UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Programa de Pós-Graduação em Computação

Tese

Avaliação Objetiva de Qualidade de Imagem e Vídeo 3D Utilizando Técnicas de Aprendizado de Máquina

Narúsci dos Santos Bastos

Pelotas, 2023

Narúsci dos Santos Bastos

Avaliação Objetiva de Qualidade de Imagem e Vídeo 3D Utilizando Técnicas de Aprendizado de Máquina

Tese apresentada ao Programa de Pós-Graduação em Computação do Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutor em Ciência da Computação.

Orientador: Prof. Dr. Bruno Zatt
Coorientadores: Prof. Dr. Guilherme Ribeiro Corrêa
Profª. Drª. Tatiana Aires Tavares

Pelotas, 2023

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

B327a Bastos, Narúsci dos Santos

Avaliação objetiva de qualidade de imagem e vídeo 3D utilizando técnicas de aprendizado de máquina / Narúsci dos Santos Bastos ; Bruno Zatt, orientador ; Guilherme Ribeiro Corrêa, Tatiana Aires Tavares, coorientadores. — Pelotas, 2023.

123 f. : il.

Tese (Doutorado) — Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2023.

1. Avaliação de qualidade de vídeo. 2. Avaliação de qualidade de imagem. 3. Vídeo estereoscópico. 4. Aprendizado de máquina. I. Zatt, Bruno, orient. II. Corrêa, Guilherme Ribeiro, coorient. III. Tavares, Tatiana Aires, coorient. IV. Título.

CDD : 005

Elaborada por Maria Inez Figueiredo Figas Machado CRB: 10/1612

Narúsci dos Santos Bastos

Avaliação Objetiva de Qualidade de Imagem e Vídeo 3D Utilizando Técnicas de Aprendizado de Máquina

Tese aprovada, como requisito parcial, para obtenção do grau de Doutor em Ciência da Computação, Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

Data da Defesa: 27 de fevereiro de 2023

Banca Examinadora:

Prof. Dr. Bruno Zatt (orientador)
Doutor em Microeletrônica pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Bruno Boessio Vizzotto
Doutor em Ciência da Computação pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Fábio Luís Livi Ramos
Doutor em Ciência da Computação pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Marcelo Schiavon Porto
Doutor em Ciência da Computação pela Universidade Federal do Rio Grande do Sul.

Dedico com todo meu amor ao meu pai Homero e minha mãe Fátima. . . .

AGRADECIMENTOS

Ao início dessa estrada, ainda que sabendo das dificuldades, nunca pensaria que todos nós iríamos atravessar tempos tão sombrios, em que, de repente tudo parou e o medo se instalou. No entanto, apesar desse tempo de isolamento junto a esse caminho que já é solitário e tem suas dores particulares. Não posso deixar de destacar a importância das relações que foram ainda mais fortificadas, intensificadas e valorizadas. Por isso, começo agradecendo meus pais Homero e Fátima pelo amor incondicional e por serem minha principal motivação, força e apoio durante esses longos dias, que nem sempre foram bons. Ao meu marido Igor, que trilhou esta jornada comigo e me ergueu por muitas vezes, também por muitos momentos de compreensão e incentivo. Além disso, agradeço também pelo apoio como mentor, professor e agora doutor, me auxiliando muitas vezes com suas habilidades na área. Dos meus pais e meu marido só tenho gratidão por nunca terem deixado de acreditar em mim e também me lembrar constantemente de acreditar. Aos meus sogros, Sérgio e Gilda que me acolheram em seus corações e também participaram desta rede de apoio, sem a qual teria sido tudo mais difícil, ainda mais em tempos de pandemia, se tornaram ainda mais parceiros e amigos,agradeço também a minha querida cunhada Milene pela amizade e parceria. Toda gratidão a minha amiga Gabriela Gidi, que mesmo de tão longe sempre esteve comigo, em cada minuto, com palavras de apoio, incentivo e até mesmo para momentos de descontração, que no meio desta bagunça me deu o presente de amar uma nova pessoinha, a sua filha Beatriz, que sempre mostra belos sorrisos. Agradeço meu irmão Sáiron que foi sempre foi meu maior exemplo de foco e determinação e me ensinou questões fundamentais para chegar até aqui. Agradeço também a minha cunhada e irmã Mírceia pela amizade e troca, e por ser exemplo de garra e força na busca do que se quer. Além disso, por trazer ao mundo o maior amor da minha vida, minha sobrinha Antônia, que sempre alegrou meu coração em dias cinzas mesmo em pensamento, além de mover minha busca para ser uma tia/dinda a qual ela sempre terá orgulho. Agradeço aos meus avós Saur e Loraci que mesmo estando longe, em outro estado, simplesmente pelo seu amor e por serem o começo de tudo. Agradeço a minha amiga e colega Roberta Palau que viveu e compartilhou comigo as angústias dessa jornada, sempre com bons conselhos e com muita positividade. Agradeço aos meus Pets, Mamutreto, Sury e Tobias que por muitas vezes acalentaram meu coração e me fizeram sorrir em muitos momentos. Agradeço a Profª.Drª. Diana Adamatti, por ter me ajudado no início de tudo a escolher este caminho e por ser um grande exemplo de pessoa e professora. Agradeço ao Luís Mendes pela confiança e amizade durante o período de trabalho e também a equipe de T.I, que foi uma válvula de escape durante estes anos. Agradeço a minha amiga Andressa Silveira, pelo apoio e pelas longas conversas. Agradeço, ao meu Orientador Prof.Dr. Bruno Zatt por todo

apoio e ajuda durante esses anos, e por ser exemplo de gentilidade. Agradeço aos colegas do Vitech que estiveram comigo durante os primeiros anos, no qual a convivência foi interrompida pela pandemia. Agradeço ao Lucas Ikenoue pelo trabalho e ajuda técnica. Agradeço acima de tudo a Deus.

Pequenos paraísos e riscos a correr...
— HUMBERTO GESSINGER

RESUMO

BASTOS, Narúsci dos Santos. **Avaliação Objetiva de Qualidade de Imagem e Vídeo 3D Utilizando Técnicas de Aprendizado de Máquina**. Orientador: Bruno Zatt. 2023. 123 f. Tese (Doutorado em Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2023.

Décadas de pesquisa em Avaliação de Qualidade de Imagem e Vídeo promoveram a criação de uma variedade de métricas objetivas de qualidade que se correlacionam fortemente com a qualidade subjetiva de imagem e vídeo. No entanto, permanecem desafios ao considerar a Avaliação de Qualidade de Imagem e Vídeo 3D/estéreo. Múltiplas métricas objetivas de qualidade para imagens e vídeos 3D foram projetadas estendendo as conhecidas métricas 2D. Como resultado, essas soluções tendem a apresentar pontos fracos em artefatos específicos de 3D. Trabalhos recentes demonstram a eficácia das técnicas de Aprendizado de Máquina (AM) no desenvolvimento de métricas de qualidade 3D. Embora eficazes, algumas soluções baseadas em aprendizado de máquina podem levar a um alto esforço computacional e restringir sua adoção em sistemas de poder computacional limitado e/ou aplicações de baixa latência. Diante deste contexto, surgem as questões de pesquisa que esta tese busca responder. Apresentamos um estudo sobre a Avaliação de Qualidade Objetiva de Imagem e Vídeo 3D de Referência Completa considerando Aprendizado de Máquina leve. Neste estudo discretizamos o escore de qualidade visando adotar soluções baseadas em classificadores. Avaliamos quatro diferentes algoritmos baseados em Árvore de Decisão (AD) considerando diferentes conjuntos de características de imagem/vídeo. Além disso, também foram avaliados cenários adotando diferentes números de classes, tanto para Avaliação de Qualidade de Imagem quanto para Vídeo. Os classificadores foram treinados usando dados do *Waterloo IVC 3D Image Quality Database* e *Waterloo IVC 3D Video Quality Database* para determinar o escore de qualidade subjetivo medido usando o *Mean Opinion Score* (MOS). Os resultados mostram que, de maneira geral, o *RandomForest* obtém a melhor precisão. Nosso estudo demonstra a viabilidade de soluções baseadas em Árvore de Decisão como uma abordagem efetiva e leve para Avaliação de Qualidade de Imagem e Vídeo 3D.

Palavras-chave: Avaliação de qualidade de vídeo. Avaliação de qualidade de imagem. Vídeo estereoscópico. Aprendizado de Máquina.

ABSTRACT

BASTOS, Narúsci dos Santos. **Objective Assessment of Image Quality and Full Reference 3D Video Using Decision Tree Based Machine Learning Techniques**. Advisor: Bruno Zatt. 2023. 123 f. Thesis (Doctorate in Computer Science) – Technology Development Center, Federal University of Pelotas, Pelotas, 2023.

Decades of research in Image and Video Quality Assessment have led to the creation of a variety of objective quality metrics that strongly correlate with subjective image and video quality. However, challenges remain when considering quality assessment of 3D/stereo images and videos. Multiple objective quality metrics for 3D images and videos were designed by extending the 2D metrics. As a result, these solutions tend to present limitations associated with 3D-specific artifacts. Recent work has demonstrated the effectiveness of machine learning techniques in developing 3D quality metrics. While effective, some machine learning solutions demand high computational effort restricting their adoption for low latency applications and/or embedded systems. In this context, a set of important research questions arise justifying this Thesis. We present a study on the evaluation of full-reference stereoscopic objective quality considering lightweight machine learning. In this study we discretized the quality score in order to treat the quality assessment as a classification problem. We evaluated four different algorithms based on decision trees considering eight different sets of image/video characteristics. In addition, we performed tests considering different numbers of classes for both image and video quality evaluation. The classifiers were trained using data from the Waterloo IVC 3D Image Quality Database and Waterloo IVC 3D Video Quality Database to determine the subjective quality score measured using the *Mean Opinion Score* (MOS). The results show that, in general, *RandomForest* presents the best accuracy. Our study demonstrates the feasibility of decision tree solutions as an effective and lightweight approach for assessing 3D image and video quality.

Keywords: Video quality assessment. Image quality assessment. Stereoscopic video. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACR	Absolute Categorical Rating
ACR-HR	Absolute Category Rating with Hidden Reference
AD	Árvore de Decisão
AM	Aprendizado de Máquina
CRT	Cathodic Ray Tube
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
DMOS	Differences in Mean Opinion Scores
DSCQS	Double Stimulus Continuous Quality scale
DSI	Disparity Spatial Indices
DSIS	Double Stimulus Impairment Scale
DTI	Disparity Temporal Indices
DTP-VQI	Disparity Temporal Perceptual - Visual Quality Image
DTPW-SSIM	Disparity Temporal Perceptual Weight - Structural Similarity Index Measure
FR	Full-Reference
FSC	Função Sensível ao contraste
GOP	Group of Pictures
HVS	Human Visual System
IA	Inteligência Artificial
IQA	Image Quality Assessment
ITU	International Telecommunication Union
LCV	Laboratory for Computional Vision
LIVE	Laboratory for image & Video Engineering
LST-SSIM	Local Spatial - Temporal - Structural Similarity Index Measure
MAE	Mean Absolut Error

MC-SSIM	Motion Compensated - Structural Similarity Index Measure
MOS	Mean Opinion Score
MOVI	Motion-based Video Integrity Evaluation
MS-SSIM	Multiscale - Structural Similarity Index Measure
MSE	Mean Square Error
NR	No-Reference
PC	Pair Comparison
PSNR	Peak Signal-Noise Ratio
PVS	Processed Video Sequences
PW-SSIM	Perceptual Weighting - Structural Similarity Index Measure
REPT	Reduced Error Pruning Tree
RR	Reduced-Reference
SAD	Sum of Absolute Differences
SAMVIQ	Subjective Assessment Method for Video Quality
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation
SI	Temporal Information
SSCQE	Single Stimulus Continuous Quality Evaluation
SSIM	Structural Similarity Index Measure
StSD	Stereoscopic Structural Distortion
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVR	Support Vector Regression
SW-SSIM	Speed Weighted - Structural Similarity Index Measure
TI	Temporal Information
TP	Presentation Test
VIF	Visual Information Fidelity
VMAF	Video Multi-Method Assessment Fusion
VQA	Video Quality Assessment
VQEG	Video Quality Experts Group
VQM	Visual Quality Metric
VTR	Video Tape Recorder

SUMÁRIO

1 INTRODUÇÃO

Ao longo das últimas décadas observamos um crescimento expressivo no compartilhamento de imagens e vídeos digitais viabilizado pela onipresença da internet e pela proliferação de mídias sociais e serviços de *streaming*. Nesse contexto, as tecnologias de codificação de imagens e vídeos passaram a ser de uso mandatório. Como o processo de codificação e transmissão pode adicionar degradações à qualidade percebida pelo usuário final, tem se tornado cada vez mais importante desenvolver métodos e técnicas que permitam medir a qualidade percebida pelo usuário. As Avaliações de Qualidade de Imagem (*Image Quality Assessment* - IQA) e Vídeo (*Video Quality Assessment* - VQA) utilizam métricas e métodos que têm a finalidade de mensurar a qualidade das imagens e vídeos que sofreram degradações e permitir melhorias na qualidade final da imagem ou vídeo que chega ao usuário (ZEGARRA RODRÍGUEZ, 2014; REGIS, 2013). No entanto, o estudo descrito em Banitalebi-dehkordi; Pourazad; Nasiopoulos (2016) observa que a maioria destas métricas são desenvolvidas para conteúdos 2D, deixando um amplo espaço para melhorias sobre a avaliação, precisão e confiabilidade na Avaliação de Qualidade de Imagens e Vídeos 3D (BANITALEBI-DEHKORDI; POURAZAD; NASIOPOULOS, 2016).

A VQA é disposta em duas classes: subjetiva e objetiva. A avaliação subjetiva é baseada na percepção humana, dada através do julgamento do observador. A qualidade do vídeo percebido depende de detalhes como a distância da visualização, tamanho da tela, resolução, brilho, contraste, nitidez, cor e outros fatores. Os Métodos de Avaliação Subjetiva são descritos pela *International Telecommunication Union* (ITU) em ITU-R BT.500 (2002) para serviços de TV e a ITU-T P.910 (2008) para aplicações multimídia. Os Métodos de Avaliação Subjetiva, geralmente, são classificadas pelo tipo de estímulo, ou seja, o número de exibições de estímulos que serão apresentadas aos observadores. Estes métodos podem ser de estímulo único, em que é exibido somente o vídeo que está sendo analisado pelo observador, sem a presença de um vídeo de referência; e estímulo duplo, no qual são apresentados dois vídeos, o vídeo que está sendo avaliado pelo observador e o vídeo de referência. Além destes, também existem métodos de múltiplos estímulos, em que os vídeos podem ser visualizados

diversas vezes pelo observador. Os métodos classificados como estímulo único são o SSCQE (*Single Stimulus Continuos Quality Evaluation*), neste método é apresentado uma sequência de vídeo, em que o observador fornece pontuações continuamente durante a execução do vídeo, em uma escala de 1 a 100; o ACR (*Absolute Categorical Rating*), é um método de categoria absoluta, ou estímulo único, que consiste em apresentar a sequência de vídeo para observador que pontuará independentemente cada cena em uma escada de categorias. Sendo este considerado o método mais rápido dentre os descritos pela ITU-T P.910 (2008) (ZEGARRA RODRÍGUEZ, 2014). Já os métodos de estímulo duplo, podem ser o DSIS (*Double Stimulus Impairment Scale*), onde as sequências de vídeo podem ser apresentadas duas vezes, utilizado para avaliar degradações claramente visíveis, como artefatos causados por erros de transmissão. O sinal de vídeo original é apresentado por poucos segundos, logo o sinal do vídeo degradado é apresentado pelo mesmo tempo, em que o observador avalia em uma escala de 5 níveis. O método SDSCE (*Simultaneous Double Stimulus for Continuous Evaluation*), apresenta as sequências original e degradada de maneira simultânea, em que é informado ao observador qual é o vídeo de referência, este método é avaliado em uma escala de 1 a 100. De maneira geral esses métodos subjetivos utilizam a métrica de avaliação *Mean Opinion Score* (MOS), na qual, os observadores classificam a qualidade do vídeo numa escala de pontuações diferentes, como por exemplo: 5 - melhor qualidade e 1 - pior qualidade (ZEGARRA RODRÍGUEZ, 2014).

Os modelos objetivos utilizam equações e algoritmos para estimar a qualidade do conteúdo visual, com base em modelos estatísticos ou matemáticos, cujo objetivo é obter uma estimativa de qualidade mais próxima possível da Percepção Visual Humana. A VQA dispõe de diversas métricas, sendo algumas baseadas em parâmetros de estatísticos/matemáticos como o MSE (*Mean Square Error*), que compara dois sinais dando uma nota quantitativa que descreve o grau de similaridade (ou dissimilaridade) entre as imagens ou quadros dos vídeos. Já o PSNR (*Peak Signal-Noise Ratio*) define a relação entre a energia de um sinal e o ruído que afeta a representação do mesmo, entre a imagem original e degradada. Ainda que muito utilizada, a métrica PSNR não leva em consideração características do HVS (*Human Visual System*) e as condições de visualização, tendo assim, uma correlação “limitada” com as medidas subjetivas (TANJI et al., 2014). No entanto, algumas métricas que consideram o HVS, apresentam boa correlação com os valores de medidas subjetivas. Como a métrica SSIM (*Structural Similarity Index*), que avalia as diferenças estruturais entre cada imagem (referência e degradada) ou quadro dos vídeos.

As métricas objetivas podem ser classificadas de acordo com a disponibilidade da imagem de referência, podendo ser de Referência Completa (*Full-Reference* - FR), Sem Referência (*No-Reference* - NR) ou Referência Reduzida (*Reduced-Reference* - RR). Apesar de ser considerado mais apropriado para avaliar a qualidade de imagem

e vídeo, os modelos subjetivos exigem a disponibilidade de recursos humanos, o que torna a sua realização custosa, lenta e, muitas vezes, inviável (WANG et al., 2003; WANG; LU; BOVIK, 2004). Porém, estes modelos são essenciais, já que funcionam como referência para validar a capacidade dos modelos objetivos. Os valores das predições objetivas devem ser comparados em termos de proximidade com os valores dos modelos subjetivos, normalmente com o valor de MOS, sendo que uma maior correlação entre esses valores indica que a métrica objetiva é confiável para predizer a qualidade da imagem (WANG et al., 2015).

A Avaliação Objetiva de Qualidade de Vídeo é uma tarefa difícil, de acordo com a percepção humana, devido à natureza multidisciplinar complexa do problema (relacionado à fisiologia, psicologia, visão e ciência da computação) e o entendimento limitado do mecanismo do HVS (NARWARIA; LIN, 2011). As 3D-VQAs são consideradas mais complexas se comparadas com as 2D-VQAs, uma vez que no caso de vídeos 2D, alguns dos fatores que afetam a qualidade percebida, como brilho, contraste e nitidez são mais bem compreendidos. Já nos vídeos 3D, a percepção de profundidade altera o impacto que esses fatores têm sobre a qualidade geral da imagem (BANITALEBI-DEHKORDI; POURAZAD; NASIOPOULOS, 2016). Além disso, como os vídeos 3D produzem a sensação de estereoscopia decorrente da disparidade das imagens vistas pelos dois olhos (vista esquerda e direita), aumenta a possibilidade de criar diferenças no nível/tipo de degradação entre as vistas, chamada de distorção assimétrica (FANG; SUI; WANG, 2019; WANG et al., 2017).

De acordo com Fang; Sui; Wang (2019), as métricas voltadas para qualidade de imagem e vídeo estereoscópicos tendem a seguir duas abordagens. A primeira diz respeito a métodos desenvolvidos para construir modelos especificamente para 3D, envolvendo informações exclusivas da estereoscopia. Já a segunda abordagem refere-se a métodos desenvolvidos com base em métricas 2D, já consolidados na literatura, e que analisam a qualidade do par estereoscópico separadamente e agrupam os valores obtidos. Os autores salientam que quando as distorções entre as vistas são simétricas este método mostra-se suficiente para prever a qualidade 3D. O mesmo não ocorre para distorções assimétricas entre vistas.

Na literatura são encontrados diferentes modelos de 3D-VQA que buscam suprir de maneiras diferentes a necessidade de incorporar as características da percepção visual humana em um modelo de avaliação de qualidade objetiva, especialmente, quando se trata de imagens ou vídeos estereoscópicos. Trabalhos como o de Fang; Sui; Wang (2019) focam na assimetria entre as vistas com base na rivalidade binocular, que ocorre quando os estímulos apresentados aos dois olhos são diferentes e uma das vistas recebe maior estímulo e tende a ter dominância sobre a outra vista (FANG et al., 2020). Outros autores implementam modelos que consideram informações complexas das imagens e vídeos como a Informação Temporal (*Temporal Information* - TI)

e Informação Espacial (*Spatial Information* - SI) (SILVA, 2013).

Uma das principais barreiras encontradas na 3D-VQA é a busca por modelos de qualidade percebida que sejam fiéis ao modelo da percepção visual humana (GALKANDAGE et al., 2017). Encontrar um modelo ideal e um conjunto de características que apresentem boa correlação com os modelos subjetivos pode resultar em uma tarefa exaustiva. Contudo, alternativas para automatizar este processo podem ser úteis para tornar esta busca mais prática. Muitas abordagens adotam modelagem do HVS para imitar a percepção de profundidade e, em particular, a disparidade binocular induzida pelo deslocamento horizontal de recursos de imagem entre as vistas esquerda e direita.

De acordo com Gastaldo; Redi (2012), os paradigmas de Aprendizado de Máquina (AM) permitem lidar com a tarefa de VQA sobre uma perspectiva diferente das métricas tradicionais, pois têm como objetivo final "imitar" a percepção humana de qualidade ao invés de projetar um modelo explícito do HVS. Alguns trabalhos utilizam modelos de AM na definição de métricas para avaliar a qualidade de imagem como os trabalhos de Narwaria; Lin (2011) que, utiliza *Support Vector Regression* (SVR) e de Charrier; Lézoray; Lebrun (2012) que, adota *Support Vector Machine* (SVM). Além disso, em Gastaldo; Redi (2012) os autores salientam que outras técnicas de AM como Redes Neurais e *Kernel Machine* também são utilizadas para a Avaliação de Qualidade de Imagem e Vídeo, Porém essas técnicas são computacionalmente custosas.

Neste contexto, hipotetizamos que algoritmos de Aprendizado de Máquina de baixa complexidade podem ser utilizados em técnicas de Avaliação de Qualidade de Imagens e Vídeos 3D. Assim, este trabalho busca avaliar soluções utilizando as principais técnicas de classificação baseadas em Árvores de Decisão (ADs), por apresentarem baixa complexidade e serem um meio intuitivo para interpretar os resultados obtidos (GARCIA, 2003).

Além disso, buscamos analisar os modelos preditivos, que apresentam a capacidade que o algoritmo tem de predizer a classe, que neste caso é o MOS. Sendo assim, quanto maior a capacidade de predição de um algoritmo, melhor sua capacidade de inferir o MOS da imagem ou vídeo, ou seja, a qualidade percebida pelo usuário.

Esse trabalho também apresenta a influência das características extraídas das imagens 3D na eficácia dos modelos gerados por estas técnicas. Buscamos, portanto, responder às seguinte questões:

- **Questão 1 (Q1)**: É viável aplicar técnicas baseadas em AD para avaliar a qualidade de imagem 3D?
- **Questão 2 (Q2)**: É viável aplicar técnicas baseadas em AD para avaliar a qualidade de vídeo 3D?

- **Questão 3 (Q3)**: Quais características das imagens são mais relevantes na definição de modelos para predição de 3D-IQA?
- **Questão 4 (Q4)**: Qual algoritmo baseado em AD se mostra mais promissor para avaliar a qualidade de imagem e vídeo 3D?
- **Questão 5 (Q5)**: É viável treinar modelos utilizando imagens para avaliar qualidade em vídeos?
- **Questão 6 (Q6)**: O aumento no número de classes tem influência na capacidade de predição dos algoritmos baseados em AD?

Esta tese está organizada em sete Capítulos, o primeiro Capítulo apresenta a Introdução, que aborda o trabalho de forma geral, o Capítulo 2 diz respeito a Fundamentação Teórica, em que são abordados os principais temas que envolvem esta pesquisa, como o Sistema Visual Humano e suas características, degradações de vídeos e imagens, os conceitos de Aprendizado de Máquina, bem como o conceito de AD e os algoritmos utilizados neste trabalho. Ainda neste capítulo são apresentadas as métricas que foram utilizadas para avaliação dos algoritmos de AM e as Medidas de Erro usadas para avaliar estas técnicas. O Capítulo 3 refere-se a Avaliação de Qualidade de Imagem e Vídeo. Em que, discorre sobre a Avaliação Objetiva de Qualidade de Imagem e Vídeo, abordando algumas das principais Métricas Objetivas de Avaliação de Qualidade 2D e 3D. Além disso, aborda a Avaliação Subjetiva de Qualidade de Imagem e Vídeo apresentando suas definições e métodos. O Capítulo 4 discute os Trabalhos Relacionados. No Capítulo 5 são apresentadas as etapas necessárias para o desenvolvimento da metodologia deste trabalho, como a Aquisição dos Dados, o Pré-processamento, Treinamento e validação, e o Teste. O Capítulo 6 apresenta a discussão dos resultados obtidos para a Avaliação de Qualidade de Imagem 3D e Avaliação de Qualidade de vídeo 3D. Por fim, o Capítulo 7 apresenta a Conclusão e os Trabalhos Futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este Capítulo discorre sobre os principais temas necessários para o desenvolvimento e compreensão da Tese. Inicialmente, a Seção 2.1 trata do Sistema Visual Humano e suas questões. Já a Seção 2.3 aborda sobre alguns tipos de degradações possíveis quando um vídeo ou imagem passam pelo processo de compressão. A Seção 2.4 apresenta os principais conceitos de Aprendizado de Máquina. A Seção 2.5 apresenta o conceito de Árvore de Decisão e também os algoritmos de AD que foram utilizados neste trabalho. A Seção 2.6 discorre sobre as Métricas de Avaliação de Desempenho de AM e a Seção 2.7 trata sobre medidas de erro para técnicas de AM. Por fim, a Seção 2.8 apresenta as Considerações finais do Capítulo.

2.1 Sistema Visual Humano

A compreensão do sistema visual humano, a percepção visual, se faz muito importante para o desenvolvimento de métricas de Avaliação de Qualidade de Imagem e Vídeo, já que uma métrica objetiva é dita eficiente quando tem um valor de correlação mais próximo aos das medidas subjetivas, que são baseados na percepção visual humana.

2.1.1 Construção Física

Para Silverthorn (2010), o olho humano é um receptor sensorial que funciona como uma câmera fotográfica, ele foca a luz sobre a superfície sensível à luz (retina) usando uma lente (cristalino) e uma abertura (pupila), em que o tamanho pode ser ajustado de acordo com a quantidade de luz que entra. A retina tem a capacidade de traduzir a luz para o sistema nervoso, extraindo do meio ambiente o que é útil e ignorando o que é redundante (HUBEL, 1988). A visão pode ser definida como o processo em que a luz refletida nos objetos no meio externo é transformada em imagem mental, esse processo pode ser dividido em três fases:

1. A luz entra no olho e é focalizada na retina pela lente (cristalino);
2. Os fotorreceptores presentes na retina transduzem a energia luminosa em um

sinal elétrico;

3. As vias neurais da retina até o cérebro processam os sinais elétricos em imagens visuais.

A luz entra na superfície anterior do olho através da córnea, que é um disco de tecido transparente especializado que permite a penetração de raios de luz nos olhos (SILVERTHORN, 2010)(PURVES et al., 2001). Logo após a luz atravessar a abertura da pupila, ela atinge a lente, que possui duas superfícies convexas. A córnea e a lente juntas alteram a direção dos raios de luz que entram para que eles possam ser focados na retina, que é o revestimento do olho sensível à luz que contém fotorreceptores (SILVERTHORN, 2010).

Ainda que, para focar uma câmera em uma imagem é necessário mudar a distância entre a lente e o filme, nos seres humanos para focalizar o olho, a distância entre a lente e a retina não muda. Mas a forma da lente é alterada, puxando ou relaxando os tendões que a mantêm na margem. Com isso, o cristalino torna-se mais esférico para objetos próximos e achatado para objetos mais distantes (HUBEL, 1988). Esta mudança de forma é produzida por um conjunto de músculos ciliares, que controlam a posição do globo ocular, conforme é apresentado na Figura 1.

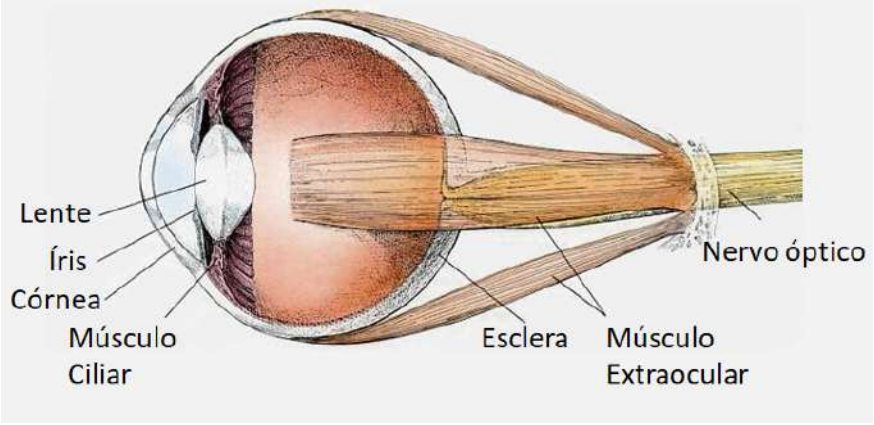


Figura 1 – O globo ocular e os músculos que controlam sua posição. A córnea e a lente focam os raios de luz na parte de trás do olho. A lente regula a focagem para objetos próximos e distantes tornando-se mais ou menos globulares (HUBEL, 1988).

Purves et al. (2001) reitera que o olho é uma esfera preenchida com fluído, fechada por três camadas de tecido, em que, somente a camada mais interna do olho, a retina, contém neurônios que são sensíveis à luz e são capazes de transmitir sinais visuais para destinos centrais. Os nervos ópticos vão dos olhos para o quiasma óptico no encéfalo, onde algumas fibras nervosas cruzam para o lado oposto. Após fazer a sinapse no corpo geniculado lateral do tálamo, os neurônios da visão finalizam seu trajeto no córtex visual do lobo occipital (SILVERTHORN, 2010).

Os neurônios no córtex visual são conhecidos por serem sintonizados em vários aspectos das correntes entrantes, como frequências espaciais e temporais, orienta-

ções e direções de movimento. Geralmente, somente a seletividade de frequência e orientação espacial são modelados por métricas de avaliação de qualidade. As correntes visuais geradas no córtex são transportadas para outras partes do cérebro para o processamento seguinte, como a detecção de movimento e cognição (WANG et al., 2003).

2.1.2 Receptores Visuais

O HVS é a principal base para interpretação da representação digital de um vídeo colorido. A visão de padrões é obtida pela distribuição de receptores sensíveis a luz ao longo da superfície da retina que contém receptores de luz. Esses receptores são chamados de bastonetes e cones, a retina possui quatro tipos de receptores, que contém um pigmento diferente, sendo um bastonete e três tipos de cones (HUBEL, 1988).

Cada olho possui cerca de 6 a 7 milhões de cones, que são localizados principalmente na porção central da retina, a fóvea, e são extremamente sensíveis às cores. Os humanos são capazes de distinguir pequenos detalhes de alta resolução com esses cones, em grande parte porque cada um deles está conectado à sua própria terminação nervosa. Porém, os cones não respondem a luz fraca, mas são responsáveis pela capacidade de ver pequenos detalhes, em alta luminosidade, e pela visão de cores (AGOSTINI, 2007; HUBEL, 1988).

Sendo assim, de acordo com Hubel (1988) para ter uma visão colorida, são necessários apenas três tipos de cones sensíveis às cores vermelho, verde e azul. Pois a cor é consequência da estimulação desigual desses cones, se por exemplo os três cones forem estimulados igualmente a sensação resultará na falta de cor ou na cor “branca”.

Já o número de bastonetes é muito maior, cerca de 75 a 150 milhões são distribuídos pela superfície da retina. Eles são utilizados para dar uma visão geral do campo de visão, captando imagens de baixa resolução. Devido ao fato de que muitos bastonetes são conectados a uma única terminação nervosa, o quantidade de detalhes discerníveis por esses receptores são reduzidas (GONZALEZ; WOODS, 2000). Além disso, os bastonetes não estão envolvidos na visualização de cores e são sensíveis a baixos níveis de iluminação, são utilizados para visão noturna.

Os bastonetes e cones traduzem a luz que entra em sinais elétricos e é o cérebro que interpreta esses sinais como cores. O HVS é capaz de distinguir milhares de cores diferentes a partir de combinações de intensidades diferentes das cores captadas pelos cones. Por outro lado, o sistema visual humano consegue distinguir mais do que cinquenta dúzias de tons de cinza, que indicam a intensidade luminosa da imagem (luminância) (GONZALEZ; WOODS, 2000).

2.1.3 Percepção de Imagens

Em geral, na base de qualquer teoria de cor existem os fenômenos associados a maneira como as cores são percebidas e diferenciadas. A luminosidade é a intensidade da luz refletida pela superfície dos objetos, enquanto que, o brilho é a quantidade de luz emitida pelas superfícies dos objetos luminosos. O brilho é um descritor subjetivo da percepção de luz, que é praticamente impossível mensurar. Ele incorpora a noção cromática de intensidade e é um dos principais fatores de descrição da sensação de cores (GONZALEZ; WOODS, 2000).

Os mecanismos e os conceitos associados aos sinais de imagens e vídeos são baseados no processo de percepção de imagem pelo ser humano. O sistema visual recebe estímulos luminosos e transfere as informações para o cérebro, que processa essas informações criando a percepção de imagens (ARTHUR, 2002). Algumas das características psicofísicas do HVS que estão relacionadas com a qualidade visual percebida são:

- **Visão central e periférica:** é o efeito causado quando um observador fixa o olhar em um ponto no seu ambiente visual. A região em volta desse ponto é apresentada com uma maior resolução espacial em relação aos outros pontos da cena. A visão periférica é a capacidade da visão de perceber o que está fora do foco principal da visão. Essas duas características possibilitam ao HVS selecionar pontos relevantes ou não na imagem.

- **Adaptação à luz e contraste simultâneo:** a adaptação à luz refere-se a capacidade do HVS de diferenciar um grande número de intensidades luminosas. O HVS opera controlando a quantidade de luz que entra no olho através da pupila, assim como mecanismos de adaptação nas células da retina que ajustam o ganho de neurônios pós-receptores na retina (WANG et al., 2003). Isto resulta na codificação do contraste do estimulo visual em vez da codificação da intensidade da luz absoluta.

  O contraste simultâneo relaciona-se com o fato de que a luminância de uma região percebida não depende somente da intensidade dos pixels, ou seja, o olho codifica o contraste relativo ao estímulo visual, e não a intensidade absoluta da luz (ARTHUR, 2002). Os neurônios respondem somente a estímulos acima de um dado valor de contraste. Por sua vez, esse quando provoca uma resposta dos neurônios é definido como o limiar de contraste. O inverso de um determinado estímulo resulta no valor da sensibilidade de contraste.

- **Funções sensíveis ao contraste:** A função de sensibilidade ao contraste (FSC) varia a frequência. Assim, a FSC modela a variação na sensibilidade do HVS sobre as diferentes frequências espaciais e temporais que estão presentes no

estímulo visual. Esta propriedade foi muito explorada para projetar aparelhos de televisão, câmeras fotográficas e de vídeo (LAMBRECHT et al., 1996). No entanto, de acordo com Wang et al. (2003), alguns modelos optam por implementar a FSC como uma operação de filtragem, enquanto outros implementam o FSC através de fatores de ponderação para sub-bandas após uma decomposição de frequência.

Em geral, o FSC espacial é modelado como uma função de passagem de banda invariante no espaço. Ainda que, o FSC seja ligeiramente passivo, a maioria dos algoritmos de avaliação de qualidade aplicam um passa baixas. Tornando assim, as métricas de avaliação de qualidade mais robustas para mudanças na distância de visualização. A sensibilidade ao contraste também é uma função da frequência temporal, que apesar de ser irrelevante para a avaliação da qualidade da imagem, foi modelada para a avaliação da qualidade do vídeo como um simples filtro temporal (WANG et al., 2003).

- **Percepção de cor:** Na origem de qualquer teoria da cor existem os fenômenos ligados à maneira como as cores são percebidas e diferenciadas. Basicamente, as cores percebidas são determinadas pela natureza da luz que é refletida pelo objeto. Sendo que, a compreensão da percepção de cores depende da maneira como a luz é codificada pelo cérebro em função da distribuição espectral. Logo, a caracterização da luz é importante para a ciência das cores. No caso, se a luz for acromática (sem cores), seu único atributo será sua intensidade, já a luz cromática engloba o espectro de energia eletromagnético de aproximadamente 400 a 700 nm. São utilizados três valores para descrever a qualidade de uma fonte de luz cromática: radiância, luminância e o brilho. A radiância é a quantidade total de energia que flui da fonte de luz naturalmente e é medida em wats (W). A luminância, mede a quantidade de energia que um observador percebe de uma fonte de luz, e é medida em lumens (lm). O brilho incorpora a ideia acromática de intensidade e é um dos fatores mais importantes para a descrição de sensação de cores (GONZALEZ; WOODS, 2000; FONSECA, 2008).

Portanto, uma vez que o sistema de visão recebe os estímulos luminosos e transfere essas informações para o cérebro, onde essas informações são processadas, é criada a percepção de imagem, a identificação e a compreensão do funcionamento das principais características do HVS. Estas características estão relacionadas com fatores de percepção de qualidade e são importantes para o desenvolvimento de medidas de Avaliação de Qualidade de Imagem e Vídeo.

2.2 Informações de Profundidade Estereoscópica

Segundo Gazzaniga (2004) uma das capacidades mais impressionantes do HVS é a capacidade de inferir a estrutura tridimensional (3D) do ambiente a partir de imagens formadas na retina. A distância entre um olho humano e o outro é em média de 6.5cm, que se movimenta em conjunto para uma mesma direção, apresentando, os dois, um ângulo de visão que é limitado. Como deslocam-se na horizontal, os elementos visuais captados por cada olho são semelhantes, no entanto, chegam a cada retina em posições longitudinais distintas, com um deslocamento pequeno, criando dois pontos de vista. A diferença entre as posições dos elementos em cada ponto de vista é chamado de disparidade binocular.

A percepção de diferentes níveis de profundidade e a distância entre os objetos que vemos, causa sensação de profundidade, que ocorre em virtude da disparidade binocular, chama-se estereopsia e só pode ser obtida com o uso de dois olhos (ZINGARELLI, 2013). Ainda, diretamente ligado a disparidade, existe a paralaxe. Esta, conforme Zingarelli (2013), é a distância horizontal entre a imagem esquerda e a direita em que os objetos aparecem em relação ao observador.

A capacidade estereoscópica dos seres humanos tem sido á principal força por trás dos esforços para o desenvolvimento de tecnologias em vídeo tridimensional. As principais características que são levadas em consideração são: a paralaxe, estereoscopia e binocularidade. Em termos técnicos, conforme Zingarelli (2013), os vídeos 3D que possibilitam a percepção de profundidade são definidos como vídeos estereoscopicos.

A paralaxe e um conceito fundamental para a produção de vídeos estereoscópicos, pois quando utilizada corretamente é capaz de gerar pontos de vista diferentes de uma mesma imagem para cada olho, tendo como consequência a formação da disparidade. Em casos que a paralaxe é utilizada incorretamente ou não calibrada, causa uma divisão incompleta das informações do par estéreo durante a visualização, introduzindo elementos visuais de uma das figuras do par na outra figura, chamado de *Crosstalk* e causa efeitos indesejáveis e desconforto na visualização estereoscópica.

A disparidade binocular, paralaxe e a estereopsia são classificadas como informações de profundidade estereoscópicas, pois com elas e possível obter as informações necessárias para o cérebro interpretar a profundidade em pares estereoscópicos.

2.3 Degradações

Durante o processo de compressão alguns tipos de degradações podem ser introduzidos, sendo em alguns casos perceptíveis ao olho humano, influenciando assim diretamente na qualidade de imagem e vídeo diante o usuário. Neste contexto, para verificar a eficiência de determinada métrica objetiva é necessário possuir amostras de

vídeos que apresentem degradações encontradas em condições reais (REGIS, 2013) (TANJI et al., 2014). Estas degradações podem ser geradas através da codificação, armazenamento, filtragem, conversão ou transformação. Alguns dos efeitos indesejados do processo de compressão são: perda de resolução, efeito de bloco, ruído de quantização e erros de blocos.

De acordo com Tanji et al. (2014) vale observar que, mesmo o vídeo comprimido sendo geralmente degradado em relação ao original, alguns processos podem reduzir o nível de degradação existente. O que significa, portanto, que para alguns tipos de degradação a codificação pode significar uma melhora na imagem. Por exemplo, quando o sinal original já está degradado o vídeo comprimido pode ser menos ruidoso do que o original, a seguir algumas degradações serão melhor detalhadas.

Blocagem: o efeito em bloco ou (*blocking*) é uma degradação dada por uma deterioração em que a imagem recebida apresenta padrões retangulares que não estavam presentes na imagem original. É causado pela perda de informação dos pixels, normalmente resultante do processo de compressão baseada em blocos (PAN et al., 2004). Sendo que as descontinuidades são mais visíveis quando a quantização é agressiva. Comumente, este tipo de degradação está presente em cenas complexas e que exigem altas taxas de compressão, a Figura 2 apresenta um exemplo do efeito de Blocagem.

Figura 2 – Artefatos: (a) Imagem original (b) blocagem. Fonte: adaptada de Arthur (2002)

Borramento: O efeito de borramento do inglês *blurring*, é a perda dos detalhes espaciais da imagem, ou seja, a redução da definição em bordas das áreas com muitos detalhes espaciais. Em cenas com alto detalhamento espacial este efeito é consequência do compromisso na alocação de bits para a descrição dos detalhes de alta frequência e na descrição do movimento (TANJI et al., 2014). O borramento é visualmente caracterizado por uma imagem turva ou desfocada como pode ser visto na Figura 3.

Ruído de quantização: Há diferentes degradações associadas ao ruído produzido pelos algoritmos de compressão. o ruído de quantização é uma das deteriorações predominantes em imagens, já que a quantização é uma parte fundamental na compres-

Figura 3 – Artefatos: (a) Imagem original (b) borramento. Fonte: adaptada de Arthur (2002)

são de vídeo. Este ruído tende a ser descorrelacionado do sinal, mas é uniformemente distribuído ao longo da imagem, como pode ser notado no exemplo apresentado na Figura 4.

Figura 4 – Artefatos: (a) Imagem original (b) ruído de quantização. Fonte: adaptada de Arthur (2002)

Ringing: Este efeito normalmente, é consequência da etapa de quantização, que busca eliminar as altas frequências da imagem, produzindo granulações ou ondulações nas bordas da imagem. Normalmente este é mais visível nas bordas do que em superfícies lisas (REGIS, 2013).

Blocos errados: Os erros do canal podem ocasionar a presença de blocos errados no vídeo reconstruído. O vídeo digital comprimido é propenso a erros de canal, por causa da utilização de ferramentas de codificação preditiva. Conforme o tamanho do GOP (*Group Of Pictures*) e o tipo de imagem afetada, um bloco errado pode se espalhar por outros quadros (TANJI et al., 2014).

2.4 Aprendizado de Máquina

No Aprendizado de Máquina (AM) os sistemas são programados para aprender com experiências anteriores. Para isso, adotam o princípio de inferência chamado de indução, em que se obtêm conclusões genéricas através de um conjunto especifico de exemplos. Assim, algoritmos de AM aprendem a induzir uma função ou hipótese que tem a capacidade de solucionar um problema através de dados que represen-

tam instâncias do problema a ser resolvido (FACELI et al., 2011). De modo geral, o Aprendizado de Máquina pode ser distinguido em três casos: não supervisionado, por reforço e supervisionado (RUSSELL, 2010).

As técnicas não supervisionadas envolve a aprendizagem de padrões baseados em dados de entrada, quando não são fornecidos valores de saída específicos. Atuam selecionando um determinado conjunto de dados e reconhecendo os padrões neles. Esses algoritmos identificam semelhanças nos dados e tomam ações baseadas na presença ou ausência de tal identidade em cada novo dado. Aprendendo com dados de teste que não são rotulados, classificados ou categorizados. As principais tarefas descritivas da aprendizagem não supervisionada são a extração de regras de associação e agrupamento (*clustering*) (AKINBO; DARAMOLA, 2021; DEVEZA, 2011).

O aprendizado por reforço busca a criação de agentes que tenham a capacidade de tomar decisões corretas de um dado ambiente sem ter qualquer conhecimento prévio sobre o mesmo. Para que o aprendizado ocorra é necessário que o agente depois de reconhecer o ambiente tome ações e observe a recompensa imediata advinda da ação e a mudança no ambiente acontece com base na ação tomada. Desta maneira, o agente não está somente interessado em ações que produzem a maior recompensa imediata, mas sim em ações que permitam o maior acúmulo de recompensa a longo prazo (ALMEIDA TEIXEIRA, 2016; AKINBO; DARAMOLA, 2021).

As técnicas supervisionadas analisam o conjunto de dados fornecido como entrada com a finalidade de “aprender” a classificar novos dados, os algoritmos operam de tal forma que desenvolverão um modelo matemático do dados que compõem as entradas (dados enviados para o sistema) e as saídas esperadas. Os dados fornecidos são categorizados como os dados de treinamento que compreendem os conjuntos de exemplos de treinamento com uma ou mais entradas. A modelagem matemática aplicada no aprendizado supervisionado utiliza um vetor de características para extração e os dados a serem treinados por uma matriz. O algoritmo que aprende determinada que a tarefa tem a capacidade de aprimorar os resultados na Precisão das saídas com o objetivo de classificar ou prever determinada tarefa e, portanto, permitir a obtenção de um bom resultado. As principais tarefas preditivas dessas técnicas supervisionadas são classificação e regressão (AKINBO; DARAMOLA, 2021; RUSSELL, 2010).

Neste trabalho iremos abordar técnicas de AM supervisionadas de classificação. A classificação é o processo que busca encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados (DEVEZA, 2011). Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertencem a fim de aprender como classificar um novo registro (aprendizado supervisionado) (CAMILO; SILVA, 2009). Tan; Steinbach; Kumar (2016) definem a classificação como a tarefa de aprender uma fun-

ção alvo f que mapeie cada conjunto de atributos x para um dos rótulos de classes y pré-determinados. Essas técnicas são apropriadas para prever ou descrever conjuntos de dados com categorias nominais ou binárias e menos efetivas para categorias ordinais, porque não consideram a ordem implícita entre as categorias. Os algoritmos de classificação incluem métodos que utilizam Árvores de Decisão, Redes Baysianas, vizinhos mais próximos, algoritmos genéticos, entre outros.

2.5 Árvore de Decisão

A técnica de Árvore de Decisão é um modelo estatístico que emprega o aprendizado supervisionado para a classificação e previsão dos dados. É considerada uma técnica que não é complexa e é de fácil interpretação, já que seu modelo também pode ser representado como um conjunto de regras do tipo se-então-senão. Dada a facilidade de compreensão do modelo, está técnica é muito utilizada em problemas de classificação (TAN; STEINBACH; KUMAR, 2016).

Os algoritmos de Árvore de Decisão são caracterizados pela utilização da técnica de Divisão e Conquista durante sua execução. A Divisão e Conquista baseia-se na sucessiva divisão do problema em vários subproblemas de menores dimensões, até que seja encontrada uma solução menor para cada subproblema. Mediante esta estratégia, os algoritmos de AD procuram dividir sucessivamente o conjunto de dados em vários subconjuntos, até que cada subconjunto complete apenas uma classe ou até que uma classe apresente uma minoria, não necessitando de novas divisões (RUSSELL, 2010; QUINLAN, 1986).

A Figura 5 representa uma AD hipotética, onde temos X_1 como o nó raiz que é o elemento que se encontra no topo da árvore, sendo a sua origem; são chamados nós todos os itens representados na árvore. Já os ramos são as ligações entre esses nós, os nós filhos seguem logo abaixo do nó pai; e por fim tem-se os nós que não possuem filhos, ditos nós folhas e representam o final das ramificações da árvore (RUSSELL, 2010).

No contexto da utilização da Árvore de Decisão como um modelo de classificação, para melhor compreensão define-se um conjunto de dados T que contém n registros, sendo que cada um desses registros contém dois tipos de atributos, o atributo classe e os atributos preditivos. O atributo classe indica a qual classe pertence o seu registro correspondente, já os atributos preditivos são analisados pelo algoritmo de AM com a finalidade de entender como se relacionam com o atributo classe (TAN; STEINBACH; KUMAR, 2016; RUSSELL, 2010). Logo, temos que um modelo de classificação de Árvore de Decisão tem na representação de sua árvore cada nó, incluindo o nó raiz, sendo um teste sobre um atributo preditivo; um ramo descendente partindo de um nó representa um provável resultado para o teste sobre o atributo que o nó em questão

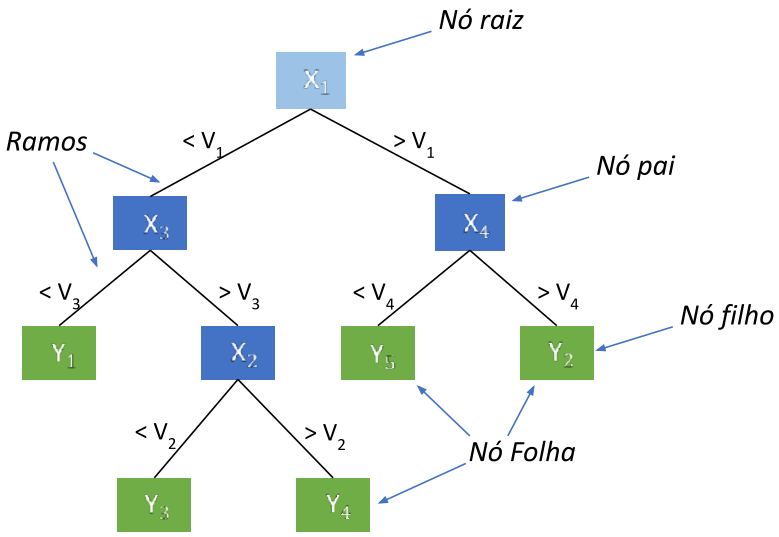


Figura 5 – Exemplo de uma Árvore de Decisão - adaptada de Russell (2010).

representa e o nó folha é considerado um atributo rótulo de classe. A classificação de um novo registro ocorre quando se percorre um dos caminhos da árvore, começando na raiz, testando os atributos nós, até alcançar alguma folha que indica a classe desse novo registro (QUINLAN, 1986).

A complexidade de uma AD pode ser medida por uma métrica que contém: o número total de nós, número total de folhas, profundidade da árvore e número de atributos usados na sua construção. O tamanho da AD deve ser relativamente pequeno, o que pode ser controlado usando técnicas de poda (BHARGAVA et al., 2013; MAIMON; ROKACH, 2014). Há diversos algoritmos de AM baseados em Árvore de Decisão, que permitem adicionar técnicas como a poda, níveis de profundidade que se deseja tratar, entre outros. Além disso, alguns algoritmos, comumente chamados de florestas, trabalham com conjuntos de AD. Estes e outros algoritmos serão melhores detalhados a seguir.

2.5.1 Algoritmos Baseados em Árvore de Decisão

Diversos algoritmos de que utilizam técnicas de AM se baseiam em Árvore de Decisão, dentre eles *J48*, *ReTree*, *RandomForest* e *ForestPA*, que serão melhor discutidos a seguir e utilizados neste trabalho.

J48*:* este algoritmo é a implementação do software WEKA para o algoritmo de Árvore de Decisão C4.5 (QUINLAN, 1986; BOUCKAERT et al., 2018). É uma AD que recebe um vetor de atributos como entrada e retorna uma decisão com base em uma série de testes lógicos sobre os valores de entrada. Este usa o método de divisão e conquista para aumentar a capacidade de predição das AD's. Com isso, sempre usa o melhor passo avaliado localmente, sem se preocupar se esse passo vai produzir a melhor solução global, dividindo um problema em vários subproblemas criando assim

sub-árvores entre a raiz e as folhas (ALVARENGA, 2014).

***RepTree*:** utiliza a lógica da AD e cria diversas árvores em diferentes iterações, através da variância das informações e usando MSE para as podas. A partir de então, ele seleciona a melhor de todas as árvores geradas, que será considerada a representante (KALMEGH, 2015). Para a poda da árvore o algoritmo utiliza a medida do Erro Quadrático Médio (*Mean Square Error* - MSE) das previsões feitas pela árvore. O RepTree, é um algoritmo de aprendizado rápido que constrói uma AD baseada no ganho de informação ou na diminuição da variância. A árvore de decisão/regressão é construída usando o ganho de informação como critério de divisão e o remove usando a poda de erro reduzida, classificando apenas uma vez os valores para atributos numéricos. Já os valores que estão ausentes são tratados usando o método C4.5, de instâncias fracionárias (KALMEGH, 2015).

***Random Forest*:** este algoritmo gera uma floresta aleatória, que combina diversas ADs em um só classificador. Esta técnica foi desenvolvida para reduzir a tendência de *overfitting* das ADs ao adicionar aleatoriedade na construção da árvore individual. Assim, garante que todas as árvores do conjunto sejam gerados a partir de uma amostra com substituição do conjunto de treinamento. A decisão final é normalmente realizada por voto majoritário, podendo ser ponderado pela probabilidade da estimativa. Durante o treinamento o algoritmo busca descobrir o número ideal de árvores que formam a floresta e profundidade máxima de cada árvore (ALI et al., 2012; AMEER et al., 2019).

***ForestPA*:** constrói um conjunto de ADs (floresta), considerando os atributos que já foram utilizados na construção de outras árvores da floresta. Além disso, o *ForestPA* impõe pesos aos atributos que participaram da construção de árvores anteriores para gerar as árvores subsequentes. O algoritmo possui ainda um mecanismo para aumentar gradualmente os pesos dos atributos que não foram testados nas árvores subsequentes (ADNAN; ISLAM, 2017; SAMAT et al., 2019).

Ainda neste contexto, existem diferentes formas de avaliar os modelos de AM. Alguns exemplos são a matriz de confusão, a Acurácia, a Precisão, o *Recall* e o *F-Measure*, que na Seção 6 será tratado como *F1-Score*. Neste trabalho essas medidas são utilizadas para avaliar os resultados obtidos e serão detalhadas na Seção 2.6.

2.6 Métricas de Avaliação de Desempenho de AM

Nesta Seção serão abordadas algumas das métricas de avaliação de desempenho para os classificadores que foram utilizados no trabalho, entre elas a Matriz de confusão, Acurácia, Precisão, *Recall* e *F-Measure*. Estas serão aprofundadas a seguir.

Matriz de confusão: Esta oferece uma medida prática do modelo de classificação ao apresentar o número de classificações corretas em comparação as classificações preditas para cada uma das classes, estruturando os resultados em duas dimensões:

classes verdadeiras e classes preditas (TAN; STEINBACH; KUMAR, 2016).

A Tabela (1) apresenta um exemplo de uma Matriz de Confusão, em que V_p e V_n representam os valores classificados corretamente (V - Verdadeiro) para as classes preditas A + e A - . Já F_p e F_n são os valores das instâncias classificadas incorretamente (F - Falso). Para ambos os casos p e n representam respectivamente (+) positivo e (-) negativo.

Tabela 1 – Exemplo de uma Matriz de Confusão para um classificador de um conjunto de dados com duas classes: A- e A+ - adaptada de Monard; Baranauskas (2003).

classe	predita A +	predita A -
verdadeira A +	V_p	F_p
verdadeira A -	F_n	V_n

Por fim, o total de exemplos é dado pela Eq.(1).

$$n = V_p + V_n + F_p + F_n \tag{1}$$

As medidas como Acurácia, Precisão, *Recall* e *F-Measure* utilizam como base os valores da Matriz de Confusão.

Acurácia: A Acurácia e a taxa de erro são métricas comumente utilizadas para a avaliação de desempenho dos classificadores. A Acurácia é a taxa de acertos do classificador, ou seja, a taxa de exemplos positivos e negativos classificados corretamente dentre todos os exemplos do conjunto de dados, conforme pode ser observado na Eq.(2). É importante observar que na presença de conjuntos de dados com classes desbalanceadas, ou seja, números desiguais entre diferentes classes, a Acurácia torna-se uma medida não confiável. Isso pode ser explicado pelo fato do viés do modelo classificar corretamente a classe majoritária e assim obter uma alta Acurácia (TAN; STEINBACH; KUMAR, 2016).

$$Ac = \frac{V_p + V_n}{n} \tag{2}$$

Precisão e *Recall*: Diferentemente da Acurácia que é mais suscetível ao desbalanceamento entre as classes, a Precisão e o *Recall* (do português Revocação) podem se tornar mais adequadas quando se trata de classes com desbalanceamento e, portanto, mais confiáveis para a avaliação dos classificadores (TAN; STEINBACH; KUMAR, 2016). A Precisão é a taxa de exemplos classificados corretamente como A + (V_p) dentre todos os exemplos classificados como A + $(V_p$ e $F_p)$, conforme descrito na Eq.(3).

$$Prec = \frac{V_p}{V_p + F_p} \tag{3}$$

O *Recall* corresponde a taxa de exemplos corretamente classificados como A+

(V_p) dentre todos os exemplos que realmente são A+ (V_p e F_p), a Eq.(4) apresenta sua conotação.

$$Rec = \frac{V_p}{V_p + F_n} \tag{4}$$

Conforme é descrito na Eq.(4), notamos que o *Recall* indica a proporção (ou a probabilidade) de um exemplo de dada classe ser classificado como tal (TAN; STEINBACH; KUMAR, 2016). Em outros termos, conforme Monard; Baranauskas (2003) o *Recall* é a capacidade do classificador em reconhecer todas as instâncias de uma classe de interesse.

***F-Measure* (Medida-F):** É a média harmônica entre as medidas de Precisão e o *Recall* dados pela Eq.(5).

$$Fmeasure = \frac{2 \times Prec \times Rec}{Prec + Rec} = \frac{2}{\frac{1}{Prec} + \frac{1}{Rec}} \tag{5}$$

Estas medidas, quando analisadas em separado, podem ser enganosas, pois uma Precisão alta geralmente indica dificultar um bom resultado para o *Recall* e vice-versa. Desta maneira, através da média harmônica entre as duas, torna-se possível mensurar um desempenho mais realista (TAN; STEINBACH; KUMAR, 2016).

2.7 Medidas de Erro para Técnicas de AM

As medidas de erro são comumente utilizadas para medir o quanto a previsão está próxima de um eventual resultado (OSISANWO et al., 2017). Neste trabalho serão apresentados o Erro Absoluto Médio (*Mean Absolut Error* - MAE) e o Erro Quadrático Médio (*Root Mean Squared Error* - RMSE), que serão detalhados a seguir.

Erro Absoluto Médio: É um critério que estima a média da magnitude dos erros em um conjunto de valores dos dados (previsões), ou seja, calcula o "erro absoluto médio" dos erros entre os valores reais e os preditos (WILLMOTT; MATSUURA, 2005).

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j| \tag{6}$$

A Eq.(6), descreve MAE onde temos n como o número de observações, y_j representa os valores reais (esperados) e $\hat{y}_j$, os valores preditos pelo modelo (AMEER et al., 2019).

Erro Quadrático Médio: Este critério também é utilizado para calcular a magnitude média do erro. É obtido através do cálculo da média das diferenças quadradas entre os valores reais e os valores preditos (WILLMOTT; MATSUURA, 2005).

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{7}$$

A Eq.(7) descreve a métrica de RMSE, em que as variáveis apresentadas na equação são as mesmas de MAE (Eq.(6)). Ambas as medidas calculam o erro médio do modelo preditivo, em relação aos dados originais tanto de treino como teste. Sendo assim, quanto menor seus valores, melhor são suas avaliações.

2.8 Considerações Finais do Capítulo

Este Capítulo apresentou os principais temas para o desenvolvimento deste trabalho. O sistema visual humano é a principal base para a interpretação da representação digital de um vídeo colorido, portanto conhecer os limites da percepção de degradações através do HVS é importante para determinar os parâmetros em sistemas de compressão, transmissão e armazenamento de vídeos. Muitos fatores podem gerar degradações nos vídeos, alguns desses erros são percebidos pelo olho humano, influenciando diretamente na qualidade do vídeo disponível para o usuário. Dentre as degradações mais comuns estão: a blocagem; borramento; ruídos de quantização; erros de bloco; e *ringing*. Além disso, esse Capítulo apresenta conceitos sobre Aprendizado de Máquina focando em Árvore de Decisão, um dos temas colocados nas questões de pesquisa deste trabalho. Foram apresentados alguns algoritmos basados em ADs e medidas de avaliação de desempenho de AM. Os assuntos abordados neste Capítulo são de importantes para a compreensão do desenvolvimento deste trabalho.

3 AVALIAÇÃO DE QUALIDADE DE IMAGEM E VÍDEO

Este Capítulo apresenta a Avaliação de Qualidade de Imagem e Vídeo, a Seção 3.1 descreve de maneira geral os principais conceitos e características da Avaliação Objetiva de Qualidade de Imagem e Vídeo, que pode ser classificada segundo a presença ou não de um vídeo de referência, sendo: Referência Completa, Referência Reduzida e Sem Referência. Logo, são apresentadas algumas das Métricas Objetivas de Qualidade de Imagens e Vídeo 2D e 3D, respectivamente nas Seções 3.2 e 3.3. A Seção 3.4 discorre sobre a Avaliação Subjetiva de Qualidade de Imagem e Vídeo. A Seção 3.5 destaca os principais Métodos Subjetivos encontrados na literatura, que são sugeridos e guiados pelas normas ITU.

3.1 Avaliação Objetiva de Qualidade de Imagem e Vídeo

A avaliação objetiva de qualidade é conhecida por ser um método realizado com a utilização de sistemas computacionais, que analisam as entradas e calculam os resultados, retornando um escore para cada vídeo processado (WANG; LU; BOVIK, 2004). O escore é validado e comparado com seu respectivo valor de MOS ou DMOS descritos na Seção 3.4. Essa avaliação é compreendida como uma modelagem matemática que permite avaliar o grau de degradação do vídeo, após algum processo de distorção. A degradação do vídeo é perceptível quando aparecem artefatos em seu conteúdo, como blocagem, borramento, distorção localizada em áreas pouco nítidas do quadro e travamentos (SILVA, 2013).

As Métricas Objetivas de Qualidade de Imagem e Vídeo podem ser classificadas de acordo com a disponibilidade do sinal original da imagem ou sinal de vídeo, podendo ser sem distorção com a qualidade perfeita. De acordo com Wang et al. (2003), essa disponibilidade pode ser utilizada como referência para comparar uma imagem ou vídeo distorcido. Uma metodologia que utiliza referência necessita ter acesso tanto aos dados degradados quanto aos originais (sem distorção), enquanto metodologias que não utilizam vídeos de referência precisam apenas dos dados degradados (DARONCO; ROESLER; LIMA, 2008).

Em geral, conforme Wang et al. (2003), a maioria das Métricas Objetivas de Qualidade de Imagem e Vídeo utilizam sinais de referências sem distorções. Sendo assim, a qualidade objetiva de vídeo é medida utilizando um algoritmo que compara o vídeo original (vídeo de referência) com o vídeo codificado (degradado). Essa comparação é realizada quadro a quadro, comparando todos os pixels dos quadros do vídeo de referência com os pixels do vídeo codificado. No caso da avaliação de imagens o processo é o mesmo, só que a comparação ocorre entre a imagem original e a imagem degradada (AGOSTINI, 2007).

As métricas objetivas podem ser divididas em categorias que dizem respeito a disponibilidade de uma imagem ou vídeo de referência. Estas categorias são:

Referência Completa (*Full Reference* - FR): O método FR assume que o vídeo de referência está totalmente disponível no momento da avaliação de qualidade (DARONCO; ROESLER; LIMA, 2008). Diante do vídeo de referência, as métricas FR são capazes de mensurar as distorções do vídeo processado, e estimar um valor que condiz a seu nível de degradação. Métricas como MSE (*Mean Square Error*), PSNR (*Peak Signal to Noise Ratio*), SSIM (*Structural Similarity*), VQM (*Visual Quality Metric*) e MOVI (*Motion-based Video Integrity Evaluation*) são amplamente utilizadas na literatura (ROMANI, 2015).

As medidas FR são realizadas mediante comparações entre os pixels da imagem original e da imagem degradada. Sendo assim, esse método tem como objetivo reunir a maior quantidade possível de informações que sejam úteis, refletindo assim, em um método de qualidade robusto e eficaz (ARTHUR, 2002).

Segundo a ITU-T J.143 (2000) uma maneira de avaliação de qualidade utilizando a técnicas de FR é fazer uma comparação entre o vídeo de entrada (referência) na entrada do sistema e o vídeo processado (degradado) na saída do sistema, a Figura 6 apresenta esse modelo.

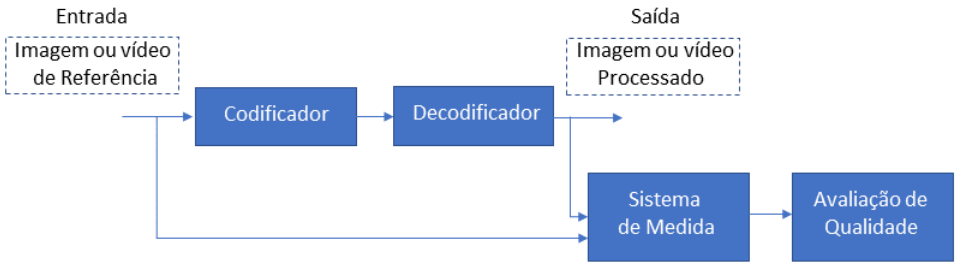


Figura 6 – Avaliação Objetiva de Qualidade de Imagem e Vídeo com Referência Completa - Adaptado de ITU-T J.143 (2000).

A comparação entre os sinais de entrada e de saída pode exigir um processo de alinhamento espacial e temporal para compensar qualquer deslocamento ou corte de imagem vertical ou horizontal. Também pode exigir correção para qualquer compen-

sação ou ganhar diferenças nos canais de luminância e crominância. Por serem objeto de estudo desta Tese, daremos maior atenção às métricas de Referência Completa.

Referência Reduzida (*Reduced Reference* - RR): A Avaliação de Qualidade de Imagem e Vídeo de Referência Reduzida não admite a disponibilidade completa do sinal de referência, é disponibilizado somente a informação uma parte do sinal, que está disponível através de um canal de dados auxiliares (WANG et al., 2003).

Ainda que, não tenha uma referência disponível, essas métricas utilizam um canal de comunicação adicional para a transmissão de informações que auxiliam na avaliação de qualidade. Estas informações adicionais são características extraídas do vídeo de referência e permitem que o custo de transmissão seja menor do que todo o vídeo de referência (DARONCO; ROESLER; LIMA, 2008).

Um tipo diferente de abordagem de "*double ended*", ou estímulo duplo, utiliza sistemas de medição nos pontos A e B conforme pode ser visto na Figura 7, em que os parâmetros específicos são extraídos da referência e dos sinais processados.

Deste modo, os dados de referência relativos a esses parâmetros no ponto A são sinalizados para o sistema de medição no ponto B a fim de permitir uma comparação entre os parâmetros em cada extremidade da cadeia. Isso, juntamente com o conhecimento do vídeo de saída, pode fornecer uma indicação de qualidade do sinal. Os parâmetros podem incluir blocagem, informação de sinal espacial, temporal e ruído. Perturbações específicas, como congelamento de quadros e perda de imagem também podem ser detectadas. Essa metodologia oferece capacidade de uso onde o decodificador e o codificador são separados fisicamente (ITU-T J.143, 2000).

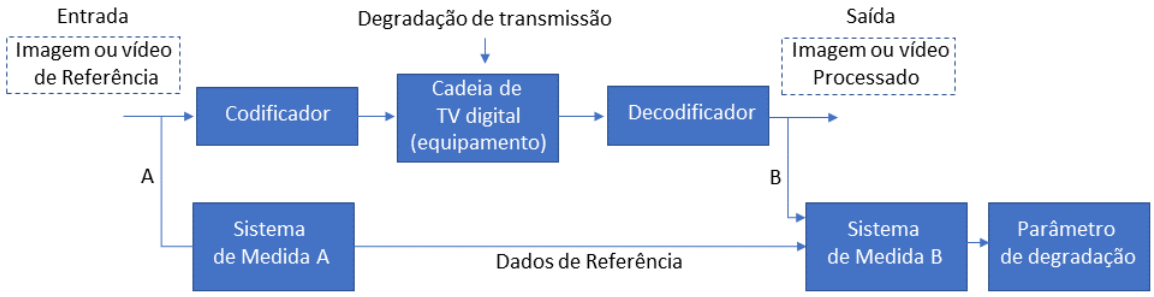


Figura 7 – Avaliação Objetiva de Qualidade de Imagem e Vídeo com Referência Reduzida - Adaptado de ITU-T J.143 (2000).

Sem Referência (*No Reference* - NR): Utilizado para analisar um vídeo degradado sem acesso ao vídeo de referência (Figura 8), ou seja, realizam a avaliação de vídeo somente através do vídeo processado (ROMANI, 2015). Essas métricas devem conseguir assumir as distorções e degradações presentes no vídeo, aumentando assim muito a complexidade no desenvolvimento do método.

A falta de referência significa que a medição pode estar sujeita a erros ocasionados

pelo conteúdo da imagem que se assemelham aos parâmetros de comprometimento específicos que estão sendo detectados. Tal como acontece com a técnica que utiliza informações de referência reduzidas, as medidas referem-se aos parâmetros que indicam deficiências na imagem, mas não se correlacionam diretamente com a Avaliação Objetiva de Qualidade de Imagem (ITU-T J.143, 2000).

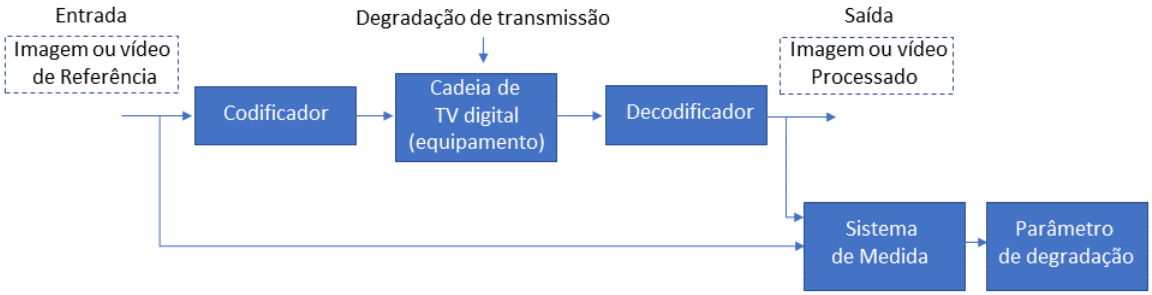


Figura 8 – Avaliação Objetiva de Qualidade de Imagem e Vídeo Sem Referência - Adaptado de ITU-T J.143 (2000).

3.2 Métricas Objetivas de Avaliação de Qualidade 2D

Nesta Seção serão abordadas algumas das principais Métricas Objetivas de Avaliação de Qualidade de Imagem e Vídeo 2D de Referência Completa. As métricas apresentadas podem ser descritas se referindo somente a vídeos ou imagens, mas normalmente são aplicadas aos dois casos.

3.2.1 SAD (*Sum of Absolute Differences*)

Esta métrica mede a similaridade entre a imagem ou vídeo de referência e o degradada. É calculada através da diferença absoluta entre cada amostra dos vídeos que estão sendo comparados. Sua definição é dada pela Eq.(8), onde f_{ij} e F_{ij} são os valores das amostras do vídeo de referência e degradado, respectivamente. i e j representam a posição das amostras (ZHU; XIONG, 2009).

$$SAD = \sum_{i=1}^{N} \sum_{j=1}^{M} |f(i,j) - F(i,j)| \tag{8}$$

3.2.2 PSNR (*Peak Signal-to-Noise Ratio*)

O PSNR fornece a estimativa da qualidade do vídeo usando a diferença entre os pixels do vídeo degradado em relação aos do vídeo de referência. Esta métrica é derivada através da definição do Erro Quadrático Médio, à medida que o MSE se aproxima de zero, o valor de PNSR tende a aumentar (HUYNH-THU; GHANBARI, 2008; WANG et al., 2003). PSNR e MSE são definidos pelas Equações (10) e (9),

respectivamente, onde M e N representam a largura e altura da imagem em *pixels*. Já $f(i,j)$ corresponde ao sinal de referência e $F(i,j)$ ao sinal degradado no pixel (i,j) (LIOTTA et al., 2013; WANG; BOVIK, 2009). O valor de MAX_I corresponde ao maior valor possível de uma amostra da imagem (FONSECA, 2008). Considerando amostras de 8 bits, MAX_I é definido como 255 (FONSECA, 2008). Já para imagens em cores com três valores de amostra por pixel, como o RGB, a definição do PSNR segue a mesma e o MSE passa a ser dividido pelo tamanho da imagem (W e H) e por 3, ou seja, $W \times H \times 3$.

$$MSE = \frac{1}{WH} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} [X(x,y) - Y(x,y)]^2 \tag{9}$$

$$PSNR = 10log_{10} \frac{MAX_I^2}{MSE(i)} \tag{10}$$

Apesar de PSNR ser uma das métricas mais utilizadas na literatura, este não leva em consideração o sistema visual humano, fato que, pode levar a uma menor correlação com valores de testes subjetivos (PUNCHIHEWA; BAILEY; HODGSON, 2003).

3.2.3 *Structural Similarity Index Measure* (SSIM)

Esta métrica é utilizada para prever a qualidade percebida de imagens e vídeos, baseada na similaridade estrutural. Sendo a primeira versão desenvolvida por LIVE (*Laboratory for image & Video Engineering*) [1] e sua versão completa desenvolvida em parceria com o Laboratory for Computional Vision (LCV)[2] da Universidade de Nova York. De acordo com Wang; Lu; Bovik (2004) é um modelo baseado na percepção, que considera a degradação da imagem como uma mudança percebida na informação estrutural. A informação estrutural em uma imagem diz respeito aos atributos que refletem a estrutura do objeto da cena, que é independente da média da luminância e contraste da imagem (GURAV; PATIL, 2016). O valor de SSIM é obtido através da Eq.(11).

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{11}$$

As medidas dos parâmetros de luminância, contraste e distorção estrutural podem ser dados separadamente, conforme as Eqs.(12,13 e 14).

$$L(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C1} \tag{12}$$

[1] http://live.ece.utexas.edu/
[2] http://www.cns.nyu.edu/ lcv/

$$C(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{13}$$

$$S(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{14}$$

Temos que $L(x,y)$ representa a luminância, $C(x,y)$ os valores de contraste e $S(x,y)$ a distorção. Além das variáveis definidas para a Eq.(11) tem-se agora os valores do desvio-padrão, dado por σ_x e σ_y. Desta forma, os valores de $SSIM(x,y)$ resulta em um valor decimal entre -1 e 1, indicando que o valor for igual a 1 as amostras são idênticas.

Os valores de μ_x e μ_y são obtidos através do cálculo de médias locais para cada pixel; σ_x e σ_y são os valores do desvio-padrão (estimadores não padronizados). Já C_1 e C_2 são constantes que estabilizam a divisão quando os denominadores forem pequenos. Que são dados por:

$$C_1 = (K_1L)^2 e C_2 = (K_2L)^2 \tag{15}$$

Considerando L como o alcance dinâmico dos valores de pixel (para 8 bits/pixel de imagens em níveis de cinza, L=255) e C_1 e C_2 adquirem valores muito pequenos.

Esta métrica também é utilizada como base da metodologia de várias métricas que aproveitam a sua característica de avaliação estrutural, como por exemplo: MS-SSIM (*Multiscale* - SSIM), baseado em multi-escalas de estrutura de similaridade; *SW-SSIM* (Speed Weighted - SSIM), baseada em diferenças de pesos; MC-SSIM (*Motion Compensated* - SSIM), baseado em compensação de movimento; e LST-SSIM (*Local Spatial - Temporal* - SSIM) baseada em localização de características de bordas de movimento (ROMANI, 2015).

3.2.4 *Visual Information Fidelity* (VIF)

O critério de VIF é também conhecido como o índice de Sheikh; Bovik (2005), pressupõe que uma imagem distorcida é a saída de um canal de comunicação que introduz erros na imagem de referência e passa através da mesma. Em Sheikh; Bovik (2005), os autores propõem uma extensão do critério VIF de imagens estáticas para vídeos utilizando derivadas temporais (SESHADRINATHAN et al., 2010). Esta métrica baseia-se em algumas características do HVS, como a função de propagação do ponto óptico, a função de sensibilidade de contraste e ruído neural interno.

A formulação do modelo VIF para vídeo permanece a mesma que para imagens estáticas (SHEIKH; BOVIK, 2006). A medida de fidelidade de informação visual é dada por:

$$\mathrm{VIF} = \frac{\sum_{j \in channels} I(\overset{\rightarrow N,j}{C}; \overset{\rightarrow N,j}{F} | s^{N,j})}{\sum_{j \in channels} I(\overset{\rightarrow N,j}{C}; \overset{\rightarrow N,j}{E} | s^{N,j})} \tag{16}$$

Na Eq.(16) temos $I(\overset{\rightarrow N,j}{C}; \overset{\rightarrow N,j}{F} | s^{N,j})$ e $I(\overset{\rightarrow N,j}{C}; \overset{\rightarrow N,j}{E} | s^{N,j})$ que representam, respectivamente, a informação que idealmente pode ser extraída pelo cérebro de um determinado canal na referência e os vídeos de teste. A medida de fidelidade de informação visual é a fração da informação da imagem de referência que pode ser extraída do sinal de teste. Logo, os canais são somados e $\overset{\rightarrow N,j}{C}$ representa N elementos da RF Cj. Em que, descreve o coeficiente do canal j e assim por diante. O VIF fornecido em Eq.(16) é calculado para uma coleção de $N \times M$ coeficientes wavelet de cada subbanda que podem representar uma subbanda inteira de uma imagem ou uma região espacialmente localizada de coeficientes de subbanda. No primeiro caso, o VIF é um número que quantifica a fidelidade da informação para toda a imagem, enquanto no último caso, uma abordagem de janela deslizante pode ser usada para calcular um mapa de qualidade que pode ilustrar visualmente como a qualidade visual do teste imagem varia ao longo do espaço.

3.3 Métricas Objetivas de Avaliação de Qualidade 3D

A Avaliação de Qualidade de Imagem e Vídeo vem sendo amplamente estudada e desenvolvida, no entanto, a maior parte dos estudos se concentram na área dos vídeos de duas dimensões. Neste contexto, muitas das métricas 2D são estendidas para avaliar as imagens e vídeos estereoscópicos, esta extensão é feita com a adição de informações de profundidade ou disparidade. De acordo com Galkandage et al. (2017), estes métodos deixam a desejar quando correlacionados com resultados subjetivos, pois não incorporam de modo satisfatório as características do HVS que envolvem a estereoscopia. A fim de melhorar a avaliação de qualidade para vídeos 3D, alguns autores propõem métricas que se baseiam em fatores humanos utilizados na percepção de profundidade.

3.3.1 Disparity Temporal Perceptual Weight - SSIM (DTPW-SSIM)

Para o desenvolvimento desta métrica o autor Regis (2013), propõe inicialmente uma abordagem PW-SSIM *Perceptual Weighting - SSIM*, que incorpora à métrica SSIM características da informação espacial perceptiva. Já que a avaliação dos vídeos é realizada utilizando informação perceptual, para fazer uma ponderação em cada região da imagem, considerando que as regiões mais importantes em um vídeo são as que tem a maior informação espacial perceptual. Para obter os valores de DTPW-SSIM são acrescentadas informações de disparidade, ou seja, referentes

à profundidade. O mapa de disparidade, $D(f(x, y, n))$, é definido conforme segue a Eq.(17).

$$\mathrm{D(f(x,y,n))} = |\mathrm{f_l(x,y,n)} - \mathrm{f_r(x,y,n)}| \forall (\mathrm{x,y,n}) \tag{17}$$

f_l e f_r são funções escalares, que correspondem respectivamente as vistas esquerda e direita. As variáveis x, y, n contém a informação de posição do pixel.

A adição da informação de disparidade nas métricas 2D é feita através da média ponderada dos valores das medições objetivas com o mapa de disparidade. Esta adição é dada conforme descrito em Eq.(18).

$$\mathrm{DPW\text{-}SSIM}(f,h) = \frac{\sum_{i=1}^{B} \mathrm{SSIM}(f_i,h_i)\mathrm{SI}_i(f)D_i(f)}{\sum_{i=1}^{B}[\mathrm{SI}_i(f)D_i(f)]} \tag{18}$$

O método proposto utiliza a informação espacial perceptiva (SI - *Spatial Perceptual Information*) como maneira de ponderar as regiões visuais mais importantes. Esta ponderação é obtida calculando a magnitude dos vetores de gradiente no vídeo original, através da máscara de Sobel. Logo é gerado um quadro atual em que os valores dos pixels são as magnitudes dos gradientes. Na sequência, esse quadro é particionado em blocos 8X8 pixels e para cada bloco é calculado o SI_i conforme a Eq.(19).

$$\mathrm{SI_i} = (\frac{1}{P}\sum_{j=1}^{P}(\mu_i - \nabla f_j)^2)^{\frac{1}{2}} \tag{19}$$

Para μ_i representa o valor médio da magnitude do gradiente em um bloco e P é o número de pixels no bloco.

A Eq.(21) apresenta a definição da métrica DTPW, que é baseada em modificações da métrica DPW. Os autores ainda sugerem que os vídeos apresentam um componente temporal que não é considerado nos algoritmos de avaliação. Para considerar este componente temporal, a mudança temporal de um vídeo é estimada pela diferença de luminância dos pixels que estão na mesma posição espacial em quadros sucessivos. Os cálculos são dados através da Eq.(20).

$$\begin{aligned} D_{f,n} &= |f_{n+1} - f_n|, \\ D_{h,n} &= |h_{n+1} - f_n| \end{aligned} \tag{20}$$

Por fim, a métrica DTPW utiliza regiões com grandes alterações perceptivas, em que, o índice de qualidade global é a média entre o índice de qualidade espacial (DPW-SSIM) e temporal (DTP-VQI), conforme segue:

$$\mathrm{DTPW\text{-}SSIM} = \frac{\mathrm{DPW\text{-}SSIM} + \mathrm{DTP\text{-}VQI}}{2} \tag{21}$$

Sendo que a qualidade temporal é estimada por meio do índice de PW-SSIM entre

as diferenças dos quadros $(D_{f,n}, D_{h,n})$ adicionando a disparidade $D_n(f)$.

$$\text{DTP-VQI} = \frac{1}{\text{N}-2}\frac{\sum_{n=0}^{N-2}\text{PW-SSIM}(D_{f,n}, D_{h,n})D_n(f)}{\sum_{n=0}^{N-2}D_n(f)} \tag{22}$$

Para obter uma medida de qualidade percebida o autor propõe a adição de informação temporal e de disparidade ao índice SSIM. O autor Regis (2013) destaca que uma das principais questões para a utilização do índice de similaridade foi que, ao comparar a informação espacial perceptiva dos vídeos originais e processados, obteve-se uma boa aproximação do valor da qualidade percebida pelo HVS sobre a degradação de borramento (REGIS, 2013).

3.3.2 Stereoscopic Structural Distortion (StSD)

A *Stereoscopic Structural Distortion*, é uma métrica de qualidade de vídeo estereoscópico de referência completa para prever a qualidade percebida do vídeo compactado, que leva em consideração as características do HVS, como a disparidade binocular. Esta métrica avalia a distorção estrutural estereoscópica dos vídeos degradados em relação ao vídeo original, considerando a diferença estrutural e o borramento. De maneira geral, a métrica proposta por Silva et al. (2013) é composta por três blocos de construção: distorção estrutural, borramento e medição de complexidade do conteúdo. Além disso, é desenvolvida em duas etapas, uma que introduz as características que quantificam os artefatos e outra é o processo de agregação destas características.

1. Seleção de características da imagem:

Medida de distorções estruturais: O coeficiente de correlação entre os vídeos original e degradado é calculado para medir as distorções estruturais. Inicialmente, cada vista do par estereoscópico é reduzida para a resolução 480 x 270 (resolução original de 960 x 1080). Em seguida, cada quadro do vídeo é dividido em K blocos de tamanho 13 X 13, este processo é realizado tanto para o vídeo original quanto para o vídeo degradado. Para cada bloco 13 x 13 do par estereoscópico é calculada a diferença estrutural D_k como em Eq.(24).

$$D_k = (\sigma_{xy} + C1/(\sigma_x \times \sigma_y + C1) \tag{23}$$

$$\begin{aligned} D_k = \quad & 1/(13\times 13)\sum_{13}^{j=1}(y_{i,j}-\bar{x})(y_{i,j}-\bar{y}) + C1) \\ & /(\sqrt{\tfrac{1}{13\times 13}\sum_{13}^{i=1}\sum_{13}^{j=1}(x_{i,j}-\bar{x})^2} \\ & \times\sqrt{\tfrac{1}{13\times 13}\sum_{13}^{i=1}\sum_{13}^{j=1}(y_{i,j}-\bar{y})^2}) \end{aligned} \tag{24}$$

Onde temos que $x_{i,j}$ e $y_{i,j}$ representam os valores de pixel da $i-$ésima linha e j-ésima coluna do bloco K do quadro do vídeo original e degradado. A constante $C1$ simula a quebra da Lei de Weber para estímulos de baixa intensidade e também para

evitar a divisões por zero. Neste contexto os autores adotam que $C1$ é 25. Uma vez que D_k é calculado para cada bloco de um quadro do vídeo, a distribuição dos valores de D_k no quadro $D = D_k : k = 1...k$ é caracterizado por sua média aparada (dm) e a distorção média dos blocos com maior distorção (dh).

$$dm = \bar{D}_t, D_t = \{d|d \ni D, d > q_c \bigcap d < q_{1-c}\} \tag{25}$$

Nas Eq. (25) e (26) temos q_c e $q_1 c$ representam o c-ésimo e $(1-c)$-ésimo quantil da distribuição D e a constante c tem valor de 2.0.

Como a baixa correlação significa uma alta diferença estrutural, a distorção média dos blocos com maior distorção dh é dada conforme a Eq. (26).

$$dh = \bar{D}_s, D_s = \{d|d \ni D, d < q_c\} \tag{26}$$

Por fim, a representação da distorção estrutural de um quadro de vídeo com um só valor, dm e dh são combinados conforme a Eq. (27).

$$d_{s,n} = 1 - (0.5 \times d_m + 1.5 \times d_h) \tag{27}$$

$$d_S = d_L + d_R \tag{28}$$

Na Eq.(28), d_L e d_R correspondem, respectivamente, a distorção estrutural das vistas esquerda e direita, sendo calculados por sequência. Logo, d_S representa a distorção estrutural percebida do vídeo estereoscópico degradado que é quantificado como a soma das distorções estruturais dos dois pontos de vista.

Medida de borramento assimétrico: O borramento é definido como a perda de magnitude da borda em áreas visualmente significativas da imagem. Para detectar estas áreas em determinado quadro do vídeo original, os autores aplicam filtro de Sobel, definido pela Eq.(29).

$$S_{h,o} = H * X, S_{v,o} = H' * X \tag{29}$$

A Eq.(29), o X representa o quadro original e $S_{h,o}$ e $S_{v,o}$ os mapas de bordas horizontais e verticais. O filtro de Sobel é denotado por H (SILVA et al., 2013). O mapa de magnitude resultante da borda do vídeo original do quadro (S_o) é calculado conforme a Eq.(30).

$$S_o = |S_{h,o}| + |S_v, o| \tag{30}$$

De maneira semelhante, o mapa de magnitude da borda para o vídeo degradado (S_c) é calculado. Em seguida, a diferença entre as magnitudes de borda do vídeo original

e degradado $(\triangle\, e_{i,j})$ é calculado somente para as áreas perceptivelmente importantes da imagem. As áreas mais significantes da imagem são definidas pelas posições dos pixels em que a magnitude da borda é maior que a média do mapa de magnitude da borda da imagem original (S_o). Além disso, o limite no qual a diferença da magnitude da borda é mais perceptível, é considerado como sendo a metade do desvio padrão do mapa de magnitude de borda do vídeo original (σ_{So}).

O desfoque na posição do pixel $(i,j), B_{i,j}$ é calculado conforme segue nas Equações (31) e (32):

$$\triangle e_{i,j} = \begin{cases} S_o(i,j) - S_C(i,j) & S_o(i,j) > \bar{S_o} \\ 0 & S_o(i,j) \leq \bar{S_o} \end{cases} \tag{31}$$

$$B_{i,j} = \begin{cases} \triangle e_{i,j} & \triangle e_{i,j} > \sigma S_o/2 \\ 0 & \triangle e_{i,j} \leq \sigma S_o/2 \end{cases} \tag{32}$$

A soma de $B_{i,j}$ sobre um determinado quadro de vídeo n é definida como o desfoque total, b_n, do quadro. Logo, um único valor é obtido para uma sequência de vídeo, através do cálculo da média de todos os quadros. No entanto, para o vídeo estereoscópico, existirão dois valores por sequência, que correspondem ao desfoque das vistas esquerda e direita (b_L e b_R). Para calcular o desfoque que é percebido no vídeo estereoscópico comprimido (b_L), é considerado o menor desfoque das duas vistas (Eq. (33)).

$$b_s = min(b_L, b_R) \tag{33}$$

Medida de complexidade do conteúdo: O cálculo da complexidade do conteúdo também é considerado na métrica através das medidas de informação perceptiva espacial e temporal (SI - *Spatial Indices* e TI - *Temporal Indice*). Neste contexto, TI e SI são calculados somente para a vista esquerda original do par estéreo. Além disso, são calculados os mapas de disparidade associados a cada uma das sequências (DSI - *Disparity Spatial Indices* e DTI - *Disparity Temporal Indice*), a fim de caracterizar a complexidade da disparidade da cena. Para obter a complexidade geral do conteúdo (C_s) de uma determinada sequência de vídeo, são feitas combinações de candidatos de SI,TI, DSI e DTI.

$$\begin{aligned} \mathrm{C_{S,1}} &= \log_{10}(\mathrm{SI.TI}) \\ C_{S,2} &= \log_{10}(\mathrm{DSI.DTI}) \\ C_{S,3} &= \alpha_1.\log_{10}(\mathrm{SI}) + \alpha_2.\log_{10}(\mathrm{TI}) \\ &\quad + \alpha_3.\log_{10}(\mathrm{DSI}) + \alpha_4.\log_{10}(\mathrm{DTI}) \\ C_{S,4} &= \alpha_1.\log_{10}(\mathrm{SI.TI}) + \alpha_2.\log_{10}(\mathrm{DSI.DTI}) \end{aligned}$$

2. Agregação da medida de distorção: A métrica baseia-se principalmente na distorção estrutural (d_S), por sua forte correlação com os escores subjetivos. Os autores Silva et al. (2013), apresentam um processo de treinamento, onde o primeiro passo é mapear os valores de d_S para uma escala perceptual através da regressão logística de d_S com $DMOS$. O mapeamento logístico de d_S é definido pela Eq.(34) onde D representa a distorção estrutural percebida, e as constantes a_1, a_2 e a_3 tem seus valores encontrados, através da técnica de AM, de Regressão Logística.

$$D = \frac{a_1}{1 + exp(-a_2.(d_s - a_3))} \tag{34}$$

A próxima etapa do processamento do treinamento é minimizar o p_E (Erro de predição), que é uma função da medida de borrão assimétrica (b_S) e da complexidade do conteúdo (c_S). Os autores supõem que a medida de distorção estrutural não é suficiente para prever DMOS com Precisão nessas regiões. E utilizam a compensação seletiva para aumentar a Precisão geral da previsão. Isso pode ser feito truncando b_S para zero nas regiões, conforme a Eq. (35).

$$B = \begin{cases} 0 & D < b_1 U D > b_2 \\ b_s & b_1 \leq D \leq b_2 \end{cases} \tag{35}$$

As constantes b_1 e b_2 na Eq.(35) são definidas no processo de treinamento e validação (SILVA et al., 2013). Assumindo que B e c_S estão linearmente relacionados com p_E, a relação entre essas características é dada conforme Eq.(36).

$$P_E = w_1 + w_2.B + w_3.C_s \tag{36}$$

Logo, é adicionado p_E dos estímulos de teste à medida de distorção estrutural (D) para estimar a medida de qualidade final dada por M, conforme em Eq.(37).

$$M = D + P_E \tag{37}$$

Ao final os autores sumarizam todos os valores com base nos limiares calculados (SILVA et al., 2013). Obtendo assim um único valor para a métrica, conforme a Eq. (38).

$$\begin{aligned} \mathrm{StSD} = {} & 0.7343/1 + \exp(-15.778(d_s - 0.14)) \\ & + 0.2096 + 0.0085.B \\ & - 0.1799 \log_{10}(\mathrm{SI} \times \mathrm{TI}) \\ & + 0.0789 \log_{10}(\mathrm{DSI} \times \mathrm{DTI}) \end{aligned} \tag{38}$$

3.3.3 Human Visual system based 3D (HV3D)

A ideia da métrica proposta por Banitalebi-dehkordi; Pourazad; Nasiopoulos (2016) é simular o HVS ao fundir informações das vistas esquerda e direita para criar uma visão ciclopeana, o estímulo ciclópico diz respeito a maneira que o espectador com visão estéreo percebem o centro do seu campo visual fundindo entre os dois olhos, como se fossem vistos por um olho ciclópico, para que ocorra este estimulo é fundamental que ocorra disparidade binocular (BANITALEBI-DEHKORDI; POURAZAD; NASIOPOULOS, 2016). Além disso, esta métrica leva em consideração a sensibilidade do sistema visual humano ao contraste e a disparidade de ambas as vistas. Uma estratégia de agrupamento é utilizada para o efeito das variações temporais da qualidade de vídeo.

1. Qualidade da visão ciclopeana: Para criar um modelo de visão binocular formando uma visão ciclopeana os autores combinam as áreas correspondentes à ambas as vistas, esquerda e direita. Para isso, a informação de luminância de cada uma das vistas é dividida em blocos $m \times m$. Para cada bloco da vista esquerda é encontrado o bloco mais semelhante na vista direita, esta busca é feita com a utilização de informações do mapa de disparidade de cada vista do par estéreo.

Para gerar o esquema de visão ciclopeana o melhor bloco correspondente em cada pixel deve ser encontrado. Tendo o valor da disparidade por bloco, é possível encontrar as coordenadas do bloco correspondente na outra vista. Com isso, as informações dos melhores blocos de ambas as vistas são fundidas, com a utilização da transformada 3D-DCT (*Discrete Cosine Transform*). Para cada par de blocos correspondentes, são gerados dois blocos $m \times m$, com as informações dos coeficientes DCT dos blocos fundidos. Através da aplicação da máscara de modelagem do FSC aos blocos 3D-DCT, são atribuídos maiores valores às frequências que são mais importantes para o HVS (Eq. 39).

$$XC = C \times X \tag{39}$$

Na Eq. (39), o XC representa o modelo da visão ciclopeana no domínio DCT para o par de blocos correspondentes nas duas vistas. X é coeficiente 3D-DCT de baixa frequência da visão fundida e por fim, C representa o modelo de mascaramento. Uma vez que o modelo de visão ciclopeana é obtido, a sua qualidade é calculada como segue na Eq.(40).

$$Q_c = \left(\sum_{i=1}^{N} \frac{\mathrm{SSIM}(\mathrm{IDCT}(\mathrm{XC_i}), \mathrm{IDCT}(\mathrm{XC'_i}))}{N} \right)^{\beta_1} \tag{40}$$

Onde XC_i, na Eq. (40), representa o modelo de visão ciclopeana para o i-ésimo par de bloco correspondente na vista 3D distorcida e IDCT representa o inverso da 2D-

DCT, N é o total de blocos em cada vista, β_1 um expoente constante e SSIM é o índice de similaridade estrutural. Neste contexto, os valores de SSIM são fornecidos com base em teste subjetivos apresentados em (BANITALEBI-DEHKORDI; POURAZAD; NASIOPOULOS, 2016).

2. Qualidade do mapa de profundidade: O comprimento de um bloco quadrado na tela pode ser totalmente projetado na fóvea ocular e é calculado conforme a Eq. (41).

$$K = 2 \times d \times tan(\alpha) \tag{41}$$

Na Eq.(41), o K representa o comprimento do bloco, d é a distância de visualização adequada da tela e α é a metade do ângulo do olho do observador na maior acuidade visual. A distância adequada de um visualizador a partir da exibição é decidida com base no tamanho da exibição. O intervalo de 2α está entre $0,5$ e 2 graus. A nitidez da visão cai rapidamente para além deste intervalo. O comprimento do bloco (K) pode ser traduzido em unidades de pixel conforme a Eq.(42).

$$k = \frac{h \times K}{H} = \frac{2 \times d \times h \times tan(\alpha)}{H} \tag{42}$$

Dada a Eq.(42), onde k é o comprimento do bloco na tela, dado em pixels, H é a altura da exibição (em [mm]), e a resolução vertical da exibição. A variância da disparidade local é calculada sobre uma área de tamanho do bloco, que pode ser totalmente projetada na fóvea do olho ao observar uma exibição 3D a partir de uma distância de visualização típica. Para o cálculo da variância do mapa de profundidade local do i-ésimo bloco, um bloco externo $k \times k$ é considerado de tal modo que o bloco $m \times m$ está localizado em seu centro, e $\alpha_{d_i}^2$ é definido como em Eq.(43).

$$\sigma_{d_1}^2 = \frac{1}{k \times k - 1} \sum_{j,i=1}^{k} (M_d - R_{j,i})^2 \tag{43}$$

O valor de M_d é a medida dos valores de profundidade de cada bloco $k \times k$ no mapa de profundidade de referência normalizado. Rj representa o valor da profundidade do pixel (j, l) no bloco $k \times k$.

O VIF é utilizado para comparar a qualidade do mapa de profundidade do conteúdo 3D distorcido em relação ao de referência, conforme segue:

$$\begin{gathered} Q_D = (VIF(D, D'))^{\beta_2} \times \\ \left(\sum_{i=1}^{N} \frac{\sigma_{d_1}^2}{N.max(\sigma_{d_j}^2 | j=1,2,...,N)} \right)^{\beta_3} \end{gathered} \tag{44}$$

Na Eq.(44), a variável D é o mapa de profundidade da visualização 3D de referência, e D' é o mapa de profundidade da visualização 3D distorcida, VIF é o índice de

fidelidade da informação visual, β_2 e β_3 são constantes, N é o número total de blocos e $\sigma_{d_1}^2$ é a variância local do bloco i no mapa de profundidade da vista de referência 3D. A última parte da Eq.(44) realiza uma soma sobre as variâncias locais normalizadas. O termo de variância é calculado para cada bloco de acordo com Eq.(43) e normalizado para o valor de variância para todos os blocos. O somatório é então dividido por N, o número total de blocos, para fornecer um valor médio de variância local normalizado no intervalo de [0,1].

3. Agregação das medidas de qualidade: Por fim, após a avaliação de qualidade da visão ciclopeana distorcida e do mapa de profundidade, a forma final da métrica HV3D é definida como em Eq.(45).

$$HV3D = Q_c \times Q_D \tag{45}$$

Como diferentes quadros de um vídeo têm influência diferente no julgamento humano de qualidade, a qualidade geral de uma sequência de vídeo é obtida através de pesos que são atribuídos às pontuações de qualidade de cada quadro, de acordo com sua influência na qualidade geral (BANITALEBI-DEHKORDI; POURAZAD; NASIOPOULOS, 2016).

3.3.4 2D-TO-3D

Nesta métrica, proposta por Fang; Sui; Wang (2019), os autores propõem um modelo de predição para avaliar a qualidade perceptual dos vídeos 3D estereoscópicos degradados assimetricamente. A métrica foi desenvolvida em dois estágios o primeiro os autores avaliam a qualidade percebida dos vídeos de vista única (*single view*) utilizando abordagens de Avaliação de Qualidade de Imagem/Vídeo 2D. No segundo estágio é projetado um modelo baseado na rivalidade binocular, em que são consideradas informações espaciais e temporais dos vídeos para integrar a qualidade perceptual do vídeo 2D de ambas as vistas na Avaliação de Qualidade de Vídeo 3D.

1. Informação Espacial: Os autores utilizam SI para estimar a energia de borda e combiná-la com a estimativa de energia de rivalidade binocular. Inicialmente, filtram o plano de luminância de cada quadro de vídeo de vista única com o operador Scharr e logo calculam a magnitude do gradiente em cada quadro filtrado pelo Scharr. Posteriormente as informações espaciais dos vídeos 2D de vista única são representados calculando a média das informações espaciais de cada quadro. A magnitude do gradiente é dado pela Eq.(46)

$$G = (G_x^2 + G_y^2)^{\frac{1}{2}} \tag{46}$$

Os valores de G_X e G_y, mostrados na Eq.(46), denotam as máscaras de convolução horizontal e vertical com o operador de Scharr (FANG; SUI; WANG, 2019). Logo SI para o i-ésimo quadro compactado a esquerda e à direita são denotados através

da Eq.(47).

$$SI_{i,c,l} = \frac{1}{W \times H} \sum G_{i,c,l}$$
$$SI_{i,c,r} = \frac{1}{W \times H} \sum G_{i,c,r} \tag{47}$$

A Eq.(47), tem que $G_{i,c,l}$ e $G_{i,c,r}$ são as magnitudes dos gradientes dos quadros da imagem degradada da vista esquerda e direita e os somatórios são sobre os mapas de gradiente completos. Sendo W e H a largura e altura do vídeo, respectivamente. Logo é calculado o valor de SI em termos de visualização através da Eq. (48).

$$SI_l = \frac{1}{N} \sum_{i=1}^{N} SI_{i,c,l}$$
$$SI_r = \frac{1}{N} \sum_{i=1}^{N} SI_{i,c,r} \tag{48}$$

Os valores de SI_l e SI_r, dados na Eq.(48), representam o SI da vista esquerda e direita do vídeo. N é o número de quadros do vídeo estereoscópico.

2. Informação Temporal: O TI é calculado através da diferença dos valores dos pixels nos quadros sucessivos em escala de cinza, que representa a diferença do movimento. Os mapas de características da diferença de movimento para esquerda e para a direita são calculados através da Eq.(49)

$$M_{i,c,l}(x,y) = I_{i,c,l}(x,y) - I_{i-1,c,l}(x,y)$$
$$M_{i,c,r}(x,y) = I_{i,c,r}(x,y) - I_{i-1,c,r}(x,y) \tag{49}$$

NA Eq.(49), onde $I_{i,c,l}$ e $I_{i,c,r}$ são os pixels no x-ésima linha e y-ésima coluna do i-ésimo quadro degradado da vista esquerda e direita no tempo. Em seguida, o TI para os i-ésimos quadros degradados são indicados conforme a seguir:

$$TI_{i,c,l} = \frac{1}{W \times H} \sum M_{i,c,l}$$
$$TI_{i,c,r} = \frac{1}{W \times H} \sum M_{i,c,r} \tag{50}$$

O cálculo de TI em termos de visualização é dado pela Eq.(51)

$$TI_l = \frac{1}{N} \sum_{i=2}^{N} TI_{i,c,l}$$
$$TI_r = \frac{1}{N} \sum_{i=2}^{N} TI_{i,c,r} \tag{51}$$

A representação de TI das vistas esquerda e direita é dada respectivamente por

TI_l e TI_r e o número de quadros do vídeo estereoscópico é dado por N.

3. 2D-to-3D predição de qualidade: Depois dos valores de SI e TI serem obtidos para as vistas esquerda e direita, o índice geral de dominância do vídeo é calculado como:

$$\begin{aligned} gl &= \frac{1}{2}\sqrt{SI_l^2 + TI_l^2} \\ gr &= \frac{1}{2}\sqrt{SI_r^2 + TI_r^2} \end{aligned} \tag{52}$$

De acordo com gl e gr, os pesos atribuídos às vistas esquerda e direita são calculados pela Eq.(53).

$$\begin{aligned} W_l &= \frac{g_l^2}{g_l^2 + g_r^2} \\ W_r &= \frac{g_r^2}{g_l^2 + g_r^2} \end{aligned} \tag{53}$$

Por fim, é calculada a média ponderada das vistas esquerda e direita para estimar a qualidade do vídeo 3D, como a Eq.(54).

$$Q^{3D} = W_l \times Q_l^{2D} + W_r \times Q_r^{2D} \tag{54}$$

Os valores de Q_l^{2D} e Q_r^{2D}, na Eq.(54), representam a qualidade de vídeo 2D das vistas esquerda e direita.

Esta métrica, bem como as outras apresentadas nessa Seção, devem ser correlacionadas com os valores de testes subjetivos, pois só assim podem ser ditas apropriadas ou não para avaliar a qualidade de uma imagem ou vídeo. Para isso, a Seção 3.4 apresenta com maiores detalhes a Avaliação Subjetiva de Qualidade de Imagem e Vídeo.

3.4 Avaliação Subjetiva de Qualidade de Imagem e Vídeo

As medidas subjetivas são obtidas através de avaliações envolvendo seres humanos, que normalmente são orientados a visualizar um determinado conjunto de vídeos e atribuir uma nota (escores) para cada um desses vídeos de acordo com sua percepção de qualidade (DARONCO; ROESLER; LIMA, 2008). Esses modelos também servem como referência de desempenho das avaliações dos modelos objetivos (TANJI et al., 2014).

O processo de avaliação subjetiva consiste em uma sequência de atividades, em que um (ou mais) avaliador, observador do vídeo, interpreta e determina uma nota para a qualidade do vídeo apresentado, de acordo com uma escala de valores predefinidos e também com os objetivos da avaliação previamente descritos para os observadores

(DARONCO; ROESLER; LIMA, 2008).

Assim como as Métricas Objetivas de Avaliação de Qualidade, os Métodos Subjetivos de Avaliação de Qualidade são guiados por normas internacionais, que recomendam como devem ser realizadas as etapas do processo de avaliação (DARONCO; ROESLER; LIMA, 2008). Em geral, esses métodos se diferenciam nos seguintes aspectos: quanto à quantidade de vezes em que as sequências de vídeos são apresentadas; quanto à disponibilidade de apresentação apenas dos vídeos degradados; quanto à escala ser contínua ou discreta; e quanto ao tempo de duração que a sequência de vídeo é apresentada (TANJI et al., 2014).

A avaliação subjetiva pode ser representada através do MOS e do DMOS. O MOS é a nota dada através da média aritmética e o desvio padrão calculados através das classificações dadas por cada sujeito para um determinado estímulo. Já o DMOS é a diferença entre as notas atribuídas às imagens/vídeos degradado e o original (FONSECA, 2008).

A recomendação da VQEG (*Video Quality Experts Group*) define o MOS como a Eq.(55) e o DMOS como a Eq.(56) .

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N} \tag{55}$$

Na Eq.(55) o R representa as classificações individuais de cada sujeito e N o número total de sujeitos.

$$\mathrm{DMOS} = \mathrm{MOS(PVS)} - \mathrm{MOS(REF)} + 5 \tag{56}$$

Na Eq.(56), definida pela VQEG, o PVS (*Processed Video Sequences*) é a sequência do vídeo processado e $MOS(REF)$ é o MOS do vídeo de referência. De acordo com as recomendações do VQEG, quanto maior for o valor do DMOS, melhor é a qualidade do vídeo avaliado.

Os procedimentos utilizados em experimentos de Avaliação de Qualidade Subjetiva são descritos nas recomendações BT.500 e BT.500-11 da ITU-T, para serviços de TV, e P.910 da ITU-T, para aplicações multimídia conforme a Tabela 2.

Tabela 2 – Normas ITU para Avaliação de Qualidade de Vídeo.

Norma	Aplicação
ITU-R Rec. BT.500 (Anexo A)	Metodologia para avaliação da qualidade de vídeo em televisores (ITU-R BT.500, 2002).
ITU-T Rec. P.910 (Anexo B)	Metodologia para avaliação subjetiva de vídeo em aplicações multimídia (ITU-T P.910, 2008)
ITU-T J.144	Técnicas para avaliação objetiva de vídeo para televisão a cabo na presença de uma referência (ITU-T J.144, 2004)

Além disso, alguns aspectos importantes das metodologias devem ser levados em consideração, dentre eles a presença ou não de um vídeo de referência, que são definidos da mesma maneira que nas métricas objetivas; o número de exibição de estímulos (um estímulo ou dois); e também a escala de votação, que difere dependendo da metodologia utilizada, como por exemplo, podem ser em escalas contínuas ou discretas (DARONCO; ROESLER; LIMA, 2008). Neste contexto, quanto ao estímulo, temos que:

1. Estímulo único (*Single stimulus*): é exibido apenas o vídeo que está sendo analisado pelo observador no momento, sem a presença de um vídeo de referência.

2. Estímulo duplo (*Double stimulus*): são apresentados dois vídeos, o que está sendo avaliado no momento e o vídeo de referência, ao mesmo tempo. Pode ainda, ser aplicado sem envolver um vídeo de referência, com o objetivo de avaliar dois vídeos degradados.

A seguir, na Seção 3.5 serão apresentados diferentes Métodos Subjetivos de Avaliação de Qualidade de Imagem e Vídeo que utilizam ambos os estímulos.

3.5 Métodos Subjetivos de Avaliação de Qualidade de Imagem e Vídeo

Nesta Seção serão abordados os principais Métodos Subjetivos de Avaliação de Qualidade de Imagem e Vídeo. Estes métodos se distinguem principalmente pela números de estímulos apresentados ao visualizador e a escala de pontuação.

3.5.1 *Double Stimulus Impairment Scale* (DSIS)

Neste método os observadores têm conhecimento da sequência que está sendo apresentada, e cada sequência é mostrada somente uma vez. Inicialmente, é mostrada uma sequência com o vídeo de referência, logo o vídeo degradado. Os observadores classificam as sequências utilizando uma escala discreta com cinco níveis, que varia do muito irritante ao imperceptível (CHEN; WU; ZHANG, 2015; CHIKKERUR et al., 2011).

A ITU-R BT.500 (2002) recomenda que a utilização de uma escala de avaliação de cinco notas conforme a Figura 9. Os avaliadores devem utilizar um formulário que indique claramente a escala, com quadros numerados e um campo para registrar a nota dada.

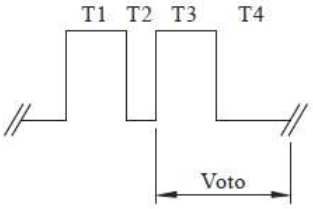
Ao início de cada sessão, deve-se dar explicações aos observadores sobre o tipo de avaliação, a escala de avaliação, a sequência e a temporização (imagem de referência, cinza, imagem de avaliação, período de votação). O intervalo e o tipo de

5		Imperceptível
4		Perceptível, mas não incomoda
3		Incomoda ligeiramente
2		Incomoda
1		Incomoda muito

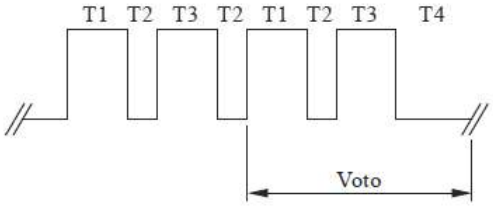
Figura 9 – Exemplo do formulário com a escala de avaliação com cinco notas para DSIS - Adaptado de (ITU-R BT.500, 2002).

degradação que serão avaliados devem ser ilustrados com imagens diferentes das utilizadas nos testes, mas de sensibilidade equivalente. Além disso, os observadores devem ser convidados a basear sua avaliação na impressão geral dada pela imagem e a expressar essas avaliações nos mesmos termos para definir a escala subjetiva.

A ITU-R BT.500 (2002) recomenda ainda, que os observadores sejam orientados a visualizar a imagem durante os períodos $T1$ e $T3$ (Figura 10). Já a votação deve ser autorizada somente durante o $T4$, conforme mostrado na Figura 10.



a) Variante I



b) Variante II

Figura 10 – Modelo da estrutura da apresentação do material de teste do método DSIS. fonte: (ITU-R BT.500, 2002).

A Figura 10, apresenta as fases da apresentação, em que $T1 = 10s$, corresponde a imagem de referência; $T2 = 3s$, ao cinza mediano produzido por um nível de cerca de 200mV; $T3 = 10s$, condição a ser avaliada; e $T4 = 5 - 11s$ corresponde ao cinza mediano. A ITU-R BT.500 (2002) observa que prolongar o período $T1$ e $T3$, além dos 10s, não irá melhorar a capacidade do avaliador para julgar as sequências.

Por fim, na sessão de teste o vídeo de referência e o vídeo degradado devem ser

apresentados em uma sequência pseudo-aleatória e, de preferência, em diferentes sequências para cada sessão. Em qualquer caso, a mesma imagem ou a sequência de teste deve ser apresentada em duas ocasiões sucessivas com os mesmos níveis de degradação, ou com diferentes níveis.

Os intervalos das degradações devem ser escolhidos para que a maioria dos observadores utilizem todas as notas; deve-se obter uma pontuação média total aproximada de 3. Uma sessão não deve durar mais de meia hora, incluindo explicações e preliminares; a sequência de teste deve ser iniciada com várias imagens indicando o intervalo das degradações.

3.5.2 *Double Stimulus Continuous Quality Scale* (DSCQS)

Este método é considerado cíclico uma vez que o observador é solicitado a visualizar um par de imagens, ambas da mesma fonte. Porém, uma é transmitida pelo sistema que está sendo avaliado e a outra diretamente da fonte. Assim como o método DSIS as sessões de teste devem ter duração de até 30 minutos, em que o observador apresenta uma sequência de pares, contendo a referência e o vídeo degradado, como exemplo: referência, vídeo degradado, referência, vídeo degradado, em ordem aleatória, apresentando todas as combinações de pares possíveis (CHIKKERUR et al., 2011). No final das sessões, a pontuação média é calculada para cada condição de teste e para cada imagem de teste.

A apresentação do material de teste é constituída de várias apresentações como: a variante I, possui somente um observador, o avaliador pode alternar livremente os sinais entre A e B para cada apresentação, até que obtenha a medida de qualidade mental associada a cada sinal; a variante II, admite vários observadores em simultâneo, antes de registrar o resultado, o par de condições é mostrado uma ou mais vezes em um período de tempo semelhante, permitindo que o observador adquira a medida mental da qualidade associada aos vídeos. Logo, cada par é novamente mostrado uma ou mais vezes, durante a gravação dos resultados, o número de repetições irá depender da duração das sequências de teste. A ITU recomenda uma sequência de 10s com duas repetições.

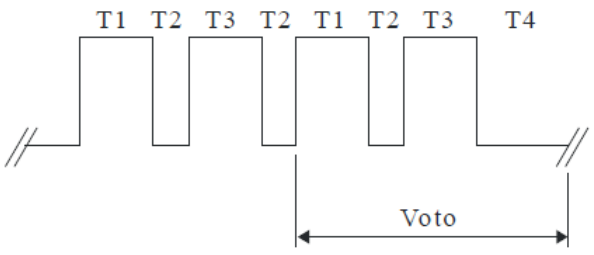


Figura 11 – Modelo da estrutura de apresentação do material de teste do método DSCQS. Fonte: ITU-R BT.500 (2002).

A Figura 11 apresenta a estrutura de apresentação do material proposta pela ITU-R BT.500 (2002), onde os tempos correspondem aos valores: $T1 = 10s$, sequência de teste A; $T2 = 3s$, refere-se ao cinza médio produzido por um nível de vídeo de 200 mV; $T3 = 10s$, é uma sequência de teste B; e $T4 = 5 - 11s$, é o cinza médio.

Sobre a escala de avaliação, esse método exige a avaliação de duas versões de cada imagem de teste. Onde cada par de teste é composto por uma imagem de referência e por uma imagem que pode ou não ser degradada. A imagem sem degradação é incluída como referência, mas os observadores não têm conhecimento de quais são as imagens de referência. Durante o teste, a posição da referência é alterada de maneira pseudo-aleatória.

Os observadores são instruídos a avaliar a qualidade geral da imagem de cada apresentação fazendo uma marca em uma escala vertical, um exemplo da escala pode ser observado na Figura 12.

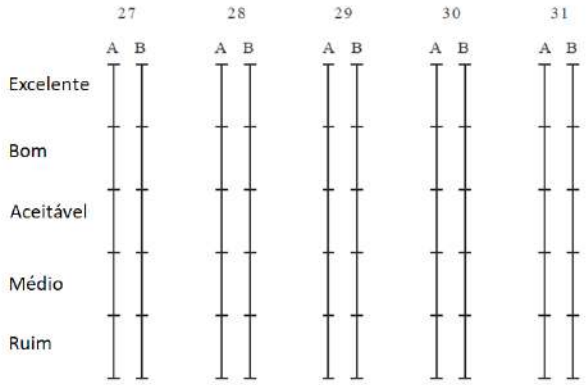


Figura 12 – Exemplo da escala contínua de avaliação de qualidade para o método DSCQS - Adaptado de ITU-R BT.500 (2002).

As escalas verticais devem ser impressas em pares a fim de respeitar a apresentação da sequência de teste. As escalas são contínuas para evitar erros de quantificação, mas são divididas em cinco segmentos como na Figura 12, que correspondem à escala de qualidade normal de cinco notas da ITU-R BT.500 (2002).

A ITU-R BT.500 (2002) recomenda que os resultados obtidos através do método DSCQS não sejam reconhecidos como resultados absolutos, e sim, como a diferença dos resultados entre a referência e o vídeo de teste. Sendo errado associar os resultados a um único termo de descrição de qualidade, mesmo os vindos do próprio protocolo DSCQS, como exemplo: excelente, bom, aceitável.

3.5.3 *Single Stimulus Continuous Quality Scale* (SSCQE)

Esta metodologia é de estímulo único, na qual os observadores visualizam a sequência de teste somente uma vez, sem a referência. Normalmente é de longa duração (20-30 minutos), os observadores são orientados a avaliar instantaneamente

a qualidade percebida em uma escala contínua DSCQS, variando de bom a excelente (CHIKKERUR et al., 2011). A ITU-R BT.500 (2002) recomenda que a avaliação de qualidade contínua seja registrada em um sistema de registro eletrônico conectado a um computador, com a seguinte configuração: o mecanismo deslizante não deve partir de uma posição pré-definida; a distância de deslocamento linear deve ser de 10cm; fixo ou montado em um console; as amostras devem ser registradas duas vezes por segundo. A apresentação dos testes refere-se a realização completa de um teste de Avaliação de Qualidade de Vídeo. Esta pode ser dividida em uma sessão de teste utilizada para atender os requisitos de duração máxima e outra para avaliar a qualidade com todos os pares do segmento do programa / parâmetros de qualidade. No caso do número de pares ser limitado, deve ser realizada a apresentação de teste repetindo a mesma sessão de teste.

Na aplicação do teste simples, um único segmento de programa pode ser usado e apenas um parâmetro de qualidade deve ser considerado. Os observadores devem ter o conhecimento claro de que a distância de deslocamento do mecanismo deslizante do aparelho corresponde à escala de qualidade contínua conforme a Figura 12.

O número de participantes deve ser de pelo menos 15, não "*experts*" e com características recomentadas atualmente na sessão 2.5 da ITU-R BT.500. Os dados devem ser coletados de todas as sessões de teste. Desta forma, é possível obter um único gráfico do índice de qualidade médio em função do tempo, como uma média da avaliação de qualidade de todos os observadores por segmento de programa, parâmetro de qualidade ou sessão de teste completa.

3.5.4 *Simultaneous Double Stimulus for Continuous Evaluation* (SDSCE)

O método de avaliação continua para estímulos duplos simultâneos, avalia a fidelidade, comparando o sinal de vídeo original com o degradado. O comportamento pode ser avaliado através de erros em velocidades de transmissão muito baixas. Os observadores assistem às sequências de vídeo de referência e o degradado ao mesmo tempo, sendo que as duas podem ser apresentadas no mesmo monitor, monitores diferentes ou dois monitores alinhados, desde que as apresentações sejam simultâneas. Os observadores são solicitados a verificar as diferenças entre as duas sequências e julgar a fidelidade das informações de vídeo movendo o mecanismo deslisante de um dispositivo de votação de celular. Quando a fidelidade é perfeita, o cursor deve estar no topo da escala (codificado 100), quando a fidelidade é nula, o cursor deve estar na parte inferior da escala (codificado 0) (COAQUIRA BEGAZO, 2012). Neste caso, os observadores devem ter o conhecimento de qual é a referência sendo convidados a expressar sua opinião, durante todo o tempo que estão observando as sequências. A ITU-R BT.500 (2002) considera que o método de estímulo duplo é especialmente útil para quando não é possível proporcionar estímulos de teste que englobam toda a

qualidade (ITU-R BT.500, 2002).

Um grupo de observadores deve visualizar duas sequências ao mesmo tempo: a referência e a condição de teste. Se as sequências forem do formato de imagem normalizada ou menor, as duas sequências podem ser vistas juntas no mesmo monitor, em outros casos, dois monitores alinhados devem ser usados.

Uma vez realizado o teste, um ou mais conjunto de dados estão disponíveis com todos os votos de cada sessão (S), que representam todo o material de voto do teste de apresentação (TP - *Presentation Test*). A primeira comprovação da validade dos dados deve ser realizada verificando que cada par de segmento vídeo / condições de teste tenha sido apresentado e que um número equivalente de votos foram dados a cada um deles. Existem três maneiras de processar os dados durante a execução do teste: 1. Análise estatística de cada segmento de vídeo separado; 2. Análise estatística de cada condição de teste separado; e 3. Análise estatística global de todos os segmentos vídeo / condições de teste em pares.

A confiabilidade dos observadores pode ser avaliada qualitativamente, verificando seu comportamento quando os pares de referência ou a referência são apresentadas. Neste caso, é esperado que os observadores proporcionem avaliações muito próximas de 100, confirmando assim que compreenderam corretamente sua tarefa e que seus votos não são aleatórios.

3.5.5 *Absolute Categorical Rating* (ACR)

O método de índices por categorias absolutas é um julgamento de categorias em que as sequências de teste são apresentadas uma por vez e se classificam independentemente, em uma escala de categoria, ou seja, estímulo único.

Este método especifica que depois de cada apresentação os observadores devem avaliar a qualidade da sequência apresentada. Se um tempo de votação constante for usado, então o tempo de votação deve ser igual ou inferior a 10s. O tempo de apresentação pode ser reduzido ou aumentado conforme o conteúdo do material de teste.

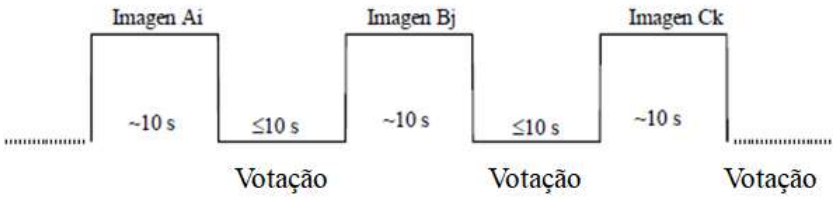


Figura 13 – Modelo da estrutura de apresentação do estímulo no método ACR - Adaptado de ITU-R BT.500 (2002).

A Figura 13 mostra que Ai é a sequência A na condição de teste i; Bj é a sequência B na condição de teste j; Ck corresponde a sequência C em condições de teste K.

A ITU-R BT.500 (2002) recomenda que para avaliar a qualidade global deve-se utilizar a escala de cinco níveis, conforme visto na Figura 14.

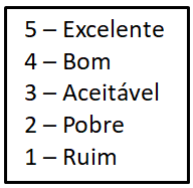


Figura 14 – Exemplo da escala de apreciação do método ACR - Adaptado de ITU-R BT.500 (2002).

Caso seja necessário uma avaliação mais detalhada, uma escala de nove níveis deve ser utilizada. Essas dimensões podem ser úteis para obter mais informações sobre diferentes fatores de qualidade de percepção quando o índice de qualidade global é quase o mesmo para certos sistemas em teste, embora os sistemas sejam claramente percebidos como diferentes.

Para o método ACR, o número necessário de repetições é obtido repetindo as mesmas condições de teste em diferentes momentos do teste.

3.5.6 *Absolute Category Rating with Hidden Reference* (ACR-HR)

As sequências de vídeos degradados são apresentadas uma de cada vez. O método incluí a referência oculta, em que uma versão da sequência do vídeo original é apresentada para cada sequência de vídeo degradado, sem que os observadores tenham conhecimento. A escala de avaliação é semelhante a ACR. Os observadores atribuem uma classificação global, com escala de cinco níveis de “péssimo” a “excelente” (COAQUIRA BEGAZO, 2012).

A ITU-R BT 500 (2002) define o ACR-HR como um julgamento de categorias nas quais as sequências de teste são apresentadas uma de cada vez e são marcadas de forma independente em uma escala de categorias. O procedimento de teste deve englobar uma versão de referência de cada sequência de teste mostrada como qualquer outro vídeo de teste, ou seja, com referência oculta. Durante a análise de dados, um escore de qualidade diferente (DMOS) deve ser computado entre cada sequência de teste e sua referência correspondente (oculta). Assim como para o ACR, o ACR-HR é representado através da Figura 13, que corresponde o mesmo diagrama de tempo de apresentação do ACR. O tempo de votação é constante, sendo igual ou menor que 10s, aumentando ou reduzindo conforme o conteúdo do material do teste.

Na Eq.(57) temos que (DV) é o diferencial de pontuação do usuário, calculado por pessoa e por sequência de vídeo processado (PVS). A referência oculta (REF) é

utilizada para calcular o DV. Onde V refere-se a pontuação ACR do usuário.

$$DV(PVS) = V(PVS) - V(REF) + 5 \tag{57}$$

Conforme a Eq.(57), um DV de 5 indica uma qualidade “excelente” e um DV de 1 uma qualidade “ruim”. No entanto, qualquer DV pode ser superior a 5, para tanto, a ITU-R BT 500 (2002) recomenda a aplicação de uma função de encolhimento de 2 pontos (Eq.58), utilizada para evitar que as pontuações do usuário do ACR-HR influenciem inadequadamente o escore da opinião geral (ITU-R BT.500, 2002).

$$DV\_ENCOLHIMENTO = (7 \times DV)/(2 + DV) \text{ se } DV > 5 \tag{58}$$

Para o método ACR, o número necessário de repetições é obtido repetindo as mesmas condições de teste em diferentes momentos do teste. Já o método ACR-HR só deve ser usado com um vídeo de referência em que um especialista considera a qualidade “boa” ou “excelente” na escala de cinco níveis, a mesma escala do ACR.

3.5.7 *Degradation Category Rating* (DCR)

O índice por categorias de degradação é de estímulo duplo, ou seja, suas sequências de teste são em pares. O primeiro estímulo é a referência, o segundo é a sequência de vídeo degradado. Ambas as sequências podem ser apresentadas em série, uma seguida da outra, ou ainda de forma conjunta, no mesmo monitor. A classificação das sequências é feita separadamente em uma escala contínua de qualidade que varia desde “muito incômodo” a “imperceptível” (COAQUIRA BEGAZO, 2012).

O diagrama de tempo da apresentação do estímulo é ilustrado na Figura 15 proposta pela ITU-R BT.500 (2002). Se um tempo de votação constante for usado, o tempo de votação deve ser igual ou inferior a 10s. O tempo de apresentação pode ser reduzido ou aumentado de acordo com o conteúdo do material de teste.

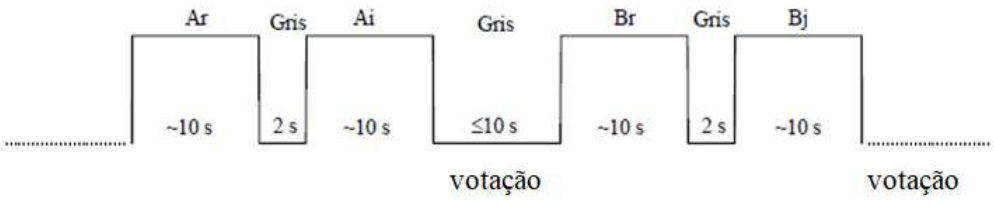


Figura 15 – Modelo da estrutura de apresentação do estímulo em DCR - Adaptado de ITU-R BT.500 (2002).

Na Figura 15 temos que Ai é a sequência A em condição de prova i; Ar, Br são sequências de A e B, respectivamente em formato de fonte de referência; Bj é a sequência de B em condição de prova j.

Neste método os observadores avaliam as degradações do segundo estímulo em relação a referência. A ITU-R BT.500 (2002) recomenda a utilização de uma escala de cinco níveis, como segue na Figura 16.

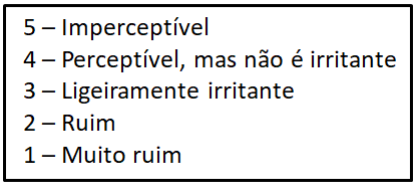


Figura 16 – Exemplo da escala de apreciação do método DCR - Adaptado de ITU-R BT.500 (2002).

Deve-se repetir as mesmas condições de prova em diferentes momentos do teste até que se obtenha o número necessário de iterações.

3.5.8 *Pair Comparison* (PC)

O método de comparação por pares, é utilizado para comparar degradações produzidas por dois sistemas diferentes, sobre o mesmo sinal de vídeo original, ou seja, são avaliadas as sequências de teste de uma mesma cena, em condições de degradação diferentes. Todos os pares das sequências devem ser exibidos nas duas ordens possíveis, exemplo: AB e BA. Ao final da exibição de cada par, os avaliadores indicam sua preferência em relação a uma das sequências apresentadas (REGIS, 2013).

O diagrama de tempo da apresentação do estímulo é ilustrado na Figura 17. Se um tempo de votação constante for usado, então o tempo de votação deve ser igual ou inferior a 10s. O tempo de apresentação deve ser de cerca de 10s e pode ser reduzido ou aumentado de acordo com o conteúdo do material de teste.

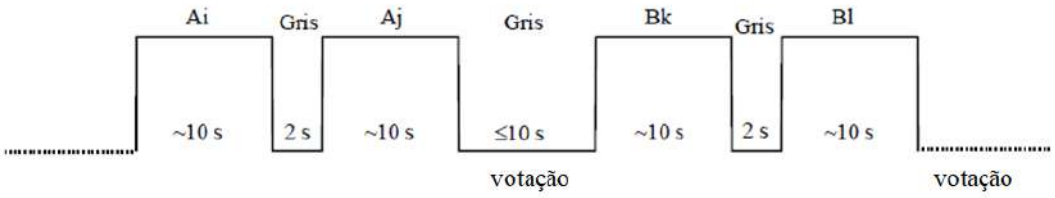


Figura 17 – Modelo da estrutura de apresentação do estímulo em DCR - Adaptado de ITU-R BT.500 (2002).

Ao usar resoluções reduzidas, pode ser conveniente exibir cada par de sequências simultaneamente no mesmo monitor. Em geral, no método PC não é necessário considerar o número de repetições, já que o próprio método implica na apresentação repetida das mesmas condições, embora em pares diferentes. Uma variação do método PC utiliza uma escala para cada categoria, a fim de apreciar em maior grau as diferenças entre pares de sequências.

3.6 Considerações Finais

Este Capítulo apresentou os principais temas sobre Avaliação de Qualidade de Imagem e Vídeo, trazendo os conceitos da Avaliação Objetiva de Qualidade de Imagem e Vídeo. Em seguida, algumas métricas desenvolvidas para 2D e 3D, já estabelecidas na literatura, foram apresentadas com maiores detalhes. Além disso, esse Capítulo também apresentou a Avaliação Subjetiva de Qualidade de Imagem e Vídeo, discorrendo-se sobre os seus principais conceitos. Também foram apresentados os principais Métodos Subjetivos de Avaliação de Qualidade de Imagem e Vídeo. A compreensão desses conceitos, bem como o entendimento de algumas Métricas Objetivas de Avaliação e Métodos Subjetivos são importantes para um melhor entendimento do desenvolvimento desse trabalho e dos próximos capítulos.

4 TRABALHOS RELACIONADOS

Alguns estudos têm proposto novas métricas objetivas para avaliar a qualidade percebida das imagens 3D. Os autores utilizam diferentes abordagens para melhor representar o HVS considerando uma diversidade de modelos de degradação. Além das imagens, os vídeos ocupam um lugar de destaque no dia-a-dia das pessoas, aplicados a diferentes finalidades, como para o entretenimento, informação e aprimoramento profissional (SOUZA BARBIERI; GOULARTE, 2020). Por isso, existem diversos estudos que tratam de Avaliação de Qualidade de Vídeo. Os trabalhos apresentados propõem Métricas de Avaliação de Qualidade de Imagens e Vídeos de Referência Completa, sendo alguns para 2D e outros para 3D, porém todos incluem alguma técnica baseada em AM.

A métrica proposta por Silva et al. (2013) a *Stereoscopic Structural Distortion* (StSD), é uma métrica de avaliação de qualidade para vídeos estereoscópicos e de referência completa, melhor detalhada na Seção 3.3.2. O modelo leva em consideração informações estruturais da imagem, borramento e informações complexas como a informação temporal (*Temporal Information* - TI) e a informação espacial (*Spatial Information* - SI). Os resultados dos testes subjetivos deste estudo são utilizados para treinar e validar valores que serão adicionados à etapa de medida de complexidade do conteúdo . Na primeira etapa do treinamento é realizado o mapeamento dos valores de distorção estrutural por meio de regressão logística com os valores de DMOS. Já na segunda etapa, o treinamento é utilizado para minimizar erros de previsão. O principal objetivo da etapa de treinamento, de acordo com os autores, é estimar valores para constantes utilizadas na regressão logística. Além disso, após a análise dos efeitos das diferentes combinações dos recursos treinados, a melhor representação da complexidade do conteúdo é selecionada. Esta seleção é chamada de validação. Após a validação, os valores são combinados gerando um único índice. Os autores, usam métricas de medida de correlação para comparar a performance desta métrica em relação as outras. As médidas de correlação utilizadas são: *Pearsons linear correlation coefficient* - CC , *Spearman rank order correlation coefficient* - ROCC, *Average Absolute Error* - AAE, RMSE e *Outlier Ratio* - OR e as métricas comparadas são a

SSIM-Ddl, CSVQ, PHVS-3D, BEQM, NRIM. Nesta análise de correlação os autores puderam observar e afirmar que sua métrica apresentou os melhores resultados para todas as métricas, indicando que ela tem uma boa capacidade de avaliar a qualidade percebida pelo HVS (SILVA, 2013).

Em Narwaria; Lin (2011), os autores analisam um estudo com Aprendizado de Máquina para avaliação de qualidade baseadas em decomposição de valor singular (*Singular Value Decomposition - SVD*). Os autores apresentam um trabalho em dois estágios, seguidos por uma análise aprofundada do SVD para avaliação de qualidade visual. Os valores singulares e vetores formam as características que foram selecionadas para VQA. Logo, utilizam o AM para o processo de agrupamento de recursos. Com a finalidade de resolver as limitações das técnicas de agrupamento, como a média, soma simples e soma de Minkowski. Os autores defende o uso de AM para o agrupamento dos recursos por ser mais sistemático e orientado por dados.

Os experimentos sugerem que o método proposto supera os oito regimes existentes. A validação é realizada com dez bancos de dados disponíveis publicamente (oito para imagens com um total de 4042 imagens de teste e duas para vídeo com um total de 228 vídeos).

Por fim, para o agrupamento, que resulta em um índice único, os autores propõem a utilização da técnica de AM para a fusão dos recursos. Onde, os parâmetros dos modelos relacionados (pesos) são estimados por meio de treinamento a partir dos dados disponíveis. Para mostrar que a métrica proposta pelos autores apresentam bons resultados, os valores foram comparados utilizando métricas de medidas de correlação, como Correlação de Pearson (precisão), Coeficiente de Spearman (Monotonicidade) e RMSE. Os autores comparam a Métrica proposta por eles Qvector em Narwaria; Lin (2011), com outras 8, que são: PSNR, SSIM, Narwaria; Lin (2010), MSVD, VIF, IFC, VSNR e Q, diante os cálculos de correlação os autores concluem que a métrica proposta por eles, apresenta os melhores resultados em termos gerais.

Já Charrier; Lézoray; Lebrun (2012), propõem uma Medida de Qualidade de Imagem Baseada em Aprendizado de Máquina (*Machine Learning-based Image Quality Measure - MLIQM)* que inicialmente classifica a qualidade utilizando a técnica *Support Vector Machine* (SVM). Para avaliar a qualidade das imagens, um vetor de recurso contendo as características visuais que descrevem o conteúdo das imagens é construído. Além disso, mostram que para o processo de classificação, dois conjuntos distintos foram gerados a partir dos bancos de dados, o conjunto para treino e para teste. Com cinclo classes de qualidade, baseadas na escala de 5 níveis de votação dos testes subjetivos conforme sugerido pela ITU ((CHARRIER; LÉZORAY; LEBRUN, 2012). Os autores apresentam resultados comparados, em termos de métricas de correlação (Correlação de Pearson, Coeficiente de Spearman e *Kendal*), com os algoritmos: MSSSIM, VIF, PSNR e VSNR. Os autores concluem que o LMIQM apresenta

os melhores resultados e produz uma significativa melhoria dos coeficientes de correlação com os testes subjetivos (CHARRIER; LÉZORAY; LEBRUN, 2012).

A VMAF (*Video Multi-Method Assessment Fusion*), que é uma Métrica de Avaliação de Qualidade de Vídeo de Referência Completa desenvolvida pela Netflix sendo voltada principalmente para vídeos de *streaming*.

Esta métrica, de acordo com Rassool (2017), foi desenvolvida para se correlacionar fortemente com as pontuações subjetivas do MOS. Utilizando técnicas de AM, para o treinamento do modelo de estimativa de qualidade foram usadas amostras de valores MOS. A VMAF busca aproximar a percepção humana da qualidade do vídeo. Para isso, concentra-se na degradação da qualidade devido à compactação e ao redimensionamento. A métrica estima a pontuação de qualidade percebida computando pontuações de vários algoritmos de avaliação de qualidade e agrupando estes valores usando uma máquina de vetor de suporte (SVM). Três métricas de fidelidade de imagem e um sinal temporal foram escolhidos como recursos para o SVM: 1. SNR anti-ruído, 2. Medida de Perda Detalhada, 3. VIF (*Visual Information Fidelity*). Como resultados os autores sugerem que a métrica VMAF apresenta uma forte correlação entre o MOS e a sua pontuação objetiva, com um valor de corrrelação de 0.948, sendo considerado um bom preditor (RASSOOL, 2017).

A Tabela 3, mostra quais trabalhos utilizam técnicas de AM e o tipo de visualização (2D ou 3D). Podemos observar que a maioria das métricas que utilizam técnicas de AM são desenvolvidas para imagens e vídeos 2D.

Tabela 3 – Métricas Objetivas de Avaliação de Qualidade, que utilizam técnicas de AM.

Métrica	**Visualização**	**Técnica de AM**
StSD (SILVA et al., 2013)	3D	Regressão Logística
Narwaria; Lin (2011)	2D	SVR
Charrier; Lézoray; Lebrun (2012)	2D	SVM
VMAF (RASSOOL, 2017)	2D	SVM

As métricas de avaliação de qualidade voltadas para 3D são mais complexas, uma vez que envolvem características da percepção humana menos compreendidas. Nota-se que alguns autores buscam explorar diferentes abordagens para prever a percepção humana frente a imagens estéreo.

O trabalho de Silva et al. (2013) é voltado para vídeos 3D, que utiliza características especificas da estereoscopia, e envolve Aprendizado de Máquina. Já os outros trabalhos como de Narwaria; Lin (2011), Charrier; Lézoray; Lebrun (2012) Rassool (2017) tratam de métricas para imagens/vídeos 2D utilizando técnicas de AM. O trabalho descrito em Silva et al. (2013) apresenta uma Métrica Objetiva de Avaliação de Qualidade para vídeos 3D de Referência Completa com a utilização do AM. No entanto, os autores não enfatizam o uso de AM em seu trabalho, deixando claro que não é o ponto central da métrica. O modelo, no entanto, é baseado em três características: distorção

estrutural, borramento e medida da complexidade do conteúdo, a qual se baseia no índice de informação temporal e espacial do vídeo. Essas informações permitem a obtenção dos dados de disparidade da cena, adicionando ao modelo características de estereoscopia.

Observa-se que existe amplo espaço para a pesquisa voltada à aplicação de aprendizado de máquina para 3D-IQA e 3D-VQA, especialmente considerando técnicas de baixa complexidade como aquelas baseadas em árvore de decisão. Já que a maioria dos trabalhos envolvem técnicas mais complexas de AM como SVM, SVR, no entanto, nenhum deles avalia o potencial das AD's. Além disso, nem todos estes trabalhos são específicos para imagens e vídeos 3D e de Referência Completa. A principal relação encontrada nestes trabalhos com esta tese é a utilização de técnicas de AM na utilização de Avaliação de Qualidade de Referência Completa.

De modo geral, estes trabalhos são métricas que geram um único índice e, por isso, demais comparações com nosso trabalho não são viáveis, já que buscamos avaliar a capacidade de predição de algumas técnicas baseadas em AM. Além disso, estas métricas não apresentam detalhes aprofundados sobre o processo de treinamento e teste das técnicas. No entanto, nos auxiliaram na compreensão da utilização das técnicas de AM para VQA e IQA.

4.1 Considerações Finais do Capítulo

Neste Capítulo foram apresentados alguns trabalhos considerados importantes para o desenvolvimento desta tese para uma melhor compreensão dos próximos Capítulos. Os trabalhos descritos envolvem Avaliação de Qualidade de Imagem e Vídeo 2D e 3D. Embora nem todos os trabalhos tratem especificamente de métricas objetivas 3D, nos interessa entender o processo de desenvolvimento durante a utilização de AM para agregação de VQA ou IQA.

5 METODOLOGIA

Este Capítulo apresenta a metodologia desenvolvida para a realização deste trabalho, que se baseia nas etapas para a aplicação de técnicas de Aprendizado de Máquina, conforme apresentada na Figura 18.

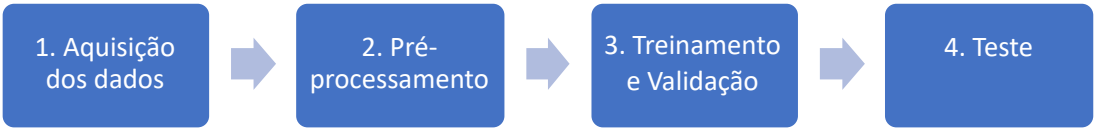


Figura 18 – Fluxograma dos processos baseado em ML utilizado para o desenvolvimento deste trabalho.

A primeira etapa trata da aquisição dos dados, discorrendo desde a seleção da base dos testes subjetivos com imagens e vídeos 3D até a extração das características das imagens selecionadas. Já a segunda etapa é o pré-processamento que trata dos principais passos realizados para a adequação dos dados para que pudessem ser utilizados como entrada do sistema de AM. Em seguida, o passo três apresenta detalhes sobre o treinamento dos modelos e por fim o passo 4 apresenta os resultados dos testes dos modelos treinados. Estas etapas serão discutidas com maiores detalhes nas seções a seguir.

5.1 Aquisição dos Dados

Esta etapa foi dividida em duas seções, a primeira Seção 5.1.1 trata sobre a base dos dados utilizada. Logo, a segunda Seção 5.1.2 aborda a etapa de extração das características de imagens e vídeos.

5.1.1 Base de Dados

Na etapa de aquisição dos dados, o primeiro passo foi a seleção do banco de dados de avaliação de qualidade subjetiva para vídeos 3D estereoscópico que foi utilizado. De maneira geral, a base de dados deve conter os vídeos/imagens de referência e também degradados, com seus valores dos testes subjetivos. Existem diversos banco de dados de testes subjetivos voltados para vídeos 2D. No entanto, quando se trata

especificamente de vídeos 3D estereoscópicos, os banco de dados são limitados, alguns são citados na literatura como em Silva et al. (2013), que apresenta a StSD 3D Video Database e Live 3D Video Database abordada por Chen; Kwon; Bovik (2012), não foram encontrados em nossas buscas. As bases de dados que encontram-se disponíveis são: NAMA3DS1-COSPAD1, MMSPG 3D, 3DVCL@FER Video Database e a Waterloo IVC 3D Video Quality Database. Algumas das características destas bases de dados serão apresentadas na Tabela 4.

Tabela 4 – Banco de dados de Avaliação de Qualidade subjetiva de vídeos Estereoscópicos e suas principais características - Adaptado de Wang; Wang; Wang (2017).

Base de Dados	**Característica**
MMSPG 3D Video Quality Assessment Database	Diferentes distâncias entre câmeras, sendo 6 vídeos com 5 distâncias diferentes, totalizando 30 sequências de vídeos estereoscópicos (GOLDMANN; DE SIMONE; EBRAHIMI, 2010).
NAMA3DS1-COSPADI	Compressão H.264 e JPEG2000, as vistas esquerda e direita foram codificadas separadamente com os mesmos parâmetros (processamento simétrico) totalizando 110 sequências (URVOY et al., 2012).
3DVCL@FER Video Database	Compressão H.264 e JPEG2000, distorções geométricas, taxa de quadros, perda de pacotes e quadros congelados. Ao total são 22 degradações diferentes apresentando simetria e assimetria e 8 sequências de vídeos, totalizando 176 sequências (DUMIĆ et al., 2017).
Waterloo IVC 3D Video Quality Database	Diferentes níveis de filtragem passa-baixa, compressão HEVC, variações de QP e distorções simétricas e assimétricas. Codificação de resolução mista, *pré-downsampling* e filtragem passa baixa gaussiana. A base de dados contém 10 vídeos originais e suas diferentes degradações, totalizando 704 sequências estereoscópicas (WANG; WANG; WANG, 2017).

No entanto, para a escolha da base de dados utilizadas neste trabalho consideramos alguns fatores como a disponibilidade de degradações simétricas e assimétricas e a quantidade de sequências disponíveis, dentro dessas condições optamos por utilizar a base de dados *Waterloo IVC 3D Video Quality Database* e também o banco de imagens, *Waterloo IVC 3D Video Quality Database*, ambos disponibilizam os testes subjetivos e imagens e vídeos de referência. Além disso, também fornecem documentos que apresentam detalhes sobre suas bases de dados e os testes subjetivos, que serão melhores detalhadas a seguir.

***Waterloo IVC 3D Image Quality Database*:** O banco de imagens de *Waterloo IVC 3D* possui duas fases, a Fase I (Figura 19), que foi criada a partir de 6 pares de

imagens estereoscópicas que são: *Laundry*, *Moebius*, *Dolls*, *Reindeer*, *Art* e *Book* e a Fase II (Figura 20), que é composta por 10 imagens estereoscópicas sendo elas: *CraftLoom*, *Dancer*, *Hall*, *Laboratory*, *OldTownCar*, *Persons*, *Soccer*, *Tree*, *Barrier* e *Umbrela*. Tanto a Fase I como a II contém três tipos de degradações que são: contaminação de ruído gaussiano branco aditivo com variância na faixa de [0.10-0.53], desfoque gaussiano com variância entre [2-20] e compressão JPEG com parâmetro de qualidade entre [3-10], cada uma delas com quatro níveis de distorções que foram aplicadas as imagens de vista única, empregadas para gerar pares estereoscópicos distorcidos simetricamente e assimetricamente. Ao todo a Fase I contém 78 imagens com vista única e 330 imagens estereoscópicas, já a Fase II possui 130 imagens de vista única e 460 pares estereoscópicos (WANG et al., 2015).

(a) (b) (c)

(d) (e) (f)

Figura 19 – Imagens estereoscópicas da Fase I disponibilizadas pela base de dados *Waterloo IVC 3D Image Quality Database* conforme descrito em Wang; Wang; Wang (2017). As imagens são: (a) *Laundry*, (b) *Moebius*, (c) *Dolls*, (d) *Reindeer*, (e) *Art*, (f) *Book*.

***Waterloo IVC 3D Video Quality Database*:** A base de dados *Waterloo IVC 3D Video Quality Database* foi desenvolvida em duas etapas, Fase I e Fase II. A primeira fase contém 4 vídeos 3D multi-vista, os quais são: *Balloons*, *Book*, *Kendo* e *Lovebird* (Figura 21), sequências de teste HEVC 3D comumente usadas, os vídeos apresentam uma resolução de 1024 x 768, com duração de 10s e taxa de 30.00 quadros por segundo. Apenas o vídeo *Book* apresenta uma duração de 6s e 16.00 quadros por segundo. Já a segunda fase contém 6 sequências 3D que são: *Barrier*, *Craft*, *Laboratory*, *Soccer*,*Tree* e *Dancer* (Figura 21). Os vídeos desta fase têm a resolução de 1920 x 1080, com uma duração de 10s e 30.00 quadros por segundo. Ambas as fases apresentam valores de testes de qualidade subjetiva e apresentam degradações que incluem compressão simétrica e assimétrica. Os vídeos correspondentes a Fase I incluem vídeos 3D obtidos a partir da codificação com quantização simétrica e assimétrica e diferentes níveis de filtragem passa-baixa. Cada vídeo de visualização única

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

(e) (f) (g) (h)

(i) (j)

Figura 20 – Imagens estereoscópicas da Fase II disponibilizadas pela base de dados *Waterloo IVC 3D Image Quality Database* conforme descrito em Wang; Wang; Wang (2017), são: (a) *CraftLoom*, (b) *Dancer*, (c) *Hall*, (d) *Laboratory*, (e) *OldTownCar*, (f) *Persons*, (g) *Soccer*, (h) *Tree*, (i) *Barrier*, (j) *Umbrella*.

foi compactado usando o codificador HEVC com cinco níveis de Parâmetro de Quantização (QP) de domínio de transformação com QP's= {25,35,40,45,50}. Os vídeos foram compactados de forma simétrica e assimétrica, ou seja, para causar assimetria as vistas foram compactadas com diferentes QP's. Além disso, para cada combinação de QP, foi aplicado o filtro passa-baixa gaussiano (*Gaussian low-pass filtering* - GLPF) com quatro níveis diferentes sendo $\sigma = \{0, 3.5, 7.5, 11.5\}$, que foram aplicados às vistas com QP's mais altos (com baixa qualidade). Ao todo são 176 sequências 3D no banco de dados, referentes a Fase I.

A segunda fase dispõe de 6 vídeos com as mesmas condições iniciais da Fase I, no entanto adota três condições de pré processamento: *pre-downsampling* por 2.4 e um pré processamento usando o GLPF com $\sigma = 2.5$, que foram aplicados a cada vídeo com vista única. Além disso, os níveis de QP são diferentes dos utilizados na Fase I, para Fase II os QP's são: {25,30,35,40,45}. Para os QP's mais altos foi aplicado o GLPF com $\sigma = 2.5, 5.5$. No total, a Fase II conta com 528 sequências de vídeos, que somados à Fase I, dispõem de um total de 704 sequências de vídeos estereoscópicos derivados de 10 vídeos estereoscópicos originais, ou seja, que não sofreram nenhum tipo de degradação.

A base de dados fornece os valores dos testes de qualidade subjetiva dos vídeos estereoscópicos. Ao total os testes contaram com 57 participantes. Os autores optaram por utilizar um método de estímulo único, a fim de minimizar as interrupções na experiência da visualização 3D no caso de métodos com estímulo duplo. Para

Figura 21 – Imagens dos vídeos disponibilizados pela base de dados *Waterloo IVC 3D Video Quality Database* conforme descrito em Wang; Wang; Wang (2017), que são: (a) *Balloons*, (b) *Book*, (c) *Kendo*, (d) *Barrier*, (e) *Craft*, (f) *Lovebird*, (g) *Laboratory*, (h) *Dancer*, (i) *Tree*, (j) *Soccer*.

isso adotaram um protocolo de um só estímulo com escala categórica numérica de 11 graus (*Single Stimulus Numerical Categorical scale* - SSNCS) disposta na ITU-R BT.1082 (R-REP BT.1082-1, 1990; ITU-R BT.500, 2002). Neste caso, os sujeitos foram instruídos a dar pontuações altas (próxima a 10 pontos) para os vídeos originais (alta qualidade), para os vídeos com degradações moderadas as pontuações devem ser na faixa intermediaria e para vídeos com fortes degradações os valores das pontuações devem ser baixos (próximos de zero). Logo os valores MOS de cada vídeo foram redimensionados para preencher um intervalo de 1 a 100 que foram computados após a remoção de valores discrepantes, esses valores foram denominados por Wang; Wang; Wang (2017) de MOS-3DVQ.

5.1.2 Extração de Características

A extração das características dos vídeos estereoscópicos tem quatro vídeos como entrada, sendo estes os pares de vista original e degradada. A Figura 22, ilustra a maneira como o extrator trata os vídeos de entrada, o conjunto VU (Vista Única) adquire as informações que correspondem a cada vista separadamente, informações como média, desvio padrão e variância. Já o Conjunto VE (Vista Esquerda) e o Conjunto VD (Vista Direita) e tratam de obter os valores das métricas utilizadas para imagens e vídeos 2D, como PSNR, SSIM, SAD e MSE. Essas métricas envolvem especificamente informações de cada vista considerando o vídeo original e o vídeo degradado. O Conjunto PV (Pares de Vistas) engloba características que envolvem tanto os pares de vistas originais quanto os das vistas degradadas. Alguns desses valores são adquiridos através do cálculo de métricas desenvolvidas para imagens/vídeos 2D e adaptadas para imagens 3D, calculando o valor para cada vista e depois realizando o agrupamento desses valores através da média entre eles, obtendo um único escore. É o caso das métricas denominadas PSNR3D, SSIM3D, SAD3D e MSE3D. O Conjunto PV (Pares de Vistas) também engloba características originalmente concebidas para imagens e vídeos 3D, como no caso das diversas características derivadas da métrica StSD.

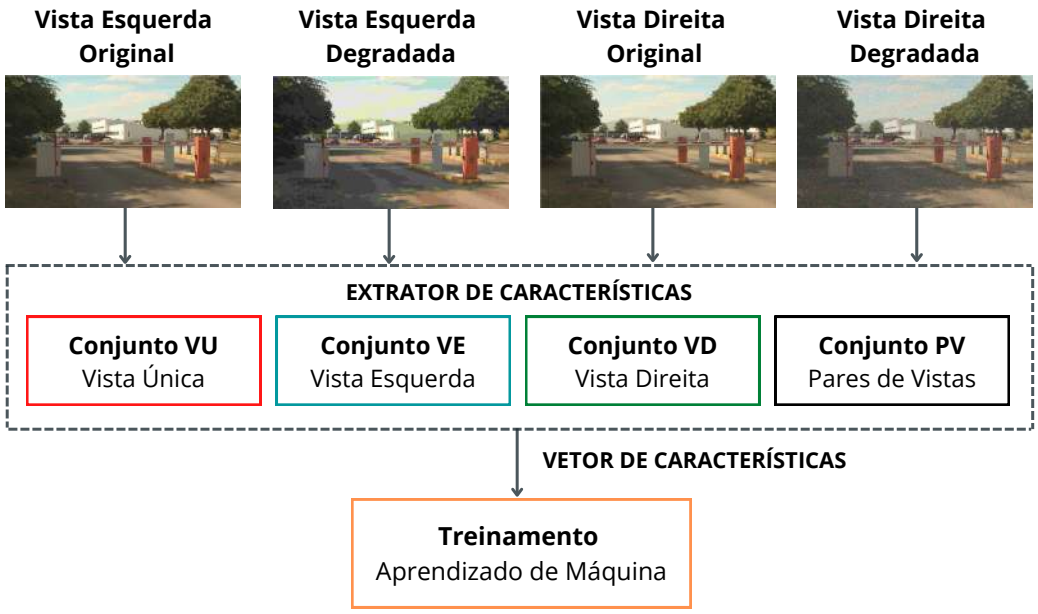


Figura 22 – Modelo conceitual da extração de características das imagens 3D.

A Tabela 5 sumariza 22 características, extraídas das imagens e vídeos, porém, no total são 44 características, já que são considerados os pares de vistas, incluindo o par original e o degradado, tendo então como entrada 4 vistas.

Os valores de cada característica são armazenados na forma de vetores correspondendo a sua imagem e rotulados com o valor do MOS, fornecidos pela base de

Tabela 5 – Características extraídas dos vídeos estereoscópicos. VU, VD, VE e PV são conjuntos descritos na Figura 22. Já VU/E e VU/D referem-se respectivamente ao Conjunto VU somente para vista esquerda (E) e direita (D).

Característica	Descrição	Vistas
Média	média das amostras de luminância	VU
Desvio Padrão	desvio padrão das amostras de luminância	VU
Variância	variância das amostras de luminância	VU
Contraste	Eq. 13	VU
PSNR	Eq. 10	VE,VD
PSNR3D	Eq. 10	PV
MSE	Eq. 9	VE,VD
MSE3D	Eq. 9	PV
SSIM	Eq. 11	VE,VD
SSIM3D	Eq. 11	PV
SAD	Eq. 8	VE,VD
SAD3D	Eq. 8	PV
StSD	Eq. 38	PV
STNx	média de Dk - Eq. 24	VE,VD
STMain	Eq.34	VE,VD
NS_m	dm da métrica StSD	VE,VD
NS_d	média de dh	VE,VD
ND_{sim}	1-NS_m+1.5*NS_d	VE,VD
$StSd_{blr}$	Eq.33	VE,VD
$StSd_{sim}$	Eq.24	VE,VD
SC_{med}	média do (S_c) da métrica StSD	VD
SC_{desv}	desvio padrão do (S_c) da métrica StSD	VD

imagens 3D. Estes vetores serviram como instâncias de entrada para o treinamento dos classificadores baseados em AM e os valores de MOS foram mapeados para valores menores, esses valores são usados como as classes de cada instância. Estas alterações serão melhores detalhadas na Seção 5.2 que trata do Pré-processamento dos dados.

5.2 Pré-processamento

Foram aplicadas diferentes técnicas de pré-processamento, a fim de deixar os dados obtidos na etapa de extração de características na melhor forma possível para serem utilizados no treinamento e validação dos modelos.

O pré-processamento constituiu nas seguintes etapas:

1. **Limpeza dos dados:** Nesta etapa foram removidos os dados considerados como irrelevantes para o processo. Estes são dados como colunas com os números (identificação) dos quadros, rótulos utilizados no processo de extração de características.

2. **Transformação dos dados:** Os valores de MOS fornecidos pela base dados variam em um intervalo de 1 a 100. Para deixar em conformidade com a ITU-R BT.500 (2002) que recomenda utilizar uma escala de cinco níveis (5 - Excelente; 4 - Bom; 3 - Aceitável; 2 - Pobre e 1 - Ruim). Este novo intervalo foi usado para definir o conjunto de 5 classes. Assim, para fins de comparações estendemos para 10 e 25 classes. Essas transformações foram criadas a partir da Eq.(59).

$$ClasseX = ARREDONDAR.PARA.CIMA(MOS * X/100; 0) \quad (59)$$

   Na Eq.(59), fui utilizado o arredondamento para cima, o MOS representa o valor do MOS original fornecido pela base de dados e X representa a classe (5,10 e 25) a ser calculada. Logo essas classes foram adicionadas aos arquivos.

3. **Separação dos dados:** Os dados foram divididos em dois conjuntos, sendo 80% para o treinamento e validação (usando validação cruzada) e 20% para o teste. Esta separação foi realizada, para impedir que o sistema de AM tivesse conhecimento sobre os vídeos e imagens utilizadas para teste.

4. **Adequação dos dados:** Os dados foram adequados para o formato *arff* (*Attribute Relation File Format*), utilizado no *software* Weka[1], ferramenta na qual os modelos foram treinados e testados. Este formato segue uma estrutura onde deve-se indicar os atributos, que são os vetores de características das imagens ou vídeos, o atributo classe, que indica a classe a ser utilizada. Essas classes também são adicionadas ao final dos vetores de características. A Figura 23 apresenta o exemplo da estrutura de um modelo arff.

5.3 Treinamento e Validação

A Figura 24 representa o esquema do AM supervisionado, onde é inicialmente apresentado o Treinamento do modelo, que tem como entrada os vetores de características das imagens ou vídeos combinados com os rótulos, neste caso os valores de MOS. Essas informações são usadas no treinamento de diferentes algoritmos com diferentes configurações para gerarem modelos preditivos. Os modelos gerados são posteriormente testados com dados nunca vistos pelo sistema (dados novos) gerando então os resultados que serão analisados no Capítulo 6.

A fim de responder às questões de pesquisa deste trabalho foram utilizados diferentes dados de entrada para os modelos treinados, que serão melhores descritos nas próximas seções.

[1] https://www.cs.waikato.ac.nz/ml/weka/index.html

```
@relation teste5C

@attribute MediaEsqOrg numeric
@attribute MediaEsqDeg numeric
@attribute MediaDirOrg numeric
@attribute MediaDirDeg numeric


@attribute class{1,2,3,4,5}

@data

95.073,95.073,97.531,97.531,5
95.073,95.669,97.531,98.088,4
95.073,96.746,97.531,98.997,3
95.073,98.88,97.531,101.04,,2
95.073,102.02,97.531,103.95,2
95.073,94.708,97.531,97.153,4
```

Figura 23 – Exemplo de dados estruturados no formato *arff*, que serve de entrada como arquivo para a ferramenta Weka.

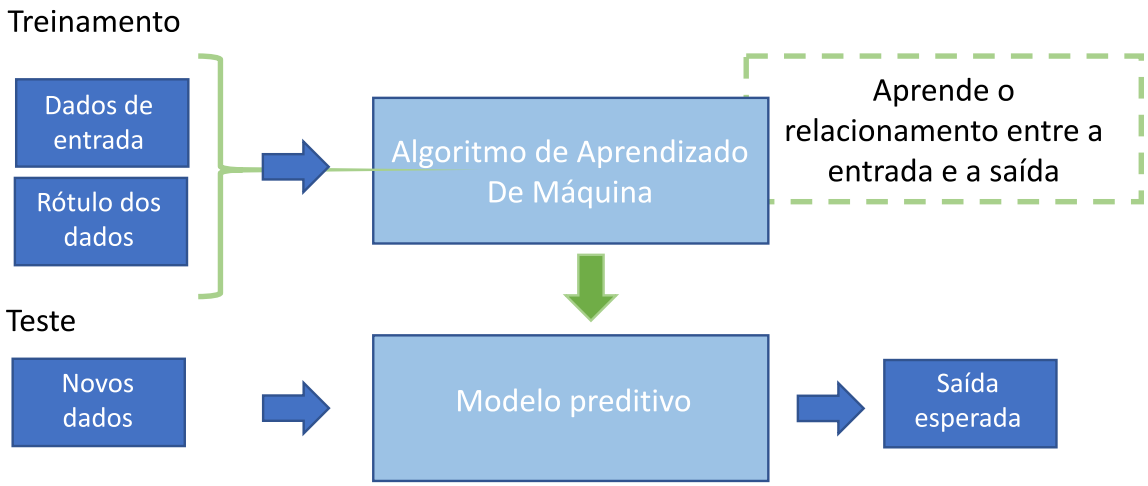


Figura 24 – Esquema do Aprendizado de Máquina Supervisionado - Adaptado de Escovedo; Koshiyama (2020).

5.3.1 Modelos Treinados com Imagens

O processo de treinamento das imagens é apresentado pela Figura 25, em que são utilizadas as características extraídas das quatro imagens. Estas imagens correspondem as vistas esquerda original e degradada e as vistas direita original e degradada. Em seguida, essas características (dados), passam pela etapa de treinamento, onde são treinados diferentes algoritmos e configurações. Gerando como saída diferentes modelos de IQA-3D, que serviram como modelos de testes posteriormente.

Para a etapa de treinamento dos modelos foram usadas 16 imagens estereoscópicas originais, com o total de 790 pares estereoscópicos degradados. Do conjunto de dados, 80% foram separados para o treinamento e validação e 20% para o teste. Utilizou-se o método de validação cruzada *k-fold* com 10 *folds* para todos os classificadores treinados. Foram treinados modelos considerando 4 algoritmos distintos

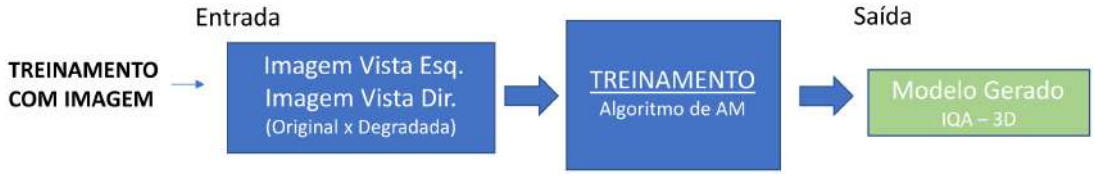


Figura 25 – Processo de treinamento com imagens.

(*J48*, *RePTree*, *RandomForest* e *ForestPA*) contendo variações de hiperparâmetros que levaram a 6 variações dos algoritmos. Estas variações são descritas na Tabela 6.

Tabela 6 – Hiperparâmetros utilizados para os algoritmos treinados com imagens.

Algoritmo	Parâmetro 1	Parâmetro 2
J48 2	minNumObj = 2	Unpruned = true
J48 10	minNumObj = 10	Unpruned = true
RepTree 2	minNum= 2	noPruning= True
RepTree 10	minNum= 10	noPruning= True
ForestPA	simpleCart = 5	-
RandomForest	OutputOutBag = true	numIterations 10

Os hiperparâmetros, conforme apresentados na Tabela 6, são separados em Parâmentro 1 e Parâmetro 2. Neste caso, para o Parâmetro 1 temos o *minNumObj* que especifica o número de instâncias por folha da árvore, o mesmo conceito é seguido para o *minNum*, o que difere entre eles é o algoritmo a que pertencem. Neste caso, com modelos treinados para imagens optamos pelo valor 2, que é o valor padrão de ambos os algoritmos e o valor 10. Já o *simpleCart* é, na verdade o parâmetro *simpleCartPruningFolds* que corresponde ao número de dobras que será realizada pelo algoritmo *ForesPA*. Temos ainda o *OutputOutBag*, que é o *outputOutOfBagComplexityStatistic* que diz se as estatísticas baseadas em complexidade devem ser geradas quando a avaliação *out-of-bag* é executada, neste caso se sim deve-se marcar como verdadeiro (*true*).

O Parâmetro 2, contém os hiperparâmetros *Unpruned*, que diz respeito a realização da poda, no caso de verdadeiro, realiza a poda para o algoritmo *J48*. A mesma definição é dada para o *noPruning* que corresponde ao algoritmo *RepTree*. Por fim o *numIterations*, que corresponde ao número de árvores que se deseja obter para uma floresta, sendo um hiperparâmetro definido para o algoritmo *RandomForest*, e no caso de imagens o valor corresponde a 10 árvores. A manipulação destes hiperparâmetros tem como objetivo buscar uma melhor predição dos modelos treinados.

Para cada variação dos algoritmos avaliados, foram treinados modelos considerando 8 cenários (C1-C8) que incluem diferentes subconjuntos de atributos (inserção ou exclusão), conforme a Tabela 7.

O Cenário 1 (C1) usa todas as características extraídas do conjunto de imagens estereoscópicas. Já o Cenário 2 (C2) exclui as métricas voltadas para 2D. No Cenário

Tabela 7 – Cenários (C1-C8) de treino e teste. VU, VD, VE e PV são conjuntos descritos na Fig.22 e definidos na Tabela 5. Já VU/E e VU/D referem-se respectivamente ao Conjunto VU somente para vista esquerda (E) e direita (D).

Característica	C1	C2	C3	C4	C5	C6	C7	C8
Média	VU	VU	-	-	-	VU	VU/E	VU/D
Desvio Padrão	VU	VU	-	-	-	VU	VU/E	VU/D
Variância	VU	VU	-	-	-	VU	VU/E	VU/D
Contraste	VU	VU	-	-	VU	VU	VU/E	VU/D
PSNR	VE,VD	VE,VD	VE,VD	-	VE,VD	VE,VD	VE	VD
PSNR3D	PV	-	-	PV	PV	-	-	-
MSE	VE,VD	VE,VD	VE,VD	-	VE,VD	VE,VD	VE	VD
MSE3D	PV	-	-	PV	PV	-	-	-
SSIM	VE,VD	VE,VD	VE,VD	-	VE,VD	VE,VD	VE	VD
SSIM3D	PV	-	-	PV	PV	-	-	-
SAD	VE,VD	VE,VD	VE,VD	-	VE,VD	VE,VD	VE	VD
SAD3D	PV	-	-	PV	PV	-	-	-
StSD	PV	PV	-	-	PV	-	-	-
STNx	VE,VD	VE,VD	-	-	VE,VD	VE,VD	VE	VD
STMain	VE,VD	VE,VD	-	-	VE,VD	VE,VD	VE	VD
NS_m	VE,VD	VE,VD	-	-	VE,VD	VE,VD	VE	VD
NS_d	VE,VD	VE,VD	-	-	VE,VD	VE,VD	VE	VD
ND_{sim}	VE,VD	VE,VD	-	-	VE,VD	VE,VD	VE	VD
$StSd_{blr}$	VE,VD	VE,VD	-	-	VE,VD	VE,VD	VE	VD
$StSd_{sim}$	VE,VD	VE,VD	-	-	VE,VD	VE,VD	VE	VD
SC_{med}	VD	VD	-	-	VD	-	-	VD
SC_{desv}	VD	VD	-	-	VD	-	-	VD

3 (C3) foram mantidas as características que se referem às métricas conhecidas na literatura para imagens e vídeos 2D. Neste caso, as métricas foram aplicadas separadamente para a vista esquerda e direita. O Cenário 4 (C4) considera as métricas desenvolvidas para 2D-IQA adaptadas para 3D. O Cenário 5 (C5) excluiu todas as características baseadas em estatísticas. No Cenário 6 (C6) foram testadas somente as características e métricas que utilizam as duas vistas, desconsiderando as métricas que utilizam as quatro vistas como entrada. No Cenário 7 (C7) foram consideradas as métricas e características que utilizam somente a vista esquerda, tanto original como degradada e o Cenário 8 (C8) considerou as métricas e recursos que utilizam somente a vista direita, tanto original como degradada.

Além da definição dos hiperparâmetros, diferentes algoritmos e modelos treinados com diferentes cenários, as imagens foram treinadas com os diferentes algoritmos e variação no número de classes, considerando 5, 10 e 25 classes.

5.3.2 Modelos Treinados com Vídeos

O processo de treinamento dos vídeos pode ser observado na Figura 26. Os modelos foram treinados utilizando características extraídas dos vídeos, quatro vídeos, que correspondem às vistas esquerda original e degradada e direita original e degradada. Estas características foram então treinadas com diferentes algoritmos de AM e configurações adequadas para vídeos. Por fim, são gerados diferentes modelos de VQA-3D, que foram testados e analisados neste trabalho.

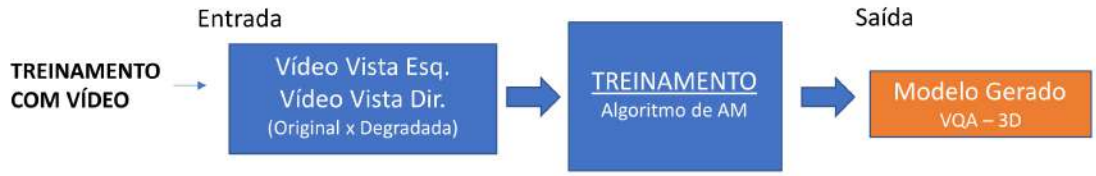


Figura 26 – Processo de treinamento com vídeos.

Os vídeos usados para o treinamento dos modelos apresentam diferentes degradações e pertencem à base de dados de *Waterloo IVC 3D Video Quality Database*, com 10 sequências originais e 704 sequências de pares estereoscópicos degradados. Do conjunto de dados, 80% foram separados para o treinamento e validação e 20% para o teste. Utilizou-se o método de validação cruzada *k-fold* com 10 *folds* para todos os classificadores treinados. Foram treinados modelos considerando 4 algoritmos distintos (*J48*, *RePTree*, *RandomForest* e *ForestPA*) com diferentes configurações dos hiperparâmetros, que levaram a 6 variações dos algoritmos. Como esses modelos tratam de vídeos, alguns hiperparâmetros tiveram que ser reajustados, pois, as características dos vídeos correspondem aos 250 quadros, fato que torna muito maior o número de instâncias quando comparados as das imagens. A Tabela 8 apresenta as configurações dos hiperparâmetros.

Tabela 8 – Hiperparâmetros utilizados para os algoritmos treinados com vídeos.

Algoritmo	Parâmetro 1	Parâmetro 2
J48 1000	minNumObj = 1000	Unpruned = true
J48 50	minNumObj = 50	-
RepTree 1000	minNum= 1000	noPruning= True
RepTree 50	minNum= 50	noPruning= True
ForestPA	simpleCart = 5	-
RandomForest	OutputOutBag = true	numIterations 10

O treinamento dos vídeos é semelhante aos das imagens. No entanto, difere principalmente por não apresentar a análise da variação entre as características, já que esta é estabelecida através do estudo das imagens. Logo, para os vídeos são considerados os 4 algoritmos e suas variações, bem como o treinamento desses algoritmos variando número de classes.

5.4 Teste

Os modelos foram testados sob os modelos treinados utilizando a validação cruzada, *K-folds* 10. A Figura 24, apresenta a forma como os testes são realizados na AM, em que os dados novos são testados sobre os modelos preditivos. Estes contemplam imagens e vídeos não conhecidos pelo sistema e são 20% do conjunto de dados. A seguir, são apresentados o funcionamento dos processos de testes com imagens e vídeos.

5.4.1 Teste com Imagens

A Figura 27 representa o processo de teste realizado com imagens. Neste caso, os dados de entrada são as características extraídas das imagens, que foram separadas para esta etapa, ou seja, o sistema de AM não tem nenhum conhecimento sobre essas imagens. Logo, o teste é realizado utilizando o modelo gerado pelo treinamento (Modelo Gerado IQA-3D), conforme a Figura 25, gerando como saída um modelo final para análise, que será discutida na Seção 6.1.

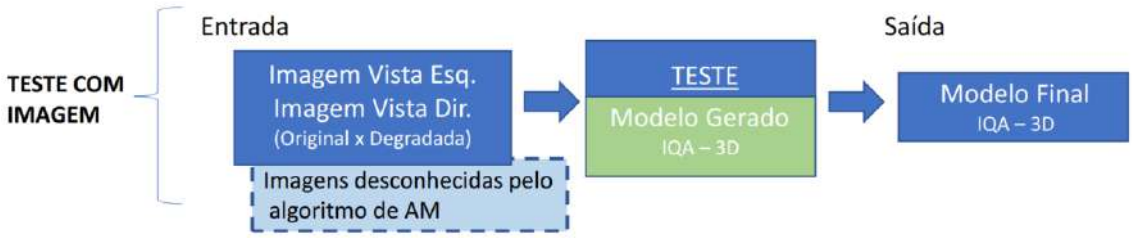


Figura 27 – Representação do processo da etapa de teste utilizando características extraídas das imagens.

5.4.2 Teste com Vídeos

Os testes com vídeos foram realizados de duas maneiras diferentes. O primeiro é representado pela Figura 28, onde as características extraídas dos quatro vídeos servem como entrada para o teste com o modelo gerado pela VQA-3D, conforme a Figura 26. Assim como as imagens, os vídeos foram separados para que o sistema de AM não tivesse conhecimento prévio sobre os dados.

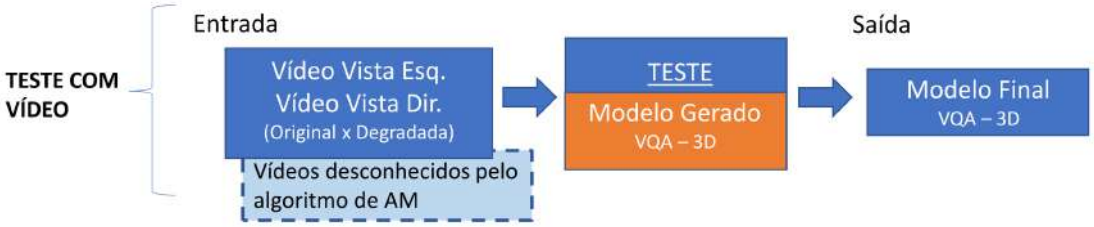


Figura 28 – Representação do processo da etapa de teste utilizando características extraídas dos vídeos.

A segunda maneira é aplicada conforme a Figura 29, em que os dados de entrada para o teste são extraídos dos vídeos e o teste é aplicado ao modelo gerado através de imagens, IQA-3D, conforme representado na Figura 25. Como saída é gerado um modelo VQA-3D. Os modelos gerados através do primeiro e segundo processo foram usados para a análise deste trabalho na Seção 6.2.

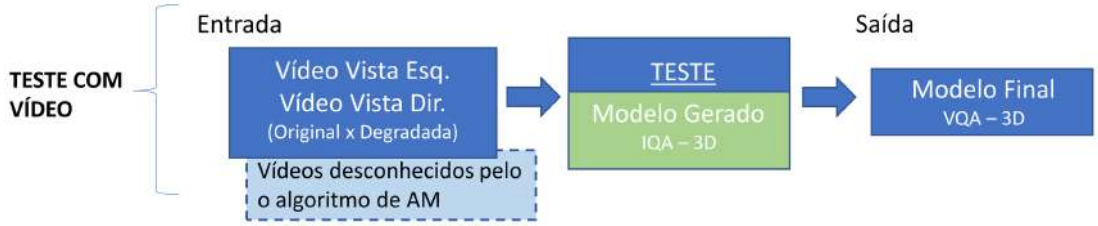


Figura 29 – Representação do processo da etapa de teste utilizando características extraídas dos vídeos, tendo o modelo gerado a partir de imagens.

5.5 Consideração Finais do Capítulo

Neste Capítulo apresentamos o desenvolvimento dos principais passos para a aplicação das técnicas de Aprendizado de Máquina conforme a Figura 18. Sendo assim, apresentamos a etapa de aquisição dos dados e os principais fatores envolvidos nela, como a escolha da base de dados e a extração de características. Logo, foram mostrados os passos da etapa de pré-processamento, importante para que os dados fiquem em conformidade com os arquivos de entrada para o sistema de AM e proporcionar maior qualidade para o treinamento. Também foram apresentadas as principais configurações e características utilizadas para o treinamento e validação das imagens e

vídeos. Por fim, a etapa de Teste onde foram mostrados os detalhes sobre as características dos testes, que terão seus valores apresentados e discutidos no Capítulo 6 a seguir.

6 RESULTADOS E DISCUSSÃO

Este Capítulo apresenta a discussão dos resultados obtidos através dos testes com os modelos treinados, é dividido em duas seções. A Seção 6.1 Avaliação de Qualidade de Imagem 3D e a Seção 6.2 Avaliação de Qualidade de Vídeo 3D. A Avaliação de Qualidade de Imagem 3D apresenta duas discussões, a primeira discorre sobre os valores dos testes com modelos treinados variando as características das imagens e também diferentes algoritmos de AD conforme já descrito na Seção 5.3.1. A segunda discute o impacto dos valores dos teste contemplando todas as características, no entanto, aplicamos diferentes classes. A discussão sobre Avaliação de Qualidade de Vídeo aborda, inicialmente, testes com modelos de treinamento com diferentes classes e a segunda discussão trata de testes com vídeos em modelos treinados com imagens, ambos utilizam todas as características extraídas das imagens e vídeos.

6.1 Avaliação de Qualidade de Imagem 3D

Esta Seção apresenta e discute os resultados dos testes com modelos treinados utilizando imagens. Todos os valores destes testes foram realizados com modelos gerados pelos algoritmos *J48*, *RepTree*,*ForestPA* e *RandomForest*, citados na Seção 2.5, e uma variação para o algoritmo *J48* e o *RepTree*.

A Seção 6.1.1 apresenta os resultados dos testes com modelos treinados variando as características extraídas das imagens. Já a Seção 6.1.2 apresenta os resultados dos testes com modelos treinados utilizando todas as características e variando os números das classes entre 5,10 e 25.

6.1.1 Testes dos Modelos Treinados com Imagens Variando os Atributos

A Tabela 9 apresenta os principais cenários dentre os oito descritos na Tabela 7. Os resultados dos cenários 3, 4, 7 e 8 não são apresentados em detalhes na tabela, pois não apresentaram modelos com predições satisfatórias, ou seja, variando seus valores de Acurácia entre 50% e 70%. Entretanto, serão discutidos brevemente na Figura 31.

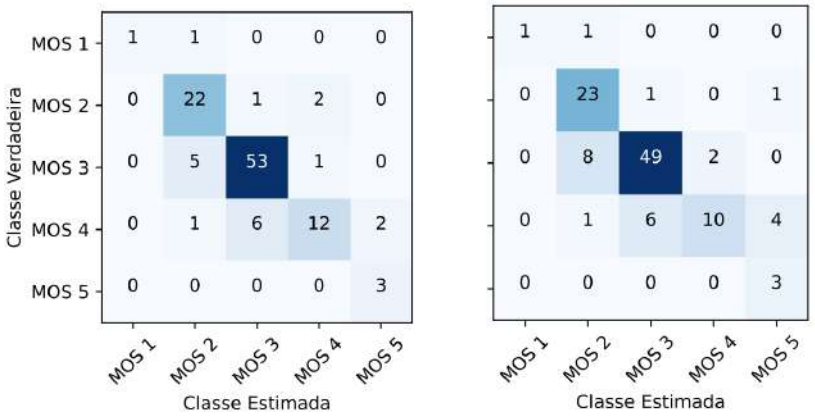
Tabela 9 – Valores de desempenho do conjunto de teste. O Tamanho (N/P) corresponde a número de árvores (N) e profundidade (P).

Cenário	Algoritmo	Acurácia	Precisão	*Recall*	*F1-score*	Tamanho (N/P)
Cenário 1	*J48 2*	0,736	0,798	0,736	0,754	1/159
	J48 10	0,673	0,716	0,673	0,684	1/59
	RepTree 2	0,673	0,690	0,673	0,677	1/167
	RepTree 10	0,627	0,632	0,627	0,611	1/47
	ForestPA	0,709	0,738	0,709	0,716	10/134
	RandomForest	**0,827**	**0,833**	**0,827**	**0,823**	100/251
Cenário 2	*J48 2*	0,689	0,724	0,682	0,692	1/175
	J48 10	0,646	0,700	0,645	0,660	1/69
	RepTree 2	0,700	0,724	0,700	0,702	1/173
	RepTree 10	0,636	0,612	0,636	0,603	1/55
	ForestPA	**0,818**	**0,832**	**0,818**	**0,815**	10/128
	RandomForest	0,800	0,821	0,800	0,798	50/245
Cenário 5	*J48 2*	0,700	0,747	0,700	0,712	1/153
	J48 10	0,673	0,716	0,673	0,684	1/57
	RepTree 2	0,700	0,725	0,700	0,704	1/165
	RepTree 10	0,627	0,632	0,627	0,611	1/47
	ForestPA	0,700	0,763	0,700	0,707	10/154
	RandomForest	**0,782**	**0,815**	**0,782**	**0,780**	100/263
Cenário 6	*J48 2*	0,646	0,670	0,645	0,647	1/179
	J48 10	0,646	0,658	0,645	0,642	1/75
	RepTree 2	0,673	0,705	0,673	0,676	1/179
	RepTree 10	0,736	0,762	0,736	0,739	1/57
	ForestPA	0,736	0,752	0,736	0,733	1/134
	RandomForest	**0,809**	**0,824**	**0,809**	**0,801**	50/257

A Tabela 9 mostra que, no Cenário 1, o teste com o modelo gerado pelo algoritmo *RandomForest* apresenta o melhor valor de Acurácia dentre todos os testes, com 0,872, indicando que o teste classificou 87,20% das instâncias corretamente. Também obteve o valor de Precisão mais alto (0,833), indicando uma baixa taxa de falsos positivos. Além disso, o *Recall* mostra que o classificador obteve sucesso em encontrar os exemplos das classes corretamente com uma frequência de 0,827. Por fim, o *F1-score* apresenta o valor 0,823, o mais próximo do ótimo (1) dentre todos os testes considerados. Ainda nesse mesmo Cenário, o teste com o modelo gerado pelo algoritmo *J48 2* tem a Acurácia de 0,736 e uma Precisão de 0,798, indicando maiores erros de predição em relação ao teste anterior. Já o *Recall* apresenta um valor de 0,736 e o *F1-score* de 0,754. Nota-se que os valores dos testes com o modelo gerado através do *RandomForest* são significativamente melhores para classificar o Cenário 1 em termos de Acurácia, Precisão, *Recall* e *F1-score*, mesmo sendo uma floresta com 100 árvores. No entanto, em termos de eficiência o modelo gerado pelo *J48 2* pode ser considerado bom, já que possui somente uma árvore de tamanho 159 e uma

Acurácia acima de 70%. Em relação aos outros índices, os valores dos testes indicam que o modelo é um bom classificador.

A Figura 30 exibe a Matriz de Confusão do teste com o modelo gerado pelo algoritmo *RandomForest* para o Cenário 1 (Figura 30(a)) e o Cenário 5 (Figura 30(b)). O Cenário 1, apresenta o teste com modelo gerado que detém a maior Acurácia, aponta que mais instâncias foram classificadas corretamente. Ambos os cenários mostram que o modelo acertou 50% as classes do MOS 1, confundido somente com MOS 2, sua classe vizinha. Já para a classe MOS 2 os modelos erraram classes mais distantes. O modelo gerado pelo Cenário 1 confundiu a classe MOS 2 com o MOS 4, e o modelo gerado pelo Cenário 5 confundiu o MOS 2 com o MOS 5. Na classe MOS 3, de maneira geral, os dois cenários obtiveram boa quantidade de acertos, confundindo poucas instâncias com a classe MOS 2 e MOS 4. A classe MOS 4, apresentou maior confusão em relação as outras classes para ambos os modelos. No entanto, o modelo gerado com o Cenário 5 apresentou maior confusão entre as classes. Por fim, a classe MOS 5 obteve 100% de acerto nos dois cenários. Na maior parte dos casos, percebe-se que os maiores erros acontecem entre as classes vizinhas, fato que pode ser justificado pelos valores das instâncias serem muito semelhantes.



(a) RandomForest - Cenário 1. (b) RandomForest - Cenário 5.

Figura 30 – Matriz de Confusão do modelo gerado pelo algoritmo *RandomForest* para os cenários 1 e 5.

No Cenário 2, observa-se que o modelo gerado pelo algoritmo *ForestPA* alcançou a maior Acurácia entre todos os outros modelos do cenário, com valor de 0,818. Já em termos de Precisão o valor é de 0,832. O *Recall* com 0,818 juntamente com o valor de *F1-score* de 0,815, mais próximo de 1, indicam que o modelo gerado pelo *ForestPA* tem uma boa capacidade de encontrar as classes esperadas. O teste com o modelo gerado pelo *RandomForest* tem a segunda maior Acurácia (0,8) e a Precisão de 0,821, apontando uma boa capacidade de Precisão diante às classes. Ainda neste caso, o valor do *Recall* com 0,800 e *F1-score* de 0,798 mostram que, em termos gerais, o modelo também tem uma boa capacidade de classificação. Pode-se observar ainda

que, neste caso do Cenário 2, o teste com o modelo gerado pelo algoritmo *ForesPA* apresenta os melhores valores em termos de Acurácia. Além disso, pode ser considerado o mais eficiente em termos das medidas de desempenho se comparado com os valores do modelo gerado pelo *RandomForest*. Em termos de tamanho, o *ForesPA* também se destaca, pois apresentou um modelo com uma floresta com 10 árvores, com a sua maior árvore de tamanho 128, enquanto o *RandomForest* apresentou uma floresta com 50 árvores e a maior com tamanho de 245.

O teste com o modelo gerado através do algoritmo *RandomForest* do Cenário 5 obteve a maior Acurácia dentre os outros modelos deste cenário, com valor de 0,782. Além disso, obteve uma boa Precisão, com 0,815, *Recall* 0,782 e *F1-score* de 0,780. Os índices indicam que o teste com o modelo gerado pelo algoritmo *RandomForest* obtém a melhor capacidade de predição. Ainda neste cenário, três testes apresentam o segundo melhor valor de Acurácia de 0,700. São para os testes com modelos gerados através dos algoritmos *J48 2*, *RepTree 2* e *ForestPA*. Observa-se que em termos de desempenho os valores são semelhantes, no entanto, o algoritmo *J48 2*, pode ser considerado uma boa opção por apresentar o valor de *F1-score* um pouco maior dentre os outros e uma árvore de tamanho 153 contra a árvore do *RepTree 2* de tamanho 165 e a floresta de *ForestPA* com 10 árvores e a maior de tamanho 154.

Por fim, no Cenário 6 o teste com o modelo gerado pelo algoritmo *RandomForest* obteve a Acurácia de 0,809, ou seja, classificou corretamente 80% das instâncias. A Precisão foi de 0,824, o *Recall* de 0,809 e o *F1-score* de 0,801, mostrando que este algoritmo tem o melhor desempenho e capacidade de predição dentre os outros do mesmo cenário. Os testes com os modelos gerados através dos algoritmos *RepTree 10* e *ForestPA* obtiveram a mesma Acurácia de 0,736, o primeiro apresenta uma Precisão de 0,762, o *Recall* de 0,736 e o *F1-score* de 0,739. Já o segundo tem o valor de Precisão de 0,752, o *Recall* de 0,736 e o *F1-score* de 0,733. Neste caso, os valores das medidas de desempenho do teste com o modelo gerado através do algoritmo *ForestPA*, indicam que seus valores têm melhor desempenho se comparado com os valores do *RepTree 10*.

De maneira análoga, através da Figura 31 observa-se que os testes com os modelos gerados através dos Cenários 1 e 2 obtiveram os dois maiores valores de Acurácia, o primeiro utilizou o algoritmo *RandomForest* e o segundo o *ForestPA*. O Cenário 1 integra todas as características, ou seja, atributos no modelo de treinamento e teste. Já o Cenário 2 exclui as métricas do conjunto PV (Tabela 5) que foram adaptadas de 2D para 3D. Baseado no comportamento dos modelos gerados para esses dois cenários, pode-se ir de acordo com a ideia de Fang; Sui; Wang (2019), que diz que as métricas 2D aplicadas a 3D não se correlacionam tão bem com o valor do MOS, já que quando retiradas do cenário a Acurácia tende a diminuir um pouco. Em outras palavras, sem essas informações o classificador é capaz de manter sua boa capacidade de predi-

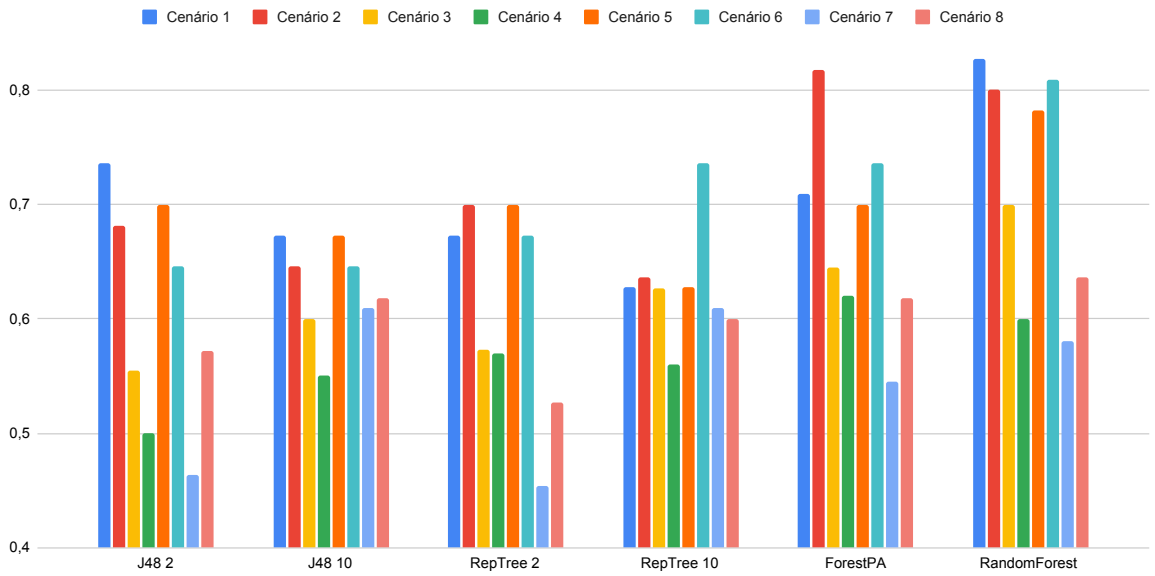


Figura 31 – Acurácia dos modelos gerados em cada cenário.

ção. Além disso, no Cenário 4, que é composto pelo Conjunto PV, apresentaram uma Acurácia significativamente baixa, confirmando que esse valores sozinhos não permitem que o classificador tenha boa capacidade de predição. Observa-se ainda, que quando os cenários são treinados e testados somente para métricas 2D, aplicadas ao Conjunto VE e VD separadamente, os valores de Acurácia também são baixos. Por outro lado, o Cenário 5 apresenta boa Acurácia na maioria dos modelos gerados por diferentes algoritmos. Este cenário omite os valores do conjunto VU, que avalia as características estatísticas das imagens. Entretanto, apresenta uma queda na Acurácia, mostrando que a retirada dessas características interferem no processo de classificação do algoritmo. O Cenário 6 é semelhante ao Cenário 2, diferindo apenas na exclusão das informações de $Th1$ e $Th2$. O valor da Acurácia do modelo gerado apresenta uma queda pequena, ou seja, as características excluídas apresentam relevância moderada para o classificador.

O Cenário 7 utilizou somente informações de VE e o Cenário 8 apenas as informações de VD. Os valores de Acurácia desses modelos mostram que só uma vista pode não fornecer informações relevantes para o classificador. Este fato pode ser apoiado pela rivalidade binocular, em que uma vista tende a exercer dominância sobre a outra quando o que é visto exerce mais estímulo sobre ela.

De maneira geral, diante da análise dos testes com os modelos gerados baseados em AD, pode-se constatar que dentre todos os algoritmos testados em cada cenário, o algoritmo *RandomForest* obteve, na maioria dos casos, os melhores resultados mostrando assim que este algoritmo apresenta a melhor capacidade de predição para os cenários usados neste trabalho.

Respondendo a questão de pesquisa **Q3**, neste estudo observou-se que todas as características quando combinadas (Cenário 1) fornecem informações relevantes para que o classificador tenha uma boa capacidade de predição. Devido a esta conclusão, optamos por apresentar os próximos estudos baseados em testes com os modelos treinados com todas as características extraídas das imagens e dos vídeos.

6.1.2 Testes com Imagens de Modelos Treinados com Imagens

A Tabela 10 apresenta os valores da avaliação de desempenho dos testes com modelos treinados para 5 classes. Nota-se que o algoritmo *RandomForest* apresenta uma maior Acurácia e os melhores valores em termos de desempenho, ou seja, a Precisão, *Recall* e *F1-score* com os valores mais altos. Podemos observar que o algoritmo *J48 2* apresenta uma boa Acurácia também, no entanto os valores de Precisão, *Recall* e *F1-score* apresentam valores menores que o *RandomForest*. Já o algoritmo *ForestPA*, apesar de apresentar uma Acurácia de 0,709, em termos de desempenho os valores são baixos em relações aos demais algoritmos, mostrando que o algoritmo apesar de ter uma boa Acurácia tem pouca capacidade de predizer corretamente as classes esperadas.

Tabela 10 – Valores de Acurácia dos testes com imagens e com os modelos treinados com imagens para 5 classes.

Algoritmo	Acurácia	Precisão	*Recall*	*F1-score*
J48 2	0,736	0,798	0,736	0,754
J48 10	0,672	0,716	0,673	0,684
RepTree 2	0,673	0,690	0,673	0,677
RepTree 10	0,627	0,632	0,627	0,611
ForestPA	0,709	0,527	0,709	0,537
RandomForest	**0,827**	**0,833**	**0,827**	**0,823**

Os valores apresentados na Tabela 11 referem-se aos testes dos modelos treinados com 10 classes. Logo, tem-se o algoritmo *RandomForest* com a melhor Acurácia e valores de Precisão, *Recall* e *F1-score*. Com a segunda maior Acurácia tem-se os valores de *ForestPA* com 0,554, diferentemente da análise com 5 classes, este algoritmo para 10 classes apresenta uma boa Precisão. Os testes realizados com os

Tabela 11 – Valores de Acurácia dos testes com imagens e com os modelos treinados com imagens para 10 classes.

Algoritmo	Acurácia	Precisão	*Recall*	*F1-score*
J48 2	0,436	0,501	0,436	0,443
J48 10	0,418	0,426	0,418	0,405
RepTree 2	0,436	0,472	0,436	0,431
RepTree 10	0,445	0,482	0,445	0,452
ForestPA	0,554	0,580	0,555	0,561
RandomForest	**0,618**	**0,653**	**0,618**	**0,625**

modelos treinados com 25 classes são apresentados na Tabela 12, onde se observa que os valores tanto de Acurácia quanto das medidas de avaliação de desempenho são baixos se comparados com as classes 5 e 10. Porém, dentro deste cenário de 25 classes temos o algoritmo *ForestPA* com maior valor de Acurácia. Entretanto, o algoritmo *j48 10*, apresenta uma Acurácia mais baixa com valor de 0,209, mas os seus valores de de desempenho como a Precisão e *F1-score* são maiores se comparados ao *ForesPA*, sendo respectivamente 0,296 e 0,261. Fato que, apesar de não ter a maior Acurácia, apresenta melhor capacidade de Precisão que o *RandomForest*.

Em termos gerais, para os testes com imagens em modelos treinados com imagens, podemos notar que para 5 classes os valores de testes apresentados são mais altos, chegando a 0,827 de Acurácia. Diferentemente dos valores preditos quando testados para 10 classes, os valores caem para uma taxa de no máximo 0,618. A principal diferença pode ser notada quando os testes são realizados com 25 classes, os valores de Acurácia assumem uma taxa muito baixa sendo o maior valor 0,218. No que diz respeito aos testes com imagens apresentados, as florestas obtém os melhores resultados.

Tabela 12 – Valores de Acurácia dos testes com imagens e com os modelos treinados com imagens para 25 classes.

Algoritmo	Acurácia	Precisão	*Recall*	*F1-score*
J48 2	0,191	0,189	0,191	0,195
J48 10	0,209	0,296	0,209	0,261
RepTree 2	0,145	0,204	0,145	0,165
RepTree 10	0,191	0,213	0,191	0,162
ForestPA	**0,218**	**0,242**	**0,218**	**0,226**
RandomForest	0,209	0,169	0,209	0,226

As Tabelas 13 e 14 apresentam os valores de RMSE e MAE dos testes para as três classes. Os cálculos seguem respectivamente conforme as Eq.(7) e (6), citadas na Seção 2.7. No entanto, os valores utilizados para os cálculos foram baseados no MOS real (classe real) e MOS esperado (classe esperada), ou seja, foi realizado um pós processamento para realizar uma nova transformação dos dados. Esta transformação, foi feita transformando os valores das classes preditas 5,10 e 25 para valores numa escala de 1 a 100 como classe real, considerando os valores originais de MOS. As Eq. (60),(61) e (62) descrevem o mapeamento de cada uma das classes, em que X representa o valor da classe predita (5,10 e 25).

$$MOSpredito5 = X(100/5) - 10 \tag{60}$$

$$MOSpredito10 = X(100/10) - 5 \tag{61}$$

$$MOSpredito10 = X(100/25) - 2 \quad (62)$$

Depois da transformação destes dados, os valores de MAE e RMSE foram calculados utilizando o MOS real e o MOS predito. Esta transformação foi utilizada para todos os outros cálculos de MAE e RMSE apresentados nas próximas seções deste trabalho.

A Tabela 13 mostra os valores de RMSE dos valores dos testes. Através do RMSE observa-se a proximidade dos valores reais e preditos pelo modelo. Neste caso, nota-se que tanto para 5 classes quanto para 10 classes os valores mais baixos são para o algoritmo *RandomForest*, mostrando uma concordância com a Acurácia. Já para 25 classes, apresenta o menor valor de RMSE para o algoritmo *J48 10*, que em termos de Acurácia o melhor valor é para o algoritmo *ForestPA*.

Tabela 13 – Valores de RMSE dos testes com imagens referente aos modelos treinados com 5,10 e 25 classes.

Algoritmo	5 classes	10 classes	25 classes
J48 2	10,160	10,253	12,67
J48 10	13,321	12,172	**10,99**
RepTree 2	10,893	12,993	14,65
RepTree 10	12,731	12,704	12,72
ForestPA	12,245	9,582	11,20
RandomForest	**9,311**	**8,032**	11,81

Tabela 14 – Valores de MAE dos testes com imagens referente aos modelos treinados com 5,10 e 25 classes.

Algoritmo	5 classes	10 classes	25 classes
J48 2	7,521	7,480	8,91
J48 10	9,458	8,569	8,32
RepTree 2	8,211	8,859	10,30
RepTree 10	9,337	8,874	9,08
ForestPA	8,493	6,393	7,83
RandomForest	**6,760**	**5,592**	**7,69**

A Figura 32 exibe o comportamento dos valores de RMSE referentes aos testes dos modelos treinados. Estes valores variam de acordo com o algoritmo e as classes utilizadas. Para o algoritmo *J48 2* os valores são semelhantes para 5 e 10 classes e tem um grande aumento para 25 classes. Já para o algoritmo *j48 10* os valores de RMSE diminuem à medida que as classes aumentam. Os valores dos testes com o algoritmo *RepTree 2* tem um comportamento totalmente oposto ao algoritmo anterior, em que, os valores de RMSE aumentam a medida que as classes aumentam. Já para o *RepTree 10* os valores se mantém muito semelhantes, tendo um pequeno declínio quando executado com 10 classes, que pode ser melhor visto na Tabela 13.

O algoritmo *ForestPA* apresenta um comportamento em “u” tendo seu menor valor em 10 classes. Por fim, o *RandomForest* apresenta um comportamento semelhante ao *ForestPA*, tendo seu menor valor quando em 10 classes. Entretanto, nota-se que o *RandomForest* tem os menores valores de RMSE para as classes 5 e 10, ainda que, para 25 classes os valores não sejam os melhores.

Os valores de MAE apresentados na Tabela 14 mostram, de forma geral, que o *RandomForest* apresenta os menores valores. Com isso, o *RandomForest* mostrou ter a melhor capacidade de predição dos dados, dentre os algoritmos testados, tanto em termos de avaliação de desempenho quanto medidas de erro.

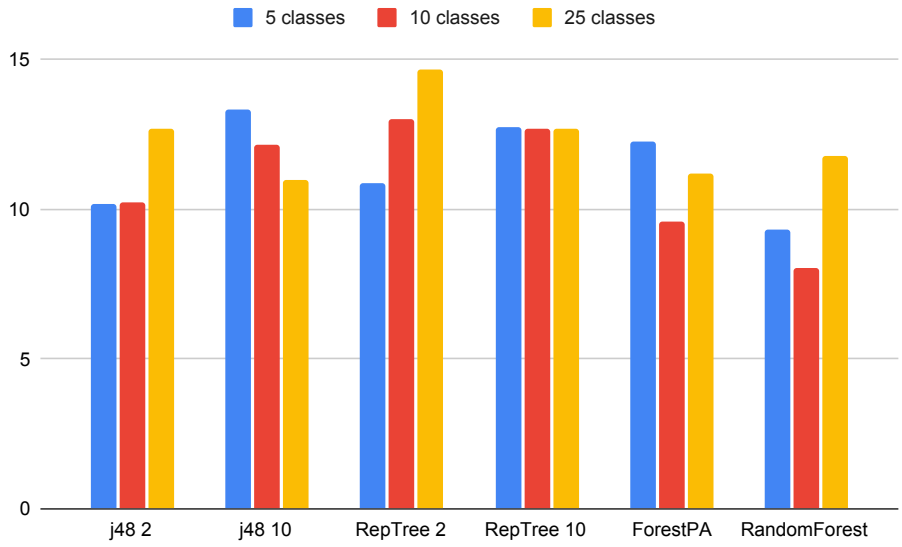
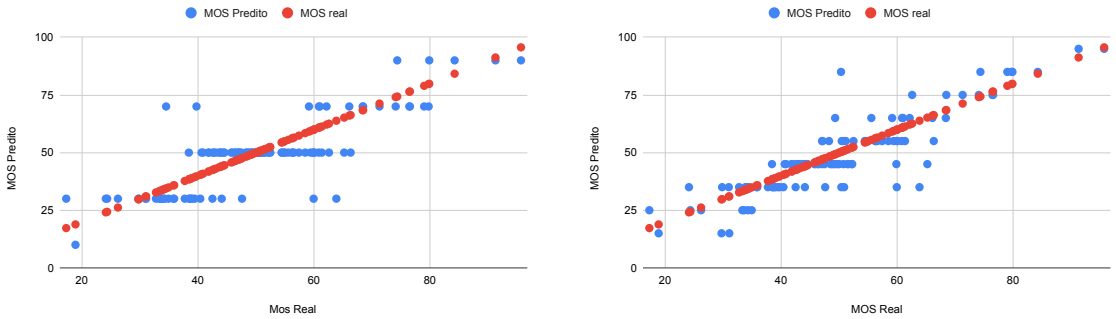


Figura 32 – Valores de RMSE para os modelos treinados e testados com imagens.

As Figuras 33(a) e 33(b) mostram a dispersão entre os valores das classes preditas e reais para o algoritmo *RandomForest*, com 5 e 10 classes, pois estes testes apresentaram maiores valores de Acurácia e menores valores de RMSE dentre os outros algoritmos testados. Nota-se que na Figura 33(a) os valores de MOS preditos apresentam maior espaçamento entre uma classe e outra, já que são somente 5 classes. No entanto, também tem maior dispersão em relação ao MOS Real. Já a Figura 33(b), os valores preditos estão menos espaçados e mais concentrados em volta do MOS real. Fato que se confirma, pois o valor do *RandomForest* para 10 classes é menor em relação a 5 classes. O mesmo não acontece quando se trata das métricas de desempenho, pois a Acurácia do *RandomForest* para 5 classes é, significativamente, maior do que para 10 classes. Porém, é importante observar que apesar de apresentar uma maior taxa de erro, os valores errados para 10 classes estão mais próximos dos valores reais.

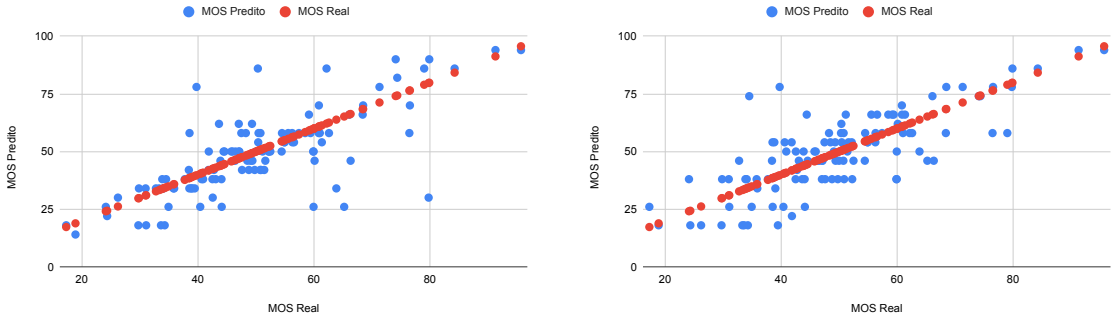
As Figuras 34(a) e 34(b) exibem a dispersão dos valores do MOS real e predito, respectivamente, para os algoritmos *ForesPA* e *j48 10* para 25 classes.



(a) (b)

Figura 33 – Dispersão entre as classes (MOS) preditas e reais do algoritmo *RandomForest*. A Figura (a) representa 5 classes e a Figura (b) representa 10 classes.

Nota-se que na Figura 34(a), os valores do MOS predito apresentam distâncias maiores em relação aos apresentados na Figura 34(b). Em que, os valores do MOS predito estão mais concentrados ao redor do MOS real. Isto, possivelmente, porque o valor de RMSE do *j48 10* é menor do que o *ForestPA*.



(a) (b)

Figura 34 – Dispersão entre as classes preditas e reais dos algoritmo *ForestPA* e *J48 10* para 25 classes. A Figura (a) representa o algoritmo *ForestPA* e a Figura (b) o algoritmo *J48 10*.

No contexto, em que os modelos treinados com imagens foram testados com imagens, tanto os valores de desempenho como os das métricas de erro apresentaram resultados satisfatórios. Com isso, diante dos resultados analisados, buscando responder à questão de pesquisa **Q1** consideramos viável aplicar técnicas baseadas em AD para avaliar a qualidade de imagem 3D, principalmente quando se tratando de florestas como os algoritmos *RandomForest* e *ForestPA*.

6.2 Avaliação de Qualidade de Vídeo 3D

Nesta Seção apresentamos os valores dos testes utilizando vídeos, os modelos treinados utilizam todas as características extraídas dos vídeos, dada a conclusão do estudo realizado na Seção anterior 6.1.1. Todos os testes foram realizados com

modelos treinados com os algoritmos *j48*, *RepTree*, *ForesPA* e *RandomForest*, com algumas variações nos hiperparâmetros. A Seção 6.2.1 apresenta os resultados dos testes com vídeos para modelos treinados com vídeos, já a Seção 6.2.2 discute os resultados dos testes com vídeos para modelos treinados com imagens.

6.2.1 Testes com Vídeos para Modelos Treinados com Vídeos

A Tabela 15 apresenta os resultados dos valores de Acurácia, Precisão, *Recall* e *F1-score* dos testes com vídeos, dos modelos também treinados com vídeos. Neste cenário, percebe-se que o algoritmo *j48 50* apesar de ter a Acurácia maior, com 0,523 em relação aos demais, o valor de *F1-score* é de 0,480. Já o algoritmo *RepTree 1000* apresenta uma Acurácia de 0,516, um pouco menor, no entanto, os valores de Precisão, *Recall* e *F1-score* apresentam menores diferenças entre si, sendo respectivamente 0,530, 0,516 e 0,515, e tem o maior valor de *F1-score* o que indica que o algoritmo teve uma melhor capacidade de predição do que o *j48 50*.

Tabela 15 – Valores de Acurácia dos testes com vídeos e com os modelos treinados com vídeos para 5 classes.

Algoritmo	Acurácia	Precisão	*Recall*	*F1-score*
J48 1000	0,509	0,490	0,510	0,465
J48 50	**0,523**	**0,518**	**0,523**	**0,480**
RepTree 1000	**0,516**	**0,530**	**0,516**	**0,515**
RepTree 50	0,500	0,512	0,500	0,484
ForestPA	0,473	0,469	0,473	0,428
RandomForest	0,501	0,532	0,501	0,445

Os valores para 10 classes, descritos na Tabela 16, mostram que o algoritmo *RepTree 1000* tem maior Acurácia, ou seja, tem maior taxa de acerto que os outros algoritmos dentro deste cenário. Além da melhor Acurácia, este algoritmo também apresenta os melhores valores em termos de desempenho.

Tabela 16 – Valores de Acurácia dos testes com vídeos e com os modelos treinados com vídeos para 10 classes.

Algoritmo	Acurácia	Precisão	*Recall*	*F1-score*
J48 1000	0,339	0,370	0,339	0,222
J48 50	0,284	0,324	0,285	0,275
RepTree 1000	**0,345**	**0,311**	**0,345**	**0,302**
RepTree 50	0,288	0,309	0,288	0,246
ForestPA	0,306	0,316	0,306	0,271
RandomForest	0,290	0,320	0,290	0,248

A Tabela 17, apresenta valores baixos para todos os valores de testes com os algoritmos. Neste contexto, o que apresenta melhores resultados é o *j48 50*, em todos os termos.

Tabela 17 – Valores de Acurácia dos testes com vídeos e com os modelos treinados com vídeos para 25 classes.

Algoritmo	Acurácia	Precisão	*Recall*	*F1-score*
J48 1000	0,103	0,118	0,103	0,098
J48 50	**0,116**	**0,129**	**0,116**	**0,111**
RepTree 1000	0,092	0,082	0,092	0,081
RepTree 50	0,083	0,09	0,083	0,074
ForestPA	0,094	0,12	0,094	0,088
RandomForest	0,093	0,088	0,094	0,080

As Tabelas 18 e 19, mostram respectivamente os valores de RMSE e MAE para os valores dos testes com modelos treinados com 5,10 e 25 classes. Observa-se que o algoritmo *RepTree 100* tem o menor valor de RMSE, indicando que dentre todos, os valores preditos para 10 classes estão mais próximos dos valores reais.

Tabela 18 – Valores de RMSE dos testes com vídeos referente aos modelos treinados com 5,10 e 25 classes.

Algoritmo	5 classes	10 classes	25 classes
J48 1000	**12,80**	12,21	12,52
J48 50	13,85	13,04	**12,21**
RepTree 1000	12,98	**12,18**	16,57
RepTree 50	14,02	13,48	16,13
ForestPA	14,34	13,19	13,11
RandomForest	13,36	12,46	12,55

Tabela 19 – Valores de MAE dos testes com vídeos referente aos modelos treinados com 5,10 e 25 classes.

Algoritmo	5 classes	10 classes	25 classes
J48 1000	**10,60**	**9,43**	10,00
J48 50	10,83	10,43	**9,63**
RepTree 1000	10,70	9,50	13,07
RepTree 50	11,40	10,71	12,91
ForestPA	11,66	10,27	10,35
RandomForest	10,87	9,95	10,10

A Figura 35, permite que seja observado o comportamento de cada algoritmo e classes diante os valores de RMSE. Inicialmente, o algoritmo *J48 2* tem maior valor de RMSE quando testado para 5 classes e seu menor valor para 10 classes. Já o algoritmo *j48 10*, mostra um comportamento de queda à medida que as classes aumentam, ou seja, quando para 25 classes tem o menor RMSE e seus valores preditos estão mais próximos dos valores reais. Já os algoritmos *Reptree 1000* e *Reptree 50*, apresentam comportamentos semelhantes entre si, sendo que os menores valores de RMSE para 10 classes e valores mais altos para 25 classes. O algoritmo *ForestPA*, indica que os valores de RMSE diminuem conforme a classe aumenta. Por fim, o al-

goritmo *RandomForest*, tem os valores de RMSE em queda quando aplicados a 10 classes.

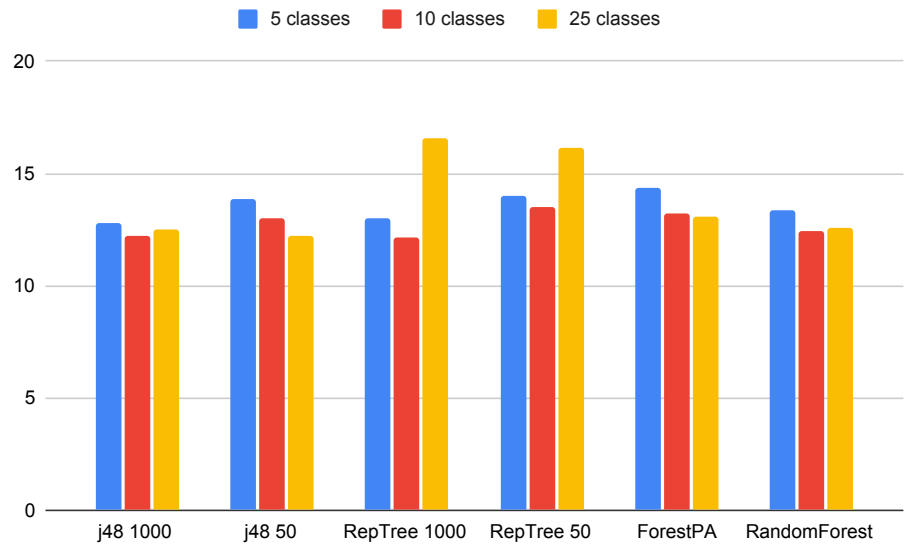


Figura 35 – Valores de RMSE para os modelos treinados e testados com vídeos.

As Figuras 36(a), 36(b) e 36(c) representam, respectivamente, a relação do valor do MOS predito e real entre as classes 5 com algoritmo *j48 1000*, classe 10 com o *RepTree 1000* e 25 classes com o algoritmo *j48 50*. Através dos valores de RMSE nota-se que apesar de ser intuitivo que, quanto maior o numero de classes o número do RMSE tende a ser menor, neste caso temos o menor valor quando o algoritmo é testado com 10 classes.

De modo geral, percebe-se que os valores de desempenho são mais baixos quando comparados com os modelos testados com imagens. Fato este que pode ocorrer devido a natureza complexa dos vídeos 3D. Os vídeos envolvem uma maior complexidade quando se trata das características, já que incluem percepções como a de movimento. Além disso, os valores de MOS fornecidos pela base de vídeos são referentes a cada sequência de vídeo e não ao nível de quadros. Nossos estudos sugerem que quando aplicados a 5 classes a capacidade de predição dos algoritmos é em torno de 50%. Sendo assim, devido a sua complexidade, a questão de pesquisa **Q2**, que questiona a viabilidade de AD para Avaliação de Qualidade de Vídeo, é considerada praticável, porém melhorias na solução são desejáveis.

6.2.2 Testes com Vídeos de Modelos Treinados com Imagens

A Tabela 20 apresenta os valores dos testes com 5 classes. Observa-se que o algoritmo *RandomForest* tem maior Acurácia, com valor de 0,408, dentre os outros algoritmos e em termos de desempenho apresenta a medida *F1-score* com valor de 0,348. Já algoritmo *J48 10*, apresenta a Acurácia de 0,379, e seu valor de *F1-score*

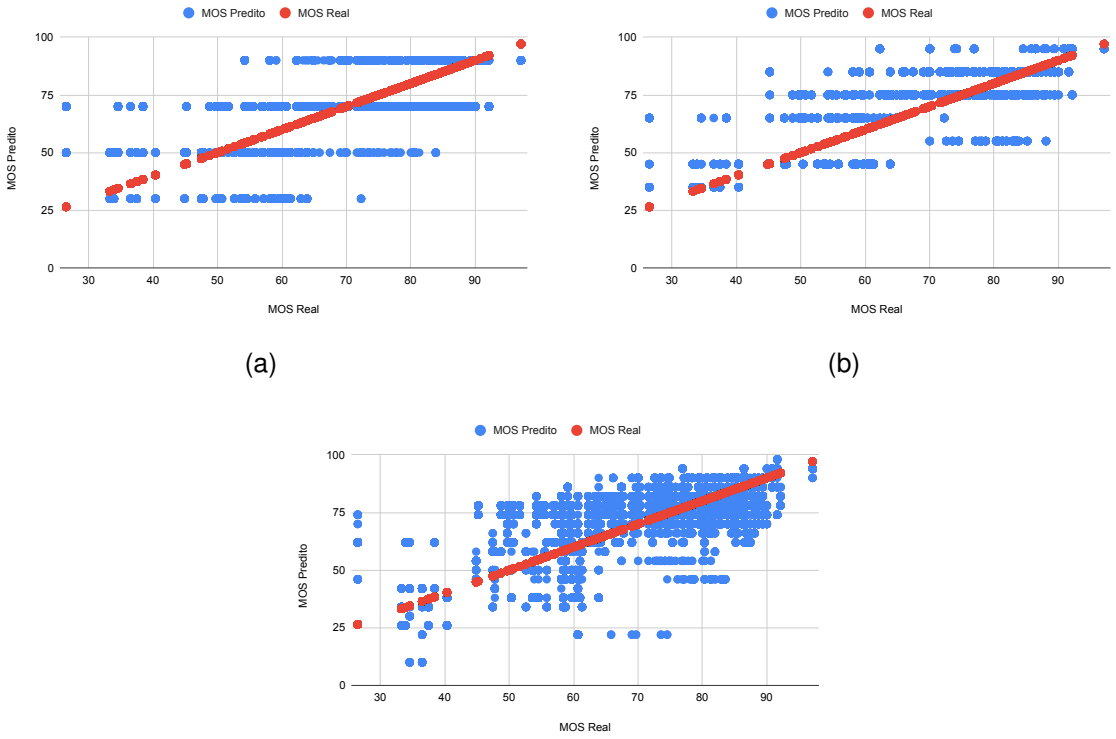


Figura 36 – Dispersão entre as classes (MOS) preditas e reais dos testes aplicados aos modelos treinados com diferentes algoritmos. As Figuras representam: (a) *J48 1000* com 5 classes; (b) *RepTree 1000* com 10 classes; (c) *j48 50* com 25 classes.

é 0,373, sendo maior do que o *RandomForest*. Com isso, entende-se que o *J48 10* tem a Acurácia menor, no entanto, sua capacidade de predição é melhor do que o algoritmo *RandomForest*. A Tabela 21, refere-se aos valores dos testes com 10 classes, onde o algoritmo *j48 10* tem a maior Acurácia, de 0,226, e a medida de *F1-score* de 0,199. Mas algoritmo *RepTree 10*, que tem uma Acurácia menor, com valor de 0,163, tem os valores de medidas de desempenho *F1-score*, de 0,273, maior em relação ao algoritmo *j48 10*,o que infere que sua capacidade de predição é melhor.

Tabela 20 – Valores de Acurácia dos testes com vídeos e com os modelos treinados com imagens para 5 classes.

Algoritmo	Acurácia	Precisão	*Recall*	*F1-score*
J48 2	0,251	0,368	0,251	0,251
J48 10	**0,379**	**0,537**	**0,379**	**0,373**
RepTree 2	0,201	0,267	0,201	0,227
RepTree 10	0,287	0,435	0,287	0,263
ForestPA	0,312	0,250	0,312	0,178
RandomForest	**0,408**	**0,572**	**0,408**	**0,348**

A Tabela 22 mostra que os valores quando são testados com 25 classes são, de modo geral, muito baixos. Entretanto, o que apresenta maior Acurácia é o algoritmo *RandomForest*. O Algoritmo *J48 10* apresenta um valor de *F1-score* maior que o

Tabela 21 – Valores de Acurácia dos testes com vídeos e com os modelos treinados com imagens para 10 classes.

Algoritmo	Acurácia	Precisão	*Recall*	*F1-score*
J48 2	0,129	0,151	0,129	0,155
J48 10	**0,226**	**0,161**	**0,226**	**0,199**
RepTree 2	0,141	0,109	0,141	0,136
RepTree 10	**0,163**	**0,270**	**0,163**	**0,273**
ForestPA	0,114	0,097	0,115	0,101
RandomForest	0,143	0,299	0,143	0,152

RandomForest, indicando que o mesmo pode classificar as classes esperadas com maior Precisão.

Tabela 22 – Valores de Acurácia dos testes com vídeos e com os modelos treinados com imagens para 25 classes.

Algoritmo	Acurácia	Precisão	*Recall*	*F1-score*
J48 2	0,036	0,061	0,036	0,051
J48 10	**0,082**	**0,105**	**0,082**	**0,086**
RepTree 2	0,047	0,095	0,047	0,041
RepTree 10	0,044	0,043	0,044	0,073
ForestPA	0,067	0,064	0,067	0,065
RandomForest	**0,089**	**0,079**	**0,089**	**0,075**

As Tabelas 24 e 23 apresentam os valores de MAE e RMSE dos valores de teste. Repara-se que os algoritmos *RepTree*, *ForesPA* e *RandomForest*, têm os menores valores de RMSE para 5 classes. Já quando os valores de testes se aplicam para 10 classes, o *RandomForest* apresenta o menor valor, sendo 19,27. E para 25 classes o algoritmo *ForesPA*, tem o menor valor com 17,46. Pode-se observar que diante este cenário os valores, tanto de medidas de desempenho quanto em termos de medidas de erro, são extremamente baixos quando comparados com os valores dos cenários anteriores.

Tabela 23 – Valores de RMSE dos testes com vídeos referente aos modelos treinados com imagens com 5,10 e 25 classes.

Algoritmo	5 classes	10 classes	25 classes
J48 2	25,92	23,51	28,86
J48 10	20,22	21,41	24,10
RepTree 2	**17,87**	25,44	26,17
RepTree 10	21,88	25,05	27,86
ForestPA	**17,87**	25,80	**17,46**
RandomForest	**17,87**	**19,27**	19,58

A Figura 37 mostra a distribuição dos valores de RMSE dos valores testado com vídeos, em um primeiro momento nota-se que o algoritmo *j48 2* tem o menor valor de RMSE para 10 classes e um valor bem maior para 25 classes. Já o *j48 10*, tem

Tabela 24 – Valores de MAE dos testes com vídeos referente aos modelos treinados com imagens com 5,10 e 25 classes.

Algoritmo	5 classes	10 classes	25 classes
J48 2	21,53	19,09	25,32
J48 10	16,01	16,96	18,79
RepTree 2	**14,33**	21,01	21,76
RepTree 10	18,30	21,07	23,42
ForestPA	17,52	20,58	**14,29**
RandomForest	**14,33**	**16,02**	16,04

os valores de RMSE aumentando á medida que as classes aumentam, sendo que o seu menor valor é para 10 classes. O algoritmo *RepTree 2*, apresenta os valores de 5 classes menores que 10 e 25 classes, sendo os valores para 10 e 25 classes bem próximos. No entanto, o *RapTree 10*, também apresenta o menor RMSE para 5 classes, porém os valores entre 10 e 25 classes variam entre eles, sendo 25 classes o mais alto. Já o algoritmo *ForestPA*, tem o menor RMSE quando a classe é 25, e um valor muito alto para 10 classes. Ao final, pode-se ver o *RandomForest*, tendo os valores de RMSE aumentando levemente quando a classe também aumenta.

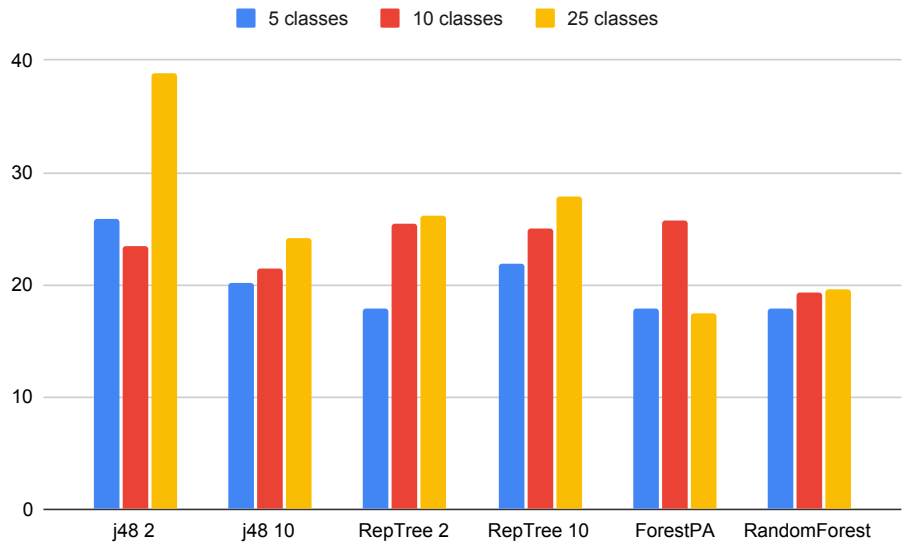


Figura 37 – Valores de RMSE para os modelos treinados com imagens e testados com vídeos.

A Figura 38, apresenta os gráficos de dispersão com os melhores valores RMSE dos testes para a 5 classes com o algoritmo *RandomForest* 38(a), 10 classes para o *RandomForest* 38(b), e 25 classes com algoritmo *ForesPA* 38(c). O menor valor de RMSE é para *ForesPA* quando testado com 25 classes, esses valores vão contra aos valores da Acurácia. Já que os valores de Acurácia, Precisão, *Recall* e *F1-score* com valores muito baixos dentre os demais apresentados na imagem. Fato mostra que, pode-se apresentar uma melhor Acurácia, mas não quer dizer que as classes preditas estão sendo classificadas mais próximas das classes reais, apesar de ter maior erro.

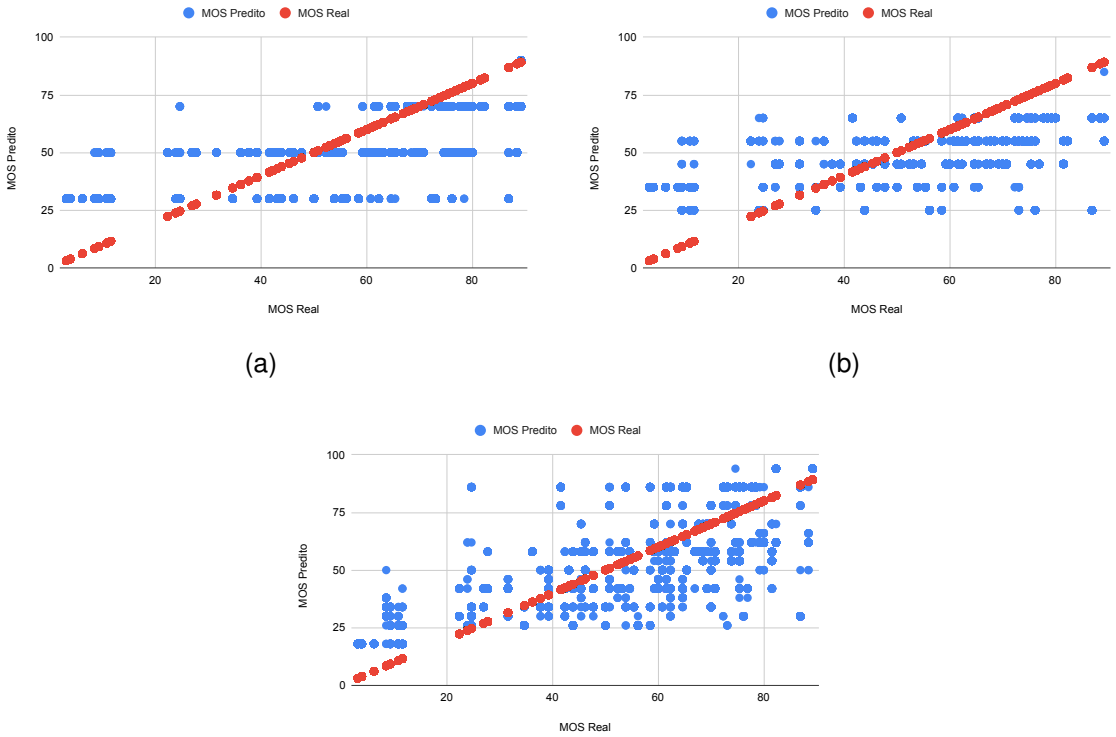


Figura 38 – Dispersão entre as classes (MOS) preditas e reais dos testes com modelos treinados com diferentes algoritmos. As Figuras: (a) *RandomForest* com 5 classes; (b) *RandomForest* com 10 classes; (c) *ForestPA* com 25 classes.

Os valores apresentados neste cenário correspondem aos testes com vídeos 3D em modelos treinados com pares de imagens estereoscópicas/3D. Os resultados apresentam valores muito baixos se comparados aos testes das seções anteriores. Pensa-se que um dos fatores que interferem na qualidade de predição dos algoritmos para esse modelo hibrido, são as degradações das imagens e vídeos, pois elas não são as mesmas. Além disso, outro fator que impede melhorias em termos gerais (todos os cenários) é o desbalanceamento dos dados frente às classes, pois os valores de MOS muito altos e baixos, são menos numerosos que os valores mais centrais.

Por tanto, diante dos resultados analisados neste estudo, respondendo a questão de pesquisa **Q5**, testar vídeos em modelos treinados com imagens não demonstrou ser uma prática adequada.

6.3 Comparação com Trabalhos Relacionados

Neste trabalho destacamos no Capítulo 4 os trabalhos relacionados, que incluem métricas de avaliação objetiva de qualidade de imagem e vídeo de referência completa que utilizam alguma técnica de AM. Dentre estas métricas temos a métrica StSD proposta por Silva (2013), a métrica proposta por Narwaria; Lin (2011), a MLIQM de-

senvolvida por Charrier; Lézoray; Lebrun (2012) e a VMAF desenvolvida pela Netflix conforme descrita por Rassool (2017). Todas as métricas apresentadas buscam avaliar o grau de degradação de uma imagem ou vídeo de maneira análoga a percepção do sistema visual humano, para isso geram um único índice. Estes autores aplicaram técnicas baseadas em AM para o desenvolvimento dessas métricas, algumas utilizam para extrair recursos e outras como técnica de agrupamento dos valores para gerar o índice. No entanto, de maneira geral, estes estudos não apresentam detalhes da metodologia de AM, como também os valores de Acurácia e medidas de desempenho. Sendo assim, nosso trabalho não pode ser comparado diretamente com estes trabalhos em termos de valores, já que analisamos e discutimos sobre a utilização de AD's aplicadas a Avaliação de Qualidade de Imagem e Vídeo 3D através de modelos preditivos e não um único índice. Nosso estudo apresenta as vantagens e desvantagens de aplicar técnicas de AD's para a avaliação de qualidade.

Os trabalhos correlatos nos mostraram que existe uma lacuna para trabalhar com Árvores de Decisão, que são técnicas leves e de fácil interpretação, além de serem bem estabelecidas na literatura. Além disso, podemos observar uma carência em estudos que abordem e discutem a avaliação de qualidade de referência completa e a utilização de técnicas de AM. Além de responder às questões de pesquisa, nosso estudo amplia o entendimento de modelos preditivos no âmbito de 3D-IQA e 3D-VQA, onde apresenta a capacidade que os algoritmos de AD's tem de predizer uma classe esperada (MOS). Sendo assim, quanto maior a capacidade de predição de um algoritmo, melhor tende a ser sua capacidade de inferir o MOS da imagem ou vídeo, ou seja, a qualidade percebida pelo usuário. Por sua vez, difere dos trabalhos relacionados, já que estes fornecem um valor único de qualidade, que pode ser correlacionado com testes subjetivos.

7 CONCLUSÃO E TRABALHOS FUTUROS

Esta tese explorou e apresentou o uso de técnicas de Aprendizado de Máquina leves baseadas em Árvore de Decisão para a Avaliação de Qualidade de Imagem e Vídeo 3D de Referência Completa. A importância da avaliação de qualidade, principalmente quando se trata de imagens e vídeos 3D, é reforçada pelo aumento expressivo do compartilhamento de imagens e vídeos digitais nas últimas décadas. Este crescimento foi viabilizado pela onipresença da internet e pela proliferação das mídias sociais e serviços de *streaming*. Neste contexto, passou a ser fundamental a utilização das tecnologias de codificação de imagens e vídeos. Como o processo de codificação e transmissão pode adicionar degradações à qualidade percebida pelo usuário final, torna-se cada vez mais necessário o desenvolvimento de métodos e técnicas para medir a qualidade percebida pelo usuário.

As Avaliações de Qualidade de Imagem e Vídeo utilizam métricas e métodos que visam medir a qualidade das imagens e vídeos que sofreram degradações, tornando possível a realização de melhorias na qualidade final da imagem que chega ao usuário. No entanto, a maioria destas métricas são desenvolvidas para conteúdos 2D, deixando um vasto espaço para melhorias sobre a avaliação, precisão e confiabilidade no que tange a Avaliação de Qualidade para Imagens e Vídeos 3D. Conforme citado na introdução deste trabalho, uma das principais barreiras encontradas na 3D-VQA é a busca por modelos de qualidade percebida que sejam fiéis ao modelo da percepção visual humana. Encontrar um modelo ideal e um conjunto de características que apresentem boa correlação com os modelos subjetivos pode resultar em uma tarefa pouco eficiente. Contudo, alternativas para automatizar este processo podem ser úteis para tornar esta busca mais prática. Muitas abordagens adotam modelagem do HVS (*Human Visual System*) para simular a percepção de profundidade e, em particular, a disparidade binocular induzida pelo deslocamento horizontal de recursos de imagem entre as vistas esquerda e direita. Os paradigmas de AM tornam possível tratar da tarefa de VQA sobre uma perspectiva diferente das métricas tradicionais, pois conseguem “simular” a percepção humana de qualidade ao invés de projetar um modelo explícito do HVS. Diante disso, destaca-se a importância de explorar técnicas de AM

no contexto de Avaliação de Qualidade de Vídeo 3D. Já que, sabe-se que a maioria dos estudos tanto com métricas objetivas como métodos subjetivos são normalmente desenvolvidos para Avaliação de Qualidade de Imagens e Vídeos 2D. Neste âmbito, surgem alguns questionamentos acerca de vídeos e imagens 3D, principalmente no que envolve Aprendizado de Máquina baseado em Árvore de Decisão. A escolha da utilização de AD's se deu através da ideia de que é uma técnica leve, de baixo custo computacional e de simples interpretação, além de ser comumente utilizada na literatura por diversas áreas.

A tese foi desenvolvida buscando responder às questões de pesquisa, que aqui serão retomadas:

- **Questão 1 (Q1)**: É viável aplicar técnicas baseadas em AD para avaliar a qualidade de imagem 3D?
- **Questão 2 (Q2)**: É viável aplicar técnicas baseadas em AD para avaliar a qualidade de vídeo 3D?
- **Questão 3 (Q3)**: Quais características das imagens são mais relevantes na definição de modelos para predição de 3D-IQA?
- **Questão 4 (Q4)**: Qual algoritmo baseado em AD se mostra mais promissor para avaliar a qualidade de imagem e vídeo 3D?
- **Questão 5 (Q5)**: É viável treinar modelos utilizando imagens para avaliar qualidade em vídeos?
- **Questão 6 (Q6)**: O aumento no número de classes tem influência na capacidade de predição dos algoritmos baseados em AD?

De modo geral, todos os experimentos foram realizados em modelos de treinamento utilizando os algoritmos de classificação baseados em AD, que são o *J48*, *RepTree*, *RandomForest* e *ForestPA*. Além disso, foram realizadas diferentes variações nos hiperparâmetros dos algoritmos, que geraram varições para o algoritmo *j48* e *ReTree*. A Avaliação de Qualidade de Imagem 3D, contempla dois estudos. O primeiro estudo trata de avaliar o impacto que a variação das características das imagens pode ter sobre a capacidade de predição dos algoritmos. Sendo assim, em relação à relevância das características para a definição de modelos de predição de 3D-IQA (questão de pesquisa **Q3**), destaca-se o Cenário 1, que considera todos os conjuntos de características extraídos das imagens. Pode-se interpretar que quanto mais características estão no conjunto de dados, melhor será o modelo gerado pela técnica de classificação. Entretanto, isso não tem, necessariamente, uma relação direta: o Cenário 2 teve resultados melhores que o Cenário 1 para alguns algoritmos (Fig.31).

A quantidade de características pode também interferir de forma negativa na capacidade de predição do modelo (*Curse of Dimensionality*) (GASTALDO; REDI, 2012). Além disso, pode existir alguma característica do conjunto que ao invés de contribuir para a qualidade do modelo, acabe prejudicando-o (SÁEZ et al., 2016). Sendo assim, o conjunto de características do Cenário 1 destaca-se em relação aos demais conjuntos, pois permitiu a geração de melhores modelos classificadores.

O segundo estudo avalia os resultados dos testes com modelos treinados variando o número de classes. Através da análise dos testes para os modelo gerados com imagens foi possível responder à questão de pesquisa **Q1**. Já que os valores de testes mostraram que os algoritmos baseados em ADs, dentre eles florestas, se mostraram promissores para a 3D-IQA, especialmente o *RandomForest*. Os valores de testes com modelos treinados com imagens, sendo *F1-score* com o melhor valor de 0,823 e o MAE mais baixo, com 5,592, se mostram com melhor capacidade de predizer a classe esperada quando utilizando o *RandomForest* tanto para a variação de características de imagens quanto para os testes com modelos treinados variando as classes.

A Avaliação de Qualidade de Vídeo 3D, apresenta, inicialmente, um estudo que permite responder a **Q2**, já que se baseia nos resultados dos testes com modelos treinados utilizando vídeos. Assim como observado nos estudos voltados a imagens, o algoritmo *RandomForest* também apresentou uma melhor capacidade de predizer as classes esperadas. No entanto, quando se trata de vídeos, os valores de desempenho são inferiores aos de imagem, se estabelecendo em torno de 50% de acerto, e dados como RMSE com menor valor de 12,18 e MAE com o menor valor de 9,43, fatos estes que pode ser dado devido à complexidade dos vídeos estereoscópicos. Ainda assim consideramos neste trabalho que as técnicas baseadas em ADs são viáveis para avaliar a qualidade de vídeos 3D. Além da complexidade dos vídeos, a da falta de mais bancos de vídeos 3D que forneçam junto os valores de testes subjetivos é um desafio já que a maioria dispõe de poucas sequências e com grande discrepância nos valores de MOS.

Ainda no contexto de Avaliação de Qualidade de Vídeo, o segundo estudo permite responder a quinta questão de pesquisa **Q5**, em que perguntamos se é viável classificar vídeos com modelos treinados com imagens, não se mostrou uma prática aplicável neste estudo. Alguns fatores podem contribuir para isso como a diferença na complexidade entre os vídeos e imagens, o movimento e o número elevado de imagens semelhantes. Também existem diferenças nas degradações das imagens e vídeos disponíveis nas bases de dados.

Diante da análise dos testes utilizando os modelos gerados tanto para imagens como vídeos, é possível responder à quarta questão de pesquisa **Q4**, em que foi constatado que o algoritmo *RandomForest* se mostra mais promissor em termos gerais. Através das medidas de desempenho e medidas de erros utilizadas para avaliar

os algoritmos, pode-se observar que o algoritmo *RandomForest* obteve melhores resultados nos testes na maioria dos casos. Indicando que o algoritmo apresenta uma boa capacidade para predizer os valores das classes esperadas, bem como a proximidade entre os valores reais e esperados.

A última questão de pesquisa,**Q6**, questiona se o aumento no número de classes pode influenciar na capacidade de predição dos algoritmos de AD. Neste estudo, apresentamos os valores dos testes com modelos treinados considerando o número de classes como 5,10 e 25. Através dos resultados, podemos constatar que o aumento das classes para 25 comprometeu de forma significativa a capacidade de predição dos algoritmos testados. Por vezes, pensa-se que quanto maior o número de classes melhor pode ser a capacidade de predição do algoritmo. No entanto, podemos notar que tanto os valores em termos de eficiência quanto em termos de erro mostram que o aumento de classe cria uma distância muito grande entre os valores de classes esperadas.

Durante o desenvolvimento deste trabalho, buscando responder às questões de pesquisas, algumas dificuldades foram encontradas, estas que apesar de tornar o trabalho mais custoso, ampliam nossa curiosidade e a busca por melhores contribuições para área de Avaliação de Qualidade de Imagem e Vídeo 3D. Dentre as dificuldades, a pouca disponibilidade de base de dados de testes subjetivos de vídeos 3D, que por consequência contribuiu para um grande desbalanceamento entre os valores de MOS, que foram utilizados como base para as classes dos dados; a falta de literatura, no contexto de AM, que exponha de forma clara as etapas de treinamento e teste. Diversos trabalhos indicam a técnica de AM utilizada, sem maiores detalhes. Sendo assim, como trabalhos futuros, sugere-se em ampliar o estudo com a utilização de outros algoritmos de AM para maiores comparações. Além disso, temos interesse na construção de uma Métrica Objetiva de Avaliação de Qualidade de Vídeo 3D, que contemple um índice único. Ainda, há de se pensar no desenvolvimento de uma base de testes subjetivos, que seria de grande contribuição para a comunidade científica e daria aporte a estudos como este.

REFERÊNCIAS

ADNAN, M. N.; ISLAM, M. Z. Forest PA: Constructing a decision forest by penalizing attributes used in previous trees. **Expert Systems with Applications**, [S.l.], v.89, p.389–403, 2017.

AGOSTINI, L. V. **Desenvolvimento de Arquiteturas de Alto Desempenho Dedicadas à Compressão de Vídeo Segundo o Padrão H. 264/AVC**. 2007. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Rio Grande do Sul.

AKINBO, R. S.; DARAMOLA, O. A. Ensemble machine learning algorithms for prediction and classification of medical images. **Algorithms, Models and Applications**, [S.l.], p.59, 2021.

ALI, J.; KHAN, R.; AHMAD, N.; MAQSOOD, I. Random forests and decision trees. **International Journal of Computer Science Issues (IJCSI)**, [S.l.], v.9, n.5, p.272, 2012.

ALMEIDA TEIXEIRA, L. de. **Métodos de Regressão para Aprendizado por Reforço**.

ALVARENGA, M. T. **Utilização da ferramenta j48 para descoberta do conhecimento em bases de dados fitossanitários, climáticos e espectrais**. 2014. Tese (Doutorado em Ciência da Computação) — Master thesis, Universidade Federal de Lavras, Minas Gerais, Brazil.

AMEER, S. et al. Comparative analysis of machine learning techniques for predicting air quality in smart cities. **IEEE Access**, [S.l.], v.7, p.128325–128338, 2019.

ARTHUR, R. **Avaliação objetiva de codecs de vídeo**. 2002. Dissertação (Mestrado em Ciência da Computação) — Universidade estadual de campinas - UNICAMP.

BANITALEBI-DEHKORDI, A.; POURAZAD, M. T.; NASIOPOULOS, P. An efficient human visual system based quality metric for 3D video. **Multimedia Tools and Applications**, [S.l.], v.75, n.8, p.4187–4215, 2016.

BHARGAVA, N.; SHARMA, G.; BHARGAVA, R.; MATHURIA, M. Decision tree analysis on j48 algorithm for data mining. **Proceedings of international journal of advanced research in computer science and software engineering**, [S.l.], v.3, n.6, 2013.

BOUCKAERT, R. R. et al. WEKA manual for version 3-9-3. **The University of Waikato, Hamilton, New Zealand**, [S.l.], 2018.

CAMILO, C. O.; SILVA, J. C. da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, [S.l.], p.1–29, 2009.

CHARRIER, C.; LÉZORAY, O.; LEBRUN, G. Machine learning to design full-reference image quality assessment algorithm. **Signal processing: Image communication**, [S.l.], v.27, n.3, p.209–219, 2012.

CHEN, M.-J.; KWON, D.-K.; BOVIK, A. C. Study of subject agreement on stereoscopic video quality. In: IEEE SOUTHWEST SYMPOSIUM ON IMAGE ANALYSIS AND INTERPRETATION, 2012., 2012. **Anais...** [S.l.: s.n.], 2012. p.173–176.

CHEN, Y.; WU, K.; ZHANG, Q. From QoS to QoE: A tutorial on video quality assessment. **IEEE Communications Surveys & Tutorials**, [S.l.], v.17, n.2, p.1126–1165, 2015.

CHIKKERUR, S.; SUNDARAM, V.; REISSLEIN, M.; KARAM, L. J. Objective video quality assessment methods: A classification, review, and performance comparison. **IEEE transactions on broadcasting**, [S.l.], v.57, n.2, p.165–182, 2011.

COAQUIRA BEGAZO, D. **Avaliação objetiva e subjetiva de qualidade de vídeo via rede IP com variação de atraso.** 2012. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo.

DARONCO, L. C.; ROESLER, V.; LIMA, J. V. de. Avaliaçao subjetiva de qualidade aplicada à codificaçao de vídeo escalável. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 14., 2008. **Proceedings...** [S.l.: s.n.], 2008. p.146–153.

DEVEZA, C. H. Minerando Padrões Sequenciais para Base de Dados de Lojas Virtuais. **Monografia (Curso de Bacharelado em Ciência da Computação), UFOP (Universidade Federal de Ouro Preto)**, [S.l.], 2011.

DUMIĆ, E. et al. 3D video subjective quality: a new database and grade comparison study. **Multimedia tools and applications**, [S.l.], v.76, n.2, p.2087–2109, 2017.

ESCOVEDO, T.; KOSHIYAMA, A. **Introdução a Data Science**: Algoritmos de Machine Learning e métodos de análise. [S.l.]: Casa do Código, 2020.

FACELI, K.; LORENA, A. C.; GAMA, J. a.; CARVALHO, A. C. P. d. L. F. d. **Inteligência artificial**: uma abordagem de aprendizado de máquina. [S.l.]: LTC., 2011.

FANG, Y. et al. Perceptual quality assessment for asymmetrically distorted stereoscopic video by temporal binocular rivalry. **IEEE Transactions on Circuits and Systems for Video Technology**, [S.l.], v.31, n.8, p.3010–3024, 2020.

FANG, Y.; SUI, X.; WANG, J. A Spatial-Temporal Weighted Method for Asymmetrically Distorted Stereo Video Quality Assessment. In: IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS (ISCAS), 2019., 2019. **Anais...** IEEE, 2019. p.1–5.

FONSECA, R. N. d. **Algoritmos para avaliação da qualidade de vídeo em sistemas de televisão digital.** 2008. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo.

GALKANDAGE, C.; CALIC, J.; DOGAN, S.; GUILLEMAUT, J.-Y. Stereoscopic video quality assessment using binocular energy. **IEEE Journal of Selected Topics in Signal Processing**, [S.l.], v.11, n.1, p.102–112, 2017.

GARCIA, S. C. **O uso de árvores de decisão na descoberta de conhecimento na área da saúde**. 2003. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Rio Grande do Sul.

GASTALDO, P.; REDI, J. A. Machine learning solutions for objective visual quality assessment. **6th international workshop on video processing and quality metrics for consumer electronics, VPQM**, [S.l.], v.12, 2012.

GAZZANIGA, M. S. **The cognitive neurosciences**. [S.l.]: MIT press, 2004.

GOLDMANN, L.; DE SIMONE, F.; EBRAHIMI, T. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In: THREE-DIMENSIONAL IMAGE PROCESSING (3DIP) AND APPLICATIONS, 2010. **Anais...** [S.l.: s.n.], 2010. v.7526, p.242–252.

GONZALEZ, R. C.; WOODS, R. E. **Processamento de imagens digitais**. [S.l.]: Edgard Blucher, 2000.

GURAV, P.; PATIL, G. Full-Reference Video Quality Assessment using Structural Similarity (SSIM) Index. **Journal of Electronics and Communication Systems**, [S.l.], v.1, n.2, 2016.

HUBEL, D. H. **Eye, brain, and vision**. [S.l.]: Scientific American Library New York, 1988. v.22.

HUYNH-THU, Q.; GHANBARI, M. Scope of validity of PSNR in image/video quality assessment. **Electronics letters**, [S.l.], v.44, n.13, p.800–801, 2008.

ITU-R BT.500, R. 500-11,“Methodology for the Subjective Assessment of the Quality of Television Pictures,” Recommendation ITU-R BT. 500-11. **ITU Telecom. Standardization Sector of ITU**, [S.l.], v.7, 2002.

ITU-T J.143, R. 500-11,“User requirements for objective perceptual video quality measurements in digital cable television,” Recommendation ITU-T J.143. **ITU Telecom. Standardization Sector of ITU**, [S.l.], v.7, 2000.

ITU-T J.144, R. 500-11,”Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference´´ Recommendation ITU-T J.144. **ITU Telecom. Standardization Sector of ITU**, [S.l.], 2004.

ITU-T P.910, R. P910. **Subjective video quality assessment methods for multimedia applications**, [S.l.], 2008.

KALMEGH, S. Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. **International Journal of Innovative Science, Engineering & Technology**, [S.l.], v.2, n.2, p.438–446, 2015.

LAMBRECHT, C. et al. Perceptual models and architectures for video coding applications. **Swiss Federal Institute of Technology/Lausanne**, [S.l.], 1996.

LIOTTA, A. et al. Instantaneous video quality assessment for lightweight devices. In: INTERNATIONAL CONFERENCE ON ADVANCES IN MOBILE COMPUTING & MULTIMEDIA, 2013. **Proceedings...** Association for Computing Machinery, 2013. p.525–531.

MAIMON, O. Z.; ROKACH, L. **Data mining with decision trees**: theory and applications. [S.l.]: World scientific, 2014. v.81.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, [S.l.], v.1, n.1, p.32, 2003.

NARWARIA, M.; LIN, W. Objective image quality assessment with singular value decomposition. In: FIFTH INTERNATIONAL WORKSHOP ON VIDEO PROCESSING AND QUALITY METRICS FOR CONSUMER ELECTRONICS (VPQM), 2010. **Anais...** [S.l.: s.n.], 2010.

NARWARIA, M.; LIN, W. SVD-based quality metric for image and video using machine learning. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, [S.l.], v.42, n.2, p.347–364, 2011.

OSISANWO, F. et al. Supervised machine learning algorithms: classification and comparison. **International Journal of Computer Trends and Technology (IJCTT)**, [S.l.], v.48, n.3, p.128–138, 2017.

PAN, F. et al. A locally-adaptive algorithm for measuring blocking artifacts in images and videos. In: IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS (IEEE CAT. NO. 04CH37512), 2004., 2004. **Anais...** [S.l.: s.n.], 2004. v.3, p.III–925.

PUNCHIHEWA, A.; BAILEY, D. G.; HODGSON, R. A survey of coded image and video quality assessment. In: IMAGE AND VISION COMPUTING NEW ZEALAND, 2003. **Proceedings...** [S.l.: s.n.], 2003. p.326–331.

PURVES, D. et al. **Invitación a la Neurociência**. [S.l.]: Editorial Médica Panamericana SA, 2001.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, [S.l.], v.1, n.1, p.81–106, 1986.

R-REP BT.1082-1, R. 1082-1,Studies Toward the Unification of Picture Assessment Methodology, Report ITU-R BT. 1082-1. **ITU Telecom. Standardization Sector of ITU**, [S.l.], 1990.

RASSOOL, R. VMAF reproducibility: Validating a perceptual practical video quality metric. In: IEEE INTERNATIONAL SYMPOSIUM ON BROADBAND MULTIMEDIA SYSTEMS AND BROADCASTING (BMSB), 2017., 2017. **Anais...** [S.l.: s.n.], 2017. p.1–2.

REGIS, M. C. D. **Métrica de Avaliação Objetiva de Vídeo Usando a Informação Espacial, a Temporal ea Disparidade**. 2013. Tese (Doutorado em Ciência da Computação) — Ph. D. dissertation, Federal University of Campina Grande–UFCG.

ROMANI, E. **Avaliação de qualidade de vídeo utilizando modelo de atenção visual baseado em saliência**. 2015. Dissertação (Mestrado em Ciência da Computação) — Universidade Tecnológica do Paraná.

RUSSELL, S. J. **Artificial intelligence a modern approach**. [S.l.]: Pearson Education, Inc., 2010.

SAMAT, A. et al. Evaluation of ForestPA for VHR RS image classification using spectral and superpixel-guided morphological profiles. **European journal of remote sensing**, [S.l.], v.52, n.1, p.107–121, 2019.

SESHADRINATHAN, K.; SOUNDARARAJAN, R.; BOVIK, A. C.; CORMACK, L. K. A subjective study to evaluate video quality assessment algorithms. In: IS&T/SPIE ELECTRONIC IMAGING, 2010. **Anais...** [S.l.: s.n.], 2010.

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, 2005. **Anais...** [S.l.: s.n.], 2005. p.23–25.

SHEIKH, H. R.; BOVIK, A. C. Image information and visual quality. **IEEE Transactions on image processing**, [S.l.], v.15, n.2, p.430–444, 2006.

SILVA, V. D.; ARACHCHI, H. K.; EKMEKCIOGLU, E.; KONDOZ, A. Toward an impairment metric for stereoscopic video: A full-reference video quality metric to assess compressed stereoscopic video. **IEEE transactions on image processing**, [S.l.], v.22, n.9, p.3392–3404, 2013.

SILVA, W. B. d. **Métodos sem referência baseados em características espaço-temporais para avaliação objetiva de qualidade de vídeo digital**. 2013. Tese (Doutorado em Ciência da Computação) — Universidade Tecnológica Federal do Paraná.

SILVERTHORN, D. U. **Fisiologia humana**: uma abordagem integrada. [S.l.]: Artmed, 2010.

SOUZA BARBIERI, T. T. de; GOULARTE, R. Investigating Subjectivity Criterion for Multi-video Summarization. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 2020. **Proceedings...** Association for Computing Machinery, 2020. p.137–144.

SáEZ, J. A.; GALAR, M.; LUENGO, J.; HERRERA, F. INFFC: An iterative class noise filter based on the fusion of classifiers with noise sensitivity control. **Information Fusion**, [S.l.], v.27, p.19–32, 2016.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. [S.l.]: Pearson Education India, 2016.

TANJI, M. et al. **Método híbrido para avaliação objetiva de qualidade de vídeo digital no padrão H. 264**. 2014. Tese (Doutorado em Ciência da Computação) — Universidade Estadual de Campinas.

URVOY, M. et al. NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. In: FOURTH INTERNATIONAL WORKSHOP ON QUALITY OF MULTIMEDIA EXPERIENCE, 2012., 2012. **Anais...** [S.l.: s.n.], 2012. p.109–114.

WANG, J. et al. Quality prediction of asymmetrically distorted stereoscopic 3D images. **IEEE Transactions on Image Processing**, [S.l.], v.24, n.11, p.3400–3414, 2015.

WANG, J. et al. Blind quality prediction of stereoscopic 3D images. **IS and T International Symposium on Electronic Imaging Science and Technology**, [S.l.], v.2017, n.14, p.70–76, 2017.

WANG, J.; WANG, S.; WANG, Z. Asymmetrically compressed stereoscopic 3D videos: Quality assessment and rate-distortion performance evaluation. **IEEE Transactions on Image Processing**, [S.l.], v.26, n.3, p.1330–1343, 2017.

WANG, Z.; BOVIK, A. C. Mean squared error: Love it or leave it? A new look at signal fidelity measures. **IEEE signal processing magazine**, [S.l.], v.26, n.1, p.98–117, 2009.

WANG, Z.; LU, L.; BOVIK, A. C. Video quality assessment based on structural distortion measurement. **Signal processing: Image communication**, [S.l.], v.19, n.2, p.121–132, 2004.

WANG, Z.; SHEIKH, H. R.; BOVIK, A. C. et al. Objective video quality assessment. **The handbook of video databases: design and applications**, [S.l.], v.41, p.1041–1078, 2003.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. **Climate research**, [S.l.], v.30, n.1, p.79–82, 2005.

ZEGARRA RODRÍGUEZ, D. **Proposta da métrica eVSQM para avaliação de QoE no serviço de streaming de vídeo sobre TCP.** 2014. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo.

ZHU, C.; XIONG, B. Transform-exempted calculation of sum of absolute Hadamard transformed differences. **ieee transactions on circuits and systems for video technology**, [S.l.], v.19, n.8, p.1183–1188, 2009.

ZINGARELLI, M. R. U. **ReyGlyph - codificação e reversão esteroscópixa anagliífica**. 2013. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo.

Anexos

ANEXO A – Recomendação ITU-R BT.500

A.1 características comuns de teste

A.1.1 Condições gerais de ambiente

A Itu-r bt.500 (2002) descreve as condições gerais para observação, em que apresenta dois ambientes: observação em laboratório, que tem por objetivo proporcionar condições críticas para comprovar o funcionamento do sistema; observação doméstica, que reproduz um ambiente próximo ao doméstico. Estes parâmetros são selecionados para definir um ambiente ligeiramente mais crítico que as situações normais de observação nas casas dos telespectadores.

A. Ambiente em laboratório: A ITU estabelece regras para as condições de observação dos observadores, devem ser organizadas levando em consideração:

   1. relação entre a luminância da tela inativa e o valor máximo da luminância;
   2. relação entre a luminância da tela, quando só é apresentado o nível preto em uma sala completamente escura e a correspondente do branco mais intenso;
   3. brilho e contraste;
   4. ângulo máximo de observação;
   5. relação entre a luminância do fundo do receptor de imagens ou o valor máximo da luminância da imagem;
   6. cromaticidade do fundo; e outras iluminações da sala.

B. Ambiente doméstico: Dentre as regras estabelecidas pela ITU, para ambientes domésticos, estão incluídas:

   1. relação entre a luminância da tela e o valor máximo da luminância;
   2. brilho e contraste;
   3. ângulo máximo de observação com respeito ao normal;
   4. tamanho da tela para formato de imagem 4/3 ou 16/9;
   5. processamento do monitor;

6. resolução do monitor;
7. valor máximo da luminância;
8. luminância do meio ambiente da tela (Luz incidente do ambiente projetado na tela).

A.1.2 Resolução do monitor

A resolução dos monitores profissionais, equipados com CRT (*Cathodic Ray Tube*), normalmente satisfaz os padrões necessários para realizar avaliações subjetivas em sua faixa operacional de luminância. No entanto, nem todos os monitores podem atingir um valor de pico de luminância de 200 cd / m2. Para verificar as resoluções máximas e mínimas, a ITU sugere o uso de um determinado valor de luminância. Se os aparelhos de televisão domésticos com CRT convencional forem utilizados para realizar as avaliações subjetivas, a resolução pode ser inadequada, dependendo do valor da luminância. Nesse caso, é recomendado verificar as resoluções máxima e mínima para o valor de luminância usado. Uma análise visual permite verificar a resolução. O limiar visual é considerado entre -12 e -20 dB.

A.1.3 Contraste do monitor

O contraste pode ser fortemente influenciado pela iluminância ambiental. Os monitores (CRT) profissionais dificilmente usam tecnologias para melhorar o contraste em um ambiente de alta iluminação. Sendo possível que não sigam o padrão de contraste sugerido, em caso de ambientes de alta iluminação. Os monitores domésticos utilizam tecnologias para obter um melhor contraste em um ambiente de alta iluminação. Para calcular o contraste de um determinado CRT, é preciso conhecer o coeficiente de reflexão da tela K. No melhor dos casos, o coeficiente de reflexão da tela é aproximadamente $K = 6\%$. Com um ambiente difuso, iluminância I de 200 lux e um valor de $K = 6\%$, um reflexo de luminosidade de $3,82cd/m^2$ das áreas da tela inativa é calculado através da Eq.(63).

$$L_{reflected} = \frac{1}{\pi}K \tag{63}$$

Com os valores indicados, a luminância refletida (cd/m^2) é quase 2% da luminância incidente (lux).Considera-se que os CRTs não apresentam reflexos de espelho sobre o vidro frontal. cuja influência exata sobre o contraste é difícil de quantificar porque depende das condições de iluminação. A relação de contraste é expressa por, RC, em (64).

$$RC = L_{min}/L_{max} \tag{64}$$

- L_{min} representa a luminância de áreas inativas sob iluminação ambiente (com

os valores indicados: $L_{min} = L_{zonasinativas} \times L_{refletida} = 3,82cd/m^2$)

- L_{max} é a luminância de áreas brancas sob iluminação ambiente (com valores indicados: $L_{max} = L_{branco} \times L_{refletida} = 200 + 3,82cd/m^2$)

Sendo que com esses valores se determina um $RC = 0,018$ muito próximo do valor de 0,02 indicado nas seções 2.1.1.1 e 2.1.2.1 da BT-R REC.500.

A.1.4 Fontes de sinal

A fonte de sinal fornece diretamente a imagem de referência, e a entrada para o sistema submetido a teste. Para a norma de TV a qualidade deve ser ótima. Pois, a ausência de defeitos na referência do par apresentado é essencial para obter resultados estáveis. As imagens e sequências digitais são as fontes de sinais mais reprodutíveis e podem ser trocadas entre laboratórios, para melhorar as comparações do sistema.

A.1.5 Faixa de condições e ancoragem

As sessões de avaliação devem incluir as faixas completas dos fatores submetidos à avaliação. Porém, pode haver uma aproximação com uma faixa mais restrita, apresentando também certas condições que são situadas nos extremos das escalas.

A.1.6 Observadores

Os observadores podem ser classificados como "*expert*" (experientes) ou "não *expert*" (sem experiencia). O observador *expert* já possui o conhecimento prévio das degradações da imagem que será apresentada no teste. Já o observador não expert não tem nenhum conhecimento sobre o teste. Em geral, a ITU recomenda que os observadores não estejam diretamente familiarizados com o vídeo que será exibido no teste. Além disso, recomenda que o número de observadores dependa da sensibilidade e viabilidade do procedimento de teste, e também do tamanho previsto para o resultado que se busca. Como exemplo, Itu-r bt.500 (2002) denota que para estudos de alcance limitado, de carácter exploratório, podem ser empregados menos de 15 observadores. Porém, os estudos devem considerar e informar o nível de experiência dos observadores na avaliação de qualidade. Os experimentos devem incluir o maior número de detalhes possíveis sobre as características das equipes de avaliação, para facilitar a investigação.

A.1.7 Instruções para a avaliação

De acordo com a Itu-r bt.500 (2002), os observadores devem estar completamente familiarizados com o método de avaliação, o fator de qualidade, os tipos de degradações que podem ocorrer, a escala de classificação e a sequência de vídeo e tempo. A ITU recomenda que as sequências de treinamento que demonstram a amplitude

e o tipo de degradação que serão avaliadas, sejam realizadas com imagens ilustrativas, diferentes das que serão utilizadas no teste, porém, a sensibilidade deve ser comparável.

A.1.8 Sessão de teste

A Itu-r bt.500 (2002) recomenda que uma sessão de teste dure no mínimo 30 minutos. Ao início da primeira sessão deve-se realizar, pelo menos, cinco apresentações do treinamento para estabilizar a opinião dos observadores, esses dados não devem ser levados em consideração para a avaliação de qualidade. As apresentações devem ser feitas em ordem aleatórias, no entanto, a ordem das condições de teste devem ser dispostas de maneira que efeitos de fadiga ou adaptação não interfiram nos teste.

A.1.9 Apresentação dos resultados

Devido a variação de alcance, a ITU considera inadequado a interpretação dos dados a partir da maioria dos métodos de avaliação em termos absolutos, por exemplo, a qualidade de uma sequência de vídeo. Para cada parâmetro deve ser dado um intervalo de confiança de 95% da distribuição estatística dos resultados da avaliação. Além do mais, é recomendado que os resultados sejam apresentados juntamente com informações como: detalhes da configuração do experimento; detalhes dos materiais da avaliação; tipo de fonte de imagem e de monitores; número e tipo de observadores; sistema de referência utilizado; nota da média global do experimento; pontuação média original e ajustada, e intervalo de confiança de 95%; caso haja a eliminação de um ou mais observadores durante um teste.

ANEXO B – Recomendação ITU-T P.910

B.1 características comuns de teste

B.1.1 Condições de observação

A Tabela 25 lista as condições típicas de observação utilizadas para a avaliação da qualidade do vídeo. Os conjuntos de parâmetros reais usados na avaliação devem ser especificados. Para comparar os resultados dos testes, todas as condições de observação em todos os laboratórios devem ser definidas e devem ser iguais para o mesmo tipo de teste.

Tanto o tamanho como o tipo de monitor utilizado devem ser adequados para a aplicação sob investigação. Ao apresentar sequências através de um sistema baseado em PC, as características de exibição devem ser especificadas, como, por exemplo, a densidade de pontos e densidade da tela, tipo de placa de vídeo usada e etc.

Com relação ao formato da visualização, é preferível usar toda a tela para a visualização das sequências. No entanto, quando, por algum motivo, as sequências forem exibidas em uma janela da tela, a cor de fundo da tela deve ser 50% cinza, o que corresponde a Y = U = V = 128 (U e V sem sinal).

Tabela 25 – Condições de observação

Parâmetro	**Valores**
Distância de visão	1-8 H
Luminância máxima da tela	100-200 cd/m
Relação da luminância da tela inativa para a luminosidade do pico	<= 0,05
Relação da luminância da tela, ao exibir apenas um nível de preto em um ambiente totalmente escuro, ao que corresponde ao pico branco	$<= 0,1$
Relação da luminância do fundo detrás do monitor de imagem para a luminância de pico da imagem	$<= 0,2$
Cromaticidade do fundo	D_{65}
Iluminação do ambiente de fundo	$<= 20$ lux

B.1.2 Sistema de processamento e reprodução

Existem dois métodos para obter imagens de teste a partir de gravações de origem:

1. através de transmissões ou reproduções de gravações de vídeos em tempo real,

realizadas por sistemas em teste, enquanto os sujeitos observam e respondem;

2. processar as gravações de origem *offline* através do dispositivo em teste e gravar a saída para produzir um novo conjunto de gravações.

No segundo caso, um VTR (*Video Tape Recorder*) digital deve ser utilizado para minimizar as degradações que podem ocorrer no processo de gravação. Em qualquer caso, considerando que as degradações introduzidas pelos esquemas de codificação de baixa taxa de bits são geralmente mais evidentes do que as degradações inseridas através da modulação, podem ser usados VTRs de qualidade profissional, como D2, MII e BetacamSP (ITU-T P.910, 2008). Pode ser utilizado um CRT, LCD, plasma, projetor ou outro tipo de monitor, levando em consideração o tipo de aplicação e a finalidade do experimento. O tamanho e o tipo de monitor utilizado deve ser apropriado para o aplicativo que está sendo investigado. Os monitores devem ser ajustados de acordo com os procedimentos definidos em ITU-R BT.814-1.

B.1.3 Observadores

A Itu-t p.910 (2008) destaca que o número possível de sujeitos em um teste de observação varia de 4 a 40. Quatro é o mínimo absoluto por razões estatísticas, enquanto é difícil obter maiores vantagens com mais de 40 indivíduos.

O número real para um determinado teste deve ser estabelecido, na prática, de acordo com a validade necessária e a necessidade de generalizar uma amostra para uma população maior.

Geralmente, pelo menos 15 observadores devem participar do experimento. Eles não devem intervir diretamente nas avaliações de qualidade de imagem como parte de seu trabalho habitual e não devem ser avaliadores "*experts*". No entanto, nos estágios iniciais do desenvolvimento de sistemas de comunicação de vídeo e em experimentos testes (treinamentos) realizados antes de um teste maior, pequenos grupos de "*experts*" (4-8) ou outros sujeitos críticos podem fornecer resultados indicativos.

Normalmente, antes de uma sessão, os observadores devem ser examinados para determinar sua acuidade visual normal, acuidade corrigida ao normal e sua visão de cor normal. No que diz respeito à acuidade, nenhum erro deve ser feito na linha 20/30 de um diagrama de visão normalizado [*b-Snellen*]. O diagrama deve ser classificado para a distância de observação do teste e o teste de acuidade deve ser realizado no mesmo local onde as sequências do vídeo serão observadas (por exemplo, suportando o diagrama de visão no monitor). No que diz respeito à cor, não deve perder mais de 2 placas [*b-Beck*] de um total de 12.

B.1.4 Instruções aos observadores e sessão de instrução

Antes de iniciar o experimento, o assunto deve ser explicado, assim como o cenário da aplicação do sistema em teste também. E uma descrição do tipo de avaliação, da escala de opinião e do tipo de apresentação dos estímulos deveram ser entregues por escrito. O intervalo e o tipo de deficiências devem ser apresentados em testes preliminares, que podem conter sequências de vídeo diferentes das utilizadas nos testes reais. A ITU observa que não se deve deduzir que a pior qualidade observada na sessão de treinamento corresponde necessariamente ao menor grau subjetivo da escala. As perguntas sobre o procedimento ou o significado das instruções devem ser respondidas com cuidado para não influenciar as avaliações e somente antes da sessão começar.

B.1.5 Análises estatístico e resultados

De acordo com a Itu-t p.910 (2008) os resultados devem ser relatados juntamente com os detalhes do *layout* experimental. Para cada combinação das variáveis de teste, o valor médio e o desvio padrão da distribuição estatística das notas de avaliação devem ser indicados.

A confiabilidade dos observadores deve ser calculada a partir dos dados e o método utilizado para avaliar essa confiabilidade deve ser relatado. Na ITU-R BT.500-9, são apresentados alguns critérios relativos à confiabilidade subjetiva. É considerável analisar a distribuição cumulativa de notas de opinião. Uma vez que as distribuições cumulativas não dependem de linearidade, elas podem ser particularmente úteis para dados cuja linearidade é duvidosa, como os obtidos através dos métodos ACR e DCR, juntamente com escalas por categorias sem gradações (por exemplo, classificação por categoria).

Para avaliar a importância dos parâmetros de teste, a ITU-P910 recomenda que técnicas clássicas de análise de variância sejam utilizadas. Se o teste for orientado para a avaliação da qualidade do vídeo como uma função de um parâmetro, pode ser conveniente usar técnicas de ajuste de curva para interpretar os dados.

No caso das comparações em pares, o método de cálculo da posição de cada estímulo em uma escala de intervalo, quando a diferença entre os estímulos corresponde à diferença nas preferências.