CNNs in the Frequency Domain for Image Super-resolution

Yingnan Liu and Randy Clinton Paffenroth

Worcester Polytechnic Institute, 100 Institute Rd, Worcester, MA, USA 01609

ABSTRACT

This paper develops methods for recovering high-resolution images from low-resolution images by combining ideas inspired by sparse coding, such as compressive sensing techniques, with super-resolution neural networks. Sparse coding leverages the existence of bases in which signals can be sparsely represented, and herein we use such ideas to improve the performance of super-resolution convolutional neural networks (CNN). In particular, we propose an improved model in which CNNs are used for super-resolution in the frequency domain, and we demonstrate that such an approach improves the performance of image super-resolution neural networks. In addition, we indicate that instead of numerous deep layers, a shallower architecture in the frequency domain is sufficient for many types of image super-resolution problems.

Keywords: Image Reconstruction, Compressive Sensing, Convolutional Neural Network, Frequency Domain

1. INTRODUCTION

Sparse coding and compressive sensing are techniques that are widely used for signal reconstruction. They assumes that the relationship between the compressed signal y_s and the original signal y_t is an under-determined linear system $y_s = Ry_t$.[1] It can be shown theoretically that if there is a domain where y_t is sparse (i.e., has only a small number of nonzero coefficients), then y_t can be recovered from only a small number of linear measurements y_s.[1] Although many algorithms have been developed in the compressed sensing domain, they are generally restricted to linear transformations where prior knowledge of the sensing matrix R is required. A stacked denoising autoencoder (SDA) was developed as an initial attempt to address this problem using neural network-based algorithms.[2] Nonetheless, we propose that the quality of the reconstructed high-resolution image can be improved by novel a combination ideas from compressive sensing and neural networks.

In particular, a popular set of techniques for image-related problems are convolutional neural networks (CNNs). A recent breakthrough by Dong et al[3] provides a single image super-resolution convolutational neural network (SRCNN), which is an end-to-end approach with pixel-wise output, making CNNs practical for image reconstruction problems. Based on previous studies, a scalable single image SRCNN called ReconNet[4] was developed combining the ideas of under-determined linear systems and SRCNN. In this algorithm, a linear layer is embedded in the framework of a image super-resolution CNN and such convolutional architectures result in better performance than previous SDAs.

1.1 Contribution

In this paper, we propose several neural network architectures, both shallow and deep, in *the frequency domain* for image super-resolution. In particular, we demonstrate how the sparse coding of images in the frequency domain improves the efficiency of scalable single image super-resolution CNNs. Based upon these ideas, we propose deep and shallow neural networks in which both the input data to the CNN and the lose-function are defined in the frequency domain, and we call these algorithms *SparseFnets*. The advantage of our model is demonstrated by comparing SparseFnet with other models and our work is novel in three ways:

- First, our method stands on the assumption that the representations of images in the frequency domain can be more effectively processed by CNNs. In particular, our model is trained purely in the frequency domain and is never exposed to the image domain.
- Second, based upon the advantageous properties of working in the frequency domain, our model can be effective even with a shallower architecture. Accordingly, the parameters in the simplified network can be trained more efficiently.
- Third, our resulting high-resolution images are more faithful to the original imagery, with substantially lower error.

2. RELATED WORK

2.1 Compressive Sensing

Compressive sensing is a signal process technique for acquiring and reconstructing signals. It provides an efficient alternative to traditional methods which require the acquisition of data with a high sampling rate. Classically the Nyquist-Shannon sampling theorem states that a signal $y_t \in \mathbb{R}^N$ can be reconstructed by $M (M \ll N)$ linear measurements if the sampling rate is more than twice the highest frequency. However, if y_t is sparse in a certain basis Ψ, in which the restricted isometry property or similar assumptions hold,[5] then y_t can be reconstructed from a compressed signal $y_s \in \mathbb{R}^M$ and a sensing matrix $R \in \mathbb{R}^{M \times N}$.[5] Specifically, such methods assume that y_t and y_s can be represented as

$$y_t = \Psi\omega, \; R\Psi\omega = y_s, \tag{1}$$

where ω is the sparse representation of the original signal in a basis Ψ, and choosing Ψ as the Fourier basis leads to many natural signals being sparse in that domain. The choice of R is crucial and it has been shown that Bernoulli and Gaussian random matrices are good candidates.[6] Based on these assumptions, standard approaches use l_1 regularization and the desired sparse solution can be obtained by solving the convex optimization problem

$$\min_{\omega \in \mathbb{R}^M} \|\omega\|_1 \;\; \text{s.t.} \;\; R\Psi\omega = y_s. \tag{2}$$

With the optimal solution ω^*, the original signal can be reconstructed as $y_t = \Psi\omega^*$. Classical algorithms for compressive sensing solve the problem by iteratively reweighting the l_1 minimization.[7] Reflecting on the success of compressive sensing methods, we are inspired to extend current CNN based super-resolution techniques by leveraging appropriate sparse representations.

2.2 Single Image Super-Resolution CNN

CNNs have become a very popular technique for image processing problems. Based on previous work, Chao et al[3] proposed a SRCNN, providing an end-to-end mapping from low-resolution patch inputs to high-resolution pixel output. This SRCNN uses the neural network as an optimized pipeline consisting of three operations; patch extraction and representation; non-linear mapping in high dimensional space; and reconstruction of high dimensional representations. Following the breakthroughs based on SRCNNs, various authors have worked on adapting compressive sensing techniques to CNNs. Inspired by compressive sensing techniques, a scalable single image super-resolution CNN was introduced by Kulkarni, et al.[4] Here, a deep neural network is used to solve the under-determined linear system in place of the traditional l_1-minimization method. They state that the image signal is not exactly sparse with respect to the domain Ψ, and a convolutional architecture which refines the image estimation in each iteration is used to estimate the sensing measurement R. In their work, a method called ReconNet is developed. A significant benefit of ReconNet is better performance with few parameters, somewhat reducing the computational cost. A linear layer is used to up-scale the image patches and then connected with convolutional layers for image reconstruction.

3. PROPOSED METHOD

In contrast to the ideas in Kulkarni, et al.,[4] we propose a different explanation for scalable single image super-resolution CNNs. The reason that the convolutional architecture works well without the l_1-norm regularization is that their method is learning the sparsity transform and the sensing transformation simultaneously. However, we propose that the sparsity transform does not need to be learned since the high-resolution image can be recovered by sparse representations accurately in the frequency domain.[8] Accordingly, with the help from Fourier transformation, CNNs can yield better reconstruction quality for high-resolution images *even when it is trained purely in the frequency domain and is never exposed to the image domain.* Furthermore, we challenge the need for the deep architectures used in previous studies since numerous convolutional layers suffer from exploding/vanishing gradients.[9] Unlike previous studies on deep residual-learning[10] and gradient clipping,[11] we propose to reduce the computational cost and increase the efficiency by a shallower architecture in the frequency domain, while still maintaining state-of-the-art image super resolution performance.

Our model outperforms traditional compressive sensing methods by learning a non-linear transformation instead of defining a linear sensing matrix R. It also improves upon previous scalable single image super-resolution CNNs by using measurements in the sparse frequency domain. In particular, the representation in the frequency domain can be easily found by $x = Fy_s$, $y = Fy_t$ where F denotes the Fourier transformation. Finding the measurement R_f between Fourier representations and training the model for $y = R_f^{-1}x$ can help to improve the performance.

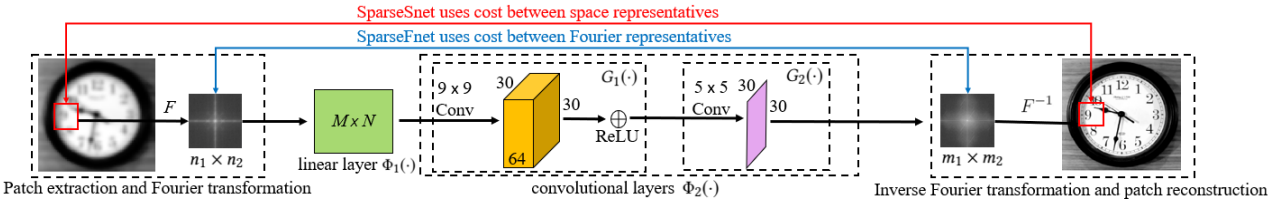

Figure 1. The shallow architecture of our proposed models. The first linear layer maps input patches to high dimensional feature space. The second convolutional layer extracts features non-linearly. The third convolutional layer acts as an averaging filter of features and reconstructs high dimensional patches.

3.1 Architecture of Proposed Models

In the following, we describe the basic structures and introduce three proposed models, namely a shallow neural network (Shallownet), a sparse domain neural network in the frequency domain with spatial cost (SparseSnet), and a sparse domain neural network in the frequency domain with frequency cost (SparseFnet).

The framework of our models is shown in Figure 1. Before training the model, we pre-process the data by extracting overlapped patches from the entire image and applying Fourier transformation on the patches. The model is then trained with Fourier representatives of the small image patches. Our proposed architecture contains a linear layer $\Phi_1(\cdot)$ and convolutional layers $\Phi_2(\cdot)$. The linear layer takes small patches from the low dimensional space as inputs and feeds preliminary reconstructed patches on the high dimensional space to the convolutional layers. Then the convolutional layers apply nonlinear filters to further improve the reconstruction accuracy. Specifically, a compressed signal $x \in \mathbb{R}^{n_1 \times n_2}$ $(N = n_1 \times n_2)$ is projected to the high dimensional space by the linear mapping $\Phi_1(\cdot)$. A preliminary reconstructed signal $y_1 \in \mathbb{R}^{m_1 \times m_2}$ $(M = m_1 \times m_2)$ is obtained by $y_1 = \Phi_1(x)$, $\Phi_1(x) = W_1 \cdot x + b_1$ where W_1 is a $M \times N$ dimensional linear filter and b_1 is a M dimensional bias vector. The convolutional part of our architecture, $\Phi_2(\cdot)$, has two convolution layers, $G_1(\cdot)$ and $G_2(\cdot)$, which are both in the projected high dimensional space. These two layers are used for feature extraction and patch reconstruction respectively. The first layer $G_1(\cdot)$ extracts 64 features from the preliminary reconstructed signal y_1 by non-linear filters, and ReLU is chosen to be the activation function. The layer is expressed as $y_2 = G_1(y_1)$, $G_1(y_1) = \max(0, W_2 \cdot y_1 + b_2)$ where W_2 contains 64 filters of size 9×9 and b_2 is a $64 \times 9 \times 9$ dimensional bias vector. Then $G_2(\cdot)$ acts as an averaging filter for patch reconstruction. The output of the second non-linear layer is the final reconstructed patch $y \in \mathbb{R}^{m_1 \times m_2}$. Each pixel in y can be viewed as a weighted average of the 64 features from the previous layer $G_1(\cdot)$. We define this layer by $y = G_2(y_2)$, $G_2(y_2) = W_3 \cdot y_2 + b_3$ where W_3 contains a filter of size 5×5, and b_3 is the corresponding bias vector of size $m_1 \times m_2$. The settings for these parameters in our experiments are $n_1 = n_2 = 10$, $m_1 = m_2 = 30$.

Shallownet is the baseline of our proposed models. It can be viewed as a simplified version of ReconNet. We skip the operations of Fourier transformation after patch extraction and before patch reconstruction in Figure 1. The input signal x and the output y are both image patches in the space domain. Suppose the operation of the network can be written as $\Phi_\theta = \Phi_2(\Phi_1(\cdot)) \in \mathbb{R}^{M \times N}$ where θ denotes the parameters in the network. The high-resolution patch can be recovered by

$$y = y_t, x = y_s, y = \Phi_\theta(x), \min_\theta \|A - \Phi_\theta(x)\|_2^2 \tag{3}$$

where y_s and y_t are the compressed patches and the reconstructed patches in the space domain respectively and A is the ground truth for the high-resolution patches. In the case where the neural network is equivalent to regular compressive sensing approaches, Φ_θ for Shallownet can be expressed as a sparse sensing matrix. In other words, the network learns the projection for the sparsity domain and the sensing measurement simultaneously. However, if the sparsity projection is given, then the algorithm can be more efficient with the same architecture. Thus, we propose SparseSnet and SparseFnet in the frequency domain.

As described in Section 2.1, the Fourier transformation is a good candidate for the sparsity domain. For both SparseSnet and SparseFnet, the input and the output patches are pre-processed by the Discrete Cosine Transformation (DCT) denoted by F. SparseSnet performs the optimization in the space domain, while SparseFnet optimizes in the frequency domain directly. The optimization of SparseSnet is

$$y = Fy_t, x = Fy_s, y = \Phi_\theta(x), \min_\theta \|A - F^{-1}\Phi_\theta(x)\|_2^2. \tag{4}$$

The network uses the inverse DCT on high dimensional Fourier representative patches to compute the cost. Although SparseSnet is in the frequency domain, it uses gradient descent with repect to the cost in the space domain.

SparseFnet is more straight forward in the frequency domain. The network learns Φ_θ from

$$y = Fy_t, x = Fy_s, \Omega = FA, y = \Phi_\theta(x), \min_\theta \|\Omega - \Phi_\theta(x)\|_2^2, \tag{5}$$

where Ω is the Fourier representation of the ground truth A. *Here the cost in each iteration is computed between Fourier representations directly.* An estimate of Ω is output in the frequency domain after which the inverse DCT is applied to recover the representation in the space domain. Unlike SparseSnet, SparseFnet never sees the space domain. This network is built on the assumption that the DCT projection provides an alternative basis on which the signal is approximately sparse. The projection for the sparsity domain is found during pre-training process, and the operation Φ_θ learns only the sensing measurement in the frequency domain. If the assumption of sparsity stands, then SparseFnet should perform better than Shallownet and SparseSnet.

4. EXPERIMENTAL RESULTS

Several experiments were conducted to demonstrate the effectiveness of our proposed techniques. Model performance is compared with respect to the convergence progress and reconstruction quality. By comparing the convergence progress, we demonstrate that utilizing the frequency domain and the shallow architecture help to provide a more stable model. We then show that our shallow models in the frequency domain yield exceptional reconstruction quality both visually and numerically.

4.1 Training Details

Our goal is to obtain a neural network which represents the unknown compressed measurement. The labels for training the neural network are the ground truth of high-resolution images. The objective is $min\frac{1}{2N}\sum_{i=1}^{N}\|f(x_i)-x_i\|_2^2$ where f represents the operation of the neural network and x_i represents the input. We use mean square error (MSE) for the peak signal-to-noise ratio (PSNR) statistic. PSNR in decibels (dB) is defined as $PSNR=10\log_{10}(\frac{I^2}{MSE})$, where I is the maximum pixel value of the data type.[12] Here, we use 8-bit images with $I=255$. Higher PSNR reflects better reconstruction quality.

The training data are generated from 300×300 images. We use the Python package *skimage* to down-scale our training images at measurement rates (MRs) of 0.25, 0.1 and 0.06 respectively. For example, using reduction factor of 0.5 both horizontally and vertically, a 150×150 image is generated for MR$=0.25$. We then extract overlapped patches for the network. Patches of size 10×10 are extracted from compressed images. The neural networks are trained with Adam optimizer. Learning rate is set to be 10^{-3}. The minibatch size is set to be 128. We train the neural networks for 100 epochs.

4.2 Baseline

Two algorithms from the literature, ReconNet[4] and O-NL-SDA,[2] are implemented for comparison with our models and serve as baselines. We implement these algorithms on the same platform as our algorithms to control the performing environment. The model architectures strictly follow those of the respective authors with activation functions and parameters chosen as they suggest. Specifically, it is indicated in the previous work on ReconNet that it outperforms former methods both with and without a BM3D denoiser. Accordingly, we make our comparisons in the absence of BM3D.

In addition to baseline models and our proposed models, we adapt the deep architecture from ReconNet to our frequency models to make comparison between architecture depths. We name the deeper versions D-SparseSnet and D-SparseFnet. The framework is shown in Figure 2.

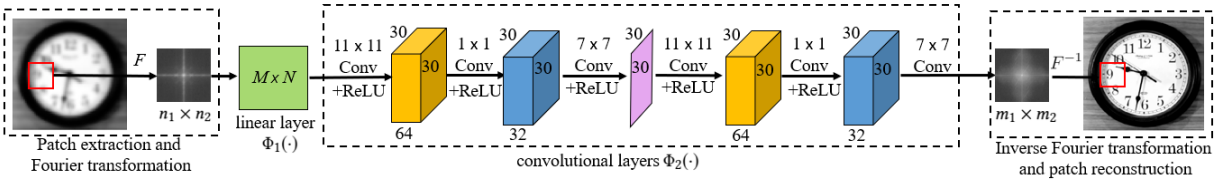

Figure 2. The framework for D-SparseSnet and D-SparseFnet. The deeper architecture is inherited from ReconNet.

4.3 Convergence Properties

The progress of convergence illustrates that sparse representations provided by the frequency domain result in more efficient outcomes with shallower architecture. Figure 3 shows the decay of MSE over training epochs for ReconNet, Shallownet, SparseSnet and SparseFnet at different measurement rates. At all measurement rates, the convergence progress of ReconNet shows the largest fluctuation. As the measurement rate decreases from 0.25 to 0.06, the instability of ReconNet becomes more severe. On the contrary, our models, which are in the frequency domain, have smoother MSE decay curves. In particular, SparseFnet has exceptional performance with rapid and stable convergence at all measurement rates.

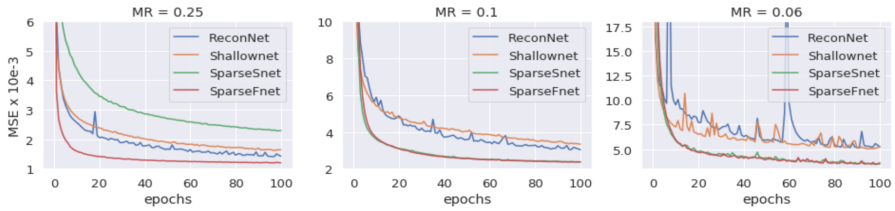

Figure 3. Decay of training MSE($\times 10^{-3}$) for different algorithms at different measurement rates. The baseline model ReconNet shows the largest fluctuation, while SparseFnet has rapid and stable convergence progress in the frequency domain.

4.4 Reconstruction Quality

Six testing images are reconstructed at MR=0.25, 0.1 and 0.06. Figure 4 shows some outcomes from different algorithms at MR=0.1. The differences in reconstruction quality can be recognized visually. SparseSnet and SparseFnet give sharper details while models in the space domain suffer from blurring effects. Table 1 presents the testing PSNR (in dB) from the seven algorithms at three measurement rates for all the six images. The highest PSNR are achieved by models in the frequency domain for all the testing images, and by shallower architecture in 17 out of the 18 experimental trials.

Both in the frequency domain, though the performances of SparseSnet and SparseFnet are similar with slight differences in general, SparseFnet has better result than SparseSnet as it uses Fourier representatives in the loss function directly. The highest PSNR is achieved by SparseFnet in 11 trails, and by SparseSnet in 6 trials. In addition, SparseSnet needs inverse Fourier transformation in the training process to use space representatives in the loss function. SparseFnet reduces the computational cost significantly by 25% of the running time when we fed 10^5 image patches in the training process.

The superiority of our shallow architecture can be found by comparing proposed models with their deeper versions. Shallownet yields better result than ReconNet in 14 trials. Both SparseSnet and SparseFnet outperform D-SparseSnet and D-SparseFnet respectively in 17 out of 18 trails. We can conclude that the frequency domain and the shallower architecture result in better performance when used together.

Figure 4. Reconstructed testing images with different algorithms at measurement rate of 0.1.

5. CONCLUSION

In this paper, an improved scalable single image super-resolution neural network in the frequency domain is proposed. Inspired by the sparsity assumption in the frequency domain from standard compressive sensing methods, a new architecture is built. A primary feature of our model is that the neural network is completely hidden from the original space domain. Moreover, we simplify the architecture of deep neural networks for image reconstructing problems. The shallower architecture is demonstrated to be just as accurate with lower computational cost. Our experiment results illustrate that the proposed SparseFnet significantly outperforms previous scalable single image super-resolution neural networks.

Table 1. Testing PSNR (in dB) with different algorithms at different measurement rates. Higher PSNR reflects better reconstruction quality. Models in the frequency domain yield better reconstruction quality in all the cases comparing with models in the space domain. Shallow architecture outperforms deep architecture.

	Flag			Tree			Bell Tower		
Algorithms	MR 0.25	MR 0.10	MR 0.06	MR 0.25	MR 0.10	MR 0.06	MR 0.25	MR 0.10	MR 0.06
O-NL-SDA	31.253	29.667	28.553	33.763	32.725	31.911	35.768	34.321	32.850
ReconNet	33.315	31.040	29.764	35.783	34.081	33.247	37.762	35.860	34.836
Shallownet	33.200	31.709	30.290	35.850	34.007	33.264	37.665	36.173	34.984
D-SparseSnet	34.882	32.405	31.003	37.317	35.239	33.950	38.836	36.983	35.651
D-SparseFnet	34.925	32.423	31.221	37.297	35.188	33.910	38.788	36.939	35.797
SparseSnet	34.980	32.487	31.276	37.534	**35.294**	34.071	38.788	**37.040**	**35.899**
SparseFnet	**35.126**	**32.490**	**31.278**	**37.555**	35.277	**34.107**	**38.875**	37.014	35.884
	Rose			Flower			Building		
O-NL-SDA	38.182	35.504	33.416	38.769	35.999	34.832	33.245	30.424	29.665
ReconNet	40.587	38.360	36.914	40.459	38.257	35.683	36.152	33.718	31.008
Shallownet	40.617	38.712	37.183	40.690	41.110	37.111	36.137	34.183	31.222
D-SparseSnet	41.837	39.906	38.241	40.445	39.893	38.421	37.421	35.367	33.008
D-SparseFnet	41.648	39.736	38.457	40.981	39.688	**39.784**	37.555	35.491	33.093
SparseSnet	**41.876**	**39.930**	38.555	41.013	39.980	39.415	37.900	**35.600**	33.317
SparseFnet	41.760	39.881	**38.638**	**41.414**	**39.993**	39.415	**37.963**	35.586	**33.326**

ACKNOWLEDGMENTS

This research is sponsored by Nanocomp Technologies Inc. and we would like to thank NASA contract number: 80LARC18C0007 for their support. We would also like to give special thanks to Bob Casoni (Quality Assurance Manager), Rachel Stephenson, PhD (Research Scientist) and Nick Dixon(Quality Engineer), who provided great help with this research.

REFERENCES

[1] Candès, E. J., Romberg, J., and Tao, T., "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory.* **52**(2), 489–509 (2006).

[2] Mousavi, A., Patel, A. B., and Baraniuk, R. G., "A deep learning approach to structured signal recovery," *Proc. Annual Allerton Conference on Communication, Control, and Computing*, 1336–1343 (2015).

[3] Dong, C., Loy, C. C., He, K., and Tang, X., "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(2), 295–307 (2016).

[4] Kulkarni, K., Lohit, S., Turaga, P., Kerviche, R., and Ashok, A., "Reconnet: Non-iterative reconstruction of images from compressively sensed measurements," *Proc. Conference on Computer Vision and Pattern Recognition*, 449–458 (2016).

[5] Donoho, D. L., "Compressed sensing," *IEEE Trans. Inf. Theory.* **52**(4), 1289–1306 (2006).

[6] Baraniuk, R., Davenport, M., DeVore, R., and Wakin, M., "A simple proof of the restricted isometry property for random matrices," *Constr. Approx.* **28**(3), 253–263 (2008).

[7] Mallat, S. G. and Zhang, Z., "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.* **41**(12), 3397–3415 (1993).

[8] Candès, E. J. et al., "Compressive sampling," *Proc. International Congress of Mathematicians* **3**, 1433–1452 (2006).

[9] Bengio, Y., Simard, P., and Frasconi, P., "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw. Learn. Syst.* **5**(2), 157–166 (1994).

[10] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

[11] Kim, J., Kwon Lee, J., and Mu Lee, K., "Deeply-recursive convolutional network for image super-resolution," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1637–1645 (2016).

[12] Huynh-Thu, Q. and Ghanbari, M., "The accuracy of psnr in predicting video quality for different video scenes and frame rates," *Telecomm Syst.* **49**(1), 35–48 (2012).