A Proper Transform for Satisfying Benford's Law and Its Application to Double JPEG Image Forensics

Ali Taimori[1], Farbod Razzazi[2], Alireza Behrad[3], Ali Ahmadi[4] and Massoud Babaie-Zadeh[5]
[1,2]Dept. of Elec. and Comp. Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[3]Faculty of Engineering, Shahed University, Tehran, Iran
[4]Dept. of Elec. and Comp. Engineering, K. N. Toosi University of Technology, Tehran, Iran
[5]Dept. of Elec. Engineering, Sharif University of Technology, Tehran, Iran
[1]a.taimori@srbiau.ac.ir, [2]razzazi@sr.iau.ac.ir, [3]behrad@shahed.ac.ir, [4]ahmadi@eetd.kntu.ac.ir, [5]mbzadeh@sharif.edu

Abstract— *This paper presents a new transform domain to evaluate the goodness of fit of natural image data to the common Benford's Law. The evaluation is made by three statistical fitness criteria including Pearson's chi-square test statistic, normalized cross correlation and a distance measure based on symmetrized Kullback-Leibler divergence. It is shown that the serial combination of variance filtering and block 2-D discrete cosine transform reveals the best goodness of fit for the first significant digit. We also show that the proposed transform domain brings reasonable fit for the second, third and fourth significant digits. As an application, the proposed transform domain is utilized to detect image manipulation by distinguishing single compressed images from doubly compressed ones.*

Keywords— *Benford's Law, significant digits statistics, image forensics, double JPEG, discrete cosine transform, variance filter.*

I. Introduction

Benford's Law (BL), also called the first digit phenomenon, was originally discovered by S. Newcomb in 1881. After that, F. A. Benford credited the law on a wide range of datasets in 1938 [1]. Briefly, this phenomenon states that the distribution of the first non-zero significant decimal digit from observations of a discrete random variable in many cases obeys a logarithmic distribution rather than uniform probability mass function (pmf) [2], as might be expected based on the basic probability theory. At first, BL was tested and validated on real-world data like rivers' length, mathematical and physical constants, population and the death rate by F. A. Benford [1]. Afterward, this law has been developed and applied in forensic science, e.g. auditing fraudulent financial data in accounting [3], investigating fraud in election, detecting manipulation of multimedia signals such as images and videos [4, 5].

Applying BL to natural datasets may be formulated as follows. Suppose we have a real random variable X with n observations that may be represented as observations vector $\mathbf{x} = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix}^T$ and $D_m(X)$ represents the m^{th} significant decimal digit of X. The common BL describes the behavior of probability mass function as:

$$\Pr\left(D_m(X) = d\right) = \begin{cases} \log_{10}\left(1 + \frac{1}{d}\right) & m = 1, d = 1, 2, \cdots, 9 \\ \sum\limits_{k=10^{m-2}}^{10^{m-1}-1} \log_{10}\left(1 + \frac{1}{10k+d}\right) & m > 1, d = 0, 1, \cdots, 9 \end{cases} \quad (1)$$

Significant decimal digits can be calculated by string processing based algorithms or may be determined explicitly as [2]:

$$D_m(X) \triangleq \begin{cases} 0 & ; X = 0 \\ \left\lfloor 10^{m-1} S(X) \right\rfloor - 10 \left\lfloor 10^{m-2} S(X) \right\rfloor & ; X \neq 0 \end{cases} \quad (2)$$

in which $S(X)$ is defined as the decimal significand function:

$$S(X) \triangleq \begin{cases} 0 & : X = 0 \\ 10^{\log_{10}|X| - \lfloor \log_{10}|X| \rfloor} & ; X \neq 0 \end{cases} \quad (3)$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

Benford's law has four basic and useful properties that help us to find whether Significant Digits (SDs) of the random variable X follow BL [2]. These properties are known as the uniform distribution, scale-invariance, base-invariance and sum-invariance characterizations. The uniform distribution property is the most powerful tool among them. This property states that significant digits distribution of the random variable X is Benford if and only if the random variable $Y = \log_{10}|X| \mod 1$ is uniform in the range $[0, 1)$. In short, the uniform distribution property is expressed as:

$$D_m(X) \sim B_m \Leftrightarrow Y \sim U(0, 1) \quad (4)$$

where B_m is the m^{th} digit Benford distribution.

Up to now, a few transform domains have been identified to map the SDs probabilities of image data onto Benford's law. Gradient Magnitude (GM) [6], 2-D Discrete Cosine Transform (DCT) [7] and 2-D Discrete Wavelet Transform (DWT) [8] are three prominent transforms in this context. But, these transforms cannot precisely map significant digits probabilities of image data onto the common Benford's law. In order to resolve this problem at least for the First Significant Digits (FSDs), a parameterized logarithmic distribution, called Generalized Benford's Law (GBL) [4, 9], was proposed as:

$$\Pr\left(D_1(X) = d\right) = N \log_{10}\left(1 + \frac{1}{s + d^q}\right) \quad ; \forall d = 1, 2, \cdots, 9 \quad (5)$$

in which the coefficient N is a normalization factor and the parameters s and q are used to cover the best matching for diverse images. It is obvious that the common BL for $m = 1$ is a special case of GBL employing the parameters N, s and

978-1-4673-5604-6/12/$31.00 ©2012 IEEE