Probabilistic Multiview Depth Image Enhancement Using Variational Inference

Pravin Kumar Rana, *Student Member, IEEE,* Jalil Taghia, Zhanyu Ma, *Member, IEEE,* and Markus Flierl, *Member, IEEE*

***Abstract*—An inference-based multiview depth image enhancement algorithm is introduced and investigated in this paper. Multiview depth imagery plays a pivotal role in 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 inter-view inconsistency. This inconsistency affects the quality of view synthesis negatively. This paper enhances the multiview depth imagery at multiple viewpoints by probabilistic weighting of each depth pixel. First, our approach classifies the color pixels 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 subclustering. Clustering based on generative models is used for assigning probabilistic weights to each depth pixel. Finally, these probabilistic weights are used to enhance the depth imagery at multiple viewpoints. Experiments show that our approach consistently improves the quality of virtual views by 0.2 dB to 1.6 dB, depending on the quality of the input multiview depth imagery.**

***Index Terms*—Multiview video, multiview depth consistency, virtual view synthesis, free-viewpoint television, Dirichlet mixture model.**

I. Introduction

CONSISTENT and precise geometry information on natural scenes is highly desirable for many computer vision algorithms and visual media applications. Free-viewpoint television (FTV) is one of such emerging immersive visual media applications [1]. It will enable users to experience a dynamic natural 3D-depth impression while freely choosing their viewpoint of a real world scene. FTV is able to display a large number of views from different viewpoints at the receiver side in order to have a seamless free-viewpoint experience while maintaining a realistic depth perception of natural 3D scenes [2]. 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 [3]. However, these requirements may be significantly reduced by using geometry information of 3D scenes, for example, depth images [4].

Usually, depth images are quantized representations where each depth value is stored as an eight bit single-channel gray value between zero and 255. Each pixel in the depth image represents the shortest distance between the corresponding object point in the natural scene and the given camera plane. For a given small subset of multiview video (MVV) imagery and its corresponding set of multiview depth (MVD) images, an arbitrary number of views can be rendered by using depth-image-based rendering (DIBR) [5]. This is why depth images are critical for FTV. Moreover, multiview depth imagery may also be used to compress the multiview video data more efficiently compared to conventional compression schemes. But for that, advanced compression of multiview depth data is needed. Various methods have been developed to obtain depth information [6]. There are basically two ways to acquire depth images from dynamic natural scenes: 1) active depth sensing and 2) passive depth sensing. Active depth sensing uses special sensors such as time-of-flight cameras to generate real-time depth images from discrete depth measurements of the natural scene. Depth is estimated by measuring the phase delay of the infrared light reflected by the dynamic scene [6]. These depth estimates are burdened by noise and holes, specially due to optical imperfection and scene environment reflectance [7]. The resulting depth images have also low spatial resolution and require upsampling to match the high-resolution video [8]. Further, depth measurements are less reliable for texture-rich regions and large distances.

Numerous methods have been proposed to improve the quality of the actively acquired depth maps. The vast majority of these methods assume and exploit the correlation between the depth maps and the corresponding color images to compute the missing depth values and to denoise depth maps, such as [9], [10], and [11]. Upsampling of high-quality depth maps may be achieved by assuming co-occurrences of depth and image boundaries in the associated high-resolution color image (e.g., [8]). In cases where an object in a scene moves fast, time-of-flight cameras are also affected by motion blur as ordinary cameras [7]. To address this issue, a method for motion blur detection and deblurring has recently been proposed in [12].

Acquisition of complete geometry information for a natural scene by using a time-of-flight camera from a single viewpoint is not possible. A feasible solution is to use multiple depth sensing cameras from different viewpoints. However, this approach gives rise to the interference problem. Multiple time-of-flight cameras can interfere with each other because each

Pravin Kumar Rana, Jalil Taghia and Markus Flierl are with the ACCESS Linnaeus Centre, School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, 100 44, Sweden, e-mail:{pravin.kumar.rana, jalil.taghia, markus.flierl}@ee.kth.se.

Zhanyu Ma is with the Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing, China. e-mail:mazhanyu@bupt.edu.cn.

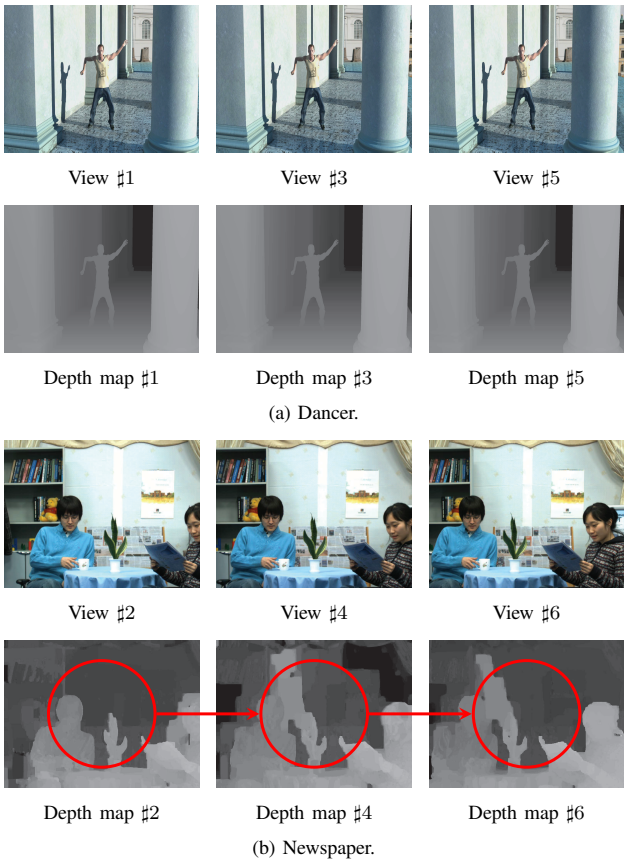


Fig. 1. Inter-view inconsistency among multiview depth maps at different viewpoints for the multiview imagery as provided by [18]. Note, the Dancer test data set is a synthetic test material and has a consistent depth representation across all viewpoints in (a). On the other hand, the estimated depth maps of the Newspaper test data set show inter-view depth inconsistency across all viewpoints in (b), where red circles mark prominent inconsistent areas in the depth maps. (Best viewed in color.)

time-of-flight camera emits its own light. Hence, the resulting depth image quality degrades significantly at multiple viewpoints [13]. A plane-sweeping based algorithm is proposed in [14] to tackle the interference problem for a multiple time-of-flight cameras setup. The fusion of depth estimates from multiple viewpoints has been investigated in [15] and [16] to obtain high quality 3D model reconstruction. The calibration of the multiple time-of-flight cameras is another challenging issue when acquiring geometry information of a natural dynamic scene [17]. Furthermore, depth sensing suffers from two other major disadvantages. First, the active depth sensors have to be placed at slightly different positions than the corresponding video cameras. Second, these sensors are mostly limited to indoor environments.

On the other hand, passive depth sensing provides full resolution and viewpoint-aligned depth images by utilizing camera-captured images from natural dynamic environments. Full resolution depth maps significantly improve the depth perception and the quality of rendered views [19]. Depth images obtained by passive sensing is the focus of our study in this paper. Passive depth estimates are usually obtained by establishing stereo correspondences among two or more camera images at different viewpoints with the help of a matching criterion [20]. The accuracy of stereo matching affects the resulting depth estimates. Stereo matching has been an active research topic for many decades and a number of optimization techniques are used to refine depth estimates, for example, graph-cut [21], belief propagation [22], and modified plane sweeping [23]. Despite using a number of optimization techniques, the resulting depth maps usually lack temporal consistency because depth estimation does not exploit temporal coherence among view frames. This results in temporal inconsistency and flicking.

In order to obtain overall geometry information about a scene, multiple depth images are estimated using stereo-matching algorithms at multiple viewpoints. However, as depth estimation at each viewpoint is independent in passive sensing, the resulting depth maps at different viewpoints usually lack inter-view consistency as shown in Fig. 1. For example, in case of a 1D parallel camera array, depth map values of a unique 3D point should be the same in all depth maps, but located at different positions in the maps at a given instant of time. In such a camera setting, all optical centers of the cameras are parallel to each other and all rotation matrices are identical. Therefore, depth observations at different viewpoints should be consistent, and related areas in different viewpoints should show the same depth values, but shifted, as shown in Fig. 1(a). This is not always the case in Fig. 1(b), as the estimated depth maps at different viewpoints reflect strong inter-view depth inconsistencies.

DIBR-based view synthesis algorithms may use multiple views acquired from different viewpoints and their corresponding depth images. The inconsistencies of depth values at different viewpoints affect the quality of rendered views. Incorrect depth values lead to erroneously chosen pixels for the view interpolation. This leads to perceptually annoying artefacts in the rendered view [24]. Hence, inter-view depth inconsistencies affect the quality of synthesized views negatively [25]. Furthermore, as depth images are crucial for many FTV data representations such as [26], [27], and [28], inter-view depth inconsistencies may hamper coding as well.

The enhancement of passively estimated multiview depth imagery at multiple viewpoints is the goal of this paper. Techniques developed to improve active depth sensing, focus usually on modeling sensor noise and measurement errors. In general, these modelling techniques are not suitable to improve the estimation errors from stereo matching algorithm (e.g., [11] and [29]). On the other hand, many methods have been proposed to improve the temporal inconsistency in the estimated multiview depth imagery, for example, [30], [31], [32], and [33].

Our goal is to improve the inter-view depth consistency among estimated depth images at multiple viewpoints. In [34], we proposed a general model-based framework for multiview depth map enhancement which improves depth maps at their respective viewpoints by utilizing color information from view imagery. The idea to improve stereo matching results by using color information has been mentioned in [35] and later investigated by [36]. Recently, many researchers exploited color classification to improve depth estimates, enforce depth

smoothness, and delineate sharp depth boundaries, for example, [37], [38], and [39].

Our initial enhancement framework in [34] mainly consists of two processing steps: multiview color classification and multiview depth classification. First, color clustering is performed on the concatenated view imagery. The clustering is carried out in an unsupervised fashion using a generative clustering approach based on Gaussian mixture models (GMMs). The model parameters are estimated in a Bayesian framework by variational inference (VI) [40], which allows automatic determination of the number of color clusters. 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. The choice of the generative model has significant influence on the clustering performance. As an extension to [34], we investigate in [41] the Dirichlet mixture model as the generative model in the color-clustering stage [42]. Contrary to [34], the clustering is preformed on the xyz chromaticity space of the view imagery. The use of the chromaticity space reduces the effects of illumination for the clustering. The choice of the Dirichlet distribution is motivated by two facts. First, a normalized vector in the xyz chromaticity space has non-negative elements and its l_1 norm equals one. These properties fit the definition of the Dirichlet distribution nicely [42], [43]. Second, the learning inference in DMM requires the estimation of fewer parameters, when compared to that of GMM. This implies less model complexity at superior clustering performance.

Although using DMM considerably reduces the model complexity when compared to GMM, the high computational demand for the enhancement framework is still a concern. Therefore, our objective of this paper is to effectively reduce the computational complexity of the framework, while maintaining its overall performance. In contrast to our fragmented pervious work [34] and [41], the first contribution of this paper is the use of superpixels instead of image pixels for color classification. A superpixel is a group of perceptually meaningful and homogeneous neighboring image pixels [44]. Superpixels capture image redundancy. This fact helps to reduce the number of feature vectors for color classification significantly. Hence, computational time is saved as well [44].

By taking advantage of reduced computational demand for color classification, we propose an algorithm for fully probabilistic multiview depth enhancement (PROMDE). In contrast to [34] and [41], which use K-means and mean-shift, respectively, depth subclassification is facilitated by variational inference using Gaussian mixture models. This is the second main contribution of our paper.

In our initial work [34] , we use discriminative clustering for the depth subclassification stage. Such methods require the prior knowledge of the exact number of clusters. The success of such methods depends highly on this prior knowledge. But note, in general, depth is noisy and determining the necessary number of clusters is a nontrivial task. In [41], we use unsupervised mean shift clustering for depth subclassification. Here, the final number of clusters is sensitive to the minimum number of pixels in a cluster. That is, clusters containing less than a specified number of pixels will be merged with its neighboring cluster [45]. To overcome these problems in this work, we use a generative-based clustering approach with a mixture of Gaussian distributions. The learning is carried out in a fully Bayesian inference, which allows automatic determination of the model complexity.

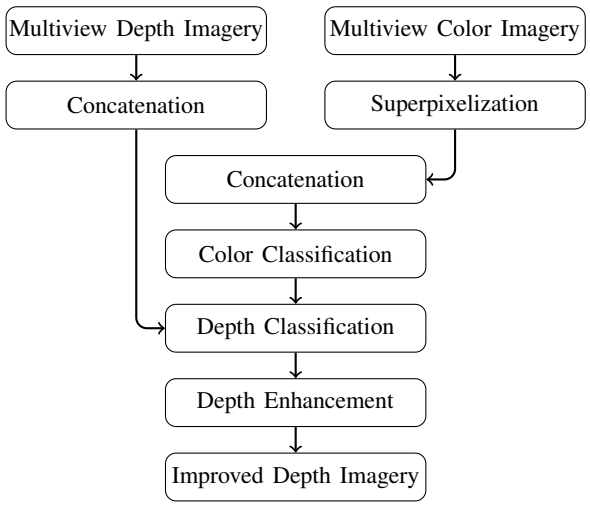


Fig. 2. Probabilistic multiview depth image enhancement.

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 data. In our initial work [34] and [41], the resulting mean of each subcluster has been used for enhancement. In the present work, Bayesian inference for depth subclustering provides us a way to gain insight in inter-view depth inconsistencies at multiple viewpoints. The resulting responsibilities act as probabilistic weights for the input depth pixels. This is the third main contribution of our paper. Finally, our proposed multiview depth enhancement algorithm can be used either on a server or on a client to improve the rendering quality and the interactive free-viewpoint experience. Experimental results on virtual view rendering will demonstrate the advantage of PROMDE.

The rest of the paper is organized as follows: Section II describes our multiview depth image enhancement. Section III presents the experimental assessment of our approach. Our conclusions are summarized in Section IV.

II. Multiview Depth Image Enhancement

Our proposed multiview depth image enhancement framework mainly consists of four steps: (1) superpixelization, (2) multiview color classification, (3) depth classification, and (4) depth image enhancement as depicted in Fig. 2. We assume that the multiview video imagery of pixel resolution $R{=}H \times W$ is independently captured for a given natural dynamic scene using projective cameras at N discrete viewpoints, where each image is H pixels in height and W pixels in width. Usually, each captured view of the scene is an image in YUV color space [46]. To make the procedure insensitive to the absolut luminance, we use a chromatic color representation [47]. We

Fig. 3. Two concatenated Newspaper views with approximately M=1000 superpixels as obtain by using SLIC [44]. The color space has been changed to RGB for better visualization. (Best viewed in color.)

transform these views from YUV color space to the XYZ color space. In this space, the virtual primary colors are denoted by X, Y, and Z, respectively [48]. The chromaticity of a pixel in a view at any viewpoint n is described by a vector of three coefficients, i.e., $\mathbf{v}_{nr} = [x_1,\ x_2,\ x_3]^{\mathsf{T}}$, $x_k \in [0, 1]$, whose entries sum to one, here $n = 1, \ldots, N$, and $r = 1, \ldots, R$. The chromaticity coefficients are defined as [48]

$$x_1 = \frac{X}{X+Y+Z},\ x_2 = \frac{Y}{X+Y+Z},\ x_3 = \frac{Z}{X+Y+Z}. \tag{1}$$

In the following, we will explain and motivate the individual steps of our approach in detail.

A. Superpixelization and Concatenation of View Imagery

The computational complexity of most classification algorithms is proportional to the number of feature samples. Our earlier approaches in [34] and [41] mainly use image color pixel vectors as features for color classification. This demands high computational resources, specially for high resolution imagery. Generally, there is no prior knowledge on the observed data. For proper classification, we need a sufficient amount of feature vectors such that the learning can be performed from the data under consideration. This may incur additional computational demand. Thus, in order to decrease this computational demand by reducing the number of feature samples for classification, we need to find an efficient way to compute image features which capture image redundancy efficiently while adhering well to object boundaries.

One way is to group image pixels perceptually into atomic regions, known as superpixels (e.g., [44]). The use of superpixels as a preprocessing step before classification, significantly decreases the number of feature samples while preserving the classification accuracy [44]. The interest in using superpixels is increasing in the computer vision community. As a result, many superpixel algorithms are proposed [44]. In this paper, we use Simple Linear Iterative Clustering (SLIC) to generate superpixels [44]. This algorithm is one of the state-of-the-art methods and an adaption of the K-means.

The SLIC algorithm performs local clustering of image pixels in the 5D feature space which is defined by the values L, a, b of the CIE Lab color space and the x, y pixel coordinates. For the clustering, the algorithm uses the idea of iteratively evolving local clusters and cluster centers, which is a special case of K-means with two major modifications. First, a novel weighted distance measure is proposed in [44], which combines color and spatial proximity, and provides control over the size and compactness of the superpixels. Second, SLIC limits the search space to a region proportional to the superpixel size during the distance calculation. This allows the algorithm to achieve a computational complexity which is linear in the number of pixels and independent of the number of superpixels.

Let a view at viewpoint n be represented by a set of R pixels $\mathbf{V}_n = \{\mathbf{v}_{nr}\}_{r=1}^{R}$, then the SLIC algorithm returns a set of superpixels $\mathbf{S}_n = \{\mathbf{s}_{n\omega}\}_{\omega=1}^{\Omega_n}$ with Ω_n as the number of desired superpixels in $\mathbf{V}_n$. Each superpixel $\mathbf{s}_{n\omega}$ is a set of pixels with index ω in view n. Similar to the K-means, once each pixel has been associated to a superpixel center, i.e., the mean color vector, it will be assigned this mean color vector which is a point in xyz chromaticity space $\mathbf{s}_{n\omega} = [x_1,\ x_2,\ x_3]^{\mathsf{T}}$.

In order to describe each color uniquely in a given natural scene through classification, we exploit the inherent inter-view similarity in the acquired multiview imagery from N viewpoints. For this, all N views with a desired number of superpixels Ω are concatenated to a single view with $N\Omega$ superpixels. This can be represented by the following superset of the superpixel mean color vectors from all N views

$$\mathbf{S} = \{\mathbf{S}_n\}_{n=1}^{N}. \tag{2}$$

By reassigning labels in (2), the concatenated image can be represented by the following set

$$\mathbf{S} = \{\mathbf{s}_m\}_{m=1}^{M} \tag{3}$$

where $m = n\omega$ and $M = \sum_{n=1}^{N} \Omega_n$ is the number of total superpixels in the concatenated views from N viewpoints. Fig. 3 is an example of such concatenated views from two different viewpoints with M superpixels, where each pixel is replaced by its corresponding superpixel vector $\mathbf{s}_m$.

B. Multiview Color Classification

In order to exploit the inherent per-pixel association between multiview view and multiview depth imagery for improving the depth at multiple viewpoints simultaneously, the underlying color clusters in the multiview view imagery need to be known. This is facilitated by color classification of the multiview imagery. In this subsection, we discuss the details of color classification. In the subsequent subsection, the results of color classification are utilized to classify depth images at multiple viewpoints using the per-pixel association between color and depth pixels. In general, the goal of classification is to assign each input data to one of a finite number of discrete groups of similar examples within the data, know as clusters [40]. Intuitively, a cluster comprises a group of data points whose inter-point distances are smaller when compared with the distances to points outside of the cluster [40].

1) Classification Methods: In a broad sense, we can categorize classification approaches to be either discriminative (e.g., [49]) or generative (e.g., [50]). Within discriminative approaches, the K-means enjoys the status of one of the simplest and most popular clustering algorithms, but it also suffers from two major drawbacks [40]. First, it does not

consider the spatial proximity of different superpixels, and second, it assumes a known number of clusters. [40].

Clustering methods with generative models are of special interest. In many cases, we can incorporate domain knowledge to uncover clusters with desirable patterns. In clustering algorithms based on generative models, the clustering task is performed by modelling the underlying distribution of the data with a mixture of known distributions. The parameters of the mixture model are often estimated by maximum likelihood estimation which involves often an application of the expectation-maximization algorithm. An example for a popular generative model is the mixture of Gaussian distributions whose effect is analogous to the use of Euclidean-type distances as the chosen measure of distortion from the discriminative point of view. Although this framework considers spatial proximity, it has its own limitations: 1) it suffers from singularities when one of the Gaussian components collapses onto a specific data point, for example, the log-likelihood function goes to infinity; 2) it suffers from over-fitting; and 3) similar to the K-means algorithm, the number of clusters has to be known. Usually in practice, the number of clusters is unknown and determining it imposes its own challenges.

In Bayesian inference for generative clustering, the number of clusters is treated as a random variable together with the parameters of the mixture components (e.g., [51]). As the exact inference is not analytically tractable, we have to resort to approximations. Two most prominent strategies in statistics and machine learning are Markov chain Monte Carlo (MCMC) sampling and variational inference [40], [52], [53]. In MCMC sampling, we collect samples from the exact posterior while the approximation arises from the use of a finite number of samples due to limited computational resources [40, Ch.11]. However, MCMC methods may converge slowly and their convergence can be difficult to diagnose. In practice, this often limits their use in small-scale problems. Variational inference based methods provide an alternative to computationally costly sampling-based methods. Variational inference replaces sampling and gives a deterministic approximation to the posterior distribution [54]. In this paper, we employ a generative clustering approach which uses variational inference for parameter estimation.

2) Dirichlet Mixture Models with Variational Inference: Our goal is to classify the chromaticity of superpixels. So, our choice of the generative model for classification is based on the properties of the feature domain. Each feature of a superpixel $\mathbf{s}_m$ is a three-dimensional vector in xyz chromaticity space which contains only nonnegative elements. These elements are in the interval [0,1] and sum to one. Obviously, $\mathbf{s}_m$ is not Gaussian distributed. Based on the properties of $\mathbf{s}_m$, a more reasonable choice is to model the underlying distribution of $\mathbf{s}_m$ by a Dirichlet distribution (e.g., [40]). The Dirichlet distribution has a probability density function of the form

$$\mathrm{Dir}\left(\mathbf{s}_m|\mathbf{u}\right) = \frac{\Gamma\left(\sum_{k=1}^{K} u_k\right)}{\prod_{k=1}^{K}\Gamma\left(u_k\right)} \prod_{k=1}^{K} x_k^{u_k-1},\ u_k > 0, \quad (4)$$

where $K = 3$, $\sum_{k=1}^{K} x_k = 1$, $0 \le x_k \le 1$, $\mathbf{u} = [u_1, \ldots, u_K]^{\mathsf{T}}$ is the parameter vector, and $\Gamma(\cdot)$ is the Gamma function. The

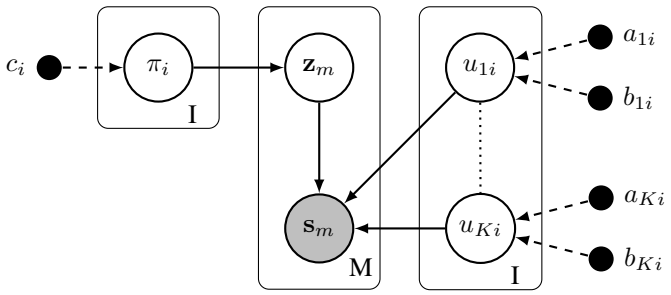


Fig. 4. Directed acyclic graph to represent the relationship of the variables in the Bayesian inference of a DMM, where the parameters of Dirichlet distributions are assumed to be mutually independent. Nodes denote random variables, edges denote possible dependencies, and plates denote replications.

distribution is parameterized by the parameter vector $\mathbf{u}$. When $u_k > 1$, it is unimodally distributed. Here, we consider only a Dirichlet distribution with all its parameters greater than one. We use a finite mixture of multivariate Dirichlet distributions to capture the underlying distribution of $\mathbf{S}$ as [55]

$$p\left(\mathbf{S}|\mathbf{\Pi}, \mathbf{U}\right) = \prod_{m=1}^{M}\sum_{i=1}^{I} \pi_i \mathrm{Dir}\left(\mathbf{s}_m|\mathbf{u}_i\right),\ 0 \le \pi_i \le 1, \quad (5)$$

where I denotes the number of mixture components, $\mathbf{\Pi} = [\pi_1, \ldots, \pi_I]^{\mathsf{T}}$ is the vector of mixture weights with $\sum_{i=1}^{I} \pi_i$=1, and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_I]$ is the parameter matrix. In the following, we adopt the Bayesian approach with variational inference to estimate the model parameters [42]. This leads to an analytically tractable solution that can be easily used in practice.

To facilitate Bayesian estimation with variational inference for Dirichlet mixture models, the following conjugate priors are introduced over π_i and $\mathbf{u}_i$ as

$$\pi_i \sim \mathrm{Dir}\left(c_{0,i}\right), \quad (6)$$

$$u_{ki} \sim \mathrm{Gam}\left(a_{0,ki}, b_{0,ki}\right). \quad (7)$$

In the above expression, $\mathrm{Dir}\left(c_{0,i}\right)$ is the Dirichlet distribution with $c_{0,i}$ as the hyperparameter for the prior distribution over π_i. $\mathrm{Gam}\left(a_{0,ki}, b_{0,ki}\right)$ is the Gamma distribution with the shape parameter $a_{0,ki}$ and the inverse scale parameter $b_{0,ki}$, which are regarded as the hyperparameters for the prior distribution over u_{ki}.

In this model, an I-dimensional indication vector $\mathbf{z}_m = [z_{m1}, \ldots, z_{mI}]^{\mathsf{T}}$ is assigned to each $\mathbf{s}_m \in \mathbf{S}$. Only one element in the indication vector is equal to 1 and the remaining elements are zeros. Thus, $z_{mi} = 1$ indicates that the mth superpixel is generated from the ith mixture component. Therefore, $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_M]$ is the indication matrix for M superpixels. Fig. 4 shows a graphical representation of the relationship between the variables in the Bayesian inference of a Dirichlet mixture model.

By treating all the parameters in (5) as random variables and assuming that $\mathbf{S}$ is conditionally independent of $\mathbf{\Pi}$ given $\mathbf{Z}$, while having $\mathbf{Z}$ independent of $\mathbf{U}$, the conditional probability of $\mathbf{S}$ given $\{\mathbf{Z}, \mathbf{U}\}$ can be written as

$$p\left(\mathbf{S}|\mathbf{Z}, \mathbf{U}\right) = \prod_{m=1}^{M}\prod_{i=1}^{I} \mathrm{Dir}\left(\mathbf{s}_m|\mathbf{u}_i\right)^{z_{mi}}. \quad (8)$$

By considering the indication $\mathbf{Z}$ as random, the conditional probability of $\mathbf{Z}$ given $\mathbf{\Pi}$ is

$$p\left(\mathbf{Z}|\mathbf{\Pi}\right)=\prod_{m=1}^{M}\prod_{i=1}^{I}\pi_{i}^{z_{mi}}. \tag{9}$$

Regarding our model description, we can represent the joint distribution of $\mathbf{S}$ and all the latent variables $\{\mathbf{U},\mathbf{\Pi},\mathbf{Z}\}$ by

$$p\left(\mathbf{S},\mathbf{U},\mathbf{\Pi},\mathbf{Z}\right)=p\left(\mathbf{S}|\mathbf{Z},\mathbf{U}\right)p\left(\mathbf{Z}|\mathbf{\Pi}\right)p\left(\mathbf{\Pi}\right)p\left(\mathbf{U}\right). \tag{10}$$

Working with the exact posterior is not analytically tractable. The idea behind variational methods is to approximate $p(\mathbf{U},\mathbf{\Pi},\mathbf{Z}\mid\mathbf{S})$ with a distribution $q(\mathbf{U},\mathbf{\Pi},\mathbf{Z})$ that belongs to a constrained family of distributions, indexed by a variational parameter. The goal is to choose a member of that family which is as close as possible to the exact posterior distribution [54]. In variational inference this is carried out by maximizing a lower bound introduced on the log marginal likelihood (the model evidence) [54]. In this work, we follow a framework as described in [42] for maximizing the variational lower bound and deriving the required sequential update equations. We proceed by providing a summary of the method.

By considering a fully factorized approximation for the latent switch variables $\mathbf{Z}$, the component parameters $\mathbf{\Pi}$ and $\mathbf{U}$, then the variational distribution $q(\mathbf{U},\mathbf{\Pi},\mathbf{Z})$ can be expressed as, [42],

$$q(\mathbf{U},\mathbf{\Pi},\mathbf{Z})=q(\mathbf{U})q(\mathbf{\Pi})q(\mathbf{Z}). \tag{11}$$

Optimization of the variational factors $q(\mathbf{Z})q(\mathbf{\Pi})q(\mathbf{U})$ is performed by maximizing the lower bound on the model evidence. The lower bound is given by

$$\mathcal{L}=\mathbb{E}_{q}\left[\ln p(\mathbf{S},\mathbf{U},\mathbf{\Pi},\mathbf{Z})\right]-\mathbb{E}_{q}\left[\ln q(\mathbf{U},\mathbf{\Pi},\mathbf{Z})\right]. \tag{12}$$

The operator $\mathbb{E}_{q}\left[\cdot\right]$ takes the expectation of variables in its argument with respect to the variational variable distribution $q(\cdot)$. The optimization of the variational posterior distribution $q(\mathbf{Z})q(\mathbf{\Pi},\mathbf{U})$ involves cycling between optimization of $q(\mathbf{Z})$ and $q(\mathbf{\Pi},\mathbf{U})$, which is analogous to the expectation and the maximization steps in the maximum-likelihood expectation-maximization algorithm.

First, we use the current distributions over the model parameters to evaluate the responsibilities $r_{mi}^{(S)}$ as

$$r_{mi}^{(S)}=\mathbb{E}_{q}\left[z_{mi}\right]=\frac{\rho_{mi}}{\sum_{j=1}^{I}\rho_{mj}}, \tag{13}$$

$$\begin{aligned}\ln\rho_{mi}=&\,\psi\left(c_{i}\right)-\psi\left(\mathbf{c}^{\mathsf{T}}\mathbf{1}_{I}\right)+\left(\mathbf{u}_{i}-\mathbf{1}_{K}\right)^{\mathsf{T}}\ln\mathbf{s}_{m}\\&+\sum_{k=1}^{K}\left[\psi(\sum_{k=1}^{K}\overline{u}_{ki})-\psi(\overline{u}_{ki})\right]\\&\times\overline{u}_{ki}\left(\mathbb{E}_{q}\left[\ln u_{ki}\right]-\ln\overline{u}_{ki}\right),\end{aligned} \tag{14}$$

$$\overline{u}_{ki}=\frac{a_{ki}}{b_{ki}}, \tag{15}$$

$$\mathbb{E}_{q}\left[\ln u_{ki}\right]=\psi\left(a_{ki}\right)-\ln b_{ki}, \tag{16}$$

where, $\overline{u}_{ki}$ is calculated from the previous iteration, $\psi(\cdot)$ is the digamma function defined as $\psi(x)=\frac{\partial\ln\Gamma(x)}{\partial x}$, the superscript S represents the superpixel, and $\mathbf{1}_{I}$ denotes an I dimensional vector with all elements equal to one. The resulting responsibilities are used to re-estimate the variational distribution over the parameters. The corresponding posteriors are given by

$$\pi_{i}\sim\text{Dir}\left(c_{i}\right), \tag{17}$$

$$u_{ki}\sim\text{Gam}\left(a_{ki},b_{ki}\right), \tag{18}$$

where

$$c_{i}=c_{0,i}+\sum_{m=1}^{M}r_{mi}^{(S)}, \tag{19}$$

$$b_{ki}=b_{0,ki}-\sum_{m=1}^{M}r_{mi}^{(S)}\ln s_{km}, \tag{20}$$

$$a_{ki}=a_{0,ki}+\sum_{m=1}^{M}r_{mi}^{(S)}\overline{u}_{ki}\left[\psi\left(\sum_{k=1}^{K}\overline{u}_{ki}\right)-\psi\left(\overline{u}_{ki}\right)\right]. \tag{21}$$

The procedure is guaranteed to converge as the lower bound is convex in each of the factors [56] and there is only one lower bound being maximized during the updating steps.

3) Color Classification: The responsibilities $r_{mi}^{(S)}$ play an important role in the classification as they express how responsible each mixture component is in explaining the data. In other words, each element $r_{mi}^{(S)}\in[0,1]$ represents the probability that $\mathbf{s}_{m}$ is generated from the ith cluster. Let $\mathcal{R}^{(S)}=[\mathbf{r}_{1}^{(S)},\ldots,\mathbf{r}_{M}^{(S)}]$ denote the responsibility matrix, where $\mathbf{r}_{m}^{(S)}=[r_{m1}^{(S)},\ldots,r_{mI}^{(S)}]^{\mathsf{T}}$ are non-negative and sum to one. Thus, we assign each superpixel to the cluster which gives the largest probability. Members of the ith cluster of superpixels $\mathcal{S}^{(i)}$ can be extracted from $\mathbf{S}$ as

$$\mathcal{S}^{(i)}=\{\underline{\mathbf{s}}_{m}^{(i)}\}_{m=1}^{M} \tag{22}$$

with

$$\underline{\mathbf{s}}_{m}^{(i)}=\begin{cases}\mathbf{s}_{m}, & \text{if } r_{mi}^{(S)}>r_{mj}^{(S)},\ \forall\ i\neq j,\ (i,j=1,\ldots,I);\\ \varnothing, & \text{otherwise,}\end{cases} \tag{23}$$

where $\varnothing$ is the empty set. Note that a superpixel is a set of perceptually grouped pixels. Therefore, a superpixel in a given cluster represents the pixels which belong to the specified superpixel. We further note that all pixels in a superpixel will be assigned the responsibility of that superpixel. Let $\mathcal{R}^{(C)}=[\mathbf{r}_{11}^{(C)},\ldots,\mathbf{r}_{NR}^{(C)}]$ denote the responsibility matrix, where, the superscript C represents the color vector pixel, $\mathbf{r}_{nr}^{(C)}=[r_{nr1}^{(C)},\ldots,r_{nrI}^{(C)}]^{\mathsf{T}}$ and each element $r_{nri}^{(C)}$ represents the probability that the color pixel $\mathbf{v}_{nr}\in\{\mathbf{V}_{n}\}_{n=1}^{N}$ which belongs to $\mathbf{s}_{m}^{(i)}$ is generated from the ith cluster. By assigning each color pixel to the cluster which gives the largest probability, we extract the members of the color cluster $\mathcal{C}^{(i)}$ from $\{\mathbf{V}_{n}\}_{n=1}^{N}$ as

$$\mathcal{C}^{(i)}=\{\mathbf{v}_{11}^{(i)},\ldots,\mathbf{v}_{NR}^{(i)}\}, \tag{24}$$

where

$$\mathbf{v}_{nr}^{(i)}=\begin{cases}\mathbf{v}_{nr}, & \text{if } r_{nri}^{(C)}>r_{nrj}^{(C)},\forall i\neq j\quad(i,j=1,\ldots,I);\\ \varnothing, & \text{otherwise.}\end{cases} \tag{25}$$

C. Multiview Depth Image Classification and Enhancement

With color clusters $\{\mathcal{C}^{(i)}\}_{i=1}^{I}$ for a given multiview imagery, we can improve the quality of depth maps and specially the inter-view depth consistency at multiple viewpoints, simultaneously. For this, depth images from N viewpoints are concatenated to a single depth $\mathbf{D} \in \mathbb{R}_{+}^{H\times NW}$ which can be represented as a set of all N depth images

$$\mathbf{D} = \{\mathbf{D}_n\}_{n=1}^{N}, \tag{26}$$

where $\mathbf{D}_n = \{d_{nr}\}_{r=1}^{R} \in \mathbb{R}_{+}^{H\times W}$ is the depth map at the viewpoint n which can be considered as a set of discrete depth values $d_{nr} \in \{0,\ldots,255\}$. For simplicity, we consider the following mapping

$$\mathbf{D} \in \mathbb{R}_{+}^{H\times NW} \mapsto \underline{\mathbf{D}} \in \mathbb{R}_{+}^{1\times NR}, \tag{27}$$

where $\underline{\mathbf{D}} = \{d_1,\ldots,d_{NR}\}$ is such that for each color pixel $\mathbf{v}_{nr}$, there is an associated depth value d_{nr}. In order to obtain members of the ith depth cluster $\mathcal{D}^{(i)}$, we utilize this per-pixel depth association with color pixels by defining

$$\mathcal{D}^{(i)} = \{\tilde{d}_1^{(i)},\ldots,\tilde{d}_{NR}^{(i)}\}, \tag{28}$$

where

$$\tilde{d}_{nr}^{(i)} = \begin{cases} d_{nr}, & \text{if } \; r_{nri}^{(C)} > r_{nrj}^{(C)}, \forall i \neq j \quad (i,j=1,\ldots,I); \\ \varnothing, & \text{otherwise.} \end{cases} \tag{29}$$

Fig. 5 shows such color clusters and the 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 natural scene.

In natural scene imagery, foreground and background object points can have a similar color, but foreground object points have different depth values when compared to background object points. This leads to ambiguities among the members of a depth cluster as compared to the associated color cluster. The members of $\mathcal{C}^{(i)}$ have similar colors, whereas members of $\mathcal{D}^{(i)}$ may have different depth values. For 1D parallel camera arrangements, a given object point with a given chromaticity, which is visible from N viewpoints, should have the same depth value in all N depth images. However, such points usually have different depth values in $\mathcal{D}^{(i)}$ due to the inter-view inconsistency across multiple viewpoints. This adds additional ambiguity to depth clusters and motivates us to resolve this by considering further subclassification of each depth cluster $\mathcal{D}^{(i)}$.

We consider a way to gain insight about inter-view inconsistencies at multiple viewpoints by using the depth responsibilities for each depth pixel. We notice that inconsistent depth values for an observed object point at multiple viewpoints will be assigned lower responsibilities when compared to consistent depth values. Further, we need to put emphasis on depth pixel values as well as their positions at multiple viewpoints. For this purpose, we define the following feature domain

$$\widehat{\mathcal{D}}^{(i)} \;\; = \;\; \{\mathbf{d}_\phi\}_{\phi\in\Phi^{(i)}}, \tag{30}$$

where each feature vector $\mathbf{d}_\phi = [d_\phi,\ h_\phi,\ w_\phi]^{\mathsf{T}}$ consists of depth pixel value $d_\phi \in \mathcal{D}^{(i)}\ \forall\ \phi$, and its location information $h_\phi \in \{1,\ldots,H\}\ \forall\ \phi$ and $w_\phi \in \{1,\ldots,W\}\ \forall\ \phi$ with respect to the viewpoint to which d_ϕ belongs. The set $\Phi^{(i)}$ denotes the set of indices of members of $\mathcal{D}^{(i)}$. Note that the defined feature vector $\mathbf{d}_\phi$ is not limited by the number of images from different viewpoints. It is only limited by the constraints of the 1D parallel camera arrangement. However, our method can easily adapt to other multi-camera arrangements such as the 2D camera array or the circular camera array by defining a new feature vector in an appropriate feature domain. In [44], the authors extended the idea of superpixels to compute 3D SLIC supervoxels for a video sequence. Our proposed depth enhancement framework can also be extended to improve temporal depth coherence by applying the 3D supervoxels on concatenated temporally successive frames and defining feature vectors that reflect the constraints of scene geometry, object point motion, and visibility [57].

As the elements of the feature vector $\mathbf{d}_\phi$ are discrete geometric values sampled from a continuous distribution with quantization noise, we model the underlying distribution of $\widehat{\mathcal{D}}^{(i)}$ by a mixture of multivariate Gaussian distributions as

$$p(\widehat{\mathcal{D}}^{(i)} \mid \boldsymbol{\mu},\boldsymbol{\Lambda},\tau) = \prod_{\phi\in\Phi^{(i)}} \sum_{l=1}^{L} \tau_l \mathcal{N}(\mathbf{d}_\phi \mid \boldsymbol{\mu}_l, \boldsymbol{\Lambda}_l^{-1}), \tag{31}$$

where L is the number of mixture components, $\tau = [\tau_1,\ldots,\tau_L]^{\mathsf{T}}$, $0 \leq \tau_l \leq 1$, $\sum_{l=1}^{L}\tau_l = 1$, denotes the vector of mixture weights, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_L]^{\mathsf{T}}$ represents the mean vectors, and $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1,\ldots,\boldsymbol{\Lambda}_L]^{\mathsf{T}}$ denotes the precision (i.e., inverse covariance) matrices. The model parameters are estimated in a Bayesian framework by variational inference [40, Ch.10]. In the first step, the following priors are used for the model parameters

$$\tau_l \sim \mathrm{Dir}(\alpha_{0,l}), \tag{32}$$

$$\boldsymbol{\Lambda}_l \sim \mathcal{W}(\mathbf{W}_{0,l}, \nu_{0,l}), \tag{33}$$

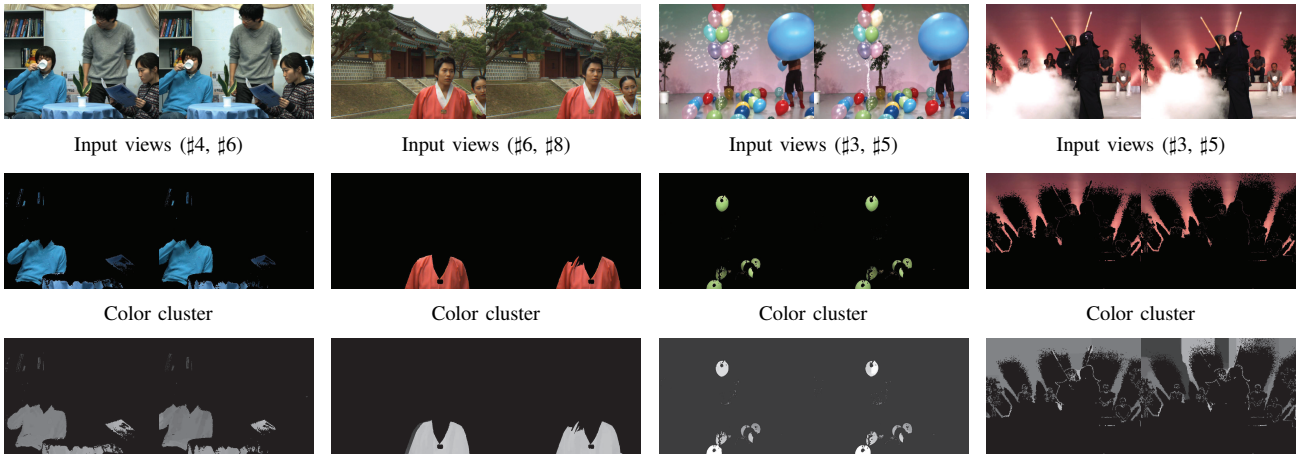
$$\boldsymbol{\mu}_l \sim \mathcal{N}(\boldsymbol{m}_{0,l}, (\beta_{0,l}\boldsymbol{\Lambda}_l)^{-1}). \tag{34}$$

In the above expressions, $\mathrm{Dir}(\alpha_{0,l})$ is the symmetric Dirichlet distribution of dimension L, with the hyperparameter $\alpha_{0,l}$. The Dirichlet distribution is the conjugate prior of the categorical distribution. $\mathcal{W}(\mathbf{W}_{0,l},\nu_{0,l})$ is the Wishart distribution with scale matrix $\mathbf{W}_{0,l}$ and degree of freedom $\nu_{0,l}$, which is the conjugate prior of the precision matrix for a multivariate Gaussian distribution. Associated with each observation $\mathbf{d}_\phi$, there is a corresponding latent switch variable $\mathbf{z}_\phi = [z_{\phi 1},\ldots,z_{\phi L}]^{\mathsf{T}}$ consisting of the binary elements $z_{\phi l}$. Indicating the set of switch variables by $\underline{\mathbf{Z}} = \{\mathbf{z}_\phi\}_{\phi\in\Phi^{(i)}}$, the conditional distribution of $\underline{\mathbf{Z}}$ given τ is

$$p(\underline{\mathbf{Z}} \mid \tau) = \prod_{\phi\in\Phi^{(i)}} \prod_{l=1}^{L} \tau_l^{z_{\phi l}}. \tag{35}$$

Again, we use first the current distributions over the model parameters to evaluate the responsibilities $r_{\phi l}^{(D)}$ for i as

$$r_{\phi l}^{(D)} = \frac{\varrho_{\phi l}}{\sum_{\varsigma=1}^{L}\varrho_{\phi\varsigma}} \tag{36}$$



(a) Newspaper. (b) Lovebird1. (c) Balloons. (d) Kendo.

Fig. 5. Example of color classification and corresponding depth classification. Concatenated imagery from two viewpoints is shown. The classification of depth pixels is achieved by utilizing the per-pixel association of depth pixels with color pixels. (Best viewed in color.)

$$\ln \varrho_{\phi l} = \mathbb{E}_q[\ln \pi_l] + \frac{1}{2}\mathbb{E}_q[\ln |\mathbf{\Lambda}_l|] - \frac{\Delta}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_q[(\mathbf{d}_\phi - \boldsymbol{m}_l)^\mathsf{T}\mathbf{\Lambda}_l(\mathbf{d}_\phi - \boldsymbol{m}_l)], \tag{37}$$

where $\Delta = 3$ which is the dimension of the feature vector $\mathbf{d}_\phi$. Next, these responsibilities are employed to re-estimate the variational distribution over the parameters. The corresponding posteriors are given by

$$\tau_l \sim \mathrm{Dir}(\alpha_l), \tag{38}$$

$$\mathbf{\Lambda}_l \sim \mathcal{W}(\mathbf{W}_l, \nu_l), \tag{39}$$

$$\boldsymbol{\mu}_l \sim \mathcal{N}(\boldsymbol{m}_l, (\beta_l \mathbf{\Lambda}_l)^{-1}), \tag{40}$$

where

$$\alpha_l = \alpha_{0,l} + F_l, \;\; F_l = \sum_{\phi \in \Phi^{(i)}} r_{\phi l}^{(D)} \tag{41}$$

$$\nu_l = \nu_{0,l} + F_l, \tag{42}$$

$$\beta_l = \beta_{0,l} + F_l, \tag{43}$$

$$\bar{\mathbf{d}}_l = \frac{1}{F_l}\sum_{\phi \in \Phi^{(i)}} r_{\phi l}^{(D)} \mathbf{d}_\phi, \tag{44}$$

$$\mathbf{G}_l = \frac{1}{F_l}\sum_{\phi \in \Phi^{(i)}} (\mathbf{d}_\phi - \bar{\mathbf{d}}_l)(\mathbf{d}_\phi - \bar{\mathbf{d}}_l)^\mathsf{T}, \tag{45}$$

$$\mathbf{m}_l = \frac{1}{\beta_l}(\beta_{0,l}\boldsymbol{m}_{0,l} + F_l \bar{\mathbf{d}}_l), \tag{46}$$

$$\mathbf{W}_l^{-1} = \mathbf{W}_{0,l}^{-1} + F_l \mathbf{G}_l + \frac{\beta_{0,l} F_l}{\beta_{0,l} + F_l}(\bar{\mathbf{d}}_l - \boldsymbol{m}_{0,l})(\bar{\mathbf{d}}_l - \boldsymbol{m}_{0,l})^\mathsf{T}. \tag{47}$$

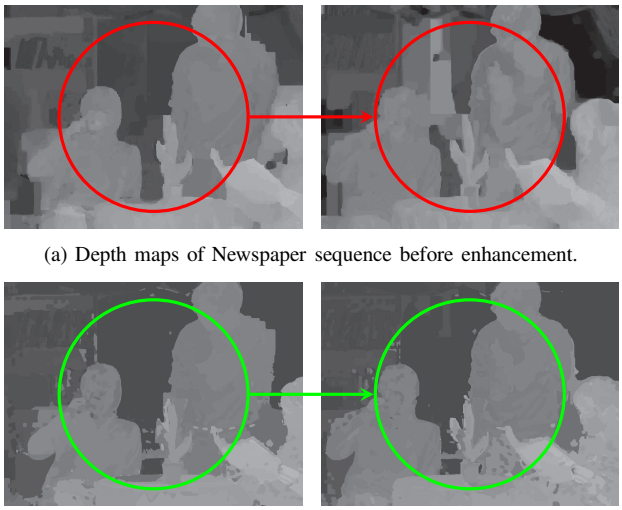
For detail, please see [40].

Let $\mathcal{R}^{(D)} = [\mathbf{r}_\phi^{(D)}]_{\phi \in \Phi^i}$ represent the responsibility matrix for all depth pixels which are obtained by subclassification of the depth cluster $\mathcal{D}^{(i)}$ using variational inference with a Gaussian mixture model, where $\mathbf{r}_\phi^{(D)} = [r_{\phi 1}^{(D)}, \ldots, r_{\phi L}^{(D)}]^\mathsf{T}$. Each element $r_{\phi l}^{(D)}$ represents the probability that $\mathbf{d}_\phi$ is generated from the lth cluster of the depth values. Thus, we assign each depth pixel d_ϕ in $\mathcal{D}^{(i)}$ to the depth subcluster which gives the largest probability. Members of the lth depth subcluster $\mathcal{D}^{(il)}$ can be extracted from $\mathcal{D}^{(i)}$ as

$$\mathcal{D}^{(il)} = \{\tilde{d}_\phi^{(il)}\}_{\phi \in \Phi^i} \tag{48}$$

with

$$\tilde{d}_\phi^{(il)} = \begin{cases} d_\phi, & \text{if } r_{\phi g}^{(D)} > r_{\phi h}^{(D)}, \; \forall \; g \neq h, \; (g, h = 1, \ldots, L); \\ \varnothing, & \text{otherwise.} \end{cases} \tag{49}$$



(a) Depth maps of Newspaper sequence before enhancement.

(b) Depth maps of Newspaper sequence after enhancement.

Fig. 6. Example for inter-view consistent depth maps as obtained by using our probabilistic depth enhancement for the MPEG Newspaper sequence. After the enhancement in (b), more regions in the depth maps are consistent when compare to the original MPEG depth maps without enhancement. The red circles mark areas with inter-view inconsistency in the depth maps before enhancement. The green circles mark corresponding areas with improved inter-view consistency after enhancement. (Best viewed in color.)

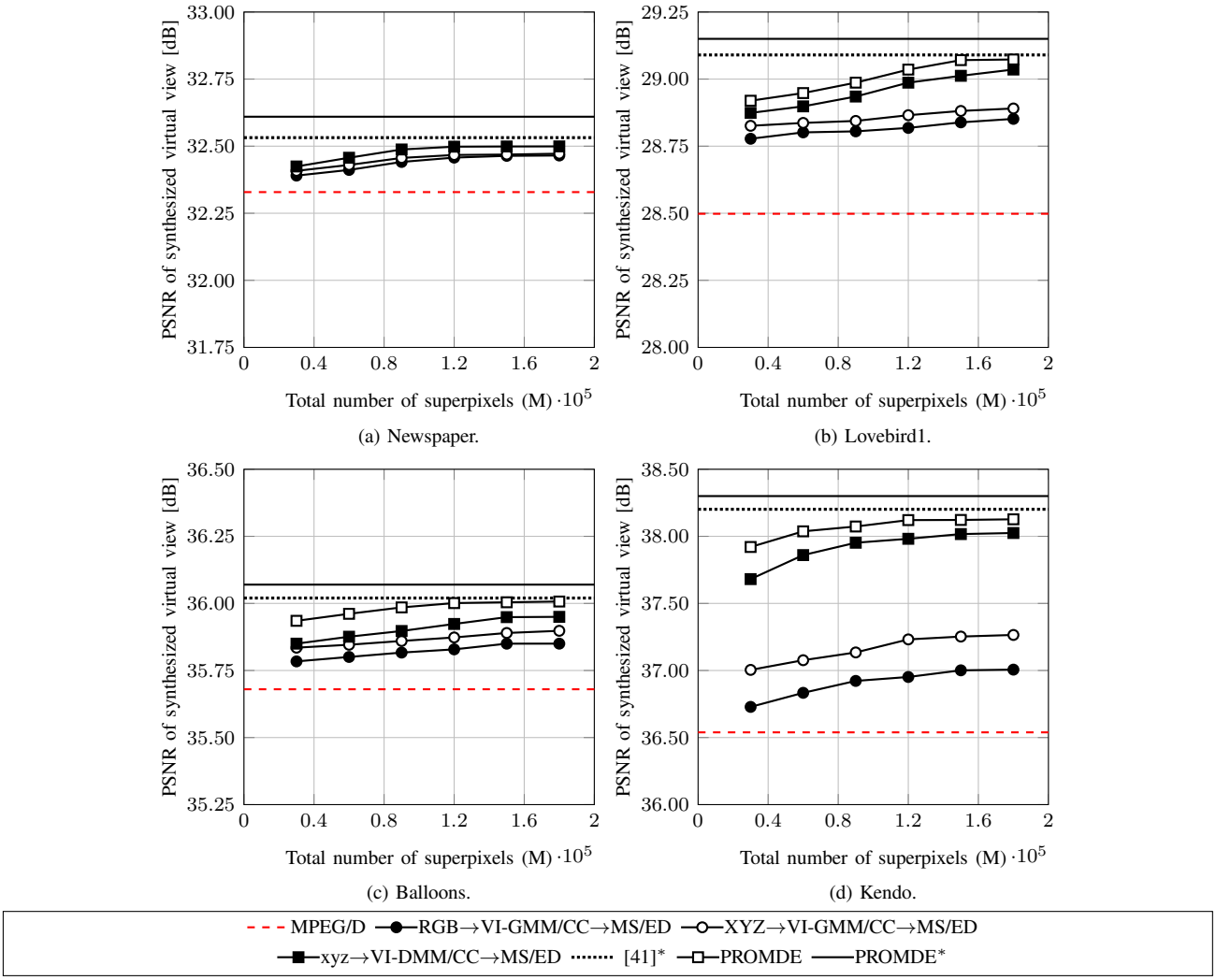


Fig. 7. Total number of superpixels vs. objective quality of virtual views as generated by VSRS 3.5 using the small baseline setting as given in Table I. The virtual views as obtained by using MPEG depth images without enhancement (MPEG/D) are used as the baseline for comparison. The use of depth images enhanced by PROMDE improves the quality of view synthesis when compared to the enhancement approach xyz→VI-DMM/CC→MS/ED, where the enhanced depth (ED) is generated by mean-shift (MS) rather than VI-GMM. This inferior approach is similar to our previous work [41] which does not use superpixels. Two more enhancement schemes which use VI-GMM color classification are used for comparison: RGB→VI-GMM/CC→MS/ED uses color clusters in RGB color space and the resulting depth is processed by mean-shift. XYZ→VI-GMM/CC→MS/ED utilizes color classification in XYZ color space while reusing mean-shift for depth subclustering. The results of the algorithm in [41] and the proposed algorithm without superpixels are represented by [41]* and PROMDE*, respectively. Note, all enhancement schemes use superpixels as pre-processing step before color classification to lower the computational complexity, except the schemes marked by an asterisk $(*)$.

We consider a set $\Upsilon^{(il)}$ which denotes the set of indices of members of $\mathcal{D}^{(il)}$. We replace the depth values in the depth subcluster $\mathcal{D}^{(il)}$ by the responsibility-weighted mean

$$\widehat{d}^{(il)} = \frac{\sum_{v \in \Upsilon^{(il)}} r_v^{(D)} d_v}{\sum_{v \in \Upsilon^{(il)}} r_v^{(D)}}, \tag{50}$$

where $r_v^{(D)}$ is the largest responsibility of the depth pixel d_v. Note that the simple mean of all depth values within $\mathcal{D}^{(il)}$ would be sensitive to depth inconsistency and noise.

The method is guaranteed to converge. This can be explained as follows. Using variational inference, we have an explicit lower bound on the marginal likelihood of data (evidence). During optimization, the lower bound value increases with each iteration. Further, the learning procedure is guaranteed to converge as the lower bound is convex in each of the factors (model parameters) [40, Ch.10]. In fact this is one of the key advantages of using variational inference.

III. Experimental Results

FTV experience allows the user to 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. The inconsistent and inaccurate depth values from different viewpoints affect the position and hence, intensity of the virtual view pixels [24]. Therefore, the quality of depth images has direct impact on view synthesis [24]. The quality of views rendered by DIBR-based view synthesis algorithms can be significantly improved by the improving the quality of the depth imagery [25]. Thus, rendering results provide a means to evaluate the inter-view

inconsistency in the multiview depth imagery. Hence, the performance of the proposed probabilistic depth enhancement is assessed through the effect on the subjective and objective quality of virtual views. In the experiments, we use four multiview test sets and the corresponding multiview depth imagery as provided by MPEG [18]: Newspaper, Lovebird1, Balloons, and Kendo. The spatial resolution of the test imagery is 1024×768. The MPEG multiview depth imagery has the same resolution and is estimated by a passive depth sensing method [58]. Each evaluation experiment for the proposed scheme mainly consists of two steps: (1) improvement of depth images at multiple viewpoints using the proposed approach and (2) virtual view synthesis with the help of the improved depth images. For this purpose, the MPEG view synthesis reference software (VSRS) is employed [58]. This reference software 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 [59]. The virtual views are generated by using the 1D parallel synthesis mode of VSRS 3.5 with half-pel precision.

TABLE I
EXPERIMENTAL MVV AND MVD VIEWPOINT SETTINGS.

MPEG Data	Small Baseline		Large Baseline	
	Input Views	Virtual View	Input Views	Virtual Views
Newspaper	♯4, ♯6	♯5	♯3, ♯7	♯4, ♯5, ♯6
Lovebird1	♯6, ♯8	♯7	♯4, ♯8	♯5, ♯6, ♯7
Kendo	♯3, ♯5	♯4	♯1, ♯5	♯2, ♯3, ♯4
Balloons	♯3, ♯5	♯4	♯1, ♯5	♯2, ♯3, ♯4

Due to the input requirements of VSRS 3.5, we restrict our evaluation experiments to improve depth images at two viewpoints, i.e., $N = 2$, for each multiview test imagery. For this, our algorithm starts with a specified number of superpixels Ω. The number mixture components for both VI-DMM color classification and VI-GMM depth subclassification is initialized by a large value, for example, $I = L = 100$. The resulting responsibilities from depth subclassification are used to improve the depth estimates at two chosen viewpoints. Fig. 6 shows an example of depth maps improved by our proposed method. Next, a virtual view at a given viewpoint is synthesized by using VSRS 3.5 with the improved depth images. The objective quality of these synthesized views is measured in terms of the peak signal-to-noise ratio (PSNR) with respect to the captured view from a real camera at the same viewpoint. We consider the quality of virtual views generated by using MPEG depth maps and enhanced MPEG depth maps by our pervious work [41] with superpixels. The experiments are performed using two different camera baseline settings: the small camera baseline and the large camera baseline. For each test data set, Table I shows the experimental setting of the input viewpoints and the corresponding synthesized virtual viewpoints for the two camera baseline settings.

A. Small Camera Baseline Setting

For the small baseline setting as given in Table I, Fig. 7 shows the average luminance PSNR (in dB) of virtual views over the number of total superpixels for the four test sets. The quality of views obtain by using enhanced MPEG depth images is better than the quality of views synthesized by using MPEG depth images without enhancement. Graphs show that probabilistic enhancement of depth imagery is advantageous. When compared to [41], the advantage by pre-processing with superpixels. Further, we noticed that PROMDE offers improvements over mean-shift subclustering of depth. The choice of the generative model and the color space for classification also influence the depth enhancement. Color classification using VI-DMM outperforms classification based on VI-GMM. The graphs in Fig. 7 also indicate that the choice of color space affects the performance of depth enhancement. For example, depth images improved by using VI-GMM color classification in XYZ space give better results than VI-GMM in RGB space. For both methods, superpixels and mean-shift depth subclassification are used. The performance of our proposed probabilistic depth enhancement even without superpixels improves the quality of depth maps significantly when compared with the [41] and MPEG depth images. When comparing the improvements among the multivew test sets, note that they depend on the quality of the input reference depth maps.

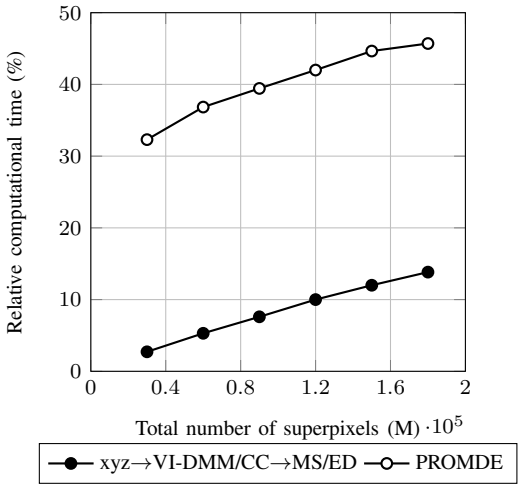


Fig. 8. Total number of superpixels vs. relative computational time in percent with respect to the computational time of [41] without superpixels. The performance of our algorithm PROMDE is shown. The method xyz→VI-DMM/CC→MS/ED denotes the extension of [41] by superpixels. Here, the relative computational time is defined as the ratio of the computational time required to enhance depth maps by our depth enhancement algorithm which uses superpixels as pre-processing to the computational time required to enhance the same depth maps by using the method as proposed in [41]. The number of initial clusters and the number of iterations are the same in all experiments. The same concatenated views from two different viewpoints are used in all experiments. The resolution of each view is 1024×768 pixels. With this imagery, the computational time of the algorithm [41] is approximately 90 min using a MATLAB implementation on a 64-bit Windows operating system and an Intel Core i7 CPU.

We observe that our algorithm does offer objective gains for the test data. The gain for Kendo is the largest. This gain mainly comes from the VI-DMM color classification. For

Kendo, the VI-DMM classification returns a well-clustered background. The corresponding depth clusters, as shown in Fig. 5, reflect a significant inter-view inconsistency. The mean-shift based enhancement efficiently improves the inter-view consistency by assigning mean values to specified clusters [41]. For PROMDE, depth values are replaced by the responsibility-weighted mean value in each specified cluster. This considers the contribution of each depth value within the cluster. The graphs in Fig. 8 show that pre-processing with superpixels reduces the computational time significantly when compared to [41].

The visual quality of virtual views is critical for the comfort of FTV viewers. With the proposed probabilistic depth enhancement, we are able to perceptibly improve the quality of the virtual views. The improvements reported here come exclusively from the enhanced depth imagery. In particular, they remove artifacts around edges of objects as defined by chromaticity clusters. Fig. 9 compares virtual views of the test imagery. Improvements are highlighted and shown through the selected cropped regions. For Newspaper, the hand and the background are well synthesized with PROMDE. Visually annoying artifacts in Lovebird1, specially around the hair and the red jeogori sleeve of the man have been noticeably suppressed. The boundaries of the balloons are well synthesized in the Balloons test set. The synthesis quality for Kendo around the hakama and the trouser of the spectator is significantly improved, as shown in Fig. 9(d).

B. Large Camera Baseline Setting

The following experiments are devised to examine the performance and robustness of our proposed depth enhancement algorithm in large baseline scenarios. In these experiments, we first enhance the depth maps at two different viewpoints, say, n and $n+4$, using our probabilistic depth enhancement or [41] with superpixels. Second, virtual views at three different intermediate viewpoints $n+1$, $n+2$, and $n+3$ are synthesized. The three intermediate views help us to analyze the effect of the improved inter-view depth consistency on the quality of the views. Note that these experiments are performed with a fixed number of desired superpixels $M = 1.2 \times 10^5$. For each multiview test set, the large baseline setting in Table I is used. Fig. 10 shows the average luminance PSNR (in dB) of all three virtual views. In general, the bar plots show that all virtual views obtain by using our enhanced MPEG depth images offer better visual quality when compared to virtual views synthesized by using [41] with superpixels and MPEG depth maps without enhancement. Similar to the small baseline setting, the gain in quality of virtual views in the large baseline setting is also highly dependent on the quality of the input multiview depth imagery.

Fig. 11 emphasizes on the improved inter-view consistency across the viewpoints by showing selected cropped regions of the virtual views. In the synthesized virtual views of Newspaper using PROMDE, the head of the man and the background wall are more consistent across the different viewpoints when compared using the MPEG depth maps without enhancement. For Lovebird1, the artifacts around the red jeogori sleeve of the man have been noticeably suppressed across all virtual views as shown in Fig. 11(b). Using our enhanced depth maps for the Balloons imagery, we observe an improved consistency of the synthesized boundaries of the balloons in Fig. 11(c). For Kendo in Fig. 11(d), the visual artifact on the head of the spectator is consistently suppressed across all virtual views.

These examples demonstrate the efficiency of our generative models used to enhance multiview depth imagery. Our approach to improve the inter-view depth consistency permits the rendering of high-quality intermediate virtual views which allow interactive users to navigate smoothly through high-quality natural scenes.

IV. Conclusions

This paper presents a probabilistic approach to multiview depth image enhancement by using variational inference. We exploit the inherent inter-view similarity in multiview imagery through color classification of concatenated views. For color classification, a Dirichlet mixture model with variational inference is employed. The resulting color clusters are used to classify depth pixels from various viewpoints. Here, a per-pixel association between depth and color pixels has been utilized. The inter-view inconsistencies in these depth clusters inspire further subclassification with Gaussian mixture models. The depth subclassification assigns probabilistic weights to depth pixels. These depth weights are then used to repair the input depth imagery at multiple viewpoints. Pre-processing by generating superpixels reduces the overall computational complexity significantly. Both objective and subjective results confirm the benefit of our multiview depth image enhancement method for interactive users.

Acknowledgment

This work was supported in part by Ericsson AB and the ACCESS Linnaeus Center at KTH Royal Institute of Technology, Stockholm, Sweden. Zhanyu Ma was partly supported by the National Nature Science Foundation of China (NSFC) Grant No. 61402047.

References

[1] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 67–76, Jan. 2011.

[2] P. Benzie, J. Watson, P. Surman, I. Rakkolainen, K. Hopf, H. Urey, V. Sainov, and C. von Kopylow, "A survey of 3DTV displays: Techniques and technologies," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1647–1658, Nov. 2007.

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

[4] 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.

[5] 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.

[6] R. A. Jarvis, "A perspective on range finding techniques for computer vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, no. 2, pp. 122–139, Mar. 1983.

[7] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. IEEE Int. Symp. Mixed and Augmented Reality*, Basel, Switzerland, Oct. 2011, pp. 127–136.

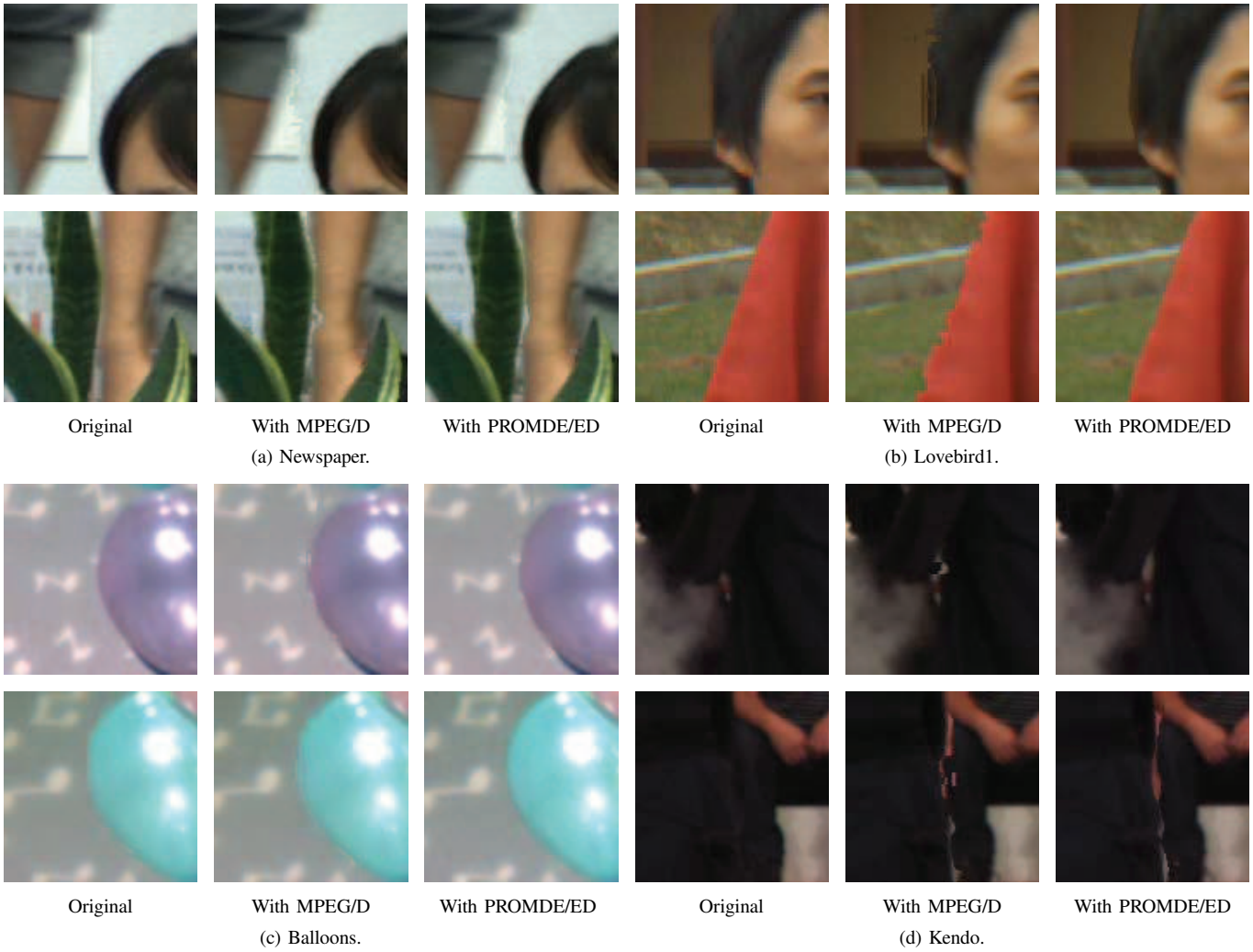


Fig. 9. Selected regions of synthesized virtual views of test sequences as generated by VSRS 3.5 using MPEG depth maps and enhanced depth maps from our depth enhancement algorithm. These virtual views are generated by using the small camera baseline setting of Table I. (Best viewed in color.)

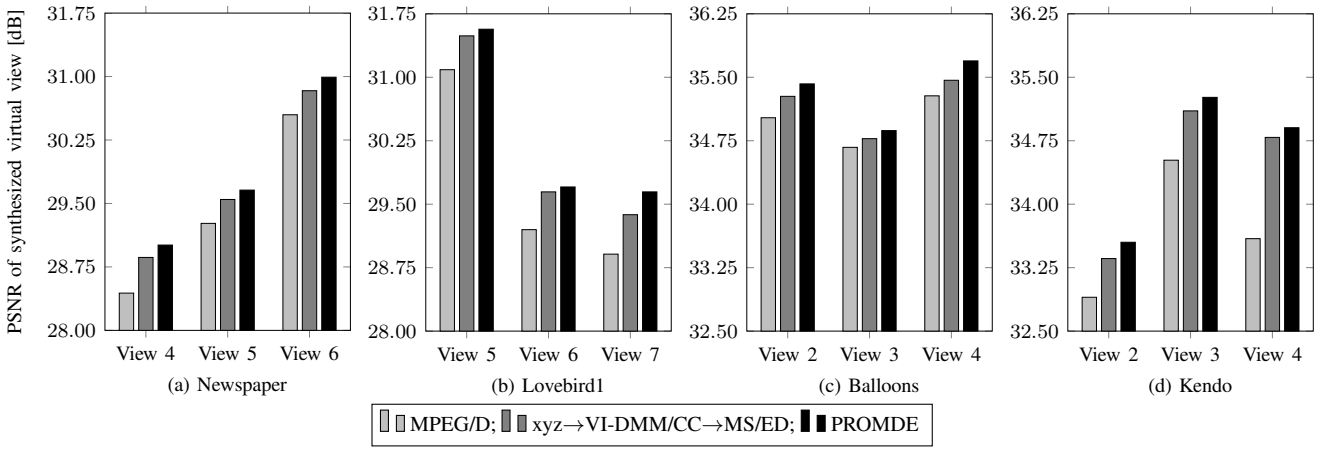


Fig. 10. The objective quality of three intermediate virtual views as generated by VSRS 3.5 using the large baseline setting as given in Table I for a total number of desired superpixels $M = 1.2 \times 10^5$. The virtual views as obtained by using MPEG depth maps without enhancement (MPEG/D) are used as the baseline for comparison. The quality of view synthesis improves across all three intermediate viewpoints by using enhanced depth maps via probabilistic depth enhancement (PROMDE) when compared to the enhancement method xyz→VI-DMM/CC→MS/ED, where the enhanced depth (ED) is generated by mean-shift (MS). This reference is the extended version of our early work [41] and includes superpixels as pre-processing step.

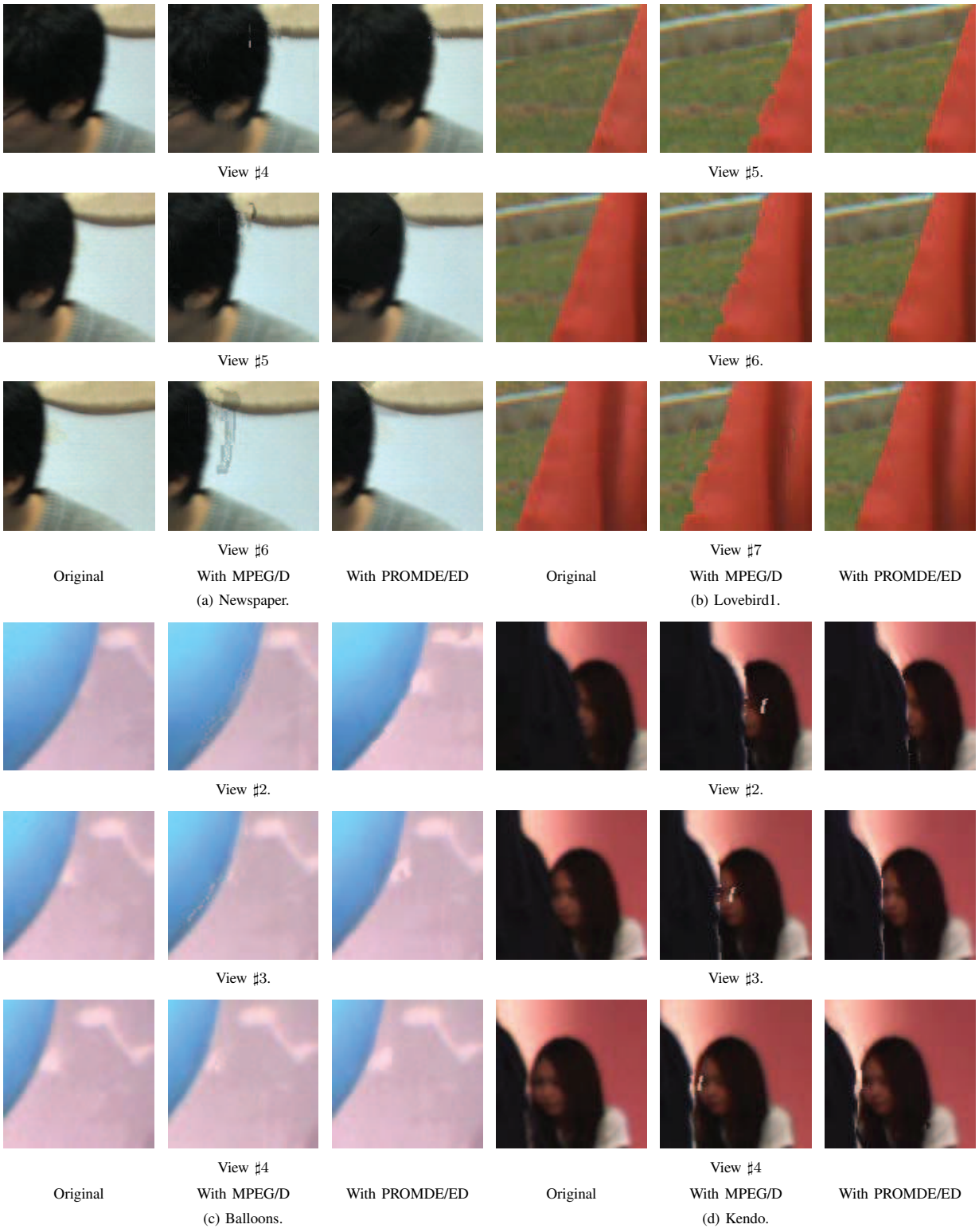


Fig. 11. For the large camera baseline setting in Table I, selected regions of synthesized virtual views at three different viewpoints are shown. VSRS 3.5 is used for rendering. The enhanced depth maps from our depth enhancement algorithm (PROMDE) exhibit more inter-view consistency when compared to the original MPEG depth maps without depth enhancement. The improved depth consistency has a direct impact on the quality of the rendered views. (Best viewed in color.)

[8] J. Park, H. Kim, Y.-W. Tai, M. Brown, and I. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. IEEE Int. Conf. Computer Vision*, Barcelona, Spain, Nov. 2011, pp. 1623–1630.

[9] M. Camplani and L. Salgado, "Efficient spatio-temporal hole filling strategy for Kinect depth maps," in *Proc. SPIE*, vol. 8290, San Francisco, CA, USA, Jan. 2012, pp. 82 900E 1–10.

[10] J. Liu, X. Gong, and J. Liu, "Guided inpainting and filtering for Kinect depth maps," in *Int. Conf. Pattern Recognition*, Tsukuba, Japan, Nov.

2012, pp. 2055–2058.

[11] J. Shen and S.-C. S. Cheung, "Layer depth denoising and completion for structured-light RGB-D cameras," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, Washington, DC, USA, Jun. 2013, pp. 1187–1194.

[12] S. Lee, "Time-of-flight depth camera motion blur detection and deblurring," *IEEE Signal Process. Letters*, vol. 21, no. 6, pp. 663–666, Jun. 2014.

[13] Y.-S. Kang and Y.-S. Ho, "High-quality multi-view depth generation using multiple color and depth cameras," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Suntec City, Singapore, Jul. 2010, pp. 1405–1410.

[14] J. Wang, C. Zhang, W. Zhu, Z. Zhang, Z. Xiong, and P. Chou, "3D scene reconstruction by multiple structured-light based commodity depth cameras," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 5429–5432.

[15] Y. M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun, "Multi-view image and ToF sensor fusion for dense 3D reconstruction," in *IEEE Int. Conf. Computer Vision Workshop*, Sept. 2009, pp. 1542–1549.

[16] J. Zhu, L. Wang, R. Yang, J. Davis, and Z. Pan, "Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1400–1414, Jul. 2011.

[17] Y. M. Kim, D. Chan, C. Theobalt, and S. Thrun, "Design and calibration of a multi-view tof sensor fusion system," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition Workshops*, Anchorage, AK, USA, Jun. 2008, pp. 1–7.

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

[19] G. Nur, S. Dogan, H. Arachchi, and A. Kondoz, "Impact of depth map spatial resolution on 3D video quality and depth perception," in *3DTV Conf.*, Tampere, Finnland, Jun. 2010, pp. 1–4.

[20] 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.

[21] 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, Sep. 2004.

[22] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, Washington, DC, USA, Jun. 2004, pp. 261–268.

[23] C. Cigla, X. Zabulis, and A. A. Aydin, "Segment-based stereo-matching via plane and angle sweeping," in *Proc. 3DTV Conf.*, Kos Island, Greece, May 2007, pp. 1–4.

[24] W.-S. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map distortion analysis for view rendering and depth coding," in *Proc. IEEE Int. Conf. Image Process.*, Cairo, Egypt, Nov. 2009, pp. 721 – 724.

[25] L. Fang, N.-M. Cheung, D. Tian, A. Vetro, H. Sun, and O. Au, "An analytical model for synthesis distortion estimation in 3D video," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 185–199, Jan. 2014.

[26] 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, Queensland, Australia, Oct. 2008, pp. 34–39.

[27] P. K. Rana and M. Flierl, "View interpolation with structured depth from multiview video," in *Proc. European Signal Process. Conf.*, Barcelona, Spain, Aug. 2011, pp. 383–387.

[28] T. Ishibashi, M. Tehrani, T. Fujii, and M. Tanimoto, "FTV format using global view and depth map," in *Proc. Picture Coding Symp.*, Krakow, Poland, May 2012, pp. 29 –32.

[29] C. Nguyen, S. Izadi, and D. Lovell, "Modeling kinect sensor noise for improved 3d reconstruction and tracking," in *Proc. Int. Conf. 3D Imaging, Modeling, Processing, Visualization and Transmission*, Zurich, Switzerland, Oct. 2012, pp. 524–530.

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

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

[32] D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1176–1190, Mar. 2012.

[33] H. Zeng and K.-K. Ma, "Content-adaptive temporal consistency enhancement for depth video," in *Proc. IEEE Int. Conf. Image Process.*, Orlando, FL, USA, Sept. 2012, pp. 3017–3020.

[34] 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.

[35] M. Drumheller and T. Poggio, "On parallel stereo," in *Proc. IEEE Int. Conf. Robotics and Automation*, vol. 3, San Francisco, CA, USA, Apr. 1986, pp. 1439–1448.

[36] 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.

[37] H. Tao and H. Sawhney, "Global matching criterion and color segmentation based stereo," in *IEEE Workshop Applicat. Comput. Vision*, Palm Springs, CA, USA, Dec. 2000, pp. 246–253.

[38] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proc. Int. Conf. Pattern Recognition*, vol. 3, Hong Kong, China, Aug. 2006, pp. 15–18.

[39] 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.

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

[41] 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.

[42] Z. Ma, P. K. Rana, J. Taghia, M. Flierl, and A. Leijon, "Bayesian estimation of dirichlet mixture model with variational inference," *Pattern Recognition*, vol. 47, no. 9, pp. 3143 – 3157, 2014.

[43] N. Bouguila, D. Ziou, and J. Vaillancourt, "Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1533–1543, Nov. 2004.

[44] 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.

[45] W. Tao, H. Jin, and Y. Zhang, "Color image segmentation based on mean shift and normalized cuts," *IEEE Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 5, pp. 1382–1389, Oct. 2007.

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

[47] G. Wyszecki and W. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd ed. New York, NY, USA: John Wiley & Sons, 2000.

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

[49] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[50] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Adv. Neural Inf. Process. Syst. II*. MIT Press, 1999, pp. 487–493.

[51] 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.

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

[53] Z. Ghahramani and M. J. Beal, "Variational inference for Bayesian mixtures of factor analysers," in *Adv. Neural Inf. Process. Syst. 12*. MIT Press, 2000, pp. 449–455.

[54] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, Nov. 1999.

[55] G. J. McLachlan and D. Peel, *Finite Mixture Models*. Wiley, 2000.

[56] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[57] H. Tao, H. Sawhney, and R. Kumar, "Dynamic depth recovery from multiple synchronized video streams," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, Kauai, HI, USA, Dec. 2001, pp. 118–124.

[58] 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.

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

Pravin Kumar Rana is a Senior Algorithm Developer at Tobii Technology AB, Stockholm, since April 2014. He is working towards his Doctorate in Telecommunications from KTH Royal Institute of Technology, Sweden. He received the Master of Science in Physics from Ranchi University, Ranchi, India in 2004 and the Master of Technology in Earth System Science and Technology from Indian Institute of Technology, Kharagpur, India in 2008. His research interest lies in image and video processing, computer vision, and application of machine learning in computer vision.

Jalil Taghia is currently a postdoctoral research fellow at Neural Information Processing Group at Technical University of Berlin since April 2014. He received his PhD in Telecommunications from KTH (Royal Institute of Technology), Sweden, in 2014. His research interest lies in probabilistic inference in machine learning and statistics including approximate methods in Bayesian inference and directional statistics

Zhanyu Ma received the MEng degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), China, and the PhD degree in electrical engineering from KTH (Royal Institute of Technology), Sweden, in 2007 and 2011, respectively. From 2012 to 2013, he has been a postdoctoral research fellow in the School of Electrical Engineering, KTH, Sweden. He has been an assistant professor at the Beijing University of Posts and Telecommunications, China, since 2013. His research interests include statistical modeling and machine learning related topics with a focus on applications in speech processing, image processing, biomedical signal processing, big data analytics, and bioinformatics.

Markus Flierl (S'01-M'04) is Associate Professor of Electrical Engineering at KTH Royal Institute of Technology, Stockholm. He received the Doctorate in Engineering from Friedrich Alexander University, Germany, in 2003. From 2000 to 2002, he visited the Information Systems Laboratory at Stanford University. From 2003 to 2005, he was a senior researcher with the Signal Processing Institute at the Swiss Federal Institute of Technology Lausanne, Switzerland. From 2005 to 2008, he was Visiting Assistant Professor at the Max Planck Center for Visual Computing and Communication at Stanford University, California. He has authored the book "Video Coding with Superimposed Motion-Compensated Signals: Applications to H.264 and Beyond". He was the recipient of the SPIE VCIP 2007 Young Investigator Award. His research interests include visual computing and communication, video representations, and mobile visual search.