Super-Resolução para Síntese de Vistas em Ambientes de Resolução Mista

Camilo Dórea, Diogo C. Garcia e Ricardo L. de Queiroz

***Resumo*— O vídeo estereoscópico de resolução mista constitui um formato comumente empregado no consumo de conteúdo 3D. No entanto, as diferenças de resolução entre as vistas podem representar uma desvantagem significativa ao adaptar esse formato às outras funções baseadas no uso de múltiplas vistas, tal como a síntese de vistas intermediárias. Nesse artigo adaptamos a metodologia padrão empregada na síntese de vistas tradicional para as particularidades das condições apresentadas pelo formato de resolução mista. Utilizando a representação de múltiplas vistas, em resolução mista, com mapas de profundidade, propomos um método de super-resolução capaz de realçar as vistas de baixa resolução com conteúdo de alta frequencia proveniente das vistas adjacentes de resolução íntegra. Resultados experimentais indicam que ambas as estratégias, a adaptação da síntese e a super-resolução da vista de baixa resolução, melhoram a qualidade subjetiva e objetiva da vista sintetizada.**

***Palavras-Chave*— Resolução mista, síntese de vista, super-resolução.**

***Abstract*— Mixed resolution stereoscopic video is a popular format aimed at 3D content consumption. Nevertheless, resolution differences among views can pose significant drawbacks if adapting such a format to other multiview functionalities, such as view synthesis. In this paper we adjust standard view synthesis methodology to mixed resolution conditions. Operating within a mixed resolution, multi-view-plus-depth scenario, we then propose an improved super resolution method capable of enhancing lower-resolution views with high-frequency content available from neighboring full-resolution views. Results on publicly available data sets show that both strategies, adaptation of synthesis to asymmetry and direct enhancement of lower-resolution input views, can improve subjective and objective quality of synthesized views.**

***Keywords*— Mixed resolution, view synthesis, super resolution.**

I. Introdução

A síntese de vistas é amplamente empregada na renderização de vistas virtuais para aplicações como *free-viewpoint television* [1] e vídeo 3D. Para melhorar a experiência do usuário, tais funcionalidades empregam múltiplas vistas da mesma cena. Um formato de representação popular que considera tanto a eficiência em termos de volume de dados e as necessidades de renderização das diversas aplicações é a representação com múltiplas vistas e mapas de profundidade (sigla MVD em inglês) [2], [3]. Essa representação consiste de múltiplas vistas da mesma cena provenientes de câmeras distintas e os mapas de profundidade correspondentes contendo a distância entre o ponto na cena e a câmera de referência. Com o uso de mapas de profundidade e parâmetros de calibração das câmeras, vistas intermediárias virtuais podem ser renderizadas em posições arbitrárias entre câmeras. De tal forma, dado um conjunto fixo de câmeras e vistas, a síntese de novas vistas permite ao usuário navegar de maneira contínua entre as vistas e melhorar as sensações de percepção de profundidade em vídeo 3D.

A eficiência de compressão de dados é também a principal preocupação de outro formato 3D comumente utilizado, o vídeo estereoscópico assimétrico [4], [5]. A assimetria refere-se às diferenças de resolução e/ou qualidade entre as vistas compondo o par estéreo. Nesses ambientes, também denominados de resolução mista, a vista de resolução íntegra é apresentada ao olho esquerdo enquanto a de resolução reduzida é reservada para o olho direito, por exemplo. Os ganhos de compressão são justificados por estudos psico-visuais [6], [7] que sugerem que a qualidade da visão estéreo não é negativamente afetada quando apenas uma das vistas estiver borrada.

O vídeo assimétrico não foi, no entanto, intencionado para algumas das aplicações e funcionalidades permitidas pelo formato MVD. Por exemplo, a navegação entre vistas pode sofrer severas degradações em qualidade caso as vistas empregadas sejam de diferentes resoluções. Para superar essas limitações à experiência do usuário e preservar tanto as vantagens de compressão quanto a compatibilidade com vídeo assimétrico, o formato de múltiplas vistas em resolução mista com mapas de profundidade (MVD-RM) além de uma técnica de super-resolução (SR) foram propostos [8]. Esse formato é ilustrado na Figura 1 e consiste em múltiplas sequências de vídeo de diferentes pontos de vista com resoluções alternadas e os mapas de profundidades correspondentes. Assumimos que as vistas de resolução reduzida são obtidas por sub-amostragens e que os mapas de profundidade estão disponíveis e que são todos de resolução íntegra. Tais mapas podem ser codificados de maneira eficiente, representando uma porcentagem pequena do volume total de dados [9].

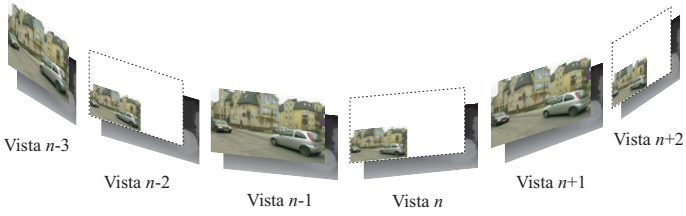

Fig. 1. O formato de múltiplas vistas em resolução mista com mapas de profundidade.

Camilo Dórea e Ricardo L. de Queiroz, Departamento de Ciência da Computação, Diogo C. Garcia, Faculdade de Engenharia do Gama, Universidade de Brasília, Brasília-DF, Brasil. E-mails: camilo@cic.unb.br, diogo@image.unb.br, queiroz@ieee.org. Este trabalho foi parcialmente financiado pelo CNPq (310375/2011-8 e 470940/2010-7).

Nesse artigo investigamos o problema de síntese de vistas sob condições de assimetria em resolução tais como apresentadas pelo formato MVD-RM. Adaptamos a metodologia padrão de síntese às condições de assimetria para melhorar a qualidade objetiva da vista virtual intermediária. Apresentamos também uma técnica de SR melhorada capaz de realçar a vista de baixa resolução com conteúdo de alta frequencia proveniente da vista adjacente de resolução íntegra. Através da técnica de SR, a recuperação de resolução de vistas em MVD-MR incrementa de maneira significativa a qualidade de síntese de vistas em ambientes de resolução mista.

A técnica de SR apresentada é similar à metodologia de SR baseada em exemplos [10] onde imagens de alta resolução são usadas para restaurar regiões que sofreram dizimação. Nessa proposta, nos baseamos em trabalho prévio [8], [11] e introduzimos uma técnica melhorada de projeção entre vistas para SR. Essa técnica, ademais, conta também com um teste de consistência baseado em profundidades superior ao apresentado em [8] ou [11]. A maioria dos trabalhos na literatura considera o uso de mútiplas vistas e a síntese de vistas em ambientes de resolução uniforme [12].

A seguir, apresentamos na seção II a síntese assimétrica de vistas para formatos MVD-RM. Na seção III, discutimos a técnica melhorada de SR e o seu uso em síntese. Ambas, técnica de síntese assimétrica e técnica de SR, podem ser empregadas conjuntamente na renderização de vistas virtuais intermediárias. Resultados experimentais são tratados na seção IV e conclusões apresentadas na seção V.

II. Síntese assimétrica de vistas

A síntese de vista é uma técnica de renderização baseada em imagens onde as vistas V_k e V_n, correspondentes às k-ésima and n-ésima câmeras, são usadas para criar uma vista virtual na posição alvo V_m. Técnicas de projeção são responsáveis por mapear cada píxel da vista de referência na vista alvo utilizando informação de geometria de cena.

Considere a localização (u, v) de um píxel da câmera de referência k juntamente com parâmetros intrínsicos de câmera $\mathbf{A_k}$, matriz de rotação $\mathbf{R_k}$, vector de translação $\mathbf{t_k}$ e o mapa de profundidade correspondente D_k. O ponto (u, v) poderá ser projetado ao espaço de coordenadas 3D globais (x, y, z) e re-projetado à posição (u', v') da vista adjacente m através de [9]:

$$(x, y, z)^T = \mathbf{R_k}{\mathbf{A_k}}^{-1}(u, v, 1)^T D_k(u, v) + \mathbf{t_k} \quad (1)$$

$$(u' * w', v' * w', w')^T = \mathbf{A_m}{\mathbf{R_m}}^{-1}[(x, y, z)^T - \mathbf{t_m}]. \quad (2)$$

Uma vez obtidas as posições correspondentes na imagem alvo, a projeção dos componentes de cor das vistas referências formarão as estimativas $V_{m|k}$ and $V_{m|n}$ da vista alvo m. Note que as coordenadas (u', v') possuem precisão sub-pixel. Técnicas de *splatting* são geralmente empregadas para tratar do arredondamento em projeções de valor fracionário e podem resultar em borramento da imagem resultante. A projeção deve também contemplar casos de colisões. Neste caso, a preferência é dada ao pixel do plano mais próximo à câmera. As estimativas da vista alvo podem ademais conter buracos causados pelo processo de projeção ou por oclusões. Esses são tratados no processo de mesclagem (*bleding*) para a formação da síntese final:

$$V_m = \alpha V_{m|k} + (1 - \alpha)V_{m|n}. \quad (3)$$

O peso α é geralmente escolhido para enfatizar as estimativas provenientes da câmera que estiver fisicamente mais próxima à m. Nesses casos, o peso é dado por $dist(m, n)/[dist(m, k) + dist(m, n)]$ onde $dist(\cdot, \cdot)$ refere-se à distância entre câmeras.

No formato MVD-RM, uma vista de baixa resolução V_n^D é obtida pela sub-amostragem, por um fator M, da vista orginal V_n. A super-amostragem (interpolação) de V_n^D pelo mesmo fator gera uma versão de baixas frequencias da imagem denominada V_n^L. Essa versão de baixa frequência V_n^L e a vista adjacente de resolução original (íntegra) V_k, serão usadas para sintetizar uma vista virtual localizada, por exemplo, à distância equidistante entre ambas vistas referências. Em tais condições de assimetria, o peso α deverá ser adaptado para enfatizar a vista de maior resolução e não necessariamente aquela de maior proximidade. Essa adatação considera a assimetria e pode ser integrada facilmente aos algoritmos padrões de síntese de vistas [13]. Na seguinte seção, tratamos do realce de V_n^L como passo prévio a uma síntese assimétrica.

III. Super-Resolução para Múltiplas Vistas

Dada uma configuração MVD-RM, o método proposto de SR obtém os detalhes de alta frequência, ausentes na vista de baixa resolução, das vistas adjacentes de resolução íntegra. O método assume que qualquer vista de resolução íntegra pode ser decomposta em componentes de baixa e alta resolução: $V_n = V_n^H + V_n^L$. Para obter a estimativa super-resolvida $\hat{V}_n$ de uma vista de baixa resolução V_n^L, devemos encontrar as estimativas de alta frequencia $\hat{V}_n^H$ que, somadas aos valores de V_n^L, formam a imagem super-resolvida $\hat{V}_n = \hat{V}_n^H + V_n^L$. A estimativa de $\hat{V}_n^H$ consiste em três passos principais:

1) a vista adjacente de resolução íntegra V_k é projetada em direção a n-ésima vista, com o auxílio dos mapas de profundidade D_k and D_n, formando $V_{n|k}$,
2) $V_{n|k}$ é sub-amostrado e logo super-amostrado para formar a versão passa baixas $V_{n|k}^L$ e
3) $\hat{V}_n^H$ é obtido através da subtração $\hat{V}_n^H = V_{n|k} - V_{n|k}^L$.

O primeiro passo do método é a projeção da vista k (de resolução íntegra) em direção à vista n (de baixa resolução) por meio das equações (1) and (2). Note que essa projeção pode ser realizada à partir da vista n (*backwards warping*), ou seja, um píxel da vista n é preenchido com os valores encontrados na vista k. A existência de ambos mapas de profundidade D_k e D_n permite-nos formular um teste de consistência e assim evitar projeções errôneas seja pela existência de oclusões ou pela imprecisão nos valores de profundidade.

Para completar o passo de projeção, introduzimos um teste de consistência baseado em profundidades. Para cada píxel (u, v) da vista n e píxel correspondente (u', v') da vista k, determinamos as posições inteiras mais próximas às coordenadas

(u', v'): $p_1 = (\lfloor u' \rfloor, \lfloor v' \rfloor)$, $p_2 = (\lceil u' \rceil, \lfloor v' \rfloor)$, $p_3 = (\lfloor u' \rfloor, \lceil v' \rceil)$ e $p_4 = (\lceil u' \rceil, \lceil v' \rceil)$, onde $\lfloor . \rfloor$ e $\lceil . \rceil$ são os operadores *floor* e *ceil*, respectivamente. (Vide ilustração na Figura 2.) Entre os quatro pontos determinados, definimos (u'_r, v'_r) como sendo o mais próximo a (u', v'). (u'_r, v'_r) é projetado de volta à vista n, resultando em coordenadas (u'', v''). Se (u'', v'') pertence ao círculo de raio de 1 píxel ao redor de (u, v), a correspondência é considerada válida e $V_{n|k}(u, v)$ é calculado por interpolação bilinear entre os valores das posições $\{p_1, p_2, p_3, p_4\}$. Se (u'', v'') estiver localizado além do círculo, a projeção não é realizada e $V_{n|k}(u, v) = V_n^L(u, v)$. Ou seja, buracos remanescentes dentro da imagem projetada, detectados através do teste de consistência, são preenchidos com conteúdo de baixa frequência. Uma vez determinado $V_{n|k}(u, v)$, a imagem $\hat{V}_n^H$ é obtida pelos passos 2 e 3 descritos acima e usada na determinação da estimativa super-resolvida $\hat{V}_n$.

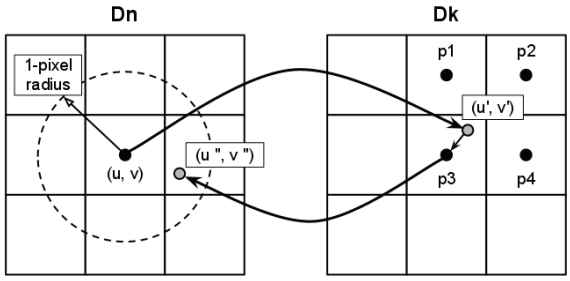

Fig. 2. Teste de consistência baseado em profundidades ilustrando uma correspondência válida entre píxeis (u, v) da vista n e (u', v') da vista k. $(u'_r, v'_r) = p_3$ que é projetado de volta à (u'', v'').

IV. Resultados Experimentais

O método proposto foi testado com três sequências MVD, disponíveis publicamente: *Poznan Street* [14], *Pantomime* [15] e *Cafe* [16]. Para cada sequência, os primeiros quadros de três câmeras com pontos de vistas adjacentes foram selecionados. As duas vistas externas foram usadas como referências para síntese e a vista intermediária para estabelecer comparações (*groundtruth*) conforme resumido na Tabela I. A síntese de vista foi realizada com o software de referência VSRS [13], versão 3.5. Um filtro de interpolação Lanczos 6-tap, de alto desempenho, foi empregado na sub- e super-amostragem com fatores $M = 2$ and $M = 4$ para formar a imagem de baixa resolução a partir de uma das imagens de referência conforme descrito na seção II. Note que filtros alternativos podem ser empregados para SR conforme apresentado em [11]. Devido a quantidade limitada de conteúdo de alta frequência nas imagens originais, as sequências foram redimensionadas conforme indicado na Tabela I.

A Tabela II apresenta valores de PSNR para a síntese em três configurações distintas. O primeiro teste considera ambas vistas de referência em resolução original (íntegra). Os outros dois testes consideram a primeira vista referência em resolução íntegra e a segunda vista como sendo de baixa resolução obtida por interpolação (V_n^L) ou super-resolvida pelo algoritmo proposto ($\hat{V}_n^H$). Os resultados indicam, como esperado, que a qualidade de síntese é superior quando ambas

TABELA I

Imagens (vistas) de referência e groundtruth usadas nos experimentos de síntese de vista.

Sequência	1ª ref.	2ª ref.	Ground-truth
Cafe, 960 × 536	2	4	3
Pantomime, 640 × 480	37	39	38
Poznan Street, 960 × 544	3	5	4

vistas de referência estão em resolução íntegra. Considerando um ambiente MVD-RM, o método proposto de SR quando aplicado à segunda imagem referência pode produzir ganhos de 0.15 dB (*Cafe*, $M = 2$) à 3.99 dB (*Pantomime*, $M = 4$) sobre a interpolação da segunda imagem referência.

TABELA II

Resultados de PSNR (luma) para síntese de vistas usando a primeira imagem referência em resolução íntegra e a segunda imagem referência como: íntegra, interpolada ou super-resolvida.

Sequência	Íntegra	Interp.	SR
M = 2			
Cafe	34.46 dB	34.15 dB	34.30 dB
Pantomime	38.65 dB	37.10 dB	38.24 dB
Poznan Street	35.58 dB	33.97 dB	34.58 dB
M = 4			
Cafe	34.46 dB	32.54 dB	33.02 dB
Pantomime	38.65 dB	33.13 dB	37.12 dB
Poznan Street	35.58 dB	31.38 dB	33.00 dB

A avaliação subjetiva do método proposto de SR para síntese assimétrica pode ser realizada com as imagens da Figura 3. Detalhes da imagem sintetizada para a sequência *Pantomime*, $M = 4$, sob as condições descritas na Tabela II são apresentadas. Melhoras significativas em aguçamento são observadas entre as imagens provenientes de (a) interpolação e (b) SR. O resultado de SR é comparável em qualidade à síntese proveniente de imagens de resolução íntegra (c).

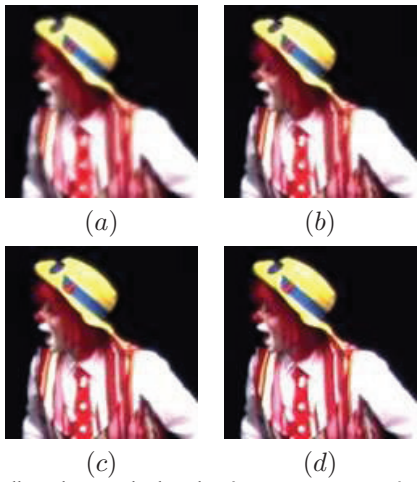

Fig. 3. Detalhes dos resultados de síntese para a sequência *Pantomime*, $M = 4$, usando a segunda imagem referência (a) em baixa resolução seguido de interpolação, (b) em baixa resolução seguido de SR, (c) em resolução íntegra e (d) *groundtruth*.

O segundo conjunto de experimentos averigua o efeito da

variação do peso α, vide equação (3), sobre a síntese assimétrica de vistas. A Figura 4 apresenta valores de PSNR para a vista sintetizada da sequência *Poznan Street*, considerando novamente a primeira imagem referência em resolução íntegra e duas condições para a segunda imagem referência: baixa resolução seguida de interpolação e baixa resolução seguida de SR. Dois fatores de sub-amostragem são empregados: $M = 2$ e $M = 4$. A variação de α de 1 a 0, progressivamente deslocando o peso da imagem referência de resolução íntegra à imagem referência interpolada ou SR, resulta na diminuição de valores de PSNR em ambas curvas. Para $M = 2$, o valor máximo de PSNR é atingido em $\alpha = 0.5$ para SR e $\alpha = 0.6$ para interpolação. Para fatores maiores de sub-amostragem, $M = 4$, o valor máximo de PSNR é atingido para valores α em torno de 0.8 ou 0.9, porém não excede o valor de PSNR atingido com $\alpha = 1.0$ de maneira significativa. Independentemente do valor de α eleito, o método proposto de SR tem desempenho superior à interpolação em todos os testes. Os resultados apresentados para a sequência *Poznan Street* são representativos de outras sequências testadas.

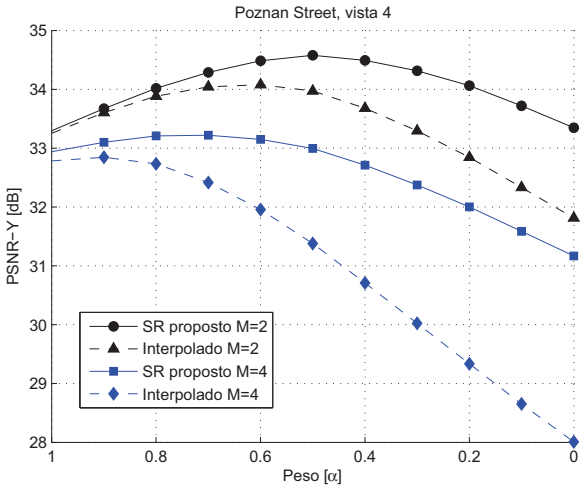

Fig. 4. Efeito da variação de peso α sobre síntese de vista usando a segunda imagem referência após interpolação e SR para a sequência *Poznan Street*.

V. Conclusões

Esse artigo aborda a síntese de vistas sob condições de assimetria, em termos de resolução, encontradas em formatos MVD-RM. Propomos um método de SR melhorado capaz de realçar imagens de baixa resolução com conteúdo de alta frequência proveniente de imagens adjacentes em resolução íntegra. Resultados comprovam os ganhos resultantes do método SR em termos de qualidade objetiva e subjetiva sobre imagens sintetizadas. Através da adaptação do peso entre vistas de referência assimétricas, os resultados de síntese podem também ser melhorados.

Possíveis trabalhos futuros incluem a escolha automática de parâmetro de peso para síntese a partir de diferenças de resolução entre imagens de referência. Outros métodos de SR, aplicados diretamente sobre o processo de síntese, devem ser pesquisados juntamente com métricas para aferir a qualidade de experiência do usuário em ambientes MVD-RM.

Referências

[1] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," *IEEE Signal Processing Magazine*, vol. 28, no. 1, January 2011.

[2] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *ICIP07*, San Antonio, USA, September 2007.

[3] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM SIGGRAPH and Trans. Graph*, pp. 600–608, 2004.

[4] M. Perkins, "Data compression of stereopairs," *IEEE Transactions on Communications*, vol. 40, no. 4, pp. 684–696, April 1992.

[5] H. Brust, G. Tech, K. Muller, and T. Wiegand, "Mixed resolution coding of stereoscopic video for mobile devices," in *3DTV-CON 2009*, Potsdam, Germany, May 2009.

[6] B. Julesz, *Foundations of cyclopean perception*, University of Chicago Press, Chicago, IL, USA, 1971.

[7] L. Stelmach, W. J. Tam, and A. Vincent, "Stereo image quality: Effects of mixed spatio-temporal resolution," *IEEE Trans. Circ. Syst. Video Tech.*, vol. 10, no. 2, pp. 188–193, March 2000.

[8] D. C. Garcia, C. C. Dorea, and R. L. de Queiroz, "Super-resolution for multiview images using depth information," in *ICIP10*, Hong Kong, China, September 2010.

[9] P. Kauff, N. Atzpadin, C. Fehn, K. Muller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image–based rendering for advanced 3DTV services providing interoperability and scalability," in *Image Commun.*, 2007.

[10] W. T. Freeman and T. R. Jones nad E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, March 2002.

[11] D. C. Garcia, C. C. Dorea, and R. L. de Queiroz, "Super-resolution for multiview images using depth information," *IEEE Trans. Circ. Syst. Video Tech.*, vol. 22, no. 9, pp. 1249–1256, 2012.

[12] K. Takahashi, M. Ishii, and T. Naemura, "Super-resolution plane-sweeping for free-viewpoint image synthesis," in *ICIP11*, Brussels, Belgium, September 2011.

[13] *Reference Softwares for Depth Estimation and View Synthesis:*, ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050, October 2009.

[14] M. Domanski, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, *Poznan Multiview Video Test Sequences and Camera Parameters*, ISO/IEC JTC1/SC29/WG11, M15386, 2009.

[15] *Tanimoto Lab at Nagoya University*, http://www.tanimoto.nuee.nagoya-u.ac.jp/.

[16] *Gwangju Institute of Science and Technology (GIST)*.