Compressão híbrida de imagens estéreo

Licius S. Kreulich, Carla L. Pagliari e Marcelo M. Perez

***Resumo*— Para tornar viável o funcionamento dos televisores auto-estereoscópicos, sem se transmitir todas as vistas, é necessário sintetizar vistas intermediárias a partir de pelo menos duas vistas, este artigo apresenta um sistema de compressão que consome uma taxa de bits inferior à necessária para se codificar as imagens de um sistema estéreo tradicional, mantendo a qualidade subjetiva da vista sintetizada.**

Palavras-Chave—compressão estéreo, 3D, multi-vistas

***Abstract*—To enable the operation of the auto-stereoscopic displays, without transmitting all necessary views, it is necessary to synthesize intermediate views from at least two views. This paper presents a compression scheme that uses a bit rate lower than the necessary to encode the images using a traditional stereo system, keeping the (perceptual) subjective quality of the synthesized views.**

Keywords—stereo compression, 3-D, multi-view

I. INTRODUÇÃO

A representação tridimensional (3D) de uma cena real permite que se perceba a profundidade da cena em questão. No entanto, o uso da tecnologia 3D ainda não se mostra possível para todas as aplicações que dela necessitam. Isso porque existem alguns desafios a serem superados, como: necessidade de *displays* especiais que dispensem o uso dos incômodos óculos, necessidade de técnicas de compressão eficientes uma vez que para tal representação (3D) pressupõe-se que seja necessário no mínimo o dobro de informação a ser armazenada e transmitida, entre outros

Uma das finalidades de um sistema estéreo é reproduzir a sensação de profundidade. Para que isto seja possível, é necessário que se capture, comprima e transmita pelo menos duas vistas da mesma cena. Também, é desejável a síntese de vistas virtuais que vão auxiliar a visualização 3D nos futuros televisores auto-estereoscópicos A idéia é transmitir a menor quantidade de informações possível mantendo-se a qualidade do sinal, bem como possibilitando a recuperação das informações necessárias para a reconstrução da informação 3D.

Como para viabilizar o funcionamento dos displays auto-estereoscópicos, sem se transmitir todas as vistas, é necessário sintetizar vistas intermediárias a partir de pelo menos duas vistas, este artigo apresenta um sistema de compressão que consome uma taxa de bits inferior à necessária para se codificar as imagens de um sistema estéreo tradicional, mantendo a qualidade subjetiva da vista sintetizada.

Licius S. Kreulich e Carla L. Pagliari, Departamento de Engenharia Elétrica, Instituto Militar de Engenharia. Marcelo M. Perez, Departamento de Ciência e Tecnologia, Exército Brasileiro. Este trabalho foi financiado pelo Prograna CAPES/Pró-Defesa.

Este artigo apresenta um sistema de compressão para imagens estéreo, onde a idéia principal é gastar menos bits ao se codificar um conjunto de informações do que ao se codificar o par estéreo original acrescido das informações de disparidades (ou profundidades) mantendo a qualidade das imagens sintetizadas após a decodificação. Como as informações de disparidades podem ser representadas como imagens em escala de cinza, codificadores de imagem como o H.264/AVC INTRA [1] e JPEG-2000 [2] são candidatos naturais não apenas para a codificação das vistas, mas também para a codificação dos mapas de disparidades e, dessa forma, serão utilizados no método proposto pelo presente trabalho. Uma extensão recente do padrão H.264/AVC, denominada Multiview Video Coding (MVC) também é empregada para codificar alguns tipos de imagens.

Uma das vistas (a vista da esquerda ou da direita) poderá ser sintetizada a partir das informações obtidas a partir da decodificação da vista complementar (direita ou esquerda) do par estéreo, dos mapas de disparidades, com o auxílio do mapa de oclusões e de um mapa composto. Este último mapa será composto da informação de textura da vista complementar apenas nas partes oclusas. Assim, em vez de se codificar e transmitir ambas as vistas, apenas uma das vistas é codificada e transmitida, juntamente com o mapa composto, o mapas de oclusões e o mapa de disparidades. A vantagem do método advém do fato de que o mapa composto é codificado com uma taxa de bits inferior à da vista original.

Este trabalho analisa o impacto da síntese de vistas intermediárias a partir deste conjunto de imagens. Uma vez geradas as vistas intermediárias, elas serão analisadas e comparadas com as vistas intermediárias geradas a partir das vistas originais não codificadas. É importante ressaltar que as imagens de referência não serão as vistas intermediárias originais, uma vez que se tem o intuito de analisar o desempenho dos codificadores, e obviamente do método proposto e não dos softwares de síntese de vistas virtuais.

Este artigo é organizado da seguinte forma. Na Seção II uma breve introdução das tecnologias de *displays* auto-estereoscópicos é realizada. A Seção III descreve o método proposto, detalhando os passos a serem realizados, além de apresentar os processos de síntese da vista complementar (vista-alvo) e das vistas intermediárias. Os resultados provenientes da codificação das vistas, dos mapas de disparidades e dos mapas compostos com os codificadores, H.264/AVC Intra e Estéreo, e JPEG-2000 são apresentados na Seção IV, bem como os resultados das vistas sintetizadas a partir do conjunto de imagens codificadas. As conclusões são realizadas na Seção V.

II. *DISPLAYS* AUTO-ESTEREOSCÓPICOS

Existe uma grande variedade de sistemas de *displays* 3D. Sendo que os mais comuns são os sistemas clássicos com duas vistas (estéreo) e óculos. Os sistemas ditos auto-estereoscópicos (sem a necessidade de uso de óculos) têm usado mais do que duas vistas (multi-vistas) para levar a sensação tridimensional ao observador. Apesar de que os *displays* exibem (emitem) mais do que uma vista ao mesmo tempo, a tecnologia garante que o espectador veja apenas um par estéreo[3].

O Sistema Visual Humano (SVH) utiliza várias informações perceptuais para perceber o mundo em 3D. Tais informações também estão presentes no mundo 2D, incluindo as oclusões, a perspectiva (ponto de vista), o tamanho aproximado dos objetos (conhecemos os tamanhos de muitos objetos do mundo real), e a perspectiva atmosférica (objetos mais distantes parecem ter cores mais pastéis [3].

Os *displays* autostereoscópicos produzem imagens 3D sem a necessidade de uso de óculos para o espectador. Os *displays* 3D, que requerem o uso de óculos para os usuários, apresentam duas imagens diferentes no mesmo plano da tela. De alguma forma, os óculos selecionam qual imagem vai para o olho esquerdo e qual vai para o direito. Os sistemas de *displays* 3D auto-estereoscópicos, baseados em multi-vistas, combinam os efeitos de paralaxe estéreo e de paralaxe de movimento para prover imagens 3D sem a necessidade de óculos [3]. Quando o espectador olha para a tela ele vê uma imagem diferente com cada olho, e se ele move a cabeça ele vê imagens diferentes. Pode-se limitar o número de imagens (vistas), criando-se compartimentos 'virtuais', ou setores. Assim, cada imagem somente é visível dentro do seu compartimento. Ainda existe o paralaxe estéreo, bem como o paralaxe de movimento. Mas, desta vez apresenta saltos quando o observador vai de um compartimento (ou setor) ao outro. Desta forma, um número inferior de vistas vai fornecer as informações de paralaxe estéreo, e de movimento. Um display autostereoscópico 3D apresenta uma imagem diferente para cada setor, produzindo a sensação de paralaxe estéreo, e de movimento a partir de um pequeno conjunto de vistas.

Mas, a necessidade de se exibir tantas vistas da mesma cena demanda uma alta banda-passante, já que o sistema tem que codificar e transmitir um número de vistas bem maior do que um sistema estéreo. Uma solução é manter o sistema estéreo, e usar o formato vídeo acrescido da informação de profundidade (video+depth format) proposto. O MPEG lançou um padrão denominado MPEG-C Part 3 [4] que comprime e transmite vídeo monocular e dados auxiliares (disparidade/profundidade). Com este formato, o receptor pode reproduzir o vídeo, ou imagens estéreo, com o auxílio de um algoritmo de síntese de vistas (comumente chamado de Depth Image-Based Rendering-DIBR) que vai criar a segunda vista do par estéreo, e posteriormente, as vistas intermediárias (entre as duas posições originais). O grande problema do uso de algoritmos de síntese de vista virtual é garantir que qualidade da vista virtual seja igual, ou similar, a da vista original daquele mesmo ponto de vista.

III. SISTEMA HÍBRIDO DE COMPRESSÃO

Hoje, existem muitas alternativas para compressão de imagens digitais a serem escolhidas. Entre elas têm-se em evidência os codificadores JPEG-2000 para imagens estáticas e H.264/MPEG-4 AVC para imagens estáticas e sequências de imagens. Ambos os padrões foram selecionados para comprimir alguns tipos de imagens do sistema de compressão proposto. Recentemente, o padrão H.264/AVC recebeu uma extensão para sequências multi-vistas e estéreo denominada Multiview Video Coding-MVC. Tal extensão apresenta um desempenho superior à sua versão monocular, e também é empregado neste artigo.

Os mapas de disparidade apresentam características diferentes da grande maioria das imagens Uma vez que os mapas representam a estrutura 3D de uma cena, as áreas dos mapas que correspondem ao mesmo objeto apresentam valores de disparidades muito próximos. Assim, a maioria dos mapas de disparidade tende a apresentar áreas com variação suave de tons de cinza E, sem esquecer de que as bordas dos objetos, que tendem a indicar mudanças de profundidade, são bem marcadas e, muitas vezes, acentuadas. A codificação livre de artefatos destas bordas é muito importante para a síntese das vistas virtuais.

A. *Trabalhos relacionados*

Recentemente, vários métodos foram propostos para comprimir mapas de disparidade/profundidade quando representados em tons de cinza.

Em [5] os mapas são codificados usando o algoritmo Multidimensional Multiscale Parser (MMP). Tal método preserva as bordas dos mapas, onde – em geral – ocorrem as mudanças de profundidade, permitindo que algoritmos de geração de vistas virtuais criem as imagens intermediárias com muita precisão. O MMP é um algoritmo baseado em blocos que combina predição com casamento aproximado de padrões recorrentes em múltiplas escalas. O algoritmo tenta representar um bloco de dados de entrada usando versões dilatadas ou contraídas de blocos anteriores. Como o MMP não se baseia em qualquer transformação para o domínio da frequência, ele não fica sujeito aos efeitos danosos de quantização nas regiões de alta freqüência como os baseados na DCT, por exemplo.

Em [6] é apresentado um algoritmo que emprega uma decomposição em quad-tree da imagem. Tal método também preserva as bordas dos mapas, embora não apresente um bom desempenho em termos de taxa-distorção quando comparado com o padrão H.264/AVC.

Este trabalho limita-se a usar os padrões H.264/AVC e JPEG-2000 para a compressão das imagens em tons de cinza, combinados com o uso do padrão JBIG2 para a codificação dos mapas de oclusão. A contribuição desta pesquisa está na geração de um mapa auxiliar, mapa composto, que substitui uma das vistas do par estéreo, e apresenta um compromisso em termos de taxa-distorção superior ao da vista original.

Um método tradicional de compressão, codifica ambas as vistas e seus respectivos mapas de disparidade para possibilitar a reconstrução de vistas intermediárias no lado

do decodificador. No entanto, tal método consome uma taxa de bits elevada.

B. *Mapa composto*

Os mapas compostos são gerados a partir de uma combinação de informações obtidas de textura das vistas esquerda e direita e do mapa de oclusões (parte superior da Fig 1). O mapa composto (parte inferior da Fig 1) consiste da informação de textura da imagem correspondente à vista complementar (esquerda ou direita) nas áreas de oclusões (representadas no tom preto nos mapas de oclusões) como mostra a Fig 1. Os mapas de oclusões são gerados com precisão de 1 pixel. Ou seja, o erro permitido no mapeamento dos pontos homólogos de acordo com os valores de disparidade do mapa da esquerda para o mapa da direita é de 1 pixel.

Foram realizados alguns testes com alguns tipos de mapas compostos. Primeiro foi feita a substituição das áreas oclusas pela informação de textura, assinalando o tom branco à área não oclusa. Analisando os resultados, foi verificado que os mapas codificados estavam com taxas de bits muito elevadas. Estes resultados são devidos ao grande salto entre as informações de textura e a área no tom branco, que representa as regiões onde não ocorreram oclusões. Desta forma, havia uma variação grande e brusca em que o degrau de quantização usado pelos codificadores também era grande, gastando-se mais bits na sua codificação. A substituição das áreas brancas pelo valor médio dos tons existentes no mapa, diminuiu significantemente o salto entre as áreas com texturas e as áreas onde não ocorreram oclusões. O novo mapa composto é representado na Fig 2 (imagem na parte inferior esquerda).

O mapa composto apresenta pouca textura e muitas áreas homogêneas, sendo assim representadas por uma quantidade de *bits* muito menor do que a necessária para representar a vista de referência codificada (esquerda ou direita).

C. *Sistema híbrido de compressão estéreo*

A Fig 2 exibe os passos realizados pelo sistema híbrido proposto. São codificados uma das vistas do par estéreo, o seu respectivo mapa de disparidades, o mapa composto, e o mapa de oclusões correspondente. A vista complementar (esquerda ou direita) é reconstruída e, a seguir, a vista intermediária pode ser gerada. A vista, o mapa de disparidades, e o mapa composto são codificados usando o padrão H.264/AVC ou o padrão JPEG-2000. O mapa de oclusões é sempre codificado usando o padrão JBIG2 [7].

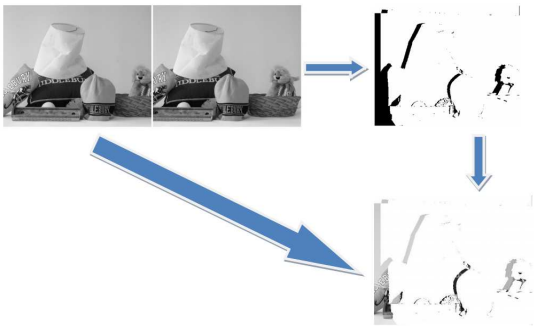


Fig 1 – Geração do Mapa Composto

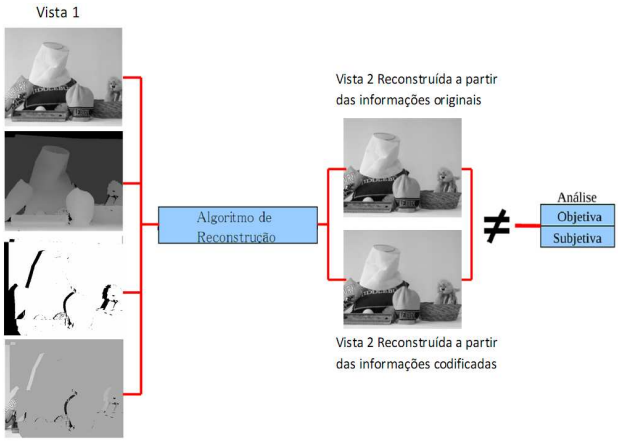


Fig 2 Sistema Híbrido de Compressão Estéreo

O método proposto utiliza para reconstrução da vista complementar (e.g. direita) do par estéreo, a vista base (referência, e.g. a vista esquerda), o mapa de oclusões, o mapa de disparidades e o mapa composto. Com isso, busca-se uma relação de compromisso que forneça uma boa qualidade para a imagem reconstruída, considerando-se os métodos de avaliação utilizados, e que forneça uma taxa de bits para transmissão menor do que se transmitir as duas vistas, e as imagens necessárias (i.e. mapas de disparidade) para a síntese das vistas virtuais.

Para isso, temos que atender a seguinte condição para as taxas de bits de cada uma das imagens e mapas:

$$bpp_{Esq} + bpp_{Disp} + bpp_{Comp} + bpp_{Ocl} < bpp_{Esq} + bpp_{Dir}$$

Para as gerações de vistas intermediárias foram usadas informações retiradas de uma das vistas do par estéreo, do mapa composto e dos dois mapas de disparidade. O resultado desse método foi comparado com a geração das vistas virtuais a partir de informações retiradas das duas vistas do par estéreo e dos dois mapas de disparidade. Os mapas de oclusões não são necessários, pois podem obtidos a partir dos mapas de disparidades. Portanto, para que o método proposto apresente o resultado esperado é necessário que atenda a seguinte condição para as taxas de bits de cada uma das imagens e mapas:

$$bpp_{Esq} + bpp_{DispEsq} + bpp_{DispDir} + bpp_{Comp} < bpp_{Esq} + bpp_{Dir} + bpp_{DispEsq} + bpp_{DispDir}$$

O algoritmo empregado na síntese de vistas é uma versão simplificada do apresentado em [8].

IV RESULTADOS

Os resultados são apresentados em função das métricas de avaliação objetiva, PSNR e SSIM, bem como em função das avaliações subjetivas. Estes resultados são referentes às codificações realizadas com os codificadores H264/AVC perfis Intra e Estéreo, e com o JPEG-2000 para as vistas originais do par estéreo, mapas de disparidades e também para os mapas compostos gerados. O mapa de oclusões é comprimido eficientemente pelo codificador JBIG2, uma vez que as informações de oclusões podem ser representadas com dois níveis apenas (preto e branco, por exemplo).

Os oito pares estéreo utilizados nas simulações, adquiridos do banco de imagens de Middlebury (http://vision.middlebury.edu) ALOE, ART, BOWLING1, BOWLING2, CONES, MIDD1, REINDEER e TEDDY foram selecionados de modo a apresentarem diferentes níveis de disparidades, diferentes tipos de estruturas geométricas (simples e complexas) e diferentes níveis de oclusões, indo de pequenas a grandes áreas oclusas.

As imagens são convertidas para o formato YUV 4:0:0 para serem processadas pelo codificador H.264/AVC, perfil High INTRA e STEREO, nível 4.0.

A definição das taxas de bits alvo foi baseada na geração de imagens codificadas a diferentes taxas de bits. Os resultados foram subjetivamente analisados, e então foi realizada a seleção das taxas de bits mínima e máxima para cada tipo (vista, mapa de disparidades e mapa composto) de imagem, variando de 0.5 bpp a 1.5 bpp. Para alguns casos foram utilizadas taxas variando de 0.2 bpp a 2.5 bpp. Estas taxas são definidas diretamente no software JPEG-2000 (http://www.kakadusoftware.com). Já para o H.264/AVC, usando-se o software de referência na versão JM 17.1, deve-se variar os valores do parâmetro QP, passo de quantização, de modo a atingir o valor mais próximo aos estipulados. Devido às diferentes características de cada tipo de imagem, os passos de quantização diferem entre elas para alcançar uma mesma taxa, de mesma forma ocorre com os mapas de disparidades e mapas compostos.

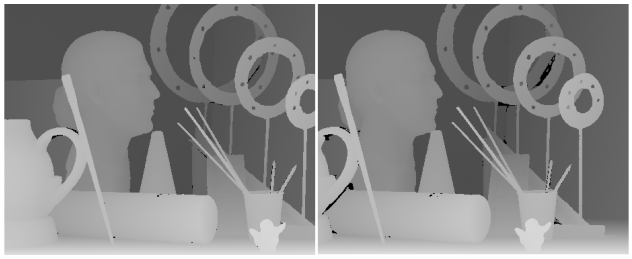
Resultados para o par estéreo ART são mostrados nas Fig 3 a 7. A Fig 3 exibe o par estéreo ART, a Fig 4 os respectivos mapas de disparidade, e a Fig 5 os mapas compostos.

O desempenho dos perfis do H.264/AVC High STEREO, e INTRA, e do padrão JPEG-2000 são exibidos nas Figs 6 e 7. Foram codificados os pares estéreo e os mapas de disparidade. Assim, pode ser avaliada a diferença de comportamento entre os perfis INTRA e STEREO, sendo que este último realiza além da predição espacial a predição inter-vistas.

No caso dos pares estéreo e mapas de disparidade, o desempenho para a vista base (e.g. vista, ou mapa da esquerda) é idêntico para ambos os perfis do H.264/AVC. Entretanto, o perfil STEREO é bastante superior em várias taxas quando codifica a vista, ou mapa complementar. O padrão JPEG-2000 apresenta desempenho inferior em todos os casos.

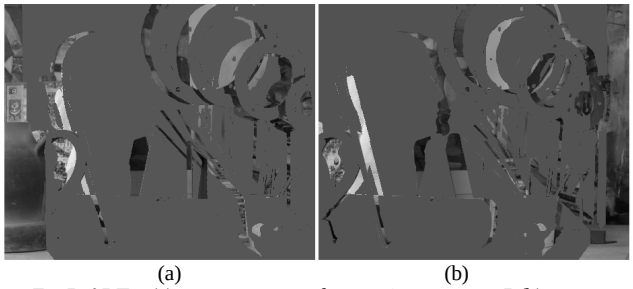
(a) (b)

Fig 3 ART, vista 1 (a) e vista 5 (b) do par estéreo



(a) (b)

Fig 4 ART– (a) mapa de disparidade da vista 1 para a vista 5 (b) mapa de disparidade da vista 5 para a vista 1



(a) (b)

Fig 5 ART – (a) mapa composto da vista 1 para a vista 5 (b) mapa composto da vista 5 para vista 1 do par estéreo com limiar 1.

A Tab 1 exibe os resultados das combinações de taxas de codificação da vista base, do mapa de disparidades, do mapa de oclusão e do mapa composto para a geração da vista complementar. Tais resultados foram selecionados por apresentarem qualidade subjetiva satisfatória, e por estarem abaixo dos limites máximos de taxas de bits para que o método proposto apresente uma taxa total menor do que a taxa total necessária para que a codificação usando modo convencional apresentasse a mesma qualidade subjetiva.

Para avaliar a eficiência do método, vistas intermediárias geradas a partir do par estéreo original e seus respectivos mapas de disparidade, na posição central da *baseline* [8] foram comparadas às vistas centrais sintetizadas a partir das vistas-base e complementares, mapas de disparidades, mapas de oclusões e mapas compostos codificados a diferentes taxas de bits.

A Fig 8(a) mostra a vista sintetizada pelo método proposto na posição central da *baseline* do par estéreo ART a partir das informações não codificadas. As Figs 8(b) e 8(c) apresentam a mesma vista gerada a partir das informações codificadas nos perfis INTRA e STEREO respectivamente. As taxas de bits empregadas foras as apresentadas na combinação 1 da TAB 1. A análise subjetiva entre as imagens mostra que as diferenças existentes não são perceptíveis. Nenhum observador foi capaz de identificar diferenças entre as imagens.

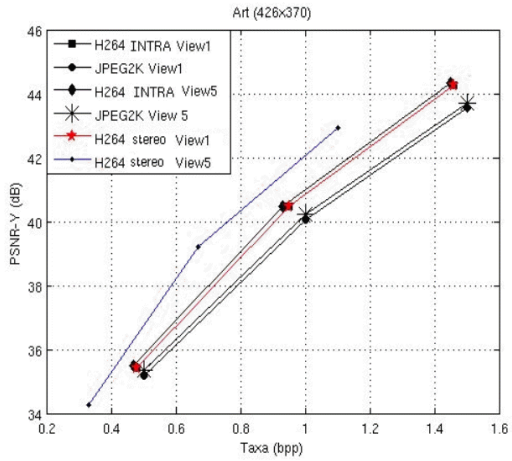


Fig 6 ART –Vistas 1 e 5 - Curvas taxa-distorção com os codificadores H.264/AVC perfis INTRA, STEREO e JPEG-2000.

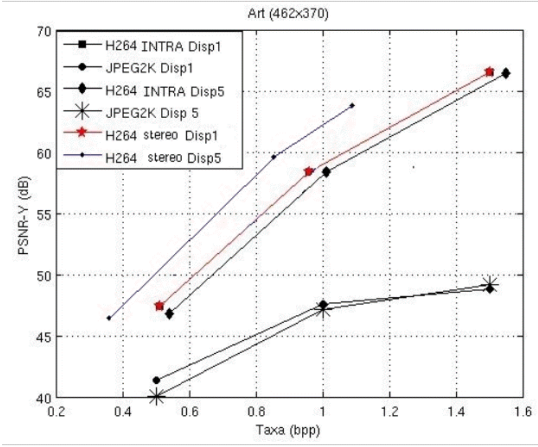


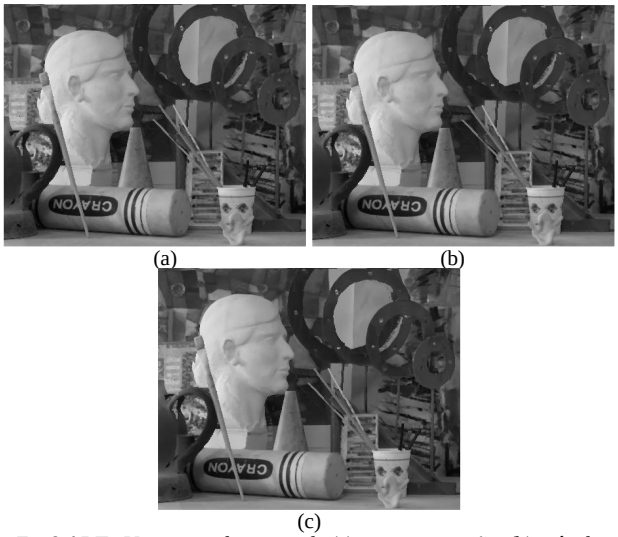
Fig 7 ART – Mapas de disparidades das vistas 1 e 5 - Curvas taxa-distorção com os codificadores H.264/AVC perfis INTRA, STEREO e JPEG-2000.

TAB 1 – Combinações de taxas de bits

Combinação	Taxa (bpp) Vista	Taxa (bpp) Mapa de Disparidades	Taxa (bpp) Mapa Composto
1	1.0	1.5	0.50
2	1.0	1.5	0.25
3	1.0	1.5	0.10

V CONCLUSÕES

As simulações corroboram com os resultados existentes na literatura em relação ao desempenho em termos de taxa-distorção dos codificadores apresentados para a compressão das vistas. Além disto, várias simulações foram realizadas para diferentes tipos de mapas de disparidade, fornecendo uma extensa análise objetiva e subjetiva dos mapas comprimidos. O codificador H.264/AVC perfil High STEREO foi utilizado na compressão de um conjunto de mapas de disparidades com diferentes características. As análises subjetivas indicaram as taxas de bits com limites superiores e inferiores para cada tipo de imagem (vistas, mapas de disparidade e compostos), permitindo a uma combinação ótima de conjuntos de taxas para cada tipo de imagem.



(a) (b)

(c)

Fig 8 ART –Vista central sintetizada (a) sem compressão; (b) método convencional; (c) método proposto.

O método proposto apresentou resultados interessantes do ponto de vista da análise perceptual, bem como do ponto de vista das métricas objetivas. O uso de um mapa composto cuja versão codificada consome menos bits do que a vista original, sem causar artefatos visualmente perceptíveis é muito promissora. As avaliações subjetivas indicam que os mapas compostos podem ser comprimidos com baixas taxas de bits, sem comprometer em termos perceptuais as vistas sintetizadas. Um resultado a ser ressaltado, é que as simulações realizadas indicam taxas mínimas desejáveis para que a compressão dos mapas de disparidade não degrade as vistas sintetizadas. Ou seja, esta afirmação reforça a importância da manutenção da informação contida no mapa de disparidades para a síntese de vistas virtuais.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro do programa CAPES/Pró-Defesa.

REFERÊNCIAS

[1] ITU-T,ISO/IEC JTC1, "Advanced video coding for generic audiovisual services," ITU-T Recommendation H.264/AVC and ISO/IEC 14496-10 (MPEG4-AVC), Version 11: 2009JPEG-2000 ISO/IEC 15444-1:2004

[2] VETRO A., YEA S., SMOLIC A. 3D Video Format for Auto-Stereoscopic Displays TR2008-057 September 2008

[3] VETRO A., Technical Report-TR2010-011, MERL, Representation and Coding Formats for Stereo and Multiview Vide, April 2010.

[4] MPEG-C, 2007 (ISO/IEC23002-3,2007)

[5] GRAZIOSI et alli, Compressing depth maps using multiscale recurrent pattern image coding, Electronics Letters, vol. 46, no. 5, pp. 340–341, 2010

[6] MERKLE P., et alli, The effects of multiview depth video compression on multiview rendering, Signal Processing: Image Communication, vol. 24, no. 1-2, pp. 73–88, 2009 "Progressive bi-level image compression," Int. Std. ISO/IEC11544, 1993

[7] JBIG2, "Progressive bi-level image compression," Int. Std. ISO/IEC11544, 1993

[8] PEREZ, M e PAGLIARI, C, Multi-Viewpoint Synthesis from Uncalibrated Stereo Cameras. ICIP (1) 2007: 221-224 2007