TEMPORAL DEPTH VIDEO ENHANCEMENT BASED ON INTRINSIC STATIC STRUCTURE

Lu Sheng *King Ngai Ngan* *Songnan Li*

Department of Electronic Engineering, The Chinese University of Hong Kong

ABSTRACT

Depth video enhancement is an essential preprocessing for various 3D applications. Despite extensive studies of spatial enhancement, effective temporal enhancement that both strengthens temporal consistency and keeps correct depth variation needs research in depth. In this paper, we propose a novel method to enhance the depth video by blending raw depth frame with estimated so-called intrinsic static structure, which defines static structure of captured scene and is estimated iteratively by a probabilistic generative model with sequentially incoming depth frames. Our experimental results show that the proposed method is effective both in static and dynamic scene and owns compatibility to work with various kinds of depth videos. And superior performance can be achieved in comparison with existing temporal enhancement approaches.

Index Terms— depth video enhancement, temporal enhancement, probabilistic model, variational approximation

1. INTRODUCTION

Depth sequences generated by current depth estimation methods, both active and passive, are subject to the temporal inconsistency problem. Apart from reasonable inconsistency comes from fast motion of foreground objects, raw depth sequence also suffers from depth flickering in static region, flips around scene discontinuities, as well as erratic depth measurements around texture-less region by stereo, or non-lambertian surface by time-of-flight camera or Kinect. Therefore, effective temporal enhancement approach should suppress inconsistency in static or slowly moving region, while maintain correct depth variations at the remaining area. In fact given camera motion parameters, moving scene can also be handled.

In spite of spatial enhancement of depth map has been extensively studied in recent years, like energy minimization methods [1, 2, 3] based on Markov Random Fields (MRF) model or Auto-regressive model or filtering methods based on high-dimensional Gaussian filtering [4, 5, 6], temporal enhancement of depth video still has room for improvement. Several existing methods [7, 8] take the temporal similarity of texture and 2D motion information into considerations, but correct depth variations cannot always be maintained around regions where depth temporal consistency is violated but texture temporal consistency is held. Depth signals should also provide essential cues for temporal enhancement.

Therefore, due to various complex and even unpredictable dynamic contents, as well as outliers in a depth video, it is not easy to exactly locate the regions where temporal consistency should be enforced by depth video itself. Instead we try to online track and incrementally refine the *intrinsic static structure*. Thus, besides spatial enhancement on individual depth frame, temporal enhancement is performed by blending with estimated intrinsic static structure, in a way that static regions will put more weights on corresponding intrinsic static structure, and non-stationary regions on input measurements, as to be elaborated in the following section.

2. PROPOSED METHOD

The intrinsic static structure can be regarded as a manifold lies on or behind the input depth map, which contains the static structure of the captured scene. Any moving or foreground object stays in front of the intrinsic static structure, while static regions or visible background area are fused into it. However, at the very beginning, reliable intrinsic static structure is not available. A reasonable initialization is just the first frame of input depth video. The intrinsic static structure is refined iteratively with input depth frames. Then each pixel's temporally enhanced depth value is calculated by a weighted combination of input depth measurement and the intrinsic static structure. The weight to the intrinsic static structure is equal to the probability that input depth belongs to this structure. Notice that no spatial/texture information or motion information are introduced here. Confidence on whether a pixel is static is just based on the previous depth sequence of this pixel. Together with other spatial enhancement approaches, an overall spatial and temporal depth video enhancement can be achieved.

2.1. A Probabilistic Generative Model

We define $\mathbf{x} = (u, v)$ as the pixel location in the image domain. The incoming depth measurement at time t of pixel $\mathbf{x}$ is $d_{\mathbf{x}}^t$. If $d_{\mathbf{x}}^t$ belongs to the intrinsic static structure, we assume that it follows a Gaussian distribution centered at $Z_{\mathbf{x}}$, as $\mathcal{N}\left(d_{\mathbf{x}}^t | Z_{\mathbf{x}}, \tau_{\mathbf{x}}^{-1}\right)$, where $\tau_{\mathbf{x}}$ denotes the precision, and is pre-defined by the user.

On the contrary, the depth measurements against moving objects or outliers in the front, follow a clutter distribution like $\mathcal{U}_f(d^t_{\mathbf{x}}|Z_{\mathbf{x}}) = U_f\left[d^t_{\mathbf{x}} < Z_{\mathbf{x}}\right]$, where $U_f\left[\cdot\right]$ is an indicator function that equals to U_f when input argument is true, and 0 otherwise. It is valid because depth value of moving object at the same line-of-sight will be always smaller than that of the intrinsic static structure.

Furthermore, it's possible that current estimation of intrinsic static structure is incorrect since it may be stuck before the true one, thus another indicator distribution is introduced as $\mathcal{U}_b\left(d^t_{\mathbf{x}}|Z_{\mathbf{x}}\right) = U_b\left[d^t_{\mathbf{x}} > Z_{\mathbf{x}}\right]$. It can naturally represent outliers that have a larger depth value than given structure. Otherwise if certain number of successive inputs fall into this category, we have enough evidence that current estimation is incorrect.

Therefore, the likelihood of $d^t_{\mathbf{x}}$ against given model is a mixture of these three densities:

$$p(d^t_{\mathbf{x}}|Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}) = \omega^1_{\mathbf{x}}\mathcal{N}(d^t_{\mathbf{x}}|Z_{\mathbf{x}}, \tau^{-1}_{\mathbf{x}}) + \omega^2_{\mathbf{x}}\mathcal{U}_f\left(d^t_{\mathbf{x}}|Z_{\mathbf{x}}\right) + \omega^3_{\mathbf{x}}\mathcal{U}_b(d^t_{\mathbf{x}}|Z_{\mathbf{x}}), \quad (1)$$

where $\sum_{i=1}^{3}\omega^i_{\mathbf{x}} = 1$. $\boldsymbol{\omega}_{\mathbf{x}}$ is the ratio of each case. Given Gaussian prior over $Z_{\mathbf{x}}$ and Dirichlet distribution over $\boldsymbol{\omega}_{\mathbf{x}}$:

$$p(Z_{\mathbf{x}}) = \mathcal{N}\left(Z_{\mathbf{x}}|\mu_{\mathbf{x}}, \lambda^{-1}_{\mathbf{x}}\right), p(\boldsymbol{\omega}_{\mathbf{x}}) = \mathrm{Dir}\left(\boldsymbol{\omega}_{\mathbf{x}}|\alpha_{1,\mathbf{x}}, \alpha_{2,\mathbf{x}}, \alpha_{3,\mathbf{x}}\right), \quad (2)$$

we can explicitly model the chance that $d^t_{\mathbf{x}}$ is inside the intrinsic static structure and the most possible $Z_{\mathbf{x}}$ by the posterior

$$p(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}|d^t_{\mathbf{x}}) = p(d^t_{\mathbf{x}}|Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}})p(Z_{\mathbf{x}})p(\boldsymbol{\omega}_{\mathbf{x}})/p(d^t_{\mathbf{x}}). \quad (3)$$

With sequentially incoming depth samples, we can update distributions of the intrinsic static structure by iteratively estimating their parameters. To solve the updating problem in real-time, we exploit a variational approximation approach to on-line track the parameters evolution.

2.2. Variational Parameter Estimation

We factorize the posterior given by Eq. 3 into independent Gaussian distribution $q(Z_{\mathbf{x}}) = \mathcal{N}\left(Z_{\mathbf{x}}|\mu'_{\mathbf{x}}, \lambda'^{-1}_{\mathbf{x}}\right)$ and Dirichlet distribution $q(\boldsymbol{\omega}_{\mathbf{x}}) = \mathrm{Dir}(\boldsymbol{\omega}_{\mathbf{x}}|\alpha'_{1,\mathbf{x}}, \alpha'_{2,\mathbf{x}}, \alpha'_{3,\mathbf{x}})$ so that

$$q(Z_{\mathbf{x}})q(\boldsymbol{\omega}_{\mathbf{x}}) \sim p(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}|d^t_{\mathbf{x}}). \quad (4)$$

The approximated posterior are found by minimizing the KL-divergence $\mathcal{KL}\left[p(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}|d^t_{\mathbf{x}})\|q(Z_{\mathbf{x}})q(\boldsymbol{\omega}_{\mathbf{x}})\right]$, which results in matching moments between true and approximated distributions [9].

The joint distribution [1] can be further written as Eq. 5, where $\alpha_0 = \sum_{i=1}^{3}\alpha_i$. The data fidelity is hence

$$p(d) = \frac{\alpha_1}{\alpha_0}\mathcal{N}\left(d|\mu, \tau^{-1} + \lambda^{-1}\right) + \frac{\alpha_3}{\alpha_0}U_b\Phi\left(\sqrt{\lambda}\left(d - \mu\right)\right) + \frac{\alpha_2}{\alpha_0}U_f\left(1 - \Phi\left(\sqrt{\lambda}\left(d - \mu\right)\right)\right), \quad (6)$$

[1] Because our discussion is based on single pixel and current frame, the notations $\mathbf{x}$ and t are omitted from related symbols in the rest part for brevity.

Algorithm 1: Intrinsic Static Structure Update Scheme

input : Input depth sequence $\{d^t|t = 1, 2, \ldots, N\}$;
Initial parameter set $\mathcal{L}_{init} = \{\mu_0, \lambda_0, \boldsymbol{\alpha}_0\}$;
output: Current parameter set $\mathcal{L} = \{\mu, \lambda, \boldsymbol{\alpha}\}$;
Intrinsic static scene depth value d_{iss};
Temporal smoothed depth value d_S;

$\mu_0 \leftarrow d^1$, $\mathcal{L} \leftarrow \mathcal{L}_{init}$ and $d_S \leftarrow d^1, d_{iss} \leftarrow d^1$;
for $t \leftarrow 2$ **to** N **do**
  $d \leftarrow d^t$;
  if $d > 0$ **then**
    estimate parameter set $\mathcal{L}'$ based on $\mathcal{L}$ by Eq. 7 and 8;
    $\rho \leftarrow \frac{\alpha'_3}{\alpha'_1}$, $\Delta_F\alpha_3 \leftarrow \alpha'^{t}_3 - \alpha^{t-F}_3$;
    if $\rho > T_\rho$ *or* $\Delta_F\alpha_3 > T_F$ **then**
      $\mu_0 \leftarrow d$, $\mathcal{L} \leftarrow \mathcal{L}_{init}$ and $d_{iss} \leftarrow d$, $d_S \leftarrow d$;
    else
      estimate $\gamma_{iss}(d)$;
      $\mathcal{L} \leftarrow \mathcal{L}'$ and $d_{iss} \leftarrow \mu$;
      `// temporal enhancement`
      $d_S \leftarrow (1 - \gamma_{iss}(d))d + \gamma_{iss}(d)d_{iss}$;

where $\Phi\left(\cdot\right)$ is the standard Gaussian cumulative density function. $p(Z, \boldsymbol{\omega}, d)$ is a weighted combination of three terms, and in each term the probabilities of Z and $\boldsymbol{\omega}$ are independent, thus it eases the computation complexity of calculating moments related to Z and $\boldsymbol{\omega}$ respectively by marginalizing the other variable.

The moments of $q(Z)$ is therefore readily computed by estimating the first and second moments of marginal distribution $p(Z|d)$:

$$\mathbb{E}_{q(Z)}[Z] = \mathbb{E}_{p(Z|d)}[Z], \quad \mathbb{E}_{q(Z)}[Z^2] = \mathbb{E}_{p(Z|d)}[Z^2]. \quad (7)$$

While the moments of $q\left(\boldsymbol{\omega}\right)$ can be estimated by matching moments similar as Eq. 7, which does not exactly minimize the KL-divergence, but allows a closed-form solution [10]:

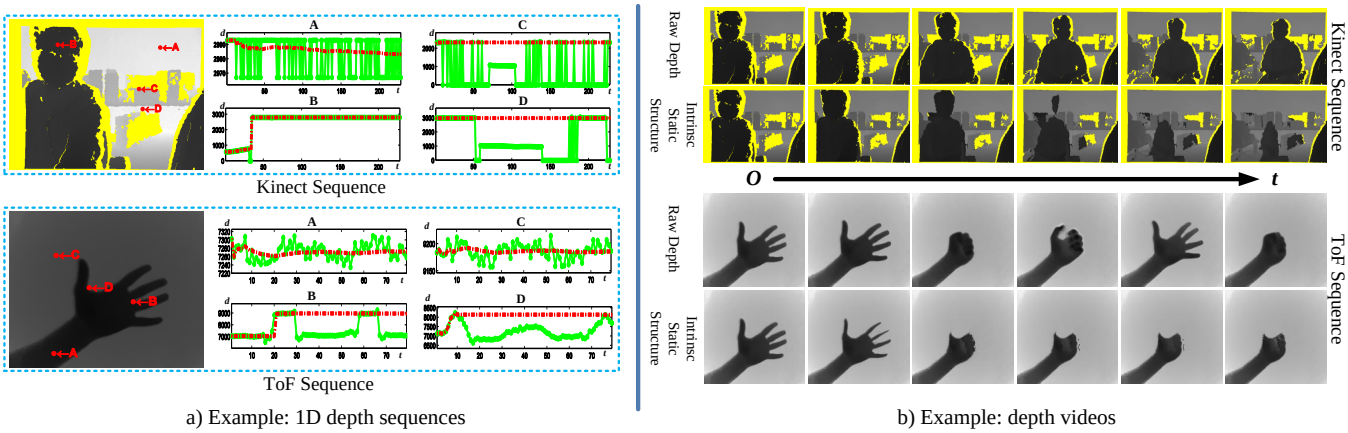
$$\mathbb{E}_{q(\boldsymbol{\omega})}[\omega_i] = \mathbb{E}_{p(\boldsymbol{\omega}|d)}[\omega_i], \quad \mathbb{E}_{q(\boldsymbol{\omega})}[\omega^2_i] = \mathbb{E}_{p(\boldsymbol{\omega}|d)}[\omega^2_i], i = 1, 2, 3. \quad (8)$$

After that, current estimation $\mathcal{L} = \{\mu, \lambda, \alpha_i|i = 1, 2, 3\}$ will be updated by $\mathcal{L}' = \{\mu', \lambda', \alpha'_i|i = 1, 2, 3\}$.

2.3. On-line Intrinsic Static Structure Update Scheme

The on-line intrinsic static structure update scheme is actually a sequential variational parameter estimation problem, see Sec. 2.2. Since our approach is sensitive to the initialization, parameter re-initialization is necessary when there is a high risk that current estimation is incorrect, as discussed in Sec. 2.1. We track the ratio $\rho = \alpha'_3/\alpha'_1$ and increment $\Delta_F\alpha_3 = \alpha'^{t}_3 - \alpha^{t-F}_3$ of α_3 over successive F frames. When they are larger than a given threshold T_ρ or T_F, the current estimation is re-initialized with parameter set $\mathcal{L}_{init} = \{\mu_0, \lambda_0, \boldsymbol{\alpha}_0\}$, where the initial μ_0 equals to d and the rest are user-given constant values. This treatment is valid because in such a case, d has a relatively larger chance to

$$p(Z,\boldsymbol{\omega},d)=\frac{\alpha_1}{\alpha_0}\mathcal{N}\left(d|\mu,\tau^{-1}+\lambda^{-1}\right)\mathcal{N}\left(Z\Big|\frac{\lambda\mu+\tau d}{\lambda+\tau},(\lambda+\tau)^{-1}\right)\mathrm{Dir}\left(\boldsymbol{\omega}|\alpha_1+1,\alpha_2,\alpha_3\right)+ \tag{5}$$
$$\frac{\alpha_2}{\alpha_0}\mathcal{U}_f\left(d<Z\right)\mathcal{N}\left(Z|\mu,\lambda^{-1}\right)\mathrm{Dir}\left(\boldsymbol{\omega}|\alpha_1,\alpha_2+1,\alpha_3\right)+\frac{\alpha_3}{\alpha_0}\mathcal{U}_b\left(d>Z\right)\mathcal{N}\left(Z|\mu,\lambda^{-1}\right)\mathrm{Dir}\left(\boldsymbol{\omega}|\alpha_1,\alpha_2,\alpha_3+1\right)$$



a) Example: 1D depth sequences

b) Example: depth videos

Fig. 1: Background model estimation. a) shows the background model estimation towards 1D depth sequence. 4 pixels (marked by A,B,C,D) are picked from each video. Raw depth sequences are green, and background sequences are represented by red dash curves. b) shows the results towards depth videos. Larger depth value is shown lighter. Best viewed in color.

behave like outliers with larger depth values. Meanwhile, $\mathcal{L}$ will stay the same when input depth measurement is absent.

The intrinsic static structure depth value d_{iss} is estimated as the expected value of $q(Z)$, *i.e.*, μ'. Examining the probability that d belongs to the intrinsic static structure as

$$\gamma_{iss}(d)=\alpha_1\mathcal{N}\left(d|\mu,\tau^{-1}+\lambda^{-1}\right)/p(d), \tag{9}$$

the temporally smoothed depth value d_S is obtained by blending d and d_{iss} with weights related to $\gamma_{iss}(d)$:

$$d_S=\left(1-\gamma_{iss}(d)\right)d+\gamma_{iss}(d)d_{iss}. \tag{10}$$

However $\gamma_{iss}(d)$ given by Eq. 9 cannot distinguish outliers from depth measurement errors with those from moving objects. Observing that the former outliers always occur sparsely in spatial domain, a binary-labeled CRF model is utilized [9, 11] to suppress sparse outliers leading to a refined $\gamma_{iss}(\mathbf{d})$. On the other hand, our approach can also handle depth missing problem. If current depth measurement is not applicable, *i.e.*, $d=0$ and other inference methods based on texture or spatial information indicate that the current pixel refers to the intrinsic static structure, *i.e.*, $\gamma_{iss}(d)>0.5$, a probable guess of missing depth value will be $d_S=d_{iss}$. Together with further spatial filtering or inpainting techniques, holes of input depth frames can be reliably filled.

The overall scheme is summarized in Alg. 1.

3. EXPERIMENTS AND DISCUSSIONS

We have implemented a MATLAB version of our algorithm to various depth videos captured by Kinect or ToF cameras, including static and dynamic scenes. Initial parameters are simply set as $\boldsymbol{\alpha}_0=[2,1,1]^T$, λ_0^{-1} is the square of 10% of the depth range of input scene. $F=3$, $T_F=2.5$ and $T_\rho=1$.

3.1. Intrinsic Static Structure Estimation

We randomly chose two available depth videos by ToF cameras [7] and Kinect to validate our algorithm. Results of 1D depth sequences and whole depth videos are shown in Fig. 1. Kinect depth videos contain severe holes, but our method can still robustly estimate the intrinsic static structure provided enough successful samples, as can be seen from sequences C, D of Kinect in Fig. 1 a). For a stationary pixel, our method is analogous to traditional sequential parameter estimation of Gaussian model, but in addition suppresses occasional outliers. From Fig. 1, our method incrementally detects and refines the intrinsic static structure model. False estimations happen due to the re-initialization criteria has not been satisfied, which can be mitigated by tuning the thresholds T_F and T_ρ or adding texture or motion information.

3.2. Temporal Enhancement

3.2.1. Static Scene Analysis

If input scene is static, intrinsic static structure is exactly the temporal enhanced depth map we wish to achieve. Additionally, $r_{\mathbf{x}}=\alpha_{1,\mathbf{x}}/\sum_{i=1}^{3}\alpha_{i,\mathbf{x}}$ also quantifies the reliability of intrinsic static structure because it illustrates the portion of depth samples that belongs to intrinsic static structure over all frames, equivalently, also indicates the reliability of the enhanced depth map. As shown in Fig. 2 a), the enhanced depth maps have superior quality than input raw depth maps. The reliability map indicates that most flat or smooth surfaces of intrinsic static structure are of high reliability. If we simply mark unreliable pixels by $r_{\mathbf{x}}\leq 0.5$ and delete them, we obtain the reliable depth map, where lots of pixels around

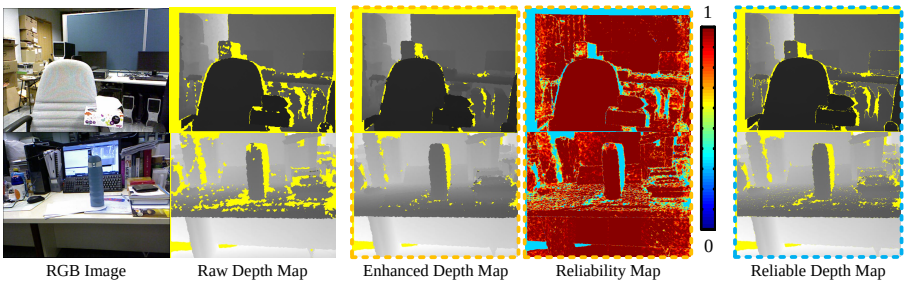


Fig. 2: a) Temporal enhancement to static scene. Data were captured by Kinect. Each row represents a static scene. b) Performance comparison with *t*-MedFilter and *TC*-Enhance. The top two sequences are captured by Kinect, the bottom two are captured by ToF camera.

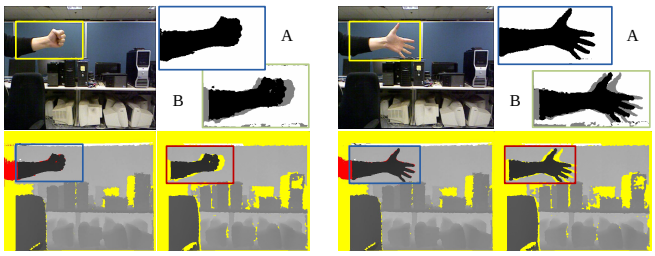


Fig. 3: Temporal enhancement results of two frames. To each frame, from top left to bottom right, RGB image, masks of static area in boxed region before (B) and after (A) hole filling and smoothing of $\gamma_{iss}(\mathbf{d})$, enhanced depth map and raw depth map.

discontinuities or occlusions are marked to be unreliable, because measurements around such regions tend to be unfaithful. Our reliability map, other than that of heuristic methods [12], is estimated by data and feasible for diverse depth data.

Fig. 2 b) also presents results of 1D depth sequences of proposed method as well as two referenced methods: temporal median filtering with depth hole filling (*t*-MedFilter), and temporal consistency enhancement (*TC*-Enhance) proposed by Fu [13], where our method can efficiently smooth the noisy and flickering depth sequences, while the rest are sensitive to outliers and cannot find the static structure of input signals.

3.2.2. *Dynamic Scene Anaysis*

Our temporal enhancement method can efficiently find the moving objects while suppress noise and flickering artifacts at static regions. Results of Kinect sequences are shown in Fig. 3. $\gamma_{iss}(\mathbf{d})$ are estimated by Eq. 9 and smoothed by a binary-labeled MRF model encoding pairwise texture similarity [9, 11, 14], which also indicates probabilities at depth holes (see (a) and (b) of each frame in Fig. 3, gray means holes). Actually if we do not perform smoothing on $\gamma_{iss}(\mathbf{d})$, it is still reasonable to fill holes just by using depth of intrinsic static structure at corresponding regions. Depth holes inside static regions are filled by intrinsic static structure, while those inside non-stationary regions keep blank[2] and are presented by red color, because we only focused on temporal enhancement.

[2]Or just use joint bilateral filtering [4] on non-stationary regions.

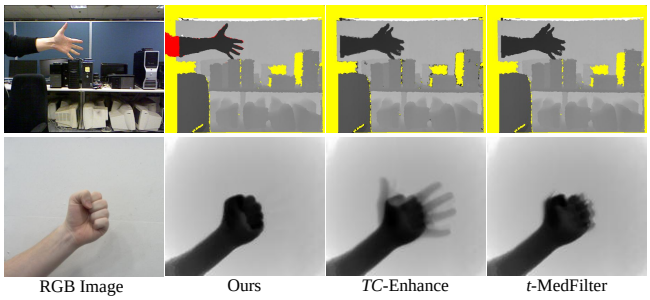


Fig. 4: Qualitative comparison of temporal enhancement of depth video with *TC*-Enhance and *t*-MedFilter. The top row is a frame capture by Kinect, the bottom row shows a frame captured by ToF camera.

Compared with *TC*-Enhance and *t*-MedFilter, our method performs better. Because it can robustly and quickly detect the static regions, we only enforce temporal consistency at detected static regions rather than at dynamic regions. Therefore, unlike the others, the proposed method will not generate apparent motion delay.

4. CONCLUSION AND FUTURE WORK

In this paper, we address the problem of robust temporal enhancement of depth video by blending input depth frame with intrinsic static structure, which underlies the target scene and can be estimated iteratively by a probabilistic generative model with efficient parameter estimation and updating scheme. Qualitative evaluation shows that our method operates well on various scenes (dynamic or static), and different sources (ToF camera or Kinect). But the lack of ground-truth hinders quantitative evaluations. In the future we would like to utilize synthesis data or high-cost device like laser scanner to provide ground-truth and conduct a thorough evaluation of our method as well as other available approaches.

There are several aspects we also would like to explore further. For instance, the combination of spatial enhancement and our temporal enhancement into a whole framework and research on more general static structure.

5. REFERENCES

[1] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *NIPS*. 2006, vol. 18, p. 291, MIT.

[2] J. Yang, X. Ye, K. Li, and C. Hou, "Depth recovery using an adaptive color-guided auto-regressive model," in *ECCV*. 2012, pp. 158–171, Springer.

[3] J. Park, H. Kim, Y.W. Tai, M.S. Brown, and I. Kweon, "High quality depth map upsampling for 3d-tof cameras," in *ICCV*. IEEE, 2011, pp. 1623–1630.

[4] J. Kopf, M.F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM TOG*, vol. 26, no. 3, pp. 96, 2007.

[5] J. Dolson, J. Baek, C. Plagemann, and S. Thrun, "Upsampling range data in dynamic environments," in *CVPR*. IEEE, 2010, pp. 1141–1148.

[6] B. Huhle, T. Schairer, P. Jenke, and W. Straßer, "Fusion of range and color images for denoising and resolution enhancement with a non-local filter," *CVIU*, vol. 114, no. 12, pp. 1336–1345, 2010.

[7] N. A. Dodgson H.-P. Seidel C. Richardt, C. Stoll and C. Theobalt, "Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos," *Computer Graphics Forum (Proceedings of Eurographics)*, vol. 31, no. 2, May 2012.

[8] Dongbo Min, Jiangbo Lu, and Minh N Do, "Depth video enhancement based on weighted mode filtering," *Image Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 1176–1190, 2012.

[9] Christopher M Bishop and Nasser M Nasrabadi, *Pattern recognition and machine learning*, vol. 1, springer New York, 2006.

[10] Thomas P Minka, *A family of algorithms for approximate Bayesian inference*, Ph.D. thesis, Massachusetts Institute of Technology, 2001.

[11] Vladlen Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," NIPS, 2011.

[12] F. Garcia, B. Mirbach, B. Ottersten, F. Grandidier, and A. Cuesta, "Pixel weighted average strategy for depth sensor data fusion," in *ICIP*. IEEE, 2010, pp. 2805–2808.

[13] Deliang Fu, Yin Zhao, and Lu Yu, "Temporal consistency enhancement on depth sequences," in *PCS, 2010*. IEEE, 2010, pp. 342–345.

[14] Ju Shen, Po-Chang Su, S.S. Cheung, and Jian Zhao, "Virtual mirror rendering with stationary rgb-d cameras and stored 3-d background," *Image Processing, IEEE Transactions on*, vol. 22, no. 9, pp. 3433–3448, 2013.