JOINT NONLOCAL SPARSE REPRESENTATION FOR DEPTH MAP SUPER-RESOLUTION

Yeda Zhang[1], *Yuan Zhou*[1,2,*], *Aihua Wang*[1], *Qiong Wu*[1] *and Chunping Hou*[1]

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072,China
[2] Electrical Engineering Department, Princeton University, Princeton, USA
Email: zhouyuan@tju.edu.cn

ABSTRACT

Depth image super-resolution reconstruction has gained significant popularity due to its practicability. However, conventional depth image super-resolution reconstruction methods access high frequency information either from a high-resolution depth image database or from a high-resolution color image of the same scene, which is limited in specific applications. In this paper, a novel joint nonlocal sparse representation model is proposed, which is able to capture the interdependency of low-resolution depth and intensity information. As a relative new and not well addressed problem, we reconstruct a high-resolution depth image from a single low-resolution depth image with a low-resolution color image as reference. Experiment results demonstrate that the proposed method outperforms many current state-of-the-art depth map super-resolution approaches on both visual effects and objective image quality.

Index Terms— depth map super-resolution, sparse representation, joint dictionary learning

1. INTRODUCTION

Sensing an accurate depth image is a fundamental task in computer vision. Rapid progress has been made during recent years. Hybrid image-depth cameras such as Microsoft Kinect and Time of Flight (ToF) cameras can provide real-time depth maps and enable a variety of different applications including image-based rendering, pose estimation and 3D scene reconstruction. However, the generated depth images made by these cameras are of low-resolution (LR) and remain afflicted with inaccurate scanning hardware and error during calculating the disparity. Therefore, acquisition of precise depth images in high-resolution (HR) format has become an important research topic in recent years. Many researchers have done a lot in this area and obtained some effect. Generally, according to whether or not high-resolution color images are used, previous works can be divided into two categories.

The first approach enhances the spatial resolution of depth images by using only LR depth images, which typically stems from natural image super-resolution (SR) reconstruction algorithms. By fusing multiple LR depth images of the same scene taken from slightly displaced vantage points [1] [2], the HR depth image can be reconstructed. The authors in [3] and [4] proposed new SR methods that utilizes the stereo view information to enhance the target depth image. Sparse representation based methods are also used in depth map super resolution [5], multi-scale dictionary training [6] improve the SR result remarkably. Deep Convolution Neural Network (DCNN) is introduced in the SR model of Li et al. [7], which take single LR depth map as network input. However, these methods ignore the relationship between the depth image and color image of the same scenario, which causing non-ideal reconstruction result.

The second approach exploits an additional pre-aligned HR color image to help upscale the depth map [8]. Diebel and Thrun [9] first introduced the depth image SR using a MRF formulation with a smoothness term according to the texture derivative from the corresponding color image. This has led to wide research interest on the topic of estimation depths from HR color images. The boundary correspondence of HR color images and depth images are used in LR depth map SR [10]. Zhuo et al. [11] use the global structure of the HR color image to estimate the depth information for indoor scene depth images. Ferstl et al. [12] upscale depth image with an anisotropic diffusion tensor calculated from a HR color image, regarding depth SR as a convex optimization problem with a higher order regularization term and solve it with a piecewise affine solution. Kiechle et al. [13] proposed a joint depth SR (J-DSR) method using the sparse representation model, where high-resolution color images are used as reference to reconstruct a high quality depth image. However, the acquisition of HR color image is limited to digital imaging system, which may be not available in specific applications. And the storage and transmission of HR color images will also put huge burden on the hardware resource. To solve these problems, we explore the upscale of a single depth image guided by a LR color image, which offers unique challenges.

In this paper, we take advantage of the fact that there is complementary information between low-resolution depth map and its corresponding low-resolution color image, and propose a joint nonlocal sparse representation (JNSR) model using a joint dictionary and introducing the sparse coding

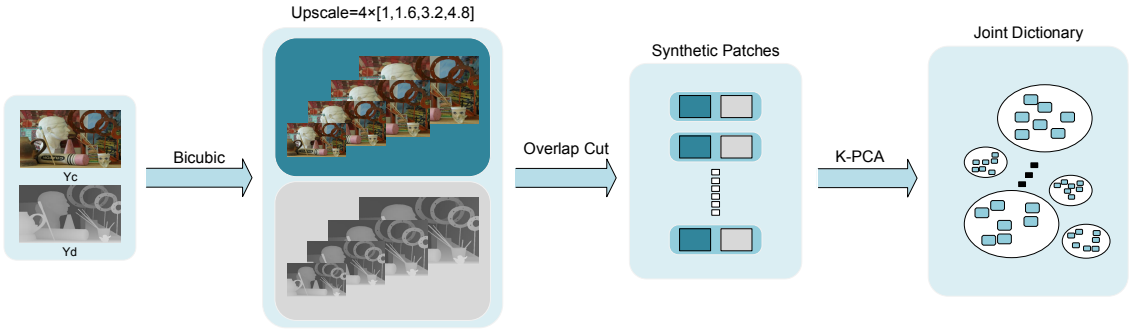


Fig. 1. The process of synthetic patches construction and joint dictionary learning

cost to reconstruct a fine quality depth image, which substantially enhance the spatial resolution of single low-quality depth map. Compared to state-of-the-art methods, the proposed method achieves higher reconstruction accuracy.

2. SUPER RESOLUTION BASED ON SPARSE REPRESENTATION THEORY

In the image degradation model, the observed low-resolution depth image $\boldsymbol{Y}_d$ is a degraded (blurred and down-sampled) version of original depth image $\boldsymbol{X}_d$. The degradation process can be expressed by:

$$\boldsymbol{Y}_d = \boldsymbol{H}\boldsymbol{D}\boldsymbol{X}_d + \boldsymbol{V} \tag{1}$$

Here, $\boldsymbol{H}$ represents a blurring filter, and $\boldsymbol{D}$ is the down-sampling operator. $\boldsymbol{X}_d$ is the unknown HR depth map to be estimated. The additional variable $\boldsymbol{V}$ denotes an random noise on the LR depth map.

In basic SR model based on sparse representation theory [15], $\boldsymbol{X}_d$ can be sparsely represented based on the HR dictionary $\boldsymbol{\Phi}_h$ via the convex l_1 - minimization:

$$\boldsymbol{\alpha}_x = \arg\min_{\boldsymbol{\alpha}} \{\|\boldsymbol{X}_d - \boldsymbol{\Phi}_h\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1\} \tag{2}$$

Where, $\boldsymbol{\Phi}_h$ is a high-resolution dictionary and constant λ denotes the regularization parameter. $\boldsymbol{\alpha}_x$ is sparse representation vector of $\boldsymbol{X}_d$. Then, we can acquire the sparse representation vector $\boldsymbol{\alpha}_y$ of the low-resolution image $\boldsymbol{Y}_d$ with respect to $\boldsymbol{\Phi}_h$:

$$\boldsymbol{\alpha}_y = \arg\min_{\boldsymbol{\alpha}} \{\|\boldsymbol{Y}_d - \boldsymbol{H}\boldsymbol{D}\boldsymbol{\Phi}_h\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1\} \tag{3}$$

The reconstructed high-resolution image, denoted by $\widehat{\boldsymbol{X}_d}$, is obtained as $\widehat{\boldsymbol{X}_d} = \boldsymbol{\Phi}_h\boldsymbol{\alpha}_y$.

3. PROPOSED JOINT NO-LOCAL SPARSE REPRESENTATION MODEL

3.1. The proposed model and joint dictionary learning

In proposed joint nonlocal sparse representation (JNSR) model, We construct synthetic characteristic image patch to learn a joint dictionary and introduce the sparse coding cost [14] to improve the accuracy of reconstructed depth map. Instead of learning an over-complete dictionary directly from the input LR depth map, we aim at learning a joint dictionary from intensity image and depth image. We replace the local sparse item with joint sparse coding cost in Eq.(3). Hence, the proposed JNSR model is defined as:

$$\boldsymbol{\alpha}_y = \arg\min_{\alpha} \{\|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{D}\boldsymbol{\Phi}_J \circ \boldsymbol{\alpha}\|_2^2 + \gamma \sum_i \|\boldsymbol{\alpha}_i - \boldsymbol{\beta}_i\|_1\} \tag{4}$$

$\boldsymbol{\Phi}_J = \{\boldsymbol{\Phi}_k | \boldsymbol{k} = \boldsymbol{1}\cdots\boldsymbol{K}\}$ is a joint dictionary contacted intensity information and depth information, where $\boldsymbol{\Phi}_k = [\boldsymbol{\Phi}_{kc}; \boldsymbol{\Phi}_{kd}]$. Here, $\boldsymbol{\Phi}_{kc}$ is intensity image sub-dictionary and $\boldsymbol{\Phi}_{kd}$ is the corresponding depth map sub-dictionary. For each given depth patch to be coded, one compact PCA sub-dictionary that is most relevant to the given patch is adaptively selected to code it [15]. $\gamma \sum_i ||\boldsymbol{\alpha}_i - \boldsymbol{\beta}_i||_1$ represents joint nonlocal regularization term valuing the sparse coding cost. Here, $\boldsymbol{\alpha}_i$ is sparse coding vector for each image patch $\mathbf{x}_i$, $\boldsymbol{\beta}_i$ is the estimation of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}$ is the concatenation of all $\boldsymbol{\alpha}_i$.

Clearly, one key procedure in the proposed JNSR model is the determination of the joint dictionary $\boldsymbol{\Phi}_J$. We attempt to recover up-scaling images based on its nonlocal self-similarity within and across scales [16]. So, we divide depth image and intensity image together with their up-scale versions into image patches to train the joint dictionary.

The joint dictionary learning process is shown in Fig.1. First, we do the synthetic image patches construction. We divide the $4\times$ bicubic LR depth map $\boldsymbol{Y}_d$ and its up-samples $\boldsymbol{Y}_d^{\boldsymbol{S}}$ at several scales ($\boldsymbol{S} = [\boldsymbol{1}, \boldsymbol{1.6}, \boldsymbol{3.2}, \boldsymbol{4.8}]$) into small patches of the same size and change these patches into single vectors

v_i^d. Here, i means i-th patch. In the same way, we get the patch vectors v_i^c of $4\times$ bicubic LR color image Y_c and its up-samples. We define $v_i = [v_i^c; v_i^d]$ as synthetic characteristic image patch vector. v_i represents the i-th column vector of matrix P.

Then, we learn several compact joint sub-dictionaries from P. To this end, we cluster P into K cluster. For each cluster, there are N numbers of depth patches and corresponding N numbers of intensity patches. We apply the principal component analysis (PCA) technique for each cluster of similar patches to learn a compact sub-dictionary. These PCA sub-dictionaries construct a large over-complete dictionary to characterize all the possible local structures of depth maps. The advantage of our proposed dictionary learning is that it can guarantee all the patches in each cluster use the same dictionary and share the same dictionary atoms.

3.2. The model solution and depth map reconstruction

How to obtain a good estimation of the unknown sparse coding vector is a key step in model solution. Let s_{ci} denotes the image patch of size $s \times s$ at the location i in color image Y_c, and s_{di} denotes the corresponding image patch in depth map Y_d. As mentioned above, for each local synthetic characteristic image patch $s_i = [s_{ci}; s_{di}]$, we exploit image patch redundancy within and across different scales to search for it similar patches.

For a given depth image patch s_{di}, we find several similar patches in the original spatial resolution level and its up-scales. Therefore, we are able to provide more high-frequency details for SR. The cross-scale searching process is also applied to the corresponding color image patch s_{ci}, as the Fig.2 shows.

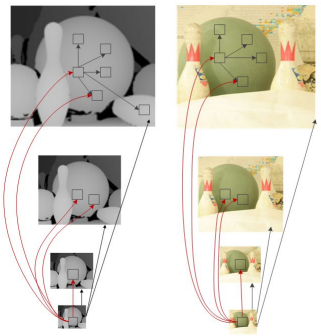


Fig. 2. The multi-scale similarity of synthetic patches

We computer Euclidean distance $dist1$ between s_{ci} and other intensity image patches respectively at the intensity image training window of size $p \times p (p > s)$ centered at position i, and computer Euclidean distance $dist2$ between s_{di} and other depth image patches respectively at the depth image training window, and then exploit the following formula:

Algorithm 1 Proposed depth map recostruction algorihm

Objective: Estimate HR depth map X_d^0;
Inputs: LR color image Y_c and LR depth map Y_d
1) Initialization: Estimate HR depth map X_d;
Set the initial estimation $\hat{X}_d^0$ and initial regularization parameter γ.
2) Outer loop: Learn dictionary and estimate parameter.
Iterate on $l = 1, 2, \cdots, L$:
a. Construct synthetic characteristic image patch $\{s_i\}$
b. Update the sub-dictionaries $\Phi_k = [\Phi_{kc}; \Phi_{kd}]$
c. Inner loop: reconstruct image.
i. Initialize $\beta_i^{(-1)} = 0$;
ii. Compute $X_d^{(0)} = \Phi_{kd} \circ \alpha_{yd}^{(0)}$;
iii. for each patch i compute $\beta_i^{(0)}$ using Eq.(6).
iv. Repeat the above three steps until convergence.
So in the l^{th} iteration, compute α_y^l using Eq.(4).
v.Image estimate update: $X_d^{(l)} = \Phi_{kd} \circ \alpha_{yd}^{(l)}$
3) Results: Output the final HR depth map X_d.

$$dist = \sqrt{dist1^2 + dist2^2} \tag{5}$$

to find K nearest similar non-local image patch pairs for synthetic image patch s_i. Then for each patch s_i, we have a set of its similar patches $s_{i,q}$, denoted by Ω_i.

Finally, we can compute β_i from the sparse codes of the patches within Ω_i. We denote the sparse codes of patch $s_{i,q}$ within set Ω_i as $\alpha_{i,q}$. Then β_i can be computed as the weighted average of $\alpha_{i,q}$ as below:

$$\beta_i = \sum_{q \in \Omega_i} \omega_{i,q} \alpha_{i,q} \tag{6}$$

Here, $\omega_{i,q}$ is the weighting vector. We set the weights similar to the nonlocal means approach [17] as:

$$\omega_{i,q} = \frac{1}{W} \exp(-\|\hat{s}_i - \hat{s}_{i,q}\|_2^2 / h) \tag{7}$$

Where, $\hat{s}_i = \Phi_J \hat{\alpha}_i$, $\hat{s}_{i,q} = \Phi_J \hat{\alpha}_{i,q}$ are the estimation of the patches s_i and $s_{i,q}$. h is a pre-determined scalar and w is the normalization factor. In the PCA sub-dictionaries, the sparse codes $\hat{\alpha}_i$ and $\hat{\alpha}_{i,q}$ can be easily computed as $\hat{\alpha}_i = \Phi_J^T \hat{x}_i$, $\hat{\alpha}_{i,q} = \Phi_J^T \hat{x}_{i,q}$.

We use bicubic interpolation of the LR depth image as initial estimation of the high-resolution depth map, then the HR X_d can be solved iteratively using Algorithm 1.

4. EXPERIMENTS RESULTS

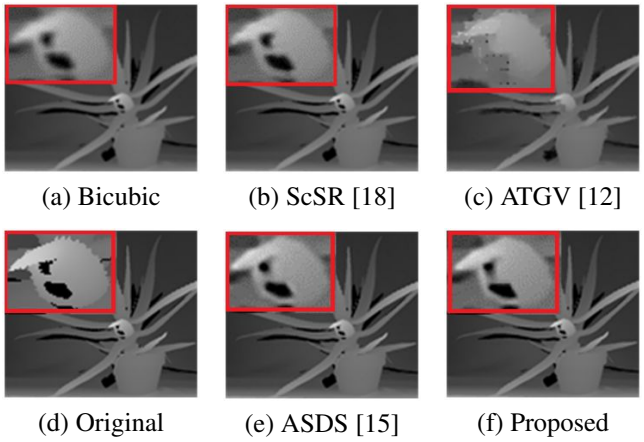
We evaluate our approach numerically on synthetic data using the well-known Middlebury stereo datasets [19], which provides aligned color images and depth maps for a number of different test scenes. In the experiments of SR, the degraded

Table 1. PSNR (dB) and SSIM results of the reconstructed HR depth map

Scale=4	Bicubic		ScSR [18]		ATGV [12]		ASDS [15]		Proposed	
art	31.84	0.9354	33.90	0.9556	32.62	0.9501	35.02	0.9600	**36.02**	**0.9709**
books	39.58	0.9787	40.41	0.9829	39.62	0.9770	42.14	0.9849	**43.03**	**0.9866**
dolls	41.14	0.9769	42.52	0.9829	42.20	0.9809	44.14	0.9849	**44.80**	**0.9874**
laundry	36.24	0.9646	37.51	0.9746	37.38	0.9704	39.29	0.9808	**40.24**	**0.9709**
moebius	40.16	0.9757	41.69	0.9822	41.13	0.9789	43.35	0.9839	**44.63**	**0.9880**
reindeer	34.89	0.9695	36.37	0.9772	36.51	0.9779	38.17	0.9804	**38.55**	**0.9827**
Average	37.31	0.9668	38.73	0.9759	38.24	0.9725	40.35	0.9783	**41.21**	**0.9827**

LR depth and color image were generated by applying a truncated 7×7 Gaussian kernel of standard deviation 1.6 to the high-resolution depth map and color image, and then downsampled by a scaling factor 4 in both horizontal and vertical. And the number of PCA sub-dictionary is $K = 64$. The regularization parameter is $\gamma = 7$.

We validate our algorithm through a large number of qualitative and quantitative comparisons against state-of-the-art SR algorithms: the sparse representation (ScSR) method [18], the adaptive sparse domain selection (ASDS-Reg) method [15], and the anisotropic total generalized variation (ATGV) method [12]. The bi-cubic interpolation method was used as basic standard. The objective comparison between the proposed JNSR and other methods are shown in Table1.



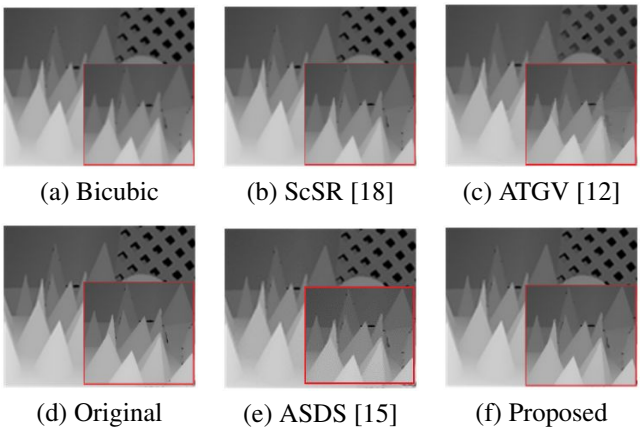
(a) Bicubic (b) ScSR [18] (c) ATGV [12]

(d) Original (e) ASDS [15] (f) Proposed

Fig. 3. $4\times$ recovery of *Aloe*

Table 1 shows parts of PSNR and SSIM results of the test methods on Middlebury stereo datasets. Our method achieves the highest PSNR and SSIM among the five algorithms over all the cases, which can improve roughly 3.90 dB in PSNR and 0.1590 in SSIM against the bi-cubic interpolation method.

Some visual results of the recovered images by different algorithms and the original depth maps are presented in Fig.3 and Fig 4. Obviously, ATGV generate the worst perceptual results. The recovered images by ScSR and ASDS possess much better visual quality than those of ATGV, but still suffer from some undesirable artifacts and distortion. Our method can reconstruct sharper edges and better structure details,showing better visual results than other methods.



(a) Bicubic (b) ScSR [18] (c) ATGV [12]

(d) Original (e) ASDS [15] (f) Proposed

Fig. 4. $4\times$ recovery of *Cones*

5. CONCLUSION

In this paper, we proposed a novel JNSR model for depth map SR guided by a LR color image. We construct synthetic characteristic image patch including intensity and depth information for SR. Experimental results show that our method is capable of recovering structured details that are missing in the LR depth image. Meanwhile, edge sharpness can be preserved with great detail due to the additional knowledge provided by nonlocal self-similarity explored in the corresponding LR color image.

6. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (No.61571326, 61471262, 61520106002) and National Natural Science Foundation of Tianjin (No.16JCQNJC00900).

7. REFERENCES

[1] A. N. Rajagopalan, Arnav Bhavsar, Frank Wallhoff, and Gerhard Rigoll, *Resolution Enhancement of PMD Range Maps*, Springer Berlin Heidelberg, 2008.

[2] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "Lidarboost: Depth superresolution for tof 3d shape scanning," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 343–350.

[3] Sigurjon Arni Gudmundsson, Henrik Aanaes, and Rasmus Larsen, "Fusion of stereo vision and time-of-flight imaging for improved 3d estimation," *International Journal of Intelligent Systems Technologies & Applications*, vol. 5, no. 3, pp. 425–433, 2008.

[4] Wei Hu, Gene Cheung, Xin Li, and Oscar Au, "Depth map super-resolution using synthesized view matching for depth-image-based rendering," vol. 49, no. 1, pp. 605–610, 2012.

[5] Kunpeng Zhu and Feng Lin, "Image super-resolution reconstruction by sparse decomposition and scale-invariant feature retrieval in micro-uav stereo vision," in *IEEE International Conference on Control & Automation*, 2014, pp. 705–710.

[6] H. Zheng, A. Bouzerdoum, and S. L. Phung, "Depth image super-resolution using multi-dictionary sparse representation," in *IEEE International Conference on Image Processing*, 2013, pp. 957–961.

[7] Fayao Liu, Chunhua Shen, and Guosheng Lin, "Deep convolutional neural fields for depth estimation from a single image," pp. 5162–5170, 2014.

[8] Kai Han Lo, Yu Chiang Frank Wang, and Kai Lung Hua, "Joint trilateral filtering for depth map super-resolution," in *Visual Communications and Image Processing*, 2013, pp. 1–6.

[9] James Diebel and Sebastian Thrun, "An application of markov random fields to range sensing.," *Advances in Neural Information Processing Systems*, pp. 291–298, 2005.

[10] Jiajun Lu and David Forsyth, "Sparse depth super resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2245–2253.

[11] Wei Zhuo, Mathieu Salzmann, Xuming He, and Miaomiao Liu, "Indoor scene structure analysis for single image depth estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 614–622.

[12] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Ruether, and Horst Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *IEEE International Conference on Computer Vision*, 2013, pp. 993–1000.

[13] M Kiechle, S Hawe, and M Kleinsteuber, "A joint intensity and depth co-sparse analysis model for depth map super-resolution," in *IEEE International Conference on Computer Vision*, 2013, pp. 1545–1552.

[14] Weisheng Dong, Lei Zhang, Guangming Shi, and Xin Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1620–1630, 2013.

[15] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization.," *Image Processing IEEE Transactions on*, vol. 20, no. 7, pp. 1838–1857, 2011.

[16] D Glasner, S Bagon, and M Irani, "Super-resolution from a single image," in *IEEE International Conference on Computer Vision*, 2009, pp. 349–356.

[17] Jianjun Yuan, "Improved anisotropic diffusion equation based on new non-local information scheme for image denoising," *IET Computer Vision*, vol. 9, no. 6, pp. 864–870, 2015.

[18] J. Yang, J Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation.," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 19, no. 11, pp. 2861, 2010.

[19] Daniel Scharstein and Chris Pal, "Learning conditional random fields for stereo," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.