 mostra o histograma da correlação de Pearson calculada para cada combinação, com intervalo de 0,01. A média e o desvio das correlações e dos erros médios estão exibidos na tabela 4.6.

Tabela 4.5: Contagem dos pixels de cada vídeo pela sua classificação como CVR, NCVR ou DIV.

Vídeo	n_{CVR}	n_{NCVR}	n_{DIV}	TOTAL
sofa-10	479519116,78	4856450,14	328433,07	484704000
sofa-20	276824727,79	193202534,48	56737,73	470084000
sofa-30	155437979,35	280460122,45	19429898,2	455328000
sofa-40	101086700,22	200678586,59	134202713,19	435968000
sofa-50	58050251,57	193966412,66	166063335,77	418080000
bike-10	499040503,69	1393496,31	0	500434000
bike-20	495634580,33	3768673,95	84745,72	499488000
bike-30	465289761,82	33115420,88	136817,3	498542000
bike-40	193167151,18	300855456,59	527392,23	494550000
bike-50	140442903,45	354902406,67	1198689,88	496544000
feet-10	466118801,03	18721198,97	0	484840000
feet-20	173315635,87	283226698,83	25665,3	456568000
feet-30	199285565,19	35365119,05	184299315,76	418950000
feet-40	112650865,16	4118851,25	266990283,59	383760000
feet-50	37178050,62	16449798,42	304482150,96	358110000
hallway-10	428965503,02	70630496,98	0	499596000
hallway-20	228050449	198345414,56	46364136,44	472760000
hallway-30	146382953,65	116356553,02	190284493,32	453024000
hallway-40	125282012,58	87587008,98	218680978,44	431550000
hallway-50	0	6558870,67	411085129,33	417644000
notebook-10	254061399,93	246577717,6	10882,47	500650000
notebook-20	68850375,75	394021776,73	13535847,52	476408000
notebook-30	36413049,15	33194269,39	387828681,46	457436000
notebook-40	38206150,81	72304707,49	321265141,71	431776000
notebook-50	3607681,75	38136280,83	373006037,42	414750000
car-10	489771712,7	5754147,69	74139,61	495600000
car-20	479399564,88	20388987,41	107447,7	499896000
car-30	207472084,99	281683514,18	1194400,82	490350000
car-40	101459454,83	379761011,66	3077533,51	484298000
car-50	59579099,46	406656176,42	14924724,12	481160000

O método foi testado em um computador com processador Pentium Core2Quad Q6600, com 2 GB de memória RAM DDR2, Windows Vista 32 bits. O processamento total de 1 quadro dura em média 2 minutos, sendo que os processos que consumiram mais tempo de processamento foram a busca dos pontos chave pelo SIFT, com média de 12 segundos por quadro do vídeo e a segmentação, com média de 100 segundos por quadro, sendo que o processamento foi executado somente por um núcleo da CPU. Com isso, o processamento de um vídeo da base de testes, que possui 250 quadros, leva em média 8 horas e 20 minutos se processado por um núcleo do processador. Como o computador utilizado tinha 4 núcleos, o tempo médio de processamento por vídeo foi de 2 horas e 5 minutos.

O método apresentado em [2] propõe utilizar as características estatísticas do mapa de disparidade e do gradiente da disparidade, além de indicadores de atividade espacial e de movimento para determinar de forma objetiva o conforto na visualização de vídeos estereoscópicos, sendo que os resultados dos testes apresentados foram obtidos para a mesma base de vídeos utilizada nesse artigo. O coeficiente

Tabela 4.6: Correlação e erro médio das notas obtidas pelo método proposto e das notas obtidas nos testes subjetivos de [1].

	Média	Desvio
Pearson	0,8590	0,1644
Spearman	0,8058	0,2049
NMRSE	0,1753	0,0427
NMAE	0,1530	0,0428

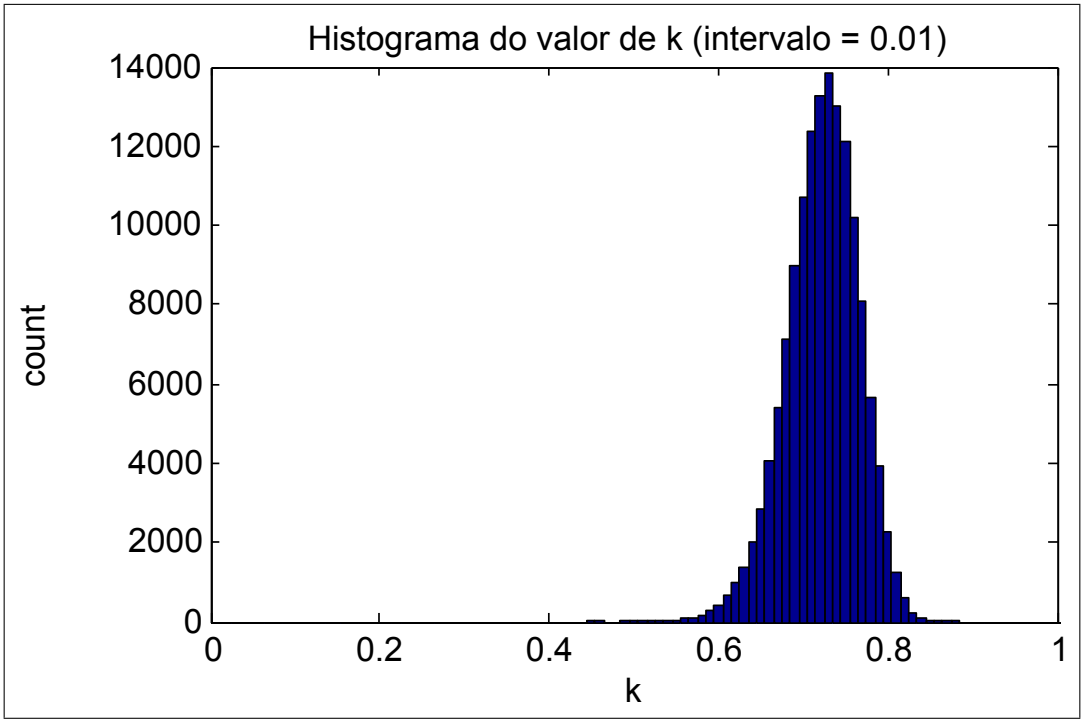

Figura 4.2: Histograma dos valores de k estimados para as C_{30}^{25} combinações, com intervalo de 0,01.

de correlação de Spearman obtido em [2] tem média de 0,76 e desvio de 0,25 para o método proposto utilizando PCA (*principal component analysis*), e média de 0,68 com desvio de 0,28 para o método proposto utilizando FFS (*forward feature selection*). Já no método proposto nesse artigo, o coeficiente de correlação de Spearman obtido é de 0,81, com desvio de 0,20, obtendo um desempenho melhor que os métodos propostos em [2].

Os resultados dos testes de validação mostram que o método proposto reproduziu bem os resultados dos testes subjetivos. Como a base de vídeos utilizada só possui variação na distância horizontal entre as câmeras, e não havia outra base disponível para estudo, somente problemas gerados pela variação horizontal das câmeras são detectados pelo método.

Além de não haver outra base disponível, não era possível gerar uma nova base

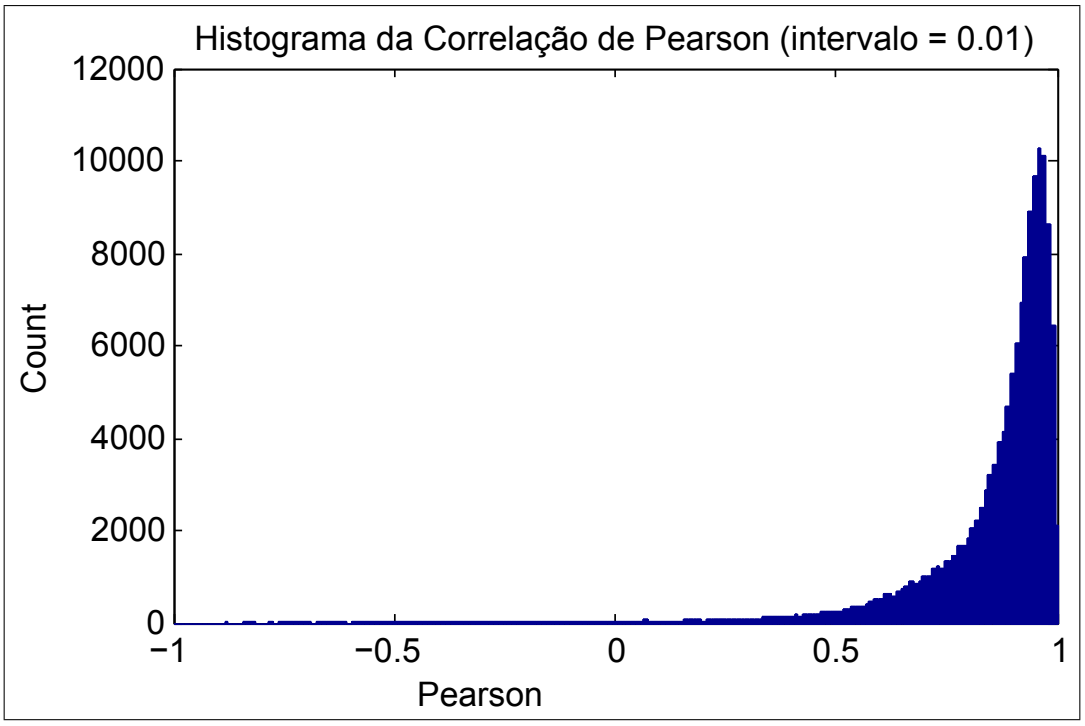

Figura 4.3: Histograma da corralação de Pearson do conjunto de testes calculado para as C_{30}^{25} combinações, com intervalo de 0,01.

para esse estudo. Uma continuação do trabalho deve incluir a geração de uma base de vídeos mais abrangente em relação às possíveis variações do arranjo de câmeras.

Capítulo 5

Conclusões

Nessa dissertação foi estudada uma forma de avaliar automaticamente a qualidade na percepção de um vídeo estéreo quando se varia a distância entre as câmeras, tentando obter um resultado o mais próximo possível dos resultados de avaliação subjetiva.

Foi apresentada a geometria da captação e da exibição de um vídeo estéreo. A partir dessa geometria foi elaborado um método de avaliação do vídeo. O método é composto primeiramente do uso da técnica SIFT para obtenção de pares de pontos correspondentes nas vistas esquerda e direita. Em seguida, são verificados se existem pontos correspondentes que podem ser reaproveitados em quadros anteriores ou posteriores ao quadro original ao qual aquele ponto pertence. Cada quadro de uma das vistas é segmentado e os pontos chaves pertencentes a cada segmento são usados para classificar o segmento com um todo. Os segmentos que não possuem pontos associados são classificados de acordo com o ponto mais próximo de seu centróide.

A classificação divide os pixels do vídeo em 3 grupos: CVR (onde o pixel encontra-se na região de conforto), NCVR (onde o pixel está na região de conflito de vergência e acomodação), e DIV (onde há divergência dos olhos). A nota final é baseada na contagem dos pixels pertencentes a cada um dos grupos.

Os resultados dos testes com a base de vídeos da EPFL mostram que o algoritmo é eficiente para a detecção automática das falhas do ajuste horizontal das câmeras. Como a base não possui outros tipos de falhas, o desenvolvimento do algoritmo ficou limitado a esse problema. Porém, é importante ressaltar que o ajuste da distância entre as câmeras é um dos pontos críticos do ajuste de cena durante a captação do conteúdo, uma vez que a distância entre as câmeras é variável de acordo com as lentes utilizadas e com o conteúdo a ser captado, e é também dependente do tamanho da tela e da resolução para o qual o conteúdo é destinado (os ajustes verticais e de rotação das câmeras são fixos e não variam de cena para cena).

5.1 Trabalhos Futuros

Alguns pontos são sugeridos como possíveis campos de estudo para trabalhos futuros:

- A detecção de falhas pode ser melhorada pois ainda não estão incluídos todos os possíveis problemas de captação dos vídeos estéreo. Para tanto é necessário utilizar-se de uma base de vídeos com outros tipos de defeitos.
- Devem ser testados outros métodos para processo de obtenção da correspondência esparsa entre os pontos estéreo, por exemplo *Speeded Up Robust Feature* (SURF) [40].
- Os parâmetros e o método de segmentação podem ser otimizados.
- Além de ser usado como um avaliador da qualidade de vídeo, o algoritmo apresentado pode ser utilizado como base para um algoritmo de correção automática da disparidade em vídeos estéreo, possibilitando a exibição de um mesmo conteúdo em diferentes plataformas mantendo sempre a qualidade em seu máximo.

Referências Bibliográficas

[1] GOLDMANN, L., SIMONE, F. D., EBRAHIMI, T. "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video", *IS&T/SPIE Electronic Imaging, 3D Image Processing (3DIP) and Applications*, 2010.

[2] MITTAL, A., MOORTHY, A. K., GHOSH, J., et al. "Algorithmic assessment of 3D quality of experience for images and videos", *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE)*, pp. 338–343, January 2011.

[3] WOODS, A., DOCHERTY, T., KOCH, R. "Image Distortions in Stereoscopic Video Systems", *Proc. SPIE, Stereoscopic Displays and Applications IV*, v. 1915, pp. 36–48, 1993.

[4] NDPGROUP. *Awareness of 3D TVs and Blu-ray Players Grows, According to The NPD Group.* Relatório técnico, Port Washington, NY, USA, April 2011.

[5] ITU-R. "Stereoscopic television based on R-and L-eye two channel signals". Rec. BT.1198, 1995.

[6] MEESTERS, L., IJSSELSTEIJN, W., SEUNTIENS, P. "Survey of perceptual quality issues in three-dimensional television systems". In: *Proc. SPEI, Stereoscopic Displays and Virtual Reality Systems X*, v. 5006, pp. 313–326, January 2003.

[7] MENDIBURU, B. *3D Movie Making - Stereoscopic Digital Cinema from Script to Screen.* 1 ed. Burlington, MA, EUA, Focal Press, 2009.

[8] MANSSON, J. "Stereovision: A Model Of Human Stereopsis", *Lund University Cognitive Studies*, v. 64, 1998.

[9] DODGSON, N. "Variation and Extrema of Human Interpupillary Distance". In: *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems XI*, v. 5291, pp. 36–46, San Jose, CA, EUA, 2004.

[10] TRUCCO, E., VERRI, A. *Introductory Techniques for 3-D Computer Vision.* Englewood cliffs, NJ, USA, Prentice Hall, 1998.

[11] JULESZ, B. *Foundations of Cyclopean perception.* Chicago, IL, USA, The University of Chicago Press, 1971.

[12] HOWARD, I., ROGERS, B. *Binocular Vision and Stereopsis.* New York, NY, USA, Oxford University Press, 1995.

[13] SOUTHARD, D. A. "Viewing model for virtual environment displays", *Journal of Electronic Imaging*, v. 4(4), pp. 413–420, October 1995.

[14] SPOTTISWOODE, R., SPOTTISWOODE, N. *The theory of stereoscopic transmission & its application to the motion picture.* University of California Press, 1953.

[15] MEESTERS, L. M. J., IJSSELSTEIJN, W. A., SEUNTIENS, P. J. H. "A survey of perceptual evaluations and requirements of threedimensional TV", *IEEE Transactions on Circuits and Systems for Video Technology*, v. 14, n. 3, pp. 381–391, 2004.

[16] SCHERTZ, A. "Source coding of stereoscopic television pictures", *IEE Inter. Conference on image processing and its applications*, pp. 462–464, 1992.

[17] LAMBOOIJ, M., IJSSELSTEIJN, W., HEYNDERICKX, I. "Visual discomfort in stereoscopic displays: A review". In: *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems XIV*, v. 6490, 2007.

[18] DEVERNAY, F., BEARDSLEY, P. "Stereoscopic Cinema". In: *Image and Geometry Processing for 3-D Cinematography*, v. 5, *Geometry and Computing*, Springer Berlin Heidelberg, 2010.

[19] VALYUS, N. *Stereoscopy.* London, UK, Focal Press, 1966.

[20] YEH, Y., SILVERSTEIN, L. "Limits of Fusion and Depth Judgment in Stereoscopic Color Displays", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, v. 32, pp. 45–60, February 1990.

[21] YANO, S., EMOTO, M., MITSUHASHI, T. "Two factors in visual fatigue caused by stereoscopic HDTV images", *Displays*, v. 25(4), pp. 141–150, November 2004.

[22] HOFFMAN, D. M., GIRSHICK, A., AKELEY, K., et al. "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue", *Journal of Vision*, v. 8(3), pp. 1–30, 2008.

[23] CHEN, W., FOURNIER, J., BARKOWSKY, M., et al. "New requirements of subjective video quality assessment methodologies for 3DTV", *Video Processing and Quality Metrics 2010 (VPQM)*, 2010.

[24] SCHARSTEIN, D., SZELISKI, R. "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms", *International Journal of Computer Vision*, v. 47(1/2/3), pp. 7–42, April-June 2002.

[25] KLAUS, A., SORMANN, M., KARNER, K. "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure", *18th International Conference on Pattern Recognition*, v. 3, pp. 15–18, 2006.

[26] ZITNICK, C. L., KANADE, T. "A cooperative algorithm for stereo matching and occlusion detection", *IEEE Trasactions Pattern Anal. Mach. Intell.*, v. 22, n. 7, pp. 675–684, July 2000.

[27] LOWE, D. "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision*, v. 60, n. 2, pp. 91–110, 2004.

[28] WITKIN, A. P. "Scale-space filtering", *International Joint Conference on Artificial Intelligence*, pp. 1019–1022, 1983.

[29] LOWE, D. G. "Object recognitionfrom local scale-invariant features", *International Conference on Computer Vision*, pp. 1150–1157, 1999.

[30] DINIZ, P., SILVA, E., NETTO, S. *Digital Signal Processing - System Analysis and Design*. Cambridge University Press, September 2010.

[31] BROWN, M., LOWE, D. G. "Invariant features from interest point groups", *British Machine Vision Conference*, pp. 656–665, 2002.

[32] HARRIS, C., STEPHENS, M. "A combined corner and edge detector", *Fourth Alvey Vision Conference*, pp. 147–151, 1988.

[33] COMANICU, D., MEER, P. "Mean shift: A robust approach toward feature space analysis", *IEEE Trans. Pattern Anal. Machine Intell.*, v. 24, pp. 603–619, 2002.

[34] MEER, P., GEORGESCU, B. "Edge detection with embedded confidence." *IEEE Trans. Pattern Anal. Machine Intell.*, v. 23, pp. 1351–1365, 2001.

[35] CHRISTOUDIAS, C., GEORGESCU, B., MEER, P. "Synergism in low-level vision." *16th International Conference on Pattern Recognition*, v. vol. IV, pp. 150–155, 2002.

[36] BERCOVITZ, J. "Image-side perspective and stereoscopy". In: *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems V*, v. 3295, 1998.

[37] ITU-R. "Methodology for the subjective assessment of the quality of television pictures". Rec. BT.500-11, 2002.

[38] ITU-R. "Subjective assessment methods for image quality in high-definition television". Rec. BT.710-4, 1998.

[39] ITU-R. "Subjective assessment of stereoscopic television pictures". Rec. BT.1438, 2000.

[40] BAY, H., ESS, A., TUYTELAARS, T., et al. "Speeded-Up Robust Features (SURF)", *Computer Vision and Image Understanding*, v. 110, n. 3, pp. 346 – 359, 2008.