A Data Minimization Algorithm for Fast and Secure Transfer of Big Datasets Using Fourier Analysis

Shervin Sharafatmandjoor, Cristian Balan,

State University of New York, Plattsburgh, NY

Abstract

Today's increased demand for fast and secure transfer of big data applications and files, has motivated researchers to develop novel algorithms within the framework of existing protocol limitations. Transferring big data datasets (e.g. Big Genomic Datasets) over the network has always been an interesting and challenging research area. This is because all data transfer protocols, such as HTTP and FTP, use the standard content-encoding schemes, namely ASCII, which offer little or no compression or data minimization. The purpose of this work is to design and implement a novel Fourier-based data minimization algorithm to decrease the required time to transfer multi-dimensional big datasets over conventional networks. In addition, we increased security in the event of a data breach. Moreover, it will introduce a generic concept that can be used by cloud-based applications to secure data that is being exchanged remotely. One and two-dimensional result show how the method could effectively be used in different data minimization scenarios.

Introduction

One of the benefits of image processing concepts and techniques has been reducing the size of data, specifically images. Image processing techniques usually include the manipulation of the original image in physical or Fourier space. Fourier based image processing techniques are generally categorized as high-pass, low-pass and band width filters [Arashpour, Alderhari]. Also, more specific sharpening filters have been developed to mimic the sharp edges of the images [Alameen, Zhang]. Each of the mentioned techniques embody advantages and disadvantages among which is losing some important information in the picture [Arashpour]. Therefore, researchers have tried to develop more sophisticated methods to conserve important features of the image while reducing the size of the field [Wang, Rasheed]. This study aims to provide an optimized Fourier-based method in the aforementioned class.

As the core part of the methodology, a novel two-step process is introduced to make sure that the image data is optimally reduced in size while being transferred in today's modern network/Cloud operations. In fact, an image that is digitized per its gray scale intensity values is first converted to the Fourier space. Then the reduced order Fourier coefficients are obtained by solving a linear set of equations using the least square concept. The reduced size output image in the Fourier space could be safely transferred over conventional network protocols and then reconstructed at the client side using the inverse discrete Fourier transform.

Theory

The two-dimensional discrete Fourier transform (DFT) and the inverse transform of a function are given below:

$$F[k,l] = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f[m,n]e^{-J2\pi(\frac{k}{M}m+\frac{l}{N}n)} \quad (1)$$

and

$$f[u,v] = \sum_{k=0}^{M-1}\sum_{l=0}^{N-1} F[k,l]e^{J2\pi(\frac{k}{M}m+\frac{l}{N}n)} \quad (2)$$

where $f[m,n]$ is a periodic $M \times N$ image function. If the original function is not naturally periodic, we can the zero padding technique in the physical domain to fulfill the requirements of the Fourier analysis.

The idea is to take the DFT of the original input image $f[M,M]$ and then reconstruct a reduced Fourier field $F'[k,l]$ with reduced mode numbers $M' = \alpha M$ and $N' = \beta N$ where $\alpha, \beta < 1$ using equation 2. Therefore,

$$f[u,v] = \sum_{k=0}^{M-1}\sum_{l=0}^{N-1} F[k,l]e^{J2\pi(\frac{k}{M}m+\frac{l}{N}n)} = \sum_{k=0}^{M'-1}\sum_{l=0}^{N'-1} F'[k,l]e^{J2\pi(\frac{k}{M}m+\frac{l}{N}n)} \quad (3)$$

In the above linear set of equations, the $f[u,v]$ that is the value of each cell pixel and $F[k,l]$ are known and the unknown coefficients are $F'[k,l]$. The final reduced order output $M' \times M$ image is the Inverse transform based on $F'[k,l]$. Since $M' < M$ and $N' < N$, the system would have more known variables than unknown and a least square algorithm could be employed to find the optimal approximation as in system 4.

$$
\begin{pmatrix}
\cos 1x_0 & \cos 2x_0 & \cos 3x_0 & \ldots & \cos mx_0 & \sin 1x_0 & \sin 2x_0 & \sin 3x_0 & \ldots & \sin mx_0 \\
\cos 1x_1 & \cos 2x_1 & \cos 3x_1 & \ldots & \cos mx_1 & \sin 1x_1 & \sin 2x_1 & \sin 3x_1 & \ldots & \sin mx_1 \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\
\cos 1x_{2m-1} & \cos 2x_{2m-1} & \cos 3x_{2m-1} & \cdots & \cos mx_{2m-1} & \sin 1x_{2m-1} & \sin 2x_{2m-1} & \sin 3x_{2m-1} & \cdots & \sin mx_{2m-1} \\
\cos 1x_{2m} & \cos 2x_{2m} & \cos 3x_{2m} & \cdots & \cos mx_{2m} & \sin 1x_{2m} & \sin 2x_{2m} & \sin 3x_{2m} & \cdots & \sin mx_{2m}
\end{pmatrix}
\times
\begin{pmatrix} a_1' \\ a_2' \\ a_3' \\ \vdots \\ a_m' \\ b_1' \\ b_2' \\ b_3' \\ \vdots \\ b_m' \end{pmatrix}
=
\begin{pmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ f(x_{2m-1}) \\ f(x_{2m}) \end{pmatrix}
\tag{4}
$$

Hence, despite low-pass, high-pass or band width filters, no specific mode value is discarded and the whole system contributes to the values of the new reduced number of modes. Obviously, in the case of $\alpha = \beta = 1$, the original input image is obtained which is trivial. As α and β get smaller, the new number of Fourier modes are smaller and then the size will be reduced accordingly by the factor $M'N'/MN$.

Results

In this section, the results of applying the data minimization algorithm are shown and discussed. As a first step, we employed the methodology to a one-dimensional case where the absolute values of the discrete Fourier modes are truncated after a cut-off point. The new reduced-sized set only includes the modified values up to the cut-off point. As shown in figure 1, the modified values are responsible to mimic the whole pattern but with less mode values. The optimum value for the cut-off point could be obtained by a compromise between the accuracy of the reconstructed field and size of the reduced-sized set.

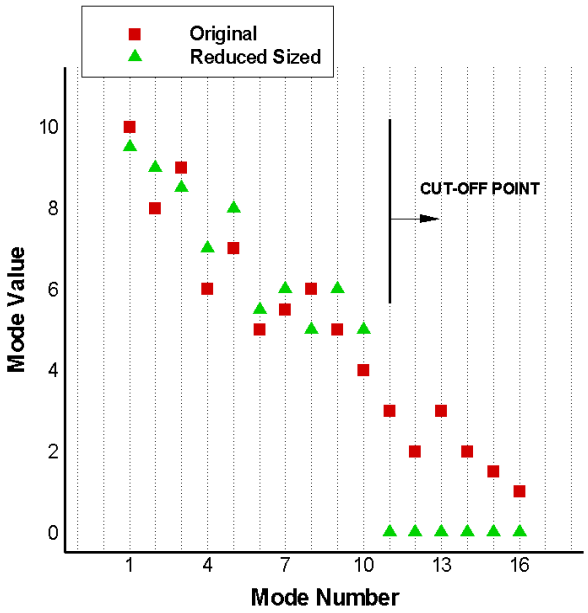

Figure 1. Mode values of the original and reduced size cases for a 32 data point one dimensional field.

Next, we can analyze a two-dimensional picture where the grayscale contour values are ranged between 0 and 255 in figure 2. The Fourier modes of the reduced-sized field are depicted in figure 3. Here the absolute values of the Fourier modes are favorably clustered around the origin starting from the first mode that is a representation of the mean value of the field and the far field values are chopped in a similar way as the one-dimensional case. It should be noted that the inevitable Gibbs phenomenon caused by the sharp edge variation on the contour is a cause of some inaccuracy.

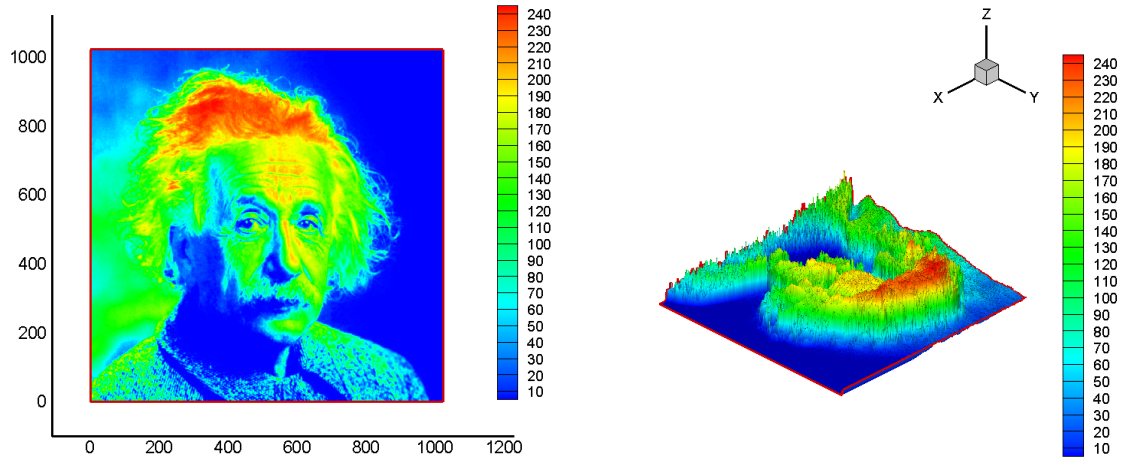

Figure 2. A sample two-dimensional grayscale contour with 1024 data points.

The computational costs of the method include the increased space and time complexities at the origin-transmitter and also at the destination-receiver to reconstruct the data from the

reduced-sized field. For the $n \times n$ two-dimensional case the time complexity is limited to $n \times log\, n$ with an optimized linear algebra solver. Considering the expensive transfer costs, a reduced-size domain will have a negligible extra complexity specially for larger data sets. Furthermore, the inherently ciphertext nature of the data in the transfer mode is a remarkable bi-product of the method.

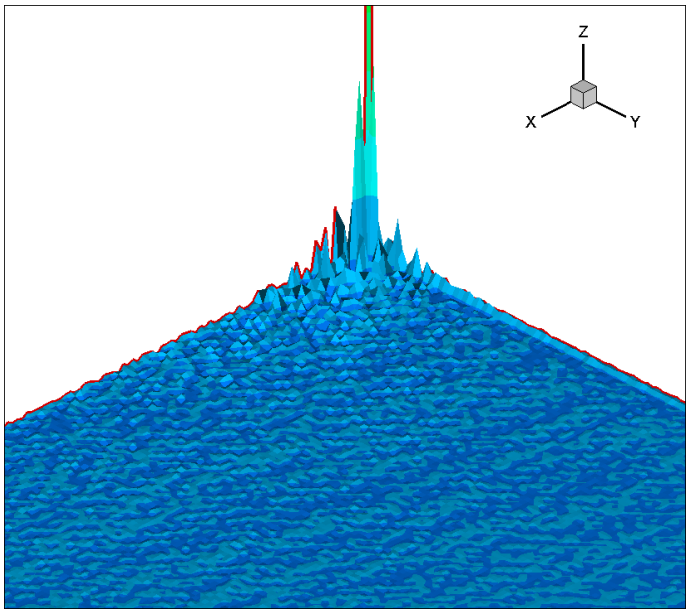

Figure 3. The reduced order Fourier mode representation of the two-dimensional grayscale case of figure 2.

Conclusions

In this work, we presented an algorithm to modify the lower modes and chop the higher modes in the Fourier space in way that the existing modes could optimally represent the whole field in the physical space.The semi-low-pass filter seems to be promising in one and two dimensional fields. However, uncertainties due to the inherent characteristics of the spectral analysis will affect the accuracy and precision of the method. The preliminary results show that with the cost of increasing the time and space complexities at the transmitter and receiver levels, the algorithm can remarkably reduce the size of the data in the transfer phase. Potential extensions and optimizations to the methodology may include introducing some smart brackets of Fourier modes instead of the pure low-pass concept and examining the idea on the other data types.

Bibliography

M. Arashpour, T, Ngo, H. Li, Scene understanding in construction and buildings using image processing methods: A comprehensive review and a case study, Journal of Building Engineering, Volume 33, January 2021, 101672.

M. Aledhari, M. Di Pierro and F. Saeed, "A Fourier-Based Data Minimization Algorithm for Fast and Secure Transfer of Big Genomic Datasets," 2018 IEEE International Congress on Big Data (BigData Congress), 2018, pp. 128-134, doi: 10.1109/BigDataCongress.2018.00024.

Z. Al-Ameen, A. Muttar and G. Al-Badrani, Improving the Sharpness of Digital Image Using an Amended Unsharp Mask FilterI.J. Image, Graphics and Signal Processing, 2019, 3, 1-9.

K. Zhang and J. U. Kang, "Graphics processing unit accelerated non-uniform fast Fourier transform for ultrahigh-speed, real-time Fourier-domain OCT," Opt. Express 18, 23472-23487 (2010).

A, Wang, W. Zhang, X, Wei, A review on weed detection using ground-based machine vision and image processing techniques, Computers and Electronics in Agriculture, Volume 158, March 2019, Pages 226-240.

M. H. Rasheed,O. M. Salih, M. M. Siddeq, M. A. Rodrigues, Image compression based on 2D Discrete Fourier Transform and matrix minimization algorithm, Array 6 (2020) 100024.