A Variational Bayesian Inference Framework for Multiview Depth Image Enhancement

Pravin Kumar Rana, Jalil Taghia, and Markus Flierl
School of Electrical Engineering
KTH Royal Institute of Technology
Stockholm, Sweden
Email: {prara, taghia, mflierl}@kth.se

***Abstract*—In this paper, a general model-based framework for multiview depth image enhancement is proposed. Depth imagery plays a pivotal role in emerging free-viewpoint television. This technology requires high quality virtual view synthesis to enable viewers to move freely in a dynamic real world scene. Depth imagery of different viewpoints is used to synthesize an arbitrary number of novel views. Usually, the depth imagery is estimated individually by stereo-matching algorithms and, hence, shows lack of inter-view consistency. This inconsistency affects the quality of view synthesis negatively. This paper enhances the inter-view consistency of multiview depth imagery by using a variational Bayesian inference framework. First, our approach classifies the color information in the multiview color imagery. Second, using the resulting color clusters, we classify the corresponding depth values in the multiview depth imagery. Each clustered depth image is subject to further sub-clustering. Finally, the resulting mean of the sub-clusters is used to enhance the depth imagery at multiple viewpoints. Experiments show that our approach improves the quality of virtual views by up to 0.25 dB.**

***Keywords*-Multiview video; depth enhancement; variational Bayesian inference; Gaussian mixture model;**

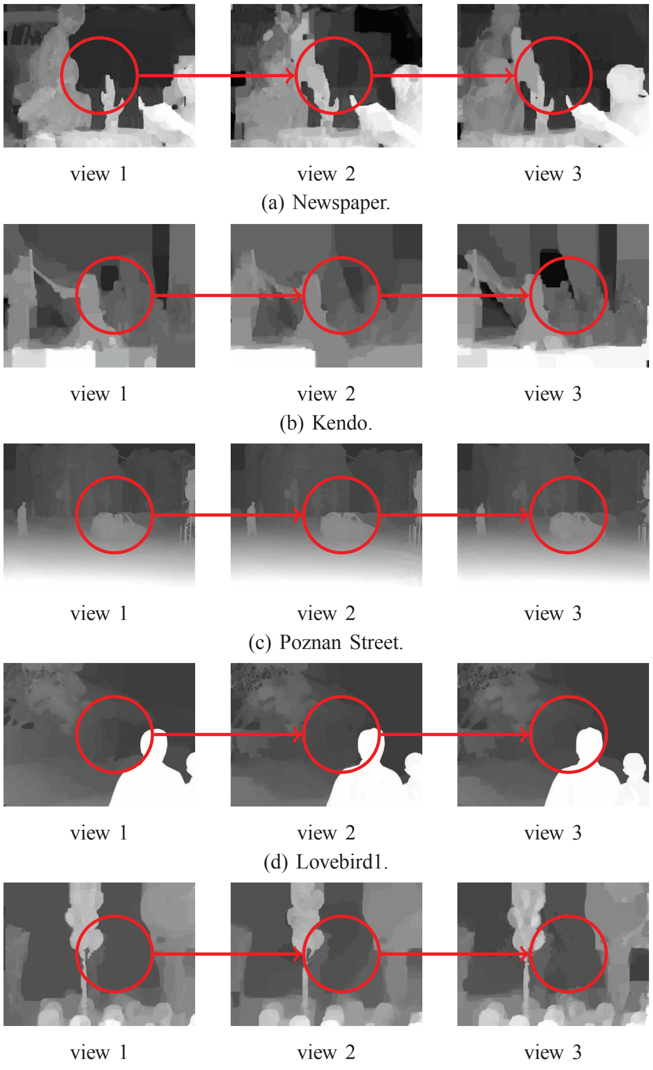
I. Introduction

Free-viewpoint television (FTV) will significantly change our current television experience [1], [2]. FTV will enable viewers to have a dynamic natural 3D-depth impression while freely choosing their viewpoint of real world scenes. This has been made possible by recent advances in electronic display technology which permits viewing of scenes from a range of perspectives [3]. Furthermore, the availability of low-cost digital cameras enables us to record easily multiview video (MVV) for FTV. MVV is a set of videos recorded by many video cameras that capture a dynamic natural scene from many viewpoints simultaneously. To provide a seamless transition among interactively selected viewpoints, we are required to store or transmit an enormous amount of MVV imagery [4]. In the future, the commercialization of FTV will further increase the demands for high-capacity multimedia transmission networks.

These requirements attracted many researcher in recent years and, as a result, many compression techniques have been proposed for MVV imagery [4], [5], [6]. As MVV is the result of capturing the same dynamic natural scene from various viewpoints, the imagery exhibits high inter-view and temporal similarities. The Moving Picture Experts Group (MPEG) proposed multiview video coding (MVC) as an extension to the existing H.264/AVC compression technology [7]. MVC exploits efficiently inherent similarities in the MVV imagery for compression. The resulting transmission cost for MVC is approximately proportional to the number of display views [8]. Therefore, a large number of views cannot be efficiently transmitted using MVC. However, limited subsets of MVV imagery can be transmitted to the receiver using existing networks. With only a limited subset of the captured color information, high quality view synthesis is not feasible [8]. However, by utilizing the scene geometry information such as depth maps, the quality can be improved significantly.

A depth map is a single channel gray scale image. Each pixel in the depth map represents the shortest distance between the corresponding object point and the given camera plane. Usually, depth maps are compressed by existing video codecs as they contain large smooth areas of constant grey levels. Given small subset of MVV imagery and its corresponding set of multiview depth images (MVD), an arbitrary number of views can be synthesized by using depth image based rendering (DIBR) [9]. The quality of these synthesized views depends significantly on the consistency of the MVD imagery. Usually, depth maps for different viewpoints are estimated independently by establishing stereo correspondences between nearby views only [10]. A number of different approaches are used for efficient depth estimation such as optimization via graph-cut [11], belief propagation [12], [13], and modified plane sweeping with segmentation [14]. Despite these optimizations, the resulting depth information at different viewpoints usually lacks inter-view consistency as shown in Fig. 1. This inconsistency affects the quality of view synthesis negatively and, hence, FTV users experience visual discomfort.

Many methods have been proposed to enhance the temporal inconsistency in MVD imagery, for example by using belief propagation [16], motion estimation [17] and by exploiting local temporal variations in the MVV imagery [18]. However, we addressed the inter-view inconsistency problem



(a) Newspaper.

(b) Kendo.

(c) Poznan Street.

(d) Lovebird1.

(e) Balloons.

Figure 1. Inter-view inconsistency among estimated multiview depth maps at different viewpoints for different multiview video imagery as provided by [15]. The red circles mark the most prevailing inconsistent areas in depth maps.

in [19] by testing evidence for depth consistency and by using a single hard threshold for testing. In the follow-up work [20], cluster adaptive thresholds are used for consistency testing, which are based on the statistics of each cluster. Recently [21] proposed content adaptive median filtering for improving temporal and inter-view consistency of depth maps by adapting to edges, motion and depth range. The methods [19], [21], and [20] warp depth maps from various viewpoints to a common viewpoint for spatial alignment. However, this warping produces errors due to the discrete values in depth maps and affects enhancement algorithms negatively [22].

In this paper, we keep the view imagery and the corresponding depth maps at their respective viewpoints. The objective of this paper is to propose and investigate a general model-based framework for depth map enhancement. First, the proposed framework uses variational Bayesian inference to perform color classification in the view imagery. Second, for each resulting color cluster, we classify the corresponding depth values from multiple viewpoints. Finally, multiple depth levels are clustered in individual sub-clusters for depth enhancement at multiple viewpoints. The resulting improved depth maps are utilized to enrich the FTV user experience by synthesizing high-quality virtual views.

The remainder of this paper is organized as follows: In Section II, we describe the proposed approach for MVD image enhancement. Section III presents the objective and subjective assessment of the proposed approach. Finally, Section IV gives concluding remarks and future directions.

II. Proposed Approach

The proposed algorithm consists of mainly four steps: (1) concatenation of view imagery, (2) multiview color classification, (3) multiview depth classification, and (4) multiview depth enhancement, as shown in Fig. 2. In rest of this section, we will explain the individual steps of the approach in detail.

To address the inconsistency problem in the estimated MVD imagery at multiple viewpoints, we assume that the MVV imagery of resolution $H \times W$ is independently captured for a given natural dynamic scene using projective cameras at N viewpoints. Usually, each captured view of the scene is an image in YUV color space [23]. We transform these views from YUV space to RGB color space [24]. This is because, the probabilistic mixture models can efficiently model pixel value distributions in the RGB space, even if the RGB space is not independent from the luminance in the capturing environment. In RGB space, a pixel in a view $\mathbf{v}_n \in \mathbf{R}^{H \times W \times 3}, n \in \{1, \ldots, N\}$, is described by a vector which comprises three primary color channels, red (r), green (g), and blue (b), i.e., $\mathbf{v}_n(p,q) = [r,g,b]^T$, where $p \in \{1, \ldots, H\}$, $q \in \{1, \ldots, W\}$ and T represents the transpose operation. Here, each color component can take values between 0 and 255.

A. Concatenation of View Imagery

The captured MVV imagery of the scene has inherent inter-view similarity. In order to have a unique model for the captured natural scene, we first exploit the inter-view similarity by concatenating views from N viewpoints to a single view $\mathbf{v} \in \mathbf{R}^{H \times NW \times 3}$,

$$\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_N], \tag{1}$$

as shown in Fig. 3(a). For simplicity, we transform

$$\mathbf{v} \in \mathbf{R}^{H \times NW \times 3} \longmapsto \overline{\mathbf{v}} \in \mathbf{R}^{3 \times M}, \tag{2}$$

where

$$\overline{\mathbf{v}} = [\overline{\mathbf{v}}_1, \ldots, \overline{\mathbf{v}}_M], \tag{3}$$

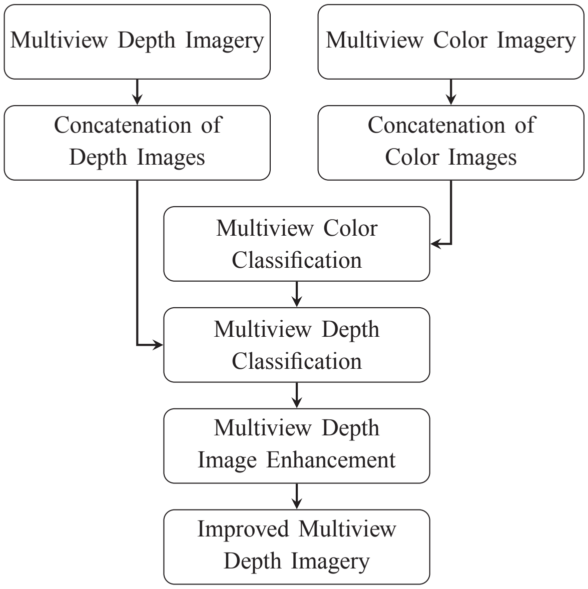

Multiview Depth Imagery
Multiview Color Imagery
Concatenation of Depth Images
Concatenation of Color Images
Multiview Color Classification
Multiview Depth Classification
Multiview Depth Image Enhancement
Improved Multiview Depth Imagery

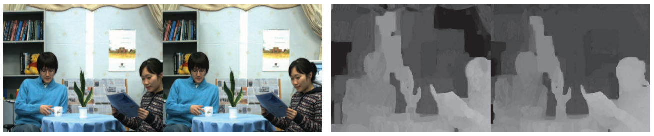
Figure 2. Block diagram of the proposed approach.

with $M = HWN$. Each $\overline{\mathbf{v}}_m, m \in \{1, \ldots, M\}$, is a point in a 3-dimensional space comprising the intensities of the r, g, and b color channels.

B. Multiview Color Classification

In this section, we discuss the proposed approach of color classification for the captured MVV imagery. We begin our discussion by considering the problem of identifying clusters of data points $\mathbf{v}$ in a multidimensional space. The goal is to partition the data set into K clusters, where we shall assume that the value of K is unknown.

Intuitively, a cluster comprises a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster. Let $\boldsymbol{\mu}_k$, where $k = 1, \ldots, K$ denote a prototype associated with the k^{th} cluster. We shall assign data points to clusters, as well as a set of vectors $\{\boldsymbol{\mu}_k\}$, such that the sum of the squares of the distances of each data point to its closest vector $\boldsymbol{\mu}_k$ becomes minimum. In other words, the problem that we are dealing with can be considered as image segmentation. As mentioned by [25], the goal of segmentation is to partition an image into regions each of which has a reasonably homogeneous visual appearance or which corresponds to objects or parts of objects. The K-means algorithm can be considered as one approach for image segmentation. However, this approach suffers from two main drawbacks: 1) it does not consider spatial proximity of different pixels; 2) the number of clusters K has to be known [26]. Alternatively, one may consider a Gaussian mixture model (GMM) for the segmentation task. GMMs are a valuable statistical tool for modeling densities. They are flexible enough to approximate any given density with high accuracy, and, in addition, they can be interpreted as a soft clustering solution. Thus, they have been widely used in both supervised and unsupervised learning, and have been extensively studied and applied in several domains. One way of making a GMM is the maximum likelihood where the expectation-maximization (EM) [27] algorithm is used to find the maximum likelihood solutions. The EM algorithm for Gaussian mixtures is similar to the K-means algorithm, where the K-means algorithm performs a hard assignment of data points to clusters, in which each data point is associated uniquely with one cluster. In general, the EM algorithm makes a soft assignment based on the posterior probabilities. There are two main problems associated with the maximum likelihood EM-based framework applied to GMM: 1) it sufferers from singularities when one of the Gaussian components collapses onto a specific data point. In this case, the log likelihood function goes to infinity; 2) it sufferers from over-fitting and, similar to the K-means algorithm, the number of components K has to be known.



(a) Concatenation of color images. (b) Concatenation of depth images.

Figure 3. Example of image concatenation of two viewpoints.

A fully Bayesian approach has been proposed in [28], where the number of components is treated as a random variable, and the reversible jump Markov chain Monte Carlo method is used for sampling. However, this method is computationally demanding. To deal with the intractable integrations appearing in the Bayesian approach, the use of a variational approximation [29], [30], [26] has been proposed that yields an iterative method similar to the formulation of EM. This general optimization method is called Variational Bayes (VB) and has been used in a number of recent works.

1) Variational Mixture of Gaussians: A detail derivation of variational mixture of Gaussians can be found in [26]. Here, we only provide a summary of this approach. Let $\overline{\mathbf{v}}$ denote a set of M independent observations, i.e., $\overline{\mathbf{v}} = \{\overline{\mathbf{v}}_1, \ldots, \overline{\mathbf{v}}_M\}$, where each observation $\overline{\mathbf{v}}_m, m \in \{1, \ldots, M\}$ is a D-dimensional feature vector ($D = 3$) comprising the intensities of the r, g and b color channels in real space $\mathbf{R}^D$. Associated with every observation $\overline{\mathbf{v}}_m$, there is a corresponding latent variable $\mathbf{z}_m = [z_{m1}, \ldots, z_{mK}]^{\text{T}}$ consisting of a 1-of-K binary vector with elements z_{mk} for $k = 1, \ldots, K$. Now, let $p(\overline{\mathbf{v}}_m)$ denote a mixture with K Gaussian components

$$p(\overline{\mathbf{v}}_m) = \sum_{k=1}^{K} \tau_k \; \mathcal{N}(\overline{\mathbf{v}}_m \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k), \tag{4}$$

where τ_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Lambda}_k$ represent the mixture weight, the mean value, and the precision of the k^{th} component.

A Bayesian mixture model is obtained by imposing priors on the parameters of the model. Typically, conjugate priors are used such that the prior and posterior will have the same functional form and, hence, optimization procedures can be carried out in an iterative manner. Therefore, a Dirichlet prior distribution is introduced over the mixing coefficients; and a Gaussian-Wishart distribution is introduced over $\underline{\boldsymbol{\mu}}$ and $\underline{\boldsymbol{\Lambda}}$ governing the mean and precision of the Gaussian components.

Bayesian model selection is obtained through maximization of the marginal likelihood. The variational approximation of the VB method suggests the maximization of a lower bound of the logarithmic marginal likelihood. A notable property of the method is that during maximization of the lower bound, if some of the components fall in the same region of the data space, then there is a strong tendency in the model to eliminate the redundant components, once the data in this region is sufficiently explained by fewer components.

The learning task in the VB approach consists of the optimization of the variational distribution of the latent variables and component parameters. Based on the optimized solution provided by variational inference [30], optimization of the posterior distribution of latent variables can be obtained by taking the expectation of terms involving latent variables in the joint distribution with respect to the component parameters. This leads to a set of responsibilities r_{mk} which tell how responsible the k^{th} component is for modeling of $\overline{\mathbf{v}}_m$. Similarly, optimization of the posterior distribution of component parameters can be obtained by taking the expectation of terms involving component parameters in the joint distribution with respect to the latent variables.

In summary, the algorithm starts with the initialization of the hyper-parameters characterizing the parameter distributions. In the next step, the current distribution over the model parameters is used to evaluate the responsibilities which is a result of the optimization of the posterior distribution of latent variables. Later, these responsibilities are used for the optimization of the variational posterior distribution over component parameters and provide the update for the hyper-parameters. This procedure continues until convergence.

Let $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_M]$ denote the responsibility matrix, where $\mathbf{r}_m = [r_{m1}, \ldots, r_{mK}]^{\mathrm{T}}$, and let $\mathbf{w} = [\mathrm{w}_1, \ldots, \mathrm{w}_K]^{\mathrm{T}}$ denote the weight mixture, where $\mathrm{w}_k = \sum_{m=1}^{N} r_{mk}$. We need to keep clusters with certain weights by which data can be sufficiently explained. The corresponding cluster indices can be obtained by

$$\{i\}_{\in K} = \max_{k} \;\; \mathbf{w} \geq 0.1 \;\; \max(\mathbf{w}), \tag{5}$$

where $\underline{I} = \{1, \ldots, i, \ldots, I\}$ and $I \leq K$. Intuitively, the set $\underline{I}$ includes indices of clusters which represent certain colors. It is notable that we do not need a very fine classification of the colors. Hence, in (5) we introduced a threshold which is data independent. The clusters which are rejected in the thresholding procedure include either none or few members which should be reassigned to their nearest clusters. Let $\{l\}_{\in K \setminus I}$ denote the index of the l^{th} cluster which is not in the set of $\underline{I}$. A set of such indices are denoted by $\underline{L} = \{1, \ldots, l, \ldots, L\}$. The reassigning members of the clusters in the set $\underline{L}$ to the nearest clusters in the set $\underline{I}$ can be done by calculating the distances of their cluster prototypes $\{\boldsymbol{\mu}_l\}$ to their closest cluster prototypes $\{\boldsymbol{\mu}_i\}$ so that the absolute value of the distance becomes minimum.

Members of the i^{th} cluster are shown as $\underline{\mathbf{Y}}^{(i)}$ which can be extracted from the observation set $\overline{\mathbf{v}}$ as,

$$\underline{\mathbf{Y}}^{(i)} = \{\mathbf{y}_1^{(i)}, \ldots, \mathbf{y}_M^{(i)}\}, \tag{6}$$

$$\mathbf{y}_m^{(i)} = \mathcal{M}_m^{(i)} \; \overline{\mathbf{v}}_m \tag{7}$$

where we have defined

$$\mathcal{M}_m^{(i)} = \begin{cases} 1, & \text{if} \;\; r_{mi} > r_{mj}, \forall j \neq i \;\; (j, i \in \{1, \ldots, K\}); \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

C. Multiview Depth Classification

For each view $\mathbf{v}_n$, we assume that the associated per-pixel depth map $\mathbf{d}_n \in \mathbf{R}^{H \times W}$ exists. Each pixel in the depth map $\mathbf{d}_n$ has a discrete value, where the value zero represents the farthest point and 255 the closest. In order to enhance inter-view consistency, we concatenate depth maps from N viewpoints to a single depth $\mathbf{d} \in \mathbf{R}^{H \times NW}$,

$$\mathbf{d} = [\mathbf{d}_1, \ldots, \mathbf{d}_N], \tag{9}$$

as shown in Fig. 3(b). Again, for simplicity, we consider the following mapping

$$\mathbf{d} \in \mathbf{R}^{H \times NW} \longmapsto \overline{\mathbf{d}} \in \mathbf{R}^{1 \times M}, \tag{10}$$

where

$$\overline{\mathbf{d}} = [\overline{\mathbf{d}}_1, \ldots, \overline{\mathbf{d}}_M], \tag{11}$$

is such that for each color pixel $\overline{\mathbf{v}}_m, m \in \{1, \ldots, M\}$, we have an associated depth value $\overline{\mathbf{d}}_m \in \{0, \ldots, 255\}$. We therefore utilize this per-pixel depth value association with color values by using $\mathcal{M}_m^{(i)}$ in order to obtain members of the i^{th} depth cluster, $\underline{\mathbf{X}}^{(i)}$,

$$\underline{\mathbf{X}}^{(i)} = \{\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_M^{(i)}\}, \tag{12}$$

$$\mathbf{x}_m^{(i)} = \mathcal{M}_m^{(i)} \; \overline{\mathbf{d}}_m. \tag{13}$$

Figure 4 shows such color clusters and associated depth clusters for concatenated color images and depth maps from two viewpoints, respectively. Note that this approach efficiently clusters similar color pixels from multiple viewpoints without making any specific assumptions about the illumination.

D. Multiview Depth Image Enhancement

The members of the cluster $\underline{\mathbf{Y}}^{(i)}$ are similar colors, whereas members of the cluster $\underline{\mathbf{X}}^{(i)}$ are different depth values. This is because a foreground and a background object point can have a similar color, but foreground object points have different depth values compared to background object points. Furthermore, if an object point with a given color is visible from N viewpoints, this point should have the same depth value in all N depth maps. However, such points usually have different depth values in the cluster $\underline{\mathbf{X}}^{(i)}$ due to the inconsistency across multiple viewpoints. This motivates us to consider further sub-clustering of each $\underline{\mathbf{X}}^{(i)}$, where the variance of each sub-cluster reflects the inconsistency in depth values from various viewpoints. Here, we apply the K-means algorithm for the purpose of sub-clustering. The K-means clustering algorithm is computationally fast and, hence, a good choice for this sub-clustering procedure. We may use again the Bayesian mixture model of Gaussians in order to perform this sub-clustering stage. This will result in a more accurate clustering, but it will entail a higher computational complexity. The K-means assigns the mean of each sub-cluster to depth pixels which fall into the specified depth sub-cluster, irrespective of the originating viewpoints.

III. Experimental Results

MPEG uses the view synthesis reference software (VSRS) for view synthesis, which is an DIBR approach for synthesizing virtual views [31], [32]. The VSRS uses two reference views, left and right, to synthesizes a virtual view at an arbitrary intermediate viewpoint by using the two corresponding reference depth maps and camera parameters. To evaluate the proposed algorithm, we compare the subjective and objective quality of the virtual views as synthesized by VSRS 3.5 with the help of MPEG provided depth maps and improved depth maps from our approach.

First, the depth imagery from two viewpoints is improved by utilizing the proposed approach with $K = 100$. For this, we concatenated only two views and the two corresponding depth maps as input to our algorithm. Second, a virtual view for a given viewpoint is synthesized by VSRS 3.5 using the improved depth maps. We synthesize these virtual views by using the 1D parallel synthesis mode with half-pel precision.

A. Objective Results

We measure the objective quality of the synthesized views in terms of the peak signal-to-noise ratio (PSNR) with respect to the captured view from a real camera at the same viewpoint. For this evaluation, we use five MVV test sets and the corresponding depth maps as provided by MPEG [15]. Table I shows a comparison of the luminance signal Y-PSNR (in dB) of the synthesized virtual view by VSRS 3.5 using (a) MPEG provided depth maps and (b) enhanced depth maps by the proposed algorithm. The presented enhancement algorithm offers improvements of up to 0.25 dB. The improvement in quality depends on the input reference depth maps from various viewpoints. The proposed algorithm is not very sensitive to the initialization for color classification. For the Balloons test data, the mean quality ± standard deviation of ten experiments with different initialization is 35.836 ± 0.068 dB. This compares to 35.68 dB when using MPEG depth maps. For experiments, the number of sub-clusters is manually fixed to 12. Table I, best results are presented.

Table I
Objective quality of the synthesized virtual views

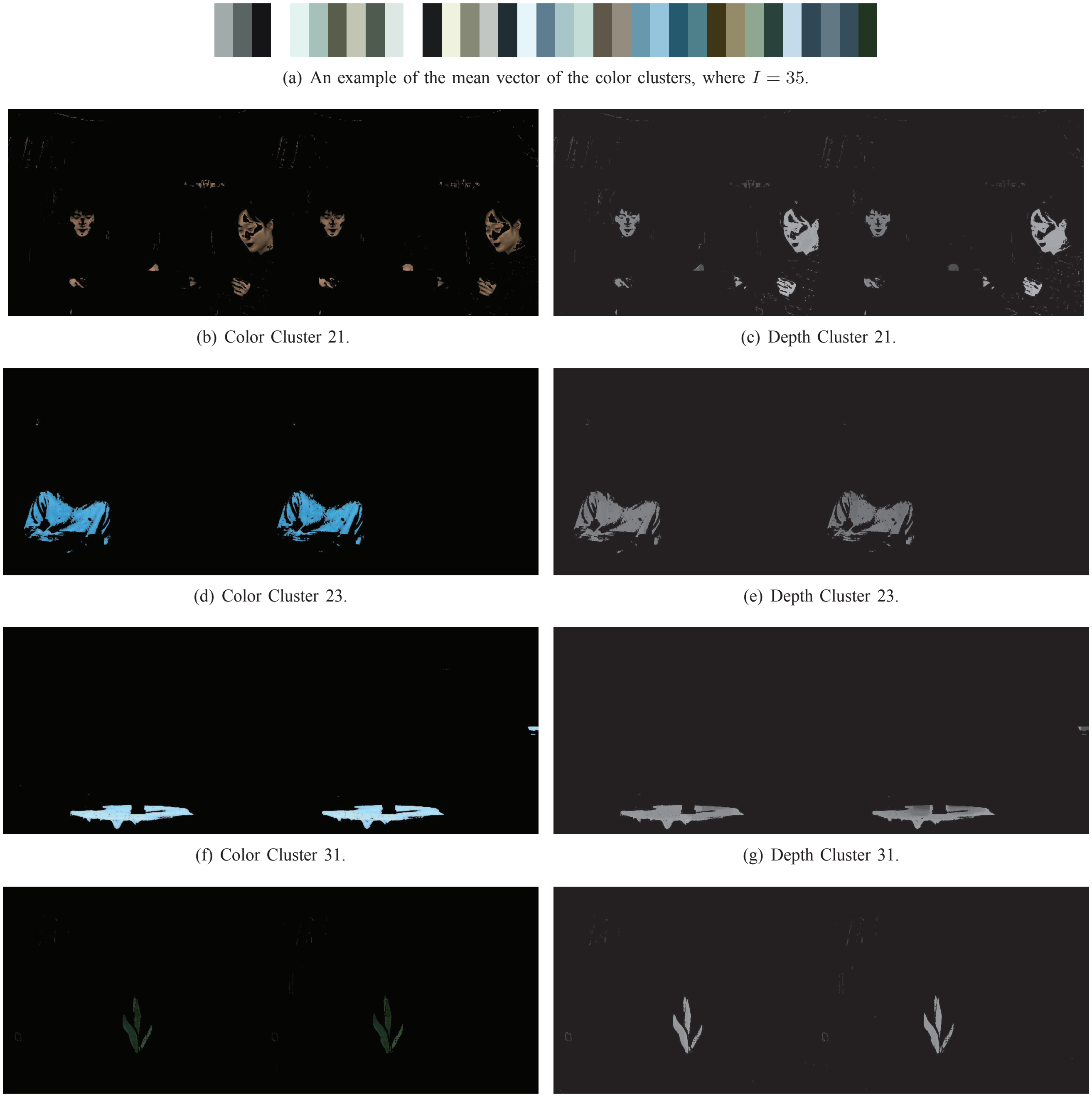
Test Sequence	Input Views	Virtual View	VSRS 3.5 [dB]	
			MPEG Depth	Improved Depth
Newspaper	4-6	5	31.98	32.10
Kendo	3-5	4	36.54	36.72
Poznan Street	3-5	4	35.56	35.58
Lovebird1	6-8	7	28.50	28.68
Balloons	3-5	4	35.68	35.93

B. Subjective Results

The visual quality of virtual view synthesis is noticeably improved by using the enhanced MVD imagery. The proposed algorithm efficiently reduces the artifacts, specially around the edges of the synthesized virtual views. Fig. 5 shows synthesized virtual views for the five test sequences, Newspaper, Kendo, Poznan Street, Lovebird1, and Balloons, respectively. For the Newspaper sequence, the blue sweater and the background are well synthesized by our proposed depth enhancement. We noticeably reduce the artifacts around the Kendo hakama and the trouser of the spectator. The visual quality of the synthesized Poznan street view improves, specially around the edges, as shown in Fig. 5. Artifacts around the hair and around the red jeogori sleeve of the man have been reduced for Lovebird1. For Balloons, the artifacts around the balloon boundaries are efficiently suppressed with the proposed enhancement. This demonstrates the efficiency of our enhancement algorithm for MVD imagery and, hence, this is a promising algorithm for improving the visual experience of FTV users.

IV. Conclusion And Future Work

We proposed a MVD image enhancement algorithm that improves inter-view depth consistency and, hence, that is able to enhance the visual experience of FTV users. The presented algorithm is based on multiview color classification by a variational Bayesian approach and uses resulting color clusters to classify depth values from various viewpoints. Here, per-pixel associations between depth and color have been exploited for the classification. Both objective and subjective results demonstrate the effectiveness of the presented algorithm. The proposed framework has the potential to address temporal depth inconsistencies by considering temporal views and depth maps from multiple viewpoints.



(a) An example of the mean vector of the color clusters, where $I = 35$.

(b) Color Cluster 21.

(c) Depth Cluster 21.

(d) Color Cluster 23.

(e) Depth Cluster 23.

(f) Color Cluster 31.

(g) Depth Cluster 31.

(h) Color Cluster 35.

(i) Depth Cluster 35.

Figure 4. Color and corresponding depth classification results.

An interesting avenue for future research is to apply our algorithm on a block-by-block basis. The motivation behind a block-by-block algorithm is to decrease the time for computation. Furthermore, we would like to consider other mixture models for efficient and fast color classification, such as Beta mixture models.

ACKNOWLEDGMENT

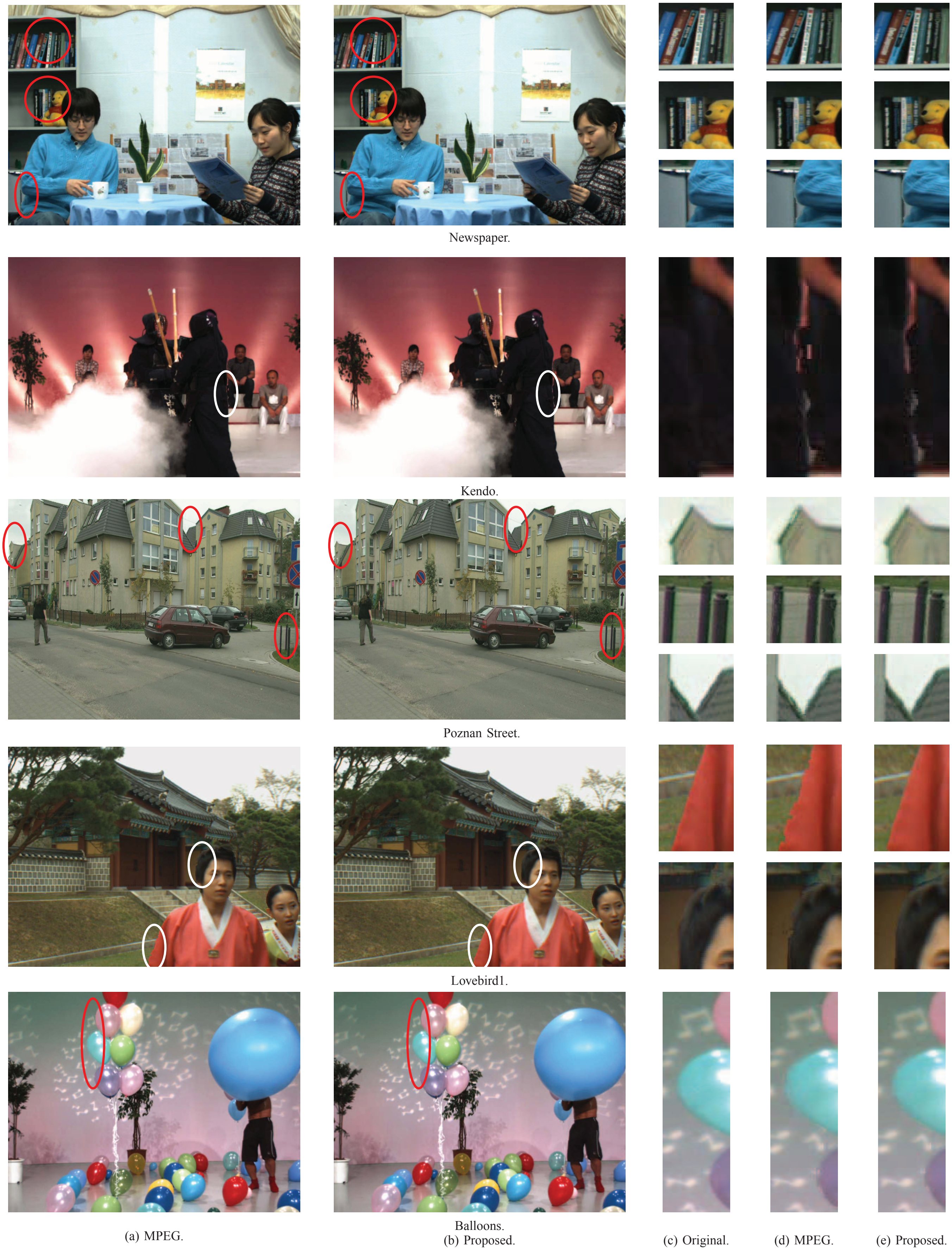
The authors thank Professor Arne Leijon for supporting this work.

REFERENCES

[1] M. Tanimoto, “Overview of free viewpoint television,” *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454–461, Jul. 2006.

[2] ——, “FTV (Free Viewpoint Television) for 3D scene reproduction and creation,” in *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, New York, USA, Jun. 2006, pp. 172–172.

[3] H. Urey, K. Chellappan, E. Erden, and P. Surman, “State of



Newspaper.

Kendo.

Poznan Street.

Lovebird1.

Balloons.

(a) MPEG.
(b) Proposed.
(c) Original.
(d) MPEG.
(e) Proposed.

Figure 5. Synthesized virtual views of the test sequences as generated by VSRS 3.5 using (a) MPEG depth maps and (b) enhanced depth maps from the proposed algorithm. (c), (d) and (e) show the white/red highlighted areas for a detailed comparison.

the art in stereoscopic and autostereoscopic displays," *Proc. IEEE*, vol. 99, no. 4, pp. 540 – 555, April 2011.

[4] M. Flierl and B. Girod, "Multiview video compression," *IEEE Signal Process. Mag.*, vol. 24, no. 6, pp. 66 –76, Nov. 2007.

[5] M. Magnor, P. Ramanathan, and B. Girod, "Multi-view coding for image-based rendering using 3-D scene geometry," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 11, pp. 1092 – 1106, Nov. 2003.

[6] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G. Akar, G. Triantafyllidis, and A. Koz, "Coding algorithms for 3DTV–A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1606 –1621, Nov. 2007.

[7] A. Vetro, T. Wiegand, and G. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard," *Proc. IEEE*, vol. 99, no. 4, pp. 626 –642, Apr. 2011.

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

[9] C. Fehn, "Depth-image-based rendering DIBR, compression, and transmission for a new approach on 3D-TV," in *Proc. of Stereoscopic Displays and Virtual Reality Systems XI*, A. J. Woods, J. O. Merritt, S. A. Benton, and M. T. Bolas, Eds., vol. 5291, no. 1. SPIE, 2004, pp. 93–104.

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

[11] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124 –1137, Sept. 2004.

[12] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787 –800, Jul. 2003.

[13] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," in *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 1, Jun. 2004, pp. I–261 –I–268.

[14] C. Cigla, X. Zabulis, and A. Alatan, "Region-based dense depth extraction from multi-view video," in *Proc. IEEE Int. Conf. Image Process.*, vol. 5, Oct. 2007, pp. V–213 –216.

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

[16] C. Cigla and A. Alatan, "Temporally consistent dense depth map estimation via belief propagation," in *3DTV Conf.: The True Vision - Capture, Transmission and Display of 3D Video*, May 2009, pp. 1 –4.

[17] S. Lee and Y. Ho, "Temporally consistent depth map estimation using motion estimation for 3DTV," in *Int. Workshop Adv. Image Technol.*, Jan. 2010, pp. 149(1 –6).

[18] D. Fu, Y. Zhao, and L. Yu, "Temporal consistency enhancement on depth sequences," in *Picture Coding Symp.*, Dec. 2010, pp. 342 –345.

[19] 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.

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

[21] 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.

[22] L. Do, S. Zinger, and P. de With, "Objective quality analysis for free-viewpoint DIBR," in *Proc. IEEE Int. Conf. Image Process.*, Hong Kong, Sept. 2010, pp. 2629 –2632.

[23] C. A. Poynton, *A technical introduction to digital video*. New York, NY, USA: John Wiley & Sons, Inc., 1996.

[24] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, Jan. 2002.

[25] D. A. Forsyth and J. Ponce, *Computer vision: a modern approach*, 1st ed. Englewood Cliffs, NJ: Prentice Hall, 2003.

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

[27] G. J. McLachlan and T. Krishnan, *The EM Algorithm and its Extensions*, 1st ed. New York: Wiley, 1977.

[28] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unknown number of components," in *J. Royal Statist. Soc.*, ser. B, vol. 59, no. 4, 1997, pp. 731 – 792.

[29] T. S. Jaakkola, "Tutorial on variational approximation methods," in *Advanced Mean Field Methods: Theory and Practice*. MIT Press, 2000, pp. 129 –159.

[30] Z. Ghahramani and M. J. Beal, "Variational inference for Bayesian mixtures of factor analysers," in *Advances in Neural Information Processing Systems 12*. MIT Press, 2000, pp. 449 –455.

[31] 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.

[32] MPEG, *View Synthesis Software Manual*, ISO/IEC JTC1/SC29/WG11, Sept. 2009, release 3.5.