UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E INFORMÁTICA INDUSTRIAL

EDUARDO ROMANI

AVALIAÇÃO DE QUALIDADE DE VÍDEO UTILIZANDO MODELO DE ATENÇÃO VISUAL BASEADO EM SALIÊNCIA

DISSERTAÇÃO

CURITIBA
2015

EDUARDO ROMANI

AVALIAÇÃO DE QUALIDADE DE VÍDEO UTILIZANDO MODELO DE ATENÇÃO VISUAL BASEADO EM SALIÊNCIA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de "Mestre em Ciências" – Área de Concentração: Telecomunicações e Redes.

Orientador: Prof. Dr. Alexandre de Almeida Prado Pohl

Co-orientador: Dr. Wyllian Bezerra da Silva

CURITIBA
2015

À minha namorada, Carla, e aos meus pais, Maristela e Amauri, com admiração, carinho e gratidão pela compreensão e apoio incondicional.

AGRADECIMENTOS

Pelo incentivo, apoio e amor agradeço à minha namorada Carla, aos meus pais Maristela e Amauri e ao meu irmão Maurício.

Ao Professor Alexandre Pohl pela amizade, incentivo e exímia orientação no decorrer desta dissertação.

Ao Professor Wyllian Bezerra pela amizade, apoio e contribuições.

Ao Professor Dubravko da Universidade de Novi Sad (Sérvia) por conceder o algoritmo de extração de saliências.

À Professora Keiko pelas dicas e colaborações.

Por fim, agradeço ao corpo docente e funcionários da UTFPR que contribuíram nesta etapa de minha vida.

“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê”.

Arthur Schopenhauer

RESUMO

ROMANI, Eduardo. AVALIAÇÃO DE QUALIDADE DE VÍDEO UTILIZANDO MODELO DE ATENÇÃO VISUAL BASEADO EM SALIÊNCIA. 75 f. Dissertação – Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial, Universidade Tecnológica Federal do Paraná. Curitiba, 2015.

A avaliação de qualidade de vídeo possui um papel fundamental no processamento de vídeo e em aplicações de comunicação. Uma métrica de qualidade de vídeo ideal deve garantir a alta correlação entre a predição da distorção do vídeo e a percepção de qualidade do Sistema Visual Humano. Este trabalho propõe o uso de modelos de atenção visual com abordagem bottom-up baseados em saliências para avaliação de qualidade de vídeo. Três métricas objetivas de avaliação são propostas. O primeiro método é uma métrica com referência completa baseada na estrutura de similaridade. O segundo modelo é uma métrica sem referência baseada em uma modelagem sigmoidal com solução de mínimos quadrados que usa o algoritmo de Levenberg-Marquardt e extração de características espaço-temporais. E, a terceira métrica é análoga à segunda, porém usa a característica *Blockiness* na detecção de distorções de blocagem no vídeo. A abordagem *bottom-up* é utilizada para obter os mapas de saliências que são extraídos através de um modelo multiescala de *background* baseado na detecção de movimentos. Os resultados experimentais apresentam um aumento da eficiência de predição de qualidade de vídeo nas métricas que utilizam o modelo de saliência em comparação com as respectivas métricas que não usam este modelo, com destaque para as métricas sem referência propostas que apresentaram resultados melhores do que métricas com referência para algumas categorias de vídeos.

Palavras-chave: Avaliação de Qualidade de Vídeo, Atenção Visual, Modelos de Saliência, Métricas Objetivas.

ABSTRACT

ROMANI, Eduardo. VIDEO QUALITY ASSESSMENT USING VISUAL ATTENTION MODEL BASED ON THE SALIENCY. 75 f. Dissertação – Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial, Universidade Tecnológica Federal do Paraná. Curitiba, 2015.

Video quality assessment plays a key role in the video processing and communications applications. An ideal video quality metric shall ensure high correlation between the video distortion prediction and the perception of the Human Visual System. This work proposes the use of visual attention models with bottom-up approach based on saliencies for video qualitty assessment. Three objective metrics are proposed. The first method is a full reference metric based on the structural similarity. The second is a no reference metric based on a sigmoidal model with least squares solution using the Levenberg-Marquardt algorithm and extraction of spatial and temporal features. And, the third is analagous to the last one, but uses the characteristic Blockiness for detecting blocking distortions in the video. The bottom-up approach is used to obtain the salient maps, which are extracted using a multiscale background model based on motion detection. The experimental results show an increase of efficiency in the quality prediction of the proposed metrics using salient model in comparission to the same metrics not using these model, highlighting the no reference proposed metrics that had better results than metrics with reference to some categories of videos.

Keywords: Video Quality Assessment, Visual Attention, Salient Model, Objective Metrics

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

3D	3 Dimensões
AQV	Avaliação de Qualidade de Vídeo
C_b	Sinal de Crominância Azul
CIF	*Common Intermediate Format*
C_r	Sinal de Crominância Vermelho
dB	Decibel
DCT	*Discrete Cosine Transform*
DMOS	*Differential Mean Opinion Score*
DMOS	*Differential Mean Opinion Score*
fps	frames per second
FR	*Full Reference*
HD	*High Definition*
IEC	*International Electrotechnical Commission*
ISO	*International Organization for Standardization*
ITU	*International Telecommunication Union*
IVP	*Image and Video Processing*
JPEG	*Joint Photographic Experts Group*
LIVE	*Laboratory for Image and Video Engineering*
LM	Levenberg-Marquardt
LST-SSIM	*Local Spatial-Temporal Structural SIMilarity*
MAD	*Mean Absolute Difference*
MADw	*Mean Absolute Difference weighted*
MC-SSIM	*Motion Compensated Structural SIMilarity*
MOS	*Mean Opinion Score*
MOVIE	*MOtion-based Video Integrity Evaluation index*
MPEG	*Moving Picture Experts Group*
MSE	*Mean Square Error*
MS-SSIM	*Multiescale Structural SIMlirarity*
MV	*Motion Vector*
NR	*No Reference*
NRLM-SMb	*No Reference metric using Levenberg-Marquardt method based on the Salient Model and Blockiness Feature*
NRLM-SM	No Reference metric using Levenberg-Marquardt method based on the Salient Model
NRVQA-LM	*No-Reference Video Quality Assessment based on Levenberg-Marquardt*
PLCC	*Pearson Linear Coefficients Correlation*
PLCC	*Pearson Linear Correlation Coefficient*
PSNR	*Peak to Signal-to-Noise Ratio*
PVS	*Processed Video Sequences*
ROI	*Region of Interest*
RR	*Reduced Reference*

SI	*Spatial Information*
SROCC	*Spearman rank order correlation coefficient*
SSIM	*Structural SIMmilarity*
SSIM-SM	*Structural Similarity based on the Salient Model*
STRRED	*Spatio-Temporal-Reduced Reference Entropic Differences*
SVH	Sistema Visual Humano
SW-SSIM	*Speed-Weighted Structural SIMilarity*
TI	*Temporal Information*
TI	*Temporal perceptual Information*
UTFPR	Universidade Tecnológica Federal do Paraná
VIPSL	*Video/Image Processing System Labs*
VQAS	*Video Quality Assessment Scores*
VQEG	*Video Quality Experts Group*
VQM	*Video Quality Metric*
WMBER	*Weighted Macro-Block Error Rate*
Y	Luminância

SUMÁRIO

1 INTRODUÇÃO

É crescente a quantidade de vídeos digitais disponíveis atualmente. Os avanços tecnológicos na área de telecomunicações propiciam a rápida geração e transmissão de conteúdo digital inserido em nosso cotidiano. A transmissão de vídeo digital é feita pela internet, com servidores que dispõembbc de vídeos armazenados ou de transmissões em tempo real. Ou pela TV digital, que passa pelo processo de implantação no Brasil desde 2007 e possui prazo de desligamento total do sinal analógico programado para o ano de 2018. Com isto, o usuário passa a ter acesso a um conteúdo com mellhor qualidade de imagem e som. Porém novos desafios são criados para a avaliação da qualidade do vídeo (AQV) recebido em cada televisor, celular, *tablet*, *notebook*, entre outros.

Avaliação da qualidade de vídeo digital é ferramenta essencial nos processos de transmissão de vídeos, uma vez que o conteúdo pode ser afetado por diversos ruídos, assim, seu objetivo é averiguar qual o nível de degradação da qualidade do vídeo causado pelos ruídos, buscando garantir uma boa qualidade de vídeo. Neste intuito, a AQV busca averiguar a qualidade do vídeo que o usuário acessa através de métricas que busquem reproduzir a percepção do Sistema Visual Humano (SVH). Neste contexto destacam-se as métricas de avaliação sem referência, NR (*No Reference*), que avaliam o vídeo recebido sem a necessidade do original, e métricas que simulam aspectos da visão humana, como os modelos de atenção que denotam regiões que atraem o foco visual e assim têm maior influência na qualidade de experiência do usuário.

Nesta dissertação são propostos três métodos de avaliação de qualidade de vídeo utilizando modelos de atenção baseados em extração de saliências dos vídeos, sendo um deles uma métrica com referência, FR (*Full Reference*) e os outros dois uma métrica NR. O método FR proposto é denominado SSIM-SM (*Structural Similarity based on the Salient Model*) e os métodos NR propostos são denotados como NRLM-SM (*No Reference Video Quality Assessment using Levenberg-Marquardt method based on the Salient Model*) e o NRLM-SMb (*No Reference Video Quality Assessment using Levenberg-Marquardt method based on the Salient Model and Blockiness Feature*).

As seções seguintes apresentam a motivação, os objetivos, o estado da arte e as contribuições da dissertação.

1.1 MOTIVAÇÃO

Nos últimos anos a inserção de modelos de saliência na avaliação de qualidade de vídeo apresenta um crescente interesse da comunidade acadêmica devido à busca de reprodução da percepção visual humana em métricas de predição de qualidade de vídeo (MAHAPATRA *et al.*, 2008; ĆULIBRK *et al.*, 2011; MA *et al.*, 2011; FENG *et al.*, 2011; WANG *et al.*, 2012; LIN *et al.*, 2012; FU *et al.*, 2013; LUO *et al.*, 2014).

As métricas objetivas de avaliação de vídeo utilizam, em sua maioria, características extraídas da sequência de vídeo para realizar a predição de qualidade. Estas características buscam reproduzir a percepção do SVH. Porém, a sequência de um vídeo apresenta muitas informações ao mesmo tempo em todas as regiões dos quadros, onde nem todas as regiões estão sendo observadas atentamente pelo espectador. Assim, uma característica extraída de um quadro como um todo pode ter seu valor influenciado por uma região em que o olho humano não fixou a atenção, e por isto, influencia menos na avaliação subjetiva do usuário, tornando o valor da respectiva característica uma fonte de divergência entre a avaliação objetiva e subjetiva. Desta forma, os modelos de saliência auxiliam a eliminar esta divergência e buscar uma avaliação mais eficiente.

Os modelos de saliência podem ser inseridos em todos os tipos de avaliação objetiva, sendo elas FR, RR (*Reduce Reference*) ou NR, já que os mapas de saliência podem ser extraídos tanto do vídeo original como do processado. Porém, os mapas de saliência podem ser obtidos de diferentes maneiras, com análise das características espaciais (tais como orientação, contraste, cor e intensidade) ou com análise de regiões salientes devido a discrepâncias de movimentos, que podem ter uma influência maior no SVH.

Deste modo, esta dissertação desenvolve três métricas objetivas de AQV, uma FR e duas NR, todas baseadas em modelos de saliência. Os mapas de saliência utilizam uma abordagem tanto espacial quanto temporal (análise de movimentos), buscando abranger o máximo de pontos de fixação da atenção visual ao decorrer do vídeo. A métrica FR utiliza a estrutura de similaridade para comparar o vídeo original e o processado com possíveis distorções. Já a métrica NR tem uma abordagem analítica sigmoidal, que pondera valores a características espaço-temporais através de um algoritmo de otimização. Além de abordarem diferentes tipos de avaliação objetiva (FR e NR), a escolha destas métricas foi feita com base na difusão dos

métodos na literatura e na eficiência de predição para grupos de vídeos no caso das métricas FR e NR, respectivamente.

1.2 OBJETIVOS

A seguir são descritos os objetivos que guiaram esta dissertação.

1.2.1 OBJETIVO GERAL

Desenvolver métricas de avaliação de qualidade de vídeo digital baseados em modelos de saliência.

1.2.2 OBJETIVOS ESPECÍFICOS

1. Analisar a influência de modelos de atenção visual baseada em saliência na avaliação de qualidade de vídeo;
2. Extrair mapa de saliências e características espaço-temporais das bases de vídeos;
3. Inserir modelo de saliência em métodos de avaliação de qualidade de vídeo;
4. Comparar a eficiência de métodos de VQA baseados em modelo de saliência com métricas disponíveis na literatura;
5. Discutir resultados e apontar novos caminhos e futuros trabalhos.

1.3 ESTADO DA ARTE

Um grande número de pesquisas vem sendo desenvolvidas na área de avaliação de qualidade de vídeo. Esta tendência é impulsionada pelo crescimento na quantidade de mídia digital e nas várias dificuldades encontradas em conseguir uma avaliação eficiente e que atue em tempo real no dispositivo de exibição do vídeo, já que este sofre efeitos de degradação em vários estágios, desde a geração, codificação, transmissão, até a reprodução no *display* de exibição (SINGH; AGGARWAL, 2014).

Uma das razões que dificultam a avaliação de qualidade é o funcionamento do SVH que é complexo, sendo afetado por diversas características cognitivas. Assim, para realizar a avaliação objetiva de qualidade de vídeo é necessário que as características do vídeo que afetam

o SVH sejam extraídas e mensuradas. O processo de extração de características de vídeos digitais é explorado nos seguintes artigos selecionados.

Babu *et al.* (2008) propuseram uma métrica NR baseada na extração de uma característica de blocagem denominada *blockiness* que utiliza a ideia de que o gradiente de borda de bloco (bloco 8 x 8 *pixels*) pode ser mascarado por uma região com alta atividade espacial, assim, pode-se observar que o *blockiness* é percebido em um quadro com baixa atividade espacial. Esta característica é observada em compressões baseadas em blocos, como ISO/IEC e padrões de códigos ITU (e.g. MPEG-1/2/3, H.261/3/4). A escala de avaliação da métrica é entre 0 e 1, onde 0 corresponde a um quadro sem *blockiness* e 1 a um cenário onde todos os blocos são visíveis. O resultado da métrica sem referência baseada em *blockiness* varia de acordo com a taxa de compressão, e apresenta forte correlação entre a percepção visual de artefatos de blocagem e as medidas resultantes da métrica.

Shabani *et al.* (2012) fazem um estudo entre diferentes características espaço-temporais salientes. As características avaliadas são divididas principalmente em dois grupos, as baseadas em detectores de estruturas e a de detectores de movimentos, sendo que no primeiro grupo são analisadas as características 3D Harris (LAPTEV *et al.*, 2007) e 3D Hessian (WILLEMS *et al.*, 2008) que utilizam um filtro 3D gaussiano simétrico, e no grupo de detectores de movimentos são abordadas as Cuboids (DOLLAR *et al.*, 2005) e as características de movimento assimétricas (SHABANI *et al.*, 2011) que localizam os eventos de movimento salientes tratando as diferenças no domínio do tempo. Como resultado, os autores apontam duas principais observações. (1) Os detectores baseados em movimento são mais eficientes que os baseados em estruturas. (2) As características de detecção de movimentos salientes utilizando filtros assimétricos têm um desempenho melhor do que os que utilizam filtros simétricos e amostragem densa.

A avaliação objetiva de qualidade de vídeo vem evoluindo muito nos últimos anos, apresentando métricas com bom desempenho de correlação com as medidas subjetivas. As métricas que apresentam melhores resultados são as métricas com referência (FR). Na sequência são apresentados alguns trabalhos que apresentam propostas de novas métricas FR.

Gao *et al.* (2010) propuseram uma métrica FR para avaliação de vídeo baseada em estrutura de similaridade utilizando um modelo de saliência com diferentes pesos. Como a métrica é FR, o desenvolvimento é feito comparando o quadro do vídeo original com o quadro do vídeo distorcido. Assim, o autor utiliza três etapas para realizar a avaliação. (1) Extração do mapa de estruturas do vídeo original e processado, desenvolvido por Wang *et al.* (2004). (2) Extração dos mapas de saliências dos vídeos processados e de referência. (3) Ponderação

das regiões salientes com diferentes pesos de acordo com seu impacto no SVH. A partir disto, é comparada a similaridade estrutural entre as regiões salientes com pesos dos vídeos e então comparam-se os resultados com a avaliação subjetiva. Os vídeos utilizados são da base de dados VIPSL (*Video/Image Processing System Labs*). Por fim, os resultados experimentais apresentaram PLCC (*Pearson Linear Correlation Coefficient*) de 0,8789.

Li e Bovik (2010) desenvolveram duas métricas FR baseadas na ideia de que diferentes regiões de uma imagem possuem diferentes significância na percepção de qualidade. Assim, os autores propuseram as métricas 3-SSIM e 3-PSNR, utilizando três modelos de componentes da imagem, a saber: os mapas de bordas, de textura e de suavidade. Cada um destes componentes foi combinado respectivamente com uma adaptação das métricas SSIM (WANG *et al.*, 2004) e PSNR, e por fim foram ponderados com um peso diferente buscando uma melhor correlação com o SVH. Os banco de vídeos utilizados para obter a correlação com as medidas subjetivas foram o banco de dados *Video Quality Experts Group* (VQEG) *Phase 1 test* (VQEG, 2000) e o banco de dados *Laboratory for Image and Video Engineering* (LIVE) *Image Quality* (SHEIKH *et al.*, 2003). Os resultados experimentais para o banco de dados VQEG *Phase* 1 apresentaram valores de PLCC de 0,796 e SROCC (*Spearman Rank Order Correlation Coefficient*) de 0,809 para a métrica 3-PSNR e valores de PLCC de 0,87 e SROCC de 0,865 para a métrica 3-SSIM.

Na maioria das situações reais de avaliação de qualidade o vídeo original não está disponível para ser comparado com o vídeo processado, sendo assim, as métricas NR têm uma função primordial nessa avaliação. Contudo, a dificuldade na predição de qualidade de vídeos com métricas NR é maior devido ao uso de informação apenas do vídeo processado. Assim, a pesquisa sobre essa técnica vem crescendo nos últimos anos, buscando melhorar este processo para obter uma avaliação mais eficiente. Alguns artigos apresentados a seguir mostram alguns avanços em métricas objetivas NR.

Silva e Pohl (2012) apresentaram uma métrica NR de avaliação objetiva baseada em um método analítico de atribuição de pesos aos descritores espaço-temporais. Estes pesos são calculados através do uso do algoritmo de Levenberg-Marquardt (LM) (MARQUARDT, 1963) com a solução de um problema de mínimos quadrados não-linear. O método proposto combina características espaciais para detecções de blocagem e borramento e características temporais com a Informação Temporal TI (*Temporal Information*) e a Média das Diferenças Absolutas MAD (*Mean Absolute Difference*). Os detectores espaciais utilizados são as características A, B e Z descritos por Wang *et al.* (2002). Em seguida, são atribuídos pesos a estas características através de β's, obtidos e otimizados através do algoritmo LM. Após aplicada esta metodologia é feita a comparação com as medidas subjetivas (DMOS - *Differential Mean Opinion Score*)

da base de dados de vídeos IVP (*Image and Video Processing*). O desempenho da métrica é comparável a métricas FR. Como um exemplo, cita-se o valor de PLCC de 0,936 para a avaliação de vídeos codificados com a técnica de compressão Dirac.

Konuk *et al.* (2013) propuseram um algoritmo de avaliação objetiva NR de qualidade de vídeo que utiliza a dimensão espacial, as características temporais, a taxa de bits (*bit rate*) e a perda de pacotes (*packet loss ratio*) para compor a avaliação. A análise da dimensão espacial é feita com a característica SI (*Spatial Information*) conforme recomendado pela ITU (1999) (*International Telecommunication Union*). O estudo da informação temporal é baseado nos vetores de movimento (MV - *Motion Vector*) e na contagem de mudança de direção destes vetores, dados por Z. Também são levados em conta o *BR* (*Bit Rate*) e a perda de pacotes dada por β. Estes parâmetros são então equacionados com diferentes pesos e seu resultado é analisado com o banco de dados LIVE. Esta abordagem possui uma correlação para os vídeos da categoria H.264, com um valor de PLCC e SROCC de, respectivamente, 0,8122 e 0,8026.

Zhu *et al.* (2014) desenvolveram um modelo NR de predição de qualidade de vídeo baseado na DCT (*Discrete Cosine Transform*). O modelo têm dois estágios: (1) medida da distorção, (2) mapeamento não linear. No primeiro caso são extraídas seis características para quantificar as distorções naturais dos quadros, como acuidade, presença de picos, suavidade e blocagem. Já no segundo caso, cada uma das características é transformada do nível de quadro para o nível de característica do vídeo através de um tratamento no domínio do tempo. Por fim, uma rede neural multicamada utiliza as características como entrada e tem como saída uma nota para a predição de qualidade do vídeo. Este método é aplicado em quatro bancos de dados, destacando os valores de 0,7855 e 0,8031, respectivamente, de PLCC e SROCC para o conjunto de todos os vídeos do banco de dados LIVE (SHEIKH *et al.*, 2003).

Na busca de uma melhor correlação da predição com as medidas subjetivas, uma das abordagens que busca aproximar a avaliação objetiva da percepção do SVH são os modelos que utilizam a atenção visual como um método de investigar as regiões. Neste cenário, os modelos de saliência se destacam nos quais são consideradas características espaciais e temporais do vídeo para definir regiões de atenção. Na sequência são apresentados dois estudos dos principais aspectos de saliência importantes para a qualidade de vídeo e imagens.

Mahapatra *et al.* (2008) estudam a influência dos movimentos na atenção visual humana e comparam esta com outras características como orientação e intensidade. Assim, é desenvolvido um modelo de saliência baseado em movimentos, integrando a informação de vetores de movimentos com a coerência espacial e temporal para gerar um mapa de atenção de movimentos. Para avaliar a eficiência do mapa de saliência de movimentos proposto foi utili-

zado um método de fixação visual, que faz o rastreamento visual, e através dos movimentos dos olhos estipula as posições onde o olho fixou a atenção. Os resultados obtidos mostram um bom desempenho do modelo saliente e também denotam uma grande influência do movimento na atenção humana, onde alguns movimentos atraem a atenção instantaneamente como um pássaro cruzando uma cena ou o movimento de mãos, e até mesmo o movimento ondulatório de uma superfície com água é percebido pelo SVH. Por fim, o autor constata que o mapa de saliência de movimentos tem maior influência no SVH do que o de características de baixo nível, como intensidade e orientação.

Winkler (2012) discute as características da visão humana, focando na anatomia e fisiologia dos componentes do sistema visual. Também explora alguns fenômenos da percepção visual com relevância em vídeos digitais. Dentre os muitos pontos abordados pelo autor, vale resaltar sua abordagem sobre o movimento dos olhos. Particurlamente, cita a pesquisa de Yarbus (1967) que mostra que os padrões de movimentos variam de acordo com a cena e com a mensagem cognitiva. Também relata o fato da direção focal não ser própria de cada pessoa, onde um número significante de pessoas podem manter o foco na mesma região da imagem (STELMACH; TAM, 1994) (ENDO *et al.*, 1994). Ainda observa os modelos de saliência como uma boa alternativa para definir as regiões mais influentes ao SVH (ITTI; KOCH, 2000), baseando-se em características como cor, intensidade, orientação, contraste e movimentos. Porém, ressalta que este modelo é puramente dirigido por estímulos, tendo uma aplicabilidade limitada à vida real, onde o conteúdo semântico afeta o movimento dos olhos (HENDERSON *et al.*, 2007).

Os modelos de saliência vêm sendo inseridos em métricas objetivas de avaliação de vídeo buscando, através de uma análise de atenção visual, aproximar-se da avaliação da percepção humana. A seguir são apresentados trabalhos que desenvolveram métricas FR baseadas em modelos de atenção.

Fu *et al.* (2013) desenvolveu um método FR de avaliação de qualidade de vídeo usando similaridade estrutural baseada em atenção visual. Este método utiliza mapas de saliência para demonstrar as regiões onde o olho humano geralmente focaliza o mapa de saliência obtido por dois modelos integrados, um espacial e outro temporal. O modelo de saliência espacial leva em conta características como orientação, crominância e intensidade, já o método de saliência temporal utiliza características de movimento baseados no modelo de *background* desenvolvido por Culibrk *et al.* (2009). Os resultados experimentais foram simulados com a base de dados EPFL-PoliMI (SIMONE *et al.*, 2009), obtendo valor de PLCC de 0,9270.

Luo *et al.* (2014) propuseram uma métrica FR baseada em modelos de saliências e informação de textura. Após extrair o mapa de textura e o mapa de saliência para ambos os

vídeos original e degradado, o método utiliza as métricas MSE (*Mean Square Error*) e SSIM para comparar as regiões de cada mapa. Em seguida, uma rede neural avalia o desempenho da métrica utilizando os dados da base LIVE. Os resultados experimentais apresentaram desempenho superior às métricas MOVIE (SESHADRINATHAN; BOVIK, 2010) e VQM (*Video Quality Metric*) (PINSON; WOLF, 2004) para as categorias de vídeo MPEG-2 e H.264, com valor de PLCC de 0,9281 e 0,9174, respectivamente.

Akamine e Farias (2014) propuseram o uso de modelos computacionais de atenção visual na avaliação de qualidade de vídeo. Todos os modelos de atenção visual utilizados são *bottom-up*, ou seja, baseadas em características presentes no vídeo. As métricas utilizadas para os testes foram a SSIM, MS-SSIM, VQM, MOVIE, Spatial-MOVIE e Temporal-MOVIE. Após a extração dos mapas de saliências das bases LIVE e IRCCyN/IVC *eyetracker* SD é feita a inserção da informação de saliência para cada métrica, gerando novos resultados que são comparados com as métricas originais. Cada métrica tem uma maneira própria de utilizar a informação dos mapas de saliências. Para SSIM é feito o uso dos mapas de erros, já para as outras métricas que não geram um mapa de características são necessárias adaptações. Também é testada a influência dos mapas de saliências subjetivos (obtidos experimentalmente) para a base IRCCyN/IVC. Com os resultados observou-se melhora de desempenho para as métricas que utilizam os modelos de saliência. Como esperado, os melhores resultados foram obtidos com o uso dos mapas de saliência subjetivos. Constatou-se que os maiores ganhos foram obtidos por métricas com apenas informação espacial (SSIM e Spatial-MOVIE).

Partindo do mesmo princípio da inserção de modelos de saliência em métricas FR, trabalhos também vêm sendo desenvolvidos com métricas NR, como apresentados na sequência.

Boujut *et al.* (2011) desenvolveram uma métrica NR baseada em mapas de saliência para avaliação de qualidade de vídeo de TV HD. A métrica proposta é denominada de WM-BER (*Weighted Macro-Block Error Rate*) e se baseia em detecção de erros em macroblocos e na extração de mapas de saliência. Esta utiliza dois mapas de saliência, um espacial e outro temporal, porém faz isto através de um método de decodificação parcial, conseguindo assim atuar em tempo real. O resultado é uma métrica com correlação com as medidas subjetivas comparável à métrica FR MSE.

Lin *et al.* (2012) introduziram uma nova métrica NR de avaliação de qualidade de vídeo baseada na região de interesse (ROI - *Region of Interest*). O método desenvolvido utiliza informações espaciais de distorções de blocagem e borramento para avaliação espacial. Já a ROI é extraída através das informações de vetores de movimentos contidas no *bitstream*, se baseando no fato do SVH ser sensível a movimentos. Assim, são estipulados pesos diferentes

para as áreas com mais e menos movimentação, para, por fim, delimitar a ROI. Utilizando uma sequencia de vídeos CIF (*Common Intermediate Format*) com resolução de 352 *x* 288 *pixels* foi obtido um valor de PLCC de 0,8113 e SROCC de 0,8042, apresentando melhor desempenho do que as métricas PSNR, SSIM (WANG *et al.*, 2004) e R-SSIM (WANG *et al.*, 2009).

Após definir o estado da arte, vale a pena destacar a inserção das métricas propostas nesta dissertação em duas das principais linhas de pesquisa nos últimos anos em avaliação de qualidade de vídeo, que são o uso de modelos de atenção visual em avaliação de qualidade de vídeo e a busca por métricas NR com boa correlação na predição de qualidade.

1.4 CONTRIBUIÇÕES

As principais contribuições desta dissertação são:

1. Desenvolvimento da métrica com referência SSIM-SM baseada na similaridade estrutural entre as regiões salientes dos vídeos originais e degradados das bases de vídeos.

2. Desenvolvimento das métricas sem referência NRLM-SM e NRLM-SMb que obedecem uma modelagem sigmoidal e utilizam características espaço-temporais salientes para o processamento de qualidade através do método iterativo de Levenberg-Marquardt.

1.5 ORGANIZAÇÃO DA DISSERTAÇÃO

A dissertação está organizada da seguinte forma:

O capítulo 2 apresenta a fundamentação teórica referente à avaliação de qualidade de vídeo e aos modelos de saliência.

O capítulo 3 descreve a metodologia no desenvolvimento das métricas FR e NR baseadas em modelos de saliência. Aborda as bases de vídeo, métodos de extração de características espaciais e temporais, extração dos mapas de saliência e os modelos estatísticos usados na comparação entre as métricas. Também apresenta as métricas propostas SSIM-SM, NRLM-SM e NRLM-SMb.

O capítulo 4 apresenta os resultados e discussões das métricas propostas através de sua correlação com medidas subjetivas e comparação com resultados de métricas disponíveis na literatura.

O capítulo 5 conclui a dissertação e expõe os caminhos para futuros trabalhos.

1.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo introduz os temas de avaliação objetiva de qualidade de vídeo e dos modelos de atenção visual dados pelos mapas de saliência. A motivação e objetivos geral e específicos da dissertação são apresentados, além de expor o estado da arte e as contribuições do trabalho. No próximo capítulo são apresentados os fundamentos teóricos necessários para o desenvolvimento das métricas propostas.

2 FUNDAMENTOS TEÓRICOS

Neste capítulo são descritos os principais fundamentos teóricos referentes à avaliação de qualidade de vídeo e à atenção visual baseada em modelos de saliências.

2.1 AVALIAÇÃO DE QUALIDADE DE VÍDEO

Cada vez mais as mídias digitais estão inclusas no cotidiano do mundo moderno, assim, a avaliação da qualidade de vídeo é uma necessidade para que o usuário do sistema de transmissão e recepção tenha acesso a um vídeo que lhe proporcione uma boa qualidade de experiência. A AQV proporciona uma reposta ao provedor (radiodifusor, provedor de internet), podendo este tomar as medidas necessárias para garantir a qualidade ao usuário final.

A AQV pode ser dividida em duas principais categorias, as métricas subjetivas e as métricas objetivas. Como os vídeos digitais são afetados por uma grande variedade de distorções durante a aquisição, processamento, compressão, armazenamento, transmissão e reprodução, as quais induzem diversos tipos de degradações visuais, a maneira que melhor se aproxima de uma avaliação ideal é a subjetiva, porém mesmo esta contém erros. Porém, na prática a avaliação subjetiva tem um alto custo e necessitam de tempo relativamente longo para ser realizada. Assim, as métricas objetivas são métodos automáticos para realizar uma avaliação da qualidade do conteúdo. Em geral, o desempenho desta métrica é aferido comparando-se as medidas objetivas obtidas com as métricas com as avaliações subjetivas (WANG *et al.*, 2004). As métricas subjetivas e objetivas são tratadas nas seções seguintes.

2.1.1 AVALIAÇÃO SUBJETIVA

A avaliação subjetiva é feita por um conjunto de pessoas escolhidas como amostragem de um grupo normal de usuários e espectadores de vídeo digital. Deste modo, esta avaliação representa uma média da opinião deste grupo de espectadores, ou seja, é diretamente correlacionado com a percepção geral de qualidade humana. Porém, como desvantagens pode-se citar

o alto custo para realizar estas avaliações, como a disponibilidade de voluntários para fazer o teste, de um espaço apropriado, de um banco de vídeos e de tempo para realizar a avaliação. Neste caso, por exemplo, não é possível realizar uma avaliação em tempo real e nem obter uma resposta de qualidade na saída do sistema (ex.: monitor na residência do usuário).

A avaliação subjetiva segue duas linhas de pesquisa divergentes:

1. Uma delas busca realizar avaliações nos ambientes naturais dos usuários e através de seus próprios meios de reprodução, buscando assim reproduzir melhor a qualidade de experiência do usuário, com os meios que possuem suas próprias influências na avaliação, como nos métodos sugeridos por Culibrk *et al.* (2010) e Han e Lee (2014), onde ambos utilizam vídeos disponíveis na internet para criar um banco de vídeos para avaliação, e disponibilizam estes para serem avaliados pelos usuários em seus respectivos meios de reprodução (*tablet, smartphones, notebooks*, TV, entre outros).

2. A outra linha de pesquisa, que é mais difundida e praticada na comunidade acadêmica, segue as recomendações da ITU-R (2004) e do VQEG (2010). Estas recomendações buscam padronizar as medições, torná-las reprodutíveis e possíveis de serem comparadas entre si. Neste cenário também é garantida a confiabilidade da métrica.

Nesta dissertação é utilizada a segunda linha de pesquisa, onde os resultados das métricas objetivas propostas são testados e comparados com os resultados de bases de vídeos que seguem as recomendações citadas (LIVE (SESHADRINATHAN *et al.*, 2010) e IVP (LI; MA, 2012)).

As recomendações da ITU-R (2004) especificam parâmetros para diversos aspectos da avaliação subjetiva. Estes são: (i) Ambiente de teste; (ii) Resolução e contraste do monitor; (iii) Fonte do sinal; (iv) Seleção do material de teste; (v) Faixa de condições e ancoragem; (vi) Escolha dos observadores; (vii) Instruções para avaliação; (viii) Seção de teste; (ix) Apresentação dos resultados.

Por fim, os escores referentes à avaliação subjetiva de um banco de dados de vídeos são armazenados como MOS (*Mean Opinion Score*) ou DMOS (*Differential Mean Opinion Score*). Enquanto a MOS é indicada para avaliação de métricas NR já que utiliza apenas o vídeo processado na avaliação, a DMOS é indicada para validação de métricas RR e FR (VQEG, 2010), pois utiliza tanto o escore do vídeo processado quanto de referência na composição de seu valor. A definição da DMOS é dada pela Equação 1.

$$\mathrm{DMOS} = \mathrm{MOS(PVS)} - \mathrm{MOS(ref)} + 5, \tag{1}$$

em que PVS (*Processed Video Sequences*) é a sequência de vídeo processada e MOS(ref) é a MOS do vídeo de referência. Assim, quanto maior o valor da DMOS, melhor é a qualidade do vídeo avaliado.

2.1.2 AVALIAÇÃO OBJETIVA

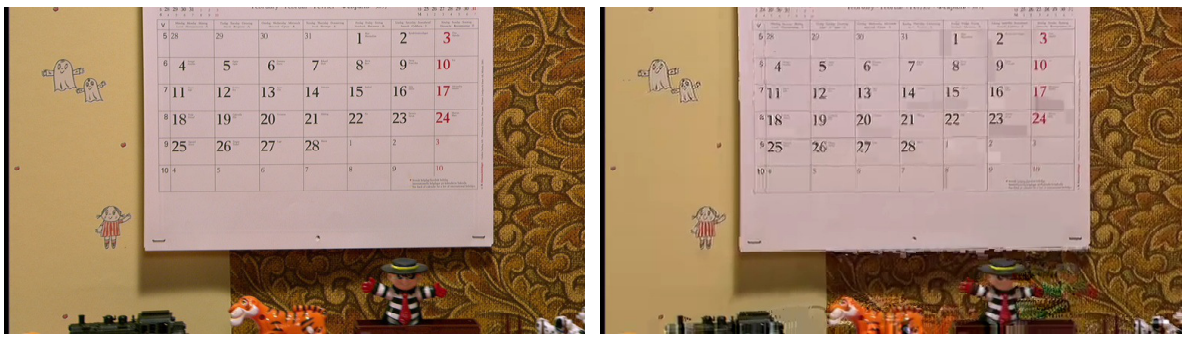
A avaliação objetiva busca realizar de forma automática a avaliação de qualidade de vídeo, retornando um escore para cada vídeo processado (WANG *et al.*, 2004). Este escore é então validado e comparado com seu respectivo valor de MOS ou DMOS. n As métricas objetivas são divididas em três principais categorias, a saber: as métricas com referência completa (FR), com referência reduzida (RR) e sem referência (NR). Estas métricas são descritas a seguir.

2.1.2.1 MÉTRICAS DE REFERÊNCIA COMPLETA (FR)

As métricas com referência completa são utilizadas em AQV em cenários onde o vídeo de referência (original) encontra-se disponível. A Figura 1 é um exemplo do mesmo quadro de um vídeo de referência e de seu vídeo processado (degradado).

A partir do vídeo de referência, as métricas FR conseguem mensurar as distorções do vídeo processado, e estimar um valor que condiz a seu nível de degradação. Algumas métricas FR são extensivamente utilizadas na literatura, tais como, MSE, PSNR, SSIM (WANG *et al.*, 2004), VQM (LI; MA, 2012) e MOVIE (SESHADRINATHAN; BOVIK, 2010).

O cálculo do erro quadrático médio (MSE) é o que chamamos de uma métrica de fidelidade ou uma métrica de dados, assim é muito utilizada na literatura para comparar a eficiência



(a) (b)

Figura 1: Exemplo de quadros original e processado da sequencia de vídeo 'mc' da base de dados LIVE. (a) Quadro 415 do vídeo de referência 'mc1_50fps.yuv'. (b) Quadro 415 do vídeo processado 'mc8_50fps.yuv'.

Fonte: Adaptado de Seshadrinathan *et al.* (2010).

de métricas. O MSE é definido pela seguinte equação.

$$\text{MSE}_f = \frac{1}{NM}\sum_{i=0}^{N}\sum_{j=0}^{M}\left[x(f,i,j) - y(f,i,j)\right]^2, \tag{2}$$

em que os termos $x(f,i,j)$ e $y(f,i,j)$ são os valores de luminância do f-ésimo quadro original e do fésimo quadro degradado, respectivamente. Os número de linhas é dado por M e o de colunas por N. MSE_f é o valor do erro médio quadrático do vídeo processado.

Baseado no MSE, a métrica PSNR (*Peak Signal-to-Noise Ratio*) é utilizada em muitos trabalhos de AQV como parâmetro para comparação de eficiência. Porém, o PSNR é uma métrica FR que não apresenta bons resultados de correlação com as medidas subjetivas, já que este não consegue mensurar adequadamente distorções estruturadas e alguns tipos de deslocamentos de imagem que afetam pouco a percepção de qualidade do SVH, porém causam grandes variações na medida do MSE (WANG; BOVIK, 2009). A Equação 3 define a métrica PSNR, que é calculada quadro a quadro e apresenta como unidade o dB.

$$\text{PSNR} = \frac{1}{F}\sum_{f=1}^{F} 20\log_{10}\left(\frac{(2^k-1)}{\sqrt{\text{MSE}_f}}\right), \tag{3}$$

em que MSE_f é dado pela Equação 2, k é o número de *bits* por *pixel* da luminância e F é o número total de quadros do vídeo.

Como um dos problemas em VQA era a avaliação de distorções estruturadas e baseadas na DCT, Wang *et al.* (2004) propuseram a métrica SSIM baseada na similaridade estrutural. A Figura 2 é um esquemático da SSIM, que utiliza a componente de luminância do vídeo de referência (Sinal x) e do processado (Sinal y) para fazer comparações de luminância, contraste e estrutura, estas então são combinadas resultando na medida de similaridade.

As comparações de luminância, contraste e estrutura são dadas pelas Equações 4, 5 e 6, respectivamente.

$$l(f,x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{4}$$

$$c(f,x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{5}$$

$$s(f,x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{6}$$

em que μ_x, μ_y são a média das entradas, a variância é dada por σ_x^2, σ_y^2 e σ_{xy} é a covariância cruzada de x e y, respectivamente. Já C_1, C_2 e C_3 são constantes de baixa magnitude.

A equação geral de SSIM para a avaliação de um vídeo com F quadros é dada pela

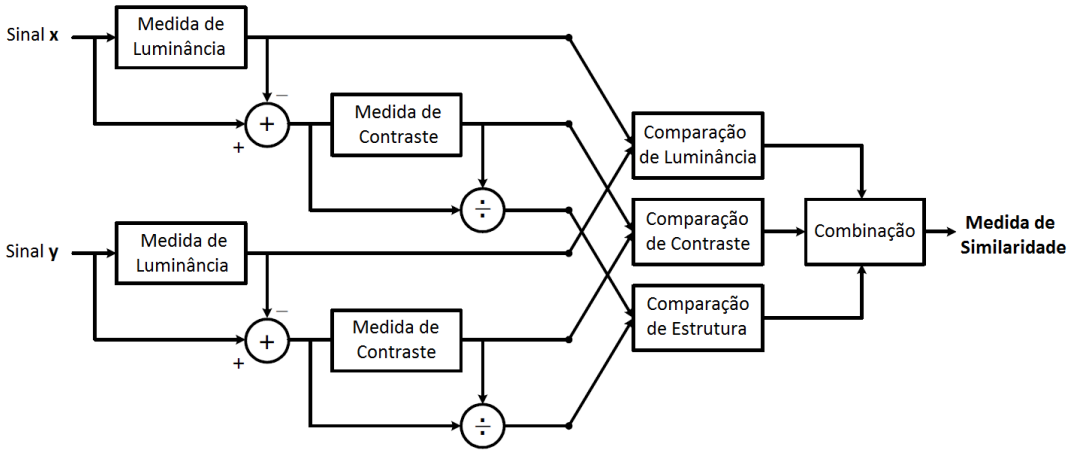


Figura 2: Diagrama esquemático da métrica SSIM.

Fonte: Adaptado de Wang *et al.* (2004).

Equação 7, onde l, c e f são as componentes de comparação de luminância, contraste e estrutura, respectivamente. O vídeo de referência é representado por x e o processado por y. Já α, β e λ são os parâmetros de ajuste de SSIM.

$$\text{SSIM} = \frac{1}{F}\sum_{f=1}^{F}[l(f,x,y)]^{\alpha}[c(f,x,y)]^{\beta}[s(f,x,y)]^{\gamma}, \tag{7}$$

em que α, β e λ são fixados com o valor de 1 para simplificar a Equação 7.

Além de ser utilizada como referência para comparação entre métricas, SSIM também é utilizada como base da metodologia de várias outras métricas que aproveitam de sua característica de avaliação estrutural, como por exemplo, nas métricas: (1) MS-SSIM - *Multi-escale SSIM* (WANG *et al.*, 2003b) baseada em multiescalas de estrutura de similaridade; (2) SW-SSIM - *Speed-Weighted SSIM* (WANG; LI, 2007) baseada em diferenças de pesos; (3) MC-SSIM - *Motion Compensated SSIM* (MOORTHY; BOVIK, 2010) baseada em compensação de movimento; e (4) LST-SSIM - *Local Spatial-Temporal SSIM* (WANG *et al.*, 2012) baseada em localização de características de bordas e movimentos.

A métrica VQM (*Video Quality Model*) foi incluída em duas recomendações da ITU como método normativo de avaliação de qualidade de vídeo. VQM foi proposta por Pinson e Wolf (2004) e é baseada nas características espaciais dos componentesY (luminância), C_b (sinal de crominância azul), C_r (sinal de crominância vermelho) e de características de movimento da

componente Y. A expressão geral de VQM é dada pela Equação 8.

$$\begin{aligned} \text{VQM} = &-0,2097 * si\_loss + 0,5969 * hv\_loss + 0,2483 * hv\_gaint + 0,0192 * chroma\_spread \\ &-2,3416 * si\_gain + 0,0431 * ct\_ati\_gain + 0,0076 * chroma\_extreme, \end{aligned} \quad (8)$$

que é composta de 7 parâmetros. Cada parâmetro tem um objetivo de detecção específico, sendo estes:

si_loss	- perda espacial por borramento,
hv_loss	- deslocamento de bordas da vertical ou horizontal para a diagonal,
hv_gaint	- deslocamento de bordas da diagonal para a vertical ou horizontal,
chroma_spread	- alteração na propagação das distribuições de cores,
si_gain	- nitidez,
ct_ati_gain	- contraste e movimentos,
chroma_extreme	- defeitos localizados de cores.

Outra métrica FR considerada eficiente por conseguir boa correlação com o SVH é a MOVIE (*MOtion-based Video Integrity Evaluation index*) baseada em componentes de movimento para verificação de qualidade. Li e Bovik (2010) desenvolveram a métrica MOVIE separando componentes para definição de qualidade espacial e temporal. A Figura 3 traz um diagrama esquemático da metodologia utilizada pela métrica MOVIE.

Primeiramente, os vídeos de referência e processados são decompostos em um canal de banda passante espaço-temporal usando filtro de Gabor. A avaliação de qualidade espacial é feita com um método inspirado na SSIM. Já a avaliação de temporal é feita usando informação de movimento da sequência de vídeo de referência. Finalmente, ambas são combinadas para obter um resultado de avaliação final, a MOVIE.

2.1.2.2 MÉTRICAS DE REFERÊNCIA REDUZIDA (RR)

As métricas de referência reduzidas (RR) utilizam apenas parte da informação contida nos vídeos de referência, não necessitando que o vídeo original esteja totalmente disponível no procedimento de avaliação. Isto traz algumas vantagens para as métricas RR em comparação com as métricas FR, como na velocidade de avaliação e disponibilidade de avaliação sem a reprodução do vídeo original, já que as características extraídas do vídeo original são obtidas

antes da transmissão e então codificadas e transmitidas ao receptor. Porém, como desvantagem, esta metodologia tem um aumento na complexidade envolvida na avaliação.

Métrica proposta por Callet *et al.* (2006) utiliza características espaciais de blocagem e borramento e características temporais de diferenças de intensidade extraídas do vídeo original, que são transmitidas codificadas para então serem verificadas pela métrica RR.

Outros autores também desenvolveram métricas RR utilizando características espaço-temporais, como, Silva *et al.* (2013) que propuseram uma métrica RR baseada na diferença de atividade dos coeficientes da DCT, e Soundararajan e Bovik (2013) que desenvolveu a métrica STRRED (*Spatio-Temporal-Reduced Reference Entropic Differences*) baseada nas diferenças de entropias espaço-temporais.

2.1.2.3 MÉTRICAS SEM REFERÊNCIA (NR)

Os métodos NR realizam a avaliação de vídeo utilizando apenas o vídeo processado, ou seja, não possuem nenhuma informação do vídeo original como no caso das métricas FR e RR. Isto faz com que a métrica tenha que ser capaz de presumir as distorções e degradações presentes no vídeo, o que aumenta muito a complexidade no desenvolvimento do método.

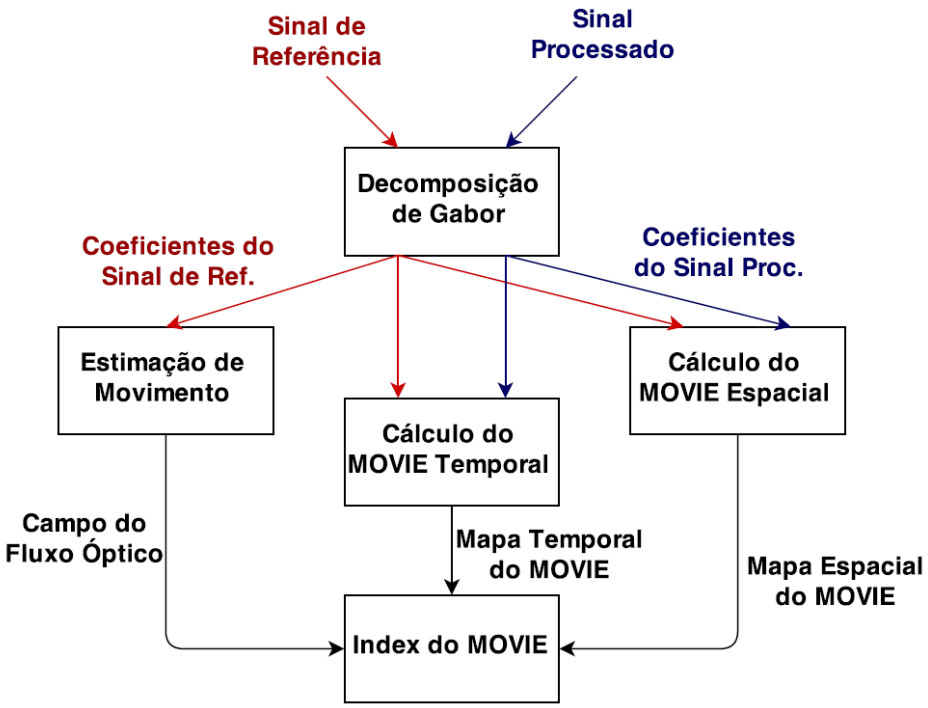


Figura 3: Diagrama do sistema de medida da métrica MOVIE.

Fonte: Adaptado de Li e Bovik (2010).

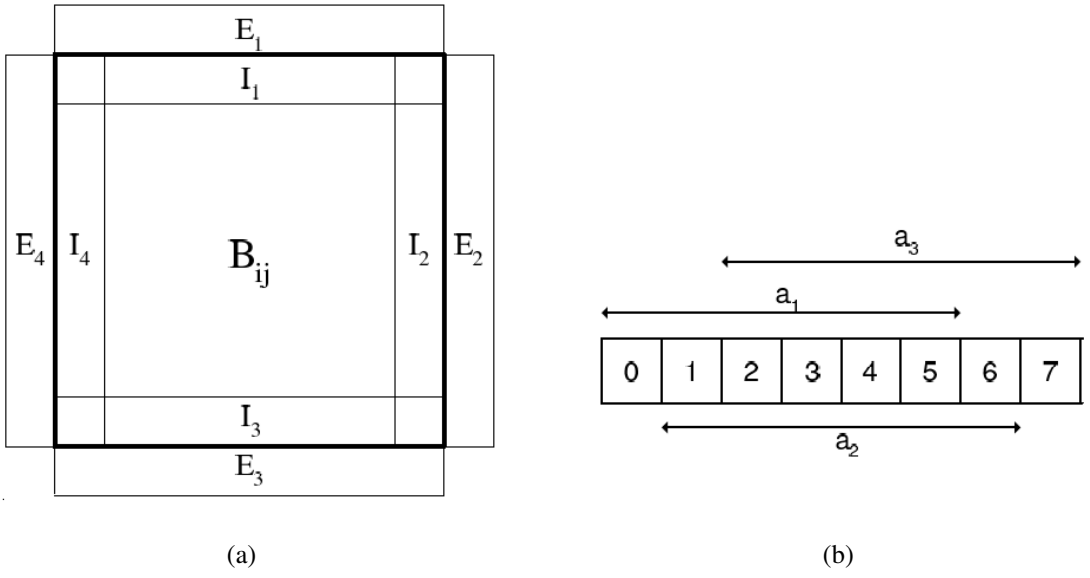
A seguir é apresentada a métrica NR Blockiness proposta por Babu *et al.* (2008) baseada em características espaciais de blocagem. Esta métrica assume que a percepção de blocagem pelo SVH é causada pela presença de bordas com baixa atividade. Assim, sendo que B_{ij} representa um bloco 8 *x* 8 com localização inicial de (i, j) em um dado quadro, e I_k representa as bordas do bloco para k = 1,...,4 como mostrado na Figura 4(a), a percepção da característica blockiness é observada quando o comprimento da borda do bloco (8*x*8) é abaixo de 5 *pixels*, então para o cálculo de blocagem é utilizado o valor de 6 *pixels*.

Para medir a atividade em uma determinada borda I_k, esta é primeiramente dividida em 3 segmentos de comprimento 6, como mostra a Figura 4(b). Estes segmentos são descritos na Equação 10.

$$\begin{aligned} a_{k1} &= I_k(n) : n = 0...,5 \\ a_{k2} &= I_k(n) : n = 1...,6 \\ a_{k3} &= I_k(n) : n = 2...,7 \end{aligned} \tag{10}$$

Então, a atividade é quantificada pelo desvio padrão, σ_{kl} para cada a_{kl} sendo l = 1,...,3. Para cada borda I_k a atividade é definida como baixa se ao menos um dos σ_{kl} estiver abaixo de um limiar ε.

A métrica é então calculada pelos seguintes passos:



(a) (b)

Figura 4: Metodologia para extração da métrica NR Blockiness. (a) Bloco 8 *x* 8 e suas bordas. (b) Uma borda do bloco 8 *x* 8.

Fonte: Adaptado de Babu *et al.* (2008).

1. Contador de blocagem inicializado com 0. (C_B=0).

2. Cada uma das 4 bordas I_k tem 3 σ_{kl}, com um total de 12 medidas de atividade.

3. A seguir é calculado um gradiente para cada a_{kl}, dados pela Equação 11, onde E_k são as bordas adjacentes às bordas do bloco como mostrado na Figura 4.

$$\begin{aligned} \Delta_{k1} &= mean(I_k(n) - E_k(n)) : n = 0..., 5 \\ \Delta_{k2} &= mean(I_k(n) - E_k(n)) : n = 1..., 6 \\ \Delta_{k3} &= mean(I_k(n) - E_k(n)) : n = 2..., 7 \end{aligned} \tag{11}$$

4. O contador C_b é incrementado em 1 se ao menos um dos segmentos satisfizer as condições de limiares conforme a Equação 12.

$$\begin{aligned} \sigma_{kl} &<= \varepsilon \\ \Delta_{kl} &<= \tau, \end{aligned} \tag{12}$$

   em que ε e τ são os limiares definidos experimentalmente por Babu *et al.* (2008) com valores de 0,1 e 2, respectivamente.

5. Por fim, a equação geral da métrica Blockiness é dada pela Equação 13.

$$B_F = \frac{C_B}{\text{número total de blocos no quadro}}, \tag{13}$$

   onde os resultados estarão no intervalo entre 0 e 1, sendo que o valor 0 corresponde a um quadro sem blocagem e 1 o cenário com todos os blocos do quadro visível.

Outra métrica NR é proposta por Silva e Pohl (2012) baseada no algoritmo de otimização de Levenberg-Marquardt (NRVQA-LM - *No-Reference Video Quality Assessment based on Levenberg-Marquardt*). NRVQA-LM é calculada por um modelo sigmoidal que utiliza a extração de características espaço-temporais como parâmetros.

A extração de características espaciais é baseada na métrica de avaliação de imagem JPEG-NR proposta por Wang *et al.* (2002) que emprega a detecção de artefatos de blocagem e borramento. O componente de luminância é dado por $y(f, j, i)$ com $i \in [1, M]$ e $j \in [1, N]$, sendo M e N o número de linhas e colunas de um quadro f, respectivamente. Em seguida são determinadas as diferenças ao longo das linhas e colunas, dadas pelas Equações 14 e 15.

$$d_h(f, i, j) = y(f, i, j+1) - y(f, i, j), \quad j \in [1, N-1], \tag{14}$$

$$d_v(f, i, j) = y(f, i+1, j) - y(f, i, j), \quad i \in [1, M-1]. \tag{15}$$

Como o formato JPEG utiliza a DCT em seu formato de compressão, os efeitos de blocagem e borramento geralmente sempre surgem devido à esta compressão. A métrica JPEG-NR é feita para avaliações de imagens. As bordas entre os blocos da DCT criam efeitos de blocagem que podem ser estimados pela Equação 18, que utiliza as componentes horizontal e vertical calculadas nas Equações 16 e 17 respectivamente.

$$B_h = \frac{1}{M\left(\lfloor \frac{N}{\tau} \rfloor - 1\right)} \sum_{i=1}^{M} \sum_{j=1}^{\lfloor \frac{N}{\tau} \rfloor - 1} |d_h(f,i,\tau j)|, \tag{16}$$

$$B_v = \frac{1}{N\left(\lfloor \frac{M}{\tau} \rfloor - 1\right)} \sum_{i=1}^{\lfloor \frac{M}{\tau} \rfloor - 1} \sum_{j=1}^{N} |d_v(f,\tau i,j)|. \tag{17}$$

$$B_f = \frac{B_h + B_v}{2}. \tag{18}$$

onde para um bloco 8 *x* 8 *pixels* o valor de τ é 8. Com o valor de B_f para cada quadro, a Equação 19 faz uma média dos valores para a sequência de vídeo, onde Q é o número total de quadros do vídeo.

$$B = \frac{1}{Q} \sum_{f=1}^{Q} B_f, \tag{19}$$

Outro artefato criado pela DCT é o de borramento na direções horizontal e vertical representados pelas Equações 20 e 21. Este efeito ocorre devido à redução da atividade espacial no processo de quantização.

$$A_h = \frac{1}{\tau - 1} \left[\frac{\tau}{M(N-1)} \sum_{i=1}^{M} \sum_{j=1}^{N-1} |d_h(f,i,j)| - B_h \right], \tag{20}$$

$$A_v = \frac{1}{\tau - 1} \left[\frac{\tau}{N(M-1)} \sum_{i=1}^{M-1} \sum_{j=1}^{N} |d_v(f,i,j)| - B_v \right]. \tag{21}$$

Combinando as componentes A_h e A_v é obtida a característica A_f para detecção de artefatos de borramento, como mostra a Equação 22

$$A_f = \frac{A_h + A_v}{2}. \tag{22}$$

Por fim, a média do valor A_f para todos os quadros do vídeo resulta na característica temporal A (Equação 23)

$$A = \frac{1}{Q} \sum_{f=1}^{Q} A_f, \tag{23}$$

Outra característica extraída para detecção de borramento, é a taxa de cruzamento por

zero (Z). A Equação 29 define Z. Para isto, são calculadas separadamente as componentes horizontais e verticais como demonstrado a seguir, onde ZC é o cruzamento por zero, ou seja, o ponto onde o sinal da função matemática muda (e.g. de negativo para positivo).

$$z_h(f,i,j) = \begin{cases} 1, & \text{se existe } ZC \text{ na direção horizontal} \\ 0, & \text{caso contrário} \end{cases}, \tag{24}$$

$$z_v(f,i,j) = \begin{cases} 1, & \text{se existe } ZC \text{ na direção vertical} \\ 0, & \text{caso contrário} \end{cases}. \tag{25}$$

A partir desta definição são equacionadas as componentes Z_h e Z_v.

$$Z_h = \frac{1}{M(N-2)} \sum_{i=1}^{M} \sum_{j=1}^{N-2} z_h(f,i,j), \tag{26}$$

$$Z_v = \frac{1}{N(M-2)} \sum_{i=1}^{M-2} \sum_{j=1}^{N} z_v(f,i,j), \tag{27}$$

A combinação entre Z_h e Z_v gera o descritor de borramento Z_f (Equação 28), e um valor médio para o vídeo é estabelecido pelo valor final de Z (Equação 29)

$$Z_f = \frac{Z_h + Z_v}{2}. \tag{28}$$

$$Z = \frac{1}{Q} \sum_{f=1}^{Q} Z_f. \tag{29}$$

Além das características espacias, a métrica NRVQA-LM também utiliza informações temporais. O descritor TI (*Temporal perceptual Information*) é uma versão da característica temporal TI_F da recomendação ITU-T P.910 (ITU-T P.910, 1999), descrita a seguir.

$$\text{TI} = \frac{1}{Q-1} \sum_{f=2}^{Q} \sigma\left[m(f,i,j)\right], \tag{30}$$

em que $m(f,i,j) = y(f,i,j) - y(f-1,i,j)$ é a diferença de movimento (luminância) entre um quadro e seu antecessor na sequência de vídeo e Q é o número total de quadros do vídeo.

Outras características temporais adotadas são o MAD e o MADw que são adaptados de Ding *et al.* (2008). Ambas as métricas utilizam análise temporal através da abordagem de quadros sucessivos $m(f,j,i)$, sendo que a MAD é definida como

$$\text{MAD}_k = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left|m(f,i,j)\right|, \tag{31}$$

em que f>1. Como cada quadro apresenta um valor, e a métrica trabalha com valores para uma sequência de vídeo, então é feita a média do valor de MAD_k para todo o vídeo, como mostrado na Equação 32.

$$\overline{\mathrm{MAD}} = \frac{1}{Q-1}\sum_{k=1}^{Q-1}\mathrm{MAD}_k. \tag{32}$$

Já o parâmetro MADw considera a razão entre o valor MAD_k de um quadro e de seu antecessor MAD_{k-1}.

$$\mathrm{MADw} = \frac{1}{Q-1}\sum_{k=1}^{Q-1}\frac{\mathrm{MAD}_k}{\mathrm{MAD}_{k-1}}. \tag{33}$$

Após a extração de características, a métrica NRVQA-LM utiliza um modelo matemático sigmoidal para realizar o cálculo dos escores objetivos. Este modelo é descrito na Equação 71.

$$NRVQA\text{-}LM = \frac{1}{1+e^{\left(\beta_1 B+\beta_2 Z+\beta_3 A+\beta_4 TI+\beta_5\overline{MAD}+\beta_6 MADw+\beta_7\right)}}, \tag{34}$$

em que A, B, Z, TI, $\overline{MAD}$ e MADw são as características citadas acima, e β_1 a β_7 são coeficientes que dão pesos diferentes a cada característica da métrica. Estes coeficientes são por fim otimizados através do algoritmo de Levenberg-Marquardt (LEVENBERG, 1944; MARQUARDT, 1963; MORÉ, 1977) que é empregado na solução de problemas de mínimos quadrados.

2.2 MODELO DE SALIÊNCIA

Recentemente, vários trabalhos em AQV vêm sendo desenvolvidos baseados na atenção visual, utilizando região de interesse e modelos de saliência para determinar regiões dos vídeos com maior influencia no SVH (ĆULIBRK *et al.*, 2011; WINKLER, 2012; LIN *et al.*, 2012; FU *et al.*, 2013; LUO *et al.*, 2014). Nesta Seção são tratados alguns fundamentos teóricos referentes aos modelos de saliência.

2.2.1 MOVIMENTOS DOS OLHOS

Os métodos de atenção visual buscam prever os movimentos do olho humano, determinando as regiões mais salientes.

Para estudar o comportamento do olho humano, utilizando rastreadores de olhar que permitem traçar gráficos dos movimentos oculares e dos pontos de fixação (MAHAPATRA *et al.*, 2008). Os resultados são utilizados no desenvolvimento de métodos que simulam o mecanismo de atenção humana.

O foco da visão humana depende apenas do espectador, porém, mesmo sendo uma

função cognitiva, os pontos de fixação em um vídeo ou imagem são semelhantes entre várias pessoas, isto ocorre porque o SVH é afetado por características similares em todas as pessoas (WINKLER, 2012). Pode-se citar, por exemplo, que o movimento de mãos, ou de um pássaro cruzando um cenário são pontos que atraem instantaneamente a atenção humana. Também observa-se que mesmo características como movimentos ondulatórios da água atraem a atenção da maioria dos espectadores.

2.2.2 ABORDAGEM BOTTOM-UP E TOP-DOWN

Pesquisas consideram o desenvolvimento da atenção visual humana como um mecanismo de duas componentes (CONNOR *et al.*, 2004), denominados como *bottom-up* e *top-down*.

A abordagem *bottom-up* é dada pela estímulo presente, que é de alguma maneira discrepante do resto do vídeo (STYLES, 2005). Como, por exemplo, um jogador com uniforme vermelho em um campo de grama verde, ou um carro em movimento em uma autoestrada. Tanto o jogador como o carro se destacam no vídeo devido a, respectivamente, diferença de cor e de movimento. Um estímulo *bottom-up* atua com um tempo de 25 a 50 ms na atenção visual (ITTI; KOCH, 2001).

O outro mecanismo de atenção visual é o *top-down*, que depende de tarefas, memórias, eventos passados e reconhecimento de padrões para formar a atenção (STYLES, 2005). Como, por exemplo, um objeto ou cenário que lembre a infância, ou que envolva algum trauma psicológico. Outro exemplo, é a atenção voltada a padrões assimilados com mais facilidade pelo cérebro, como faces e alguns objetos. Este tipo de atenção visual leva acima de 200 ms para afetar o movimento dos olhos, assim sendo um estímulo mais lento.

Ambos os mecanismos atuam de maneira conjunta no SVH, porém, como constatado por Ćulibrk *et al.* (2011), a atenção visual baseada em *bottom-up* avalia melhor os vídeos em virtude da dinâmica do processo de visualização, onde atua com mais frequência e devido ao tempo mais rápido de resposta tem um efeito mais impactante na visão humana.

2.2.3 CARACTERÍTICAS ESPACIAIS SALIENTES

As características espaciais salientes são possíveis aspectos discrepantes dentro da imagem. Como as discrepâncias são os pontos que afetam o SVH, estes devem ser detectados e mensurados. Assim, os principais aspectos espaciais percebidos pelo SVH são:

1. Contraste - a visão humana é sensível a regiões com alto contraste (WINKLER, 2012). Assim, estas regiões tendem a ser regiões de atenção.

2. Orientação - regiões com diferenças de orientação afetam o SVH (L. *et al.*, 1998).

3. Intensidade - a intensidade é um dos aspectos com grande influência no SVH (L. *et al.*, 1998).

4. Cores - as discrepâncias de cores em uma imagem podem delimitar regiões de interesse (FU *et al.*, 2013). Elas respondem pelas diferenças de tom, saturação e valor.

Fu *et al.* (2013) propôs um modelo de saliência baseado em características espaço-temporais, onde a Equação 35 caracteriza o mapa de saliências das características espaciais.

$$S_m = w_C * C + w_I * I + w_O * O + w_R * R, \tag{35}$$

em que C, I, O e R são os mapas de, cores, intensidade, orientação e contraste, respectivamente. Os parâmetros w_c, w_i, w_o e w_r são pesos para cada característica correspondente.

2.2.4 CARACTERÍSTICAS TEMPORAIS SALIENTES

Mahapatra *et al.* (2008) demonstraram que as características salientes de movimentos têm mais influência no SVH do que as espaciais, ou seja, movimentos de objetos ou de cenários afetam mais a atenção visual do que mudanças de cor, intensidade, entre outros.

Como exemplo de algumas características temporais desenvolvidas temos a intensidade de vetores de movimento (MAHAPATRA *et al.*, 2008; ROMANI *et al.*, 2014), a coerência temporal (MAHAPATRA *et al.*, 2008) e o modelo de subtração de *background* (CULIBRK *et al.*, 2009; FU *et al.*, 2013).

Os vetores de movimentos são extraídos através da informação no fluxo de dados. Cada macrobloco criado na compressão baseada na DCT possui um vetor de movimento, quando este macrobloco altera sua posição entre um quadro e outro, uma intensidade é dada ao vetor. Mahapatra *et al.* (2008) define a medida de vetor de movimento de um quadro na Equação 36.

$$I_t(x,y) = \frac{\sqrt{dx_{x,y}^2 + dy_{x,y}^2}}{MaxMAG}, \tag{36}$$

em que $dx_{x,y}$ e $dy_{x,y}$ são os componentes de vetor de movimento em x, y e $MaxMag$ é a magnitude máxima do campo do vetor de movimento.

A característica de coerência temporal é dada pela Equação 37. Esta calcula a diferença de valor entre um grupo de *pixels* entre quadros sucessivos, onde, quanto maior a entropia maior será o movimento e também a saliência apresentada.

$$C_t(x,y) = -\sum_{i=1}^{M} p_t(i).log(p_t(i)), \tag{37}$$

em que $p_t(i)$ é a probabilidade de ocorrência de um *pixel* em um local correspondente em diferentes quadros. M é o número de quadros avaliados.

2.2.5 EXTRAÇÃO DE SALIÊNCIAS

A modelo de saliências utiliza uma abordagem *bottom-up* como descrita na Seção 2.2.2, que se baseia nas características de estímulos presentes no vídeo. Nesta abordagem, as regiões salientes são caracterizadas pelas informações espaciais e temporais discrepantes, não levando em consideração características afetadas por tarefas, memórias ou outros eventos.

A extração de saliência neste trabalho é feita tanto pelas características espaciais como temporais. As informações espaciais são os parâmetros de cor, intensidade e orientação. O algoritmo de extração destas características se baseia em uma adaptação do método proposto por Itti e Koch (2001), mostrado na Figura 5. Este método passa por algumas etapas para extração das saliências propostas por Itti e Koch (2000), estas são:

1. **Extração de Características Visuais** - Nesta etapa é feita a extração das informações de cor (vermelho-verde e azual-amarelo), orientação (ângulos 0^o, 45^o, 90^o e 135^o) e intensidade, somando 7 modalidades. Nove escalas (nível 0 à 8) espaciais são criadas usando pirâmides Gaussianas (BURT; ADELSON, 1983), que consiste na filtragem progressiva e subamostragem do quadro de entrada. Em seguida, são calculadas as diferenças entre o centro (níveis centrais 2, 3 e 4) e periferia (distância de 3 e 4 níveis de profundidade dos níveis centrais) gerando seis diferenças entre níveis (2-5, 2-6, 3-6, 3-7, 4-7, 4-8). Estas diferenças geram um total de 42 mapas de características (6 para cada modalidade).

2. **Combinação de informação entre múltiplos mapas de características** - Utilizando um filtro 2D de diferenças Gaussianas é feita uma combinação para cada modalidade. Na sequência, estas são normalizadas e por fim combinadas em três mapas de conspicuidade, de cor, intensidade e orientação, levando em conta que a combinação entre caracaterísticas competem fortemente para criação do mapa de saliência, enquanto a combinação entre modalidades contribuem independemente para o mesmo.

3. **O Mapa de Saliência** - Os mapas de conspicuidade são linearmente somados em um único mapa de saliência, gerando a informação espacial de saliência baseada na aboradagem *bottom-up*

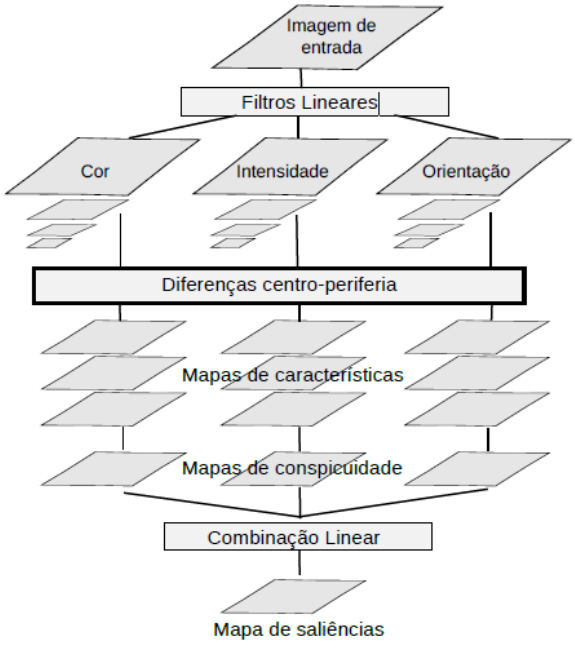


Figura 5: Modelos de extração de mapas de conspicuidade e mapas de saliência de Itti e Koch. Fonte: Adaptado de Itti e Koch (2000)

Em seguida, são exploradas as características de movimentos salientes. Conforme Mahapatra *et al.* (2008), as informações de movimento salientes têm maior influencia no SVH, o modelo utilizado neste trabalho aplica um método multiescala do modelo *background* no quadro que forma uma pirâmide Gaussiana na detecção destes movimentos. Isto permite verificar as diferenças na coerência espacial e na consistência entre escalas tanto para movimentos do objeto como da câmera. O método de *background*, mostrado na Figura 7, se baseia no uso de dois planos de fundos para cada quadro. Estes planos são extraídos através de dois filtros do tipo IIR (Resposta Infinita ao Impulso), em que inicialmente os planos de fundo apresentam o mesmo valor do quadro atual, porém com o decorrer da sequência do vídeo a consistência de movimento é aos poucos agregada aos planos de fundo, corforme mostra a Equação 38.

$$b_l(i) = (1-\alpha_l)b_l(i) + \alpha_l p(i), l \in \{1,2\}, \tag{38}$$

em que α_l é a taxa de aprendizado usada para filtrar o quadro de plano de fundo l, p(i) é o valor do *pixel* na localização i no quadro, b$_l$(i) é o valor do *pixel* na localização i no quadro de plano

de fundo l. Os dois planos de fundo são obtidos utilizando valores de α diferentes, sendo $\alpha_2 = \alpha_1/2$ conforme Culibrk *et al.* (2010).

Em seguida, um filtro temporal é aplicado para obter uma única imagem através dos 2 planos de fundo e do quadro. O filtro temporal aplicado é dado pela Equação 39 e sua função tem o formato de um chapéu mexicano conforme mostra a Figura 6.

$$f(x) = -\frac{2}{\sqrt{3}}\pi^{\frac{-1}{4}}.(1-x^2).exp^{\frac{x^2}{2}}, \tag{39}$$

em que x representa a distância Euclidiana do ponto para o centro do filtro.

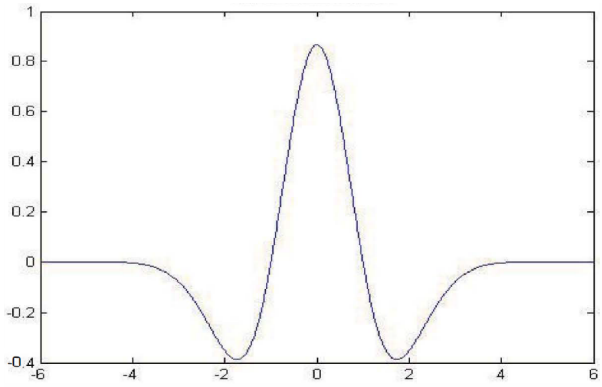


Figura 6: Filtro IIR com forma de onda denominado chapéu mexicano.
Fonte: Adaptado de Fu *et al.* (2013).

A resultante filtrada de cada derivada da pirâmide Gaussiana é somada e passada por um detector de discrepâncias, resultando assim no mapa de saliência baseada em movimentos, como observado na figura abaixo.

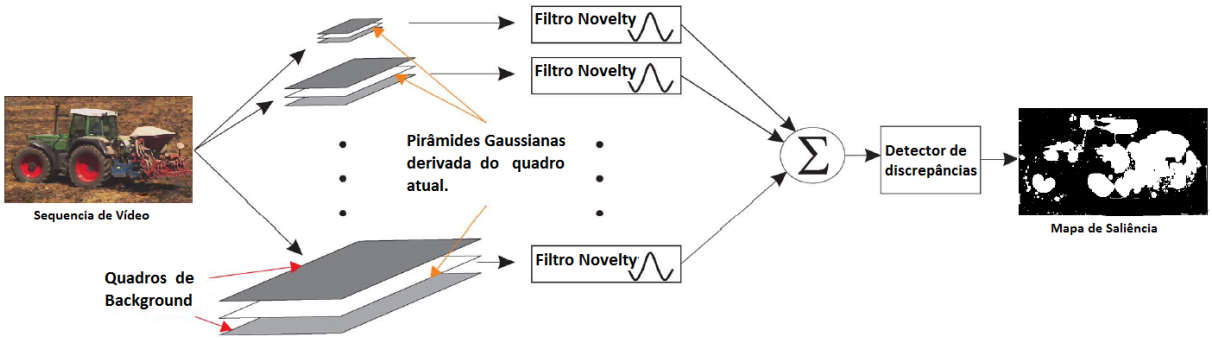


Figura 7: Diagrama esquemático da segmentação de regiões de saliência por movimentos.
Fonte: Adaptado de Ćulibrk *et al.* (2011).

2.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capítulo descreveu os principais aspectos referentes a AQV e a modelos de saliência necessário para o desenvolvimento desta dissertação. No próximo capítulo será exposta a metodologia para o desenvolvimento das métricas de avaliação de qualidade baseadas em modelos de saliência.

3 METODOLOGIA

Neste capítulo são descritos os materiais e métodos utilizados e criados no desenvolvimento da dissertação. Também são apresentadas as métricas SSIM-SM (*Structural SIMilarity based in Salient Model*), NRLM-SM (*No-Reference Metric using Levenberg-Marquardt based in Salient Model*) e NRLM-SMb (*No-Reference Metric using Levenberg-Marquardt based in Salient Model and Blockiness features*) propostas.

As seções a seguir tratam das bases de vídeo, do processamento de escores subjetivos e objetivos, do método de extração de saliências, da extração de características baseadas em saliência, e por fim, dos métodos propostos.

3.1 BASES DE VÍDEOS

As bases de vídeos utilizadas para o desenvolvimento e validação das métricas propostas foram a base LIVE (SESHADRINATHAN *et al.*, 2010) e IVP (LI; MA, 2012). Ambas as bases de vídeos são amplamente empregadas em ensaios e citadas na literatura em AQV e seguem as recomendações do VQEG (VQEG, 2010) e da ITU (ITU-T P.910, 1999).

Nas subseções seguintes são abordadas as principais características destas bases de vídeos, sendo exploradas principalmente as características de atividade espacial e temporal, calculados utilizando as seguintes equações, conforme a recomendação ITU-R (2004).

$$\mathrm{TI}_F = \max_F\{\sigma_s[m(f,i,j)]\}, \tag{40}$$

$$\mathrm{SI} = \max_F\{\sigma_s[\mathrm{Sobel}(y(f,i,j))]\}, \tag{41}$$

na qual $m(f,i,j)$ é a diferença de movimento (*i.e.*, diferença de luminância) dos *pixels* de quadros sucessivos (f e $f-1$), e $\sigma_s[m(f,i,j)]$ é o desvio-padrão de $m(f,i,j)$. Já $\sigma_s[\mathrm{Sobel}(y(f,i,j))]$ é o desvio-padrão do quadro f processado pelo filtro de Sobel.

3.1.1 BASE DE DADOS LIVE

A base LIVE (SESHADRINATHAN *et al.*, 2010) é composta por 10 conteúdos de vídeos, cada conteúdo possui um vídeo de referência sem degradações e 15 vídeos processados. Os vídeos processados são divididos em diferentes categorias, sendo elas: (1) Wireless, gerados por transmissões em rede sem fio; (2) IP, com perdas de pacotes; (3) H.264 e (4) MPEG, os últimos dois degradados pelas perdas na compressão.

Os vídeos estão no formato '.yuv' e possuem resolução de 768 *x* 432 *pixels* e duração de 10 segundos, com sete sequências de vídeos com 25 fps e oito com 50 fps, e com subamostragem no formato 4:2:0. Os escores subjetivos da base estão na escala da DMOS.

Silva (2013) realizou um estudo da distribuição espaço-temporal dada pelas características TI e SI. A Figura 8 mostra esta distribuição, onde o eixo vertical é dado pelo valor de TI e o horizontal pelo valor de SI na qual os triângulos representam cada um dos 10 vídeos de referência.

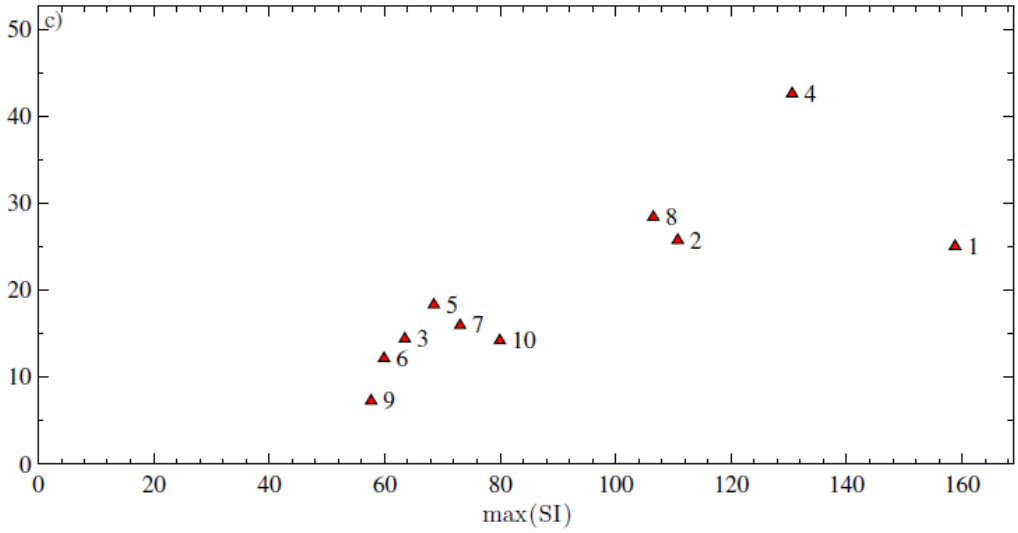


Figura 8: Diagrama TI *vs.* SI da base de dados LIVE.
Fonte: Adaptado de Silva (2013)

Através da figura acima é possível observar a distribuição espacial e temporal da base LIVE. Nota-se que os vídeos não apresentam o mesmo nível de atividades, o que é um fator importante para uma base de vídeos.

3.1.2 BASE DE DADOS IVP

A base de dados IVP (LI; MA, 2012) possui 10 sequências de vídeos com resoluções de 1920*x*1088 *pixels* e duração de 10 segundos. A partir dos vídeos de referências foram processados 128 vídeos na base, com 25 fps e formato 4:2:0. São 4 categorias de vídeos, Dirac, Ip, H.264 e MPEG-2, sendo as medidas subjetivas dadas pela DMOS.

A Figura 9 traz a distribuição das características TI e SI dos vídeos IVP da mesma forma como mostrados na Figura 8 para os vídeos da base LIVE.

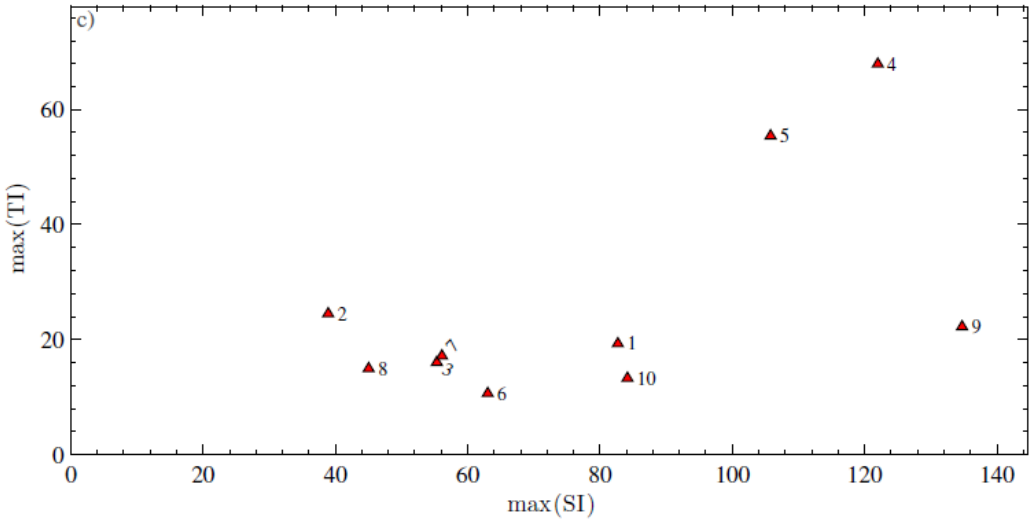


Figura 9: Diagrama TI *vs.* SI da base de dados IVP.

Fonte: Adaptado de Silva (2013)

3.2 PROCESSAMENTO DOS ESCORES

Uma métrica objetiva de AQV retorna um escore objetivo para uma sequência de vídeo avaliada. Este escore está distribuído em uma determinada faixa de valores, dependendo da métrica utilizada. Para verificar se este escore objetivo possui uma boa correlação com a MOS ou a DMOS é necessário utilizar um método de processamento para que a medida seja corretamente correlacionada. O método utilizado nesta dissertação é mostrado no diagrama esquemático da Figura 10 e segue as recomendações para processamento de escores do VQEG (VQEG, 2010).

Uma das etapas recomendadas é o mapeamento entre os escores objetivos (VQAS - *Video Quality Assessment Scores*) e os escores subjetivos (MOS e DMOS). Este mapeamento é feito para que os escores objetivos e subjetivos sejam dispostos na mesma escala e, assim, possam ser correlacionados.

Outro método recomendado é o processo de validação. Como exemplo, pode-se citar a métrica NRLM proposta nesta dissertação, onde o processo de validação é feito através do uso de dois grupos disjuntos de vídeos para cada base de vídeo. Um dos grupos é utilizado no treinamento dos coeficientes da Equação 71, enquanto o outro é o grupo de teste. Este procedimento de ajuste de características é realizado 100 vezes com diferentes grupos de treinamento e teste, e o valor da média de cada coeficiente é utilizado para o cálculo dos escores objetivos. Este procedimento é chamado de validação cruzada.

Por fim, é feita uma análise estatística para estabelecer a correlação entre as métricas objetivas e subjetivas, esta será abordada na seção seguinte.

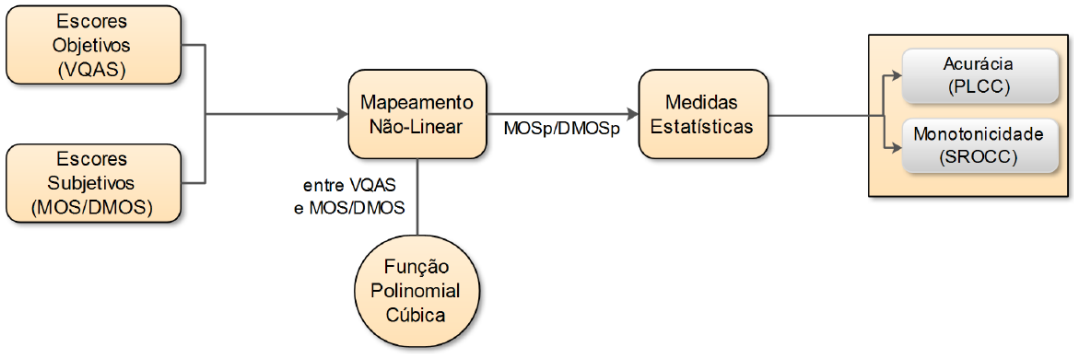


Figura 10: Diagrama da metodologia utilizada no processamento de escores.
Fonte: Adaptado de Silva (2013)

Nas métricas propostas nesta dissertação foi empregado o modelo não-linear de mapeamento baseado em uma função polinomial cúbica. A função utilizada é definida pela seguinte equação:

$$\text{MOSp} = ax^3 + bx^2 + cx + d, \tag{42}$$

na qual MOSp é o valor predito para o MOS e x é o escore subjetivo. Os coeficientes a, b, c e d são retornados pelo procedimento de ajuste de curvas. Um processo análogo pode ser feito para a DMOS.

As recomendações do VQEG também indicam os modelos estatísticos utilizados para validação. Estes são abordados na subseção seguinte.

3.2.1 MEDIDAS ESTATÍSTICAS

Em AQV as medidas estatísticas têm como papel descrever o desempenho das métricas objetivas ao serem comparadas com métricas subjetivas. Estas medidas se baseiam na análise

de alguns parâmetros como a acurácia e monotonicidade, que são utilizadas para a correlação com os escores subjetivos.

O cálculo da acurácia é feita através do uso dos coeficientes de correlação de Pearson (PLCC - *Pearson Linear Coefficients Correlation*). O resultado desta medida estatística está contido no intervalo de -1 a 1, onde quanto mais próximo o valor absoluto de PLCC for de 1 melhor é a correlação entre a métrica objetiva e a métrica subjetiva. A Equação 43 traz a definição de PLCC, que é calculado por um conjunto de escores de vídeos dado por ξ.

$$\text{PLCC} = \frac{\sum_{k=1}^{\xi} (\mu_k - \overline{\mu})(\nu_k - \overline{\nu})}{\sqrt{\sum_{k=1}^{\xi} (\mu_k - \overline{\mu})^2} \sqrt{\sum_{k=1}^{\xi} (\nu_k - \overline{\nu})^2}}, \tag{43}$$

em que μ_k e ν_k são os valores dos escores dos vídeos subjetivos e objetivos, respectivamente, e os parâmetros $\overline{\mu}$ e $\overline{\nu}$ são as médias destes conjuntos.

Para a predição da monotonicidade é usada a medida estatística de coeficientes de correlação de postos de Spearman (SROCC - *Spearman's Rank Correlation Coefficient*). Esta medida acompanha as alterações de magnitude entre os escores objetivos dados por ρ_k e os escores subjetivos γ_k. A Equação 44 define a expressão de SROCC.

$$\text{SROCC} = \frac{\sum_{k=1}^{\xi} (\rho_k - \overline{\rho})(\gamma_k - \overline{\gamma})}{\sqrt{\sum_{k=1}^{\xi} (\rho_k - \overline{\rho})^2} \sqrt{\sum_{k=1}^{\xi} (\gamma_k - \overline{\gamma})^2}}, \tag{44}$$

em que $\overline{\rho}$ e $\overline{\gamma}$ são as médias dos escores objetivos e subjetivos, respectivamente. O resultado de SROCC esta contido entre -1 e 1, e quanto mais próximo destes valores melhor é a sua correlação de postos.

3.3 EXTRAÇÃO DE SALIÊNCIAS

Esta seção descreve o modelo proposto nesta dissertação para a extração de saliências. O método se baseia no modelo de *background* descrito na seção 2.2.5, porém, são propostos algumas modificações com o intuito de aumentar a velocidade de processamento e de eliminar regiões salientes muito pequenas que não agreguem informações relevantes para a avaliação de qualidade de vídeo. Estas alterações são a implementação de um filtro e a adaptação de um modelo de detecção de saliência dupla, sendo estes descritos na sequência.

Para eliminar regiões muito pequenas (menor que 8 *pixels* agrupados) no mapa de saliência extraído através dos modelos apresentados na seção 2.2.5, um filtro de tamanho $8x8$ é utilizado. O filtro $8x8$ serve para eliminar regiões de saliência muito pequenas que não conse-

guem atrair a atenção, além de tornar o processamento do algoritmo mais veloz, já que o mapa resultante é constituído por macroblocos de 8x8 similar aos macroblocos encontrados na DCT. Este processo pode ser observado na Figura 11(c), onde o desenho esquemático mostra como o filtro atua. Este filtro faz com que o mapa de saliência (Figura 11(b)) extraído do quadro original (Figura 11(a)) seja transformado no mapa da Figura 11(d).

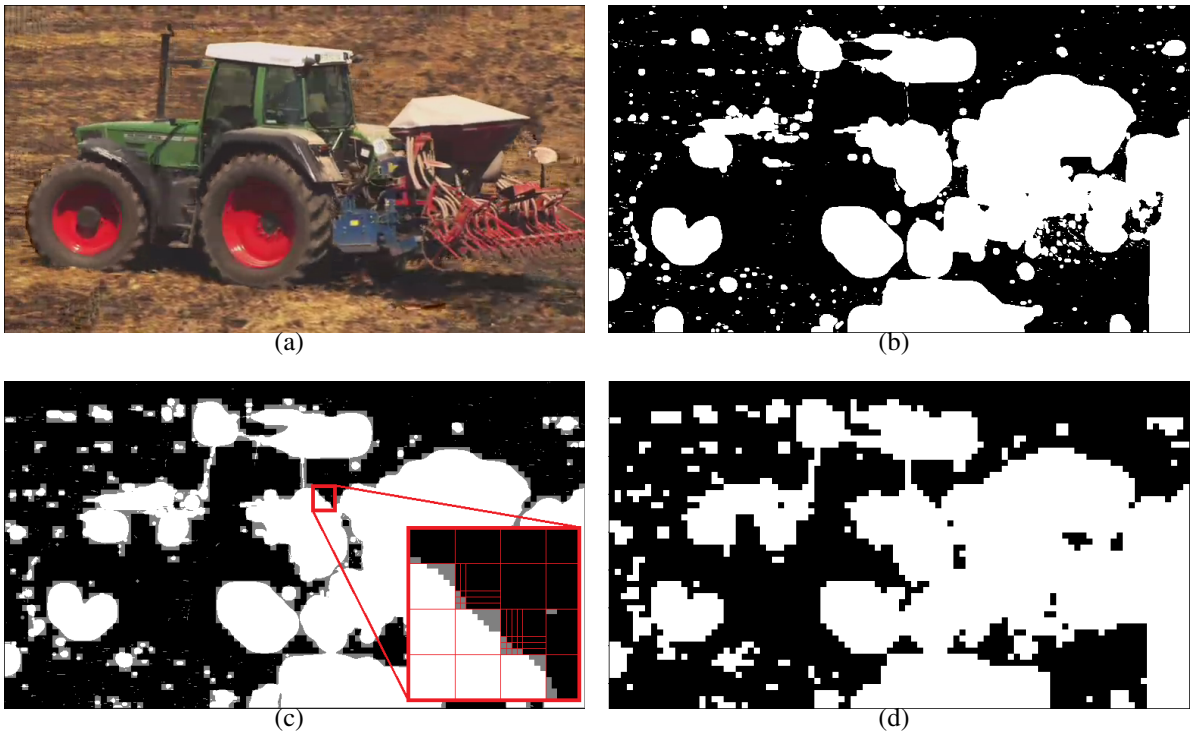


Figura 11: Processo de extração de saliências através do uso de um filtro 8x8. a) Quadro Processado. b) Mapa de saliência com o uso do modelo de Ćulibrk *et al.* (2011). c) Filtro 8×8 para extração de macro regiões salientes. d) Mapa de saliência filtrado.

Assim, um mapa de saliência para cada quadro dos vídeos é obtido. No caso das duas métricas NR propostas, que serão apresentadas na seção seguinte, o mapa de saliência utilizado para extração das características é o obtido ao fim deste processo, porém, no caso da métrica FR é utilizada uma modificação deste mapa.

A métrica FR proposta utiliza o quadro original e o quadro processado para predizer um escore objetivo de avaliação. Assim, o modelo de extração também utilizará ambos os quadros para montar um mapa de saliências com as características discrepantes. Para isto foi utilizado um modelo de detecção de saliência dupla proposta por Wang *et al.* (2012) conforme mostra a Figura 12. Primeiramente, são extraídos os *pixels* salientes dos quadros original e processado. Os mapas de saliência dos dois quadros são então combinados formando o terceiro mapa com uma união das regiões salientes.

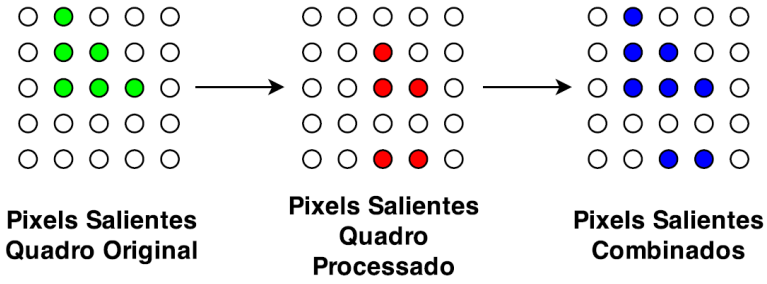


Figura 12: Ilustração de detecção de saliência dupla.

Fonte: Adaptado de Wang *et al.* (2012).

3.3.1 EXTRAÇÃO DE CARACTERÍSTICAS SALIENTES

As características espaciais abordadas na Subseção 2.1.2.3 detectam artefatos de blocagem (característica B) e borramento (características A e Z). Este trabalho propõe a divisão dos quadros em regiões salientes e não salientes para então extrair estas características separadamente em cada uma das duas regiões. Para definir estas regiões a segmentação é feita baseada nos mapas de saliência, como no exemplo da Figura 11(d).

As três componentes espaciais são extraídas por uma varredura bloco a bloco de um quadro da sequencia de vídeo, que busca detectar os artefatos gerados pela DCT no processo de compressão. O mapa de saliência exemplificado na Figura 11(d) também é formado por blocos $8x8$ *pixels* propositalmente, já que a varredura do processo de extração das características também utiliza blocos desta dimensão.

Com a informação do bloco de saliência e das características espaciais pode-se então segmentar a informação em duas partes, as contidas na região saliente e as não contidas na região de saliência, conforme a descrição da Tabela 1.

Tabela 1: Características baseadas em modelos de saliência.

Características Originais	Características das Regiões Salientes	Características das Regiões Não-Salientes
A	AS	ANS
B	BS	BNS
Z	ZS	ZNS

As características AS, BS e ZS são extraídas das regiões contidas na região de saliência, ou seja, regiões do vídeo localizadas no quadro nas mesmas coordenadas das regiões brancas da Figura 11(d). ANS, BNS e ZNS seguem a mesma ideia, porém extraindo as características localizadas das regiões não salientes, ou seja, das regiões escuras da Figura 11(d).

Considera-se MS_j como o mapa de saliência binário (os blocos salientes/brancos recebem o valor de '1' e os de regiões não salientes/escuras o valor de '0'.) exemplificado na Figura 10(d). Assim, MS_j é utilizado para o cálculo de características contidas ou não nas regiões de saliência. Para isto, um contador M_s de regiões de saliência de cada quadro é criado, contando o total de números de '1' em MS_j, como mostrado a seguir.

$$M_s = \sum_{j=1}^{M} \mathrm{MS}_j. \tag{45}$$

Assim, para extrair as características AS e ANS, utiliza-se a informação contida no mapa binário MS_j e a definição de A original dada pela Equação 23 definida no capítulo anterior. As Equações 46 e 47 definem as expressões gerais de AS e ANS.

$$\mathrm{AS} = \frac{1}{M_s} \sum_{j=1}^{M} \mathrm{A}(x_j, y_j), \qquad \text{se } \mathrm{MS}_j = 1, \tag{46}$$

$$\mathrm{ANS} = \frac{1}{M - M_s} \sum_{j=1}^{M} \mathrm{A}(x_j, y_j), \text{se } \mathrm{MS}_j = 0. \tag{47}$$

Seguindo um modelo análogo são extraídas as características BS, BNS, ZS e ZNS, baseadas nas formas originais de B e Z, definidas nas Equações 19 e 29, respectivamente. A seguir são expostas as expressões que representam suas definições.

$$\mathrm{BS} = \frac{1}{M_s} \sum_{j=1}^{M} \mathrm{B}(x_j, y_j), \qquad \text{se } \mathrm{MS}_j = 1, \tag{48}$$

$$\mathrm{BNS} = \frac{1}{M - M_s} \sum_{j=1}^{M} \mathrm{B}(x_j, y_j), \text{se } \mathrm{MS}_j = 0. \tag{49}$$

$$\mathrm{ZS} = \frac{1}{M_s} \sum_{j=1}^{M} \mathrm{Z}(x_j, y_j), \qquad \text{se } \mathrm{MS}_j = 1, \tag{50}$$

$$\mathrm{ZNS} = \frac{1}{M - M_s} \sum_{j=1}^{M} \mathrm{Z}(x_j, y_j), \text{se } \mathrm{MS}_j = 0. \tag{51}$$

Outra característica que será extraída juntamente com suas versões salientes e não salientes é o Blockiness (B_F) (BABU *et al.*, 2008). Esta característica foi descrita na Secção 2.1.2.3 como uma métrica NR e sua definição é dada pela Equação 13. Seguindo o mesmo modelo de extração utilizando o mapa de saliências, as versões saliente ($\mathrm{B}_F\mathrm{S}$) e não salientes

(B_FNS) são apresentadas a seguir.

$$B_F S = \frac{1}{M_s} \sum_{j=1}^{M} B_F(x_j, y_j), \qquad \text{se } MS_j = 1, \tag{52}$$

$$B_F NS = \frac{1}{M - M_s} \sum_{j=1}^{M} B_F(x_j, y_j), \text{se } MS_j = 0. \tag{53}$$

A seguir é apresentado o gráfico da Figura 13 com o comportamento das características espaciais acima descritas em uma sequência de vídeo da base LIVE. O eixo x é dado pelo número de quadros do vídeo (250 quadros) e o eixo y é a amplitude da informação espacial para cada característica. Nota-se alterações no comportamento nas proximidades dos quadros 100 e 200, onde ocorrem distorções ocasionadas no processo de codificação IP. Este gráfico mostra informações comportamentais das características em uma sequência de vídeo, onde estas ainda não podem ser utilizadas para fazer a predição de qualidade do vídeo, porém servem de parâmetros para o desenvolvimento das métricas que são apresentadas na próxima secção.

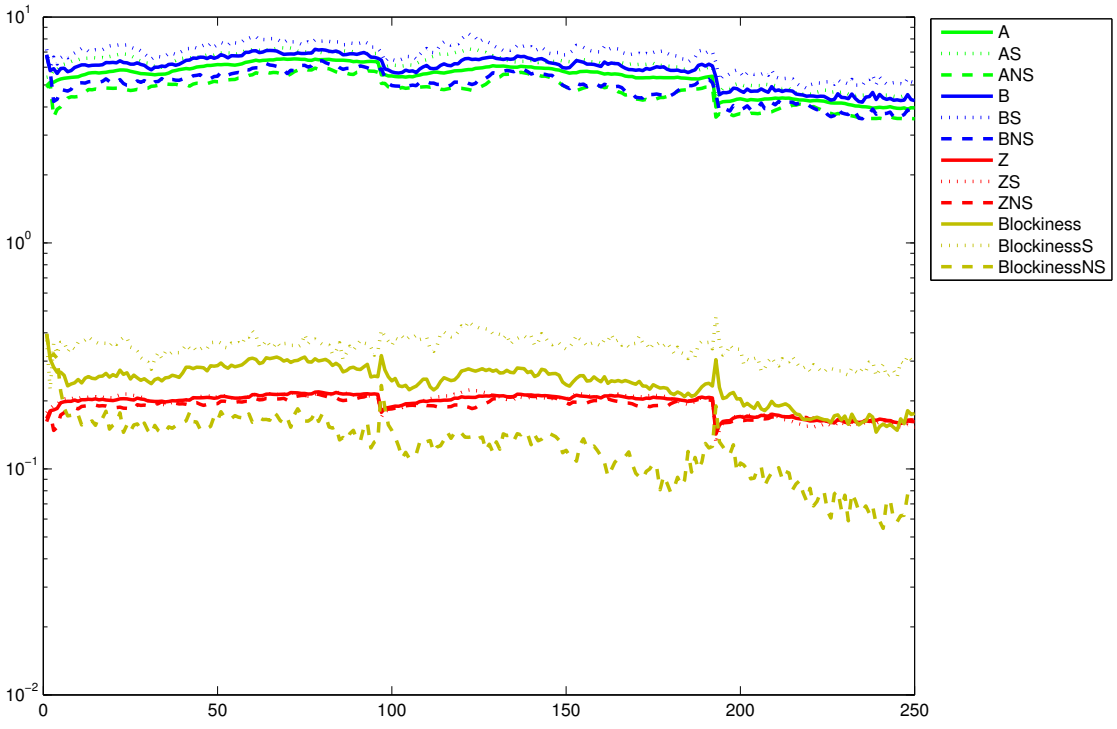


Figura 13: Comportamento das características espaciais A, B, Z e Blockiness em uma sequencia de vídeo processada da base LIVE com 250 quadros.

Os parâmetros extraídos de um quadro do vídeo por este método de segmentação buscam separar aspectos contidos em regiões que têm maior influência no SVH, procurando então aumentar a correlação das métricas que as utilizam.

3.4 MÉTODOS PROPOSTOS

Nesta seção são descritos os métodos propostos na dissertação. A métrica FR baseada em similaridade estrutural (WANG *et al.*, 2004), utilizando modelos de saliência, é a SSIM-SM (*Structurail SIMilarity metric using Salient Model*). As duas métricas NR baseadas em um modelo matemático sigmoidal e no algoritmo de Levenberg-Marquardt (MARQUARDT, 1963), utilizando modelos de saliência, são a NRLM-SM (*No-Reference metric based on Levenberg-Marquardt method using Salient Model*) e a NRLM-SMb (*No-Reference metric based on Levenberg-Marquardt method using Salient Model and Blockiness feature*).

3.4.1 SSIM-SM

A métrica FR SSIM-SM utiliza o vídeo de referência e o vídeo processado para processar os escores objetivos. A Figura 14 traz um diagrama esquemático para exemplificar a extração dos mapas de saliência utilizando ambos os vídeos.

A coluna da esquerda na Figura 14 representa o vídeo original, enquanto a coluna da direita o vídeo processado. A primeira etapa no processamento dos vídeos é a extração do mapa de saliências pelo modelo de *background* descrito na Secção 2.2.5 e exemplificado na Figura 7. A seguir, é feita a filtragem de ambos os quadros salientes com o filtro $8x8$ como demonstrado na Figura 11(c), resultando em dois quadros do mapa de saliência formados por blocos $8x8$. A próxima etapa é feita pela combinação destes 2 mapas, utilizando o processo descrito na Figura 12, onde as regiões vermelhas do mapa a direita do campo 'Combinação dos Mapas de Saliência' representa as regiões, onde apenas um dos quadros possuía saliências nessas coordenadas. Assim, é obtido um mapa saliente final, como mostrado na última figura da coluna esquerda. Por fim, para fins de análise, é sobreposto o mapa saliente no quadro processado do vídeo, demonstrando as regiões que serão utilizadas na métrica SSIM-SM.

Como a métrica SSIM-SM é baseada na estrutura de similaridade, o mapa de similaridade entre o vídeo original e processado é extraído através da Equação 7, como exemplificado na Figura 15.

Através do mapa de similaridade calcula-se o valor de SSIM geral para o quadro, dado por MSSIM, mostrado a seguir.

$$\mathrm{MSSIM}(X,Y) = \frac{1}{M}\sum_{j=1}^{M}\mathrm{SSIM}(x_j,y_j), \tag{54}$$

onde M é o número total de *pixels* do mapa de similaridade. Para criar a versão saliente e

não saliente de SSIM, é necessário calcular separadamente as regiões de interesses correspondentes, utilizando o mapa de saliência obtido através do esquema da Figura 14. Considerando o contador M_s de blocos salientes obtido na Equação 45, é calculado um valor médio de estrutura de similaridade das regiões salientes e não-salientes de um quadro do vídeo conforme, respectivamente, as Equações 55 e 56.

$$\text{MSSIM\_S} = \frac{1}{M_s}\sum_{j=1}^{M} \text{SSIM}(x_j, y_j), \qquad \text{se } \text{MS}_j = 1, \tag{55}$$

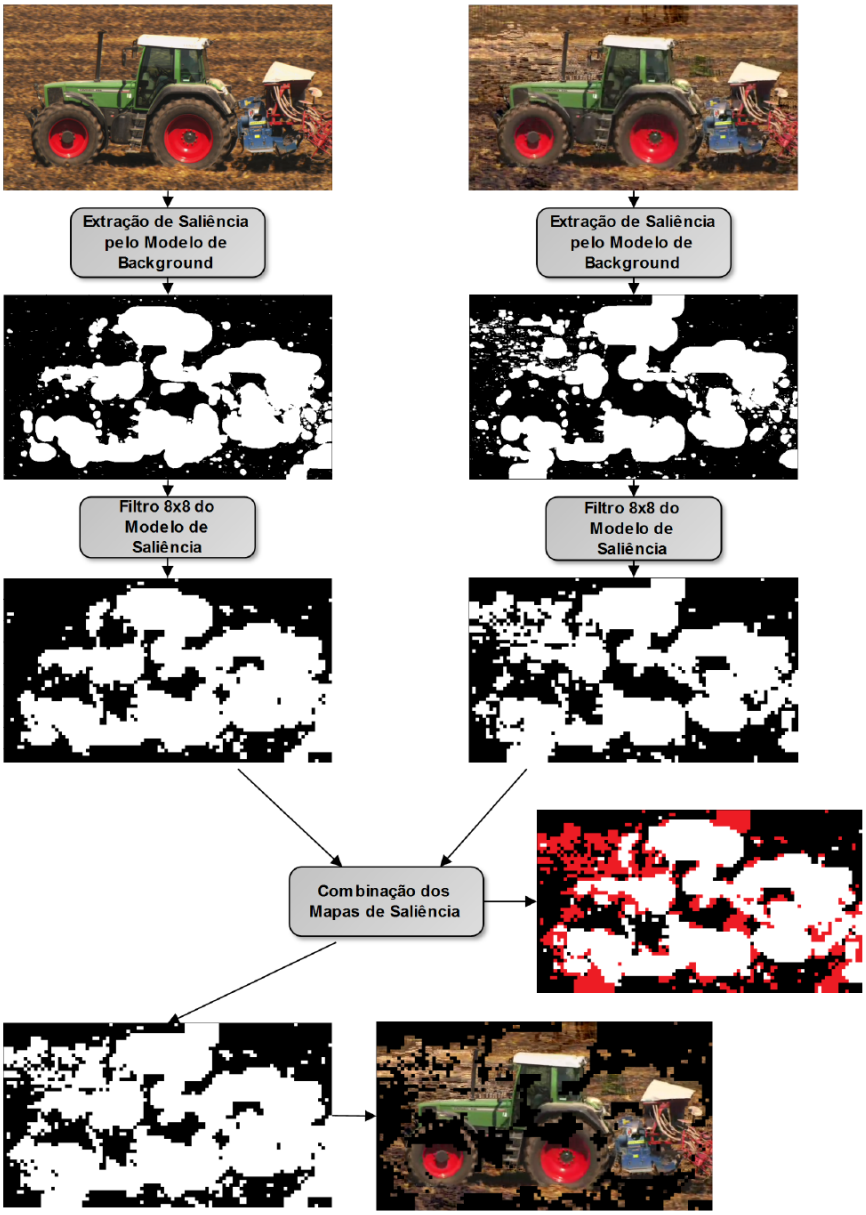


Figura 14: Diagrama esquemático da extração de regiões de saliência para a métrica SSIM.
Fonte: Imagem própria utilizando vídeo da base LIVE (SESHADRINATHAN *et al.*, 2010).

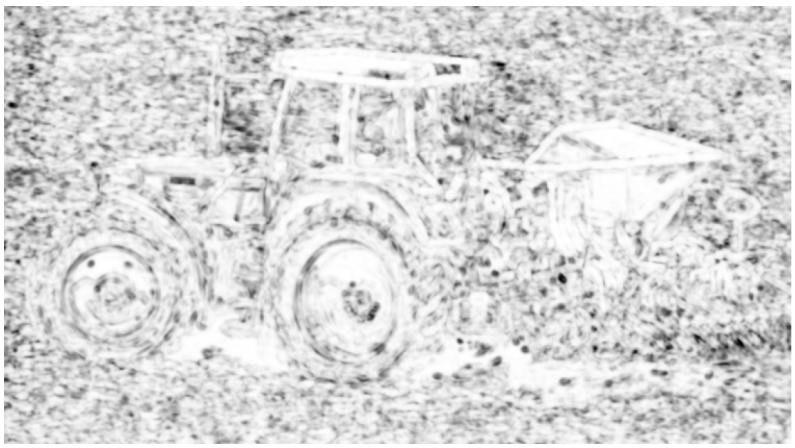


Figura 15: Mapa de estrutura de similaridade de um quadro da sequência de vídeo 'tractor' da base de dados LIVE.

Fonte: Quadro extraído da base LIVE (SESHADRINATHAN *et al.*, 2010) e mapa de similaridade baseado em Wang *et al.* (2004).

$$\text{MSSIM\_NS} = \frac{1}{M - M_s} \sum_{j=1}^{M} \text{SSIM}(x_j, y_j), \text{se MS}_j = 0. \tag{56}$$

Finalmente, para obter o escore objetivo da métrica para uma sequência de vídeo, a média de MSSIM de todos os quadros é calculada. Resultando nas seguintes expressões.

$$\text{SSIM\_std} = \frac{1}{N_f} \sum_{i=1}^{N_f} \text{MSSIM}_i, \tag{57}$$

$$\text{SSIM\_S} = \frac{1}{N_f} \sum_{i=1}^{N_f} \text{MSSIM\_S}_i, \tag{58}$$

$$\text{SSIM\_NS} = \frac{1}{N_f} \sum_{i=1}^{N_f} \text{MSSIM\_NS}_i, \tag{59}$$

onde N_f é o número total de quadros da sequencia de vídeo. A Equação 57 é a métrica padrão proposta por Wang *et al.* (2004), e as métricas propostas neste trabalho são dadas pelas, Equação 58 que é a métrica SSIM-SM proposta neste trabalho, e a Equação 59 que é a métrica complementar com a análise das regiões não salientes.

3.4.2 NRLM-SM

A métrica NRLM-SM baseada no modelo sigmoidal usando o algoritmo de Levenberg-Marquardt é baseada na Equação 71, descrita na Secção 2.1.2.3.

O processo de modelagem saliente na métrica original NRVQA-LM proposta por Silva e Pohl (2012) é análogo ao elaborado para a métrica SSIM-SM e exposto na Figura 14, com a

diferença de avaliar apenas o vídeo processado, já que o vídeo de referência não está disponível para avaliações objetivas NR.

Em um quadro, calculamos as características espaciais A, B e Z contidas nas regiões salientes e não salientes, como mostrado nas Equações 46 a 51. Para aplicar estas características em uma avaliação de um vídeo é preciso calcular o valor de cada uma delas para uma sequência do vídeo processado a ser mensurado. Este cálculo é mostrado através das Equações 61 até 69, sendo o resultado dado pela média dos valores de cada quadro no vídeo completo.

$$\mathrm{A} = \frac{1}{N_f}\sum_{i=1}^{N_f}\mathrm{A}_i, \tag{61}$$

$$\mathrm{B} = \frac{1}{N_f}\sum_{i=1}^{N_f}\mathrm{B}_i, \tag{62}$$

$$\mathrm{Z} = \frac{1}{N_f}\sum_{i=1}^{N_f}\mathrm{Z}_i, \tag{63}$$

$$\mathrm{AS} = \frac{1}{N_f}\sum_{i=1}^{N_f}\mathrm{AS}_i, \tag{64}$$

$$\mathrm{BS} = \frac{1}{N_f}\sum_{i=1}^{N_f}\mathrm{BS}_i, \tag{65}$$

$$\mathrm{ZS} = \frac{1}{N_f}\sum_{i=1}^{N_f}\mathrm{ZS}_i, \tag{66}$$

$$\mathrm{ANS} = \frac{1}{N_f}\sum_{i=1}^{N_f}\mathrm{ANS}_i, \tag{67}$$

$$\mathrm{BNS} = \frac{1}{N_f}\sum_{i=1}^{N_f}\mathrm{BNS}_i, \tag{68}$$

$$\mathrm{ZNS} = \frac{1}{N_f}\sum_{i=1}^{N_f}\mathrm{ZNS}_i, \tag{69}$$

em que N_f é o número total de quadros do vídeo.

A métrica saliente NRLM-SM utiliza as características AS, BS e ZS na sua equação sigmoidal, enquanto a versão não saliente (NRLM-NSM) usa ANS, BNS e ZNS. As Equações 71, 72 e 73 mostram a métrica original NRVQA-LM e as métricas baseadas em modelos de saliência propostas, NRLM-SM e NRLM-NSM.

$$\text{NRVQA-LM} = \frac{1}{1+e^{\left(\beta_1 B+\beta_2 Z+\beta_3 A+\beta_4 TI+\beta_5 \overline{MAD}+\beta_6 MADw+\beta_7\right)}}, \tag{71}$$

$$\text{NRLM-SM} = \frac{1}{1+e^{\left(\beta_1 BS+\beta_2 ZS+\beta_3 AS+\beta_4 TI+\beta_5 \overline{MAD}+\beta_6 MADw+\beta_7\right)}}, \tag{72}$$

$$\text{NRLM-NSM} = \frac{1}{1+e^{\left(\beta_1 BNS+\beta_2 ZNS+\beta_3 ANS+\beta_4 TI+\beta_5 \overline{MAD}+\beta_6 MADw+\beta_7\right)}}, \tag{73}$$

Para obter-se os resultados por estas métricas é necessária a divisão dos vídeos das bases de dados em dois grupos, de treinamento e de teste, e então são obtidos os escores objetivos para estes grupos disjuntos. Para obter um escore final para os grupos disjuntos são feitas 100 repetições entre os grupos e é retirada a média deste valor. Utilizando grupos disjuntos de vídeos se mantém a propriedade sem referência da métrica.

3.4.2.1 NRLM-SMB

A métrica NRLM-SMb utiliza o mesmo método de avaliação de qualidade de vídeo descrito na métrica NRLM-SM, porém substitui a característica espacial de detecção de blocagem B pela característica Blockiness (B_F) (BABU *et al.*, 2008) apresentada na Secção 2.1.2.3 e definida pela Equação 13. O Blockiness pode também ser utilizado como métrica NR como exposto nesta dissertação.

A característica B_F apresenta um valor entre 0 e 1 para cada quadro no vídeo. Assim, para atribuir apenas um valor de Blockiness para a sequencia de vídeo é calculada a média destes valores, portanto, a característica B_F será definida pela Equação 74.

$$\mathrm{B}_F = \frac{1}{N_f}\sum_{i=1}^{N_f} \mathrm{B}_{Fi}, \tag{74}$$

em que N_f é o número total de quadros do vídeo. Seguindo o mesmo procedimento são extraídas as características $\mathrm{B}_F\mathrm{S}$ e $\mathrm{B}_F\mathrm{NS}$ para a sequencia de vídeo, utilizando para isto Equações 52 e 53, resultando nas expressões a seguir.

$$\mathrm{B}_F\mathrm{S} = \frac{1}{N_f}\sum_{i=1}^{N_f} \mathrm{B}_F\mathrm{S}_i, \tag{75}$$

$$\mathrm{B}_F\mathrm{NS} = \frac{1}{N_f}\sum_{i=1}^{N_f} \mathrm{B}_F\mathrm{NS}_i, \tag{76}$$

Fazendo a substituição das informações de blocagem B por B_F é criada a métrica NRLMb, e do mesmo modo suas versões saliente NRLM-SMb e não saliente NRLM-NSMb com as características $\mathrm{B}_F\mathrm{S}$ e $\mathrm{B}_F\mathrm{NS}$, respectivamente. Estes métodos são definidos pelas equações abaixo.

$$\text{NRLMb} = \frac{1}{1+e^{\left(\beta_1 B_F+\beta_2 Z+\beta_3 A+\beta_4 TI+\beta_5 \overline{MAD}+\beta_6 MADw+\beta_7\right)}}, \tag{77}$$

$$\text{NRLM-SMb} = \frac{1}{1+e^{\left(\beta_1 B_F S+\beta_2 ZS+\beta_3 AS+\beta_4 TI+\beta_5 \overline{MAD}+\beta_6 MADw+\beta_7\right)}}, \tag{78}$$

$$\text{NRLM-NSMb} = \frac{1}{1+e^{\left(\beta_1 B_F NS+\beta_2 ZNS+\beta_3 ANS+\beta_4 TI+\beta_5 \overline{MAD}+\beta_6 MADw+\beta_7\right)}}, \tag{79}$$

A validação das métricas sigmoidais são feitas utilizando os β_s (β_1 = -0,3922 , β_2 = 41,9226, β_3 = -0,1441, β_4 = 0,0223 , β_5 = -0,5875, β_6 = 9,1590 e β_7 = -2,4752) otimizados propostos por Silva (2013). Também são utilizados apenas os resultados obtidos com grupos disjuntos de vídeos de treinamento e teste, respeitando a natureza NR da métrica.

3.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram abordados os materiais e métodos utilizados na dissertação, como as bases de vídeos, extração dos mapas de saliência e extração das características salientes. Também foram apresentados os métodos propostos no trabalho.

4 RESULTADOS E DISCUSSÕES

Neste capítulo são apresentados os resultados das métricas SSIM-SM, NRLM-SM e NRLM-SMb propostas. Estes são comparados com os resultados de suas respectivas métricas originais e versões não salientes com intuito de mostrar a eficácia do uso de modelos de saliência na avaliação de qualidade de vídeo. Ainda é feita uma comparação do desempenho das métricas com outras métricas adotadas como referência na literatura (LUO *et al.*, 2014) (LI; MA, 2012).

Para realizar estes experimentos e obter os resultados são utilizadas as bases de dados LIVE (SESHADRINATHAN *et al.*, 2010) e IVP (LIN *et al.*, 2012).

Todos resultados apresentados no trabalho seguem as recomendações do VQEG (VQEG, 2000, 2003, 2008, 2009, 2010) e da ITU (ITU-T P.910, 1999; ITU-R, 2004), ou seja, podem ser reproduzidos e comparados a outros métodos que também seguem estas recomendações.

4.1 SSIM-SM

Nesta secção são apresentados os resultados referentes à avaliação de qualidade de vídeo das bases de dados LIVE e IVP pelas métricas SSIM, SSIM-SM e SSIM-NSM. Seguindo o objetivo de desenvolver métricas que utilizem o modelo de saliência para melhorar o desempenho de métricas de VQA, os resultados avaliam principalmente a métrica SSIM-SM, fazendo a comparação com as outras duas através dos coeficientes de correlação PLCC e SROCC.

A Figura 16 traz a distribuição da correlação entre os escores objetivos e subjetivos para cada métrica e cada tipo de vídeo da base LIVE. Os gráficos também apresentam o valor de PLCC para cada correlação e ainda apresentam uma regressão não linear de terceira ordem que é mostrada pelas curvas em vermelho.

Já a Figura 17 mostra a distribuição da correlação entre os escores subjetivos para todos os vídeos da base LIVE com a sobreposição dos resultados das métricas SSIM e SSIM-SM. Observa-se uma diferença pequena na concentração dos pontos, onde a métrica SSIM-SM está um pouco mais concentrada, o que reflete em um melhor valor de PLCC.

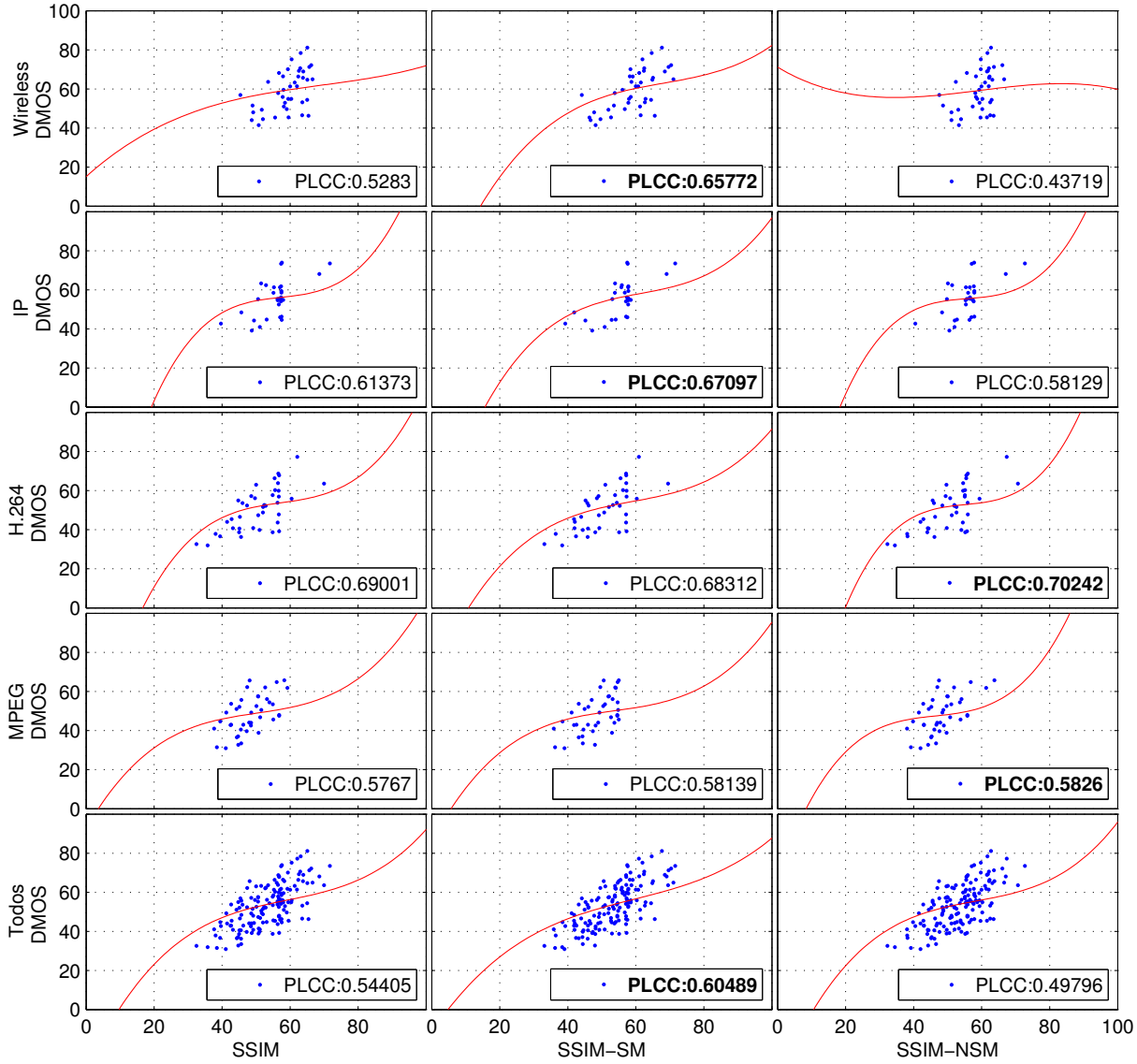


Figura 16: Comparação entre escores subjetivos e objetivos das métricas SSIM, SSIM-SM e SSIM-NSM do banco de dados LIVE.

Observando as figuras referentes às distribuições dos escores da base LIVE observa-se através da regressão não linear que os pontos para os vídeos mensurados pela métrica SSIM-SM apresentam uma maior proximidade com a diagonal de correlação entre os escores. Nota-se que esta métrica apresentou uma significativa melhora nos casos Wireless e IP, manteve basicamente o mesmo valor para os casos H.264 e MPEG e melhorou seu PLCC quando avaliados todos os vídeos. Na Tabela 2 são apresentados os valores das correlações de Pearson (PLCC) e Spearman (SROCC) entre as métricas objetivas baseadas em estrutura de similaridade (SSIM, SSIM-SM e SSIM-NSM) e a DMOS da base LIVE.

Analisando a tabela percebe-se o aumento na eficiência do método original (SSIM) utilizando o modelo de saliências principalmente nos casos dos tipos de vídeo Wireless e IP.

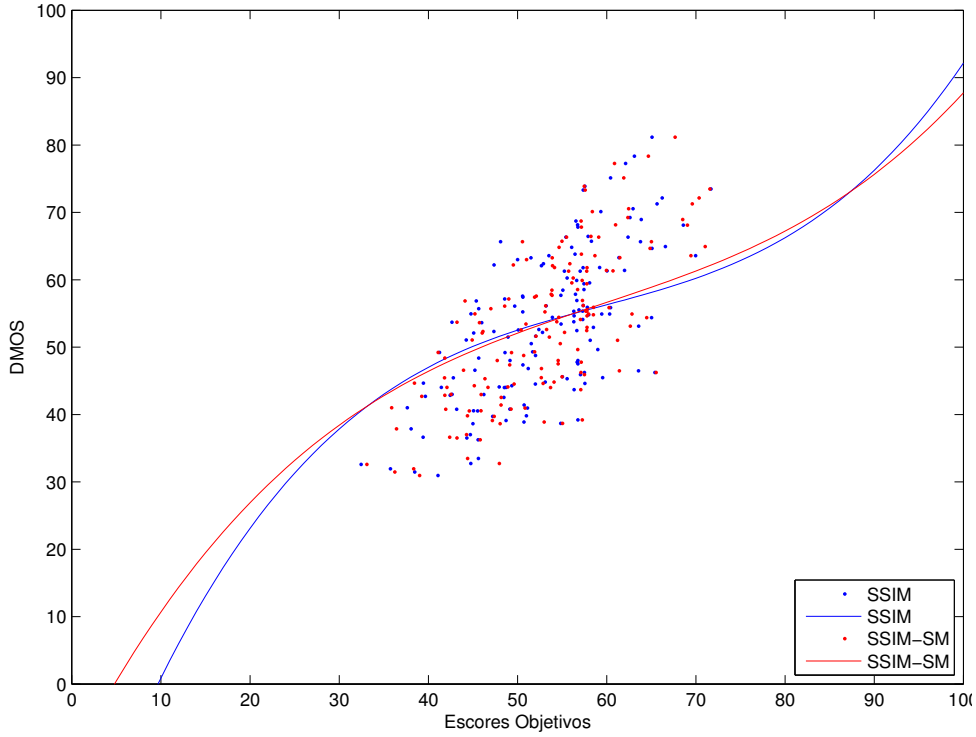


Figura 17: Comparação entre escores subjetivos e objetivos das métricas SSIM e SSIM-SM para o banco de dados LIVE.

Tabela 2: Distribuição entre os escores objetivos das métricas SSIM, SSIM-SM e SSIM-NSM e dos escores subjetivos (DMOS) da base LIVE.

	PLCC			SROCC		
	SSIM	SSIM-SM	SSIM-NSM	SSIM	SSIM-SM	SSIM-NSM
Wireless	0,5283	**0,6577**	0,4371	0,5221	**0,6345**	0,4163
IP	0,6137	**0,6709**	0,5812	0,4812	**0,5430**	0,4888
H.264	0,6900	0,6831	**0,7024**	0,6503	**0,6598**	0,6398
MPEG-2	0,5767	0,5813	**0,5826**	0,5581	**0,5699**	0,5527
Todos	0,5440	**0,6048**	0,4979	0,5248	**0,5828**	0,4714

Observa-se também que os coeficientes de postos apresentaram melhora para todos os tipos de vídeos.

Em seguida, são avaliadas as três métricas baseadas em estrutura de similaridade com a base de vídeos IVP. Os valores de PLCC e SROCC obtidos são apresentados na Tabela 3.

Como observado anteriormente para a métrica LIVE, os resultados para a base IVP apresentaram melhora com o uso do modelo de saliência, onde para todos os tipos de vídeos as correlações foram maiores para a métrica SSIM-SM. As principais melhoras ocorreram para o vídeo tipo IP. Tanto para a base LIVE como para a IVP os benefícios do modelo de saliência foram observados nos tipos de vídeos com menor correlação com a DMOS na métrica SSIM. Também se observa o efeito complementar entre as métricas SSIM-SM e SSIM-NSM, onde

Tabela 3: Distribuição entre os escores objetivos das métricas SSIM, SSIM-SM e SSIM-NSM e dos escores subjetivos (DMOS) da base IVP.

	PLCC			SROCC		
	SSIM	SSIM-SM	SSIM-NSM	SSIM	SSIM-SM	SSIM-NSM
Dirac	0,8905	**0,9012**	0,8815	0,8492	**0,8506**	0,8441
H.264	0,8752	**0,8803**	0,8724	0,8231	**0,8329**	0,8171
IP	0,5329	**0,6327**	0,4584	0,5320	**0,6127**	0,4687
MPEG-2	0,7608	**0,7710**	0,7551	0,6403	**0,6755**	0,6280
Todos	0,6521	**0,6708**	0,6411	0,6485	**0,6510**	0,6397

enquanto uma apresenta um resultado melhor que o da métrica original SSIM, o outro necessariamente esboça uma piora. Este efeito ocorre pois, as regiões de interesse (salientes) e não salientes são complementares.

4.2 NRLM-SM

As métricas NR apresentam maior complexidade no cálculo de um escore objetivo devido às dificuldades em extrair informação apenas do vídeo processado. Assim, o modelo de saliência busca auxiliar nesta extração de informação buscando dar maior atenção às áreas com maior influência no SVH. Nesta secção são apresentados os resultados para as métricas NRVQA-LM (SILVA; POHL, 2012), NRLM-SM e NRLM-NSM, onde as duas últimas são, respectivamente, as versões salientes e não salientes.

Na Tabela 4 são apresentados os valores de PLCC e SROCC para a correlação entre as medidas objetivas e a DMOS de LIVE. Os resultados apresentados são obtidos através dos métodos descritos na Seção 3.4.2.

Tabela 4: Distribuição entre os escores objetivos das métricas NRVQA-LM, NRLM-SM e NRLM-NSM e dos escores subjetivos (DMOS) da base LIVE.

	PLCC			SROCC		
	NRVQA-LM	NRLM-SM	NRLM-NSM	NRVQA-LM	NRLM-SM	NRLM-NSM
Wireless	0,3707	**0,5102**	0,3245	0,3612	**0,4827**	0,3240
IP	0,4344	**0,5057**	0,3827	0,4106	**0,4751**	0,3766
H.264	0,6128	**0,6345**	0,5980	0,6187	**0,6299**	0,5874
MPEG-2	0,8670	**0,8677**	0,8654	0,8219	**0,8386**	0,8112
Todos	0,2330	**0,2734**	0,2098	0,2240	**0,2550**	0,2123

Nota-se na Tabela 4 que os melhores resultados foram obtidos pela métrica proposta NRLM-SM, ou seja, o modelo de saliência utilizado na extração de características espaciais auxiliou na melhora do desempenho da métrica original. Também é visto que as maiores diferenças ocorreram nos tipos de vídeos Wireless e IP, que possuíam os menores valores de correlação anteriormente. Ambas as características foram observadas para na métrica FR SSIM-SM.

Outro experimento realiza o procedimento análogo para a base de dados IVP. Os resultados são expressos na Tabela 5.

Tabela 5: Distribuição entre os escores objetivos das métricas NRVQA-LM, NRLM-SM e NRLM-NSM e dos escores subjetivos (DMOS) da base IVP.

	PLCC			SROCC		
	NRVQA-LM	NRLM-SM	NRLM-NSM	NRVQA-LM	NRLM-SM	NRLM-NSM
Dirac	0,8071	**0,8112**	0,7964	0,7756	**0,7925**	0,7530
H.264	0,7880	**0,7840**	0,7950	0,7490	**0,7603**	0,7331
IP	0,6405	**0,6702**	0,6220	0,5622	**0,6312**	0,5288
MPEG-2	0,8335	**0,8358**	0,8317	0,7901	**0,8005**	0,7829
Todos	0,3054	**0,3262**	0,2896	0,2884	**0,3116**	0,2734

Novamente, os melhores resultados são apresentados pela métrica que utiliza o modelo de saliência. Observa-se também que os vídeos IP obtiveram maior benefício neste modelo de divisão de regiões de interesse, o que também aconteceu na métrica FR SSIM-SM.

Outro ponto importante é a característica complementar esperada entre as métricas NRLM-SM e NRLM-NSM que se mostrou evidente. Por fim, tanto o coeficiente de correlação de Pearson como o coeficiente de correlação de postos de Spearman obtiveram uma melhoria em seus resultados com o modelo de saliência aplicado. Assim, nota-se claramente a influência positiva das regiões de interesse em VQA.

Observando tanto os resultados da base LIVE quanto da base IVP pode-se notar uma baixa correlação quando todos os vídeos são avaliados juntamente, mostrando que a métrica NR é mais eficiente quando o tipo de vídeo é conhecido.

4.2.1 NRLM-SMB

Os experimentos mostrados nesta seção resultam nos escores objetivos das métricas propostas NRLMb, NRLM-SMb e NRLM-SMb. Todas as métricas são propostas nesta seção se baseiam na métrica NRVQA-LM com a utilização do algoritmo de Levenberg-Marquardt e extração de características espaço-temporais. Como explicitado na Seção 3.4 a métrica NRLMb é análoga a NRVQA-LM, onde a troca da característica de detecção de blocagem é feita para se buscar uma informação espacial de blocagem que analisa a interação das bordas dos blocos DCT, tanto interna como externamente. A característica usada foi a Blockiness (BABU *et al.*, 2008) como expressa na Equação 13.

Os resultados referentes ao experimento com a base de dados LIVE são expostos na Tabela 6.

Os melhores resultados são os da métrica NRLM-SMb. Seguindo o mesmo padrão das

Tabela 6: Distribuição entre os escores objetivos das métricas NRLMb, NRLM-SMb e NRLM-NSMb e dos escores subjetivos (DMOS) da base LIVE.

	PLCC			SROCC		
	NRLMb	NRLM-SMb	NRLM-NSMb	NRVQA-LM	NRLM-SMb	NRLM-NSMb
Wireless	0,3707	**0,5412**	0,3132	0,3612	**0,5152**	0,3110
IP	0,4344	**0,5280**	0,3720	0,4106	**0,4867**	0,3682
H,264	0,6128	**0,6510**	0,5999	0,6187	**0,6201**	0,5887
MPEG-2	0,8670	**0,8690**	0,8475	0,8219	**0,8442**	0,8050
Todos	0,2330	**0,2904**	0,2020	0,2240	**0,2633**	0,2087

métricas anteriores, os maiores aumentos na eficiência se concentraram nos vídeos Wireless e IP. Isto se deve ao fato da utilização de características que conseguem representar as distorções criadas nas compressões baseadas na DCT, como as características A, Z e Blockiness. Porém, vale a pena ressaltar os bons resultados de correlação para os vídeos MPEG, com PLCC de 0,8690.

O experimento com a base de dados IVP tem seus resultados expostos na Tabela 7.

Tabela 7: Distribuição entre os escores objetivos das métricas NRLMb, NRLM-SMb e NRLM-NSMb e dos escores subjetivos (DMOS) da base IVP.

	PLCC			SROCC		
	NRLMb	NRLM-SMb	NRLM-NSMb	NRVQA-LM	NRLM-SMb	NRLM-NSMb
Dirac	0,8071	**0,8154**	0,7913	0,7756	**0,7994**	0,7488
H,264	0,7880	**0,7877**	0,7885	0,7490	**0,7607**	0,7327
IP	0,6405	**0,6778**	0,6193	0,5622	**0,6514**	0,5019
MPEG-2	0,8335	**0,8344**	0,8325	0,7901	**0,8021**	0,7810
Todos	0,3054	**0,3398**	0,2787	0,2884	**0,3221**	0,2668

O mesmo efeito do modelo de saliência é observado, com um incremento na correlação em todos os tipos de vídeos.

Por fim, observando os resultados de todos os experimentos comprova-se os benefícios do uso do modelo de saliência tanto para métricas com referência como para as sem referência. Também nota-se que para alguns determinados tipos de vídeos a melhora na performance é mais significante do que para outros, sendo estes justamente os tipos de vídeos que apresentam os menores resultados de correlação.

4.3 SÍNTESE E DISCUSSÃO DOS RESULTADOS

As Tabelas 8 e 9 apresentam as correlações PLCC E SROCC entre a DMOS da base de vídeos LIVE e várias métricas objetivas. As métricas SSIM-SM, NRLM-SM e NRLM-SMb propostas neste trabalho estão em negrito, a seguir são listadas as métricas sem referência NRVQA-LM (SILVA; POHL, 2012) e JPEG-NR (WANG *et al.*, 2002), e as métricas com re-

ferência Temporal MOVIE, MOVIE, Spatial MOVIE (SESHADRINATHAN; BOVIK, 2010), MS-SSIM (WANG *et al.*, 2003a), VQM (LI; MA, 2012), PSNR e SSIM (WANG *et al.*, 2004).

As métricas estão subdivididas nas tabelas por uma linha horizontal, onde as métricas da parte superiores são métricas NR, e na parte inferior métricas FR. Esta divisão é feita para tornar as comparações mais justas, já que a complexidade envolvida no processo de predição de uma métrica NR é maior devido a ausência do vídeo de referência. As correlações destas métricas foram obtidas através da reprodução ou por valores citados na literatura para a mesma base de dados.

Tabela 8: Tabela de coeficientes PLCC entre métricas objetivas e a DMOS da base de dados LIVE.

	PLCC				
Método de Predição	Wireless	IP	H.264	MPEG-2	Todos
NRLM-SMb	0,5412	0,5280	0,6510	**0,8690**	0,2904
NRLM-SM	0,5102	0,5057	0,6345	0,8677	0,2734
NRVQA-LM[1]	0,3707	0,4344	0,6128	0,8670	0,2330
JPEG-NR[2]	0,5209	0,2247	0,4784	0,4216	0,1751
SSIM-SM	0,6577	0,6709	0,6831	0,5813	0,6048
Temporal MOVIE[3]	0,8371	0,7383	**0,7920**	0,8252	**0,8217**
MOVIE[3]	**0,8386**	**0,7622**	0,7902	0,7595	0,8116
Spatial MOVIE[3]	0,7883	0,7378	0,7252	0,6587	0,7451
MS-SSIM[4]	0,7170	0,7219	0,6919	0,6604	0,7441
VQM[5]	0,7325	0,6480	0,6459	0,7860	0,7236
PSNR	0,6690	0,4645	0,5293	0,3891	0,5621
SSIM[6]	0,5283	0,6137	0,6900	0,5767	0,5440

[1] (SILVA; POHL, 2012)
[3] (WANG *et al.*, 2002)
[3] (SESHADRINATHAN; BOVIK, 2010)
[4] (WANG *et al.*, 2003b)
[5] (PINSON; WOLF, 2004)
[6] (WANG *et al.*, 2004)

Tabela 9: Tabela de coeficientes SROCC entre métricas objetivas e a DMOS da base de dados LIVE.

	SROCC				
Método de Predição	Wireless	IP	H.264	MPEG-2	Todos
NRLM-SMb	0,5152	0,4867	0,6201	**0,8442**	0,2633
NRLM-SM	0,4827	0,4751	0,6299	0,8386	0,2550
NRVQA-LM	0,3612	0,4106	0,6187	0,8219	0,2240
JPEG-NR	0,5006	0,2102	0,4542	0,4063	0,1680
SSIM-SM	0,6345	0,5430	0,6598	0,5699	0,5828
Temporal MOVIE	**0,8114**	**0,7192**	**0,7797**	0,8170	**0,8055**
MOVIE	0,8109	0,7157	0,7664	0,7733	0,7890
Spatial MOVIE	0,7927	0,7046	0,7066	0,6911	0,7270
MS-SSIM	0,7285	0,6534	0,7051	0,6617	0,7361
VQM	0,7214	0,6383	0,6520	0,7810	0,7026
PSNR	0,6574	0,4167	0,4585	0,3862	0,5397
SSIM	0,5221	0,4812	0,6503	0,5581	0,5248

Em ambos os coeficientes de correlação, PLCC e SROCC, observa-se que a métrica FR

SSIM-SM é mais eficiente que as métricas clássicas PSNR e SSIM, porém tem um desempenho pior que métricas que exploram características espaciais e temporais dos vídeos, como a MOVIE e a VQM. Isto ocorre porque ela é baseada apenas na análise da estrutura de similaridade dos vídeos. SSIM-SM também possui um valor superior ao das métricas NR para os tipos de vídeos Wireless, IP, H.264 e todos juntos.

Em relação as métricas NR NRLM-SM e NRLM-SMb observa-se principalmente o bom desempenho destas para vídeos MPEG, em que NRLM-SMb apresentou PLCC de 0,8690 e SROCC de 0,8442, respectivamente, sendo este o melhor desempenho entre todas as métricas, superando assim até as métricas FR de melhor desempenho. Este resultado é bastante expressivo. No entanto, a métrica não atua de forma genérica e quando avalia o conjunto inteiro de vídeos apresenta PLCC e SROCC, respectivamente de, 0,2904 e 0,2633.

Comparando apenas as métricas NR, nota-se que NRLM-SMb apresenta valores de correlação mais altos para todos os tipos de vídeos. Um diferencial importante dos modelos salientes é a melhora expressiva na correlação dos vídeos IP comparando com as métricas NR que não utilizam este método.

Para a base de dados IVP é feita uma análise comparativa similar, mostradas nas Tabelas 10 e 11. As métricas objetivas exploradas são as mesmas das apresentadas na análise da base LIVE excluindo-se as métricas MOVIE e VQM.

Tabela 10: Tabela de coeficientes PLCC entre métricas objetivas e a DMOS da base de dados IVP.

Método de Predição	PLCC				
	Dirac	H.264	IP	MPEG-2	Todos
NRLM-SMb	0,8154	0,7877	**0,6778**	0,8344	0,3398
NRLM-SM	0,8112	0,7840	0,6702	**0,8358**	0,3262
NRVQA-LM	0,8071	0,7880	0,6405	0,8335	0,3054
JPEG-NR	0,6731	0,6833	0,4428	0,7572	0,2287
SSIM-SM	**0,9012**	**0,8803**	0,6327	0,7710	0,6708
MS-SSIM	0,8684	0,8502	0,5490	0,7676	0,6512
PSNR	0,8601	0,8712	0,6633	0,7203	**0,6870**
SSIM	0,8905	0,8752	0,5329	0,7608	0,6521

A métrica proposta SSIM-SM apresentou o melhor desempenho entre todas as métricas para os vídeos Dirac e H.264, com os valores PLCC e SROCC, respectivamente, de 0,9012 e 0,8506 para os vídeos Wireless, e PLCC e SROCC, respectivamente de, 0,8803 e 0,8329 para os vídeos IP.

Conforme a Tabela 10 a métrica NRLM-SMb apresentou o melhor desempenho para os vídeos IP com valor de PLCC de 0,6778, e a métrica NRLM-SM obteve PLCC de 0,8358 para os vídeos MPEG-2, superando a performance das demais métricas.

Tabela 11: Tabela de coeficientes SROCC entre métricas objetivas e a DMOS da base de dados IVP.

	SROCC				
Método de Predição	Dirac	H.264	IP	MPEG-2	Todos
NRLM-SMb	0,7994	0,7607	0,6514	**0,8021**	0,3221
NRLM-SM	0,7925	0,7603	0,6312	0,8005	0,3116
NRVQA-LM	0,7756	0,7490	0,5622	0,7901	0,2884
JPEG-NR	0,6310	0,6757	0,4420	0,6993	0,2406
SSIM-SM	**0,8506**	**0,8329**	0,6127	0,6755	0,6510
MS-SSIM	0,8126	0,8027	0,5194	0,6345	0,6450
PSNR	0,8460	0,7820	**0,6790**	0,7410	**0,6940**
SSIM	0,8492	0,8231	0,5320	0,6403	0,6485

Na Tabela 11 destacam-se os valores de SROCC de 0,8021 e 0,8005 das métricas NRLM-SMb e NRLM-SM, respectivamente, os tipos de vídeos MPEG-2.

Porém, seguindo o padrão apresentado pelos ensaios na base LIVE, as métricas NRLM-SM e NRLM-SMb apresentam uma correlação muito baixa quando são avaliados todos os tipos de vídeos ao mesmo tempo, com PLCC e SROCC de aproximadamente 0,3. Isto destaca o fato destas métricas serem especializadas, ou seja, para se obter performacne com alto nível de correlação devem ser utilizadas apenas com alguns tipos de vídeo, como MPEG-2.

Sintetizando, observa-se que as métricas propostas têm um desempenho comparado às métricas utilizadas como referência em VQA, destacando-se principalmente o desempenho da métrica SSIM-SM para análise dos vídeos Dirac e H.264 e das métricas NRLM-SM e NRLM-SMb para análise dos vídeos IP e MPEG-2.

4.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram apresentados os resultados das métricas SSIM-SM, NRLM-SM e NRLM-SMb propostas. Estes resultados são comparados tanto com as métricas originais (SSIM e NRVQA-LM), como também com métricas de referência na literatura. Os testes comprovaram a eficiência na melhora na predição ocasionada pelo uso do modelo de saliência em métricas VQA para alguns tipos de vídeos.

5 CONCLUSÃO E TRABALHOS FUTUROS

Esta dissertação propôs uma métrica FR e duas métricas NR de avaliação de qualidade de vídeo utilizando modelos de saliência para melhorar a predição dos escores objetivos. A métrica FR (SSIM-SM) é baseada na estrutura de similaridade entre o vídeo de referência (original) e o vídeo processado. As métricas NR (NRLM-SM e NRLM-SMb) são baseadas em uma modelagem sigmoidal de características espaço-temporais com solução de mínimos quadrados otimizados pelo algoritmo LM. Ambas as métricas NR usam as características espaciais de detecção de borramento A e Z (WANG *et al.*, 2002), e as características temporais TI (ITU-T P.910, 1999), MAD e $\overline{\text{MAD}}$ (SILVA; POHL, 2012). A diferença entre elas está na detecção de distorções de blocagem, onde NRLM-SM usa a característica B (WANG *et al.*, 2002) e a NRLM-SMb usa a característica Blockiness (BABU *et al.*, 2008). Todas as métricas apresentadas usam o modelo de atenção visual baseado na extração de saliências pela abordagem *bottom-up* (ĆULIBRK *et al.*, 2011) e utilizando o modelo de *background* (CULIBRK *et al.*, 2009) para separar as regiões de interesse dos vídeos, sendo estas utilizadas na extração do mapa de similaridade saliente na métrica SSIM-SM e na extração das características espaciais A, B, Z e Blockiness no caso das métricas NRLM-SM e NRLM-SMb. Os aspectos empregados nas métricas visam reproduzir com maior similaridade a percepção do SVH.

Os objetivos estipulados na Secção 1.2 foram cumpridos, destacando-se o item 3 que ambiciona a inserção dos modelos de saliência em métricas de VQA, assim foi criada uma metodologia que pode ser aplicada a qualquer métrica que utiliza características espaciais em seu contexto, seja esta FR, RR ou NR. As contribuições foram citadas na Secção 1.4, porém, ressalta-se o fato de todas as métricas propostas apresentarem resultados que melhoram a eficácia na predição da percepção do SVH comparando com as métricas originais que não utilizavam o modelo de atenção visual baseado em saliência. Para isto, foi de fundamental importância o uso dos modelos de atenção visual baseados tanto em características espaciais (contraste, orientação, intensidade e crominância) (ITTI; KOCH, 2001) como também em características de movimento detectadas com o modelo de *background* (ĆULIBRK *et al.*, 2011).

Outro aspecto importante das métricas propostas é que estas podem ser reproduzidas

através da metodologia descrita no Capítulo 3, além de seguir as recomendações do VQEG (VQEG, 2009) e ITU (ITU-R, 2004; ITU-T P.910, 1999), o que garante uma comparação justa com outras métricas usadas na literatura, tornando os resultados apresentados confiáveis.

Os resultados são mostrados no Capítulo 4. Para validar os resultados foram utilizadas as bases de dados LIVE (SESHADRINATHAN *et al.*, 2010) e IVP (LIN *et al.*, 2012). A métrica SSIM-SM apresentou resultados significantemente melhores para os tipos de vídeos Wireless e IP da base LIVE e os vídeos IP da base IVP. As melhorias foram observadas nos tipos de vídeos que apresentavam menor correlação na métrica SSIM (WANG *et al.*, 2004). A métrica NRLM-SM apresentou melhor desempenho que a original (NRVQA-LM (SILVA; POHL, 2012)) sem o uso de modelo de atenção para todos os tipos de vídeos, destacando-se melhora na predição dos vídeos Wireless e IP da base LIVE e também a alta correlação para os vídeos MPEG-2 em ambas as bases. A métrica NRLM-SMb, como versão da métrica NRLM-SM, apresentou resultados semelhantes a anterior, com uma pequena melhora na predição devido ao uso de uma característica de detecção de blocagem mais complexa, onde os tipos de vídeos com melhor desempenho foram os mesmos citados para a métrica NRLM-SM. Por fim, comparando as métricas propostas com métricas citadas na literatura, observa-se que os resultados das métricas NR apresentadas podem ser comparadas às métricas FR, destacando-se principalmente os resultados das métricas NRLM-SM e NRLM-SMb para os tipos de vídeos MPEG-2, com valores de PLCC e SROCC superiores a 0,8.

Como conclusão dos experimentos realizados observa-se que o uso dos modelos de saliência em VQA auxilia na predição dos escores objetivos, melhorando a correlação destes com as medidas subjetivas, ou seja, aumenta o nível de correlação das métricas com a percepção do SVH. Também se observa que, de maneira geral, em todas as métricas propostas em que os tipos de vídeos das métricas originais apresentavam menor correlação obtiveram, tais valores sofreram uma melhora pelo uso das saliências. Por fim, as métricas NR NRLM-SM e NRLM-SMb apresentam desempenho comparável a métricas FR, porém com uma resalva, que é o conhecimento prévio do tipo de vídeo a ser avaliado, apresentando, porém, um resultado inferior quando usado para um conjunto com todos os tipos de vídeos mesclados.

Como desenvolvimento futuro nesta linha de trabalho sugere-se: (1) a utilização de pesos no modelo de saliências (WANG; LI, 2007) para aprimorar as regiões salientes com maior e menor influência no SVH; (2) a busca de um método de extração de características temporais como TI e MAD salientes, seguindo o conceito de extração de características salientes aplicadas nas métricas NR propostas; (3) implantação das métricas NR de VQA em aparelhos móveis (celulares, *tablets, notebooks*) (RADOSAVLJEVIC *et al.*, 2014) ; (4) aplicação de métricas

objetivas em tempo real (LIU *et al.*, 2010); (5) testar o uso das métricas propostas em bases de vídeos que não seguem as recomendações do VQEG, como aquelas que utilizam vídeos da *web* (CULIBRK *et al.*, 2010); e por fim, (6) explorar a abordagem *top-down* (CONNOR *et al.*, 2004) de atenção visual e seus efeitos em métricas VQA.

Ainda, como uma aplicação mais direta da contribuição deste trabalho, sugere-se a utilização das métricas NRLM-SM e NRLM-SMb na avaliação de qualidade de vídeos MPEG-2 em *set-top boxes, tablets, notebooks, smart* TVs e celulares de alto desempenho. A otimização destas métricas ainda é necessária para o uso em aparelhos com baixo poder de processamento. Assim, buscando melhorar a qualidade de vídeo reproduzido ao usuário final, a sequência sugerida para este trabalho é a obtenção dos escores objetivos mais elevados para métricas NR propostas em tempo real.

REFERÊNCIAS

AKAMINE, W. Y. L.; FARIAS, M. C. Q. Video quality assessment using visual attention computational models. **SPIE Journal of Electronic Imaging**, v. 23, n. 6, p. 1–9, Dec 2014.

BABU, R. V.; PERKIS, A.; HILLESTAD, O. I. Evaluation and monitoring of video quality for uma enabled videos streaming systems. **Multimedia Tools and Application**, v. 37, p. 211–231, April 2008.

BOUJUT, H. *et al.* A metric for no-reference video quality assessment for hd tv delivery based on saliency maps. **IEEE International Conference on Multimedia and Expo (ICME)**, p. 1–5, July 2011.

BURT, P.; ADELSON, E. The laplacian pyramid as a compact image code. **IEEE Transactions on Communications**, v. 31, p. 532– 540, 1983.

CALLET, P. L.; VIARD-GAUDIN, C.; BARBA, D. A convolutional neural network approach for objective video quality assessment. **IEEE Transactions on Neural Networks**, v. 17, n. 5, p. 1316–1327, 2006.

CONNOR, C.; EGETH, H.; YANTIS, S. Visual attention: Bottom-up versus top-down. **Current Biol.**, v. 14, n. 19, p. R850–R852, 2004.

CULIBRK, D.; CRNOJEVIC, V.; ANTIC, B. Multiscale background modelling and segmentation. **16th International Conference on Digital Signal Processing**, p. 1–6, July 2009.

CULIBRK, D. *et al.* Mining web videos for video quality assessment. **International Conference of Soft Computing and Pattern Recognition (SoCPaR)**, p. 75–80, Dec. 2010.

ĆULIBRK, D. *et al.* Salient motion features for video quality assessment. **IEEE Transactions on Image Processing**, v. 20, n. 4, p. 948–958, April 2011.

DING, W. *et al.* Image and video quality assessment using neural network and SVM. **Tsinghua Science and Technology Journal**, v. 13, n. 1, p. 112–116, February 2008.

DOLLAR, P. *et al.* Behavior recognition via sparse spatio-temporal filters. **IEEE International Workshop VS-PETS, Beijing, China**, n. 65–72, Aug. 2005.

ENDO, C. *et al.* Analysis of the eye movements and its applications to image evaluation. **Proceedings of the Color Imaging Conference**, p. 153–155, Nov. 1994.

FENG, X. *et al.* Saliency inspired full-reference quality metrics for packet-loss-impaired video. **IEEE Transactions on Broadcasting**, v. 57, n. 1, p. 81–88, 2011.

FU, B. *et al.* Visual attention modeling for video quality assessment with structural similarity. **16th International Symposium on Wireless Personal Multimedia Communications (WPMC)**, p. 1–5, June 2013.

GAO, X. *et al.* Spatio-temporal salience based video quality assessment. **IEEE International Conference on Systems Man and Cybernetics (SMC)**, p. 1501–1505, Oct. 2010.

HAN, C.-H.; LEE, J.-S. Quality assessment of on-line videos using metadata. **IEEE International Conference on Multimedia and Expo (ICME)Acoustics, Speech and Signal Processing (ICASSP)**, p. 1385–1388, May 2014.

HENDERSON, J. *et al.* Visual saliency does not account for eye movements during visual search in real-world scenes. **Eye Movements: A Window on Mind and Brain, R. van Gompel, M. Fischer, W. Murray, and R. Hill (eds.), Amsterdam, The Netherlands: Elsevier**, p. 537–562, 2007.

ITTI, L.; KOCH, C. A saliency-based search mechanism for overt and covert shifts of visual attention. **Vision Research**, v. 40, p. 1489–1506, 2000.

ITTI, L.; KOCH, C. Computational modelling of visual attention. **Nature Reviews Neurosci.**, v. 2, n. 3, p. 194–203, March 2001.

ITU-R. **Recommendation BT-500: methodology for the subjective assessment of the quality for television pictures, Rev. 11**. Geneva, Switzerland, 2004.

ITU-T P.910. **Subjective video quality assessment methods for multimedia applications**. Standardization Sector of ITU, Geneva, Switzerland, 1999.

KONUK, B. *et al.* A spatiotemporal no-reference video quality assessment model. **20th IEEE International Conference on Image Processing (ICIP)**, p. 54–58, Sep. 2013.

L., I.; C., K.; E., N. A model of saliency-based visual attention for rapid scene analysis , 1998, 20 (11):1254-1259. **IEEE Transactions On Pattern Analysis And Machine Intelligence**, v. 20, n. 11, p. 1254–1259, 1998.

LAPTEV, I. *et al.* Local velocity-adapted motion events for spatial-temporal recognition. **Computer Vision and Image Understanding**, p. 207–229, 2007.

LEVENBERG, K. A method for the solution of certain problems in least squares. **Quarterly Applied Math**, v. 2, p. 164–168, 1944.

LI, C.; BOVIK, A. C. Content-weighted video quality assessment using a three-component image model. **Journal of Electronic Imaging**, v. 19, p. 1–9, Jan. 2010.

LI, S.; MA, L. Full-reference video quality assessment by decoupling detail losses and additive impairments. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 22, n. 99, p. 1100–1112, 2012.

LIN, X.; TIAN, X.; CHEN, Y. No-reference video quality assessment based on region of interest. **2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)**, p. 1924 – 1927, April 2012.

LIU, T. *et al.* Real-time video quality monitoring for mobile devices. **44th Annual Conference on Information Sciences and Systems (CISS)**, p. 1–6, March 2010.

LUO, Q. *et al.* Saliency and texture information based full-reference quality metrics for video qoe assessment. **IEEE Network Operations and Management Symposium (NOMS)**, p. 1–6, May 2014.

MA, L.; LI, S.; NGAN, K. Motion trajectory based visual saliency for video quality assessment. **18th IEEE International Conference on Image Processing (ICIP)**, p. 233–236, Sep. 2011.

MAHAPATRA, D.; WINKLER, S.; YEN, S. cheng. Motion saliency outweighs other low-level features while watching videos. **Proc. of SPIE-IS&T Electronic Imaging**, v. 6806, p. 1–10, 2008.

MARQUARDT, D. W. An algorithm for least-squares estimation of nonlinear parameters. **SIAM Journal on Applied Mathematics**, JSTOR, v. 11, n. 2, p. 431–441, 1963.

MOORTHY, A. K.; BOVIK, A. C. Efficient video quality assessment along temporal trajectories. **IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY**, v. 20, n. 11, p. 1653–1659, Nov. 2010.

MORÉ, J. The Levenberg-Marquardt algorithm: implementation and theory. In: WATSON, G. A. (Ed.). **Numerical Analysis**. Berlin: Springer, 1977, (Lecture Notes in Mathematics, x). cap. 10, p. 105–116.

PINSON, M.; WOLF, S. A new standardized method for objectively measuring video quality. **IEEE Transactions on Broadcasting**, IEEE, v. 50, n. 3, p. 312–322, September 2004.

RADOSAVLJEVIC, M. *et al.* Qoe-aware rate-conservative dynamic http streaming over mobile cellular networks. **IEEE Network Operations and Management Symposium (NOMS)**, p. 1–6, May 2014.

ROMANI, E.; SILVA, W. da; POHL, A. Análise comparativa de características temporais de vídeos digitais. **Simpósio da Sociedade de Engenharia de Televisão, São Paul**, p. 1–5, Aug. 2014.

SESHADRINATHAN, K.; BOVIK, A. Motion tuned spatio-temporal quality assessment of natural videos. **IEEE Transactions on Image Processing**, v. 19, n. 2, p. 335–350, February 2010.

SESHADRINATHAN, K. *et al.* Study of subjective and objective quality assessment of video. **IEEE Transactions on Image Processing**, v. 19, p. 1427–1441, June 2010.

SHABANI, A. H.; CLAUSI, D. A.; ZELEK, J. S. Improved spatio-temporal salient feature detection for action recognition. **British Machine Vision Conference**, p. 1–12, Sep. 2011.

SHABANI, A. H.; CLAUSI, D. A.; ZELEK, J. S. Evaluation of local spatio-temporal salient feature detectors for human action recognition. **Ninth Conference on Computer and Robot Vision (CRV)**, p. 468 – 475, May 2012.

SHEIKH, H. R. *et al.* **LIVE Image Quality Assessment Database**. 2003. Disponível em: <http://live.ece.utexas.edu/research/quality>.

SILVA, W. B. **Métodos Sem Referência Baseados em Características Espaço-Temporais Para Avaliação Objetiva de Qualidade de Vídeo Digital**. Tese (Doutorado em Engenharia Elétrica e Informática Industrial) — Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Março 2013. Disponível em: <http://repositorio.utfpr.edu.br/jspui/handle/1/525>.

SILVA, W. B.; POHL, A. A. P. No-reference video quality assessment method based on the Levenberg-Marquardt minimization. In: **XXX Brazilian Symposium on Telecommunications (SBrT'12)**. Brasília, Brazil: SBrT, 2012. v. 1, p. 1–4.

SILVA, W. B.; POHL, A. A. P.; FONSECA, K. V. O. A reduced-reference video quality assessment method based on the activity-difference of DCT coefficients. **IEICE Transactions on Information and Systems**, E96-D, n. 3, p. 708–718, March 2013.

SIMONE, F. D. *et al.* Subjective assessment of h.264/avc video sequences transmitted over a noisy channel. **International Workshop on Quality of Multimedia Experience**, p. 204 – 209, July 2009.

SINGH, R.; AGGARWAL, N. State of the art and research issues in video quality assessment. **Engineering and Computational Sciences (RAECS)**, p. 1–6, March 2014.

SOUNDARARAJAN, R.; BOVIK, A. C. Video quality assessment by reduced reference spatio-temporal entropic differencing. **IEEE Transactions on Circuits and Systems For Video Technology**, v. 3, n. 4, p. 684–694, April 2013.

STELMACH, L.; TAM, W. Processing image sequences based on eye movements. **Proceedings of SPIE**, v. 2179, p. 90–98, 1994.

STYLES, E. A. **Attention, Perception, and Memory: An Integrated Introduction.** 1º. ed. [S.l.]: New York: Taylor & Francis Routledge, 2005.

VQEG. **Final report from the video quality experts group on the validation of objective models of video quality assessment**. Video Quality Experts Group (VQEG): NTIA/ITS, 2000. Tech. Rep. Disponível em: <http://www.its.bldrdoc.gov/vqeg/>.

VQEG. **Final report from the video quality experts group on the validation of objective models video quality assessment, Phase II**. Video Quality Experts Group (VQEG): NTIA/ITS, 2003. Tech. Rep. Disponível em: <http://www.its.bldrdoc.gov/vqeg/>.

VQEG. **Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, Phase I**. Video Quality Experts Group (VQEG): NTIA/ITS, 2008. Tech. Rep. Disponível em: <http://www.its.bldrdoc.gov/vqeg/>.

VQEG. **Final report from the video quality experts group on the validation of reduced-reference and no-reference objective models for standard definition television, Phase I**. Video Quality Experts Group (VQEG): NTIA/ITS, 2009. Tech. Rep. Disponível em: <http://www.its.bldrdoc.gov/vqeg/>.

VQEG. **Report on the validation of video quality models for high definition video content**. Video Quality Experts Group (VQEG): NTIA/ITS, June 2010. Tech. Rep., Version 2.0. Disponível em: <http://www.its.bldrdoc.gov/vqeg/projects/hdtv/>.

WANG, A. *et al.* New no-reference blocking artifacts metric based on human visual system. In: **Proceedings of the International Conference on Wireless Communications Signal Processing (WCSP'09)**. Nanjing: IEEE, 2009. p. 1–5.

WANG, Y. *et al.* Novel spatio-temporal structural information based video quality metric. **IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY**, v. 22, n. 7, p. 989–998, July 2012.

WANG, Z.; BOVIK, A. C. Mean square error: Love it or leave it? a new look at signal fidelity measures. **IEEE Signal Processing Magazine**, p. 98–117, Jan 2009.

WANG, Z. *et al.* Image quality assessment: from error visibility to structural similarity. **IEEE Signal Processing Letters**, v. 13, n. 4, p. 600–612, 2004.

WANG, Z.; LI, Q. Video quality assessment using a statistical model of human visual speed perception. **Journal of the Optical Society of America**, v. 24, n. 12, p. B61–B69, Dec. 2007.

WANG, Z.; LU, L.; BOVIK, A. C. Video quality assessment based on structural distortion measurement. **Signal Processing: Image Communication**, v. 19, n. 2, p. 121–132, 2004.

WANG, Z.; SHEIKH, H. R.; BOVIK, A. C. No-reference perceptual quality assessment of JPEG compressed images. In: **Proceedings of the IEEE International Conference on Image Processing (ICIP'02)**. New York: IEEE, 2002. v. 1, p. I–477–I–480.

WANG, Z.; SHEIKH, H. R.; BOVIK, A. C. Objective video quality assessment. In: FURHT, B.; MARQUES, O. (Ed.). **The Handbook of Video Databases: Design and Applications**. Boca Raton, USA: CRC Press, 2003. cap. 41, p. 1041–1078.

WANG, Z.; SIMONCELLI, E. P.; BOVIK, A. C. Multiscale structural similarity for image quality assessment. In: **Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems, and Computers**. Pacific Grove, CA: IEEE Computer Society, 2003. v. 2, p. 1398–1402.

WILLEMS, G.; TUYTELAARS, T.; GOOL, L. V. An efficient dense and scale-invariant spatio-temporal interest point detector. **European Conference on Computer Vision**, p. 650–683, Oct. 2008.

WINKLER, S. Analysis of public image and video databases for quality assessment. **IEEE Jounal of Selected Topics in Signal Processing**, v. 6, n. 6, p. 616 – 626, October 2012.

YARBUS, A. L. **Eye Movements and Vision. New York**. [S.l.]: New York, USA: Plenum Press, 1967.

ZHU, K. *et al.* No-reference video quality assessment based on artifact measurement and statistical analysis. **IEEE Transactions on Circuits and Systems for Video Technology**, PP, p. 1–14, Oct. 2014.