Subjective Video Quality Assessment in Segmentation for Augmented Reality Applications
Avaliação Subjetiva de Qualidade na Segmentação de Vídeo para Aplicações em Realidade Aumentada

Silvio R. R. Sanches, Daniel M. Tokunaga, Valdinei F. Silva, Romero Tori
Universidade de São Paulo
São Paulo, Brasil
{*silviorrs,dmtokunaga,valdinei.freire*}*@usp.br, tori@acm.org*

***Abstract*—Video segmentation to extract a person in foreground has been a common task in many Augmented Reality (AR) applications. In natural environments which the background color and the light environment are not constant the segmentation method must be able to extract the element of the interest in these conditions. However, methods for segmentation in natural environment are more error prone than the traditional ones which are based on a constant color elimination. Thus, in order to use them in a large number of applications it is necessary to know how the segmentation errors are perceived by the users in order to focus on development of algorithms which avoid the more perceptible ones. In this work a video quality subjective assessment method was applied to obtain the AR user's opinions about videos with different misclassified pixels rates. The results showed that segmentation errors are perceived by AR applications users. However, the video quality was not related with the number of the misclassified pixels. In addiction, it was noted that when the errors concentrated in the element of the interest increase the score of the associated video decreases.**

***Keywords*-Bilayer Segmentation; Augmented Reality; Subjective Assessment; Segmentation Evaluation.**

***Resumo*—A segmentação de vídeos com o objetivo de extrair uma pessoa em primeiro plano é uma necessidade comum em muitas aplicações de Realidade Aumentada (RA). Quando essa tarefa deve ser realizada em ambientes não controlados, nos quais não se pode contar com homogeneidade de cor de fundo nem de iluminação, métodos de segmentação apropriados para essas condições devem ser utilizados. Tais métodos, no entanto, se mostram mais propensos a erros do que os tradicionais, baseados na subtração de uma cor de fundo homogênea. Para que a utilização desses métodos se torne viável para um maior número de aplicações, faz-se importante um estudo desses erros e seus impactos na percepção do usuário, a fim de que se possa concentrar esforços no desenvolvimento de algoritmos que reduzam erros de maior impacto. No presente trabalho foram realizadas avaliações subjetivas de qualidade, utilizando como fontes vídeos produzidos a partir de imagens segmentadas com diferentes percentuais de erros. Os resultados obtidos mostram que os erros de segmentação são perceptíveis aos usuários de aplicações de RA, mas a percepção não é necessariamente proporcional ao volume de erros. Observou-se, no entanto, que, a medida que os erros concentrados no elemento de interesse aumentam, as notas atribuídas aos vídeos decresce.**

***Palavras-chave*-Segmentação; Realidade Aumentada; Avaliação Subjetiva; Avaliação de Segmentação.**

I. Introdução

Segmentar uma imagem com o objetivo de extrair uma pessoa em primeiro plano de seu contexto original tem se tornado uma tarefa comum em muitos sistemas de Realidade Aumentada (RA). Teleconferência imersiva [1, 2] e jogos com RA [3, 4] são exemplos desses sistemas.

Em grande parte das aplicações de RA que exigem segmentação, o elemento de interesse extraído (Figura 1) é utilizado na geração de avatares. Os avatares – que são representações humanas em ambientes virtuais – podem gerados na forma plana [5], semi-plana [5, 6] ou tridimensional. Essa última exige múltiplas imagens e, em alguns casos, o auxílio de mapas de profundidade [7, 8, 9, 10, 11]. Existem, ainda, aplicações de RA em que a imagem segmentada é utilizada de forma alternativa, não necessariamente como uma textura aplicada sobre um modelo [1, 12].

Figura 1. Extração do elemento de interesse de seu contexto original.

Outra característica comum a maioria dos sistemas de RA que necessitam de segmentação é o fato de serem executados em ambientes em que a cena pode ser manipulada para que uma cor constante seja exibida como fundo [2, 7, 13]. Desse modo, o elemento de interesse pode ser isolado, de forma precisa, por meio de técnicas de segmentação como o chroma-key [14] ou a subtração de fundo [15, 16]. Existem, ainda, sistemas que utilizam equipamentos especializados, que geram mapas de profundidades para auxiliar a segmentação [7, 10].

978-0-7695-4725-1/25 $26.00 © 4725 IEEE
DOI 10.1109/SVR.2012.2

A identificação do que é primeiro plano e do que é fundo se torna mais problemática quando a aplicação de RA exige que essa tarefa seja realizada em ambientes naturais – com fundo arbitrário e sem controle de iluminação – e com captura realizada por meio de vídeo monocular [3, 17, 18] (sem o auxílio de sensores ou de câmeras adicionais calibradas). Os resultados apresentados em suas avaliações mostram que os métodos de segmentação que atuam nessas condições são mais propensos a erros [19, 20] que os baseados em cor de fundo homogênea.

Apesar de as aplicações executadas em ambiente controlado predominarem no contexto da RA, sistemas em que a segmentação deve ser realizada em ambientes naturais têm se mostrado frequente nos últimos anos [3, 17, 18, 21, 22, 23]. Pesquisas recentes têm produzido métodos que atuam em tempo real e que realizam a segmentação com base em imagens monoculares [19, 20, 24]. Aplicações como alguns sistemas de videoconferências [25], *videochats* [19, 20, 26] e jogos imersivos [3] têm adotado esses métodos, que também passaram a ser utilizados em sistemas de RA [3, 17].

Em consequência da dificuldade de segmentar uma imagem natural, a imagem resultante, que deveria conter apenas o elemento de interesse, pode apresentar-se com erros de classificação de pixels. Isto significa que pode haver pixels visíveis que pertençam ao fundo original (falsos positivos) ou pixels pertencentes ao elemento de primeiro plano podem ter sido eliminados no processo de segmentação (falsos negativos). A utilização dessas imagens pode influenciar consideravelmente na qualidade da cena exibida ao usuário, o que pode inviabilizar a utilização desses métodos em muitas aplicações.

Embora, intuitivamente, considere-se que quanto menor a quantidade de erros melhor será a qualidade da cena de RA exibida (levando-se em consideração apenas os problemas relacionados à segmentação), faz-se importante identificar o real impacto que esses erros podem causar aos usuários finais. Além disso, como podem se apresentar de formas diferentes na imagem, percentuais iguais de erros em diferentes imagens podem ser percebidos como diferentes pelos usuários. Desse modo, a percepção do observador torna-se um fator a ser considerado.

Uma forma de considerar a percepção do usuário em relação aos erros de segmentação é por meio da coleta de opiniões de pessoas a respeito da cena final exibida, construída a partir de imagens que apresentam erros de classificação de pixels. Avaliações subjetivas de qualidade de vídeo podem ser utilizadas com essa finalidade.

A aplicação de avaliações subjetivas formais de qualidade de imagem é uma abordagem comum na indústria da televisão [27] e, mais recentemente, os métodos utilizados vêm sendo adaptados para que a qualidade das imagens exibidas em aplicações multimídias também possam ser avaliadas [28, 29]. Alguns desses métodos mostram-se adequados para identificar como os erros de segmentação influenciam na qualidade da cena final, exibida em aplicações de RA.

Vários trabalhos que utilizam técnicas de avaliação subjetiva aplicada em RA são descritos no levantamento realizado em [30]. No entanto, estudos similares ao desta pesquisa – avaliação de segmentação em aplicações de RA – não foram encontrados na literatura. O levantamento desse tipo de informação pode mostrar a viabilidade da utilização de vários métodos de segmentação de imagens naturais em aplicações de RA, ainda que esses métodos não se mostrem tão robustos quanto os que agem em ambientes controlados.

Para que o experimento fosse realizado, foi necessário produzir cenas de RA em que vídeos que com diferentes tipos e percentuais de erros de segmentação fossem utilizados. Duas abordagens de segmentação de vídeos foram utilizadas para produzir esses vídeos. A primeira, conhecida como subtração de fundo [15, 16] – não robusta quando aplicada em ambientes em que a iluminação não é constante – consiste, basicamente, na comparação do quadro no tempo atual com uma imagem estática do fundo para extrair a camada de primeiro plano com base nos pixels não coincidentes dessas duas imagens, que pertencerão ao elemento de interesse.

A segunda abordagem, que é baseada na identificação da movimentação do elemento de interesse, busca reconhecer os pixels em movimento na imagem com o objetivo de eliminar os que pertencem ao fundo original. Para isso, esses métodos utilizam informações de cor, aliada, entre outras informações obtidas da imagem, a observação da coerência temporal dos quadros do vídeo [19, 20, 31].

Inserido nesse contexto, o objetivo do presente trabalho é levantar o quão perceptíveis aos usuários são os erros de segmentação, quando exibidos em uma ambiente de AR, além de verificar se as diferentes formas que esses erros se mostram na cena final, são também percebidas de forma diferente pelo usuário.

O restante do artigo está organizado da forma como segue. Na seção II é apresentada uma visão geral sobre segmentação de vídeos, no que se refere aos métodos que realizam a extração do elemento de interesse em ambientes não controlados. Os métodos de segmentação aplicados no experimento realizado são detalhados na mesma seção.

As avaliações subjetivas de qualidade de imagem e vídeos são discutidas na seção III, com ênfase no método utilizado neste trabalho. A forma de condução do experimento é detalhada na seção IV e os resultados obtidos são apresentados na seção V. Finalmente, na seção VI, são expostas as conclusões e as perspectivas de trabalhos futuros.

II. Segmentação de Vídeos

A extração de elementos de interesse em imagens para futura composição de cena é um problema estudado desde o inicio do século passado pela indústria de cinema e televisão [32]. O processo de segmentação consiste em dividir uma imagem em estruturas com conteúdo semântico para uma

aplicação específica [33]. Nas aplicações que fazem parte do contexto deste trabalho, a tarefa é finalizada quando uma pessoa em primeiro plano é isolada do seu plano de fundo original.

Segundo [34], uma imagem I_z deve ser entendida como uma combinação de um primeiro plano F_z com um fundo B_z, utilizando um canal α (alfa) que permite controlar a transparência do pixel. Desse modo, uma imagem pode ser representada pela equação $I = \alpha_z F_z + (1 - \alpha_z)B_z$, onde α_z assume valores entre [0,1]. Em processos de segmentação binária, objeto de estudo deste trabalho, α assume os valores 0 ou 1 (pixel pertence ao primeiro plano ou ao fundo, respectivamente). A determinação de valores fracionários de α com o objeto de suavizar a combinação do elemento de interesse com o novo fundo é conhecido como o problema do *matting* [35].

No contexto da segmentação binária, podem ser destacados dois métodos: um mais simplificado, apresentado em [15] e outro mais sofisticado, proposto em [20]. Os métodos citados foram utilizados nesta pesquisa pelo fato de os erros exibidos por cada um deles se mostrarem de forma diferente na imagem final.

O método considerado mais simples consiste na comparação do fundo com o quadro de vídeo atual

$$|z^t - z^{ref}| > Th \tag{1}$$

onde z^t representa um quadro de vídeo no tempo t e z^{ref} uma imagem de referência, capturada previamente, que contém apenas 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.

O método apresentado em [20], considerado mais sofisticado, é baseado em um arcabouço de minimização de energia, em que cada quadro de vídeo representa um vetor $z = (z_1, z_2, \cdots, z_n, \cdots, z_N)$ (no espaço de cores YUV), indexado por um pixel n. z^t representa um quadro no tempo t. A derivada temporal $\dot{z} = (\dot{z}_1, \dot{z}_2, \cdots, \dot{z}_n, \cdots, \dot{z}_N)$ é calculada $\dot{z}_n^t = |G(z_n^t) - G(z_n^{t-1})|$ em cada tempo t utilizando um *kernel* Gaussiano $G(.)$ com escala σ_t pixels.

Os gradientes espaciais $g = (g_1, g_2, \cdots, g_n, \cdots, g_N)$, onde $g_n = |\nabla z_n|$, são obtidos pela convolução de cada quadro com a derivada de primeira ordem do *kernel* Gaussiano com desvio padrão $\sigma_s = \sigma_t = 0.8$, mesmo valor utilizado em [20]. As observações de movimento são denotadas $m = (g, \dot{z})$ e a segmentação consiste em inferir um rótulo binário $\alpha^t \in \{F, B\}$, onde F e B representam a camada de primeiro plano e o plano de fundo, respectivamente. A probabilidade condicional é modelada por um Campo Aleatório Condicional (CRF) [36]

$$p(\alpha^1, \ldots, \alpha^t | z^1, \ldots, z^t, m^1, \ldots, m^t) \propto \exp - \left\{ \sum_{t'=1}^{t} E^{t'} \right\} \tag{2}$$

onde

$$E^t = E(\alpha^t, \alpha^{t-1}, \alpha^{t-2}, z^t, m^t). \tag{3}$$

A energia E^t associada com cada tempo t é decomposta como uma soma de 4 (quatro) termos:

$$\begin{aligned} E(\alpha^t, \alpha^{t-1}, \alpha^{t-2}, z^t, m^t) = \\ \eta V^T(\alpha^t, \alpha^{t-1}, \alpha^{t-2}) + \gamma V^S(\alpha^t, z^t) \\ + \rho U^C(\alpha^t, z) + \phi U^M(\alpha^t, \alpha^{t-1}, m^t), \end{aligned} \tag{4}$$

onde $V^T(\cdot)$ impõe a tendência de continuidade temporal, $V^S(\cdot)$ se baseia na tendência de continuidade espacial, $U^C(\cdot)$, avalia evidências baseado na distribuição de cores da imagem e $U^M(\cdot)$ usa as informações espacial e temporal para obter características de movimento. η, γ, ρ e ϕ são parâmetros de normalização [20].

III. Avaliações Subjetivas de Qualidade de Imagens e Vídeos

O aumento do número de aplicações que têm como resultado final um vídeo digital exibido ao usuário tem impulsionado pesquisas que buscam formas eficientes para avaliar a qualidade desses vídeos [37]. Um número considerável de métodos, resultados dessas pesquisas, podem ser encontrados na literatura e podem ser classificados em dois grupos [38]: subjetivos (envolvem seres humanos para avaliar a qualidade dos vídeos) e objetivos (a qualidade dos vídeos é calculada automaticamente, sem a participação de usuários).

As avaliações subjetivas têm se mostrado a forma mais eficiente de obter medições confiáveis [39]. Os métodos utilizados, tradicionalmente voltados para 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.

Alguns métodos reconhecidamente eficientes [39, 40], são particularmente populares, entre eles, o SAMVIQ (*Subjective Assessment Methodology for Video Quality*) [28, 29], que, de acordo com alguns estudos [39], tem se mostrado bastante preciso. No presente trabalho, o método SAMVIQ foi utilizado para levantar os erros de segmentação mais perceptíveis ao usuário.

A aplicação desses métodos tem sido recomendada por órgãos como a ITU[1] (*International Telecommunications Union*) e a EBU[2] (*European Broadcasting Union*), que sugerem como deve ser realizada cada etapa do processo de avaliação [41] e a configuração física do ambiente em que o teste é realizado [42].

Detalhes como o número de observadores, tamanho e o tipo de tela, que deve ser apropriado para a aplicação sendo

[1] http://www.itu.int
[2] http://www.ebu.ch

avaliada, assim como a cor do fundo da imagem, quando o sistema trabalha com imagens de tamanho reduzido fazem parte de recomendações.

O processo de avaliação realizado por meio do método SAMVIQ é organizado da seguinte forma [29]: a) o processo é aplicado a cada cena (conteúdo audiovisual), como mostrado na figura 2; 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 [29].

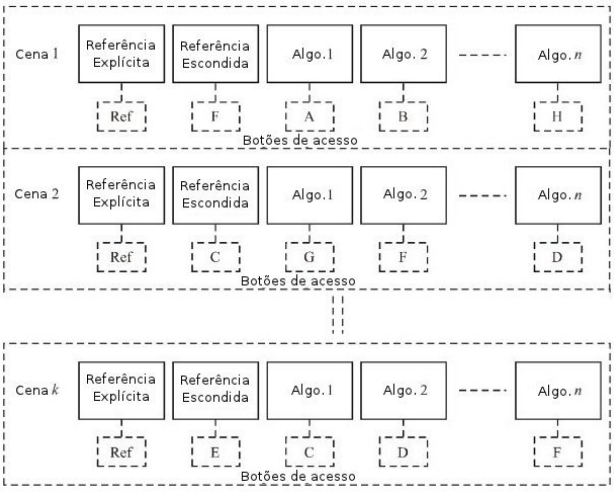


Figura 2. Exemplo de organização de um teste utilizando o método SAMVIQ, adaptado de [29].

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 [42].

Em consequência disso, foram estabelecidos alguns critérios, apresentados em [42], 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:

$$\overline{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \tag{5}$$

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 [42].

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 de confiança de 95%, que é dado por

$$[\overline{u}_{jkr} - \delta_{jkr}, \overline{u}_{jkr} + \delta_{jkr}] \tag{6}$$

onde

$$\overline{u}_{jkr} = 1.96\frac{S_{jkr}}{\sqrt{N}} \tag{7}$$

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

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\overline{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \tag{8}$$

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 [29]. 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 [29].

IV. Realização do Experimento com base no SAMVIQ

O desenvolvimento do experimento foi realizado utilizando-se o método de avaliação subjetiva de qualidade de imagem SAMVIQ [28]. Uma vez que a presente pesquisa volta-se aos métodos de segmentação que sejam aplicáveis em sistemas de RA, quando executados em ambientes não controlados e que não fazem uso de equipamentos específicos que auxiliem a segmentação, concentrou-se na avaliação de vídeos que exibissem erros produzidos por métodos aplicáveis nesse contexto.

A. *Preparação da Base de Vídeos*

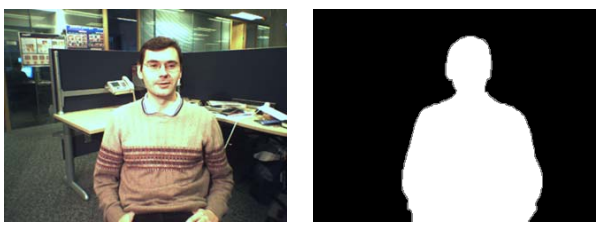
Foram utilizados como vídeos fonte 4 (quatro) sequências diferentes, chamadas de SEQ1, SEQ2, SEQ3 e SEQ4. Nas duas primeiras, o elemento de interesse na cena – a pessoa em primeiro plano – encontrava-se distante da câmera, portanto, todo seu corpo pode ser visualizado. Nas demais, a câmera foi posicionada mais próxima e apenas a parte superior do corpo pôde ser visualizada. Um quadro de vídeo de cada sequência é mostrado na figura 3.

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

Figura 3. Quadros das sequências de vídeo originais: SEQ1, SEQ2, SEQ3 e SEQ4

Todas as sequências originais utilizadas possuem um *ground truth*. Em outras palavras, para cada quadro de vídeo, existe um quadro correspondente segmentado de maneira precisa, como mostrado na figura 4. Desse modo, foi possível calcular a quantidade de erros de classificação para cada sequência. Os pixels dos quadros de vídeo do *ground truth* foram rotulados como primeiro plano (pixels branco), plano de fundo (pixels pretos) e região desconhecida (pixels cinza).

As sequências SEQ2 e SEQ4, e seus respectivos *ground truths*, foram obtidos da base de dados disponível em [43], ao passo que as demais foram capturadas e rotuladas manualmente.

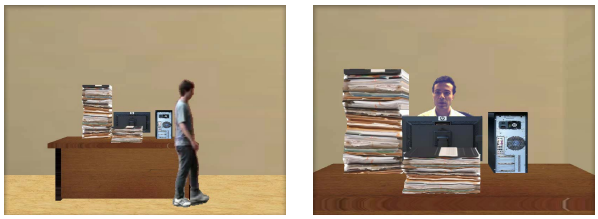


(a) (b)

Figura 4. Quadro da sequência de vídeo SEQ2 e seu respectivo *ground truth*.

A partir das sequências de vídeo originais, foram produzidas novas sequências, de curta duração (10s) e com resolução 800x600 pixels. Nesses vídeos, era possível ser visualizado o elemento de interesse do quadro original (sem o fundo) aplicado como textura sobre um plano, dentro de um ambiente virtual. O ambiente virtual desenvolvido para os testes simula o cenário de um escritório.

Obteve-se, portanto, um avatar, do tipo *billboard*, em que o plano que contém a textura permanece com a face principal sempre voltada para o usuário, independentemente do ponto de vista escolhido por ele. 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, como mostrado na figura 5.



(a) (b)

Figura 5. Exemplos de vídeos produzidos para o experimento: (a) Quadro de vídeo da SEQ1 e (b) e SEQ2.

Para cada sequência original foram produzidos 6 (seis) vídeos, que exibiam um ambiente de RA, como o mostrado na figura 5. 3 (três) desses vídeos foram segmentados utilizando o método baseado em Subtração de Fundo (Sub.) [15] e, nos outros 3 (três), o método proposto por Criminisi (Crim.) *et al* [20]. Ambos os métodos de segmentação foram detalhados na seção II. Cada vídeo produzido possui um percentual diferente de erros de segmentação, como pode ser visualizado na tabela I. Os pixels que fazem parte da região desconhecida, onde aplica-se técnicas de *matting*, foram desconsiderados na contagem dos erros.

Um vídeo de referência, que não possui erros, também foi produzido para cada sequência. Os quadros exibidos na figura 5 pertencem aos vídeos de referência gerados a partir das sequências SEQ1 e SEQ4. As referências foram segmentadas utilizando seus respectivos *ground truths*, o que resultou em sequências livres de erros de segmentação.

Os erros propositadamente introduzidos variam na faixa de 0% (vídeo de referência) até 31,85% (pior caso). Os diferentes percentuais foram alcançados, no método de subtração de fundo, alterando-se o valor do limiar Th, da equação 1, e, no método de Criminisi *et al.*, alterando os valores dos parâmetros de normalização do CRF η, γ, ρ e ϕ, da equação 4. Dois quadros desses vídeos podem ser visualizados na figura 6.

Tabela I
QUANTIDADE DE ERROS DE CLASSIFICAÇÃO DE PIXELS PRESENTES NA IMAGEM.

SEQ1			SEQ2		
N^o	Método	Erros	N^o	Método	Erros
Ref.	–	0%	Ref.	–	0%
1	Crim.	1,01%	7	Sub.	28.82%
2	Sub.	2,08%	8	Crim.	7.61%
3	Crim.	3,09%	9	Sub.	29.72%
4	Sub.	4,17%	10	Crim.	8.27%
5	Crim.	5,13%	11	Sub.	31.85%
6	Sub.	6,46%	12	Crim.	10.09%
SEQ3			SEQ4		
N^o	Método	Erros	N^o	Método	Erros
Ref.	–	0%	Ref.	–	0%
13	Sub.	8.36%	19	Crim.	12.68%
14	Crim.	9.80%	20	Sub.	12.87%
15	Sub.	10.20%	21	Crim.	15.65%
16	Crim.	11.95%	22	Sub.	15.74%
17	Sub.	13.43%	23	Crim.	17.50%
18	Crim.	14.03%	24	Sub.	19.63%

(a) (b)

Figura 6. Exemplos de vídeos produzidos para o experimento: (a) quadro de vídeo da sequência número 8 e (b) quadro de vídeo da sequência número 13, ambos descritos na tabela I .

B. Aplicação dos Testes

A avaliação dos vídeos produzidos foi realizada aplicando-se o método SAMVIQ [28], implementado na ferramenta disponível em [44]. Um total de 15 (quinze) voluntários participaram do experimento. A única restrição em relação ao perfil dos participantes, conforme definido pelo SAMVIQ, é que os voluntários não tenham a avaliação de imagem como sua ocupação principal. Procurou-se manter as mesmas condições de ambiente em todos os testes. Cada participante era posicionado a uma distância de aproximadamente 30 (trinta) centímetros da tela (LCD 19').

Inicialmente, os procedimentos foram explicados e dúvidas em relação a aplicação dos testes foram esclarecidas. Os 4 (quatro) conjuntos de vídeos foram apresentados a cada participante obedecendo a sequência descrita na tabela I. No entanto, a exibição dos vídeos (com diferentes percentuais de erros) que pertencem a um mesmo conjunto era aleatória.

Cada bateria de exibição (que possuíam 6 vídeos cada) era precedida da exibição do vídeo de referência (sem votação). Além disso, o vídeo de referência também era exibido ao participante da mesma forma que os demais (referência escondida), recebendo votação. A tela do ambiente de avaliação é mostrada na figura 7.

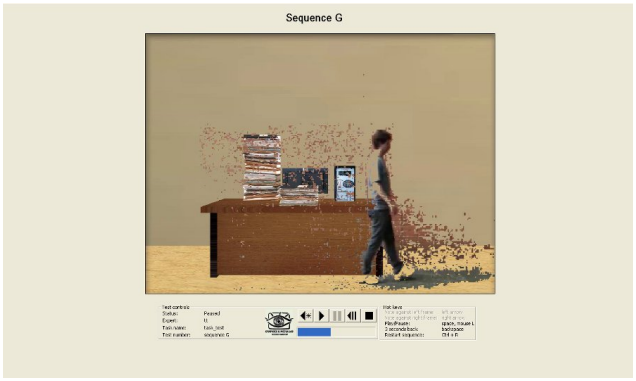


Figura 7. Ambiente de avaliação do método SAMVIQ [44]

Quando todas as versões de uma sequência eram avaliadas pelo participante, a próxima sequência era iniciada. A avaliação era realizada ao final de cada versão exibida, por meio da escolha de um valor na escala, mostrada na figura 8. O teste era finalizado quando os 24 (vinte e quatro) vídeos com seus respectivos percentuais de erros, além das referências escondidas, eram avaliados.

Cada participante teve acesso a várias versões de uma mesma sequência de vídeo – cada versão consiste em uma combinação de um dos métodos de segmentação com um valor ou conjunto de valores de parâmetros diferentes, como mostrado na tabela I. Quando todas as versões de um vídeo é avaliada pelo participante, a próxima sequência é iniciada. A avaliação é realizada a cada versão exibida, por meio da escolha de valores em uma escala contínua de qualidade, com valores de 0 a 100, agrupados em 5 itens (excelente, bom, regular, ruim e péssimo), como pode ser visualizado na figura 8.

Figura 8. Escala contínua de qualidade, utilizada pelo método SAMVIQ.

As diferentes versões podem ser selecionadas aleatoriamente pelo participante, a quem é permitido parar, rever e modificar notas que foram atribuídas anteriormente. Uma referência explícita (um vídeo sem erros de segmentação, que não é avaliado) e referências escondidas (que são avaliadas) foram introduzidas, conforme sugerido pelo SAMVIQ.

V. RESULTADOS

Inicialmente, os dados obtidos do experimento foram analisados com base nos critérios discutidos na seção III com o objetivo de identificar se pequenos percentuais de erros de segmentação são percebidos pelos usuários. Na tabela II podem ser visualizados a avaliação do usuário em relação a referência implícita de cada sequência, os intervalos de confiança direito e esquerdo, o desvio padrão e a melhor média de avaliação recebida entre as 6 (seis) variações avaliadas de uma mesma sequência.

Tabela II
DADOS DA AVALIAÇÃO DOS VÍDEOS DE REFERÊNCIA.

Ref.	Aval.	Int. Esq.	Int. Dir.	D. Padrão	Melhor
SEQ1	8.63	7.93	9.34	1.39	5.46
SEQ2	8.83	8.09	9.56	1.45	4.36
SEQ3	8.93	8.32	9.53	1.20	4.38
SEQ4	9.38	8.88	9.88	0.98	5.18

Como pode ser observado, mesmo na SEQ1 em que os percentuais de erros são menores – próximo de 1% no melhor caso – é possível afirmar, com um grau de confiança de 95%, que erros de segmentação são percebidos, ainda que ocorram em baixa quantidade.

Uma vez constatado que os erros de segmentação são percebidos, procurou-se verificar a influência da forma em que esses erros ocorrem. No gráfico exibido na figura 9 pode ser observado que a percepção do usuário não se mostra proporcional ao volume de erros. Vídeos que possuíam diferenças significativas (em relação ao percentual de erros) foram percebidos como semelhantes e vídeos com percentuais de erros próximos receberam avaliações significativamente diferentes.

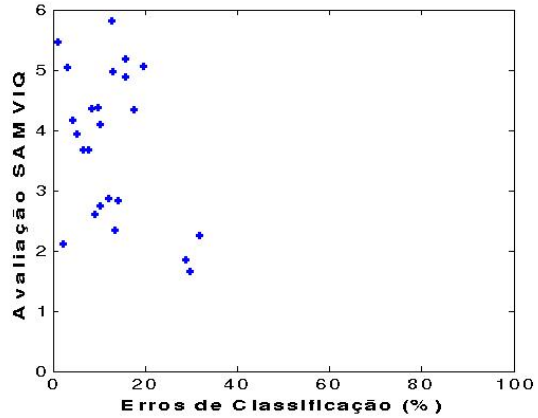


Figura 9. Relação entre a avaliação do usuário e a média de erros de classificação de pixels nas sequências de vídeo.

Dois atributos considerados relevantes, que foram alvo de investigação neste trabalho, consistem nas quantidades de falsos negativos e falsos positivos. Considera-se um falso positivo um pixel do fundo original classificado como pertencente ao elemento de interesse e, falso negativo, um pixel do elemento de interesse eliminado durante o processo de segmentação.

O gráfico exibido na figura 10 mostra a relação entre os falsos positivos e a avaliação subjetiva do usuário. Como pode ser observado, não é possível identificar tendências nas informações exibidas.

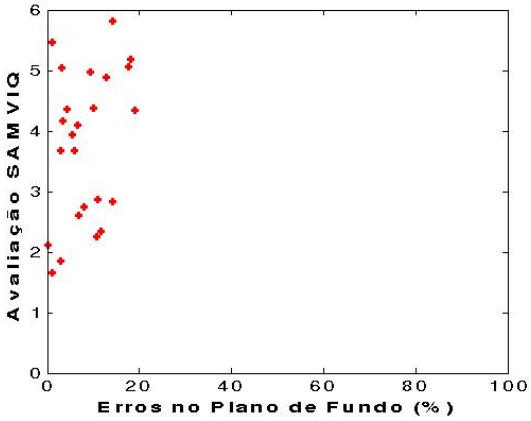


Figura 10. Relação entre a avaliação do usuário e o percentual de falsos positivos ocorridos nas sequências de vídeo.

De forma similar a análise anterior, verificou-se a existência de uma possível relação entre a avaliação subjetiva do usuário e o percentual de falsos negativos. Ao contrário do que ocorre com os falsos positivos, e de acordo com o gráfico exibido na figura 11, pode ser observada uma relação entre a ocorrência de falsos negativos e a qualidade percebida pelos usuários.

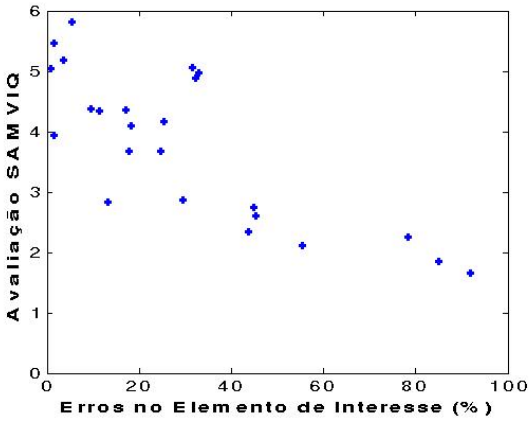


Figura 11. Relação entre a avaliação do usuário e o percentual de falsos negativos ocorridos nas sequências de vídeo.

Um fator a ser ressaltado, no entanto, é a ocorrência

de algumas avaliações próximas a 5 (cinco), consideradas boas, atribuídas a vídeos com erros próximos a 35%. Essas avaliações podem ser explicadas observando-se os quadros de vídeo mostrados na figura 12.

(a) (b)

Figura 12. Situações específicas que podem levar a avaliações inconsistentes. (a) falsos negativos são mascarados pela cor semelhante do novo fundo e (b) elemento do fundo original extraído de forma indesejada interpretado como um objeto virtual que compõe o ambiente de RA.

Duas situações específicas podem ser notadas. Na primeira, mostrada na figura 12(a), os falsos negativos ocorridos são mascarados em consequência da semelhança das cores do elemento de interesse e do novo fundo. A segunda situação, exibida na figura 12(b), ocorre quando um elemento do fundo original é extraído juntamente com o elemento de interesse. Esse elemento pode se confundir com os objetos virtuais presentes na cena e ser interpretado como pertencente ao ambiente. No entanto, os pixels pertencentes a esse elemento são, de fato, erros de segmentação.

VI. Conclusão

O experimento realizado neste trabalho teve como objetivos levantar a influência dos erros de segmentação na percepção de um usuário de sistemas de RA e verificar se a percepção varia proporcionalmente com a quantidade de erros ou se há outros fatores que a influenciam.

Para isso foi desenvolvido um experimento que utilizou o método subjetivo de avaliação de qualidade de imagem SAMVIQ. Os vídeos que serviram como base para a análise apresentavam diferentes percentuais de erros de segmentação, simulados pelo ajuste de parâmetros de dois métodos de segmentação de vídeos aplicáveis em ambientes não controlados.

Verificou-se que os erros de segmentação são percebidos pelo usuário de aplicações de RA e que a forma a qual os erros ocorrem influencia sua percepção. Constatou-se que vídeos com diferenças significativas em termos de percentual de erros foram igualmente avaliados pelos usuários e que, a medida que os erros de classificação concentrados no elemento de interesse aumentam, esses erros se tornam mais evidentes.

Os resultados obtidos do presente experimento podem ser considerados por aplicações de RA que exibam representações planas do usuário no ambiente de RA [3, 17]. Em métodos de geração de avatares que se baseiam em múltiplas capturas [21], os erros de segmentação podem se mostrar de formas diferentes das analisadas neste artigo. Além disso, uma vez que câmeras ao redor de um usuário sejam necessárias, a utilização de abordagens que agem em ambientes arbitrários não se justifica.

Pretende-se, como trabalhos futuros, realizar novos experimentos que evolvam outros domínios de aplicação, além de aplicar testes que permitam levantar novos atributos relacionados aos erros de segmentação que mais influenciam a percepção do usuário.

Agradecimentos

Silvio R. R. Sanches e Daniel M. Tokunaga agradecem à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e Valdinei F. Silva agradece à FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo – Processo 11/19280-8), pelos seus respectivos apoios financeiros. Os autores agradecem também ao Instituto Nacional de Ciência e Tecnologia – Medicina Assistida por Computação Científica (INCT-MACC – Processo 573710/2008-2 Edital MCT/CNPq Nº 015/2008), pelos equipamentos utilizados nos testes. Esta pesquisa foi aprovada pelo Comitê de Ética em Pesquisa do Hospital Universitário da Universidade de São Paulo (Registro CEP-HU/USP: 1129/11 – SISNEP CAAE: 0022.0.198.000.11.

Referências

[1] P. Kauff e O. Schreer, "An immersive 3d video-conferencing system using shared virtual team user environments," in *Proceedings of the 4th international conference on Collaborative virtual environments*, sér. CVE '02. New York, NY, USA: ACM, 2002, pp. 105–112.

[2] S.-Y. Lee, I.-J. Kim, S. C. Ahn, H. Ko, M.-T. Lim, e H.-G. Kim, "Real time 3d avatar for interactive mixed reality," in *VRCAI '04: Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry*. New York, NY, USA: ACM, 2004, pp. 75–80.

[3] R. Nakamura, L. L. M. Lago, A. B. Carneiro, A. J. C. Cunha, F. J. M. Ortega, J. L. Bernardes-Jr, e R. Tori, "3PI experiment: immersion in third-person view," in *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, sér. Sandbox '10. New York, NY, USA: ACM, 2010, pp. 43–48.

[4] P. Hämäläinen, T. Ilmonen, J. Höysniemi, M. Lindholm, e A. Nykänen, "Martial arts in artificial reality," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, sér. CHI '05. New York, NY, USA: ACM, 2005, pp. 781–790.

[5] C. G. Corrêa, D. M. Tokunaga, S. R. R. Sanches, R. Nakamura, e R. Tori, "Immersive teleconferencing system based on video-avatar for distance learning," in *Virtual Reality (SVR), 2011 XIII Symposium on*, 2011, pp. 197–206.

[6] D. M. Tokunaga, S. R. R. Sanches, L. P. Trias, R. Nakamura, J. L. Bernardes, e R. Tori, "Video-based microfacet-billboard avatar for educational immersive teleconference systems," in *SVR '09: Proceedings of XI Symposium on Virtual and Augmented Reality*. Porto Alegre, RS, Brasil: Sociedade Brasileira da Computação, 2009, pp. 199–209.

[7] K. Tamagawa, T. Yamada, T. Ogi, e M. Hirose, "Developing a 2.5-d video avatar," *Signal Processing Magazine, IEEE*, vol. 18, no. 3, pp. 35–42, 2001.

[8] S. Prince, A. Cheok, F. Farbiz, T. Williamson, N. Johnson, M. Billinghurst, e H. Kato, "3d live: real time captured content for mixed reality," *Mixed and Augmented Reality, 2002. ISMAR 2002. Proceedings. International Symposium on*, pp. 7–13, 2002.

[9] B. Goldlucke e M. Magnor, "Real-time microfacet billboarding for free-viewpoint video rendering," *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 3, pp. III–713–16 vol.2, 2003.

[10] T. Kanade e P. J. Narayanan, "Virtualized reality: Perspectives on 4d digitization of dynamic events," *IEEE Comput. Graph. Appl.*, vol. 27, no. 3, pp. 32–40, 2007.

[11] T. Shin, N. Kasuya, I. Kitahara, Y. Kameda, e Y. Ohta, "A comparison between two 3d free-viewpoint generation methods: Player-billboard and 3d reconstruction," in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*, 2010, pp. 1 –4.

[12] S. R. R. Sanches, A. C. Sementille, I. A. Rodello, e J. R. F. Brega, "The generation of scenes in mixed reality environments using the chromakey technique," in *ICAT '07: Proceedings of the 17th International Conference on Artificial Reality and Telexistence*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 296–297.

[13] D. M. Tokunaga, R. Nakamura, e R. Tori, "Non-photorealistic 3d video-avatar," in *SIGGRAPH '09: Posters*, sér. SIGGRAPH '09. New York, NY, USA: ACM, 2009, pp. 101:1–101:1.

[14] S. Gibbs, C. Arapis, C. Breiteneder, V. Lalioti, S. Mostafawy, e J. Speier, "Virtual studios: an overview," *Multimedia, IEEE*, vol. 5, no. 1, pp. 18–35, 1998.

[15] R. Qian e M. Sezan, "Video background replacement without a blue screen," *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, vol. 4, pp. 143–146 vol.4, 1999.

[16] M. Piccardi, "Background subtraction techniques: a review," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 4, 2004, pp. 3099 – 3104 vol.4.

[17] S. R. R. Sanches, D. M. Tokunaga, V. F. Silva, A. C. Sementille, e R. Tori, "Mutual occlusion between real and virtual elements in augmented reality based on fiducial markers," in *Applications of Computer Vision (WACV), 2012 IEEE Workshop on*, 2012, pp. 49 –54.

[18] T. Koyama, I. Kitahara, e Y. Ohta, "Live mixed-reality 3d video in soccer stadium," in *Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality*, sér. ISMAR '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 178–.

[19] P. Yin, A. Criminisi, J. Winn, e I. Essa, "Bilayer segmentation of webcam videos using tree-based classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 1, pp. 30–42, 2011.

[20] A. Criminisi, G. Cross, A. Blake, e V. Kolmogorov, "Bilayer segmentation of live video," in *CVPR '06: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. Washington, DC, USA: IEEE Computer Society, 2006, pp. 53–60.

[21] H. Kim, R. Sakamoto, I. Kitahara, N. Orman, T. Toriyama, e K. Kogure, "Compensated visual hull for defective segmentation and occlusion," 2007, pp. 210–217.

[22] T. Kakuta, L. B. Vinh, R. Kawakami, T. Oishi, e K. Ikeuchi, "Detection of moving objects and cast shadows using a spherical vision camera for outdoor mixed reality," in *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*, sér. VRST '08. New York, NY, USA: ACM, 2008, pp. 219–222.

[23] J. Ko, S. Lee, S. Kang, e J. Lee, "Hybrid camera based real-time human body segmentation for virtual reality e-learning system," in *Computers, Networks, Systems and Industrial Engineering (CNSI), 2011 First ACIS/JNU International Conference on*, 2011, pp. 116 –118.

[24] L. Wang, C. Zhang, R. Yang, e C. Zhang, "Tofcut: Towards robust real-time foreground extraction using a time-of-flight camera," in *Fifth International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010.

[25] C. Harrison e S. E. Hudson, "Pseudo-3d video conferencing with a generic webcam," in *Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia*, sér. ISM '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 236–241.

[26] J. Sun, W. Zhang, X. Tang, e H.-Y. Shum, "Background cut," in *ECCV 2006 – Proceedings of European Conference on Computer Vision*, 2006, pp. 628–641.

[27] V. Baroncini, "New tendencies in subjective video quality evaluation," *IEICE Transactions on Fundamentals of Electronics*, vol. E89-A, no. 11, pp. 2933–2937, 2006.

[28] F. Kozamernik, V. Steinmann, P. Sunna, e E. Wyckens, "Samviq – a new ebu methodology for video quality evaluations in multimedia," *SMPTE Motion Imaging Journal*, vol. 114, no. 4, pp. 152 – 160, 2005.

[29] ITU-R, "Recommendation ITU-R BT.1788 – methodology for the subjective assessment of video quality in multimedia applications," International Telecommunications Union, BT Series Broadcasting service (television) BT.1788, 2007.

[30] A. Dünser, R. Grasset, e M. Billinghurst, "A survey of evaluation techniques used in augmented reality studies," in *ACM SIGGRAPH ASIA 2008 courses*, sér. SIGGRAPH Asia '08. New York, NY, USA: ACM, 2008, pp. 5:1–5:27.

[31] A. Parolin, G. P. Fickel, C. R. Jung, T. Malzbender, e R. Samadani, "Bilayer video segmentation for videoconferencing applications," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, 2011, pp. 1–6.

[32] F. D. Williams, "Method of taking motion pictures," U.S. Patente 1,273,435, 23, 1918.

[33] H. Pedrini e W. R. Schwartz, *Análise de Imagens Digitais: Princípios, Algoritmos e Aplicações*, 1a ed. Thomson Learning, 2008.

[34] T. Porter e T. Duff, "Compositing digital images," in *SIGGRAPH '84: Proceedings of the 11th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1984, pp. 253–259.

[35] J. Wang e M. F. Cohen, "Image and video matting: a survey," *Found. Trends. Comput. Graph. Vis.*, vol. 3, no. 2, pp. 97–175, 2007.

[36] J. D. Lafferty, A. McCallum, e F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[37] K. Seshadrinathan, R. Soundararajan, A. Bovik, e L. Cormack, "Study of subjective and objective quality assessment of video," *Image Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1427–1441, 2010.

[38] K.-H. Thung e P. Raveendran, "A survey of image quality measures," in *Technical Postgraduates (TECHPOS), 2009 International Conference for*, 2009, pp. 1 –4.

[39] S. Péchard, R. Pépion, e P. L. Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," in *International Workshop on Image Media Quality and its Applications (IMQA2008)*, 2008.

[40] T. Tominaga, T. Hayashi, J. Okamoto, e A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, 2010, pp. 82 –87.

[41] L. C. Daronco, V. Roesler, e J. V. de Lima, "Subjective video quality assessment applied to scalable video coding and transmission instability," in *Proceedings of the 2010 ACM Symposium on Applied Computing*, sér. SAC '10. New York, NY, USA: ACM, 2010, pp. 1898–1904.

[42] ITU-R, "Recommendation ITU-R BT.500-12 – methodology for the subjective assessment of the quality of television pictures," International Telecommunications Union, BT Series Broadcasting service (television) BT.500-12, 2009.

[43] Microsoft Corporation, "Microsoft research – free research data," 2010, acessado em maio de 2012, disponível em http://research.microsoft.com/en-us/projects/i2i/data.aspx.

[44] D. Vatolin, "MSU perceptual video quality tool," 2012, acessado em fev. de 2012, disponível em http://compression.ru/video/quality_measure/perceptual_video_quality_tool_en.html.