SILVIO RICARDO RODRIGUES SANCHES

AVALIAÇÃO OBJETIVA DE QUALIDADE DE SEGMENTAÇÃO

São Paulo
2013

SILVIO RICARDO RODRIGUES SANCHES

AVALIAÇÃO OBJETIVA DE QUALIDADE DE SEGMENTAÇÃO

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Doutor em Engenharia Elétrica.

São Paulo
2013

SILVIO RICARDO RODRIGUES SANCHES

AVALIAÇÃO OBJETIVA DE QUALIDADE DE SEGMENTAÇÃO

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Doutor em Engenharia Elétrica.

Área de concentração:
Sistemas Digitais

Orientador:
Prof. Livre-Docente Romero Tori

São Paulo
2013

FICHA CATALOGRÁFICA

Sanches, Silvio Ricardo Rodrigues
Avaliação objetiva de qualidade de segmentação / S.R.R. Sanches. – São Paulo, 2013.
120 p.

Tese (Doutorado) – Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1. Computação gráfica. I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II. t.

AGRADECIMENTOS

Ao Prof. Romero Tori, pela orientação deste trabalho e pela confiança depositada.

Ao Prof. Valdinei Silva, pela disponibilidade, atenção dispensada e paciência.

À minha família e a todos que colaboraram, direta ou indiretamente, na elaboração deste trabalho: Cléber, Makoto, Ana Claudia, Ricardo, Juliana Salioni, Juliana Souza, Mariza, Daniel Lemezenski, Daniel Calife, Lucas Trias, João Bernardes, Cilene, Bruninho, Lima, Alexandre Tomoyose, Fernando Obana, Eunice, Fabio Carmo, Fabio Picchi, Missae, Pedro Câmara e Mayra.

Este trabalho foi realizado com o auxílio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por meio da concessão de Bolsa de Doutorado.

“Não sabendo que era impossível, foi lá e fez”

(Jean Cocteau/Mark Twain)

RESUMO

A avaliação de qualidade de segmentação de vídeos tem se mostrado um problema pouco investigado no meio científico. Apesar disso, estudos recentes na área resultaram em algumas métricas que têm como finalidade avaliar objetivamente a qualidade da segmentação produzida pelos algoritmos. Tais métricas consideram as diferentes formas em que os erros ocorrem (fatores perceptuais) e seus parâmetros são ajustados de acordo com a aplicação em que se pretende utilizar os vídeos segmentados. Neste trabalho apresentam-se: i) uma avaliação da métrica que representa o estado-da-arte, demonstrando que seu desempenho varia de acordo com o algoritmo; ii) um método subjetivo para avaliação de qualidade de segmentação; e iii) uma nova métrica perceptual objetiva, derivada do método subjetivo aqui proposto, capaz de encontrar o melhor ajuste dos parâmetros de dois algoritmos de segmentação encontrados na literatura, quando os vídeos por eles segmentados são utilizados na composição de cenas em ambientes de Teleconferência Imersiva.

Palavras-chave: Avaliação de Segmentação. Avaliação Objetiva. Métrica Objetiva. Métrica Perceptual. Qualidade de Segmentação. Avaliação Subjetiva.

ABSTRACT

Assessment of video segmentation quality is a problem seldom investigated by the scientific community. Nevertheless, recent studies presented some objective metrics to evaluate algorithms. Such metrics consider different ways in which segmentation errors occur (perceptual factors) and its parameters are adjusted according to the application for which the segmented frames are intended. In this work: i) we demonstrate empirically that the performance of existing metrics changes according to the segmentation algorithm; ii) we developed a subjective method to evaluate segmentation quality; and iii) we contribute with a new objective metric derived on the basis of experiments from subjective method in order to adjust the parameters of two bilayer segmentation algorithms found in the literature when these algorithms are used for compose scenes in Immersive Teleconference environments.

Keywords: Segmentation Evaluation. Objective Assessment. Objective Metric. Perceptual Metric. Segmentation Quality. Subjective Assessment

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Segmentar uma imagem, em tempo real, com o objetivo de extrair uma pessoa em primeiro plano de seu contexto original passou a ser uma tarefa comum em sistemas de Realidade Aumentada (RA) (SANCHES et al., 2012b). Essa tarefa torna-se mais problemática quando a aplicação de RA exige que a extração do elemento de interesse seja realizada a partir de vídeo monocular, capturado em ambientes com plano de fundo arbitrário e sem controle de iluminação (KOYAMA; KITAHARA; OHTA, 2003; NAKAMURA et al., 2010; SANCHES et al., 2012a).

Embora pesquisas recentes apresentem algoritmos que atuam nessas condições (CRIMINISI et al., 2006; YIN et al., 2011; PAROLIN et al., 2011; SANCHES; SILVA; TORI, 2012), os resultados apresentados em suas avaliações mostram que tais algoritmos são mais propensos a erros (CRIMINISI et al., 2006; YIN et al., 2011) que os baseados em cor de fundo homogênea. Apesar dessas limitações, alguns sistemas de videoconferência (HARRISON; HUDSON, 2008), *videochats* (CRIMINISI et al., 2006; YIN et al., 2011) e jogos imersivos (NAKAMURA et al., 2010) têm utilizado esses algoritmos, que também passaram a ser implementados em sistemas de RA (NAKAMURA et al., 2010; SANCHES et al., 2012a).

Nessas aplicações, a imagem resultante (quadro de vídeo), que deveria conter apenas o elemento de interesse, pode apresentar-se com erros de classificação de pixels. A utilização dessas imagens para construir uma nova cena pode prejudicar sua qualidade, sobretudo se os erros exibidos causarem grande incômodo ao usuário. Por esse motivo, tais erros devem ser identificados.

A melhor maneira de identificar os erros que causam maior incômodo ao usuário e, por consequência, medir a qualidade da segmentação produzida por um algoritmo é por meio de experimentos formais subjetivos, em que imagens resultantes do processo de segmentação são avaliadas por usuários (GELASCA; EBRAHIMI, 2009). Os grandes problemas associados a esses tipos de experimento são a necessidade de recrutar pessoas que ava-

liem os vídeos e a preparação do ambiente em que os testes são aplicados (a aplicação do experimento que requer alguma infraestrutura) (GELASCA; EBRAHIMI, 2009).

Existem aplicações em que avaliar a qualidade da segmentação é um procedimento que deve ser realizado com certa frequência. Exemplo disso é a necessidade de alguns sistemas de encontrar o melhor conjunto de parâmetros para que determinado algoritmo produza melhores resultados. Os algoritmos de segmentação de vídeos, inclusive os mais simplificados, são parametrizados e a escolha desses parâmetros é fundamental para a eficiência do algoritmo com relação à qualidade da segmentação obtida.

Um conjunto de parâmetros ideais deve fazer com que o algoritmo não produza os tipos de erros que causam maior incômodo ao usuário ou, pelo menos, que evite grande ocorrência desses erros. Nesses casos, seria estabelecida uma forma de predizer o que os usuários diriam a respeito da qualidade da segmentação, sem que se realize qualquer experimento subjetivo a cada conjunto de parâmetros testados. Em outras palavras, uma forma objetiva que considere a percepção do usuário para avaliar a qualidade da segmentação faz-se necessária.

Um exemplo de aplicação em que uma métrica perceptual objetiva de avaliação de qualidade pode ser utilizada são os sistemas de RA voltados a Teleconferência Imersiva (OGI et al., 2001; CORRÊA et al., 2011). Nesses sistemas, a segmentação – em alguns casos realizada a partir de imagens com fundo arbitrário – é necessária para isolar o elemento de interesse que será utilizado na geração dos avatares baseados em vídeo (vídeo-avatares[1]) que são, posteriormente, inseridos em um ambiente virtual.

1.1 Objetivos e Visão Geral da Abordagem adotada

O objetivo da presente pesquisa consiste no desenvolvimento de uma métrica objetiva capaz de avaliar a qualidade da segmentação produzida por determinado algoritmo, considerando o impacto aos usuários das diferentes formas em que os erros ocorrem. A métrica possibilita que se encontre, de forma automática, o melhor ajuste dos parâmetros de de-

[1]Segundo Ogi et al. (2001), um vídeo-avatar é uma imagem tridimensional sintetizada por computador, gerada a partir de vídeo capturado em tempo real. Outras definições, no entanto, o definem como uma representação virtual – não necessariamente tridimensional – baseada na imagem de um usuário humano, obtida por meio de um dispositivo de aquisição de vídeo e atualizada em tempo real (NAKAMURA, 2008).

terminado algoritmo, quando utilizado em um domínio específico de aplicação, os sistemas de RA voltados a Teleconferência Imersiva.

Entre as encontradas na literatura, as métricas analisadas nesta pesquisa não produzem resultados que refletem os obtidos em avaliações subjetivas. Isso se deve, possivelmente, à abordagem adotada em seu desenvolvimento. No trabalho de Gelasca e Ebrahimi (2009), que representa o estado-da-arte na área, os experimentos subjetivos que serviram de base para a geração da métrica objetiva foram realizados de forma que aos avaliadores foram apresentados vídeos que continham erros de segmentação gerados artificialmente, que podem nunca serem exibidos em cenas geradas nas aplicações.

A partir do método subjetivo proposto naquele trabalho desenvolveram-se uma métrica objetiva que, segundo os autores, pode avaliar a qualidade da segmentação, independentemente do algoritmo utilizado. Aqui foi demonstrado que além de não se mostrar eficiente na avaliação da qualidade da segmentação aplicada em sistemas de Teleconferência Imersiva, os resultados apresentados naquela métrica variam conforme o algoritmo de segmentação utilizado.

A hipótese levantada neste trabalho é que uma métrica objetiva deve ser dependente tanto da aplicação quanto do algoritmo de segmentação. Além disso, essa métrica deve ser derivada de experimentos subjetivos em que os vídeos submetidos aos avaliadores sejam gerados a partir de camadas de primeiro plano obtidas de resultados da execução de algoritmos de segmentação. Desse modo, como esses vídeos exibem erros tipicamente encontrados nas aplicações, uma métrica mais eficiente pode ser obtida da análise dos resultados dessas avaliações.

A partir dessa intuição, um método subjetivo com as características citadas acima foi desenvolvido. A métrica objetiva resultante deste trabalho foi derivada de experimentos subjetivos realizados conforme o método subjetivo proposto.

1.2 Organização do Texto

Para que favoreça seu pleno entendimento, este trabalho está organizado da forma que segue. O capítulo 2 mostra uma visão geral sobre segmentação de vídeos em duas camadas e uma classificação das abordagens mais adotadas no desenvolvimento de algoritmos. No

capítulo 3, abordam-se os principais problemas relacionados à qualidade de segmentação. Para identificar tais problemas, os métodos de segmentação que representam o estado-da-arte na área foram analisados. Ainda no capítulo 3, são apresentados as principais métricas para avaliação de qualidade de segmentação encontradas na literatura e as principais aplicações em que essas métricas podem ser utilizadas. Entre essas aplicações, o sistema de Teleconferência Imersiva em que os resultados desta pesquisa foram aplicados é discutido em detalhes.

O capítulo 4 trata dos métodos empregados no desenvolvimento dos experimentos subjetivos cujos resultados, que serviram de base para o desenvolvimento da métrica objetiva, permitiram a identificação dos erros de segmentação mais perceptíveis. Tipos de erros foram caracterizados e o conjunto de vídeos utilizados no experimento simularam todas essas formas de erros. Os detalhes da aplicação do método formal subjetivo, que utilizou esses vídeos, foram discutidos nesse mesmo capítulo.

No capítulo 5, exibe-se a análise dos resultados obtidos dos experimentos a partir dos quais se derivou a métrica objetiva. Primeiramente, as limitações da métrica apresentada por Gelasca e Ebrahimi (2009), quando aplicada em ambientes de Teleconferência Imersiva, foram expostas. Em seguida, os resultados das avaliações subjetivas foram analisados, identificando os erros mais perceptíveis e gerando uma nova métrica objetiva capaz de avaliar a qualidade dos algoritmos de segmentação. Finalmente, reservou-se o capítulo 6 para as conclusões do presente estudo, e para apresentar as perspectivas de trabalhos futuros.

2 SEGMENTAÇÃO DE VÍDEOS EM DUAS CAMADAS

A segmentação (extração de elementos de interesse em imagens) aplicada em composições de cenas é um problema que tem sido alvo de pesquisas desde o início do século passado (WILLIAMS, 1918). Produções de cinema e televisão, que até o final da década de 1970 eram apoiadas em tecnologia analógica (FOSTER, 2010), tradicionalmente utilizam métodos que permitem isolar elementos de uma imagem, que na maioria das vezes são pessoas em primeiro plano, com o objetivo de gerar cenas a partir da combinação desses elementos com novos planos de fundo (VLAHOS, 1963, 1964, 1978).

Os algoritmos mais tradicionais de segmentação (em duas camadas) de imagens ou vídeos partem do princípio de que a captura do vídeo se realiza em ambientes controlados, com fundos de cor única – normalmente azul ou verde – e iluminação devidamente direcionada para que a tonalidade do fundo se mantenha constante (VLAHOS, 1978; MISHIMA, 1994; GIBBS et al., 1998). De forma simplificada, tais algoritmos procuram isolar o elemento de interesse por meio da eliminação da cor do fundo, que é conhecida do sistema.

A partir da década de 1980, novos algoritmos começaram a surgir baseados nos recursos da tecnologia digital (GIBBS et al., 1998) e, mais recentemente, algoritmos capazes de extrair elementos de interesse não apenas em tempo real, mas a partir de imagens com planos de fundo arbitrários também passaram a ser desenvolvidos (BERGEN et al., 1992; SUN et al., 2006; CRIMINISI et al., 2006; YIN et al., 2011; WANG et al., 2010; SANCHES; SILVA; TORI, 2012).

Essa possibilidade impulsionou pesquisas com foco em outras áreas de aplicação a utilizarem imagens segmentadas (ou camadas de imagens que possuem apenas elementos de interesse da cena), principalmente as voltadas para aplicações em que os elementos de interesse são pessoas posicionadas em primeiro plano na cena. Exemplos a serem citados são os sistemas de videoconferências tradicionais (HARRISON; HUDSON, 2008), *vi-*

deochats (SUN et al., 2006; CRIMINISI et al., 2006; YIN et al., 2011; SANCHES; SILVA; TORI, 2012), jogos imersivos (NAKAMURA et al., 2010) e aplicações de Realidade Aumentada (SANCHES et al., 2012a).

A revisão bibliográfica apresentada neste capítulo tem como objetivo expor o estado-da-arte na forma de uma classificação dos algoritmos mais utilizados, capazes de extrair um elemento de interesse, em tempo real, a partir de uma sequência de imagens obtidas em ambiente não controlado.

Para isso, foram analisados os trabalhos encontrados na literatura que são voltados ao desenvolvimento de algoritmos cuja finalidade é a divisão de uma imagem de entrada em duas camadas (*bilayer*): primeiro plano (que contém o elemento de interesse) e plano de fundo, para posterior substituição do plano de fundo original. A representação de uma camada de primeiro plano resultante de um processo de segmentação em duas camadas pode ser visualizada na figura 1.

Figura 1 – Representação de uma camada de primeiro plano (SANCHES et al., 2012). O elemento de interesse é extraído do plano de fundo original tornando transparentes os pixels que pertencem ao fundo e mantendo opacos os que pertencem ao elemento de interesse

Algoritmos utilizados em aplicações que segmentam múltiplos elementos de interesse,

como compressão de vídeo (WU; CHEN, 2001), ou nas que não têm como objetivo a segmentação para substituição do fundo da cena, como identificação de pessoas para sistemas de segurança (NAM; HAN, 2006), reconhecimento de gestos (MITRA; ACHARYA, 2007; BERNARDES-JUNIOR; NAKAMURA; R.TORI, 2011) e rastreamento (YILMAZ; JAVED; SHAH, 2006) não foram analisados.

Nesses casos, embora muitas abordagens sejam aplicáveis, a precisão na separação do elemento de interesse do seu fundo original pode não ser um requisito tão rígido quanto nos algoritmos utilizados em aplicações de substituição de fundo. Aplicações desse tipo exigem algoritmos de segmentação mais precisos, para que proporcionem qualidade na combinação com o novo fundo.

Em aplicações em que a extração do elemento de interesse se realiza em ambientes controlados, os erros de classificação de pixels podem ser evitados por meio da intervenção do usuário. O direcionamento manual de luzes e a distribuição dos elementos da cena, por exemplo, podem ser adequados, para que a cor do fundo se mantenha constante, impedindo a ocorrência de sombras, reflexos ou ruídos sobre o fundo.

Em ambientes não controlados, por sua vez, o plano de fundo é arbitrário e qualquer situação que atrapalhe a segmentação deve ser tratada pelo algoritmo, evitando intervenções do usuário para modificar o ambiente. Nesses casos, como não existe o conhecimento prévio da cor do fundo, outras informações, que podem ser obtidas da sequência de imagens, passam a ser fundamentais para que um elemento de interesse seja isolado. Algumas abordagens utilizam, ainda, equipamentos específicos, ou mais de um dispositivo para obter novas informações que auxiliem a segmentação.

Grande parte dos algoritmos computacionalmente eficientes para execução em tempo real trabalha com essas informações, na forma de um conjunto de "cortes" (CRIMINISI et al., 2006). Cor, contraste, movimento e estéreo são exemplos de cortes muito utilizados. Esses cortes combinam-se probabilisticamente e aplicam-se na imagem por meio de algum arcabouço de minimização de energia, como o mostrado na seção I.2 do apêndice I. Alguns algoritmos mais simplificados, no entanto, utilizam essas informações (ou apenas uma delas) de formas alternativas.

O fato de determinadas abordagens se apoiarem em dispositivos específicos, ou de exigir calibração de mais de um dispositivo, pode restringir sua aplicabilidade. Em apli-

cações executadas em ambientes domésticos, como *videochats*, por exemplo, imagina-se que a maioria dos participantes possuam computadores e câmeras de vídeo convencionais. Essa observação sugere que uma classificação das abordagens considere dois grupos principais: as abordagens executadas a partir de captura realizada por câmeras monoculares (convencionais) e as que necessitam de entrada binocular ou de equipamento específico (não convencionais) para produzir informações que auxiliem a segmentação.

Abordagens apoiadas em vídeo monocular podem ser divididas em dois subgrupos, cujos algoritmos são classificados de acordo com a técnica adotada. São eles: subtração de fundo e arcabouço de minimização de energia. Ainda que muitos algoritmos utilizem mais de uma técnica, uma delas, normalmente, tem maior importância que as demais no processo – i.e., sua utilização isolada resulta na rotulação correta da maioria dos pixels da imagem. A classificação sugerida neste trabalho toma como base a técnica principal utilizada pelo algoritmo.

As abordagens apoiadas em equipamentos específicos utilizam esse tipo de recurso para geração de mapas de profundidade da cena. Desse modo, a distância de cada pixel em relação a um sensor acoplado à câmera constitui-se na principal informação a ser utilizada no processo de segmentação. Algoritmos que pertencem a esse grupo podem ser divididos também em dois subgrupos: os baseados em estéreo e os baseados em sensores. Nas subseções seguintes, será apresentada uma visão geral dos trabalhos que representam o estado-da-arte em segmentação em tempo real de sequência de imagens em ambientes não controlados, considerando a classificação descrita. A figura 2 exibe um diagrama que sintetiza tal classificação.

2.1 Algoritmos que utilizam Vídeo Monocular

Grande parte dos algoritmos analisados são capazes de realizar a segmentação a partir de uma imagem, capturada por uma câmera de vídeo convencional (vídeo monocular). Tais algoritmos podem ser divididos em dois subgrupos, classificados de acordo com sua abordagem principal.

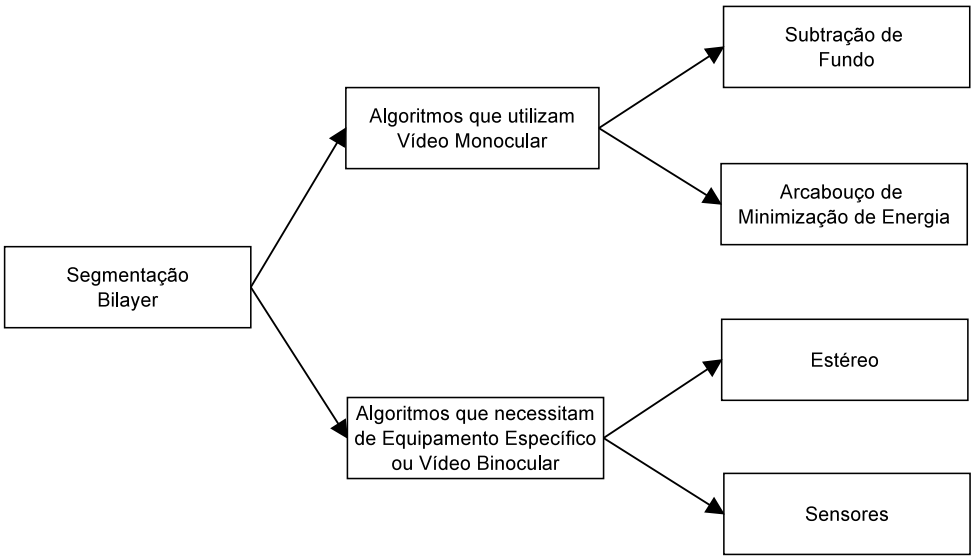


Figura 2 – Classificação dos algoritmos de segmentação de vídeos em duas camadas que atuam em ambientes não controlados (SANCHES et al., 2012). A utilização ou não de equipamento específico é o principal critério para o agrupamento dos algoritmos, seguido da técnica principal adotada como base

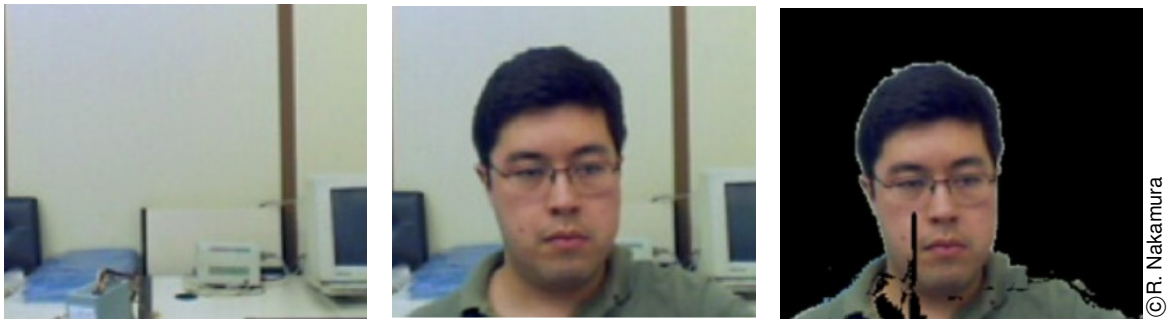
2.1.1 Subtração de Fundo

A abordagem da subtração de fundo (PICCARDI, 2004) consiste, basicamente, na comparação do quadro de vídeo no tempo atual (figura 3(b)) com uma imagem que representa um modelo do fundo (figura 3(a)). Como mostrado na figura 3(c), a camada de primeiro plano é gerada com base nos pixels não coincidentes dessas duas imagens. Esses pixels pertencerão ao elemento de interesse.

Algoritmos mais simplificados calculam a diferença do quadro atual e do anterior com base em um *threshold* (FRIEDMAN; RUSSELL, 1997; PICCARDI, 2004) ou calculam um modelo do plano de fundo por meio da média ou da mediana de alguns quadros anteriores (CUCCHIARA et al., 2003). Outros utilizam ainda informações do quadro atual, considerando também uma taxa de aprendizado (PICCARDI, 2004). Esses algoritmos, que fazem parte de métodos denominados básicos (PICCARDI, 2004), apoiam-se na história recente dos pixels e não estabelecem quaisquer correlações espaciais entre pixels vizinhos.

Algoritmos mais sofisticados, por sua vez, utilizam, por exemplo, misturas de modelos gaussianos de cores (TANG; MIAO; WAN, 2007), estimadores de densidade de *kernel* (ELGAMMAL; HARWOOD; DAVIS, 2000), estimadores de *Mean-Shift* (HAN; COMANICIU; DAVIS, 2004) ou decomposição da imagem em autoespaços (*Eigenbackground*) (OLIVER; ROSA-

RIO; PENTLAND, 2000). Desse modo, obtêm-se métodos capazes de lidar com planos de fundo que apresentam maiores variações (FRIEDMAN; RUSSELL, 1997).



(a) (b) (c)

Figura 3 – Abordagem da Subtração de Fundo (NAKAMURA, 2008). Em (a) é mostrado um modelo do fundo e em (b) um quadro de vídeo no tempo atual. Os pixels não coincidentes nas duas imagens fazem parte do elemento de interesse (c). Cores semelhantes no fundo e no elemento de interesse podem provocar erros de classificação, como mostrado em (c)

Na utilização de algoritmos baseados em subtração de fundo, a maior dificuldade encontrada não se relaciona à diferenciação dos pixels em si, mas na construção automática de um modelo do fundo e na manutenção deste modelo, que é comparado quadro a quadro com a imagem atual (TOYAMA et al., 1999). Essa abordagem, apesar de ter sua aplicação voltada tradicionalmente aos sistemas de segurança (NAM; HAN, 2006), também é utilizada em métodos que funcionam como base para aplicações de substituição de fundo (QIAN; SEZAN, 1999; KIM; AHN; KIM, 2004; HARRISON; HUDSON, 2008).

2.1.2 Arcabouço de Minimização de Energia

Muitos algoritmos que podem ser utilizados para segmentação em ambientes não controlados têm como característica comum a busca por informações que permitem mapear a movimentação do elemento de interesse por meio de arcabouços de minimização de energia (seção I.2 do apêndice I). No entanto, uma das técnicas mais aplicadas para identificação de elementos em movimento em uma sequência de imagens – o cálculo do fluxo óptico (*optical flow*) (BARRON; FLEET; BEAUCHEMIN, 1994) – é normalmente evitada devido ao seu custo computacional (CRIMINISI et al., 2006) e a impossibilidade de representar o elemento de interesse como um modelo rígido, dado que este, em grande parte das aplicações, é uma pessoa em primeiro plano (YIN et al., 2007).

Algoritmos de segmentação apresentados em trabalhos recentes identificam pixels em movimento utilizando informações de cor, aliadas a observação da coerência temporal da sequência de imagens (CRIMINISI et al., 2006; YIN et al., 2007; PAROLIN et al., 2011). Processos de aprendizado *offline*, baseados em "*ground-truths*" (na figura 15 é exibido um exemplo de *ground-truth*), também são recursos utilizados por métodos desenvolvidos a partir dessa abordagem. Obtém-se, desse modo, as probabilidades de cada pixel da imagem pertencer ao fundo ou ao elemento de interesse. Tais valores são combinados pelo modelo (CRIMINISI et al., 2006; YIN et al., 2007; SANCHES; SILVA; TORI, 2012) utilizando arcabouços de minimização de energia. Esse tipo de abordagem – predominante em métodos capazes de segmentar imagens monoculares – é detalhada na seção I.2.

Alguns algoritmos assumem que o plano de fundo seja estático e necessitam de inicialização na forma de um "plano de fundo limpo" (SUN et al., 2006; HARRISON; HUDSON, 2008), para reduzir erros de classificação provocados por regiões de alto contraste no plano de fundo. Em alguns trabalhos, as características do movimento são combinadas com informações a respeito da forma do elemento de interesse, para modelar correlações espaciais (YIN et al., 2007, 2011). Desse modo, pode-se classificar regiões da imagem pouco texturizadas, ou onde não houve movimentação (pixels dessas regiões não podem ser classificados com base apenas em informações de movimento, como mostrado na figura 4).

Figura 4 – Exemplo de mapa de movimentação de pixels. As regiões mais claras da imagem correspondem às bordas em movimento, ao passo que as áreas mais escuras são bordas estacionárias. As regiões intermediárias representam áreas não texturizadas, que permanecem ambíguas (CRIMINISI et al., 2006)

2.2 Algoritmos que necessitam de Equipamento Específico ou Vídeo Binocular

Muitas abordagens utilizadas para extração do elemento de interesse a partir de uma sequência de imagens apoiam-se em equipamentos específicos, considerados não convencionais, ou utilizam mais de um equipamento (calibrados), com o objetivo de obter novas informações que auxiliem a segmentação. Apesar de sua utilização estar restrita a algumas aplicações atualmente, o desenvolvimento desses métodos também se justifica pela possibilidade desses equipamentos tornarem-se convencionais no futuro.

Dois tipos de abordagens são comumente utilizadas para preencher esses mapas: estéreo e as baseadas em sensores. Ambas utilizam esses equipamentos com a finalidade de estimar mapas de profundidade da cena. Um mapa de profundidade é uma matriz, de tamanho correspondente ao da imagem, que contém a distância de cada pixel em relação à câmera. Na figura 5, mostra-se um mapa de profundidade da cena, que pode ser utilizado para auxiliar a segmentação.

(a) (b)

Figura 5 – Mapa de Profundidade obtido por meio de sensor do tipo TOF (*Time-of-Flight*) (IDDAN; YAHAV, 2001). Em (a) e (b) são mostrados o quadro de vídeo e o mapa de profundidade do mesmo quadro, respectivamente. Os pixels mais claros representam os mais próximos da câmera de vídeo (e do sensor) ao passo que os mais escuros são os mais distantes (SANCHES et al., 2012)

2.2.1 Estéreo

Uma das formas de estimar mapas de profundidade para resolver problemas de segmentação é por meio da utilização de algoritmos de estéreo (OHTA; KANADE, 1985; COX et al., 1996). A técnica do estéreo exige que dois vídeos sincronizados sejam utilizados como

entrada. O principal desafio em abordagens desse tipo é a localização dos pixels correspondentes (SCHARSTEIN; SZELISKI, 2002) nas imagens esquerda e direita, para que a profundidade de cada pixel possa ser calculada por meio de um processo de triangulação (OHTA; KANADE, 1985). Uma estratégia adotada para encontrar correspondência nas imagens estéreo é determinar a linha epipolar, cujo processo se mostra na figura 6.

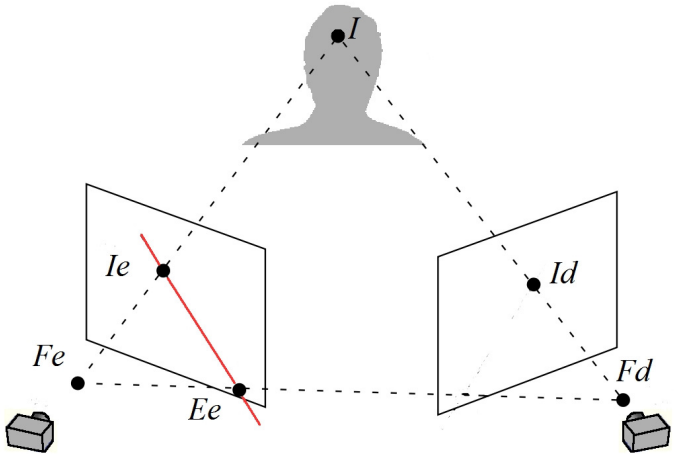


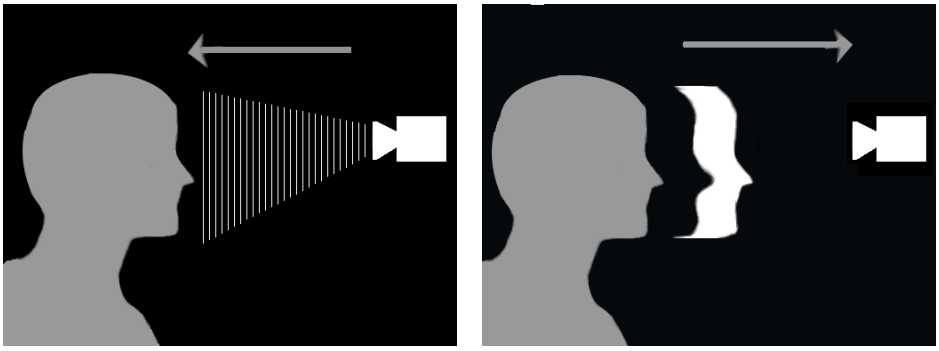
Figura 6 – Determinação da linha epipolar (SANCHES et al., 2012). Um ponto I, pertencente ao elemento de interesse é observado por duas câmeras com seus respectivos pontos focais Fe e Fd. A projeção de I sobre os planos das imagens direita e esquerda são Ie e Id. A reta $IeEe$ representa a linha epipolar. O espaço de busca aos pontos correspondentes da imagem direita passa a ser restrito a essa reta. Como os pontos Ie e Id, e suas projeções são conhecidos, a distância do ponto I pode ser calculada por um processo de triangulação (OHTA; KANADE, 1985)

Apesar de utilizarem a distância dos pixels, obtida por meio de estéreo como informação principal, alguns trabalhos aplicam também cortes de cor e contraste (KOLMOGOROV et al., 2005a, 2005b, 2006) para evitar erros de classificação, principalmente nas bordas. Outros utilizam técnicas de reconhecimento de faces (LAW; SCLAROFF, 2005), para obter a localização do elemento de interesse e desconsiderar regiões dele distantes, tornando a segmentação mais robusta.

2.2.2 Sensores

TOF (*Time-of-Flight* – Tempo de Voo) são sensores ativos que utilizam laser para medir as distâncias entre o próprio sensor e os objetos da cena (BIANCHI et al., 2009) (figura 7). Essas distâncias são utilizadas para preencher mapas de profundidades densos, utilizados por algoritmos de segmentação.

Basicamente, esses sensores utilizam luz pulsada (IDDAN; YAHAV, 2001; GVILI et al., 2003) ou luz modulada (GOKTURK; YALCIN; BAMJI, 2004). No primeiro caso, uma onda de luz constante acerta os elementos da cena e a propagação de fótons de alta frequência mede o tempo de retorno do pulso de luz. No segundo caso, a luz emitida é modulada e o TOF é medido pela detecção do atraso da fase.



(a) (b)

Figura 7 – Funcionamento do sensor TOF (SANCHES et al., 2012). (a) Gera-se uma "parede de luz", que se desloca ao longo do campo visão da câmera. Essa parede pode ser gerada, por exemplo, como um pulso de laser de curta duração, com um campo de iluminação igual ao campo de visão da câmera. (b) Quando a parede de luz atinge os objetos na cena, ela é refletida de volta para a câmera, carregando uma impressão dos objetos (GVILI et al., 2003)

Equipamentos comerciais (GEISS, 2010), que utilizam outros tipos de sensores, como os baseados em técnicas que utilizam luz estruturada (SCHARSTEIN; SZELISKI, 2003) para a aquisição de mapas de profundidade, também têm sido utilizados para resolver problemas de segmentação em aplicações com substituição de fundo.

Algoritmos que pertencem a esse grupo, além da informação de profundidade (IDDAN; YAHAV, 2001), trabalham com cortes de cor e contraste (WANG et al., 2010) ou atuam em conjunto com algoritmos de rastreamento (BLEIWEISS; WERMAN, 2009), para alcançar resultados robustos. Definir um *threshold* simples, com base na distância do pixel não é suficiente, pois os valores de profundidade obtidos, na maioria das vezes, não são precisos a ponto de alcançar qualidade aceitável para aplicações com substituição de fundo (WANG et al., 2010).

3 QUALIDADE DE SEGMENTAÇÃO

A classificação apresentada na seção 2 mostra que os algoritmos de segmentação podem ser agrupados, primeiramente, de acordo com o equipamento necessário para sua execução e, em seguida, pela abordagem adotada como ponto de partida para o seu desenvolvimento. Abordagens que se apoiam em informações adquiridas por equipamento específico (ou exigem do usuário algum tipo de calibração de equipamentos) produzem algoritmos que apresentam, atualmente, os resultados mais robustos (YIN et al., 2011). No entanto, existe um volume considerável de pesquisas que buscam resultados semelhantes, utilizando apenas equipamento convencional (vídeo monocular) para aquisição do vídeo de entrada, com o objetivo de produzir algoritmos que possam ser utilizados em número maior de aplicações.

Os algoritmos de segmentação existentes, independentemente de abordagem ou equipamento que necessitam, apesar de sofisticados, ainda não são precisos o suficiente para serem considerados uma solução geral para o problema. Várias situações que podem ocorrer durante a captura do vídeo fazem com que erros de segmentação ocasionalmente ocorram durante a execução da aplicação. Neste capítulo são discutidas em detalhes tais situações, aqui chamadas “situações-problema”, que são consideradas fontes potenciais de falhas na tarefa de segmentação.

Uma vez que as aplicações que necessitam de segmentação podem exibir ao usuário cenas geradas a partir de segmentação imperfeita, torna-se necessário encontrar formas de avaliar a qualidade dessas cenas. Segundo Gelasca e Ebrahimi (2009), a busca por uma métrica de qualidade de segmentação deve ser considerada um problema mal colocado, pois para uma mesma imagem (ou vídeo) o resultado ótimo pode ser diferente, dependendo da aplicação em que a imagem segmentada é utilizada.

Do ponto de vista do usuário, um quadro de vídeo pode ser considerado com menos

qualidade que outro, ainda que ambos possuam o mesmo percentual de erros. Em outras palavras, a forma em que os erros se apresentam na imagem segmentada deve ser considerada quando se avalia um algoritmo de segmentação. Uma medida de qualidade que permita descobrir o impacto desses erros aos usuários pode ser utilizada não apenas para determinar a aplicabilidade de determinado algoritmo, mas para auxiliar a escolha do mais adequado ou encontrar o melhor ajuste de seus parâmetros.

Ainda que os resultados obtidos de estudos sobre avaliação de segmentação historicamente não tenham recebido no meio científico a mesma atenção que as pesquisas voltadas ao desenvolvimento de novos algoritmos de segmentação (ZHANG, 1996), existe um número razoável de trabalhos que buscam soluções para o problema. Neste capítulo, essas pesquisas são apresentadas.

3.1 Principais Fontes de Erros de Segmentação

O objetivo dos algoritmos de segmentação, que dividem cada quadro de vídeo em duas camadas, que atuam em ambientes não controlados consiste na extração do elemento de interesse, sem que seja necessária a intervenção do usuário no ambiente onde a captura do vídeo se realiza. Isso significa que, além da dificuldade implícita de identificar o elemento a ser isolado em uma cena arbitrária, o algoritmo implementado no método deve tratar todas as situações desfavoráveis que podem ocorrer durante a execução de uma aplicação.

Variações na iluminação, pessoas que atravessam o fundo da cena, ou a movimentação da câmera que captura o vídeo são situações comuns em ambientes não controlados. Ocorrências desse tipo são exemplos de situações desfavoráveis, que dificultam a identificação de um elemento de interesse dentro de uma sequência de imagens.

Algumas dessas situações, no entanto, podem se tornar um problema, quando se aplica determinada abordagem. Por outro lado, essa mesma situação é contornada implicitamente por algoritmos apoiados em outra. Um exemplo disso é a situação em que uma pessoa atravessa o fundo da cena. Apesar de não representar um problema para algoritmos que utilizam mapas de profundidade – pois estes baseiam-se em informação de profundidade dos pixels –, é uma ocorrência difícil de ser contornada pelos que utili-

zam informações de movimentação do elemento de interesse como meio de identificá-lo. Nesse caso, os pixels em movimento no fundo serão considerados como pertencente ao elemento de interesse, caso nenhum tratamento adicional seja incorporado ao algoritmo.

Na tabela 1, registram-se as situações-problema e suas possíveis causas, tais como identificadas nos trabalhos analisados nesta revisão revisão bibliográfica, independentemente dos algoritmos que afetam. Nas tabelas 2 e 3, mostram-se as formas com que cada uma dessas situações-problema afeta os algoritmos de segmentação apoiados em equipamentos não convencionais e convencionais, respectivamente. O símbolo "–" (traço), na tabela, indica que a abordagem não é afetada pela situação-problema.

Importa ressaltar que se destacaram os problemas que podem ocorrer durante a execução da aplicação. A abordagem estéreo, por exemplo, exige um trabalhoso processo de calibração de duas (ou mais) câmeras (CRIMINISI et al., 2006) que antecede sua aplicação. Considera-se, neste levantamento, que tais dispositivos estejam devidamente calibrados. Do mesmo modo, considera-se também que o problema da sincronização do sensor TOF com a câmera de vídeo (BIANCHI et al., 2009) esteja resolvido. A utilização de câmeras tanto binoculares quanto com sensores TOF pré-calibradas, que podem ser encontradas no mercado, evitam problemas de calibração.

Posto que a segmentação voltada a aplicações em ambientes não controlados represente um desafio aos pesquisadores da área, para alguns dos problemas levantados há soluções eficientes. Por outro lado, várias são as situações-problema cujos erros provocados apenas se minimizam.

Entre as abordagens apoiadas em vídeo monocular, os algoritmos baseados em subtração de fundo que utilizam informações espaciais, por exemplo, vieram para solucionar muitos dos problemas que ocorriam em algoritmos mais simplificados, normalmente baseados apenas em *thresholds*. Variações na iluminação, por exemplo, desde que ocorram dentro de determinados níveis, podem ser contornadas por esses algoritmos (HARRISON; HUDSON, 2008). Quando essas variações ocorrem de forma brusca, no entanto, o problema é de difícil tratamento.

A ocorrência de grande movimentação do fundo não pode ser tratada por algoritmos puros de subtração de fundo. Nesse caso, outras informações obtidas da imagem são necessárias. Quando existe pequena movimentação, os erros de classificação provocados

Tabela 1 – Problemas que podem ocorrer quando se utilizam algoritmos de segmentação de vídeo que atuam, em tempo real, em ambientes não controlados e suas possíveis causas

Problema	Possíveis Causas
Variações na iluminação	O acender ou o apagar de lâmpadas em um escritório (SUN et al., 2006), movimentação de pessoas próxima à câmera que podem provocar sombras ou acionar o ajuste automático de branco da câmera (SUN et al., 2006; CRIMINISI et al., 2006).
Movimentação no fundo	Movimentos de cortinas, provocados por rajadas de vento (SUN et al., 2006). Movimento de nuvens, ondas do mar, galhos e folhas de árvores (PICCARDI, 2004). Objetos ou pessoas distantes que atravessam a cena (YIN et al., 2007, 2011). Objetos que se movem até a cena e depois deixam de se movimentar ou objetos presentes na cena se afastam e revelam novas partes do fundo (PICCARDI, 2004; SUN et al., 2006).
Elemento de interesse estático	Indivíduo em primeiro plano permanece imóvel em frente a câmera (YIN et al., 2011).
Grande movimentação do elemento de interesse	O elemento de interesse se movimenta além do campo de visão da câmera (BARRON; FLEET; BEAUCHEMIN, 1994; YIN et al., 2007).
Oscilações da câmera	Tremulação da câmera acoplada em um computador móvel, posicionado no colo do usuário (SUN et al., 2006; CRIMINISI et al., 2006; YIN et al., 2007, 2011).
Cores semelhantes no fundo e no elemento de interesse	Existência de objetos no plano de fundo que possuem a mesma tonalidade de parte do vestuário da pessoa em primeiro plano (figura 3(c)).
Regiões pouco texturizadas ou homogêneas	Imagens saturadas ou presença de elementos como paredes brancas e partes do céu (KOLMOGOROV et al., 2005a, 2006; CRIMINISI et al., 2006).
Intensidade da luz do ambiente	Presença de superfícies reflexivas ou de vários sensores no ambiente (KOLB; BARTH; KOCH, 2008).

são em pequeno número e os mesmos podem ser preenchidos (quando ocorrem no elemento de interesse), ou removidos (quando ocorrem no fundo) aplicando-se operadores morfológicos (LI, 2005). Em alguns casos, o problema do elemento de interesse estático, o que dificulta a construção automática do modelo do fundo, tem sido contornado por meio da inicialização do sistema com uma imagem “limpa” do plano de fundo – ou seja, excluindo-se o elemento de interesse (HARRISON; HUDSON, 2008), ou pela captação de um conjunto de imagens do fundo (KIM; AHN; KIM, 2004; LI, 2005). Alguns algoritmos baseados em subtração de fundo tratam as oscilações na câmera por meio da utilização de um modelo de plano de fundo estendido, que contém regiões além do tamanho da janela do vídeo. Para contornar o problema das cores semelhantes no fundo e no elemento de interesse, faz-se necessário o processamento em conjunto com outras técnicas.

Tabela 2 – Formas como as situações-problema podem afetar cada abordagem em um processo de segmentação baseado em equipamento convencional

Situação/Problema	Equipamento Convencional	
	Minimização de Energia	**Subtração de Fundo**
Variações na iluminação	A mudança de cor de um pixel, devido a variação da iluminação, pode ser confundida com a movimentação do elemento de interesse (CRIMINISI et al., 2006).	Podem tornar as cores do quadro atual bastante diferente das do modelo do fundo (SUN et al., 2006).
Movimentação no fundo	O objeto ou pessoa que atravessa o fundo da cena será considerado elemento de interesse (CRIMINISI et al., 2006; YIN et al., 2011).	O objeto ou pessoa que atravessa o fundo da cena será considerado elemento de interesse (CRIMINISI et al., 2006; SUN et al., 2006; YIN et al., 2007, 2011).
Elemento de interesse estático	Impossibilita a classificação dos pixels, dado que não existe movimentação na cena (utilizando apenas informação de movimentação de pixel) (YIN et al., 2011).	Pode impossibilitar a geração do modelo do fundo (PICCARDI, 2004) (quando não existe inicialização na forma de um “plano de fundo limpo”).
Grande movimentação do elemento de interesse	–	–
Oscilações da câmera	Impede a diferenciação do que é movimento do elemento de interesse e do que são alterações de cores provocadas pela movimentação da câmera.	Faz com que a imagem de referência não represente o plano de fundo nos quadros em que a posição da câmera é diferente da inicial (PICCARDI, 2004).
Cores semelhantes no fundo e no elemento de interesse	Dificulta a classificação das regiões das bordas, devido a ausência de contraste (SUN et al., 2006) (o contraste é um informação utilizada para identificação de pixels em movimento).	Pode fazer com que os pixels do fundo sejam confundidos com os do elemento de interesse, provocando erros de classificação (SUN et al., 2006).
Regiões pouco texturizadas ou homogêneas	–	–
Intensidade da luz do ambiente	–	–

Com respeito aos algoritmos baseados em arcabouços de minimização de energia, o problema da grande movimentação do elemento de interesse não foi inserido na tabela 2, dado que as soluções mais recentes não se apoiam em cálculos do fluxo óptico (CRIMINISI et al., 2006; YIN et al., 2007, 2011), em que tal situação representa um problema. A utilização de outras informações, como cor e contraste, tem sido a solução para que regiões pouco

Tabela 3 – Formas como as situações-problema podem afetar cada abordagem em um processo de segmentação baseado em equipamento não convencional

Situação/Problema	Equipamento não Convencional	
	Estéreo	**Sensores TOF**
Variações na Iluminação	–	–
Movimentação no Fundo	–	–
Elemento de interesse estático	–	–
Grande movimentação do elemento de interesse	Pode causar oclusão estéreo (GEIGER; LADENDORF; YUILLE, 1995; KOLMOGOROV et al., 2005a, 2006) (o elemento de interesse não fica visível em uma das câmeras, impossibilitando a localização de pixels correspondentes nas imagens).	O elemento de interesse pode se mover além do limite de emissão de luz em TOF (BIANCHI et al., 2009), impossibilitando o calculo de profundidade.
Oscilações da câmera	Pode afetar a calibração das câmeras (caso uma delas seja movimentada). Não representa um problema quando se utiliza um equipamento pré-calibrado.	–
Cores semelhantes no fundo e no elemento de interesse	–	–
Regiões pouco texturizadas ou homogêneas	Impede a identificação dos pixels correspondentes nas duas imagens de entrada, o que é a tarefa essencial para esse tipo de abordagem (KOLMOGOROV et al., 2005a, 2006; CRIMINISI et al., 2006).	–
Intensidade da luz do ambiente	–	Pode provocar múltiplas reflexões, causando interferências nos sinais de retorno, que são utilizados para o cálculo dos valores de profundidade dos pixels (KOLB; BARTH; KOCH, 2008).

texturizadas, ou homogêneas, possam ser classificadas, quando o elemento de interesse é estacionário (CRIMINISI et al., 2006). Oscilações na câmera e variações na iluminação são problemas que têm sido minimizados, utilizando-se informações obtidas da coerência temporal do vídeo, e com o auxílio de treinamento *offline* (KOLMOGOROV et al., 2005a; CRIMINISI et al., 2006; SUN et al., 2006). Em alguns casos, essas informações combinam-se com filtros de forma (SHOTTON et al., 2006), para estimar a geometria do elemento de interesse e mini-

mizar os problemas ocasionados por movimentação no fundo, além de reduzir ainda mais os provocados por oscilações na câmera e variações na iluminação (YIN et al., 2007, 2011).

A utilização do modelo de movimentação em conjunto com outras técnicas, como a de subtração de fundo, pode evitar o problema das cores semelhantes no fundo e no elemento de interesse (SUN et al., 2006). Entre as abordagens apoiadas em equipamento específico, ou em vídeo binocular, os algoritmos de estéreo, em que a informação de profundidade é a única, não tratam o problema da oclusão estéreo. Em alguns algoritmos mais sofisticados, adotam-se informações de cor, contraste e a coerência espacial entre os quadros, para evitar o problema (KOLMOGOROV et al., 2005a, 2006). A inclusão dessas informações evita o problema da impossibilidade de classificação nas regiões pouco texturizadas.

A aplicação conjunta de algoritmos de identificação de faces também é uma solução para os problemas das regiões poucos texturizadas, ou de oscilações da câmera (LAW; SCLAROFF, 2005). Em algoritmos que necessitam de informações de sensores, o problema das múltiplas reflexões não foi abordado em nenhum dos trabalhos analisados.

3.2 Avaliação de Qualidade de Segmentação

A avaliação de qualidade de segmentação é um problema que tem sido investigado em diferentes contextos da literatura, entre eles, na avaliação de imagens compostas a partir de objetos – em alguns casos, pessoas – extraídos do conteúdo de um vídeo (GELASCA; EBRAHIMI, 2009). Uma vez identificadas na seção 3.1 as principais fontes causadoras de erros, nesta seção são apresentadas as formas de medir a qualidade dos algoritmos.

Segundo Zhang (1996), nos métodos[1] de avaliação podem ser identificadas duas abordagens: analítica, que avalia o algoritmo, e empírica, que analisa os resultados da execução do algoritmo. A segunda tem sido a mais praticada.

Imagens resultantes dos processos de segmentação e de composição com um novo plano de fundo têm sido avaliadas de forma subjetiva ou objetiva (GELASCA; EBRAHIMI, 2009). As avaliações subjetivas têm se mostrado a forma mais eficiente de obter medi-

[1] Alguns trabalhos, sobretudo os escritos em língua inglesa, utilizam o termo “métrica” para designar uma forma objetiva de obter uma medida de qualidade de segmentação ao passo que outros utilizam o termo método. Nesta revisão bibliográfica foi adotado o mesmo termo utilizado pelos autores do trabalho original.

ções confiáveis (PÉCHARD; PÉPION; CALLET, 2008) tanto na indústria como na comunidade científica. Métodos subjetivos, que são utilizados tradicionalmente em avaliações de qualidade de codificadores de vídeo para transmissões de TV, foram, no decorrer dos anos, sendo adaptados para que pudessem ser utilizados em avaliações de imagens exibidas em aplicações multimídia, inclusive em avaliações de segmentação (SANCHES et al., 2012b). Alguns desses métodos são, inclusive, diretamente aplicados em processos de avaliação de qualidade da segmentação em que os objetos do vídeo são pessoas (SANCHES et al., 2012b).

O grande problema das avaliações subjetivas é o fato de, em geral, requererem um grande número de observadores e alguma infraestrutura para realização das avaliações. Isso torna o processo demorado e, algumas vezes, caro. Avaliar subjetivamente de forma sistemática é um procedimento que deve ser evitado (GELASCA; EBRAHIMI, 2009).

Alguns métodos objetivos para avaliação de qualidade de segmentação podem ser encontrados na literatura. O trabalho de Erdem e Sankur (ERDEM; SANKUR, 2000), por exemplo, propõe um método de avaliação de qualidade de segmentação baseado na penalidade de pixels classificados de forma incorreta, considerando erros em relação à forma e ao movimento do objeto de vídeo segmentado. Atributos semelhantes foram também considerados no trabalho proposto por Mech e Marqués (2002).

Apesar da relevância dos trabalhos anteriores, o grande impulso para pesquisas da área está diretamente relacionado com a tentativa de padronização do formato ISO/MPEG-4. Devido à capacidade do padrão de codificar formas de objetos de vídeo independentemente, pesquisas voltadas a avaliação dessas imagens segmentadas tornaram-se necessárias. O método desenvolvido a partir desses estudos considera a acurácia espacial, que se refere à quantidade de pixels classificados de forma incorreta tanto no elemento de interesse quanto no plano de fundo (falsos negativos e falsos positivos) e a coerência temporal, que se refere à diferença da acurácia espacial entre o resultado da segmentação do quadro atual e do quadro anterior (WOLLBORN; MECH, 1997; VILLEGAS; MARICHAL; SALCEDO, 1999). Outros métodos, que são refinamentos do modelo original, também foram propostos pelo mesmo grupo de pesquisa (MARICHAL; VILLEGAS, 2000; VILLEGAS; MARICHAL, 2004).

Em Correia e Pereira (2003), a segmentação de objetos individuais – o trabalho trata

também da avaliação de segmentação com a presença de múltiplos objetos no conteúdo de um vídeo – também é avaliada com base em critérios espaciais e temporais. Foram utilizados como critérios temporais: a fidelidade da forma, em que são considerados a quantidade de pixels segmentados incorretamente e suas distâncias em relação a borda do elemento de interesse; a fidelidade geométrica, em que se considera o tamanho, a posição e uma combinação do alongamento e compacidade do elemento segmentado; a similaridade de conteúdo da borda, obtida por meio de filtros combinados com informações espaciais derivadas de experimentos subjetivos; e a similaridade de dados estatísticos, que está relacionada com brilho e vermelhidão do objeto segmentado.

Os critérios temporais adotados representam uma informação perceptual temporal e uma medida de criticidade, que utiliza informações espaciais e temporais simultaneamente (CORREIA; PEREIRA, 2003). O método apresentado, segundo os autores, é limitado no que se refere a objetos com semântica complexa, como é o caso de uma pessoa.

No trabalho de Gelasca e Ebrahimi (2009), com o objetivo de encontrar uma medida objetiva, testes subjetivos foram propostos para analisar artefatos[2] que simulavam erros espaciais e temporais. A métrica objetiva desenvolvida, chamada *Perceptual Spatio-Temporal* (PST), considera artefatos produzidos no processo de avaliação que causam incômodo maior ao usuário. Quatro artefatos foram caracterizados (esses artefatos serão detalhados na seção 3.2.2) e combinados para produzir uma medida geral de incômodo.

Entre os métodos mencionados nos parágrafos anteriores, dois deles, considerados métricas, são discutidos em detalhes nas subseções seguintes. A primeira, que é resultado de pesquisas do grupo ISO/MPEG-4, é considerada popular devido a sua simplicidade. No entanto sua principal aplicação consiste na segmentação voltada ao problema da compressão de vídeos. A segunda, que pode ser considerada pertencente ao estado-da-arte na área, considera a percepção humana em sua solução e tem como uma de suas aplicações em potencial os sistemas de RA (objeto de estudo desta pesquisa).

[2]No trabalho de Gelasca e Ebrahimi (2009), define-se um artefato como uma característica perceptual relativamente pura de um erro de segmentação que foi criado artificialmente. Neste trabalho, ainda que os tipos de erros definidos não tenham sido inseridos de forma controlada nos vídeos de teste, a mesma palavra será utilizada para representá-los.

3.2.1 Métrica de Avaliação de Qualidade de Segmentação do Padrão MPEG

A forma de avaliação de qualidade de vídeo definido pelo grupo do ISO/MPEG-4 (VILLEGAS; MARICHAL, 2004) tornou-se bastante popular na comunidade científica devido, principalmente, à sua simplicidade. A métrica, que consiste de um refinamento de pesquisas anteriores realizadas pelo mesmo grupo (WOLLBORN; MECH, 1997), baseia-se em dois critérios objetivos: a precisão espacial e a coerência temporal.

Utilizando a formalização de Gelasca (2005), uma região i no quadro k pode ser definida como um conjunto de pixels $\mathcal{R}_i(k)$ com as seguintes propriedades: i) $\mathcal{R}_i(k)$ é espacialmente conectada; ii) $\mathcal{R}_i(k) \cup \mathcal{R}_j(k)$ é desconectado $\forall i \neq j$. $\mathcal{R}(k)$ indica o conjunto de todos os elementos de interesse que fazem parte da segmentação ótima, e pode ser expressa conforme apresentada na equação

$$\mathcal{R}(k) = \bigcup_{0 \leqslant j < J} \mathcal{R}_j(k) \tag{1}$$

onde J é o número de elementos de interesse presentes no quadro[3]. O conjunto de pixels segmentados no quadro k, $\mathcal{C}(k)$ é a união dos i elementos de interesse dado pela equação $\mathcal{C}_i(k)$

$$\mathcal{C}(k) = \bigcup_{0 \leqslant j < I} \mathcal{C}_i(k) \tag{2}$$

onde I é o número de elementos de interesse. O conjunto de falsos positivos $\mathcal{P}(k)$ em que os elementos não pertencem a segmentação perfeita pode ser representado por $\mathcal{P}(k) = \mathcal{C}(k) \cap \mathcal{R}'(k)$ onde $\mathcal{R}'(k)$ é o complemento de $\mathcal{R}(k)$. Do mesmo modo, os falsos negativos $\mathcal{N}(k)$ podem ser representados por $\mathcal{N}(k) = \mathcal{C}'(k) \cap \mathcal{R}(k)$.

Como forma de obter precisão espacial, Villegas e Marichal (2004) definiram que os pixels classificados de forma incorreta pertencem a uma entre duas classes: falsos positivos ou falsos negativos, cada qual com diferentes pesos associados. O método diferencia o impacto dessas duas classes na precisão espacial quando avalia a distância d do pixel até o contorno do elemento de interesse. A precisão espacial $qms(k)$ é normalizada pela

[3]Nesta pesquisa, embora o elemento de interesse trate-se de uma pessoa em primeiro plano, pode haver regiões disjuntas nesse elemento. Um exemplo é a situação em que podem estar visíveis o rosto e parte do tronco em uma única região. No entanto, a mão aparece visível e desconectada em consequência do braço permanecer fora do campo de visão da câmera.

soma das áreas dos elementos (obtidas de um *ground truth*) de acordo com a equação

$$qms(k) = \frac{qms^{+}(k) + qms^{-}(k)}{\sum_{i=1}^{N_R} R_i(k)} = \frac{\sum_{d=1}^{D_M^+} w_{+}(d) \cdot |\mathcal{P}_d(k)| + \sum_{d=1}^{D_M^-} w_{-}(d) \cdot |\mathcal{N}_d(k)|}{\sum_{i=1}^{N_R} R_i(k)} \quad (3)$$

onde D_M^+ e D_M^- são as maiores distâncias d dos falsos positivos e falsos negativos, respectivamente. N_R é o número total de regiões disjuntas do elemento de interesse no *ground truth* R. $\sum_{i=1}^{N_R} R_i(k)$ é a soma da área de todos os i elementos. $w_{+}(d)$ e $w_{-}(d)$ são os pesos dos pixels falsos positivos e falsos negativos, dados por

$$w_{+}(d) = b_1 + \frac{b_2}{d + b_3}, w_{-}(d) = f_S \cdot d \quad (4)$$

onde os parâmetros $b_1 = 20$, $b_2 = -178,125$, $b_3 = 9.375$ e $f_S = 2$ são escolhidos empiricamente, com base na análise de vários resultados, e, segundo os autores parecem concordar com uma visão subjetiva (VILLEGAS; MARICHAL, 2004). Essas funções mostram que os pesos dos falsos negativos aumentam linearmente e são maiores que os pesos dos falsos positivos se estiverem na mesma distância da borda do elemento de interesse.

Dois critérios são utilizados para estimar a coerência temporal, a estabilidade qmt e a direção qmd. A estabilidade temporal é obtida pela soma normalizada das diferenças da precisão espacial dos falsos positivos e dos falsos negativos em dois quadros consecutivos

$$qmt(k) = \frac{|qms^{+}(k) - qms^{+}(k-1)| + |qms^{-}(k) - qms^{-}(k-1)|}{\sum_{i=1}^{N_R} R_i(k)}. \quad (5)$$

Em seguida, é calculado, entre quadros consecutivos, o deslocamento do centro de gravidade $\overrightarrow{G}$ do elemento resultante da segmentação em relação à referência, com o objetivo de estimar possíveis direções $\overrightarrow{qmd}(k)$ na trajetória do objeto

$$\overrightarrow{qmd}(k) = \left[\overrightarrow{G}_E(k) - \overrightarrow{G}_R(k)\right] - \left[\overrightarrow{G}_E(k-1) - \overrightarrow{G}_R(k-1)\right] \quad (6)$$

que representa o deslocamento, do tempo $(k-1)$ para o tempo (k), dos centros de gravidade $\overrightarrow{G}$ das máscaras estimadas E e da referência R. A direção consiste na norma do vetor de deslocamento normalizado pela soma das áreas dos elementos (*bounding box*)

$$qmd(k) = \frac{||\overrightarrow{qmd}(k)||}{\frac{1}{N_R}\sum_{i=1}^{N_R} BB_i^{x,y}(k)} \quad (7)$$

onde $BB_i^{x,y}(k)$ são as dimensões horizontal e vertical do *bounding box* que representa a área do objeto i da referência R no tempo k. A métrica wqm é obtida pela combinação linear das três medidas apresentadas, conforme a equação

$$wqm(k) = w_1 \cdot qms(k) + w_2 \cdot qmt(k) + w_3 \cdot qmd(k), \;\; wqm = \frac{1}{K}\sum_k wqm(k) \qquad (8)$$

onde os pesos w são dependentes da aplicação.

3.2.2 Métrica de Avaliação de Qualidade de Segmentação PST

Entre os trabalhos relacionados à avaliação de segmentação encontrados na literatura, a métrica proposta por Gelasca e Ebrahimi (2009) consiste na única pesquisa entre as encontradas na revisão bibliográfica aqui realizada que considera a segmentação no contexto dos sistemas de RA. Naquele trabalho, os autores propõem um método formal para realizar experimentos psicofísicos voltados à avaliação subjetiva de segmentação e constroem uma métrica objetiva perceptual com base em experimentos realizados a partir do método subjetivo proposto.

Uma vez gerada a métrica objetiva que, segundo os autores, é capaz de avaliar a qualidade da segmentação de forma geral (fora do contexto de uma aplicação específica), novos experimentos foram realizados para encontrar o melhor ajuste de seus parâmetros quando utilizados em diferentes domínios de aplicação, inclusive sistemas de RA[4].

O método subjetivo proposto, com pequenas variações, foi aplicado em todos os experimentos realizados no trabalho. Os vídeos de teste (que continham erros de segmentação), que eram exibidos aos avaliadores, foram gerados a partir de artefatos sintéticos, para que os erros pudessem ser facilmente descritos. O método consiste basicamente de 5 (cinco) passos:

1. Instruções Orais: têm como objetivo familiarizar os participantes com o ambiente do experimento, com a tarefa de avaliação a ser realizada e com as sequências de vídeo originais (ou vídeos-fonte) utilizadas nos experimentos;

[4]No trabalho de Gelasca (2005), os sistemas de RA considerados na pesquisa têm as características do apresentado em Marichal et al. (2002).

2. Treinamento: são exibidas as sequências originais, um vídeo de referência (sem erros de segmentação) e sequências de vídeos que contenham artefatos em grande quantidade. O objetivo é que o avaliador tenha noção dos limites inferior e superior em relação ao nível de incômodo provocado por um artefato ou conjunto de artefatos;

3. Avaliação preliminar: as avaliações subjetivas são realizadas em um subconjunto dos vídeos que contêm os artefatos;

4. Avaliação Subjetiva: os testes subjetivos são efetivamente executados (utilizando a base de vídeos completa[5]); e

5. Entrevista: Após os experimentos subjetivos, as percepções dos participantes a respeito dos artefatos são levantadas por meio de entrevistas.

As avaliações, segundo Gelasca e Ebrahimi (2009), devem ser realizadas de acordo com as recomendações da *International Telecommunications Union*[6] (ITU), descritas em ITU-T (2008), utilizando o método de estímulo único em que as notas são atribuídas em uma escala contínua (0-100). O método descrito nas recomendações ITU-R BT-500 (ITU-R, 2002) foi utilizado para filtrar os dados da avaliação, eliminando resultados discrepantes. Como parte do experimento, os avaliadores eram orientados a executar uma entre duas tarefas: atribuir um valor numérico do incômodo detectado, em relação a uma referência, ou emitir uma opinião sobre o quão “forte” ou visível um conjunto de artefatos se mostra em um vídeo com erros de segmentação.

O método subjetivo proposto foi utilizado para entender como os erros objetivos são percebidos pelas pessoas e, desse modo, gerar uma métrica perceptual objetiva. Os fatores envolvidos na derivação da métrica são mostrados no diagrama da figura 8.

As sequências de teste podem ser entendidas como uma combinação de um *ground truth* e de artefatos. Inicialmente, as regiões classificadas de forma incorreta são identificadas pela sobreposição *ground truth* ao vídeo em análise. Essas regiões são classificadas e quantificadas na forma de tipos de erros objetivos (ou artefatos). A ligação entre o *ground truth* e o bloco “objetivo”, no diagrama, é pontilhada, uma vez que a referência não necessariamente é utilizada para obter os erros objetivos. Esses erros objetivos são, então,

[5]No experimento realizado no trabalho de Gelasca e Ebrahimi (2009), uma bateria de testes continha de 150 a 180 vídeos.

[6]http://www.itu.int

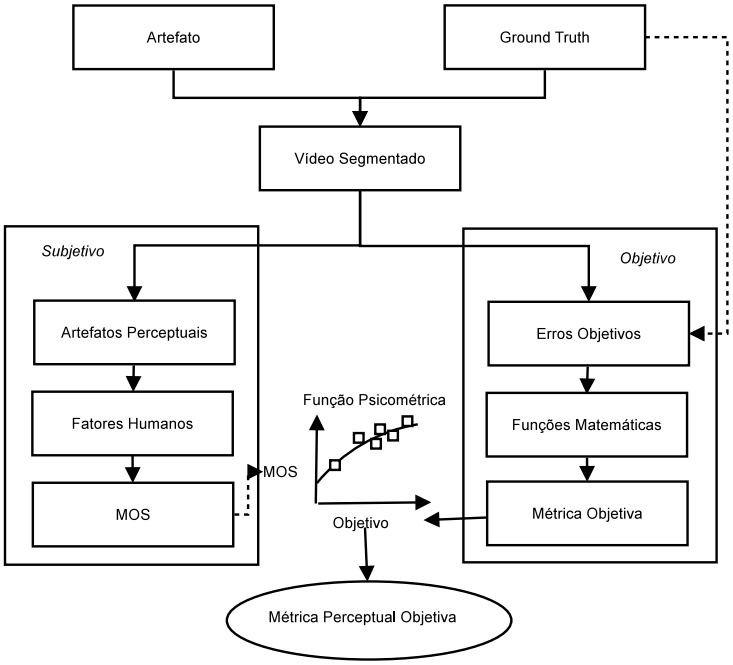


Figura 8 – Fatores envolvidos no processo de geração da métrica objetiva a partir da avaliação subjetiva de qualidade de segmentação de vídeo, adaptado de (GELASCA, 2005)

combinados, por meio de alguma fórmula matemática, para encontrar uma qualidade global que possibilite definir uma métrica objetiva. O objetivo é encontrar funções perceptuais que relacionem as medidas objetivas com a qualidade global da segmentação percebida pelos usuários (*Mean Opinion Score* – MOS).

Como pode ser observado na figura 8, os artefatos presentes no vídeo segmentado em análise tornam-se perceptuais quando analisados pelo sistema visual humano. A relação entre os erros objetivos e os resultados da avaliação subjetiva é feita por meio de uma função psicométrica (MAXWELL; DELANEY, 2003), uma vez que, segundo os autores, são capazes de modelar a percepção humana.

Em relação aos erros de segmentação, quatro artefatos que representam todos os possíveis erros espaciais foram caracterizados: Regiões Adicionadas $\mathcal{A}_r$, que são os erros de classificação ocorridos no plano de fundo que são desconectados do elemento de interesse; Plano de Fundo Adicionado $\mathcal{A}_b$, que são erros no plano de fundo conectados à borda do elemento de interesse; Buracos internos $\mathcal{H}_i$, que são os erros, que ocorrem no elemento de interesse, desconectados da borda; e os Buracos de Borda $\mathcal{H}_b$, que ocorrem quando os erros de classificação no elemento de interesse se mostram conectados a

borda (GELASCA; EBRAHIMI, 2009).

Outro passo importante no desenvolvimento do experimento subjetivo trata da escolha do conjunto de vídeos a serem utilizados nos experimentos. Os vídeos originais foram obtidos de bases vídeos de uso livre para pesquisa[7]. Em seguida, os artefatos foram introduzidos nesses vídeos modificando-se a segmentação ideal obtida com o auxilio do *ground truth*. Para considerar erros espaciais, os artefatos definidos foram estudados variando seu tamanho, posição e forma. O aspecto temporal foi analisado, por exemplo, variando-se a posição de determinado artefato ao longo da execução de uma sequência de teste. Na figura 9, alguns artefatos submetidos aos avaliadores na fase de avaliação subjetiva podem ser visualizados.

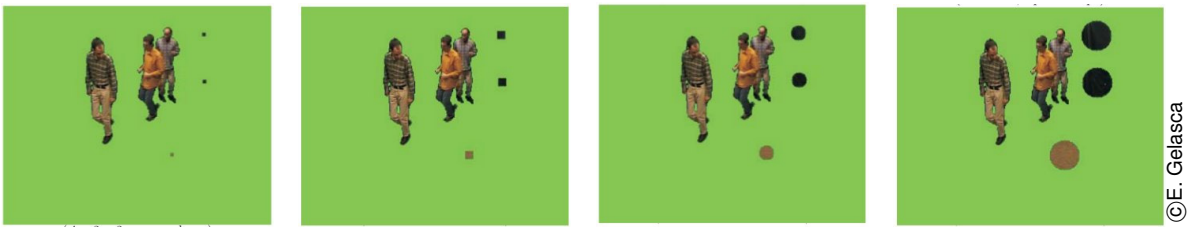


Figura 9 – Exemplos de artefatos submetidos aos avaliadores na fase de avaliação subjetiva cujos resultados foram utilizados na geração da métrica PST (GELASCA, 2005)

Desse modo, a métrica objetiva proposta baseia-se em dois tipos de erros: objetivos e perceptuais. Em relação aos erros objetivos, a partir do artefato $\mathcal{A}_r$, o erro espacial relativo $S_{A_r}(k)$, para todas as regiões j adicionadas no quadro k $\mathcal{A}_r^j(k)$ é dado por

$$S_{A_r}(k) = \frac{\sum_{j=1}^{N_{Ar}} |\mathcal{A}_r^j(k)|}{|n(k)|} \tag{9}$$

onde $|\ .\ |$ é o operador de cardinalidade do conjunto, $n(k)$ é a soma dos pixels da referência e do resultado da segmentação e N_{Ar} é o número total de regiões adicionadas. Do mesmo modo, para os buracos internos j $\mathcal{H}_i^j(k)$, o erro espacial relativo é dado por

$$S_{H_i}(k) = \frac{\sum_{j=1}^{N_{Hi}} |\mathcal{H}_i^j(k)|}{|n(k)|} \tag{10}$$

onde N_{Hi} é o número total de buracos internos no objeto.

Para os demais tipos de erros, que são conectados à borda do elemento de interesse,

[7] http://www.tele.ucl.ac.be/PROJECTS/art.live/artlive.html

um peso D^j também é considerado

$$D^j = 1 + \frac{\overline{d^j} + \sigma_d^j}{d_{max}^j} \tag{11}$$

onde d é a distância em pixels até o contorno do objeto. A média $\overline{d}$ e o desvio padrão σ_d são calculados e, em seguida, normalizados pelo diâmetro máximo d_{max} do conjunto de pixels que formam o erro em que o pixel erroneamente classificado pertence. O diâmetro máximo é calculado pelo máximo das distâncias entre qualquer pixel do conjunto e a borda do elemento de interesse.

Utilizando a equação 11, obtém-se o erro espacial relativo para os erros de borda $S_{A_b}(k)$ para j regiões adicionadas

$$S_{A_b}(k) = \frac{\sum_{j=1}^{N_{Ab}} D_{Ab}.|\mathcal{A}_b^j(k)|}{|n(k)|}. \tag{12}$$

Do mesmo modo, para os buracos de borda j $\mathcal{H}_b^j(k)$, o erro espacial relativo $S_{H_b}(k)$ é dado por

$$S_{H_b}(k) = \frac{\sum_{j=1}^{N_{Hb}} D_{Hb}.|\mathcal{H}_b^j(k)|}{|n(k)|}. \tag{13}$$

Segundo Gelasca (2005), o efeito mais indesejado em relação a qualidade de segmentação está relacionado a variação abrupta dos erros espaciais entre quadros consecutivos, o chamado *flickering*. Uma movimentação não suave de qualquer erro espacial deteriora de forma considerável a qualidade percebida pelo usuário.

Para que este problema fosse considerado, foram calculados erros temporais $F(k)$ para cada tipo de artefato $\Lambda = [\mathcal{A}_r, \mathcal{A}_b, \mathcal{H}_i, \mathcal{H}_b]$, de acordo com a equação

$$F_\Lambda(k) = \frac{||\Lambda(k)| - |\Lambda(k-1)||}{|\Lambda(k)| + |\Lambda(k+1)|} \tag{14}$$

Como mostra a equação 14, quando um artefato desaparece subitamente (efeito surpresa (SENDERS, 1997)), existe uma penalização. Desse modo, para considerar-se a qualidade e a estabilidade dos resultados, o erro relativo espaço-temporal $ST(k)$ é dado por

$$ST_\Lambda(k) = S_\Lambda(k).\frac{1 + F_\Lambda(k)}{2} \tag{15}$$

Outro efeito considerado na métrica é o chamado efeito memória (INAZUMI et al., 1999).

Depois de algum tempo, o ser humano se acostuma com certa qualidade visual, julgando-a mais aceitável se a mesma qualidade persistir determinado tempo. Existe ainda o efeito "expectativa" (INAZUMI et al., 1999), que mostra que uma segmentação de qualidade no início pode criar uma boa impressão a respeito da qualidade geral (ou vice-versa). Esse efeito é modelado de acordo com a equação

$$ST_\Lambda(k) = \frac{1}{k}\sum_{k=1}^{K} w_t(k) ST_\Lambda(k) \tag{16}$$

em que o peso $w_t(k)$, que modela o efeito expectativa, foi definido empiricamente por meio de testes subjetivos (GELASCA, 2005) da forma

$$w_t(k) = (\alpha . e^{\frac{k-30}{\beta}} + \lambda) \tag{17}$$

com $\alpha = 0.02$, $\beta = 7.8$ e $\lambda = 0.0078$. k representa o quadro atual.

Os valores de ST para cada artefato foram plotados juntamente com os valores obtidos da avaliação subjetiva (GELASCA, 2005), para ajustar várias curvas psicométricas (que descrevem a percepção humana a respeito dos erros). A função que melhor se ajustou aos artefatos foi a função de Weibull W. Obteve-se, portanto, quatro métricas perceptuais (PST_Λ)

$$W(x, S, k) = 1 - e^{-(Sx)^k}, \text{onde } x = ST_\Lambda, \; PST_\Lambda = W(ST_\Lambda, S, k) \tag{18}$$

onde os parâmetros S e k foram obtidos por meio de experimentos subjetivos, detalhados em Gelasca e Ebrahimi (2009) e Gelasca (2005).

Finalmente, a métrica perceptual é obtida pela combinação das 4 (quatro) métricas. Uma simples combinação linear, como mostrado em Gelasca (2005), representa o incômodo total, de acordo com a equação

$$PST = a \times PST_{A_r} + b \times PST_{A_b} + c \times PST_{H_i} + d \times PST_{H_b} \tag{19}$$

onde os valores de a, b, c e d foram obtidos por meio de experimento subjetivo, utilizando combinações de artefatos sintéticos. Como a qualidade da segmentação é considerada dependente da aplicação, os melhores valores para esses parâmetros para cada aplicação investigada foram obtidos por meio de uma regressão por mínimos quadráticos linear, utilizando os dados da avaliação subjetiva em que as sequências de teste foram obtidas

da execução de algoritmos de segmentação.

Nas aplicações de RA, apenas o artefato “Regiões adicionadas” foi considerado pouco percebido, uma vez que a atenção dos usuários está voltada ao elemento real presente na cena. Os valores $a = 6.71$, $b = 8.39$, $c = 12.57$ e $d = 8.74$ se mostraram os mais adequados para avaliação de segmentação aplicada aos sistemas de RA.

3.3 Principais Aplicações

Uma vez que as aplicações que se baseiam em algoritmos como os apresentados no capítulo 2 podem construir cenas a partir de elementos de interesse extraídos de forma imperfeita, em todas essas aplicações uma métrica objetiva para avaliação de qualidade de segmentação pode ser utilizada para encontrar o algoritmo mais adequado ou para encontrar o melhor ajuste de seus parâmetros.

Um exemplo clássico são os programas de televisão, exibidos ao vivo, como ocorre nos informativos de previsão do tempo apresentados em telejornais. O fundo original da imagem, que possui um apresentador em primeiro plano, é substituído pelo mapa de determinada região do país (figura 10).

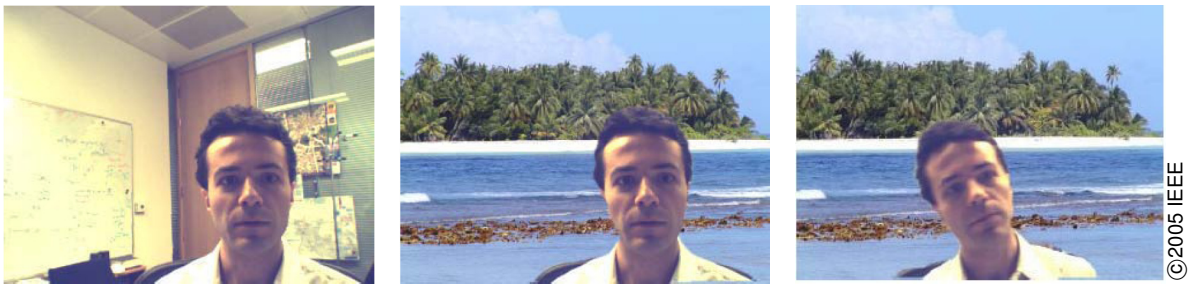
Figura 10 – Substituição de fundo utilizada em telejornais para informar a previsão do tempo (CRIMINISI et al., 2006). O fundo original da cena é substituído por um mapa de determinada região do país. Algoritmos que agem em ambientes não controlados podem ser utilizados nesse tipo de aplicação

Nesse caso, o ambiente normalmente é formado por cor única e a segmentação ocorre utilizando-se de algoritmos que se baseiam na eliminação da cor do fundo (GIBBS et al., 1998). Algoritmos que atuam em ambientes não controlados, no entanto, também podem

ser utilizados nesse tipo de aplicação (IDDAN; YAHAV, 2001; CRIMINISI et al., 2006), possibilitando, inclusive, que a captura do vídeo se realize em ambientes externos (GVILI et al., 2003).

Do mesmo modo, podem-se encontrar pesquisas voltadas a sistemas de videoconferência que realizam a segmentação da imagem dos participantes, com o objetivo de preservar o local da captura do vídeo (KOLMOGOROV et al., 2005a), ou de produzir uma nova imagem com efeitos 3D (HARRISON; HUDSON, 2008). Os *videochats*, que surgem com a popularização das conexões de rede de alta velocidade, também são um grupo de aplicações em potencial. Ao contrário das videoconferências tradicionais, em que os sinais de áudio e vídeo são, em alguns sistemas, transmitidos via satélite, os *videochats* são executados quase sempre em computadores pessoais (inclusive *laptops*), utilizando a Internet como meio de comunicação.

Essas aplicações podem realizar a segmentação e a substituição de fundo como forma de redução de banda ou de obtenção de privacidade (figura 11) (KIM; AHN; KIM, 2004; KOLMOGOROV et al., 2005a; CRIMINISI et al., 2006; SUN et al., 2006; YIN et al., 2007, 2011; PAROLIN et al., 2011; SANCHES; SILVA; TORI, 2012). Mesmo os telefones móveis 3G podem adicionar esse recurso aos seus serviços (WU; BOULANGER; BISCHOF, 2008).



(a) (b) (c)

Figura 11 – Substituição de fundo como forma de obtenção de privacidade em videoconferências (KOLMOGOROV et al., 2005a). O plano de fundo original (a), que é arbitrário, é substituído por uma nova imagem, antes dos quadros de vídeo serem enviados pela rede aos demais participantes (b) e (c)

Além das citadas, tornam-se aplicações em potencial os sistemas de Realidade Aumentada (SANCHES et al., 2012a) e os jogos imersivos (WANG et al., 2006; NAKAMURA et al., 2010), em que, a representação humana no ambiente virtual (avatar) se constrói com base na imagem do usuário. Nesses sistemas, em muitos casos, não se realiza uma simples substituição do fundo da cena. A imagem segmentada pode ser utilizada em modelo bidi-

mensional, aplicada em um modelo geométrico ou volumétrico (OGI et al., 2003), ou pode-se combinar várias camadas de imagem, que contenham o mesmo elemento de interesse em diferentes pontos de vista, para sintetizar um modelo 3D do avatar (MATUSIK et al., 2000). A precisão na separação do elemento de interesse do fundo original, nesses casos, também é fator preponderante.

Esquemas de compressão de vídeo (WU; CHEN, 2001), que segmentam múltiplos elementos de interesse em tempo real para identificar objetos (e pessoas), sistemas que necessitam do rastreamento (YILMAZ; JAVED; SHAH, 2006) de pessoas, como os de segurança, em que indivíduos devem ser detectados e isolados para diminuir a área de atuação de algoritmos de análise de comportamento humano (NAM; HAN, 2006), e sistemas de reconhecimento de gestos (BERNARDES-JUNIOR, 2010), em que as mãos do usuário precisa ser isolada do fundo, são exemplos de aplicações que utilizam métodos de segmentação que atuam em ambientes não controlados.

Nesses casos, no entanto, o objetivo principal não é a segmentação para substituição do fundo da cena (NAM; HAN, 2006). Embora muitas abordagens sejam aplicáveis, a precisão na separação do elemento de interesse do seu fundo original não é um requisito tão rígido quanto nos métodos utilizados em aplicações de substituição de fundo.

Entre as aplicações descritas nesta seção, a presente pesquisa tem como foco medir a qualidade da segmentação em sistemas de RA voltados a Teleconferência Imersiva, para encontrar o melhor ajuste dos parâmetros dos algoritmos de segmentação implementados. Por meio desses algoritmos é que serão extraídos os elementos de interesse utilizados na geração do avatar. O sistema é parte do projeto “Vídeo-Avatar: Augmented Reality Teleconferencing System” (CORRÊA et al., 2011) cujo objetivo principal é o desenvolvimento de um sistema que possa ser utilizado na educação a distância. Suas funcionalidades possibilitam que um instrutor, representado por um vídeo-avatar, interaja com o ambiente virtual e comunique-se com estudantes remotos, que acessam o sistema via Internet, visualizando, em 3D, o ambiente virtual com o vídeo-avatar nele inserido.

O sistema permite, inclusive, que o aluno altere seu ponto de vista, ainda que limitado a determinado ângulo, e ative recursos, como visão estereoscópica. O grande desafio é proporcionar aos participantes a sensação de que o professor esteja realmente presente no ambiente de ensino virtual. Explicações sobre como funciona uma plataforma de petróleo

ou um passeio do professor por um templo histórico, por exemplo, reconstruído virtualmente tornam-se assim mais realistas e envolventes, mesmo que o aluno não faça uma imersão no ambiente.

Em relação aos seus aspectos técnicos, o projeto foi elaborado de forma que as tarefas custosas computacionalmente e que exijam equipamentos mais sofisticados sejam executadas nos módulos de *software* presumidamente localizados em uma instituição de ensino (módulo servidor). Aos estudantes remotos, que executam o módulo cliente, são exigidos apenas equipamentos de baixo custo (*webcams* e computadores pessoais convencionais).

As técnicas de remoção de fundo implementadas como parte do módulo servidor permitem que a imagem do instrutor real (isolada do fundo original) seja utilizada na geração do avatar. Os algoritmos disponíveis, no entanto, partem do princípio de que o ambiente onde o vídeo é capturado possui iluminação constante e fundo de cor única. Dessa forma, o elemento de interesse (instrutor) é extraído, a cena virtual e o avatar são gerados e o quadro de vídeo final é enviado aos alunos. Essa restrição (ambiente de captura previamente preparado) é perfeitamente aceitável, dado que a estruturação do ambiente fica a cargo do instrutor ou da instituição.

Uma possibilidade levantada como evolução do projeto trata da viabilidade de determinado aluno, em algum momento, assumir o papel do professor. Nesse caso, a existência de uma representação do aluno, inserida em um ambiente virtual torna-se necessária. Na figura 12 podem ser visualizados os módulos servidor e cliente do sistema, em que esta nova funcionalidade está prevista. A imagem do aluno é capturada e enviada ao servidor. Em seguida, o quadro de vídeo é segmentado, o avatar é gerado, inserido no ambiente sintético. Após a composição da cena, a nova imagem é distribuída aos demais alunos.

O grande desafio na utilização de vídeo-avatares para representar um aluno conectado ao sistema está relacionado ao problema da segmentação dos quadros de vídeo. Ao contrário do que ocorre com o instrutor, a captura do vídeo desse aluno é normalmente realizada em um ambiente doméstico e utilizando uma câmera de vídeo convencional. A escolha dos melhores parâmetros dos vários algoritmos que atuam em ambientes não controlados implementados no sistema, considerando o aspecto perceptual, é uma tarefa fundamental.

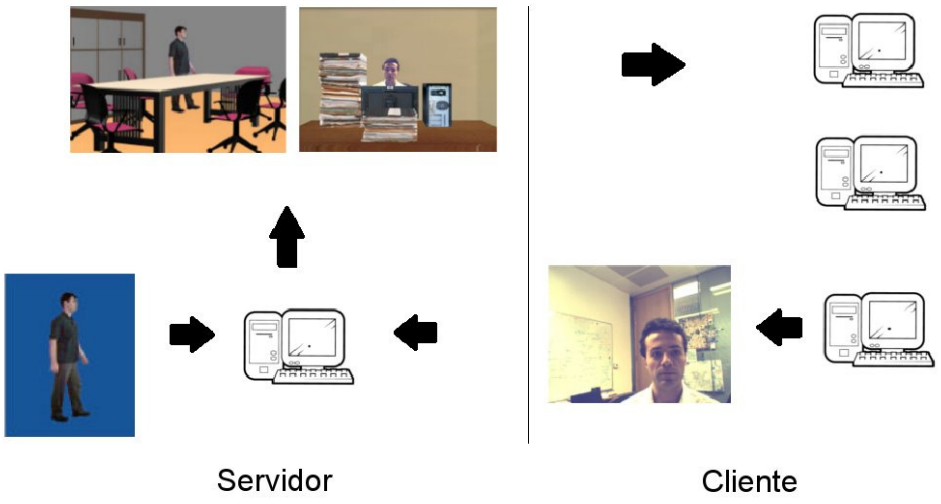


Figura 12 – Módulos do sistema com a funcionalidade em que o aluno pode ser inserido no ambiente de RA. A imagem do aluno é capturada e enviada ao módulo servidor, onde é processada. Após a composição da cena, a nova imagem é distribuída aos demais alunos (SISCOUTTO, 2003; CORRÊA et al., 2011)

4 MÉTODO SUBJETIVO E REALIZAÇÃO DOS EXPERIMENTOS

Neste capítulo, são expostas as etapas que representam o método subjetivo desenvolvido nesta pesquisa, seguidas de sua aplicação no contexto dos sistemas de Teleconferência Imersiva. Organizados em seções, são detalhados os critérios utilizados nas escolhas e definições necessárias em cada uma das etapas.

4.1 Desenvolvimento do Método Subjetivo

O método subjetivo de avaliação de qualidade de segmentação aqui apresentado consiste na sequência de passos mostrados na figura 13.

Como pode ser observado no diagrama, faz-se necessária a seleção de um conjunto de vídeos que contenha o elemento de interesse a ser isolado no processo de segmentação. Nesses vídeos, o ambiente em que esse elemento está inserido deve apresentar as características do local em que será realizada a captura do vídeo na aplicação em investigação. Além desses vídeos-fonte, são também necessárias implementações[1] de algoritmos de segmentação de vídeos capazes de extrair elementos de interesse na aplicação em questão. Por meio desses algoritmos é que os vídeos-fonte devem ser segmentados, como mostrado no diagrama da figura 13.

Uma característica dos algoritmos de segmentação que representam o estado-da-arte em diferentes domínios de aplicação, inclusive o investigado neste trabalho, é o fato de serem desenvolvidos com base em abordagens que permitem ajustes de um ou mais pa-

[1] A métrica objetiva derivada a partir dos resultados do método subjetivo é dependente do algoritmo, portanto, cada conjunto de vídeos gerados a partir de implementações de um algoritmo de segmentação específico resulta em uma métrica diferente.

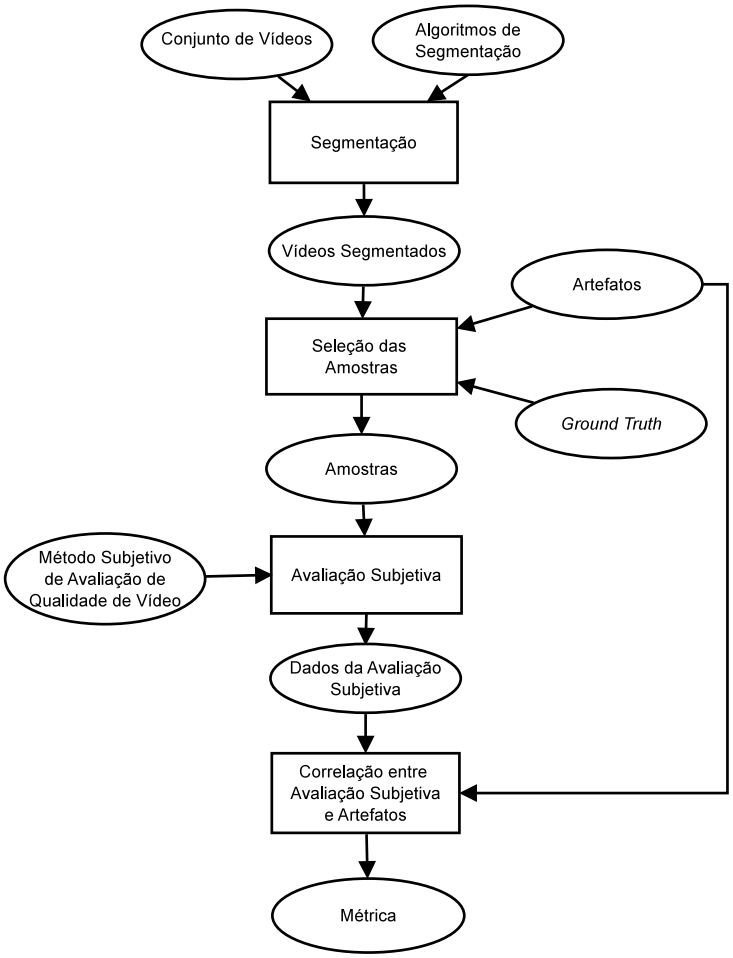


Figura 13 – Diagrama de blocos representando os métodos utilizados no desenvolvimento desta pesquisa

râmetros. Tais parâmetros são definidos previamente ou durante a execução da aplicação, para que o algoritmo adapte-se ao ambiente em que se captura o vídeo e produza melhores resultados.

Uma vez que os algoritmos são parametrizados e que cada conjunto de parâmetros produzem resultados diferentes, ainda que possa haver pouca variação, pode-se definir várias combinações de parâmetros como entrada. A partir dessas combinações, novos vídeos (mais precisamente, camadas de vídeos), que contêm diferentes formas de erros, são produzidos como resultado da execução de diferentes algoritmos de segmentação, parametrizados com diferentes combinações de valores de parâmetros. Essas camadas de vídeo serão, representados pelo bloco “Vídeos Segmentados” da figura 13, são, posteriormente, utilizadas em experimentos subjetivos.

Como a quantidade de avaliações realizadas por uma pessoa em um experimento subjetivo deve ser limitada[2], apenas algumas das camadas de vídeo produzidas devem ser selecionadas e, a partir dessa seleção, devem ser geradas cenas que simulem o ambiente da aplicação em investigação. Em outras palavras, o bloco "Amostras" exibido no diagrama da figura 13 deve representar cenas compostas da combinação de um elemento de interesse com um cenário que represente o ambiente de uma aplicação específica. Os *ground truths* correspondentes aos vídeos-fonte cujas camadas foram selecionadas podem ser utilizados para quantificar os erros de segmentação contidos nessas camadas de imagens, o que pode ser necessário, caso um critério baseado em erro objetivo seja escolhido para selecioná-las.

Uma tarefa importante no processo consiste na definição dos tipos de erros que, hipoteticamente, causam grande incômodo aos usuários. Esses erros, aqui chamados de "Artefatos" (figura 13), devem considerar tanto características espaciais quanto temporais, uma vez que tratam-se de sequências de quadros e não de imagem estática. Tais artefatos devem simular os erros de segmentação que ocorrem em uma situação real.

Definidos os artefatos e gerados os novos vídeos, um método formal de avaliação de qualidade de vídeo deve ser utilizado para que as opiniões dos usuários em relação aos artefatos presentes nos vídeos possam ser colhidas e utilizadas para encontrar níveis de incômodo relacionados aos próprios artefatos. Realizadas as avaliações subjetivas, deve-se encontrar uma forma de correlacionar os resultados dessas avaliações (bloco "Dados da Avaliação Subjetiva") com a ocorrência dos artefatos, para que os que causam maior incômodo possam ser identificados e utilizados no desenvolvimento da métrica objetiva.

As seções seguintes apresentam em detalhes as principais tarefas envolvidas na aplicação do método.

4.2 Seleção dos Vídeos-Fonte

O primeiro passo no processo de aplicação do método consiste na seleção dos vídeos dos quais os elementos de interesse devem ser extraídos. Conforme discutido na seção 3.1,

[2]Segundo a ITU-R (2009), uma pessoa deve permanecer entre 30 e 60 minutos em um processo de avaliação. Trabalhos na área de testes psico-visual (FARIAS, 2004), sugerem que esse tempo não ultrapasse 30 minutos.

os erros de classificação de pixels ocorrem, principalmente, devido a dificuldade dos algoritmos em lidar com as “situações-problema” que ocorrem no ambiente em que se captura o vídeo. Por esse motivo, foram selecionados como base para os experimentos realizados neste trabalho várias sequências de vídeo que simulam algumas daquelas situações, considerando as possíveis características do ambiente onde a captura do vídeo se realiza, na aplicação de RA aqui investigada.

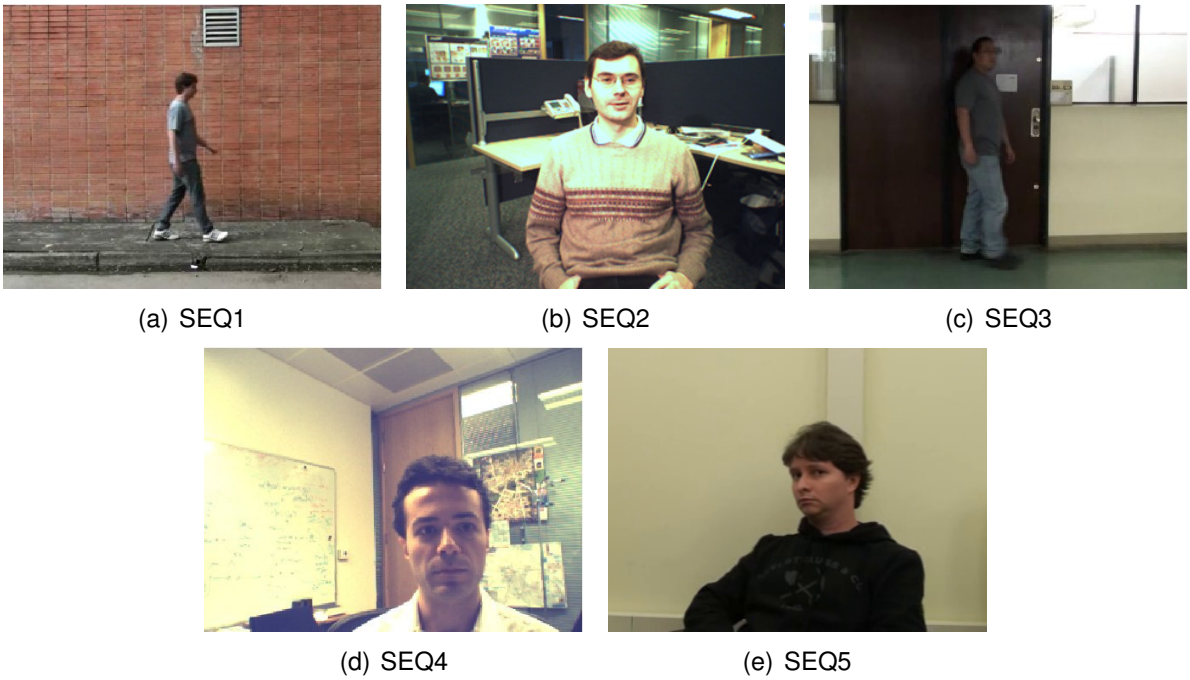
O sistema de Teleconferência Imersiva descrito na seção 3.3, em que os resultados desta pesquisa serão aplicados, exige algoritmos de segmentação como os descritos na tabela 2 (baseado em equipamento convencional). Das situações-problema associadas a esses algoritmos, cabe as seguintes considerações:

- Variações na Iluminação: imagina-se que, no contexto da aplicação investigada neste trabalho, as variações de iluminação devem ocorrer. No entanto, espera-se que ocorram de forma branda. Essa situação-problema foi considerada nesta pesquisa.
- Cores semelhantes no fundo e no elemento de interesse: situação comum em ambientes não controlados, que se pode fazer presente no ambiente da aplicação em questão. Essa situação também foi considerada nesta pesquisa.
- Movimentação no Fundo: não foi tratada nesta pesquisa, pois os métodos de segmentação analisados exigem plano de fundo estático. Considera-se que um único usuário esteja presente na cena e que não exista movimentação no fundo.
- Elemento de interesse estático: como existe a possibilidade da inicialização do sistema com uma imagem “limpa” do plano de fundo e os métodos mais recentes que se baseiam em arcabouços de minimização de energia não trabalham apenas com informação de movimento, essa situação-problema não foi considerada nesta pesquisa.
- Oscilações da câmera: considera-se que a câmera permanecerá estática durante a execução da aplicação quando o usuário (aluno) assumir o papel de professor durante a execução da aplicação. Essa situação não foi considerada nesta pesquisa.

Considerando as observações acima, foram utilizados como vídeos-fonte 5 (cinco) sequências, denominadas SEQ1, SEQ2, SEQ3, SEQ4 e SEQ5. Os vídeos-fonte SEQ2,

SEQ3 e SEQ4 apresentam em seu conteúdo duas situações-problema: variações na iluminação e cores semelhantes no fundo e no elemento de interesse. Tais situações são consideradas de maior ocorrência na aplicação em questão. Duas sequências de vídeos-fonte, SEQ1 e SEQ5 não apresentam nenhuma das situações-problemas analisadas neste trabalho.

Os vídeos-fonte SEQ2 e SEQ4 foram obtidos de uma base de dados de uso livre para pesquisas[3] ao passo que os vídeos-fonte SEQ1, SEQ3 e SEQ5 foram produzidos para esta pesquisa. Quadros desses vídeos-fonte são mostrados na figura 14.



(a) SEQ1 (b) SEQ2 (c) SEQ3

(d) SEQ4 (e) SEQ5

Figura 14 – Quadros das sequências de vídeo originais (vídeos-fonte) SEQ1, SEQ2, SEQ3, SEQ4 e SEQ 5

Nas sequências SEQ1 e SEQ3, o elemento de interesse na cena – a pessoa em primeiro plano – encontra-se distante da câmera, portanto, todo seu corpo pode ser visualizado. Nas demais, a câmera foi posicionada mais próxima e apenas a parte superior do corpo pôde ser visualizada[4]. Imagina-se que essas duas situações possam ocorrer durante a execução da aplicação.

Todos os vídeos-fonte possuem um *ground truth*. Em outras palavras, para cada qua-

[3] http://research.microsoft.com/en-us/projects/i2i/data.aspx

[4] Estudos voltados a videoconferências mostram que, ainda que os vídeos permitam visualizar apenas a parte superior do corpo dos participantes, a comunicação pode ser tão efetiva quanto a presencial (NGUYEN; CANNY, 2009)

dro de vídeo, existe um quadro correspondente segmentado de maneira precisa. Os *ground truths* associados aos vídeos-fonte SEQ1, SEQ3 e SEQ5, que foram produzidos para este experimento, foram rotulados (segmentados) manualmente.

Os pixels dos quadros de vídeo dos *ground truths* consistem em um *trimap*, pois são rotulados como primeiro plano, plano de fundo e região desconhecida (figura 15(c)), que são os que fazem parte das regiões que sofrem influência das cores tanto do plano de fundo como do elemento de interesse (normalmente localizada nas bordas do elemento de interesse), onde aplicam-se técnicas de *matting* (WANG; COHEN, 2007). Na figura 15 são exibidos um quadro da sequência SEQ2 e seu respectivo *ground truth*.

(a) (b) (c)

Figura 15 – Quadro da sequência de vídeo SEQ2 e seu respectivo *ground truth*

4.3 Algoritmos de Segmentação

Uma vez que a presente pesquisa volta-se à avaliação da qualidade da segmentação no contexto das aplicações de Teleconferência Imersiva, foram utilizados algoritmos que sejam aplicáveis no sistema descrito na seção 3.3. Esses algoritmos podem ser executados em ambientes não controlados e não fazem uso de equipamentos específicos para auxiliar a segmentação.

Como descrito na seção 4.1, os tipos de erros são produzidos por meio da execução desses algoritmos. Nesta pesquisa, foram selecionados 4 (quatro) algoritmos que se baseiam em plano de fundo estático, aqui denominados Qian (QIAN; SEZAN, 1999), Stau (STAUFFER; GRIMSON, 2000), Crim (CRIMINISI et al., 2006) e Sanc (SANCHES; SILVA; TORI, 2012). Os dois primeiros são baseados na tradicional abordagem de subtração de fundo, discutida na seção 2.1.1, ao passo que os demais são desenvolvidos com base em ar-

cabouços de minimização de energia, discutidos na seção 2.1.2 e no apêndice I (seção I.2).

Em Qian e Stau, o principal parâmetro a ser ajustado refere-se ao limiar (*threshold*), que é sensível a iluminação do ambiente em que se captura o vídeo. Nos métodos Crim e Sanc, os parâmetros principais controlam a influência de cada um dos 4 (quatro) termos que servem como entrada para o arcabouço de minimização de energia utilizado pelo algoritmo, conforme detalhado nas seções I.3.2 e I.3.3, do apêndice I.

O algoritmo Qian exige inicialização na forma de um plano de fundo "limpo" (em que o elemento de interesse não esteja presente) e os algoritmos Crim e Sanc devem ser inicializados com duas amostras de cores: as presentes no plano de fundo e as do elemento de interesse. No método Stau, por sua vez, um modelo do fundo pode ser obtido de forma automática, porém, sua inicialização com um modelo de fundo pré-capturado produz melhores resultados, sobretudo, na segmentação dos primeiros quadros.

Os algoritmos de segmentação baseados na técnica de subtração de fundo tradicionalmente são utilizados em aplicações de RA, no entanto, por se mostrarem mais robustas, abordagens baseadas em arcabouços de minimização de energia têm sido adotadas por algumas aplicações (SANCHES et al., 2012a). Os 4 (quatro) algoritmos de segmentação utilizados neste trabalho são expostos em detalhes nas seções I.3.1, I.3.2, I.3.3 e I.3.4, do apêndice I.

4.4 Método de Avaliação Subjetiva de Qualidade de Vídeo

Para que sejam colhidas as opiniões de usuários em relação a qualidade dos vídeos produzidos a partir de segmentação imperfeita faz-se necessária a aplicação de um método de avaliação subjetiva[5] de qualidade vídeo, como discutido na seção 4.1. Alguns métodos formais reconhecidamente eficientes (PÉCHARD; PÉPION; CALLET, 2008) são populares tanto na indústria quanto na comunidade científica, entre eles, o SAMVIQ (*Subjective Assessment Methodology for Video Quality*) (KOZAMERNIK et al., 2005; ITU-R, 2007), que, de acordo com alguns estudos (PÉCHARD; PÉPION; CALLET, 2008), tem se mostrado bastante preciso.

[5]O experimento subjetivo realizado nesta pesquisa foi aprovado (CAAE: 0022.0.198.000-11) pelo Comitê de Ética em Pesquisa (CEP) do Hospital Universitário da Universidade de São Paulo (Anexo A).

Devido a essa precisão, o método SAMVIQ foi utilizado neste trabalho para levantar os artefatos mais perceptíveis ao usuário.

A aplicação desses métodos na indústria tem sido recomendada por órgãos como a ITU e a EBU[6] (*European Broadcasting Union*), que sugerem tanto o modo como deve ser realizada cada etapa do processo de avaliação quanto a configuração física do ambiente em que os testes devem ser realizados (ITU-R, 2009). Detalhes como o número de observadores e a distância desses observadores da tela; o tamanho, o tipo e a intensidade da luz emitida da tela, que deve ser apropriado para a aplicação sendo avaliada; assim como a cor do fundo da imagem, quando o sistema trabalha com imagens de tamanho reduzido; fazem parte dessas recomendações (apêndice II), que foram obedecidas neste trabalho.

No processo de avaliação realizado por meio do método SAMVIQ o usuário tem acesso a várias versões de um mesmo vídeo – no caso, cada versão contém um tipo ou uma combinação de artefatos. A produção dos vídeos exibidos aos avaliadores, que simulam um ambiente de RA e apresentam vários tipos de artefatos, é descrita na seção 4.6. A definição dos artefatos, que são observados nos vídeos, é discutida na seção 4.5.

Quando todas as versões de determinado vídeo são avaliadas pelo observador, o conteúdo da versão seguinte pode ser acessado. Cada versão de um vídeo é mostrada isoladamente e avaliada por meio da escolha de valores em uma escala de qualidade contínua (ITU-R, 2007), exibida na figura 16.

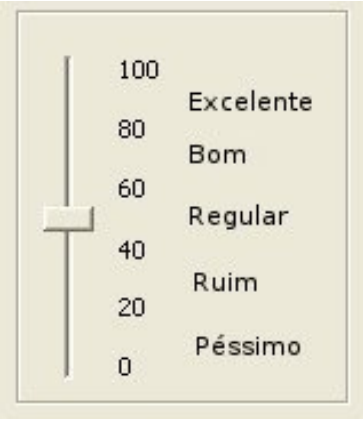


Figura 16 – Escala de qualidade contínua exibida ao avaliador durante a execução das avaliações subjetivas

Na escala, cada observador move o controle deslizante sobre a grade contínua, que vai

[6] http://www.ebu.ch

de 0 (zero) a 100 (cem) e são agrupadas em 5 (cinco) itens, arranjados linearmente (excelente, bom, regular, ruim, péssimo). Os usuários assistem aos vídeos sentados à distância de aproximadamente 30 (trinta) centímetros, como especificado pelas recomendações da ITU (ITU-R, 2007), que são adotadas pelo SAMVIQ.

As diferentes versões são selecionadas aleatoriamente pelo usuário, que pode parar, rever e modificar o resultado de cada versão de uma sequência desejada. Esse método inclui uma referência explícita (nesse caso, um vídeo sem erros de segmentação), que não é avaliado, e referências escondidas, que são avaliadas. As referências escondidas referem-se aos próprios vídeos de referência são inseridos no grupo de vídeos avaliados (figura 29). Antes do processo de avaliação há uma breve sessão de treinamento para que o avaliador se familiarize com o ambiente de teste e com a interface da ferramenta de *software* utilizada.

O método sugere também medidas para a iluminação do ambiente, que foram mantidas em todos os testes. Conforme as recomendações da ITU (ITU-R, 2009), 15 (quinze) avaliadores foram recrutados para cada bateria de testes. A forma de análise dos resultados também é padronizada e fazem parte das recomendações da ITU (ITU-R, 2009). Detalhes adicionais sobre o método SAMVIQ são descritos na seção II.1 do apêndice II.

4.5 Definição dos Artefatos

Uma tarefa importante na aplicação do método subjetivo e, consequentemente, no desenvolvimento da métrica objetiva consiste na definição dos artefatos que representam as diferentes formas em que os erros de segmentação apresentam-se na cena final. Esses artefatos, que devem estar presentes no conteúdo dos vídeos analisados nos experimentos subjetivos, serão, em uma etapa posterior, combinados e associados a um peso, para representar o nível geral de incômodo relacionado aos erros de segmentação.

No entanto, antes que se definissem novos artefatos foi realizada uma análise das métricas de avaliação de segmentação existentes, com ênfase nos trabalhos que apresentem as historicamente mais utilizadas e as que representam o estado-da-arte, como as descritas na seção 3.2. Como essas métricas foram desenvolvidas para aplicações de diferentes domínios, nem todas as abordagens são diretamente aplicáveis em avaliação

de segmentação voltada aos sistemas de RA investigado neste trabalho. Porém, algumas dessas métricas definem artefatos e sugerem formas de combiná-los e, por esse motivo, essas pesquisas foram consideradas e analisadas.

Como detalhado na seção 3.2.2, quatro artefatos $\mathcal{A}_r$, $\mathcal{A}_b$, $\mathcal{H}_i$ e $\mathcal{H}_b$ foram definidos na métrica PST (GELASCA; EBRAHIMI, 2009). A partir desses artefatos, quatro métricas PST_{A_r}, PST_{A_b}, PST_{H_i} e PST_{H_b} foram derivadas. Essas métricas foram combinadas para representar o incômodo geral relacionado aos erros de segmentação, resultando na métrica PST. Para efeito de análises, as métricas PST_{A_r}, PST_{A_b}, PST_{H_i} e PST_{H_b} aqui serão tratadas como artefatos e serão incluídas ao grupo de artefatos definidos nesta seção.

Entre os novos artefatos aqui definidos existem alguns considerados mais simples, como os descritos em Villegas, Marichal e Salcedo (1999) e outros mais elaborados, como os da PST. Conforme justificado na seção 4.1, foram consideradas tanto características espaciais quanto temporais em sua definição. Alguns deles possuem, ainda, parâmetros cujos valores devem ser definidos.

O primeiro artefato definido neste trabalho considera a influência dos pixels classificados de forma incorreta localizados na região do plano de fundo (falso negativo), calculado de acordo com a equação

$$E_{\mathcal{N}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{p=1}^{P} pix(p,k) \in \mathcal{N}(k), \tag{20}$$

onde $pix(p)$ é o pixel na posição p do quadro k. Os falsos positivos $E_{\mathcal{P}}$, que representam a soma dos erros que ocorrem no plano de fundo podem ser calculados de forma similar, segundo a equação

$$E_{\mathcal{P}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{p=1}^{P} pix(p,k) \in \mathcal{P}(k). \tag{21}$$

O erro total $E_{\mathcal{T}}$, que representa todos os pixels classificados de forma incorreta, é dado por

$$E_{\mathcal{T}} = E_{\mathcal{N}} + E_{\mathcal{P}}. \tag{22}$$

Alguns artefatos foram definidos para medir a influência da distância (euclidiana) dos falsos positivos ao elemento de interesse. $D_{\mathcal{P}in(dt)}$ representa os falsos positivos distantes do elemento de interesse até dt pixels, definido pela equação

$$D_{\mathcal{P}in(dt)} = \frac{1}{K}\sum_{k=1}^{K}\sum_{p=1}^{P} pix(p) \in \mathcal{P}(k),\ \forall\ pix(p) < dt \tag{23}$$

onde $dt \in \{80, 90, 100, 110, 120\}$, valores definidos com base na proporção média entre o tamanho da janela e a região ocupada pelo elemento de interesse. Da mesma forma, podem ser calculada $D_{\mathcal{P}out(dt)}$, que representa os falsos positivos mais que dt pixels distantes do elemento de interesse, de acordo com a equação

$$D_{\mathcal{P}out(dt)} = \frac{1}{K}\sum_{k=1}^{K}\sum_{p=1}^{P} pix(p) \in \mathcal{P}(k),\ \forall\ pix(p) > dt. \tag{24}$$

Outra forma de incômodo considerada neste trabalho trata da influência de artefatos em forma de componentes conectados (blobs) no plano de fundo. Esse artefato pode ser calculado conforme a equação

$$B_{\mathcal{P}large(size)} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{B}_{\mathcal{P}}(k),\ \forall\ |\mathcal{B}_{\mathcal{P}}| > size \tag{25}$$

onde $\mathcal{B}_{\mathcal{P}}$ é um conjunto de falsos positivos conectados, $|\ .\ |$ representa o operador de cardinalidade e $size \in \{5, 10, 15, 20\}$, representa quantidades de pixels de um componente conectado. De forma similar podem ser calculados $B_{\mathcal{P}small(size)}$, que são falsos positivos conectados com cardinalidade menor que $size$, dado pela equação

$$B_{\mathcal{P}small(size)} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{B}_{\mathcal{P}}(k),\ \forall\ |\mathcal{B}_{\mathcal{P}}| < size, \tag{26}$$

$B_{\mathcal{N}large(size)}$, que são falsos negativos conectados com cardinalidade maior que $size$, dado por

$$B_{\mathcal{N}large(size)} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{B}_{\mathcal{N}}(k),\ \forall\ |\mathcal{B}_{\mathcal{N}}| > size \tag{27}$$

e $B_{\mathcal{N}small(size)}$, que são os falsos positivos conectados com cardinalidade menor que $size$, definido pela equação

$$B_{\mathcal{N}small(size)} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{B}_{\mathcal{N}}(k),\ \forall\ |\mathcal{B}_{\mathcal{N}}| < size. \tag{28}$$

Alguns dos artefatos foram definidos com o objetivo de calcular a influência do as-

pecto temporal na qualidade da segmentação. Para que se considere erros temporais, foram analisados artefatos como $T_{\mathcal{N}(pc)}$, que representa os erros que ocorrem nos pixels da posição p e não ultrapassam um percentual pc dos K quadros da sequência de vídeo

$$T_{\mathcal{N}(pc)} = \frac{1}{K}\sum_{k=1}^{K} pix(p,k) \in \mathcal{N}(k) \text{ and } T_{\mathcal{N}(pc)} < pc \tag{29}$$

onde $pc \in \{40, 50, 60, 70, 80, 90\}$. De forma similar pode ser calculado $T_{\mathcal{P}(pc)}$, que representa os mesmos erros temporais em relação aos falsos positivos, dados pela equação

$$T_{\mathcal{P}(pc)} = \frac{1}{K}\sum_{k=1}^{K} pix(p,k) \in \mathcal{P}(k) \text{ and } T_{\mathcal{P}(pc)} < pc. \tag{30}$$

Erros espaciais e temporais também foram calculados com base no conceito de "falso blob". Um falso blob espacial em relação aos falsos negativos $FS_{\mathcal{N}}$ é calculado pela convolução de uma imagem binária com um kernel M_s, de acordo com a equação

$$FS_{\mathcal{N}}s = \sum_{k=1}^{K} M_{\mathcal{N}(k)} * M_s \tag{31}$$

onde $*$ é o operador de convolução, $M_{\mathcal{N}(k)}$ é uma imagem binária

$$pix(p)_{M_{\mathcal{N}(k)}} = \begin{cases} 1 & , \text{ if } pix(p) \in \mathcal{N} \\ 0 & , \text{ if } pix(p) \notin \mathcal{N} \end{cases} \tag{32}$$

e M_s é o *kernel*

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{33}$$

Falsos blobs relacionados aos falsos positivos $FS_{\mathcal{P}}s$ podem ser calculados de forma similar, conforme a equação

$$FS_{\mathcal{P}}s = \sum_{k=1}^{K} M_{\mathcal{P}(k)} * M_s. \tag{34}$$

Outra forma de calcular níveis de incômodo relacionados a falsos blobs é por meio do artefato $FS_{\mathcal{N}}g$ que foi obtido pela convolução de $M_{\mathcal{N}(k)}$ de acordo com a equação

$$FS_{\mathcal{N}}g = \sum_{k=1}^{K} M_{\mathcal{N}(k)} * M_g\sigma \tag{35}$$

onde M_g é um *kernel* gaussiano centralizado com desvio padrão $\sigma = 0.8$. $FS_{\mathcal{P}}g$ foi obtido de maneira similar conforme a equação

$$FS_{\mathcal{P}}g = \sum_{k=1}^{K} M_{\mathcal{P}(k)} * M_g\sigma. \tag{36}$$

Além dos citados, foram também definidos atributos relacionados ao aspecto temporal da ocorrência de falsos blobs. Considera-se uma sequência de vídeo como uma matriz tridimensional H x W x K, onde H representa a altura, W a largura de um quadro e K a quantidade de quadros. Um "quadro temporal" Qt pode ser definido como uma imagem H x K, como mostrado na figura 17.

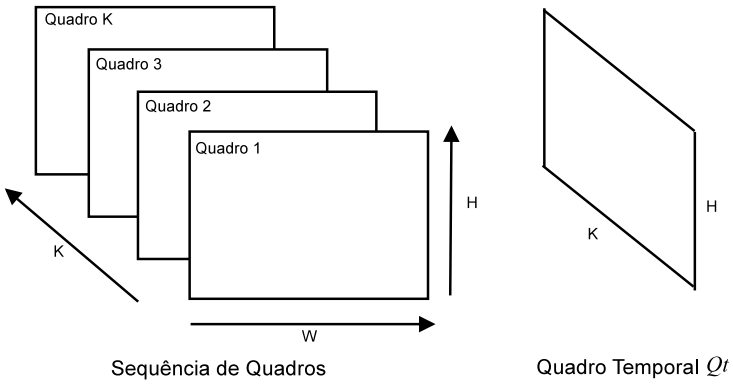


Figura 17 – Representação de um quadro temporal

Desse modo, uma sequência contêm W quadros temporais. Os falsos blobs temporais falso negativos são dados por

$$FT_{\mathcal{N}}s = \sum_{w=1}^{W} Qt_{\mathcal{N}(w)} * M_s\sigma \tag{37}$$

onde $Qt_{\mathcal{N}}$ é uma imagem binária criada a partir dos falsos negativos. $FT_{\mathcal{P}}s$, que são os blobs temporais falso positivos

$$FT_{\mathcal{P}}s = \sum_{w=1}^{W} Qt_{\mathcal{P}(w)} * M_s\sigma, \tag{38}$$

$FT_{\mathcal{N}}g$, que são os blobs temporais falso negativos obtidos da convolução com o *kernel* gaussiano

$$FT_{\mathcal{N}}s = \sum_{w=1}^{W} Qt_{\mathcal{N}(w)} * M_g\sigma \tag{39}$$

e $FT_{\mathcal{P}}g$, que são os blobs temporais falso positivos obtidos da convolução com o *kernel* gaussiano

$$FT_{\mathcal{P}}g = \sum_{w=1}^{W} Qt_{\mathcal{P}(w)} * M_g\sigma \tag{40}$$

foram obtidos de forma similar.

4.6 Preparação da Base de Vídeos e Execução das Avaliações Subjetivas

Uma vez definidos o conjunto de vídeos-fonte, os algoritmos de segmentação, o método de avaliação subjetiva de qualidade de vídeo e os artefatos a serem analisados, os próximos passos tratam da construção dos novos vídeos (que simulam cenas de RA com erros de segmentação) e da aplicação do método de avaliação subjetiva descrito na seção 4.4.

A partir das sequências de vídeos-fonte e com o auxílio dos *ground truths* foram produzidas novas sequências, utilizando camadas de primeiro plano obtidas da segmentação dos vídeos-fonte por meio dos métodos descritos na subseção 4.3. Como descrito na seção 4.3, todos os algoritmos utilizados, ou exigem ou permitem algum tipo de inicialização para que produzam melhores resultados.

Em consequência disso, os algoritmos Qian e Stau foram inicializados com um modelo do fundo “limpo”, obtido conforme detalhado na seção I.3.1, ao passo que nos algoritmos Crim e Sanc, o primeiro quadro da sequência e seu respectivo *ground truth* foram utilizados para isolar o elemento de primeiro plano do quadro e obter os histogramas das cores do fundo e do elemento de interesse, que foram utilizados para inicializar o termo de cor, como detalhado no apêndice I, nas seções I.3.2 e I.3.3.

No algoritmo Qian, para que se alcançassem diferentes formas de erros, foram utilizados na execução desses dois métodos, diferentes valores do limiar (equação 50) que controla uma faixa de tolerância na comparação das cores do modelo do fundo e do quadro em análise. No algoritmo Stau, os erros também foram obtidos alterando-se o valor do limiar que têm essa mesma finalidade, como mostrado na equação 67. Foram utilizados como entrada para o método um parâmetro do limiar no intervalo (1-100), totalizando 100 (cem) variações de erros para cada algoritmo.

Nos algoritmos Crim e Sanc, alterou-se os valores dos 4 (quatro) parâmetros de normalização do Campo Aleatório Condicional (equação 54) utilizado no modelo. Os conjuntos de parâmetros utilizados para alimentar esses algoritmos foram escolhidos aleatoriamente no intervalo (0,0 - 0,2). Foram produzidos um total de 1000 (mil) combinações desses parâmetros, a partir dos quais os métodos Crim e Sanc foram configurados e executados sobre os vídeos-fonte descritos na seção 4.2.

Uma vez que a utilização de todas as combinações de parâmetros produziriam um grande número de novos vídeos e que apenas uma pequena quantidade devem ser submetidos a avaliação, torna-se necessário encontrar um forma de reduzir a quantidade de amostras de camadas de primeiro plano, que serão utilizadas para gerar os vídeos a serem submetidos aos experimentos subjetivos.

Nesta pesquisa, a escolha dessas amostras foi guiada pelo espalhamento do artefato mais comum $E_{\mathcal{T}}$, que representa a quantidade de pixels classificados de forma incorreta (equação 22). Desse modo, as amostras selecionadas continham esse artefato variando na faixa de 0%, que representa o vídeo de referência (seção 4.4), até 31,85% (pior caso).

No trabalho de Gelasca e Ebrahimi (2009), em que a métrica de avaliação PST é apresentada, os artefatos foram produzidos e inseridos artificialmente nos vídeos, o que leva a acreditar que alguns deles podem não se apresentar exatamente da forma como foram exibidos ao avaliador. Em outras palavras, a métrica objetiva apresentada naquele trabalho foi derivada da análise dos resultados de avaliações subjetivas de cenas que dificilmente ocorreriam durante a execução da aplicação (figura 9).

Diferentemente da abordagem adotada no desenvolvimento da PST, neste trabalho, os vídeos submetidos aos avaliadores foram gerados a partir de camadas de primeiro plano obtidas dos resultados da execução dos algoritmos descritos nas seções 4.3 e I.3. Desse modo, os vídeos exibidos aos avaliadores na fase de avaliações subjetivas exibem artefatos tipicamente encontrados nas aplicações.

Uma vez definido um grupo de amostras a serem utilizadas, para que facilitasse a seleção de voluntários, essas amostras foram divididas em 4 (quatro) baterias de testes, para realização dos experimentos subjetivos. Na primeira, foram selecionados 4 (quatro) grupos de 6 (seis) amostras. Cada grupo é formado por camadas de primeiro plano geradas a partir da segmentação dos vídeos-fonte SEQ1, SEQ2, SEQ3 e SEQ4, respectivamente,

e representam 6 (seis) variações do artefato $E_{\mathcal{T}}$.

Três dessas variações foram segmentadas utilizando o algoritmo Qian e, as outras três, são resultados da execução do algoritmo Crim. As variações do artefato $E_{\mathcal{T}}$, presentes no conteúdo das camadas utilizadas nessa bateria, podem ser visualizados na tabela 4, em que as colunas N^o e Alg representam um identificador da camada de primeiro plano no experimento e o algoritmo de segmentação utilizado para gerar a camada, respectivamente.

Tabela 4 – Ocorrência do artefato $E_{\mathcal{T}}$, que representa a média dos erros de classificação de pixels, presentes nos vídeos dos testes da bateria 1

	SEQ1			SEQ2			SEQ3			SEQ4	
N^o	Alg.	$E_{\mathcal{T}}$	N^o	Alg.	$E_{\mathcal{T}}$	N^o	Alg.	$E_{\mathcal{T}}$	N^o	Alg.	$E_{\mathcal{T}}$
Ref.	–	0	Ref.	–	0	Ref.	–	0	Ref.	–	0
1	Crim	1,01	7	Qian	28.82	13	Qian	9.13	19	Crim	12.68
2	Qian	2,08	8	Crim	7.61	14	Crim	9.80	20	Qian	12.87
3	Crim	3,09	9	Qian	29.72	15	Qian	10.20	21	Crim	15.65
4	Qian	4,17	10	Crim	8.27	16	Crim	11.95	22	Qian	15.74
5	Crim	5,13	11	Qian	31.85	17	Qian	13.43	23	Crim	17.50
6	Qian	6,46	12	Crim	10.09	18	Crim	14.03	24	Qian	19.63

Na segunda bateria de testes, as mesmas amostras de camadas de primeiro plano foram utilizadas para gerar os novos vídeos. No entanto, essas camadas foram combinadas com diferentes planos de fundo, como discutido no final desta seção.

Na terceira bateria de testes, também foram selecionados 4 (quatro) grupos de 6 (seis) amostras. Cada grupo é formado por camadas de primeiro plano geradas a partir da segmentação somente dos vídeos-fonte em que o elemento de interesse encontra-se mais próximo da câmera, SEQ2, SEQ4 e SEQ5. Deste último foram selecionados 2 (dois) grupos de amostras. Novamente, cada grupo contém 6 (seis) variações do artefato $E_{\mathcal{T}}$. Assim como na bateria 1, três dessas variações são resultados da execução do algoritmo Crim e as demais, geradas a partir da execução do algoritmo Qian. Na tabela 5, são exibidas variações do artefato $E_{\mathcal{T}}$, presentes no conteúdo das camadas utilizadas nessa bateria.

As camadas de primeiro plano utilizadas na bateria 4 são resultados da execução de dois algoritmos não utilizados nas amostras das baterias anteriores, os algoritmos Sanc e Stau. Nesta bateria, foram selecionados apenas 3 (três) grupos de 6 (seis) amostras. Cada

Tabela 5 – Ocorrência do artefato $E_{\mathcal{T}}$, que representa a média dos erros de classificação de pixels, presentes nos vídeos dos testes da bateria 3

	SEQ5			SEQ2			SEQ4			SEQ5	
N^o	Alg.	$E_{\mathcal{T}}$	N^o	Alg.	$E_{\mathcal{T}}$	N^o	Alg.	$E_{\mathcal{T}}$	N^o	Alg.	$E_{\mathcal{T}}$
Ref.	–	0	Ref.	–	0	Ref.	–	0	Ref.	–	0
1	Crim	0.33	7	Qian	6.47	13	Crim	5.80	19	Qian	0.09
2	Qian	1.57	8	Crim	9.96	14	Qian	29.04	20	Crim	0.57
3	Crim	2.52	9	Qian	6.82	15	Crim	6.14	21	Qian	1.75
4	Qian	3.53	10	Crim	10.37	16	Qian	29.20	22	Crim	2.84
5	Crim	4.50	11	Qian	7.23	17	Crim	6.35	23	Qian	3.36
6	Qian	5.62	12	Crim	10.60	18	Qian	29.34	24	Crim	4.31

grupo é formado por camadas de primeiro plano geradas a partir da segmentação dos vídeos-fonte SEQ2, SEQ4 e SEQ5, respectivamente, e, assim como nas demais baterias, representam 6 (seis) variações do artefato $E_{\mathcal{T}}$. Três dessas variações são resultados da execução do algoritmo Sanc e as demais foram obtidas da execução do algoritmo Stau. As variações do artefato $E_{\mathcal{T}}$, presentes no conteúdo das camadas utilizadas nessa bateria, são exibidas na tabela 6.

Tabela 6 – Ocorrência do artefato $E_{\mathcal{T}}$, que representa a média dos erros de classificação de pixels, presentes nos vídeos dos testes da bateria 4

	SEQ5			SEQ2			SEQ4	
N^o	Alg.	$E_{\mathcal{T}}$	N^o	Alg.	$E_{\mathcal{T}}$	N^o	Alg.	$E_{\mathcal{T}}$
Ref.	–	0	Ref.	–	0	Ref.	–	0
1	Sanc	1,25	7	Stau	2.80	13	Sanc	5.20
2	Stau	1.04	8	Sanc	8.54	14	Stau	6.33
3	Sanc	2.49	9	Stau	3.62	15	Sanc	5.50
4	Stau	2.54	10	Sanc	9.02	16	Stau	6.50
5	Sanc	4.46	11	Stau	4.40	17	Sanc	6.53
6	Stau	5.09	12	Sanc	9.62	18	Stau	7.01

Uma vez que as amostras de camadas de primeiro plano foram selecionadas e agrupadas em baterias de testes, os vídeos exibidos aos avaliadores nas avaliações subjetivas podem ser gerados a partir da combinação dessas amostras com um novo plano de fundo.

Como a aplicação em que a qualidade da segmentação deve ser avaliada trata-se de um sistema de Teleconferência Imersiva, a maioria dos vídeos utilizados nos experimentos simulam um ambiente dessa natureza.

Nesses vídeos, o elemento de interesse pode ser visualizado como uma textura sobre um plano, dentro de um ambiente virtual, o que caracteriza uma cena de um ambiente de RA. Obteve-se, portanto, um avatar, do tipo *billboard* (RHEE et al., 2007; CORRÊA et al., 2011), em que o plano que contém a textura permanece com a face principal sempre voltada para o usuário, independentemente do ponto de vista por ele escolhido. Todos os vídeos, no entanto, foram gerados a partir de um único ponto de vista, variando apenas os valores dos eixos z das coordenadas do ambiente virtual. O ambiente virtual desenvolvido para os testes simula um cenário que pode ser utilizado em sistemas de Teleconferência Imersiva, como pode ser observado na figura 18.

Figura 18 – Exemplos de vídeos produzidos para o experimento. Quadros dos vídeos baseados nos vídeos-fonte SEQ4 e SEQ1 em que um ambiente virtual é visualizado como plano de fundo

Os vídeos que simulam um sistema de Teleconferência Imersiva como o descrito na seção 3.3, foram utilizados nas baterias 1, 3 e 4. Para a bateria 2 foram produzidos vídeos em que o fundo original foi substituído por uma cor constante (cinza R=127, G=127 e B=127), como o mostrado na figura 19. A justificativa para a produção de vídeos com plano de fundo neutro será apresentada na seção 5.1. De acordo com especialistas em psicofísica, o fundo cinza pouco afeta o observador humano (GELASCA; EBRAHIMI, 2009), possibilitando que sua atenção se prenda ao elemento de interesse presente na cena.

Na composição dos vídeos, os pixels da camada de primeiro plano correspondentes à região desconhecida do *ground truth* (figura 15(c)) foram desconsiderados na análise dos artefatos. Nessa região, foi aplicada transparência de 50% do pixel da camada de

(a) (b)

Figura 19 – Exemplos de vídeos produzidos para o experimento. Quadros dos vídeo baseados nos vídeos-fonte SEQ3 e SEQ2 em que uma cor constante é visualizada como plano de fundo

elemento de interesse e 50% do pixel do novo fundo, com o objeto de suavizar as bordas do elemento de interesse na composição.

Em relação aos seus aspectos técnicos, os novos vídeos produzidos são de curta duração (10s) e possuem resolução 640x480 pixels. Tomou-se o cuidado para que todo o elemento de interesse não fosse obstruído pelos elementos virtuais que compõem a cena e, desse modo, todos os pixels pertencentes a esse elemento permanecesse visível em todos os quadros.

Definidos os vídeos utilizados em todas as baterias de teste, os experimentos subjetivos, cujo processo de aplicação é detalhado na seção 4.4 e no apêndice II, foram realizados. Importa ressaltar que os avaliadores não foram diretamente questionados sobre o nível de incômodo proporcionado pelos erros de segmentação (essa informação foi obtida da análise dos dados). Ao invés disso, cada participante era instruído a emitir sua opinião a respeito da qualidade dos vídeos exibidos.

A ferramenta de *software* MSU perceptual video quality[7], que possui implementações de métodos de avaliação subjetiva, incluindo o SAMVIQ, foi utilizada na aplicação das avaliações subjetivas. A interface gráfica utilizada pelos usuários nos experimentos é exibida na figura 20.

Somadas todas as baterias de testes dos experimentos subjetivos, um total de 90 (noventa) vídeos foram avaliados por 39 (trinta e nove) voluntários, com idades entre 20 (vinte) e 64 (sessenta e quatro) anos. Desses voluntários, 27 (vinte e sete) eram do sexo mas-

[7] http://compression.ru/video/quality_measure/perceptual_video_quality_tool_en.html

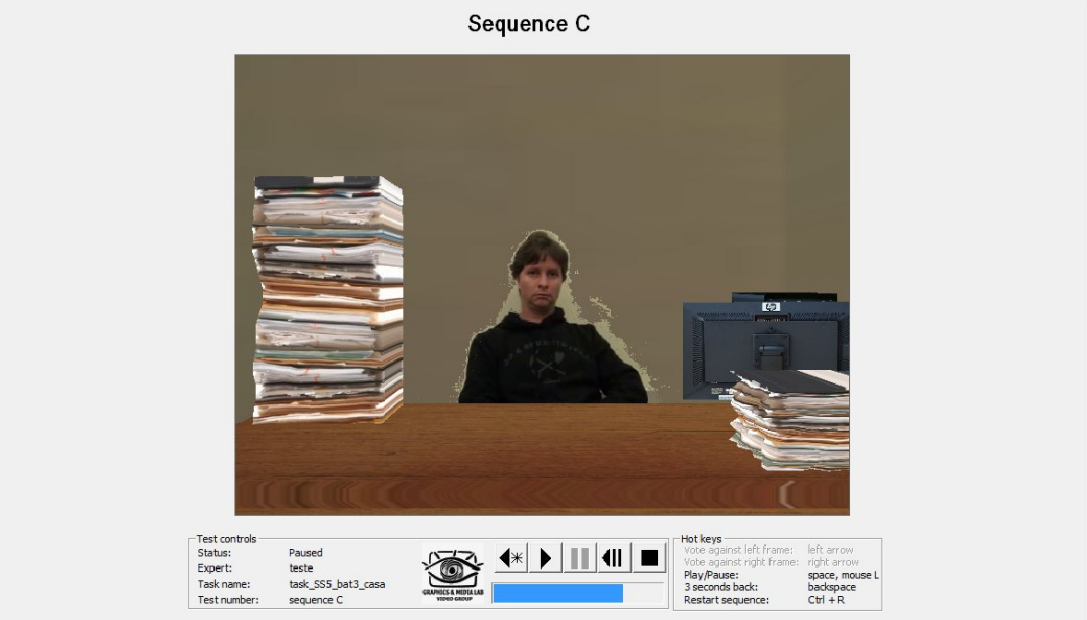


Figura 20 – Interface gráfica da implementação do método SAMVIQ

culino e 12 (doze) do sexo feminino. 4 (quatro) deles participaram de 3 (três) baterias de testes, 13 (treze) voluntários participaram de duas baterias de testes e os demais participaram de uma única bateria. Importa ressaltar que houve um intervalo de, no mínimo 45 (quarenta e cinco) dias, entre duas baterias de testes consecutivas. Detalhes como faixa etária, gênero e as baterias em que participaram os voluntários podem ser encontrados no apêndice II.

Os dados gerados das avaliações subjetivas consistem em valores que representam o nível de incômodo de cada um dos 90 (noventa) vídeos avaliados (somadas todas as baterias de teste). Cada valor é obtido da média (ITU-R, 2002) das avaliações dos 15 (quinze) voluntários que participaram da bateria em que o vídeo foi analisado. Uma vez que os vídeos possuem um nível de incômodo associado, foi verificado para cada um desses vídeos a ocorrência dos artefatos definidos na seção 4.5, inclusive os definidos na PST.

De posse desses dados, deve-se encontrar uma forma de correlacionar a ocorrência dos artefatos com os dados obtidos dos resultados das avaliações subjetivas. Esse processo, que possibilitará a obtenção da métrica objetiva será detalhado no capítulo 5.

5 ANÁLISE DOS RESULTADOS E DEFINIÇÃO DA MÉTRICA OBJETIVA

Realizados os experimentos com base no método subjetivo apresentado na seção 4.1, os dados que representam as opiniões dos usuários podem ser analisados com o objetivo de encontrar sua correlação com a ocorrência dos artefatos definidos na seção 4.5 e, como consequência, identificar o nível de incômodo por eles provocados. Essa correlação traz informações fundamentais, necessárias na definição da métrica objetiva.

Uma vez que o método subjetivo, exibido na figura 13, foi aplicado utilizando-se vídeos que simulam um sistema de Teleconferência Imersiva (esses vídeos foram exibidos aos avaliadores), a métrica objetiva derivada desses experimentos tem como sua aplicação a avaliação da segmentação produzida quando determinado algoritmo for utilizado no contexto desses sistemas.

Observando a figura 18, que simula o ambiente da aplicação, é possível identificar o avatar em posições distintas em relação à câmera: mais próximo, como na figura 18(a), e mais distante dela, como na figura 18(b). Quando esse comportamento do avatar pode ser conhecido *a priori*, os sistemas de Teleconferência Imersiva podem ser diferenciados de acordo com essa característica.

No sistema de Teleconferência Imersiva descrito na seção 3.3, por exemplo, essa característica pode ser considerada. Quando o aluno assume o papel do professor e o algoritmo de segmentação em ambiente não controlado é acionado pelo sistema, três possíveis cenários podem ser identificados e conhecidos *a priori*: i) o avatar (elemento real da cena) permanece sempre próximo da câmera, de forma que apenas parte de seu seu corpo pode ser visualizado, durante todo o tempo de execução da aplicação, como mostrado na figura 18(a); ii) o avatar permanece sempre distante da câmera, de forma que todo seu corpo pode ser visualizado, durante todo o tempo de execução da aplicação, como mos-

trado na figura 18(b); iii) o avatar alterna entre posições como as exibidas nas figuras 18(a) e 18(b), apresentando-se próximo ou distante da câmera.

Quando a informação sobre o comportamento do elemento de interesse – nesse caso, o avatar do aluno – pode ser obtida *a priori*, é possível considerar tal característica para refinar a métrica objetiva, para que produza melhores resultados. Esses detalhes e todo o processo de análise dos resultados obtidos da aplicação do método subjetivo, além da formalização da própria métrica, são expostos neste capítulo. Inicialmente, uma análise preliminar é realizada com o objetivo de verificar a aplicabilidade da métrica PST em sistemas de Teleconferência Imersiva com as características do descrito na seção 3.3.

5.1 Aplicabilidade da Métrica PST

A primeira análise a ser realizada tem como objetivo testar a aplicabilidade da métrica PST, apresentada em Gelasca e Ebrahimi (2009) e que representa o estado-da-arte na área.

Como discutido na seção 3.2.2, segundo Gelasca e Ebrahimi (2009), é possível obter uma métrica objetiva que avalie a qualidade de algoritmos de segmentação tanto em um cenário geral, quando não se conhece *a priori* a aplicação em que a imagem segmentada será utilizada, quanto no contexto de determinadas aplicações, quando os quadros de vídeo segmentados são utilizados em aplicações como vigilância por vídeo, compressão de vídeos e Realidade Aumentada.

Naquele trabalho, os autores definem quatro artefatos espaciais que englobam todas as possibilidades de erros de segmentação em um quadro. Considerando fatores espaciais e temporais relacionados a percepção dos usuários, obtidos de experimentos subjetivos, foram encontrados pesos para cada artefato, o que resultou em quatro métricas (aqui consideradas artefatos, uma vez que são combinadas para gerar uma nova métrica).

Os testes realizados nesta seção procuram verificar se a métrica objetiva PST pode ser utilizada para avaliar a segmentação produzida pelos algoritmos de segmentação descritos na seção 4.3, quando aplicados em sistemas de Teleconferência Imersiva. Importa ressaltar que o nível de incômodo produzido pelos erros de segmentação são representados pelos valores obtidos das avaliações subjetivas, conduzidas de acordo com o método SAMVIQ (KOZAMERNIK et al., 2005; ITU-R, 2007).

Ainda que a pesquisa aqui realizada tenha como foco a qualidade da segmentação no contexto de uma aplicação específica, inicialmente foi verificada a possibilidade de utilizar a métrica PST ajustada para avaliações em que não se conhece a aplicação, como proposto em Gelasca e Ebrahimi (2009). Para isso, foram utilizados os artefatos (PST_{A_r}, PST_{H_i}, PST_{A_b} e PST_{H_b}) e os respectivos pesos a, b, c e d, propostos na métrica PST (equação 19), para avaliações nessas condições. Os artefatos e pesos foram analisados de acordo com o seguinte procedimento, que foi reproduzido nas demais análises realizadas nesta seção:

- os dados das avaliações subjetivas relativas à ocorrência dos artefatos da PST foram agrupados utilizando como critério o algoritmo executado para segmentar a camada de primeiro plano que foi utilizada na geração do vídeo a ser avaliado no experimento subjetivo (D_{Crim} e D_{Qian});
- para cada grupo de dados, aplicou-se uma regressão linear com os artefatos (PST_{A_r}, PST_{H_i}, PST_{A_b} e PST_{H_b}) e os valores que representam o nível de incômodo desses artefatos, obtidos da avaliação subjetiva;
- encontrou-se, portanto, os valores ótimos para os pesos (a, b, c, d) e os respectivos intervalos de confiança;
- para cada algoritmo, foi verificado se os valores dos pesos (a, b, c, d) definidos na PST encontram-se dentro do intervalo de confiança.

Os resultados obtidos desse processo podem ser visualizados na tabela 7 em que a coluna P_{Gel} mostra os valores dos pesos (a, b, c, d) definidos na PST para avaliações em um cenário geral, quando não se conhece a aplicação em que serão utilizadas as camadas de primeiro plano geradas. As colunas P_{Crim} e P_{Qian} representam os pesos ótimos para os algoritmos Crim e Qian, respectivamente, obtidos pelo procedimento descrito no parágrafo anterior (regressão linear). As demais colunas representam as bordas esquerda e direita dos intervalos de confiança associados aos pesos (P_{Crim} e P_{Qian}), considerando níveis de confiança iguais a 99%, 95% e 85%, respectivamente. A coluna “Pos” representa as posições do peso P_{Gel} em relação as bordas dos intervalos.

Foram considerados, nesta análise, os dados da bateria 3, uma vez que são associados ao conjunto de vídeos que foram produzidos com o fundo cinza, exibindo o elemento de interesse fora do contexto de qualquer aplicação.

Tabela 7 – Valores dos pesos calculados para os algoritmos Crim e Qian, seus respectivos intervalos de confiança e os pesos P_{Gel} sugeridos no método PST para avaliar segmentação em um cenário geral

			Int. Confiança (99%)			Int. Confiança (95%)			Int. Confiança (85%)		
Peso	P_{Gel}	P_{Crim}	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.
a	2,86	8,78	-16,37	33,94	dentro	-8,21	25,78	dentro	-2,84	20,40	dentro
b	4,50	9,75	-6,16	25,66	dentro	-1,00	20,50	dentro	2,40	17,10	dentro
c	4,77	0,27	-4,92	5,45	fora	-3,24	3,77	fora	-2,13	2,66	fora
d	5,82	1,71	-1,11	4,54	fora	-0,20	3,63	fora	0,41	3,02	fora
			Int. Confiança (99%)			Int. Confiança (95%)			Int. Confiança (85%)		
Peso	P_{Gel}	P_{Qian}	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.
a	2,86	25,52	-43,11	94,16	dentro	-20,86	71,89	dentro	-2,84	20,40	dentro
b	4,50	-8,90	-75,95	58,14	dentro	-54,21	36,39	dentro	2,40	17,10	dentro
c	4,77	2,46	-48,76	53,68	dentro	-32,15	37,07	dentro	-2,13	2,66	dentro
d	5,82	0,35	-15,20	15,90	dentro	-10,16	10,86	dentro	0,41	3,02	fora

Como pode ser observado na tabela 7, dois dos pesos definidos na PST encontram-se fora do intervalo de confiança, considerando os dados D_{Crim}. Isso demonstra que a métrica PST, da forma como foi concebida (capaz de avaliar a qualidade da segmentação independentemente do algoritmo utilizado), não se mostra eficiente para avaliar a segmentação produzida pelo algoritmo Crim. Em relação ao algoritmo Qian, apenas com o nível de confiança reduzido a 85%, um único peso se mostra fora do intervalo.

Verificada a impossibilidade de realizar avaliações em um cenário geral utilizando a PST, no passo seguinte, foi avaliado o desempenho do método quando ajustado para uma aplicação em especial, os sistemas de RA[1]. Importa ressaltar que a Teleconferência Imersiva trata-se de uma aplicação de RA. Como exibido na seção 3.2.2, a PST considera os mesmos artefatos (PST_{A_r}, PST_{H_i}, PST_{A_b} e PST_{H_b}), no entanto, define novos pesos, otimizados para avaliar a qualidade da segmentação no contexto desse tipo de sistema. Os resultados obtidos dessa análise podem ser visualizados na tabela 8.

Considerando as avaliações associadas ao algoritmo Crim, dois pesos P_{Gel} (segunda

[1]O ambiente de RA simulado nos experimentos subjetivos realizados no desenvolvimento da métrica PST tem as características do apresentado em Marichal et al. (2002).

Tabela 8 – Valores dos pesos calculados para os algoritmos Crim e Qian, seus respectivos intervalos de confiança e os pesos sugeridos no método PST para avaliar segmentação em Teleconferência Imersiva

			Int. Confiança (99%)			Int. Confiança (95%)			Int. Confiança (85%)		
Peso	P_{Gel}	P_{Crim}	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.
a	6,71	6,33	-1,06	13,71	dentro	0,93	11,72	dentro	2,46	10,20	dentro
b	8,39	17,63	2,38	32,89	dentro	6,48	28,79	dentro	9,64	25,63	fora
c	12,57	-2,70	-7,36	1,96	fora	-6,10	0,71	fora	-5,14	-0,25	fora
d	8,74	9,49	4,72	14,26	dentro	6,00	12,98	dentro	6,99	11,99	dentro
			Int. Confiança (99%)			Int. Confiança (95%)			Int. Confiança (85%)		
Peso	P_{Gel}	P_{Qian}	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.
a	6,71	23,95	-3,65	51,56	dentro	3,76	44,15	dentro	9,48	38,43	fora
b	8,39	-8,95	-27,72	9,81	dentro	-22,68	4,78	fora	-18,80	0,89	fora
c	12,57	-12,25	-24,04	-0,47	fora	-20,87	-3,63	fora	-18,43	-6,07	fora
d	8,74	6,12	2,60	9,64	dentro	3,54	8,70	fora	4,27	7,97	fora

coluna da tabela) encontram-se fora do intervalo com um nível de confiança de 85% e, com níveis de confiança iguais a 99% e 95%, apenas um dos pesos se mostra fora do intervalo. Considerando os dados D_{Qian}, todos os pesos encontram-se fora do intervalo quando o nível de confiança é igual a 85% e ainda existem pesos fora dos seus respectivos intervalos nos demais níveis.

Esses resultados demonstram que a métrica PST, ajustada para avaliar a segmentação no contexto das aplicações de RA, não se mostra eficiente, sobretudo quando se considera os dados associados ao algoritmo Qian. Foram utilizados, nesta análise, os dados das baterias 1, 3 e 4, em que o fundo dos vídeos avaliados simulam um sistema de RA voltado a Teleconferência Imersiva.

Retomando a discussão apresentada no início deste capítulo, existe a possibilidade de diferenciar as aplicações de acordo com determinadas características comuns. Um exemplo de característica trata-se do comportamento do elemento de interesse, que pode ser conhecido *a priori* em determinados sistemas. Considerando que o avatar permaneça sempre na mesma distância em relação à câmera (no caso, posicionado próximo a ela), restringiu-se os dados analisados para que fossem considerados apenas os vídeos que representassem tal comportamento.

Os resultados exibidos na tabela 9 mostram que existem pesos fora dos intervalos com todos os níveis de confianças testados, tanto para as avaliações associadas ao algoritmo

Tabela 9 – Valores dos pesos calculados para os algoritmos Crim e Qian, seus respectivos intervalos de confiança e os pesos sugeridos no método PST para avaliar segmentação em sistemas de Teleconferência Imersiva com determinada característica

			Int. Confiança (99%)			Int. Confiança (95%)			Int. Confiança (85%)		
Peso	P_{Gel}	P_{Crim}	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.
a	6,71	-11,55	-30,09	6,99	dentro	-24,85	1,74	fora	-20,97	-2,14	fora
b	8,39	10,57	-6,32	27,47	dentro	-1,54	22,69	dentro	1,99	19,16	dentro
c	12,57	3,16	-5,92	12,25	fora	-3,35	9,68	fora	-1,45	7,78	fora
d	8,74	7,72	2,05	13,39	dentro	3,65	11,79	dentro	4,84	10,60	dentro
			Int. Confiança (99%)			Int. Confiança (95%)			Int. Confiança (85%)		
Peso	P_{Gel}	P_{Qian}	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.
a	6,71	17,15	0,12	34,18	dentro	4,93	29,36	dentro	8,50	25,80	fora
b	8,39	-9,46	-21,15	2,24	fora	-17,85	-1,07	fora	-15,40	-3,52	fora
c	12,57	-8,79	-15,69	-1,89	fora	-13,74	-3,84	fora	-12,29	-5,28	fora
d	8,74	7,35	5,18	9,52	dentro	5,79	8,90	dentro	6,25	8,45	fora

Crim (D_{Crim}) quanto para as associadas ao algoritmo Qian (D_{Qian}.) Com o nível de confiança igual a 85%, todos os pesos associados ao algoritmo Qian encontram-se fora de seus respectivos intervalos.

Ainda que esta pesquisa tenha elegido os algoritmos Qian e Crim como objetos de análise, com o objetivo de comprovar os resultados obtidos, uma análise adicional foi realizada para verificar a aplicabilidade da PST na avaliação da qualidade da segmentação produzida por outros dois algoritmos, Sanc e Stau. Ambos os algoritmos, detalhados nas seções I.3.3 e I.3.4, respectivamente, possuem características que os tornam aplicáveis em sistema de Teleconferência Imersiva. Os resultados obtidos desta análise são apresentados na tabela 10.

Quando o nível de confiança é igual a 85%, ambos os algoritmos possuem pesos fora de seus respectivos intervalos. Para as avaliações associadas ao algoritmo Stau, um dos pesos encontra-se fora do intervalo quando o nível de confiança é igual a 95%. Importa ressaltar que, nesta análise, foram utilizados apenas os dados da bateria 4. Esses dados representam vídeos gerados a partir de camadas de primeiro plano obtidas da execução dos algoritmos em questão. Como existem menos dados em relação à análise anterior, os intervalos de confiança tendem a ser mais abrangentes.

Tabela 10 – Valores dos pesos calculados para os algoritmos Sanc e Stau, seus respectivos intervalos de confiança e os pesos sugeridos no método PST para avaliar segmentação em sistemas de Teleconferência Imersiva com determinada característica

			Int. Confiança (99%)			Int. Confiança (95%)			Int. Confiança (85%)		
Peso	P_{Gel}	P_{Sanc}	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.
a	6,71	3,02	-81,40	87,45	dentro	-47,89	53,93	dentro	-29,58	35,63	dentro
b	8,39	14,66	-37,23	66,55	dentro	-16,63	45,95	dentro	-5,38	34,70	dentro
c	12,57	-6,51	-38,60	25,57	dentro	-25,86	12,84	dentro	-18,90	5,88	fora
d	8,74	11,51	-15,44	38,47	dentro	-4,74	27,77	dentro	1,11	21,92	dentro
			Int. Confiança (99%)			Int. Confiança (95%)			Int. Confiança (85%)		
Peso	P_{Gel}	P_{Stau}	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.	Esq.	Dir.	Pos.
a	6,71	-46,49	-199,26	106,29	dentro	-138,62	45,65	dentro	-105,49	12,52	dentro
b	8,39	52,20	-46,02	150,41	dentro	-7,03	111,42	dentro	14,26	90,13	fora
c	12,57	12,44	-28,77	53,65	dentro	-12,41	37,29	dentro	-3,48	28,36	dentro
d	8,74	-5,61	-29,38	18,16	dentro	-19,95	8,72	fora	-14,79	3,57	fora

5.2 Das Avaliações Subjetivas para a Métrica Objetiva

A métrica PST, proposta em Gelasca e Ebrahimi (2009), como demostrado na seção 5.1, não se mostrou eficiente para avaliar a qualidade da segmentação produzida pelos algoritmos descritos na seção 4.3, quando utilizados em sistemas de Teleconferência Imersiva com as características daquele descrito na seção 3.3. Desse modo, para que uma forma objetiva de ajustar os parâmetros dos algoritmos citados seja encontrada, uma nova métrica deve ser desenvolvida. Com base nos dados obtidos nas avaliações subjetivas, as análises que possibilitaram a definição dessa métrica são apresentadas nesta seção.

5.2.1 Dependência do Algoritmo e Ordenação dos Artefatos

Dado que a métrica PST propõe fornecer objetivamente o nível de incômodo provocado pelos erros de segmentação, independentemente do algoritmo, a primeira análise realizada em relação ao novo conjunto de artefatos (os definidos na seção 4.5 somados aos propostos na PST) consiste em descobrir se os que causam maior incômodo permanecem os mesmos e não variam de acordo com o algoritmo. Como diferentes algoritmos são utilizados para extrair as camadas de primeiro plano utilizadas para gerar o vídeo avaliado nos experimentos subjetivos, os dados associados à segmentação realizada pelos algoritmos

Crim (D_{Crim}) e Qian (D_{Qian}) foram analisados separadamente.

Para selecionar e ordenar os artefatos de acordo com o nível de incômodo por eles provocados, uma busca gulosa (BENDALL; MARGOT, 2006), em vários passos, foi realizada, depois que um conjunto de artefatos era avaliado por meio de regressão linear. De forma mais detalhada, esse processo corresponde aos procedimentos descritos nos próximos parágrafos.

Considerando os dados resultantes do cálculo da ocorrência dos artefatos nos vídeos analisados organizados em tabela, as colunas representam os artefatos e as linhas representam os vídeos. Cada linha da tabela consiste no valor calculado para o artefato referente àquela coluna. Os dados da avaliação subjetiva, obtidos pela média ITU-R, são representado por uma coluna, uma vez que existe um valor para cada vídeo analisado.

Os dados da tabela foram divididos em dois grupos, utilizando como critério o algoritmo executado para segmentar a camada de primeiro plano que gerou o vídeo avaliado (D_{Crim} e D_{Qian}). Os dados da coluna da avaliação subjetiva foram divididos da mesma forma.

Para cada grupo de dados[2], aplicou-se uma regressão linear entre os artefatos e os dados da média da avaliação subjetiva, realizando uma busca gulosa nos parâmetros. Obteve-se, desse modo, um peso e um erro relacionado a cada artefato analisado. A partir desse erro, o artefato que causa maior incômodo pode ser identificado.

No passo seguinte, o mesmo procedimento é realizado e uma nova regressão linear é aplicada entre os artefatos e os dados da média das avaliações subjetivas, realizando nova busca gulosa. No entanto, os erros são calculados, nesse segundo passo, considerando grupos de dois artefatos na regressão. Nos passos seguintes, são considerados grupos de três, quatro, cinco, até que um número determinado de artefatos do conjunto analisado seja atingido.

Nesta análise foram incluídos os dados associados aos vídeos com planos de fundo similares ao cenário da aplicação de Teleconferência Imersiva e não houve restrições quanto ao conteúdo em relação ao comportamento do elemento de interesse. Os dados das baterias 1 e 3, que representam vídeos gerados a partir de camadas de primeiro plano segmentadas pelos algoritmos Crim e Qian foram considerados nesta análise. Na tabela 11, são

[2]A técnica *leave-one-out* foi utilizada em algumas análises em razão de haver uma base de dados reduzida.

exibidos, ordenadamente, os 10 (dez) atributos que causam maior incômodo ao usuário, separados por algoritmo.

Tabela 11 – Artefatos que causam maior incômodo ao usuário resultado da análise dos dados das aplicações de Teleconferência Imersiva em que não há restrições quanto ao comportamento do avatar

Ord.	D_{Crim}	Descrição do Artefato	D_{Qian}	Descrição do Artefato
1	$E_{\mathcal{N}}$	Erros no Elemento de Interesse (médio)	$E_{\mathcal{N}}$	Erros no Elemento de Interesse (médio)
2	$D_{\mathcal{P}out(110)}$	Erros no Plano de Fundo distantes do Elemento de Interesse mais que 110 pixels	$T_{\mathcal{N}(70)}$	Erros Temporais no elemento de interesse em menos de 70% dos quadros
3	$T_{\mathcal{N}(60)}$	Erros Temporais no elemento de interesse em menos de 60% dos quadros	$T_{\mathcal{P}(50)}$	Erros Temporais no plano de fundo em menos de 50% dos quadros
4	PST_{A_b}	Erros no plano de fundo conectados no elemento de interesse	$B_{\mathcal{N}small(5)}$	Componentes conectados no elemento de interesse menores que 5 pixels
5	PST_{H_i}	Erros no elemento de interesse desconectados da borda	$FT_{\mathcal{N}}$	Falsos Blobs Temporais Falsos Negativo
6	$T_{\mathcal{P}(70)}$	Erros Temporais no plano de fundo em menos de 70% dos quadros	PST_{H_b}	Erros no elemento de interesse conectados na borda
7	$T_{\mathcal{P}(80)}$	Erros Temporais no plano de fundo em menos de 80% dos quadros	$B_{\mathcal{N}small(15)}$	Componentes conectados no elemento de interesse menores que 15 pixels
8	$FS_{\mathcal{N}}s$	Falsos Blobs Falso Negativos	$D_{\mathcal{P}in}(120)$	Falsos Positivos distantes do elemento de interesse até 120 pixels
9	$FS_{\mathcal{N}}g$	Falsos Blobs Falso Negativos Gaussiano	$E_{\mathcal{T}}$	Erro Total (médio)
10	$B_{\mathcal{P}large(5)}$	Componentes conectados no elemento de interesse maiores que 5 pixels	$B_{\mathcal{N}small(10)}$	Componentes conectados no elemento de interesse menores que 10 pixels

O artefato $E_{\mathcal{N}}$ (Erros no Elemento de Interesse), como pode ser observado, apresenta-se como o que causa maior incômodo, tanto para os dados originados do algoritmo Crim quanto para os do algoritmo Qian. Os três artefatos seguintes, por sua vez, variam de acordo com o algoritmo utilizado, embora artefatos relacionados a erros temporais concentrados no elemento de interesse se mostrem presentes em ambas as análises. De modo geral, os artefatos relacionados a erros visíveis no elemento de interesse são predominantes na tabela 11.

Apesar de haver coincidência na ocorrência de artefatos, os resultados indicam que uma métrica objetiva que possa ser utilizada neste contexto deve ser dependente do algoritmo, e não generalizada para um domínio de aplicação, como propõe a métrica PST.

Considerando a discussão sobre características dos sistemas de Teleconferência Imersiva apresentada no início deste capítulo, o conhecimento sobre o comportamento do elemento de interesse na cena também pode ser utilizado para que uma métrica específica seja obtida para os sistemas com essa característica.

Entre os sistemas em que o avatar permanece na mesma distância em relação a câmera durante todo o tempo de execução da aplicação, analisou-se os casos em que esse elemento se encontra próximo da câmera, como o exibido na figura 18(a), eliminando-se da análise os dados em que os vídeos associados não possuíam essa característica. Os dados das baterias 1 e 3, que representam tais situações, foram utilizados nesta análise cujos resultados são exibidos na tabela 12.

Quando comparados aos resultados da análise anterior, nos dados D_{Crim} pode ser observado que os artefatos que causam mais incômodo considerando essa característica da aplicação, apesar de haver mudanças entre eles, essencialmente não existe grande variação. Isso também ocorre em relação ao algoritmo Qian, ainda que, nesse algoritmo, a variação seja maior. Assim como na análise anterior (tabela 11), os artefatos relacionados a erros visíveis no elemento de interesse são predominantes.

5.2.2 Quantidade de Artefatos

Outro fator relevante na análise dos artefatos consiste na definição de quantos deles serão considerados na composição da métrica. Poucos artefatos podem não ser suficiente para representar corretamente a percepção do usuário, ao passo que um número grande pode ser desnecessário ou, ainda, não ser ideal na formalização da métrica.

Para que um número adequado de artefatos fosse encontrado, à medida que os conjuntos eram testados na regressão linear, os erros em relação à avaliação subjetiva eram armazenados. Como a regressão era realizada com combinações de dois, três, quatro artefatos e assim sucessivamente, torna-se possível relacionar a ocorrência do erro em relação à média ITU-R (erro médio), obtida do experimento subjetivo, com cada conjunto.

Tabela 12 – Artefatos que causam maior incômodo ao usuário resultado da análise dos dados das aplicações de Teleconferência Imersiva em que o avatar permanece sempre próximo da câmera

Ord.	D_{Crim}	Descrição do Artefato	D_{Qian}	Descrição do Artefato
1	$E_{\mathcal{N}}$	Erros no Elemento de Interesse (médio)	PST_{H_b}	Erros no elemento de interesse conectados na borda
2	PST_{A_b}	Erros no plano de fundo conectados no elemento de interesse	PST_{H_i}	Erros no elemento de interesse desconectados da borda
3	PST_{A_r}	Erros no plano de fundo desconectados do elemento de interesse	$B_{\mathcal{N}large(5)}$	Componentes conectados no elemento de interesse maiores que 5 pixels
4	PST_{H_i}	Erros no elemento de interesse desconectados da borda	$B_{\mathcal{N}large(15)}$	Componentes conectados no elemento de interesse maiores que 15 pixels
5	$B_{\mathcal{P}small(5)}$	Componentes conectados no plano de fundo menores que 5 pixels	$B_{\mathcal{N}small(20)}$	Componentes conectados no elemento de interesse menores que 20 pixels
6	$B_{\mathcal{N}large(5)}$	Componentes conectados no elemento de interesse maiores que 5 pixels	PST_{A_b}	Erros no plano de fundo conectados no elemento de interesse
7	$B_{\mathcal{P}small(15)}$	Componentes conectados no plano de fundo menores que 15 pixels	$B_{\mathcal{N}large(10)}$	Componentes conectados no elemento de interesse maiores que 10 pixels
8	$B_{\mathcal{N}large(10)}$	Componentes conectados no elemento de interesse maiores que 10 pixels	PST_{A_r}	Erros no plano de fundo desconectados do elemento de interesse
9	$FT_{\mathcal{N}}s$	Falsos Blobs Temporais Falso Negativos	$T_{\mathcal{N}(80)}$	Erros Temporais no elemento de interesse em menos de 40% dos quadros
10	$T_{\mathcal{N}(40)}$	Erros Temporais no elemento de interesse em menos de 40% dos quadros	$E_{\mathcal{N}}$	Erros no Elemento de Interesse (médio)

O gráfico que confronta a quantidade de artefatos, considerando D_{Crim} e D_{Qian}, com o erro em relação à média ITU-R pode ser visualizado na figura 21.

Como pode ser observado, o erro médio se torna maior à medida que a quantidade de artefatos considerados para representar os erros objetivos na correlação com os dados subjetivos se torna muito grande[3]. Uma métrica que considere apenas 1 (um) artefato apresenta o menor erro médio para o algoritmo Crim, ainda que 6 (seis) deles produzam

[3]Importa ressaltar que a técnica *leave-one-out* foi utilizada nas análises, uma vez que a quantidade de dados é reduzida.

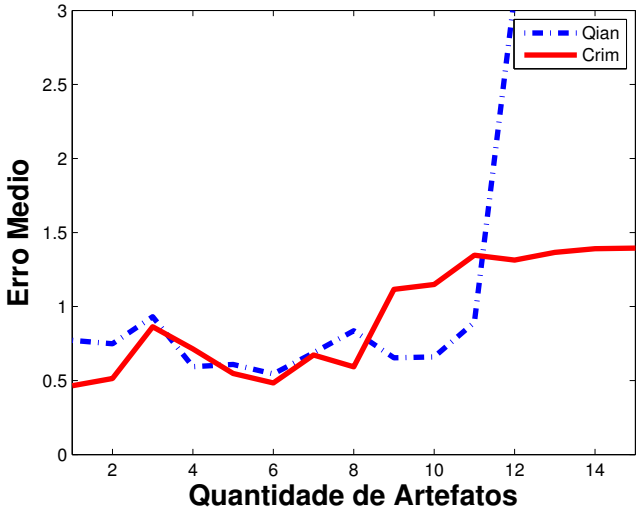


Figura 21 – Gráfico confrontando a quantidade de artefatos e o erro médio, resultado da análise dos dados das aplicações de Teleconferência Imersiva em um cenário sem restrições quanto ao comportamento do avatar

praticamente o mesmo erro. Para o algoritmo Qian, 6 (seis) artefatos são necessários para produzir o menor erro médio.

Da mesma forma, o gráfico que confronta a quantidade de artefatos com o erro em relação à avaliação subjetiva foi gerado com o objetivo de encontrar quantos deles serão considerados na métrica, quando aplicações em que o comportamento do avatar (permanece sempre próximo a câmera) é conhecido *a priori*. Esse gráfico pode ser visualizado na figura 22.

Diferentemente dos resultados apresentados no gráfico da figura 21, a figura 22 mostra que, para os dados D_{Crim}, houve variação na quantidade de artefatos necessários para definição da métrica. De acordo com o gráfico, o menor erro médio é obtido quando o terceiro artefato é adicionado ao conjunto.

Um fator a ser considerado nesta última análise, no entanto, pode haver influenciado a diferença entre os resultados apresentados pelos gráficos das figura 21 e 22. Quando se considerou o comportamento do elemento de interesse como característica da aplicação, os dados dos vídeos em que essa característica não se apresentava foram retirados da análise.

A diminuição da quantidade de artefatos necessários para compor a métrica, obser-

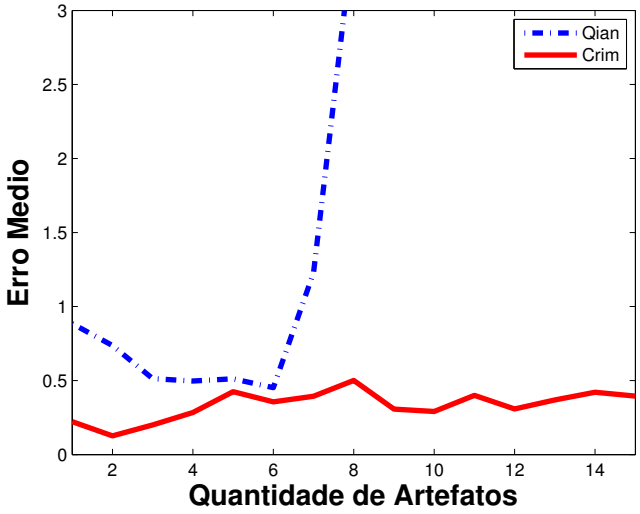


Figura 22 – Gráfico confrontando a quantidade de artefatos e o erro médio resultado da análise dos dados associados a Teleconferência Imersiva em que uma característica específica, o comportamento do elemento de interesse, foi considerado na análise dos dados

vada na figura 22, pode ter sido causada pela diminuição do conjunto de dados e não em decorrência da troca de aplicação. Uma nova análise foi realizada com o objetivo de testar essa hipótese.

Para isso, foi reproduzida a análise cujo resultado foi exibido no gráfico da figura 21 utilizando a mesma quantidade de dados da análise em que os resultados são mostrados na figura 22. O gráfico com esses resultados pode ser visualizado na figura 23.

Como pode ser observado no gráfico, a quantidade ideal de artefatos, em relação aos dados associados ao algoritmo Crim, permaneceu inalterada, ao passo que a redução da base de dados resultou na diminuição da quantidade ideal de artefatos em relação aos dados associados ao algoritmo Qian. Desse modo, para a definição da quantidade ideal de artefatos pode ser necessária maior quantidade de dados.

5.2.3 Transferência de Pesos e Artefatos

No experimento realizado na seção 5.1, dois conjuntos de pesos ideais associados aos algoritmos P_{Crim} e P_{Qian} foram calculados. O teste estatístico realizado na mesma seção mostrou que esses pesos são significativamente diferentes de P_{Gel} (conjunto de pesos definidos na métrica PST). O passo seguinte consiste em avaliar o quanto cada conjunto

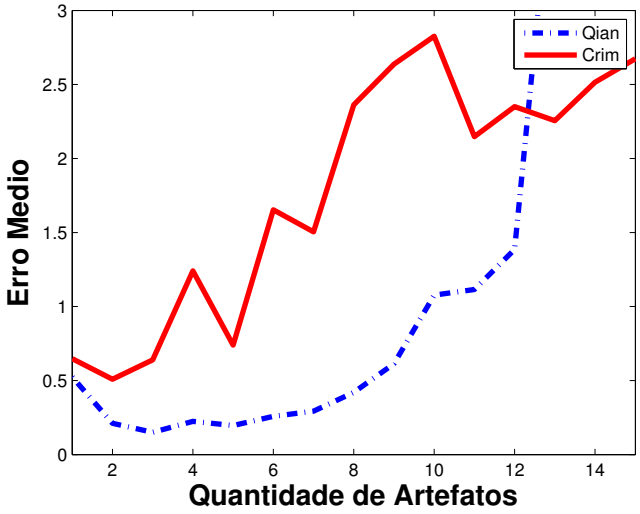


Figura 23 – Gráfico confrontando a quantidade de artefatos com o erro médio. Nesta análise foram considerados os dados das aplicações de Teleconferência Imersiva em um cenário sem restrições quanto ao comportamento do avatar e com a base de dados reduzida

de pesos influencia a predição da avaliação.

Na métrica PST foram definidos 4 (quatro) artefatos (PST_{A_r}, PST_{A_b}, PST_{H_i} e PST_{H_b}) e, associados a esses artefatos, foram definidos pesos que foram ajustados para um aplicação específica, inclusive os sistemas de RA P_{Gel}. Desse modo, de acordo com a métrica PST, existem um conjunto de artefatos que causam maior incômodo ao usuário e um conjunto de pesos associados a esses artefatos.

Na seção 5.2.1, foram identificados os artefatos que causam maior incômodo, considerando separadamente os dados relacionados aos algoritmos Crim e Qian. Nesta análise, esses dados são denominados A_{Crim} e A_{Qian}.

Considerando cada artefato do conjunto A_{Qian}, associados aos pesos P_{Qian}, calcula-se os erros – aqui denominados 1º conjunto de erros – em relação aos valores obtidos do processo de avaliação subjetiva (média ITU-R), considerando os dados do algoritmo Qian (D_{Qian}). De forma similar, utilizando o conjunto A_{Crim}, associado ao conjunto de pesos P_{Qian}, calculam-se os erros (2º conjunto de erros) em relação aos dados (D_{Qian}).

Aplicando-se o teste “t de Student” nos dois conjuntos de erros resultantes, verificou-se a rejeição ou não da hipótese nula com 3 (três) diferentes níveis de significância (1%, 5% e 15%). O teste descrito no parágrafo anterior, considerando 1% de significância,

corresponde a primeira linha da tabela 13.

De forma similar, foram realizados 24 (vinte e quatro) testes, combinando dois conjuntos de erros obtidos de uma relação Artefatos/Pesos/Dados, como pode ser visualizado na tabela 13. Nesses testes, foram considerados os dados correspondentes a aplicações de Teleconferência Imersiva em que não existem restrições quanto ao comportamento do elemento de interesse.

Tabela 13 – Testes "t de Student" aplicados em conjuntos de erros obtidos das combinações Artefatos/Pesos/Dados, considerando as aplicações de Teleconferência Imersiva em que não existe restrições quanto ao comportamento do elemento de interesse

1º Conjunto de Erros			2º Conjunto de Erros				
Artef.	Pesos	Dados	Artef.	Pesos	Dados	Signif.	Hipótese
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Qian}	D_{Qian}	1%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	P_{Qian}	D_{Qian}	1%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Crim}	D_{Qian}	1%	rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	P_{Gel}	D_{Qian}	1%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Crim}	D_{Crim}	1%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Crim}	D_{Crim}	1%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Qian}	D_{Crim}	1%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Gel}	D_{Crim}	1%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Qian}	D_{Qian}	5%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	P_{Qian}	D_{Qian}	5%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Crim}	D_{Qian}	5%	rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	A_{Gel}	D_{Qian}	5%	rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Crim}	D_{Crim}	5%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Crim}	D_{Crim}	5%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Qian}	D_{Crim}	5%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Gel}	D_{Crim}	5%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Qian}	D_{Qian}	15%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	P_{Qian}	D_{Qian}	15%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Crim}	D_{Qian}	15%	rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	P_{Gel}	D_{Qian}	15%	rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Crim}	D_{Crim}	15%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Crim}	D_{Crim}	15%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Qian}	D_{Crim}	15%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Gel}	D_{Crim}	15%	não rejeitada

Como pode ser observado, existem rejeições da hipótese nula em todos os percentuais de significância testados. Conclui-se, por exemplo, que existem diferenças significativas (1% de significância) entre os erros obtidos da combinação dos artefatos de Crim, pesos de Crim e dados de Qian, quando comparados aos erros obtidos da combinação dos artefatos de Qian e pesos de Qian testados sobre os dados de Qian. Com 5% e 15% de significância, dois testes apresentam rejeição da hipótese nula.

Assim como nas análises das seções anteriores, os mesmos testes foram reproduzidos considerando apenas os dados em que os vídeo simulam aplicações em que o comportamento do elemento de interesse (permanece sempre próximo a câmera) é conhecido *a priori*. Os resultados dessa análise podem ser visualizados na tabela 14.

Como pode ser observado, ainda que uma determinada característica da aplicação seja considerada, as hipóteses rejeitadas, exibidas na tabela 14, essencialmente, são similares às observadas na análise anterior (tabela 13).

5.2.4 Análises Individuais

Outro tipo de análise realizada nesta pesquisa trata da influência de cada artefato considerando individualmente os participantes dos experimentos subjetivos. O objetivo dessa análise consiste em verificar se os artefatos que causam maior incômodo obtidos da média das avaliações permanecem os mesmos, se forem considerados suas ocorrências nas avaliações individuais.

Participaram dos experimentos subjetivos 39 (trinta e nove) voluntários, que emitiram opiniões em uma, duas ou três baterias de teste. Cada bateria de teste necessita de pelo menos 15 (quinze) avaliadores, como sugerido nas recomendações da ITU-R (ITU-R, 2009), portanto, esse número de avaliadores foi utilizado em cada bateria. O mesmo modelo de equipamento foi mantido em todas as baterias e o local em que se realizaram os experimentos foi configurado conforme as recomendações da ITU (ITU-R, 2009).

A baterias 1, 2 e 3 foi composta de 24 (vinte e quatro) vídeos a serem analisados, ao passo que, na bateria 4, 18 (dezoito) vídeos foram exibidos. Desse modo, nas três primeiras baterias foram emitidas 24 (vinte e quatro) opiniões de 15 (quinze) avaliadores, somando 360 (trezentos e sessenta) avaliações por bateria. Essas avaliações somadas às

Tabela 14 – Testes “t de Student” aplicados em conjuntos de erros obtidos das combinações Artefatos/Pesos/Dados, considerando as aplicações de RA em que o elemento de interesse permanece sempre na mesma distância em relação a câmera

1º Conjunto de Erros			2º Conjunto de Erros				
Artef.	Pesos	Dados	Artef.	Pesos	Dados	Signif.	Hipótese
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Qian}	D_{Qian}	1%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	P_{Qian}	D_{Qian}	1%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Crim}	D_{Qian}	1%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	P_{Gel}	D_{Qian}	1%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Crim}	D_{Crim}	1%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Crim}	D_{Crim}	1%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Qian}	D_{Crim}	1%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Gel}	D_{Crim}	1%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Qian}	D_{Qian}	5%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	P_{Qian}	D_{Qian}	5%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Crim}	D_{Qian}	5%	rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	A_{Gel}	D_{Qian}	5%	rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Crim}	D_{Crim}	5%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Crim}	D_{Crim}	5%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Qian}	D_{Crim}	5%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Gel}	D_{Crim}	5%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Qian}	D_{Qian}	15%	rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	P_{Qian}	D_{Qian}	15%	não rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Crim}	P_{Crim}	D_{Qian}	15%	rejeitada
A_{Qian}	P_{Qian}	D_{Qian}	A_{Gel}	P_{Gel}	D_{Qian}	15%	rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Crim}	D_{Crim}	15%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Crim}	D_{Crim}	15%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Qian}	P_{Qian}	D_{Crim}	15%	não rejeitada
A_{Crim}	P_{Crim}	D_{Crim}	A_{Gel}	P_{Gel}	D_{Crim}	15%	não rejeitada

270 (duzentos e setenta) avaliações da bateria 4 – 18 (dezoito) avaliações de 15 (quinze) voluntários – totalizaram 1350 (mil e trezentos e cinquenta) avaliações.

Nesta análise foram utilizadas apenas as avaliações referentes aos dados D_{Crim} e D_{Qian} que simulavam o ambiente de Teleconferência Imersiva. Cada grupo de dados é composto por 540 (quinhentos e quarenta) avaliações. Considerando a média das avaliações do grupo, foram identificados os 4 (quatro) artefatos que causam maior incômodo em

cada um deles.

Em seguida, as 540 (quinhentos e quarenta) avaliações relacionadas ao algoritmo Crim foram agrupadas por avaliador e, para cada conjunto de avaliações de um mesmo avaliador, foi encontrado os artefatos que causam maior incômodo. Obteve-se, portanto, um conjunto de 4 (quatro) artefatos (que causam maior incomodo) para cada avaliador. Analisando todos esses conjuntos, foi calculada a frequência de cada artefato. Os resultados obtidos desta análise são mostrados na tabela 15.

Tabela 15 – Artefatos que causam maior incômodo aos usuários dos grupos Crim e Qian, obtidos da média das avaliações do grupo e da frequência dos atributos nas avaliações individuais, considerando aplicações de Teleconferência Imersiva sem restrições relacionadas as características do sistema

Algoritmo Crim				
Artefatos (Média das avaliações subjetivas)		Artefatos (Mais frequentes)		Freq.
PST_{A_r}	Erros no plano de fundo desconectados do elemento de interesse	PST_{A_r}	Erros no plano de fundo desconectados do elemento de interesse	21
PST_{A_b}	Erros no plano de fundo conectados no elemento de interesse	PST_{A_b}	Erros no plano de fundo conectados no elemento de interesse	19
$T_{\mathcal{N}(80)}$	Erros Temporais no elemento de interesse em menos de 80% dos quadros	$T_{\mathcal{N}(80)}$	Erros Temporais no elemento de interesse em menos de 80% dos quadros	14
$E_{\mathcal{N}}$	Erros no Elemento de Interesse (médio)	$B_{\mathcal{P}large(10)}$	Componentes conectados no plano de fundo maiores que 10 pixels	13
Algoritmo Qian				
Artefatos (Média das avaliações subjetivas)		Artefatos (Mais frequentes)		Freq.
$E_{\mathcal{N}}$	Erros no Elemento de Interesse (médio)	PST_{H_b}	Erros no elemento de interesse conectados na borda	20
PST_{A_r}	Erros no plano de fundo desconectados do elemento de interesse	PST_{A_r}	Erros no plano de fundo desconectados do elemento de interesse	19
PST_{A_b}	Erros no plano de fundo conectados no elemento de interesse	$E_{\mathcal{N}}$	Erros no Elemento de Interesse (médio)	18
$B_{\mathcal{N}small(5)}$	Componentes conectados no elemento de interesse menores que 5 pixels	PST_{H_i}	Erros no elemento de interesse desconectados da borda	17

Como pode ser observado na tabela, em relação ao grupo Crim, não existem alterações nos primeiros artefatos, que se mantêm na mesma ordem. Em relação ao algoritmo

Qian, as variações são maiores.

Assim como em análises anteriores, o mesmo procedimento foi realizado considerando apenas os dados das aplicações de Teleconferência Imersiva em que o avatar permanece sempre próximo da câmera. Os resultados dessa análise se mostraram semelhantes aos da tabela 15.

5.3 Definição da Métrica Objetiva

Diante das várias análises realizadas nesta seção, o passo seguinte consiste na definição da própria métrica. Os resultados obtidos neste trabalho mostraram que uma solução genérica, que possa ser utilizada em um domínio de aplicação, pode não ser precisa. A métrica objetiva deve ser específica não apenas para uma aplicação, mas para um determinado algoritmo de segmentação. Além disso, pode-se, ainda, refiná-la considerando a característica da aplicação.

A partir dessas observações, a métrica M, derivada do método subjetivo proposto neste trabalho, pode ser definida de acordo com a equação

$$M_{(Alg,Apl,Car)} = \sum_{i=1}^{I} (pes_i \times art_i)_{Alg} \tag{41}$$

onde Alg representa o algoritmo de segmentação em que os parâmetros devem ser ajustados, Apl consiste no domínio de aplicação em que as camadas de primeiro plano segmentadas de pelo algoritmo serão utilizadas e Car trata-se de uma característica da aplicação, que pode ser conhecida *a priori*. Os pesos são denotados por um vetor $pes = (pes_1, pes_2, \ldots, pes_i, \ldots, pes_I)$, assim como os artefatos $(art_1, art_2, \ldots, art_i, \ldots, art_I)$.

Aplicando-se nova a métrica no contexto das aplicações de RA voltadas a Teleconferência Imersiva, em que não existe o conhecimento prévio de determinada característica do sistema, pode ser definida uma métrica para ajuste de parâmetros do algoritmo Crim de acordo com a equação

$$M_{(Crim,RA)} = a \times E_{\mathcal{P}} + b \times D_{\mathcal{P}out(110)} + c \times T_{\mathcal{N}(60)} + d \times PST_{A_b} + e \times PST_{H_i} + f \times T_{\mathcal{P}(70)} \tag{42}$$

onde $a = -0,051$, $b = 0$, $c = -6,319$, $d = 17,938$, $e = -2,175$ e $f = 2,143$ e represen-

tam os pesos obtidos no processo de correlação da avaliação subjetiva com os artefatos, quando 6 (seis) artefatos são considerados na regressão linear (seção 5.2.1). A quantidade de artefatos considerado na métrica, nesse caso 6 (seis), foi analisada na seção 5.2.2. De forma similar, a equação 43 define a métrica que ajusta parâmetros do algoritmo Qian.

$$M_{(Qian,RA)} = a \times E_{\mathcal{P}} + b \times T_{\mathcal{N}(70)} + c \times T_{\mathcal{P}(50)} + d \times B_{\mathcal{N}small(5)} + +e \times FT_{\mathcal{P}} + f \times PST_{H_b} \quad (43)$$

onde $a = -0,022$, $b = 0,003$, $c = 10,872$, $d = 0$, $e = 0$ e $f = 5,400$.

As discussões apresentadas no início deste capítulo sobre a possibilidade de identificar características da aplicação, que podem ser conhecidas *a priori*, permite que uma métrica objetiva específica seja produzida para avaliar a qualidade de algoritmos de segmentação em tais aplicações. Desse modo, considerando as aplicações de Teleconferência Imersiva em que o elemento de interesse permanece sempre próximo à câmera, a métrica que avalia a qualidade do algoritmo Crim pode ser definida pela equação

$$\begin{aligned} M_{(Crim,RA,Prox)} = \quad & a \times E_{\mathcal{P}} + b \times PST_{A_b} + c \times PST_{A_r} + d \times PST_{H_i} \\ & + e \times B_{\mathcal{P}small(5)} + f \times B_{N\mathcal{N}}large(5) \end{aligned} \quad (44)$$

onde $a = -0,153$, $b = 14,33$, $c = -17,334$, $d = 6,972$, $e = 0$ e $f = 0,002$. A métrica objetiva que avalia o algoritmo Qian pode ser obtida da mesma forma, conforme a equação

$$\begin{aligned} M_{(Qian,RA,Prox)} = \quad & a \times PST_{H_b} + b \times PST_{H_i} + c \times B_{\mathcal{N}large(5)} + d \times B_{\mathcal{N}large(15)} \\ & + e \times B_{\mathcal{P}small(5)} + f \times PST_{A_b} \end{aligned} \quad (45)$$

onde $a = 8.824$, $b = 0,059$, $c = 0,027$, $d = -0,030$, $e = 0,002$ e $f = -4,053$. A quantidade de artefatos considerado permaneceu inalterada em relação as métricas anteriores, uma vez que a diminuição no número considerado ideal apresentada nos gráficos das figuras 22 e 23 ocorreu em consequência da quantidade de dados analisados.

Definidas as novas métricas, ainda cabem algumas considerações a respeito de sua utilização. Primeiramente, ressalta-se que foram considerados apenas os artefatos definidos na seção 4.5, somados aos apresentados na PST. Embora o conjunto seja grande, uma vez que se considera variações de parâmetros de alguns deles, ainda existem inúmeras possibilidades de representar formas de erros.

A quantidade de artefatos mostrou-se variante conforme a quantidade de dados anali-

sados. Em consequência disso, os 6 (seis) considerados na métrica podem não representar a quantidade que torna a métrica mais precisa possível. Além disso, a busca gulosa realizada em vários passos não se mostrou uma solução ótima (por exemplo, $D_{\mathcal{P}out(110)}$ revela-se irrelevante após a adição de mais artefatos nos termos que representa a métrica).

A proximidade entre P_{Gel} e P_{Crim} pode indicar uma direção para definir uma métrica que avalie a qualidade da segmentação produzida por grupos de algoritmos. Artefatos da PST, por exemplo, foram eleitos entre os que causam mais incômodo relacionados ao algoritmo Crim em todas as suas métricas.

Finalmente, apesar de diferentes algoritmos exigirem métricas específicas, em um cenário sem restrições quanto ao comportamento do elemento de interesse ambos elegem os artefatos $E_{\mathcal{N}}$, $T_{\mathcal{N}(pc)}$ e $T_{\mathcal{P}(pc)}$, indicando um caminho para um subconjunto comum. Essa constatação foi ratificada pelos resultados das análises individuais.

Ainda que a discussão acima deva ser considerada, as métricas obtidas como resultados desta pesquisa se mostram mais eficientes que a PST, principalmente em relação ao algoritmo Qian.

6 CONCLUSÕES

O problema abordado nesta pesquisa consiste em encontrar uma forma de avaliar a qualidade da segmentação, sem que se realize experimentos subjetivos. A segmentação, neste contexto, refere-se à camada de primeiro plano obtida da execução de algoritmos que dividem cada quadro de um vídeo em duas camadas (*bilayer*). Considera-se, inclusive, que essa camada seja utilizada para composição de cenas em sistemas de Teleconferência Imersiva. Em outras palavras, apresenta-se uma métrica objetiva, dependente da aplicação, que considera a percepção do usuário na avaliação da qualidade da segmentação. Sua utilização tem como finalidade encontrar o melhor conjunto de parâmetros de determinado algoritmo.

Além da apresentação da nova métrica, demonstra-se, neste trabalho, que as encontradas na literatura não são eficientes quando utilizadas no contexto da aplicação de Teleconferência Imersiva, ainda que um dos trabalhos seja voltado às aplicações de RA. Apesar de os artefatos propostos naquele trabalho não se mostrarem totalmente irrelevantes quando analisados separadamente, quando esses artefatos são combinados conforme sugerido naquela métrica, os resultados observados mostram-se desalinhados com os obtidos nos experimentos subjetivos.

Diferentemente da que representa o estado-da-arte, a métrica desenvolvida neste trabalho foi derivada da análise de resultados de experimentos subjetivos em que os vídeos submetidos aos avaliadores foram gerados a partir de camadas de primeiro plano obtidas da execução de algoritmos de segmentação. Conforme aqui demonstrado, essa abordagem mostrou-se eficiente. O próprio método subjetivo com base no qual os experimentos foram conduzidos trata-se de uma contribuição desta pesquisa.

Finalmente, foi demostrado que para cada algoritmo de segmentação utilizado nesta pesquisa, novos artefatos podem melhor representar o nível de incômodo produzido pelos

erros de segmentação. Esses artefatos associados a respectivos pesos foram utilizados para compor a nova métrica objetiva.

Pretende-se, como trabalhos futuros, ampliar o número de experimentos subjetivos para que novos artefatos possam ser definidos e investigados. Pretende-se, ainda, analisar outras formas de correlacionar os valores obtidos das avaliações com a ocorrência dos artefatos. Novos experimentos também devem ser realizados considerando outros domínios de aplicação.

REFERÊNCIAS BIBLIOGRÁFICAS

BARRON, J.; FLEET, D.; BEAUCHEMIN, S. Performance of optical flow techniques. *International Journal of Computer Vision*, Kluwer Academic Publishers, v. 12, p. 43–77, 1994. ISSN 0920-5691.

BENDALL, G.; MARGOT, F. Greedy-type resistance of combinatorial problems. *Discrete Optimization*, v. 3, n. 4, p. 288 – 298, 2006. ISSN 1572-5286.

BERGEN, J.; BURT, P.; HINGORANI, R.; PELEG, S. A three-frame algorithm for estimating two-component image motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, Los Alamitos, CA, USA, v. 14, n. 9, p. 886–896, 1992. ISSN 0162-8828.

BERNARDES-JUNIOR, J.; NAKAMURA, R.; R.TORI. Comprehensive model and image-based recognition of hand gestures for interaction in 3D environments. *International Journal of Virtual Reality*, v. 10, p. 11–23, 2011.

BERNARDES-JUNIOR, J. L. *Modelo abrangente e reconhecimento de gestos com as mãos livres para ambientes 3D*. Tese (Doutorado) — Escola Politécnica da Universidade de São Paulo, 2010.

BIANCHI, L.; DONDI, P.; GATTI, R.; LOMBARDI, L.; LOMBARDI, P. Evaluation of a foreground segmentation algorithm for 3d camera sensors. In: FOGGIA, P.; SANSONE, C.; VENTO, M. (Ed.). *Image Analysis and Processing - ICIAP 2009*. Berlin / Heidelberg: Springer, 2009, (Lecture Notes in Computer Science, v. 5716). p. 797–806. ISBN 978-3-642-04145-7.

BLEIWEISS, A.; WERMAN, M. Fusing time-of-flight depth and color for real-time segmentation and tracking. In: *Proceedings of the Workshop on Dynamic 3D Imaging – DAGM 2009*. Berlin / Heidelberg: Springer–Verlag, 2009, (Dyn3D '09). p. 58–69. ISBN 978-3-642-03777-1.

BOYKOV, Y.; KOLMOGOROV, V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 26, n. 9, p. 1124–1137, 2004. ISSN 0162-8828.

BOYKOV, Y. Y.; JOLLY, M.-P. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. **Proceedings of the IEEE International Conference on Computer Vision**. Los Alamitos, CA, USA: IEEE Computer Society, 2001. v. 1, p. 105–112. ISBN 0-7695-1143-0.

CIPRA, B. A. An introduction to the ising model. *The American Mathematical Monthly*, Mathematical Association of America, v. 94, n. 10, p. 937–959, 1987. ISSN 00029890.

CORRÊA, C. G.; TOKUNAGA, D. M.; SANCHES, S. R. R.; NAKAMURA, R.; TORI, R. Immersive teleconferencing system based on video-avatar for distance learning. **Proceedings of the XIII Symposium on Virtual Reality – SVR 2011**. Washington, DC, USA: IEEE Computer Society, 2011. p. 197–206. ISBN 978-0-7695-4445-8.

CORREIA, P.; PEREIRA, F. Objective evaluation of video segmentation quality. *IEEE Transactions on Image Processing*, v. 12, n. 2, p. 186–200, feb 2003. ISSN 1057-7149.

COX, I. J.; HINGORANI, S. L.; RAO, S. B.; MAGGS, B. M. A maximum likelihood stereo algorithm. *Comput. Vis. Image Underst.*, Elsevier Science Inc., New York, NY, USA, v. 63, n. 3, p. 542–567, 1996. ISSN 1077-3142.

CRIMINISI, A.; CROSS, G.; BLAKE, A.; KOLMOGOROV, V. Bilayer segmentation of live video. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition – CVPR '06**. Washington, DC, USA: IEEE Computer Society, 2006. v. 1, p. 53–60. ISBN 0-7695-2597-0. ISSN 1063-6919.

CUCCHIARA, R.; GRANA, C.; PICCARDI, M.; PRATI, A. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 25, n. 10, p. 1337–1342, 2003. ISSN 0162-8828.

ELGAMMAL, A. M.; HARWOOD, D.; DAVIS, L. S. Non-parametric model for background subtraction. In: VERNON, D. (Ed.). *European Conference on Computer Vision – ECCV 2000, Part II*. London: Springer, 2000, (Lecture Notes in Computer Science, v. 1843). p. 751–767. ISBN 3-540-67686-4.

ERDEM, C. E.; SANKUR, B. Performance evaluation metrics for object-based video segmentation. **Proceedings of the X European Signal Processing Conference – EUSIPCO**. [S.l.: s.n.], 2000. v. 2, p. 917–920.

FARIAS, M. *No-Reference and Reduced Reference Video Quality Metrics: New Contributions*. Tese (Doutorado) — University of California, 2004.

FOSTER, J. The green screen handbook: Real-world production techniques. In: ____. Chichester, GB: John Wiley and Sons Ltd, 2010. cap. Mattes and Compositing Defined, p. 3–15. ISBN 0470521074.

FRIEDMAN, N.; RUSSELL, S. Image segmentation in video sequences: A probabilistic approach. **Proceedings of the 13th Conf. Uncertainty in Artificial Intelligence – UAI'97**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. p. 175–181. ISBN 1-55860-485-5.

GEIGER, D.; LADENDORF, B.; YUILLE, A. Occlusions and binocular stereo. *International Journal of Computer Vision*, Springer Netherlands, v. 14, p. 211–226, 1995. ISSN 0920-5691.

GEISS, R. M. *Visual Targed Tracking*. U. S. Patent 2010/0197399 A1. Aug 2010.

GELASCA, E.; EBRAHIMI, T. On evaluating video object segmentation quality: A perceptually driven objective metric. *IEEE Journal of Selected Topics in Signal Processing*, v. 3, n. 2, p. 319 –335, april 2009. ISSN 1932-4553.

GELASCA, E. D. *Full-reference objective quality metrics for video watermarking, video segmentation and 3d model watermarking*. Tese (Doutorado) — École Polytechnique Fédérale de Lausanne, 2005.

GIBBS, S.; ARAPIS, C.; BREITENEDER, C.; LALIOTI, V.; MOSTAFAWY, S.; SPEIER, J. Virtual studios: an overview. *IEEE Multimedia*, v. 5, n. 1, p. 18–35, Jan-Mar 1998. ISSN 1070-986X.

GOKTURK, S. B.; YALCIN, H.; BAMJI, C. A time-of-flight depth sensor – system description, issues and solutions. **Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop – CVPRW'04**. Washington, DC, USA: IEEE Computer Society, 2004. v. 3, p. 35. ISBN 0-7695-2158-4.

GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall, Inc., 2002. 793 p. ISBN 0201180758.

GREIG, D. M.; PORTEOUS, B. T.; SEHEULT, A. H. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society*, v. 51, n. 2, p. 271–279, 1989.

GVILI, R.; KAPLAN, A.; OFEK, E.; YAHAV, G. Depth keying. *SPIE Elec. Imaging*, v. 5006, p. 554–563, 2003.

HAN, B.; COMANICIU, D.; DAVIS, L. Sequential kernel density approximation through mode propagation: applications to background modeling. **Proceedings of the Asian Conference on Computer Vision – ACCV 2004**. [S.l.: s.n.], 2004.

HARRISON, C.; HUDSON, S. E. Pseudo-3d video conferencing with a generic webcam. **Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia – ISM '08**. Washington, DC, USA: IEEE Computer Society, 2008. p. 236–241. ISBN 978-0-7695-3454-1.

IDDAN, G. J.; YAHAV, G. Three-dimensional imaging in the studio and elsewhere. **Proceedings of the SPIE**. Bellingham, Washington USA: Society of Photo-Optical Instrumentation Engineers (SPIE), 2001. v. 4298, n. 1, p. 48–55. ISSN 0277-786X.

INAZUMI, Y.; HORITA, Y.; KOTANI, K.; MURAI, T. Quality evaluation method considering time transition of coded video quality. **Proceedings of the International Conference on Image Processing – ICIP 99**. Washington, DC, USA: IEEE Computer Society, 1999. v. 4, p. 338–342. ISBN 0-7803-5467-2.

ITU-R. *Recommendation ITU-R BT.500-11 – Methodology for the subjective assessment of the quality of television pictures*. [S.l.], 2002.

ITU-R. *Recommendation ITU-R BT.1788 – Methodology for the subjective assessment of video quality in multimedia applications*. Geneva, Switzerland, 2007.

ITU-R. *Recommendation ITU-R BT.500-12 – Methodology for the subjective assessment of the quality of television pictures*. Geneva, Switzerland, 2009.

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

KIM, J. H.; AHN, S. C.; KIM, H.-G. Teleconference system with a shared working space and face mouseinteraction. In: AIZAWA, K.; NAKAMURA, Y.; SATOH, S. (Ed.). *Proceedings of the 5th Pacific Rim Conference on Advances in Multimedia Information Processing*. Berlin, Heidelberg: Springer-Verlag, 2004, (Lecture Notes in Computer Science, v. 3332). p. 665–671. ISBN 978-3-540-23977-2.

KOLB, A.; BARTH, E.; KOCH, R. Tof-sensors: New dimensions for realism and interactivity. **Proceedins of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops – CVPRW '08**. Washington, DC, USA: IEEE Computer Society, 2008. p. 1–6. ISSN 2160-7508.

KOLMOGOROV, V.; CRIMINISI, A.; BLAKE, A.; CROSS, G.; ROTHER, C. Bi-layer segmentation of binocular stereo video. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition – CVPR '05**. Washington, DC, USA: IEEE Computer Society, 2005. v. 2, p. 407–414. ISBN 0-7695-2372-2. ISSN 1063-6919.

KOLMOGOROV, V.; CRIMINISI, A.; BLAKE, A.; CROSS, G.; ROTHER, C. *Probabilistic fusion of stereo with color and contrast for bi-layer segmentation*. Cambridge, Mar 2005. MSR-TR-2005-35. Disponível em: <http://research.microsoft.com/pubs/70156/StereoSegmentation_tr.pdf>.

KOLMOGOROV, V.; CRIMINISI, A.; BLAKE, A.; CROSS, G.; ROTHER, C. Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 28, n. 9, p. 1480–1492, Sept. 2006. ISSN 0162-8828.

KOLMOGOROV, V.; ZABIN, R. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 26, n. 2, p. 147–159, feb 2004. ISSN 0162-8828.

KOYAMA, T.; KITAHARA, I.; OHTA, Y. Live mixed-reality 3d video in soccer stadium. **Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality – ISMAR '03**. Washington, DC, USA: IEEE Computer Society, 2003. p. 178. ISBN 0-7695-2006-5.

KOZAMERNIK, F.; STEINMANN, V.; SUNNA, P.; WYCKENS, E. Samviq – a new ebu methodology for video quality evaluations in multimedia. *SMPTE Motion Imaging Journal*, v. 114, n. 4, p. 152–160, april 2005.

KUMAR, S.; HEBERT, M. Discriminative random fields: A discriminative framework for contextual interaction in classification. **Proceedings of the Ninth IEEE International Conference on Computer Vision – ICCV '03**. Washington, DC, USA: IEEE Computer Society, 2003. p. 1150. ISBN 0-7695-1950-4.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. **Proceedings of the Eighteenth International Conference on Machine Learning – ICML '01**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. p. 282–289. ISBN 1-55860-778-1.

LAW, K.; SCLAROFF, S. Foreground object segmentation from binocular stereo video. *Intelligent Robots and Computer Vision XXIII: Algorithms, Techniques, and Active Vision*, SPIE, v. 6006, n. 1, p. 60060C, 2005.

LI, M. *Towards Real-Time Novel View Synthesis Using Visual Hulls*. Tese (Doutorado) — Universität des Saarlandes, 2005.

MARICHAL, X.; MACQ, B.; DOUXCHAMPS, D.; UMEDA, T. et al. The art.live architecture for mixed reality. **Proceedings of the International Virtual Reality Conference 2002 (IVRC 2002)**. Laval, France: [s.n.], 2002.

MARICHAL, X.; VILLEGAS, P. Objective evaluation of segmentation masks in video sequences. **Proceedings of the European Conference on Signal Processing (EUSIPCO '2000)**. [S.l.: s.n.], 2000. v. 4, p. 2193–2196.

MATUSIK, W.; BUEHLER, C.; RASKAR, R.; GORTLER, S. J.; MCMILLAN, L. Image-based visual hulls. **Proceedings of the 27th annual conference on Computer graphics and interactive techniques – SIGGRAPH '00**. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 2000. p. 369–374. ISBN 1-58113-208-5.

MAXWELL, S. E.; DELANEY, H. D. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. 2. ed. [S.l.]: Routledge Academic, 2003. 1104 p. ISBN 0805837183.

MECH, R.; MARQUÉS, F. Objective evaluation criteria for 2d-shape estimation results of moving objects. *EURASIP J. Appl. Signal Process.*, Hindawi Publishing Corp., New York, NY, United States, v. 2002, n. 4, p. 401–409, 2002. ISSN 1110-8657.

MISHIMA, Y. *Soft edge chroma-key generation based upon hexoctahedral color space*. U.S. Patent 5,355,174. Out. 1994. 11-10-1994.

MITRA, S.; ACHARYA, T. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, v. 37, n. 3, p. 311–324, may 2007. ISSN 1094-6977.

MORTENSEN, E.; BARRETT, W. Toboggan-based intelligent scissors with a four-parameter edge model. **Proceedings of the IEEE Computer Society Conference on**

Computer Vision and Pattern Recognition. Washington, DC, USA: IEEE Computer Society, 1999. v. 2, p. 2452–2458. ISBN 0-7695-0149-4.

NAKAMURA, R. *Vídeo-Avatar com detecção de colisão para realidade aumentada e jogos*. Tese (Doutorado) — Escola Politécnica da Universidade de São Paulo, 2008.

NAKAMURA, R.; LAGO, L. L. M.; CARNEIRO, A. B.; CUNHA, A. J. C.; ORTEGA, F. J. M.; BERNARDES-JR, J. L.; TORI, R. 3PI experiment: immersion in third-person view. **Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games – Sandbox '10**. New York, NY, USA: ACM, 2010. p. 43–48. ISBN 978-1-4503-0097-1.

NAM, W.; HAN, J. Motion-based background modeling for foreground segmentation. **Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks – VSSN '06**. New York, NY, USA: ACM, 2006. p. 35–44. ISBN 1-59593-496-0.

NGUYEN, D. T.; CANNY, J. More than face-to-face: empathy effects of video framing. **Proceedings of the 27th international conference on Human factors in computing systems – CHI '09**. New York, NY, USA: ACM, 2009. p. 423–432. ISBN 978-1-60558-246-7.

OGI, T.; YAMADA, T.; KURITA, Y.; HATTORI, Y. Y.; HIROSE, M. Usage of video avatar technology for immersive communication. **Proceedings of the First International Workshop on Language Understanding and Agents for Real World Interaction – ACL 2003**. [S.l.: s.n.], 2003. p. 24–31.

OGI, T.; YAMADA, T.; TAMAGAWA, K.; KANO, M.; HIROSE, M. Immersive telecommunication using stereo video avatar. **Proceedings of the Virtual Reality 2001 Conference – VR '01**. Washington, DC, USA: IEEE Computer Society, 2001. p. 45. ISBN 0-7695-0948-7.

OHTA, Y.; KANADE, T. Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI-7, n. 1, p. 139–154, March 1985.

OLIVER, N.; ROSARIO, B.; PENTLAND, A. A bayesian computer vision system for modeling human interactions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 22, n. 8, p. 831–843, ago. 2000. ISSN 0162-8828.

PAROLIN, A.; FICKEL, G. P.; JUNG, C. R.; MALZBENDER, T.; SAMADANI, R. Bilayer video segmentation for videoconferencing applications. **Proceedings of the IEEE International Conference on Multimedia and Expo – ICME 2011**. Washington, DC, USA: IEEE Computer Society, 2011. p. 1–6. ISBN 978-1-61284-348-3. ISSN 1945-7871.

PÉCHARD, S.; PÉPION, R.; CALLET, P. L. Suitable methodology in subjective video quality assessment: a resolution dependent paradigm. **Proceedings of the International Workshop on Image Media Quality and its Applications – IMQA2008**. [S.l.: s.n.], 2008.

PEDRINI, H.; SCHWARTZ, W. R. *Análise de Imagens Digitais: Princípios, Algoritmos e Aplicações*. 1. ed. [S.l.]: Thomson Learning, 2008. 508 p. ISBN 9788522105953.

PICCARDI, M. Background subtraction techniques: a review. **Proceedings of the IEEE International Conference on Systems, Man and Cybernetics**. Washington, DC, USA: IEEE Computer Society, 2004. v. 4, p. 3099–3104. ISSN 1062-922X.

PORTER, T.; DUFF, T. Compositing digital images. **Proceedings of the 11th annual conference on Computer graphics and interactive techniques – SIGGRAPH '84**. New York, NY, USA: ACM Press, 1984. p. 253–259. ISBN 0-89791-138-5.

QIAN, R.; SEZAN, M. Video background replacement without a blue screen. **Proceedings of the International Conference on Image Processing – ICIP 99**. Washington, DC, USA: IEEE Computer Society, 1999. v. 4, p. 143–146.

RHEE, S.-M.; ZIEGLER, R.; PARK, J.; NAEF, M.; GROSS, M.; KIM, M.-H. Low-cost telepresence for collaborative virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, v. 13, n. 1, p. 156 –166, 2007. ISSN 1077-2626.

ROTHER, C.; KOLMOGOROV, V.; BLAKE, A. "Grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, v. 23, n. 3, p. 309–314, 2004.

SANCHES, S. R. R.; NAKAMURA, R.; SILVA, V. F.; TORI, R. Bilayer segmentation of live video in uncontrolled environments for background substitution: An overview and main challenges. *IEEE Latin America Transactions*, v. 10, p. 2138–2149, 2012.

SANCHES, S. R. R.; SILVA, V.; TORI, R. Bilayer segmentation augmented with future evidence. In: MURGANTE, B.; GERVASI, O.; MISRA, S.; NEDJAH, N.; ROCHA, A.; TANIAR, D.; APDUHAN, B. (Ed.). *Computational Science and Its Applications – ICCSA 2012*. [S.l.]: Springer Berlin / Heidelberg, 2012, (Lecture Notes in Computer Science, v. 7334). p. 699–711. ISBN 978-3-642-31074-4.

SANCHES, S. R. R.; TOKUNAGA, D. M.; SILVA, V. F.; SEMENTILLE, A. C.; TORI, R. Mutual occlusion between real and virtual elements in augmented reality based on fiducial markers. **Proceedings of IEEE Workshop on Applications of Computer Vision – WACV 2012**. Washington, DC, USA: IEEE Computer Society, 2012. p. 49–54. ISSN 1550-5790.

SANCHES, S. R. R.; TOKUNAGA, D. M.; SILVA, V. F.; TORI, R. Subjective video quality assessment in segmentation for augmented reality applications. **Proceedings of the XIII Symposium on Virtual Reality – SVR 2012**. Washington, DC, USA: IEEE Computer Society, 2012. p. 46 –55.

SCHARSTEIN, D.; SZELISKI, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, Kluwer Academic Publishers, Hingham, MA, USA, v. 47, p. 7–42, Apr 2002. ISSN 0920-5691.

SCHARSTEIN, D.; SZELISKI, R. High-accuracy stereo depth maps using structured light. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, IEEE Computer Society, Los Alamitos, CA, USA, v. 1, p. 195, 2003. ISSN 1063-6919.

SENDERS, J. W. Distribution of visual attention in static and dynamic displays. **Proceedings of the Human Vision and Electronic Imaging II**. San Jose, CA: SPIE, 1997. v. 3016, p. 186–194.

SHOTTON, J.; WINN, J.; ROTHER, C.; CRIMINISI, A. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. p. I: 1–15, 2006.

SISCOUTTO, R. A. *Proposta de Arquitetura para Teleconferência Baseada na Integração de Vídeo Avatar Estereoscópico em Ambiente Tridimensional*. Tese (Doutorado) — Escola Politécnica de Universidade de São Paulo, 2003.

STAUFFER, C.; GRIMSON, W. E. L. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 22, n. 8, p. 747–757, 2000. ISSN 0162-8828.

SUN, J.; ZHANG, W.; TANG, X.; SHUM, H.-Y. Background cut. In: LEONARDIS, A.; BISCHOF, H.; PINZ, A. (Ed.). *European Conference on Computer Vision – ECCV 2006*. Berlin / Heidelberg: Springer, 2006, (Lecture Notes in Computer Science, v. 3952). p. 628–641. ISBN 3-540-33834-9.

TANG, Z.; MIAO, Z.; WAN, Y. Background subtraction using running gaussian average and frame difference. In: MA, L.; RAUTERBERG, M.; NAKATSU, R. (Ed.). *Entertainment Computing – ICEC 2007*. Berlin / Heidelberg: Springer, 2007, (Lecture Notes in Computer Science, v. 4740). p. 411–414.

TOYAMA, K.; KRUMM, J.; BRUMITT, B.; MEYERS, B. Wallflower: Principles and practice of background maintenance. **Proceedings of the Seventh IEEE International Conference on Computer Vision**. Los Alamitos, CA, USA: IEEE Computer Society, 1999. v. 1, p. 255. ISBN 0-7695-0164-8.

VILLEGAS, P.; MARICHAL, X. Perceptually-weighted evaluation criteria for segmentation masks in video sequences. *IEEE Transactions on Image Processing*, v. 13, n. 8, p. 1092 –1103, aug 2004. ISSN 1057-7149.

VILLEGAS, P.; MARICHAL, X.; SALCEDO, A. Objective evaluation of segmentation masks in video sequences. **Proceedings of the Workshop on Image Analysis for Multimedia Interactive Services – WIAMIS'99**. [S.l.: s.n.], 1999. p. 85–88.

VLAHOS, P. *Composite photography utilizing sodium vapor illumination*. U.S. Patent 3,095,304. Jun. 1963.

VLAHOS, P. *Composite color photography*. U.S. Patent 3,158,477. Nov. 1964.

VLAHOS, P. *Comprehensive electronic compositing system*. U.S. Patent 4,100,569. Jul. 1978.

WANG, J.; COHEN, M. F. Image and video matting: a survey. *Found. Trends. Comput. Graph. Vis.*, Now Publishers Inc., Hanover, MA, USA, v. 3, n. 2, p. 97–175, 2007. ISSN 1572-2740.

WANG, L.; ZHANG, C.; YANG, R.; ZHANG, C. Tofcut: Towards robust real-time foreground extraction using a time-of-flight camera. **Proceedings of the Fifth International Symposium on 3D Data Processing, Visualization and Transmission – 3DPVT**. [S.l.: s.n.], 2010. p. 1–8.

WANG, S.; XIONG, X.; XU, Y.; WANG, C.; ZHANG, W.; DAI, X.; ZHANG, D. Face-tracking as an augmented input in video games: enhancing presence, role-playing and control. **Proceedings of the SIGCHI conference on Human Factors in computing systems – CHI '06**. New York, NY, USA: ACM, 2006. p. 1097–1106. ISBN 1-59593-372-7.

WILLIAMS, F. D. *Method of Taking Motion Pictures*. U.S. Patent 1,273,435. Jul. 1918.

WOLLBORN, M.; MECH, R. *Refined procedure for objective evaluation of vop generation algorithms*. Tech. Report ISO/IECJTCI/SC29/WG11 M3448, 1997.

WU, Q.; BOULANGER, P.; BISCHOF, W. F. Robust real-time bi-layer video segmentation using infrared video. **Proceedings of the Canadian Conference on Computer and Robot Vision – CRV '08**. Washington, DC, USA: IEEE Computer Society, 2008. p. 87–94. ISBN 978-0-7695-3153-3.

WU, Z.; CHEN, C. A new foreground extraction scheme for video streams. **Proceedings of the ninth ACM international conference on Multimedia – MULTIMEDIA '01**. New York, NY, USA: ACM, 2001. p. 552–554. ISBN 1-58113-394-4.

YILMAZ, A.; JAVED, O.; SHAH, M. Object tracking: A survey. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 38, n. 4, p. 13, 2006. ISSN 0360-0300.

YIN, P.; CRIMINISI, A.; WINN, J.; ESSA, I. Tree-based classifiers for bilayer video segmentation. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR '07**. Los Alamitos, CA, USA: IEEE Computer Society, 2007. v. 0, p. 1–8. ISBN 1-4244-1179-3.

YIN, P.; CRIMINISI, A.; WINN, J.; ESSA, I. Bilayer segmentation of webcam videos using tree-based classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, Los Alamitos, CA, USA, v. 33, n. 1, p. 30–42, 2011. ISSN 0162-8828.

ZHANG, Y. A survey on evaluation methods for image segmentation. *Pattern Recognition*, v. 29, n. 8, p. 1335–1346, 1996. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/0031320395001697>.

Apêndice I – CONCEITOS E ALGORITMOS

Neste apêndice são apresentados alguns conceitos relacionados aos algoritmos de segmentação de vídeos em duas camadas discutidos na seção 2. Além disso, são detalhados os algoritmos utilizados para segmentar os vídeos utilizados nos experimentos subjetivos descritos na seção 4.6.

I.1 Segmentação Binária, Transparência de Pixels e Representação do Elemento de Interesse

Segundo Gonzalez e Woods (2002), uma imagem (ou quadro de vídeo) pode ser definida como uma função bidimensional $z(x, y)$, onde x e y são coordenadas no plano espacial, e a amplitude de z, em cada par de coordenadas (x, y), é sua intensidade naquele ponto. Em imagens digitais, os valores de (x, y) e da amplitude de z são finitos. O processo de segmentação consiste na subdivisão dessa imagem em estruturas com conteúdo semântico relevante para uma determinada aplicação (PEDRINI; SCHWARTZ, 2008). Em outras palavras, o que determina o nível dessa subdivisão consiste no problema a ser resolvido, pois o processo apenas se finaliza quando o elemento de interesse para a aplicação em questão estiver isolado (GONZALEZ; WOODS, 2002). Uma prática comum em processos de segmentação é tratar o elemento de interesse, que foi extraído do seu contexto original, como uma camada de imagem. Para tornar a representação de uma camada de primeiro plano possível, faz-se necessária a utilização de formatos de pixel que permitam controlar sua transparência, como mostrado na figura 1.

A tarefa de estimar níveis de transparência, conhecida como “problema do *matting*” foi definida matematicamente em Porter e Duff (1984), por meio da introdução do canal alfa,

uma solução para controlar a interpolação linear das cores de duas camadas de imagens. Efeitos como suavização de bordas, além da preservação da transparência de objetos translúcidos, podem ser obtidos com esse tipo de recurso.

Segundo Porter e Duff (1984), a imagem I_z é modelada como uma combinação de uma camada de primeiro plano F_z e uma de fundo B_z, utilizando-se o canal alfa α_z como na equação

$$I_z = \alpha_z F_z + (1 - \alpha_z) B_z \tag{46}$$

onde α_z pode ser qualquer valor entre [0,1]. Se $\alpha_z = 1$ ou 0, o pixel pertence à camada de primeiro plano e à camada de fundo, respectivamente. Aos pixels cujas tonalidades são influenciadas pelas duas camadas – o que ocorre com frequência em objetos transparentes ou nas bordas de objetos opacos – valores intermediários de alfa devem ser estimados para que a separação do elemento de interesse seja mais precisa (WANG; COHEN, 2007).

Na equação 46, restringindo-se o valor de alfa a assumir apenas os valores 0 ou 1, transforma-se o problema do *matting* em outro problema clássico: a segmentação binária, objeto de estudo deste trabalho, em que cada pixel pertence totalmente à camada de primeiro plano ou a camada de fundo (WANG; COHEN, 2007).

Segundo Wang e Cohen (2007), a maioria das pesquisas que buscam soluções para o problema do *matting* não trata o problema da segmentação binária. Algoritmos de *matting* são frequentemente custosos do ponto de vista computacional, uma vez que são normalmente voltados para composição de imagens estáticas ou vídeos pré-gravados. Por esse motivo, muitos métodos não têm compromisso com seu tempo de execução, pois podem ser aplicados *offline*.

Os métodos de segmentação desenvolvidos para essas aplicações (*offline*), normalmente, utilizam, além da imagem original, uma máscara da mesma imagem, chamada trimap, que pode ser produzida manualmente pelo usuário, ou estimada por qualquer método de segmentação binária, que não é necessariamente parte do método principal, responsável pelo matting (WANG; COHEN, 2007).

Um *trimap* é composto por três regiões: primeiro plano, plano de fundo e regiões desconhecidas, em que o pixel não pertence nem totalmente ao fundo, nem totalmente ao elemento de interesse (WANG; COHEN, 2007). Apenas nessas regiões ambíguas (ou

desconhecidas) atuam os algoritmos que estimam valores intermediários de alfa.

Por outro lado, aplicações executadas em tempo real, baseadas ou não em *trimaps*, exigem que todo o processo seja automático e que a solução para o problema da segmentação em duas camadas seja de rápida execução e inclua estimativas de transparência de pixels na geração da camada de primeiro plano.

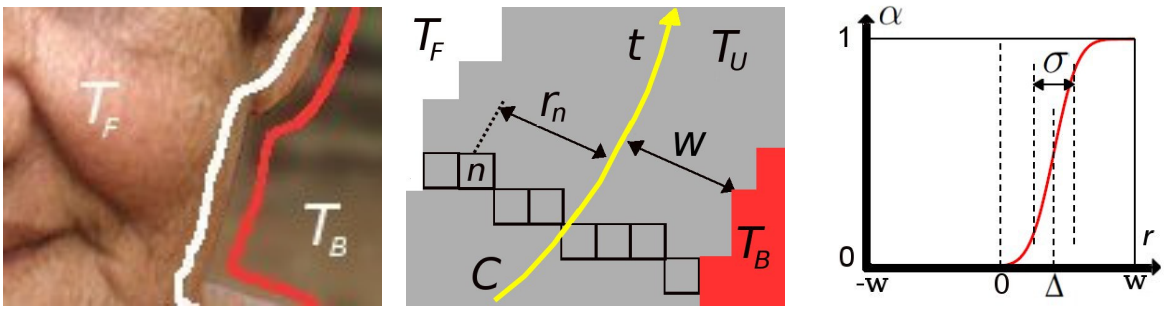
Uma técnica muita utilizada para estimar transparência de pixels em métodos de segmentação para aplicações de tempo real é conhecida como *border matting* (ROTHER; KOLMOGOROV; BLAKE, 2004).A suavização nas bordas do elemento de interesse, que permite produzir cenas compostas com qualidade aceitável para aplicações de substituição de fundo, pode ser obtida por meio da técnica.

Resumidamente, o algoritmo de *border matting* toma como base uma polilinha C, mostrada em amarelo na figura 24(b), que contorna o elemento de interesse. O conjunto de pixels que pertencem a C pode ser obtido automaticamente a partir da segmentação binária, que produz uma borda rígida.

Um *trimap* $\{T_B, T_U, T_F\}$ é calculado (figura 24(a)), onde T_B e T_F são os conjuntos de pixels que pertencem ao plano de fundo e ao elemento de interesse respectivamente. T_U é o conjunto de pixels em uma faixa de tamanho $\pm w$ pixels, de ambos os lados de C. O objetivo é calcular um mapa de transparência α_n, $n \in T_U$ utilizando um modelo baseado no proposto em (MORTENSEN; BARRETT, 1999), que define a forma como α varia dentro de T_U. Os parâmetros do contorno C, $t = 1, \ldots, T$ têm periodicidade T, a medida que a curva C é fechada (ROTHER; KOLMOGOROV; BLAKE, 2004).

Um índice $t(n)$ é atribuído para cada pixel $n \in T_U$, como mostrado na figura 24(b). Os valores de α são obtidos por meio de uma função g: $\alpha_n = g(r_n; \Delta_{t(n)}, \sigma_{t(n)})$, onde r_n é a distância do pixel n até C (figura 24(c)).

Os parâmetros Δ, σ determinam, respectivamente, o centro e a largura da transição de 0 até 1 no conjunto de valores possíveis de α. Todos os pixels com o mesmo índice t tem os mesmos valores de parâmetros Δ_t e σ_t. Os parâmetros $\Delta_1, \sigma_1, \ldots, \Delta_t, \sigma_t$ são estimados por meio de funções de minimização de energia, detalhadas em (ROTHER; KOLMOGOROV; BLAKE, 2004).



(a) (b) (c)

Figura 24 – Técnica do Border Matting. (a) Imagem original sobreposta pelo *trimap*. (b) Notação para a parametrização do contorno C, obtido da segmentação binária, e do mapa de distâncias. Para cada pixel em T_U são atribuídos valores do parâmetro t, do contorno, e da distância r_n de C. (c) Função g, que define a atribuição dos valores de α (ROTHER; KOLMOGOROV; BLAKE, 2004)

I.2 Segmentação como um problema de minimização de energia

Entre as características comuns identificadas em abordagens recentes de segmentação de vídeos em duas camadas, importa ressaltar o fato de muitas soluções tratarem a segmentação como um problema de minimização de energia.

Greig, Porteous e Seheult (1989) foram os primeiros a descobrirem que algoritmos de fluxo máximo/mínimo para otimização combinatorial podem ser utilizados também para minimizar funções de energia em visão computacional (BOYKOV; KOLMOGOROV, 2004). Em um processo de atribuição de rótulos a pixels de uma imagem, a partir de um conjunto de pixels P e de um conjunto de rótulos L, o objetivo é encontrar um rótulo f (i.e., realizar um mapeamento de P em L), que minimize determinada função de energia (KOLMOGOROV; ZABIN, 2004).

Para uma divisão da imagem em duas camadas, o conjunto L possui dois rótulos: elemento de interesse e plano de fundo (segmentação binária). Níveis de transparência de pixels são determinados, normalmente, em um passo posterior a segmentação binária, por meio de técnicas como *border matting* (ROTHER; KOLMOGOROV; BLAKE, 2004), discutida na seção I.1.

A função de energia utilizada no primeiro trabalho de Greig, Porteous e Seheult (1989), e por muitos métodos de segmentação atuais, descritos na seção 2.1.2, pode ser repre-

sentada da forma

$$E(f) = \sum_{p \in P} D_p(f_p) + \sum_{p,q \in N} V_{p,q}(f_p, f_q) \quad (47)$$

onde $N \subset PXP$ é o conjunto dos pixels que possuem relação de vizinhança. O termo $D_p(f_p)$ é uma função derivada dos dados observados, que mede o custo para atribuição do rótulo f_p ao pixel p. O termo $V_{p,q}(f_p, f_q)$ é responsável pela medição do custo para atribuir os rótulos f_p, f_q aos pixels adjacentes p, q, e é utilizado para manutenção das descontinuidades na imagem (KOLMOGOROV; ZABIN, 2004). Aos pixels vizinhos com contraste alto são atribuídos custos menores, pois existe maior probabilidade de pertencerem a conjuntos diferentes. Um exemplo de imagem rotulada é mostrado na figura 25.

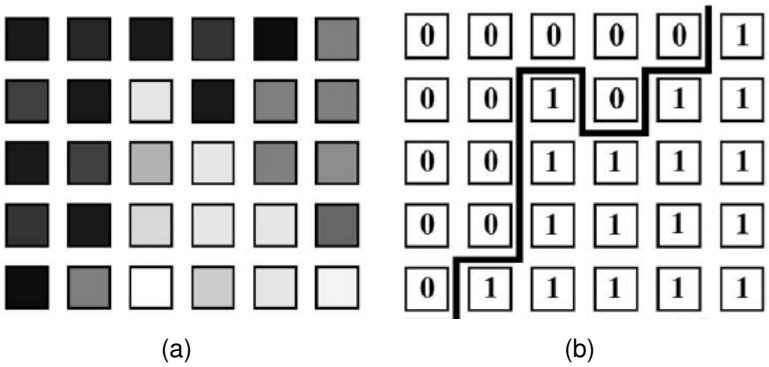


Figura 25 – Exemplo de imagem rotulada, adaptada de (BOYKOV; KOLMOGOROV, 2004). A imagem (a) representa um conjunto de pixels P com intensidades observadas I_p, para cada $p \in P$. Em (b) é mostrada a atribuição de um rótulo $f_p \in \{0, 1\}$ para cada pixel $p \in P$. As linhas mais espessas, mostradas em (b), representam rótulos de descontinuidades entre pixels vizinhos

Uma abordagem, aplicável em tempo real, bastante utilizada para minimizar energia consiste em transformar a segmentação binária em um problema de corte em grafos (KOLMOGOROV; ZABIN, 2004). A ideia básica é construir um grafo específico para determinada função de energia ser minimizada, de modo que o corte mínimo no grafo minimize também a energia. Grande parte dos trabalhos utilizam um arcabouço geral, proposto em Boykov e Jolly (2001), para essa finalidade.

Segundo Boykov e Jolly (2001), dado um grafo direcionado $G = (V, \varepsilon)$ com arestas de pesos não negativos e dois vértices terminais s (*source*) e t (*sink*), um corte ***s-t*** $C = (S, T)$ é um particionamento dos vértices em V em dois conjuntos disjuntos S e T, de modo que $s \in S$ e $t \in T$. O custo total do corte é a soma dos custos de todas as arestas que partem

de S e chegam em T (KOLMOGOROV; ZABIN, 2004)

$$c(S,T) = \sum_{u \in S, v \in T, (u,v) \in \varepsilon} c(u,v) \tag{48}$$

O problema do corte mínimo é encontrar um corte C com o menor custo, o que é equivalente ao cálculo do fluxo máximo de s até t. Existem muitos algoritmos que resolvem esse problema em tempo polinomial, como o do fluxo máximo otimizado, proposto em Boykov e Kolmogorov (2004).

Importa observar que um corte $C = (S,T)$ é um processo de atribuição de rótulos f que mapeia o conjunto de vértices $V - \{s,t\}$ em $\{0,1\}$, onde $f(v) = 0$ implica $v \in S$ e $f(v) = 1$ implica $v \in T$. Isso significa que um corte é um particionamento binário de um grafo visto como uma atribuição de rótulos com dois valores possíveis (KOLMOGOROV; ZABIN, 2004). Um exemplo de um grafo utilizado como estrutura auxiliar para minimização de energia aplicada a segmentação de vídeos é mostrado na figura 26.

O conjunto de vértices V é formado pelos pixels $P \cup \{s - t\}$, e o custo entre $\{s,t\}$ para elementos em P é justamente a função D_p, ao passo que o custo entre elementos de P é justamente $V_{p,q}$.

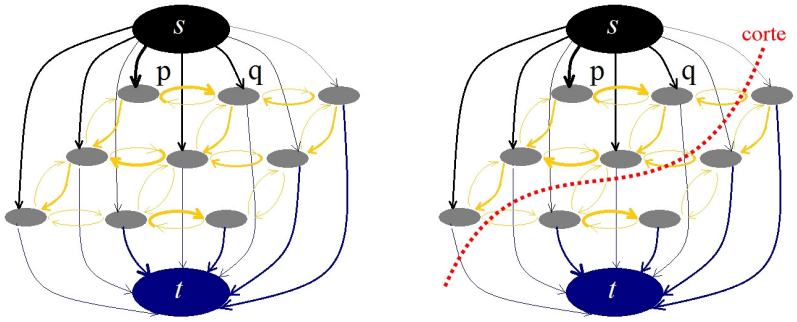


Figura 26 – Exemplo de grafo utilizado em segmentação binária, adaptado de (BOYKOV; KOLMOGOROV, 2004). Os custos das arestas são representados por sua espessura. Um grafo de corte similar foi utilizado pela primeira vez na visão computacional em (GREIG; PORTEOUS; SEHEULT, 1989), para restauração de imagens binárias. O grafo G é mostrado em (a) e o corte em G pode ser visualizado em (b)

I.3 Algoritmos de Segmentação utilizados

No contexto da segmentação binária, foram utilizados neste trabalho quatro métodos, dois deles baseado na técnica de Subtração de Fundo e outros dois baseados em arcabouço de minimização de energia. Uma breve descrição desses métodos é mostrado nesta seção

I.3.1 Algoritmo de Qian e Sezan (1999)

O algoritmo apresentado por Qian e Sezan (1999) consiste na técnica de subtração de fundo na sua forma mais simplificada. Dado que cada quadro de vídeo é representado por uma matriz de pixels

$$z = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,Y} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,Y} \\ \vdots & \vdots & \ddots & \vdots \\ z_{X,1} & z_{X,2} & \cdots & z_{X,Y} \end{bmatrix}, \tag{49}$$

a segmentação consiste na comparação do modelo do fundo z^{ref} com o quadro de vídeo atual z_t

$$\alpha_{z^t} = \begin{cases} 1 & , \text{ if } |z^t - z^{ref}| > Th \\ 0 & , \text{ if } |z^t - z^{ref}| \leqslant Th \end{cases} \tag{50}$$

onde z^t representa um quadro de vídeo no tempo t e z^{ref} uma imagem de referência, capturada previamente, que contém apenas o fundo da cena (sem a presença do elemento de interesse). Th representa um limiar que permite que pequenas variações na cor do pixel sejam desconsideradas, quando comparada com a imagem de referência.

Embora não houvesse uma imagem “limpa” (sem o elemento de interesse) do plano de fundo dos vídeos utilizados no experimento, os modelos de fundo foram obtidos utilizando as sequências SEQ1, SEQ2, SEQ3, SEQ4 e SEQ5 e seus respectivos *ground truths* da seguinte forma. Cada pixel do modelo do fundo corresponde a média dos valores desse pixel obtidos dos quadros da sequência em que o pixel não se encontrava ocultado pelo elemento de interesse.

Em seguida, a imagem resultante, que representa o modelo do fundo, foi percorrida pixel a pixel – da esquerda para a direita e de cima para baixo –, preenchendo os pixels correspondentes as partes do fundo que não ficaram visíveis em nenhum dos quadros com

o valor do pixel visível mais próximo. Na figura 27 pode ser visualizado três quadros das sequências SEQ4 e SEQ5, seguidos de seus respectivos modelos do fundo obtidos pelo processo descrito.

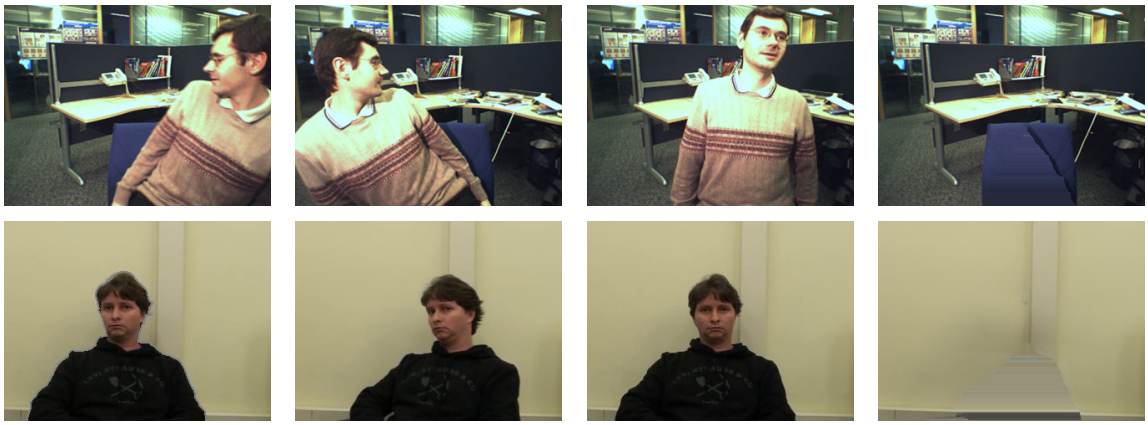


Figura 27 – Modelos de fundo utilizados no Experimento. Três quadros de vídeo das sequências SEQ4 e SEQ5, seguidos dos respectivos modelos de fundo da sequência

I.3.2 Algoritmo de Criminisi et al. (2006)

O algoritmo apresentado em (CRIMINISI et al., 2006) é baseado em um arcabouço de minimização de energia em que a matriz de pixels z, que representa um quadro de vídeo, encontra-se no espaço de cores YUV e um frame no tempo t é denotado z^t. A derivada temporal é denotada $\dot{z} = [\dot{z}_{x,y}]_{X \times Y}$ e calculada

$$\dot{z}^t = |G(\mathbf{0}; \sigma_T) * z^t - G(\mathbf{0}; \sigma_T) * z^{t-1}| \tag{51}$$

em cada tempo t, onde $G(\mathbf{0}; \sigma_T)$ é um *kernel* gaussiano 2D centralizado com desvio padrão σ_T e $*$ é o operador de convolução. Os gradientes espaciais $g = [g_{x,y}]_{X \times Y}$ são calculados pela convolução dos quadros de vídeo com a derivada de primeira ordem do *kernel* gaussiano, que possui desvio padrão σ_S,

$$g^t = \sqrt{\left(\frac{\partial G(\mathbf{0}; \sigma_S)}{\partial x} * z^t\right)^2 + \left(\frac{\partial G(\mathbf{0}; \sigma_S)}{\partial y} * z^t\right)^2}. \tag{52}$$

Assim como em Criminisi et al. (2006), foi utilizado $\sigma_S = \sigma_T = 0.8$.

As derivadas espaço-temporais são calculadas apenas para o canal Y. As observações de movimento no tempo t são denotadas $m^t = (g^t, \dot{z}^t)$. Dado uma sequência de dados

da imagem $z^1, z^2, \ldots, z^t$ e uma sequência de dados de movimento $m^1, m^2, \ldots, m^t$, a segmentação consiste em inferir um rótulo binário $\alpha^t_{x,y} \in \{F, B\}$ para cada pixel do quadro em análise. F e B denotam primeiro plano e plano de fundo, respectivamente.

O modelo probabilístico para a extração da camada de primeiro plano apresentado por Criminisi et al. (2006) se baseia em um arcabouço de minimização de energia e consiste na extensão do modelo descrito em Boykov e Jolly (2001), Rother, Kolmogorov e Blake (2004), Kolmogorov et al. (2005a). O modelo consiste em um Campo Aleatório Condicional (*Conditional Randon Field* (CRF)) (LAFFERTY; MCCALLUM; PEREIRA, 2001) com termos independentes determinados discriminativamente. Em outras palavras, ao invés de trabalhar com distribuições conjuntas, distribuições condicionais são consideradas (KUMAR; HEBERT, 2003). A probabilidade condicional é modelada pelo CRF da forma:

$$p(\alpha^1, \ldots, \alpha^t | z^1, \ldots, z^t, m^1, \ldots, m^t) \propto \exp - \left\{ \sum_{t'=1}^{t} E^{t'} \right\} \tag{53}$$

onde $E^t = E(\alpha^t, \alpha^{t-1}, \alpha^{t-2}, z^t, m^t)$.

A energia E^t associada ao tempo t consiste na soma de quatro termos:

$$\begin{aligned} E(\alpha^t, \alpha^{t-1}, \alpha^{t-2}, z^t, m^t) = & \\ \eta V^T(\alpha^t, \alpha^{t-1}, \alpha^{t-2}) + \gamma V^S(\alpha^t, z^t) & \\ + \rho U^C(\alpha^t, z) + \phi U^M(\alpha^t, \alpha^{t-1}, m^t), & \end{aligned} \tag{54}$$

em que os dois primeiros termos são conhecidos a priori e os dois segundos são observações. η, γ, ρ e ϕ são parâmetros de normalização.

O termo temporal $V^T(\cdot)$, que é obtido a priori, impõe uma tendência para continuidade temporal dos rótulos. Uma cadeia de Markov de segunda ordem é utilizada no arcabouço de minimização de energia para que seja incorporada a intuição de que um pixel que pertencia ao plano de fundo no tempo $t-2$ e pertencia ao elemento de interesse no tempo $t-1$ provavelmente continuará pertencendo ao elemento de interesse no tempo t. As transições temporais são aprendidas de uma base de vídeos rotulados. O termo temporal é dado por:

$$V^T(\alpha^t, \alpha^{t-1}, \alpha^{t-2}) = \sum_{m=1}^{X} \sum_{n=1}^{Y} [-\log p(\alpha^t_{x,y} | \alpha^{t-1}_{x,y}, \alpha^{t-2}_{x,y})]. \tag{55}$$

O termo espacial $V^S(\cdot)$ é um termo Ising (CIPRA, 1987) que impõe a tendência para continuidade espacial dos rótulos. Esse termo é inibido pelo alto contraste. C consiste no conjunto de pares de pixels vizinhos em um quadro, z_i representa os valores do pixel i no espaço de cores YUV e α_i é o rótulo binário. O termo Ising é representado por

$$V^S(\alpha, z) = \sum_{i,j \in C} [\alpha_i \neq \alpha_j] \left(\frac{\epsilon + e^{-\mu ||z_i - z_j||^2}}{1 + \epsilon} \right). \tag{56}$$

O parâmetro de contraste μ é dado por $\mu = (2\langle ||z_i - z_j||^2 \rangle)^{-1}$, onde $\langle \cdot \rangle$ são os valores esperados dos pares de vizinhos em uma imagem. À constante de "diluição" ϵ foi atribuído o valor $\epsilon = 1$, como em (KOLMOGOROV et al., 2005a).

O termo de cor $U^C(\cdot)$ avalia a evidência para atribuição de rótulos com base nas distribuições de cores do primeiro plano e do plano de fundo. As probabilidades são modeladas como histogramas no espaço de cores YUV. Neste trabalho, as probabilidades de cores foram aprendidas do primeiro quadro de vídeo, utilizando-se seu respectivo *ground truth*. O termo de cor é definido como:

$$U^C(\alpha, z) = -\sum_{m=1}^{X} \sum_{n=1}^{Y} \log p(z_{x,y} | \alpha_{x,y}). \tag{57}$$

O termo de movimento $U^M(\cdot)$ utiliza as derivadas espaciais e temporais $m = (g, \dot{z})$ para obter as características dos movimentos do elemento de interesse. Segundo Criminisi et al. (2006), a história recente da segmentação de um pixel pertence a uma das quatro classes: FF, BB, FB and BF. As características dos movimentos observadas da imagem $m^t_{x,y} = (g^t_{x,y}, \dot{z}^t_{x,y})$ no tempo t estão condicionadas as combinações dos rótulos $\alpha^{t-1}_{x,y}$ and $\alpha^t_{x,y}$. A derivada temporal $\dot{z}^t_{x,y}$ é calculada dos quadros $t-1$ e t, portanto, depende do resultado da segmentação desses quadros.

As probabilidades do movimento são aprendidas dos vídeos *ground-truth* e armazenadas como histogramas 2D para serem utilizadas no processo como parte da energia total da forma

$$U^M(\alpha^t, \alpha^{t-1}, m^t) = -\sum_{x=1}^{X} \sum_{y=1}^{Y} \log p(m^t_{x,y} | \alpha^t_{x,y}, \alpha^{t-1}_{x,y}). \tag{58}$$

A energia E^t, modelada como um CRF, é descrita como em Kumar e Hebert (2003):

$$E^t = \sum_{i \in \mathcal{S}} \left[A_i(\alpha_i^t, \mathbf{o}^t) + \sum_{j \in \mathcal{N}_i} I_{ij}(\alpha_i^t, \alpha_j^t, \mathbf{o}^t) \right],$$

onde $\mathcal{S}$ é o conjunto de pixels de um quadro, $\mathbf{o} = (\hat{\alpha}^{t-1}, \hat{\alpha}^{t-2}, z^t, m^t)$ é a observação no tempo t, $\mathcal{N}_i$ é a vizinhança do pixel i, A_i e I_{ij} são as potenciais associação e interação, respectivamente. Finalmente, a energia total é minimizada por meio do algoritmo de corte em grafo (*graph cut*), apresentado em Kolmogorov e Zabin (2004).

I.3.3 Algoritmo de Sanches, Silva e Tori (2012)

Embora derivado do algoritmo de Criminisi et al. (2006), o apresentado em (SANCHES; SILVA; TORI, 2012) foi utilizado neste experimento pelo fato dos erros decorrentes da segmentação se mostrarem de formas diferentes na imagem resultante.

Diferentemente do algoritmo original, que faz uso das características de movimento U^M considerando apenas os quadros passados, o de Sanches, Silva e Tori (2012) utiliza derivadas temporais de forma bidirecional, obtendo informações de quadros “futuros”. Como em muitas aplicações executadas tempo real algum atraso é esperado, o atraso de um quadro – da forma como explorado no trabalho citado – pode ser imperceptível para o usuário. A figura 28 mostra como é feito o relacionamento entre as variáveis utilizadas no CRF.

No modelo original a observação é definida $m^t = (g^t, \dot{z}^t)$ e a energia minimiza a probabilidade $p(g^t, \dot{z}^t | \alpha^t, \alpha^{t-1})$. No trabalho de Sanches, Silva e Tori (2012), espera-se um novo quadro para que uma nova evidência $\dot{z}^{t+1}$ possa ser observada e $p(g^t, \dot{z}^t, \dot{z}^{t+1} | \alpha^t, \alpha^{t-1})$ seja minimizada, de acordo com a equação 54.

A importância de cada evidência observada foi calculada por meio da análise da entropia, utilizando uma base de vídeos (com 38 sequências de vídeo) e seus respectivos *ground truths*. Desse modo, as observações $\alpha^{t-1}, \dot{z}^t, \dot{z}^{t+1}$ e g^t foram combinadas para testar a influência de cada uma delas. A combinação das evidências $\dot{z}^t, \dot{z}^{t+1}$ e g^t se mostrou a mais eficiente e os resultados obtidos na segmentação por meio do algoritmo estendido foram melhores quando comparados ao original.

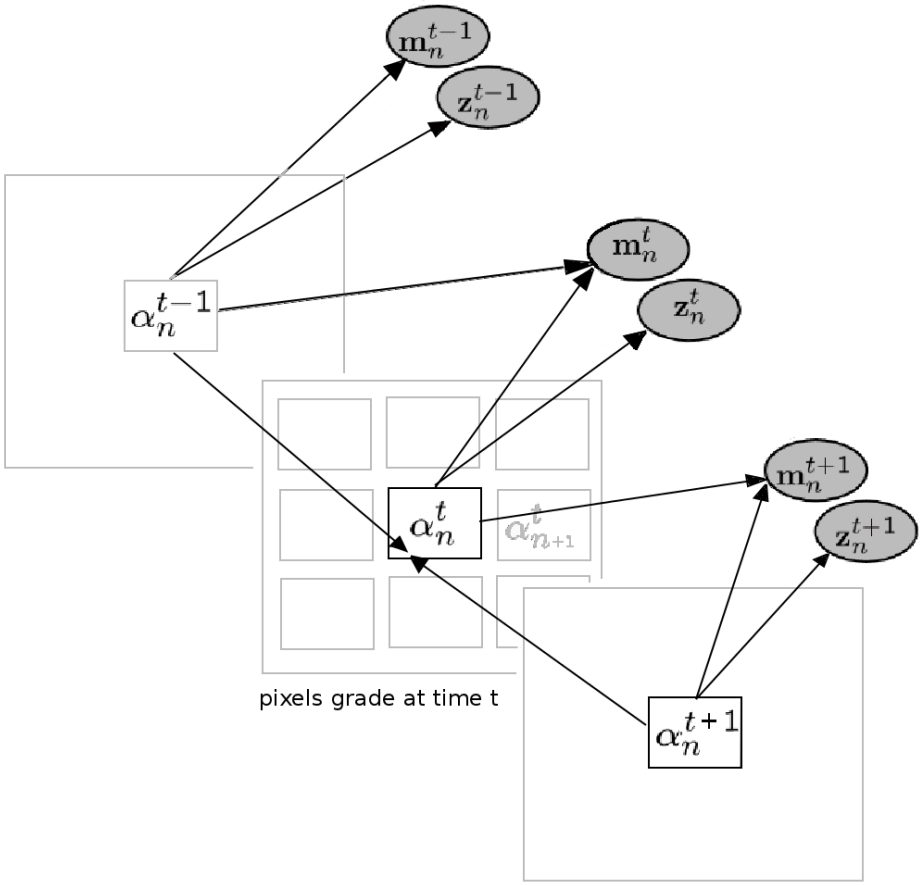


Figura 28 – Campo Aleatório Condicional utilizado no trabalho de Sanches, Silva e Tori (2012)

As probabilidades relacionadas ao termo de cor $U^C(\cdot)$, assim como em Criminisi et al. (2006), foram aprendidas do primeiro quadro de vídeo, utilizando-se o primeiro quadro do seu *ground truth* para segmentá-lo de forma precisa e obter as distribuições de cores do elemento de interesse e do plano de fundo.

Importa ressaltar que, nos algoritmos de Criminisi et al. (2006) e Sanches, Silva e Tori (2012), os parâmetros α^{t-1} e α^{t-2} são desconhecidos no tempo $t = 1$ e, portanto, tem-se a energia $E^1 = E(\alpha^1|z^t, m^t)$. Da mesma forma, no tempo $t = 2$ apenas os termos V^S, U^C e U^M são conhecidos e a energia $E^2 = E(\alpha^2, \alpha^1|z^t, m^t)$ é minimizada.

I.3.4 Algoritmo de Stauffer e Grimson (2000)

O algoritmo apresentado por Stauffer e Grimson (2000) modela os valores de um pixel como uma mistura de gaussianas. Baseado na persistência e na variância de cada gaussiana da mistura, determina-se qual gaussiana pode corresponder as cores do plano de fundo. Valores de pixels que não se alinham com as distribuições do plano de fundo são consideradas primeiro plano até que haja uma gaussiana com evidências suficientes para incluí-los como uma nova mistura do plano de fundo.

Segundo Stauffer e Grimson (2000), os valores de um pixel em particular pode ser considerado um "processo de pixel". Isso significa que existem vetores de valores para a cor do pixel. Em qualquer tempo t, o que é conhecido sobre um pixel em particular $\{x_0, y_0\}$ é sua história recente

$$\{X_1, \ldots, X_t\} = \{I(x_0, y0, i) : 1 \leqslant i \leqslant t\} \tag{59}$$

onde I é o quadro de vídeo. A história recente $\{X_1, \ldots, X_t\}$ de cada pixel é modelada como uma mistura de distribuições gaussianas. As probabilidades de de observação de um determinado valor de pixel é dada por

$$P(X_t) = \sum_{t=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}), \tag{60}$$

onde K é o número de distribuições. $\omega_{i,t}$, $\mu_{i,t}$ e $\Sigma_{i,t}$ são, respectivamente, os pesos, a média e a matriz de covariância da $i$nésima gaussiana da mistura no tempo t. η é uma função densidade de probabilidade gaussiana, representada pela equação

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t - \mu_t)^T \sum^{-1} (X_t - \mu_t)} \tag{61}$$

Para diminuir o custo computacional, uma matriz de covariância é definida de forma simplificada, como na equação

$$\sum_{k,t} = \mu_k^2 I. \tag{62}$$

Desse modo, assume-se que cada canal de cor é independente e possui a mesma variância. O modelo é atualizado por meio de uma aproximação K-means *online* em que cada nova amostra de pixel é comparada com as distribuições Gaussianas. Uma comparação é

verdadeira quando o valor de um pixel se encontra dentro de 2,5 desvios padrão de uma distribuição. Caso nenhuma das distribuições corresponda a amostra de pixel em análise, a distribuição menos provável é substituída por outra, armazenando o seu valor médio, uma variância (inicialmente elevada) e um peso inicializado com valor baixo. Os pesos das K distribuições são ajustados de acordo com a equação

$$w_{k,t} = (1 - \alpha)w_{k,t-1} + \alpha(M_{k,t}) \tag{63}$$

onde α é a taxa de aprendizado e $M_{k,t}$ assume o valor 1, para as distribuições em que a comparação foi verdadeira, e o valor 0, para as demais distribuições. Em seguida, os pesos são novamente normalizados e os parâmetros μ e σ das distribuições em que a comparação foi falsa permanecem inalterados. Os parâmetros das distribuições que correspondem as novas observações são atualizados da seguinte forma

$$\mu_t = (1 - \rho)\mu_{t-1} + \rho X_t \tag{64}$$

$$\sigma_t^2 = (1 - \rho)\sigma_{t-1}^2 + \rho(X_t - \mu_t)^T(X_t - \mu_t) \tag{65}$$

onde ρ corresponde a taxa de aprendizado, dada por

$$\rho = \alpha\eta(X_t|\mu_k, \sigma_k) \tag{66}$$

O passo seguinte trata da definição de um método que decida quais gaussianas do modelo melhor representa o plano de fundo. Primeiramente, as gaussianas são ordenadas com base nos valores de w/σ. Desta forma, as distribuições que apresentam maior peso e menor desvio padrão (evidência e consistência) ficam posicionadas no topo da lista. Em seguida, as primeiras B distribuições são escolhidas como modelo do plano de fundo, onde

$$B = argmin_b \left(\sum_{k=1}^{b} w_k > T \right) \tag{67}$$

onde T é uma porção mínima de dados, cujo valor é escolhido de forma empírica, que deve ser considerada como fundo.

Apêndice II – INFORMAÇÕES E DADOS DAS AVALIAÇÕES SUBJETIVAS

Neste apêndice são apresentadas informações complementares que podem auxiliar o entendimento dos experimentos subjetivos realizados nesta pesquisa. São discutidos detalhes sobre a aplicação do método SAMVIQ e sua forma de análise dos resultados, conforme recomendada pela (ITU-R, 2009). Ainda neste apêndice, são expostas as configurações de visualização, empregadas durante os testes, e as especificações do sistema multimídia adotado (informações relacionadas aos equipamentos que influenciam a visualização). Finalmente, um relatório detalhado dos votos dos participantes é apresentado, no formato de tabelas.

II.1 Método de Avaliação de Qualidade de vídeo SAMVIQ

O SAMVIQ (KOZAMERNIK et al., 2005; ITU-R, 2007) consiste em um método formal de avaliação de qualidade de vídeo utilizados em aplicações multimídia. Sua utilização é recomendada por organizações, como a ITU e EBU, que sugerem o modo como deve ser realizada cada etapa da avaliação e a configuração física do ambiente de teste.

Detalhes como o número de observadores, tamanho e o tipo de tela, que deve ser apropriado para determinada aplicação, assim como a cor do fundo que completa a imagem, nas situações em que o sistema trabalha com imagens que não ocupam todo espaço da tela.

O processo de avaliação realizado por meio do método SAMVIQ é organizado da seguinte forma (ITU-R, 2007): a) o processo é aplicado a cada cena (conteúdo audiovisual),

como mostrado na figura 29; b) para cada cena, é possível visualizá-la e avaliá-la em qualquer ordem. Cada sequência (cena processada ou sem processamento) pode ser executada em qualquer ordem; c) na passagem de uma cena para outra, as sequências devem ser randomizadas; d) quando uma sequência é iniciada pela primeira vez, ela deve ser executada até o final antes de ser avaliada; e) a próxima cena só deve ser exibida quando todas das sequências teste da atual estiverem avaliadas; f) o teste é finalizado quando todas as sequências de todas as cenas são avaliadas. As notas são escolhidas em uma escala que vai de 0 a 100.

O método SAMVIQ se mostra apropriado no contexto de aplicações multimídia por ser possível combinar diferentes características de processamento de imagem (codificadores, formatos, taxa de atualização, etc). A palavra algoritmo, na figura, representa uma ou a combinação de algumas dessas características (ITU-R, 2007).

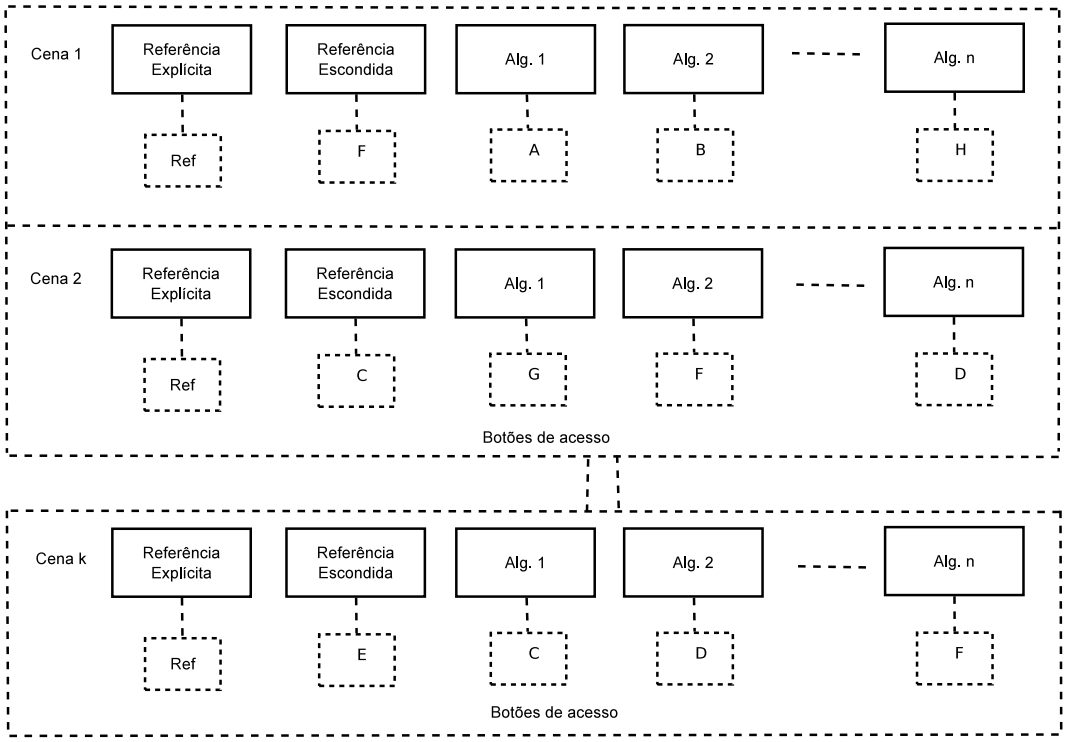


Figura 29 – Exemplo de organização de um teste utilizando o método SAMVIQ, adaptado de ITU-R (2007)

Pelo menos 15 (quinze) observadores devem realizar cada bateria de testes. Esses usuários também são submetidos a um teste de visão, baseado em cartões Ishihara (ITU-R, 2007), implementado na própria ferramenta. Os resultados desse teste, porém, não serão exibidos ao usuário. Em caso de não aprovação no teste de visão, a avaliação

desse usuário não é considerada na análise dos resultados.

Como os testes realizados por meio da SAMVIQ produzem distribuições de valores inteiros em uma escala que vai de 0 a 100, haverá variações dessas distribuições devido às diferenças de julgamento entre os observadores e o efeito que pode ser produzido por condições associadas com uma experiência específica, por exemplo, o uso de várias imagens ou de vídeos (ITU-R, 2009).

Em consequência disso, foram estabelecidos alguns critérios, apresentados em (ITU-R, 2009), para analisar os resultados obtidos da aplicação das avaliações. Nesse contexto, um teste consiste em um número de apresentações L e cada apresentação representa uma entre as várias condições de teste J, aplicada a uma entre várias sequências de teste K. Em alguns casos, cada combinação de sequências de teste e condições de teste pode ser repetida um número R de vezes. O primeiro passo na análise dos resultados é o cálculo da média dos escores, $\overline{u}_{jkr}$ para cada uma das apresentações:

$$\overline{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \tag{68}$$

onde u_{ijkr} é a nota do observador i para a condição de teste j, da sequência k, na repetição r e N é o número de observadores. Similarmente, pode-se calcular as notas médias globais, $\overline{U}_j$ e $\overline{U}_k$, correspondentes a cada condição de teste e para cada sequência de teste (ITU-R, 2009).

Quando se apresenta os resultados de um teste, todas as pontuações médias devem ter um intervalo de confiança associado, que é derivado do desvio padrão e do tamanho de cada amostra. Propõe-se usar o intervalo com 95% de confiança, que é dado por

$$[\overline{u}_{jkr} - \delta_{jkr}, \overline{u}_{jkr} + \delta_{jkr}] \tag{69}$$

onde

$$\overline{u}_{jkr} = 1.96\frac{S_{jkr}}{\sqrt{N}} \tag{70}$$

O desvio padrão para cada apresentação, S_{jkr}, é dado por

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\overline{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \tag{71}$$

Em relação a análise dos observadores, imagina-se que cada participante deve ter um método estável e coerente para votar em uma relativa degradação de qualidade em cada cena e algoritmo. Os critérios de rejeição verificam se o nível de coerência das notas de um observador segue a média de todos os observadores para uma determinada sessão (ITU-R, 2007). Isso é calculado utilizando uma correlação – com base nos coeficientes de correlação de Pearson e de rango de Spearman – das notas individuais em relação as notas médias correspondentes dos demais observadores (ITU-R, 2007).

II.2 Configuração do Ambiente dos Experimentos

Com o objetivo de fornecer um meio para avaliar a qualidade dos vídeos, do ponto de vista dos observadores (neste caso, usuários de sistemas de RA), a recomendação ITU-R (2009), adotada pelo SAMVIQ, sugere um ambiente de visualização que se assemelhe com o doméstico. Os parâmetros que configuram esse ambiente, no entanto, foram escolhidos para simular um ambiente levemente mais crítico, comparado as situações de visualização domésticas mais típicas. Esses parâmetros encontram-se listados na tabela 16.

Segundo a ITU-R (2007), o tamanho e o tipo de tela devem ser escolhidos conforme a aplicação sob investigação. Uma vez que várias tecnologias para visualização são utilizadas em aplicações multimídia, todas as informações pertinentes relativas ao sistema (por exemplo, fabricante, modelo e especificações), devem ser informados. Quando sistemas baseados em computadores pessoais são utilizados para apresentar os vídeos, as características dos sistemas, por exemplo, placa de vídeo, também deve ser informada. Essas informações podem ser visualizadas na tabela 17.

[1] Os valores de luminância e iluminância foram ajustados utilizando um Fotômetro Sekonic L-398A Studio Deluxe III

[2] Os valores na unidade de medida candelas por metro quadrado (cd/m^2) foram convertidos em valor de exposição (Exposure Value – EV) e utilizado o valor EV = 10.

Tabela 16 – Condições de visualização, recomendadas pela ITU, utilizadas na avaliação de qualidade dos vídeos

Parâmetro	Valor[1]
Distância de Visualização	1-8 H (30 cm)
Luminância máxima da tela[2]	70-250 cd/m^2
Razão entre luminância da tela inativa e luminância máxima	$\leqslant$ 0,05
Razão entre luminância da tela quando exibindo uma tela preta em uma sala completamente escura e luminância máxima de um ponto branco	$\leqslant$ 0,1
Razão entre luminância do ambiente atrás da tela e luminância máxima dos vídeos	$\leqslant$ 0,2
Iluminação da sala	$\leqslant$ 20 lux

Tabela 17 – Configuração do sistema multimídia utilizado nos testes

Parâmetro	Especificação
Tipo de Tela	LCD
Tamanho da Tela	19’
Placa de Vídeo	nVidia GeForce 6150SE
Fabricante	LG
Modelo	W1952TQ
Imagem	Resolução 1440 x 824

II.3 Relatório dos Votos

Nesta seção são apresentados os relatórios dos votos dos voluntários que participaram da avaliação subjetiva. Nas tabelas 18, 19, 20 e 21 são exibidos os resultados das avaliações, separadas por baterias de teste. Na tabelas 22 são exibidos os dados sobre os voluntários, como o gênero, a idade e as baterias de teste dos experimentos subjetivos em que participaram.

Tabela 18 – Relatório dos Votos dos Testes da Bateria 1

Vídeo-Fonte	Nº	Média ITU-R	Int. Conf. Esq.	Int. Conf. Dir.	Desv. Padrão
SEQ1	Ref.	8,63	7,93	9,34	1,39
SEQ1	1	5,46	4,60	6,32	1,71
SEQ1	2	2,11	1,55	2,67	1,10
SEQ1	3	5,05	4,19	5,92	1,72
SEQ1	4	4,17	3,33	5,01	1,66
SEQ1	5	3,93	3,30	4,57	1,25
SEQ1	6	3,67	2,81	4,54	1,71
SEQ2	Ref.	8,83	8,09	9,56	1,45
SEQ2	7	1,85	1,08	2,61	1,51
SEQ2	8	3,67	2,85	4,49	1,62
SEQ2	9	1,65	0,99	2,30	1,30
SEQ2	10	4,36	3,58	5,14	1,55
SEQ2	11	2,25	1,54	2,97	1,42
SEQ2	12	4,09	3,23	4,96	1,71
SEQ3	Ref.	8,93	8,32	9,53	1,20
SEQ3	13	2,61	1,96	3,27	1,30
SEQ3	14	4,38	3,28	5,48	2,18
SEQ3	15	2,74	1,98	3,50	1,50
SEQ3	16	2,87	2,14	3,61	1,46
SEQ3	17	2,35	1,73	2,97	1,22
SEQ3	18	2,83	2,16	3,50	1,33
SEQ4	Ref.	9,38	8,88	9,88	0,98
SEQ4	19	5,81	4,93	6,69	1,74
SEQ4	20	4,97	4,19	5,74	1,54
SEQ4	21	5,18	4,26	6,10	1,82
SEQ4	22	4,89	4,00	5,77	1,75
SEQ4	23	4,34	3,41	5,27	1,83
SEQ4	24	5,07	4,13	6,00	1,84

Tabela 19 – Relatório dos Votos dos Testes da Bateria 2

Vídeo-Fonte	Nº	Média ITU-R	Int. Conf. Esq.	Int. Conf. Dir.	Desv. Padrão
SEQ1	Ref.	8,85	8,45	9,25	0,79
SEQ1	1	5,96	4,95	6,97	1,99
SEQ1	2	2,71	1,69	3,73	2,01
SEQ1	3	5,28	4,26	6,30	2,02
SEQ1	4	4,38	3,26	5,50	2,21
SEQ1	5	4,97	3,79	6,16	2,34
SEQ1	6	4,69	3,61	5,77	2,13
SEQ2	Ref.	9,48	9,16	9,80	0,63
SEQ2	7	1,83	0,60	3,06	2,43
SEQ2	8	4,87	3,71	6,03	2,29
SEQ2	9	1,52	0,63	2,41	1,76
SEQ2	10	4,90	3,64	6,16	2,49
SEQ2	11	1,70	0,64	2,76	2,10
SEQ2	12	4,36	3,37	5,35	1,96
SEQ3	Ref.	8,91	8,20	9,63	1,41
SEQ3	13	3,37	2,20	4,53	2,30
SEQ3	14	4,36	3,19	5,53	2,32
SEQ3	15	3,39	2,34	4,44	2,08
SEQ3	16	3,71	2,52	4,89	2,34
SEQ3	17	3,12	2,06	4,18	2,10
SEQ3	18	4,09	2,86	5,33	2,44
SEQ4	Ref.	9,47	9,14	9,79	0,64
SEQ4	19	5,73	4,53	6,94	2,38
SEQ4	20	4,05	3,07	5,04	1,95
SEQ4	21	5,56	4,37	6,75	2,35
SEQ4	22	4,19	3,00	5,37	2,35
SEQ4	23	5,26	4,06	6,46	2,37
SEQ4	24	4,30	3,05	5,55	2,46

Tabela 20 – Relatório dos Votos dos Testes da Bateria 3

Vídeo-Fonte	Nº	Média ITU-R	Int. Conf. Esq.	Int. Conf. Dir.	Desv. Padrão
SEQ5	Ref.	8,41	7,08	9,74	2,63
SEQ5	1	6,55	5,40	7,70	2,27
SEQ5	2	4,96	3,88	6,04	2,14
SEQ5	3	5,89	4,63	7,14	2,47
SEQ5	4	3,76	2,89	4,63	1,72
SEQ5	5	6,25	5,04	7,47	2,4
SEQ5	6	3,51	2,73	4,29	1,54
SEQ2	Ref.	8,26	6,97	9,55	2,54
SEQ2	7	3,26	2,36	4,16	1,79
SEQ2	8	4,88	3,74	6,02	2,26
SEQ2	9	2,97	2,22	3,72	1,48
SEQ2	10	4,68	3,56	5,8	2,22
SEQ2	11	3,25	2,41	4,1	1,68
SEQ2	12	4,35	3,26	5,45	2,16
SEQ4	Ref.	8,73	7,81	9,65	1,82
SEQ4	13	4,51	3,49	5,52	2,01
SEQ4	14	1,36	0,84	1,88	1,02
SEQ4	15	4,43	3,45	5,42	1,94
SEQ4	16	1,45	0,93	1,97	1,03
SEQ4	17	4,37	3,57	5,18	1,59
SEQ4	18	1,46	0,96	1,96	0,99
SEQ5	Ref.	9,16	8,47	9,85	1,37
SEQ5	19	7,85	6,97	8,72	1,72
SEQ5	20	6,89	5,94	7,83	1,86
SEQ5	21	5,73	4,89	6,57	1,66
SEQ5	22	6,13	5,12	7,14	1,99
SEQ5	23	4,69	3,81	5,57	1,74
SEQ5	24	5,74	4,7	6,78	2,05

Tabela 21 – Relatório dos Votos dos Testes da Bateria 4

Vídeo-Fonte	Nº	Média ITU-R	Int. Conf. Esq.	Int. Conf. Dir.	Desv. Padrão
SEQ5	Ref.	9,05	8,37	9,73	1,34
SEQ5	1	5,78	4,77	6,79	1,99
SEQ5	2	5,90	4,84	6,96	2,09
SEQ5	3	4,85	3,87	5,82	1,92
SEQ5	4	5,02	4,03	6,01	1,95
SEQ5	5	4,89	3,75	6,03	2,25
SEQ5	6	4,03	2,95	5,10	2,13
SEQ2	Ref.	8,52	7,89	9,15	1,24
SEQ2	7	4,69	3,57	5,81	2,21
SEQ2	8	4,43	3,39	5,47	2,05
SEQ2	9	4,27	3,19	5,36	2,14
SEQ2	10	5,21	4,18	6,25	2,04
SEQ2	11	3,82	2,83	4,81	1,95
SEQ2	12	4,73	3,62	5,83	2,18
SEQ4	Ref.	8,19	7,38	8,99	1,59
SEQ4	13	3,62	2,70	4,54	1,82
SEQ4	14	4,81	3,94	5,69	1,73
SEQ4	15	3,61	2,95	4,26	1,29
SEQ4	16	5,39	4,60	6,19	1,57
SEQ4	17	3,68	2,87	4,49	1,60
SEQ4	18	4,90	3,98	5,82	1,81

Tabela 22 – Dados sobre os voluntários. Identificador, gênero, idade e baterias de teste dos experimentos subjetivos em que participaram

Ident. Voluntário	Gênero	Idade	Bateria 1	Bateria 2	Bateria 3	Bateria 4
1	F	38	X	X		
2	F	40		X		
3	M	32	X	X	X	
4	M	30		X	X	
5	M	28		X	X	
6	M	33	X	X		
7	M	40		X		
8	M	41	X	X	X	
9	F	64	X	X		X
10	M	33	X	X		
11	F	32	X	X	X	
12	M	39	X	X		
13	F	55		X	X	
14	M	28	X	X		
15	M	22		X		
16	F	30	X			
17	M	27	X			X
18	F	40	X			
19	M	36	X			
20	M	41	X		X	
21	M	36	X			
22	M	24			X	X
23	M	31			X	
24	M	20			X	
25	M	25			X	
26	M	34			X	
27	M	39			X	
28	M	35			X	X
29	M	37			X	X
30	M	40				X
31	M	24				X
32	F	34				X
33	F	37				X
34	M	42				X
35	F	64				X
36	F	22				X
37	M	24				X
38	M	36				X
39	F	36				X

Anexo A – APROVAÇÃO DO COMITÊ DE ÉTICA

São Paulo, 6 de agosto de 2011.

Il^mo(a). S^r(a).
Prof. Dr. Romero Tori
Departamento de Engenharia de Computação e Sistemas Digitais - PCS
Escola Politécnica
UNIVERSIDADE DE SÃO PAULO

REFERENTE: **Projeto de Pesquisa** "Segmentação de vídeo em tempo real para aplicações em Teleimersão" - **Pesquisador(a) responsável:** Prof. Dr. Romero Tori **– Co-Autor(es):** Silvio Ricardo Rodrigues Sanches - **Registro CEP-HU/USP**: 1129/11 – **SISNEP CAAE:** 0022.0.198.000-11.

Prezado(a) Senhor(a)

O Comitê de Ética em Pesquisa do Hospital Universitário da Universidade de São Paulo, em reunião ordinária realizada no dia 5 de agosto de 2011, analisou o Projeto de Pesquisa acima citado, considerando-o como **APROVADO,** bem como o seu **Termo de Consentimento Livre e Esclarecido.**

Lembramos que cabe ao pesquisador elaborar e apresentar a este Comitê, relatórios anuais (parciais ou final, em função da duração da pesquisa), de acordo com a Resolução nº 196/96 do Conselho Nacional de Saúde, inciso IX.2, letra "c".

O primeiro relatório está previsto para 5 de agosto de 2012.

Atenciosamente,

Dr. Maurício Seckler
Coordenador do Comitê de Ética em Pesquisa
Hospital Universitário da USP

COMITÊ DE ÉTICA EM PESQUISA DO HOSPITAL UNIVERSITÁRIO DA USP
Avenida Professor Lineu Prestes, 2565 – Cidade Universitária – 05508-000 São Paulo – SP
Tels.: (11) 3091-9457 – Fax: (11) 3091-9479 - E-mail: cep@hu.usp.br