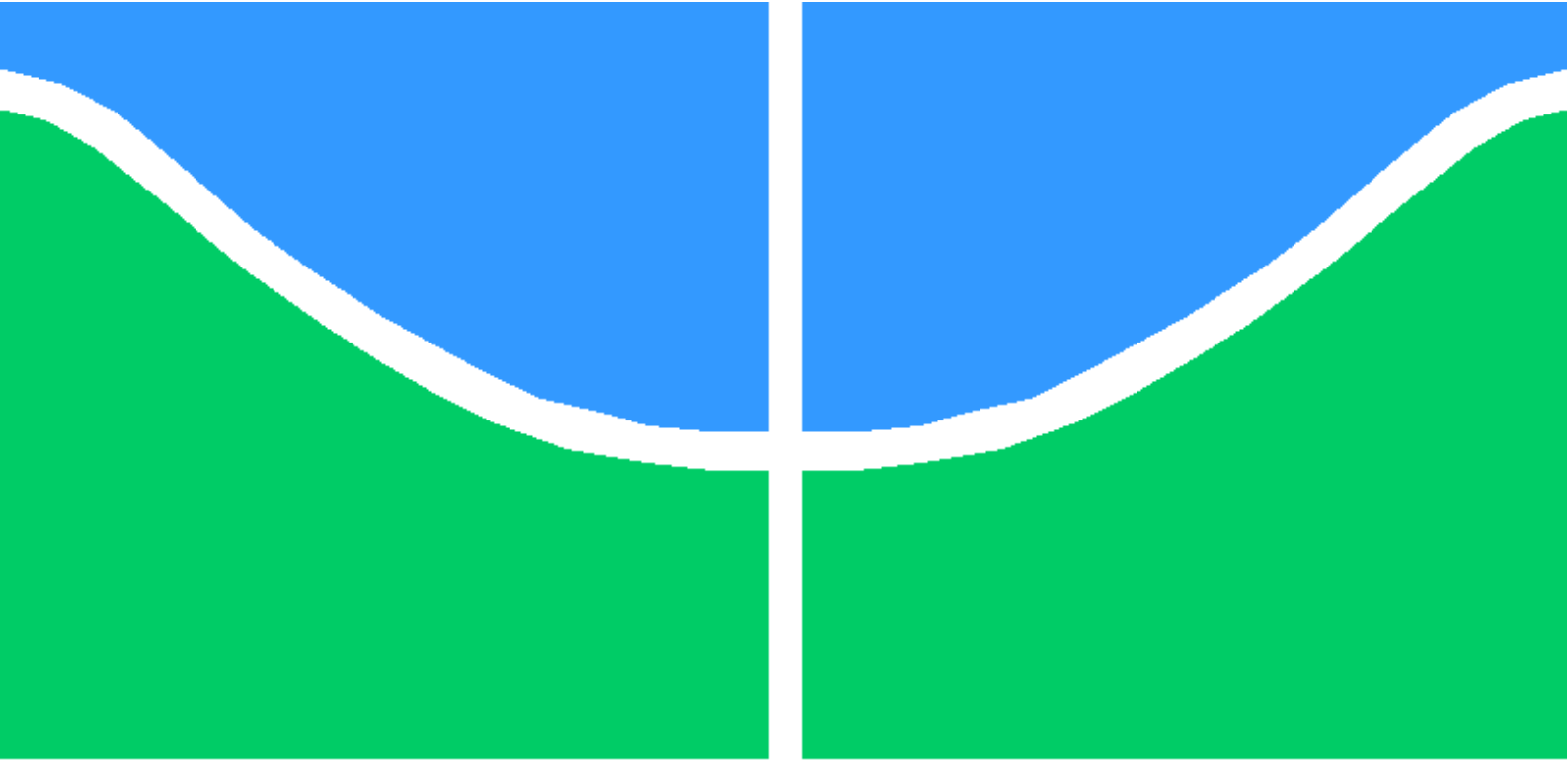


DISSERTAÇÃO DE MESTRADO

MÉTRICAS OBJETIVAS BASEADAS EM PROJEÇÕES PARA AVALIAÇÃO DE QUALIDADE EM NUVENS DE PONTOS

Eric de Menezes Torlig

Brasília, Dezembro de 2018

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASILIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO

MÉTRICAS OBJETIVAS BASEADAS EM PROJEÇÕES PARA AVALIAÇÃO DE QUALIDADE EM NUVENS DE PONTOS

Eric de Menezes Torlig

Relatório submetido ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Mestre em Engenharia Elétrica

Banca Examinadora

Prof. Ricardo L. de Queiroz, ENE/UnB _____
Orientador

Prof. Eduardo Antônio Barros da Silva, UFRJ _____
Examinador externo

Prof. Eduardo Peixoto Fernandes da Silva, PGEA/UnB _____
Examinador interno

RESUMO

Este estudo propõe um novo arcabouço para a mensuração e avaliação da qualidade visual de conteúdos tridimensionais representados através de nuvens de pontos, baseado na projeção bidimensionais dos conteúdos sob análise. A escolha do número e orientação das projeções é feita de maneira a cobrir a superfície do conteúdo em análise da maneira mais uniforme possível. As projeções são então mensuradas através de métricas objetivas comuns para a avaliação de conteúdo bidimensional. Por meio de experimentos subjetivos, o desempenho do arcabouço é explorado em conjunto com diversas métricas comumente utilizadas em análise de imagens ou vídeos, e comparado com métricas baseadas em pontos já utilizadas para a avaliação de conteúdo tridimensional. Em relação a outras métricas já existentes, o desempenho do arcabouço proposto se mostra consideravelmente superior em prever a qualidade subjetiva percebida por pessoas.

Palavras-chave: Nuvem de pontos, métrica objetiva.

ABSTRACT

This study proposes a novel framework for the measuring and evaluation of visual quality of three dimensional content represented by point clouds, based on two dimensional projections of the contents under evaluation. The choice of the number and orientation of the projections is done so as to cover content surface as uniformly as possible. Projections are then evaluated using objective metrics usual to two dimensional content. By performing subjective experiments, the framework's performance is explored combined with several metrics common in analysis of images or video, and compared with point-based metrics normally used in three-dimensional content evaluation. Relative to other previously existing metrics, the framework's performance is considerably superior at predicting subjective quality as perceived by human beings.

Key-words: Point cloud, objective metric

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

Símbolos Latinos

V	Conjunto de vértices	
C	Conjunto de cores	

Símbolos Gregos

θ	Azimute	[rad]
ϕ	Ângulo polar	[rad]

Grupos Adimensionais

π	Pi

Subscritos

min	Mínimo
max	Máximo
inf	Infimum
sup	Supremum
sim	Similaridade

Sobrescritos

$\hat{}$	Valor estimado

Siglas

PC	*Point Cloud*
PSNR	*Peak Signal-to-Noise Ratio*
PSNR-HVS	*PSNR Human Visual System*
PSNR-HVS-M	*PSNR-HVS Multiscale*
P-PSNR	*Projective Peak Signal-to-Noise Ratio*
SSIM	*Structural Similarity Index*
MSSIM	*Multi-scale Structural Similarity Index*
MSE	*Mean Squared Error*
RMSE	*Root Mean Squared Error*
VIF	*Visual Information Fidelity*
VIFP	*Pixel-based Visual Information Fidelity*
NSS	*Natural Scene Statistics*
ITU	*International Telecommunication Union*
ITU-T	*ITU Telecommunication Standardization Sector*
ACR	*Absolute Category Rating*
ACR-HR	*Absolute Category Rating with Hidden Reference*
DSIS	*Double Stimulus Impairment Scale*
MOS	*Mean Opinion Score*
OD	*Octree Depth*
QP	*Quality Parameter*
UNB	Universidade de Brasília
EPFL	*École Polytechnique Fédérale de Lausanne*
ANOVA	*Analysis of Variance*
OR	*Outlier Ratio*
PCC	*Pearson's Correlation Coefficient*
SROCC	*Spearman's Rank-order Correlation Coefficient*
SS	*Sum of Squares*
DF	*Degrees of Freedom*
MS	*Mean Squares*

Capítulo 1

Introdução

Nesta seção são estabelecidos conceitos necessários para a compreensão de como conteúdo visual tridimensional pode ser representado através de nuvens de pontos e como esse tipo de conteúdo pode ter sua qualidade avaliada.

1.1 Contextualização

O sistema visual humano é naturalmente adaptado à percepção de conteúdos tridimensionais. No entanto, historicamente a maior parte do conteúdo de mídias visuais foi bidimensional [1, 2]. Conteúdo bidimensional não se aproveita da totalidade da capacidade sensorial humana, perdendo oportunidades de comunicar riqueza de detalhes e imersão em níveis próximos ao de interações em pessoa.

O aumento da disponibilidade de conteúdo tridimensional ocorreu principalmente nos últimos anos. Acompanhando essa tendência, tecnologias voltadas a esse tipo de conteúdo começaram a se desenvolver recentemente [3].

1.2 Motivação

Durante o desenvolvimento de qualquer técnica de processamento de conteúdo visual, como compressão de vídeo, uma etapa necessária é a comparação relativa da degradação da fidelidade entre duas imagens. Isto é, para determinar entre dois processos qual é o mais vantajoso em termos de qualidade, é necessário obter uma medida de quanto o conteúdo modificado por cada processo tem sua qualidade degradada em relação ao conteúdo original, no contexto da visão humana.

Quanto mais correlacionada com a percepção subjetiva humana, maior a garantia de que as conclusões da análise objetiva de técnicas de processamento sejam significativas e tenham resultados consistentes com a realidade.

1.3 Definição do problema

Maneiras de se disponibilizar, processar e avaliar a qualidade de conteúdos de duas dimensões já foram bem exploradas na literatura [4, 5, 6, 7, 8, 9].

No entanto, processamento de conteúdo de três dimensões é um campo com poucos padrões definidos e sem consenso sobre a melhor maneira de realizar todas as tarefas necessárias para a análise e desenvolvimento de novas tecnologias.

Além do mais, propostas existentes falham tanto em incorporar à medida da qualidade aspectos relacionados à geometria e à cor dos conteúdos avaliados, como em manter uma correlação próxima com a qualidade visual subjetiva percebida por seres humanos [10, 11, 12].

1.4 Objetivos do projeto

Propõe-se uma nova técnica computacionalmente simples e eficiente para a avaliação objetiva de qualidade de imagens tridimensionais no formato de nuvens de pontos que é capaz de incorporar aspectos visuais de cor e geometria, ao mesmo tempo que mantém uma correlação robusta com avaliações subjetivas de qualidade.

1.5 Apresentação do manuscrito

No Capítulo 2 será feita uma revisão bibliográfica sobre o tema de estudo. Em seguida, o Capítulo 3 descreve a metodologia empregada no desenvolvimento do projeto. Resultados experimentais são discutidos no capítulo 4, seguido das conclusões no Capítulo 5.

Capítulo 2

Revisão Bibliográfica

2.1 Introdução

Avaliação objetiva e subjetiva da qualidade de nuvens de pontos ainda são problemas abertos [4, 5, 10, 11, 12, 6, 7, 8, 9]. Particularmente, a combinação das avaliações de geometria e de cor tem sido difícil.

Para compreender a relação entre métricas objetivas e conteúdos tridimensionais, principalmente aqueles representados através de nuvens de pontos, é interessante observar a mesma relação acerca de conteúdos bidimensionais. Ao longo deste capítulo são abordados temas relacionados com a representação e análise de conteúdo bidimensional. Em seguida, são traçados paralelos entre o caso bidimensional e tridimensional, em termos de análise do conteúdo. Por fim, é explorado o estado da arte na análise de qualidade de conteúdo tridimensional.

2.2 Representação de imagens

2.2.1 Gráficos raster

No campo da computação gráfica, um gráfico *raster* (gráfico em varredura), também chamado de *bitmap* (mapa de bits) é uma estrutura de dados usada para representar imagens através de uma matriz de pontos regularmente e densamente espaçados em uma área retangular. Cada elemento nessa matriz tem um valor definido, representando alguma característica visual (e.g. cor, intensidade luminosa) de um elemento mínimo que compõe uma imagem, o *pixel*, do inglês *picture element*.

Esse tipo de representação obtém vantagem do modo como imagens digitais são visualizadas em praticamente qualquer equipamento eletrônico de mídia moderno. De computadores a TVs e *smartphones*, a adoção de telas compostas por elementos emissores de luz individuais arranjados em uma estrutura regular é ubíqua. Isso cria uma relação de um para um entre a representação de imagens através de gráficos *raster* e a maneira como o hardware fornece o conteúdo diretamente ao

usuário. De fato, a relação de proximidade entre gráficos *raster* e visualização de conteúdo digital é inerente, ao ponto de que representações alternativas de conteúdo gráfico normalmente requerem a conversão para gráficos *raster* (rasterização) antes que possam ser disponibilizadas a usuários.

2.2.2 Gráficos vetoriais

Em oposição a gráficos *raster*, que definem diretamente o valor de cada posição de uma imagem, gráficos vetoriais descrevem apenas alguns pontos presentes na imagem e a maneira como eles se conectam. É possível estabelecer que dois pontos estejam conectados por uma linha reta, ou uma curva polinomial de terceira ordem, por exemplo. Também é possível especificar a cor, largura e estilo (e.g. pontilhada, rajada), entre outras características da linha. Combinando esses elementos, se formam objetos visuais gradualmente mais complexos, desde simples polígonos e letras a objetos físicos completos.

Uma vantagem de representar uma imagem através de um gráfico vetorial é a resolução ser independente de escala. Ampliar ou diminuir a imagem não causa o surgimento de artefatos de pixelização (quando *pixels* individuais podem ser identificados em uma imagem), como é comum em imagens *raster*. Gráficos vetoriais também são capazes de representar imagens de maneira mais eficiente em termos de armazenamento: como a imagem não precisa ser descrita explicitamente em termos de cada região presente (i.e. gráficos vetoriais são uma representação esparsa, ao invés de densa), no total menos *bits* precisam ser escritos para representá-la.

No entanto, a principal desvantagem acerca de gráficos vetoriais decorre da natureza dos dispositivos que são usados para visualizar imagens. Em praticamente qualquer mídia eletrônica moderna, o princípio de funcionamento é inerentemente análogo a gráficos *raster*. Dessa maneira, mesmo uma imagem vetorial precisa ser convertida para uma imagem *raster* antes de poder ser visualizada na tela de algum dispositivo eletrônico.

2.3 Representação de volumes

Com o surgimento de conteúdo digital tridimensional, foi necessário o desenvolvimento de maneiras de representá-lo de maneiras que possibilitassem seu consumo direto por seres humanos. Nesta seção, são exploradas algumas maneiras de representar conteúdos dessa natureza comumente aplicadas atualmente.

2.3.1 Nuvens de pontos

Nuvens de pontos (*point clouds* ou PCs) são uma maneira de baixa complexidade e alta eficiência de captura, codificação e visualização de conteúdo tridimensional. Um determinado objeto é representado por nuvens de pontos listando-se, com a desejada precisão, cada posição espacial que é ocupada pelo objeto em questão. Se cada ponto p_i tem sua posição definida por uma tupla

$v_i = (x_i, y_i, z_i)$, a geometria de um objeto pode ser descrita por um conjunto V, tal que

$$V = \{v_1, v_2, ..., v_n\} = \begin{Bmatrix} (x_1, y_1, z_1) \\ (x_2, y_2, z_2) \\ \vdots \\ (x_n, y_n, z_n) \end{Bmatrix}. \tag{2.1}$$

É possível também listar de maneira similar algum atributo (por exemplo, cores, vetores normais ao ponto) que se deseja representar acerca do objeto. No caso de valores de cor representados no espaço RGB, cada ponto p_i tem sua cor determinada por uma tupla $c_i = (r_i, g_i, b_i)$, em que r_i, g_i e b_i são valores inteiros de 0 a 255, diretamente proporcionais à intensidade luminosa dos canais vermelho, verde e azul da imagem, respectivamente[1]. Chega-se então ao conjunto

$$C = \{c_1, c_2, ..., c_n\} = \begin{Bmatrix} (r_1, g_1, b_1) \\ (r_2, g_2, b_2) \\ \vdots \\ (r_n, g_n, b_n) \end{Bmatrix} \tag{2.2}$$

que, junto com o conjunto V, é capaz de formar uma representação visual completa de um objeto tridimensional.

É interessante notar que, baseando-se nas Equações 2.1 e 2.2, os conjuntos V e C podem facilmente ser representados através de notação matricial. Assim, obtém-se uma lista de pontos (coordenadas espaciais) e uma lista de atributos cujos itens são pareados um a um com os pontos da lista de posições. Isso resulta em uma representação esparsa do conteúdo visual do objeto de interesse. Em contraste, imagens *raster* são uma representação densa do conteúdo em questão, i.e. cada posição possível de ser representada no espaço visual tem um valor atribuído com alguma grandeza (por exemplo, cor).

Na maioria dos casos, conteúdo tridimensional representa apenas a superfície de objetos. Por isso, há uma grande redução dos requisitos de armazenamento e processamento ao se adotarem representações esparsas para esse tipo de conteúdo, já que a maior parte do volume espacial não é ocupado.

Quanto à representação das coordenadas espaciais dos pontos ocupados, não existe uma limitação prévia do formato que deve ser seguido. A maioria dos sistemas adota a representação através de coordenadas cartesianas com precisão de ponto flutuante. Este trabalho adota uma representação similar no sistema cartesiano, no entanto, optou-se por limitar o escopo à precisão inteira, fazendo uso do conceito de *voxels*.

2.4 Visualização de nuvens de pontos

Toda aplicação gráfica tem em comum uma etapa de visualização, ou renderização. A melhor maneira de visualizar um conteúdo depende da finalidade da aplicação. Isso é verdade principal-

[1] Considerando-se 8 bits de precisão por canal para a representação do sinal de cor.

mente para nuvens de pontos. Uma aplicação que requer o máximo de qualidade visual pode se beneficiar de uma abordagem baseada em *ray tracing*[13], em que os possíveis caminhos de raios de luz entre o objeto e o observador são exaustivamente explorados. No entanto, essa abordagem é extremamente custosa em termos computacionais. Aplicações que requerem desempenho e funcionamento rápido, como visualizações interativas em tempo real, ou *streaming*, precisam de abordagens alternativas.

Aplicações que permitem menor fidelidade visual em troca de desempenho costumam adotar abordagens baseadas em rasterização, ou seja, convertem diretamente os elementos de representação da nuvem de pontos em *pixels*. Aplicações baseadas em rasterização permitem que a visualização seja feita de maneira mais rápida e mais flexível. A seguir são discutidas duas maneiras de renderização baseadas em rasterização de nuvens de pontos, primeiro através de *splats* e em seguida através de *voxels*.

Em ambos os casos, independentemente da natureza do conteúdo, nessa etapa o conteúdo é fornecido ao usuário através de imagens bidimensionais. Isso revela uma possível relação subjacente, pelo menos em termos de maneira de consumo e percepção do conteúdo, entre conteúdos de natureza a princípio diferentes (bidimensional e tridimensional).

2.4.1 Renderização através de splats

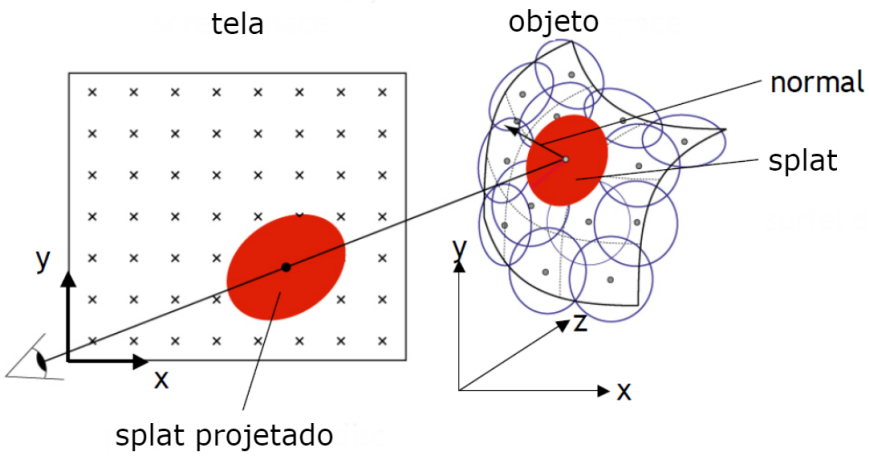
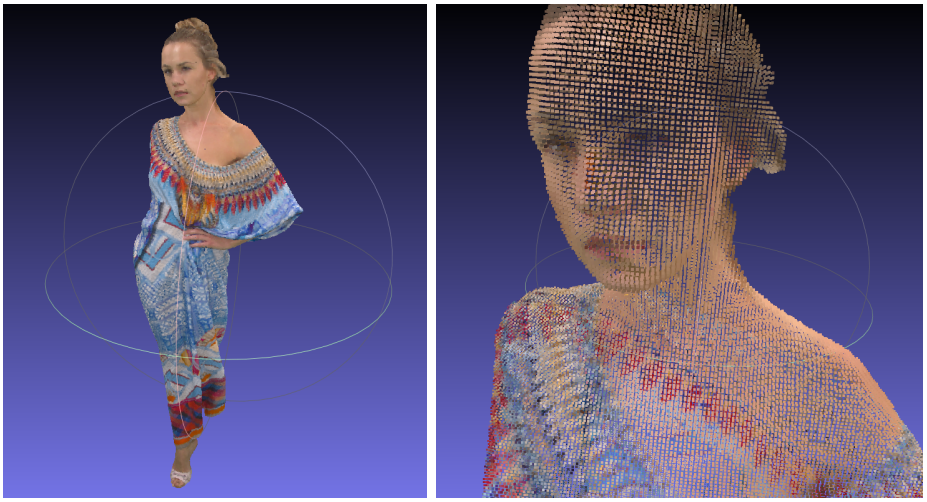


Figura 2.1: Exemplo de projeção e subsequente rasterização de um *splat*. Imagem disponível em http://www.cs.rug.nl/ roe/courses/acg/rendering

Algoritmos baseados em *splats* são alguns dos mais utilizados na visualização de nuvens de pontos. O elemento mínimo de uma nuvem de pontos, o ponto adimensional, é um objeto abstrato, sem volume ou área associados, tendo apenas posição definida. A visualização através de *splatting* tem a ideia de considerar um ponto como uma amostra de uma superfície orientada. Isto é, cada ponto pode estar associado a um objeto que, além de posição, tem área, formato, orientação e cor definidos, chamado de *splat*[14]. Quando pontos suficientes são tomados em conjunto, a união de seus respectivos *splats* forma uma descrição completa da superfície do objeto modelado.

Existem diversas maneiras possíveis de se definirem os *splats* respectivos de cada ponto em uma nuvem de pontos. Para se determinar a orientação do *splats*, pode-se levar em consideração a normal no ponto em questão, ou até as normais em uma dada vizinhança ao redor do mesmo. O mesmo pode ser feito para determinar o tamanho ou formato do *splats*, dependendo do objetivo da aplicação. Outra opção é adotar valores fixos e pré-determinados dos mesmos. Normalmente, *splats* são feitos usando círculos ou elipses, mas outros formatos são possíveis. Após uma representação adequada ter sido adotada, a representação é rasterizada, sendo convertida a uma projeção bidimensional da superfície do objeto, agora composta por *splats*. Esse processo é demonstrado na Figura 2.1.



(a) (b)

Figura 2.2: Nuvem de pontos visualizada através de *splats* quadrados, a partir de diferentes distâncias. Para um mesmo tamanho de elemento primitivo, à medida que se aproxima do modelo, lacunas surgem na geometria.

Em princípio, *splats* podem ocupar qualquer posição no espaço e ter qualquer orientação, independentemente de outros *splats* em sua vizinhança geométrica. Uma consequência disto é que, caso procedimentos adicionais não sejam incorporados no processo de visualização (como filtragem ou limitações às posições/orientações possíveis), artefatos desagradáveis, como buracos, *aliasing* e sobreposição de *splats*, podem surgir e prejudicar a qualidade percebida na imagem.

Especialmente, buracos e espaços entre *splats* consecutivos são perceptíveis. Esse tipo de artefato pode ser ainda exacerbado caso a aplicação de visualização não utilize *splats* com tamanhos que reagem ao nível de proximidade entre observador e imagem (i.e. *zoom*). Nesse caso, expandir a imagem torna os pontos relativamente mais distantes entre si, enquanto que o tamanho do *splats* na imagem continua o mesmo, causando o efeito de que o objeto fica cada vez menos denso e mais translúcido, até que pontos individuais conseguem ser distinguidos e o espaço entre eles é visto claramente. Esse efeito é demonstrado na Figura 2.2.

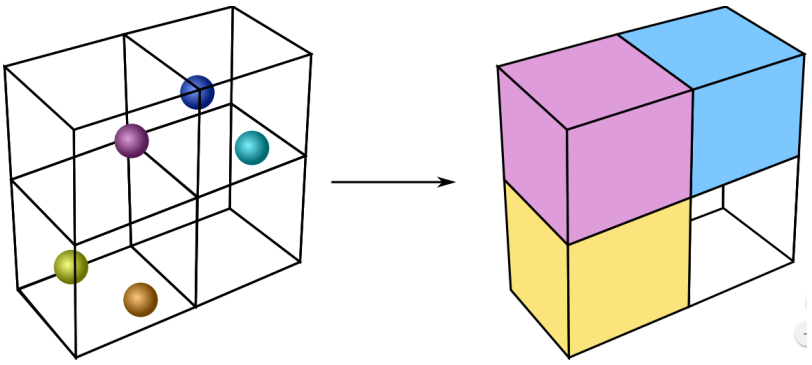


Figura 2.3: Exemplo de interpolação de cores durante o processo de voxelização

2.4.2 Renderização através de Voxels

Assim como imagens bidimensionais são compostas pela união de elementos mínimos (denominados de *pixels*) organizados regularmente em uma matriz, objetos volumétricos podem ser descritos em um espaço regularmente amostrado em *voxels* (do inglês *volume element*).

Para se representar conteúdo visual através de voxels, é preciso limitar o domínio espacial a um volume conhecido e estabelecer um nível de resolução geométrico. O mais comum é se trabalhar com volumes cúbicos com lados de dimensão igual a uma potência de 2, e resolução espacial igual a 1. Ou seja, cada voxel tem dimensões iguais a $1 \times 1 \times 1$ e ocupa uma posição inteira em uma grade tridimensional regular de dimensões $W \times W \times W$ (em que $W = 2^L$, $L \in \mathbb{N}$), capaz de conter até W^3 voxels.

É comum, no entanto, que inicialmente uma dada nuvem de pontos se encontre não voxelizada. Ou seja, seus pontos podem ocupar qualquer posição real no espaço tridimensional. O processo de representar tal nuvem de pontos através de *voxels*, denominado voxelização, é realizado percorrendo cada voxel do volume de representação e atribuindo ao mesmo um valor de cor dependendo dos pontos que ocupam posições contidas em seu volume. Uma descrição visual do processo se encontra na Figura 2.3. *Voxels* não ocupados não tem cor atribuída, equivalente a serem completamente transparentes. No caso de mais de um ponto se encontrar ocupando um mesmo *voxel*, a cor atribuída a tal *voxel* é calculada como a média dos pontos em seu interior.

É de especial interesse a restrição do conteúdo voxelizado a posições inteiras pela semelhança como imagens bidimensionais são representadas. Isso permite a visualização de maneira simples e rápida do conteúdo em um contexto bidimensional, através da projeção do conteúdo tridimensional.

2.5 Projeções

Apesar de humanos serem adaptados à vida em um ambiente com três dimensões espaciais, o aparato visual humano é inerentemente baseado em representações de duas dimensões desse mesmo ambiente, já que a própria luz é interceptada no olho humano pela superfície da retina [15]. De fato, como mencionado na Seção 2.4, mesmo conteúdos tridimensionais são disponibilizados em

formato bidimensional. Esse processo em que um objeto volumétrico é representado através de uma imagem bidimensional é denominado projeção.

De maneira geral, projeções são uma representação matemática, ou um mapeamento, de um conjunto a um subconjunto dele. Diversas técnicas de projeções existem, e podem incluir diversos campos matemáticos e aplicações. No entanto, o próprio conceito de projeção tem sua origem no ramo da geometria, que é o foco do presente estudo. Para este trabalho, dois tipos de projeção são mais relevantes: projeções perspectivas e projeções ortográficas.

2.5.1 Projeção perspectiva

Projeções perspectivas são o tipo de projeção mais similar ao o funcionamento da visão humana. Elas são obtidas traçando-se linhas de visão entre o objeto a ser presentado e um determinado ponto de vista virtual no espaço. Entre o objeto e o ponto de vista há um plano de projeção. Nos pontos do plano cruzados por cada uma das linhas de visão se armazena a imagem do ultimo ponto do objeto pelo qual aquela linha de visão passou. Isso resulta em uma imagem bidimensional da superfície tridimensional visível a partir do ponto vista escolhido, contida no plano de projeção.

Devido à característica convergente das linhas de visão em direção ao ponto de vista, projeções perspectivas tem o efeito de representarem objetos, ou partes de objetos, mais próximas ao plano de projeção com um tamanho aparente maior que o de objetos ou suas partes mais distantes. Apesar de ser natural observar esse comportamento no mundo real, imagens observadas dessa maneira podem distorcer algumas características do objeto original de maneiras indesejadas. Para este estudo, optou-se pelo uso de outra opção de projeção, as projeções ortográficas.

2.5.2 Projeção ortográfica

Projeções ortográficas são um tipo de projeção paralela. Neste tipo de projeção as linhas de visão são traçadas paralelas entre si, diferente de projeções perspectivas, em que linhas de visão convergem para um ponto de vista virtual. Em projeções ortográficas, as linhas são traçadas perpendiculares ao plano de projeção escolhido.

Essa característica é especialmente interessante quando a projeção ortogonal é usada para se visualizar objetos compostos por *voxels*. Caso o plano de projeção seja paralelo a alguma das faces dos *voxels*, existe uma relação direta entre os *pixels* da imagem projetada e os *voxels* visíveis no objeto original. Isso é ilustrado na Figura 2.4, em que cada uma das projeções mostradas coincide com uma face de um cubo em volta do objeto visualizado. Determinar o *voxel* correspondente a um pixel de coordenadas conhecidas também é simples, bastando encontrar o *voxel* com as mesmas coordenadas cuja terceira coordenada (no eixo perpendicular ao plano de projeção) é a mais próxima desse plano. Por exemplo, caso se assuma que o plano contenha os eixos x e y e esteja localizado na posição 0 do eixo z, o pixel na posição (x_n, y_n) é uma imagem do *voxel* de coordenadas (x_n, y_n, z) com o menor z.

Outra característica interessante de projeções ortográficas é que elas equivalem a projeções

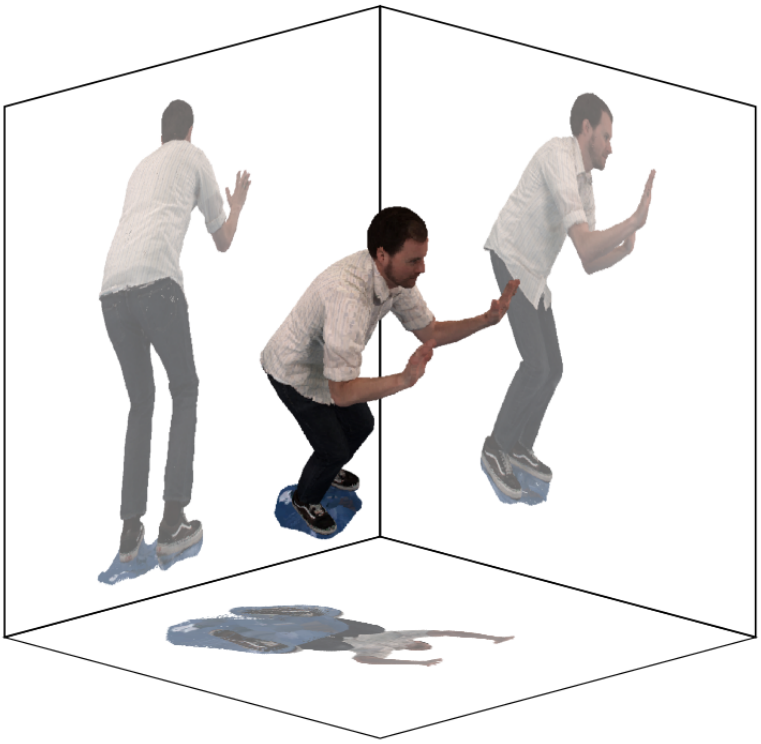


Figura 2.4: Projeções ortográficas visualizadas quando os planos de projeção se encontram em direções ortogonais entre si

perspectivas em que o ponto de vista se localiza a uma distância infinita do objeto sendo projetado. Ou seja, quando observados de distâncias cada vez maiores, objetos tendem de uma projeção perspectiva a uma projeção ortográfica.

2.6 Espaços de cores

Tanto imagens bidimensionais como conteúdos tridimensionais dependem de um sistema de representação de cores definido para que informações de cor possam ser transmitidas ou armazenadas.

O sistema visual humano envolve mais variáveis do que apenas a intensidade e o comprimento de onda da luz que atinge os receptores localizados nos olhos. Fatores como a diferença de luminosidade entre partes de uma imagem, velocidade de movimento, posição relativa entre objeto e observador podem afetar como uma pessoa percebe cor ou luz [16, 17].

No entanto, na maioria dos casos, modelos simples são suficientes para oferecer conteúdo visual em diversas mídias. Logo após o surgimento das primeiras fotografias, por volta do início do século XIX, já havia pesquisa no campo de fotografia em cores, com resultados experimentais sendo obtidos já desde 1840 [18].

Tentativas iniciais de se reproduzir cores foram baseadas principalmente em projetar luz sobre

anteparos preparados com substâncias químicas capazes de reagir a diferentes cores do espectro visual. No entanto, já em 1860 havia-se percebido que o aparato visual humano não dispõe de sensores para cada valor possível do espectro de luz visível e que, com a mistura de apenas algumas cores específicas, era possível representar, se não todas, a maior parte das cores que seres humanos são capazes de perceber [19].

Isso levou tanto a indústria visual e a comunidade acadêmica à representação de cores através de sistemas denominados espaços de cores. Um espaço de cor é um modelo abstrato em que cada cor possível é representada através de tuplas de números. Cada elemento da tupla indica a quantidade de um certo componente de cor que o sinal de cor em questão apresenta.

A modelagem através de espaços de cores oferece ferramentas matemáticas para se lidar com imagens. Cores individuais podem ser tratadas como pontos em um sistema de coordenadas específico, e sistemas alternativos mais convenientes para determinadas tarefas podem ser usados, com a possibilidade de se alternar livremente entre diferentes espaços de cores, tanto em uma como em outra direção [20]. Além disso, valores respectivos a cada uma das coordenadas podem ser tratados individualmente e independentemente das outras coordenadas usadas para descrever a imagem. A cada uma dessas coordenadas é dado o nome de canal (e.g. canal vermelho, no espaço RGB). A seguir são discutidos dois espaços de cores mais relevantes para a representação de cores de conteúdos tridimensionais através de nuvens de pontos.

2.6.1 Espaço RGB

O espaço de cores RGB é um sistema de representação de cores através da combinação aditiva de três componentes primários de cor, independentes de nível de luminosidade: vermelho, verde e azul [21]. Variando as quantidades de cada componente de cor, é possível representar qualquer tom de cor entre os componentes primários utilizados, além de suas variações de luminosidade entre branco puro e preto puro.

O motivo da escolha das cores vermelho, verde e azul está relacionado com os mecanismos de visão que ocorrem no olho humano, no nível celular. Apesar de existirem duas teorias complementares que descrevem esse processo em maior detalhe (teoria tricromática [22] e teoria do processo oponente [23]), o espaço RGB está ligado principalmente com conceitos da teoria tricromática, que descreve um primeiro estágio da visão humana [24].

Na retina, parte do olho humano que converte luz em impulsos elétricos neurológicos, há a ocorrência de dois tipos especializados de células: bastonetes, mais sensíveis a luz em luminosidades baixas (independentemente de comprimento de onda), e cones, que são excitados em função, além da intensidade, do comprimento de onda de luz incidente. Se observa também que há uma subdivisão dos cones em outras três especializações: cones dos tipos S, M e L [25, 26].

Cones do tipo S demonstram serem excitados principalmente na faixa entre 400 nm e 500 nm. Já cones M são excitados principalmente entre 450 nm e 630 nm, e cones L entre 500 nm e 700 nm. Essas faixas não apresentam limites nítidos, e há sobreposições consideráveis entre elas, principalmente entre os cones M e L.

Historicamente, os cones dos tipo S, M e L passaram a ser associados com cores específicas, respectivamente azul, verde e vermelho, mesmo em face de apresentarem um comportamento mais complexo do que simplesmente serem excitados por essas cores específicas. Especialmente, cones do tipo L demonstram um pico de excitação para cores mais próximas de amarelo-esverdeado do que para o vermelho. No entanto, a nomenclatura não é totalmente sem justificativa. De fato, cada uma das cores do padrão RGB está associada com o cone que apresenta maior sensibilidade a ela.

2.6.2 Espaço YUV

O espaço de cores YUV é uma representação alternativa ao RGB em que um canal é usado para representar exclusivamente a luminância (ou luma) presente na imagem, enquanto dois outros representam a crominância azul e vermelha, respectivamente. Isto é, um dos canais descreve completamente a intensidade luminosa (ou brilho) da imagem, enquanto os outros dois, em combinação, representam o tom de cor da imagem.

Existem diversas variações do padrão YUV, além de outros espaços de cores similares, como Y'UV, YCbCr e Y'CbCr, sendo comum ocorrer alguma confusão na nomenclatura desses sistemas, com alguns nomes sendo usados de maneira intercambiável em alguns contextos. A presença do símbolo apóstrofo (') seguido ao símbolo de um canal denota que aquele canal passa por compressão (ou correção) gama, que é uma escala não linear que tem por objetivo aproximar a percepção humana de diferença luminosa [21].

Especialmente, quando o canal Y não passa por compressão gama (escala linear de intensidade luminosa) ele é denominado de luminância. Já quando ocorre compressão gama (canal Y') o canal é denoninado de luma. Sendo assim, a única diferença entre os sistemas YUV e Y'UV é referente ao canal de intensidade luminosa. A mesma diferença ocorre entre os sistemas YCbCr e Y'CbCr, com o canal Y sendo idêntico entre os sistemas YUV e YCbCr.

Quanto aos canais de crominância, no sistema YUV (e Y'UV), os canais U e V são definidos, respectivamente, como a diferença entre o valor de azul e de intensidade luminosa, e a diferença entre o valor de vermelho e de intensidade luminosa. Já nos sistemas YCbCr (e Y'CbCr) os canais Cb e Cr são obtidos através do desvio da cor cinza no eixo azul-amarelo e no eixo vermelho-ciano, respectivamente.

Os sistemas de cores YCbCr, YUV e suas variantes adotam ideias compatíveis com a etapa da visão humana denominada de processo oponente. Após a aquisição de luz colorida através das células da retina, antes de ser transmitido pelo nervo ótico, o sinal visual gerado no olho é processado por neurônios especializados que tem ativações reguladas pelas diferenças de excitação que os bastonetes e cada um dos tipos de cones apresentam [27].

O processamento dos sinais gerados pelos cones envolve principalmente dois tipos de neurônios: as células retinais bipolares e as células retinais ganglionares. Células bipolares efetivamente regulam os sinais emitidos por cones e bastonetes, e os transmitem às células ganglionares, que processam diferenças de contraste ou cor ao longo do tempo ou do campo visual [28]. Parte dos

neurônios ganglionares é excitada por cones L e S mas inibida por cones M (diferenças no eixo vermelho-verde), enquanto que outra parte é excitada pelos cones L e M mas inibidos por cones S (diferenças no eixo azul-amarelo), de maneira similar ao sistema YUV.

Adotar um modelo de representação de cores com funcionamento próximo ao exibido pelo organismo humano permite que algoritmos de compressão se aproveitem do fato de que nem todos os detalhes de uma imagem são igualmente percebidos por seres humanos. Assim, é possível descartar informação pouco perceptível e alocar mais recursos para representar detalhes mais relevantes, resultando em uma qualidade observada maior, a uma taxa de bits menor.

A conversão do espaço RGB para YUV pode ser feita usando as seguintes fórmulas:

$$Y' = W_R R + W_G G + W_B B \tag{2.3}$$

$$U = U_{\text{MAX}} \frac{B - Y'}{1 - W_B} \tag{2.4}$$

$$V = V_{\text{MAX}} \frac{R - Y'}{1 - W_R} \tag{2.5}$$

em que, de acordo com o padrão BT.601 [20],

$$W_R = 0.299 \tag{2.6}$$

$$W_B = 0.114 \tag{2.7}$$

$$W_G = 1 - W_R - W_B = 0.587 \tag{2.8}$$

$$U_{\text{MAX}} = 0.436 \tag{2.9}$$

$$V_{\text{MAX}} = 0.615 \tag{2.10}$$

Adotando uma representação matricial, tem-se que

$$\begin{bmatrix} Y' \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.14713 & -0.28886 & 0.436 \\ 0.615 & -0.51499 & -0.10001 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.11}$$

e que, inversamente,

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.13983 \\ 1 & -0.39465 & -0.58060 \\ 1 & 2.03211 & 0 \end{bmatrix} \begin{bmatrix} Y' \\ U \\ V \end{bmatrix} \tag{2.12}$$

.

2.7 Codificação de nuvens de pontos

Como foi mencionado na Seção 2.6, e principalmente na Seção 2.6.2, existe a possibilidade de se escolher a representação de um conteúdo de maneira a dedicar mais informação para representar

informações mais relevantes para a percepção humana de qualidade, enquanto que detalhes menos importantes são descartados.

O mesmo pode ser feito com informação respectiva a geometria de modelos de objetos. Intuitivamente, deseja-se uma representação em que seja possível escolher o nível de detalhamento da estrutura geométrica da representação. Deve ser possível gradualmente descartar detalhes mais finos, ao mesmo tempo que a estrutura geral não é descaracterizada de maneira demasiada.

A seguir, são exploradas as *octrees*, uma estrutura comumente usada para a codificação de nuvens de pontos, que oferece a capacidade representar níveis graduais de detalhe.

2.7.1 Octrees

No contexto de ciência da computação, uma árvore é uma estrutura de dados em que elementos (ou nós) se relacionam hierarquicamente entre si através elos. Árvores são compostas por um nó inicial, denominado raiz, acima de todos os outros, além de seus nós filhos. Cada nó subsequente pode ter um ou mais filhos. Quando um nó não tem nenhum filho, ele é denominado de nó folha [29].

Octrees são um tipo específico de árvore em que cada nó tem exatamente 0 ou 8 filhos [30]. Essa característica é interessante para aplicações relacionadas com geometrias tridimensionais. Como $8 = 2^3$, e em um espaço tridimensional existem três direções ortogonais, ao se subdividir o espaço em dois, na direção de cada uma de suas coordenadas, obtém-se 8 octantes. Dessa maneira, é possível facilmente representar a geometria de objetos contidos na região delimitada pela união desses octantes através de uma estrutura baseada em *octrees*.

De maneira mais específica, se cada nó da *octree* representa um octante, um valor binário designado ao nó em questão indica se aquele octante está ocupado ou não. O octante pode ser então subdividido em mais 8 octantes, se repetindo o processo para determinar cada uma de suas regiões ocupadas. No caso de um dos sub-octantes se encontrar não ocupado, o presente nó é considerado como não tendo nenhum filho (nó folha). Ao final do processo, cada nível da *octree* é representado por um byte, cujo cada bit indica se o n-ésimo octante estava ocupado. O byte seguinte representa essa mesma informação respectiva ao primeiro bit ocupado do nível anterior da *octree*, e assim por diante, até o último bit ocupado do nível anterior. Quando todos os bits de um nível foram considerados, o processo continua para os bits e bytes dos próximos níveis. A ordem de percorrimento da *octree* é arbitrária, bastando apenas que o codificador e o decodificador estejam de acordo quanto a ela.

2.8 Métricas objetivas de qualidade

A qualidade visual de mídias geralmente é avaliada através do uso ou de métricas subjetivas ou de métricas objetivas. Avaliações subjetivas consomem muito tempo e são caras. Devido a isso, fazem-se necessárias métricas objetivas eficientes, que consigam prever com exatidão a qualidade de algum conteúdo, ou o nível de distorção ao qual ele está sujeito. No caso de nuvens de pontos,

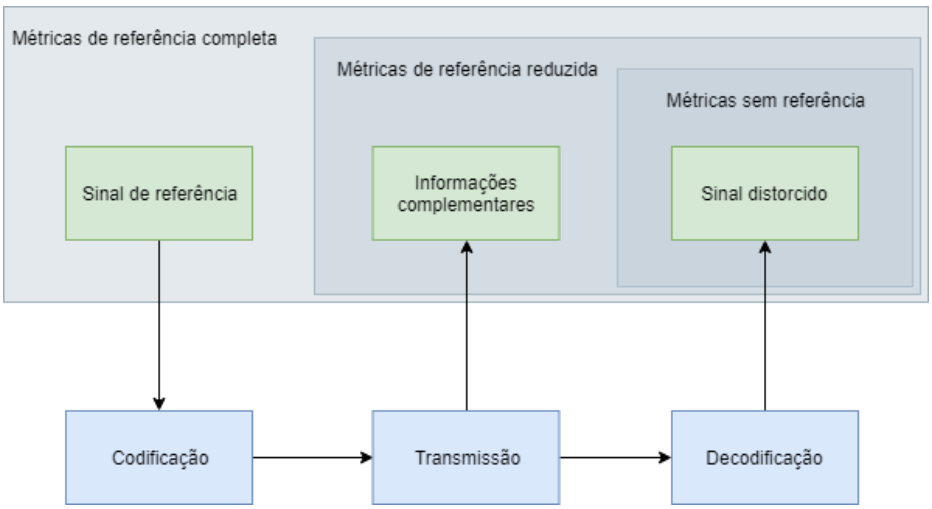


Figura 2.5: Diagrama demonstrando o fluxo de informação entre diferentes categorias de métricas objetivas de qualidade.

avaliação objetiva de qualidade ainda é um problema aberto [10].

Métricas objetivas podem ser classificadas em três categorias distintas, em função da quantidade de informação disponível acerca dos conteúdos envolvidos na análise. Um diagrama demonstrando a relação entre o compartilhamento de informação em cada uma dessas categorias pode ser observado na Figura 2.5.

Métricas de referência completa avaliam a qualidade de um sinal recebido (que passou por algum processo de distorção) através de suas diferenças em relação a ao sinal original, antes de sofrer modificações, denominado sinal de referência. Este método assume que todas as informações acerca de ambos os sinais (ou pelo menos os dois sinais em sua íntegra) estão disponíveis. Em alguns casos, isso pode se tornar um impedimento.

Métricas de referência reduzida podem utilizar informações de ambos os sinais, mas não é necessário que eles sejam utilizados inteiramente. Em casos em que ter acesso completo a algum dos sinais é impossível ou impraticável, métodos dessa natureza permitem alguma mensuração de qualidade, ainda que com exatidão reduzida. Estes métodos também costumam ser computacionalmente mais eficientes que métodos de referência completa, já que precisam de menos dados.

Existem ainda métricas sem referência. Esses modelos tem o objetivo de estimar a qualidade de sinais distorcidos sem o uso de qualquer informação acerca do sinal original. Normalmente, são observadas características internas do sinal recebido, como a variação de *pixels*, ou dados acerca da transmissão em si, como vetores de movimento, parâmetros de quantização e outros metadados, ou ainda uma combinação dessas informações para determinar se há ocorrência de artefatos desagradáveis no conteúdo recebido. Esses métodos costumam ser os mais rápidos, sendo que algumas variações não requerem nem a codificação do sinal recebido. No entanto, essas métricas também são as que oferecem o menor poder preditivo de qualidade [31].

Avaliação objetiva da qualidade de nuvens de pontos é geralmente realizada através de métricas de referência completa. Métricas da distorção da cor de nuvens de pontos são baseadas em métricas

convencionais aplicadas a conteúdos bidimensionais. Já o estado da arte de métricas de referência completa para a avaliação distorções geométricas de nuvens de pontos podem ser separadas em duas categorias: as baseadas em distância e as baseadas em normais. Apenas um tipo de métrica é classificado atualmente como baseado em normais, as denominadas métricas plano a plano. Já as métricas classificadas como baseadas em distância consistem nos seguintes tipos: métricas ponto a ponto, métricas ponto a plano, e métricas ponto a malha. Cada um desses tipos de métricas é explorado na Seção 2.8.1 e na Seção 2.8.2.

Tanto métricas de degradação de geometria e de cor costumam calcular suas medidas de erro de maneira simétrica. Isto é, o erro é obtido calculando-se primeiro com um dos dois conteúdos utilizados na métrica (ou a versão original da nuvem de pontos ou sua versão distorcida) como referência e o outro como conteúdo de teste. O primeiro valor calculado é armazenado, e o cálculo é feito novamente com as duas versões da nuvem de pontos trocadas: se primeiro a versão original foi adotada como referência e a versão distorcida foi tida como conteúdo sob teste, agora a versão original será o teste, enquanto que a versão distorcida é a referência, e vice-versa. O valor de erro final é escolhido como o valor máximo entre os dois valores de erro calculados.

2.8.1 Métricas de cor baseadas em pontos

Métricas de distorção de cor em nuvens de pontos são realizadas se associando pontos do conteúdo sob análise com seus respectivos pontos no conteúdo de referência. Tipicamente para isso se usa o algoritmo de busca do vizinho mais próximo.

Em seguida, se calcula a degradação de cor como se cada tupla de cor de pontos correspondentes tivessem a mesma relação de *pixels* de pares de imagens bidimensinais em métricas de referência completa convencionais, como as descritas na Seção 2.8.3. No entanto, como não são necessariamente levadas em consideração relações de proximidade no ordenamento dos pontos, é mais comum o uso de métricas que atuam somente na escala de *pixels* individuais (em oposição a métricas como o SSIM, que considera regiões da imagem em seu cálculo).

É possível, por exemplo, utilizar a PSNR para calcular a distorção de cor dessa maneira. Pode-se usar tanto valores de cor no espaço RGB ou em qualquer outro que a aplicação exigir. Utilizando o espaço YUV no padrão ITU-R, recomendação BT.709-3 [32], uma medida do erro de cor é calculada através de uma média ponderada das diferenças nos canais de luma e crominância [33]:

$$\mathrm{PSNR}_{\mathrm{YUV}} = \left(6 \cdot \mathrm{PSNR}_{\mathrm{Y}} + \mathrm{PSNR}_{\mathrm{U}} + \mathrm{PSNR}_{\mathrm{V}}\right)/8. \tag{2.13}$$

2.8.2 Métricas de geometria baseadas em pontos

As métricas discutidas a seguir são baseadas no cálculo de medidas de erro individuais para cada ponto presente na nuvem de pontos sob análise. Para se obter valor referente à degradação geométrica da nuvem de pontos como um conjunto, é preciso calcular algum valor através dos erros individuais obtidos. Opções comuns são o erro total (soma de todos os erros individuais), o erro

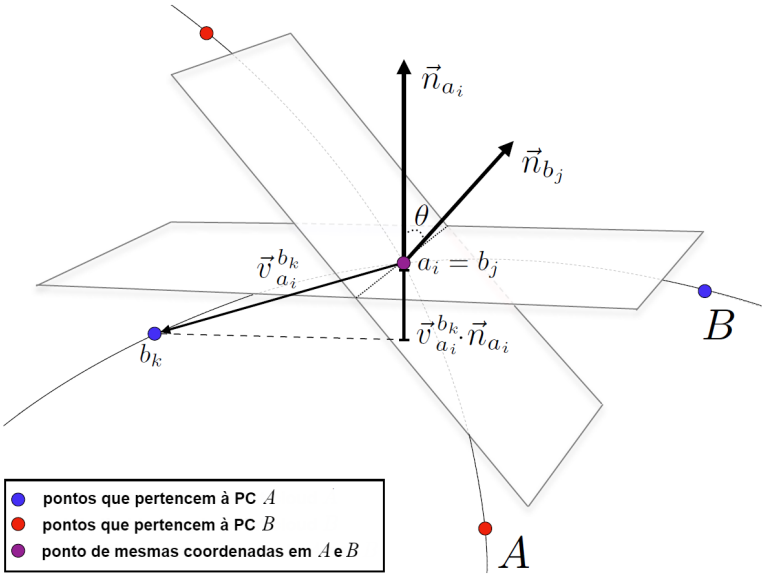


Figura 2.6: Relação entre os pontos de interesse em duas PCs durante o cálculo de métricas baseadas em geometria.

quadrático médio (MSE), a raiz do erro quadrático médio (RMSE), ou a distância de Hausdorff [34], esta última sendo definida como

$$d_{\mathbb{H}}(X,Y) = \max\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(y,x)\}, \tag{2.14}$$

em que $d(x,y)$ é alguma medida de distância entre dois elementos dos conjuntos X e Y, e sup e inf denotam respectivamente o *supremum* e o *infimum* de um subconjunto em relação ao conjunto em que está contido. No entanto, para conjuntos ordenados e finitos o *infimum* e o *supremum* coincidem respectivamente como o elemento mínimo e o elemento máximo do subconjunto sob análise [35]. Dessa maneira, no presente contexto a distância de Hausdorff pode ser intuitivamente compreendida como a maior distância observada entre cada ponto das duas nuvens de pontos e seu respectivo par mais próximo (dada uma métrica d de distância) na outra nuvem de pontos.

2.8.2.1 Métrica ponto a ponto

Métricas ponto a ponto são baseadas na distância geométrica entre pontos associados dos conteúdos sob análise e de referência. Geralmente o valor de erro é relacionado com o deslocamento do ponto do conteúdo sob análise em relação a seu vizinho mais próximo no conteúdo de referência. Ou seja, seguindo a nomenclatura presente na Figura 2.6, para cada ponto b_k do conteúdo sob análise (B), seu vizinho mais próximo, a_i no conteúdo de referência (A) é selecionado. Em seguida, alguma métrica de distância entre os dois pontos, normalmente a distância euclideana, é calculada:

$$\mathrm{E}(b_k, a_i) = \left\| \overrightarrow{v}_{a_i}^{b_k} \right\|_2 . \tag{2.15}$$

O erro referente à nuvem de pontos B pode então ser calculado como a soma, ou a média, das distâncias entre todos os pontos do conteúdo analisado e seus respectivos vizinhos mais próximos

em A:

$$\mathrm{MSE}(B, A) = \frac{\sum_{b_k \in B} \mathrm{E}(b_k, a_i)}{|B|}, \tag{2.16}$$

em que $|B|$ é a cardinalidade (número de elementos) do conjunto B.

2.8.2.2 Métrica ponto a plano

Métricas ponto a plano são baseadas no erro projeção de um ponto, que pertence a um conteúdo sob análise, em relação ao vetor normal de um ponto associado no conteúdo de referência. Isto é, após se identificar, para cada ponto b_k no conteúdo analisado (B) seu vizinho mais próximo, a_i, no conteúdo de referência (A), o erro é projetado sobre a normal $\overrightarrow{n}_{a_i}$ através da fórmula

$$\hat{\mathrm{E}}(b_k, a_i) = \overrightarrow{n}_{a_i} \cdot \overrightarrow{v}_{a_i}^{b_k}, \tag{2.17}$$

novamente podendo-se obter o valor médio do erro através de

$$\mathrm{MSE}(B, A) = \frac{\sum_{b_k \in B} \hat{\mathrm{E}}^2(b_k, a_i)}{|B|}, \tag{2.18}$$

com ambas equações seguindo a nomenclatura da Figura 2.6.

A interpretação por trás da métrica ponto a plano é baseada no fato de que custos maiores ocorrem devido a pontos que desviam da superfície local aproximada do objeto de referência. Essa métrica requer que pelo menos um dos conteúdos tenha normais conhecidas. No caso das normais do conteúdo de referência sejam conhecidas, o cálculo da métrica se dá normalmente. Caso apenas os vetores normais de um dos conteúdos sejam conhecidos e deseje-se usar o conteúdo sem normais como referência, ainda é possível utilizar essa métrica estimando-se as normais do conteúdo sem os vetores calculando-se a média dos vetores presentes nos vizinhos mais próximos correspondentes no outro conteúdo.

2.8.2.3 Métrica plano a plano

Métricas plano a plano são baseadas na similaridade angular de planos tangentes que correspondem a pontos associados entre a referência e o conteúdo sob análise. O valor de erro oferece uma aproximação da dissimilaridade entre superfícies locais correspondentes. Para cada ponto b_j que pertence ao conteúdo sob análise (B), seu vizinho mais próximo, a_i no conteúdo de referência (A) é identificado. Através dos vetores normais correspondentes a cada um dos pontos, pode-se calcular a similaridade angular dos planos tangentes aos mesmos. Isso é realizado calculando-se o ângulo $\hat{\theta}$ entre os vetores normais $\overrightarrow{n}_{b_j}$ e $\overrightarrow{n}_{a_i}$. O ângulo efetivo adotado é restringido ao menor dos dois ângulos entre as duas normais, de modo que

$$\theta = \min\{\hat{\theta}, \pi - \hat{\theta}\}, \tag{2.19}$$

com π em radianos.

A similaridade angular $\text{sim}(\theta)$ é então calculada como

$$\text{sim}(\theta) = 1 - \frac{2\theta}{\pi}, \tag{2.20}$$

sendo a imagem da função limitada ao conjunto fechado $[0, 1]$. Finalmente, alguma média das similaridades individuais pode ser calculada, de maneira similar como foi feito nas Equações 2.16 e 2.18.

Esta métrica é baseada na premissa de que o sistema visual humano naturalmente interpola um conjunto de pontos visualizado para inferir o objeto em questão. O plano tangente serve como uma aproximação linear da superfície local do conteúdo. Portanto, a similaridade angular entre planos tangentes de pontos associados entre o conteúdo analisado e o conteúdo de referência oferece uma aproximação da dissimilaridade entre superfícies locais correspondentes entre os dois objetos.

Uma desvantagem dessa métrica é que ela requer que as normais, tanto do objeto de referência como do objeto sob análise, sejam conhecidas. Caso não estejam disponíveis, os vetores precisam ser estimados. Assim, o desempenho desta métrica, tanto em termos computacionais como em termos de correlação com qualidade subjetiva observada, fica limitado ao desempenho do algoritmo de estimativa de normais utilizado.

2.8.2.4 Métrica ponto a malha

Métricas ponto a malha envolvem a representação das nuvens de ponto de interesse através de malhas poligonais (*meshes*), um processo denominado neste contexto de reconstrução de superfície. Inicialmente, o conteúdo de referência é reconstruído através de um *mesh*. Em seguida, para cada ponto do conteúdo de teste, é calculada a menor distância para a superfície mais próxima do *mesh* de referência. Considerando que não existe uma maneira única de se gerar uma malha de um conjunto de pontos, as notas objetivas obtidas dependem consideravelmente do algoritmo de reconstrução de superfície selecionado. Assim, métricas ponto a malha são consideradas soluções sub-ótimas para a avaliação de qualidade de nuvens de pontos. Neste trabalho, tais métricas não serão mais investigadas daqui em diante.

2.8.3 Métricas baseadas em imagens

Ao longo das Seções 2.2, 2.5 e 2.6, pôde-se notar que o desenvolvimento das mídias visuais modernas está intimamente ligado ao funcionamento da visão no organismo humano. Adicionalmente, como também foi mencionado nas Seções 2.4 e 2.5, devido à própria natureza do sistema visual humano, nuvens de pontos (e outros conteúdos naturalmente tridimensionais) ainda são ligados a conceitos e aspectos de imagens bidimensionais.

Em luz dessas observações, é intuitivo considerar que a análise da qualidade visual de algum conteúdo, especialmente, esteja conectada com a fisiologia humana, e que medidas dessa qualidade se beneficiem de maior correlação com a real qualidade subjetiva percebida à medida que elas se tornam cada vez mais embasadas no funcionamento do sistema visual humano. Essa premissa leva a crer que, assim como parte da visão humana se relaciona com imagens bidimensionais projetadas,

basear a avaliação qualidade de conteúdos tridimensionais na projeção e no subsequente tratamento das imagens resultantes pode ser vantajoso.

A seguir, são exploradas algumas métricas classicamente utilizadas na análise de imagens bidimensionais.

2.8.3.1 PSNR

A relação sinal-ruído de pico (*peak signal-to-noise ratio* - PSNR) é uma das métricas de qualidade de sinais mais utilizadas, principalmente no contexto de processamento de imagens. Ela foi uma das primeiras métricas capaz de traduzir diferenças entre sinais ou imagens para uma escala objetiva e de fácil comparação entre conteúdos.

A PSNR é definida através da fórmula

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}} \right), \tag{2.21}$$

em que MAX é o valor máximo na escala adotada que o sinal pode assumir (255 para imagens de 8 bits, por exemplo), e MSE é o erro quadrático médio entre o sinal de referência e o sinal analisado. Para uma imagem de referência R e sua aproximação ruidosa I, ele é calculado como

$$\mathrm{MSE} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} [R(i,j) - I(i,j)]^2. \tag{2.22}$$

Mensurar a proporção entre o erro médio e o sinal de pico garante que diferenças na escala de sinais não afetam a métrica. Além disso, como é comum sinais apresentarem uma faixa dinâmica ampla, adotar a escala logarítmica proporciona uma métrica com valores em um intervalo mais conveniente e gerenciável.

No entanto, a PSNR apresenta uma grande variação, mesmo para conteúdos similares. A PSNR é, por exemplo, consideravelmente sensível a translações espaciais: deslocamentos da ordem de um pixel entre uma imagem e sua referência já são suficientes para provocar uma queda da PSNR. Ademais, nem sempre ocorre uma correlação direta entre PSNR e qualidade observada. É possível que conteúdos praticamente idênticos tenham PSNRs consideravelmente diferentes e, em certas situações, imagens com melhor qualidade subjetiva podem apresentar PSNR piores do que uma imagem com artefatos de distorções mais perceptíveis. Para evitar esse tipo de comportamento, é importante limitar a comparação através de PSNR a apenas imagens com conteúdos similares e distorcidas por procedimentos de mesma natureza.

2.8.3.2 SSIM

O índice de similaridade estrutural (*structural similarity index* - SSIM) foi proposto como uma melhoria em relação à PSNR, capaz de prever com melhor acurácia a qualidade percebida do conteúdo medido. O SSIM propõe calcular a similaridade entre pares de imagens através

medidas intuitivamente relacionadas com características perceptuais que, se preservadas, espera-se observar uma qualidade visual maior do que caso contrário. Especificamente, essas grandezas são denominadas luminância ($l(x,y)$), contraste ($c(x,y)$) e estrutura ($s(x,y)$), para uma imagem distorcida y e sua referência x. Elas são definidas como

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}, \tag{2.23}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \tag{2.24}$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}, \tag{2.25}$$

em que μ_x, μ_y, σ_x^2, σ_y^2, σ_{xy} são, respectivamente, o valor médio dos valores na imagem x, o valor médio dos valores na imagem y, a variância dos valores da imagem x, a variância dos valores da imagem y e a covariância entre os valores das imagem x e y. Além disso, $c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$ e $c_3 = c_2/2$, três coeficientes que estabilizam a divisão em casos em que o denominador é muito pequeno, e L é o intervalo dinâmico de valores possíveis nas imagens. Em imagens de 8 bits, $L = 255$. Em geral k_1 e k_2 são escolhidos como 0.01 e 0.03, respectivamente.

O SSIM é proporcional à média geométrica ponderada das medidas obtidas de luminância, contraste e estrutura. Assim,

$$\mathrm{SSIM}(x,y) = l(x,y)^\alpha \times c(x,y)^\beta \times s(x,y)^\gamma, \tag{2.26}$$

com α, β e γ arbitrários. Caso sejam escolhidos todos como iguais a 1,

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}. \tag{2.27}$$

O SSIM costuma ser calculado não sobre a imagem como um todo, mas aplicado a janelas que cobrem subregiões da imagem completa. Podem-se escolher janelas de qualquer tamanho, que podem ser tomadas através de deslocamentos sucessivos de um ou mais pixels, até que toda a imagem seja coberta. O SSIM pode ser calculado para apenas um dos canais presentes na imagem, normalmente o canal de luma, ou para mais canais. Nesse caso, o SSIM é calculado separadamente para cada canal, e a métrica total da imagem é obtida através de uma média entre os índices de cada canal.

2.8.3.3 VIF e VIFP

O índice de fidelidade de informação visual (VIF), e sua realização baseada em *pixels*, o VIFP, são métricas de referência completa para a avaliação de qualidade em imagens. O VIF é baseado nas chamadas estatísticas de cena natural (NSS) e na noção de como o sistema visual humano extrai informação de imagens. É adotado um critério de fidelidade que quantifica a informação de Shannon [36] compartilhada entre as imagens distorcida e de referência, relativamente à informação

contida na imagem em si. Dessa maneira, são adotados três modelos em conjunto para a obtenção do VIF: um modelo de NSS, um modelo de degradação de imagem e um modelo do sistema visual humano [37].

Cenas naturais se referem ao conjunto de informações presentes em um ambiente físico que são percebidas por agentes através de seus sentidos, especificamente aqueles que seres humanos estão acostumados a observar no modo natural de operação de seus órgãos sensoriais [38]. Esse conjunto de cenas pode incluir ambientes como ruas em uma cidade, o interior de uma casa ou plantas em um jardim, por exemplo. Tem-se interesse principalmente nos aspectos visuais de uma cena. Apesar de serem uma fração pequena dos sinais visuais possíveis, cenas naturais formam uma grande parte da mídia consumida, e modelos estatísticos robustos já foram desenvolvidos para modelar essa classe de sinais.

A maior parte dos processos de distorção observados em sistemas reais modifica essas estatísticas e torna as cenas não naturais. Medir esse desvio estatístico, através da quantidade da informação compartilhada entre o sinal distorcido e o sinal de referência, portanto, pode indicar a qualidade observada em conteúdos que se encaixem nesse modelo. Além disso, é possível determinar a quantidade de informação total presente na imagem de referência. Dessa maneira, pode-se calcular a perda de informação relativa à quantidade de informação originalmente presente.

No contexto do índice VIF, imagens de cenas naturais perfeitas (sem qualquer distorção ou ruído) são modeladas como uma fonte estocástica, especificamente o modelo de mistura de escala gaussiana (GSM) no domínio da transformada Wavelet, que então é distorcida por um canal (operador de distorção), fornecendo as imagens a serem avaliadas.

Campos aleatórios (RFs) são generalizações de processos estocásticos que, em vez de serem parametrizados por um índice unidimensional (seja discreto ou contínuo), são parametrizados por vetores multidimensionais, ou pontos em uma superfície multidimensional [39]. Um GSM é um RF que pode ser expressado como o produto de dois RFs independentes [40]. Ou seja, um GSM $\mathcal{C}$ tal que $\mathcal{C} = \{\overrightarrow{C_i} : i \in I\}$ pode ser escrito como

$$\mathcal{C} = \mathcal{S} \cdot \mathcal{U} = \{S_i \cdot \overrightarrow{U_i} : i \in I\}, \tag{2.28}$$

em que I denota um conjunto de índices espaciais para o RF. $\mathcal{S} = \{S_i : i \in I\}$ é um RF de escalares positivos, enquanto que $\mathcal{U} = \{\overrightarrow{U_i} : i \in I\}$ é um RF de vetores com distribuição gaussiana de média zero e covariância igual a $\mathbf{C}_U$. $\overrightarrow{C_i}$ e $\overrightarrow{U_i}$ são vetores M dimensionais.

A distorção na cena natural é modelada como um ganho de sinal acompanhado de ruído aditivo no domínio Wavelet, tal que

$$\mathcal{D} = \mathcal{G}\mathcal{C} + \mathcal{V} = \{g_i\overrightarrow{C_i} + \overrightarrow{V_i} : i \in I\}, \tag{2.29}$$

em que $\mathcal{C}$ é o RF do sinal de referência e $\mathcal{D} = \{\overrightarrow{D_i} : i \in I\}$, $\mathcal{G} = \{g_i : i \in I\}$ e $\mathcal{V} = \{\overrightarrow{V_i} : i \in I\}$ são, respectivamente, o RF do sinal distorcido, um campo determinístico de ganho escalar e um RF estacionário de ruído aditivo gaussiano, de média zero e variância $\mathbf{C}_V = \sigma_v^2\mathbf{I}$, em que $\mathbf{I}$ é a matriz identidade.

O RF $\mathcal{V}$ é branco (potência uniforme para todas as frequências) e independente de $\mathcal{S}$ e de $\mathcal{U}$.

Isso significa que o modelo, apesar de simples, captura dois tipos importantes de ruídos, ruído branco, devido à presença do RF $\mathcal{V}$, e suavização (*blur*), devido ao campo escalar de atenuação $\mathcal{G}$. Uma motivação mais detalhada desse modelo pode ser encontrada em [41].

O modelo de sistema visual humano é dual ao modelo de NSS, com muitos aspectos dele já sendo incluídos na descrição do modelo NSS. Dentre os aspectos não incluídos no modelo NSS, há a função de espalhamento ótico, a função de sensitividade ao contraste e o ruído neural interno, entre outros. Uma comparação de desempenho mais detalhada entre diferentes modelos de NSS e visão humana se encontra em [41].

O modelo de visão humana incluído na implementação padrão do VIF leva em conta apenas o ruído neural interno, o que já é suficiente para aumentar consideravelmente o desempenho preditivo da métrica. Ele pode ser modelado como ruído aditivo gaussiano branco:

$$\mathcal{E} = \mathcal{C} + \mathcal{N}, \tag{2.30}$$

$$\mathcal{F} = \mathcal{D} + \mathcal{N}, \tag{2.31}$$

em que $\mathcal{E}$ e $\mathcal{F}$ são os sinais visuais dos quais o cérebro humano extrai informações cognitivas acerca da imagem de referência e da imagem imagem distorcida, respectivamente. $\mathcal{N} = \{\overrightarrow{N_i} : i \in I\}$ é um RF, com vetores gaussianos multivariados descorrelacionados $\overrightarrow{N_i}$, de média zero e covariância $\mathbf{C}_N = \sigma_n^2\mathbf{I}$.

Cada um desses modelos considera apenas uma das sub-bandas da decomposição *wavelet* de escala-espaço-orientação como um GSM. Por exemplo, para a imagem de referência, cada sub-banda é particionada em blocos de M coeficientes cada, sem sobreposição. Assume-se que cada bloco é independente dos outros. Cada bloco é então modelado como o vetor $\overrightarrow{C_i}$. Assim, se observa que $\mathcal{C}$ segue uma distribuição normal condicionada a $\mathcal{S}$, e que os vetores C_i são condicionalmente independentes entre si, dado $\mathcal{S}$ [40].

Para se obter o valor final do VIF, basta então se considerar a informação para cada uma das sub-bandas presentes. Portanto, para o conjunto de todas as sub-bandas J, tem-se

$$\mathrm{VIF} = \frac{\sum_{j\in J} \mathrm{I}(\overrightarrow{C}^{N,j}; \overrightarrow{F}^{N,j}|s^{N,j})}{\sum_{j\in J} \mathrm{I}(\overrightarrow{C}^{N,j}; \overrightarrow{E}^{N,j}|s^{N,j})}, \tag{2.32}$$

em que $\overrightarrow{C}^{N,j}$ representa N elementos da RF $\mathcal{C}_j$ (RF $\mathcal{C}$ da sub-banda j), com definições similares para $\overrightarrow{F}^{N,j}$ e $\overrightarrow{E}^{N,j}$. Na fórmula 2.32, $\mathrm{I}(\overrightarrow{C}^{N,j}; \overrightarrow{F}^{N,j}|s^{N,j})$ é a informação mútua de Shannon [36] entre $\overrightarrow{C}^{N,j}$ e $\overrightarrow{F}^{N,j}$ dada uma realização $s^{N,j}$ de S^N (N elementos do RF $\mathcal{S}$) na banda j. A mesma relação se mantém para $\overrightarrow{E}^{N,j}$, no denominador. Em termos gerais, a informação mútua entre duas variáveis aleatórias X e Y se dá por

$$\mathrm{I}(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right), \tag{2.33}$$

em que $p(x,y)$ é a função de probabilidade conjunta entre X e Y, enquanto que $p(x)$ e $p(y)$ são as respectivas funções de probabilidade marginal.

Vale notar que a informação mútua só pode ser calculada diretamente entre RFs com parâmetros que se assumem conhecidos, e não entre realiações dessas RFs (i.e. uma imagem contendo uma cena natural). No entanto, é possível estimar os parâmetros relevantes de cada RF, dada uma realização e respeitando-se as devidas condições. Por exemplo, em [42] é demonstrada uma série de métodos para estimar os parâmetros s_i^2 e $\mathbf{C}_U$ do modelo de fonte do sinal quando as RFs são ergódicas. Em [43] também se propõe obter $\mathcal{G}$ e σ_v^2, respectivos ao modelo de distorção, através regressão linear entre a entrada do modelo (a imagem de referência) e a saída (imagem de teste). Se considera que ambos os parâmetros são constantes entre todos os blocos espaço-temporais que dividem cada canal do sinal. Também em [43] se propõe estimar o parâmetro σ_n^2 empiricamente, variando-o até se obter o melhor desempenho para os dados disponíveis.

2.9 Testes subjetivos de qualidade

Como foi mencionado na Seção 2.8, a qualidade visual de mídias é avaliada principalmente através de métricas objetivas. De fato, avaliações subjetivas costumam apresentar mais custos associados. Entretanto, em casos em que a avaliação objetiva não é um problema resolvido, ou pelo menos caso se deseje validar alguma métrica objetiva nova, é necessário validar o poder preditivo de métricas propostas, levando em consideração principalmente como o desempenho da métrica proposta se compara com outras alternativas existentes. Isso é feito com a realização de testes subjetivos de qualidade.

Existem duas classes de avaliações subjetivas. A primeira classe, denominada avaliação de qualidade, estabelece o desempenho de sistemas de mídia sob condições ótimas. A segunda classe, denominada avaliação de degradação, estabelece a capacidade de sistemas de manter a qualidade sob condições sub-ótimas (e.g. canais de transmissão ruidosa, *codecs* que introduzem artefatos de compressão).

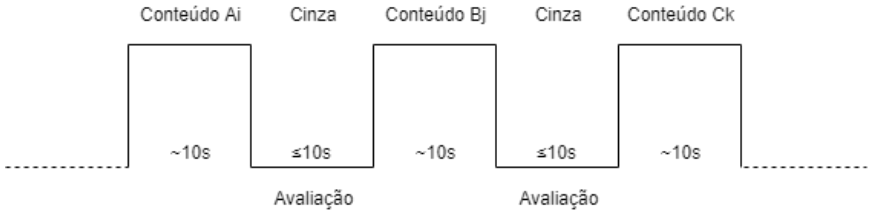
Testes subjetivos procedem com participantes sendo informados sobre o tipo de avaliação que deverão desempenhar. É necessário fornecer aos participantes informações acerca do tipo de conteúdo que será analisado, no que o participante deve focar ao avaliar o conteúdo, o funcionamento da escala de avaliação, e como o experimento deve ocorrer (por exemplo, quantas sequências de conteúdo devem ser observadas, tempo permitido para a avaliação, se o participante deve realizar alguma tarefa de maneira específica ou permanecer passivo durante a avaliação).

A maior organização responsável por propor e padronizar testes subjetivos de conteúdos de telecomunicação é a ITU-T. A ITU-T recomenda que exemplos práticos do tipo de conteúdo a ser avaliado, que não devem ser usados nos testes em si, sejam mostrados aos participantes antes do início da aquisição real de avaliações. Isso pode ser feito através de uma rodada de avaliações de treinamento cujos dados não serão considerados. Também é recomendado que avaliações, em vez de adotar uma escala puramente numérica, sejam baseadas em ideias subjetivas relacionadas com os termos linguísticos usados para descrever qualidade.

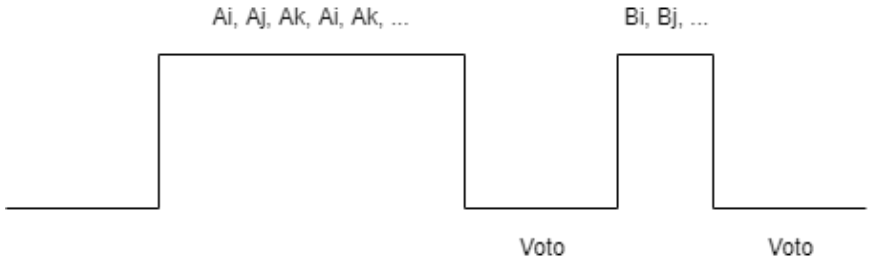
Além de definir condições ambientes ideais para a realização de testes, a ITU-T também propõe esquemas experimentais específicos desenvolvidos para medir aspectos de qualidade específicos.

Alguns deles são discutidos a seguir.

2.9.1 Absolute Category Rating



(a) Variante I



(b) Variante II

Figura 2.7: Diagramas de realizações típicas de duas variantes de testes de qualidade subjetiva do tipo ACR.

O método de avaliação de categoria absoluta (ACR), também chamado de método de estímulo único, é um teste subjetivo de qualidade em que os conteúdos a serem avaliados são mostrados aos participantes e então avaliados de maneira independente, um por vez, em uma escala categórica. A avaliação indica a qualidade observada no conteúdo que acabou de ser mostrado. Para cada conteúdo, é calculada a nota de opinião média (MOS) através das médias das notas fornecidas pelos participantes.

As notas escolhidas pelos participantes devem ser uma escala de 5 categorias. Cada categoria representa um nível subjetivo de qualidade, e tem um valor numérico associado para calculo do MOS. A escala é indicada por:

5. Excelente
4. Boa
3. Razoável
2. Pobre
1. Ruim

É possível incluir na sequência de conteúdos mostrados uma versão não distorcida do conteúdo de referência, porém não identificada ao avaliador. Nessa variante, denominada ACR com referência

oculta (ACR-HR), a nota de cada conteúdo é calculada de maneira relativa a sua respectiva referência. Em vez do MOS, se obtém o DMOS, a média das notas diferenciais de cada avaliador (DV), tal que, para algum conteúdo distorcido k e para algum participante n,

$$DV_n(k) = PS_n(k) - RS_n(k) + 5, \tag{2.34}$$

em que $PS_n(k)$ é a nota ACR dada pelo participante n ao conteúdo k e $RS_n(k)$ é a nota dada pelo mesmo participante à versão de referência do conteúdo distorcido em questão. O DMOS daquele conteúdo é então a média de DV, entre todos os participantes:

$$DMOS(k) = \frac{\sum_{n=1}^{N} DV_n(k)}{N}. \tag{2.35}$$

Seguindo a Equação 2.34, uma nota diferencial igual a 5 indica uma qualidade excelente, enquanto que uma nota igual a 1 indica qualidade ruim. Em geral, valores acima de 5 (casos em que algum conteúdo distorcido foi considerado como tendo uma qualidade maior do que sua referência) não são descartados, sendo incluídos normalmente no cálculo do DMOS.

A sequência de conteúdos apresentada na sessão de testes deve ser aleatória, de preferência de maneira que dois avaliadores não observem a mesma sequência. Antes e depois da visualização de cada conteúdo, são mostradas telas em um tom de cinza intermediário. Isso tem o objetivo de evitar que possíveis efeitos de fadiga alterem de maneira indesejada a avaliação dos participantes. Há duas variantes quanto à estrutura de apresentações. O processo de ambas as variantes se encontra representado na Figura 2.7.

Na primeira variante, cada conteúdo é mostrado uma única vez. Uma sequência típica inclui uma visualização inicial em um tom intermediário de cinza por algum tempo determinado, a visualização do conteúdo a ser analisado, e novamente uma visualização de cinza, após a qual o participante registra sua avaliação. Após o período de avaliação, o próximo conteúdo é mostrado da mesma maneira, e o ciclo de visualizações se repete, até que todos os conteúdos sejam avaliados.

Na segunda variante, conteúdos são mostrados da mesma maneira que na primeira variante. Após o fim do primeiro ciclo, é anunciado ao participante que outro ciclo se iniciará. Novamente os conteúdos são mostrados e o participante deve avaliá-los. Esse processo se repete mais uma vez e a sessão de testes termina. O primeiro ciclo tem o objetivo de estabilizar as avaliações do participante, e dados coletados nessa etapa não devem ser levados em consideração. As avaliações finais são calculadas com as médias das avaliações realizadas no segundo e no terceiro ciclo de visualizações. É importante que em nenhuma das sequências geradas em cada ciclo um mesmo conteúdo esteja localizado na mesma posição que em outro ciclo. Os mesmos dois conteúdos também não podem ocorrer seguidos um do outro em mais de um ciclo.

2.9.2 Double Stimulus Impairment Scale

O método de escala de distorção de estímulo duplo (DSIS) é um teste subjetivo de degradação baseado na comparação direta entre um conteúdo distorcido e uma versão de referência do mesmo conteúdo, livre de distorções. Sessões de avaliação tem duração de aproximadamente 30 minutos

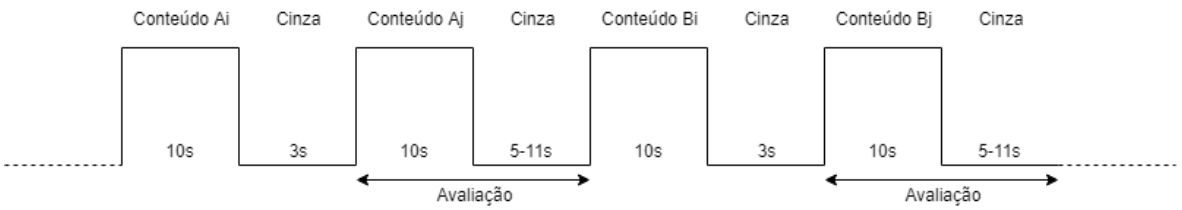


Figura 2.8: Diagrama de uma realização típica de testes de qualidade subjetiva do tipo DSIS.

e contam com um participante no papel de avaliador por vez. O método DSIS é caracterizado por períodos alternados de observação e votação, que se repetem de maneira cíclica até que todos os conteúdos sob análise sejam avaliados pelo participante. É comum a inclusão do próprio conteúdo de referência na sequência de conteúdos que devem ser avaliados pelos participantes, sem a sua identificação para os mesmos. Esse tipo de variante experimental é identificado pelo termo "referência oculta" (DSIS-HR).

A sequência de conteúdos mostrada deve ser escolhida de modo cobrir todos os níveis de degradações dispostos na escala adotada no experimento, pelo menos para a maioria dos participantes. Ao final da aquisição de dados, a nota média obtida entre todos os conteúdos e todos os participantes seja aproximadamente 3. Indica que a faixa de distorções mostrada no experimento foi adequada.

A sequência deve ter ordem aleatória evitando que participantes diferentes observem os conteúdos na mesma sequência ou, pelo menos, de maneira que um mesmo participante não observe o mesmo conteúdo duas vezes seguidas, seja com o mesmo nível de degradação ou com degradações diferentes. Isso pode ser obtido arranjando as sequências, por exemplo, através de um quadrado latino [44, 45]. Quadrados latinos são arranjos de símbolos em uma matriz quadrada de lado n, de tal maneira que cada símbolo ocorre exatamente uma vez em cada coluna e cada linha [46]. Uma maneira de aplicar o quadrado latino a sequências de um experimento é tratar cada eixo como um passo da realização do experimento (usuário versus conteúdo de referência, por exemplo) e cada símbolo como outra variável de interesse (uma permutação da sequência de degradações). Isso permite projetar experimentos reduzindo a ocorrência de combinações repetidas de variáveis independentes (controladas) entre realizações do experimento.

Evitar repetições na ordens de observações de diferentes participantes diminui a influência que a relação de proximidade entre diferentes distorções possa ter nos resultados, e reduz a chance de possíveis vieses ocorrerem. Um exemplo de realização de um teste DSIS se encontra diagramado na Figura 2.8.

Ao final da série de sessões de avaliações, as notas atribuídas pelos participantes a cada conteúdo têm suas médias calculadas. As notas variam em uma escala de distorção com as seguintes opções

5. Imperceptível

4. Perceptível, mas não incomoda

3. Incomoda levemente

2. Incomoda

1. Incomoda muito

Esse tipo de escala costuma apresentar resultados mais estáveis para pequenas distorções do que para grandes distorções. É possível realizar o método DSIS com uma escala reduzida a apenas uma faixa da escala completa (e.g. de "incomoda levemente" a "imperceptível"), apesar de que se recomenda o uso da escala toda. Em casos em que uma resolução maior das avaliação subjetiva dos participantes, é possível ainda estender, sem muita alteração do design experimental, a escala de avaliação de 5 pontos para 9 pontos, da seguinte maneira:

9. Imperceptível

8.

7. Perceptível, mas não incomoda

6.

5. Incomoda levemente

4.

3. Incomoda

2.

1. Incomoda muito

Existem duas variantes quanto ao formato das apresentações no método DSIS. Na variante I, o conteúdo analisado e sua respectiva versão de referência são mostrados apenas uma vez cada, e em seguida o participante registra sua nota. Na variante II, o par de conteúdos é mostrado duas (ou múltiplas) vezes para cada participante, que então registra sua nota. A variante II requer mais tempo para sua realização, e normalmente é utilizada quando o objeto de análise envolve distorções com detalhes muito sutis.

2.10 Trabalhos prévios

Uma investigação preliminar [9] tentou estudar a relevância de artefatos de formato e de cor para a qualidade geral, porém sem um estudo estatístico aprofundado. A questão de como avaliar distorções de geometria em conteúdo tridimensional foi previamente discutida em um estudo sobre avaliação de qualidade em vídeos estéreo [47], que sugeriu que avaliações de qualidade são dependentes da estrutura e do conteúdo da cena, e propôs que apenas alguns níveis de qualidade geométrica poderiam ser distinguidos.

Em trabalhos recentes [10, 11], foi proposta uma metodologia realista para se avaliar a qualidade do atributo geométrico de nuvens de pontos. Avaliações subjetivas foram feitas sobre conteúdos

sujeitos a distorções de ruído gaussiano e de compressão baseada em *octrees*, que afetam apenas a geometria. Os modelos distorcidos foram visualizados como nuvens de pontos, sem conversão para representações baseadas em malhas (*meshes*), e avaliados usando um protocolo interativo, através de um arranjo composto por um *desktop* mais um monitor, no primeiro artigo [10], e depois através um dispositivo de realidade aumentada (AR), no segundo artigo [11]. Outro estudo [7] realizou avaliações subjetivas de conteúdos codificados através de *octrees* e através de grafos, mantendo os atributos de cor dos modelos originais inalterados. As nuvens de pontos foram visualizadas representando pontos individuais como cubos, cujo tamanho era ajustado automaticamente em função dos pontos vizinhos mais próximos. A visualização foi realizada de maneira passiva durante a avaliação. O efeito de artefatos não-sintéticos também é discutido em outro estudo [48], apesar de que restrito a métricas ponto-a-ponto. Métricas baseadas na distância de ponto-a-plano [4, 10] se mostram mais robustas para avaliar defeitos no formato.

Também foi feita uma avaliação de qualidade de algoritmos de remoção de ruído em nuvens de pontos [6]. Participantes visualizavam os conteúdos processados por um processo *Screened Poisson* de reconstrução de superfície [49]. Outro estudo [12] adotou um procedimento similar. Nuvens de pontos sem cor foram codificadas através de poda de *octrees* e visualizadas como malhas poligonais, criadas através do mesmo algoritmo de reconstrução de superfícies (*Screened Poisson*).

Avaliação da qualidade de cor foi inicialmente discutida em um estudo [9] dentro de um escopo limitado e de maneira mais extensa em outro estudo [48]. O primeiro sinteticamente adiciona ruído ao sinal de cor e não propõe uma métrica objetiva, enquanto o segundo considera apenas uma métrica SNR ponto-a-ponto para medir a qualidade de cor.

Em face das questões levantadas pelos estudos mencionados, o presente trabalho foca em um novo método combinado de avaliação de qualidade de cor e de geometria em nuvens de pontos voxelizadas. Para validar esse novo *framework* proposto, foi realizada uma série de experimentos para determinar a correlação entre as métricas obtidas e a percepção humana subjetiva de qualidade visual em conteúdo tridimensional, comparando com o desempenho de outras métricas atualmente empregadas para medir a qualidade visual desse tipo de conteúdo. Nos próximos Capítulos são descritos os experimentos realizados e os resultados obtidos são analisados.

Capítulo 3

Desenvolvimento

3.1 Introdução

Neste capítulo o processo de elaboração do *framework* de métricas objetivas proposto é descrito em detalhe. Em seguida, são descritos os experimentos realizados para verificar a correlação entre as métricas obtidas através do *framework* e a percepção subjetiva de qualidade visual de nuvens de pontos.

3.2 Representação de nuvens de pontos

O primeiro passo do método proposto de avaliação de qualidade visual de nuvens de pontos é gerar múltiplas projeções ortográficas, de diferentes pontos de vista, do conteúdo voxelizado sob análise. A escolha do número e posições dos pontos de vista deve garantir a extração eficiente da maior quantidade de informação possível acerca do conteúdo original.

Sem conhecimento *a priori* da geometria e da importância de diferentes partes do objeto, idealmente deve-se obter uma amostragem uniforme dos *voxels* presentes no conteúdo. Em outras palavras, projeções devem ser tomadas de direções uniformemente espaçadas entre si ao redor do objeto. Isso é equivalente a amostrar uniformemente a superfície de uma esfera ou, em termos de coordenadas esféricas, o plano composto pelas coordenadas θ (ângulo polar) e ϕ (ângulo azimutal). Isso só é garantido de maneira exata para uma quantidade limitada de números de amostras. Especificamente, os únicos arranjos de amostras que se distribuem uniformemente através da superfície de uma esfera são aqueles que coincidem com os vértices de um sólido platônico inscrito em tal esfera. Consequentemente, apenas projeções tomadas desses pontos de vista amostram uniformemente as direções possíveis ao redor de objetos [50]. Para números diferentes de amostragens, a configuração ótima é dependente da tarefa em questão [51, 52, 53, 54]. No presente caso, se considera que quanto menor a variação da área da esfera associada com cada ponto de amostragem, melhor a distribuição.

Na Figura 3.1, encontra-se um exemplo de projeções de um modelo tridimensional de uma

Figura 3.1: Seis projeções ortográficas igualmente espaçadas ao redor de um modelo humano.

pessoa, tomadas a partir dos 6 vértices de um octaedro virtual circunscrevendo o volume representado. Outro exemplo de projeções do mesmo modelo se encontra na Figura 3.9, porém a partir de 4, 8, 14 e 40 pontos de vista. Nos dois últimos casos, o espaçamento entre os pontos de vista é aproximadamente uniforme. Além disso, nos 4 casos os pontos de vista não se alinham mais com a grade de *voxels*. Assim, para rasterizar a imagem é preciso interpolar os *voxels* que se encontram em posições não inteiras, efetivamente tendo que se voxelizar o modelo novamente.

Existem maneiras de se aproximar o comportamento uniforme de amostragem da esfera, principalmente quando o número de amostras se aproxima do infinito. É possível, por exemplo, amostrar as direções aleatoriamente, ou amostrar através do algoritmo da espiral de Fibonacci [50]. Um exemplo de amostragem usando a espiral de Fibonacci está disposto na Figura 3.2. No entanto, para a maioria dos casos, um número pequeno de projeções se mostra suficiente. Na Figura 3.3 é mostrado o percentual de *voxels* de uma nuvem de pontos que é visto de alguma projeção (i.e. não fica ocluso em pelo menos alguma projeção) em função do número de pontos de vista utilizados para se realizarem as projeções. Nos casos em que o número de pontos de vistas coincidiu com o número de vértices de algum dos sólidos platônicos, as projeções foram tomadas de acordo. Nos demais casos, foi realizada uma amostragem aproximadamente uniforme através da espiral de Fibonacci. Se observa que a taxa com que o número de *voxels* vistos aumenta é cada vez menor à medida que o número de ponto de vistas usados para se obter projeções do modelo aumenta também. Isso revela que caso haja um custo maior associado a amostrar mais pontos de vista, um número relativamente menor de amostras pode ser mais vantajoso.

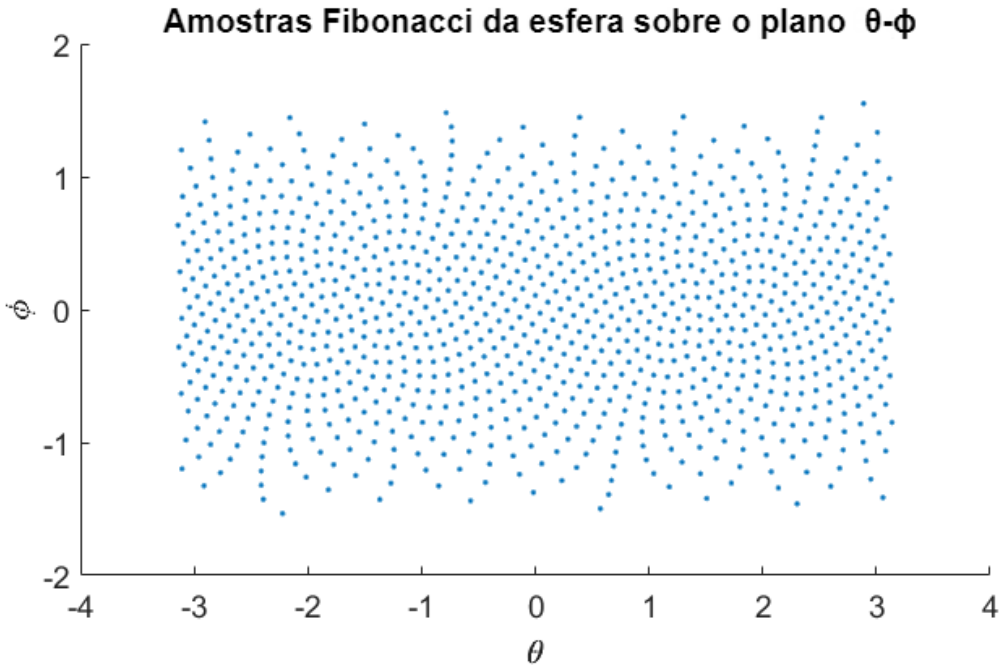


Figura 3.2: Pontos de amostragem distribuídos ao redor das esfera de acordo com o algoritmo da espiral de Fibonacci e projetados no plano $\theta \times \phi$.

Especialmente, o uso de 6 projeções se revelou conveniente. Nesse caso, os pontos de vista coincidem com os vértices de um tetraedro regular. Isso ainda implica que os planos de projeções formam exatamente um cubo ao redor do objeto avaliado, o que permite que as projeções se alinhem paralelamente com as faces dos *voxels* que compõem o conteúdo em questão, mantendo uma relação direta entre *voxels* do objeto e *pixels* de suas projeções, sem a necessidade de qualquer interpolação ou outro tipo de pré-processamento. Isso pode explicar o comportamento da PSNR média das projeções à medida que o número de pontos de vista utilizados varia, demonstrado na Figura 3.4. Além da eventual convergência do valor de PSNR a partir de 15 projeções, se observa um pico da PSNR para certos números de pontos de vista no início do gráfico. Principalmente para 6 e 12 projeções, o valor de PSNR é mais alto.

Valores maiores de PSNR indicam um erro relativamente menor entre a imagem sob análise e sua referência. De maneira análoga, quando o número de pontos de vista não permite que as projeções tomadas se alinhem com os vértices de um octaedro (6 vértices) ou um icosaedro (12 vértices), o erro medido entre as versões de teste e de referência aumenta. Pode se observar que dos sólidos regulares, o octaedro e o icosaedro são os que apresentam o maior número de vértices alinhados com as faces de um cubo, assumindo orientações compatíveis. Isso indica que a interpolação necessária para a projeção de conteúdos voxelizados fora dos eixos ortogonais de seu sistema de coordenadas introduz erro nas imagens geradas, pelo menos quando a métrica adotada é a PSNR. Assim, o uso de 6 pontos de vista garante a maior fidelidade entre o conteúdo de teste e sua versão de referência. Portanto, escolheu-se fixar em 6 o número de projeções no cálculo das métricas usadas nos experimentos subjetivos realizados neste estudo.

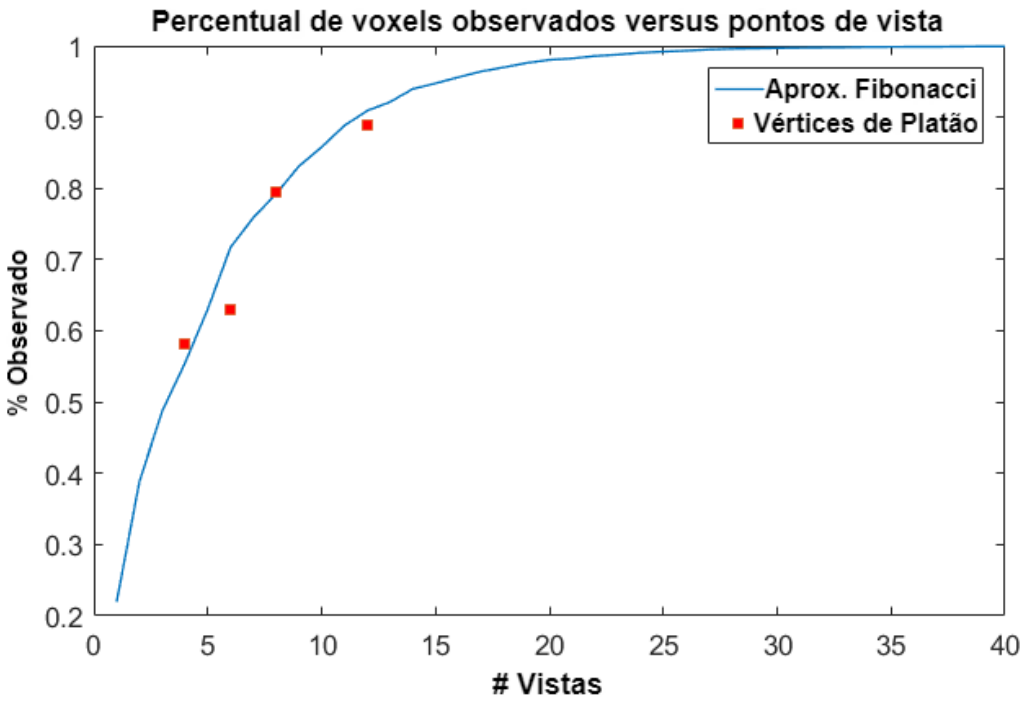


Figura 3.3: Proporção de *voxels* vistos para um determinado modelo em função no número de pontos de vista projetados.

3.3 Métrica objetiva das projeções

De posse de um conjunto de projeções de um objeto tridimensional de referência e outro conjunto de projeções a partir dos mesmos pontos de vista da versão distorcida desse objeto, é possível realizar uma comparação entre os dois conjuntos de imagens através de qualquer métrica objetiva existente para a qualidade de imagens bidimensionais. Foram exploradas 6 métricas diferentes, entre elas PSNR e SSIM (e variações delas) e VIFP.

Para cada par de projeções, respectivas ao conteúdo de referência e o conteúdo distorcido, obtém-se uma medida do nível de distorção. A métrica final para o conteúdo tridimensional é calculada a partir da média da métrica entre todos os pares. Caso haja informação acerca da relevância de cada projeção, uma média ponderada pode ser empregada.

3.4 Validação experimental

Foram realizados dois experimentos independentes para validar a correlação da métrica proposta com a percepção humana de qualidade visual em conteúdos tridimensionais estáticos. O primeiro experimento consistiu de participantes voluntários interagindo com modelos tridimensionais de pessoas através de um visualizador e atribuindo notas para diversas características de versões diferentes de cada conteúdo. O segundo experimento contou com modelos tanto de pessoas como objetos. Assim como no primeiro, cada participante interagia com o conteúdo disponibili-

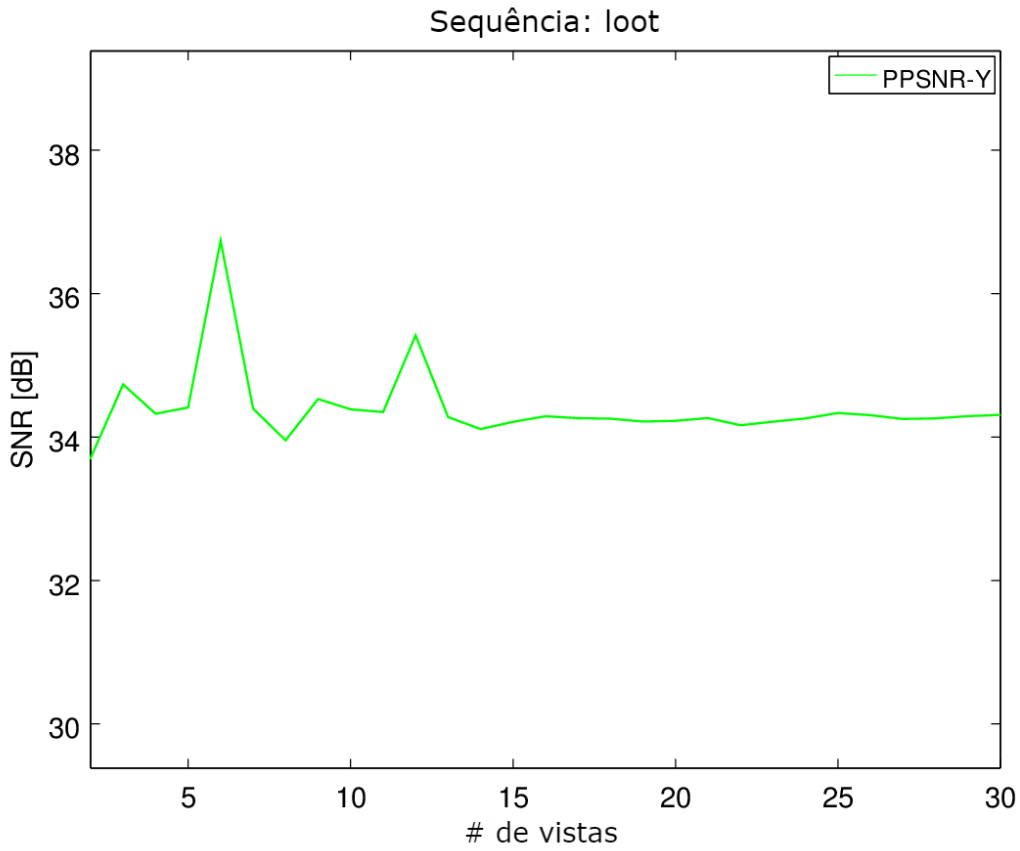


Figura 3.4: PSNR médio entre diferentes projeções em função do número de pontos de vista utilizados.

zado. No entanto, apenas uma única nota era atribuída a pares de conteúdos, referente à diferença de qualidade geral observada entre cada conteúdo do par.

3.4.1 Experimento ACR

Este experimento foi desenvolvido como uma modificação do método de avaliação subjetiva ACR-HR [55]. Cada participante observava em uma janela quadrada uma projeção ortográfica de uma sequência com duas nuvens de pontos. Cada nuvem de pontos apresentava cinco versões diferentes, uma de referência oculta e outras 4 versões degradadas através de diferentes processos.

A visualização era interativa, e os participantes podiam, para uma determinada nuvem de pontos, alternar livremente e sem limite de tempo entre suas versões disponíveis, escolher o ponto de vista (rotacionando e transladando o conteúdo) e o nível de ampliação da projeção. Cada versão era disponibilizada em uma ordem aleatória e sem identificação para o participante. A interface gráfica do programa utilizado para a visualização é mostrada na Figura 3.5.

A partir do momento que se sentisse confortável com sua aferição, o participante selecionava, através da interface gráfica, uma nota para a qualidade observada em cada versão do conteúdo. Foi considerada a escala ACR de 5 categorias:

Figura 3.5: Interface gráfica do visualizador utilizado nos experimentos preliminares.

5. Excelente
4. Boa
3. Razoável
2. Pobre
1. Ruim

Após confirmar suas avaliações, as notas dadas pelo participantes para todas as versões da nuvem de pontos avaliada eram registradas simultaneamente, e o próximo conteúdo era mostrado para ele.

Foram utilizados dois conteúdos diferentes, cada um com 4 níveis de distorção e mais um de referência, somando 10 avaliações por participante. No total, 12 voluntários participaram do experimento.

Figura 3.6: Interface gráfica do visualizador utilizado nos experimentos com estímulo duplo.

3.4.1.1 Conteúdos utilizados

Como conteúdos a serem avaliados no experimento, foram usadas duas nuvens de pontos. Ambas as nuvens de pontos foram obtidas extraindo um único quadro de uma sequência de vídeo tridimensional (nuvem de pontos dinâmicas). O primeiro conteúdo, denominado de *Ricardo*, apresenta um modelo tridimensional da metade anterior do torso de uma pessoa. O segundo conteúdo, denominado de *Loot*, apresenta um modelo completo do corpo de uma pessoa.

As versões sob avaliação da nuvem de pontos *Ricardo* foram geradas usando um algoritmo de compressão com compensação de movimento [56] seguindo quatro níveis diferentes de quantização, que degradavam tanto cor como geometria.

A sequência *Loot* foi avaliada sob quatro degradações diferentes:

- Alta qualidade, que sofreu apenas uma leve distorção de cor [56].
- Baixa qualidade, que sofreu uma distorção considerável de cor.
- Alta qualidade passa-baixas, que sofreu distorções de cor e de geometria leves. A distorção de geometria foi obtida por um processo de filtragem passa-baixas composto por uma sub-amostragem seguida de uma super-amostragem de mesmo nível.
- Baixa qualidade passa-baixas, que sofreu uma distorção considerável de cor e a mesma distorção de geometria descrita na versão anterior.

3.4.2 Experimento DSIS

Este experimento consistiu de sessões de avaliação subjetiva de maneira similar ao experimento anterior. Neste caso, em vez de observar um conteúdo de cada vez, participantes viam pares de projeções do mesmo ponto de vista, lado a lado. Em um dos lados, havia uma projeção do conteúdo original. Do outro lado, a projeção de uma versão distorcida do conteúdo. Era usado apenas um nível de distorção por vez. O lado em que a imagem de referência era mostrada era aleatório, permanecendo o mesmo até o final da seção de cada participante. Ambos os lados eram devidamente identificados para o usuário através da interface.

O participante podia então interagir com os pares de conteúdo, novamente podendo rotacionar, transladar e ampliar eles livremente e sem limite de tempo. Toda interação era aplicada igualmente a cada uma das duas projeções, resultando sempre em imagens equivalentes, exceto pela presença de distorção em um delas. Uma representação da interface se encontra disposta na Figura 3.6.

Após o período de observação, o participante selecionava uma nota, de 1 (ruim) a 5 (bom), para o grau de distorção observado entre a imagem de referência e a imagem distorcida. O usuário então confirmava o envio de sua avaliação e o próximo conteúdo era mostrado para ele, em uma ordem aleatória. Cada participante observou 8 níveis de distorção para 10 conteúdos diferentes (incluindo tanto pessoas como objetos), totalizando 80 pares por participante, com 20 pessoas tendo participado do experimento.

3.4.2.1 Conteúdos utilizados

No total, 7 conteúdos diferentes foram utilizados no decorrer do experimento. Projeções demonstrativas das nuvens de pontos estão dispostas na Figura 3.7. Projeções vistas de 6 pontos de vista diferentes (formando um cubo ao redor do objeto) dos mesmos conteúdos estão dispostas nas Figuras 3.10, 3.11, 3.12, 3.13, 3.14 e 3.15. Os conteúdos foram escolhidos de forma a apresentar uma ampla gama de características diferentes. Tanto objetos inanimados e corpos humanos foram incluídos, ambas as classes contendo variações de níveis de detalhes geométricos e de cor. Os conteúdos longdress_vox10_1300 (*longdress*), loot_vox10_1200 (*loot*) redandblack_vox10_1550 (*redandblack*) e *statue_Klimt* foram obtidos do repositório MPEG e apresentam modelos humanos. Já os conteúdos *romanoillamp11* e *biplane* foram obtidos do repositório JPEG enquanto o conteúdo *amphoriskos12* foi obtido da plataforma Sketchfab[1], sendo esses últimos modelos de objetos inanimados. A aquisição desses modelos pode se dar de diversas maneiras. Por exemplo, os modelos *longdress*, *loot* e *redandblack* foram gerados filmando-se pessoas realizando ações em tempo real no interior de uma estrutura com câmeras arranjadas em uma esfera ao redor do volume sendo modelado. Para este experimento, foram utilizados apenas um dos quadros de cada uma das sequências de vídeo.

O passo seguinte da preparação dos conteúdos foi processar as nuvens de pontos dos modelos de forma a reduzir os possíveis fatores de influência dos resultados. Principalmente o número de pontos precisou ser padronizado em algumas das nuvens de pontos utilizadas. Nem todos os

[1] https://sketchfab.com/

(a) *amphoriskos* (b) *biplane* (c) *longdress* (d) *loot*

(e) *redandblack* (f) *romanoillamp* (g) *statue_Klimt*

Figura 3.7: Nuvens de pontos de referência usadas no experimento DSIS. O conteúdo *"statue_Klimt"* (g) foi utilizado apenas para o treinamento dos participantes.

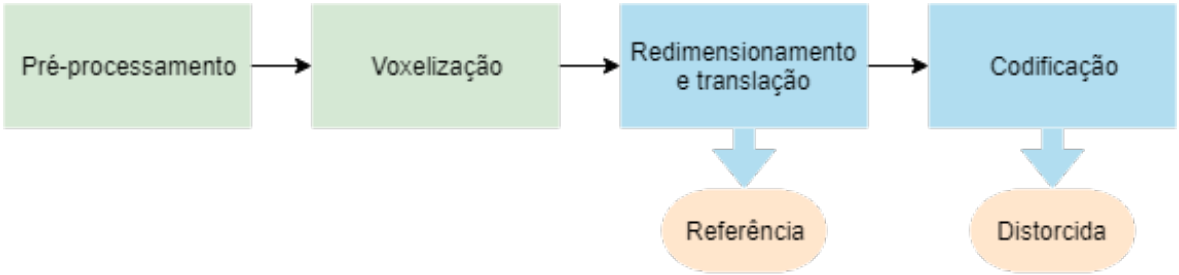


Figura 3.8: Etapas do processamento dos conteúdos visualizados no experimento.

conteúdos precisaram passar por todos os processos empregados. Na figura 3.8 são demonstradas as etapas de processamento. Em verde são mostradas etapas pelas quais apenas alguns dos conteúdos passaram, enquanto que em azul estão os processos pelos quais todos os conteúdos passaram.

A etapa de pré-processamento serviu para garantir que o número de pontos não variasse demasiadamente entre conteúdos. O conteúdo *biplane*, principalmente, precisou passar por essa etapa. Ele está disponível em múltiplas versões correspondentes a capturas de diferentes partes do objeto representado (um avião biplano). Para o experimento, um modelo completo foi reconstruído a partir das diferentes varreduras. Isso resultou em uma nuvem de pontos com cerca 106 milhões de pontos, muito acima da média dos outros conteúdos e acima das capacidades de desempenho aceitável do visualizador. Para reduzir o número de pontos a um limite aceitável, o programa CloudCompare13[2] foi utilizado. Foi feita uma subamostragem uniforme da nuvem de pontos ori-

[2]Aplicação disponível em http://www.cloudcompare.org/

Tabela 3.1: Descrição geométrica de cada conteúdo de referência. Além do número total de pontos em cada modelo, são especificadas as distâncias mínimas e máximas entre todos os pares de vizinhos mais próximos do modelo. São também especificadas as dimensões (após normalização) em cada uma das direções ortogonais do sistema de coordenadas cartesiano.

Conteúdos:	*amphoriskos*	*biplane*	*longdress*	*loot*	*redandblack*	*romanoillamp*	*statue_Klimt*
Pontos:	828,820	773,447	857,966	805,285	757,691	636,097	482,941
Min NN:	0.000977501	0.000977516	0.00101107	0.00101936	0.00103515	0.000977516	0.000977516
Max NN:	0.00239442	0.0470835	0.00226096	0.00203872	0.00253568	0.0693761	0.0100166
X/Y/Z:	0.60/1/0.68	0.65/0.23/1	0.40/1/0.20	0.35/1/0.41	0.44/1/0.30	1/0.45/0.51	0.30/1/0.29

ginal, com uma distância máxima permitida entre amostragens mais próximas igual a 0.009 (na escala interna da própria nuvem de pontos, já que não é assumida nenhuma unidade de distância na representação do modelo). Outro conteúdo que sofreu alterações em seu número de pontos foi o modelo *amphoriskos*. Neste caso, foi preciso aumentar o número de pontos presente. Isso foi feito utilizando um processo de reconstrução de superfície de Poisson, também através do programa CloudCompare. Foi utilizada uma amostra por nó, mantendo-se os valores padrão das outras configurações. Foram utilizados os vetores normais originais da nuvem de pontos. De posse da malha reconstruída, 1 milhão de pontos foram aleatoriamente amostrados através do mesmo programa. Nenhum dos outros conteúdos passou por esta etapa de pré-processamento.

A etapa de voxelização garante que todas as nuvens de pontos fiquem restritas a uma geometria com pontos regularmente espaçados. Isso evita que o visualizador utilizado ou que a compressão aplicada nas nuvens de pontos introduzam vieses. Especificamente, como todos os modelos humanos do conjunto de dados utilizado já era originalmente voxelizado, os modelos de objetos inanimados foram convertidos a grades de *voxels* quantizadas com precisão de 10 bits para que a representação geométrica contínua desses conteúdos não afetasse as avaliações.

Em seguida, ocorre a etapa de redimensionamento e translação das nuvens de pontos. Isso garante que todas as nuvens de pontos se encontrem na mesma faixa dinâmica de posições. O *codec* utilizado para a introdução de distorções retorna nuvens de pontos na faixa de posições que vai de -0.5 a 0.5 em cada direção, enquanto que os conteúdos originais ocupam uma faixa de 0 a 1023. Como o experimento requer a visualização simultânea dos dois conteúdos, é preciso que eles estejam em posições e escalas equivalentes quando mostrados. Como padronização, todos os conteúdos foram redimensionados de acordo e transladados para a origem, antes da codificação. Os conteúdos de referência são obtidos diretamente desta etapa de redimensionamento e translação. Informação sobre as características geométricas desses conteúdos está disposta na Tabela 3.1.

A etapa de codificação é responsável por produzir as versões distorcidas dos conteúdos a serem usadas no teste subjetivo. A codificação aplicada nas nuvens de pontos de referência foi feita através do software *opensource* disponibilizado como âncora em uma das chamadas de propostas para compressão de nuvens de pontos emitida pelo MPEG[3]. Essa codificação segue um esquema

[3]Disponível em https://github.com/cwi-dis/cwi-pcl-codec

Tabela 3.2: Pontos remanescentes e taxa (bpp) de geometria e color para cada conteúdo de teste codificado.

Conteúdo	**Profundidade de *octree***	**Percentual de pontos remanescentes**	**Geometria (bpp)**	**Cor (bpp)**		
				QP = 10	*QP = 50*	*QP = 90*
amphoriskos	*OD = 08*	16.61%	0.400	0.078	0.234	0.652
	OD = 09	53.92%	1.561	0.188	0.612	1.764
	OD = 10	100%	5.006	0.301	1.004	2.889
biplane	*OD = 08*	8.04%	0.142	0.069	0.191	0.430
	OD = 09	32.69%	0.618	0.209	0.686	1.623
	OD = 10	100%	2.890	0.589	2.101	4.926
longdress	*OD = 08*	7.76%	0.169	0.047	0.134	0.358
	OD = 09	29.63%	0.649	0.125	0.414	1.178
	OD = 10	100%	2.520	0.347	1.169	3.423
loot	*OD = 08*	7.84%	0.173	0.034	0.078	0.210
	OD = 09	29.99%	0.662	0.073	0.213	0.636
	OD = 10	100%	2.556	0.182	0.561	1.716
redandblack	*OD = 08*	8.13%	0.182	0.039	0.093	0.258
	OD = 09	31.09%	0.699	0.084	0.249	0.773
	OD = 10	100%	2.694	0.199	0.632	2.037
romanoillamp	*OD = 08*	12.14%	0.282	0.055	0.159	0.447
	OD = 09	42.47%	1.059	0.136	0.491	1.488
	OD = 10	100%	3.827	0.289	1.124	3.492
statue_Klimt	*OD = 08*	15.00%	0.324	0.098	0.286	0.722
	OD = 09	50.56%	1.384	0.240	0.792	2.147
	OD = 10	100%	4.552	0.413	1.392	3.889

de compressão através de *octrees*. Cores são codificadas utilizando o algoritmo JPEG após serem mapeadas a uma grade bidimensional, percorrendo a *octree* em ordem de profundidade. Para se obter uma ampla faixa de distorções, foram aplicadas codificações em 3 níveis de qualidade de geometria e 3 níveis de qualidade de cor: geometrias com *octree* de 8-bits, 9-bits e 10-bits, e cores com parâmetro de qualidade JPEG (QP) igual a 10, 50 e 90. Tanto para a geometria como para a cor, quanto maior o parâmetro utilizado, se espera obter uma qualidade visual maior. Foram feitas todas as combinações entre os níveis de degradação utilizados, fornecendo 9 degradações diferentes para cada conteúdo de referência. Outros parâmetros de degradação presentes no *codec* utilizado não foram explorados, com todas as outras configurações mantendo seus valores padrão. Esta etapa resulta nos conteúdos distorcidos a serem avaliados. Todos os 9 níveis de distorção dos conteúdos de referência foram observados e avaliados por cada participante. Na Tabela 3.2 estão listados o número de bits por pontos de cada modelo degradado e a porcentagem correspondente de pontos remanescentes. Em alinhamento com o esperado, se observa que a distribuição de bits, tanto em termos de geometria e de cor, varia consideravelmente dados a profundidade na *octree* e o valor de QP, dependendo do conteúdo.

Para se calcularem as notas objetivas das métricas ponto a ponto, ponto a plano e baseadas em cor, foi usado o programa de avaliação de compressão de nuvens de pontos adotado pelo MPEG em sua versão 0.12 [57, 58]. No caso da métrica baseada em cor, o programa fornece os valores $\mathrm{PSNR_Y}$, $\mathrm{PSNR_U}$, and $\mathrm{PSNR_V}$, que são então combinados através da Equação 2.13, resultando na

degradação de cor total do conteúdo. Para as métricas ponto a ponto e ponto a plano, os valores totais de degradação de geometria foram baseados no erro quadrático médio (MSE) e na distância de Hausdorff dos erros individuais.

As métricas plano a plano foram calculadas através do *software* proposto em [59], em usa versão 1.0[4]. Como tanto as nuvens de ponto distorcidas como suas versões de referência continham vetores normais previamente associados a suas coordenadas, a metodologia proposta por Hoppe et al. [60] foi utilizada para estimar as normais. Este método se baseia no ajuste de planos através de mínimos quadrados para o conjunto dos 12 pontos mais próximos na vizinhança de cada ponto de interesse ajustado pelo plano. Foi utilizada a implementação realizada na *Point Cloud Library* (PCL) [61].

3.4.2.2 Equipamentos e ambiente

Os experimentos se realizaram em dois laboratórios durante aproximadamente o mesmo período: na Universidade de Brasília (UnB), em Brasília, Brasil, e na École Polytechnique Fédérale de Lausanne (MMSPG - EPFL), em Lausanne, Suíça. Nos dois laboratórios, foi utilizado um arranjo com computador pessoal e um monitor Apple Cinema Display de 27 polegadas e resolução de 2560 *pixels* na horizontal por 1440 *pixels* na vertical, de modelo A1316. Participantes observavam os conteúdos através do visualizador descrito na Seção 3.4.2, e eram capazes de rotacionar, transladar e redimensionar os conteúdos usando um mouse. Para avaliar os conteúdos observados, botões de rádio presentes na interface gráfica do visualizador eram selecionados, também com uso do mouse.

No MMSPG, experimentos se deram em uma sala que cumpre os requisitos para avaliação de representação visual de dados da recomendação ITU-R BT.500-1316. A sala foi equipada com luzes neon com temperatura de cor de 6500 K. A cor das paredes e das cortinas era de tom cinza médio. A luminosidade da tela foi regulada para 120 cd/m2 seguindo o perfil CIE D65, e a luz ambiente foi ajustada para o nível de 15 lux incidentes de maneira perpendicular à tela, medidos de acordo com a recomendação ITU-R BT.2022. Na UnB, a sala de testes se encontrou isolada de luz natural, sem acesso a janelas para o exterior. A iluminação foi composta por luzes fluorescentes de temperatura de cor de 4000 K, e a cor das parede era branca.

[4]https://github.com/mmspg/point-cloud-angular-similarity-metric

(a) 4 vistas

(b) 8 vistas

(c) 14 vistas

(d) 40 vistas

Figura 3.9: Projeções ortográficas igualmente espaçadas ao redor de um modelo humano.

(a) vista superior (b) vista posterior (c) vista direita

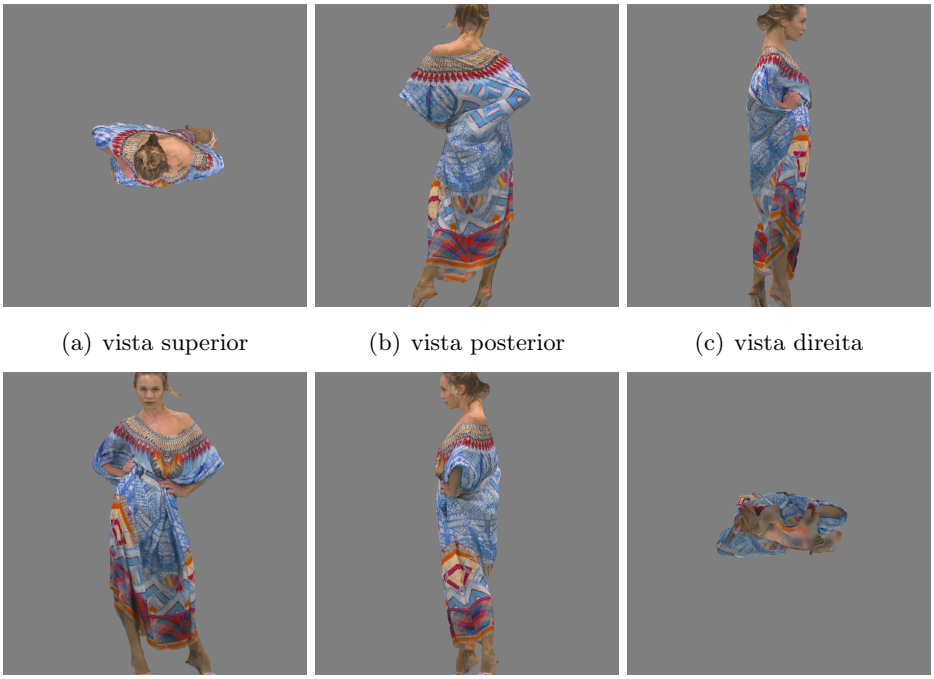
(d) vista anterior (e) vista esquerda (f) vista inferior

Figura 3.10: Projeções ao redor da nuvem de pontos de referência do conteúdo *amphoriskos*.

(a) vista superior (b) vista posterior (c) vista direita

(d) vista anterior (e) vista esquerda (f) vista inferior

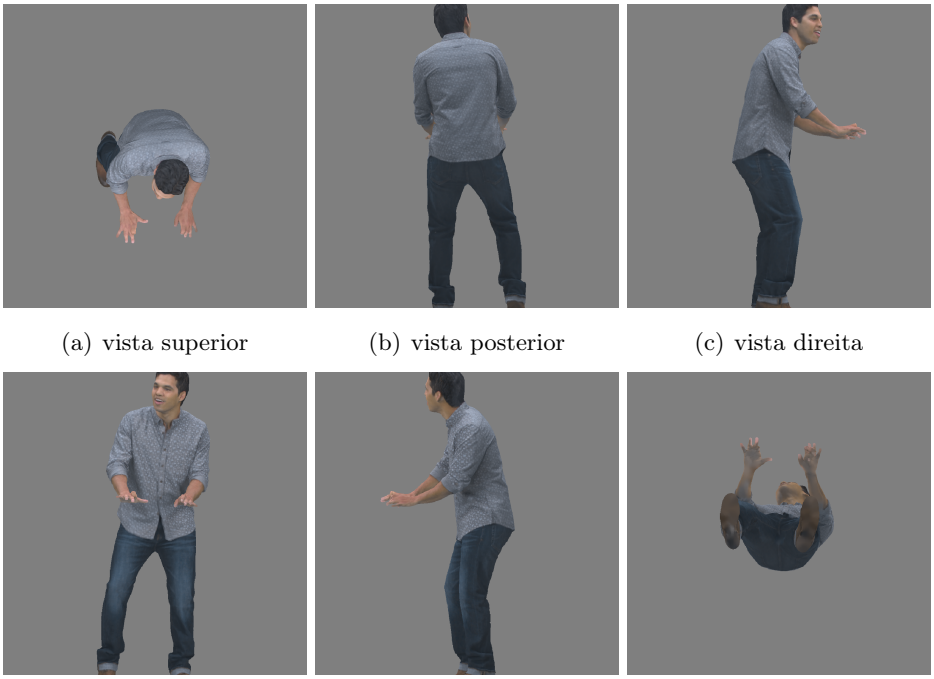
Figura 3.11: Projeções ao redor da nuvem de pontos de referência do conteúdo *biplane*.



(a) vista superior (b) vista posterior (c) vista direita

(d) vista anterior (e) vista esquerda (f) vista inferior

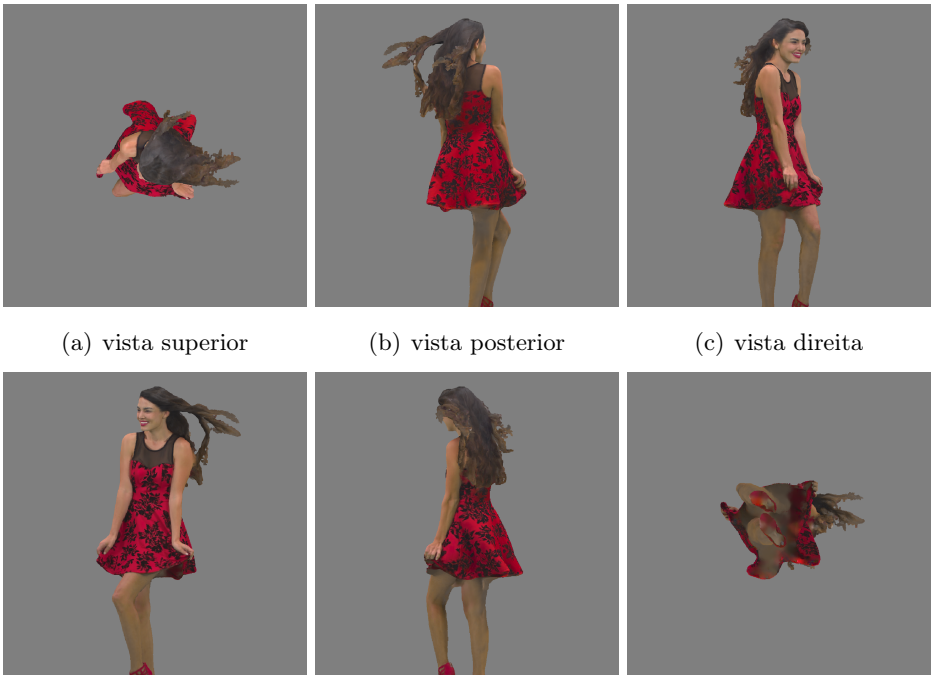
Figura 3.12: Projeções ao redor da nuvem de pontos de referência do conteúdo *longdress*.



(a) vista superior (b) vista posterior (c) vista direita

(d) vista anterior (e) vista esquerda (f) vista inferior

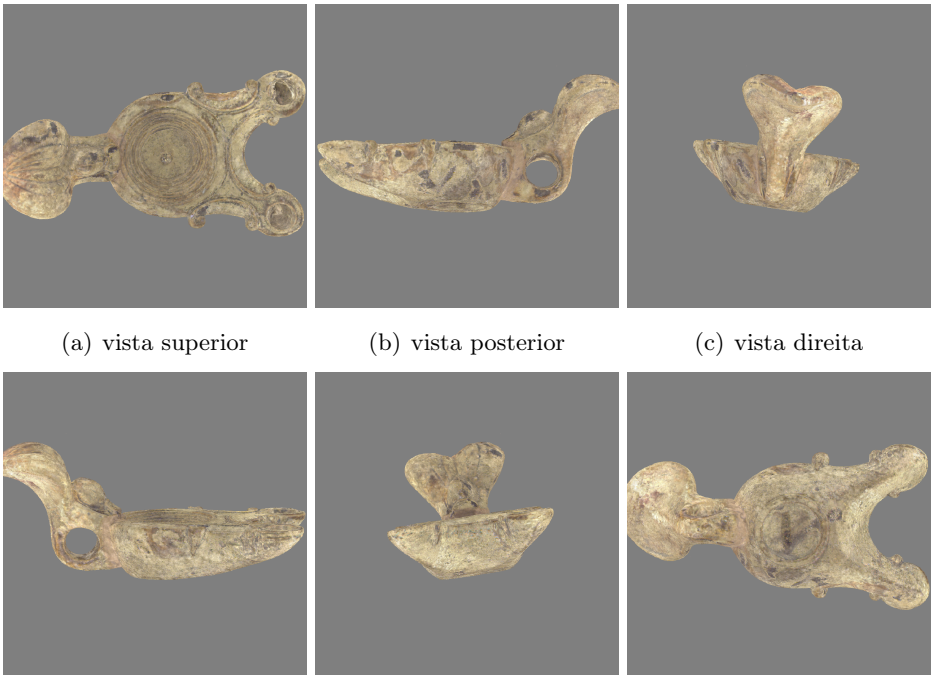
Figura 3.13: Projeções ao redor da nuvem de pontos de referência do conteúdo *loot*.



(a) vista superior (b) vista posterior (c) vista direita

(d) vista anterior (e) vista esquerda (f) vista inferior

Figura 3.14: Projeções ao redor da nuvem de pontos de referência do conteúdo *redandblack*.



(a) vista superior (b) vista posterior (c) vista direita

(d) vista anterior (e) vista esquerda (f) vista inferior

Figura 3.15: Projeções ao redor da nuvem de pontos de referência do conteúdo *romanoillamp*.

Capítulo 4

Resultados Experimentais

4.1 Introdução

Neste Capítulo são apresentados os resultados obtidos através dos experimentos subjetivos realizados ao longo deste estudo. São discutidos tanto os comportamentos dos dados respectivos às notas subjetivas em si, quanto as relações entre as notas e as métricas objetivas exploradas.

Especialmente são investigadas as relações com as métricas objetivas obtidas através do *framework* proposto (métricas projetivas). Como as métricas projetivas exploradas são baseadas na aplicação de métricas bidimensionais já estabelecidas para a avaliação de imagens, as métricas projetivas são denominadas pelo nome da respectiva métrica bidimensional acrescido do prefixo "P" (e.g. a versão projetiva da métrica PSNR é denominada P-PSNR).

A seguir, na Seção 4.2, são mostrados os resultados do experimento baseado no método *Absolute Category Rating with Hidden Reference* (ACR-HR). Logo após, são mostrados os resultados acerca do experimento baseado no método *Double Stimulus Impairment Scale*, na Seção 4.3.

4.2 Experimento ACR-HR

Uma análise inicial da Figura 4.1 demonstra que a P-PSNR consegue discriminar um sinal de alta qualidade dentre outros. Os resultados também demonstram uma correlação positiva entre DMOS e P-PSNR com respeito ao sinal original. As avaliações de qualidade aparentaram ser altamente dependentes do conteúdo presente na cena observada, e níveis de qualidade intermediária são correlacionados de maneira menos consistente com a P-PSNR [10].

A sequência *Loot* revelou um comportamento inesperado quando usuários tenderam a preferir a versão "alta qualidade passa baixas" em vez da versão "alta qualidade", mesmo a primeira introduzindo mais distorção e apresentando uma P-PSNR inferior. Isso pode ser justificado por um viés causado pelo conteúdo da cena [10]. Todas as outras relações entre as notas mantiveram o comportamento esperado de correlação com a P-PSNR.

Os resultados foram realizados com um número relativamente pequeno de participantes. Por

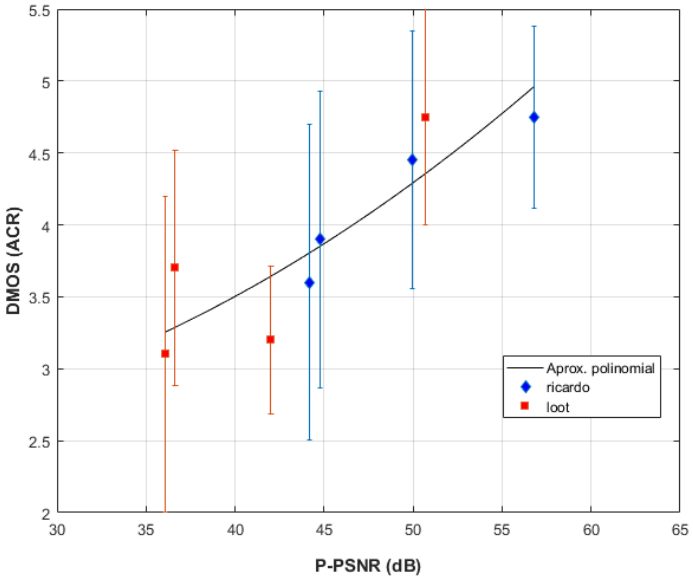


Figura 4.1: Comparação entre a PSNR projetada em 6 vistas e as notas subjetivas dos participantes obtidas no experimento baseado em ACR-HR.

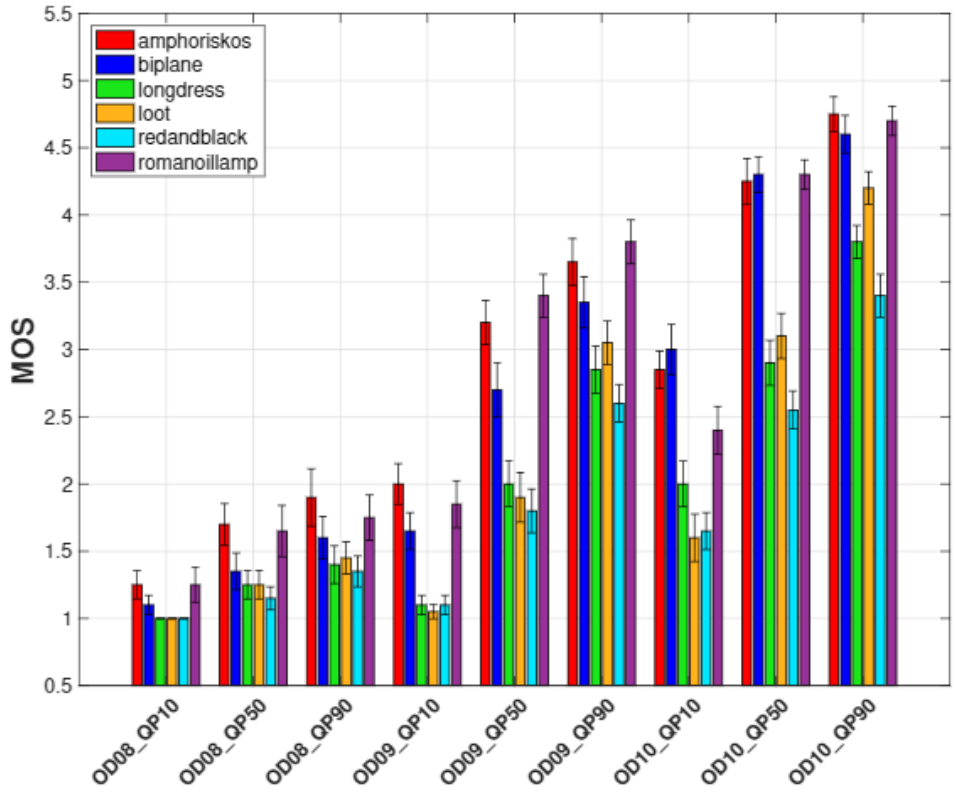
consequência foi encontrada uma variância significativa nos dados obtidos. Apesar de os resultados não serem estatisticamente fortes, eles indicam o potencial da métrica proposta e justificam um estudo mais extenso.

4.3 Experimento DSIS

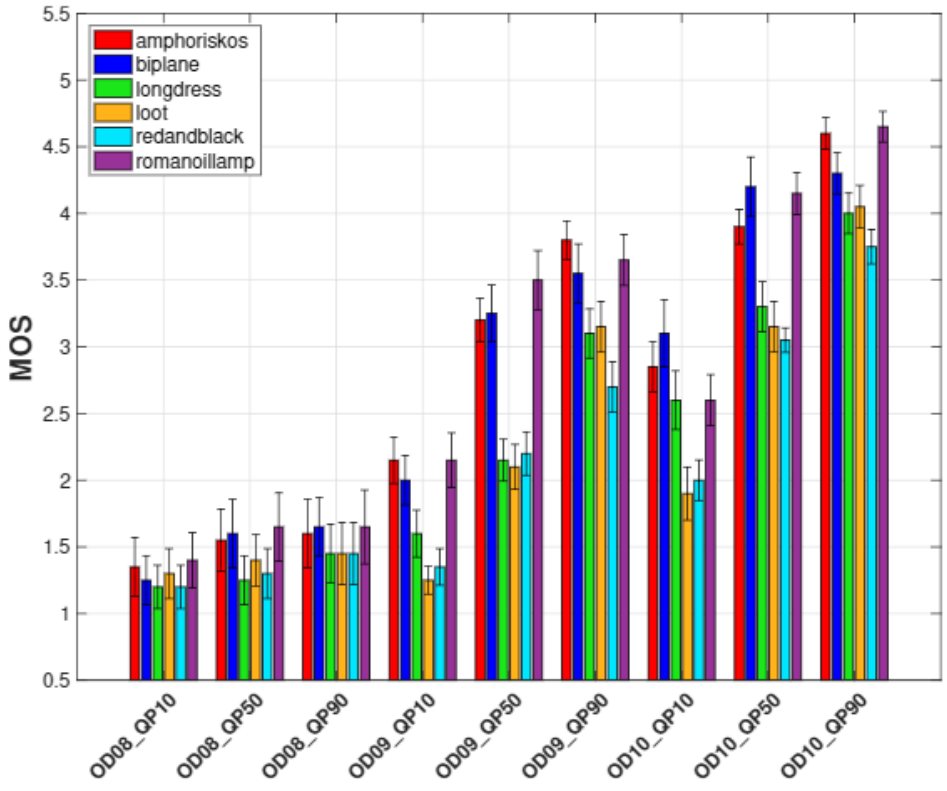
Nesta seção, apresentam-se e discutem-se as notas subjetivas coletadas durantes os experimentos com esquema DSIS e os resultados de desempenho de métricas objetivas de qualidade no estado da arte baseadas em projeção e baseadas em pontos. Para se referir às métricas pré-existentes baseadas em pontos, nas tabelas e figuras desta seção se usam as abreviaturas po2point, po2plane e pl2plane para indicar se a métrica em questão é baseada nas distâncias ponto-a-ponto, ponto-a-plano e plano-a-plano, respectivamente. A métrica de cor baseada em pontos explorada, denominada $\mathrm{PSNR}_{\mathrm{YUV}}$, é calculada através da fórmula definida na Equação 2.13. Para se referir às métricas baseadas em projeções, propostas neste estudo, o prefixo P é omitido nas tabelas e figuras em função de clareza visual, e a métrica é identificada pela métrica 2D aplicada nas projeções.

4.3.1 Análise das notas subjetivas

Notas subjetivas obtidas dos dois laboratórios se mostraram estatisticamente distintas. Por isso, a análise comparativa entre as métricas foi feita de maneira separada entre os dois conjuntos de dados. Além disso, as notas também se mostraram estatisticamente distintas entre tipos diferentes de conteúdos. Assim, a análise foi separada também em três conjuntos de dados, para os conjuntos

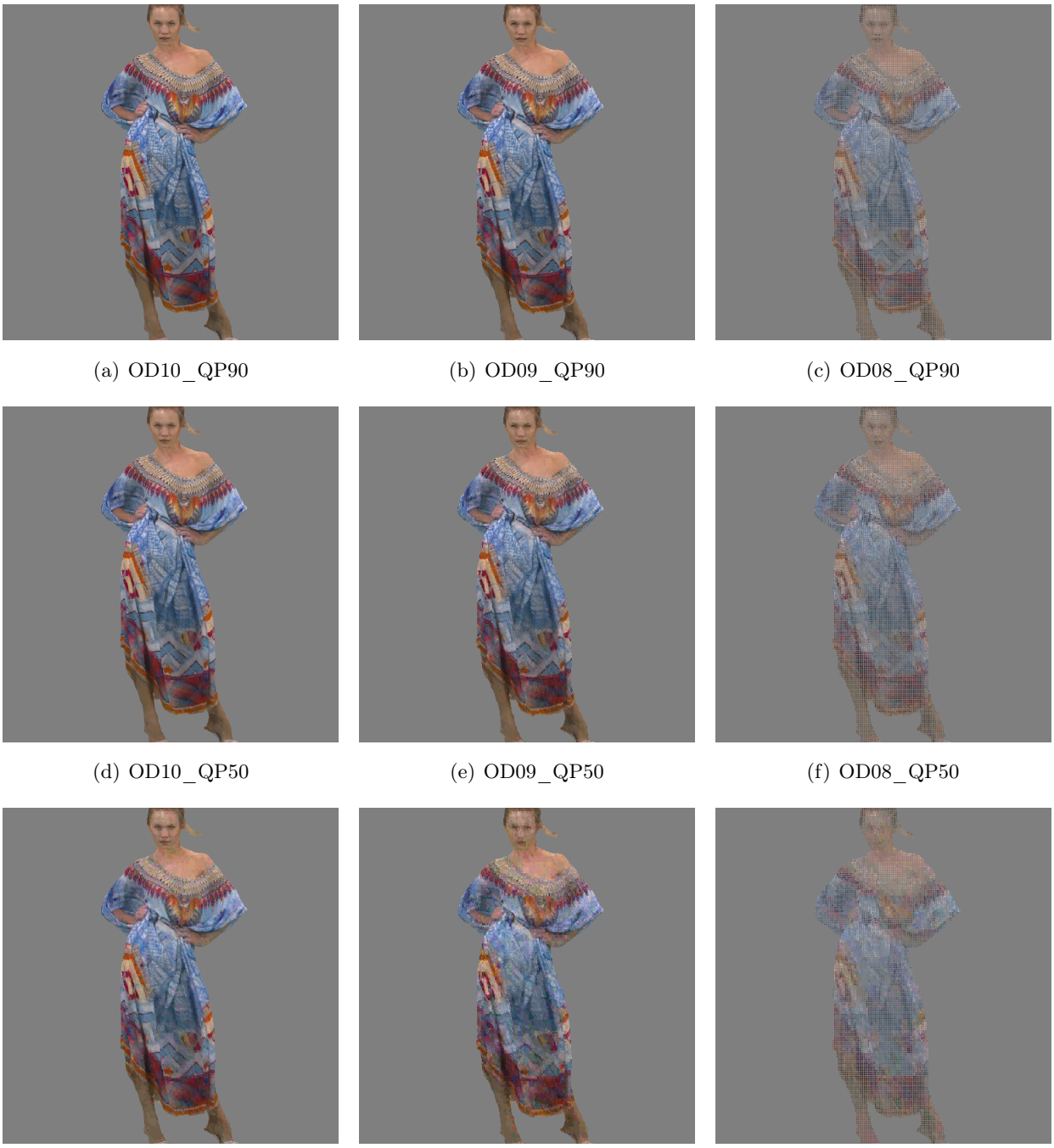


(a) EPFL



(b) UNB

Figura 4.2: Avaliações subjetivas de cada conteúdo, separadas por degradação.



(a) OD10_QP90 (b) OD09_QP90 (c) OD08_QP90

(d) OD10_QP50 (e) OD09_QP50 (f) OD08_QP50

(g) OD10_QP10 (h) OD09_QP10 (i) OD08_QP10

Figura 4.3: Diferentes níveis de distorção aplicados ao conteúdo *longdress*.

de dados de cada laboratório: o conjunto de dados completo (i), dados referentes a conteúdos mostrando corpos humanos (ii) e dados referentes a conteúdos mostrando objetos inanimados (iii).

Os resultados acerca das notas subjetivas referentes às 6 nuvens de pontos descritas na Seção 3.4.2.1 são mostrados na Figura 4.2. Cada subfigura indica o local de onde os dados foram obtidos. São mostrados os histogramas das notas de opinião médias (MOS) de cada conteúdo e cada tipo de degradação, com seus respectivos intervalos de confiança. Foi adotada a seguinte

convenção de nomenclatura: dada uma profundidade de *octree* (OD) igual a XX $\in \{08, 09, 10\}$, que determina nível de qualidade da geometria da nuvem de pontos, e dada um parâmetro de qualidade JPEG (QP) igual a YY $\in \{10, 50, 90\}$, que denota o nível de qualidade de cor da nuvem de pontos, o conteúdo é denominado ODXX_QPYY. Exmplos das diferentes combinações de distorção do conteúdo *longdress* se encontram na Figura 4.3.

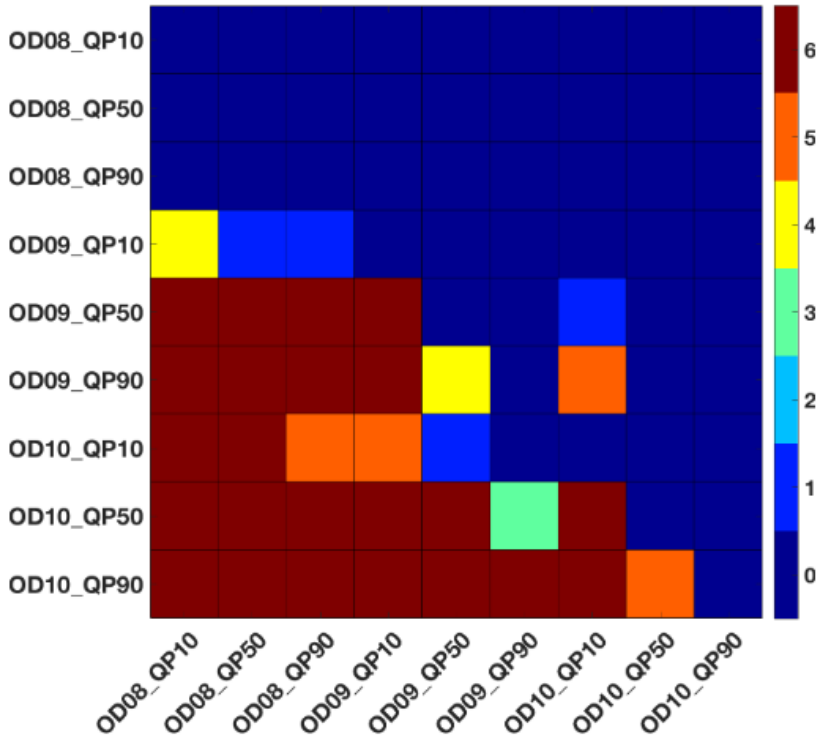
Baseado nos resultados indicados na Figura 4.2, as notas subjetivas variam entre os tipos de degradação, mesmo para um mesmo conteúdo. Em particular, se nota que para as versões de geometria mais esparsa dos conteúdos (OD = 08), independentemente do conteúdo em si, melhorias na qualidade de cor (aumento do QP) causam um aumento relativamente mais lento da nota subjetiva média. As avaliações aumentam mais rapidamente quando a resolução da geometria é maior. Isso indica que quando a geometria do conteúdo tem uma resolução muito baixa, como na Figura 4.3f, a percepção de qualidade é consideravelmente limitada, independentemente de melhorias na cor. Esse é o caso pelo menos quando uma geometria pouco densa se traduz na falta de espaços preenchidos pelo modelo (i.e. buracos entre os pontos).

Neste estudo não se exploraram efeitos que diferentes métodos de visualização podem ter na percepção subjetiva dos avaliadores, como quando geometrias mais esparsas são interpoladas de modo a preencher espaços vazios entre os pixels. No visualizador utilizado, o tamanho máximo dos voxels/pixels renderizados foi mantido em um tamanho reduzido de modo a evitar efeitos de borramento (*blurring*) do conteúdo de referência. Assim, se observaram lacunas nos modelos principalmente para uma profundidade de *octree* igual a 08. Para nuvens de pontos de *octree* de profundidade igual a 09, tais artefatos só se tornavam visíveis quando o avaliador inspecionava os modelos com um nível de ampliação (*zoom*) considerável.

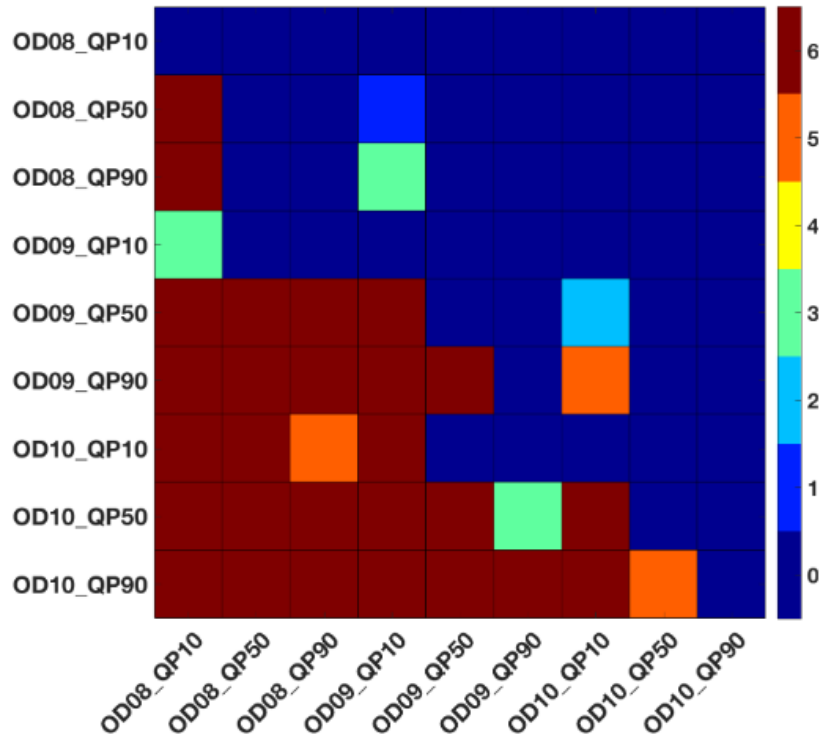
Outra razão que explica os comportamentos das avaliações de conteúdos com os menores níveis de qualidade de geometria é o uso da estrutura de *octree* como base para a compressão. À medida que se reduz a resolução de geometria de uma *octree*, um número cada vez maior de pontos da nuvem de pontos original se encontra em nós folha da árvore. Considerando que a cor de um ponto após compressão é determinada pela combinação dos pontos que se encontravam em um mesmo nó folha antes da compressão, inerentemente os detalhes de cor de uma nuvem de pontos são limitados pela resolução (e consequentemente, pela degradação) de sua geometria.

Outra informação que pode ser deduzida da Figura 4.2 é que, para um dado tipo de degradação, o perfil de variação da qualidade percebida varia consideravelmente dependendo do tipo de conteúdo em questão. Especificamente, avaliadores tenderam a ser mais críticos em relação a conteúdos que representavam humanos do que aos conteúdos representando objetos inanimados. Também se observam desvios menores nas notas de conteúdos que pertencem ao mesmo tipo, indicando que comportamentos de avaliações similares ocorrem dentro de cada um dos grupos.

Ao se analisar as taxas de *bits* dos conteúdos codificados, como estão descritas na Tabela 3.2, e comparar com as notas registradas na Figura 4.2, é possível concluir que taxas de *bits* mais altas não necessariamente resultam em uma maior qualidade visual percebida. Por exemplo, em todos os conteúdos participantes de ambos os locais de teste demonstraram uma preferência significativa pela combinação entre melhor qualidade de cor (QP = 90) com uma qualidade média de geometria



(a) EPFL



(b) UNB

Figura 4.4: Matriz de diferença de significância com nível de confiança de 5% das preferências subjetivas dos participantes nos testes, comparadas com cada outra combinação de distorções.

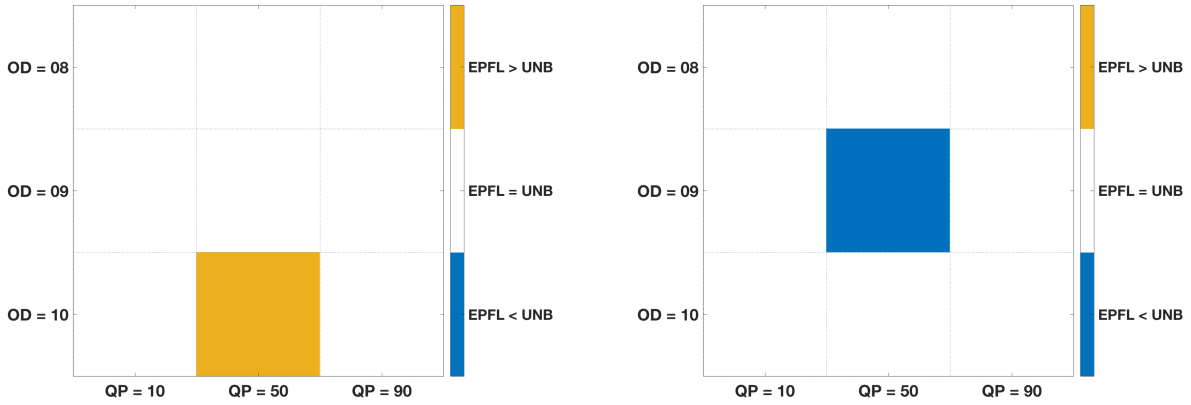
(OD = 09), em vez da melhor qualidade de geometria (OD = 10) combinada com a pior qualidade de cor (QP = 10), mesmo a primeira combinação apresentando uma taxa de bits menor do que a segunda para todos os conteúdos avaliados.

Apesar dos valores de taxa de *bits* por ponto (bpp) serem dependentes do *codec* utilizado, as observações feitas sugerem que uma melhor relação entre qualidade visual e recursos exigidos (tanto para espaço de armazenamento como transmissão) pode ser atingida com a apropriada alocação de *bits* para a representação de geometria e cor. Para corroborar as observações, foi realizado um teste-t unicaudal com nível de confiança de 5%. O teste foi aplicado separadamente aos dados obtidos em cada laboratório. A hipótese nula assumiu que uma nota média, obtida através da média das avaliações de todos os conteúdos dadas um determinado nível de geometria e de cor, é a mesma para qualquer outra combinação de níveis de geometria e de cor. Os resultados estão dispostos na Figura 4.4. Em cada uma das sub-figuras, a cor do ladrilho de posição (X, Y) representa quantas vezes, de 6 comparações no total, a combinação ODYY_QPYY foi preferida contra a combinação ODXX_QPXX.

Considerando os dados obtidos na EPFL, as combinações OD09_QP50 e OD09_QP90, em 6 comparações com a combinação OD10_QP10, foram preferidas respectivamente 1 e 5 vezes. Um padrão similar é observado com os dados obtidos na UnB, com as combinações mencionadas sendo preferidas 2 e 5 vezes contra a combinação OD10_QP10, respectivamente. De maneira menos pronunciada, ainda considerando os dados provenientes da UnB, uma preferência por taxas de *bits* menores foi observado até mesmo em para níveis mais baixos de geometria, com as combinações OD08_QP50 e OD08_QP90 sendo preferidas em vez da combinação OD09_QP10 1 e 3 vezes, respectivamente.

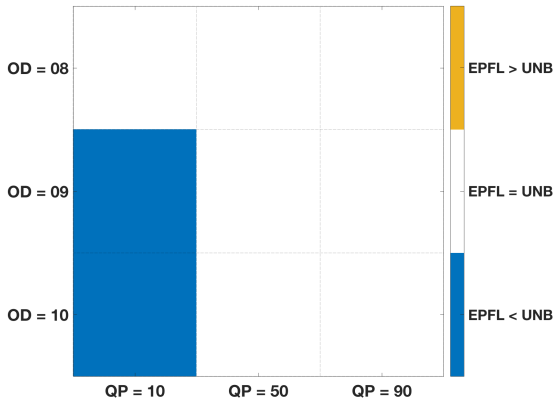
É importante constatar que houve diferenças entre as notas obtidas em cada laboratório. Por exemplo, observando a Figura 4.2, participantes na EPFL demonstraram uma rejeição maior a degradações mais intensas de cor (QP = 10) do que participantes na UnB, especialmente com nuvens de pontos contendo corpos humanos. Um teste t unicaudal com nível de confiança de 5% foi realizado para determinar se os comportamentos dos participantes diferiram de maneira significante, estatisticamente. Resultados estão dispostos na Figura 4.5. A hipótese nula nesse caso considerou que a MOS calculada para cada conteúdo degradado era a mesma entre os dois laboratórios. De acordo com os resultados do teste, em casos que a hipótese nula foi rejeitada participantes na EPFL apresentaram avaliações com notas menores do que participantes na UnB. Essa diferença é observada principalmente para níveis de qualidade de cor intermediários ou menores, com a única exceção sendo o conteúdo *amphoriskos* com qualidade de cor média e qualidade de geometria inferior, que foi avaliado com uma nota mais baixa na UnB. É interessante notar que o conteúdo *redandblack* recebeu notas maiores na UnB do que na EPFL na maioria das combinações de distorções.

Também foi realizado um teste ANOVA multivariado para corroborar as observações feitas. Resultados estão presentes na Tabela 4.1. Os valores p obtidos sugerem que tanto o laboratório onde o experimento foi realizado (UnB versus EPFL), como o tipo de conteúdo (corpos humanos versus objetos inanimados) e os níveis de geometria e de cor levaram a conjuntos de dados que são

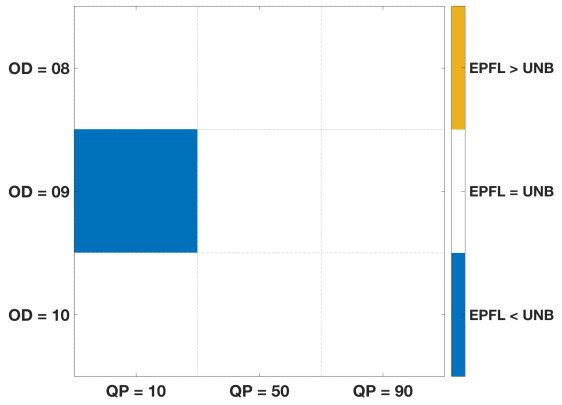


(a) *amphoriskos*

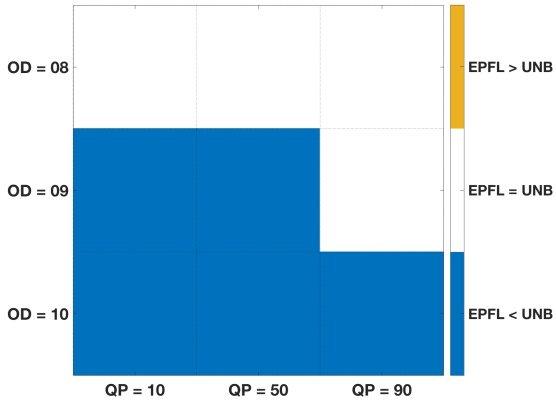
(b) *biplane*



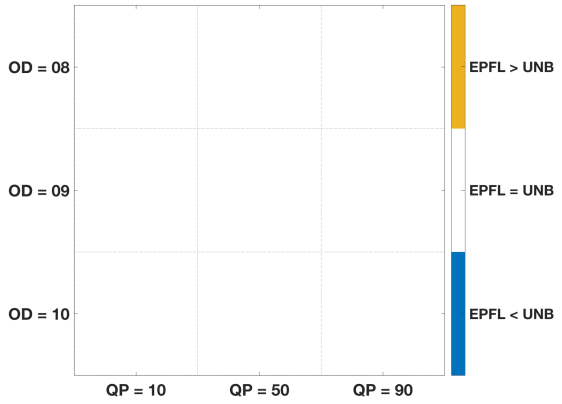
(c) *longdress*



(d) *loot*



(e) *redandblack*



(f) *romanoillamp*

Figura 4.5: Matrizes de diferença de significância com nível de confiança de 5% indicando se participantes do experimento em um laboratório avaliaram a qualidade visual, acerca de uma dada degradação de um conteúdo em particular, de maneira significativamente mais alta ou mais baixa em relação a participantes do teste no outro laboratório.

estatisticamente distintos entre si, com um intervalo de confiança de 5%.

4.3.2 Comparação entre métricas objetivas

Para a comparação das métricas objetivas investigadas, se considerou que os dados obtidos nos testes realizados nos dois laboratórios eram estatisticamente distintos entre si, como foi demonstrado na Seção 4.3.1. Sendo assim, o *benchmarking* foi realizado nos dois conjuntos de dados separadamente. Além disso, dado que as notas subjetivas se mostraram estatisticamente distintas também para cada tipo de conteúdo, a análise também foi feita de maneira separada para 3 conjuntos de dados:

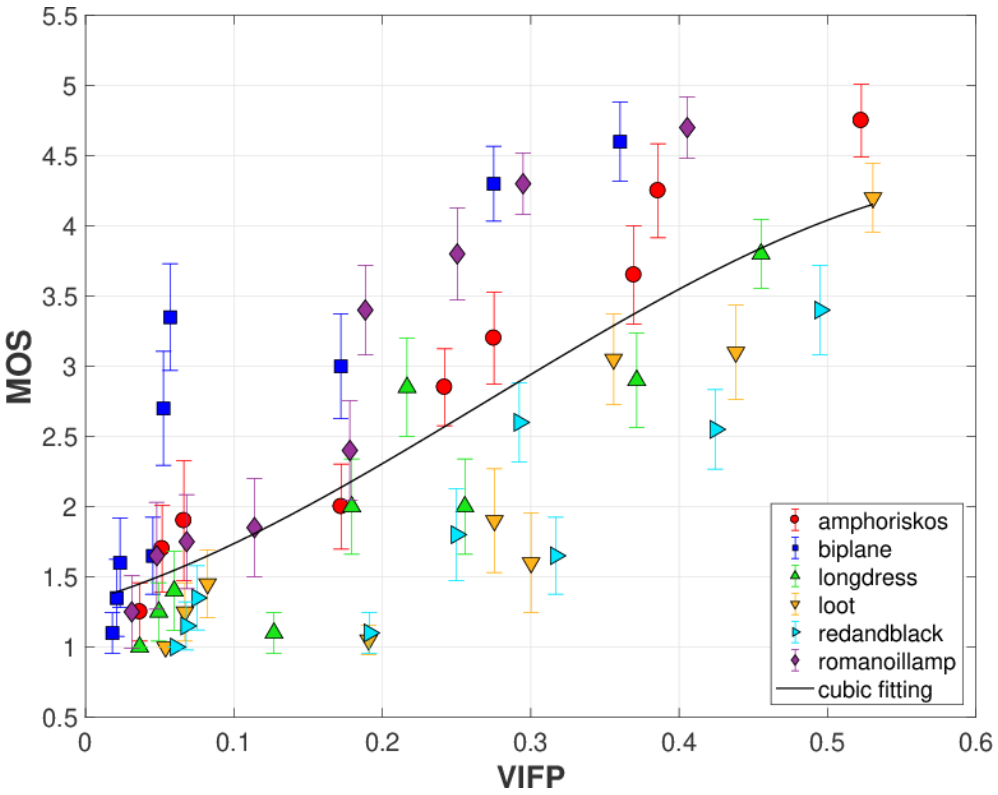
1. O conjunto de dados completo, incluindo todos os conteúdos
2. Dados respectivos somente a conteúdos contento objetos inanimados
3. Dados respectivos somente a conteúdos contendo corpos humanos

Na Tabela 4.2, estão dispostos os índices de desempenho para cada métrica objetiva de qualidade comparada com as notas subjetivas obtidas na EPFL, consideradas como valores de referência. Para cada par de métrica e conjunto de dados, foram comparados o coeficiente de correlação de Pearson [62] (PCC), o coeficiente de correlação de postos de Spearman [63] (SROCC), a raiz do erro quadrático médio (RMSE) e a proporção de *outliers* (OR), este último definido como a proporção de dados que se encontram fora do intervalo de confiança da curva ajustada. De maneira geral, as métricas baseadas em projeção demonstram um desempenho melhor que as métricas baseadas em pontos. A métrica de melhor desempenho para o conjunto de dados completo foi a baseada em VIFP, apesar de que com uma baixa correlação entre métrica objetiva e notas subjetivas.

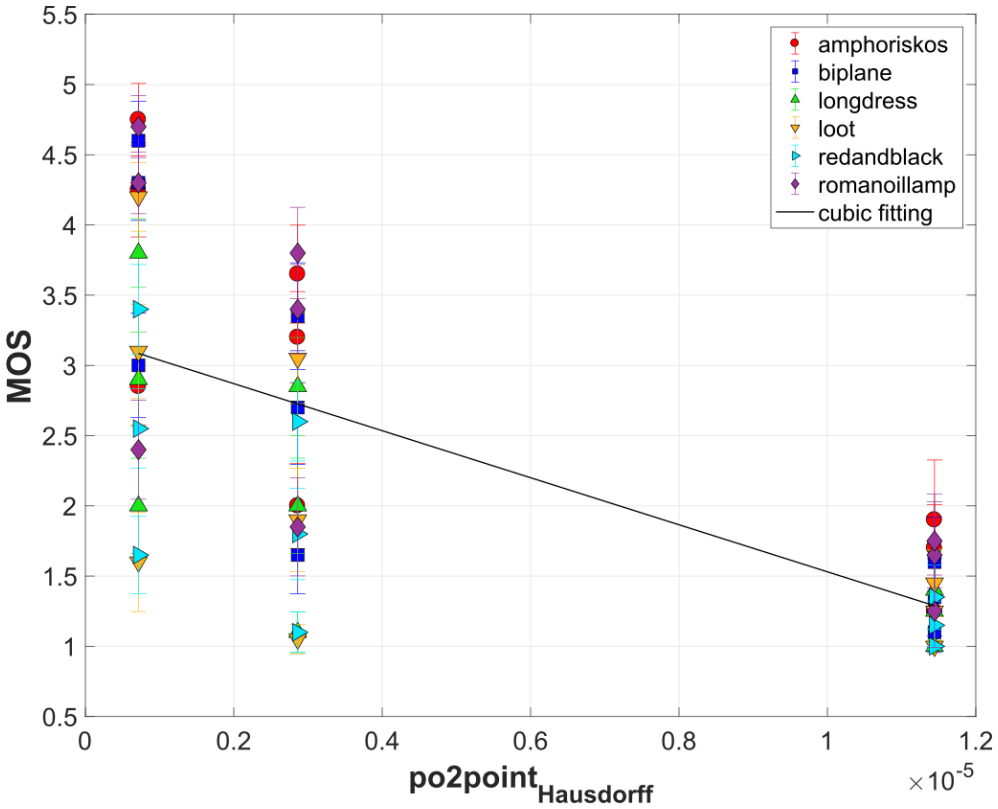
A correlação melhora drasticamente quando se comparam as métricas referentes apenas para os determinados tipos de conteúdos separadamente. Particularmente, tanto o MSSIM como o VIFP demonstram alto poder preditivo, em ambos os conjuntos de dados. Especialmente, o MSSIM apresenta um resultado relativamente maior para objetos inanimados e o VIFP para corpos humanos.

Na Figura 4.6, são apresentados gráficos da distribuição de notas objetivas da métrica baseada em projeções e baseada em pontos de melhor desempenho versus notas subjetivas de todos os conteúdos, junto com um ajuste de curvas cúbico, para cada métrica. Também são apresentados gráficos das distribuições de notas objetivas versus notas subjetivas para os conteúdos contendo objetos inanimados na Figura 4.7 e versus conteúdos contendo corpos humanos na Figura 4.8.

Métricas baseadas em pontos se demonstram limitadas por não conseguirem examinar simultaneamente degradações de cor e de geometria. Nas Figuras 4.6b, 4.7b, 4.8b, 4.9b e 4.10b, cada conteúdo está associado com uma nota determinada inteiramente pela profundidade da *octree*. No entanto, à medida que a qualidade de cor é elevada e as notas subjetivas aumentam, a métrica é incapaz de discriminar entre as versões, atribuindo a mesma nota objetiva para conteúdos de qualidades perceptivelmente distintas. Em contraposição, as métricas baseadas em projeções não

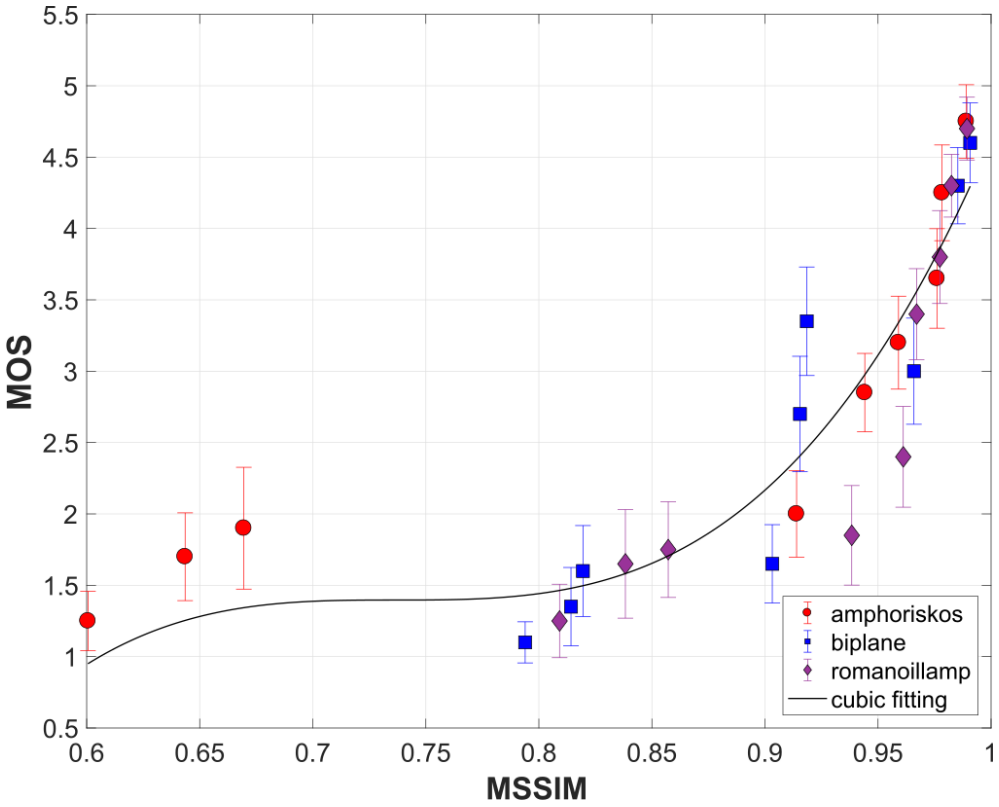


(a) Baseada em projeções.

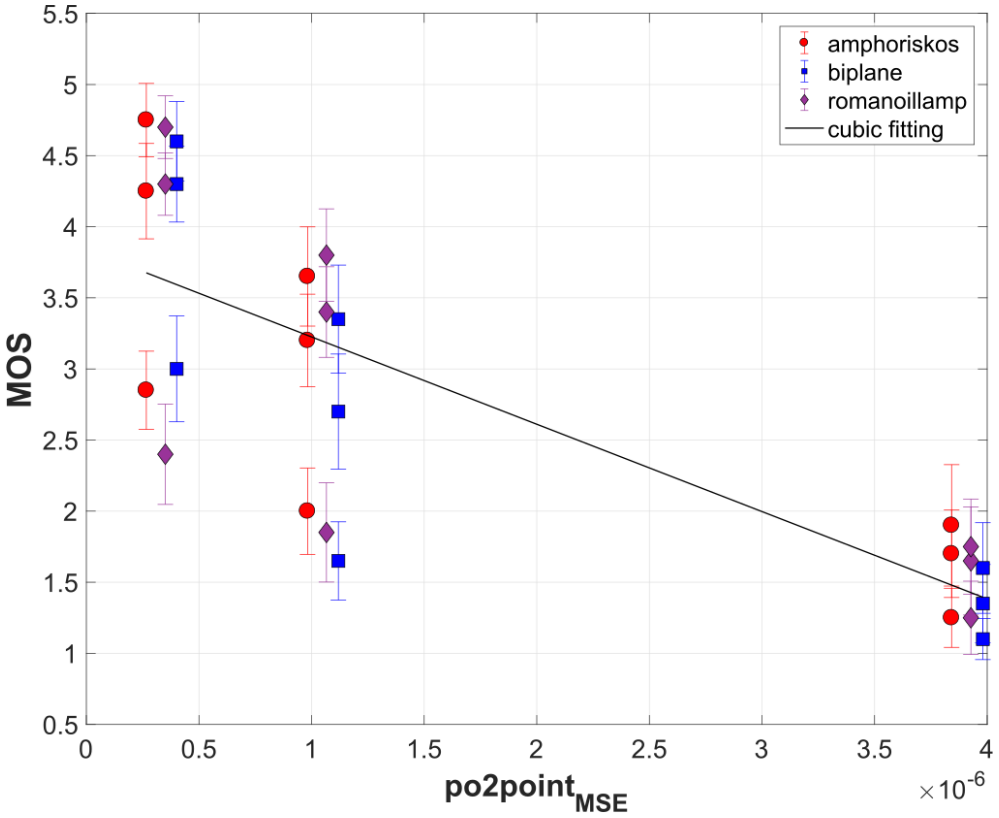


(b) Baseada em pontos.

Figura 4.6: Métricas de maior correlação com notas subjetivas respectivas a todos os conteúdos provenientes da EPFL.

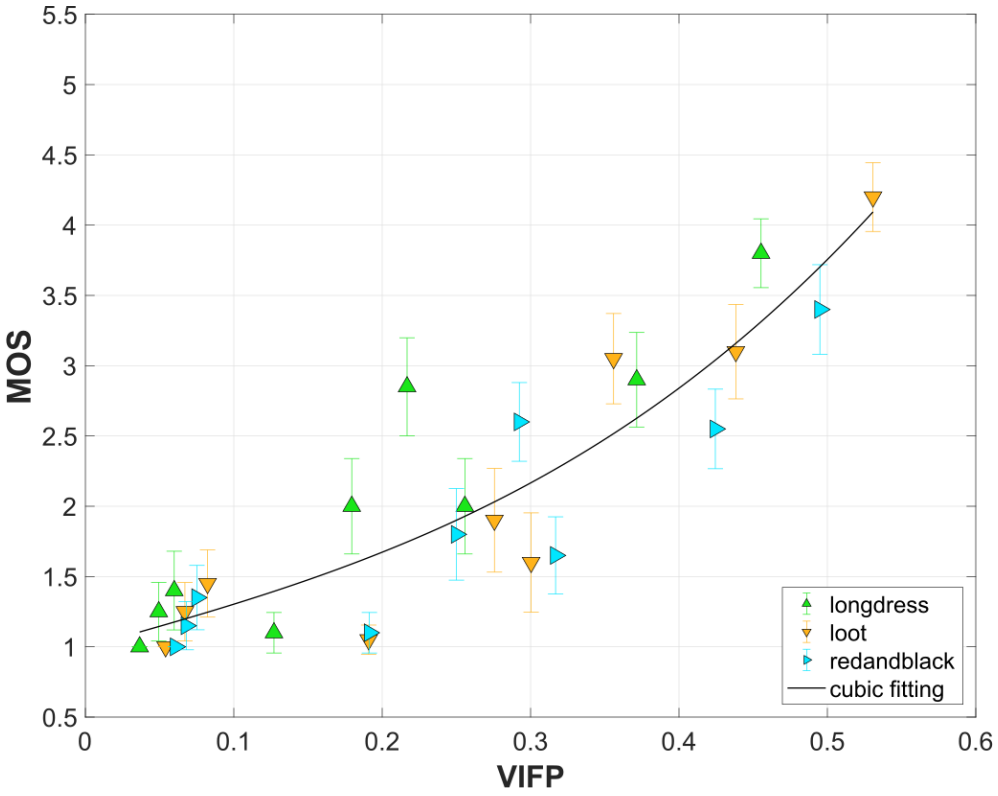


(a) Baseada em projeções.

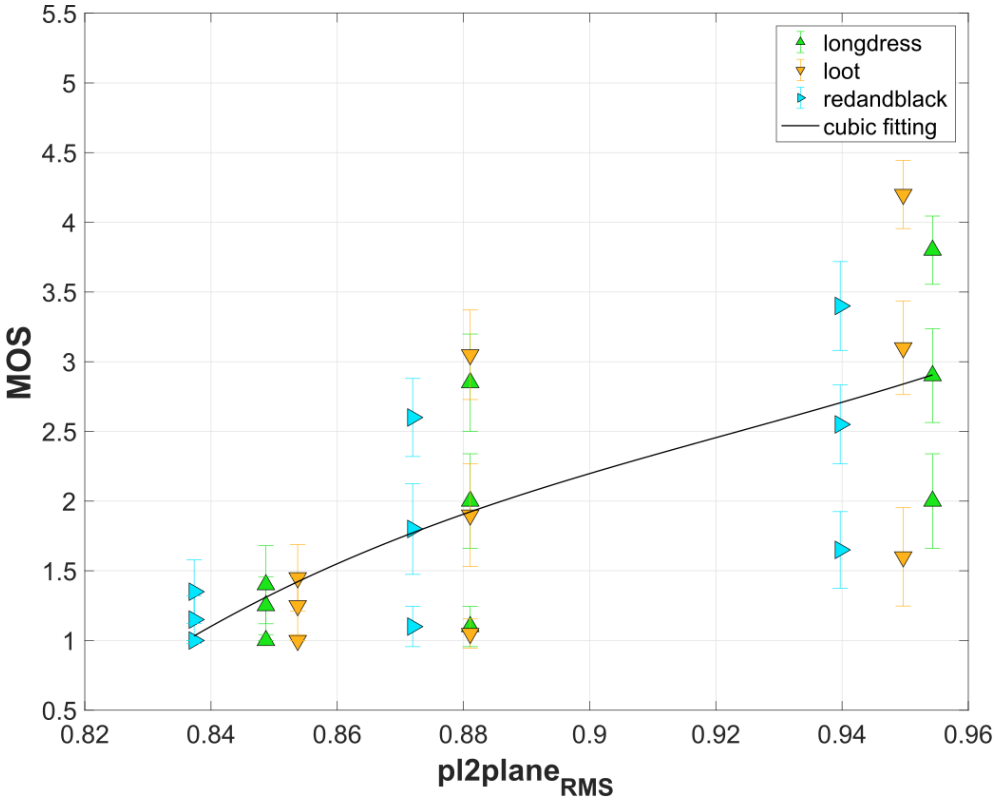


(b) Baseada em pontos.

Figura 4.7: Métricas de maior correlação com notas subjetivas respectivas a conteúdos contendo objetos inanimados provenientes da EPFL.

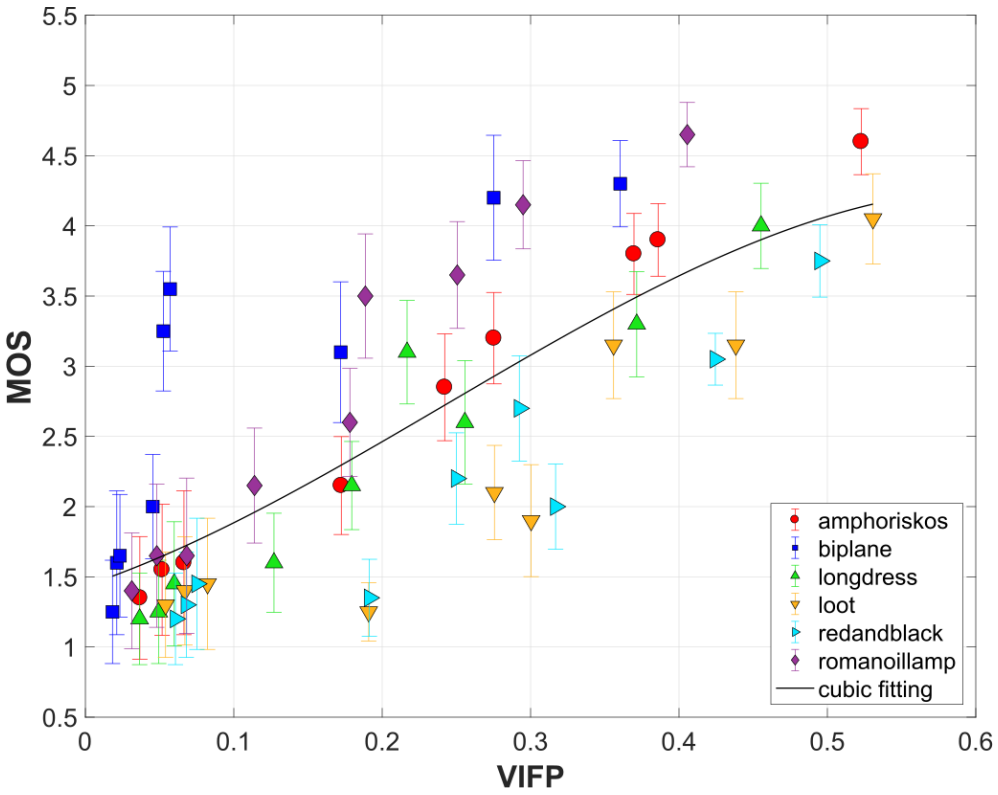


(a) Baseada em projeções.

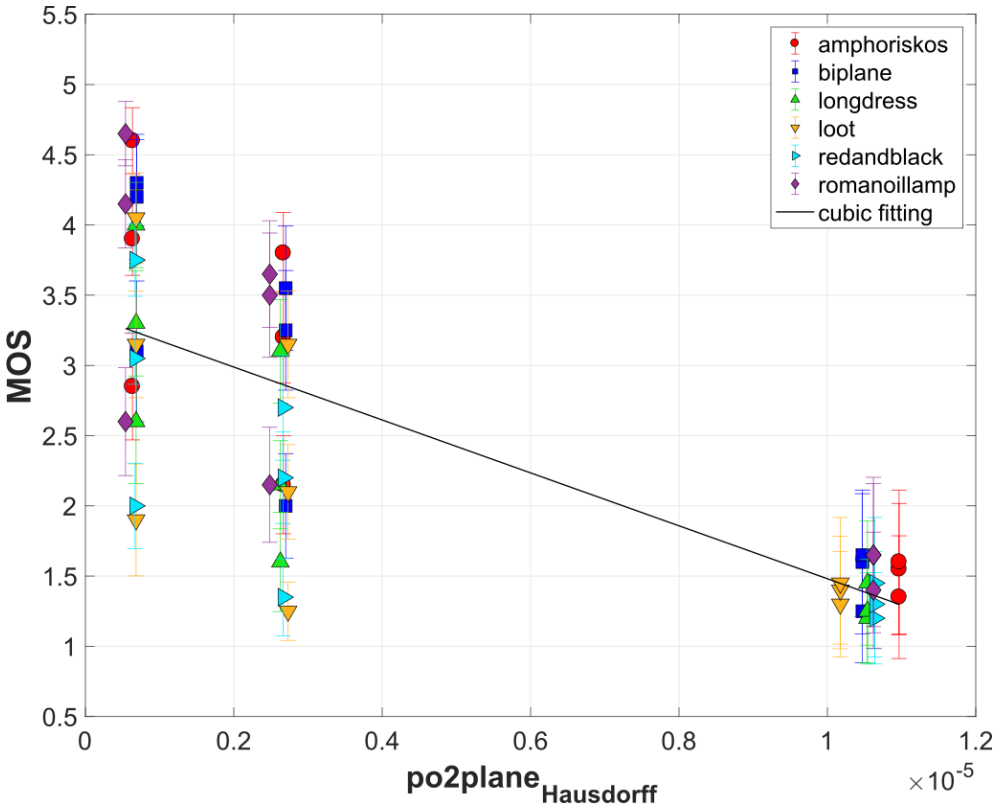


(b) Baseada em pontos.

Figura 4.8: Métricas de maior correlação com notas subjetivas respectivas a conteúdos contendo corpos humanos provenientes da EPFL.

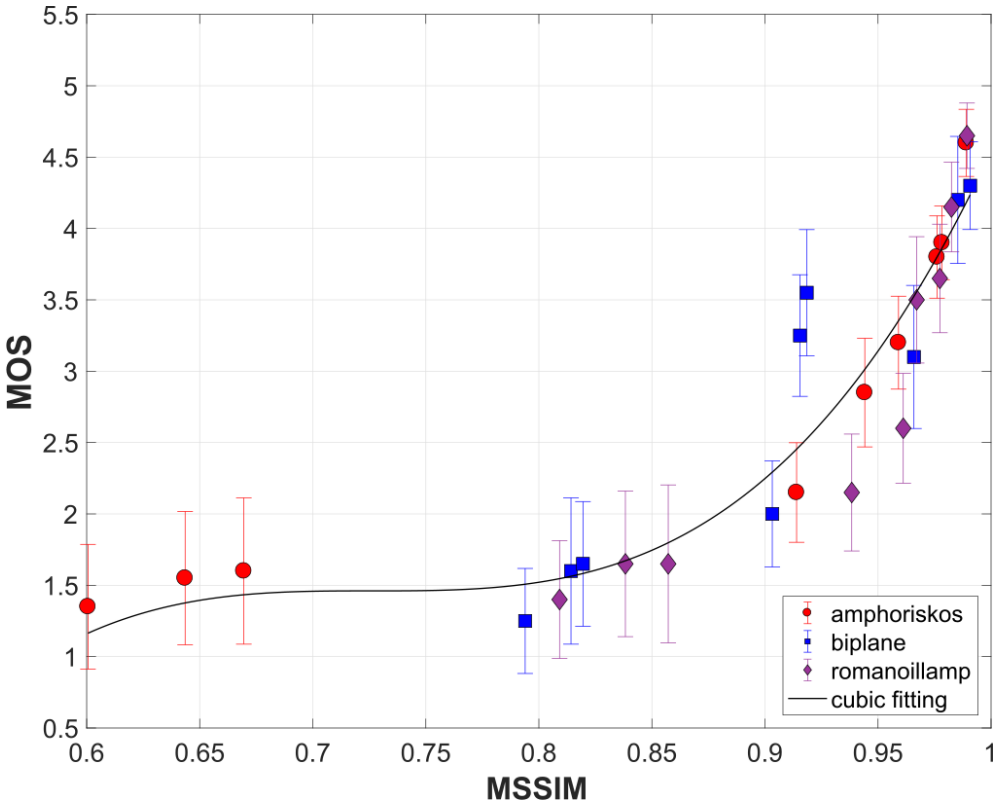


(a) Baseada em projeções.

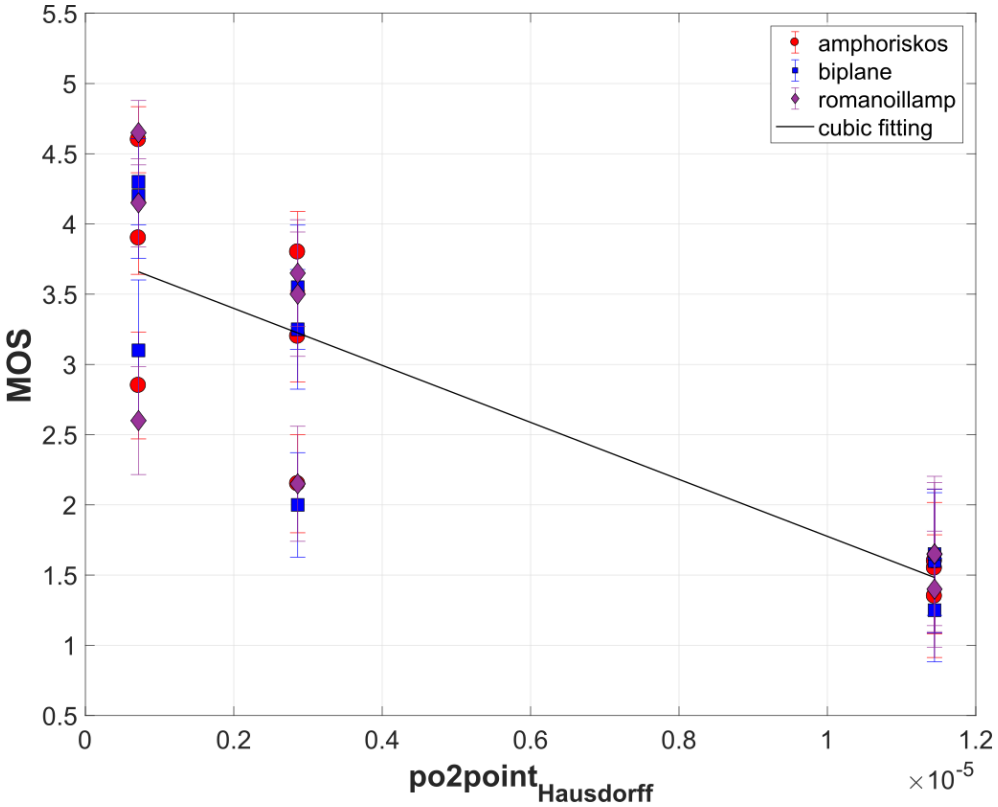


(b) Baseada em pontos.

Figura 4.9: Métricas de maior correlação com notas subjetivas respectivas a todos os conteúdos provenientes da UnB.

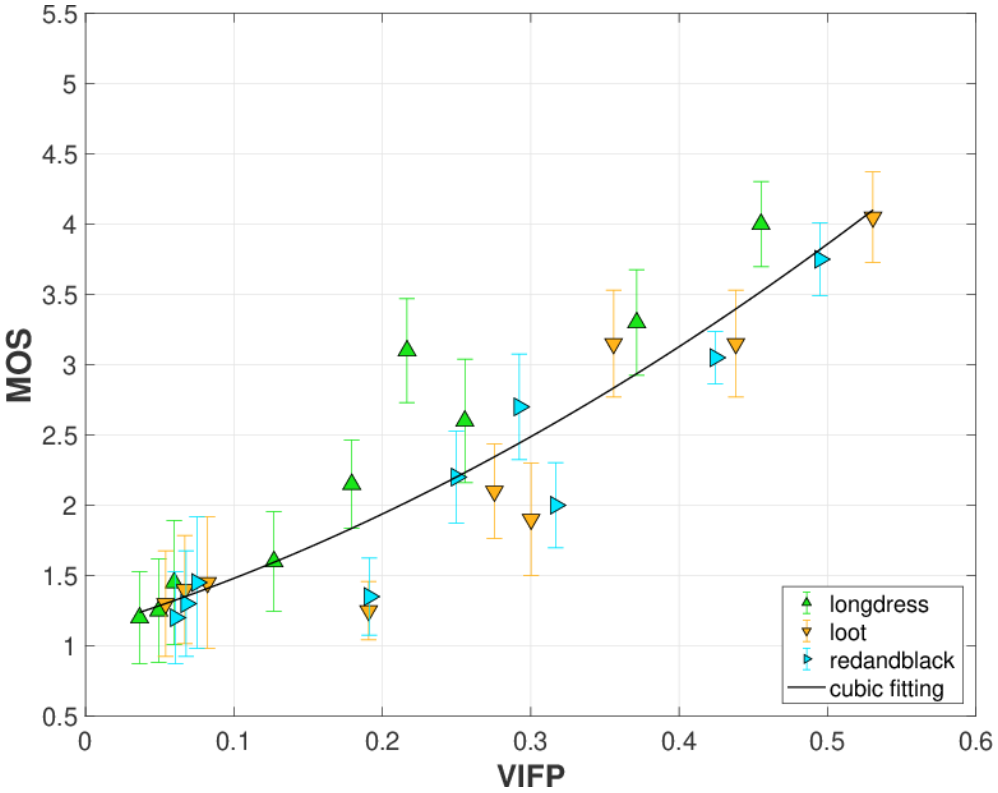


(a) Baseada em projeções.

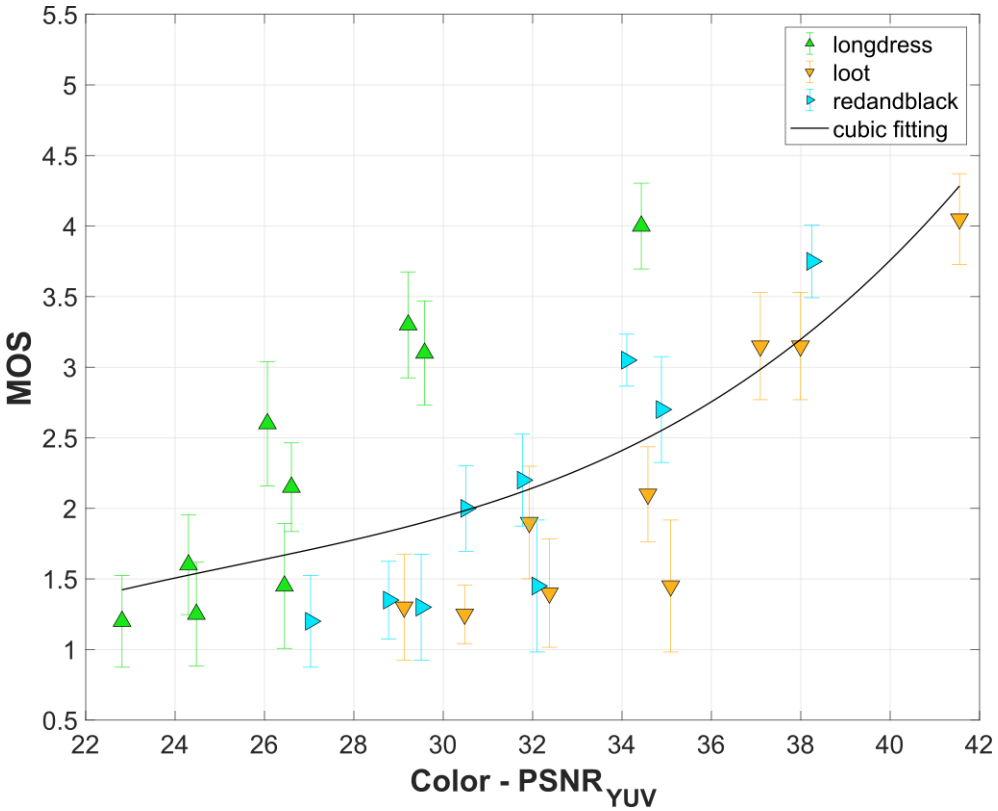


(b) Baseada em pontos.

Figura 4.10: Métricas de maior correlação com notas subjetivas respectivas a conteúdos contendo objetos inanimados provenientes da UnB.



(a) Baseada em projeções.



(b) Baseada em pontos.

Figura 4.11: Métricas de maior correlação com notas subjetivas respectivas a conteúdos contendo corpos humanos provenientes da UnB.

demonstram dificuldade em diferenciar entre esses conteúdos e prever com uma boa correlação a qualidade subjetiva dos conteúdos.

A correlação entre as métricas objetivas e os dados de notas subjetivas obtidos na UnB estão dispostos na Tabela 4.3. Foram usados as mesmas relações estatísticas empregadas na Tabela 4.2. Gráficos da distribuição de notas objetivas em função de notas subjetivas de todos os conteúdos se encontram na Figuras 4.9, enquanto que na Figura 4.10 se encontra a distribuição apenas para conteúdos contendo objetos inanimados, e na Figura 4.11 a distribuição é feita considerando apenas conteúdos com corpos humanos. Ocorre um comportamento similar ao observado para o conjunto de dados obtidos na EPFL. A correlação para todas as métricas é mais forte quando dados referentes a objetos e a modelos humanos são analisados separadamente, e em ambos os casos as duas métricas com maior correlação com as notas subjetivas são novamente o MSSIM e o VIFP.

O RMSE e o OR observados para os dados obtidos na UnB se mantiveram menores com respeito aos valores respectivos aos dados obtidos na EPFL. Uma explicação para esse comportamento são os intervalos de confiança relativamente maiores encontrados nos dados obtidos na UnB, com o intervalo de confiança médio dos mesmos sendo 26,95% maior que os dos dados EPFL.

Outra característica observada é que uma acurácia maior é encontrada para conteúdos representando objetos inanimados. Observadores humanos tendem a avaliar conteúdos que contém seres humanos de maneira mais fina do que outros tipos de conteúdo. Assim, pequenas degradações afetam relativamente mais a qualidade percebida de conteúdos contendo pessoas do que os de outros tipos. Nenhuma das métricas leva essa diferença em consideração, resultando em comportamentos e relações diferentes entre a qualidade prevista através das métricas e a qualidade subjetiva percebida. Isso é confirmado pelos resultados de um teste ANOVA, presentes na Tabela 4.1, que demonstram que as notas não são estatisticamente equivalentes.

É interessante notar que as observações sobre o desempenho relativo das métricas projetivas comparadas se mantêm para dados obtidos em ambos os locais de realização dos experimentos. Mesmo com diferenças estatísticas entre as notas subjetivas em função do local do experimento, a melhor métrica encontrada era a mesma entre os dois locais. Isso é evidência de que a correlação entre métricas projetivas e qualidade visual percebida pode ser robusta entre diferentes populações.

Ressalta-se que, apesar de durante a avaliação subjetiva os participantes terem tido um acesso interativo ao conteúdo, com escolha livre de ponto de vista, apenas 6 pontos de vista distintos e específicos foram usados para calcular a métrica objetiva. Ainda assim, isso foi o suficiente para prever a qualidade visual dos conteúdos. É possível que uma correlação maior seja observada para uma escolha diferente de pontos de vista para as projeções, tanto em maior número ou em arranjos diferentes entre si.

Outro fator que pode ter influenciado os resultados obtidos é a presença do plano de fundo cinza nas projeções. Como ele é incluído no cálculo das métricas objetivas, é possível que essa seja a causa das diferenças observadas em métricas objetivas de conteúdos de notas subjetivas próximas (i.e., *amphoriskos* e *romanoillamp*).

Por fim, uma outra extensão possível do arcabouço proposto é o cálculo contínuo e em tempo real da métrica objetiva enquanto um observador interage com o objeto sendo avaliado, a partir da mesma visualização fornecida para o observador.

Tabela 4.1: ANOVA multivariado.

Fonte	SS	DF	MS	F	p
Laboratório de teste	8.82	1	8.817	13.92	0.0002
Tipo de conteúdo	249.42	1	249.424	393.81	0
Degradação de geometria	1460.84	2	730.422	1153.24	0
Degradação de cor	618.86	2	309.429	488.55	0
Erro	1363.64	2153	0.633		
Total	3701.58	2159			

Tabela 4.2: *Benchmarking* das métricas objetivas considerando os dados obtidos na EPFL como valores de referência. As métricas comparadas encontram-se separadas entre baseadas em projeções (tal qual no *framework* proposto) e baseadas em pontos (já previamente estabelecidas). Os índices de correlação entre cada métrica e as notas subjetivas são calculados para 3 conjuntos de dados (todos os conteúdos, conteúdos de objetos e conteúdos de pessoas).

	Métrica	**Conjunto completo**				**Objetos inanimados**				**Corpos humanos**			
		PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
Baseada em projeção	PSNR	0.520	0.497	0.981	0.741	0.797	0.786	0.735	0.630	0.744	0.739	0.633	0.704
	PSNR-HVS	0.570	0.564	0.943	0.741	0.845	0.841	0.650	0.630	0.797	0.773	0.572	0.667
	PSNR-HVS-M	0.601	0.585	0.918	0.741	0.866	0.851	0.609	0.593	0.822	0.795	0.539	0.667
	SSIM	0.494	0.497	0.998	0.778	0.873	0.838	0.593	0.704	0.847	0.815	0.503	0.630
	MSSIM	0.677	0.682	0.845	0.685	**0.929**	**0.934**	**0.451**	**0.556**	0.814	**0.861**	0.550	0.667
	VIFP	**0.754**	**0.717**	**0.754**	**0.648**	0.906	0.932	0.516	0.593	**0.905**	**0.861**	**0.402**	**0.519**
Baseada em pontos	$\text{po2point}_{\text{MSE}}$	0.672	0.597	0.850	0.667	**0.795**	0.822	**0.738**	**0.630**	0.651	0.702	0.719	0.704
	$\text{po2point}_{\text{Hausdorff}}$	**0.683**	**0.725**	**0.839**	**0.648**	0.793	**0.824**	0.741	**0.630**	0.651	0.707	0.719	0.704
	$\text{po2plane}_{\text{MSE}}$	0.656	0.598	0.866	0.704	0.763	0.755	0.786	0.741	0.637	0.689	0.730	0.741
	$\text{po2plane}_{\text{Hausdorff}}$	**0.683**	0.686	**0.839**	**0.648**	0.792	0.778	0.743	0.667	0.652	0.686	0.718	0.741
	$\text{pl2plane}_{\text{RMS}}$	0.679	0.676	0.843	0.759	0.707	0.702	0.861	0.778	**0.756**	0.653	**0.620**	0.630
	$\text{pl2plane}_{\text{MSE}}$	0.675	0.676	0.847	0.759	0.662	0.753	0.912	0.852	0.701	**0.715**	0.676	**0.593**
	Color - PSNR_{YUV}	0.539	0.491	0.967	0.833	0.669	0.753	0.904	0.852	0.702	**0.715**	0.675	**0.593**

Tabela 4.3: *Benchmarking* das métricas objetivas considerando os dados obtidos na UnB como valores de referência. O mesmo esquema de organização da Tabela 4.2 é seguido.

	Métrica	**Conjunto completo**				**Objetos inanimados**				**Corpos humanos**			
		PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
Baseada em projeção	PSNR	0.582	0.545	0.874	0.667	0.799	0.794	0.683	0.481	0.756	0.747	0.616	0.444
	PSNR-HVS	0.623	0.608	0.840	0.648	0.835	0.850	0.625	0.519	0.805	0.783	0.558	0.407
	PSNR-HVS-M	0.652	0.629	0.814	0.630	0.853	0.862	0.592	0.444	0.830	0.806	0.524	0.444
	SSIM	0.566	0.570	0.886	0.667	0.880	0.893	0.539	0.593	0.865	0.831	0.471	0.370
	MSSIM	0.739	0.738	0.724	0.537	**0.940**	**0.961**	**0.389**	**0.222**	0.859	0.886	0.482	0.370
	VIFP	**0.784**	**0.740**	**0.667**	**0.519**	0.877	0.884	0.545	0.444	**0.919**	**0.890**	**0.370**	**0.296**
Baseada em pontos	$\text{po2point}_{\text{MSE}}$	0.747	0.652	0.714	0.556	0.843	0.792	0.610	**0.481**	0.728	0.758	0.645	0.519
	$\text{po2point}_{\text{Hausdorff}}$	0.757	**0.775**	0.702	0.537	**0.844**	**0.839**	**0.609**	**0.481**	0.728	0.757	0.645	0.519
	$\text{po2plane}_{\text{MSE}}$	0.736	0.670	0.727	**0.500**	0.824	0.798	0.643	0.519	0.713	0.740	0.659	0.556
	$\text{po2plane}_{\text{Hausdorff}}$	**0.758**	0.749	**0.701**	0.537	**0.844**	0.806	0.610	0.481	0.730	0.762	0.643	0.519
	$\text{pl2plane}_{\text{RMS}}$	0.520	0.461	0.918	0.815	0.654	0.596	0.859	0.741	0.685	0.607	0.685	0.593
	$\text{pl2plane}_{\text{MSE}}$	0.666	0.664	0.801	0.704	0.629	0.678	0.882	0.778	0.771	**0.781**	0.599	**0.444**
	Color - PSNR_{YUV}	0.672	0.664	0.795	0.704	0.629	0.678	0.883	0.778	**0.773**	**0.781**	**0.597**	**0.444**

Capítulo 5

Conclusões

Uma métrica objetiva para avaliação de qualidade que inerentemente leva em consideração distorções de cor e de geometria foi proposta. Experimentos de avaliação subjetiva verificaram uma forte correlação entre o *framework* proposto e a qualidade subjetiva percebida.

Outras métricas propostas na literatura, além de não incorporar os aspectos de cor e de geometria simultaneamente, apresentam correlação menor com avaliações subjetivas. Isso evidencia a superioridade do sistema proposto em relação a métodos existentes no campo de processamento de nuvens de pontos.

Também foram feitas observações interessantes acerca do comportamento dos avaliadores durante os testes subjetivos. O principal artefato que prejudicou a percepção da qualidade dos conteúdos analisados para os participantes no experimento foi o surgimento de lacunas entre os *pixels* das imagens projetadas. Também se percebe que participantes tendem a ser mais críticos quando o conteúdo em questão representa pessoas.

Se revela a possibilidade de se explorar a percepção humana sob mais fatores. Especialmente, em trabalhos futuros deseja-se determinar a relação entre qualidade percebida, distorções de cor e distorções geométricas que não introduzam lacunas na superfície dos modelos ou que diminuam a densidade de pontos/*voxels* da nuvem de pontos.

Esforços já estão sendo realizados para a elaboração de experimentos subsequentes com um *framework* de visualização com melhor desempenho, em uma plataforma *web* com aceleração gráfica. Isso permitirá a aquisição de dados com mais participantes, aumentando a significância estatística dos padrões de comportamento observados e o uso de conteúdos mais variados.

Também deseja-se incluir em futuros estudos novas metodologias de compressão no estado da arte, propostas ao longo do desenvolvimento deste trabalho. Além de uma validação subsequente da métrica proposta, poderão se traçar considerações acerca dos desempenhos relativos de cada *codec*, em termos de suas relações de *rate-distortion*.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] ZONE, R. *3-D revolution: The history of modern stereoscopic cinema.* [S.l.]: University Press of Kentucky, 2012.

[2] SCHRÖTER, J. *3D: History, Theory and Aesthetics of the Transplane Image.* [S.l.]: Bloomsbury Publishing USA, 2014.

[3] KLINGER, B. Three-dimensional cinema. *Convergence: The International Journal of Research into New Media Technologies*, SAGE Publications, v. 19, n. 4, p. 423–431, jul 2013. Disponível em: <https://doi.org/10.1177/1354856513494177>.

[4] TIAN, D. et al. Geometric distortion metrics for point cloud compression. In: *Proc. IEEE Intl. Conf. Image Processing.* [S.l.: s.n.], 2017.

[5] ALEXIOU, E.; EBRAHIMI, T. On subjective and objective quality evaluation of point cloud geometry. In: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX).* [S.l.: s.n.], 2017. p. 1–3.

[6] JAVAHERI, A. et al. Subjective and objective quality evaluation of 3D point cloud denoising algorithms. In: *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW).* [S.l.: s.n.], 2017. p. 1–6.

[7] JAVAHERI, A. et al. Subjective and objective quality evaluation of compressed point clouds. In: *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP).* [S.l.: s.n.], 2017. p. 1–6.

[8] GOLDMANN, L.; EBRAHIMI, T. 3d quality is more than just the sum of 2d and depth. In: *Proc. IEEE Intl. Workshop on Hot Topics in 3D.* [S.l.: s.n.], 2010.

[9] ZHANG, J. et al. A subjective quality evaluation for 3D point cloud models. In: *2014 International Conference on Audio, Language and Image Processing.* [S.l.: s.n.], 2014. p. 827–831.

[10] ALEXIOU, E.; EBRAHIMI, T. On the performance of metrics to predict quality in point cloud representations. In: *Proceedings of SPIE.* [S.l.: s.n.], 2017. (Applications of Digital Image Processing XL, v. 10396).

[11] ALEXIOU, E.; UPENIK, E.; EBRAHIMI, T. Towards subjective quality assessment of point cloud imaging in augmented reality. In: *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP).* [S.l.: s.n.], 2017. p. 1–6.

[12] ALEXIOU, E. et al. Point cloud subjective evaluation methodology based on 2D rendering. In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. [S.l.: s.n.], 2018.

[13] GUENNEBAUD, G.; BARTHE, L.; PAULIN, M. Splat/mesh blending, perspective rasterization and transparency for point-based rendering. In: *SPBG*. [S.l.: s.n.], 2006. p. 49–57.

[14] ZWICKER, M. et al. Surface splatting. In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*. ACM Press, 2001. Disponível em: <https://doi.org/10.1145/383259.383300>.

[15] BRITANNICA, E. Sensory reception: human vision: structure and function of the human eye. *Encyclopedia Brittanica*, v. 27, p. 179, 1987.

[16] CARLSON, N. R. *Physiology of Behavior (11th Edition)*. [S.l.]: Pearson, 2012. ISBN 0205239390.

[17] STONE, J. V. Footprints sticking out of the sand. part 2: Children's bayesian priors for shape and lighting direction. *Perception*, SAGE Publications, v. 40, n. 2, p. 175–190, jan 2011. Disponível em: <https://doi.org/10.1068/p6776>.

[18] BAATZ, W. *Photography (Crash Course Series)*. [S.l.]: Barrons Educational Series Inc, 1997. ISBN 0764102435.

[19] MAXWELL, J. C. On the theory of compound colours, and the relations of the colours of the spectrum. *Philosophical Transactions of the Royal Society of London*, The Royal Society, v. 150, n. 0, p. 57–84, jan 1860. Disponível em: <https://doi.org/10.1098/rstl.1860.0005>.

[20] ITU-R BT.601-7. *Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios*. March 2011. International Telecommunications Union.

[21] POYNTON, C. *Digital Video and HD: Algorithms and Interfaces (The Morgan Kaufmann Series in Computer Graphics)*. [S.l.]: Morgan Kaufmann, 2002. ISBN 1558607927.

[22] YOUNG, T. The bakerian lecture: On the theory of light and colours. *Philosophical Transactions of the Royal Society of London*, The Royal Society, v. 92, n. 0, p. 12–48, jan 1802. Disponível em: <https://doi.org/10.1098/rstl.1802.0004>.

[23] HERING, E. Outlines of a theory of the light sense. Harvard University Press, 1964.

[24] ROWE, M. H. Trichromatic color vision in primates. *Physiology*, American Physiological Society, v. 17, n. 3, p. 93–98, jun 2002. Disponível em: <https://doi.org/10.1152/nips.01376.2001>.

[25] WYSZECKI, G.; STILES, W. S. *Color Science: Concepts and Methods, Quantitative Data and Formulae (The Wiley Series in Pure and Applied Optics)*. [S.l.]: Wiley-Interscience, 1982. ISBN 0471021067.

[26] HUNT, R. W. G. *The Reproduction of Colour*. [S.l.]: Wiley, 2004. ISBN 0470024259.

[27] MASLAND, R. H. The fundamental plan of the retina. *Nature Neuroscience*, Springer Nature America, Inc, v. 4, n. 9, p. 877–886, sep 2001. Disponível em: <https://doi.org/10.1038/nn0901-877>.

[28] BLOJ, M.; HEDRICH, M. Color perception. In: *Handbook of Visual Display Technology*. [S.l.]: Springer Berlin Heidelberg, 2015. p. 1–7.

[29] LANGSAM, Y.; AUGENSTEIN, M.; TENENBAUM, A. M. *Data Structures using C and C++*. [S.l.]: Prentice Hall New Jersey, 1996.

[30] MEAGHER, D. Geometric modeling using octree encoding. *Computer graphics and image processing*, Elsevier, v. 19, n. 2, p. 129–147, 1982.

[31] SHAHID, M. et al. No-reference image and video quality assessment: a classification and review of recent approaches. *EURASIP Journal on Image and Video Processing*, Springer Nature, v. 2014, n. 1, aug 2014. Disponível em: <https://doi.org/10.1186/1687-5281-2014-40>.

[32] ITU-R BT.709-6. *Parameter values for the HDTV standards for production and international programme exchange*. June 2015. International Telecommunications Union.

[33] OHM, J.-R. et al. Comparison of the Coding Efficiency of Video Coding Standards-Including High Efficiency Video Coding (HEVC). *IEEE Trans. Cir. and Sys. for Video Technol.*, IEEE Press, Piscataway, NJ, USA, v. 22, n. 12, p. 1669–1684, 12 2012. ISSN 1051-8215. Disponível em: <http://dx.doi.org/10.1109/TCSVT.2012.2221192>.

[34] MUNKRES, J. *Topology (2nd Edition)*. [S.l.]: Pearson, 2000. ISBN 0131816292.

[35] ENCYCLOPAEDIA of Mathematics (Encyclopaedia of Mathematics, 10 Volume Set). [S.l.]: Kluwer, 1994. ISBN 1556080107.

[36] COVER, T. M.; THOMAS, J. A. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. [S.l.]: Wiley-Interscience, 2006. ISBN 0471241954.

[37] SHEIKH, H.; BOVIK, A. Image information and visual quality. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. [S.l.]: IEEE, 2004.

[38] GEISLER, W. S.; PERRY, J. S.; ING, A. D. Natural systems analysis. In: ROGOWITZ, B. E.; PAPPAS, T. N. (Ed.). *Human Vision and Electronic Imaging XIII*. [S.l.]: SPIE, 2008.

[39] VANMARCKE, E. *Random Fields: Analysis And Synthesis (Revised And Expanded New Edition)*. [S.l.]: Wspc, 2010. ISBN 9812563539.

[40] WAINWRIGHT, M. J.; SIMONCELLI, E. P.; WILLSKY, A. S. Random cascades on wavelet trees and their use in analyzing and modeling natural images. In: ALDROUBI, A.; LAINE, A. F.; UNSER, M. A. (Ed.). *Wavelet Applications in Signal and Image Processing VIII*. [S.l.]: SPIE, 2000.

[41] SHEIKH, H. R.; BOVIK, A. C.; VECIANA, G. de. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans Image Process*, v. 14, n. 12, p. 2117–2128, Dec 2005.

[42] STRELA, V.; PORTILLA, J.; SIMONCELLI, E. P. Image denoising using a local gaussian scale mixture model in the wavelet domain. In: ALDROUBI, A.; LAINE, A. F.; UNSER, M. A. (Ed.). *Wavelet Applications in Signal and Image Processing VIII*. [S.l.]: SPIE, 2000.

[43] SHEIKH, H. R.; BOVIK, A. C. A visual information fidelity approach to video quality assessment. In: *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. [S.l.: s.n.], 2005. p. 23–25.

[44] BAILEY, R. A. *Design of Comparative Experiments (Cambridge Series in Statistical and Probabilistic Mathematics)*. [S.l.]: Cambridge University Press, 2008. ISBN 9780521683579.

[45] HINKELMANN, K.; KEMPTHORNE, O. *Design and Analysis of Experiments Set (Wiley Series in Probability and Statistics)*. [S.l.]: Wiley, 2008. ISBN 0470385510.

[46] DENES, J.; KEEDWELL, A. D. *Latin Squares and Their Applications*. [S.l.]: Academic Press Inc, 1974. ISBN 012209350X.

[47] GOLDMANN, L.; SIMONE, F. D.; EBRAHIMI, T. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In: *Proc. SPIE 7526, Three-Dimensional Image Processing (3DIP) and Applications*. [S.l.: s.n.], 2010. v. 7526, p. 75260S.

[48] MEKURIA, R.; BLOM, K.; CESAR, P. Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 27, n. 4, p. 828–842, April 2017.

[49] KAZHDAN, M.; HOPPE, H. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, ACM, v. 32, n. 3, p. 29, 2013.

[50] GONZALEZ. Measurement of areas on a sphere using fibonacci and latitude-longitude lattices,. *Mathematical Geosciences*, v. 42, n. 1, p. 49–64, November 2009.

[51] CONWAY, J.; SLOANE, N. J. A. A note on optimal unimodular lattices. *Journal of Number Theory*, Elsevier, v. 72, n. 2, p. 357–362, 1998.

[52] GREGORY, M. J. et al. A comparison of intercell metrics on discrete global grid systems. *Computers, Environment and Urban Systems*, Elsevier, v. 32, n. 3, p. 188–203, 2008.

[53] WILLIAMSON, D. L. The evolution of dynamical cores for global atmospheric models. *Journal of the Meteorological Society of Japan. Ser. II*, Meteorological Society of Japan, v. 85, p. 241–269, 2007.

[54] SAFF, E. B.; KUIJLAARS, A. B. Distributing many points on a sphere. *The mathematical intelligencer*, Springer, v. 19, n. 1, p. 5–11, 1997.

[55] ITU-T. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. *Recommendation ITU-T P.1401*, 2012.

[56] QUEIROZ, P. A. C. R. L. de. Motion-compensated compression of dynamic voxelized point clouds. *IEEE Trans. on Image Processing vol. 26, no. 8, pp. 3886–3895*, August 2017.

[57] TIAN, D. et al. *Evaluation Metrics for Point Cloud Compression*. Geneva, Switzerland: [s.n.], January 2017. ISO/IEC JTC1/SC29/WG11 input document MPEG2016/M39966.

[58] TIAN, D. et al. *Updates and Integration of Evaluation Metric Software for PCC*. Hobart, Australia: [s.n.], April 2017. ISO/IEC JTC1/SC29/WG11 input document MPEG2017/M40522.

[59] ALEXIOU, E.; EBRAHIMI, T. Point cloud quality assessment metric based on angular similarity. In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*. [S.l.: s.n.], 2018. p. 1–6.

[60] HOPPE, H. et al. Surface Reconstruction from Unorganized Points. In: *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 1992. (SIGGRAPH '92), p. 71–78. ISBN 0-89791-479-1. Disponível em: <http://doi.acm.org/10.1145/133994.134011>.

[61] POINT Cloud Library (PCL). Http://pointclouds.org/.

[62] PEARSON, K. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, JSTOR, v. 58, p. 240–242, 1895.

[63] SPEARMAN, C. The proof and measurement of association between two things. *The American journal of psychology*, JSTOR, v. 15, n. 1, p. 72–101, 1904.