Overcomplete Compressed Sensing of Ray Space for Generating Free Viewpoint Images

Qiang Yao, Keita Takahashi, Toshiaki Fujii
Graduate School of Engineering, Nagoya University, Nagoya, Japan
E-mail: yaoqiang@fujii.nuee.nagoya-u.ac.jp Tel: +81-52-789-3163

***Abstract*—Free Viewpoint Image (FVI) technique has gained a great popularity because it enables people to freely choose the viewpoints from which they watch the targets and scenes. Generation of FVI requires all the information of a 3-D virtual space. Ray space is a direct representation of complete 3-D information, and FVIs can be generated simply by cutting ray space. However, in the straightforward construction of a ray space, a numerous number of images have to be captured in advance which triggered data explosive from data acquisiton to data storage and transmission. In this paper, we focus on compressed sensing to sparsely capture a ray space at encoder and reconstruct it at decoder. Specifically, we propose to adopt overcomplete dictionaries which are produced by learning methods, to sparsely represent the ray space data. Experimental results show that the proposed dictionary is much better than the structured dictionary in previous work or orthogonal basis in terms of the reconstruction quality of the ray space from compressively sensed data.**

I. Introduction

Recently, Free Viewpoint Image (FVI) technique has obtained increasing attention due to the fact that FVI enables people to freely choose the viewpoints from which they watch the targets and scenes [1]. In order to generate FVIs, Depth Image Based Rendering (DIBR) could be employed [2], where the corresponding depth information has also to be either captured directly by depth cameras or estimated from color images by using stereo methods. However, the resolution of depth cameras is always not as high as the resolution of color images, and unfortunately the depth information estimated by stereo matching methods in practical situations is always not as accurate as it is expected.

Another method of generating FVIs is to employ ray space [3], [4] (light field [5] in other literature). The ray space is constructed from the images that are captured by many cameras horizontally arranged in front of the objects. An example is shown in **Figure 1**, where v and w denote the image coordinate and u denotes the viewpoint coordinate. If the interval between each viewpoint image is quite small, which means the sampling density along the u axis is quite high, the ray space can contain all the information of a 3-D virtual space. In such a case, by simply cutting the ray space, images from any viewpoint or FVI can be generated. For instance, as shown in **Figure 1**, intersection 1 and intersection 2 correspond to different viewpoints.

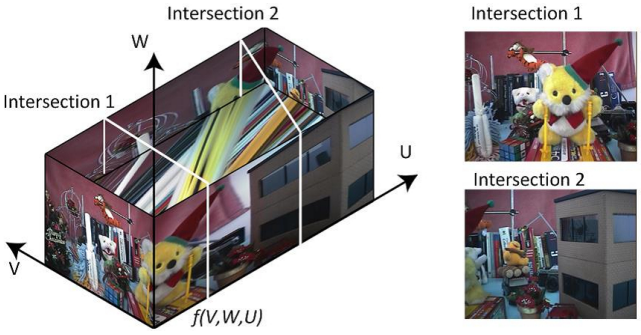


Fig. 1. Generation of FVI by cutting ray space

However, to meet the requirement of viewpoint density is quite difficult. The Plenoptic Sampling theory [6] states that the difference between the maximum and minimum disparities from two adjacent cameras has to be at most one pixel. Generally, an objective does not always share the same depth but has some range of depth. Furthermore, the overlaps between two or more objectives are always unavoidable. Thus, the viewpoints are supposed to be arranged quite closely in order to keep the disparity range in one pixel. Therefore, the composition of ray space requires a large number of images, which has posed a great obstacle from data acquisition to data storage and transmission. Especially, as the increase of viewpoints and the image resolution, the amount of data will be extremely large and no longer acceptable in real application. Meanwhile, it is known that due to the high redundancy of ray space, most of data will be thrown away in compression, and only a small part will be left for transmission. Thus, all of mentioned above drive us to think whether it is possible to capture only the necessary data in acquisition and recover all afterwards, which leads us to resort to compressed sensing [7], [8].

Babacan et.al. [9] applied compressed sensing methods to develop a light field acquisition system. Their method relied on the disparity information that was used as priors for the light field reconstruction. In contrast, our work does not rely on any disparity information but tries to explore the basic structure of the ray space itself. In our previous work [10], we adopted structured Gabor dictionaries to represent ray space data and applied compressed sensing method to sparsely sense a ray

This work was supported by Grant-in-Aid for Scientific Research (C) with Number 24560450.

space at encoder and reconstruct it at decoder. In this paper, we newly adopt dictionaries which are shaped by learning methods. Experimental result shows that sparser representation of ray space can be achieved and the performance of overcomplete compressed sensing is also improved by using the proposed dictionary. Besides, the size of the proposed dictionary can be greatly reduced compared to the previous structured dictionary, which reduces the computation cost.

The paper is organized as follows. In section 2, sparse representation and traditional compressed sensing are introduced. In section 3, after briefly presenting the concept of ray space, we propose to construct overcomplete dictionaries by learning to sparsely represent the ray space data. Besides, we also extend the compressed sensing from orthogonal basis to overcomplete dictionaries. Section 4 describes the simulation procedure, and experimental results are presented to illustrate that proposed dictionary performs much better than previous works. A brief conclusion and future direction are mentioned in the last section.

II. Compressed Sensing and Sparse Representation

In this section, we briefly review the background of compressed sensing and sparse representation. In the next section we will further apply them to the ray space sensing problem.

Traditionally, compressed sensing [7], [8] enables a sparse signal to be recovered from fewer nonadaptive and linear measurements by certain optimization tool with sparsity promotion. Assume the target signal required to capture is noted as $y_{N\times 1}$. Then, $y_{N\times 1}$ can be analyzed by the equation $y_{N\times 1} = \Phi_{N\times N} x_{N\times 1}$, where $\Phi_{N\times N} = [\phi_1, \phi_2, ..., \phi_N]$ is a transforming matrix or compressed matrix and $x_{N\times 1}$ is the corresponding coefficients. Strictly speaking, the sparsity of $y_{N\times 1}$ is defined as the number of nonzero elements in $x_{N\times 1}$. More practically, an approximated definition is adopted as follows. There is x_s representing the s largest elements in x, thus $y_s = \Phi_{N\times N} x_s$ is an approximation of y with the error $\epsilon = \| y - y_s \|_{l_2}$. If the error ϵ is smaller than a certain threshold, the signal $y_{N\times 1}$ is named as $s-$sparse signal and the sparsity of signal y is s. Besides, it is regulated that the sparsity of y is higher as s becomes smaller. Next, the sensing matrix is defined as $\Psi_{P\times N}$ with $P < N$. In compressed sensing, less data can be directly obtained by projecting signal $y_{N\times 1}$ to a sensing matrix $\Psi_{P\times N}$. The obtained measurement $z_{P\times 1}$ can be written as $z_{P\times 1} = \Psi_{P\times N} y_{N\times 1}$, which reduces the data size from N to P in the acquisition. In the reconstruction stage, $y_{N\times 1}$ can be recovered from the sparsely sensed measurements $z_{P\times 1}$ using optimization with sparsity promotion.

Basically, there are two requirements in traditional compressed sensing. One is high sparsity of signal $y_{N\times 1}$ in a certain transform domain $x_{N\times 1}$ by the matrix $\Phi_{N\times N}$ and the other is the low mutual coherence between sensing matrix Ψ and compressed matrix Φ. The mutual coherence between the two matrices is defined as $\mu(\Psi, \Phi) = \max\{\psi_i^T \phi_j, 1 \le i \le P, 1 \le j \le N\}$, where all the columns in Ψ and Φ are l_2 normalized. The upper bound of $\mu(\Psi, \Phi)$ is 1 when there is the same column in both matrices. Commonly, in order to keep the mutual coherence low, the sensing matrix $\Psi_{P\times N}$ is set to be random because random matrices are regarded to be irrelevant to any other matrices. Recently, however, Rauhut in [11] and Candes in [12] proposed and proved theoretically that an overcomplete dictionary could also be applied in compressed sensing for approximation recovery, even if the mutual coherence reached the maximum value. Therefore, in this work, we keep the sensing matrix as random one and take an overcomplete dictionary by learning as the compressed matrix in order to seek sparser representation of signal $y_{N\times 1}$.

The essence of sparse representation by dictionary learning [13], [14], [15] is that an observation of signal $Y_{N\times L} = [y_1, y_2, ..., y_L]$, where each column $y_i \in R^N$ is a signal, can be sparsely represented by a few of elements from a dictionary $D_{N\times M} = [d_1, d_2, ..., d_M], d_i \in R^N, M > N$ which can be adaptively learned from a set of training data. Each column d_i in dictionary $D_{N\times M}$ is called an atom, and $X_{M\times L}$ is comoposed of the corresponding coefficients of atoms in dictionary, resulting the equation $Y_{N\times L} = D_{N\times M} X_{M\times L}$, where $X_{M\times L}$ is a sparse matrix. The whole procedure of overcomplete dictionary learning consists of two parts. One is sparse coding and the other is dictionary update. In the sparse coding stage, the dictionary $D_{N\times M}$ is assumed to be known and fixed while obtaining sparse representation $X_{M\times L}$ by greedy pursuit algorithm, such as Orthogonal Matching Pursuit (OMP) [16]. In the part of the dictionary update, the obtained sparse representation $X_{M\times L}$ is employed with the signal $Y_{N\times L}$ for updating the dictionary $D_{N\times M}$. Thus, dictionary is updated gradually by each iteration k.

III. Sparsity Pursuit of EPI and Overcomplete Sensing of Ray Space

In this section, we firstly give a brief description of ray space and its basic unit, epipolar plane image (EPI), and point out the unique feature of EPI. Next, we propose to adopt dictionary learning to exploit the feature of EPI. Furthermore, the learned overcomplete dictionary is adopted as the compressed matrix in the compressed sensing framework and applied to the problem of the ray space sampling.

A. Ray space and EPI

Consider a ray space shown in **Figure 2**, where v and w are the image coordinate while u is the coordinate of viewpoint. As the basic unit of ray space, an epipolar plane image (EPI) is actually the route of rays on each $u - v$ plane, and it is composed of several straight lines with different slopes. The slope of EPI directly reflects the depth information of the object in real space. Since EPIs have quite single structures, we assume that EPIs can be recovered from the sparsely sensed measurements. In the following part, we focus on how to sense and reconstruct EPIs, but it is equivalent to the process for the entire ray space.

B. Sparsity Pursuit of EPI by Learning Method

It is assumed that one EPI is partitioned into L blocks in size $\sqrt{N} \times \sqrt{N}$ and each block is unfolded to one vector

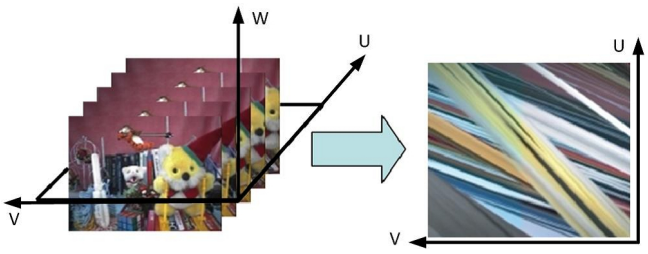


Fig. 2. The ray space and EPI

$y_{N\times 1}$to be stacked in a signal matrix $Y_{N\times L} = [y_1, y_2, ...y_L]$, where L is the number of signal. Thus, we would like to shape one dictionary, $D_{N\times M} = [d_1, d_2, ..., d_M], d_i \in R^N$, where M is the number of atoms. Since the dictionary is overcomplete, we set $M > N$. To enforce sparsity, $Y_{N\times L}$ is supposed to be approximated by only few combinations of atoms in $D_{N\times M}$ with corresponding coefficients in $X_{M\times L} = [x_1, x_2, ..., x_L]$. It can be specifically formulated as

$$\begin{aligned} &arg \min_{D_{N\times M}, X_{M\times L}} \|Y_{N\times L} - D_{N\times M}X_{M\times L}\|_F^2 \\ &\text{s.t. } \|x_i\|_{l_0} \le s, x_i \in R^M, 1 \le i \le L \end{aligned} \tag{1}$$

where $\| \bullet \|_F$ is frobenius norm and $\| \bullet \|_{l_0}$ is the l_0 quasi norm. In addition, s represents the number of non-zero elements in x_i and $s << M$.

Since there are two unknowns, $D_{N\times M}$ and $X_{M\times L}$, in Eq. (1), an alternate projection method is adopted. Firstly, $D_{N\times M}$ is fixed to obtain sparse representation $X_{M\times L}$, and it can be formulated as

$$arg \min_{x_i, 1\le i\le L} \| y_i - D_{N\times M}x_i \|_{l_2}, \text{s.t. } \| x_i \|_{l_0} < s. \tag{2}$$

It is a NP-hard problem to find the globally optimal solution for this problem with l_0 regularization. Thus, we adopt OMP to find the suboptimal solution $X_{M\times L}$. Next, in the stage of dictionary update, the optimization problem is written as

$$\begin{aligned} &arg \min_{D_{N\times M}} f(D_{N\times M}), \\ &\text{where } f(D_{N\times M}) = \| Y_{N\times L} - D_{N\times M}X_{M\times L} \|_{l_2}^2, \\ &\text{s.t. } d_i^T d_i = 1, 1 \le i \le M. \end{aligned} \tag{3}$$

We take the derivative of f with respect to $D_{N\times M}$ and set it to be zero, and it is easy to obtain

$$\frac{\partial f}{\partial D_{N\times M}} = -2(Y_{N\times L} - D_{N\times M}X_{M\times L})X_{M\times L}^T = 0. \tag{4}$$

Thus, finally we have

$$D_{N\times M} = Y_{N\times L}X_{M\times L}^T(X_{M\times L}X_{M\times L}^T)^{-1}. \tag{5}$$

At the end of one iteration, each column of the dictionary is l_2 normalized. Therefore, after several iterations, the dictionary $D_{N\times M}$ is expected to be well shaped for sparse representation of $Y_{N\times L}$.

Obviously, the main purpose of dictionary learning is to shape the dictionary so that the approximation error of given signal is minimized, thus the dictionary can fully represent the feature of EPI. In addition, the size of dictionary, M, can be preset before learning, we can keep the dictionary size compact to reduce the computation cost.

C. Overcomplete Compressed Sensing of Ray Space

Next, the overcomplete dictionary mentioned in Section 3-B is adopted in compressed sensing framework. Since the dictionary is no longer orthogonal but overcomplete, we refer to this extended model as overcomplete compressed sensing.

In the sensing process, an EPI $Y_{N\times L}$ is measured after it is projected by a sensing matrix, $\Psi_{P\times N} = [\psi_1, \psi_2, ...\psi_p]^T$, where $\psi_i \in R^N, 1 \le i \le P$. The projected measurements $Z_{P\times L} = [z_1, z_2, ...z_L]$, where $P < N$, can be formulated as $Z_{P\times L} = \Psi_{P\times N}Y_{N\times L}$. Thus, the number of samples that are actually measured is reduced from NL to PL. Note that L was the number of training data in the previous learning procedure, but here we refer to it as the number of EPI blocks in the whole ray space which we want to obtain.

Next, in the recovery stage, we consider to recover y_i from z_i, and finally $Y_{N\times L}$. However, y_i is not directly recovered, but the sparse representation x_i is recovered in the first place. The approximated sparse representation $\hat{x}_i$ is explored from z_i and the merged matrix $(\Psi D)_{P\times M}$ by optimization with sparsity promotion as follows:

$$\begin{aligned} &\hat{x}_i = arg \min_{x_i, 1\le i\le L} \| z_i - (\Psi D)_{P\times M}x_i \|_{l_2}, \\ &\text{s.t. } \| x_i \|_{l_0} \le \epsilon \end{aligned} \tag{6}$$

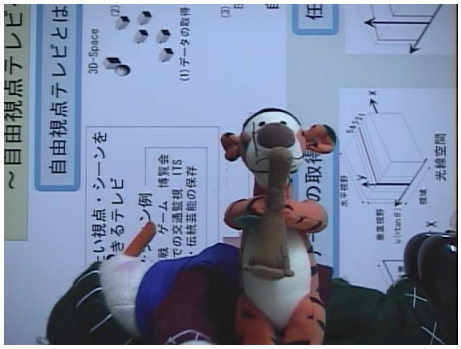
Similarly, in order to find the globally optimal solution for this problem with l_0 regularization, the exhaustive sweep has to be conducted through all the possible supports, $\binom{s}{N}$ and it is also a NP-hard problem. Thus, OMP is employed to search the sub-optimal solution. Then, from the recovered $\hat{x}_i$, the EPI $\hat{y}_i$ can be obtained from $\hat{y}_i = D_{N\times M}\hat{x}_i$.

IV. Experimental Results

We firstly checked whether the proposed dictionary can have a better representation of EPIs. Secondly, we simulated the sensed measurements by using random sensing matrix, and then obtained the recovered EPIs by adopting OMP method. The recovered results by the proposed dictionary and other dictionary or basis were compared to show the advantage of the proposed dictionary.

Two data sets, named as 'Fuzzy' and 'Kuma' were employed throughout the experiments in this paper. There were 64 viewpoints closely arranged and the resolution of each viewpoint image was 640×480. One viewpoint images (No.25) from both data sets are shown in **Figure 3**. All the 64 images were aligned together for each data set to construct the ray space with the resolution of $640 \times 480 \times 64$. Next, EPIs were generated by cutting the ray space horizontally, and the resolution of an EPI was 640×64, indicating that there were 480 EPIs in the constructed ray space. EPIs from both ray spaces are shown in **Figure 4**. It can be observed that the EPI from 'Fuzzy', which included more objects with different depth, is more complex than the EPI from 'Kuma'. Each EPI

(a) FVI in 'Fuzzy'



(b) FVI in 'Kuma'

Fig. 3. FVIs in two ray spaces

(a) EPI in 'Fuzzy'

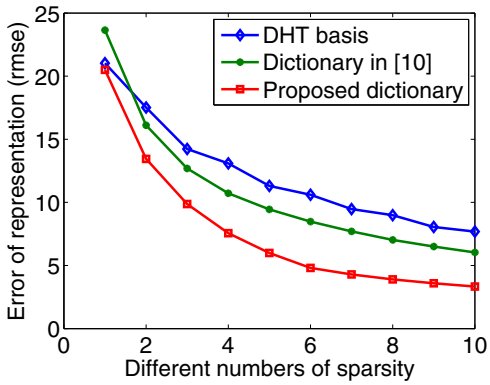
(b) EPI in 'Kuma'

Fig. 4. EPIs in two ray spaces

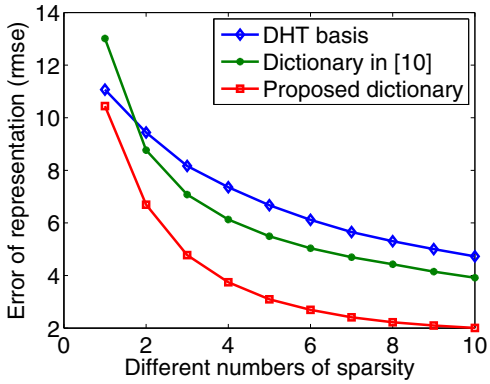
was divided into 8×8 blocks and each block was unfolded to be a 1×64 vector y_i.

Three candidates, orthogonal basis (DHT), structured dictionary in [10], and the proposed dictionary were compared in terms of sparsity. The size of orthogonal basis is 64×64, and only the largest s coefficients in X were used for the approximation. The structured dictionary is highly redundant with the size of 64×740 for 'Fuzzy'and 64×610 for 'Kuma' respectively.* The size of the proposed dictionary is only 64×128, which is much smaller than the structured dictionary, and it is learned by 20 iterations from 7200 random-chosen EPI blocks. In **Figure 5**, the horizontal axis represents the desired sparsity s in sparse representation, while the vertical axis corresponds to the error which is evaluated by root mean square error (RMSE). Clearly, from the graphs, the proposed

*The original size of structure dictionary was huge, beyond 3000, due to the large parameter space. To keep the dictionary size acceptable, we analyzed a statistics in sparse coding and kicked out the atoms which were rarely used in representation.



(a) Result for 'Fuzzy' EPI



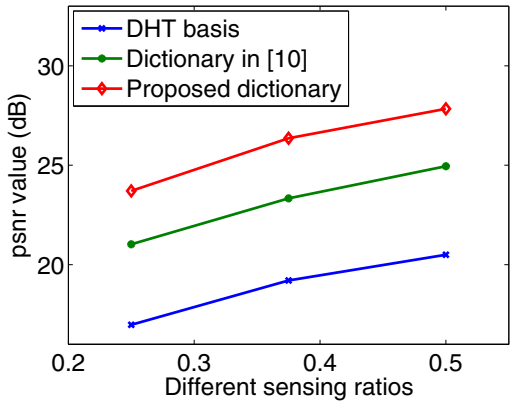
(b) Result for 'Kuma' EPI

Fig. 5. Sparsity representation of EPI in orthogonal basis, structured dictionary and proposed dictionary

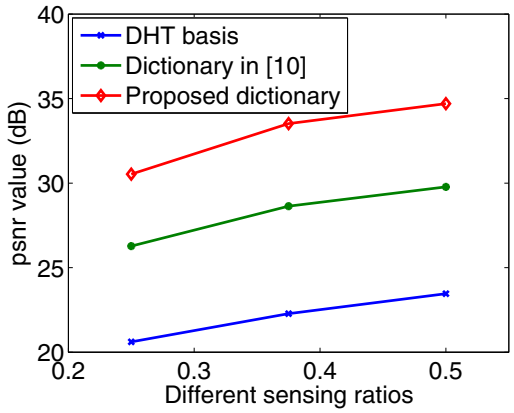
dictionary performed the best for both ray spaces, achieving the least error for the same spasity.

Next, the three candidates mentioned above were compared in the framework of compressed sensing. A Gaussian random matrix was adopted as the sensing matrix for all the cases, and the sensing ratio R, which is the fraction between the number of sensed data and the number of original pixels, was set to $1/4$, $3/8$ and $1/2$ respectively. The desired sparsity was set to $s = 6$ in the recovery of a randomly selected EPI, and the reconstruction errors (PSNR values) were compared among the three candidates. Since the sensing procedure was random, we conducted sensing and recovery for 10 times and the average reconstruction errors are presented in **Figure 6**. From the graph, the proposed dictionary can achieve 3-5 dB improvement in average compared to the structured dictionary, and much better result than orthogonal basis.

All the recovered EPIs are stacked together to reconstruct the whole ray space. **Figures 7** and **8** present the generated FVIs from the ray spaces that are reconstructed by using different compressed matrices. The sensing ratio was set as $R = 1/2$. It is shown that the generated FVIs by the proposed dictionary have better subjective quality with higher PSNR values than the other two candidates, which is consistant with the result of sparsity comparison in **Figure 5**. Furthermore, the



(a) Result for 'Fuzzy' EPI



(b) Result for 'Kuma' EPI

Fig. 6. Recovery results of two EPIs by using various of sensing ratios

reconstruction qualities of the two ray spaces are different; 'Kuma' is reconstructed better than 'Fuzzy' for the same condition. This is due to the complexity difference between the two ray spaces; 'Fuzzy' has more objects with different depths than 'Kuma', resulting in more complex structures in the ray space.

Finally, we show a comparison among the three cases with respect to the computational time required for the recovery of one EPI. We used a PC with a 3.20GHz Intel(R) Core(TM) i7 CPU and 3.0 GB main memory, and developed the software using Matlab 2013a without parallelization. As for 'Fuzzy' EPI, the orthogonal basis, the structured dictionary [10], and the proposed dictionary consume 1.03s, 69.89s and 12.62s, respectively. As for 'Kuma' EPI, the computational time of three candidates are 1.11s, 55.84s, and 12.64s respectively. Obviously, the orthogonal basis requires the smallest computational cost but it produces the worst reconstruction result. Compared to the structured dictionary, the proposed dictionary requires less computational cost and achieves better reconstruction quality.

V. Conclusions

In this paper, we discussed the compressed sensing problem of ray space for generation of FVIs. Instead of the structured dictionaries adopted in our previous work [10], we proposed to shape dictionaries by learning so that the features of EPIs can be better and more sparsely represented. Besides, the size of dictionary can be greatly reduced so that computation burden can also be alleviated. Finally, in the simulation, the learned dictionary outperformed the structured dictionary. Our future work is to further analyze the features of ray space to explore sensing matrices better than random ones.

References

[1] M. Tanimoto, M.-P. Tehrani, T. Fujii, and T. Yendo,"Free-Viewpoint TV,", IEEE Signal Processing Magazine, Vol. 28, No. 1, pp.67–76, 2011
[2] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," SPIE Stereoscopic Displays and Virtual Reality Systems XI Vol. 5291, pp.179–187, 2004
[3] T. Fujii, T. Kimoto, and M. Tanimoto, "Ray space coding for 3D visual communication," Proc. Picture Coding Symposium (PCS1996), Vol. 2, pp.447–451, 1996
[4] T. Fujii and M. Tanimoto, "Free viewpoint TV system based on ray-space representatio," Proc. SPIE ITCom, Vol. 4864, pp.175–189, 2002
[5] M. Levoy, P. Hanrahan, "Light field rendering," Proc. SIGGRAPH' 96, pp.31–42, 1996
[6] L. Mcmillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," Proc. SIGGRAPH' 95, pp.39–46, 1995
[7] D.-L. Donoho, "Compressed sensing," IEEE Transactions on Information Theory, Vol. 52, No. 4, pp.1289–1306, 2006
[8] E.J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," IEEE Transactions on Information Theory, Vol. 52, No. 2, pp.489–509, 2006
[9] S.D. Babacan, R. Ansorge, M. Luessi, et.al. "Compressive light field sensing," IEEE Transactions on Image Processing, Vol. 21, No. 12, pp.4746–4757, 2012
[10] Q. Yao, T. Fujii, "Compressed Sensing of Ray Space for Free Viewpoint Image(FVI) Generation," ITE journal, (unpublished)
[11] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed Sensing and Redundant Dictionaries," IEEE Transaction on Information and Theory, Vol. 54, No. 5, pp.2210–2219, 2008
[12] E.J. Candes, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," Applied and Computational Harmonic Analysis, Vol. 31, pp.59–73, 2011
[13] K. Engan, S.O. Aase, and H.-H. John, "Method of Optimal Directions for Frame Design," 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 5, pp.2443–2446, 1999
[14] A. Michal, E. Michael, and B. Alfred, "K-SVD An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," IEEE Transaction on Signal Processing, Vol. 54, No. 11, pp:4311–4322, 2006
[15] T. Ivana, F. Pascal, "Dictionary Learning," IEEE Signal Processing Magazine, Vol. 28, No. 2, pp:27–38, 2011
[16] Y.-C. Pati, R. Rezaiifar, P.-S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, Vol. 1, pp.40-44, 1993

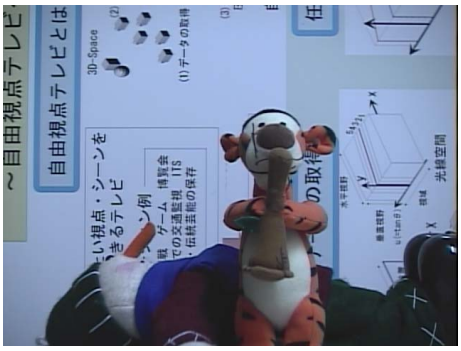
(a) original FVI

(b) DHT basis (21.97dB)

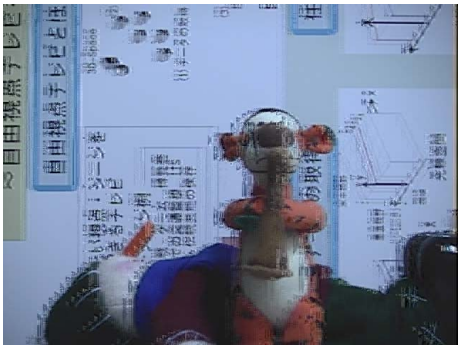
(c) Structured dictionary in [10] (24.39dB)

(d) Proposed dictionary (30.47dB)

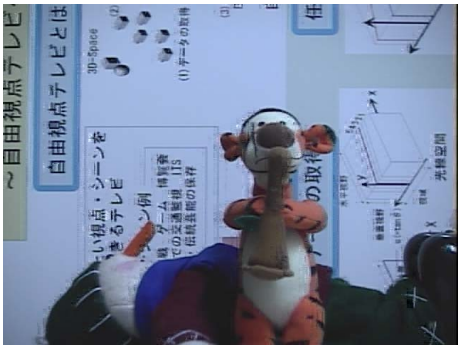
Fig. 7. 'Fuzzy' FVI from reconstructed ray space by using differnent compressed matrices



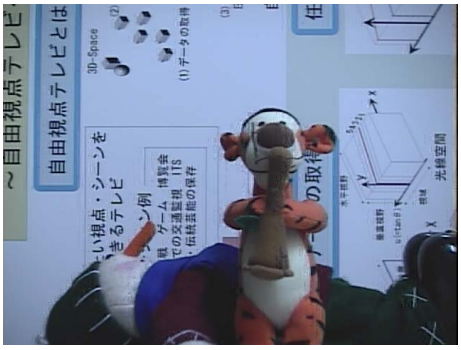
(a) original FVI



(b) DHT basis (23.09dB)



(c) Structured dictionary in [10] (30.03dB)



(d) Proposed dictionary (34.62dB)

Fig. 8. 'Kuma' FVI from reconstructed ray space by using different compressed matrices