Vídeo com Ponto De Vista Livre Usando Decodificador em Rede

Thacio G. Scandaroli e Ricardo L. de Queiroz

***Resumo*— Vídeos multi-vistas podem ser utilizados para difusão do conteúdo de sistemas de ponto de vista livre, o que requer uma grande banda de transmissão e uma alta complexidade computacional na síntese de vista e estimação de profundidade. Neste trabalho, discutimos uma estratégia de configuração com um decodificador multi-vista embutido na rede que realiza estimação de profundidade e síntese de vista. Desta maneira, o dispositivo que exibe a imagem sintetizada ao usuário pode ter baixa complexidade computacional, como tablets e celulares. Nós conduzimos experimentos nesta arquitetura genérica para estudar a taxa de transmissão do conteúdo e a qualidade da vista sintetizada. Nossos experimentos indicam que a transmissão apenas das vistas sintetizadas é a melhor opção para sistemas com canal de retorno e com baixo número de usuários. Nossos resultados também indicam que estimar a profundidade no decodificador se equipara em taxa-distorção ao cenário em que o conteúdo multi-vista e os mapas de profundidade são transmitidos pela rede. Este trabalho abre uma nova discussão de configuração de sistemas de ponto de vista livre com decodificador em rede e de estimação de profundidade no decodificador.**

***Palavras-Chave*— Sistemas multi-vistas, Ponto de vista livre, Codificação de vídeo, Transmissão.**

***Abstract*— Multiview video can be used for free-viewpoint television (FTV) broadcasts, which require a great amount of bandwidth for data transmission and heavy computational complexity for view rendering and depth estimation. We discuss a network strategy wherein the decoder can make use of an in-network multiview decoder. In that, view synthesis can be computed at either the encoder or decoder node and this reflects on the data sent over the network. In this manner, we can use a low-complexity free-viewpoint device, such as a tablet. We conducted tests in this generic architecture to study the bandwidth required for the transmission of the content and on the quality of the rendered view. Our experiments indicate that transmission of the rendered view is the best option for systems with a feedback channel and with a small number of users. Our results also show that estimating depth information at the decoder using the decompressed views may perform close to the scenario in which multiview-plus-depth data is sent over the network. This work opens a new discussion on the setup for FTV systems with depth estimation at the decoder.**

***Keywords*— Multiview systems, Free viewpoint television, Video Coding, Transmission.**

I. Introdução

Vídeos multi-vistas [1] possibilitam novas aplicação como 3DTV [2] e televisão de ponto de vista livre (FTV) [3], esta sendo uma aplicação emergente que possibilita ao usuário o controle interativo dos pontos de vista da cena exibida. Aplicações multi-vistas aumentam drasticamente a largura de banda necessária em comparação à sistemas de vídeo tradicionais, então a compressão possui um papel importante em tais sistemas. Há diversos esforços sendo realizados para padronizar os formatos e técnicas de compressão de vídeos multi-vistas. As atividades de padronização do grupo MPEG já especificaram compressão eficiente de vídeos multi-vista, e agora o MPEG iniciou um grupo *ad-hoc* focado em 3DTV/FTV. As atividades mais recentes têm foco na reconstrução de alta qualidade de novas vistas. O grupo MPEG disponibilizou os softwares de referência para estimação de profundidade (DERS, *Depth Estimation Reference Software*) e síntese de vista (VSRS, *View Synthesis Reference Software*). O DERS estima o mapa profundidade de uma dada câmera utilizando as imagens de duas câmeras adjacentes à ela. Com os mapas de profundidade de duas câmeras, o VSRS sintetiza uma vista virtual que representa uma câmera posicionada entre duas câmeras utilizando baseado em mapas de profundidade e imagens (DIBR, *depth-image-based rendering*) [5] por meio da projeção de pixels.

A codificação de vídeo multi-vista (MVC) [6] já foi especificada no padrão H.264/MVC que extende o padrão H.264/AVC [7] com o acréscimento de predição entre vistas, explorando assim a redundância entre câmeras. O MVC utiliza tanto predição temporal como entre vistas para atingir uma compressão mais eficiente em sistemas multi-vistas. As informações de sistemas multi-vistas podem ser representadas por vídeos multi-vista e seus mapas de profundidade (MVD, *multiview-plus-depth*). A síntese de vista é sensível a qualidade do mapa de profundidade [8] e a compressão dos mapas podem ocasionar em artefatos na imagem sintetizada. Assim, pesquisas anteriores focaram em uma melhor compressão dos mapas de profundidade que ocasione menos artefatos na síntese de vista. [9], [10]

Para o controle interativo de pontos de vista, é necessário a estimação de profundidade e a síntese de vista, que demandam um alto poder computacional. Além disso, pode ser preciso um equilíbrio entre recursos como poder computacional e largura de banda disponível. Em difusão de sistemas FTV, um gerador de conteúdo computacionalmente potente pode transmitir o conteúdo MVD para vários receptores, cada um sintetizando seu próprio ponto de vista. Em um cenário contrastante como em uma conferência de vídeo FTV entre duas pessoas, pode haver um canal de retorno entre codificador e decodificador, de forma que o decodificador pode requisitar vistas e o codificador pode sintetizá-las e transmitir apenas tais vistas. Neste trabalho, nós apresentamos uma arquitetura genérica que pode acomodar diferentes configurações para sistemas FTV e nós analisamos o equilíbrio de taxa de transmissão por qualidade da vista sintetizada, pois a estimação de profundidade e a síntese de vista podem ser realizadas tanto no codificador ou

Thacio G. Scandaroli e Ricardo L. de Queiroz¸ Universidade de Brasília, Brasília-DF, Brasil, E-mails: thacio@gmail.com, queiroz@image.unb.br.

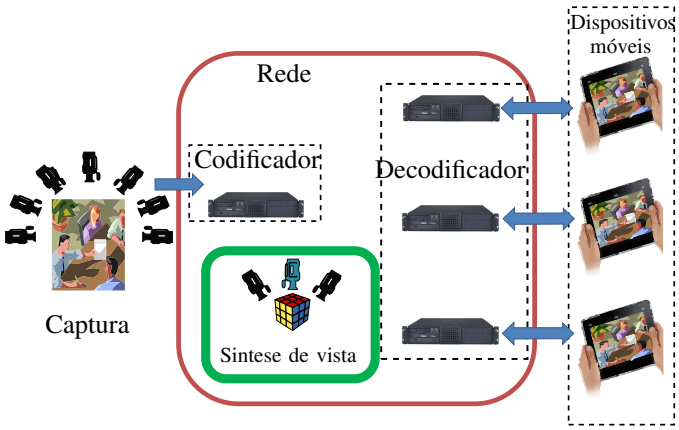


Fig. 1. Arquitetura genérica para sistemas FTV. O conteúdo é capturado e enviado para o codificador. As informações dos vídeos são codificadas e transmitidas pela rede para um decodificador na rede, que decodifica e transmite para o dispositivo móvel. A síntese de vista e estimação de profundidade podem ser realizadas no codificador ou no decodificador.

no decodificador.

A. Arquitetura de Sistemas FTV

Uma arquitetura genérica para sistemas FTV é ilustrada na Figura 1. Ela é baseada em três blocos principais: a captura da cena, a rede e o dispositivo móvel. O primeiro bloco captura o conteúdo multi-vista da cena. A rede é constituída por um bloco codificador e um bloco decodificador. A geração dos mapas de profundidade e síntese de vista podem ser realizadas em qualquer bloco com potência computacional adequada. Normalmente, estes processos são realizados no bloco codificador ou decodificador. Note que o dispositivo móvel pode ser utilizado como decodificador dependendo de sua potência computacional ou da aplicação do sistema. Cada dispositivo móvel pode escolher o seu próprio ponto de vista e o conteúdo multi-vista pode estar disponível para todos os receptores. A princípio, o ponto de vista é decidido pelo dispositivo móvel e o processo computacionalmente complexo de síntese de vista é realizado por algum equipamento ao alcance dele. Entretanto, se o dispositivo móvel tiver baixa complexidade computacional, é improvável que ele seja capaz de estimar a profundidade e sintetizar uma nova vista utilizando as informações multi-vista recebidas. Se o codificador transmitir os mapas de profundidade de cada vista, a síntese de vista é simplificada em troca de uma taxa de transmissão maior. É sugerido a introdução de um decodificador na rede capaz de receber conteúdos multi-vistas do codificador, com ou sem a transmissão dos mapas de profundidade, e sintetizar uma nova vista. Deste modo, dispositivos móveis se comunicariam com o decodificador e receberiam apenas a vista desejada. O decodificador da rede precisa re-codificar a vista sintetizada com uma qualidade proporcional à banda de transmissão disponível na rede local. Desta maneira, difusão de conteúdos FTV podem ser usados em *tablets* e outros equipamentos menores, transferindo os complexos processos de estimação de profundidade e síntese de vista para a rede.

A estimação de profunidade e síntese de vista podem ser realizadas em diferentes blocos da arquitetura, isto leva a três cenários diferentes.

B. Cenários para Sistemas FTV

Caso haja um canal de retorno, o decodificador pode informar ao codificador qual vista está sendo exibida ao usuário pelo dispositivo móvel. Logo, a geração dos mapas de profundidade e a síntese de vista podem ser realizadas no bloco codificador a apenas a vista sintetizada é transmitida ao dispositivo móvel, reduzindo a banda de transmissão utilizada pela rede. Este cenário é ilustrado na Figura 2(a).

Sem o canal de retorno, a síntese deve ser realizada no decodificador ou no dispositivo móvel, e o conteúdo multi-vista deve ser transmitido pela rede. Isto leva a dois casos distintos.

No primeiro, ilustrado na Figura 2(b), as imagens das vistas são transmitidas sem os mapas de profundidade. Logo, a estimação de profundidade e a síntese de vista são realizadas no decodificador ou no dispositivo móvel.

No segundo caso, ilustrado na Figura 2(c), a profundidade é estimada no codificador e o conteúdo MVD é transmitido. A nova vista é sintetizada no decodificador ou no dispositivo móvel.

Note que se a síntese de vista for realizada no decodificador da rede, a vista virtual deve ser codificada e enviada para o dispositivo móvel por uma rede local, resultando em mais perdas de qualidade. Se na rede local existir uma banda de transmissão suficientemente grande, a máxima qualidade atingida com a síntese de vista no bloco decodificador será no caso em que a codificação da vista sintetizada for realizada com compressão sem perdas. Neste caso, a síntese de vista no decodificador na rede ou no dispositivo móvel resultam na mesma qualidade do vídeo final, sendo este o caso considerado neste trabalho. Os três cenários para a arquitetura proposta são:

(a) Transmissão da vista sintetizada com estimação de profundidade e síntese de vista no bloco codificador. Um canal de retorno é necessário entre o codificador e o decodificador.
(b) Transmissão das imagens das vistas com estimação de profundidade e síntese de vista realizados no decodificador na rede ou no dispositivo móvel.
(c) Transmissão do conteúdo MVD com estimação de profundidade no codificador e síntese de vista no decodificador na rede ou no dispositivo móvel.

A Figura 2 ilustra esses cenários para um sistema FTV de cinco vistas.

II. Experimento

Foi realizado um experimento em um sistema FTV de cinco vistas. Foram usados 90 quadros das sequências de teste *Pantomime* e *Champagne* [11]. São consideradas as vistas 35, 37, 39, 41 e 43 destas duas sequências. O sistema é capaz de sintetizar qualquer vista intermediária entre as câmeras 35 e 43. A estimação de profundidade necessita da imagem de duas câmeras adjacentes, logo, para gerar o mapa de profundidade da vista 35, as imagens das vistas 33 e 37 são usadas como referência. Da mesma maneira, a geração do mapa de profundidade da vista 43 ocorre utilizando as

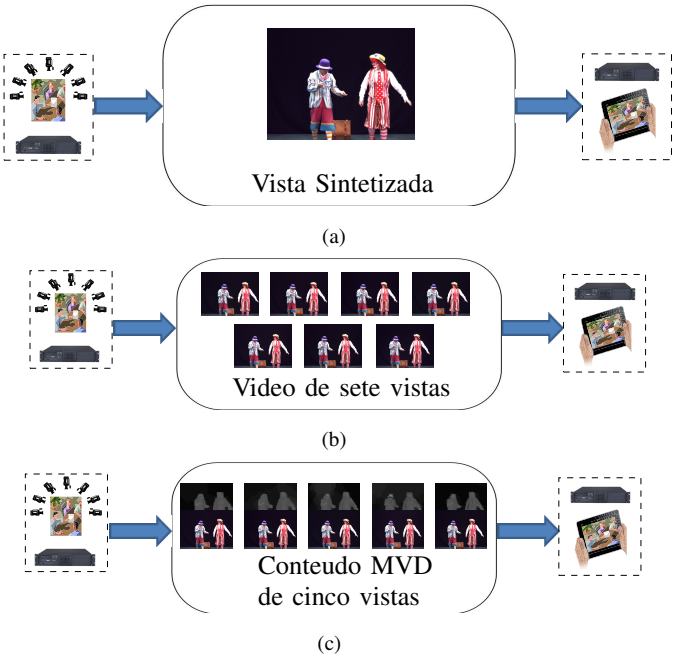


Fig. 2. Possibilidades de cenários de um sistema FTV de cinco vistas: (a) Transmissão da vista sintetizada, com síntese no codificador; (b) Estimação de profundidade e síntese no decodificador, baseado nas imagens das vistas transmitidas; (c) Síntese no decodificador baseado na transmissão das imagens e dos mapas de profundidade das vistas.

imagens das vistas 41 e 45. Embora este sistema use cinco vistas para sintetizar novas vistas, as vistas vizinhas 33 e 45 são necessárias apenas para o propósito de estimação de profundidade. Logo sete vistas são necessárias para o sistema. A síntese de vista é realizada com o software de referência VSRS 3.5 [4] e a estimação de profundidade é feita no modo automático do software de referência DERS 5.1 [4]. Os vídeos são codificados no padrão H.264 com o software de referência JMVC 8.3.1 para codificação do tipo AVC e MVC. Para codificação multi-vista, predição temporal e entre vistas são realizadas.

Para estudar a qualidade de sintetização do sistema, as vistas 35 e 37 são utilizadas para sintetizar a vista 36 das sequências de testes, e a PSNR entre a vista sintetizada e a imagem original da vista 36 é calculada. Note que não são necessárias todas as vistas para a sintetização da vista 36, mas todas as vistas precisam ser transmitidas nos cenários nos quais não há canal de retorno. A vista sintetizada com estimação de profundidade e síntese de vista com os vídeos originais sem perdas produz uma PSNR de 38.16 dB para a sequência Pantomime e 32.55 dB para a sequência Champagne. Neste experimento, os QPs utilizados para a codificação das imagens são: 6, 10, 14, 18, 22, 30, 38 e 46. Para os mapas de profundidade, os QPs usados são: 10, 18, 38, 46.

A. *Síntese no Codificador*

No cenário (a), ilustrado na Figura 2(a), apenas a vista 36 sintetizada é transmitida pela rede. Então, a PSNR entre a imagem original e descomprimida da vista 36 sintetizada é calculada.

B. *Estimação de Profundidade e síntese no Decodificador*

No cenário (b), ilustrado na Figura 2(b), as imagens das sete vistas (33, 35, 37, 39, 41, 43, 45) são transmitidas pela rede. Como mencionado anteriormente, as imagens das vistas 33 e 45 são enviadas apenas pelo propósito de estimação de profundidade das vistas 35 e 43. As imagens das sete vistas são codificadas em MVC com o mesmo parâmetro de quantização (QP). A estimação de profundidade é realizada com os vídeos descomprimidos, então a vista 36 é sintetizada. A taxa de bits considerada é a taxa de bits total necessária para o envio das imagens das sete vistas.

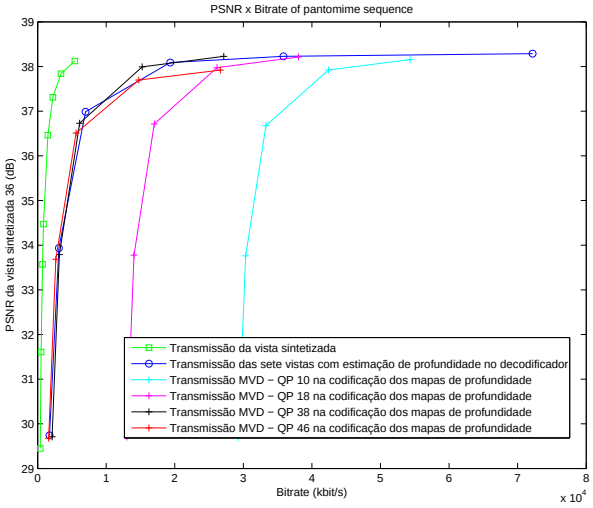
C. *Transmissão do Conteúdo MVD com Síntese no Codificador*

No cenário (c), o conteúdo MVD é transmitido pela rede para síntese no decodificador. As imagens e mapas de profundidade das vistas são codificados separadamente em MVC e elas podem ser codificados com diferentes QPs. Para obter os dados de PSNR por taxa de bits, é fixado um QP para a compressão dos mapas de profundidade enquanto é variado o QP das imagens das vistas. Variando o QP dos mapas de profundidade, é obtido uma de PSNR por taxa de bits para cada QP dos mapas de profundidade. A síntese da vista 36 é realizada utilizando as informações decodificadas. A taxa de bits considerada é a taxa de bits total necessária para enviar o conteúdo MVD das cinco vistas. Esta abordagem é útil na análise dos efeitos de compressão dos mapas de profundidade na síntese de vista.

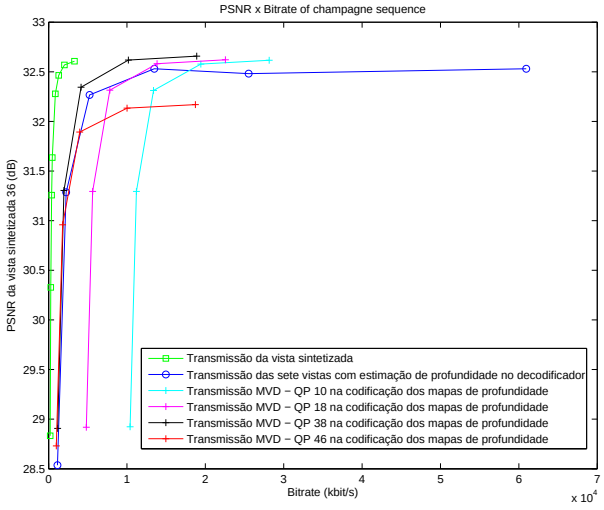
III. Resultados

A Figura 3 mostra a comparação entre os três cenários. Na figura 3(a), o cenário (a) necessita, para uma PSNR de 38.1 dB, de uma taxa de bits de aproximadamente 74% menor que os outros dois cenários. Na Figura 3(b), a taxa necessária para o cenário (a) para uma PSNR de 32.6 dB é aproximadamente 80% menor se comparada com a melhor transmissão do conteúdo MVD do cenário (c). Este cenário só é adequado para aplicações FTV com um baixo número de usuários, como vídeo conferências. Caso o número de usuários que tenham escolhidos diferentes pontos de vista seja aproximadamente igual ou ultrapasse o número de vistas do sistema, é esperado que os outros cenários sejam mais eficientes que este.

Para o cenário (c), é possível observar que a qualidade das imagens das vistas são mais importantes para o desempenho taxa-distorção da vista sintetizada do que a qualidade dos mapas de profundidade, como observado em [12]. A PSNR da vista sintetizada com diferentes taxas de compressão para os mapas de profundidade apresentaram pequenas diferenças, embora a taxa de bits dos mapas de profundidade afetem significativamente a taxa de bits total do sistema. Na Figura 3(a), o cenário (b) teve uma performance parecida com o cenário (c) para baixas e médias taxas de bits, mas foi superado nos casos de alta taxa de bits para os mapas de profundidade. Na Figura 3(b), o cenário (b) teve sua performance superada pelo cenário (c) para altos valores de PSNR, isto ocorreu pelo fato de que, para a sequência Champagne, a estimação de profundidade utilizando imagens de referência descomprimidas causaram



(a) Pantomime



(b) Champagne

Fig. 3. Comparação de um sistema FTV de cinco vistas para as sequências Pantomime e Champagne em diferentes cenários. As curvas são obtidas com a taxa total de bits necessária para cada cenário e as informações PSNR são calculadas entre a imagem original e sintetizada da vista 36.

distorções na qualidade da vista sintetizada final. No geral, o cenário (b) alcançou uma melhor performance no sentido taxa-distorção em relação ao cenário (c).

Esta seção considerou um sistema FTV de cinco vistas. Em sistemas FTV práticos, é esperado um número maior de câmeras, podendo ser utilizados na captura de 3 a 100 câmeras, ou mais. Neste caso, a estimação de profundidade no codificador deve obter uma performance melhor do que o cenário em que o conteúdo MVD é transmitido. Em um sistema FTV em que as câmeras são alinhadas e espaçadas apenas em um direção, a horizontal, assim como nas sequências de teste utilizadas neste capítulo, o cenário (b) necessitaria de apenas duas vistas a mais para estimação de profundidade no decodificador, enquanto no cenário (c), a taxa de bits pros mapas de profundidade aumentariam com o aumento do número de câmeras. As pesquisas atuais têm seu foco na melhora da compressão dos mapas de profundidade, reduzindo artefatos gerados na síntese de vista devidos a compressão. Trabalhos futuros podem focar na melhor estimação de profundidade utilizando imagens descomprimidas, pois isto levaria a ganhos no sistema podendo ter uma performance melhor do que cenário (c).

IV. Conclusões

Foi proposto uma arquitetura genérica que acomoda diferentes configuração para sistemas multi-vistas com síntese de vista. A estimação de profundidade e a síntese de vista podem ser realizadas tanto no lado do codificador ou do decodificador, criando assim três diferentes cenários. O primeiro estima a profundidade e sintetiza a vista virtual no lado do codificador, havendo a transmissão apenas da vista sintetizada para o usuário. Este cenário é propício para aplicações com canal de retorno e um baixo número de usuários, como vídeo conferências. Entretanto, ele pode não ser eficiente para aplicação em que o número de usuários se aproxime ou ultrapasse o número de câmeras do sistema. No segundo cenário, o codificador transmite as imagens de todas as câmeras. No decodificador, a profundidade é estimada utilizando esses vídeos descomprimidos e então a vista virtual é sintetizada. No terceiro cenário, a profundidade é estimada no codificador e o conteúdo MVD é transmitido pela rede. A vista virtual é sintetizada no decodificador utilizando essas informações. Experimentos foram realizados em um sistema FTV de cinco vistas, no qual renderizar a vista virtual no codificador produz o melhor resultado em taxa-distorção. Os outros dois cenários produziram resultados próximos. Trabalhos futuros podem focar na melhor estimação de profundidade utilizando vídeos descomprimidos como referência, pois o segundo cenário pode produzir melhores resultados que o cenário MVD para sistemas com um grande número de vistas.

Referências

[1] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview imaging and 3DTV," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 10 –21, nov. 2007.

[2] A. Smolic and P. Kauff, "Interactive 3-D video representation and coding technologies," *Proc. of the IEEE*, vol. 93, no. 1, 2005.

[3] M. Tanimoto, M.P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," *IEEE Signal Process. Magazine*, vol. 28, no. 1, pp. 67 –76, jan. 2011.

[4] ISO/IEC JTC1/SC29/WG11, "Reference softwares for depth estimation and view synthesis," Doc. M15377, April 2008.

[5] S.C. Chan, H.-Y. Shum, and K.-T. Ng, "Image-based rendering and synthesis," *IEEE SP Magazine*, vol. 24, no. 6, 2007.

[6] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Trans. on Circ. and Sys. for Video Tec.*, vol. 17, no. 11, 2007.

[7] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. on Circ. and Sys. for Video Tec.*, vol. 13, no. 7, 2003.

[8] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, P. H. N. de With, and T. Wiegand, "The effects of multiview depth video compression on multiview rendering," *Signal Process.: Image Commun.*, vol. 24, pp. 73–88, January 2009.

[9] V. Jantet, C. Guillemot, and L. Morin, "Object-based layered depth images for improved virtual view synthesis in rate-constrained context," in *IEEE ICIP*, sept. 2011, pp. 125 –128.

[10] K.-J. Oh, A. Vetro, and Y.-S. Ho, "Depth coding using a boundary reconstruction filter for 3-D video systems," *IEEE Trans. on CSVT*, vol. 21, no. 3, pp. 350 –359, march 2011.
[11] "http://www.tanimoto.nuee.nagoya-u.ac.jp/"
[12] K. Klimaszewski, K. Wegner, and M. Domanski, "Distortions of synthesized views caused by compression of views and depth maps," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*