Image Completion for View Synthesis Using Markov Random Fields and Efficient Belief Propagation

Julian Habigt and Klaus Diepold

View synthesis is a process for generating novel views from a scene which has been recorded with a 3-D camera setup. It has important applications in 3-D post-production and 2-D to 3-D conversion. However, a central problem in the generation of novel views lies in the handling of disocclusions. Background content, which was occluded in the original view, may become unveiled in the synthesized view. This leads to missing information in the generated view which has to be filled in a visually plausible manner. We present an inpainting algorithm for disocclusion filling in synthesized views based on Markov random fields and efficient belief propagation. We compare the result to two state-of-the-art algorithms and demonstrate a significant improvement in image quality.

Published in:	Proc. IEEE International Conference on Image Processing (ICIP 2013)
Date of Conference:	15-18 Sep. 2013
Pages:	2131 - 2134
Publisher:	IEEE
DOI:	10.1109/ICIP.2013.6738439
WWW:	http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6738439

IMAGE COMPLETION FOR VIEW SYNTHESIS USING MARKOV RANDOM FIELDS AND EFFICIENT BELIEF PROPAGATION

Julian Habigt and Klaus Diepold

Technische Universität München, Institute for Data Processing,
Arcisstr. 21, 80333 Munich, Germany
jh@tum.de, kldi@tum.de

ABSTRACT

View synthesis is a process for generating novel views from a scene which has been recorded with a 3-D camera setup. It has important applications in 3-D post-production and 2-D to 3-D conversion. However, a central problem in the generation of novel views lies in the handling of disocclusions. Background content, which was occluded in the original view, may become unveiled in the synthesized view. This leads to missing information in the generated view which has to be filled in a visually plausible manner. We present an inpainting algorithm for disocclusion filling in synthesized views based on Markov random fields and efficient belief propagation. We compare the result to two state-of-the-art algorithms and demonstrate a significant improvement in image quality.

***Index Terms*—** DIBR, View Synthesis, Inpainting, Hole-Filling, MRF

1. INTRODUCTION

View synthesis is an important tool for the generation of content for 3-D television [1]. In traditional stereoscopic setups, the scene is filmed with two cameras and then reproduced on a 3-D television screen, which can produce two separate pictures for the eyes of the viewer. This approach has several drawbacks. The baseline of the setup, i.e., the distance between the two cameras, has to be fixed during the production of the 3-D content and cannot be changed afterwards. When this content is shown on screens of different sizes, e.g., in a cinema or on a mobile device, the common baseline leads to an incorrect reproduction of the perceived depth of the scene [2]. Furthermore, current autostereoscopic displays, i.e., displays which don't require the viewer to wear glasses to see 3-D content need a much higher number of views of the same scene, e.g., 28 or more. It is therefore necessary to be able to generate virtual views of a scene once the scene has been recorded.

One technique to generate such virtual views which has gained momentum in recent years is called depth image-based rendering. There, the virtual view is generated from the image of one or more cameras and corresponding depth maps. A central problem in the generation of novel views lies in the handling of disocclusions. Background content, which was occluded in the original view by objects that were closer to the camera, may become unveiled in the virtual view. In a setup with two or more cameras, these so-called disocclusions may be partially filled with content from another camera, yet some disocclusions usually still remain [3]. Even more challenging, when there is only one view and a corresponding depth map, there is no other other information available to fill the holes in the resulting view and the holes may have to be filled with synthetically generated content. This is, for example, the case in 2-D to 3-D conversion or in proposed transmission schemes with only one texture and depth map such as ATTEST [4].

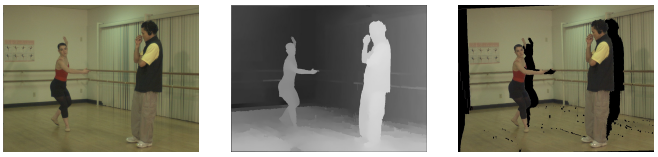


Fig. 1: Given an input image and a corresponding depth-map, we can generate virtual views, a process which is called depth image-based rendering (DIBR). However, disocclusions appear where background is unveiled which was occluded by foreground objects in the original view.

A number of solutions have been proposed for this problem. One way to avoid it entirely would be the modification of the depth-map. Disocclusions appear at regions in the image where there is a steep gradient in the depth-map, i.e., at the borders of foreground objects. Zhang et al. [5] proposed a technique where the depth-map is filtered to remove these steep gradients. The result is a virtual view which doesn't contain any holes, at the cost of an incorrect reproduction of the depth which may lead to visible errors [6]. Another way is the use of so-called inpainting techniques. Inpainting describes a process where holes in images are filled with synthesized content in a visually plausible manner, so that the viewer doesn't recognize that the content has been generated artificially. Recently, there has been quite extensive research on the adaptation of the inpainting algorithm by Criminisi et al. [7] for disocclusion filling. Criminisi's algorithm can be categorized into the group of so-called exemplar-based techniques, i.e., the algorithm uses patches of the image itself and copies these into the hole, thus exploiting the redundancy of natural images. Criminisi discovered that the order in which this filling process is executed determines the quality of the output image. He therefore introduced a confidence and a priority term with the intention to steer the filling process into the direction of isophotes, i.e., lines with constant luminance. However, using this algorithm directly for disocclusion filling in the context of view synthesis leads to very poor results [8]. Therefore, several modifications have been proposed.

Oh et al. [9] explicitly modified the boundaries of the holes to only incorporate background pixels. Daribo and Saito [8] proposed a depth-based modification to Criminisi's priority term to prioritize background pixels over foreground pixels. Gautier et al. [10] replaced the color gradient in the priority term with a structure tensor based on the color of the texture and the structure of the depth map.

Criminisi's inpainting technique is a greedy algorithm, i.e., once a patch has been copied into the hole, it won't be changed regardless of the patches that follow in its neighbourhood. Komodakis and Tzir-

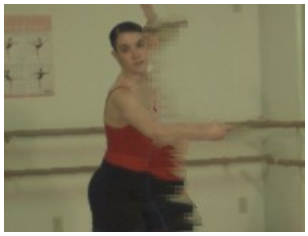
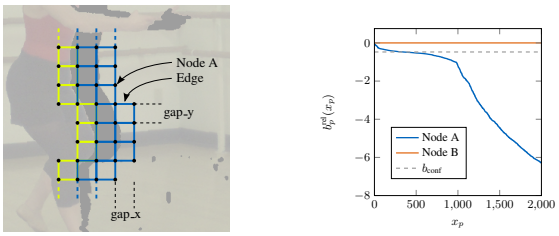


Fig. 2: Unfortunately, Komodakis and Tziritas's algorithm cannot be used directly to fill disocclusions. The most obvious problem is bleeding of foreground objects into the background.



(a) (b)

Fig. 3: Figure (a) shows a schematic visualization of the distribution of the nodes over the disocclusion. Nodes marked with yellow edges lie over foreground content and will be assigned a node potential $V_p = 0$. Figure (b) showcases the relative beliefs of a node at the border of the hole (Node A) and of any interior node (Node B) before the message passing step.

itas [11] recognized this as a potential drawback and therefore introduced an inpainting algorithm based on the solution of a Markov random field. They demonstrate that this technique has the potential to significantly outperform the method of Criminisi in terms of visual quality of the inpainting result. In this contribution, we therefore propose an adaptation of the algorithm of Komodakis and Tziritas for disocclusion filling for view synthesis. The results compare favorably to the state of the art. We start by shortly reviewing the algorithm of Komodakis and Tziritas, introduce our extensions to make it applicable to view synthesis, show some of the results and compare it to the state of the art before concluding this paper.

2. ALGORITHM

2.1. Komodakis and Tziritas's algorithm

To make this paper self-contained, we will start by a brief summary of the algorithm of Komodakis and Tziritas so that we can introduce our extensions that make it applicable to view synthesis. However, for the sake of brevity, we would like to refer the reader to [11] for further details. For clarity, we try to adapt the notation of [11] as closely as possible.

The hole-filling task is treated as a discrete, global optimization problem with a well-defined objective function. It therefore doesn't require any ad-hoc heuristics, such as the isophote continuation in Criminisi's algorithm, which may not be adequate in a general setting. First, we have to separate our image I_0 into a source region S and a target region T, i.e., the hole(s) to be filled. In the view-synthesis problem, the location of the holes is determined by the scene geometry, and during the process of mapping the texture of the original view to the virtual view, we can simultaneously generate a mask which specifies the location of the holes. The results of this warping process can be seen in Figure 1.

The image is then partitioned into small, overlapping patches of size $w \times h$ with a spacing of gap_x and gap_y, respectively. Note that $\text{gap}_x < w$ and $\text{gap}_y < h$. The goal of the inpainting algorithm is then to find suitable patches from S which can be filled into the holes T. To this end, Komodakis and Tziritas proposed a Markov network which consists of nodes ν at the positions of the patches inside and at the border of the hole. Each of these nodes has a set of labels associated with it which comprises candidate patches from S to be inserted at the position of the node. The association of any of the labels to a node p incurs specific costs which are defined as the node potential

$$V_{I,p}(x_p) = \sum_{dp \in \left[-\frac{w}{2}\,\frac{w}{2}\right] \times \left[-\frac{h}{2}\,\frac{h}{2}\right]} \mathcal{M}(p+dp)(I_0(p+dp) - I_0(x_p+dp))^2,$$

which describes how well the patch x_p matches any available content from S, and a pairwise potential $V_{p,q}$, i.e., how well the patch matches the other patches in its 4-connected neighbourhood. $\mathcal{M}$ denotes a mask which is zero inside T, 1 else. The goal of the optimization problem is then to minimize the total energy of the MRF

$$\mathcal{F}(\hat{x}) = \sum_{p \in \nu} V_p(\hat{x}_p) + \sum_{(p,q) \in \varepsilon} V_{pq}(\hat{x}_p, \hat{x}_q),$$

for which Komodakis and Tziritas proposed a *priority-belief propagation* algorithm. In belief propagation, messages are exchanged along the edges ε of connected nodes about the confidence in the association of a patch to a neighbouring node, which then in turn defines the belief $b_p(x_p)$ each node has in its set of labels. As the number of possible patches is quite high in the setting of image completion, the computational cost of this BP-algorithm would be prohibitive. Komodakis and Tziritas therefore added a method called *dynamic label pruning* based on priority. If a node has only a small set of labels in which it has a belief higher than a given confidence threshold, i.e., $b_p^{\text{rel}}(x_p) \geq b_{\text{conf}}$ with $b_p^{\text{rel}}(x_p) = b_p(x_p) - b_p^{\max}(x_p)$, it will be assigned a high priority, that means it is quite confident about the assignment of its patch. On the other hand, if a node has similar beliefs in all of its labels, it may be considered indetermined and will be given a low priority. Nodes with high priority will be the ones to first get rid of all labels in which they have a low belief and then send efficient messages. Figure 3b shows the distribution of the relative beliefs of a node with high priority which is usually located at the border of the hole. An interior node has a low priority as its node potential is zero and therefore has the same belief in all of its labels.

2.2. Extensions for view synthesis

Komodakis and Tziritas's algorithm is not directly applicable to the disocclusion problem in view synthesis as a naïve application leads to very poor results, as shown in Figure 2. The most obvious problem is that there occurs bleeding of foreground objects into the background, which should be avoided. We therefore present our extensions which will deal with this problem. As stated in the introduction, disocclusions occur at steep depth gradients, where there is a jump between a foreground object to the background of a scene. When we move the virtual camera to the right, the disocclusions will appear on the right side of foreground objects. We therefore adapt the idea of [10] and others to steer the filling process into the opposite direction of the camera movement. In our setting, we achieve this by modifying the node potential of all nodes that are on the side of the disocclusion opposite to the camera movement, e.g., on the

left side. These nodes, in Figure 3a marked as yellow, are given a node potential $V_p = 0$. The algorithm thereby treats these just like interior nodes and they will get the lowest priority. As the MRF now doesn't have any support on the left side of the hole, the inpainting task has become somewhat similar to the texture synthesis task described in [11]. It is therefore necessary to introduce another term $V^0_{pq}(x_p, x_q) = w_0$ if $x_p - x_q \neq p - q$, $V^0_{pq}(x_p, x_q) = 0$, else, to the cost function which enforces the coherence of the image by penalizing the filling of non-adjacent patches.

Furthermore, we modify the node potential

$$V_p(x_p) = V_{I,p}(x_p) + \lambda_D V_{D,p}(x_p)$$

and the pairwise potential to not only accommodate for visual similarity between neighbouring nodes but also for similarity in depth. To this end, we add another term to both potentials which calculates the SSD

$$V_{D,p}(x_p) = \sum_{dp \in \left[-\frac{w}{2}\ \frac{w}{2}\right] \times \left[-\frac{h}{2}\ \frac{h}{2}\right]} \mathcal{M}(p+dp)\left(\mathcal{D}_0(p+dp) - \mathcal{D}_0(x_p+dp)\right)^2$$

in the depth map $\mathcal{D}$, weighted by a factor λ_D. Thereby, we make sure that candidate patches are selected from similar depth ranges as the nodes which ensures consistency of the image and also improves the efficiency of the algorithm because it dramatically reduces the number of contemplable labels for each node.

3. EVALUATION

To evaluate the performance of our algorithm, we use the well-known Multiview Video-plus-Depth sequence *Ballet* from Microsoft Research [12] because of its large baseline and because it allows us to make a fair comparison with two state-of-the-art algorithms [8, 10]. For our evaluation we take the view from Camera No. 5 and create a virtual view which would be seen from Camera No. 4. We can therefore use the image from Camera No. 4 as a ground truth reference. We use the MPEG View Synthesis Reference Software (VSRS) [13] in version 3.5 to generate the virtual view and use our algorithm to fill the disocclusions. As an objective measure for the quality of the inpainting result, we use SSIM. For completeness, we have also included the PSNR values, even though we think that PSNR is hardly a suitable measure to judge the quality of an inpainting algorithm. We also provide both measures for the regions which have been inpainted, only. The parameters of our algorithm have been chosen on the basis of the recommendations in [11] and therefore weren't specifically tuned to the sequence; with the exception of the newly introduced parameter λ_D which was set to 3 to accommodate for the difference in the number of channels between the image and the depth map. The results can be seen in Table 1 and in Figure 4.

Table 1: Objective evaluation of the inpainting result

	[8]	[10]	proposed
PSNR$_Y$ [dB]	30.3	31.4	**33.2**
PSNR$_Y$ holes only [dB]	24.2	24.0	**26.2**
SSIM	0.87	0.88	**0.93**
SSIM holes only	0.68	0.69	**0.73**

4. CONCLUSION & ACKNOWLEDGEMENTS

We have presented a new algorithm for disocclusion handling in view synthesis. A global optimization approach on Markov random fields incorporating the information of the depth map not only leads to consistent inpainting results but also to a higher algorithmic efficiency due to rigorous label pruning based on depth range. Objective evaluation on a standard dataset shows a significant improvement in image quality compared to the state of the art.

We would like to thank Josselin Gautier for providing the software and results for his algorithm and the Microsoft Research team for providing the *Ballet* data set.

5. REFERENCES

[1] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Müller, and M. Lang, "Three-dimensional video postproduction and processing," *Proc. IEEE*, vol. 99, no. 4, pp. 607–625, Apr. 2011.

[2] R. Held and M. Banks, "Misperceptions in Stereoscopic Displays: A Vision Science Perspective," in *Proc. 5th Symp. Applied Perception Graphics and Visualization*, Los Angeles, CA, 2008, pp. 23–32.

[3] S.E. Chen and L. Williams, "View interpolation for image synthesis," in *Proc. 20th Annu. Conf. and Exhibition Computer Graphics and Interactive Techniques*, Anaheim, CA, 1993, pp. 279–288.

[4] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *Proc. SPIE*, San Jose, CA, 2004, vol. 5291, pp. 93–104.

[5] L. Zhang and W.J. Tam, "Stereoscopic Image Generation Based on Depth Images for 3D TV," *IEEE Trans. Broadcast.*, vol. 51, no. 2, pp. 191–199, June 2005.

[6] I. Daribo and H. Saito, "Bilateral depth-discontinuity filter for novel view synthesis," in *Proc. IEEE Int. Workshop Multimedia Signal Processing*, Saint-Malo, France, 2010, pp. 145–149.

[7] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sept. 2004.

[8] I. Daribo and H. Saito, "A Novel Inpainting-Based Layered Depth Video for 3DTV," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 533–541, June 2011.

[9] K. Oh, S. Yea, and Y. Ho, "Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-D video," in *Proc. 27th Picture Coding Symp.*, Chicago, IL, 2009, pp. 1–4.

[10] J. Gautier, O. Le Meur, and C. Guillemot, "Depth-based image completion for view synthesis," in *Proc. 3DTV Conf. The True Vision Capture Transmission and Display of 3D Video*, Antalya, Turkey, 2011, pp. 1–4.

[11] N. Komodakis and G. Tziritas, "Image completion using efficient belief propagation via priority scheduling and dynamic pruning," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2649–2661, Nov. 2007.

[12] C.L. Zitnick, S.B. Kang, and M. Uyttendaele, "High-quality video view interpolation using a layered representation," *ACM Trans. Graphics*, vol. 23, no. 3, pp. 600–608, Aug. 2004.

[13] M. Tanimoto, T. Fujii, and K. Suzuki, "View synthesis algorithm in View Synthesis Reference Software 2.0 (VSRS2.0)," Tech. Rep., ISO/IEC JTC1/SC29/WG11 M16090, Lausanne, Switzerland, Feb. 2008.

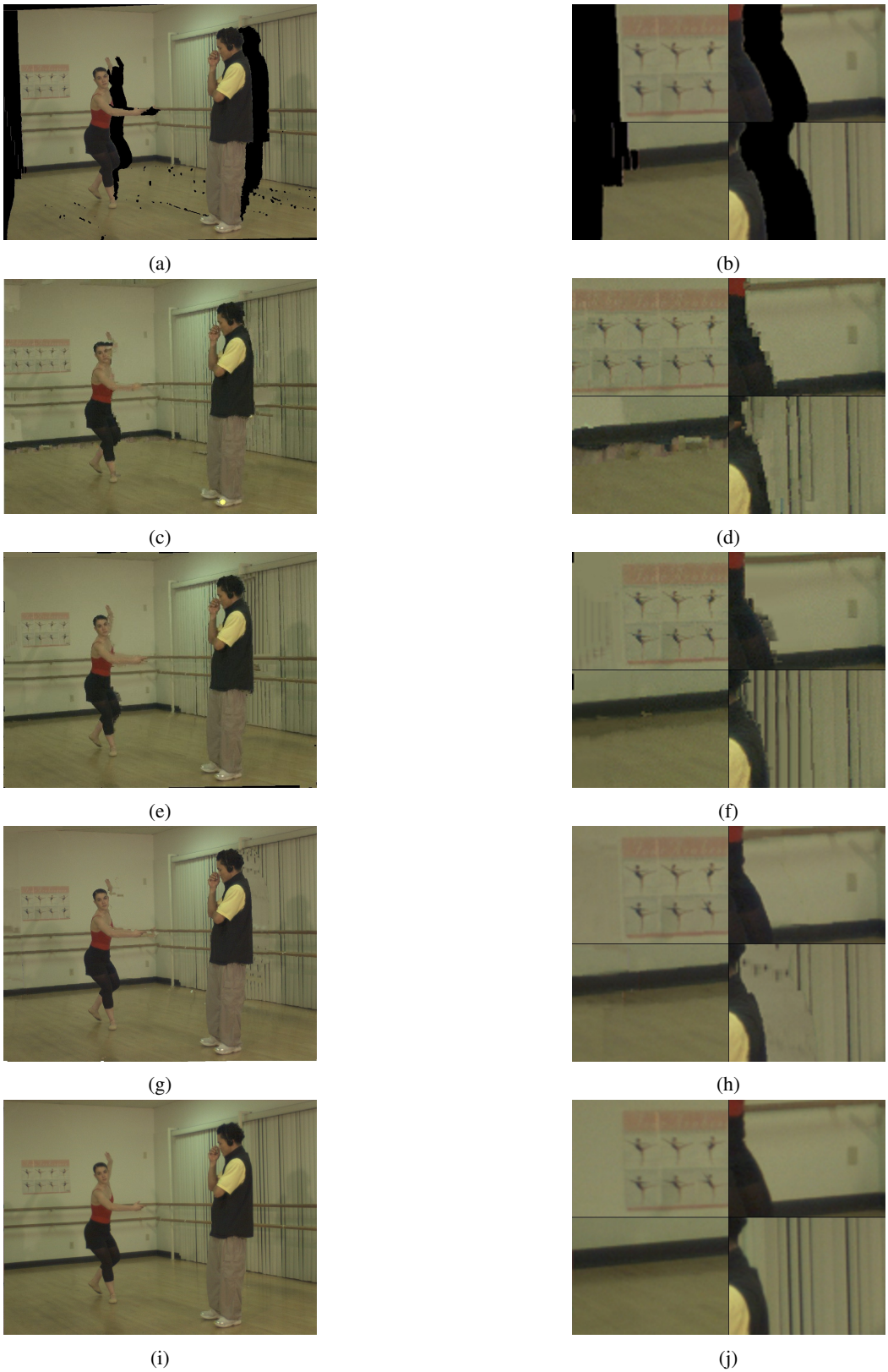


Fig. 4: Inpainting results for the first frame of the *Ballet* sequence of two state-of-the-art algorithms and our proposed algorithm. (a), (b) Synthesized View; (c), (d) Result of Daribo's method [8]; (e), (f) Result of Gautier's method [10]; (g), (h) Proposed method; (i), (j) Ground truth (Camera No. 4).