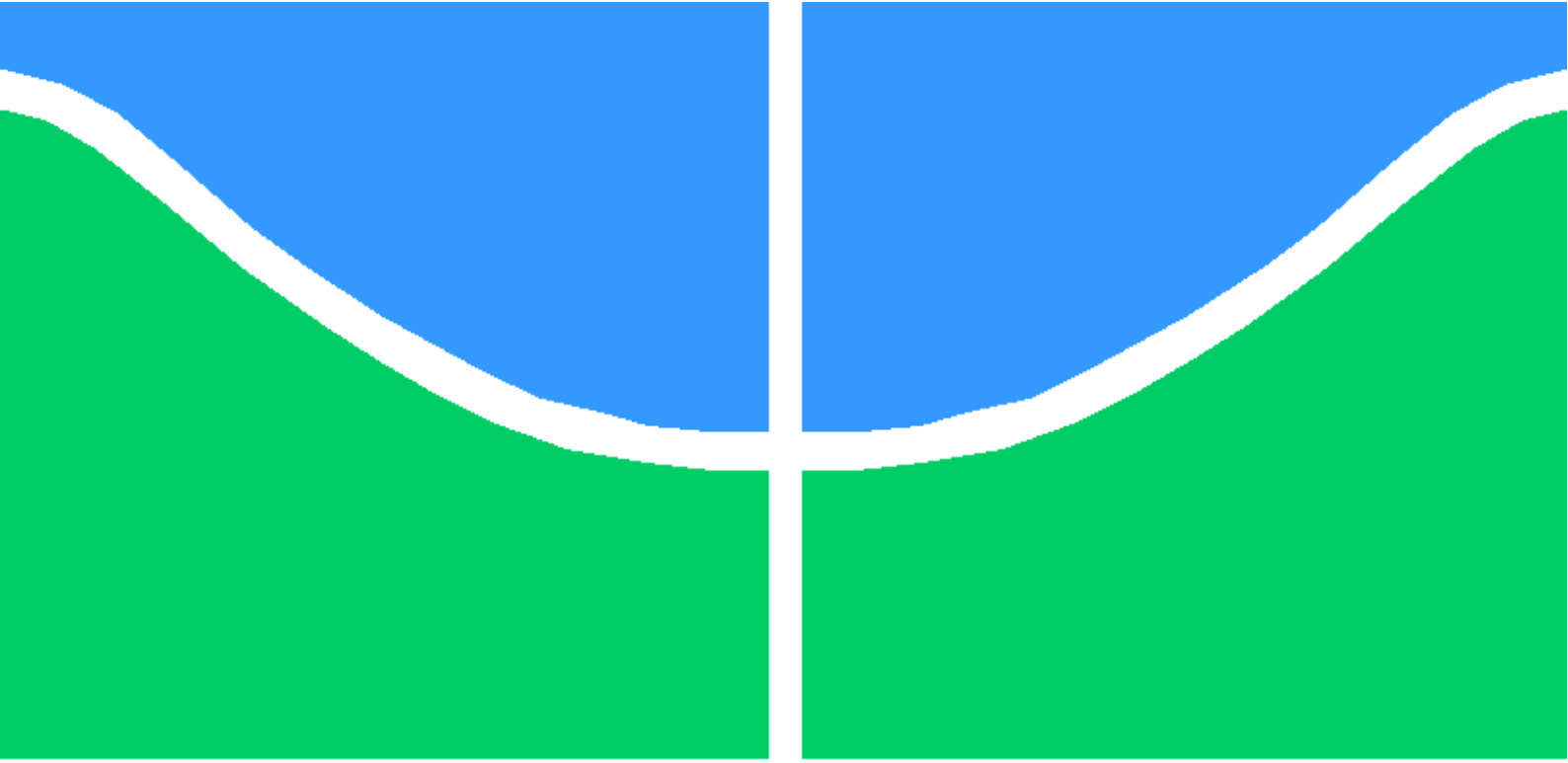


TRABALHO DE CONCLUSÃO DE CURSO

MÉTRICAS DE QUALIDADE DE VÍDEO COM CARACTERÍSTICAS TOP-DOWN DE ATENÇÃO VISUAL

Rodrigo Cerqueira Gonzalez Pena

Brasília, agosto de 2014

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASILIA
Faculdade de Tecnologia

TRABALHO DE CONCLUSÃO DE CURSO

MÉTRICAS DE QUALIDADE DE VÍDEO COM CARACTERÍSTICAS TOP-DOWN DE ATENÇÃO VISUAL

Rodrigo Cerqueira Gonzalez Pena

Relatório submetido ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Engenheiro de Eletricista

Banca Examinadora

Prof[a]. Mylène C. Q. Farias, ENE/UnB _____
Orientadora

Prof. Alexandre Zaghetto, CIC/UnB _____
Examinador interno

Prof. João Luiz Carvalho, ENE/UnB _____
Examinador interno

Dedicatória

Aos meus avós, que sempre desejaram ver o dia da minha graduação. Ao meu pai e ao meu tio, que apostaram inicialmente na Engenharia Elétrica, mesmo que depois tenham encontrado algo mais satisfatório. Por fim, ao leitor que porventura se deparar com este trabalho. Espero que encontre o que procura.

Rodrigo Cerqueira Gonzalez Pena

Agradecimentos

Agradeço à minha família, por todo incentivo e suporte em tudo o que me propus a fazer até hoje. Devo a eles o fato de acreditar que o aprendizado constante é o caminho para o sucesso e a felicidade. Agradeço à professora Mylène Farias, que, mesmo tendo uma das agendas mais ocupadas que eu já vi, sempre encontrou tempo para orientar e revisar este trabalho, bem como para sanar dúvidas e discutir assuntos não relacionados diretamente ao que foi feito aqui. Agradeço também ao Grupo de Processamento Digital de Sinais (GPDS) por ter providenciado o espaço e equipamentos necessários para que grande parte deste trabalho pudesse ser realizada. Agradeço finalmente a todos os colegas, professores e funcionários com quem tive o prazer de conviver no Brasil e na França. Vocês tornaram esses anos de estudo não só melhores, mas também inesquecíveis.

Rodrigo Cerqueira Gonzalez Pena

RESUMO

Propomos neste trabalho a criação de um modelo de saliência para vídeos baseado em um modelo de saliência para imagens com um dos melhores desempenhos na literatura. Os mapas do modelo de saliência ponderam os mapas de erro (ou de qualidade) de métricas objetivas de qualidade e mostramos que essa integração da atenção visual com a avaliação de qualidade traz melhora de desempenho na estimação da qualidade de vídeos percebida pelos seres humanos.

ABSTRACT

We propose in this work the creation of a saliency model for videos based in a saliency model for images that has one of the best performances in the literature. The maps from the saliency model weigh the error maps (or quality maps) of objective quality metrics and we show that this integration of visual attention and quality assessment improves the prediction performance of video quality as perceived by human beings.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

Capítulo 1

Introdução

No começo do século XX o contato das pessoas com a mídia do vídeo era restrita à tela de cinema. Após o advento da televisão, programações em vídeo podiam ser vistas de dentro da casa das pessoas, o que foi se expandindo até virar um meio de comunicação de massas. Hoje em dia a televisão ainda existe, o cinema se modernizou e, com o aparecimento da Internet e os avanços da microeletrônia, o número de *displays* multimídia cresceu de forma exponencial. O aumento do fluxo de dados trouxe com ele um crescimento no número de serviços, codecs e compressores de vídeo. De forma a garantir a Qualidade de Experiência (QoE) e a Qualidade de Serviço (QoS), avaliações da qualidade dos vídeos processados devem ser incorporadas à validação dos serviços ofertados. E devem ser incorporadas de forma objetiva, automática, sem que seja necessário fazer a avaliação através da opinião de pessoas (o que é lento e caro). Porém, longe de se propor avaliar a qualidade artística dos vídeos, o que se propõe a estimar é o grau de ausência de defeitos e artefatos que impeçam a compreensão de um dado vídeo. Mesmo com essa restrição, a estimação de qualidade é ainda uma tarefa díficil para métricas objetivas, pois, mais do que identificar alterações em um sinal de vídeo, os cálculos devem levar em conta apenas o que é percebido como defeito pelo sistema visual humano. Percepção aqui é a palavra chave. Se os seres humanos percebecem imagens como um fluxo de dados sem estruturação espacial, uma simples métrica de fidelidade de sinal como a PSNR (*Peak Signal-to-Noise Ratio*) seria suficiente para estimar a qualidade de um vídeo. No entanto, observemos a Figura 1.1. Winkler e Mohandas [1] dão como exemplo esta imagem para indicar como a percepção humana pode variar. O mesmo artefato foi inserido em ambas as imagens. Na imagem da esquerda o padrão inserido se encontra na região inferior, onde estão as rochas. Na imagem da direita, este padrão foi inserido no céu. Ambas tem o mesmo valor de PSNR, mas a fotografia à direita é muito mais incômoda quanto à presença de defeitos na imagem. Como na primera imagem os defeitos foram inseridos em uma região texturizada, temos um efeito de mascaramento desses artefatos. No segundo caso, a região onde se encontra os defeitos é originalmente uniforme, havendo um efeito de facilitação da percepção dos artefatos.

Apesar dessa clara limitação de desempenho da PSNR, ela continua sendo muito usada nesses tipos de avaliação. Entretanto, recentemente maiores esforços têm sido direcionados à criação de métricas baseadas no sistema visual humano e na alta estruturação espacial das imagens naturais. Atualmente, uma das métricas mais populares na avaliação da qualidade de imagens é a SSIM,

Figura 1.1: O mesmo padrão foi inserido às duas imagens. Ambas têm o mesmo valor de PSNR, mas a percepção de qualidade entre as duas é completamente diferente [1]. À esquerda, o padrão é mascarado e, à direita, a percepção do padrão é facilitada.

proposta por Wang *et al.* [14]. Esta métrica realiza medidas estruturais, de cor e de luminância, baseadas em propriedades do sistema visual humano, para fornecer um índice de qualidade para uma imagem. Esta abordagem apresenta desempenhos melhores que os das métricas de fidelidade de dados, como a PSNR. No contexto de vídeos, temos como exemplo a métrica VQM, proposta por Pinson e Wolf [15], cujo alto desempenho fez com que se tornasse padrão norte-americano na avaliação de vídeos. Apesar desses esforços, as previsões feitas por essas métricas não é perfeita e ainda há muito que ser melhorado. Uma das apostas nesse sentido é pela consideração de onde se encontra o foco de atenção das pessoas ao observar uma cena [1]. Isso porque, fora das regiões de interesse em uma imagem, a sensibilidade da percepção humana à presença de defeitos é atenuada.

A atenção visual humana é um assunto bastante estudado por psicólogos e neurocientistas há muito tempo. Evolutivamente, o sistema visual humano desenvolveu maneiras de filtrar a enorme quantidade de informações que é fornecida ao olho humano a cada instante de forma que seja possível processar e retirar significado da cena observada. Desde o primeiro quarto do século XX experimentos são feitos para rastrear onde se encontram as fixações de olhar de voluntários observando uma cena. Diferentemente do que se poderia imaginar, ao invés de percorrer a cena por completo, os olhos tendem a concentrar suas fixações em regiões de interesse, ou salientes. Além disso, como percebido por Yarbus [4], os padrões de fixação podem variar conforme a tarefa demandada à pessoa que observa a cena. Apesar desses estudos, o modelamento computacional da atenção só começou a ser estabelecido na década de 1980, com a "Teoria da Integração de Características" de Treisman e Gelade [16], trabalho no qual muitos dos modelos computacionais de atenção visual se baseiam até hoje. Vejamos a Figura 1.2, que também é um exemplo da grande quantidade de informações que uma cena pode conter. Para onde um observador olharia se o mesmo esivesse procurando alguém na multidão? Muito provavelmente fixaria o olhar em uma face de cada vez. E se o alvo procurado estivesse usando roupas azuis? Podemos imaginar como a estratégia de procura seria diferente. E se não fosse dada nenhuma tarefa ao observador? A resposta a essas perguntas será melhor discutida no Capítulo 4, mas podemos adiantar que

alguns modelos computacionais afirmam que há características intrínsecas na cena que atraem naturalmente a atenção do espectador [17]. Outros trabalhos afirmam que é a tarefa dada durante a observação que governa a maior parte das fixações de olhar [18], [19]. Há também estudos que indicam que, independentemente da tarefa, o foco de atenção tende a se concentrar no centro das imagens [6], [20].

Figura 1.2: Como o sistema visual se comporta quando procuramos alguém na multidão?

Neste trabalho fazemos uma revisão do que vem sido estudado nos domínios da estimação de qualidade de vídeo e do modelamento computacional da atenção visual humana. Além disso, propomos também uma maneira de integrar a atenção visual a métricas objetivas de qualidade de vídeo. Esta integração é feita através da ponderação, pelos mapas de saliência, dos mapas de erro gerados pelas métricas de qualidade. Conduzimos vários testes para avaliar a estratégia integração e, por fim, mostramos que há melhoras de desempenho observadas ao se considerar a atenção visual na avaliação da qualidade de vídeos.

Capítulo 2

Atenção Visual

Atenção é um mecanismo cerebral de seleção de estímulos. Embora este processo seletivo esteja presente em todos os sentidos, encontramos na atenção visual um exemplo claro do esforço necessário para se identificar estímulos relevantes em meio à enorme quantidade de dados apresentada ao olho humano. Estima-se que o fluxo de dados visuais recebidos pelo olho é da ordem de $10^8 - 10^9$ bits por segundo [21], o que justifica as adaptações evolutivas ocorridas para conseguir processar e dar sentido a essa vasta quantidade de informações. Particularmente no caso dos seres humanos, a seletividade da atenção é facilitada pelo desenvolvimento de uma fóvea (área central do olho) de alta resolução circundada por uma periferia de baixa resolução. A presença destas estruturas faz com que a atenção visual humana seja frequentemente comparada a um holofote em um quarto escuro [20]. O processo de focalização da atenção visual funcionaria então, pelo menos em primeira análise, através da orientação da fóvea do olho em direção a regiões de interesse em uma cena. Os mecanismos que geram esta orientação, bem como as maneiras de simulá-los computacionalmente são os objetos de estudo da área de Modelamento da Atenção Visual. Esta comunidade de pesquisa, com extensa atividade nos últimos 25 anos, tem por inspiração a enorme quantidade de trabalho realizada por psicólogos, neuropsicólogos e neurocientistas computacionais no estudo da atenção durante as últimas décadas [18].

Experimentos sobre os padrões de fixação do olhar humano sobre uma cena são feitos desde a década de 1920, quando os primeiros dispositivos não-intrusivos de rastreamento de olhar (*eye-trackers*, em inglês) foram inventados por G. T. Buswell [22]. Estes primeiros dispositivos disparavam raios de luz que eram refletidos nos olhos de um voluntário e por fim atingiam um filme, o que registrava a posição amostrada do olho naquele instante. *Eye-trackers* modernos contam com um apoio de cabeça (para evitar a movimentação desta) e câmeras apontadas para os olhos do observador (ver Figura 2.1a). Estas câmeras rastreiam, a uma dada taxa de amostragem, a orientação ocular do voluntário e cálculos computacionais são realizados para descobrir onde, na imagem apresentada, estão localizadas as fixações. Algumas das pesquisas mais importantes com *eye-trackers* na metade do século XX foram realizadas por Yarbus [4]. Em seu trabalho, Yarbus afirma que os movimentos oculares refletem diretamente os processos mentais humanos e que é fácil determinar a ordem e frequência dos elementos que atraem a atenção de uma pessoa através da análise das fixações de seus olhos em uma cena. Apesar de vários resultados importantes obtidos

(a) *Eye-tracker*

(b) Fixações

Figura 2.1: Exemplo (a) de dispositivo rastreador de olhar (*eye-tracker*) moderno [2] e (b) de uma imagem com as fixações de um observador superimpostas [3]

nas pesquisas de Yarbus, hoje em dia o consenso é de que a atenção visual está sempre levemente à frente (entre 100 e 250 ms) da movimentação do olho [23].

Na área de modelamento computacional da atenção visual, grande parte dos modelos são baseados na "Teoria de Integração de Características" (*Feature Integration Theory*, em inglês) desenvolvida por Treisman e Gelade [16]. Neste trabalho, os pesquisadores propuseram que a informação visual é analisada de forma paralela pelo cérebro humano, que combina diferentes caacterísticas (*features*) visuais para identificar regiões salientes em uma cena. Saliência é um termo que caracteriza a propriedade que certas regiões de uma cena tem de se destacar em relação às regiões vizinhas [18]. Inspirados na teoria de Treisman e Gelade, Koch e Ullman [17] criaram um modelo de combinação das características visuais (cor, intensidade e orientação) e introduziram o conceito de mapa de saliência. Um mapa de saliência é um mapa topológico que representa as regiões notáveis de uma cena [18]. A primeira implementação e verificação completa do modelo de Koch e Ullman foi feita por Itti *et al.* [5] e desde então uma extensa série de modelos foi proposta seguindo a ideia de combinar características visuais individuais de uma cena em um mapa de saliência [20].

Entretanto, como observado por Yarbus [4], a atenção visual humana não depende apenas de características visuais intrínsecas à cena observada, mas também das ações demandadas aos observadores durante o experimento. Em testes realizados em cima da pintura *The Unexpected Visitor*, de Ilya Repin (ver Figura 2.2), Yarbus registrou as fixações oculares de voluntários sob diferentes tarefas. Ele observou que o padrão de fixações variou consideravelmente entre quando, por exemplo, se pediu aos observadores que obsevassem livremente a cena e quando a tarefa dada foi avaliar a idade das pessoas na pintura. Essas observações, juntamente com aquelas presentes na Teoria de Integração de Características [16], sugerem que existem dois grandes mecanismos que gerem a atenção visual humana, chamados de fatores *bottom-up* e fatores *top-down*. Os fatores *bottom-up* são exemplificados pelas características presentes na Teoria de Integração de Características. Eles são derivados exclusivamente de propriedades inerentes à cena observada (cor, contraste, etc) e são também chamados de fatores automáticos, reflexivos ou guiados por estímulo [20]. Em contrapartida, fatores *top-down* são guiados por fatores cognitivos como conhecimento, expectativas e objetivos. Este último tipo de mecanismo também é conhecido como "guiado por objetivos" e "voluntário".

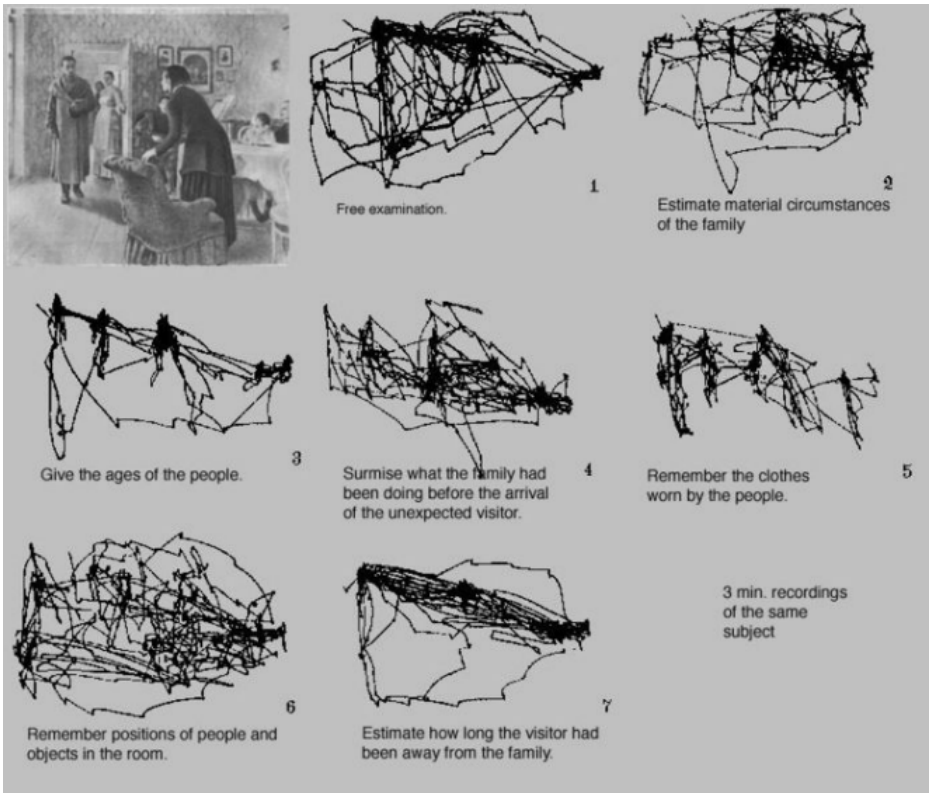


Figura 2.2: Exemplo da influência das tarefas sobre o padrão de fixações do olho humano [4]

Por fim, cabe comentar que o modelamento computacional como tem sido feito até presentemente modela apenas o que é conhecido como atenção ostensiva (*overt attention*). A atenção ostensiva é aquela à que se refere a analogia do holofote citada no começo deste capítulo. Ela consiste no processo de direcionar a fóvea para um estímulo visual em uma cena e, com isso, fixá-lo. A esta última se opõe a chamada atenção encoberta (*covert attention*), que se traduz na habilidade humana de prestar atenção em algo que não está sendo olhado diretamente. Um exemplo

disso é quando se dirige um carro: a pessoa mantém os olhos na estrada, mas está constantemente monitorando as placas e sinais de trânsito [18] no caminho. Embora tenhamos definidos os dois tipos de atenção, atualmente apenas se sabe como medir a atenção ostensiva (o que é feito com os dispositivos de ratreamento de olhar). Além disso, os mecanismos comportamentais e as funções da atenção encoberta ainda não são completamente conhecidos [18]. Embora estes dois conceitos não sejam mutuamente exclusivos, fica implícito neste trabalho que os modelos de atenção apresentados se referem a estudos sobre a atenção ostensiva.

2.1 Modelamento Bottom-up

A atenção *bottom-up* é rápida, involuntária e provavelmente se comporta como um sistema de malha aberta (*feed-forward*) [18]. Treisman e Gelade [16] citam um exemplo ilustrativo deste tipo de mecanismo: quando se observa uma cena com apenas uma barra horizontal em meio a várias barras verticais, o foco de atenção é imediatamente atraído para a barra horizontal. Há uma grande quantidade de estudos realizados sobre os fatores *bottom-up* e estes são mais fáceis de modelar computacionalmente do que os fatores *top-down* (já que dependem apenas de características intrínsecas à cena observada). Entretanto, como apontam Henderson e Hollingworth [24], a maioria das fixações são orientadas pelas tarefas. Assim, o modelamento *bottom-up* pode explicar apenas uma fração pequena dos movimentos oculares no contexto da atenção visual.

Seguindo o que foi aprendido com a Teoria de Integração de Características [16], 3 características (*features*) visuais têm sido tradicionalmente usadas nos modelos computacionais *bottom-up*: intensidade, cor e orientação. A implementação desses canais é costumeiramente feita de maneira similar à proposta por Itti *et al.* [5]. O canal de intensidade é calculado como a média dos três canais de cor e passa por um processamento baseado em respostas neuronais. A característica de cor é formada por canais vermelho-verde e azul-amarelo, também baseados em comportamentos neuronais. A terceira destas características, orientação, é obtida com a convolução da imagem da cena com filtros de Gabor orientados. Na Figura 2.3 podemos ver um exemplo de mapa de saliência gerado utilizando estas 3 características visuais.

Pensando no modelamento da saliência para vídeos, Itti *et al.* foram os primeiros a utilizar um canal de caracterísa de movimento [25]. A implementação deste canal foi feita através da aplicação de máscaras direcionais aos quadros. Outros pesquisadores propuseram a utilização de outros canais de caraterística visual. Alguns exemplos neste contexto são: linhas horizontais, transformadas wavelet, viés pelo centro, *optical flow*, tremulação (*optical flow*), simetria e contraste de texturas [18]. As três primeiras características visuais citadas nesta seção (intensidade, cor e orientação) fazem parte de modelos que se inspiram em conceitos cognitivos, tendo base nos achados da Psicologia e da Neurociência. Segundo Borji *et al.* [18], a vantagem do uso de modelos de atenção cognitivos é que os mesmos ampliam nossa visão das estruturas biológicas envolvidas na atenção visual. Entretanto, modelos de atenção *bottom-up* também podem, por exemplo, ser baseados na Teoria da Informação, postulando que a saliência local serve para maximizar a informação amostrada na cena observada. Foi o que fizeram Bruce e Tsotsos [26] ao usar a auto-informação de

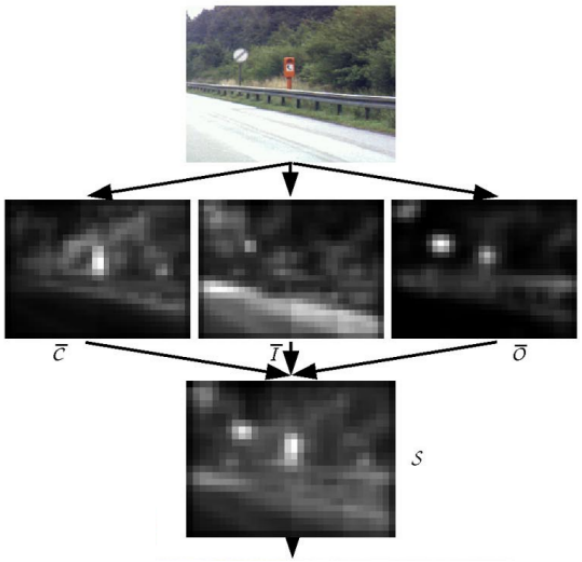


Figura 2.3: Exemplo de mapa de saliência gerado no trabalho de Itti *et al.* [5]. C representa as características de cor, I as de intensidade e O as de orientação. No mapa S, pixels com intensidade mais próxima do branco representam regiões mais salientes.

Shannon para calcular a saliência de regiões de uma imagem. Outras abordagens, como por grafos ou por análise espectral também podem ser encontradas na literatura [18]. Apesar disso, faltam ainda no campo de estudos do modelamento da atenção visual humana, maneiras consistentes de relacionar a interação entre a atenção *bottom-up* e a atenção *top-down*.

2.2 Modelamento Top-down

Em oposição ao caso *bottom-up*, Itti e Koch [27] observaram que os fatores *top-down* da atenção visual humana são lentos, orientados pelas tarefas e funcionam como sistemas de malha fechada. Como visto no trabalho de Yarbus [4] (Figura 2.2), esses fatores têm um peso grande na determinação de para onde os seres humanos desviam o olhar observando uma cena. Os modelos computacionais em geral procuram levar em conta informações *top-down* de 3 diferentes maneiras [18]: 1) Em tarefas de busca, é dada mais importância às características visuais do objeto procurado; 2) Alguns modelos investigam a influência do contexto ou essência (*gist*) da cena e 3) Quando a tarefa é mais complicada, como as interativas (dirigir um carro, por exemplo), um modelamento específico desta tarefa é feito.

Um fato relacionado ao o mecanismo de atenção *top-down* é de que não há consenso sobre se seres humanos fixam a atenção em objetos ou em regiões espaciais. De fato, a tendência atual é de que a unidade de atenção não é nenhum nem outro, admitindo a opção de que os mesmos não são mutuamente exclusivos. No entanto, o argumento de que seres humanos trabalham e pensam através de objetos [28], levou alguns modelos computacionais a utilizarem o modelamento de objetos como blocos fundamentais de formação da atenção *top-down*. Exemplo disso pode ser

visto no algoritmo bayesiano proposto por Borji *et al.* [19]. Uma última abordagem a ser citada é a utilização de técnicas de *Machine Learning* e reconhecimento de padrões. Os modelos com essa abordagem são treinados usando dados reais de fixação ocular. Eles aprendem os pesos a ser dados às características *bottom-up* e *top-down* de forma que as regiões salientes de um vídeo ou imagem coincidam com as regiões fixadas pelos voluntários nos experimentos com *eye-tracker*. Peters e Itti [29] treinaram um classificador para capturar a associação entre uma cena e as localizações preferenciais de fixação ocular enquanto voluntários jogavam video-game. Já Judd *et al.* [6], treinaram um modelo contendo características de baixo, médio e alto níveis.

Claramente podemos perceber a diferença de complexidade entre os modelamentos *bottom-up* e *top-down*. No caso deste último, há que se levar em conta a realimentação de informações e frequentemente são usados mecanismos de treinamento adaptados para tarefas específicas. A necessidade de adaptar o modelo a cada tarefa torna a abordagem *top-down* menos generalizável que a *bottom-up*. Entretanto, a importância dos fatores *top-down* na definição do mecanismo de atenção humana torna a adição dessas informações necessárias. Além disso, Borji e Itti [18] sugerem que o desenvolvimento de modelos que levem em conta demandas de tarefas variantes no tempo é uma direção promissora para a pesquisa futura em atenção visual.

2.3 Modelo de Saliência de Judd

O modelo de saliência mais importante para este trabalho é o descrito por Judd *et al.* [6], um modelo para imagens. Fazendo uso de vários modelos de saliência bem estabelecidos na literatura, a inspiração para a criação do algoritmo foi a de adicionar informações *top-down* a características de médio e baixo níveis (*bottom-up*) de maneira ótima. A combinação dos 33 canais do modelo (cor, orientação, detecção de objetos, entre outros) foi feita com técnicas de *Machine Learning*, em um treinamento com um extenso banco de dados de 1003 imagens observadas livremente por 15 voluntários. Borji e Itti classificam este modelo como um modelo de "Reconhecimento de Padrões" [18]. Apesar do alto desempenho, modelos deste tipo, escrevem Borji e Itti, tornam-se dependentes dos dados, lentos e, de certa maneira, "caixas-pretas" (no sentido em que não se aprende como cada parte do modelo contribui para a saliência final). O modelo de Judd possui um dos melhores desempenhos de estimativa de saliência entre os modelos presentes na literatura [13], [20].

Como já indicado acima, o modelo de saliência de Judd conta com 33 canais característicos, que podem ser divididos em três grupos, segundo o nível da característica representada: Baixo-nível, Médio-nível e Alto-nível. As características de Baixo-nível, que também podem ser vistas como características *bottom-up*, compõem a maior parte do grupo de canais, sendo 28 ao total. Os primeiros 13 destes canais correspondem à energia local de filtros em pirâmides direcionáveis (*steerable pyramids*) [30] (ver Figura 2.4, imagens 1 a 13). O canal 14 corresponde a um modelo simples de saliência descrito por Torralba [31], baseado em *subband pyramids* (Figura 2.4, canto inferior esquerdo). Do trabalho de Itti e Koch [27], mais especificamente da versão descrita por Walther [32], as informações de intensidade, orientação e contraste de cor formam os canais 15–17 (Figura 2.4, imagens 14 a 16). Os canais 18–23 são formados pelos canais vermelho, verde e azul da

imagem, bem como as probabilidades de cada um destes canais (Figura 2.4, imagens 20 a 25). Por fim, os canais 24–28 correspondem à probabilidade de cada cor, calculada através do histograma de cores 3D, da imagem filtrada por um filtro de mediana a 5 escalas diferentes (Figura 2.4, imagens 26 a 30).

Há apenas uma característica Médio-nível, representada no modelo pela detecção do horizonte da cena [33] (ver Figura 2.4, imagem 19). Esta característica se apoia no fato de que a maioria dos objetos em uma cena se encontram pousados sobre a superfície da Terra. Logo, o horizonte é um lugar natural para serem humanos olharem à procura de objetos salientes.

Na criação da base de dados de 1003 imagens e posterior experimentação rastreando o olhar de voluntários, Judd *et al.* [6] perceberam que as pessoas fixavam o olhar, muito consistentemente, em pessoas, faces e alguns outros objetos, particularmente carros. De fato, 10% das fixações observadas recaem sobre faces [20]. As características de Alto-nível do modelo são, então, representadas por um detector de faces Viola Jones [8] e detectores de pessoas e carros, como descritos por Felzenszwalb [34] (Figura 2.4, imagens 18, 32 e 33, respectivamente). Essas características são *top-down* no sentido que funcionam como guias de atenção na tarefa proposta aos voluntários durante o experimento (livre observação, com um teste de memória ao final).

Há ainda uma outra característica presente no modelo que não foi incluída por Judd *et al.* nos 3 grupos apresentados anteriormente. Analisando as fixações dos voluntários sobre as imagens da base de dados, os pesquisadores perceberam que há uma forte tendência das pessoas de fixarem o olhar em regiões próximas ao centro de uma imagem. De fato, de todas as fixações registradas, 40% se encontram entre os 11% pixels mais centrais da imagem e 70% se encontram entre os 25% pixels mais centrais. Devido a este viés, foi adicionado ao modelo de saliência um canal formado pela distância de cada pixel ao centro da imagem (Figura 2.4, imagem 17), completando os 33 canais presentes no modelamento.

A combinação dos 33 canais foi treinada, usando a técnica de *Support Vector Machines* (SVM). Após o treinamento, um conjunto de 33 ganhos, ou pesos, é retornado para ser utilizado na soma ponderada dos mapas de cada canal. Além disso, os canais são normalizados para terem média zero e variância unitária. O diagrama de blocos do modelo completo é apresentado na Figura 2.5. Um exemplo do mapa de saliência final, após o treinamento, pode ser visto na Figura 2.6.

Nos testes que seguiram o treinamento, se confirmou a importância da distância ao centro da imagem. Por outro lado, foi mostrado também que o modelo completo tem desempenho melhor que apenas o canal de distância ao centro sozinho. Na Figura 2.7, retirada do trabalho de Judd *et al.* [6], são dispostas curvas ROC (*Receiver Operating Characteristic*) para diversas partes do modelo com relação a testes realizados sobre a base de dados criada. Como explicado no Anexo I, quanto maior a área abaixo de uma curva ROC, melhor é o desempenho do modelo utilizado na estimativa da saliência. Isto posto, vemos na Figura 2.7 que a distância ao centro (*Center*, a curva ciano) tem melhor desempenho que todos os outros canais combinados, mas ainda não é melhor que o modelo completo. Ainda assim, cada canal apresentado estima a saliência de uma imagem significativamente melhor do que mapas gerados aleatoriamente (*Chance*, no gráfico).

Em uma última análise, Judd *et al.* [6] mediram quais canais agregam mais ao desempenho

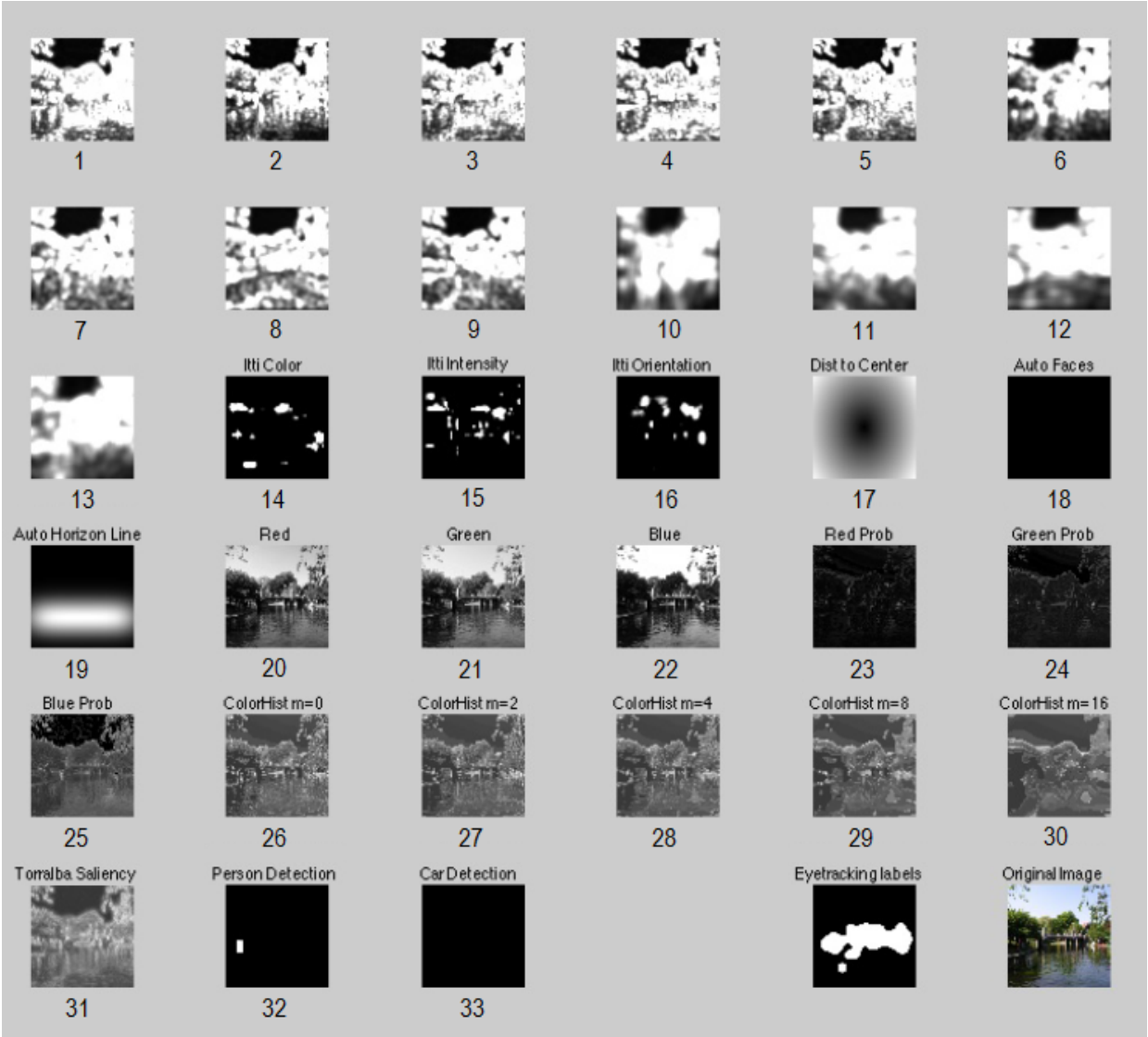


Figura 2.4: Os 33 canais característicos presentes no modelo de Judd *et al.* [6]. A imagem "*EyeTrackingLabels*" é uma segmentação das regiões salientes na imagem original de acordo com fixações reais capturadas por *eye-tracker*. A última figura no canto inferior direito é a imagem de referência.

final quando combinados com o canal de distância ao centro. Essa medida foi feita comparando-se o desempenho do canal de distância ao centro sozinho com o desempenho deste combinado com cada um dos outros canais e observando-se qual combinação resulta na maior diferença. Neste contexto, podemos classificar os diversos canais do modelo segundo a melhora trazida da seguinte maneira:

$$Modelo\ de\ Torralba > Canais\ de\ cor > Horizonte > Detectores\ de\ objeto > Modelo\ de\ Itti \tag{2.1}$$

Na conclusão de um dos capítulos de sua tese de doutorado [20], Judd ainda comenta que a forte tendência da atenção visual humana em direção ao centro das imagens é devida tanto ao viés fotográfico (tendência humana de posicionar objetos mais interessantes no centro de fotografias)

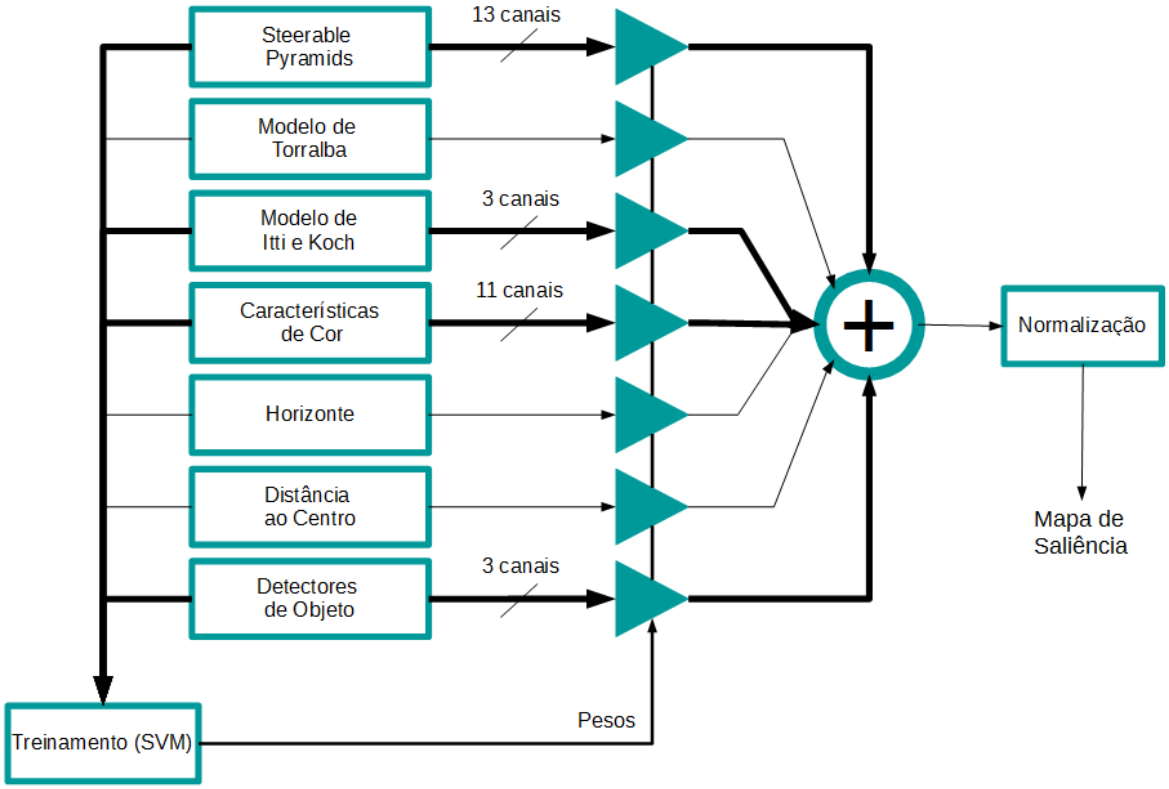


Figura 2.5: Diagrama de blocos do modelo de Judd *et al.* [6].

quanto pelo fato da estratégia de observação das pessoas estar enviesada para o centro. Judd sugere ainda que, num trabalho futuro, a adição de um detector de textos melhoraria ainda mais o desempenho do modelo, já que 11% das fixações das pessoas nos experimentos se localizavam sobre textos. O modelo explicado nesta seção é a base do modelo de atenção visual desenvolvido neste trabalho.

2.4 Precisão do Modelo de Atenção

Para afirmar que uma abordagem específica modela satisfatoriamente a atenção visual humana, é necessário que esteja bem definida a "verdade" (do inglês, *ground-truth*), que deve servir de base para todas as comparações de desempenho. No contexto de modelos de atenção que geram mapas de saliência, essa base são os chamados mapas de fixação, obtidos através de experimentos com seres humanos. Nesses experimentos, pessoas em um grupo de teste são convidadas, uma a uma, a observar imagens ou vídeos presentes em uma dada base de dados. A tarefa dada pelo pesquisador aos participantes do experimento pode variar (livre observação, procurar por algo específico, memorizar a cena, etc), mas, independentemente da tarefa, as fixações oculares e movimentos sacádicos dos olhos desses participantes são gravadas com o auxílio de um dispositivo *eye-tracker* (rastreador de olhos). O *eye-tracker* consiste em um suporte para a cabeça, que impede que esta se mova, havendo possibilidade apenas de movimento ocular, e câmeras apontadas para

(a) Um exemplo de imagem (b) Seu mapa de saliência

Figura 2.6: Exemplo de um mapa de saliência segundo [6]

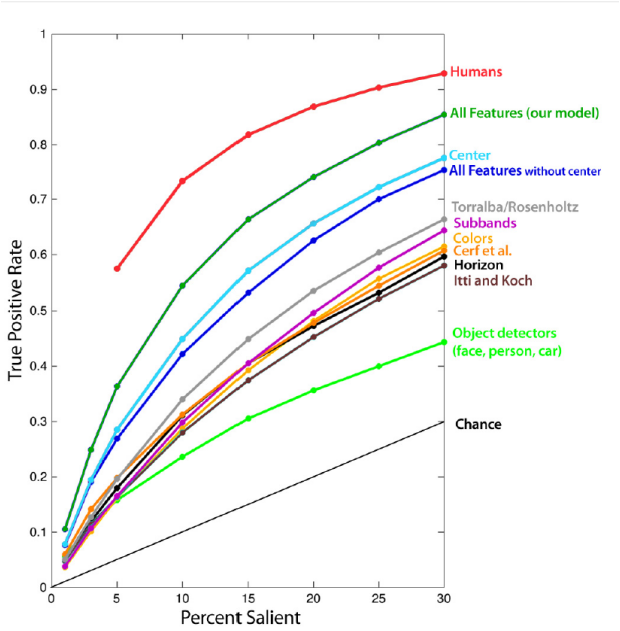


Figura 2.7: Resultados de desempenho medido pela curva ROC para elementos do modelo de Judd *et al.* [6]

os olhos do participante, as quais gravam a posição relativa dos olhos, que após alguns cálculos computacionais pode ser traduzida em um ponto na imagem observada representando uma amostra de fixação ocular. Várias dessas fixações são amostradas para cada imagem ou quadro de vídeo na base de dados, formando os já citados mapas de fixação para cada imagem ou quadro. É calculada a média destes mapas, com relação ao grupo de participantes do experimento, obtendo-se um conjunto de mapas de fixação. Pode-se, ainda, realizar uma operação de "foveação" dos mapas de fixação para que estes levem em conta o cone de observação da fóvea humana, que faz com que nós humanos focalizemos uma pequena região circular, e não apenas um ponto ou pixel, a cada instante de tempo. Esta operação, que acaba por transformar os mapas de fixação em mapas de saliência nos moldes já mostrados neste trabalho, em geral é feita através da convolução do mapa de fixações com um filtro gaussiano passa-baixas, cuja variância depende da abertura angular da

fóvea humana e da distância entre as imagens e os observadores durante o experimento.

Conjuntos como esses de mapas de saliência "verdadeiros" costumam ser parte integrante de bases de dados utilizadas nos testes de modelos de atenção visual [10], [11], [12]. O que resta então é encontrar uma maneira de comparar os mapas gerados pelo modelo de atenção de interesse e aqueles presentes na base de dados escolhida, de maneira que obtenha-se um número que indique quão grande é a similaridade entre eles. Quão maior for essa similaridade, mais acurado será então o modelo de atenção visual proposto, visto que os mapas na base de dados foram gerados a partir de observações reais de fixações oculares humanas. No Anexo I, apresentamos uma explicação mais detalhada de como pode ser feita essa comparação. Mas, para a compreenssão deste trabalho, basta que se saiba que foi usada uma métrica chamada AUC, a mais popular na comunidade de modelamento computacional da atenção visual humana. Esta métrica fornece índices compreendidos entre 0 e 1. O valor 1 indica que os mapas de saliência do modelo predizem perfeitamente a localização das fixações nos mapas de fixação dos experimentos com *eye-tracker*. O valor 0 é de certa forma complementar pois indica que o mapa de saliência do modelo prevê as regiões não-fixadas nos mapas de fixação (um mapa é o oposto do outro).

A título de exemplo, a Tabela 2.1 mostra o resultado da aplicação da métrica AUC na avaliação de alguns modelos, importantes na literatura [18], de saliência visual para imagens. Os resultados foram tirados de um *Benchmark* disponível em [13] que foi iniciado por Judd como parte de sua tese de doutorado [20]. O banco de dados de imagens usado na aplicação da métrica também foi criado no contexto dessa tese.

Tabela 2.1: Exemplos de desempenhos de modelos computacionais de saliência para imagens[13]

Modelo	BMS [35]	Judd [6]	Centro	AIM [26]	Torralba [31]	IttiKoch [32]	Aleatório
AUC	0,8257	0,8093	0,7830	0,7654	0,6837	0,5951	0,5030

Os resultados na Tabela 2.1 e no *Benchmark* em [13] foram organizados de forma decrescente e, no momento em que este trabalho está sendo escrito, o modelo de saliência para imagens com melhor desempenho na comunidade é o BMS [35], que está mostrado na tabela. O modelo de Judd [6], entretanto tem desempenho bem próximo ao BMS, sendo este mais um dos motivos para ter se escolhido este modelo como base do trabalho aqui desenvolvido. Pode-se perceber também que, devido ao "viés do fotógrafo", o modelo Centro, que corresponde a mapas cujos pixels possuem intensidade inversamente proporcional a sua distância ao centro geométrico, tem mais importância, segundo a AUC, do que modelos clássicos do meio, como o de Itti e Koch [32]. No final da escala de desempenho se encontra o modelo Aleatório, que é o limite inferior de desempenho acima do qual todos os modelos de atenção devem estar. O modelo Aleatório consiste simplesmente em se escolher aleatoriamente, em uma imagem, quais pixels são fixados ou não. Desta maneira, o valor de $AUC = 0,5$ é o pior valor que um modelo de saliência pode obter.

Capítulo 3

Qualidade de Vídeo

Antes que possamos fazer uma revisão dos trabalhos realizados no domínio da qualidade de vídeo e das técnicas de estimação normalmente utilizadas, é necessário que determinemos precisamente a que nos referimos quando falamos na qualidade de um vídeo. Por mais que, informalmente, o termo "qualidade" possa ser tomado como a qualidade artística das filmagens ou do conteúdo dos vídeos, neste contexto sua definição é restringida. A qualidade de um vídeo é definida aqui como o grau de ausência de artefatos ou defeitos que impeçam o olho humano de compreender a cena que é observada. Em princípio, a avaliação da qualidade poderia ser feita perguntando a opinião de voluntários em testes subjetivos. Contudo, este seria um processo lento e caro, imcompatível com o rápido crescimento da demanda por novos serviços. Este fato motiva o desenvolvimento de métricas objetivas para fazer a avaliação da qualidade. Porém, esta é uma tarefa difícil, devido à complexidade dos sistemas de vídeo e do modelamento da percepção visual humana. Além disso, a natureza audiovisual dos conteúdos assistidos nas mais diversas plataformas requer que medidas de qualidade sejam feitas tanto para o áudio quanto para o vídeo. Os principais esforços tem sido feitos nestas duas áreas individualmente, embora testes mostrem que a avaliação conjunta descreve melhor a percepção humana de qualidade [36]. Dito isso, este trabalho se concentrará apenas nas medidas de qualidade visual. Estas podem ser separadas em 3 grupos, de acordo com a quantidade de informação do vídeo de referência necessária para poder estimar a qualidade do vídeo de teste: *Full-reference*, *No-Reference* e *Reduced-Reference*.

3.1 Métricas *Full-Reference*

As métricas *Full-reference*, em português "Referência Completa", são também conhecidas pela sigla FR. Como o nome sugere, métricas deste tipo requerem que o vídeo, ou imagem de referência, esteja completamente disponível, geralmente sem defeitos ou compressão aplicada. Isso se deve ao fato das métricas FR estimarem a qualidade do vídeo de teste através de uma comparação quadro a quadro com o vídeo original (ou de referência). Considerando a aplicabilidade prática em sistemas de vídeo, a necessidade do vídeo original se torna uma restrição pesada [1]. De forma geral, métricas FR demandam também alinhamentos espaciais e temporais precisos. Além disso,

não costumam responder bem a desvios de brilho, contraste ou cor entre o vídeo de teste e o de referência, necessitando geralmente de uma fase de calibração antes da aplicação da métrica. O uso desse tipo de métrica é em geral mais adequado a medidas *offline* de qualidade de vídeo. Esse é o caso, por exemplo, de ajustes de codecs e testes de laboratório, nos quais a precisão na estimativa é mais importante que ter resultados em tempo real [1]. Métricas FR baseadas no sistema visual humano possuem os melhores desempenhos de estimação de qualidade [36]. Normalmente, em uma métrica deste tipo são realizados os seguintes processamentos: processamento de cor, decomposição em múltiplos canais, contraste percebido, mascaramento espacial e temporal, entre outros.

Todas as métricas de qualidade utilizadas neste trabalho são classificadas como *Full-Reference*, embora no futuro os princípios da solução proposta possam ser aplicados a métricas *Reduced-Reference* e *No-Reference*. Algumas delas são métricas de dados (MSE, PSNR) enquanto outras são baseadas no sistema visual humano (SSIM, MS-SSIM e VQM). Além disso, das métricas utilizadas, quatro (MSE, PSNR, SSIM e MS-SSIM) foram concebidas em sua origem como métricas de qualidade de imagens e apenas uma (VQM) é adaptada, desde sua concepção, para vídeos. Todas essas métricas são detalhadas a seguir.

3.1.1 *Mean Squared Error* (MSE)

A sigla MSE significa Média dos Erros Quadráticos em português e é uma métrica que originalmente mede a fidelidade entre sinais através da avaliação do erro, ou distorção entre eles. Na aplicação da métrica, em geral se supõe que um sinal é puro, original e o outro é uma versão distorcida deste. Neste sentido, a MSE pode ser vista como uma medida de qualidade do sinal distorcido. Embora a métrica possa ser definida para um sinal com um número arbitrário de dimensões, a versão de interesse aqui é a bidimensional, já que a MSE será usada para avaliar quadros de vídeos. Considerando x o sinal (imagem) de referência e y o sinal distorcido, ambos com tamanho $M \times N$, o índice MSE é dado pela seguinte equação:

$$MSE = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (x(i,j) - y(i,j))^2. \tag{3.1}$$

O sinal de erro relacionado a esta métrica é comumente definido como $e(i,j) = x(i,j) - y(i,j)$. No entanto, neste trabalho, quando mencionamos um sinal (ou mapa) de erros no contexto da métrica MSE, nos referimos ao mapa de erros quadráticos $e(i,j) = (x(i,j) - y(i,j))^2$. Outra maneira de interpretar esta métrica é como sendo proporcional ao quadrado da distância Euclideana entre dois vetores de dimensão $(M \cdot N)$.

Juntamente com a PSNR, a MSE é uma das métricas mais utilizadas no campo do processamento de imagens e vídeos [1]. Algumas razões para tanto, além da tradição, são o fato desta ser uma métrica simples, com claro significado físico e ter propriedades estatísticas interessantes [7]. Por exemplo, a MSE é aditiva para fontes de distorção independentes e, fisicamente, é a forma natural de se definir a energia do sinal de erro. Apesar destas vantagens a MSE pode falhar quando usada para estimar a percepção humana de qualidade de imagem, como mostrado na Figura 3.1. Nesta figura, algumas imagens com distorções dramaticamente diferentes recebem os mesmos va-

lores de MSE. Pode-se observar que várias imagens com aproximadamente o mesmo valor de MSE (Figuras 3.1b a 3.1g) não possuem a mesma qualidade (de acordo com a percepção humana).

A origem dos erros de estimação observados na Figura 3.1 se encontra em algumas suposições feitas quando se usa a métrica MSE. Em primeiro lugar, supõe-se que a fidelidade das imagens é independente de relações espaciais ou temporais entre as amostras da imagem original. Em segundo lugar, supõe-se que essa fidelidade também é independente da relação entre a imagem original e o mapa de erro. Além disso, não é levada em consideração a influência dos sinais (positivo ou negativo) no mapa de erro. Por último, fica implicito no cálculo da MSE que todas as amostras dos sinais são igualmente importantes na avaliação da fidelidade. A alta estruturação espacial dos sinais de imagens naturais impede que a aplicação direta da MSE produza bons resultados na estimação da qualidade de imagens. Wang e Bovik fazem uma avaliação extensa das vantagens e desvantagens da MSE nesse contexto [7].

3.1.2 PSNR

A PSNR, ou *Peak Signal-to-Noise Ratio* (Razão Sinal-Ruído de Pico, em português) pode ser definida, a partir da MSE, da seguinte maneira:

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX_x^2}{MSE}\right) \tag{3.2}$$

na qual MAX_x é o maior valor de intensidade de pixel possível na imagem de referência x. Por exemplo, em representações de 8 bits por pixel, $MAX_x = 255$. A PSNR é uma conversão da MSE para uma escala logarítmica, sendo útil se as imagens comparadas têm diferentes faixas dinâmicas [7]. Embora seja bastante usada no campo de processamento de imagens, em princípio a PSNR não adiciona nenhuma informação nova em relação à MSE. Um problema técnico na utilização da PSNR é que ela não é definida caso os sinais x e y sejam iguais.

3.1.3 SSIM

Proposta por Wang *et al.* [14], a métrica SSIM (*Structural-Similarity based Image quality Metric*) procura levar em consideração a alta dependência espacial dos pixels de uma imagem na estimação de qualidade da mesma. Sua construção se baseia no princípio de que seres humanos são adaptados a extrair informações estruturais das cenas que observam. A métrica é baseada em 3 medidas comparativas: luminância, contraste e estrutura. Considerando novamente x a imagem de referência e y a imagem de teste, a medida de luminância é dada por:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{{\mu_x}^2 + {\mu_y}^2 + C_1}, \tag{3.3}$$

na qual μ_x representa a intensidade média dos pixels em uma janela 8×8 da imagem x e C_1 é uma constante adicionada para evitar instabilidade quando ${\mu_x}^2 + {\mu_y}^2$ se aproxima de zero. Wang *et al.* [14] definem esta constante como $C_1 = (K_1 L)^2$. L é o valor máximo da faixa dinâmica dos pixels de x e K_1 é uma constante muito menor que 1.

A medida de contraste entre x e y é dada pela seguinte equação:

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{{\sigma_x}^2 + {\sigma_y}^2 + C_2}, \tag{3.4}$$

na qual σ_x é o desvio padrão das intensidades dos pixels de uma janela 8×8 em x, o que é usado aqui como uma medida de contraste. Analogamente ao caso anterior, $C_2 = (K_2L)^2$, com $K_2 \ll 1$.

A terceira medida comparativa, a estrutural, é definida como a correlação entre o sinais x e y depois que estes são normalizados para terem média zero e variância unitária. Sua fórmula é dada por:

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{3.5}$$

na qual σ_{xy} é a covariância, dentro de uma janela 8×8, entre os sinais x e y e C_3 é uma constante adicionada pelos mesmos motivos explicados para as medidas de luminância e contraste.

Após as medidas comparativas, a combinação das informações colhidas para uma dada janela 8×8 nas imagens é feita da seguinte maneira:

$$SSIM_{window}(x,y) = [l(x,y)]^\alpha \cdot [c(x,y)]^\beta \cdot [s(x,y)]^\gamma, \tag{3.6}$$

na qual $\alpha > 0$, $\beta > 0$ e $\gamma > 0$ são parâmetros usados para ajustar a importância relativa das 3 componentes. Normalmente, se faz as simplificações $\alpha = \beta = \gamma = 1$ e $C_3 = C_2$ [14] e a combinação passa a ser escrita da seguinte maneira:

$$SSIM_{window}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x\sigma_y + C_2)}{({\mu_x}^2 + {\mu_y}^2 + C_1)({\sigma_x}^2 + {\sigma_y}^2 + C_2)}. \tag{3.7}$$

As estatísticas locais são calculadas repetidamente para janelas 8×8, transladando estas um pixel de cada vez até que seja coberta toda a extensão das imagens. Isso resulta em um mapa, $SSIM_{map}$ com o tamanho de x, contendo um índice SSIM relativo a cada janela. A agregação deste mapa em um índice único, representando a qualidade de y em relação a x é feita através de uma simples média:

$$SSIM = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} SSIM_{map}(i,j), \tag{3.8}$$

na qual $SSIM_{map}$ é uma imagem de dimensões $M \times N$. Neste trabalho, quando mencionarmos, nos capítulos seguintes, o índice SSIM, estaremos nos referindo ao valor obtido através da Equação 3.8. A métrica SSIM é simétrica ($SSIM(x,y) = SSIM(y,x)$) e tem valores limitados ao intervalo $[-1,1]$. Observamos na Figura 3.1 que esta métrica é um claro progresso em relação à MSE. Além disso, sua definição genérica sugere que ela pode ter extensa aplicabilidade, mesmo em campos outros que não o processamento de imagens [7]. Como desvantagem, esta métrica possui certa sensibilidade a translações relativas, escalonamentos e rotações. Entretanto, uma versão com transformadas *wavelet*, chamada CW-SSIM foi proposta para lidar com estes tipos de situações [37].

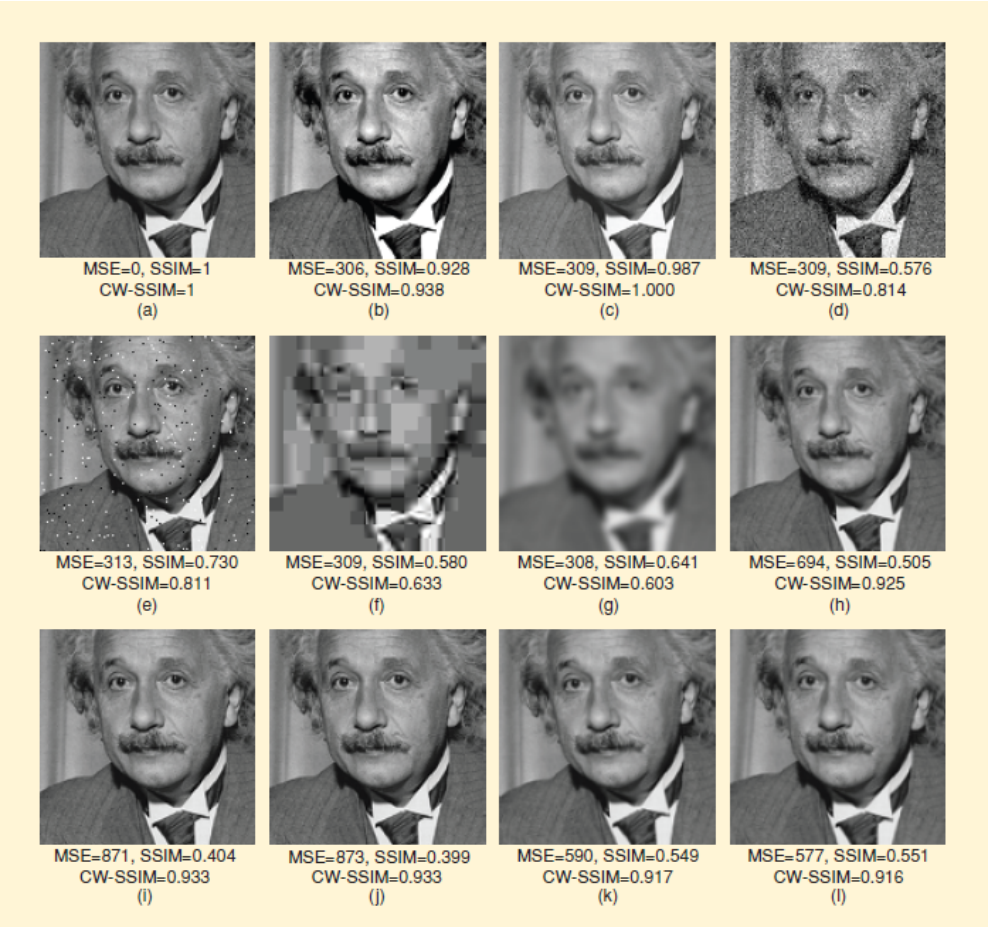


Figura 3.1: Exemplo de distorções que não são propriamente captadas pela métrica MSE mas são captadas pela métrica SSIM [7] . Temos: (a) Imagem de referência, (b) aumento de contraste, (c) mudança de luminância, (d) ruído gaussiano, (e) ruído impulsional, (f) compressão JPEG, (g) borramento, (h) *zoom out*, (i) translação à direita, (j) translação à esquerda, (k) rotação anti-horária e (l) rotação horária.

3.1.4 MS-SSIM

A métrica MS-SSIM é a versão em múltiplas escalas da SSIM. Também foi proposta por Wang *et al.*, como uma alternativa mais flexível na incorporação de variações nas condições de visualização [38]. No algoritmo da métrica, são aplicadas, iterativamente, filtragens passa-baixas e subamostragens aos sinais de entrada x e y. Essas subamostragens reduzem a escala por um fator de 2 a cada iteração. À escala final neste processo se dá o nome S. As medidas de contraste e estrutura são feitas a cada escala k (entre S e a escala original) e a medida de luminância é feita apenas na escala S. A combinação das medidas comparativas, para uma dada janela de análise, percorrendo todas as S escalas é definida pela seguinte equação:

$$MS\text{-}SSIM_{window}(x,y) = [l(x,y)]_S^{\alpha} \prod_{k=1}^{S} [c_j(x,y)]_k^{\beta_j} \cdot [s_j(x,y)]_k^{\gamma_j}. \tag{3.9}$$

Assim como no caso da métrica SSIM, a varredura das janelas de análise produz um mapa de valores MS-SSIM do tamanho da imagem x a cada escala. A agregação deste mapa em um índice único é feita através do cálculo da média do mesmo:

$$MS\text{-}SSIM = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} MS\text{-}SSIM_{map}(i,j). \tag{3.10}$$

na qual $M \times N$ são as dimensões da imagem de referência x. Observe que a métrica SSIM pode ser vista como um caso especial da MS-SSIM com $S = 1$. Como padrão, Wang *et al.* [38] definiram $S = 5$, $\beta_1 = \gamma_1 = 0,0448$, $\beta_2 = \gamma_2 = 0,2856$, $\beta_3 = \gamma_3 = 0,3001$, $\beta_4 = \gamma_4 = 0,2363$, $\alpha = \beta_5 = \gamma_5 = 0,1333$. Os pesquisadores perceberam que, escolhendo este conjunto de parâmetros, a MS-SSIM apresenta consistentemente melhor desempenho que o estado-da-arte das métricas de qualidade de imagens à época, incluindo a própria SSIM.

3.1.5 VQM

VQM (*Video Quality Model*) é o nome dado ao modelo geral de estimação de qualidade de vídeo proposto por Pinson e Wolf, da *National Telecommunications and Information Administration* (NTIA) [15]. Na época em que foi proposto, era o único estimador de qualidade de vídeo de alto desempenho em ambos os testes com vídeos de 525 e 625 linhas. Consequentemente, foi adotado, em 2003, como Padrão Norte-Americano pelo *American National Standards Institute* (ANSI). A construção do algoritmo da métrica VQM se dá, de acordo com Winkler e Mohandas [1], a partir de uma abordagem de engenharia. Este tipo de abordagem se baseia, em primeira instância, na extração e análise de características (*features*) ou artefatos do vídeo. Esta abordagem é, de fato, a mesma utilizada pela métrica SSIM e tem ganhado popularidade nos últimos anos. Em linhas gerais, a VQM consiste em dividir a sequência de vídeo em blocos espaço-temporais e medir as características de quantidade e orientação de atividade em cada bloco. As características extraídas dos vídeos de referência e de teste são, por fim, comparadas e a estimação de qualidade é retornada como um índice numérico.

O algoritmo da métrica VQM se divide em 4 etapas. A primeira etapa do algoritmo, a **calibração**, consiste em determinar e corrigir deslocamentos espaciais e temporais e desvios (*offsets*) de contraste e brilho entre os vídeos original e de teste. Em seguida, é feita a **extração de características** (*features*) de qualidade dos blocos espaço-temporais em cada vídeo. Estas características medem mudanças perceptuais nas propriedades espaciais, temporais e de crominância das sequencias de vídeos. Em um terceiro passo, as características extraídas são comparadas entre os dois vídeos e resultam na criação de 7 **parâmetros de qualidade**. A comparação das características de cada vídeo é feita usando a distância Euclideana, a razão entre as características ou o logaritmo da razão, dependendo da característica analisada. Dos 7 parâmetros de qualidade obtidos na comparação, 4 são baseados em características extraídas de gradientes espaciais da componente de luminância dos vídeos. Outros 2 são baseadas em características extraídas de um vetor formado pelas componentes de crominância e o último se origina de medidas de contraste e movimento, ambas feitas sobre o canal de lminância dos vídeos. Os 7 parâmetros são descritos a seguir.

O parâmetro SI_{loss} detecta perda de informação espacial (por borramento, por exemplo). O parâmetro HV_{loss} detecta a transição de bordas orientadas horizontal e verticalmente para bordas com orientação diagonal. Esta transição será notada se bordas as horizontais e verticais sofrerem maior borramento que as diagonais no vídeo de teste. O parâmetro HV_{gain} detecta a transição no sentido contrário do parâmetro HV_{loss}. O aparecimento de mais bordas horizontais e verticais é típico de vídeos com artefatos de blocagem. O parâmetro $Chroma_{spread}$ detecta alterações no espalhamento da distribuição das amostras de cores nos vídeos. O parâmetro SI_{gain} mede ganhos de qualidade devido a maior nitidez de bordas ou outras melhorias no vídeo de teste. Este é o único dos 7 parâmetros que quantifica melhoria de qualidade em relação ao vídeo original. O parâmetro $CT_{ATIgain}$ mede a quantidade de detalhamento espacial e de movimento presentes nos blocos espaço-temporais dos vídeos. Mudanças nestas características são notadas com a presença de ruído ou blocos de erro no vídeo de teste. Durante o cálculo deste parâmetro um filtro de detecção de movimento chamado "Valor Absoluto da Informação Temporal" (ATI, na sigla em inglês) é aplicado aos canais de luminância dos vídeos comparados. A importância desse filtro para este trabalho será evidenciada no Capítulo 4. O último dos 7 parâmetros, $Chroma_{extreme}$ detecta defeitos locais de cor, como aqueles produzidos por erros de transmissão digital.

Finalmente, o último passo do algoritmo da métrica VQM, a **estimativa de qualidade**, consiste em conferir um índice de qualidade único ao vídeo de teste. Isto é feito através de uma simples combinação linear dos 7 parâmetros de qualidade descritos no parágrafo anterior. A combinação é dada pela seguinte equação:

$$VQM = -0,2097 \cdot SI_{loss} + 0,5969 \cdot HV_{loss} + 0,2483 \cdot HV_{gain} + 0,0192 \cdot Chroma_{spread} - 2,3416 \cdot SI_{gain} + 0,0431 \cdot CT_{ATIgain} + 0,0076 \cdot Chroma_{extreme} \quad (3.11)$$

Os valores retornados pela métrica VQM variam de 0 (defeitos imperceptíveis entre os vídeos) a aproximadamente 1 (defeitos extremamente perceptíveis). A métrica foi completamente implementada em software por Pinson e Wolf [39] e é livremente disponível.

3.2 Métricas *No-Reference*

Métricas *No-Reference* (NR), em oposição às métricas *Full-Reference* não necessitam de nenhuma informação do vídeo de referência para avaliar a qualidade do vídeo de teste. Por este motivo, também são conhecidas como métricas "cegas" [40]. Esse tipo de avaliação de qualidade é feito por nós seres-humanos a todo momento, sem dificuldades, entretanto, ela é uma tarefa difícil para um algoritmo objetivo. Embora possuam, em geral, desempenhos piores que o de métricas FR, as métricas NR têm a vantagem de ser bastante convenientes em sistemas de telecomunicações, por exemplo, já que requerem menos banda para efetuar suas medidas. Apesar disso, os maiores desenvolvimentos na direção de métricas de qualidade universal são origem em abordagens *Full-Reference*. Por outro lado, métricas *No-Reference* não sofrem do problema de alinhamento espaço-temporal das métricas *Full-Reference*. Hoje em dia, a maioria das métricas NR se concentram em medir artefatos ou características de distorção, como blocagem e borramento [36]. Após essas medidas, algum método de agregação (*pooling*) é usado para retornar um índice de qualidade.

Este índice porém, representa uma estimação de qualidade absoluta, o que dificulta, por exemplo, a quantificação perda de qualidade sofrida no processo de transmissão de um vídeo [40]. Isto porque, por não ter acesso ao vídeo original antes do processo, as métricas NR devem recorrer a outros meios para descobrir qual era o estado de qualidade anterior ao do vídeo-teste avaliado.

3.3 Métricas *Reduced-Reference*

O terceiro grupo, o das métricas *Reduced-Reference* (RR), representa um compromisso entre as restrições das métricas FR e a dificuldade de modelamento das métricas NR. Métricas RR usam informações do vídeo de referência para avaliar o vídeo de teste, mas de maneira reduzida. Em geral, são extraídas características (*features*) do vídeo original como a quantidade de movimento ou o detalhamento espacial [1] e são essas características que são transmitidas para a avaliação de qualidade. As métricas *Reduced-Reference* sofrem do problema de alinhamento, mas em geral de forma mais leve do que as métricas FR, já que apenas as características extraídas devem passar pelo processo de calibragem. Adicionalmente, Para uso em sistemas que requerem medidas em tempo real, as métricas RR são mais convenientes que as FR. Por outro lado, as métricas NR são mais flexíveis, pois as métricas RR necessitam de acesso a um vídeo de referência em pelo menos uma etapa da sua cadeia de implementação. Um exemplo de estudo e implementação de uma métrica de referência reduzida pode ser vista na tese de doutorado de Engelke [40].

3.4 Integrando Atenção Visual às Métricas de Qualidade

A consideração da atenção visual nos cálculos das métricas objetivas de vídeo é apontada por Winkler e Mohandas [1] como uma tendência para os desenvolvimentos nesta área. O interesse por esse tipo de integração tem base na ideia de que, fora das regiões salientes de um vídeo, a sensibilidade humana às degradações presentes diminui. A maioria das métricas de qualidade (incluindo as apresentadas anteriormente neste capítulo) não leva esse fato em consideração e dá a mesma importância a todas as regiões do vídeo na avaliação das degradações. Há esforços no sentido de incorporar informações da fóvea (*foveation*) às métricas de qualidade [41], mas estes modelos em geral assumem que as fixações se encontram no centro dos quadros dos vídeos. De fato, na literatura, em geral se leva em conta a atenção visual de duas maneiras diferentes. A primeira considera apenas alguns aspectos da atenção visual mas de forma direta na cadeia de cálculo da métrica de qualidade. A segunda, trata a integração de forma paralela, gerando mapas de qualidade através de uma dada métrica objetiva e ponderando-os através de mapas de saliência. You *et al.* [42] comentam que esta última abordagem ignora o fato dos mecanismos de atenção e de percepção de qualidade estarem intrinsicamente conectados nos seres humanos. De fato, You *et al.* consideram essa conexão ao propor a métrica de Qualidade de Vídeo Foveada Guiada pela Atenção (AFViQ) e conseguem desempenhos excelentes nas estimativas de qualidade. Na construção da métrica AFViQ, a atenção visual não é vista de forma externa, mas sim como parte integrante da cadeia do algoritmo. No entanto, a exemplo dos trabalhos de Redi *et al.* [43] e de Liu e Heynderickx [44], a abordagem da ponderação pode também oferecer boas melhorias na estimação das métricas

de qualidade. É esta última abordagem que é adotada neste trabalho.

Liu e Heynderickx [44] definiram a métrica ponderada $WMET$ através da seguinte equação:

$$WMET = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} SM(i,j) \cdot Map_{MET}(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} SM(i,j)} \tag{3.12}$$

na qual x é a imagem original, SM seu mapa de saliência , y a imagem de teste, MET uma métrica de qualidade e Map_{MET} é o mapa de qualidade, ou de erro, dado pela métrica MET para a comparação entre as imagens x, e y. Redi *et al.* [43] também trabalharam com a multiplicação ponto-a-ponto dos mapas de atenção e os mapas de qualidade, mas sua estratégia de agregação (*pooling*) foi diferente da média ponderada. Eles utilizaram uma abordagem estatística e uma rede neural para transformar o mapa ponderado em um índice de qualidade. Os pesquisadores também experimentaram com a utilização de funções do mapa de saliência para realizar a ponderação espacial e observaram melhoras de desempenho em relação ao caso em que o mapa de saliência é usado diretamente.

Tanto Redi *et al.* quanto Liu e Heynderickx trabalharam com métricas de qualidade de imagem e uma adaptação deve ser feita para que possamos utilizar os príncipios de seus algoritmos na análise de vídeos. Essas adaptações serão discutidas nos Capítulos 4 e 5.

3.5 Desempenho de uma Métrica de Qualidade

Enquanto métricas de fidelidade medem apenas quão fiel um sinal de teste é a um sinal de referência, métricas de qualidade se propõem a estimar a percepção humana da qualidade de um sinal. Sendo assim, contuma-se medir o desempenho de uma métrica de qualidade comparando os resultados desta com a avaliação subjetiva da qualidade de vídeos feita por pessoas. A hipótese neste caso é que a opinião dos seres humanos quanto a um vídeo reflete sua percepção da qualidade do mesmo.

Nos experimentos subjetivos, é pedido a um grupo de voluntários que assista a uma sequência de vídeos e dê uma nota de qualidade para cada um. Esta nota geralmente está restrita ao intervalo de 1 a 5 ou de 1 a 9. Para cada vídeo, é computada a média das notas com relação a todos os voluntários, valor conhecido como Pontuação Média das Opiniões (MOS, ver Tabela 3.1). Ao final do experimento, a qualidade dos vídeos na base de dados é representada por um vetor de valores MOS. Um exemplo de aquisição desse tipo de resultado pode ser visto no trabalho de Le Meur at al. [10]. Devido a diferentes interesses e expectativas dos indivíduos, a medida da opinião tende a ser ruidosa. Entretanto, escolhendo-se um número suficiente grande de voluntários (15 a 30, em geral) a medida tende a se estabilizar. Em alguns casos, quando a métrica objetiva que se pretende validar assume que os vídeos de referência tem máxima qualidade, é interessante se calcular o valor MOS diferencial, ou DMOS [11]. O valor DMOS nada mais é que o valor MOS do vídeo de teste subtraído do valor MOS do vídeo de referência correspondente. Assim, quando os valores MOS são medidos em uma escala de 1 a 5, os valores DMOS se encontrarão em uma escala de 0 a 4 (ver Tabela 3.2). Alternativamente, como feito por Seshadrinathan *et al.* [11], os valores DMOS

podem ainda ser normalizados para apresentar média zero e variância unitária, além de passarem também por um processo de remoção de *outliers*.

Tabela 3.1: Valores de *Mean Opinion Scores* (MOS) usando uma escala de 1 a 5.

MOS	Qualidade	Defeitos
5	Excelente	Imperceptíveis
4	Boa	Perceptíveis mas não incômodos
3	Razoável	Levemente incômodos
2	Baixa	Incômodos
1	Inaceitável	Muito incômodos

Tabela 3.2: Valores de *Differential Mean Opinion Scores* (DMOS) usando uma escala de 0 a 4.

DMOS	Significado
3,1-4,0	Maioria dos voluntários insatisfeita
2,1-3,0	Muitos voluntários insatisfeitos
1,1-2,0	Alguns voluntários satisfeitos
0,7-1,0	Muitos voluntários satisfeitos
0,0-0,6	Maioria dos voluntários satisfeita

Os testes subjetivos são feitos em geral seguindo padrões especificados pela União International de Telecomunicações (ITU). A recomendação BT.500-11 [45] da ITU, por exemplo, especifica dois métodos: avaliação com escala contínua de qualidade e estímulo único (SSCQE) e a avaliação com escala contínua de qualidade e estímulo duplo (DSCQS). No método SSCQE, os observadores avaliam a qualidade dos vídeos distorcidos sem ter como referência o vídeo original. Já no método DSCQS, os observadores avaliam tanto os vídeos distorcidos quanto os originais. A aderência da comunidade científica a essas recomendações contribui para que os resulados dos experimentos subjetivos sejam largamente aceitos como medidas da percepção humana de qualidade [40]. No entanto, as mesmas recomendações exigem um projeto detalhado e cuidadoso dos experimentos subjetivos, o que torna estes processos caros e demorados. Ainda assim, estes experimentos são extremamente importantes para validar métricas objetivas mais rápidas, de maneira a permitir o seu uso em aplicações reais.

A validação de uma métrica objetiva de qualidade ocorre, geralmente, em duas etapas. Primeiramente se escolhe uma base de dados de vídeos sobre a qual experimentos subjetivos de qualidade foram realizados, resultando em um vetor de valores MOS (ou DMOS). A métrica objetiva é então aplicada a todos os vídeos nessa base de dados. Na segunda etapa, são comparados os resultados retornados pela métrica objetiva com os valores MOS (ou DMOS) na base de dados. Essa comparação é feita em geral por uma medida de correlação, de forma que quanto maior for a correlação, mais precisa é a métrica objetiva na estimação da percepção humana de qualidade. No Anexo II detalhamos as duas medidas de correlação utilizadas neste trabalho: a correlação de Pearson e a correlação de Spearman.

Capítulo 4

Proposta de solução

O modelo de atenção visual utilizado neste trabalho se apoia em grande parte no modelo apresentado por Judd [6]. A razão desta escolha é o fato deste ser bastante preciso no que diz respeito à previsão das áreas salientes de uma imagem [13]. Além disso, a sua implementação, disponível online, tem um caráter modular, permitindo que modificações sejam feitas em partes específicas do modelo. No entanto, este modelo foi concebido para imagens e, para podermos utilizá-lo para sinais de vídeo, foi necessário realizar uma adaptação. Para treinamento desta adaptação foi utilizado o banco de dados da IRCCyN [10].

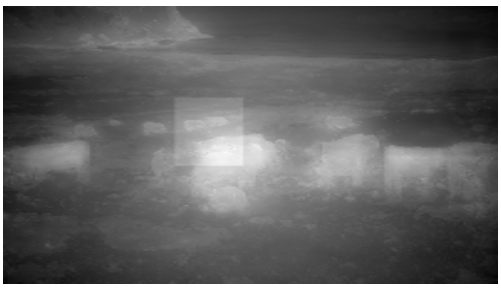
Na avaliação de qualidade de vídeos, os testes de desempenho foram realizados utilizando um outro banco de dados (LIVE Video Quality Database [11], [12]) e um conjunto de métricas de qualidade (MSE, PSNR, SSIM, MS-SSIM e VQM). A informação dos mapas de saliência foi integrada com os resultados das métricas de qualidade e foram medidas as diferenças no desempenho das métricas decorrentes dessa integração. É justamente na busca de uma melhora no desempenho que se encontra o objetivo principal deste trabalho. Seguindo, então, o padrão observado na literatura [43], [46], [44], [15], duas populares metodologias de cálculo de correlação (Anexo II) foram utilizadas para verificar a correspondência estatística entre as métricas objetivas de qualidade neste trabalho e a pontuação média de opiniões (MOS) de um grupo de voluntários [11], [12] em relação aos vídeos na base de dados LIVE.

4.1 Geração do Mapa de Saliência Espacial (Estático)

A implementação encontrada do algoritmo de Judd [47] foi primeiramente otimizada ao máximo em termos do tempo de execução do código. Isto foi feito pensando na aplicação final deste trabalho, que trata da estimação de qualidade para sinais de vídeos. Como a quantidade de quadros em um vídeo, no contexto aqui tratado, é em geral igual a 25-30 vezes a sua duração em segundos, o tempo de execução do código é fundamental para que os resultados dos testes a serem realizados possam ser obtidos em tempo hábil. A otimização foi realizada de modo que não fossem alterados os resultados intermediários de nenhum dos 33 canais do modelo.

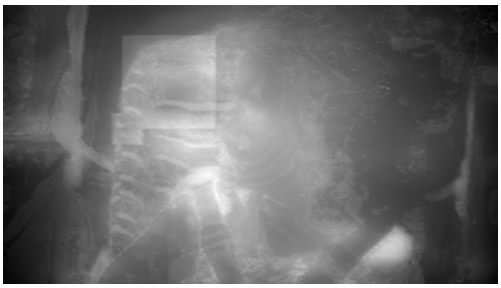
Uma vez ajustado o tempo de execução, havia ainda um problema que poderia potencialmente

(a) Imagem sem faces humanas



(b) Mapa de saliência de (a)

(c) Imagem com face humana



(d) Mapa de saliência de (c)

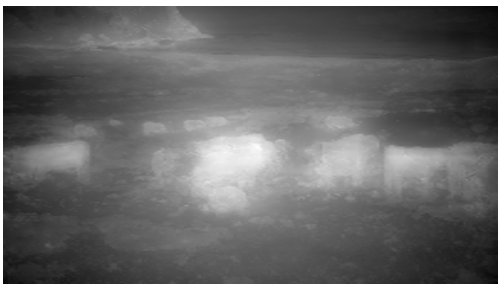
Figura 4.1: Exemplo de (b) falso-positivo e (c) falso-negativo na detecção de face feita pelo algoritmo de Viola e Jones [8]. As detecções de face podem ser notadas pelos retângulos cinzas nos mapas de saliência.

afetar o desempenho da solução final. O detector de faces utilizado por Judd, baseado no algoritmo de Viola e Jones [8], frequentemente fazia previsões falhas identificando faces onde não havia nenhuma. Havia também a ocorrência de falsos-negativos, quando não se detectava faces nos lugares onde elas deveriam ser identificadas. Um exemplo desses comportamentos pode ser visto na Figura 4.1. Na Figura 4.1a, a imagem trata-se apenas de um grupo de vacas num campo aberto e mesmo assim uma face é identificada próxima ao centro da cena (observe o quadrado cinza próximo ao centro da Figura 4.1b). No caso da Figura 4.1c, que contém a face de uma menina, há também um falso-positivo (na cadeira de bebê), porém mais impotante é que a própria face humana na fotografia não foi identificada.

Justamente devido à implementação do modelo de Judd [6] ser modular, é fácil trocar apenas o módulo que identifica as faces na imagem por outro mais acurado. Uma opção encontrada foi o algoritmo proposto por Zhu e Ramanan [9]. Este algoritmo realiza simultaneamente a detecção de faces, a estimação de pose e a estimação de ponto de referência obtendo um desempenho muito melhor que o modelo de Viola e Jones na detecção de faces [9]. A abordagem deste modelamento é similar à proposta por Felzenszwalb *et al.* [34], que é o algoritmo de detecção de pessoas e carros utilizado pelo modelo de Judd [6]. Dessa forma, a mudança aqui feita coloca os três algoritmos utilizados para detecção de objetos (faces, pessoas e carros) em um mesmo nível de complexidade e desempenho.

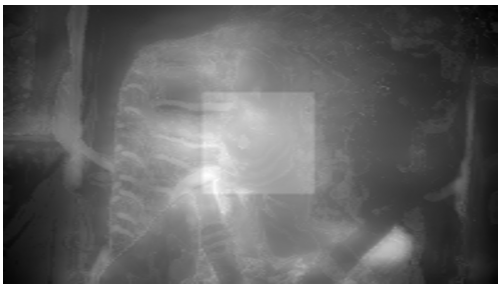
Um exemplo do resultado no mapa de saliência obtido usando este algoritmo de detecção de faces pode ser visto na Figura 4.2. Daqui em diante, os mapas de saliência calculados com as modificações descritas nesta seção passarão a ser referenciados como “mapas de saliência estáticos”.

(a) Imagem sem faces humanas



(b) Mapa de saliência de (a)

(c) Imagem com face humana



(d) Mapa de saliência de (c)

Figura 4.2: Exemplo do que foi consertado com o novo detector de faces [9]. Na primeira imagem nenhuma face é identificada, enquanto que na segunda a face da menina é localizada (ver quadrado na imagem).

4.2 Geração do Mapa de Saliência de Movimento

A presença de movimento é algo que não é levado em consideração em modelos de atenção visual para imagens, mas que é muito importante para sinais de vídeo. Intuitivamente, é fácil imaginar que objetos se movendo mais rapidamente que outros em um vídeo tendem, em princípio, a chamar mais atenção. Pensando nisso, a adaptação concebida para generalizar o modelo foi a de adicionar um mapa, ou canal, de saliência dinâmico ao mapa de saliência estático já descrito anteriormente.

Assim como no caso do modelamento espacial da atenção visual humana, podemos distinguir a consideração de informação temporal aos modelos de saliência como sendo *bottom-up* ou *top-down* [18]. A primeira consiste em adicionar um canal de movimento [25], baseado em características *bottom-up* como mudanças locais de intensidade de pixels, ou mudança de orientação de estruturas entre dois quadros consecutivos de um vídeo. A segunda abordagem, por ser *top-down*, tenta modelar o dinamismo temporal de uma tarefa, ou seja, como as etapas de uma tarefa se desenvolvem com o tempo. Como exemplo deste último tipo de abordagem, Borji *et al.* [19], gravaram ações e movimentos oculares de voluntários em ambientes interativos (video-games 2D e 3D) e modelaram, de forma Bayesiana, a sequência de ações e objetos fixados pelas pessoas conforme a tarefa de jogar os jogos selecionados se desenvolvia.

Optou-se aqui pela primeira abordagem, por ser mais simples e mais fácil de generalizar para a aplicação na estimação da qualidade de vídeos. Mas, diferentemente do que foi feito anteriormente na literatura [25], o canal de movimento foi simplificado. O canal de movimento proposto consiste

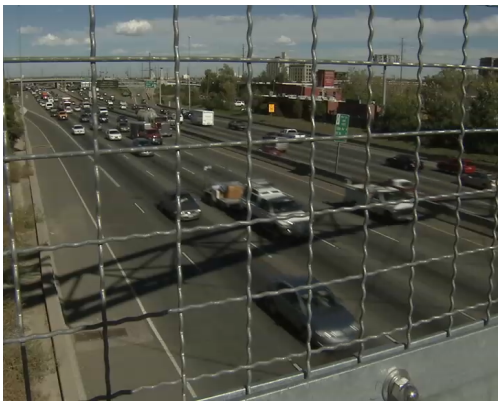
em uma simples aplicação do algoritmo de *Block Matching* a cada dois quadros do vídeo analisado. Na versão deste algoritmo usada neste trabalho, um quadro i qualquer do vídeo analisado é dividido em blocos de tamanho 15×15 pixels e a área de busca é definida de modo que o pixel central de cada bloco possa se mover 7 unidades verticalmente e horizontalmente. O método de busca utilizado é o chamado Exaustivo, o que significa que o algoritmo vasculhará toda a área de busca, pixel a pixel, antes de retornar um resultado. Para cada bloco no quadro i, calcula-se a MSE com relação a cada bloco no quadro i - *1* pertencente à área de busca definida. O bloco mais semelhante no quadro i - *1* é definido então como aquele que minimiza a MSE na área de busca e o vetor de movimento é calculado com direção e módulo (este definido como a distância Euclideana entre os centros dos blocos pareados). Para transformar o mapa de vetores de movimento em um mapa de saliência de movimento, são apenas armazenados os módulos dos vetores para cada bloco, o que resulta em um mapa bidimensional em tons de cinza. Neste mapa, os blocos de pixels que se movimentam mais são mais salientes. Em seguida, o mapa é redimensionado para ficar do mesmo tamanho do quadro i ao qual está relacionado.

Na Figura 4.3 é apresentado um exemplo da geração dos mapas de saliência de movimento utilizando o *Block Matching*. Na Figura 4.3a é apresentado o quadro original, enquanto que na Figura 4.3b é apresentado o canal de movimento correspondente. Observa-se que o algoritmo como descrito acima produz mapas de movimento com ruído. O ruído pode ser causado por uma variação brusca na luminância entre dois quadros. Para gerar versões mais "limpas" do canal de movimento, foram adicionadas duas opções de tratamento dos mapas calculados pelo algoritmo descrito acima. Na primeira delas, antes de redimensionar o mapa para o tamanho do quadro de vídeo analisado, é aplicada uma operação morfológica de abertura [1], eliminando pequenas "ilhas" no mapa que não representam o movimento de um objeto bem definido no vídeo (ver resultado na Figura 4.3c). Na segunda opção de tratamento, após o redimensionamento do mapa de movimento original, é aplicado um filtro gaussiano passa-baixas ao mapa. Desta forma, a imagem é borrada e atenuamos a presença de pequenas regiões com muito movimento que não pertencem nenhum a objeto no vídeo (Figura 4.3d).

4.3 Treinamento

Depois que os dois canais (estático e de movimento) do mapa de saliência final de cada quadro de um vídeo são gerados, é necessário descobrir uma maneira ótima de combinar ambos em um mapa único. Semelhante ao realizado por Judd [6], utilizamos uma técnica de *Machine Learning* baseada em *Support Vector Machines* (SVM). Para treinamento, utilizamos uma base de dados de vídeos, contendo dados de rastreamento de olhar adquiridos em experimentos com observadores humanos. O treinamento retorna os pesos que devem ser atribuídos aos mapas de saliência estático e de movimento.

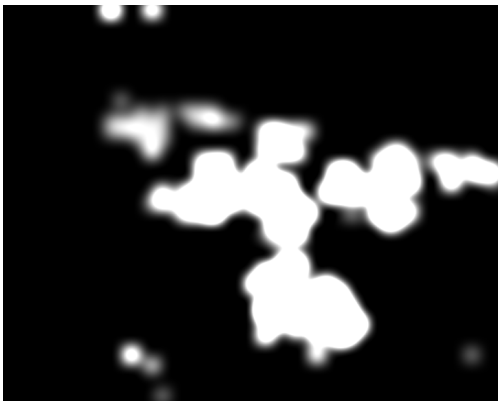
[1] A abertura de uma imagem A por um elemento estruturante B é definida como a erosão de A por B seguida pela dilatação do resultado por B [48]. A operação de abertura remove completamente de A regiões que não possam conter o elemento estruturante, o que suaviza contornos e quebra conexões finas. Neste trabalho, B foi um quadrado com 2 pixels de lado.



(a) Quadro de referência

(b) Canal de movimento padrão

(c) Canal de movimento sofrendo operação de abertura



(d) Canal de movimento sofrendo filtragem gaussiana passa-baixas

Figura 4.3: Exemplo dos mapas de movimento obtidos através das três opções descritas

SVM é um modelo de aprendizado supervisionado, sendo um classificador binário, linear e não-probabilístico. Exemplos em diferentes categorias são classificados de maneira que o espaço que os separa seja o maior possível. O código de treinamento reconhece os mapas de fixação dos experimentos subjetivos na base de dados e designa pesos para os mapas [47] estático e de movimento do modelo aqui apresentado. Isso é feito de forma que regiões salientes no mapa combinado correspondam à classe de regiões fixadas realmente por humanos e, em contrapartida, regiões não salientes correspondam a regiões não-fixadas.

A base de dados utilizada aqui é a do Instituto de Pesquisa em Comunicações e Cibernética de Nantes (IRCCyN, na sigla em francês) [10], vinculado à Escola Politécnica da Universidade de Nantes. Nela, estão contidos 20 vídeos de referência, bem como 4 vídeos com diferentes distorções para cada um dos vídeos originais. Além disso, foram feitos experimentos subjetivos com vídeos de referência de forma a obter mapas de saliência subjetivos, que são baseados nas fixações dos humanos no experimento. São esses mapas que foram tomados como *ground-truth* no treinamento usando SVM. Também são disponibilizados, na base de dados, *Mean Opinion Scores* (MOS) para cada vídeo original e distorcido. Os MOS foram obtidos em um experimento onde os voluntários julgaram a qualidade dos vídeos. Exemplos de quadros de cada vídeo de referência são apresentados na Figura 4.4.

No que concerne a detalhes do treinamento, a cada ensaio o conjunto de vídeos de referência na base de dados IRCCyN é separado, aleatoriamente, em dois grupos: os que serão utilizados para treinar os pesos de cada mapa e os que serão utilizados para testar os pesos escolhidos na etapa de treinamento. Nos dois grupos de vídeos, são escolhidos, novamente de forma aleatória, 33% do total de quadros para serem amostrados e utilizados nos cálculos. Desta forma, este subconjunto se comporta como uma base de imagens e não mais de vídeos. Para todos esses quadros, os mapas de saliência estático e de movimento são calculados previamente para a base de dados inteira. Na SVM, os mapas de saliência referentes aos quadros amostrados são comparados aos mapas de fixação disponíveis na base de dados do LIVE. Ao final do ensaio, o treinamento retorna a dupla de pesos que fornece a melhor combinação dos mapas estático e de movimento, tomando como referência os mapas de fixação reais. O desempenho do mapa final (obtido com a soma ponderada dos mapas estático e de movimento) é avaliado por meio da métrica AUC (*Area Under the ROC Curve*, explicada no Anexo I)[47]. Esse ensaio completo foi repetido 5 vezes. A dupla de pesos que obteve o melhor desempenho foi utilizada no resto deste trabalho para combinar os canais estático e de movimento em um só mapa. A combinação é feita de modo que o mapa de saliência final seja normalizado e contenha índices no intervalo entre 0 e 1.

4.4 Integração Com as Métricas de Qualidade

Uma vez modelada a saliência de cada quadro do vídeo, podemos integrar os mapas de saliência a métricas de qualidade de vídeo. As métricas de qualidade fornecem, não exatamente da mesma maneira, um mapa espacial de erros. A ideia da integração feita aqui é introduzir os mapas de saliência nessa etapa dos algoritmos de estimação de qualidade, multiplicando-os ponto-a-ponto

Figura 4.4: Vídeos na base de dados IRCCyN [10].

aos mapas de erro. Essa integração funciona de maneira a dar maior importância aos erros nas regiões salientes de cada quadro dos vídeos analisados.

Mas essa ponderação não precisa ser feita de forma direta com o mapa de saliência. Como mostrado por Redi *et al.* [43], o uso de funções-peso, baseadas nos mapas de saliência, pode apresentar resultados melhores do que o uso dos mapas de saliência de forma direta. Neste contexto, a função-peso mais simples é aquela descrita pela seguinte equação:

$$WF_1(i,j) = SM(i,j),\ 1 \leq i \leq M\ e\ 1 \leq j \leq N. \tag{4.1}$$

Nesta equação, SM representa o mapa de saliência de um dado quadro de dimensões $M \times N$ em um vídeo qualquer. Mesmo sendo idêntica ao mapa de saliência, por questão de padronização na etapa de integração com as métricas, esta função-peso será identificada como WF_1. A função WF_1, portanto, não representa uma mudança no que já havia se discutido quanto à técnica de

integração da atenção visual às métricas de qualidade.

A segunda função-peso é dada pela seguinte equação [43], [2]:

$$WF_2(i,j) = SM(i,j) + 1,\ 1 \leq i \leq M\ e\ 1 \leq j \leq N. \tag{4.2}$$

Como SM é normalizado para o intervalo $[0,1]$, a função WF_2 não possui valores nulos mas mantém as distâncias relativas entre os maiores e menores índices do mapa de saliência. Desta maneira, assim como a função WF_1, WF_2 privilegia as regiões salientes (primeiro plano), mas evita anular a influência do plano de fundo [43].

A terceira e última função-peso utilizada neste trabalho é a descrita abaixo:

$$WF_3(i,j) = \begin{cases} 1 - SM(i,j) & \text{se } SM(i,j) < 0.5 \\ SM(i,j) & \text{caso contrário} \end{cases},\ 1 \leq i \leq M\ e\ 1 \leq j \leq N. \tag{4.3}$$

A função WF_3, como a função anterior, foi usada por Redi *et al.* [43]. O intuito desta função-peso é privilegiar tanto o primeiro plano (regiões salientes) quanto o plano de fundo dos quadros do vídeo, atenuando a as áreas de transição entre estes dois planos. Redi *et al.* [43] observaram uma melhora na integração da atenção visual com diferentes métricas de qualidade quando foram usadas as funções WF_2 e WF_3 no lugar da simples WF_1. Foi este resultado que motivou a introdução de ponderações diferentes daquela feita simplesmente pela aplicação direta do mapa de saliência.

Em seguida, é necessário combinar ou agregar a informação dos mapas de erro ponderados de forma a obter um único valor para cada quadro do vídeo analisado. Adotamos o procedimento de Liu e Heynderickx [44], dado pela seguinte equação:

$$WMET = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} WF_k(i,j) \cdot Map_{MET}(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} WF_k(i,j)} \tag{4.4}$$

Para uma métrica de qualidade MET dentre as utilizadas (MSE, SSIM, etc.) e uma função-peso WF_k (com $k \in \{1,2,3\}$), o índice $WMET$ é a média ponderada do mapa de qualidade, ou de erros, gerado originalmente por MET, com pesos dados pela função WF_k.

Este método de agregação (ou *pooling*) espacial é diferente do usado por Redi *et al.* [43], o qual envolvia uma abordagem estatística com redes neurais. Entretanto, o mesmo foi testado para WF_1 em [44] , [2] e [46], pelo menos, nos quais mapas de saliência foram usados diretamente na integração. Na Figura 4.5, apresentamos o diagrama de blocos completo da nossa proposta de integração da atenção visual humana a métricas objetivas de qualidade.

A aplicação deste algoritmo de integração é bastante direta para as métricas de qualidade MSE, PSNR, SSIM e MS-SSIM, já que no cálculo dos índices dessas métricas apenas um mapa que poderia ser considerado como um mapa de erros é gerado. Entretanto, a métrica VQM não gera mapas de erro, mas sim um mapa de "comparação" e um mapa do "valor absoluto da informação temporal" (ATI, na sigla em inglês). Seguindo a abordagem proposta por Akamine e Farias [46], substituimos o mapa ATI na VQM pelas funções-peso definidas acima, já que o mapa ATI funciona como as funções-peso, dando mais importância a certas áreas do quadro que outras.

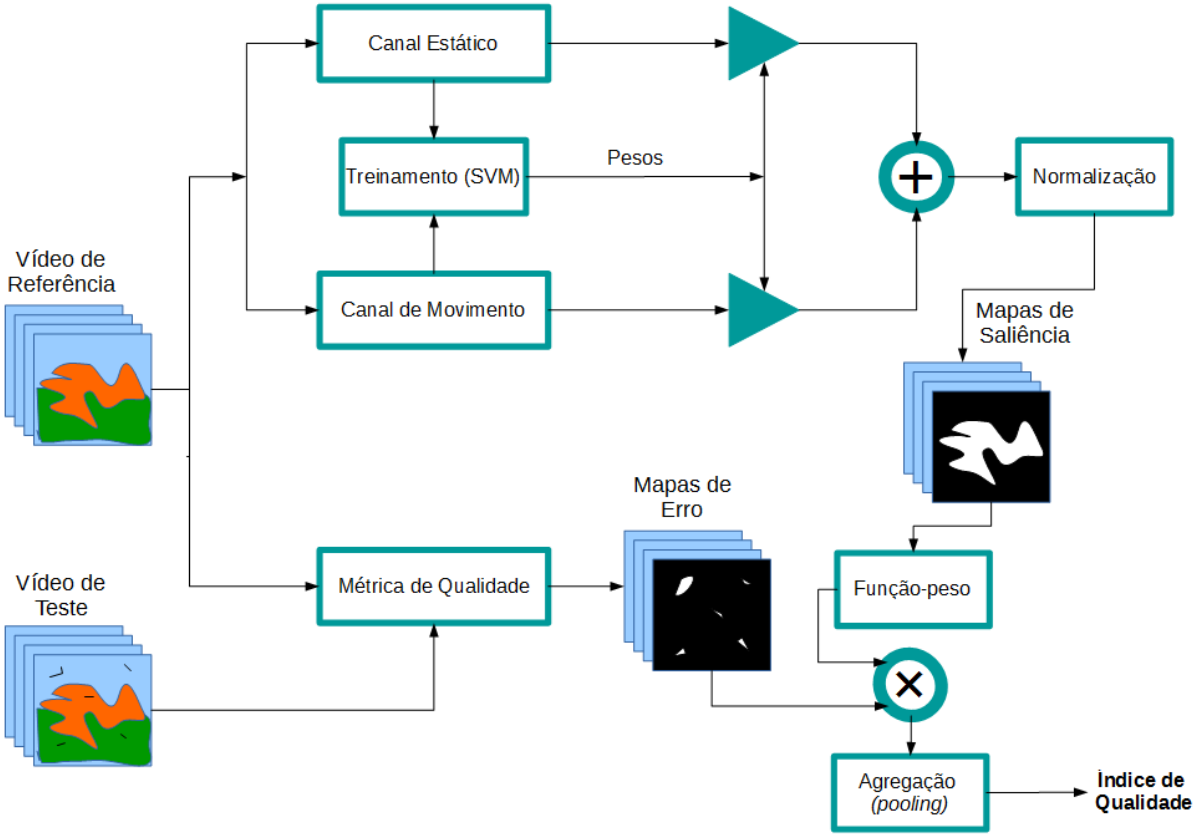


Figura 4.5: Diagrama de blocos da integração proposta neste trabalho.

Por fim, assim como é padrão na avaliação de métricas de qualidade, a eficácia da integração foi medida calculando-se os coeficientes de correlação de Pearson e Spearman (ver Anexo II) entre os índices ponderados das métricas e os valores MOS da base de dados. Já que o modelo de saliência foi treinado utilizando a base de dados do IRCCyN [10], para os testes foi utilizada uma base de dados diferente, a *LIVE Video Quality Database* [11], [12]. A base de dados LIVE possui 10 vídeos de referência. Sobre cada um destes há 15 tipos distorções diferentes, sendo 4 relativas a distorções de comunicações sem fio, 3 relativas a distorções de IP, 4 relativas a distorções de compressão MPEG-2 e 4 relativas a distorções de compressão H.264. Sendo assim, há um total de 150 vídeos na base de dados.

Além disso, para cada uma dessas 150 sequências de vídeos, um valor de DMOS é disponibilizado, resultante de testes subjetivos realizados com voluntários [11], [12]. O nome DMOS significa apenas que foi calculada a diferença entre o MOS do vídeo de teste e o MOS do vídeo de referência. Uma amostra dos quadros de cada vídeo na base de dados LIVE pode ser vista na Figura 4.6.

Figura 4.6: Vídeos na base de dados LIVE [11], [12].

Capítulo 5

Resultados

Os resultados neste capítulo encontram-se divididos em duas grandes seções, seguindo a solução discutida no capítulo anterior. Na primeira seção é descrita a síntese final do modelo de atenção visual apresentado no Capítulo 4, definindo de como os mapas de saliência propostos foram formados e disponibilizados. Na segunda seção apresentamos e analisamos os testes das propostas de integração implementadas. Também definimos a relevância das propostas de aperfeiçoamento e adaptação das métricas de qualidade utilizadas (MSE, PSNR, SSIM, MS-SSIM e VQM).

5.1 Desempenho do Modelo de Saliência Proposto

Na Tabela 5.1 são apresentados os resultados da aplicação da métrica AUC (Anexo I) para diferentes casos de combinação dos canais estático e de movimento do modelo de saliência proposto. A métrica AUC trata os mapas de saliência como classificadores binários. Regiões do mapa acima do limiar são classificada como salientes e regiões abaxo desse limiar como não-salientes. Fazendo variar o nível desse limiar ao mesmo tempo para os mapas de saliência do modelo proposto e os mapas de saliência subjetivos da base de dados IRCCyN, podemos traçar a curva da taxa de falsos-positivos vs. a taxa de verdadeiros-positivos. Quanto maior a área abaixo desta curva, que é o próprio índice AUC, mais próximo dos mapas de saliência subjetivos os mapas de saliência propostos se encontram. Foram incluidos nos testes os casos em que se escolhe apenas um destes mapas para representar o mapa de saliência final. O símbolo '+' na Tabela 5.1 representa a utilização de uma soma ponderada ótima para combinação dos dois canais (estático e de movimento). Lembrando que 5 diferentes treinamentos foram realizados, nos moldes descritos na Seção 4.3. Em outras palavras, os valores de AUC apresentados correspondem aos melhores valores de AUC obtidos entre cinco pares de pesos obtidos nos cinco diferentes treinamentos utilizando a base de dados IRCCyN [10]. Este mesmo procedimento foi utilizado por Judd [6], [20].

Nos casos onde não há combinação (escolha de apenas um dos canais), há apenas uma pequena diferença em como foi escolhido o valor de AUC apresentado na Tabela 5.1. Para manter os testes no mesmo protocolo, realizamos testes com cinco amostragens aleatórias dos vídeos na base de dados para cada um destes casos. Entretanto, como neste contexto não haviam pesos sendo

treinados (já que não há combinação dos canais), não foi tomado o melhor AUC entre os cinco ensaios. Em vez disso, os valores de AUC apresentados nas 4 últimas linhas da Tabela 5.1 são as médias dos valores de AUC para esses cinco ensaios. Esta abordagem tende a eliminar o viés introduzido caso se tenha escolhido aleatoriamente um conjunto particularmente bom ou ruim de vídeos durante algum dos cinco testes.

Na Seção 4.2 foram detalhadas três tentativas de se gerar o canal de movimento (envolvendo a aplicação, ou não, de uma operação morfológica de abertura ou de uma filtragem gaussiana passa-baixas). Os resultados destas três tentativas são apresentados aqui. As tentativas são indicadas em parênteses após o nome "Movimento" na Tabela 5.1.

Tabela 5.1: Medida de desempenho (AUC) dos Estimadores de Mapas de Saliência.

Método de Definição dos Mapas de Saliência	AUC
Estático + Movimento	0,9389
Estático + Movimento (Abertura)	0,9183
Estático + Movimento (Gauss.)	0,9152
Somente Estático	0,9096
Somente Movimento	0,5037
Somente Movimento (Abertura)	0,4963
Somente Movimento (Gauss.)	0,5020

Valores de AUC próximos de 1 representam uma excelente estimativa de saliência (em relação àquela obtida com *eye-tracking*) e valores próximos de $0,5$ representam modelos que se comportam como mapas gerados de forma aleatória (Anexo I). A primeira conclusão que podemos tirar da Tabela 5.1 é que houve de fato uma melhora obtida quando se introduziu um canal de movimento ao canal estático. Além disso, qualquer das tentativas apresentadas de se usar apenas o canal estático ou o de movimento para representar a saliência total dos vídeos é pior do que quando utilizamos ums combinação ótima dos dois canais. Isto é motivador no sentido que pode-se imaginar que as adaptações feitas no modelo de Judd [6] contribuirão para aperfeiçoar as métricas de qualidade para vídeos de forma melhor do que a aplicação do modelo de Judd sozinho.

Com relação à influência das transformações realizadas no canal de movimento (abertura e filtragem gaussiana), oberva-se que as mesmas não contribuem para um mapa mais acurado. Atuando sozinhos como mapas de saliência, os canais de movimento processados não tem desempenho melhor que a versão não processada. Além disso, quando combinados otimamente com o canal estático, aquele que mais contribui para a melhora no desempenho do mapa final é justamente o canal de movimento sem processamento posterior. Este resultado foi fundamental para decidir que as opções de canal de movimento submetidos a tratamentos não fossem utilizadas na integração com as métricas objetivas de qualidade.

Observe também que quando os canais de movimento são usados sozinhos como preditores da saliência visual dos vídeos (últimas três linhas da Tabela 5.1), obtemos um desempenho igual ao de mapas aleatórios. Ou seja, se fosse montado um mapa no qual cada pixel é escolhido como saliente ou não de forma aleatória, seu desempenho segundo a métrica AUC seria o mesmo que

usar os canais de movimento como preditores da saliência dos vídeos. A partir deste fato, pode-se questionar: como é que a combinação do canal estático com o de movimento é melhor que usar apenas o canal estático? Quando é usado sozinho, o canal de movimento proposto se comporta de forma a sugerir que o pixel que mais se moveu entre dois quadros de um vídeo é o mais saliente, sem considerar qualquer outra característica. Isso é claramente uma suposição incorreta, já que esse grande movimento pode ser efeito de uma mudança de iluminação brusca em uma pequena região do vídeo, ou, ainda, esse pixel com grande movimento pode fazer parte de um plano com baixa variabilidade de cor (exemplo: vídeo de um céu azul sem nuvens). Portanto, este movimento seria imperceptível e outras variáveis teriam uma influência muito maior na atenção visual humana. Essas outras variáveis podem ser, por exemplo, a linha de horizonte, o centro do quadro ou a presença de rostos. Entretanto, quando combinado com o mapa estático, o canal de movimento se comporta como deve e dá maior peso às regiões do quadro do vídeo com mais movimento, sendo a intensidade desse peso calculada de forma ótima atráves do treinamento por SVM.

O modelo final de saliência considerado para a integração com as métricas objetivas foi escolhido de acordo com os resultados apresentados na Tabela 5.1. Os pesos escolhidos para combinar os canais de saliência estático e de movimento são aqueles que conseguiram o melhor desempenho ($AUC = 0,9389$). O canal estático, CE, é um mapa normalizado, com valores entre 0 e 1. Já o canal de movimento, CM, não é normalizado, devido a problemas em encontrar uma constante de normalização para o canal. Considerando um quadro de um vídeo com resolução $M \times N$, o mapa de saliência final, SM é dado pela seguinte expressão:

$$SM(i,j) = 0,7739 \cdot CE(i,j) + 0,0399 \cdot CM(i,j),\ 1 \leq i \leq M\ e\ 1 \leq j \leq N, \tag{5.1}$$

no qual (i,j) são as posições espaciais. Após o cálculo da Equação 5.1, o mapa $SalMap$ é normalizado para o intervalo entre 0 e 1 para ser usado na próxima etapa da solução apresentada neste trabalho.

5.2 Desempenho das Tentativas de Integração

Depois de terminados os testes com os modelos de saliência e escolhido o modelo definitivo para o resto deste trabalho, podemos aplicar os mapas de atenção às métricas objetivas MSE, PSNR, SSIM, MS-SSIM e VQM, de forma a melhorar o desempenho na estimação da qualidade de vídeos. Embora as estratégias de integração já tenham sido delineadas através das Equações 4.1 a 4.4, das métricas citadas acima, apenas a VQM foi concebida para vídeos, fornecendo um índice de qualidade por vídeo analisado. As outras são métricas para imagens e fornecem um índice de qualidade para cada quadro de cada vídeo testado. Torna-se importante então, no contexto das métricas para imagens, criar uma maneira de fazer a agregação (ou *pooling*, em inglês) dos índices de cada quadro de modo que, a exemplo da métrica VQM, obtenha-se um índice por vídeo.

Num primeiro momento, foram adotadas como estratégias de agregação Médias, Máximos e Mínimos. Detalhadamente, imaginando-se um vídeo com Q quadros, a aplicação da métrica SSIM neste vídeo retornaria um vetor com Q índices SSIM, um para cada quadro. A primeira técnica de agregação, das Médias, calcula o índice SSIM do vídeo como sendo a média dos Q índices SSIM

individuais de cada quadro. Já a segunda técnica (dos Máximos) escolhe o maior dentre todos os Q índices SSIM. Por fim, a técnica dos Mínimos escolhe como índice SSIM do vídeo o menor dentre os mesmos Q índices. Os resultados dispostos nas tabelas 5.2 a 5.4 são referentes apenas ao uso da métrica SSIM. As outras métricas para imagens se comportaram na maior parte das vezes da mesma maneira nos testes de agregação (*pooling*), sendo apontados os casos em que isso não aconteceu.

Na Tabela 5.2 são apresentados os coeficientes de correlação das três técnicas aplicadas a cada vídeo na base de dados IRCCyN [10]. Assim como convencionado na Seção 4.4, a letra W à frente da métrica indica que a mesma está sendo ponderada e o número ao final do nome indica por qual função-peso se dá essa ponderação. Por exemplo, o nome WSSIM2 indica a aplicação da métrica SSIM ponderada pela função-peso WF2 (vide Seção 4.4). Vale ressaltar que os valores apresentados na Tabela 5.2 devem apenas ser vistos como indicativos do potencial de melhoramento que a integração da atenção visual pode trazer às métricas objetivas de qualidade. Estes dados não devem ser tomados em seu valor absoluto, devido ao viés causado pelo treino do modelo de saliência ter sido realizado na base de dados IRCCyN.

Tabela 5.2: Desempenho da integração com o SSIM sobre a base de dados IRCCyN.

Método de agregação (*pooling*)	Pearson	Spearman
SSIM (Médias)	0,5480	0,6837
WSSIM1 (Médias)	0,6345	0,7559
WSSIM2 (Médias)	0,5760	0,7086
WSSIM3 (Médias)	0,5394	0,6711
SSIM (Máximos)	0,2910	0,3705
WSSIM1 (Máximos)	0,2925	0,3671
WSSIM2 (Máximos)	0,2916	0,3695
WSSIM3 (Máximos)	0,2895	0,3708
SSIM (Mínimos)	0,7356	0,6812
WSSIM1 (Mínimos)	0,8023	0,8165
WSSIM2 (Mínimos)	0,7736	0,7501
WSSIM3 (Mínimos)	0,7253	0,6447

Conforme detalhado no Anexo II, as correlações de Pearson e Spearman variam entre -1 e 1. Os extremos do intervalo são indicativos de uma alta correlação entre as variáveis comparadas, enquanto que o valor 0 indica que não há nenhuma correlação entre as mesmas. Tendo isto em mente, podemos ver que o melhor desempenho foi obtido com a ponderação pela função-peso $WF1$. Além disso, dentre as técnicas de agregação, a de melhor resultados é a dos Mínimos. Isto significa que o menor índice de qualidade dentre os quadros de um vídeo é o melhor estimador da opinião subjetiva da qualidade desse vídeo. Este comportamento é interessante pois indica que o método

de agregação utilizado pode ter um papel fundamental no desempenho da solução final.

Entretanto, para validar a técnica os testes foram realizados em outra base de dados, já que os resultados tendem a ser naturalmente melhores na base de dados em que o modelo de saliência foi treinado. Como explicado no Capítulo 4, a base escolhida para esta validação foi a *LIVE Video Quality Database* [11], [12]. Para explorar mais a influência dos métodos de agregação, foram introduzidas outras técnicas: Medianas, Modas, Médias Geométricas e Médias Harmônicas. Os nomes das técnicas são auto-explicativos e as mesmas atuam sobre o vetor de índices de qualidade representando cada quadro de um dado vídeo de maneira análoga ao método das Médias. Novamente fazendo uso das correlações de Pearson e Spearman, os desempenhos dos métodos de agregação para a métrica SSIM são mostrados na Tabela 5.3.

Mais um vez, se observa uma variação importante quanto ao método de agregação utilizado. Diferentemente do que foi observado na Tabela 5.2, desta vez o melhor método é o das Médias, evidenciando a importância de verificar o desempenho do método em um banco de dados diferente do banco de dados de treinamento. O método das Médias foi o melhor para todas as outras métricas de qualidade observadas, exceto para a PSNR, na a qual o método das Médias Harmônicas teve melhor desempenho. O comportamento da ponderação parece se repetir: em geral função-peso $WF1$ tem melhor desempenho. Já a função-peso $WF3$ parece ser igual ou pior que o caso sem nenhuma ponderação.

A diferença observada na Tabela 5.2 entre o método das Médias e o método dos Mínimos motivou a realização de mais testes inspirados neste último método. Em vez de calcularmos o índice mínimo dentre os quadros de um dado vídeo, separamos uma porcentagem dos índices contendo os menores valores. Em seguida, calculamos a média desse subconjunto. Assim, considerando, por exemplo, uma porcentagem de 30% e um vídeo com Q quadros, aplicamos a métrica para cada quadro. Dos índices gerados, seria calculada a média dos 30% índices de menores valores. Os resultados desses novos testes são apresentados na Tabela 5.4.

Embora a última abordagem descrita seja claramente melhor que o método dos Mínimos, ela não consegue superar o método de agregação das Médias. De fato, conforme se aumenta a porcentagem escolhida, mais próximo do desempenho do método das Médias se aproxima o método das Médias dos Menores. Isso parece óbvio, em segunda análise, já que aumentar a porcentagem de menores significa se aproximar mais do tamanho total do vetor de índices de qualidade, fazendo com que a média dos menores tenda à média geral. De acordo com esses resultados, convencionamos que o método de agregação a ser utilizado para adaptar as métricas de qualidade de imagens para vídeos é o das Médias (a exemplo de [46]). Daqui para frente quando forem citados os nomes MSE, PSNR, SSIM ou MS-SSIM, no contexto de estimação da qualidade de um vídeo, está implícito que o método de agregação utilizado é o das Médias.

Depois de definida a maneira de se adaptar as métricas de imagens para vídeos, podemos comparar como as ponderações pelas diferentes funções-peso propostas (Seção 4.4) se comportam para cada uma das métricas objetivas usadas neste trabalho. Na Tabela 5.5, são apresentadas as correlações de Pearson e Spearman para cada uma das métricas testadas, bem como suas versões ponderadas. Nesta tabela, são apresentados os resultados dos métodos de integração dos mapas

Tabela 5.3: Desempenho da integração com o SSIM sobre a base de dados LIVE.

Método de agregação (*pooling*)	Pearson	Spearman
SSIM (Médias)	0,6258	0,6947
WSSIM1 (Médias)	0,6518	0,7084
WSSIM2 (Médias)	0,6341	0,6995
WSSIM3 (Médias)	0,6260	0,6926
SSIM (Máximos)	0,4213	0,6016
WSSIM1 (Máximos)	0,4077	0,6161
WSSIM2 (Máximos)	0,4168	0,6061
WSSIM3 (Máximos)	0,4228	0,6014
SSIM (Mínimos)	0,2294	0,2299
WSSIM1 (Mínimos)	0,2308	0,2391
WSSIM2 (Mínimos)	0,2297	0,2357
WSSIM3 (Mínimos)	0,2312	0,2303
SSIM (Medianas)	0,5666	0,6205
WSSIM1 (Medianas)	0,5881	0,6268
WSSIM2 (Medianas)	0,5739	0,6235
WSSIM3 (Medianas)	0,5669	0,6200
SSIM (Modas)	0,1338	0,0482
WSSIM1 (Modas)	0,2557	0,1874
WSSIM2 (Modas)	0,1377	0,1260
WSSIM3 (Modas)	0,2972	0,2540
SSIM (Médias Geométricas)	0,5904	0,6640
WSSIM1 (Médias Geométricas)	0,6134	0,6750
WSSIM2 (Médias Geométricas)	0,5978	0,6664
WSSIM3 (Médias Geométricas)	0,5907	0,6628
SSIM (Médias Harmônicas)	0,5472	0,6203
WSSIM1 (Médias Harmônicas)	0,5674	0,6257
WSSIM2 (Médias Harmônicas)	0,5538	0,6235
WSSIM3 (Médias Harmônicas)	0,5474	0,6209

Tabela 5.4: Desempenho da integração com o SSIM sobre a base de dados LIVE (outros testes).

Método de agregação (*pooling*)	Pearson	Spearman
SSIM (Médias dos 50% menores)	0,5217	0,5914
WSSIM1 (Médias dos 50% menores)	0,5400	0,5999
WSSIM2 (Médias dos 50% menores)	0,5276	0,5973
WSSIM3 (Médias dos 50% menores)	0,5218	0,5901
SSIM (Médias dos 40% menores)	0,5085	0,5819
WSSIM1 (Médias dos 40% menores)	0,5248	0,5876
WSSIM2 (Médias dos 40% menores)	0,5138	0,5839
WSSIM3 (Médias dos 40% menores)	0,5088	0,5787
SSIM (Médias dos 30% menores)	0,4946	0,5615
WSSIM1 (Médias dos 30% menores)	0,5086	0,5649
WSSIM2 (Médias dos 30% menores)	0,4992	0,5639
WSSIM3 (Médias dos 30% menores)	0,4951	0,5623
SSIM (Médias dos 20% menores)	0,4724	0,5376
WSSIM1 (Médias dos 20% menores)	0,4849	0,5381
WSSIM2 (Médias dos 20% menores)	0,4767	0,5389
WSSIM3 (Médias dos 20% menores)	0,4724	0,5391
SSIM (Médias dos 10% menores)	0,4465	0,5089
WSSIM1 (Médias dos 10% menores)	0,4547	0,5057
WSSIM2 (Médias dos 10% menores)	0,4499	0,5099
WSSIM3 (Médias dos 10% menores)	0,4469	0,5077

(a) Quadro de um vídeo.

(b) Seu mapa de saliência.

Figura 5.1: Exemplo de um quadro de um vídeo na base de dados LIVE e seu mapa de saliência. São mapas deste tipo que serviram como ponderadores das métricas objetivas de qualidade.

de saliência para métricas de qualidade de vídeo.

Observa-se na Tabela 5.5 que houve uma queda no desempenho das métricas MSE e PSNR quando as mesmas foram ponderadas pelo modelo de atenção visual proposto neste trabalho. Entretanto, observa-se que a MSE e a PSNR são as métricas que apresentam o pior desempenho geral e, desta forma, a ponderação das métricas pela sua saliência não foi capaz de melhorar o seu desempenho. Mas isso felizmente não é constatado com relação às outras métricas e vê-se que o objetivo final, de aperfeiçoar as estimativas de qualidade, foi conseguido. Já para a métrica MS-SSIM o resultado é particularmente satisfatório, atingindo melhoras da ordem de 11% (Pearson) e 8% (Spearman) com relação ao caso não ponderado. Quanto ao efeito causado pela aplicação das diferentes funções-peso, o que vinha sendo notado desde as tabelas anteriores se repete aqui: a função-peso que melhor pondera as métricas de qualidade é a WF_1 e a pior é a WF_3. Os resultados aqui apresentados estão abaixo do esperado, uma vez que em trabalhos anteriores [49], onde se utilizou modelos de atenção puramente *bottom-up*, se obteve níveis de desempenho semelhantes para a mesma base de dados de vídeos (LIVE).

Baseando-se nesta última observação e aproveitando a modularidade do modelo de atenção visual proposto neste trabalho, alguns últimos testes foram realizados para se descobrir quais partes desse modelo contribuem mais para as melhoras observadas nos desempenhos das métricas SSIM, MS-SSIM e VQM após a ponderação pelo mapas de saliência. O modelo de atenção visual completo foi separado em quatro blocos: *Bottom-up*, Centro e Horizonte, Canal de Movimento e Detectores de Objeto (ver Figura 5.2). O primeiro deles se refere à parcela do Canal Estático que contém mapas vindos de modelos *bottom-up* de atenção visual (vide Seção 2.3). Já o segundo, se refere aos canais de Centro e Horizonte utilizados em [6], os quais, segundo Judd, têm um papel extremamente importante na atenção visual humana. O terceiro bloco, por sua vez, é simplesmente o Canal de Movimento do modelo de saliência deste trabalho. Por fim, o quarto bloco corresponde às características top-down do modelo completo, ou seja, os detectores de objeto. A escolha da separação nesses quatro blocos foi feita devido a cada um destes ter, intrinsicamente, uma conotação clara e distinta no modelo completo de atenção, facilitando as análises posteriores.

Os últimos ensaios consistiram em testar cada um destes blocos individualmente, ponderando as métricas objetivas de qualidade e gerando tabelas análogas à Tabela 5.5. A primeira disposição

Tabela 5.5: Desempenho das diferentes métricas.

Métrica	Pearson	Melhora (%)	Spearman	Melhora (%)
MSE	0,5421	–	0,5420	–
WMSE1	0,5280	-2,6%	0,5412	-0,1%
WMSE2	0,5381	-0,7%	0,5418	≈ 0%
WMSE3	0,5438	0,3%	0,5420	0%
PSNR	0,5400	–	0,5224	–
WPSNR1	0,5364	-0,6%	0,5174	-1,0%
WPSNR2	0,5390	-0,2%	0,5216	-0,2%
WPSNR3	0,5400	0%	0,5216	-0,2%
SSIM	0,6258	–	0,6947	–
WSSIM1	0,6518	4,2%	0,7084	2,0%
WSSIM2	0,6341	1,3%	0,6995	0,7%
WSSIM3	0,6260	≈ 0%	0,6926	-0,3%
MS-SSIM	0,6258	–	0,6947	–
WMS-SSIM1	0,6978	11,5%	0,7555	8,8%
WMS-SSIM2	0,6944	11,0%	0,7509	8,1%
WMS-SSIM3	0,6931	10,8%	0,7453	7,3%
VQM	0,7144	–	0,7028	–
WVQM1	0,7345	2,8%	0,7234	2,9%
WVQM2	0,7201	0,8%	0,7106	1,1%
WVQM3	0,6935	-2,9%	0,6823	-2,9%

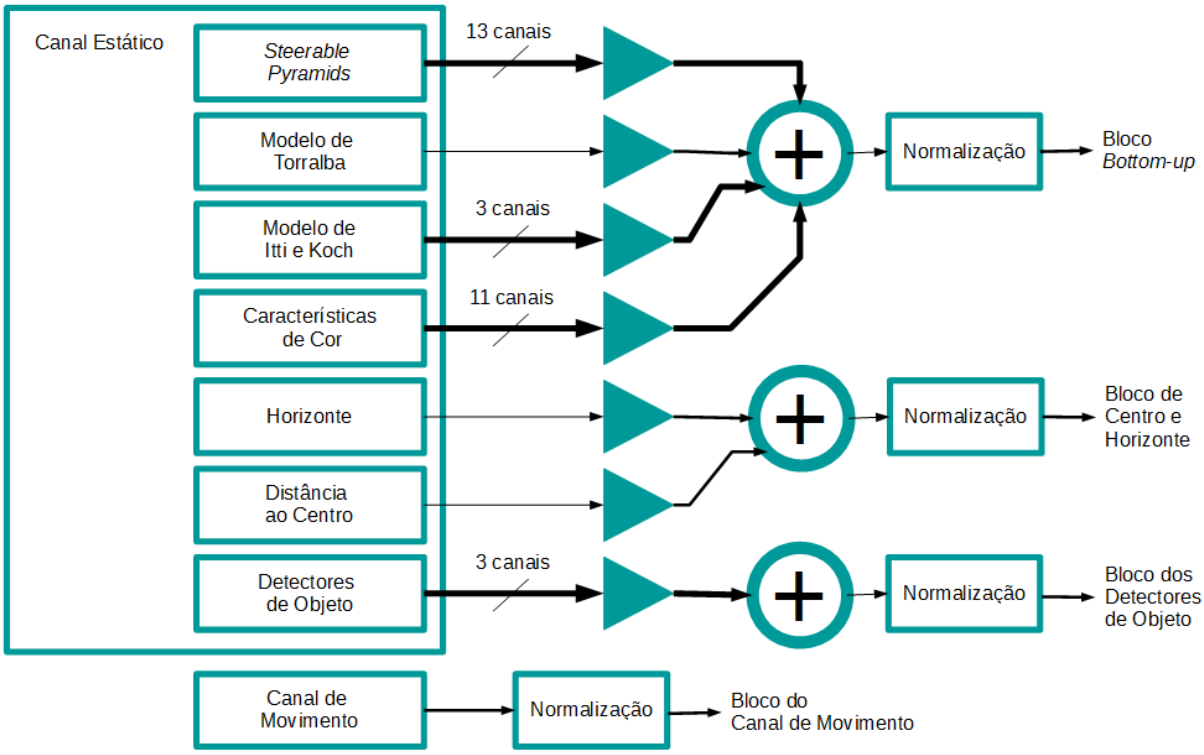


Figura 5.2: Diagrama de blocos da separação do modelo de saliência completo feita para os últimos testes.

de dados desses novos testes se encontra na Tabela 5.6 e se refere ao bloco *Bottom-up* do modelo de saliência deste trabalho. Um exemplo do mapa de saliência gerado por este bloco pode ser visto na Figura 5.3. Observa-se destes resultados que o bloco *Bottom-up* melhorou mais o desempenho da métrica SSIM do que o modelo completo de saliência, conseguindo resultados semelhantes a este último em relação à MS-SSIM. Por outro lado, este bloco piorou o desempenho das métricas MSE e PSNR e, além disso, conseguiu piorar bastante a métrica VQM.

A Tabela 5.7 apresenta os resultados do uso dos canais de Centro e Horizonte como mapas de saliência na ponderação das métricas objetivas de qualidade. Um exemplo dos mapas de saliência gerados por este bloco pode ser visto na Figura 5.4. Em geral, o comportamento aqui observado é parecido com os da tabela anterior, à exceção de que não é observado nenhum desempenho melhor que os da Tabela 5.5. Pode ser observada uma queda no desempenho da métrica VQM ponderada, apesar do fato dos canais aqui considerados terem um papel de peso no modelo de atenção visual humana.

Os resultados do próximo teste são apresentados na Tabela 5.8. Aqui, o Canal de Movimento do modelo completo de saliência é usado como ponderador das métricas de qualidade. Um exemplo do mapa gerado por este bloco pode ser visto Figura 5.5. Relembrando o que foi observado na Tabela 5.1, este canal não é um bom estimador da saliência de um vídeo. Entretanto, a ponderação por ele realizada consegue melhorar um pouco o desempenho das métricas MSE e PSNR (o que não havia sido ainda observado) e também o das métricas SSIM e MS-SSIM. No

(a) Quadro de um vídeo.

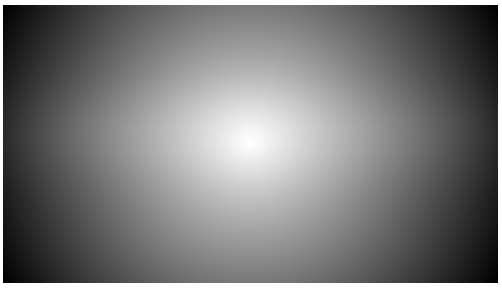
(b) Seu mapa de saliência.

Figura 5.3: Exemplo de um quadro de um vídeo na base de dados LIVE e seu mapa de saliência (este formado apenas pela parte *bottom-up* do modelo completo de saliência).

Tabela 5.6: Desempenho das diferentes métricas (Mapas de Saliência = Mapas *Bottom-up*).

Métrica	Pearson	Melhora (%)	Spearman	Melhora (%)
MSE	0,5421	–	0,5420	–
WMSE1	0,5088	-6,1%	0,5415	-0,1%
WMSE2	0,5335	-1,6%	0,5408	-0,2%
WMSE3	0,5465	0,8%	0,5374	-0,8%
PSNR	0,5400	–	0,5224	–
WPSNR1	0,5328	-1,3%	0,5158	-1,3%
WPSNR2	0,5375	-0,5%	0,5217	-0,1%
WPSNR3	0,5407	0,1%	0,5216	-0,2%
SSIM	0,6258	–	0,6947	–
WSSIM1	0,6637	6,1%	0,7151	2,9%
WSSIM2	0,6353	1,5%	0,7001	0,8%
WSSIM3	0,6142	-1,9%	0,6871	-1,1%
MS-SSIM	0,6258	–	0,6947	–
WMS-SSIM1	0,6943	10,9%	0,7521	8,3%
WMS-SSIM2	0,6936	10,8%	0,7501	8,0%
WMS-SSIM3	0,6894	10,16%	0,7439	7,1%
VQM	0,7144	–	0,7028	–
WVQM1	0,5851	-18,1%	0,5874	-16,4%
WVQM2	0,6983	-2,3%	0,6906	-1,7%
WVQM3	0,7041	-1,4%	0,6875	-2,2%

(a) Quadro de um vídeo.



(b) Seu mapa de saliência.

Figura 5.4: Exemplo de um quadro de um vídeo na base de dados LIVE e seu mapa de saliência (este formado apenas pelos canais de Centro e Horizonte do modelo completo de saliência).

Tabela 5.7: Desempenho das diferentes métricas (Mapas de Saliência = Mapas de Centro e Horizonte).

Métrica	Pearson	Melhora (%)	Spearman	Melhora (%)
MSE	0,5421	–	0,5420	–
WMSE1	0,5425	0,1%	0,5371	-0,9%
WMSE2	0,5427	0,1%	0,5412	-0,1%
WMSE3	0,5463	0,8%	0,5437	0,3%
PSNR	0,5400	–	0,5224	–
WPSNR1	0,5403	0,1%	0,5182	-0,8%
WPSNR2	0,5406	0,1%	0,5215	-0,2%
WPSNR3	0,5432	0,6%	0,5236	% 0,2
SSIM	0,6258	–	0,6947	–
WSSIM1	0,6438	2,9%	0,7055	1,6%
WSSIM2	0,6317	0,9%	0,6979	0,5%
WSSIM3	0,6243	-0,2%	0,6936	-0,2%
MS-SSIM	0,6258	–	0,6947	–
WMS-SSIM1	0,6988	11,7%	0,7548	8,7%
WMS-SSIM2	0,6946	11,0%	0,7492	7,8%
WMS-SSIM3	0,6918	10,5%	0,7466	7,5%
VQM	0,7144	–	0,7028	–
WVQM1	0,6958	-2,6%	0,6856	-2,4%
WVQM2	0,7069	-1,1%	0,7002	-0,4%
WVQM3	0,7115	-0,4%	0,7010	-0,3%

(a) Quadro de um vídeo.

(b) Seu mapa de saliência.

Figura 5.5: Exemplo de um quadro de um vídeo na base de dados LIVE e seu mapa de saliência (este formado apenas pelo Canal de Movimento do modelo completo de saliência).

caso desta última, o desempenho chega a ser igual a quando se usa o modelo completo de saliência como ponderador. Essa melhora pode vir do fato que o mapa de estimação de movimento adiciona informação temporal à estimação de qualidade, o que não é levado em consideração pelas métricas SSIM e MS-SSIM (já que elas foram concebidas para imagens). Ainda assim, o Canal de Movimento como ponderador consegue os piores desempenhos observados até agora para a métrica VQM. Este último resultado não era esperado, já que, no caso desta métrica, o Canal de Movimento substitui o mapa ATI (ver Seção 3.1.5), que funciona justamente como um filtro de detecção de movimento no algoritmo ma VQM.

Finalmente, o que se obteve do último dos testes é resumido na Tabela 5.9. Aqui os mapas de saliência são formados pela parte do Canal Estático que busca introduzir informações top-down baseadas na tendência humana de desviar a atenção em direção a pessoas, faces e, com menos intensidade, carros. Os mapas do bloco de Detectores de Objetos conseguem melhorar, ainda mais que o Canal de Movimento, a estimação de qualidade das métricas MSE e PSNR. Adicionalmente, há melhoras também nas métricas SSIM e MS-SSIM (esta com valores quase tão altos quanto os observados anteriormente). Infelizmente, quando novamente se examina o caso da métrica VQM, a aplicação dos mapas do bloco de Detectores de Objetos falha enormemente em melhorar a estimativa de qualidade desta métrica.

Observamos a partir dos resultados que só houveram melhoras significativas para as métricas MSE e PSNR quando os mapas ponderadores utilizados foram o Canal de Movimento e os mapas do bloco de Detectores de Objetos. As métricas SSIM e MS-SSIM sofreram melhoras em todos os casos, ocorrendo o pico de desempenho da primeira quando o bloco *Bottom-up* do Canal Estático foi usado como ponderador e havendo, no caso da segunda, melhoras significativas em todos os testes. A métrica VQM, por fim, teve comportamentos extremos, tendo sido um pouco melhorada quando o modelo completo de saliência foi utilizado, mas mostrando desempenhos péssimos quando, alternativamente, se usou um dos quatro blocos descritos para realizar a ponderação.

Tabela 5.8: Desempenho das diferentes métricas (Mapas de Saliência = Canal de Movimento).

Métrica	Pearson	Melhora (%)	Spearman	Melhora (%)
MSE	0,5421	–	0,5420	–
WMSE1	0,5603	3,4%	0,5517	1,8%
WMSE2	0,5469	0,9%	0,5434	0,3%
WMSE3	0,5398	-0,4%	0,5362	-1,1%
PSNR	0,5400	–	0,5224	–
WPSNR1	0,5416	0,3%	0,5196	-0,5%
WPSNR2	0,5411	0,2%	0,5224	0,0%
WPSNR3	0,5377	-0,4%	0,5206	-0,3%
SSIM	0,6258	–	0,6947	–
WSSIM1	0,6369	1,8%	0,6993	0,7%
WSSIM2	0,6305	0,8%	0,6983	0,5%
WSSIM3	0,6249	-0,1%	0,6930	-0,2%
MS-SSIM	0,6258	–	0,6947	–
WMS-SSIM1	0,6992	11,7%	0,7508	8,1%
WMS-SSIM2	0,6955	11,1%	0,7495	7,9%
WMS-SSIM3	0,6911	10,4%	0,7457	7,3%
VQM	0,7144	–	0,7028	–
WVQM1	0,4461	-37,6%	0,4419	-37,1%
WVQM2	0,7075	-1,0%	0,7066	0,5%
WVQM3	0,6182	-13,5%	0,6300	-10,4%

(a) Quadro de um vídeo.

(b) Seu mapa de saliência.

Figura 5.6: Exemplo de um quadro de um vídeo na base de dados LIVE e seu mapa de saliência (este formado apenas pelo mapa dos detectores de objeto do modelo completo de saliência).

Tabela 5.9: Desempenho das diferentes métricas (Mapas de Saliência = Mapas de detecção de objetos).

Métrica	Pearson	Melhora (%)	Spearman	Melhora (%)
MSE	0,5421	–	0,5420	–
WMSE1	0,5676	4,7%	0,5659	4,4%
WMSE2	0,5482	1,1%	0,5466	0,8%
WMSE3	0,5421	0%	0,5413	-0,1%
PSNR	0,5400	–	0,5224	–
WPSNR1	0,5604	3,8%	0,5424	3,8%
WPSNR2	0,5450	0,9%	0,5294	1,3%
WPSNR3	0,5397	$\approx$ 0%	0,5229	0,1%
SSIM	0,6258	–	0,6947	–
WSSIM1	0,6339	1,3%	0,7028	1,2%
WSSIM2	0,6280	0,4%	0,6959	0,2%
WSSIM3	0,6257	$\approx$ 0%	0,6948	$\approx$ 0%
MS-SSIM	0,6258	–	0,6947	–
WMS-SSIM1	0,6913	10,5%	0,7509	8,1%
WMS-SSIM2	0,6920	10,6%	0,7478	7,6%
WMS-SSIM3	0,6918	10,5%	0,7477	7,6%
VQM	0,7144	–	0,7028	–
WVQM1	0,5715	-20,0%	0,6451	-8,2%
WVQM2	0,7079	-0,9%	0,6985	-0,6%
WVQM3	0,6595	-7,7%	0,6638	-5,5%

Capítulo 6

Conclusão

Neste trabalho foram apresentadas prospostas de adaptação de um modelo de atenção visual de alto desempenho e de integração deste modelo com métricas objetivas de qualidade de vídeo. O Canal Estático do modelo de saliência aqui proposto foi composto basicamente pelo algoritmo de atenção visual para imagens de Judd *et al.* [6]. Entretanto, foi feita uma contribuição a esse algoritmo ao se trocar o detector de faces Viola Jones [8] por um mais preciso, proposto por Zhu e Ramanan [9]. Embora tenha sido observado que o Canal Estático sozinho estima muito bem a saliência de vídeos (ver Tabela 5.1), a combinação com o Canal de Movimento discutido na Seção 4.2 consegue melhorar esse desempenho. Essa melhora foi conseguida mesmo com um algoritmo muito simples de estimação de movimento (*Block-Matching*), o que indica a importância, para a atenção visual humana, do movimento nos vídeos. Houve tentativas de tornar os mapas de movimento menos ruidosos através da aplicação de uma operação morfológica (abertura) ou uma filtragem passa-baixas gaussiana, mas elas não se traduziram em melhor desempenho de estimação de saliência para o modelo final (Canal Estático + Canal de Movimento). Ainda sobre o Canal de Movimento, vale observar que ele não deve ser usado sozinho como estimador da saliência de um vídeo, já que observamos que, nesta situação, os mapas de movimento têm o mesmo desempenho de mapas gerados aleatoriamente.

Excluindo o caso da aplicação do Canal de Movimento sozinho, vemos que na Tabela 5.1 os desempenhos medidos foram bastante altos, maiores inclusive do que qualquer um encontrado no *Benchmark* para imagens compilado por Judd *et al.* [20], [13]. Devido a este fato, são indicados testes futuros com bases de dados de vídeo maiores e mais diversas para verificar os valores AUC absolutos da Tabela 5.1. A última observação quanto ao modelo de atenção visual proposto concerne a parcela *top-down* do mesmo. Judd *et al.* [6] utilizaram os detectores de pessoas, faces e carros como elementos *top-down* de seu modelo de atenção visual pois eles perceberam que esses objetos eram muito fixados pelas pessoas, quando a tarefa dada era de memorizar a cena observada. Mesmo que esta tarefa possa se relacionar à avaliação de qualidade, o modelamento específico da tarefa de avaliação de qualidade pode trazer melhores desempenhos quando a atenção visual for integrada a métricas de qualidade. O modelamento dessa tarefa pode ser feita, por exemplo, como no trabalho de Borji *et al.* [19], no qual a saliência de cenas de video-game foi modelada através de uma rede Bayesiana treinada com dados do *eye-tracker* e do *joystick* simultaneamente.

Na segunda parte do trabalho, que se refere às tentativas de melhoramento das estimativas das métricas de qualidade, a estratégia do uso de funções-peso baseadas nos mapas de saliência (ver Seção 4.4) não foi tão frutífera quanto esperado. A ponderação feita pela função WF_1 foi melhor que aquelas feitas pelas funções WF_2 e WF_3 em quase todos os casos, lembrando que a função WF_1 é o próprio modelo de saliência proposto neste trabalho, sem modificações. Esses resultados podem ser devidos ao uso da Equação 4.4 para fazer a agregação (*pooling*) em cada quadro de vídeo. Este método funciona bem quando o mapa de saliência é usado de forma direta para ponderar os mapas das métricas de qualidade [44] , [2], [46]. Entretanto, quando Redi *et al.* [43] introduziram a estratégia das funções-peso, o método de agregação a cada imagem envolveu uma análise estatística e o uso de uma rede neural. Concluimos, então, que as melhoras trazidas pelo uso das funções-peso WF_2 e WF_3 só será observado se um método de *pooling* quadro-a-quadro mais complexo for empregado.

A maioria das métricas de qualidade utilizadas (MSE, PSNR, SSIM e MS-SSIM) eram métricas concebidas para imagens e um método de agregação entre quadros também teve que ser proposto (ver Figura 4.5). A técnica de agregação que obteve resultados mais correlacionados com a percepção humana de qualidade foi a das Médias. Nesta técnica, o índice de qualidade de um vídeo é definido como a média dos índices de qualidade dos quadros desse vídeo. No entanto, não afirmamos aqui que esta é a melhor técnica possível, já que a variabilidade de desempenho observada entre as diversas técnicas de agregação sugere que pode haver uma estratégia melhor que o uso de uma simples média aritmética. No caso da métrica VQM, os mapas de saliência subtituíram os mapas ATI no cálculo do parâmetro $CT_{ATIgain}$ e o *pooling* foi feito seguindo normalmente o algoritmo da métrica.

Dentre as métricas objetivas selecionadas para este trabalho, a que obteve maior crescimento percentual de desempenho com a inserção das informações da atenção visual humana foi a MS-SSIM. A melhora de desempenho observada foi de 11,5%. Em valores absolutos, a métrica VQM, que já era a melhor inicialmente, foi a que teve melhor desempenho depois de ponderada pelos mapas de saliência do modelo proposto. Não observamos melhora de desempenho da MSE e da PSNR após a integração. Apesar de comprovado o aperfeiçoamento da estimativa de qualidade para algumas das métricas, as melhoras não foram muito diferentes das observadas por Akamine e Farias [46], [49], que utilizaram um modelo menos complexo, *bottom-up*, de atenção visual. Dividimos então o modelo de saliência proposto em 4 blocos (ver Figura 5.2) e medimos o desempenho da integração com as métricas de qualidade feita por cada um deles. Observamos que, para as métricas MSE e PSNR, apenas o bloco do Canal de Movimento e o bloco de Detectores de Objeto trouxeram melhorias. A MS-SSIM teve sua estimativa de qualidade melhorada em todos os 4 casos. O bloco *Bottom-up* trouxe melhorias maiores que o modelo de saliência completo no caso da métrica SSIM. Apesar disso, todos os 4 blocos falharam em melhorar o desempenho da métrica VQM, fato que apenas o modelo completo conseguiu.

Conseguimos atingir o objetivo de aperfeiçoar as estimativas das métricas de qualidade analisadas, mas, por que mesmo utilizando um bom modelo de atenção visual os resultados absolutos ficaram aquém do esperado? Talvez a resposta esteja no que foi indicado por You *et al.* [42]: a atenção visual e a avaliação de qualidade não são processos separados no cérebro humano. A

integração proposta neste trabalho é feita calculando a qualidade e a saliência de forma paralela e só ao final ponderando uma pela outra. A métrica AFViQ proposta por You *et al.* [42] considera a atenção visual como parte integrante fundamental do algoritmo da métrica da qualidade e, com isso, apresenta coeficientes de correlação com a opinião humana maiores do que os observados neste trabalho. Assim, podemos considerar que esta interdependência entre a atenção visual e a avaliação de qualidade deve ser considerada em tentativas futuras para construir métricas de qualidade mais precisas. Além disso, a natureza *Full-Reference* das métricas de qualidade utilizadas neste trabalho é uma séria restrição a sua utilização em vários campos de aplicação práticos, como em serviços de telecomunicações, nos quais a banda utilizada é fator crítico do sistema. Portanto, além de tentar conseguir desempenhos melhores baseados em métricas *Full-Reference*, devemos incentivar maiores esforços no desenvolvimento de métricas *No-Reference* (ou pelo menos *Reduced-Reference*), que são mais flexíveis e de mais aplicáveis fora do ambiente de laboratório.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] WINKLER, S.; MOHANDAS, P. The evolution of video quality measurement: From PSNR to hybrid metrics. *IEEE Transactions on Broadcasting*, IEEE, v. 54, p. 1–8, 2008.

[2] NINASSI, A. et al. Does where you gaze your attention on an image affect your perception of quality? Applying visual attention to image quality metrics. In: IEEE. *Proc. International Conference on Image Processing (ICIP)*. San Antonio, TX USA, 2007. v. 2, p. 169–172.

[3] RAJASHEKAR, U. et al. GAFFE: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing*, IEEE, v. 17, p. 564–573, 2008.

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

[5] ITTI, L.; KOCH, C.; NIEBUR, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 20, n. 11, p. 1254–1259, 1998.

[6] JUDD, T. et al. Learning to predict where humans look. In: IEEE. *International Conference on Computer Vision (ICCV)*. Kyoto, 2009. p. 2106–2113.

[7] WANG, Z.; BOVIK, A. C. Mean squared error: Love it or leave it? - a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, IEEE, v. 25, n. 1, p. 98–117, 2009.

[8] VIOLA, P.; JONES, J. Robust real-time face detection. *International Journal of Computer Visison (IJCV)*, Kluwer Academic Publishers, v. 57(2), p. 137–154, 2004.

[9] ZHU, X.; RAMANAN, D. Face detection, pose estimation, and landmark localization in the wild. In: IEEE. *Conference on Computer Vision and Pattern Recognition (CVPR)*. Providence, RI USA, 2012. p. 2879–2886.

[10] ENGELKE, U. et al. Modelling saliency awareness for objective video quality assessment. In: *Second International Workshop on Quality of Multimedia Experience (QoMEX)*. Trondheim: IEEE, 2010. p. 212–217.

[11] SESHADRINATHAN, K. et al. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, IEEE, v. 19, n. 6, p. 1427–1441, 2010.

[12] SESHADRINATHAN, K. et al. A subjective study to evaluate video quality assessment algorithms. In: SPIE. *Proceedings Human Vision and Electronic Imaging*. San Jose, CA USA, 2010. v. 7527.

[13] JUDD, T.; DURAND, F.; TORRALBA, A. A benchmark of computational models of saliency to predict human fixations. In: *MIT Technical Report*. [S.l.: s.n.], 2012.

[14] WANG, Z. et al. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, IEEE, v. 13, n. 4, p. 600–612, 2004.

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

[16] TREISMAN, A. M.; GELADE, G. A feature integration theory of attention. *Cognitive Psychology*, v. 12, p. 97–136, 1980.

[17] KOCH, C.; ULLMAN, S. Shifts in selective visual attention: Towards an underlying neural circuitry. *Human Neurobiology*, v. 4, n. 4, p. 219–227, 1985.

[18] BORJI, A.; ITTI, L. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 35, p. 185–207, 2013.

[19] BORJI, A.; SIHITE, D. N.; ITTI, L. What/where to look next? modeling top-down visual attention in complex interactive environments. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, IEEE, v. 44, p. 523–538), 2014.

[20] JUDD, T. *Understanding and Predicting Where People Look in Images*. Tese (Doutorado) — Massachusetts Institute of Technology, 2011.

[21] KOCH, K.; MCLEAN, J. et al. How much the eye tells the brain. *Current Biology*, v. 25, n. 14–16, p. 1423–1434, 2006.

[22] BUSWELL, G. T. *Fundamental reading habits: A study of their development*. [S.l.]: University of Chicago Press, 1922.

[23] HOFFMAN, J. E. Visual attention and eye movements. In: PASHLER, H. (Ed.). *Attention*. [S.l.]: Psychology Press, 1998. p. 119–154.

[24] HENDERSON, J. M.; HOLLINGWORTH, A. High-level scene perception. *Ann. Rev. Psychology*, v. 50, p. 243–271, 1999.

[25] ITTI, L. et al. Realistic avatar eye and head animation using a neurobiological model of visual attention. In: SPIE. *Proc. Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI*. San Diego, CA USA, 2004. v. 5200, p. 64–78.

[26] BRUCE, N.; TSOTSOS, J. Attention based on information maximization. *Journal of Vision*, ARVO, v. 7, n. 9, 2007.

[27] ITTI, L.; KOCH, C. Computational modelling of visual attention. *Nature Reviews Neuroscience*, Nature, v. 2, p. 1–11, 2001.

[28] EINHAUSER, W. et al. Objects predict fixations better than early saliency. *Journal of Vision*, ARVO, v. 14, p. 787–835, 2008.

[29] PETERS, R. J.; ITTI, L. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2007.

[30] SIMONCELLI, E. P.; FREEMAN, W. T. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In: IEEE. *Proc. International Conference on Image Processing (ICIP)*. Washington, DC USA, 1995. v. 3, p. 444–447.

[31] TORRALBA, A. et al. Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, American Psychological Association, 2006.

[32] WALTHER, D. Modeling attention to salient proto-objects. *Neural Networks*, Elsevier, v. 19, p. 1395–1407, 2006.

[33] RUSSELL, B. C. et al. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, v. 77, p. 157–173, 2008.

[34] FELZENSZWALB, P. F. et al. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 32, p. 1627–1645, 2010.

[35] ZHANG, J.; SCLAROFF, S. Saliency detection: A boolean map approach. In: *IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2013.

[36] YOU, J. et al. Perceptual-based objective quality metrics for audio-visual services? a survey. *Signal Process. Image Commun.*, v. 25, n. 7, p. 482–501, 2010.

[37] WANG, Z.; SIMONCELLI, E. P. Translation insensitive image similarity in complex wavelet domain. *IEEE Transactions on Broadcasting*, IEEE, v. 54, p. 1–8, 2008.

[38] WANG, Z.; SIMONCELLI, E. P.; BOVIK, A. C. Multi-scale structural similarity for image quality assessment. In: *IEEE Asilomar Conference Signals, Systems and Computers*. [S.l.: s.n.], 2003.

[39] PINSON M. E WOLF, S. *Video Quality Metric Software, Version 2.* `http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm`.

[40] ENGELKE, U. *Modelling Perceptual Quality and Visual Saliency for Image and Video Communications*. Tese (Doutorado) — Blekinge Institute of Technology, 2010.

[41] WANG, Z. et al. Foveated wavelet image quality index. San Diego, CA USA, v. 4472, p. 42–52, 2001.

[42] YOU, J.; EBRAHIMI, T.; PERKIS, A. Attention driven foveated video quality assessment. *IEEE Transactions on Image Processing*, IEEE, v. 23, n. 1, p. 200–213, 2014.

[43] REDI, J. A. et al. How to apply spatial saliency into objective metrics for jpeg compressed images? In: IEEE. *16th International Conference on Image Processing (ICIP)*. Cairo, 2009. p. 961–964.

[44] LIU, H.; HEYNDERICKX, I. Studying the added value of visual attention in objective image quality metrics based on eye movement data. In: IEEE. *16th International Conference on Image Processing (ICIP)*. Cairo, 2009. p. 3097–3100.

[45] Internatonal Telecommunication Union. *Methodology for the subjective assessment of the quality of television pictures*. [S.l.]: ITU-R, Rec. BT.500-11, 2002.

[46] AKAMINE, W. Y. L.; FARIAS, M. C. Q. Incorporating visual attention models into video quality metrics. In: SPIE. *Proc. Image Quality and System Performance XI*. San Francisco, CA USA, 2014. v. 9016.

[47] JUDD, T. et al. *Learning to Predict Where Humans Look*. `http://people.csail.mit.edu/tjudd/WherePeopleLook/`.

[48] GONZALEZ, R. C.; WOODS, R. E.; EDDINS, S. L. *Digital Image Processing Using MATLAB*. Upper Saddle River, NJ USA: Pearson Prentice Hall, 2004.

[49] AKAMINE, W. Y. L.; FARIAS, M. C. Q. Video quality assessment using visual attention computational models. Unpublished. 2014.

[50] GREEN, D.; SWETS, J. *Signal Detection Theory and Psychophysics*. New York, NY USA: John Wiley, 1966.

[51] PETERS, R. et al. Components of bottom-up gaze allocation in natural images. *Vision Research*, Elsevier, v. 45, p. 2397–2416, 2005.

ANEXOS

I. MÉTRICAS PARA AVALIAR O DESEMPENHO DE MODELOS DE SALIÊNCIA

A avaliação da precisão de um modelo de saliência é em geral feita através de uma comparação. A base para essa comparação, ou o que é chamado em inglês de *ground-truth*, são mapas de saliência obtidos a partir de experimentos subjetivos. Nesses experimentos, voluntários são convidados a assistir vídeos ou observar imagens e seus olhares são rastreados. O resultado desse rastreio é um mapa de fixações para cada quadro ou imagem. Pontos em um mapa de fixações representam amostras das posições nas imagens nas quais os voluntários fixaram o olhar. Esses mapas de fixações podem ser transformados em mapas de saliência através de um processo chamado "'foveação" (*foveation*, em inglês), nomeado em referência à fóvea humana. A "foveação" consiste em simular a área focalizada pela fóvea humana e geralmente é implementada por uma filtragem gaussiana sobre o mapa de fixações. A variância do filtro gaussiano depende da abertura angular da fóvea humana e da distância entre as imagens e os voluntários durante os experimentos subjetivos.

Definida a base, a comparação pode ser feita de diversas maneiras, mas um grupo delas é mais popular na literatura [18]. Imaginando que os mapas de saliência são distribuições de probabilidade, podemos usar a divergência de Kullback-Leibler para medir a distância entre os mapas comparados. Se por outro lado, considerarmos os mapas como sendo variáveis aleatórias, métricas usadas para compará-los podem ser o Coeficiente de Correlação (CC) ou a *Normalized Scanpath Saliency* (NSS). Por fim, ainda pode-se imaginar os mapas como classificadores binários (separando a imagem em regiões salientes e regiões não-salientes) e neste caso avaliamos o desempenho do classificador através da Área Abaixo da Curva ROC (AUC, na sigla em inglês).

Detalhamos nas próximas seções as métricas AUC e NSS. A primeira foi efetivamente usada neste trabalho para avaliar o modelo de saliência proposto. Já a segunda só está presente para indicar como uma métrica diferente avalia a precisão de um modelo de saliência.

I.1 AUC

A sigla AUC se refere à Área Abaixo da Curva ROC. ROC é uma sigla em inglês vinda do domínio da teoria de detecção de sinais para se referir à Característica de Operação do Receptor. A obteção desta curva característica tem sido usada desde a Segunda Guerra Mundial, quando foi empregada na análise de sinais de radar [50]. Como apontam Borji e Itti [18], a AUC se tornou a métrica mais popular na comunidade de modelamento computacional da atenção visual humana para avaliar a eficácia dos modelos criados. A ideia central da métrica é considerar os mapas de saliência de um modelo como um classificador binário. Para um dado limiar na intensidade de níveis de cinza, pixels no mapa são considerados como fixados (salientes) se suas intensidades forem maiores que a deste limiar e não-fixados caso contrário. Fazendo variar o limiar e comparando com os resultados do mapa de saliência subjetivo (*ground-truth*), podemos traçar a curva da taxa

de falsos-positivos versus a taxa de verdadeiros-positivos. É justamente esta última curva que é definida como sendo a Característica de Operação do Receptor (ROC).

Se os mapas de saliência modelados forem exatamente iguais aos mapas experimentais, haverá apenas verdadeiros-positivos, independentemente do limiar fixado, e com isso a área abaixo da curva ROC (AUC) será igual a 1. Alternativamente, se os mapas modelados forem exatamente o oposto dos mapas experimentais, teremos $AUC = 0$, o que em si também é um bom resultado, já que os modelos estarão apenas a uma operação matemática de estimar perfeitamente a atenção humana. Desta maneira, o pior resultado possível é $AUC = 0,5$, o que corresponde a afirmar que os modelos avaliados se comportam tão bem quanto mapas gerados aleatoriamentes. Em mapas deste tipo, cada pixel é escolhido de forma aleatória como sendo saliente ou não. Podemos encontrar na literatura diversos modelos com desempenho bem melhor que o de mapas aleatórios. No entanto, até o presente momento, o modelo de saliência para imagens com o melhor desempenho é o BMS [35] com $AUC = 0,8257$ [13].

Finalmente, uma propriedade interessante da AUC é que ela é invariante a transformações na forma de uma função monótona crescente aplicadas aos mapas de saliência [18]. Com isso, não precisamos nos preocupar com o efeito da normalização dos mapas sobre o desempenho, por exemplo.

I.2 NSS

Apresentamos a descrição desta métrica para ilustrar como poderia ser alternativamente tratada a questão da avaliação de um modelo de atenção visual. Como já foi indicado, a sigla significa *Normalized Scanpath Saliency* [51] e seu algoritmo consiste em primeiramente normalizar os mapas de saliência originados pelo modelo para que os mesmos tenham média zero e desvio padrão unitário. Em seguida, para cada posição (x_h, y_h) fixada na imagem (dada pelos experimentos subjetivos), o índice NSS para um mapa de saliência S é dado pela seguinte equação:

$$NSS = \frac{1}{\sigma_S} - (S(x_h, y_h) - \mu_S) \tag{I.1}$$

na qual μ_S é a média de S e σ_S é seu desvio padrão.

Assim, um índice $NSS = 1$ significa que as fixações reais, obtidas nos experimentos, estão em regiões dos mapas a um desvio padrão acima da média, indicando que o modelo de saliência escolhido é um ótimo preditor das fixações dos olhos humanos. Em oposição, se $NSS \leq 0$, então o modelo analisado se comporta exatamente como mapas gerados aleatoriamente.

Diferentemente da AUC, a NSS não é invariante a transformações aplicadas nos mapas [18]. Além disso, a necessidade de se ter a posição das fixações reais a cada instante de tempo (em oposição a uma mapa de fixações para cada imagem ou quadro de vídeo observados nos experimentos) torna a aplicação da NSS mais complicada que a da AUC.

II. ESTATÍSTICAS USADAS PARA AVALIAR AS MÉTRICAS DE QUALIDADE

Correlações são importantes ferramentas estatísticas para analisar a dependência de duas variáveis entre si. Neste trabalho duas medidas de correlação foram importantes para definir quão bem as métricas objetivas de qualidade podem estimar a qualidade percebida por seres humanos: os coeficientes de correlação de Pearson e de Spearman.

II.1 Correlação de Pearson

A correlação de Pearson, também conhecida como coeficiente de correlação produto-momento de Pearson, é uma medida de quão linearmente correlacionadas estão duas variáveis, X e Y. Esta medida leva o nome do matemático inglês que a desenvolveu, Karl Pearson (1857-1936).

Em sua definição, o índice de correlação de Pearson, r, é a covariância entre duas variáveis X e Y , dividida pelo produto dos desvios padrões dessas mesmas variáveis e é dado pela seguinte expressão:

$$r = \frac{\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i)(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i)^2}\sqrt{\sum_{i=1}^{n}(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i)^2}}, \tag{II.1}$$

na qual n é o tamanho das amostras X e Y.

A primeira propriedade importante da correlação de Pearson é que seu valor está contido no intervalo de números reais entre 1 e -1. Se r é igual a qualquer um destes extremos, a interpretação é que as variáveis X e Y são perfeitamente linearmente correlacionadas. O valor negativo indica apenas que enquanto X cresce, Y decresce. Devido a essa simetria, temos que um valor $r = 0$ implica não haver nenhuma correlação linear entre as váriáveis.

Outra propriedade importante desta medida de correlação é o fato da mesma ser invariante a translações e escalonamento. Ou seja, caso uma transformação linear seja aplicada a alguma das variáveis(X ou Y), o índice de correlação permanece o mesmo que no caso de nenhuma transformação linear ser aplicada. Este fato foi importante neste trabalho pois permitiu que os resultados das métricas de qualidade fossem diretamente correlacionados com os valores MOS das bases de dados sem que nenhuma adaptação prévia necessitasse ser realizada.

Por fim cabe notar que, da mesma forma que outras estatísticas popularmente utilizadas, a correlação de Pearson não é robusta à presença de *outliers*. Além disso, por evidenciar apenas relações lineares entre variáveis, esta medida não consegue identificar variáveis altamente correlacionadas de forma não-linear.

Na literatura, abreviações comuns desta correlação são PPMCC, PCC e r de Pearson.

II.2 Correlação de Spearman

A correlação de Spearman foi desenvolvida no começo do século XX pelo psicólogo Charles Spearman (1863-1945). É uma versão ranqueada da correlação de Pearson e mede quão bem a relação entre duas variáveis, X e Y, pode ser descrita como uma função monótona.

Para a apresentação da fórmula desta correlação consideremos o seguinte: sejam duas variáveis, X e Y, cada uma com tamanho n. Os valores em X e Y são rearranjados do menor para o maior e, em seguida, convertidos para o *rank* que os mesmos representam dentro dos vetores. Esses ranks são armazenados nos novos vetores x e y. Em outras palavras, o menor valor em X, corresponderá a um valor em x igual a 1 e, em contrapartida, o maior valor de em X, corresponderá a um valor em x igual a n. Valores repetidos em X e Y recebem como *rank* a média dos índices das posições que os mesmos ocupam nos vetores já reorganizados. Feito isso, o índice de correlação ρ de Spearman é definido como:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1)}, \tag{II.2}$$

na qual $(x_i - y_i)$ é feito mantendo-se a combinação relativa de pares de valores original entre X e Y.

Assim como no caso da correlação de Pearson, o coeficiente de Spearman pode apresentar valores reais entre -1 e 1. O sinal de ρ indica apenas a direção da tendência de crescimento de uma variável com relação à outra e um sinal negativo indica que Y tende a decrescer quando X cresce. Um sinal positivo indica a relação oposta. Finalmente, $\rho = 0$ indica que Y não tende nem a crescer ou a decrescer quando X cresce, uma marca da falta de correlação entre as duas variáveis. Correlação perfeita no contexto da medida de Spearman ($\rho = -1$ ou $\rho = 1$) indica que Y é uma perfeita função monótona de X.

Diferentemente da correlação de Pearson, a correlação de Spearman não se limita a identificar dependência linear entre duas variáveis, mas sim a identificar o tipo de dependência descrita por uma função monótona. À excessão disso, a correlação de Spearman mantém as outras propriedades descritas acima para a medida de Pearson.

Na literatura, a correlação de Spearman também é comumente identificada como SCC ou ρ de Spearman.