

VIEWPOINT INTERPOLATION: DIRECT AND VARIATIONAL METHODS

Sergi Pujades Frédéric Devernay

PRIMA team, INRIA Grenoble Rhône-Alpes, France

ABSTRACT

We address the topic of novel view synthesis from a stereoscopic pair of images. Most current techniques are composed of 3 stages: reconstruction of correspondences between the views, estimation of the per-pixel or per-image weight of each view in the final view, and rendering. Rendering methods can be classified into *direct* methods, which define the final image as a function of the original images, and *variational* methods, where the synthesized image is expressed as the solution minimising an energy. The existing literature is mainly focused on the correspondence issue, but few publications address the questions of which weights or blending factors are better, or which rendering method should be used. In this paper, we experiment different combinations of blending factors and rendering methods, in order to evaluate the influence of these two factors on the final image quality.

Index Terms— Viewpoint interpolation, image-based rendering, blending factors, variational method.

1. INTRODUCTION

Novel view synthesis from a stereoscopic pair of images has been extensively studied, for example for content creation for glasses-free 3D displays from binocular stereoscopic content. Those techniques proceed in general in 3 stages [1, 2]: the estimation of correspondences between the novel and the reference views, the estimation of the contribution (or weight) of each view in the final view, and the rendering method. The stereo correspondence estimation problem has been largely explored [3], but few works have formally studied which is the *correct* leverage between the contributions of each view. We will call this leverage the “*blending factors*” of each view. The most common blending factors considered in the literature are the local deformation of the image introduced by the change in the view point, and the distance between the new view and the reference ones. Yet, several questions arise: How is each blending factor choice formally supported? How does the synthesized view quality depend on these blending factors? Ultimately, what blending factors should be preferred?

In the view interpolation domain, methods can be clustered in 2 groups. The *direct* methods and the *variational*

methods. Most state of the art methods are direct: the color of a pixel in the final image is given as a function of the colors of the corresponding pixels in the reference images. In the variational methods, the final image minimises an energy, often corresponding to the maximum a posteriori (MAP) derived from a generative model. Using the Bayesian formalism, the blending factors between the views can be formally derived. However, those optimization techniques require heavier computations than the direct ones. This paper tries to answer whether the results obtained with these methods compensate for their computational complexity, and analyse the impact of the blending factors and the used method on the final result. Those methods are evaluated on Lambertian and non-Lambertian scenes in order to see, in which case, which choice is better.

2. PRIOR WORK

Viewpoint interpolation methods belong to the largely studied field of image-based rendering [4]. *Unstructured Lumigraph* [5] introduced the desirable properties that an ideal image rendering method should have: “*use of geometric proxies*”, “*unstructured input*”, “*epipolar consistency*”, “*minimal angular deviation*”, “*continuity*”, “*resolution sensitivity*”, “*equivalent ray consistency*”, and finally “*real-time*”. Moreover, the method explains how to weight each of reference images when rendering the final view. The authors also present a direct rendering method taking into account all these desirable properties. The “*minimal angular deviation*” property is computed from the angles between the optical rays of the rendered and reference images. The “*resolution sensitivity*” property is enforced by computing an approximation of the Jacobian of the planar homography between the rendered and the reference image. The balance between these factors is tuned manually, and depends on the scene. More precisely, “*resolution sensitivity*” has in most experiments a tiny weight compared to the “*minimal angular deviation*”.

Existing methods addressing the problem of view interpolation from a stereoscopic pair of images use mainly two kinds of blending factors: most of them consider the normalised distance α between the new and the reference images. The two associated blending factors are α and $(1 - \alpha)$ [6, 7, 8]. Other methods measure the deformation of the image [9], using the Jacobian of the planar homography between the new and the reference images.

This work was done within the Action 3DS project, founded by the French *Caisse des dépôts et consignations*

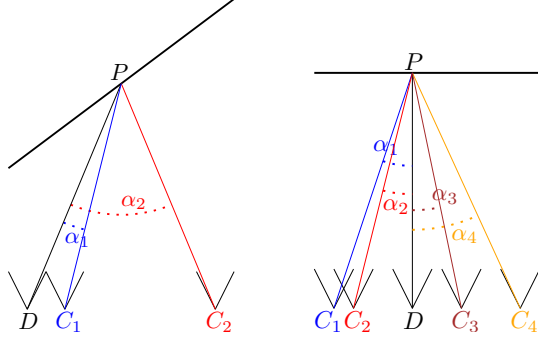


Fig. 1. Rendering view D from C_i using [10]. Left: resolution sensitivity will prefer C_2 to C_1 , even-though angular distance would prefer C_1 over C_2 . Right: flat scene parallel to views: all views will have the same blending factor, even-though their angular distance is different for each of them.

While addressing the problem of generating super-resolved images from multiple views, Wanner and Goldluecke [10] proposed a very general variational method to generate images at new viewpoints. They present a generative model describing the image formation process and establishing the energy corresponding to its maximum a posteriori, using the Bayesian formalism. This formalisation brings them to derive the blending factors of the images as the determinant of the Jacobian of the 2D transformation between the final image and the reference one. While these blending factors take into account the “*resolution sensitivity*”, the “*minimal angular deviation*” is overlooked. It doesn’t appear in the equations and Fig. 1 illustrates two configurations showing the contradictions between [10] and [5].

Alternative methods are proposed by [11, 12] based on the *deformable meshes* from Gal et al.[13]. Their main hypothesis is that artifacts introduced by mesh deformations are visually more acceptable than those produced by image blending. However, in case of important deformations of the images, it would be interesting to merge two images generated with these techniques. Unfortunately this question is avoided in their works.

So, in the literature, different blending factors are used, but to our knowledge no study directly compares their performance.

3. VIEWPOINT INTERPOLATION METHODS

The viewpoint interpolation methods considered here have all the same input: a rectified pair of stereoscopic images, two disparity maps which establish per-pixel correspondences between the views, and a value $\alpha \in [0, 1]$ corresponding to the position of the interpolated view with respect to the original views, being 0 the left image position, and 1 the right image position. The interpolated view has the same orientation and parameters as the rectified views.



Fig. 2. Example images from the datasets: *aloe* from [14] and *tarot*, *bracelet* and *amethyst* from [15].

3.1. Direct Methods

The direct methods first compute the inverse transformations going from the final image to each reference image using the disparity maps and the α value. Pixels without a correspondence are labeled as invalid. This typically happens at disocclusions. The final image is generated using the color of the corresponding pixels in the reference images. The warped pixel coordinates are in general non-integer, and the color is obtained by bilinear interpolation.

We studied 4 different kinds of weights for the blending factors. First, we consider the *classic* weight $((1 - \alpha), \alpha)$ taking into account the “*minimal angular deviation*”. The second weight also fulfills this property. It is $((1 - \alpha)^2, \alpha^2)$. The third chosen weight ignores the “*minimal angular deviation*” and assigns to each image the same weight. This is done independently of the α value and the local deformation of the image transformation. The fourth weight is proportional to the local deformation of the transformation of the image, as described in [10], $|\det D\tau|^{-1}$. In our case τ is the transformation given by the disparity and we compute its deformation by finite differences. In all 4 cases, weights are normalized so that their sum is 1. If one of the two corresponding pixels is marked as invalid, we assign a 0 weight to it, and a 1 weight to the other pixel. If both corresponding pixels are invalid, we mark the pixel as invalid. If one or both of the corresponding pixels are valid, we assign to the final pixel the weighted sum of the values. Notice that at this stage some pixels of the rendered image are labeled as invalid. We explain how we handle those cases in section 3.3.

3.2. Variational Methods

Variational methods find the image minimising an energy. Those energies are usually composed two parts: the data term

TAROT	view 2		view 3		view 8		view 9		view 14		view 15	
<i>Direct Methods</i>	PSNR	DSSIM	PSNR	DSSIM	PSNR	DSSIM	PSNR	DSSIM	PSNR	DSSIM	PSNR	DSSIM
$\alpha(1 - \alpha)$	33.65	25	31.76	35	30.19	42	30.84	36	32.60	29	33.81	24
$\alpha^2(1 - \alpha)^2$	33.49	26	31.47	38	30.21	42	30.78	36	32.38	31	33.58	26
Constants	32.20	27	31.24	34	30.16	42	30.88	36	31.61	33	32.64	26
Deformation	<i>31.81</i>	29	<i>30.94</i>	35	29.96	44	30.67	37	31.42	34	32.40	27
<i>Variational Methods</i>												
$\alpha(1 - \alpha)$	33.60	25	31.65	35	29.95	45	30.45	39	31.96	33	33.10	27
Deformation ([10])	32.48	27	31.31	35	29.78	45	30.29	39	<i>31.22</i>	<i>36</i>	<i>32.33</i>	28
BRACELET	view 2		view 3		view 8		view 9		view 14		view 15	
<i>Direct Methods</i>												
$\alpha(1 - \alpha)$	36.12	14	32.87	28	33.81	24	33.69	24	33.71	23	35.67	16
$\alpha^2(1 - \alpha)^2$	36.06	15	32.66	29	33.81	24	33.71	24	33.46	25	35.48	16
Constants	34.20	20	32.50	30	33.79	24	33.67	25	33.21	26	34.59	19
Deformation	33.68	23	32.20	32	33.46	25	33.40	26	<i>32.94</i>	28	<i>34.25</i>	<i>21</i>
<i>Variational Methods</i>												
$\alpha(1 - \alpha)$	36.27	14	32.92	27	33.46	26	33.30	27	33.80	24	35.77	16
Deformation ([10])	34.52	19	32.44	30	<i>33.11</i>	27	<i>33.15</i>	27	33.06	27	34.47	20

Table 1. Numerical results of the comparison between real and rendered images for the datasets *tarot* and *bracelet*. For each view and method, the PSNR is in dB (the bigger, the better the signal) and the DSSIM scaled with 10^{-4} (the smaller, the most similar the images are). Views (2, 3, 14 and 15) are close to the reference views. Views (8 and 9) are central viewpoints. The value of the best method for each view is in bold, and the worse in italic.

and the prior (or regularizer).

$$E(u) = E_{\text{data}}(u) + \lambda E_{\text{prior}}(u). \quad (1)$$

In the Bayesian formalism, the first term can be obtained by modeling the image formation process. A common hypothesis is to consider the sensor noise as additive Gaussian. The energy of the maximum a posteriori (MAP) matches the minimum of a least squares overdetermined system. Let u be the final rendered image, v_i be the reference image number i , and the Ω_i be the definition domain of the images. m_i is a binary image indicating if a pixel in i is visible in the final image, and τ_i is the 2D transformation from the reference image i into the final image. ω_i is the weight of each pixel of image i . The data term can then be formulated as:

$$E_{\text{data}}(u) = \sum_{i=1}^n \frac{1}{2} \int_{\Omega_i} \omega_i m_i ((u \circ \tau_i) - v_i)^2. \quad (2)$$

For the variational methods, we have chosen to use two different weights. The first is the one proposed by [10] corresponding to the deformation of the transformation of the images:

$$\omega_i = |\det D\tau_i|^{-1} \quad (3)$$

The second weight is the *classic* $((1 - \alpha), \alpha)$, taking into account the “*minimal angular deviation*”. Although this second weight is not formally derived from a known generative model, the energy is well defined and can be minimized. Again, all weights are normalized so that their sum is 1.

3.3. Prior or regularizer

In the Bayesian formalism, a prior is used to constrain the solution, and to decide in cases where no information is avail-

able for some areas, or when several candidates for a solution are possible. In the direct methods, this phenomenon also arises if none of the reference images can propose information for the desired area. Note that inpainting methods [16] may also fulfil this role, but they are outside of the scope of this study.

A *classical* prior in computer vision is the total variation [17]: $E_{\text{prior}}(u) = \int_{\Gamma} |Du|$. It has the important property to provide a convex energy and to be well suited for the gradient descend minimization techniques. In the direct methods, we fill the remaining invalid pixels with a hole filling technique propagating valid information from the nearest neighbors. Its behaviour is similar to the role played by the prior [17] during the minimization, so it seems a fair choice in order to provide a reasonable comparison. We highlight that our goal is not to obtain the best possible images, but to provide a fair comparison between *direct* and *variational* methods.

3.4. Experiments

In order to compare the different methods and the impact of the blending factors in the final result we have used multiple datasets from the *Middlebury Stereo Dataset* [14] and the *Stanford Lightfield Archive* [15]. Fig. 2 shows examples of the used images. The first one proposes very Lambertian scenes (color does not depend on the viewing angle). Each scene has 6 aligned views. We use the most distant ones (1st and 6th) to generate the other 4. The second database provides more challenging scenes, including highly specular reflections, transparencies and inter-reflections. Each dataset has 256 views per scene arranged in a regular 16x16 grid. We have selected the central row (the 8th) and used its most

AMETHYST	view 2		view 3		view 8		view 9		view 14		view 15	
<i>Direct Methods</i>	PSNR	DSSIM	PSNR	DSSIM	PSNR	DSSIM	PSNR	DSSIM	PSNR	DSSIM	PSNR	DSSIM
$\alpha(1 - \alpha)$	33.84	1247	32.67	1265	30.66	1292	30.88	1299	33.60	1305	34.86	1305
$\alpha^2(1 - \alpha)^2$	33.81	1255	32.61	1282	30.69	1294	30.84	1300	33.42	1325	34.79	1313
Constants	32.16	1218	31.62	1239	30.63	1288	30.91	1294	32.67	1273	33.13	1274
Deformation	31.56	1239	31.07	1259	30.27	1307	30.54	1313	32.00	1293	32.36	1296
<i>Variational Methods</i>												
$\alpha(1 - \alpha)$	32.80	1093	31.39	1116	30.30	1160	30.66	1168	33.77	1142	34.98	1141
Deformation ([10])	31.87	1107	30.96	1126	30.27	1163	30.63	1170	33.01	1150	33.68	1152

Table 2. Numerical results for the *amethyst* dataset. Same measures as in Table 1 are displayed.

distant viewpoints (1st and 16th) to generate the other 14. This way we can compare the original images with the generated ones. To do so we have used two state of the art measures. The *Peak Signal to Noise Ratio* (PSNR), in dB, the bigger, the better is the signal. And the *Structural SIMilarity* (SSIM) developed to measure the visual similarity between two images. We report results with a distance based on SSIM: $DSSIM = \frac{1-SSIM}{2}$, having no units. The smaller, the more similar are the images.

In our study we have used standard methods for the disparity maps computation [18] and [19]. Disparity maps obtained with [19] are not dense but we have completed them using the hole filling technique in [18]. The λ parameter in eq. 1 was empirically set to 0.15 for all experiments.

3.5. Results

We present the obtained results for the datasets *tarot*, *bracelet* and *amethyst*, using the disparities computed with SGBM [19]. Results on *Middlebury Dataset* (*aloe*, ...) are very similar to those from *tarot* and are not presented. Results obtained with disparities from [18] are very close to the obtained with [19] and are not presented.

Input images are coded in sRGB space. It is advised to convert them into RGB-linear when operating with pixel values. We tested using the sRGB and the linear versions and very similar results were obtained. Running times for direct methods are around 30ms for 1024x1024 images (real-time). Variational methods were solved using a GPU implementation of FISTA [20, 21]. Running time was about 1s for 1024x1024 images (not real-time).

The dataset *tarot* has a crystal ball with transparencies, violating the Lambertian model. However, colors on the rest of the scene do not change from one view to the other. The dataset *bracelet* has some specular highlights on the metal, but the color difference between the left and right views is very small. The dataset *amethyst* is more complex. It has inter-reflections and highlight effects.

In Tab. 1, the difference between the methods for a fixed scene is small. PSNR and DSSIM values are very close. In *bracelet*, results of view 2 and 15 are slightly better when taking into account the “*minimal angular deviation*” ($\alpha, (1 - \alpha)$) and ($\alpha^2 (1 - \alpha)^2$) both in direct and variational methods.

This was expected as those methods are capable of better rendering the highlights. However the improvement quickly dissipates as we render the next views (3 or 14). No significant difference between the direct and variational methods can be reported.

Notice how PSNR values on Tab. 1 (*tarot* and *bracelet*) and Tab. 2 (*amethyst*) are on the same order of magnitude; but DSSIM values are higher on Tab. 2 than on Tab. 1. We believe that this difference appears due to the fact that the scene is more complex, together with a greater capability of DSSIM to measure the visual quality. Again, for each view, tendencies are similar as in Tab.1. However we report a significant difference between the DSSIM values of direct and variational methods. For complex scenes, variational methods are capable of better reconstructing the structure of the image.

4. DISCUSSION AND CONCLUSION

In this paper we have presented a study of the influence of the blending factors and the method in the result of the viewpoint interpolation techniques. We have compared several blending factors and two groups of methods, direct and variational. Conducted experiments show that for Lambertian scenes, the choice of the blending factors has almost no impact. Moreover, results obtained with direct or variational methods are equivalent, so direct methods should be preferred thanks to their simplicity. In the non-Lambertian scenes, the choice of the blending factors which account for the “*minimal angular deviation*” property produce slightly better results, both with direct and with variational methods, although the improvements are only visible when rendering new images that are close to a reference view. However, the variational methods are capable of better render the structure of the image, obtaining improved results. In this case variational methods with blending factors which take into account “*minimal angular deviation*” should be preferred.

In future work, it would be interesting to continue this study for more general camera and synthesized view configurations: multiple input views and general position of the reference and final views.

5. REFERENCES

- [1] Sammy Rogmans, Jiangbo Lu, Philippe Bekaert, and Gauthier Lafruit, “Real-time stereo-based view synthesis algorithms: A unified framework and evaluation on commodity GPUs,” *Signal Processing: Image Communication*, vol. 24, no. 1-2, pp. 49–64, 2009, Special issue on advances in three-dimensional television and video.
- [2] Frédéric Devernay and Sylvain Duchêne, “New view synthesis for stereo cinema by hybrid disparity remapping,” in *International Conference on Image Processing (ICIP)*, Hong Kong, Sept. 2010, pp. 5–8.
- [3] Daniel Scharstein and Richard Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *Int. J. Comput. Vision*, vol. 47, no. 1-3, pp. 7–42, Apr. 2002.
- [4] Heung-Yeung Shum, Shing-Chouw Chan, and Sing Bing Kang, *Image-based rendering*, Springer, 2007.
- [5] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen, “Unstructured Lumigraph rendering,” in *Proc. SIGGRAPH*. ACM, 2001, pp. 425–432.
- [6] Joon Hong Park and Hyun Wook Park, “Fast view interpolation of stereo images using image gradient and disparity triangulation,” *Signal Processing: Image Communication*, vol. 18, no. 5, pp. 401–416, 2003.
- [7] Aljoscha Smolic, Karsten Muller, Kristina Dix, Philipp Merkle, Peter Kauff, and Thomas Wiegand, “Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems,” in *Proc. ICIP*. IEEE, 2008, pp. 2448–2451.
- [8] Frédéric Devernay and Adrian Ramos Peon, “Novel view synthesis for stereoscopic cinema: detecting and removing artifacts,” in *Proceedings of the 1st international workshop on 3D video processing*. ACM, 2010, pp. 25–30.
- [9] Maxime Lhuillier and Long Quan, “Image interpolation by joint view triangulation,” in *CVPR*. IEEE, 1999, vol. 2.
- [10] Sven Wanner and Bastian Goldluecke, “Spatial and angular variational super-resolution of 4D light fields,” in *Proc. ECCV*, pp. 608–621. Springer, 2012.
- [11] Tao Yan, Rynson WH Lau, Yun Xu, and Liusheng Huang, “Depth mapping for stereoscopic videos,” *IJCV*, pp. 1–15, 2013.
- [12] Manuel Lang, Alexander Hornung, Oliver Wang, Steven Poulakos, Aljoscha Smolic, and Markus Gross, “Non-linear disparity mapping for stereoscopic 3d,” *TOG*, vol. 29, no. 4, pp. 75, 2010.
- [13] Ran Gal, Olga Sorkine, and Daniel Cohen-Or, “Feature-aware texturing,” in *Proceedings of the 17th Eurographics conference on Rendering Techniques*. Eurographics Association, 2006, pp. 297–303.
- [14] Heiko Hirschmuller and Daniel Scharstein, “Evaluation of cost functions for stereo matching,” in *CVPR*. IEEE, 2007, pp. 1–8.
- [15] V. Vaish and A. Adams, “The (New) Stanford Light Field Archive,” <http://lightfield.stanford.edu>, 2008.
- [16] C. Guillemot and O. Le Meur, “Image inpainting : Overview and recent advances,” *Signal Processing Magazine, IEEE*, vol. 31, no. 1, pp. 127–144, Jan 2014.
- [17] Leonid I Rudin, Stanley Osher, and Emad Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.
- [18] Mikhail Sizintsev and Richard P Wildes, “Coarse-to-fine stereo vision with accurate 3d boundaries,” *Image and Vision Computing*, vol. 28, no. 3, pp. 352–366, 2010.
- [19] Heiko Hirschmuller, “Accurate and efficient stereo processing by semi-global matching and mutual information,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 2, pp. 807–814.
- [20] Amir Beck and Marc Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIIMS*, vol. 2, no. 1, pp. 183–202, 2009.
- [21] B. Goldluecke, E. Strelakovsky, and D. Cremers, “The natural vectorial total variation which arises from geometric measure theory,” *SIIMS*, 2012.