STATISTICAL METHODS FOR INTER-VIEW DEPTH ENHANCEMENT

Pravin Kumar Rana, Jalil Taghia, and Markus Flierl

School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

ABSTRACT

This paper briefly presents and evaluates recent advances in statistical methods for improving inter-view inconsistency in multiview depth imagery. View synthesis is vital in free-viewpoint television in order to allow viewers to move freely in a dynamic scene. Here, depth image-based rendering plays a pivotal role by synthesizing an arbitrary number of novel views by using a subset of captured views and corresponding depth maps only. Usually, each depth map is estimated individually at different viewpoints by stereo matching and, hence, shows lack of inter-view consistency. This lack of consistency affects the quality of view synthesis negatively. This paper discusses two different approaches to enhance the inter-view depth consistency. The first one uses generative models based on multiview color and depth classification to assign a probabilistic weight to each depth pixel. The weighted depth pixels are utilized to enhance depth maps. The second one performs inter-view consistency testing in depth difference space to enhance the depth maps at multiple viewpoints. We comparatively evaluate these two methods and discuss their pros and cons for future work.

Index Terms — Multiview depth maps, depth map enhancement, inter-view consistency, variational Bayesian inference.

1. INTRODUCTION

Free-viewpoint television (FTV) is an emerging visual media application that will allow viewers to have a dynamic natural depth impression while freely choosing their viewpoint to watch the telecast of real world scenes [1]. FTV is able to display a large number of views from a range of different perspectives to have a seamless free-view experience of natural 3-D scenes. The availability of low-cost digital cameras permits us to record easily multiview video (MVV) for FTV. MVV is a set of videos recorded by an array of video cameras that capture a dynamic natural scene from many viewpoints simultaneously. However, this entails a demand for high camera density around the natural scene and a need of high storage and transmission capacity for the vast amount of captured imagery at multiple viewpoints [2]. These requirements may be greatly reduced by using geometry information of 3-D scenes, for example, depth maps [2]. The reason is that given a small set of MVV and its corresponding set of multiview depth (MVD) maps, an arbitrary number of views can be generated by using depth image-based rendering (DIBR) [3].

Depth maps are quantized gray scale images where each pixel in the depth map represents the shortest distance between the corresponding object point in 3-D world and the given camera plane. The gray value zero represents the farthest object points and the largest value the closest. The estimation of depth maps by using camera captured images of natural dynamic scenes is a challenging task. For example, stereo matching algorithms obtain depth maps by accurately establishing stereo correspondences between two or more images captured at different viewpoints with the help of a matching criterion. The accuracy of stereo matching and the resulting depth estimates are affected by many physical factors as mention in [4]. Although, a number of optimization techniques are used to refine depth estimates [4], the depth maps usually lack temporal consistency, since depth estimation does not exploit temporal similarities among neighbouring frames. This results in temporal inconsistency and flickering. Many methods have been proposed to repair temporal inconsistencies in MVD imagery (e.g., [5]).

To obtain overall geometry information about a natural scene, depth maps may be independently estimated using stereo matching at multiple viewpoints. Consequently, the resulting depth maps at different viewpoints normally lack inter-view consistency. For example, the depth value of a unique 3D point is the same in all depth maps for a 1D-parallel camera array setting, but located at different positions in the maps. Therefore, depth observations at different viewpoints should be consistent, and related areas in different viewpoints should show the same depth values, but shifted. In the 1D-parallel camera array setting, all optical centers of the cameras are parallel to each other and all corresponding rotation matrices are identical. DIBR based view synthesis may use multiple views from different viewpoints and their corresponding depth maps. Hence, inter-view depth inconsistencies affect the quality of synthesized views negatively, and FTV users feel the resulting visual discomfort. The inter-view depth consistency is also critical for FTV data representations (e.g. [6]).

With recent MPEG activities on 3D video [7], the problem of inter-view depth inconsistency has become an active research topic (e.g., [8, 9]). In [10, 11], inter-view depth consistency testing (IVDCT) of MVD imagery is described, where the resulting consistency information is utilized to improve the quality of view synthesis. The framework in [12, 13] exploits the conditional dependency between color and depth pixels to enhance the inter-view depth consistency. The framework exploits the dependency through color and depth classification of the MVV imagery by utilizing a generative model where the model parameters are estimated in a Bayesian framework by variational inference (VI) [14]. In this paper, we briefly describe the recent improvements of the IVDCT as presented in [10, 11], and that of the probabilistic framework as presented in [13]. We also present a comparative experimental study between the two enhancement methods. This allows us to discuss the benefits and limitations of the specific algorithms.

In the following, we outline the inference-based framework in Section 2 and the IVDCT method in Section 3. Section 4 discusses the comparative experimental results. Conclusions are given in Section 5.

2. PROBABILISTIC INTER-VIEW DEPTH ENHANCEMENT

Our goal is to improve the inter-view consistency among depth maps at multiple viewpoints, and with that, to enhance the quality of FTV. For this, we proposed in earlier work a general model-

978-1-4799-4758-4/14/$31.00 ©2014 IEEE

based framework in [12, 13] which improves depth maps at their respective viewpoints by utilizing the color information from in the corresponding view imagery. The idea to improve stereo matching results by using color information has been investigated by many researchers (e.g., [15, 16]). Our initial framework [12] consists of two steps: multiview color classification and multiview depth classification. First, the framework performs a color classification of the concatenated view imagery by utilizing a generative model based on a Gaussian mixture model (GMM). The model parameters are estimated in a Bayesian framework by variational inference (VI) [14]. Second, for each resulting color cluster, we classify the corresponding depth values from multiple viewpoints. Finally, multiple depth levels are assigned to individual sub-clusters for depth enhancement at multiple viewpoints. In [13], we improve the performance of [12] by utilizing VI for Dirichlet mixture models (VI-DMM) to perform color classification in the xyz chromaticity space of the view imagery, and by using mean-shift clustering for depth subclassification. The choice of the Dirichlet distribution is motivated by two facts. First, a normalized vector in the xyz chromaticity space has nonnegative elements and its l_1 norm equals to one. These properties fit the definition of the Dirichlet distribution nicely [14, 17]. Second, VI-DMM reduces the model complexity significantly when compared to VI-GMM. The choice of xyz chromaticity space also makes the procedure insensitive to the absolute luminance.

Although, VI-DMM reduces model complexity, high computational demand is still a concern in [13]. Therefore, in contrast to [12, 13], we use the concept of superpixels instead of image pixels for the color classification step [18]. A superpixel is a group of perceptually meaningful and homogeneous neighboring image pixels [19]. Superpixels capture image redundancy and help to reduce the number of feature vectors for color classification. The use of superpixels also reduces significantly the overall computational complexity of the framework, when compared to [12, 13]. This allows us to propose a fully probabilistic multiview depth enhancement method in [18], where depth subclassification is also performed by VI-GMM, which is not supported in [12, 13].

With Bayesian inference to model learning, we use both our prior knowledge (by assigning proper prior distributions) and given data to estimate the posterior. In our work, Bayesian learning is carried out using variational inference. As we learn the model in a fully Bayesian inference, we have a set of posterior probabilities, also known as responsibilities. They determine the contribution of data points when explaining the current data. The use of the fully Bayesian inference for depth sub-clustering provides us a way to gain insight about the inter-view depth inconsistencies at multiple viewpoints. At a later stage, it helps us to improve consistency by using the resulting responsibilities as probabilistic weights for depth pixels.

After obtaining the color clusters for a given multiview image set, we exploit the conditional dependency between color and depth. For this, depth images from the given viewpoints are concatenated to a single depth map. This single concatenated depth map is such that for each color pixel there is an associated depth value. Therefore, the structure of the color clusters can be imposed on the depth pixels. However, the resulting depth pixel partition does not lead to homogeneous depth clusters. The members of a given color cluster have similar colors, whereas the members of an imposed depth cluster may have distinct depth values. For example, as foreground and background object points may have similar colors, foreground object points have different depth values when compared to background object points. An object point with a given color which is visible from all viewpoints should have the same depth value in all depth maps. However, such points usually have different depth values in the imposed cluster due to inter-view inconsistencies. This ambiguity is the motivation for the further sub-clustering of each imposed depth cluster.

Our main assumption is that the inconsistent depth values for an observed object point at multiple viewpoints will be assigned lower responsibilities when compared to consistent ones. Further, to capture the inter-view relation of depth pixels, we emphasize their positions as well. For this, we propose a feature vector which consists of the depth pixel value and its location information from the corresponding viewpoint. As the discrete location values in the feature vector are sampled from a continuous distribution with quantization noise, we model the feature vector by a mixture of multivariate Gaussian distributions where the model parameters are estimated by Bayesian inference [14]. This gives a responsibility value for each depth pixel. It is basically the probability that the associated depth value is generated from the specified depth cluster. Thus, we assign each depth pixel to a depth sub-cluster which gives the largest probability. Finally, we replace the depth values in each depth sub-cluster by the corresponding responsibility-weighted mean. Here, we use the largest responsibility of the depth pixel. Note that the arithmetic mean of all depth values within a sub-cluster would be more sensitive to depth inconsistency and noise.

3. INTER-VIEW DEPTH CONSISTENCY TESTING AND ENHANCEMENT

In the previous section, we used a probabilistic model for the depth values. For this approach, we design a special statistical test to assess the consistency among the inter-view depth values. For given k viewpoints, we first create the k depth hypotheses by warping depth maps from k viewpoints to a single viewpoint, say the principal viewpoint p. We use 3D warping [3] to achieve spatial alignment. For every principal pixel in the principal viewpoint image, we define a loop difference vector $\mathbf{\Delta}$ by using all k depth hypotheses as

$$\mathbf{\Delta} = [\Delta_{12}, \Delta_{23}, \dots, \Delta_{k1}]^{\mathsf{T}} \in \mathbb{R}^k, \tag{1}$$

where $\Delta_{ij} = \hat{d}_p(i; x, y) - \hat{d}_p(j; x, y)$ is the inter-view depth difference between warped depth values $\hat{d}_p(i; x, y)$ and $\hat{d}_p(j; x, y)$ at the viewpoint p from view i and j respectively, where $i, j = 1, \dots, k$. Note, $\mathbf{\Delta}$ satisfies the *zero-sum constraint*, $\mathbf{1}^{\mathsf{T}}\mathbf{\Delta} = 0$, for any principal pixel. Here $\mathbf{1}$ is the k-dimensional vector with each element equal to one. With that, we can define the energy of the loop difference vector

$$E_k(\mathbf{\Delta}) = \mathbf{\Delta}^T \mathbf{\Delta}. \tag{2}$$

This loop energy captures the individual inter-view depth differences. It is only zero if all k inter-view depth differences are zero.

Now, we test the inter-view depth consistency with respect to an *inter-view consistency threshold* ϑ. If $\mathrm{E}_k(\mathbf{\Delta}) \leq \vartheta$, we accept all associated k depth hypotheses as consistent at the principal pixel. Subsequently, we assume that all the corresponding depth pixels have a consistent description and relate to the same object point in 3-D world. Using this information and the perspective projection, the corresponding depth values in the reference depth maps can be used to determine an improved depth estimate. Moreover, if $\mathrm{E}_k(\mathbf{\Delta}) = 0$, all the depth hypotheses are assumed to be perfectly consistent. Otherwise, if $\mathrm{E}_k(\mathbf{\Delta}) > \vartheta$, we reject all k depth hypotheses and assume that we do not have a consistent depth representation of the associated 3-D point.

In contrast to [10, 11], to select ϑ with a desired consistency quality, we consider a basic error event where only one depth hypothesis out of all available depth hypotheses is erroneous. For such events, all the depth hypotheses are equal to the true depth value $d_p(x, y)$ at the principal pixel, except the depth hypothesis from one viewpoint, say l, $\hat{d}_p(l; x, y) = d_p(x, y) + \mu$, where μ is the deviation from the true depth. The vector for a such event is $\mathbf{\Delta}_\mathsf{l} = [0, \ldots, 0, -\mu, \mu, 0, \ldots, 0]^\mathsf{T} \in \mathbb{R}^k$. With

$$\vartheta \equiv \mathrm{E}_k(\mathbf{\Delta}_\mathsf{l}) = 2\mu^2 \; \forall \; k, \tag{3}$$

we only allow the least possible error.

As each principal pixel has an associated energy $\mathrm{E}_k(\mathbf{\Delta})$, the testing gives the inter-view consistency information across multiple viewpoints. If the test fails with available k depth hypotheses, we repeat the consistency analysis and test with $(k-1)$ out of k available depth hypotheses. However, there are k ways to select $(k-1)$ out of k available depth hypotheses and to define the corresponding k unique loop difference vectors of dimension $(k-1)$. We therefore perform k consistency checks and test with k different depth difference vectors. If multiple tests out of k tests with threshold ϑ are successful, then we accept only the test with the lowest energy. If all k test with $(k-1)$ depth hypotheses fail, we repeat the process of consistency analysis and testing again with a reduced number of depth hypotheses until the desired consistency is achieved with $k \geq 2$. When all tests fail, we mark the associated principal pixel by a mask that allows other techniques, such as [20], to decide the best depth value.

To have inter-view consistent depth maps across k viewpoints, IVDCT is first utilized to obtain inter-view consistency information at viewpoint p which coincides with one of the reference viewpoints, i.e., $p = i$, where $i = 1, \ldots, k$. Next, the resulting consistency information at i is used to update the depth pixel at p. By updating, we mean that we replace the previous depth pixel value at i by a new improved depth value. It is determined by averaging the chosen depth hypotheses as obtained by consistency information. However, if the viewpoints are irregularly spaced, the depth value is updated by weighted baseline averaging of the chosen depth hypotheses. The enhanced depth values are then used to update the corresponding depth map value in the viewpoint i. A similar procedure is applied to update the depth maps at each viewpoint. The resulting depth maps show improved inter-view consistency across all the viewpoints. The resulting depth pixel value at i is then used as an improved input when testing the next viewpoint. We repeat this process until each viewpoint satisfies our stopping criterion which is the difference between the loop energies from recent iterations. When the loop energy does not change anymore with further iterations, we stop the process. This gives the MVD imagery with improved inter-view consistency.

The fundamental approach of this work is to consider a certain set of depth differences, and not the absolute depth values. We consider the loop difference vector as evidence. Due to the zero-sum constraint, the autocorrelation matrix of the evidence is singular. With the threshold constraint (i.e. the constraint on the variance), we are able to find a subspace of the evidence in which the zero-sum constraint is satisfied at a lower variance.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The FTV user will enjoy either camera captured views or virtual views at a time. Depth images are employed for generating novel virtual views by DIBR at viewpoints where real cameras are missing. Therefore, the quality of depth images has direct impact on the view synthesis. Hence, the performance of both depth enhancement algorithms is assessed through the effect on the objective quality of virtual views. In the experiments, we use four MVV data sets and the corresponding estimated MVD imagery of as provided by MPEG [7]: Newspaper, Lovebird1, Balloons, and Kendo. Each evaluation experiment for the proposed scheme mainly consists of two steps: (1) improvement of depth maps at multiple viewpoints using one of the proposed enhancement methods and (2) virtual view synthesis with the help of the improved depth maps. For the latter, the MPEG view synthesis reference software (VSRS) is employed [21]. The VSRS is a DIBR approach which takes two views, left and right, to render a view at a given intermediate viewpoint by using the two corresponding depth images and camera parameters. The virtual views are generated by using the 1D parallel synthesis mode of VSRS 3.5 with half-pel precision.

Due to computational complexity, we restrict our enhancement algorithms to use depth maps from only three viewpoints and one time instance (i.e., single frame). Our probabilistic depth enhancement algorithm starts with a specified number of superpixels. The number of superpixels in all experiments is fixed to 225000 superpixels for each concatenated three-view imagery. In our IVDCT experiments, μ is defined to be directly proportional to the standard deviation of Δ_{ij} and an additional factor. This factor is based on the number of available depth hypotheses. Table 1 shows a comparison of PSNR values (in dB) for the synthesized virtual views as generated by VSRS 3.5 when using (a) MPEG depth maps (MPEG/D), (b) IVDCT enhanced depth maps (IVDCT/ED) [22], and (c) enhanced depth maps as obtained by the probabilistic depth enhancement framework (PROMDE/ED) [18].

In general, both enhancement algorithms offer improvements in the quality of view synthesis when compared to conventional MPEG depth maps. However, the improvement in quality is likely to increase with an increasing number of reference viewpoints used for the testing. It also depends on the quality of the input reference depth maps at various viewpoints. Moreover, the objective improvement in the quality of view synthesis offered by both depth enhancement methods over MPEG depth maps is scene-dependent. The probabilistic framework gives the best quality for Kendo, but is inferior for Newspaper when compared to IVDCT. This is mainly due to the color classification results as achieved by VI-DMM in xyz chromaticity space. For the other sequences, the objective performance of both depth enhancement methods is similar.

One drawback of our probabilistic framework is the computational complexity of Bayesian learning in order to obtain probabilistic weights for depth enhancement. On the other hand, IVDCT is a pixel-based statistical approach which allows to test and improve the inter-view depth consistency in real time. The computational complexity and memory usage of the consistency testing algorithm is significantly lower. The quality of view synthesis is also limited by the choice of the rendering algorithm. In particular, VSRS 3.5 can not fully exploit our consistency information.

5. CONCLUSIONS

We described and compared two different inter-view depth enhancement algorithms for multiview depth imagery. The first is a probabilistic approach which exploits the inherent inter-view similarity in multiview imagery through color classification and Bayesian learning. The resulting probabilistic weights from the learning of depth models are then used to repair the input depth

Table 1. Comparison of objective quality of the synthesized virtual views using depth maps as enhanced by IVDCT and PROMDE.

Test	Input (Virtual) Views			VSRS 3.5 [dB]		
Sequence	VSRS	IVDCT	PROMDE	(a) MPEG/D	(b) IVDCT/ED	(c) PROMDE/ED
Kendo	3, (4), 5	1, 3, 5	1, 3, 5	36.5	37.0	38.0
Balloons	3, (4), 5	1, 3, 5	1, 3, 5	35.7	36.0	36.0
Lovebird1	6, (7), 8	4, 6, 8	4, 6, 8	28.5	29.0	29.1
Newspaper	4, (5), 6	2, 4, 6	2, 4, 6	32.3	33.2	32.6

imagery. The second improves the inter-view depth consistency by testing multiple depth hypotheses from various viewpoints. With the improved depth consistency, we are able to enhance the visual experience of FTV. Experimental results demonstrate the effectiveness of both depth enhancement methods. However, for FTV scenarios which require computationally inexpensive depth enhancement, inter-view consistency testing offers advantages. MVD imagery with more reliable and inter-view consistent depth values is offered by the probabilistic depth enhancement framework. But this approach comes with high computational cost. Selecting the feature vector elements for depth sub-clustering is still challenging and an open problem. The probabilistic framework has potential to improve temporal depth coherence by concatenating temporally successive frames. On the other hand, the use of the resulting inter-view consistency information from IVDCT is not limited to depth consistency enhancement. It can also be used to improve the quality of view synthesis by selecting reliable color pixels from the view imagery [10, 11]. Future research could combine the advantages of inter-view consistency testing and probabilistic depth enhancement.

6. REFERENCES

[1] M. Tanimoto, M. Tehrani, T. Fujii, and T. Yendo, "FTV for 3-D spatial communication," *Proc. IEEE*, vol. 100, no. 4, pp. 905–917, Apr. 2012.

[2] K. Müller, P. Merkle, and T. Wiegand, "3-D video representation using depth maps," *Proc. IEEE*, vol. 99, no. 4, pp. 643–656, Apr. 2011.

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

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, pp. 7–42, Apr. 2002.

[5] C. Cigla and A. Alatan, "Temporally consistent dense depth map estimation via belief propagation," in *3DTV Conf.*, Potsdam, Germany, May 2009, pp. 1–4.

[6] K. Müller, A. Smolic, K. Dix, P. Kauff, and T. Wiegand, "Reliability-based generation and view synthesis in layered depth video," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Cairns, Australia, Oct. 2008, pp. 34–39.

[7] MPEG, "Call for proposals on 3D video coding technology," ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Tech. Rep. N12036, Mar. 2011.

[8] E. Ekmekcioglu, V. Velisavljević, and S. Worrall, "Content adaptive enhancement of multi-view depth maps for free viewpoint video," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 2, pp. 352–361, Apr. 2011.

[9] R. Li, D. Rusanovskyy, M. M. Hannuksela, and H. Li, "Joint view filtering for multiview depth map sequences," in *Proc. IEEE Int. Conf. Image Process.*, Orlando, USA, Sept. 2012, pp. 1329–1332.

[10] P. K. Rana and M. Flierl, "Depth consistency testing for improved view interpolation," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, St. Malo, France, Oct. 2010, pp. 384–389.

[11] ——, "Depth pixel clustering for consistency testing of multiview depth," in *Proc. European Signal Process. Conf.*, Bucharest, Romania, Aug. 2012, pp. 1119–1123.

[12] P. K. Rana, J. Taghia, and M. Flierl, "A variational Bayesian inference framework for multiview depth image enhancement," in *Proc. IEEE Int. Symp. Multimedia*, Irvine, California, USA, Dec. 2012, pp. 183–190.

[13] P. K. Rana, Z. Ma, J. Taghia, and M. Flierl, "Multiview depth map enhancement by variational Bayes inference estimation of Dirichlet mixture models," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, Vancouver, Canada, May 2013, pp. 1528–1532.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. New York: Springer, 2006.

[15] M. Okutomi, O. Yoshizaki, and G. Tomita, "Color stereo matching and its application to 3-D measurement of optic nerve head," in *Proc. IAPR Int. Conf. Pattern Recognition*, The Hague, Netherlands, Aug. 1992, pp. 509–513.

[16] C. Dorea and R. De Queiroz, "Depth map reconstruction using color-based region merging," in *Proc. IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 1977–1980.

[17] Z. Ma, "Non-Gaussian statistical models and their applications," Ph.D. dissertation, KTH Royal Institute of Technology, Stockholm, 2011.

[18] P. K. Rana, J. Taghia, Z. Ma, and M. Flierl, "Probabilistic multiview depth image enhancement using variational inference," *IEEE J. Sel. Topics Signal Process.*, 2014, submitted.

[19] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[20] M. Bertalmio, A. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, Kauai, HI, USA, Dec. 2001, pp. 355–362.

[21] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," ISO/IEC JTC1/SC29/WG11, Archamps, France, Tech. Rep. M15377, Apr. 2008.

[22] P. K. Rana and M. Flierl, "Inter-view depth consistency testing in depth difference subspace," *IEEE Trans. Image Process.*, 2014, submitted.