Multiple-view Video Coding Using Depth Map in Projective Space

Nina Yorozu, Yuko Uematsu, and Hideo Saito

Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa, 223-8522, Japan
{yorozu,yu-ko,saito}@hvrl.ics.keio.ac.jp

Abstract. In this paper a new video coding by using multiple uncalibrated cameras is proposed. We consider the redundancy between the cameras view points and efficiently compress based on a depth map. Since our target videos are taken with uncalibrated cameras, our depth map is computed not in the real world but in the Projective Space. This is a virtual space defined by projective reconstruction of two still images. This means that the position in the space is correspondence to the depth value. Therefore we do not require full-calibration of the cameras. Generating the depth map requires finding the correspondence between the cameras. We use a "plane sweep" algorithm for it. Our method needs only a depth map except for the original base image and the camera parameters, and it contributes to the effectiveness of compression.

1 Introduction

According to the development of digital image processing technique, multiple-view video captured with multiple cameras has high demand for a lot of new media: broadcasting, live sport events, cinema production, and so on. The "Eye-Vision" system [1] used in the live broadcasting of American football is famous and landmark example of the multiple-view video researches. In the field of cinema production, "Matrix" [2] employed the novel technique for generating the computer graphics based on real videos and created the scene where virtual camera was panning around the actor for a moment. Moreover, for such as 3D-TV and free-viewpoint TV (FTV), there are many related researches [3][4][5] that generate free-viewpoint images from real images taken by multiple cameras. By free-viewpoint images, viewers can easily change their interactive viewpoints without concerning about the real camera position. As noted by Tanimoto [6] and Smolic *et al.* [7], these types of videos have many advantages in many fields.

On the other hand, streaming distribution of a movie is provided on the Internet at present. It is expected that streaming distribution of a multiple-view videos will also start in the future, and multiple-view videos coding (MVC) will become very important.

In the case of MVC, the redundancy between viewpoints should also be taken into consideration besides spatial or time redundancy. Many techniques have been proposed and they can be classified into some approaches.

G. Bebis et al. (Eds.): ISVC 2009, Part II, LNCS 5876, pp. 651–661, 2009.

Object base coding is applied by MPEG-4. In this coding, the scene is constructed by synthesizing each object. This requires the objects in the scene to be separated in advance. It is difficult to apply this coding to natural images.

Disparity compensation is the most popular approach and the technique is an extended method of a single-view video coding. The images taken by other viewpoints are treated as the encoding target and are used just as reference images for coding. Therefore, disparity information such as motion vectors and residual signals, i.e., prediction error, are encoded and transmitted to the decoder side. Though it has effectiveness for time redundancy, this is a minor benefit for distantly-positioned cameras.

View synthesis and *View interpolation* use techniques from the field of image-based rendering to predict coding target images. These approaches have been developed in order to use the fact that the disparity of an object between two views depends on the geometric relation between their cameras. This means that the objects closer to the camera move much more than the objects far from the camera when moving the position of view. Therefore, one of the most popular and general algorithms is to use depth instead of disparity vectors, as proposed by Martinian *et al.*[8], Shimizu *et al.* [9], Tsung *et al.* [10] and Ozkalayci *et al.* [11]. Many of them focus on the color matching between the input images to generate a higher-accuracy depth map. These approaches require full-calibration of the multiple cameras to get a depth map, however, such a calibration of many cameras is very time consuming task. This is one of the difficulties for practical use of multiple-view shooting. Moreover easy segmentation of images is also necessary to generate a depth map.

In this paper, a new video coding method based on a depth map is proposed. The targets are multiple-view videos which are taken by multiple cameras. We focus on the redundancy between the cameras to compress large-volume videos. In this method, we use only an original image and a depth map of a base camera, which is one of the multiple cameras, and predict the images taken by the other cameras. In contrast with a conventional method based on a depth map, our method does not require full-calibration to generate a depth map, because it is computed in the Projective Space that is a virtual space. This space is defined by projective reconstruction of two images. It means that the depth value of our depth map is corresponding to the position in the Projective Space not in the real world. Therefore we do not require full-calibration of the cameras.

For getting correspondence between the cameras to generate the depth map, we apply *Plane Sweep* algorithm [12]. We assume that two virtual planes are defined in the Projective Space so that every target object lies between the planes. The space between the two planes is divided by parallel multiple planes. By projecting each pixel of each plane onto the input images and matching the colors found in the images, then, pixel-to-pixel correspondences between the cameras are obtained.

Many related researches have applied the Joint Multiview Video Model (JMVM) [13] that is the reference software for MVC to evaluate their methods. However, in this paper, we examine the effectiveness of compression by using multiple (still) images captured at a same time and apply the entropy coding as the multiple-view video coding method. If this experiment for the multiple still images achieves good result, our method will also have good performance for multiple videos.

In the section 2, we explain detailed our method. In the section 3, we apply our method to multiple-view video coding, and demonstrate the effectiveness of compression. The conclusions are given in the section 4.

2 Method

The outline of our method is shown in Fig. 1. In our method, we use three uncalibrated cameras: base camera, reference camera and input camera.

The base camera, which is one of the cameras, is used as a basis for coding, and the "base image" is taken with the base camera. The images taken with the other two cameras are called as "reference image" and "input image", respectively. The target images of coding are the reference image and the input image. Therefore those two images are predicted by using the base image.

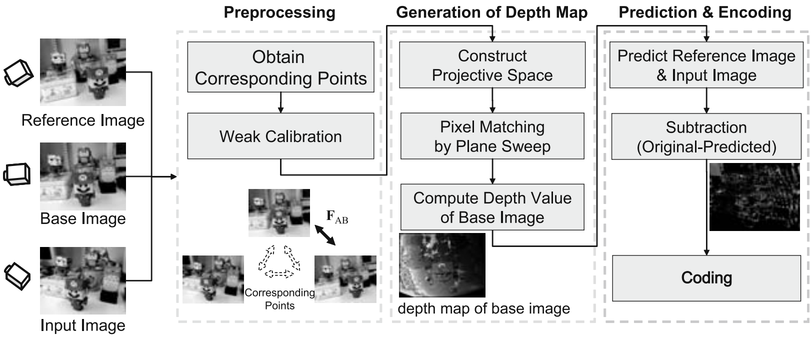

Preprocessing
Generation of Depth Map
Prediction & Encoding
Reference Image
Base Image
Input Image
Obtain Corresponding Points
Weak Calibration
F_AB
Corresponding Points
Construct Projective Space
Pixel Matching by Plane Sweep
Compute Depth Value of Base Image
depth map of base image
Predict Reference Image & Input Image
Subtraction (Original-Predicted)
Coding

Fig. 1. Outline of Proposed Method

Our method is divided into a preprocessing and two main processing, "Depth Map Generation" and "Prediction & Coding". In the preprocessing, F-matrix that represents the epipolar geometry between every pair of images is obtained. By finding eight or more pair of corresponding points, the F-matrix is computed. That is usually called as the weak calibration.

In the "Depth Map Generation", a depth map of the base image is generated by constructing a Projective Space, which is a virtual 3D space. For constructing the Projective Space, the base and reference images are utilized with projective reconstruction.

In the "Prediction & Coding", the reference and input images are predicted by using the computed depth map of the base image. And then, the difference between the original images and the predicted images of the reference and input images are encoded. The details will be described in the following sections.

2.1 Generation of Depth Map

Our method does not require full-calibration and generate a depth map in the Projective Space, which is a 3D virtual space. Using the F-matrix obtained in the preprocessing, the Projective Space is constructed from the base image and the reference image. Then, the position in the Projective Space not in the real world is corresponding to each depth value of our depth map. For getting correspondence between the cameras to generate the depth map, we apply *Plane Sweep* algorithm [12]. A detailed flow is shown in Fig. 2.

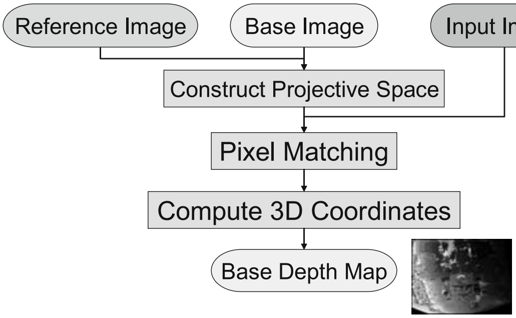


Fig. 2. Flow of Generation of Depth Map

Construct Projective Space The Projective Space is constructed from two images, base image and reference image, taken by two cameras shown in Fig. 3. Since this technique is based on the projective reconstruction, the parallelism of axes is unkept.

When epipolar geometry between two images (cameras) is established, the relationship between the Projective Space and two images are respectively

$$\boldsymbol{P}_A = [\mathbf{I}|\mathbf{0}], \quad \boldsymbol{P}_B = [-\frac{[\boldsymbol{e}_B]_\times \boldsymbol{F}_{AB}}{\|\boldsymbol{e}_B\|^2}|\boldsymbol{e}_B] \tag{1}$$

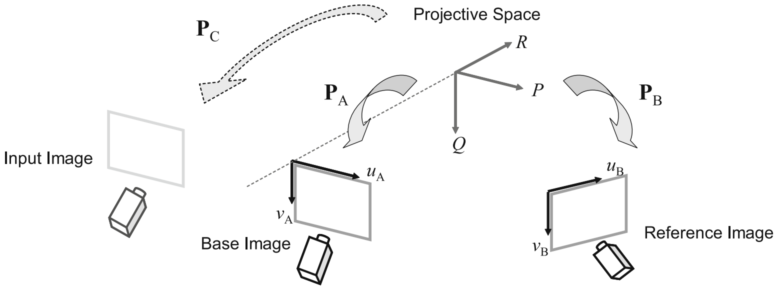


Fig. 3. Projective Space

where $\boldsymbol{P}_A$ and $\boldsymbol{P}_B$ are the projection matrices to the base image and the reference image, $\boldsymbol{F}_{AB}$ is a F-matrix of the base image to the reference image, and $\boldsymbol{e}_B$ is an epipole on the reference image. Consider $\boldsymbol{X_P}(P, Q, R)$ as a point in the Projective Space, $\boldsymbol{x_A}(u_A, v_A)$ as on the base image, $\boldsymbol{x_B}(u_B, v_B)$ as on the reference image, we can write

$$\boldsymbol{M}\tilde{\boldsymbol{X_P}} = \boldsymbol{0} \tag{2}$$

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{p^1_A} - u_A\boldsymbol{p^3_A} \\ \boldsymbol{p^2_A} - v_A\boldsymbol{p^3_A} \\ \boldsymbol{p^1_B} - u_B\boldsymbol{p^3_B} \\ \boldsymbol{p^2_B} - v_B\boldsymbol{p^3_B} \end{bmatrix} \tag{3}$$

$\boldsymbol{p^i_A}$C$\boldsymbol{p^i_B}$ are the ith column vector of $\boldsymbol{P}_A$ and $\boldsymbol{P}_B$. Then, we obtain $\tilde{\boldsymbol{X_P}}(P, Q, R, 1)$ by the singular value decomposition of $\boldsymbol{M}$.

If more than six corresponding points are detected among the three images, the base image, the reference image and the input image, the projection matrix $\boldsymbol{P}_C$ from the Projective Space to the input image is obtained. The projection matrices $\boldsymbol{P}_A$, $\boldsymbol{P}_B$ and $\boldsymbol{P}_C$ are used for the pixel matching.

Pixel Matching. All pixels in the base image are matched to the reference image and the input image, and their 3D coordinates in the Projective Space are computed. The 3D coordinates represent the depth value of our depth map. For getting correspondence between two images, we employ *Plane Sweep* algorithm as shown in Fig. 4.

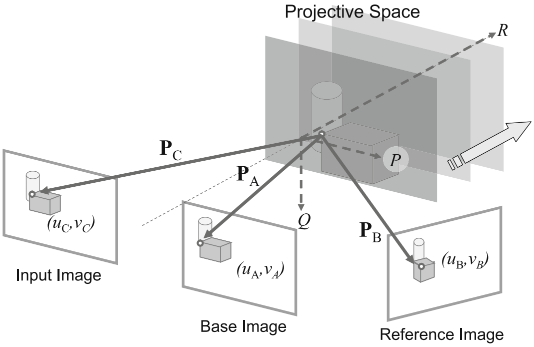


Fig. 4. Plane Sweep

The space is divided by parallel multiple planes. We assume that every target object lies between the planes. By projecting each pixel of each plane onto the three images (base, reference, input) using $\boldsymbol{P}_A$, $\boldsymbol{P}_B$, $\boldsymbol{P}_C$ and matching the colors found in the images, then, pixel-to-pixel correspondences between the images are obtained. The 3D coordinate$\boldsymbol{X_P}(P, Q, R)$ in the Projective Space can be computed by the matched pair of pixels. Therefore we consider R as the depth value in the depth map.

2.2 Prediction and Coding

In our method, the basis of coding is the base image, and the target of coding is the reference image and the input image. We predict the reference image and the input image only by using the information of the base image such as the depth map and the intensities.

For coding, we make the subtraction images by subtracting the original image from the predicted image of the reference image and the input image.

A detailed flow is shown in Fig. 5.

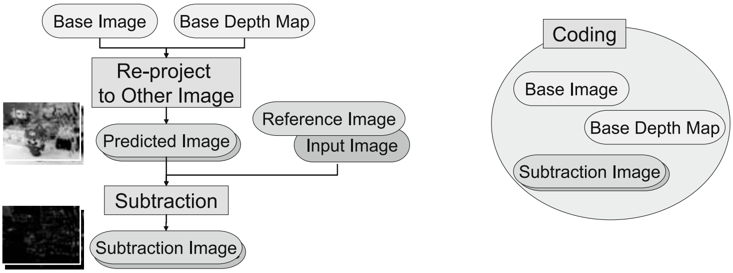


Fig. 5. Flow of Prediction and Coding

Prediction Other Images. By the definition of the Projective Space, we can consider $\boldsymbol{X_P}(P, Q, R)$ as a point in the Projective Space, $\boldsymbol{x_A}(u_A, v_A)$ as on the base image. Therefore, the relationship between 2D coordinate on the base image and 3D coordinate in the Projective Space is described as follows

$$u_A = P/R, \quad v_A = Q/R \tag{4}$$

When the depth map of the base camera is obtained, the 3D coordinates of all pixels can be computed, because $\boldsymbol{x_A}(u_A, v_A)$ and R is known. Then, by projecting every point onto the reference image and the input image, each image is predicted.

Image Coding. After getting the subtraction images of the reference image and the input image, they are encoded with the original image and the depth map of the base image. As described above, if the predicted images are quite accurate, the subtraction images should be similar to 0. Therefore more accurate prediction makes more efficient coding. In our method, Entropy Coding is applied as the multiple-view video coding method. The entropy E of the image is represented as follows

$$E = -\sum_i S_i \log S_i \tag{5}$$

where S_i is the probability of the color value i ($0 \leq i \leq 255$). In the same way, the entropy is computed for each image; base image, depth map and subtraction images of the reference image and input image.

3 Experimentation and Discussion

We applied our method to following two test color sequences. In both cases, we used three non-calibrated cameras and set them as shown in Fig. 6. In this experiment, the corresponding points are manually selected for the weak calibration.

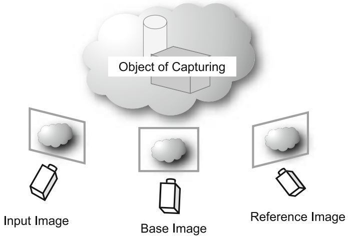
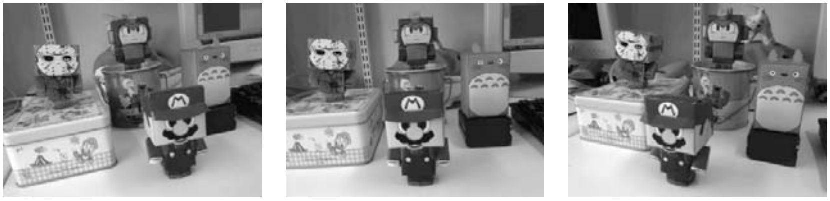
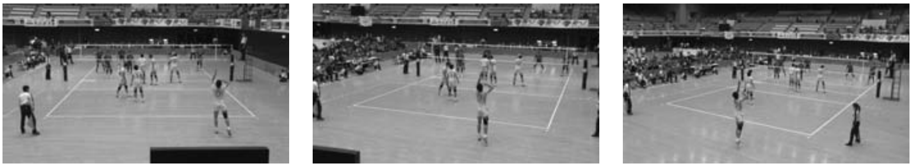


Fig. 6. Experimental Scene



(a)Input Image (b)Base Image (c)Reference Image

Fig. 7. Images for Test Sequences “on the desk”



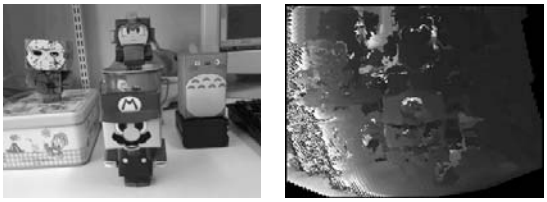
(a)Input Image (b)Base Image (c)Reference Image

Fig. 8. Images for Test Sequences “volleyball”

- “on the desk” : a scene that some paper-crafts are put on the desk (with a 320 × 240 resolution, as shown in Fig. 7)
- “volleyball” : a scene of a volleyball game (with a 480 × 270 resolution, as shown in Fig. 8)

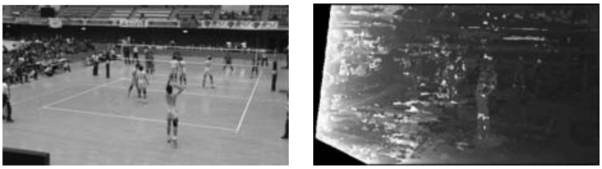
3.1 Depth Map

Generated depth maps are shown in Fig. 9, 10. Our depth map is represented in the Projective Space not in the real world. Since the axis of R may not be perpendicular to the image plane, as described in Sec. 2.1, the depth map is visually different from a general depth map. The area of the depth map is the common area of three images (base, reference, input), because the map is generated by getting correspondence among them.



(a)Original (b)Depth Map

Fig. 9. Depth Map of a base camera “on the desk”



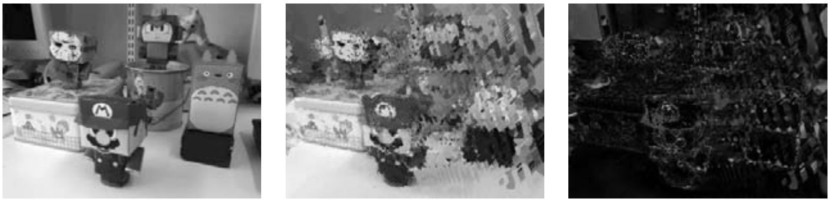
(a)Original (b)Depth Map

Fig. 10. Depth Map of a base camera “volleyball”

3.2 Prediction and Subtraction

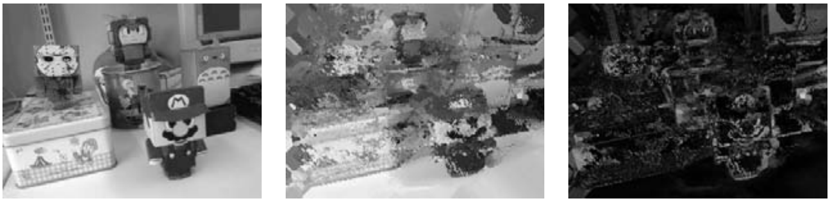
As described in the section 2.2, the results of the prediction and subtraction are shown as Fig. 11 to Fig. 14. The area, that is not common area captured by three cameras, is interpolated with neighbor colors and the whole image is predicted.

As shown in the subtraction images (c), the difference value is quite small in every image. This is because our predicted images have high accuracy. As described in Sec. 2.2, the accurate prediction can increase the coding efficiency, because the subtraction image becomes almost 0. The quantitative evaluation of the coding is described in the next section.



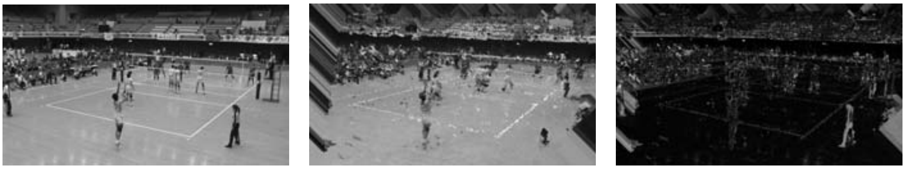
(a)Original (b)Predicted (c)Subtraction

Fig. 11. Reference Image of “on the desk”



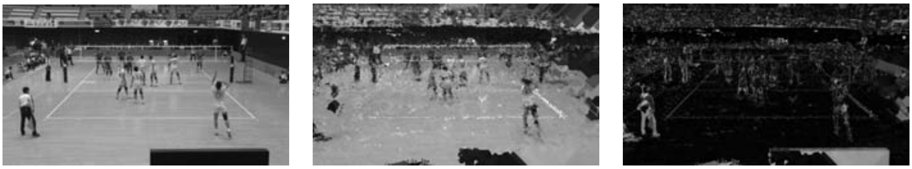
(a)Original (b)Predicted (c)Subtraction

Fig. 12. Input Image of “on the desk”



(a)Original (b)Predicted (c)Subtraction

Fig. 13. Reference Image of “volleyball”



(a)Original (b)Predicted (c)Subtraction

Fig. 14. Input Image of “volleyball”

3.3 Comparison Entropy

The calculation result of the entropy is shown in Table 1. Even though our method needs only a depth map except for the original base image and the

Table 1. Comparison Entropy

	Entropy [bit/pixel]				
	Base Image	Depth Map	Reference Image (Subtraction Image)	Input Image (Subtraction Image)	Total
“on the desk”					
original	22.1	-	22.1	22.2	66.4
proposed	22.1	6.5	(18.2)	(19.1)	65.9
“volleyball”					
original	20.8	-	21.0	21.2	63.0
proposed	20.8	7.0	(17.6)	(17.4)	62.8

camera parameters, it has the effectiveness of compression. We use three cameras in this experiment, however, it can achieve high and efficiency compression if we use more cameras. This is because, our method employs the Plane Sweep algorithm which uses color matching of every pixel of all cameras. The more cameras are utilized, therefore, the higher accuracy of pixel matching is obtained.

4 Conclusion

In this paper, a new video coding method based on a depth map is proposed. The target of our method is multiple-view images which are taken with multiple cameras. We consider the redundancy between the viewpoints of the cameras and efficiently compress large-volume image data.

Only using a single original image and a depth map, our method could predict the images taken with the other cameras. Applying our method to multiple-view video coding, we demonstrated the effectiveness of the compression. Even though our method needs only the depth map and the original image, it achieved effective compression than using raw images.

One advantage of our method is that we do not require full-calibration of the cameras in contrast with conventional method using a depth map. In our method, the depth map is generated in the Projective Space that is a virtual 3D space defined by projective reconstruction of two images. Therefore, we need only weak-calibration, which represents epipolar geometry of the cameras. This is a big advantage, because any uncalibrated videos (images) can be easily applied to our method.

In our future work, we plan to use more cameras to make more effective depth map and apply our method to the video sequence. By applying general feature detection technique to obtaining corresponding points between the cameras in computing F-matrix, we can easily extend to full automatic system.

Acknowledgments

This research was supported by National Institute of Information and Communications Technology (NICT).

References

1. Eye Vision, `http://www.pvi-inc.com/eyevision/`
2. Manex Entertainment Inc.: Matrix, `http://www.mvfx.com`
3. Chen, S.E., Williams, L.: View interpolation for image synthesis. IEEE Trans. on Pattern Analysis and Machine Intelligence 20, 218–226 (1998)
4. Inamoto, N., Saito, H.: Fly through view video generation of soccer scene. In: IWEC CWorkshop Note, May 2002, pp. 94–101 (2002)
5. Nozick, V., Saito, H.: On-line free-viewpoint video: From single to multiple view rendering. International Journal of Automation and Computing 5, 257–265 (2008)
6. Tanimoto, M.: Overview of free viewpoint television. Signal Proceedings: Image Communication 21, 454–461 (2006)
7. Smolic, A.: 3d video and free viewpoint video -technologies, applications and mpeg standards. In: Proc. ICME 2006, July 2006, pp. 2161–2164 (2006)
8. Martinian, E., et al.: View synthesis for multivew video compression. In: Proc. PSC 2006, April 2006, pp. SS3–4 (2006)
9. Shimizu, S., et al.: View scalable multiview video coding using 3-d warping with depth map. IEEE Trans. Circuits Syst. Video Technol. 17, 1485–1495 (2007)
10. Tsung, P.K., Lin, C.Y., Chen, W.Y., Ding, L.F., Chen, L.G.: Multiview video hybrid coding system with texture-depth synthesis. In: IEEE International Conference on Multimedia and Expo., April 2008, vol. 26, pp. 1581–1584 (2008)
11. Ozkalayci, B., Serdar Gedik, O., Aydin Alatan, A.: Multi-view video coding via dense depth estimation. In: 3DTV Conference, May 2007, pp. 1–4 (2007)
12. Collins, R.: A space-sweep approach to true multi-image matching. In: Proceedings of IEEE Computer Society Conference on CVPR, pp. 358–363 (1996)
13. MPEG-4 Video Group: Joint multiview video model (jmvm) 1.0