International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 8, No. 5 (2015), pp. 87-92
http://dx.doi.org/10.14257/ijsip.2015.8.5.09

# Refinement of Kinect Sensor's Depth Maps Based on GMM and CS Theory

Qian Zhang, ShaoMin Li, Wenfeng Guo, Pei Wang and Jifeng Huang

*College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China*
*qian83@126.com*

## *Abstract*

*As the Microsoft's Kinect sensor can generate a real-time dense depth map with relatively commercial available, it is widely used in depth map capturing. However, there are some artifacts like holes, instability of the raw input data, which seriously affect the application. To solve this problem, in this paper, we propose a novel depth map refinement method based on by GMM and CS theory which enable the kinect sensor generate a dense depth map, the background large holes are filled without blurring, and the edges of the objects are sharpened, median filter is used to remove noise. Experiments on captured indoor data demonstrate the effectiveness of the method especially in the edge area and occlusion area that our method can obtain better results.*

***Keywords:*** *depth image, Gaussian mixture model, hole filling*

## 1. Introduction

With the rapid development of computer graphics and digital video processing technology, 2-dimensional video cannot satisfy people's growing demand, three-dimensional televisions towards a more natural and live visual home entertainment experience, which attracted intensive research interests and has become the next logical development [1]. The multi-view video system consists of many texture views captured by cameras[2], since multi-view camera sets acquired the same scene from different viewpoints, they contains large amount of redundant information. Restricted to the limited bandwidth, we cannot transmit all the captured pictures. The common solution is to synthesize the intermediate views; the traditional method is using DIBR (depth image based rendering).

Multi-view Video plus Depth (MVD) data format enables advanced video synthesis method in which depth map plays an important role in the view synthesis for 3D video applications, depth map is a gray image, the darker regions in the depth map represent far-off objects and lighter regions represent closer objects. With smooth regions and sharp edges in depth map, the most important information rely in the edge region, so we need to find efficient method to estimate and compress an accurate depth map with edge parts.

At present, the depth map can be acquisition from depth camera and depth estimation algorithm. For example, computing the depth map algorithmically by stereo matching, or using depth-sensing devices or range-finding cameras. Among them, the depth-sensing devices including TOF camera, stereo camera, laser scanner, camera light structure, *etc.* [3]. These sensors are using light volatility to measure the distance from the object's surfaces to the camera. But their principle is different. TOF camera detection distance according to the object surface reflection phase shift, however, the depth camera is based on the observed picture, using the depth estimation algorithm to obtain depth information, to produce a gap. On the other side, most of the depth estimation algorithm of 3d video system is based on

ISSN: 2005-4254 IJSIP

energy minimization framework of global method. Energy function contains data item and smooth item. Data item or matching cost item usually gets from the difference of matching point strength or color. Generally, neighborhood pixels parallax around the object boundary is smooth, smooth item commonly used algorithm of minimum energy is incredibly spread [4, 5] and graph cut [6, 7]. For the depth map estimation, MPEG put forward View Synthesis reference software (VSRS), by stereo matching estimation algorithm to obtain the depth of the data. Although these methods can give the scene a more accurate representation of geometry, which is complicated in calculation, so it's hard real-time implementation. In general, not like the traditional image sensor, the depth camera is expensive and depth detection distance is small, while stereo matching algorithm is usually with good effect, but high complexity.

In 2010, Microsoft introduced X360 as kinect's external device, which mainly has a camera, infrared camera, infrared projectors, a set of mic and motor. Microsoft's kinect sensor has the highest influence on the robotics and mechatronics communities in the last three years, which can generates a dense depth map and color image of the scene , the resolution is typically 640×480 at real-time rate of 30 frames/sec, which is comparable to a time-of-flight camera [8].

Compressive sensing theory is a hot research area in image processing, such as image denoising, image coding, *etc.*, CS theory enable certain images can be reconstruct from far fewer measurements which breaks the limitation of Nyquist sampling criterion and has worldwide used [9-10]. Compressed sensing reconstruction algorithm can be divided into two categories: the convex optimization algorithm, greedy algorithm. The convex optimization algorithm by adding constraints to obtain the sparse representation, and the reconstruction method including the basis pursuit method and variation total algorithm, which can get good reconstruction result. Greedy algorithm is based on the orthogonal matching pursuit method. The reconstruction precision is lower than convex optimization algorithm [11].

As we all known that the conditional depth map compression method such as H.264 using prediction method to estimate the neighboring area, which only considering the horizontal, vertical or diagonal edges ,which cannot efficiently predict edge blocks for depth map with arbitrary edge shapes and reduce the extra bits [12]. On the other hand, CS method can be more efficient to compress depth map and preserve the edge area than traditional method.

In this paper, we propose an improved depth map coding method incorporated with the compressed sensing theory to improve the quality of intermediate view, which plays an important role in the 3D video applications [13], and the efficient compression of depth map is need to reduce the extra bits.

## 2. Optimization of depth map based on GMM

Despite the kinect sensor has apparent advantages, there are still some disadvantages in the Kinect generated depth maps. The kinect sensor consisting of an infrared projector and an infrared camera at distance of 7.5cm. The projector using an IR laser with 830nm wavelength capturing the scene and calculate correspondence between projector and the camera. But in some special surfaces like glass wall which deflect infrared light result in losing depth value in such areas. The area that losing depth value is referred to as a "hole" with zero depth value [14, 15], as their depth cannot be captured.

In the real scene, usually exist multiple foreground objects take on the characteristics of the multimodal, so, we need to use multiple probability model to describe the change. Gaussian mixture model can use more distribution to analysis more changeable object, it is the effective method to solve the problem of multiple modal. The method is based on the distribution of each pixel establish gaussian mixture model [16-17] to show the pixel color change, so distinguish the background or foreground based on variance and weights.

Foreground and background separation process based on GMM model: First of all, defined several gaussian models in the process of initialization, such as, the mean, the variance and the weights of the gaussian model, *etc.*, and to calculate the Yilmaz distance. Secondly, the pixels in each frame is processed separately, if it match with the best model, it is classified as the model, the model parameters will be update at the same time, if not match, set a GMM model for this pixels, and initialization parameters, instead of the original unreliable model. Finally choose the best background model, in preparation for the background object extraction.

Assuming that the characteristics of each pixel Expressed in *K* gaussian model, *K* values are usually 3 ~ 5. After get a new frame ,update the gaussian mixture model ,make the current each pixel in the image match with the gaussian mixture model, if successful matched, that point belong to the background, otherwise is in the foreground. If use the variable *Xt* to represent the color value of each pixel, that *K* of the probability distribution function of gaussian function can be represented as:

$$p(X_t) = \sum_{i=1}^{K} \omega_{(i,t)} \eta(X_t, \mu_{i,t}, \sum{}_{i,t}) \tag{1}$$

Among them, $\eta(X_t, \mu_{i,t}, \sum{}_{i,t})$ represent the moment of *t* the *ith* gaussian distribution,

$\sum{}_{i,t}$ is the variance, $\mu_{i,t}$ is the average.

$$\eta(X_t, \mu_{i,t}, \sum{}_{i,t}) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\sum{}_{i,t}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(Xt-\mu_{i,t})^T \sum_{i,t}^{-1}(Xt-\mu_{i,t})} \tag{2}$$

We capturing the first frame of depth map and texture map as the initial maps and use frame difference method for the background update, after that we using background depth value to fill the hole in the current frame.

## 3. Overview of Compressive Sensing

Nyquist sampling criterion prove that signal can be exactly reconstructed from sampling signal at sampling rate greater than two times bandwidth. CS theory break the Nyquist sampling criterion, and predict that many signals can be represented or approximated with only a few coefficients in a suitable basis [18-20].

Suppose a spare signal *X*, $X \in R^N$,is a *K*-sparse when it has at most *K* nonzero elements, and the number $K \ll N$.Consider *X* can be represented by a set of orthonormal basis $\psi = [\psi_1, \psi_2, \cdots, \psi_N]$, $X = \psi\theta$, $\theta$ is a coefficient vector.

We can express the sampling process as follows:

$$Y = \Phi X \tag{3}$$

Where *Y* is the measurement, and $\Phi$ is the M x N measurement matrix that takes *M* number of measurement from *X*.

According to the CS theory, we can recover the *K*-sparse signal *X* by estimating *X* from the measurement *Y*. So we have to find the sparsest solution from minimization problem as the following description:

$$\tilde{\theta} = \arg\min_{\theta} \|\theta\|_1, \quad s.t. \quad Y = \Phi X \tag{4}$$

by solving the minimization *L1*-norm problem ,we can perfectly reconstruct *X* by $\tilde{X} = \psi\tilde{\theta}$

The main title (on the first page) should begin 1 3/16 inches (7 picas) from the top edge of the page, centered, and in Times New Roman 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Please initially capitalize only the first word in other titles, including section titles and first, second, and third-order headings (for example, “Titles and headings” — as in these guidelines). Leave two blank lines after the title.

## 4. Depth Map Reconstruction and View Rendering

Depth map plays an important role in the view rendering in 3D video applications [21].We find that depth map usually has large smooth areas and very sharp edges, which result in sparse gradient, so We incorporate an additional total variation (TV) minimization for reconstruct the depth map for preserving the edges and avoid introducing noise to the results. We reconstruct the data by Using the reconstructed disparity map $\hat{d}$ computed by performing above method, we can get the desired intermediate view $I_s$ by linear interpolation, as follows:

$$I_s(x,y) = \begin{cases} I\ (x+\alpha*\hat{d}\ ,y) \\ \quad if\ (x+\alpha*\hat{d}\ <= ncols) and (x-\alpha*\hat{d}_L > 0) \\ 0 \\ \quad esle \end{cases} \tag{5}$$

## 5. Results

To test the efficiency of our method, in this section, we present simulation results to show that our algorithm is good at staircase artifact; we use a Microsoft Kinect depth sensor connected to a computer running the Windows 7 operating system. The depth maps and the corresponding color images, the color image is a three channel RGB image and the depth map is a single channel gray image. The proposed method is implementing in MATLAB, C, and OpenCV.

The results of the proposed algorithm can be seen in Figure 1 and it is clear that all the holes have been filled and the edges have been refined. In Figure 2, the view rendering quality of the proposed algorithm is compared with the method that not using the CS theory, we can see that the proposed method outperforms. The subjective quality difference can be clearly observed among the enlarged portions of above images, shown as Figure 2, from left to right, are the enlarged portions of the original images (Figure 2 (b), Figure 2(f)), the intermediate views reconstructed using our method (Figure 2 (d), Figure 2(h)), respectively. Note that, the edge and texture can be preserved well in our method, which indicates that the depths change smoothly and consistently, so that our method generates intermediate views with higher quality.

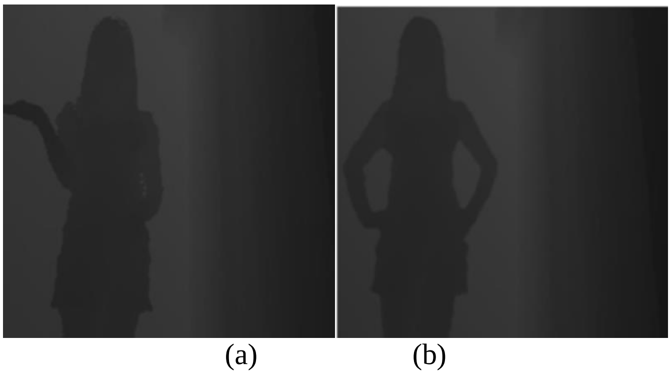

Figure 1. The Depth Map Obtained by Our Method

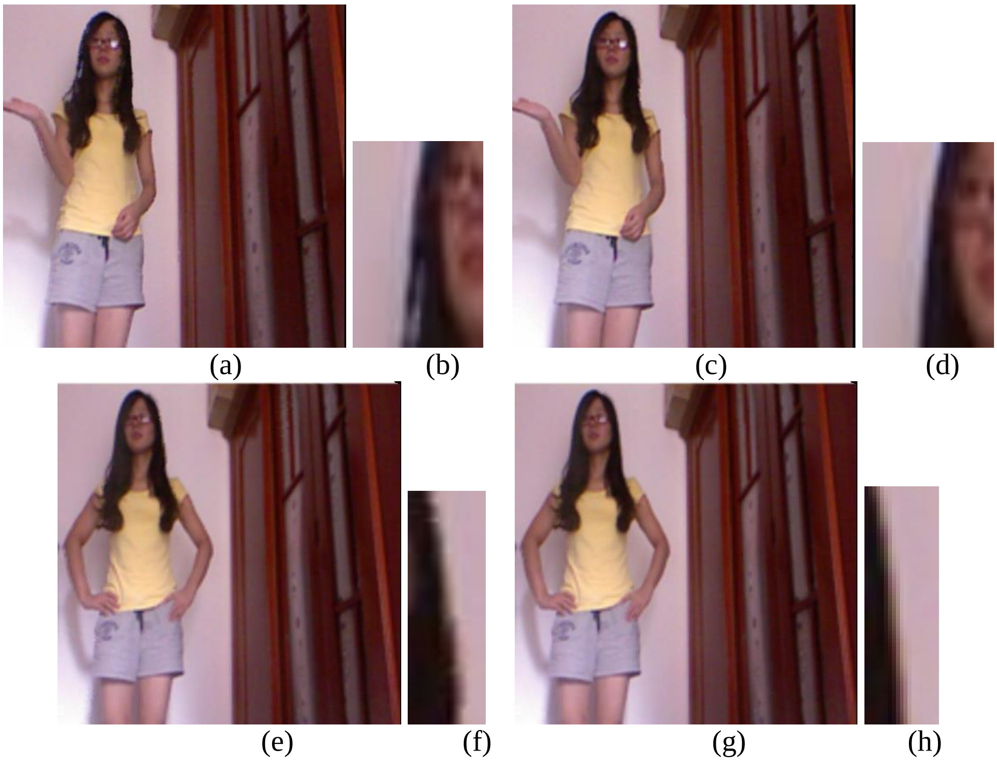

Figure 1. Reconstructed View Obtained by (a), (e) Original method, (c), (g) Our Method, (b), (d), (f), (h) the Enlarged Portions

## 6. Conclusion

With further research about depth extraction and multi-view render, we investigated a depth extraction method by Kinect for the view rendering. By using GMM algorithm and CS theory, large gaps could be filled without blurring between foregrounds and background and efficiently estimate the depth map with arbitrary edge shapes. Experiments show that this algorithm effectively filling the holes in the depth map.

## ACKNOWLEDGEMENTS

This research was jointly sponsored by Natural Science Foundation of Shanghai (No. 15ZR1431500), Key Laboratory of Advanced Display and System Applications (Shanghai University), Ministry of Education, China (Project No. P200803), General research project of Shanghai Normal University (Grant No. SK201225), Shanghai municipal education commission of scientific research innovation projects (Grant No. 14ZZ125), Funding scheme for training young teachers in Shanghai Colleges,.

## References

[1] W. Motusik, *et al.*, “3DTV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes”, ACM, **(2004)**, pp. 23-24.
[2] E. Cooke, P. Kauff and T. Sikora, “Multi-view synthesis: A novel view creation approach for free viewpoint video”, signal processing-image communication, vol. 21, no. 6, **(2009)**, pp. 476-492.
[3] J. Fu, D. Miao, W. Yu, S. Wang, Y. Lu and S. Li, “Kinect-like depth data compression”, IEEE Transactions on Multimedia, vol. 15, no. 6, **(2013)**, pp. 1340-1352.
[4] R. K. Gupta and S.-Y. Cho, “Stereo correspondence using efficient hierarchical belief propagation”, Neural Computing and Applications, vol. 21, no. 7, **(2012)** October, pp. 1585-1592.
[5] K. Zhang and Z. Gao, “The technology of fast 3d reconstruction based on stereo vision”, Key Manufacturing Automation Technology and Application, vol. 579, no. 580, **(2014)**, pp. 654-658.
[6] G. Yu, J. Liu, X. Xie and J. Zeng, “A global stereo matching algorithm based on adaptive support-weight and Graph cut”, Applied Mechanics and Materials, vol. 151, **(2012)**, pp. 612-616.
[7] M. Huan, K. Wang, W. Zuo and Z. Li, “Template based stereo matching using graph-cut”, Proceedings 2011 International Conference on Instrumentation, Measurement, Computer, Communication and Control, **(2011)**, pp. 303-306.
[8] K. R. Vijayanagar, M. Loghman and J. Kim, “Real-Time Refinement of Kinect Depth Maps using Multi-Resolution Anisotropic Diffusion”, Mobile Network, Appl, **(2013)**, pp. 9.
[9] E. J. Candes, “Compressive sampling”, in Int. Congress of Mathematics, Spain, **(2006)**.
[10] E. Candes and J. Romberg, “Sparsity and incoherence in compressive sampling”, Inverse Problems, vol. 23, **(2007)**, pp. 969.
[11] Z. Yanmeng and S. Jiamxin, “Method for Image Denoising Based on Compressed Sensing Total Variation Algorithm”, video engineering, vol. 38, no. 5, **(2014)**, pp. 5-8.
[12] T. M. Cho, Y. Lee and J. Shin, “A Homogenizing Filter for Depth Map Compressive Sensing Using Edge-Awarded Method”, ICTC, **(2013)**, pp. 591-595.
[13] S. Tao, Y. Chen, M. M. Hannuksela, Y.-K. Wang, M. Gabbouj and H. Li, “Joint texture and depth map video coding based on the scalable extension of H.264/AVC”, in 2009. ISCAS 2009 IEEE International Symposium on Circuits and Systems, **(2009)**, pp. 2353-2356.
[14] M. Bertalmio, G. Sapiro, V. Caselles and C. Ballester, “Image inpainting,” in Proc. ACM SIGGRAPH, **(2000)**, pp. 417–424.
[15] Z. Tauber, Z.-N. Li and M. Drew, “Review and preview: Disocclusion by inpainting for image-based rendering,” IEEE Trans. Syst., Man, Cybern., Part C: Appl. Rev., vol. 37, no. 4, **(2007)** July, pp. 527–540.
[16] W. Lijuan, “Moving objects detection based on opencv and Gaussian mixture model”, Electronic test, vol. 9, **(2009)**.
[17] W. Liangsheng and C. Yinhang, “An Improved Background Subtraction Using Adaptive Gaussian Mixture Models”, Journal of northern Jiaotong Universiy, vol. 27, no. 6, **(2013)**, pp. 22-25.
[18] E. Cand`es, J. Romberg and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information”, IEEE Trans. Inform. Theory, vol. 52, **(2006)**, pp. 489–509.
[19] D. Donoho, “Compressive sensing”, IEEE Trans. Inform. Theory, vol. 52, **(2006)**, pp. 1289–1306.
[20] E. Cand`es and M. Wakin, “An introduction to compressive sampling”, IEEE Signal Processing Magazine, vol. 25, **(2008)**, pp. 21–30.
[21] S. Tao, Y. Chen, M. M. Hannuksela, Y.-K. Wang, M. Gabbouj and H. Li, “Joint texture and depth map video coding based on the scalable extension of H.264/AVC”, in 2009 ISCAS 2009 IEEE InternationalSymposium on Circuits and Systems, **(2009)**, pp. 2353-2356.
[22] Q. Zhang, P. An, Z.-Y. Zhang, H. Wang, Y-F Wu and G-Y Jiang, “New reconstruction method for intermediate views from multiple Views”, imaging science journal, vol. 58, no. 2, **(2010)**, pp. 89-95.