UNIVERSIDADE FEDERAL DE PERNAMBUCO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA

A CONFORMIDADE À LEI DE NEWCOMB-BENFORD DE QUALIFICADORES DE PONTOS DE INTERESSE EM IMAGENS DIGITAIS

DISSERTAÇÃO DE MESTRADO

Aluno: Felipe Maia (fm@cin.ufpe.br)
Orientador: Sílvio de Barros Melo (sbm@cin.ufpe.br)

2012

Universidade Federal de Pernambuco

Centro de Informática

Felipe Maia

A CONFORMIDADE À LEI DE NEWCOMB-BENFORD DE QUALIFICADORES DE PONTOS DE INTERESSE EM IMAGENS DIGITAIS

Dissertação submetida ao Centro de Informática, sob orientação do Professor Sílvio de Barros Melo, Ph.D como requisito parcial para obtenção do título de Mestre do programa de Pós Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco.

Orientador: Sílvio de Barros Melo

Recife/PE

2012

Agradecimentos

Dedico este trabalho e agradeço à minha amada Michelle, pelo zelo, compreensão e incentivo. Agradeço a minha família, em especial meus pais e meu irmão Bruno Maia, pelo incentivo e educação que me foi proporcionado durante toda a minha vida, sem os quais eu não chegaria onde cheguei. Agradeço aos colegas e amigos pela colaboração e incentivo mútuo. Obrigado ao professor Silvio Melo, pela paciência, dedicação e clareza que desprendeu desde o início de meu curso de graduação: sua orientação e contribuição foram fundamentais para o término e sucesso deste trabalho. Obrigado a Diogo Brandão Borborema Henriques pelo compartilhamento de sua pesquisa, referências e de suas opiniões em relação ao trabalho. Agradeço a meus amigos e colegas de trabalho pelo incentivo e ajuda, em especial a Adonis Tavares da Silva pelo incentivo em todo o processo do trabalho.

“All my life I've had one dream, to achieve my many goals.”
- HOMER SIMPSON

Resumo

As sequências de números aleatórios advindas de situações reais são geralmente modeladas através de funções contínuas de densidade que associam valores de probabilidade a pontos na reta real correspondentes aos números das sequências. O agrupamento dos números aleatórios de acordo com o dígito mais significativo para algumas sequências do mundo real tem revelado um fenômeno já observado no século XIX: a chamada Lei de Benford. Esta lei afirma que as mantissas dos logaritmos desses números estão distribuídas segundo uma uniforme. Sequências tais como área da superfície de rios, população de cidades, razão de números da sequência de Fibonacci, lista de números que aparecem em documentos financeiros, valores em declarações de imposto de renda, tamanho das manchas solares, e muitas outras grandezas seguem esta lei. Esta propriedade presente em algumas grandezas tem sido útil na identificação de patologias nos dados. Neste trabalho, empregamos os métodos estatísticos mais utilizados na área para demonstrar que os qualicadores de pontos de interesse, como o detector de Harris, aplicados a imagens digitais comuns são grandezas que se conformam à Lei de Benford. O detector de Harris extrai um valor de cada pixel da imagem baseado em derivadas de segunda ordem das cores, e é utilizado para classificar os chamados pontos de interesse que, dentre as muitas aplicações, possibilita o rastreamento de objetos num vídeo e a calibração de uma câmera em Realidade Aumentada. Os experimentos com as sequências de coeficientes extraídas de um conhecido banco de imagens confirmam que os seletores de pontos de interesse se adequam à Lei de Benford: a conformidade do detector de Harris é tão significante, que concluímos que na literatura ela é a grandeza extraída de dados reais que melhor se adequa à Lei até o momento. No trabalho também discutimos o estado da arte e as limitações nas medidas de conformidade utilizadas na maioria dos ambientes aplicados.

Palavras chave: Detector de Harris, Lei de Benford, Lei de Newcomb-Benford, Lei dos Dígitos Significativos, Seleção de pontos de interesse.

Abstract

Sequences of random numbers arising from real situations are usually modeled using continuous density functions that associate probability values to points on the real line corresponding to the numbers of sequences. The grouping of random numbers, according to the most significant digit for some real-world sequences, has revealed a phenomenon already observed in the nineteenth century: the so-called Benford's Law. This law states that the probability of occurrence of numbers is such that all mantissa of their logarithms are equally likely. Sequences such as surface area of rivers, population of cities, list of numbers that appear in financial statements, value statements on tax declarations, size of sunspots, and many other quantities follow this law. This property present in some quantities has been helpful in the identification of pathologies in the data. In this paper, we use statistical methods commonly used in the area to demonstrate that feature selection qualifiers, such as the Harris Corner Detector, applied to digital images are quantities that conform to Benford's Law. The Harris detector extracts a value for each pixel of the image based on second derivatives of the pixels, and is used to classify features that among many applications, enables tracking objects in a video stream and calibration of a camera in Augmented Reality. Experiments with sequences of feature qualifiers extracted from a known image database confirm that the feature selectors follow Benford's Law: the compliance of the Harris detector is so significant, that we conclude that in the literature it is the quantity extracted from actual data that best suits the Law so far. In this work we also discuss the state of the art and limitations of the compliance measures used in most of the applications in the area.

Key-Words: Harris Corner Detector, Benford's Law, Newcomb-Benford's Law, Significant Digit Law, First Digit Law, Feature Selection.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O crescimento do conhecimento e da cultura humana tem sido apontado como sendo "exponencial", e as medidas desse crescimento em geral estão associadas à geração de novas tecnologias. A idéia da aceleração exponencial das mudanças tecnológicas foi popularizada pelo matemático americano Vernor Vinge na década de 80 em sua ficção científica Marooned in Realtime (1986). No entanto, desde 1965 que o co-fundador da Intel Gordon Moore observou que a quantidade de componentes em circuitos integrados dobrava a cada dois anos desde a invenção do transistor em 1958. Naturalmente este crescimento estava também associado à demanda criada pelo crescimento populacional e da retroalimentação do próprio desenvolvimento tecnológico.

Nos últimos anos, com o advento da internet e a necessidade do processamento cada vez mais rápido de informações, viu-se aumentar vertiginosamente a demanda por produtos e processos associados às tecnologias de apresentação visual e, nesse contexto, desenvolveram-se duas áreas de pesquisa que hoje são essenciais em Ciência da Computação: a Realidade Virtual e a Realidade Aumentada. A primeira permite uma imersão do usuário no mundo sintético, inteiramente produzido através das técnicas de computação gráfica; e a segunda utiliza vídeos produzidos do mundo real por câmeras comuns, e insere em tempo real elementos sintéticos, que podem ser de alta complexidade gráfica, como objetos de geometria complexa 3D devidamente iluminados de acordo com a iluminação do ambiente real, ou mesmo elementos simples, como textos com informações turísticas.

A produção de objetos sintéticos 3D na cena real depende da reconstrução 3D (geração de modelos sintéticos) de pelo menos parte da cena real, para que se permita a síntese e a implantação dos objetos adicionais de forma fisicamente coerente com a cena real, conferindo ao usuário a sensação de plausibilidade física na imersão que se pretende na Realidade Aumentada. A reconstrução dos modelos da cena real, por sua vez, depende da identificação nas diversas imagens de certos objetos, através de sua geometria e posição, bem como de parâmetros da própria câmera, empregando conhecimentos desenvolvidos principalmente na década de 90 na área de Visão Computacional, que se assenta basilarmente na Geometria Projetiva e na Álgebra Linear.

A reconstrução 3D compreende uma série de etapas, todas envolvidas no processo de inferir sistemas de coordenadas, parâmetros de câmeras e objetos a partir de uma sequência de imagens ou de um vídeo. Nesse processo, uma etapa-chave é a identificação de pontos de interesse ao longo das imagens, tecnicamente chamada de "*feature tracking*" ou rastreamento de pontos de interesse. O conceito de *features* se baseia na ideia de que em vez de olhar a imagem como um todo, pode ser vantajoso selecionar alguns pontos especiais na imagem e realizar uma análise local sobre estes.

Esta abordagem funciona bem desde que um número suficiente de tais pontos possam ser detectados, e que estes pontos possuam características estáveis e distinguíveis, podendo ser precisamente localizados em uma imagem subsequente. Entre os detectores de *features*, o detector de *features* de Harris é uma abordagem clássica que procura por cantos (*corners*) em uma imagem. Mais precisamente, o detector de Harris é um parâmetro extraído da matriz de covariância de um dado pixel, e que o classifica de acordo com sua eficácia quanto a ser usado como *feature*. Os autovalores ou o mínimo dos autovalores dessa mesma matriz podem também ser utilizados como qualificadores de pontos de interesse.

O conjunto de qualificadores de pontos de interesse de todos os pixels de uma imagem qualquer (aleatória) pode ser modelado como uma sequência de números aleatórios. Embora não se conheça uma distribuição paramétrica que governe o comportamento de uma tal sequência, existe uma lei, descoberta no final do século XIX, que limita a ocorrência dos dígitos mais significativos dos números da sequência, quando esta for formada por números não enviesados, ou advindos de grandezas naturais: a chamada Lei de Newcomb-Benford.

Neste trabalho nós analisamos se três qualificadores de *features*, especificamente o detector de Harris, ambos os autovalores e o mínimo autovalor da matriz de covariância, de forma extremamente significativa a Lei de Newcomb-Benford. Para tal, nós investigamos o estado da arte da Lei e suas aplicações, investigando os trabalhos na área de imagens digitais e suas metodologias utilizadas. Então definimos uma metodologia de análise baseada nos trabalhos anteriores, composta de diversos testes/medidas. Comprovamos experimentalmente que os três classificadores analisados seguem à Lei, e comparando os resultados obtidos com os da literatura, concluímos que o classificador de pontos de interesse de Harris é a grandeza extraída de dados reais que melhor se adequa à Lei dentre todas as grandezas conhecidas na literatura até o presente momento.

O restante deste trabalho está organizado da seguinte forma: no Capitulo 2 são apresentados os conceitos gerais referentes à Lei de Newcomb-Benford, juntamente com o relato do estado da arte da área, incluindo suas aplicações, e por fim são abordados os métodos de análise de conformidade de dados experimentais à Lei.

O Capítulo 3 apresenta os conceitos básicos dos pontos de interesses em imagens digitais, a definição do detector de Harris e seu uso aplicado na seleção de pontos de interesse em uma imagem.

No Capítulo 4 é exposta a metodologia empregada na realização deste trabalho, incluindo o estado da arte das metodologias de aderência à Lei de Newcomb-Benford, o procedimento adotado para a análise, além dos detalhes dos experimentos e da implementação.

No Capítulo 5 são expostos os resultados das análises da conformidade dos seletores de pontos de interesse, com seus julgamentos e comentários, e então é realizada uma comparação direta dos resultados obtidos com os relatados pela literatura.

Por fim, o Capitulo 6 realiza o fechamento do documento, ressaltando os resultados, as contribuições realizadas, e propondo os trabalhos futuros a serem realizados.

2 Lei de Newcomb-Benford

A Lei do Dígito Significativo é a observação empírica de que em muitas tabelas contendo dados numéricos de ocorrência natural, os dígitos mais significativos não são uniformemente distribuídos como é intuitivamente esperado, e sim seguem uma distribuição logarítmica particular. A primeira referência escrita relatando este fenômeno é um artigo de duas páginas de Simon Newcomb publicado em 1881 no American Journal of Mathematics (NEWCOMB, 1881), que elicita:

"The law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally probable."

Traduzindo, isto significa que a probabilidade de ocorrência da mantissa do logaritmo de qualquer número é equiprovável. Newcomb notou que as primeiras páginas das tabelas de logaritmo se desgastam muito mais rápido que as últimas, e após diversas pequenas heurísticas, concluiu a lei das mantissas equiprováveis.

Cinquenta e sete anos depois, o físico Frank Benford redescobriu a lei (BENFORD, 1938) e a apoiou com mais de vinte mil observações de vinte tabelas distintas incluindo dados de áreas de rios, constantes físicas, massas atômicas, e números retirados de artigos e jornais.

Devido ao fato do artigo de Newcomb ter passado despercebido, e do artigo de Benford ter recebido bastante atenção, a Lei do Dígito Significativo ficou conhecida na literatura como Lei de Benford. Neste trabalho, os termos Lei de Benford, Lei de Newcomb-Benford, NB-Lei e Lei do Dígito Significativo, se referem a este mesmo fenômeno.

Neste capítulo, apresentamos inicialmente os conceitos gerais referentes à Lei de Newcomb-Benford, depois revisamos o estado da arte da área, incluindo suas aplicações, e finalmente abordamos os métodos de análise de conformidade de dados experimentais à Lei.

2.1 Conceitos Gerais

Qualquer número $x \in \mathbb{R}$ pode ser reescrito como $x = m_b \cdot b^n$, com $m_b \in [1/b, 1)$, para algum $n \in \mathbb{N}$, onde $b \in \{2,3,..\}$ representa a base do sistema numérico, e m_b é chamada a mantissa de x. $D_k^{(b)}(x)$ denota o k-ésimo dígito significativo de $x \in \mathbb{R}^*$na base b, ou seja, $D_2^{(10)}(\pi) = 1$. Quando no sistema decimal ($b = 10$) o sobrescrito é omitido.

Seja m_b a mantissa de um número real x, a versão mais geral da Lei do Primeiro Dígito (HILL, 1995c) é dada por:

$$prob\left(m \leq \frac{t}{b}\right) = \log_b(t),\ t \in [1, b) \quad (2.1)$$

Note que t acaba representando a probabilidade acumulada do dígito d quando $t = d$.

Partindo desta definição é possível deduzir que a probabilidade do digito d ser o dígito mais significativo de um número real na base decimal é dada por:

$$prob(D_1 = d) = \log_{10}\left(1 + \frac{1}{d}\right) \qquad d = 1,2,\dots,9 \quad (2.2)$$

Para calcular a probabilidade de um digito d aparecer na posição n (D_n), chegamos à fórmula:

$$prob(D_n = d) = \sum_{i=10^{n-2}}^{10^{n-1}-1} log_{10}(1 + (10i + d)^{-1}) \quad (2.3)$$

A probabilidade da sequência de dígitos $(d_1, d_2, \dots, d_n)$ serem os dígitos mais significativos, por sua vez é dada por:

$$prob(D_1 = d_1, \dots, D_k = d_k) = \log_{10}\left(1 + \left(\sum_{i=1}^{k} d_i \cdot 10^{k-i}\right)^{-1}\right) \quad (2.4)$$

Por exemplo, a probabilidade do algarismo "2" seguido do "0" ("20") serem os dígitos mais significativos de um número real é de $\log_{10}\left(1 + \frac{1}{20}\right) \cong 0.021$. A Figura 2-1 expõe as probabilidades dos dígitos 1,2,...,9 aparecerem na posição mais significativa.

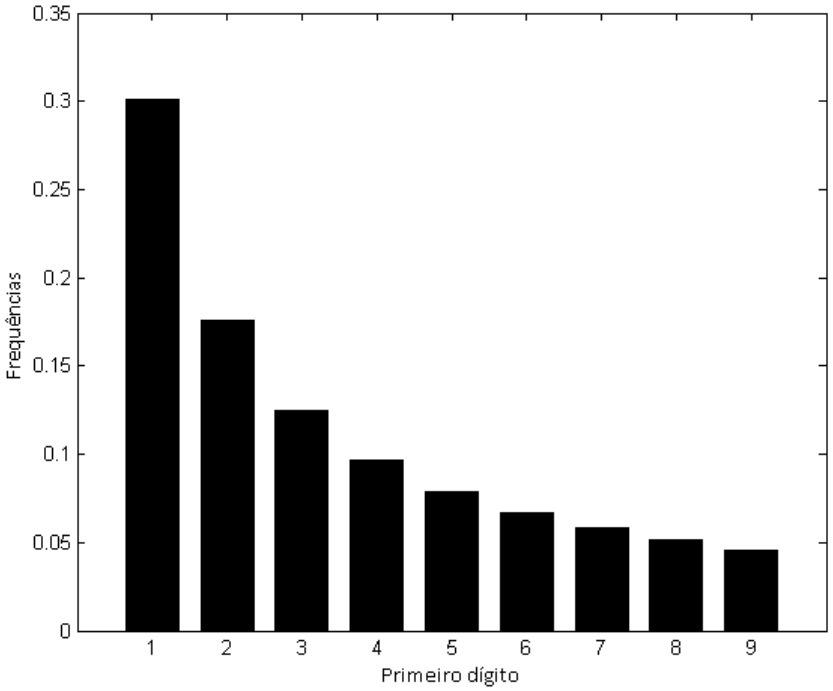

Figura 2-1: Probabilidade dos dígitos (1,..9) aparecerem como dígito mais significativo de acordo com a NB-Lei

Podemos perceber que os dígitos significativos são dependentes entre si, e que essa dependência entre dígitos decai rapidamente à medida que a distância entre os dígitos aumenta, e segue-se facilmente da equação genérica (2.4) que a distribuição do n-ésimo dígito mais significativo aproxima-se exponencialmente de uma distribuição uniforme à medida que $n \to \infty$ (HILL, 1995c).

Uma sequência de números que atenda à NB-Lei é invariante escalar. Isto quer dizer que a proporção de cada dígito mais significativo não se altera quando se multiplicam todos os elementos por uma constante. A intuição disto parte de que a conformidade não deve ser sujeita à unidade de medida adotada e sua prova pode ser encontrada em (HILL, 1995b). O matemático Terry Tao fornece uma explicação mais simples para o fenômeno (TAO, 2009). Este autor ainda conclui que a invariância escalar acontece por este ser um caso particular do fato que medidas de Haar são únicas. Quando escalamos uma dada sequência em base 10 por 2, temos que a proporção de números começando com os dígitos 2 e 3 se igualará a proporção de números começando com o dígito 1.

Similarmente à invariância escalar, os números de uma sequência Benford são invariantes à base. Isto significa que eles continuarão obedecendo à NB-Lei caso sofram uma mudança de base. A prova também pode ser encontrada em (HILL, 1995b).

2.2 Estado da Arte

Como já vimos anteriormente, a Lei de Newcomb-Benford estipula uma frequência na qual os dígitos aparecem quando surgem de uma lista numérica ordenada de fonte natural, sendo relatada inicialmente por Simon Newcomb, ao observar livros de tabelas de logaritmos (NEWCOMB, 1881). Em seu estudo, Newcomb afirma que os logaritmos ocorrem com maior frequência nos dígitos mais baixos, diminuindo gradualmente ao se afastar do dígito 1 até o dígito 9. Através de um breve ensaio sobre as razões esperadas entre números que acontecem naturalmente, ele conclui seu trabalho afirmando que a probabilidade na qual os números ocorrem é tal que as mantissas de seus logaritmos são equiprováveis.

Poincaré contribui para a formalização deste estudo em um pequeno ensaio em sua obra Calcul des Probabilités (POINCARÉ, 1912). O autor argumenta que, observados números consecutivos em uma lista suficientemente grande de logaritmos, intui-se que, em dada posição significativa, considerando os dígitos 0,1,…,9, a ocorrência de números pares ou impares são eventos equiprováveis. Avaliando uma função que retorna “1” caso o dígito observado seja par ou “-1” caso o dígito observado seja ímpar, Poincaré demonstra que a média de tal função tende a zero. Ao final de seu ensaio o autor exprime a necessidade de se fundamentar uma teoria sobre as probabilidades das ocorrências dos dígitos numa tabela numérica. Franel apresenta

algumas correções ao trabalho de Poincaré, e confirma que a probabilidade de qualquer dígito de fato tende a 1/10 quando a posição do dígito observada tende ao infinito (FRANEL, 1917).

A linha de pensamento que inevitavelmente levaria à formulação da lei dos dígitos é retomada por Weyl ao discutir sobre distribuição dos números em módulo um (WEYL, 1916). Dentre suas contribuições, explicita-se seu teorema e os critérios para determinar equidistribuição módulo um, fundamentando o caminho para a formulação de sequências de baixa discrepância, que são sequências tais que, para cada *n* e os seus primeiros *n* elementos, elas se aproximam de uma uniforme.

Cerca de meio século depois da publicação inicial, Frank Benford, relatou de forma independente o mesmo fenômeno (BENFORD, 1938), desenvolvendo sua pesquisa em torno da distribuição do primeiro dígito significativo de diversas sequências de números provenientes de fontes tidas como naturais. Dentre os dados estudados por Benford, incluem-se, além dos previamente já informados, dados de taxas de mortalidade, taxas de população de cidades, calor específico de elementos, além de outros.

Apesar dos dados estudados por Benford possuírem um certo desvio das frequências esperadas, os dados claramente convergem para uma distribuição logarítmica das probabilidades de ocorrências dos dígitos. Ele também apresentou casos não conformes com a NB-Lei. Em (WEISSTEIN, 2012), destaca-se que dos valores de uma base contendo mais de 54 milhões de constantes físicas, 30% deles começam com o dígito “1”.

Outros autores como Levy (LÉVY, 1939) e Robbins (ROBBINS, 1953) elaboram melhor a teoria estabelecida por Weyl, e outros como Goudsmith (GOUDSMIT e FURRY, 1944) e (HSÜ, 1948) trabalham na fundamentação deixada por Benford. Porém é com Pinkham (PINKHAM, 1961) onde surge um avanço significativo na estruturação de uma lei das probabilidades dos dígitos; é estabelecida uma forte relação entre a Lei de Newcomb-Benford e a invariância escalar. Fazendo uma analogia a uma agulha em um disco flutuante de raio unitário, Pinkham demonstra que os dígitos que ocorrem com probabilidade logarítmica são os únicos que mantém sua probabilidade de ocorrência constante, quando multiplicados por uma escala.

Alguns anos mais tarde Knuth aponta uma falha no trabalho de Pinkham (KNUTH, 1969). Pinkham assume que existe uma certa função de probabilidade contínua e Knuth demonstra que tal função não existe. Apesar das correções, o resultado de Pinkham é mantido. A referência da invariância escalar com a probabilidade de ocorrência dos dígitos é extremamente pertinente; descoberta que Pinkham atribui a R. Hamming, que demonstra os efeitos da computação de pontos flutuantes nas

mantissas dos números cujos dígitos ocorrem com probabilidade logarítmica (HAMMING, 1970).

Nota-se que em sua publicação, Pinkham cita apenas dois livros e cinco trabalhos tratando sobre a distribuição dos dígitos (incluindo o trabalho original de Benford e sem fazer qualquer menção a Newcomb). No entanto, poucos anos mais tarde, Raimi (RAIMI, 1976) revisa uma vasta gama de literatura lidando diretamente com este fenômeno. Além de incluir a publicação de Newcomb e outro não mencionado por Pinkham, Raimi estuda outros 27 trabalhos sobre a lei, indicando uma grande explosão de publicações desde 61, dentre estas uma menção por Feller em seu livro (FELLER, 1966).

Raimi comenta sobre a falta de visibilidade do trabalho de Newcomb na área, mas menciona que o próprio Newcomb diz que o fenômeno já era conhecido em seu tempo, e que seria trabalhoso renomear o que na época já era conhecido como Lei de Benford, apenas para mais tarde descobrir que uma segunda mudança de nome é necessária: “Existe amplo precedente para nomear leis e teoremas não pelos seus descobridores, ou então metade da análise seria nomeada segundo Euler.” (tradução livre) (RAIMI, 1976).

Apesar de existir certa visibilidade do assunto, boa parte do entendimento sobre o fato estudado não se dá até a década de 90 com pesquisas publicadas por Hill (HILL, 1988) (HILL, 1995a) (HILL, 1995c). O pesquisador solidifica boa parte do conhecimento gerado durante todo século passado, além de provar de forma definitiva que a distribuição em questão é a única que pode atender aos princípios de invariância de base e escala através da sua definição da σ-algebra de mantissas.

Ao se juntar com Berger, Hill publica um compêndio dedicado à NB-Lei (BERGER e HILL, 2011). Nele estão contidas definições, características e provas dentre outros conteúdos, apenas referentes ao assunto em questão. Uma lista completa e atualizada do estado da arte da área, contendo publicações referentes à Lei de Newcomb-Benford pode ser acessada em (BERGER e HILL, 2012).

2.2.1 Aplicações

Uma das primeiras potenciais aplicações da NB-Lei surgiu no início do século XX com Boring. Em seu trabalho, o autor estuda de um ponto de vista psicofísico como pessoas atribuem probabilidades a eventos sobre os quais são leigas (BORING, 1920). A teoria melhor aceita na época era de que se a pessoa não tem qualquer motivo para viés, ela admite que todos os eventos são equiprováveis.

Em contextos mais próximos das ciências puras, Hamming propôs uma série de aplicações de hardware e software que podem se beneficiar do uso da NB-Lei (HAMMING, 1970). Outros autores tomam caminhos similares mostrando que

algoritmos bem difundidos podem ter seus erros de cálculo minimizados pelas operações de ponto flutuante quando a lei de Newcomb-Benford é aplicada. Berger e Hill estudam os erros esperados nas operações de ponto flutuante do método de Newton (BERGER e HILL, 2007). A mesma dupla, junto com Kaynar e Ridder estudaram o comportamento de cadeias de Markov, mostrando que de fato obedecem a NB-Lei, e também mostrando como os erros de underflow, overflow e round-off podem ser minimizados nos cálculos computacionais através deste fenômeno (BERGER et al., 2011).

Aplicações de cunho mais exótico também podem ser encontradas quando se estuda a ocorrência de dígitos. Shao e Ma observaram a complacência de algumas propriedades de pulsares com a lei de Newcomb-Benford (SHAO e MA, 2011). Dentre as características das pulsares estudadas encontram-se período baricêntrico e velocidade de rotação (e suas derivadas em relação ao tempo), notando-se uma conformidade com a lei.

Sistemas dinâmicos também foram bastante explorados no contexto da NB-Lei. Sob o incentivo da vasta aplicabilidade dos sistemas dinâmicos como modelos para sistemas físicos e sociais, diversos autores buscaram encontrar relações entre os modelos construídos e a lei dos dígitos. Tolle, et. al. avaliaram a distribuição dos dígitos de alguns modelos de autômato celular e de dinâmicas de fluidos (TOLLE, BUDZIEN e LAVIOLETTE, 2000). Os autores partiram sem qualquer viés de conformidade com o fenômeno e testaram os modelos sob condições consideradas ótimas pelos mesmos. Conclui-se que os modelos de autômato celular produzem dígitos uniformemente distribuídos, mas que este fato já era conhecido e esperado. Os autores, no entanto, ficaram surpresos em descobrir o grau de conformidade tanto em modelos de gases como de líquidos. Outros autores adotaram linhas similares de pesquisas avaliando a conformidade da lei de Benford em sistemas discretos unidimensionais (SNYDER, CURRY e DOUGHERTY, 2001) (BERGER, BUNIMOVICH e HILL, 2004) e também em sistemas com comportamento exponencial (BERGER, 2006).

Varian foi o primeiro a utilizar a Lei na área de análise de dados como forma de validar dados no contexto sócio-econômico, avaliando resultados obtidos em uma simulação de um sistema de previsão de crescimento urbano para a área da Baía de São Francisco (MORGAN et al., 1972).

No final da década de 80, o uso da NB-Lei começa a aparecer de forma significativa na contabilidade e auditoria. Carslaw foi o precursor desta conduta ao estudar o comportamento de dados financeiros de empresas da Nova Zelândia (CARSLAW, 1988). Carslaw indica que existe uma suspeita de que gerentes e administradores estejam arredondando valores de balancetes para melhorar o desempenho das

empresas. Segundo ele, seres humanos tendem a memorizar apenas o primeiro dígito de um número e, portanto, um número na forma $n10^k$ possui alto valor cognitivo.

Para verificar se existia manipulação de dados visando esses valores, Carslaw verificou a quantidade de zeros que apareciam como segundo dígito mais significativo nos valores que constavam nas demonstrações contábeis das empresas avaliadas. Como base para sua análise o autor utilizou a NB-Lei como frequência esperada. Ele detectou não apenas que o dígito 0 mostrava um excesso na segunda posição, mas que havia uma carência do dígito 9 na segunda posição e considerou isso como evidência suficiente para suportar a hipótese que havia manipulação nos dados.

Thomas replicou o trabalho de Carslaw em firmas americanas (THOMAS, 1989). Assim como o autor tomado como base, Thomas utilizou a NB-Lei para encontrar excesso e falta de dígitos na primeira e segunda posição, concluindo que existe evidência suficiente para suportar a hipótese de que existe manipulação tanto no ganho (arredondamento para cima) quanto na perda (arredondamento para baixo).

Nigrini é um dos autores atuais que mais assiduamente defende o uso da Lei de Newcomb-Benford como procedimento analítico. Ele juntou algumas noções do trabalho de Carslaw com outras do trabalho de Thomas e elaborou em sua tese técnicas para detectar desvios no imposto de renda (NIGRINI, 1992). Em um outro trabalho, ele demonstrou que os dados coletados de imposto de renda possuem conformidade com a NB-Lei (NIGRINI, 1996).

Anos mais tarde Nigrini aplica novamente a Lei para detectar desvios em dados fornecidos por companhias petrolíferas (NIGRINI; MITTERMAIER, 1997), demonstrando como esta pode ser usada como ferramenta de revisão analítica no auxílio ao planejamento de auditoria.

Com uma abordagem um pouco diferente, Busta e Weinberg estudaram a possibilidade de se criar um sistema de apoio à decisão baseado na NB-Lei e redes neurais (BUSTA; WEINBERG, 1998). Como os autores não dispunham de dados reais, eles utilizaram bases de dados simulados, misturando dados selecionados de uma sequência pura (NB) com amostras de uma distribuição ruidosa em uma proporção predeterminada.

Seguindo uma linha similar, Bhattacharya, Xu e Kumar (BHATTACHARYA; XU; KUMAR, 2011) propõem um sistema de suporte à decisão baseado em redes neurais. Novamente um procedimento de revisão analítica é utilizado para classificar os dados simulados por sua conformidade com a NB-Lei. Também foi utilizada uma técnica de otimização baseada em algoritmo genético para escolher qual modelo de rede neural melhor classifica a conformidade de um conjunto de dados como sendo NB-Lei. Diferentemente de seus precursores, os autores deste utilizam conjuntos com mais

elementos, além de testar novas entradas nas redes neurais, mantendo apenas aquelas que foram consideradas de sucesso.

2.2.2 Aplicações em imagens digitais

Jolion foi o primeiro a estudar a distribuição dos dígitos em imagens digitais (JOLION, 2001). Neste trabalho foi demonstrado que apesar dos valores dos pixels das imagens não seguirem a lei de Benford, as magnitudes dos gradientes, além da decomposição piramidal baseada na transformação de Laplace, obedecem à lei do dígito significativo. O autor então sugeriu como aplicação um método de codificação baseado em entropia que utiliza-se da probabilidade Benford esperada.

Acebo e Sbert (ACEBO; SBERT, 2005) propuseram o uso da lei de Benford para determinar se imagens geradas sinteticamente foram renderizadas baseadas em métodos fisicamente realistas. No entanto, o fato de que muitas imagens não seguem a lei de Benford no domínio de *pixels* coloca esta aplicação em questão.

Sanches e Marques (SANCHES; MARQUES, 2006) mostraram que o primeiro dígito da magnitude dos gradientes de imagens médicas de ressonâncias magnéticas, tomografias computadorizadas e ultrassons obedecem à lei de Benford. Eles então propõem um algoritmo de reconstrução baseado na NB-lei que não requer ajuste de parâmetros regulatórios, obtendo bons resultados.

Fu, Shi e Su (FU; SHI; SU, 2007) mostraram que a distribuição do dígito mais significativo dos coeficientes do bloco-DCT (transformada discreta de cosseno) segue satisfatoriamente a NB-lei e que os coeficientes quantizados JPEG seguem uma distribuição similar à lei logarítmica de Benford quando a imagem JPEG for comprimida somente uma única vez. Eles então propuseram um modelo empírico paramétrico para formular essa lei observada. Além disso, eles demonstraram que essa distribuição é muito sensível à compressão JPEG dupla, propondo então a aplicação da NB-lei na análise forense de imagens.

Em um trabalho relacionado, porém independente, Pérez-González, Heileman e Abdallah (PÉREZ-GONZÁLEZ; HEILEMAN; ABDALLAH, 2007) apresentaram uma generalização da lei de Benford para o dígito mais significativo. Esta generalização se baseia em manter os dois primeiros termos da expansão de Fourier da função densidade de probabilidade dos dados no domínio logarítmico modulado. Além disso, demonstraram que imagens no domínio da transformada discreta de cosseno (DCT) seguem esta generalização. Eles então propuseram a aplicação da lei de Benford na área de esteganálise de imagens para verificar se uma imagem contem uma mensagem escondida.

Qadir, Zhao e Ho (QADIR; ZHAO; HO, 2010) analisaram a NB-lei aplicada a imagens que utilizam o padrão de compressão JPEG 2000. Eles demonstraram experimentalmente

que imagens no domínio da transformada discreta de *wavelet* (DWT) seguem a lei do dígito significativo. Eles então propuseram a aplicação da NB-lei para estimar o fator de qualidade utilizado na compressão JPEG 2000 de uma imagem. Posteriormente, Qadir et al. (QADIR; ZHAO; HO; CASEY, 2011) propuseram a aplicação da NB-lei a imagens no domínio DWT para identificar imagens naturais que cotenham captação de brilho exacerbado (*glare*).

Heijer e Eiben (HEIJER; EIBEN, 2011) utilizaram a NB-lei como uma medida da qualidade estética para a evolução sem supervisão de arte sintética revolucionaria gerada por computação genética. A NB-lei é uma das três medidas de qualidade e a arte que evolui de acordo com sua avaliação possui características distintas das demais.

2.3 Critérios de Conformidade

O primeiro teste de conformidade de dados naturais à Lei do Digito Significativo foi realizado pelo próprio Frank Benford em seu trabalho original (BENFORD, 1938), utilizando a diferença entre as probabilidades esperadas e observadas (MORGAN et al., 1972). O desvio das frequências, também conhecido por distância ou diferença entre frequências, ε é dado por:

$$\varepsilon = \sum |P_o - P_e|/2 \quad (2.5)$$

onde P_o e P_e representam as probabilidades observadas e esperadas, respectivamente. A divisão por 2 ocorre pois, como a soma das probabilidades é obrigatoriamente 1, a última célula de probabilidade é linearmente dependente das outras. Isto se espelha nos desvios, sempre que é gerado um excesso em algum dígito, uma ausência da mesma intensidade é refletida em um ou mais dígitos: quando olhamos para o somatório, consideramos os desvios duas vezes (uma para o excesso e outra para a falta).

No entanto, as medidas de conformidade mais difundidas na literatura são baseadas nos testes estatísticos χ^2 de Pearson, no teste Z, e no teste Komolgorov-Smirnov (K-S). O teste χ^2 foi o primeiro a ser utilizado (Diaconis, 1977), no entanto seu autor estava mais preocupado com o desenvolvimento teórico do que aplicações em geral. O teste Z foi introduzido por Carslaw (Carslaw, 1988) no contexto da NB-Lei e possui natureza local, analisando os desvios em um único dígito por vez. Os testes χ^2 e K-S, por outro lado, possuem natureza global e analisam todos os dígitos.

A vasta maioria dos autores verifica a conformidade apenas para o primeiro dígito. Algumas aplicações surgiram trabalhando com os demais dígitos, porém são poucas em número e ainda assim raramente vão além dos dois primeiros.

Uma descrição detalhada do teste Z aplicado à NB-lei (Thomas, 1989), sendo calculado por:

$$Z = \frac{|P_o - P_e| - \frac{1}{2n}}{\sqrt{\frac{P_o(1-P_o)}{n}}} \quad (2.6)$$

onde P_o e P_e representam as probabilidades observadas e esperadas respectivamente, n é o tamanho total da amostra, e o termo $\frac{1}{2n}$ é um fator de correção de continuidade e só é utilizado quando ele é menor que o primeiro termo do numerador.

O teste estatístico χ^2 de Pearson, por sua vez é dado por:

$$\chi^2 = \sum \frac{(O_o - O_e)^2}{O_e} \quad (2.7)$$

onde O_{O_d} e O_{E_d} representam respectivamente a quantidade de elementos observadas e a quantidade de elementos esperadas para o dígito.

Já a estatística do teste Kolmogorov-Smirnov (PETITT, STEVENS, 1977) é dada por: $\max(\sum|P_o - P_e|)$, onde P_o e P_e representam as probabilidades observadas e esperadas respectivamente.

Nigrini e Mittermaier sugeriram (NIGRINI; MITTERMAIER, 1997) uma distância de conformidade para a NB-Lei chamada *M.A.D (Mean Absolute Deviaton)*. O *M.A.D* é calculado dividindo o somatório das diferenças absolutas pela quantidade de dígitos, equivalendo-se à distância ε (Equação 2.5) por uma diferença de escala.

Busta e Weinberg (BUSTA; WEINBERG, 1998) utilizaram-se de redes neurais para a classificação da conformidade dos dados com a NB-Lei, que consideraram em sua análise elementos da estatística descritiva como a frequência de ocorrência dos dígitos das primeiras e segundas posições, média, mediana, desvio padrão, curtose e obliquidade além de valores de estatísticas calculadas como Z e χ^2. Estendendo o trabalho de Busta e Weinberg, Bhattacharya, Xu e Kumar (BHATTACHARYA; XU; KUMAR, 2011) reaplicaram os critérios mais bem sucedidos de seus antecessores além de avaliar novos critérios como χ^2, Kolmogorov-Smirnov (K-S) discreto, distância de Kullbak-Lieber, entropia de Shannon, distância euclidiana, coeficiente de relação de Pearson e o alpha de Judge-Schechter.

Focando no poder de testes estatísticos, Steele e Chaseling (STEELE; CHASELING, 2006) demonstraram que para distribuições de tendência, dentre os testes avaliados, os que apresentaram melhor desempenho foram o K-S discreto, Anderson-Darling (A^2) discreto e o Cramér-von Mises (W^2) discreto. O teste χ^2 teve desempenho insatisfatório quando comparado com os outros testes e o teste Z não foi avaliado.

Em sua tese de mestrado, Wong (Wong, 2010) avaliou os poderes de diversos testes estatísticos para detecção de desvios em sequências NB. O autor simulou várias naturezas de desvios em proporções crescentes como forma de avaliação, medindo o poder do teste em cada caso. Os resultados de Steele e Chaseling foram comprovados, A^2 e W^2 apresentaram os melhores poderes enquanto o teste χ^2 se mostrou novamente menos poder que os outros (o K-S não foi avaliado).

Na metodologia (Capítulo 4) fazemos um estudo aprofundado de quais testes de conformidade são usuais na área de aplicação em foco.

3 Seleção de pontos de interesse e o detector de cantos de Harris

Na área de visão computacional, o conceito de pontos de interesse, também chamados de *features,* tem sido amplamente utilizado para resolver diversos problemas nas áreas de reconhecimento de objetos, reconstrução 3D, registro e categorização de imagens, rastreamento visual, entre outras. O conceito se baseia na ideia de que em vez de olhar a imagem como um todo, pode ser vantajoso selecionar alguns pontos especiais na imagem e realizar uma análise local sobre estes. Esta abordagem funciona bem desde que um número suficiente de tais pontos possam ser detectados, e que estes pontos possuam características estáveis e distinguíveis, podendo ser precisamente localizados em uma imagem subsequente.

Na seleção de tais *features* em uma imagem, obviamente, se por exemplo, escolhermos um ponto qualquer em uma parede branca, não será fácil identifica-lo novamente em uma próxima imagem; se todos os pontos da parede são indistinguíveis, teremos muita dificuldade em identificá-los. No entanto, ao escolhermos um ponto de característica única, nós temos grande chance de encontrá-lo novamente. Na prática, o ponto ou *feature* que escolhemos deve ser único, ou relativamente único quando comparado a sua vizinhança, e deve ser parametrizado de forma que possa ser comparado a outros pontos da imagem subsequente.

Retornando ao nosso exemplo da parede branca, intuitivamente nós devemos escolher pontos que tenham uma mudança significante – por exemplo, uma derivada de valor alto. Os pontos que possuem uma derivada de valor alto estão associados a uma borda ou linha de algum tipo, porém pode ocorrer de todos os pontos na direção da linha possuírem as mesmas características. No entanto, se derivadas em direções ortogonais no ponto são ambas altas, existe uma chance muito maior deste ponto ser único. Por essa razão, a maioria das *features* ou pontos rastreáveis são chamados de cantos (*corners*). A definição mais comumente utilizada de um *corner* foi fornecida por Harris (HARRIS; STEPHENS, 1988).

3.1 O detector de cantos de Harris

Para definir a noção de *corners* em imagens, Harris olha para a mudança direcional média de intensidade em uma janela ao redor do ponto de interesse. Se considerarmos um vetor de deslocamento (u,v), a mudança média de intensidade é dada por:

$$R = \sum (I(x+u, y+v) - I(x,y))^2$$

O somatório é realizado sobre a vizinhança definida ao redor do pixel considerado. Esta mudança média de intensidade pode então ser computada em todas as possíveis

direções, o que leva à definição de que um *corner* é um ponto onde a mudança de intensidade é alta em mais de uma direção. A partir desta definição, o teste de Harris é realizado como se segue. Primeiro nós obtemos a direção com a maior mudança média de intensidade. Em seguida, nós verificamos se a mudança média de intensidade na direção ortogonal também é alta. Se este é o caso, então nós temos um *corner*.

Matematicamente, esta condição pode ser testada através de uma aproximação da fórmula precedente utilizando a expansão de Taylor de segunda ordem:

$$R \approx \sum \left(I(x,y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v - I(x,y) \right) = \sum \left(\left(\frac{\partial I}{\partial x}u \right)^2 + \left(\frac{\partial I}{\partial y}v \right)^2 + 2\frac{\partial I}{\partial x}\frac{\partial I}{\partial y}uv \right)$$

Podemos então reescrevê-la na forma de matriz:

$$R \approx \begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} \sum \left(\frac{\delta I}{\delta x} \right)^2 & \sum \frac{\delta I}{\delta x}\frac{\delta I}{\delta y} \\ \sum \frac{\delta I}{\delta x}\frac{\delta I}{\delta y} & \sum \left(\frac{\delta I}{\delta y} \right)^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

Esta matriz é a matriz de covariância que caracteriza a taxa de variação de intensidade em todas as direções, também conhecida por matriz de Harris. Pode ser mostrado que os dois autovalores da matriz de covariância dão a máxima mudança média de intensidade, juntamente com a mudança média de intensidade na direção ortogonal. Então segue que se ambos os autovalores forem baixos, nós estamos em uma região da imagem relativamente uniforme. Se um dos autovalores for alto e o outro baixo, nós estamos em uma aresta. Finalmente, se ambos os autovalores forem altos, então nós estamos em uma localização de *corner*. Portanto, a condição para que um ponto seja aceito como um *corner* é de que o menor dos autovalores da matriz de covariância seja maior que um dado limiar.

A definição original do detector de *corners* de Harris utiliza-se de algumas propriedades da teoria da decomposição de valores singulares de maneira a evitar o custo explicito de computar os autovalores. Essas propriedades são:

- O produto dos autovalores de uma matriz é igual a seu determinante
- A soma dos autovalores é igual à soma da diagonal da matriz (conhecida como traço da matriz)

Então podemos verificar se ambos os autovalores são altos pelo cálculo final da fórmula de Harris:

$$Det(C) - k \cdot Traço(C)^2 \quad (3.1)$$

É fácil então a verificar que a formula de Harris retorna valores altos se ambos os autovalores também forem altos. O parâmetro ***k*** é um coeficiente de peso que pode ser determinado experimentalmente. Dessa forma, um ponto é considerado uma boa *feature* a ser rastreada se seu coeficiente de Harris ultrapassa um valor limiar.

Posteriormente, Shi e Tomasi (SHI, TOMASI, 1994) propuseram que bons *corners* resultam desde que o menor dos autovalores da matriz de covariância seja maior que um valor de limiar mínimo. Ao utilizar diretamente os autovalores, o uso de um dos parâmetros arbitrários, ***k***, é evitado. O método de Shi e Tomasi não só foi suficiente, mas também, em muitos casos, produziu resultados mais satisfatórios quando comparados ao coeficiente de Harris.

3.2 Seleção de pontos de interesse utilizando o detector de Harris

De maneira a selecionar pontos de interesse utilizando o detector de cantos de Harris, se faz necessário partir de uma imagem original (Figura 3-1). Então, para todos os pontos (*pixels*), os coeficientes de Harris (Equação 3.1) são calculados, mostrados na Figura 3-2 (invertido e com contraste exacerbado para facilitar a inspeção visual).

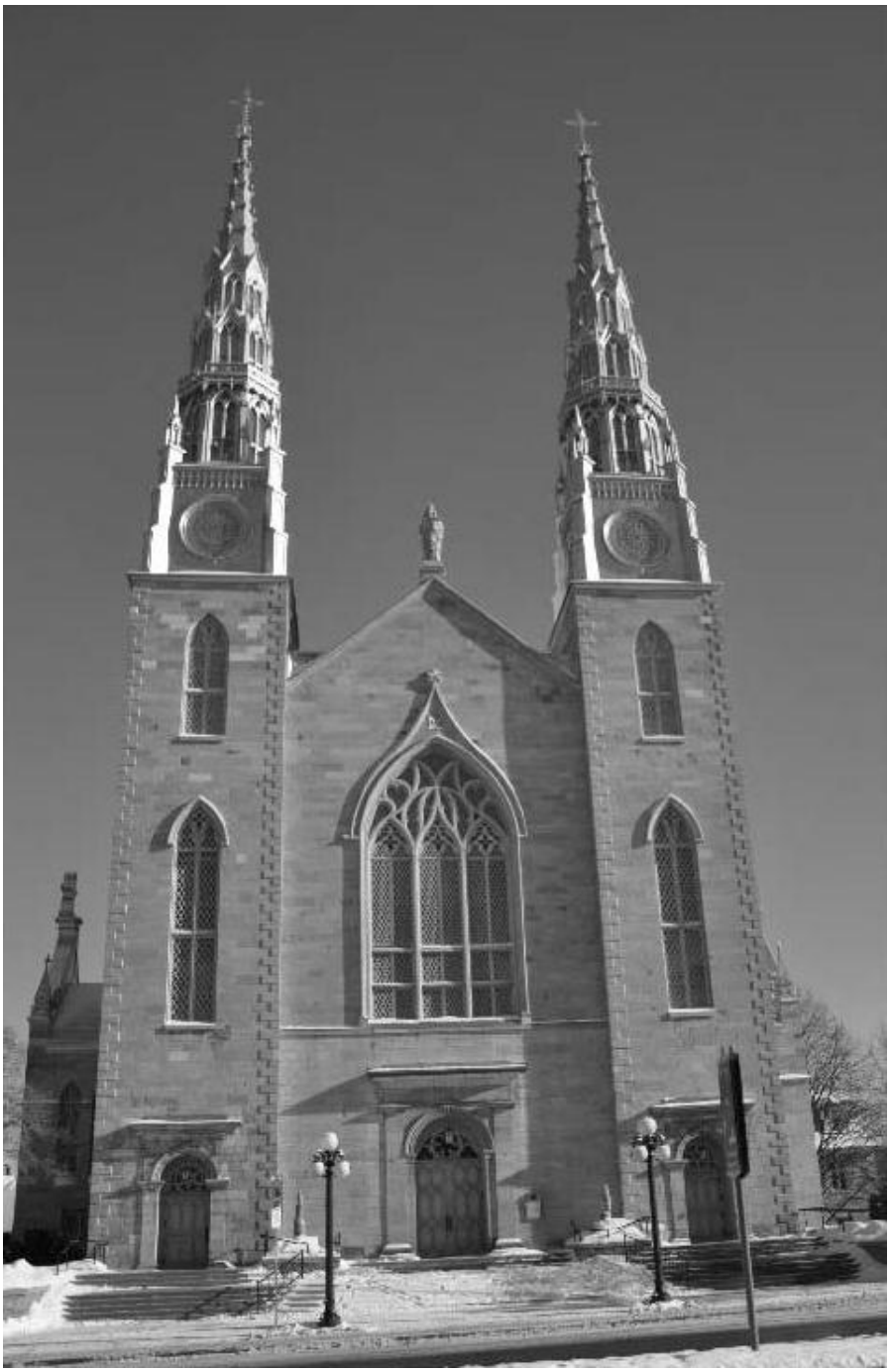

Figura 3-1: Imagem original

Figura 3-2: Valores dos coeficientes de Harris calculados na imagem original (com contraste exacerbado para melhor visualização)

Com os coeficientes calculados, é aplicado um valor de limiar para identificar os *corners* da imagem: pontos com coeficiente superior ao limiar são considerados *corners* enquanto os inferiores são desconsiderados. O resultado pode ser visualizado na Figura 3-3, onde um mapa binário (invertido para melhor visualização) com os pontos classificados como corners é mostrado. Nesta imagem, é possível perceber que diversos aglomerados de *corner pixels* são obtidos, o que contradiz o fato que é desejável detectar *features* únicas bem localizadas.

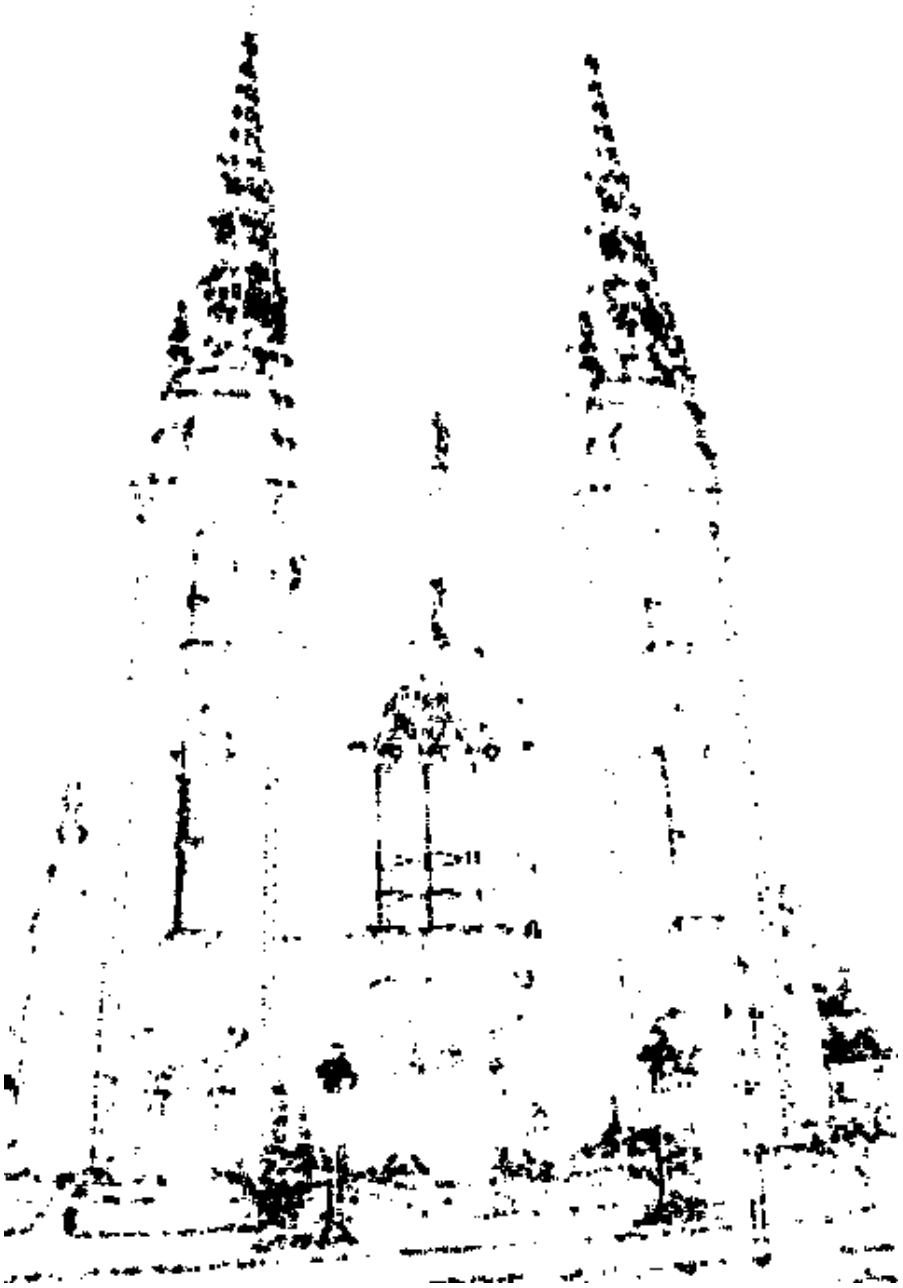

Figura 3-3: Limiar aplicado aos coeficientes de Harris, só os pontos com valores acima do limiar são candidatos à *features*

Faz-se necessário, portanto, mais um passo para eliminar os *corners* muito próximos entre si, selecionando os pontos com coeficiente mais forte. Para tal, após identificar os corners que satisfazem o valor de limiar, os coeficientes são ordenados de acordo com seu valor, e então uma distância mínima entre as *features* é forçada, gerando os pontos de interesse finais (Figura 3-4).

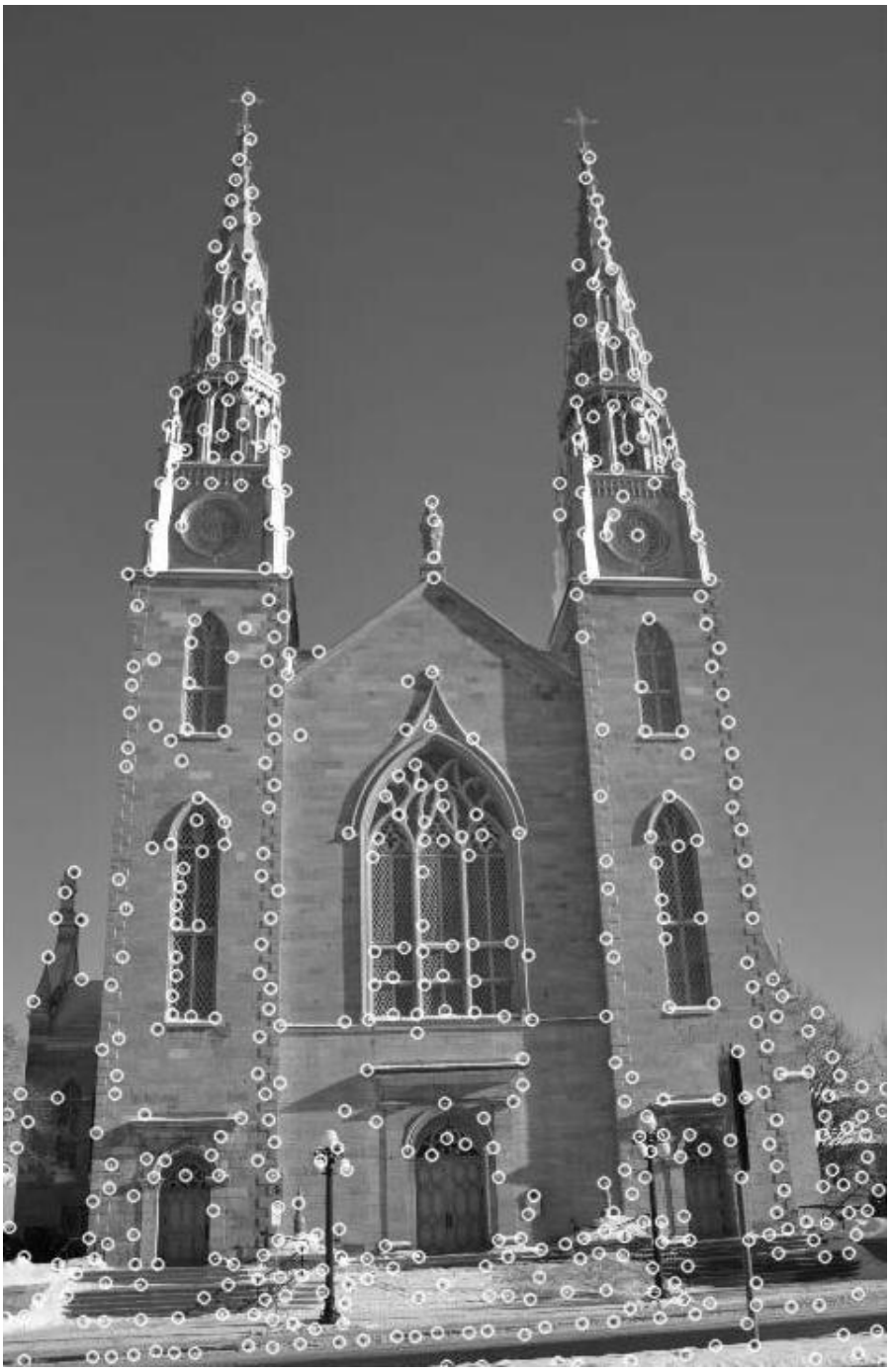

Figura 3-4: Features finais selecionadas na imagem

4 Metodologia

Este trabalho tem como objetivo estudar como os detectores de canto, em específico o *Harris Corner Detector* (coeficiente de Harris) e os autovalores, se comportam em relação à Lei do Dígito Significativo em imagens digitais. Devido ao fato de que não existe um teste de conformidade à lei de Benford universalmente aceito, nós primeiro investigamos como a análise de conformidade é realizada nos principais trabalhos relacionados à NB-lei com foco nos trabalhos realizados na área de imagens digitais, depois o método adotado é explicado em detalhes, e finalmente expomos nossa implementação dos experimentos.

4.1 Estado da Arte das Metodologias de Aderência à NB-Lei

Benford (BENFORD, 1938) foi o primeiro pesquisador a analisar a conformidade da lei do dígito mais significativo a dados reais. Ao analisar vinte grupos distintos de dados contendo um total de 20.229 observações, Benford utilizou-se da distância entre as frequências esperadas e as observadas para ordenar os grupos analisados. A Figura 4-1 apresenta uma cópia da tabela principal do artigo de Benford:

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

Figura 4-1: Tabela apresentada no trabalho original de Benford, onde são apresentados os percentuais encontrados do primeiro dígito para cada grupo analisado

A menor distância calculada por Benford teve valor de 2.8% (em um grupo com 100 observações) e a maior distância foi de 35.4% (em um grupo com 91 observações). A média de todos os grupos obteve distancia de 3%. Benford não chegou a classificar quais dos grupos eram considerados conformes com a lei ou não, não gerando um limiar de comparação, no entanto, a literatura que se seguiu, baseando-se nesta tabela, considera que diversos desses grupos são grandezas conformes à Lei, como população, áreas de rios, itens de jornais, etc. A Figura 4-2 mostra as diferenças totais (de todos os dígitos) entre as freqüências observadas e as esperadas para as grandezas destacadas por Benford:

SUMMATION OF DIFFERENCES BETWEEN OBSERVED AND THEORETICAL FREQUENCIES

Order	Item	Nature	Difference	Order	Item	Nature	Difference
1	D	Newspaper Items	2.8	11	N	Cost Data, Concrete	12.4
2	F	Pressure Lost, Air Flow	3.2	12	S	$n^1 \cdots n^8, n!$	13.8
3	G	H.P. Lost in Air Flow	4.8	13	L	Design Data Generators	16.6
4	R	Street Addresses, A.M.S.	5.4	14	B	Population, U. S. A.	16.6
5	P	Am. League, 1936	6.6	15	I	Drainage Rate of Rivers	21.6
6	Q	Black Body Radiation	7.2	16	K	$n^{-1}, \sqrt{n} \cdots$	22.8
7	O	X-Ray Voltage	7.4	17	H	Molecular Wgts.	23.2
8	M	*Readers' Digest*	8.4	18	E	Specific Heats	24.2
9	A	Area Rivers	9.8	19	C	Physical Constants	34.9
10	T	Death Rates	11.2	20	J	Atomic Weights	35.4

Figura 4-2: Tabela apresentada no trabalho original de Benford onde os grupos são organizados de acordo com sua conformidade à NB-lei de acordo com a diferença entre as frequências observadas e esperadas

Carslaw (CARSLAW, 1988) foi o precursor do uso da lei de Newcomb-Benford para a validação de dados contábeis. Ao analisar a declaração de renda de diversas empresas da Nova Zelândia, Carslaw se utilizou da estatística Z para avaliar individualmente a frequência de cada digito e realizou o teste χ^2 como teste global analisando a conformidade da distribuição inteira à NB-lei. O grupo de dados analisado por Carslaw continha 805 observações retiradas das declarações de 220 empresas no período de 1981 a 1985. A Figura 4-3 apresenta a tabela divulgada no artigo por Carslaw contendo os dados da análise para o primeiro dígito.

TABLE 1

FREQUENCY OF FIRST DIGITS IN INCOME NUMBERS

($n = 805$)

		Ordinary Income		*Net Income*	
Digit	*Expected Frequency[a] Percent*	*Observed Deviation Percent*	*Z-statistic*	*Observed Deviation Percent*	*Z-statistic*
0	—	—	—	—	—
1	30.1	+2.5	+1.54	−0.9	−0.56
2	17.6	−2.8	−2.09*	−2.3	−1.76
3	12.5	+1.1	+0.94	+2.3	+1.97*
4	9.7	+0.5	+0.48	+0.5	+0.48
5	7.9	−1.1	−1.16	−0.9	−0.95
6	6.7	−0.1	−0.11	+0.7	+0.91
7	5.8	−0.2	−0.24	−0.6	−0.73
8	5.1	+0.2	+0.26	+0.6	+0.77
9	4.6	−0.1	−0.14	+0.6	+0.81
	100.0	—		—	
χ^2		=7.64		9.33	

[a] Computed via Feller's formula.

* Significant at the .05 percent level.

Figura 4-3: Tabela apresentada no trabalho de Carslaw, onde são sumarizados o desvio entre as probabilidades esperadas e obtidas no primeiro dígito, juntamente com as estatísticas-Z para cada dígito e o valor do teste χ^2

Carslaw focou sua análise nas ocorrências do segundo dígito em específico, verificando uma discrepância entre o numero de zeros e noves esperados pela NB-lei e os reportados pelas empresas, como mostra a Figura 4-4. Carslaw concluiu que havia uma anormalidade nos dados reportados pelas empresas, sendo encontrada particularmente uma manipulação artificial de arredondamento, visando gerar valores no formato $n10^k$, considerados de maior valor cognitivo.

TABLE 4

FREQUENCY OF SECOND DIGITS FOR INCOME NUMBERS FOR OWNER AND MANAGER CONTROLLED FIRMS

		Owner Controlled Firms					
		Domestic ($n=244$)		*Overseas* ($n=126$)		*Manager Controlled Firms* ($n=434$)	
Digit	*Expected Frequency Percent*	*Observed Deviation Percent*	*Z-statistic*	*Observed Deviation Percent*	*Z-statistic*	*Observed Deviation Percent*	*Z-statistic*
0	12.0	+9.3	+4.47***	−0.5	−0.18	+4.1	+2.63**
1	11.4	−1.1	−0.54	−0.6	−0.22	+0.6	+0.39
2	10.9	+0.6	+0.30	−4.7	−1.72*	−0.5	−0.33
3	10.4	−1.4	−0.71	+2.7	+1.01	+0.2	+0.14
4	10.0	−2.2	−1.15	−1.5	−0.57	−1.2	−0.83
5	9.7	−2.3	−1.22	+2.6	+1.00	−1.9	−1.34
6	9.3	+3.0	+1.61	−3.9	−1.53	+1.8	+1.30
7	9.0	+0.8	+0.44	+2.4	+0.96	−0.5	−0.37
8	8.8	−3.2	−1.93*	+1.2	+0.48	+0.6	+0.44
9	8.5	−3.5	−1.79*	+2.3	+0.94	−3.2	−2.39**
	100.0	—		—		—	
χ^2		=29.77**		8.80		15.45	

* Significant at the .10 level.
** Significant at the .05 level.
*** Significant at the .01 level.

Figura 4-4: Tabela apresentada no trabalho de Carslaw, onde são analisadas as frequências encontradas para o segundo digito, e é encontrada uma discrepância particularmente alta nos dígitos nove e zero pelo teste Z

Na aplicação da NB-lei a imagens digitais, Jolion (JOLION, 2001) aplicou o teste estatístico de Kolmogorov-Smirnov (K-S) a uma base de 221 imagens, avaliando a conformidade dos valores de cinza dos pixels, a magnitude dos gradientes, além da decomposição piramidal baseada na transformada de Laplace, como é exposto na Figura 4-5. Nenhuma informação é apresentada sobre as imagens utilizadas nos experimentos, não sendo fornecido o numero médio de observações extraídas das imagens.

Image nature	Min	Max	Mean	Standard deviation
Gray level	0.0036	1	0.455	0.381
Gradient	0	1	0.983	0.105
Laplacian (level 0)	0	1	0.9549	0.146
Laplacian (level 1)	0.0292	1	0.9663	0.124
Laplacian (level 2)	0.0706	1	0.9847	0.089

Figura 4-5: Tabela apresentada no trabalho de Jolion: Adequações entre a NB-Lei e os valores da distribuição das imagens, seus gradientes e da decomposição piramidal baseada na transformada de Laplace (níveis 0, 1 e 2). Os valores mostrados na tabela são os níveis de significância do teste Kolmogorov-Smirnov

Sanches e Marques (SANCHES; MARQUES, 2006) também utilizaram o teste K-S para avaliar a qualidade de adequação do gradiente de imagens médicas à lei do primeiro

dígito. A análise foi realizada em três grupos de imagens distintos: ressonâncias magnéticas, tomografias computadorizadas e ultrassons, totalizando 476 imagens, no entanto a quantidade de observações por imagem também não é informada. Dois testes K-S foram realizados para cada grupo, calculando a medida de adequação à distribuição Benford e a uma distribuição uniforme. Na Figura 4-6 nós vemos a tabela contendo as médias geométricas das probabilidades de rejeição P_e da hipótese nula de que a distribuição segue a distribuição Benford/Uniforme.

P_e	Benford	Uniform
CT	$7.06E-14$	$4.41E-01$
MRI	$1.06E-11$	$6.84E-01$
US	$2.38E-13$	$2.34E-01$

Figura 4-6: Tabela apresentada no trabalho de Sanches e Marques onde são apresentadas as médias geométricas das probabilidades de rejeição P_e de acordo com o teste estatístico de Kolmogorov-Smirnov, computadas sobre 476 imagens

Acebo e Sbert (ACEBO; SBERT, 2005) avaliaram a conformidade de valores de pixels de imagens sintéticas à NB-lei, propondo o mensuramento da qualidade de adequação à lei de Benford de acordo com o que chamam de divergência χ^2. Eles, aparentemente inadvertidamente, aplicaram os valores de frequência ao teste χ^2 ao invés dos valores de observações obtidos, modificando o teste e removendo o seu valor estatístico, porém tornando-o em uma medida independente do tamanho da amostra. Um total de 18 imagens são analisadas e apresentadas diretamente no trabalho juntamente com a divergência χ^2 calculada, no entanto, o número de observações extraídas das imagens não é apresentado. A Figura 4-7 apresenta uma das análises realizada no trabalho.

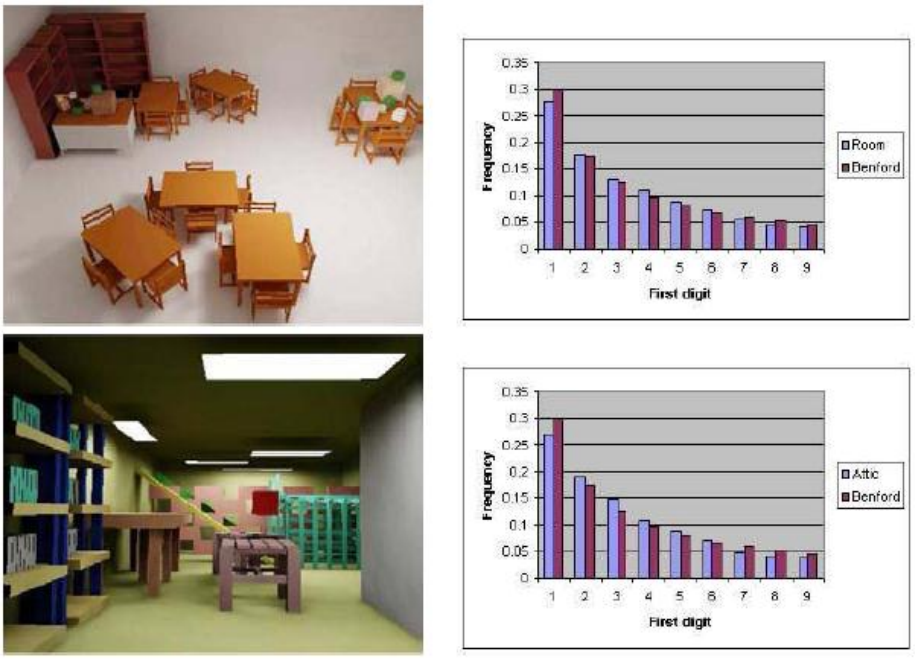

Figura 4-7: Figuras apresentadas no trabalho de Acerbo e Sbert. Na esquerda são apresentadas duas imagens sintéticas geradas através de radiosidade; no meio é apresentado o gráfico da adequação dos valores de pixel à NB-Lei para o primeiro dígito. As imagens obtiveram respectivamente uma divergência χ^2 de 0.00703 e 0.01549.

Fu, Shi e Su (FU; SHI; SU, 2007) utilizam-se do mesmo método de qualidade de adequação proposto por Acebo e Sbert para analisar a conformidade dos coeficientes do bloco-DCT em imagens. Fu, Shi e Su utilizaram como base de imagens a UCID, a mesma base que utilizamos em nossos experimentos e descrita em detalhes na Seção 4.3.2. No trabalho é apresentado um gráfico sintetizando a adequação do bloco-DCT à NB-lei, mostrado na Figura 4-8

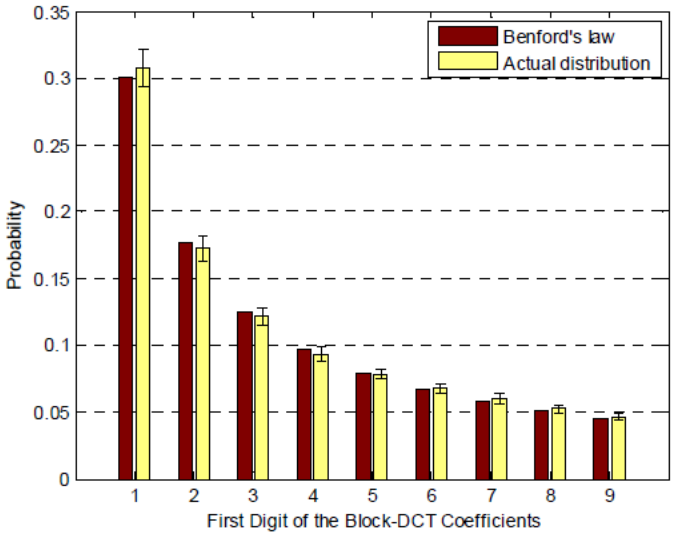

Figura 4-8: Gráfico apresentado no trabalho de Fu, Shi e Su, onde é apresentada a média das frequências encontradas para o primeiro digito do bloco-DCT para as 1.338 imagens. As barras vermelhas são as frequências esperadas pela lei de Benford, as barras amarelas representam a média obtida, contendo barras de erro representando o desvio padrão.

Qadir et al. em seus dois trabalhos (QADIR; ZHAO; HO, 2010)(QADIR; ZHAO; HO; CASEY, 2011) também se utilizam da divergência χ^2 como medida de qualidade de adequação à lei do dígito significativo. Eles também realizaram seus experimentos em cima da base de imagens UCID, verificando a conformidade da transformada discreta de wavelet (DWT) em imagens à Lei de Newcomb-Benford. Assim como Fu, Shi e Su, eles também apresentaram um gráfico sintetizando a adequação à NB-lei para o primeiro dígito, como é mostrado na Figura 4-9:

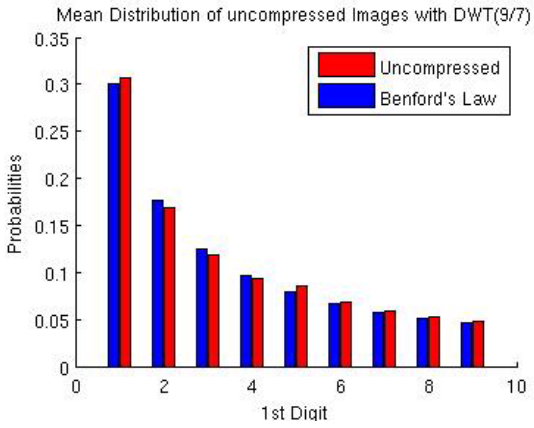

Figura 4-9: Gráfico apresentado no trabalho de Qadir et al.: em azul estão os valores esperados pela lei de Benford e em vermelho está representada a média encontrada dos primeiros dígitos da DWT sem compressão das 1.338 imagens contidas na UCID.

Do exposto acima tem-se uma clara ideia de que a demonstração da conformidade à NB-Lei de certas grandezas extraídas de imagens digitais, ou mesmo de outros modelos do mundo real, é apenas experimental, seja para a comunidade de Processamento de Imagens ou para outras áreas aplicadas, por várias razões. Entre elas, que a expressão probabilística que define a NB-Lei resulta em valores irracionais para elementos que não são potências de 10, o que de início, invalida o rigorismo numa afirmação de conformidade em sequências finitas. Outra razão é que ainda não existem estatísticas ou métricas que sejam plenamente adequadas para este fim, seja por limitações na sua aplicabilidade, como no caso do teste χ^2 de Pearson para amostras muito grandes, ou na ausência de um intervalo de confiança, como no caso das diversas outras métricas, inclusive as que atuam diretamente nas mantissas.

Uma outra alternativa seria utilizar um modelo matemático gerador das amostras e provar analiticamente que este modelo satisfaz a equação probabilística da NB-Lei, mesmo que assintoticamente. No entanto, os modelos disponíveis na literatura para imagens digitais são muito simplificados, não satisfazendo nem de longe a generalidade que se pretende neste trabalho. Assim sendo, a modelagem estatística ainda é a que apresenta mais recursos para a verificação e a validação da conformidade, tendo em vista o caráter aleatório das cores que os pixels podem assumir numa imagem qualquer.

As exposições da conformidade à NB-Lei nos artigos da área sugerem que seus autores se valem da inspeção visual sobre gráficos e planilhas para concluírem que suas grandezas são conformes. Gráficos de barras das frequências dos dígitos na primeira posição onde há visualmente "pequena diferença" nas alturas das barras são os recursos mais comuns utilizados nesta literatura, constituindo-se talvez na aplicação de uma metodologia empregada pelo próprio Frank Benford em seu artigo seminal. Em muitos casos nem mesmo investigam o grau de aleatoriedade, independência ou adequação do tamanho das amostras para consubstanciarem os resultados ou mesmo para inferirem qual é a proporção de ruídos presentes nos dados.

Diante deste cenário, propusemos aqui uma metodologia que empregou as principais técnicas utilizadas na literatura específica e que propiciaram uma comparação direta de conformidade do coeficiente de Harris e dos autovalores da matriz de Harris contra a conformidade das outras grandezas dessa literatura, comparação tal que se mostrou favorável às primeiras.

4.2 Procedimento adotado

Nosso objetivo é avaliar como os detectores de canto, em particular o coeficiente de Harris e os autovalores da matriz de covariância, se comportam em relação à Lei do Dígito Significativo. Para tal, nós compilamos na seção anterior a metodologia utilizada pela literatura com foco na aplicação em imagens. Percebemos que não existe um método único difundido, e que o método mais utilizado na análise de imagens, a divergência χ^2, é uma aplicação inusitada do teste χ^2 de Pearson, sem valor estatístico, como exposto anteriormente.

Testes estatísticos são desenvolvidos para trabalhar com amostras de uma população. Como em nosso caso nós temos acesso a toda população, ou seja, todos os valores de uma grandeza em uma imagem, estes testes tem seu valor reduzido para nosso uso. Nigrini (NIGRINI, 1999) classifica esse fenômeno como excesso de poder: quando o tamanho dos dados, ou seja, a quantidade de observações a serem avaliadas, se torna cada vez maior, os testes estatísticos se tornam cada vez mais rigorosos; segundo Nigrini, a partir de 10.000 observações, até pequenas diferenças se tornam significantes; para grandes grupos de dados (acima de 1.000.000 de observações), diferenças imperceptíveis em um gráfico são consideradas significantes, rejeitando a hipótese nula. O mesmo fenômeno é relatado por Krakar (KRAKAR; ZGELA, 2009) ao descrever o teste χ^2. No texto, é dito que para conjuntos de dados com mais de 10.000 amostras o valor da estatística é quase sempre superior ao valor crítico, fazendo com que o auditor imagine que o conjunto não seja conforme com a NB-Lei. Os autores em (LUQUE; LACASA, 2009) citam novamente o problema de excesso de poder com os teste Z e χ^2. Este fenômeno afeta os testes Z, χ^2, e Kolmogorov-Smirnov.

A base de imagens que utilizamos, UCID, possui tamanho de imagens de 512x384 pixels, gerando um total de 196.608 possíveis observações para o coeficiente de Harris e o dobro deste valor para os autovalores. Este tamanho de dados é significantemente alto, afetando a utilidade dos testes estatísticos mencionados.

Portanto, devido às limitações individuais de cada teste, a falta de consenso quanto a um teste universal difundido, e por motivos de comparação com outros trabalhos, nós aplicamos diversos métodos para avaliar a conformidade de nossos experimentos à Lei de Newcomb-Benford:

- Diferença entre as frequências observadas e esperadas ε (Equação 2.5), utilizada no trabalho original de Benford;
- Teste estatístico χ^2 de Pearson, utilizado por Carslaw e diversos outros autores (Equação 2.7). Este teste, porém, é suscetível ao alto tamanho de observações extraídas das imagens, como foi previamente discutido;
- Divergência χ^2 proposta por Acerbo e Sbert e difundida na aplicação da Lei de Benford à imagens, onde as frequências são aplicadas diretamente à formula χ^2 (Equação 2.7), no lugar do valor de observações;
- Teste estatístico Kolmogorov-Smirnov (PETITT, STEVENS, 1977), utilizado por Jolion, Sanches e Marques na análise da NB-Lei em imagens. Assim como o teste χ^2 de Pearson, o teste Kolmogorov-Smirnov é suscetível ao tamanho da amostra;

Nós também fizemos uso extensivo de gráficos para melhor compreender a conformidade da lei de Newcomb-Benford dos coeficientes de Harris e os autovalores em imagens digitais.

4.3 Experimentos e Implementação

Nós avaliamos a conformidade da Lei de Newcomb-Benford para três grandezas:

- Coeficiente de Harris;
- Ambos autovalores da matriz de covariância;
- Mínimo autovalor da matriz de covariância;

Para cada grandeza nós examinamos a conformidade à Lei em três analises distintas:

- Analise do primeiro dígito, assim como é realizada em praticamente todos os trabalhos relacionados à NB-Lei. As probabilidades esperadas para o primeiro dígito são obtidas via a Equação 2.2 e são apresentadas na Figura 2-1;
- Análise do segundo dígito, como no trabalho de Carslaw, porém inédita na área de Processamento de Imagens. As probabilidades esperadas para o segundo dígito, por sua vez, podem ser calculadas via a Equação 2.3;
- Análise dos dois primeiros dígitos, também inédita na área de Processamento de Imagens. As probabilidades esperadas para os dois primeiros dígitos podem ser calculadas pela Equação 2.4;

Dessa forma, desenvolvemos um total de nove análises.

4.3.1 Implementação

Nossa implementação está dividida em duas partes: a primeira, implementada utilizando a linguagem de computação C++ e a biblioteca OpenCV de Visão Computacional, calcula o valor dos autovalores da matriz de covariância e do Coeficiente de Harris para cada pixel da imagem de entrada e então salva os valores de mantissa em um arquivo para futura análise. A segunda, implementada em MATLAB, por sua vez, realiza as análises de conformidade à NB-lei, gerando os quatro índices de qualidade de adequação explicitadas na seção anterior, assim como os gráficos para análise visual.

O cálculo das grandezas analisadas: Coeficiente de Harris, autovalores da matriz de covariância, e mínimo dos autovalores foram calculados respectivamente pelas seguintes funções e parâmetros da biblioteca OpenCV:

void cornerHarris(const Mat& src, Mat& dst, int blockSize=3, int apertureSize=3, double k=0.04, int borderType=BORDER_DEFAULT);

void cornerEigenValsAndVecs(const Mat& src, Mat& dst, int blockSize=3, int apertureSize=3, int borderType=BORDER_DEFAULT);

void cornerMinEigenVal(const Mat& src, Mat& dst, int blockSize=3, int apertureSize=3, int borderType=BORDER_DEFAULT);

Estes parâmetros utilizados são os parâmetros padrão (*default)* da biblioteca. Nós experimentamos com outros valores para os parâmetros '*blockSize', 'apertureSize',* e *'k',*

porém não encontramos mudanças significativas nos resultados obtidos. Após o cálculo dos respectivos valores, nós extraímos os valores de mantissa, inclusive de valores negativos, e geramos a tabela de observações/probabilidades; as amostras que assumem valor ‘0’ são desconsideradas, visto que zeros não sendo tratados pela NB-Lei, o que gera uma pequena diferença no tamanho das amostras extraídas de cada imagem.

4.3.2 Base de Imagens

Em todos os experimentos nós utilizamos a base de imagens UCID (*Uncompressed Image Database*) (SCHAEFER; STICH, 2003), uma base de imagens sem compressão composta de 1.338 imagens no formato *.tif* com altura e largura de 512x384 pixels. As imagens desta base de dados são bem diversas entre si, como vemos na Figura 4-10, contendo cenas de paisagens, locais, e objetos diversos, sendo popularmente utilizadas pela comunidade de pesquisa de processamento de imagens. Vale ressaltar que o cômputo dos qualificadores de ponto de interesse na literatura é geralmente realizado com a imagem em escala de cinza, que é o caso de nossa implementação.

Figura 4-10: Exemplo de imagens contidas na base de imagens UCID

5 Análise dos resultados

Neste capítulo são apresentados os resultados dos experimentos previamente planejados na Seção 4.3. Para cada grandeza nós analisamos a conformidade à Lei de Newcomb-Benford para o primeiro dígito, para o segundo dígito e para os dois primeiros dígitos.

Os resultados de cada análise estão sumarizados em dois gráficos e uma tabela. O primeiro gráfico apresenta as frequências extraídas de cada uma das 1.338 imagens, totalizando 1.338 linhas em um único gráfico. O segundo, apresentado em formato de linha e de barras, expressa os valores mínimos, máximos, médios, desvio padrão, além das frequências esperadas pela NB-Lei para comparação, sumarizando o encaixe da grandeza à Lei. A tabela, por sua vez, apresenta o resultado dos quatro testes/medidas realizados, também incluindo mínimo, máximo, média e desvio padrão, além do número de imagens que foram aprovadas pelo teste com nível de significância 0.05 (válidos somente para o teste χ^2 de Pearson e para o teste Kolmogorov-Smirnov). Para todos os testes/medidas apresentados, quanto menor o valor obtido, maior é a conformidade à Lei do Dígito Significativo.

5.1 Coeficiente de Harris

Nesta seção iremos apresentar os resultados da análise da conformidade do coeficiente de Harris à NB-Lei.

5.1.1 Análise do primeiro dígito

Na Figura 5-1 visualizamos 1.338 linhas representando as probabilidades de todas as 1.338 imagens do primeiro dígito do coeficiente de Harris. No gráfico, percebemos que praticamente todas as linhas se concentram na mesma localidade, gerando um desvio-padrão inferior a meio por cento. Todas as imagens se assemelham em sua conformidade à NB-Lei para o primeiro dígito.

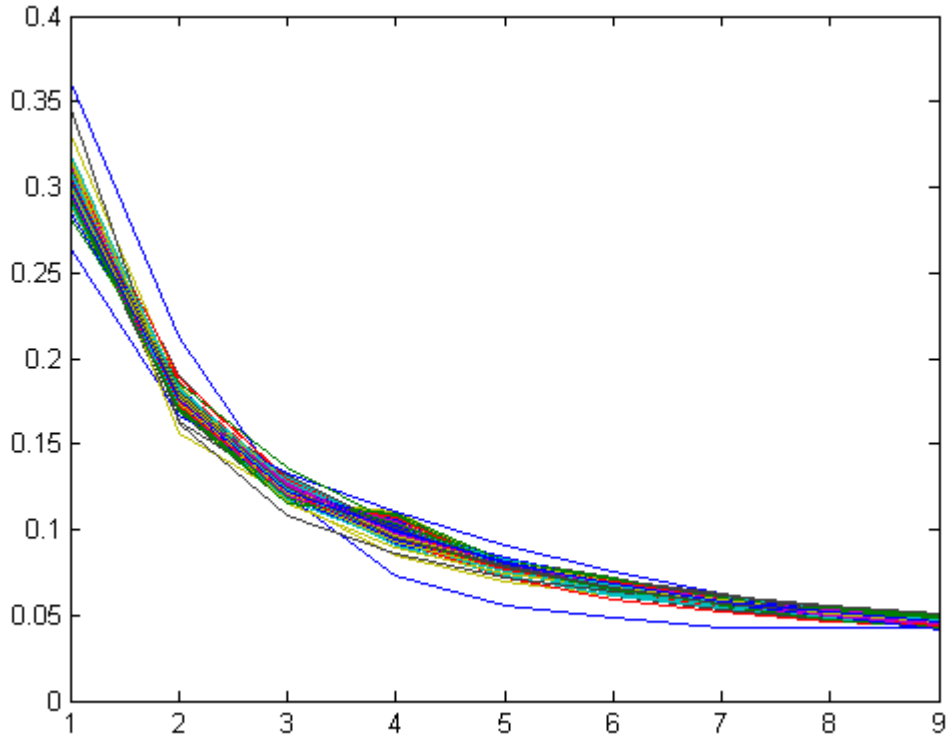

Figura 5-1: Probabilidades do primeiro dígito do coeficiente de Harris para cada uma das 1.338 imagens UCID

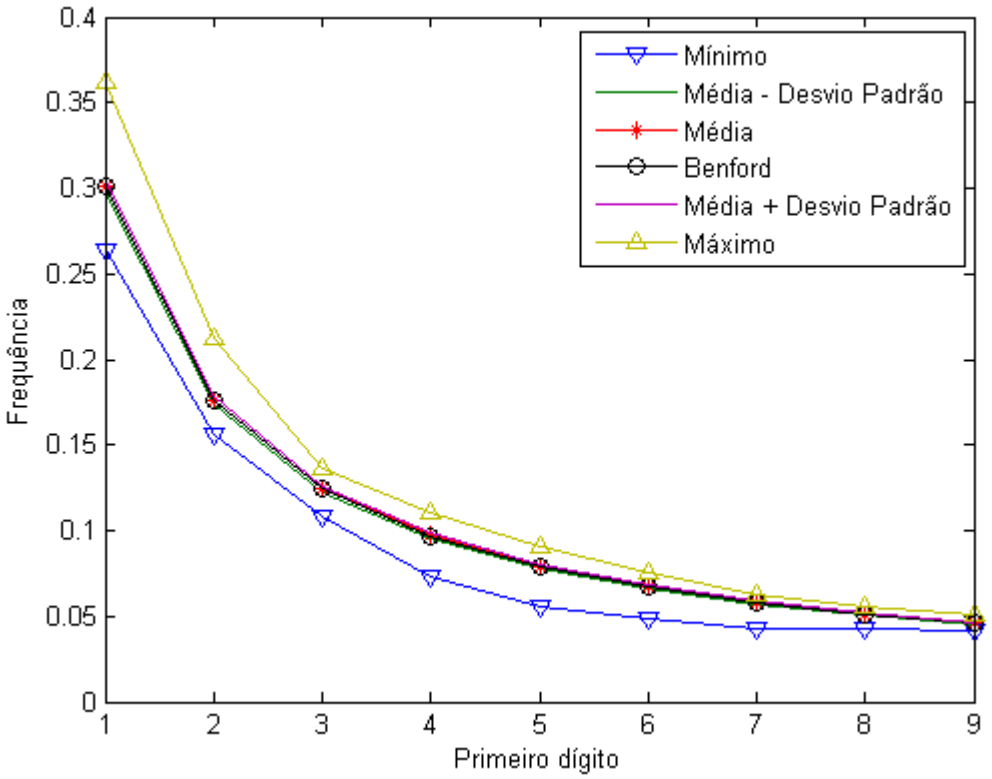

Figura 5-2: Síntese da conformidade do coeficiente de Harris à NB-Lei para o primeiro dígito

Nas Figura 5-2 e Figura 5-3, por sua vez, visualizamos a adequação do coeficiente de Harris à Lei. Percebemos que as probabilidades esperadas para a Lei de Benford (mostradas em preto), se encaixam perfeitamente com a média obtida das imagens (mostrada em vermelho) e inclusive com o desvio-padrão (em verde e roxo).

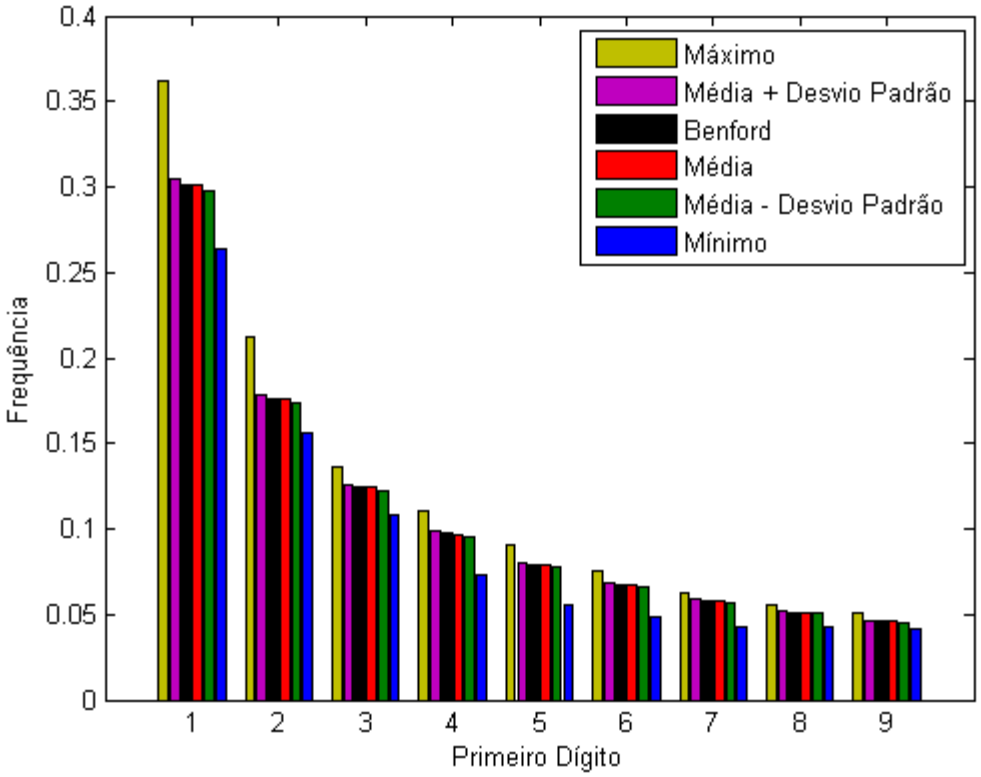

Figura 5-3: Síntese da conformidade do coeficiente de Harris à NB-Lei para o primeiro dígito (gráfico em barras)

Tabela 5-1: Análise da conformidade do coeficiente de Harris à NB-Lei para o primeiro dígito

Conformidade	Média	Desvio Padrão	Mínimo	Máximo	Aceitação
Distância ε	0.45%	0.45%	0. 069%	9.69%	N/A
χ^2 de Pearson	43.26	259.28	1.075	8473.71	741
Divergência χ^2	0.00022	0.0013	5.55e-06	0.043	N/A
K-S	0.0034	0.0043	0.00037	0.096	1061

Na Tabela 5-1 visualizamos a síntese dos testes/medidas de conformidade do coeficiente de Harris para o primeiro dígito. Percebemos que a média e o desvio padrão da diferença ε entre as probabilidades esperadas e obtidas são menores que meio por cento. Como podemos ver na tabela, os testes estatísticos χ^2 de Pearson e Kolmogorov-Smirnov, mesmo sendo extremamente sensíveis ao alto tamanho de

amostras utilizadas, aceitam respectivamente 55% e 79% das imagens como conformes, com nível de significância 0.05.

5.1.2 Análise do segundo dígito

Na Figura 5-4 podemos ver novamente 1.338 linhas representando as probabilidades do coeficiente de Harris para cada uma das imagens, desta vez para o segundo dígito. No gráfico, percebemos que a maioria das linhas estão situadas na mesma localidade, porém vemos que alguns dígitos específicos apresentam um maior desvio, em especial o dígito "5".

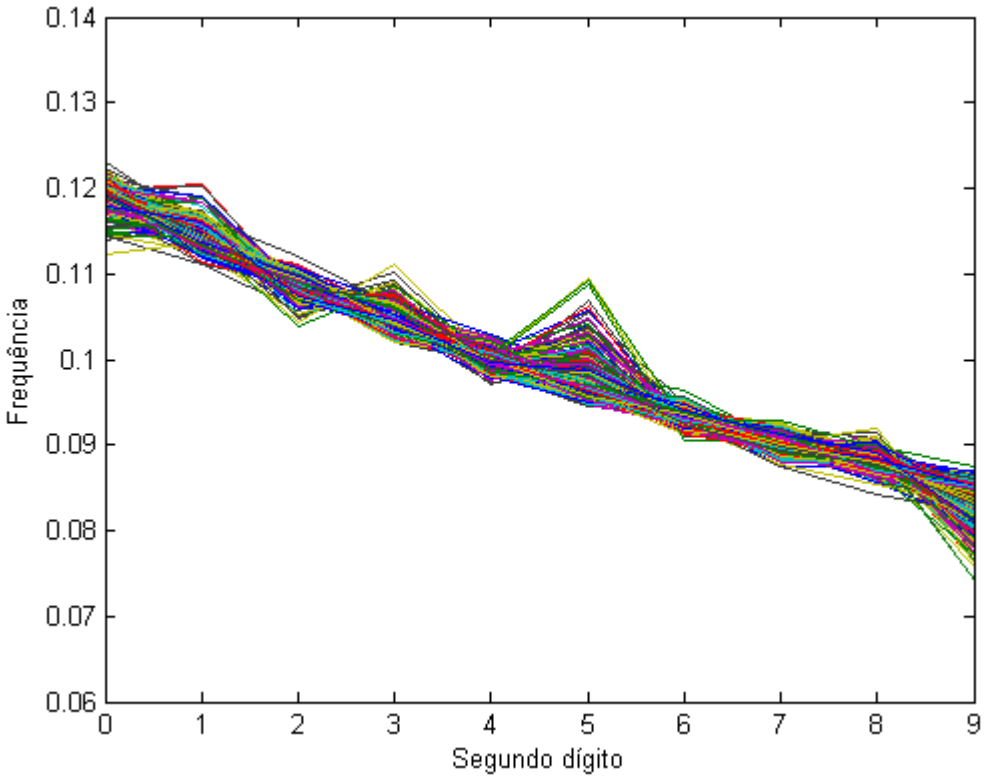

Figura 5-4: Probabilidades do segundo dígito do coeficiente de Harris para cada uma das 1.338 imagens UCID

É importante notarmos duas diferenças em comparação ao gráfico do primeiro dígito (Figura 5-1): a primeira, que o segundo dígito pode assumir o valor "0", havendo mais um grau de liberdade; a segunda, que como as probabilidades para o segundo dígito são mais uniformes do que as do primeiro dígito, as diferenças são mais ampliadas.

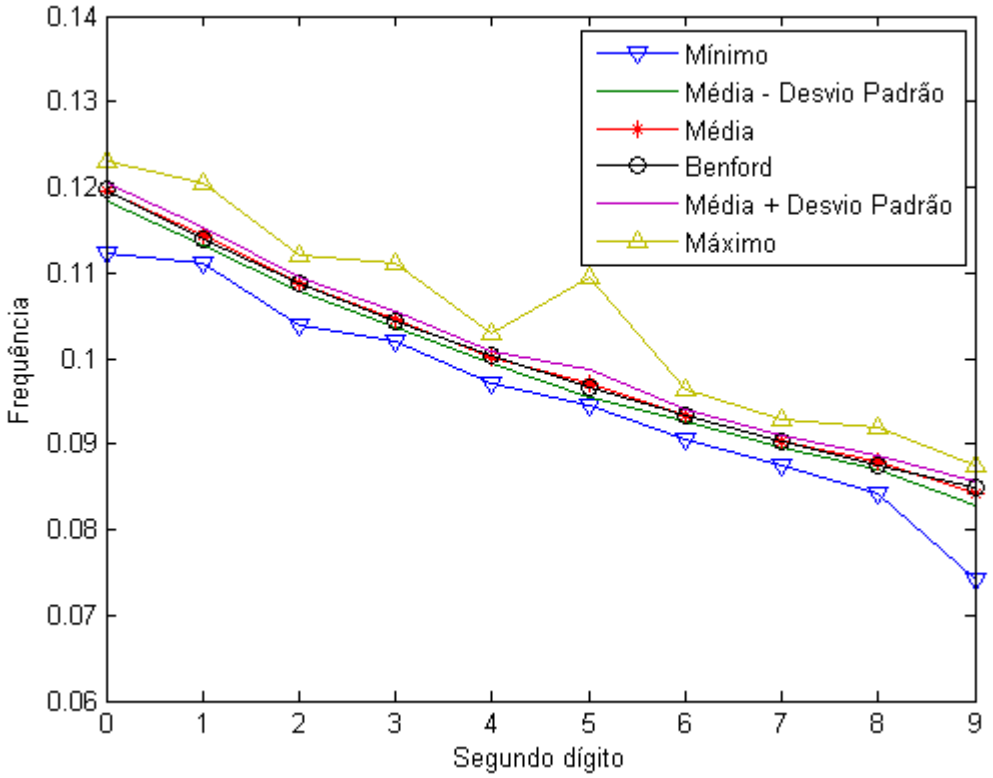

Figura 5-5: Síntese da conformidade do coeficiente de Harris à NB-Lei para o segundo dígito

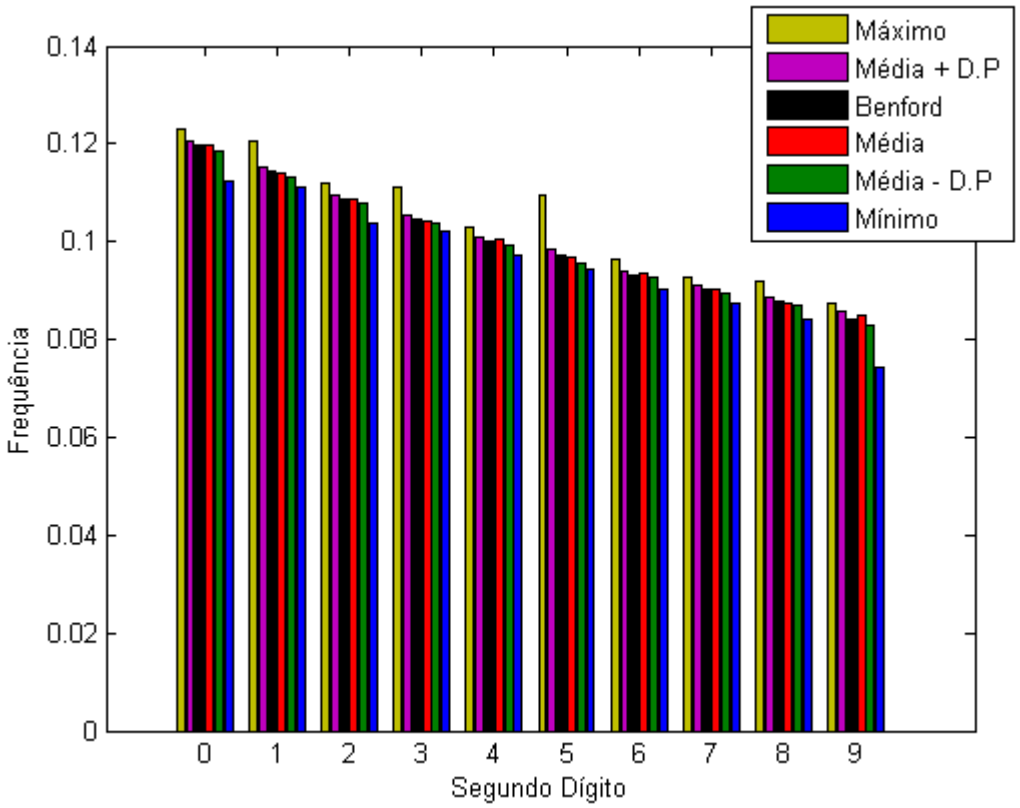

Figura 5-6: Síntese da conformidade do coeficiente de Harris à NB-Lei para o segundo dígito (gráfico em barras)

Nas Figura 5-5 e Figura 5-6 visualizamos a adequação do segundo dígito do coeficiente de Harris à Lei. Percebemos que o encaixe para o segundo dígito à Lei também é praticamente perfeito, sendo inclusive melhor do que para o primeiro dígito.

Tabela 5-2: Análise da conformidade do coeficiente de Harris à NB-Lei para o segundo dígito

Conformidade	Média	Desvio Padrão	Mínimo	Máximo	Aceitação
Distância ε	0.36%	0.23%	0.091%	2.22%	N/A
χ^2 de Pearson	22.88	49.67	0.98	624.37	1001
Divergência χ^2	0.00012	0.00028	5e-06	0.0037	N/A
K-S	0.00206	0.0013	0.00039	0.011	1266

Na Tabela 5-2 visualizamos a síntese dos testes/medidas de conformidade do coeficiente de Harris para o segundo dígito. Ao compararmos seus valores com os obtidos para o primeiro dígito (Tabela 5-1), concluímos que em todos os casos a conformidade do segundo dígito é melhor que a do primeiro. Os testes estatísticos χ^2 de Pearson e Kolmogorov-Smirnov, apesar de todas suas limitações técnicas que desfavorecem o julgamento para grandezas com alta quantidade de amostras, aceitam respectivamente 74.8% e 94.6% das imagens como conformes com nível de significância 0.05.

5.1.3 Análise dos dois primeiros dígitos

Na Figura 5-7 podemos ver mais uma vez que as linhas situam-se na mesma localidade, assim como para o primeiro (Figura 5-1) e para o segundo dígito (Figura 5-4). Também percebemos que algumas sequências específicas de dígitos possuem um desvio ainda mais exacerbado do que vimos para o segundo dígito, como é o caso em especial da sequencia “45”, que possui segundo dígito “5”, que como vimos na análise do segundo dígito também apresenta desvio maior.

Desta vez precisamos atentar ao fato de que os dois primeiros dígitos podem assumir 90 valores distintos, de “10” a “99”, e que as probabilidades são distribuídas entre eles, fazendo com que um erro alto em somente uma sequência seja mitigado pelas outras 89.

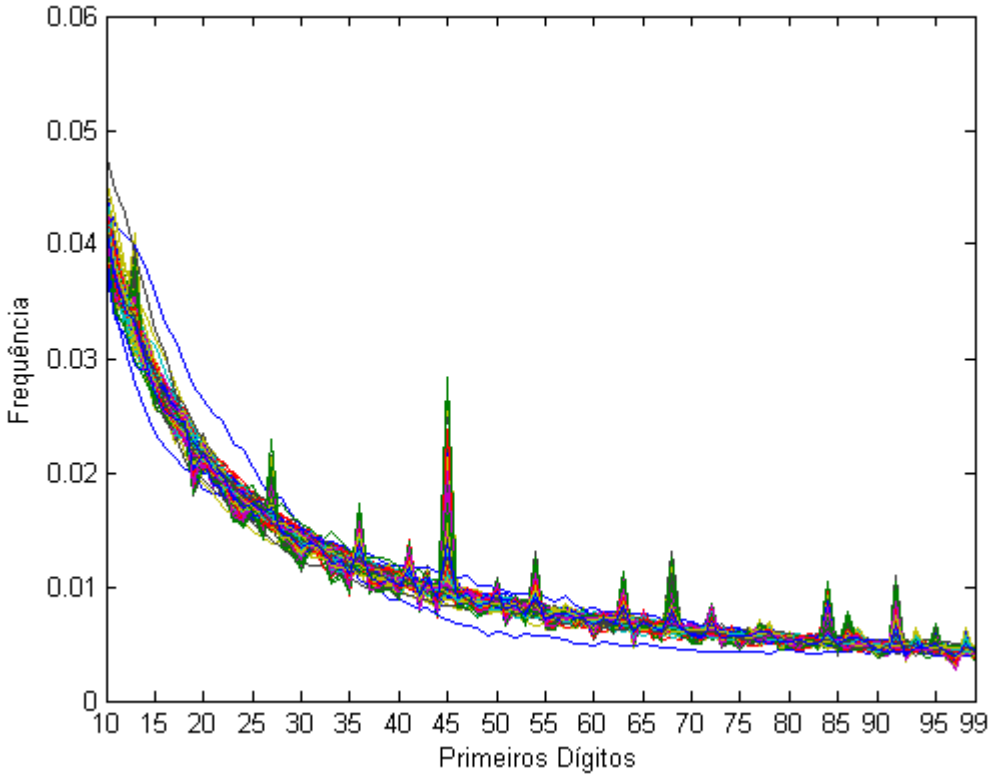

Figura 5-7: Probabilidades dos dois primeiros dígitos do coeficiente de Harris para cada uma das 1.338 imagens

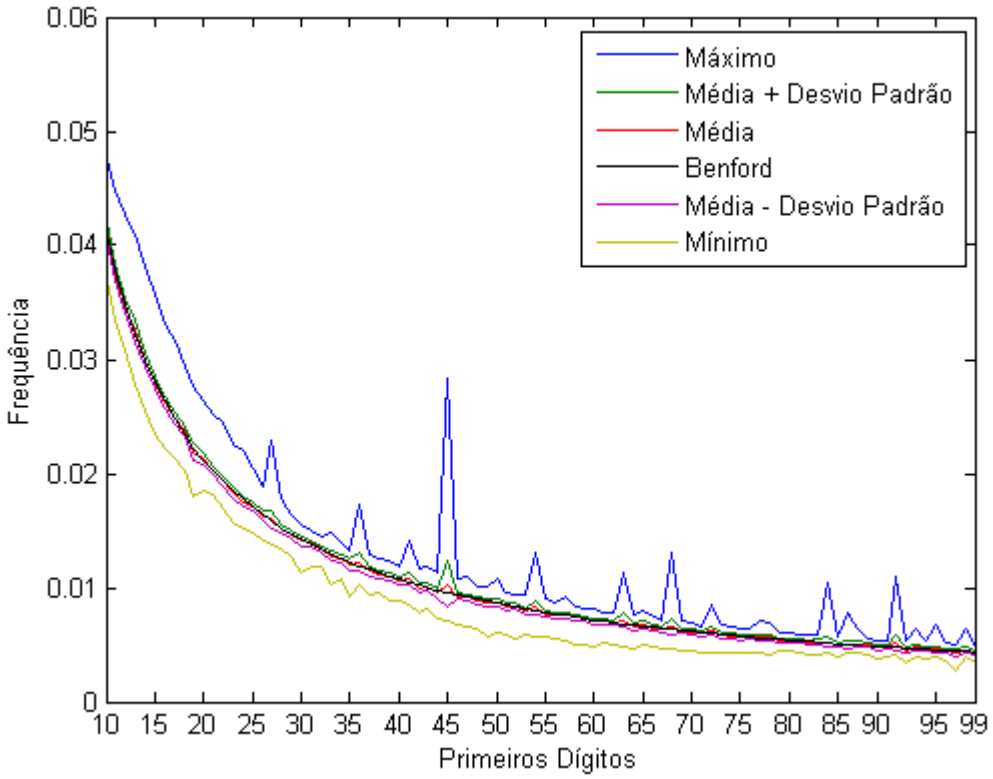

Figura 5-8: Síntese da conformidade do coeficiente de Harris à NB-Lei para os dois primeiros dígitos

Na Figura 5-8 visualizamos a adequação dos dois primeiros dígitos do coeficiente de Harris à Lei. Percebemos que o encaixe à Lei para os dois primeiros dígitos também é excelente, embora pior do que para o primeiro (Figura 5-2) e o segundo (Figura 5-5) dígito individualmente, apresentando maiores irregularidades.

Tabela 5-3: Análise da conformidade do coeficiente de Harris à NB-Lei para os dois primeiros dígitos

Conformidade	Média	Desvio Padrão	Mínimo	Máximo	Aceitação
Distância ε	1.23%	0.883%	0.6138%	9.9492%	N/A
χ^2 de Pearson	391.75	1091.87	55.56	12787.16	584
Divergência χ^2	0.0021	0.0062	0.00028	0.079	N/A
K-S	0.0039	0.0045	0.00078	0.099	1004

Na Tabela 5-3 visualizamos a síntese dos testes/medidas de conformidade do coeficiente de Harris para os dois primeiros dígitos. Ao compararmos seus valores com os obtidos para o primeiro (Tabela 5-1) e segundo dígito (Tabela 5-2), podemos inferir que a adequação à Lei de Benford para os dois dígitos é mais errática, porém, ainda assim, excelente. Os testes estatísticos χ^2 de Pearson e Kolmogorov-Smirnov, aceitam respectivamente 43.6% e 75% das imagens como conformes com nível de significância 0.05.

5.2 Ambos autovalores

Nesta seção iremos apresentar os resultados da análise da conformidade de ambos autovalores à NB-Lei. É importante ressaltar que a análise de ambos os autovalores considera dois valores para cada pixel, dobrando o tamanho da amostra analisada quando comparada ao tamanho da amostra do coeficiente de Harris e do mínimo dos autovalores.

5.2.1 Análise do primeiro dígito

Na Figura 5-9 visualizamos 1.338 linhas representando as probabilidades de todas as 1.338 imagens do primeiro dígito, desta vez para ambos os autovalores. No gráfico, percebemos que as 1.338 imagens se assemelham muito em sua conformidade à NB-Lei. Ao compararmos com o detector de Harris (Figura 5-1), percebemos que as conformidades dos autovalores são um pouco mais dispersas, gerando um desvio padrão um pouco maior.

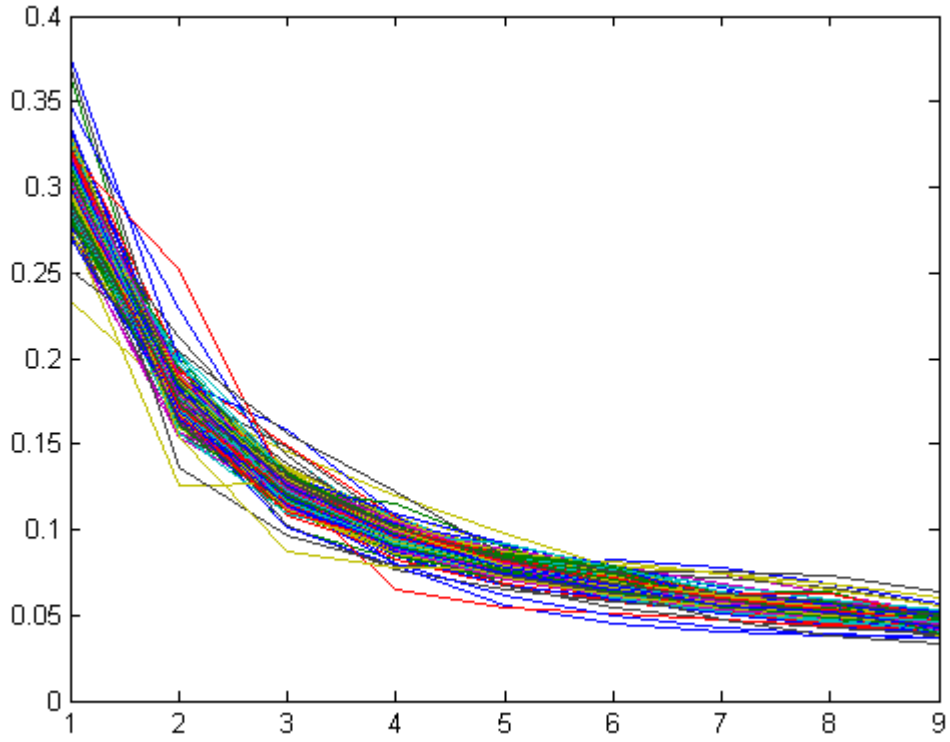

Figura 5-9: Probabilidades do primeiro dígito de ambos os autovalores para cada uma das 1.338 imagens UCID

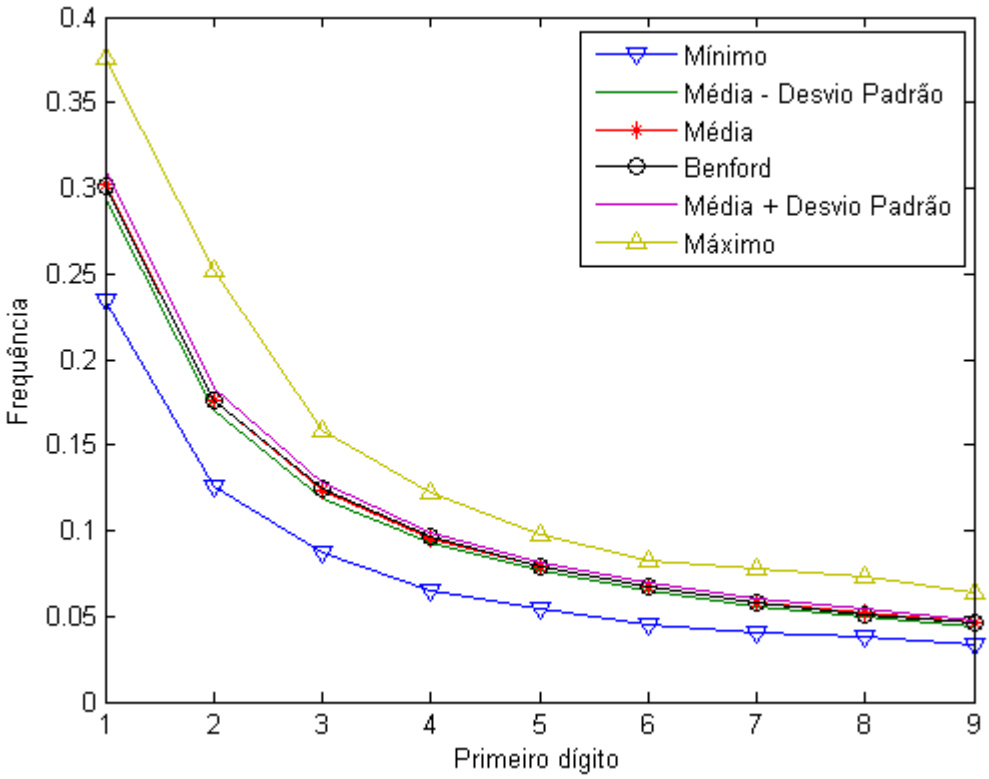

Figura 5-10: Síntese da conformidade de ambos os autovalores à NB-Lei para o primeiro dígito

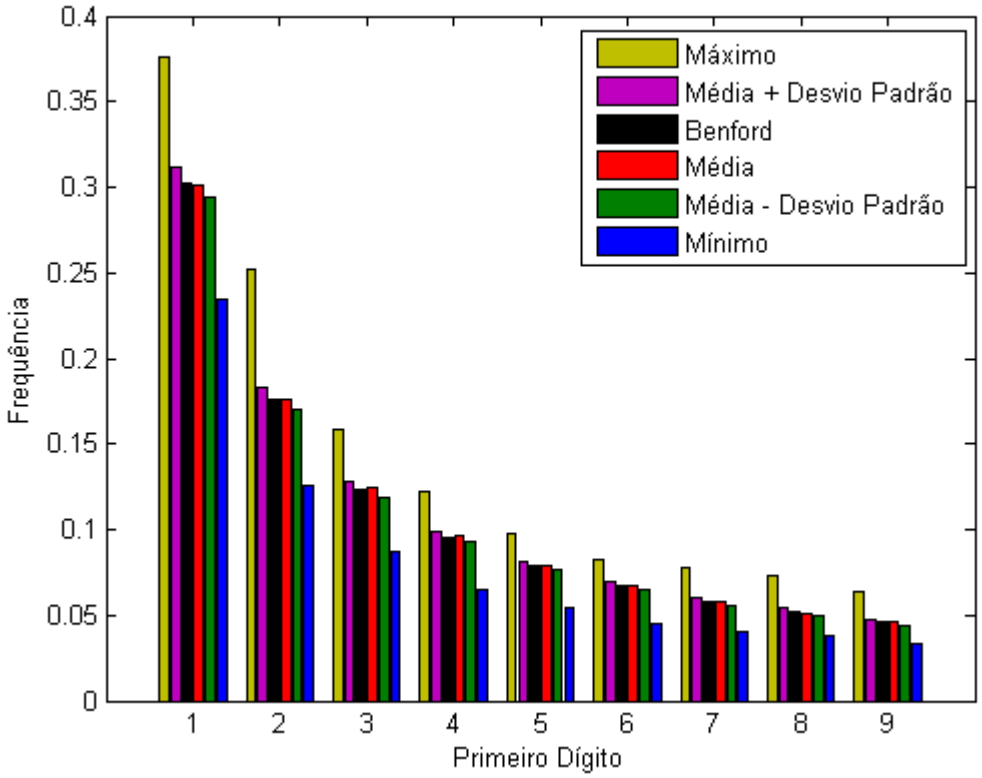

Figura 5-11: Síntese da conformidade de ambos os autovalores à NB-Lei para o primeiro dígito (gráfico em barras)

Nas Figura 5-10 e Figura 5-11, visualizamos a adequação de ambos autovalores à Lei. Percebemos novamente que, assim como no coeficiente de Harris (Figura 5-2 e Figura 5-3), as probabilidades esperadas para a Lei de Benford (mostradas em preto), se intersectam perfeitamente com a média obtida das imagens (mostrada em vermelho). Já o desvio padrão é ligeiramente maior se comparado ao coeficiente de Harris.

Tabela 5-4: Análise da conformidade de ambos os autovalores à NB-Lei para o primeiro dígito

Conformidade	Média	Desvio Padrão	Mínimo	Máximo	Aceitação
Distância ε	1.15%	1.023%	0.091%	10.28%	N/A
χ^2 de Pearson	485.49	1500.35	2.52	23737.18	53
Divergência χ^2	0.0012	0.0038	6.42e-06	0.0603	N/A
K-S	0.0094	0.00909	0.00037	0.102	191

Na Tabela 5-4 visualizamos a síntese dos testes/medidas de conformidade de ambos os autovalores para o primeiro dígito. Ao compararmos com o coeficiente de Harris (Tabela 5-1), percebemos que a conformidade de ambos os autovalores é ligeiramente inferior, mas que ainda é excelente. Como podemos comprovar na tabela, os testes

estatísticos χ^2 de Pearson e Kolmogorov-Smirnov, são sensíveis ao tamanho da amostra (para ambos os autovalores é o dobro), a aceitação cai vertiginosamente, aceitando respectivamente somente 3.9% e 14% das imagens como conformes, com nível de significância 0.05.

5.2.2 Análise do segundo dígito

Na Figura 5-12 podemos ver novamente as 1.338 linhas, desta vez representando as probabilidades do segundo dígito de ambos os autovalores para cada uma das imagens. No gráfico, percebemos que a maioria das linhas estão situadas na mesma localidade, e vemos também que alguns dígitos específicos apresentam um maior desvio, em especial o dígito "1" e o dígito "3".

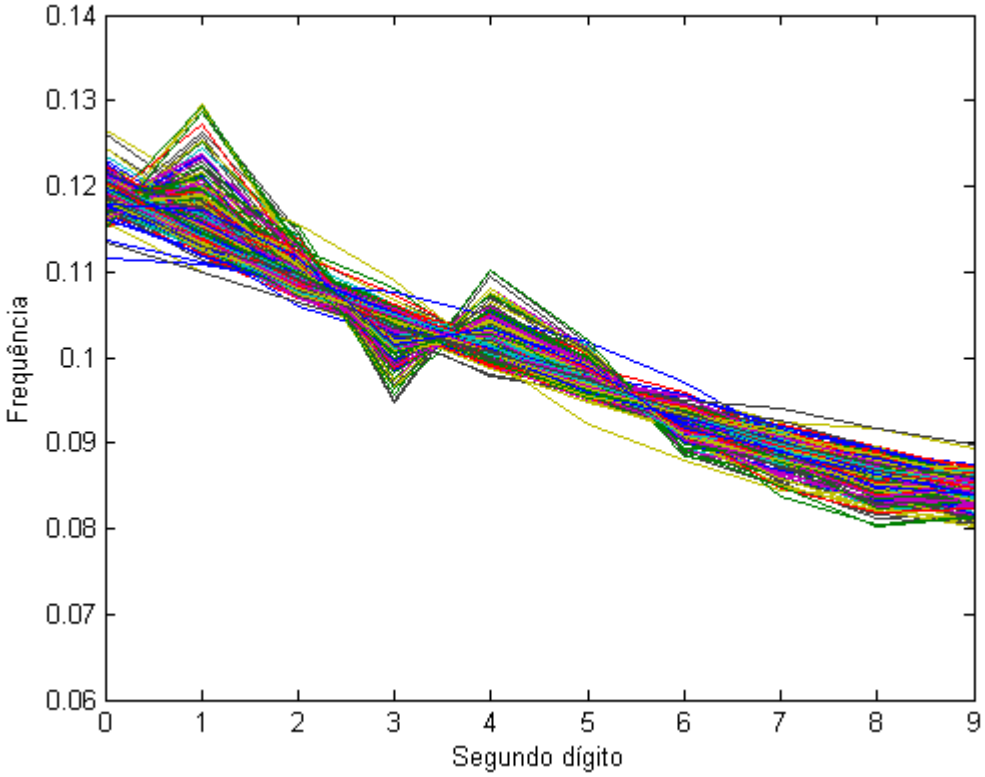

Figura 5-12: Probabilidades do segundo dígito de ambos os autovalores para cada uma das 1.338 imagens UCID

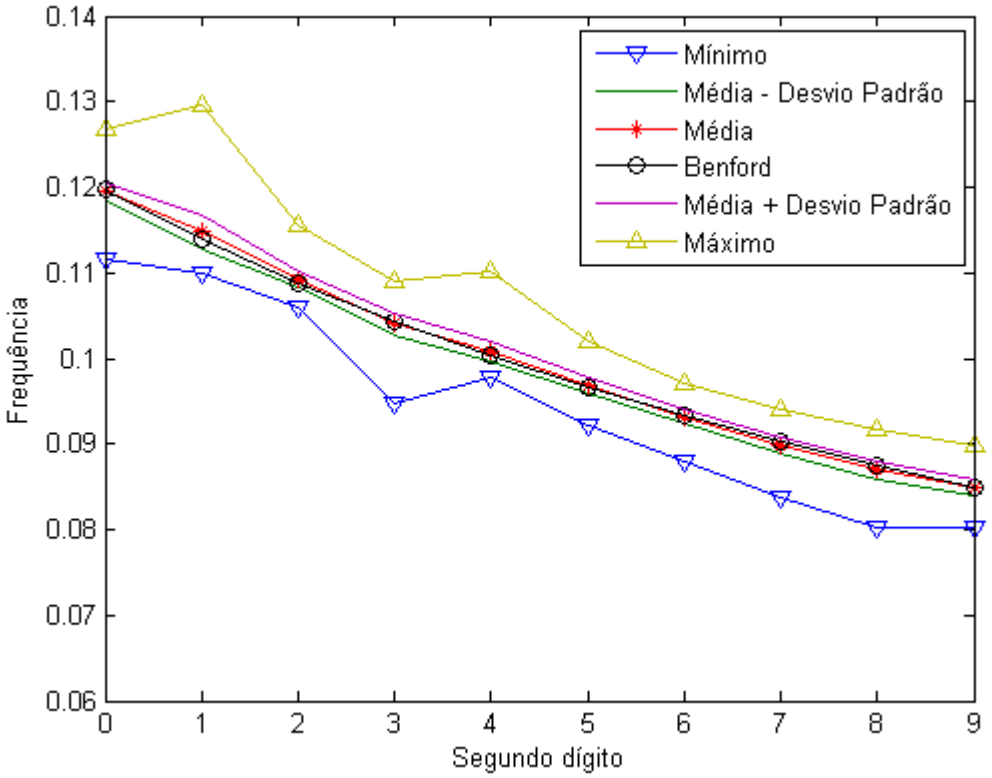

Figura 5-13: Síntese da conformidade de ambos os autovalores à NB-Lei para o segundo dígito

Nas Figura 5-13 e Figura 5-14 visualizamos a adequação do segundo dígito de ambos os autovalores à Lei. Percebemos que o encaixe para o segundo dígito de ambos os autovalores à Lei também é excelente, sendo pior do que para o coeficiente de Harris (Figura 5-5 e Figura 5-6), mas melhor do que para o primeiro dígito (Figura 5-10 e Figura 5-11).

Na Tabela 5-5 visualizamos a síntese dos testes/medidas de conformidade de ambos os autovalores para o segundo dígito. Ao compararmos seus valores com os obtidos para o primeiro dígito (Tabela 5-4), concluímos novamente que em todos os casos a conformidade do segundo dígito é melhor que a do primeiro. Ao compararmos ao coeficiente de Harris (Tabela 5-2), percebemos uma pior adequação, mas ainda excelente. Os testes estatísticos χ^2 de Pearson e Kolmogorov-Smirnov, novamente sendo afetados pela alta quantidade de amostras, aceitam respectivamente 50% e 69.9% das imagens como conformes com nível de significância 0.05.

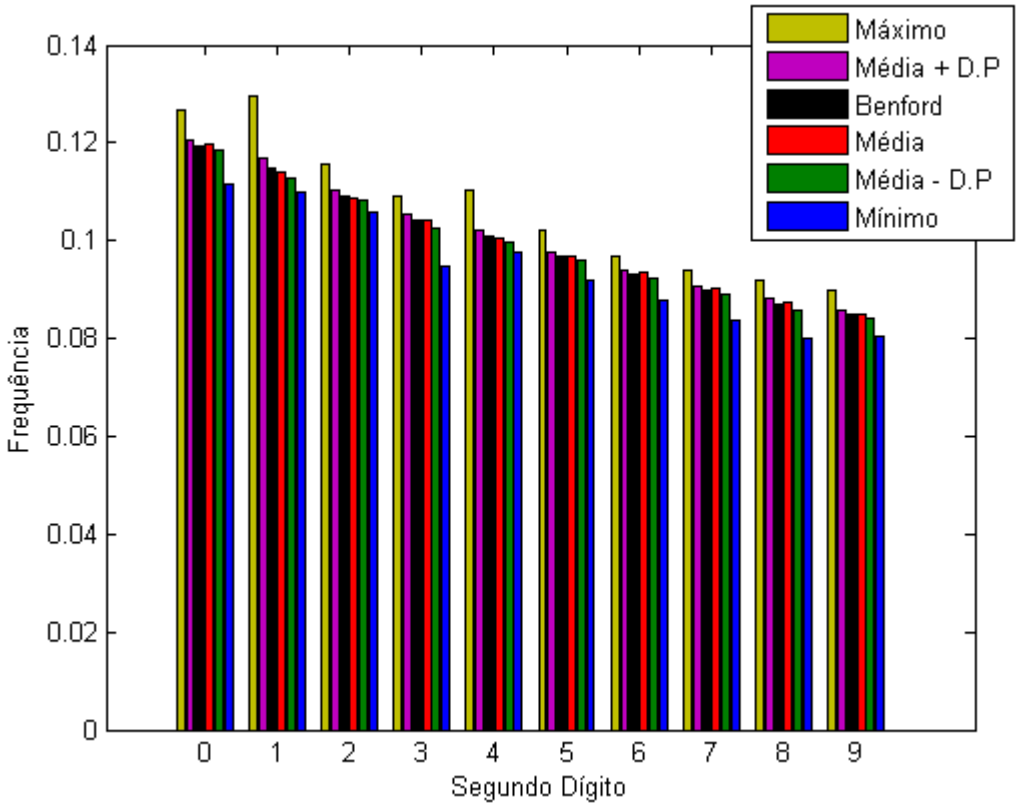

Figura 5-14: Síntese da conformidade de ambos os autovalores à NB-Lei para o segundo dígito (gráfico em barras)

Tabela 5-5: Análise da conformidade de ambos os autovalores à NB-Lei para o segundo dígito

Conformidade	Média	Desvio Padrão	Mínimo	Máximo	Aceitação
Distância ε	0. 38%	0. 37%	0. 065%	3.41%	N/A
χ^2 de Pearson	55.84	163.15	1.15	1976.83	674
Divergência χ^2	0.00015	0.00046	2.93e-06	0.0055	N/A
K-S	0.0026	0.0025	0.00032	0.024	1009

5.2.3 Análise dos dois primeiros dígitos

Na Figura 5-15 podemos ver que as linhas se apresentam mais irregularmente do que em todos os gráficos anteriores, mas continuam seguindo à NB-Lei. Novamente percebemos que algumas sequências específicas de dígitos apresentam picos.

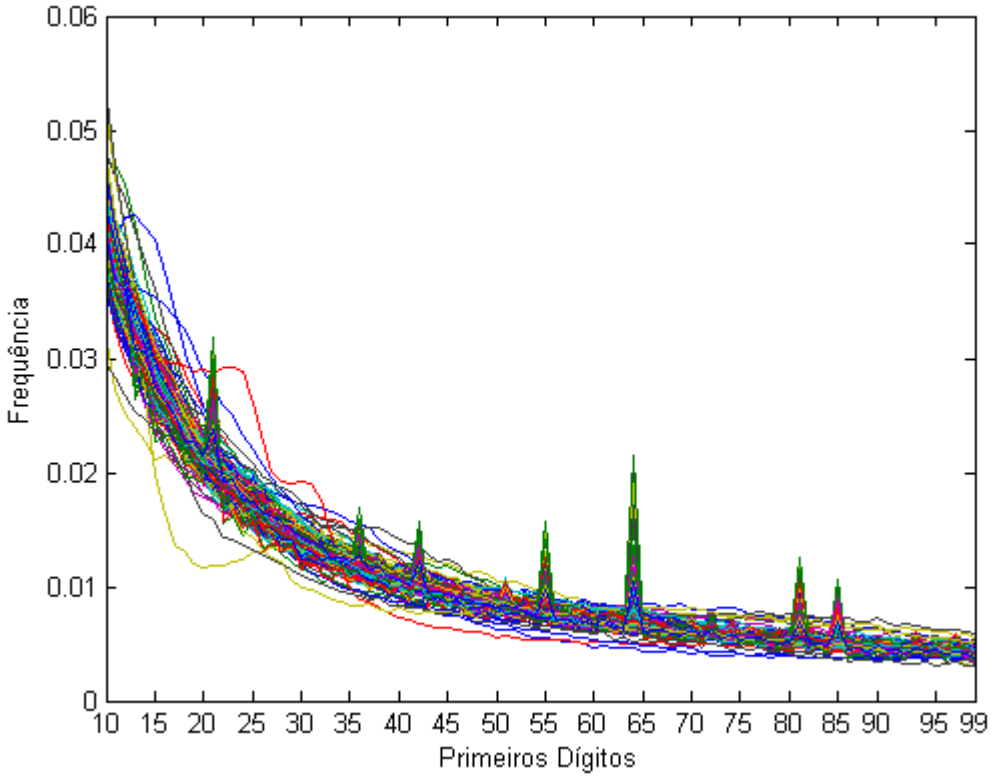

Figura 5-15: Probabilidades dos dois primeiros dígitos de ambos autovalores para cada uma das 1.338 imagens

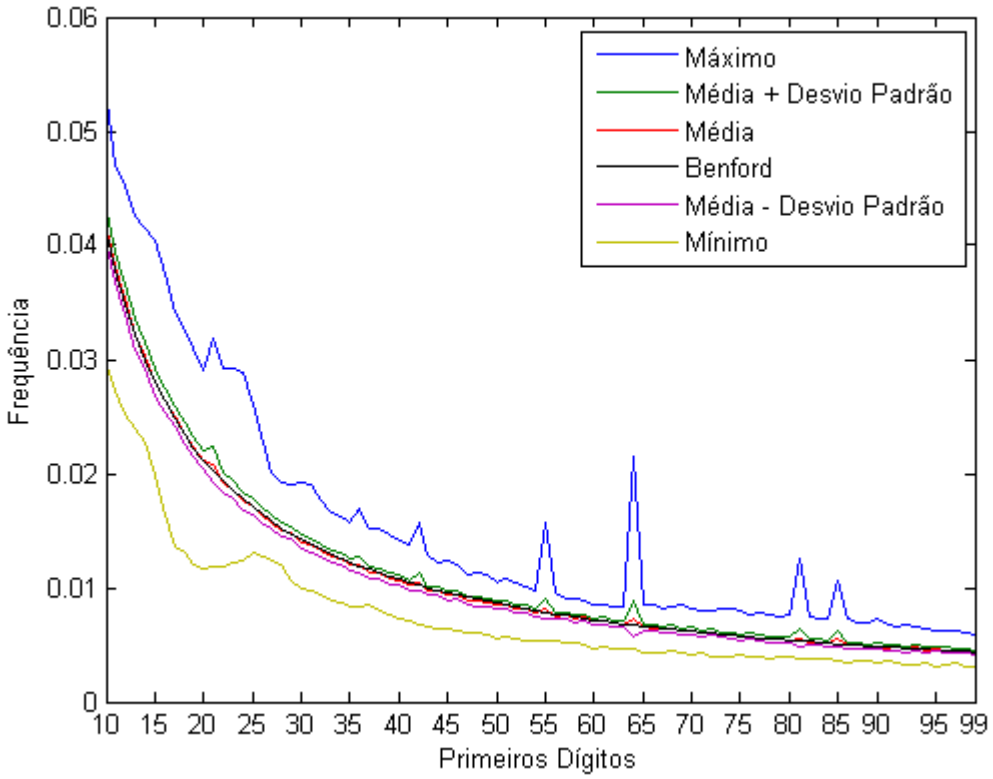

Figura 5-16: Síntese da conformidade de ambos autovalores à NB-Lei para os dois primeiros dígitos

Na Figura 5-16 visualizamos a adequação dos dois primeiros dígitos de ambos autovalores à Lei. Percebemos que o encaixe dos autovalores à Lei para os dois primeiros dígitos também é excelente, embora pior que o encaixe do coeficiente de Harris, apresentando maiores irregularidades.

Tabela 5-6: Análise da conformidade de ambos os autovalores à NB-Lei para os dois primeiros dígitos

Conformidade	Média	Desvio Padrão	Mínimo	Máximo	Aceitação
Distância ε	1.57%	1.14%	0.49%	11.62%	N/A
χ^2 de Pearson	1044.12	2536.81	65.35	29966.72	68
Divergência χ^2	0.0027	0.0068	0.00016	0.076	N/A
K-S	0.0099	0.0093	0.00066	0.11	154

Na Tabela 5-6 visualizamos a síntese dos testes/medidas de conformidade de ambos autovalores para os dois primeiros dígitos. Ao compararmos seus valores com os obtidos para o coeficiente de Harris (Tabela 5-3), podemos inferir que a adequação à Lei de Benford para os autovalores é mais errática, mas ainda excelente. Os testes estatísticos χ^2 de Pearson e Kolmogorov-Smirnov, mais uma vez afetados pelo alto tamanho de amostras, aceitam respectivamente somente 5% e 11.5% das imagens como conformes com nível de significância 0.05.

5.3 Mínimo dos autovalores

Nesta seção iremos apresentar os resultados da análise da conformidade do mínimo dos autovalores à NB-Lei. É importante ressaltar a grandeza do mínimo dos autovalores utiliza-se de um operador incomum na área da Lei, do mínimo entre dois valores.

5.3.1 Análise do primeiro dígito

Na Figura 5-17 visualizamos as 1.338 linhas representando as probabilidades do mínimo dos autovalores para o primeiro dígito. No gráfico, as linhas se distribuem mais dispersamente do que se comparadas ao coeficiente de Harris (Figura 5-1) e a ambos os autovalores (Figura 5-9), mas ainda seguem à Lei do Dígito Significativo muito bem.

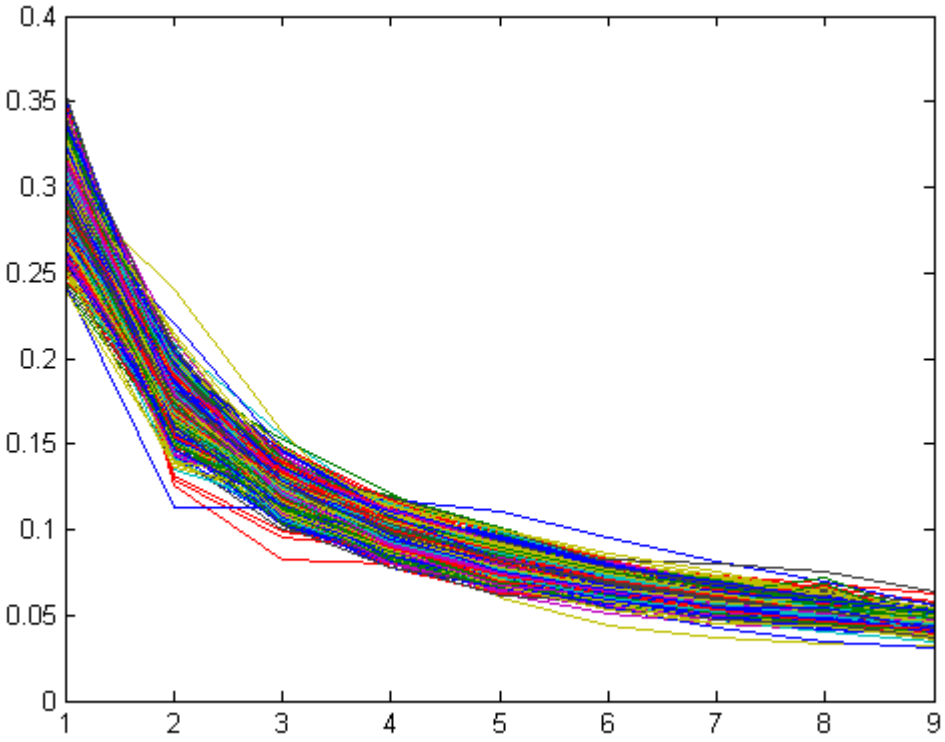

Figura 5-17: Probabilidades do primeiro dígito do mínimo dos autovalores para cada uma das 1.338 imagens

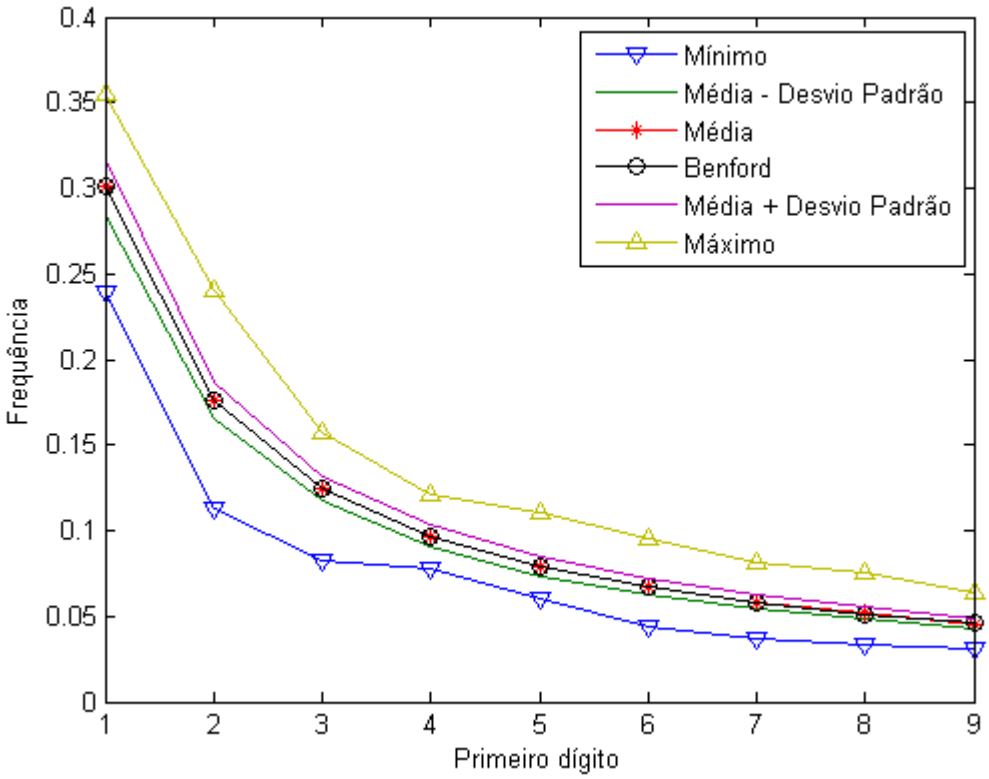

Figura 5-18: Síntese da conformidade do mínimo dos autovalores à NB-Lei para o primeiro dígito

Nas Figura 5-18 e Figura 5-19, visualizamos a adequação do mínimo dos autovalores à Lei. Percebemos novamente que, assim como no coeficiente de Harris (Figura 5-2 e Figura 5-3) e em ambos os autovalores (Figura 5-10 e Figura 5-11), as probabilidades esperadas para a Lei de Benford (mostradas em preto), se intersectam perfeitamente com a média obtida das imagens (mostrada em vermelho). Já o desvio padrão é ligeiramente maior quando comparado aos dois.

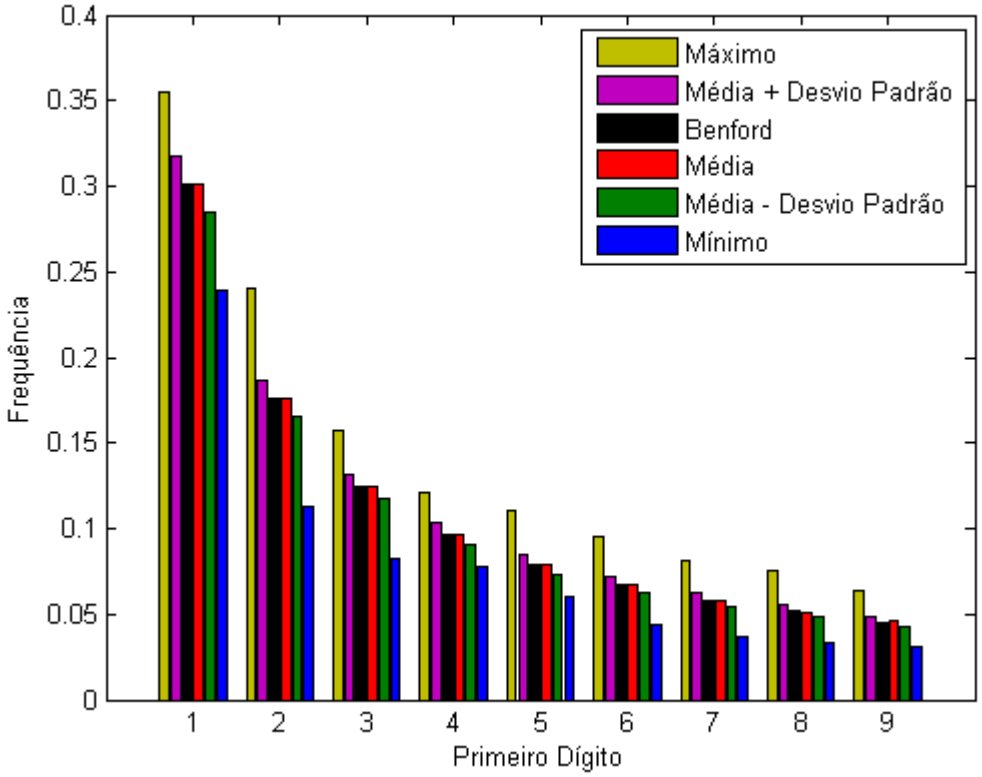

Figura 5-19: Síntese da conformidade do mínimo dos autovalores à NB-Lei para o primeiro dígito (gráfico em barras)

Tabela 5-7: Análise da conformidade do mínimo dos autovalores à NB-Lei para o primeiro dígito

Conformidade	Média	Desvio Padrão	Mínimo	Máximo	Aceitação
Distância ε	2.21%	1.66%	0.17%	13.208%	N/A
χ^2 de Pearson	750.75	1204.51	4.14	15992.83	46
Divergência χ^2	0.0038	0.0061	2.11e-05	0.081	N/A
K-S	0.019	0.015	0.00069	0.13	142

Na Tabela 5-7 visualizamos a síntese dos testes/medidas de conformidade do mínimo dos autovalores para o primeiro dígito. Ao compararmos com o coeficiente de Harris

(Tabela 5-1) e ambos os autovalores (Tabela 5-4), percebemos que a conformidade de ambos os autovalores é inferior, mas que ainda é excelente. Os testes estatísticos χ^2 de Pearson e Kolmogorov-Smirnov, neste caso, aceitam respectivamente 3.4% e 10.6% das imagens como conformes, com nível de significância 0.05.

5.3.2 Análise do segundo dígito

Na Figura 5-20 vemos as 1.338 linhas representando as probabilidades do segundo dígito do mínimo dos autovalores para cada uma das imagens. No gráfico, percebemos que a maioria das linhas estão situadas na mesma localidade, e vemos também que assim como em ambos os autovalores (Figura 5-12), alguns dígitos específicos apresentam um maior desvio.

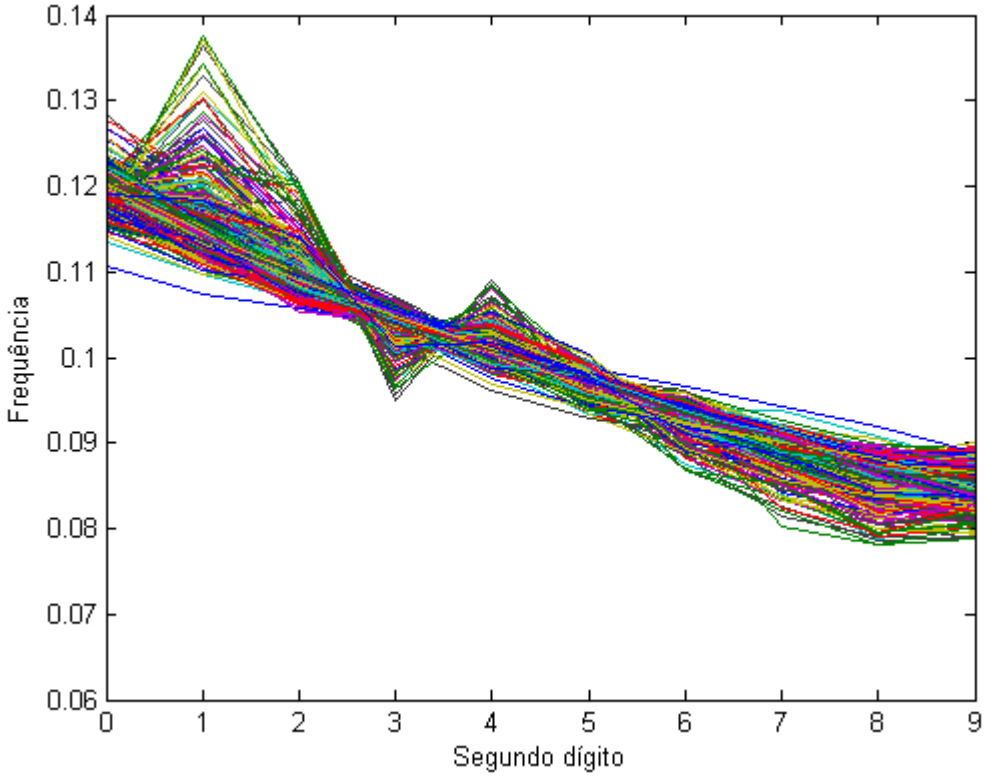

Figura 5-20: Probabilidades do segundo dígito do mínimo dos autovalores para cada uma das 1.338 imagens UCID

Nas Figura 5-21 e Figura 5-22 observamos a adequação do segundo dígito do mínimo dos autovalores à Lei. Percebemos que o encaixe para o segundo dígito do mínimo dos autovalores à Lei é excelente, porém pior que o do coeficiente de Harris (Figura 5-5 e Figura 5-6) e de ambos os autovalores (Figura 5-13 e Figura 5-14).

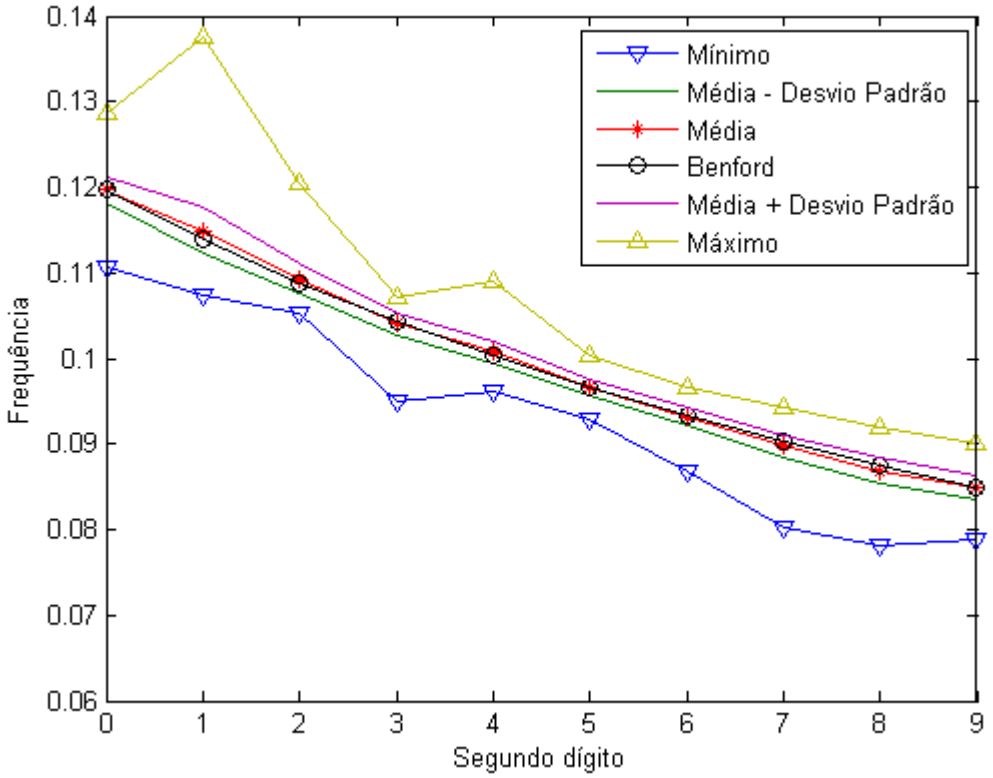

Figura 5-21: Síntese da conformidade do mínimo dos autovalores à NB-Lei para o segundo dígito

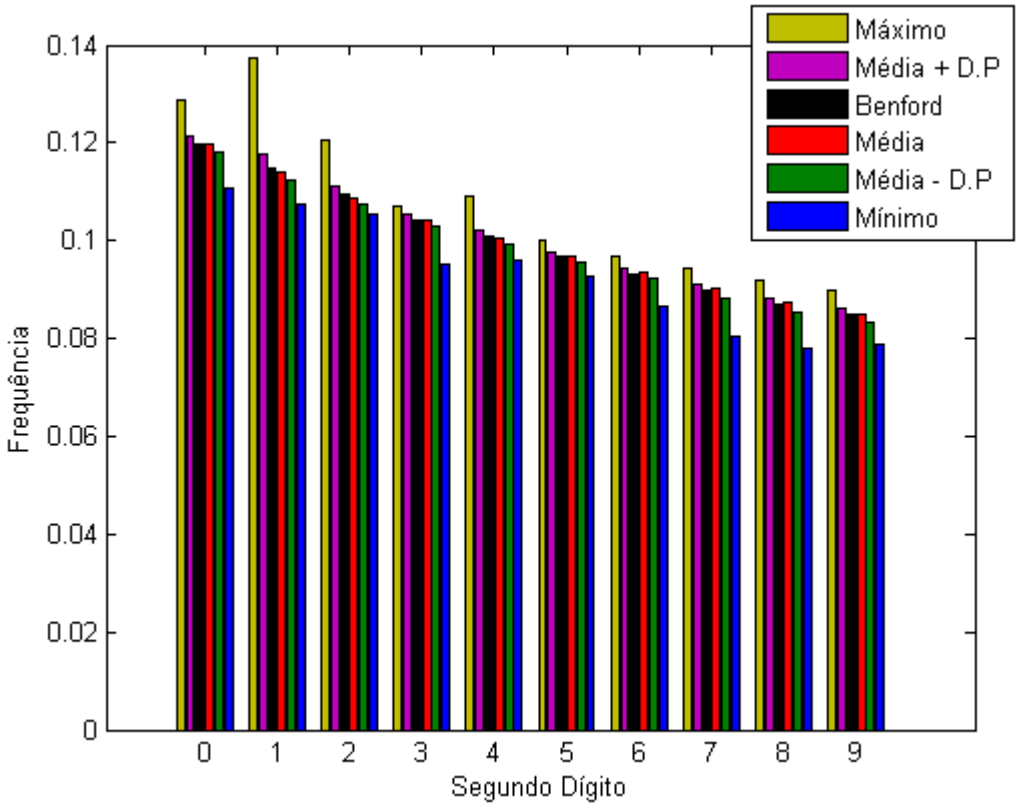

Figura 5-22: Síntese da conformidade do mínimo dos autovalores à NB-Lei para o segundo dígito (em barras)

Tabela 5-8: Análise da conformidade do mínimo dos autovalores à NB-Lei para o segundo dígito

Conformidade	Média	Desvio Padrão	Mínimo	Máximo	Aceitação
Distância ε	0.51%	0.45%	0.103%	4.25%	N/A
χ^2 de Pearson	49.61	137.077	1.29	1729.57	651
Divergência χ^2	0.00026	0.00077	6.59e-06	0.0099	N/A
K-S	0.00401	0.0038	0.00043	0.031	936

A Tabela 5-8 apresenta a síntese dos testes/medidas de conformidade do mínimo dos autovalores para o segundo dígito. Ao compararmos ao coeficiente de Harris (Tabela 5-2) e a ambos os autovalores (Tabela 5-5), percebemos uma pior adequação, mas ainda excelente. Os testes estatísticos χ^2 de Pearson e Kolmogorov-Smirnov aceitam respectivamente 48% e 69.9% das imagens como conformes com nível de significância 0.05.

5.3.3 Análise dos dois primeiros dígitos

Na Figura 5-23 podemos ver que as linhas se apresentam mais irregularmente do que em todos os gráficos anteriores, mas continuam seguindo à NB-Lei. Ao compararmos com ambos os autovalores (Figura 5-15), percebemos que os picos se distribuem nas mesmas sequencias.

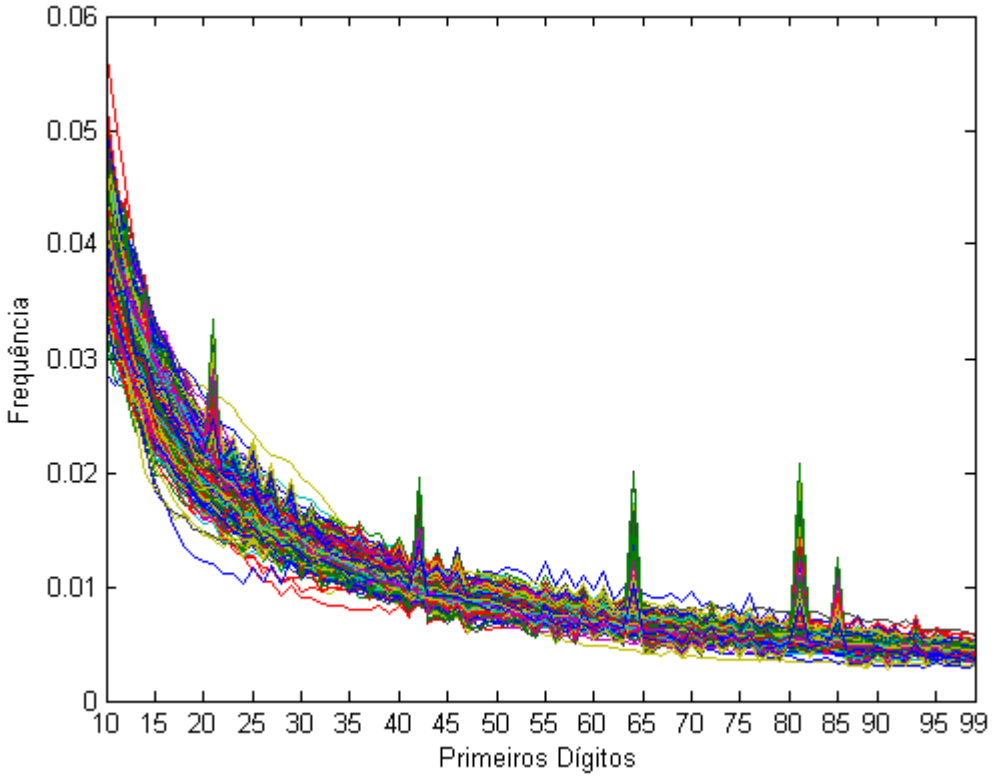

Figura 5-23: Probabilidades dos dois primeiros dígitos do mínimo dos autovalores para todas 1.338 imagens

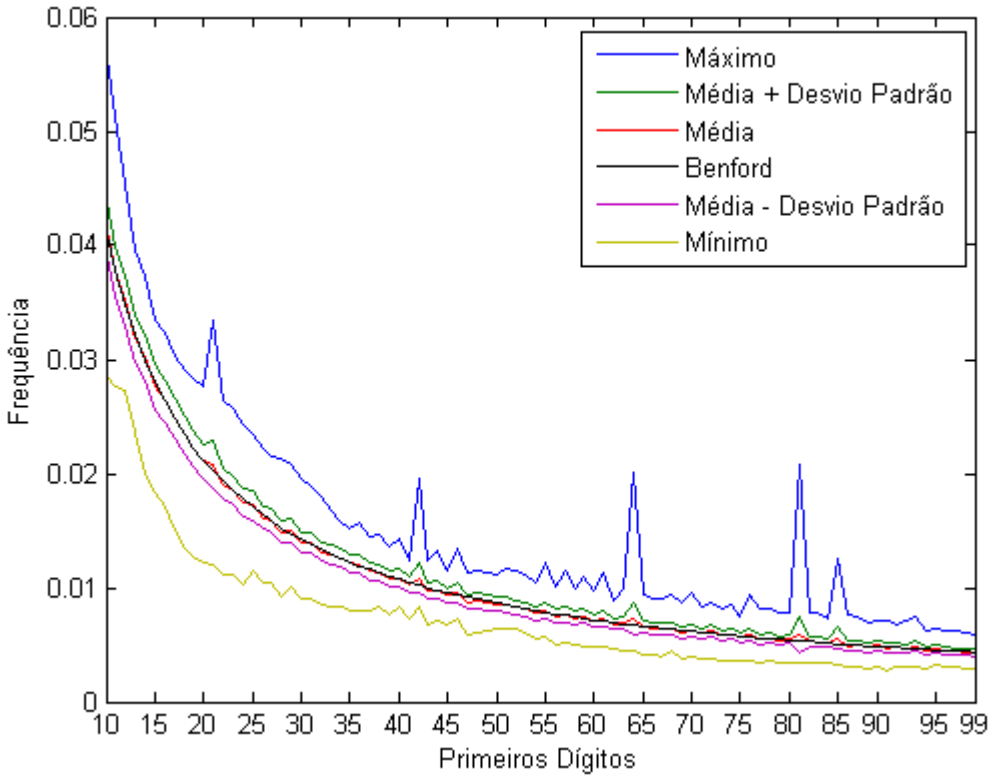

Figura 5-24: Síntese da conformidade do mínimo dos autovalores à NB-Lei para os dois primeiros dígitos

Na Figura 5-24 visualizamos a adequação dos dois primeiros dígitos do mínimo dos autovalores à Lei. Percebemos que o encaixe do mínimo dos autovalores à Lei para os dois primeiros dígitos ainda é excelente, embora pior do que o encaixe das duas outras grandezas previamente analisadas (Figura 5-8 e Figura 5-16), apresentando ainda maiores irregularidades.

Tabela 5-9: Análise da conformidade do mínimo dos autovalores à NB-Lei para os dois primeiros dígitos

Conformidade	Média	Desvio Padrão	Mínimo	Máximo	Aceitação
Distância ε	2.72%	1.77%	0.72%	14.02%	N/A
χ^2 de Pearson	1257.76	1990.86	64.67	19606.8	81
Divergência χ^2	0.0065	0.0107	0.00033	0.109	N/A
K-S	0.019	0.015	0.00085	0.13	122

Na Tabela 5-9 visualizamos a síntese dos testes/medidas de conformidade do coeficiente de Harris para os dois primeiros dígitos. Ao compararmos seus valores com os obtidos para outras grandezas analisadas (Tabela 5-3 e Tabela 5-6), podemos inferir que das três, esta é a que pior se adequa, mas sua adequação ainda é excelente. Os testes estatísticos χ^2 de Pearson e Kolmogorov-Smirnov, aceitam respectivamente 6% e 9.1% das imagens como conformes com nível de significância 0.05.

5.4 Síntese de Resultados

Nas seções anteriores foram apresentados os resultados obtidos da a análise de conformidade à Lei de Newcomb-Benford para as três grandezas: detector de Harris, ambos autovalores e o mínimo dos autovalores da matriz de covariância. Para cada grandeza foi avaliada separadamente sua conformidade ao primeiro dígito, ao segundo dígito, e aos dois primeiros dígitos. Observamos que para todas as grandezas, a maior aderência à Lei se deu no segundo dígito, seguido pelo primeiro dígito e finalmente os dois primeiros dígitos foram os menos aderentes. Além disso, a grandeza que apresentou melhor conformidade em todas as análises foi o detector de Harris, seguido por ambos os autovalores, e finalmente o mínimo dos autovalores foi a grandeza com pior conformidade analisada. Iremos então comparar os resultados obtidos com alguns dos trabalhos relatados na Seção 4.1. Como os trabalhos informados só realizam a análise para o primeiro dígito, com exceção de Carslaw, as comparações são feitas só para o primeiro dígito.

Benford (BENFORD, 1938), obteve como a menor distância ε o valor de 2.8% para o grupo de itens numéricos retirados de jornais (grupo D) (Figura 4-2); as grandezas avaliadas obtiveram média de 0.45%, 1.15% e 2.21%. O valor máximo obtido foi de 9.69%, 10.28% e 13.2%. Podemos comparar este valor máximo com o do grupo de populações de Benford (grupo B), que obteve distância de 16.6%, e é considerado pela

literatura como aderente à Lei. Podemos então afirmar que nossa pior aderência foi melhor do que a aderência de um grupo tido universalmente como conforme à Lei e que a média foi melhor do que a melhor aderência analisada por Benford.

Podemos também comparar diretamente nossos resultados com os obtidos por Fu, Shi e Su (FU; SHI; SU, 2007) e por Qadir, Zhao e Ho (QADIR; ZHAO; HO, 2010), uma vez que são utilizadas a mesma base de imagens e a mesma medida, a divergência χ^2. Ao comparar nossos gráficos de adequação (Figura 5-3, Figura 5-11, Figura 5-19) aos gráficos obtidos por estes trabalhos (Figura 4-8 e Figura 4-9), é notável que a adequação das grandezas aqui analisadas é significativamente superior à adequação do bloco-DCT e da transformada DWT. Além disso, ambos os trabalhos informam a divergência χ^2 calculada para a probabilidade média de todas as imagens: 0.0034 e 0.0016 respectivamente; as nossas grandezas obtiveram os valores de: 4.87e-06, 4.54e-05 e 2.7e-05; a diferença de adequação dada por esta métrica, considerando as interpretações dos valores dos testes na literatura aplicada, é extremamente significativa, sendo superior a uma ordem de grandeza.

6 Considerações finais e conclusão

Este trabalho realizou uma análise minuciosa da conformidade da Lei de Newcomb-Benford aos qualificadores de pontos de interesse, em específico o detector de Harris, os autovalores da matriz de covariância, e o mínimo dos autovalores da matriz de covariância. Fundamentamos nossa análise com diversos testes/medidas de conformidade difundidas na literatura, além de comparações diretas com os principais trabalhos relacionados. Concluímos então que todos os três qualificadores examinados possuem excelente conformidade à Lei do Dígito Significativo, e que a conformidade do detector de Harris é tão significativa, que podemos concluir que na literatura ela é a grandeza extraída de dados reais que melhor se adequa à NB-Lei até o momento.

6.1 Principais contribuições

A principal contribuição deste trabalho foi a demonstração de que as três grandezas qualificadoras de pontos de interesse analisadas apresentam alta conformidade à Lei de Newcomb-Benford, e que o detector de Harris é a grandeza não simulada que melhor se adequa à Lei até o presente momento na literatura.

A comprovação de que imagens reais possuem alta conformidade para as três grandezas sugere que imagens sintetizadas por métodos fisicamente realistas devam obter também alta conformidade. Além disso, acreditamos que estes resultados obtidos podem ser aplicados como método de classificação de imagens, conseguindo separar imagens reais de imagens sintéticas e desenhos ou gráficos.

Ademais, ao analisarmos a aderência do mínimo dos autovalores à Lei, comprovamos experimentalmente que o operador de seleção do mínimo entre dois valores relacionados, embora piore levemente os níveis de conformidade, não inviabiliza a aderência à Lei. O mínimo entre os dois autovalores, mesmo aplicando o operador mínimo, possui conformidade média melhor do que a melhor conformidade apresentada no trabalho original de Benford.

6.2 Trabalhos futuros

O primeiro trabalho futuro a ser realizado é a investigação do comportamento das conformidades para outros tipos de imagens. Faz-se necessária uma investigação de quais propriedades encontradas nas imagens afetam a conformidade. Isto inclui, porém não se limita a: análise de ruídos; análise de operadores comuns em processamento de imagens, como filtros de alta e baixa frequência; compressão com perda de qualidade; composição ou adulteração de imagens; tamanho das imagens; e claridade. A partir desta investigação, estudaremos as possíveis aplicações práticas da análise de conformidade das grandezas estudadas à NB-Lei.

Outra investigação que se faz necessária é analisar detalhadamente a causa dos picos verificados nos gráficos para alguns dígitos específicos, como é o caso do pico da

sequência “45” na Figura 5-7. Tal investigação concluirá se estes picos são causados devido a propriedades das imagens, ou se são resultantes das complexas operações de ponto flutuante utilizadas durante o cálculo.

Ademais, seria interessante investigar como a conformidade se comporta localmente em partes de uma imagem. Para tal podemos dividir uma imagem em seções não sobrepostas e verificar a conformidade de cada seção individualmente, comparando os resultados obtidos entre si e entre o resultado da imagem inteira.

Finalmente, destacamos a necessidade da criação de um modelo estatístico/probabilístico que de fato prove a conformidade de uma grandeza à Lei do Dígito Significativo.

7 Referências

ACEBO, E.; SBERT, M. Benford's law for natural and synthetic images. Computational Aesthetics Eurographics Association, p. 169–176, 2005.

BENFORD, F. The law of anomalous numbers. Proceedings of the American Philosophical Society, v. 78, n. 4, p. 551-572, 1938.

BERGER, A. Benford's law in power-like dynamical systems. Stochastics and Dynamics, v. 5, n. 4, p. 587–607, 2006.

BERGER, A. et al. Finite-state Markov Chains Obey Benford's Law. SIAM J. on Matrix Analysis and Applications, v. 32, n. 3, p. 665-684, 2011.

BERGER, A.; BUNIMOVICH, L. A.; HILL, T. P. One-dimensional dynamical systems and Benford's Law. American Mathematical Socisety, v. 357, n. 1, p. 197-219, 2004.

BERGER, A.; HILL, T. P. Newton's Method Obeys Benford's Law. The American Mathematical Monthly, v. 114, n. 7, p. 588–601, 2007.

BERGER, A.; HILL, T. P. A basic theory of Benford's Law. Probability Surveys, v. 8, p. 1-126, 2011.

BERGER, A.; HILL, T. P. VIEW CHRONOLOGICAL. BENFORD ONLINE BIBLIOGRAPHY, 2012. Disponivel em: <http://www.benfordonline.net/list/chronological>. Acesso em: Janeiro 2012.

BHATTACHARYA, S.; XU, D.; KUMAR, K. An ANN-based auditor decision support system using Benford's law. Decision Support Systems, v. 50, n. 3, p. 576-584, 2011.

BORING, E. G. The Logic of the Normal Law of Error in Mental Measurement. American Journal of Psychology, v. 31, n. 1, p. 1-33, 1920.

BUSTA, B.; WEINBERG, R. Using Benford's law and neural networks as a review procedure. Managerial Auditing Journal, MCB UP Ltd, v. 13, n. 6, p. 356-366, 1998.

CARSLAW, C. A. P. N. Anomalies in income numbers: Evidence of goal oriented behavior. Accounting Review, v. 63, n. 2, p. 321–327, 1988.

DIACONIS, P. The Distribution of Leading Digits and Uniform Distribution Mod1. The Annals of Probability, v. 5, n. 1, p. 72–81, 1977.

FELLER, W. Introduction to Probability Theory. New York: Wiley, v. 2, 1966. p. 1056-1061

FRANEL, J. A propos des tables de logarithmes. Festschrift der Naturforschenden Gesellschaft, v. 62, p. 286-295, 1917.

FU, D.; SHI, Y. Q.; SU, W. A generalized Benford's law for JPEG coefficients and its applications in image forensics. Security, Steganography, and Watermarking of Multimedia Contents IX. [S.l.]: [s.n.]. 2007.

GOUDSMIT, S. A.; FURRY, W. H. Significant Figures of Numbers in Statistical Tables. Nature, v. 154, n. 3921, p. 800-801, 1944.

HAMMING, R. W. On the Distribution of Numbers. Bell Systems Technical Journal, v. 49, n. 8, p. 1609-1625, 1970.

HARRIS, C.; STEPHENS, M., A Combined Corner and Edge Detector, in 'Proceedings of the 4th Alvey Vision Conference' , p. 147-151, 1988.

den HEIJER, E.; EIBEN, A.E.; , "Using aesthetic measures to evolve art," Evolutionary Computation (CEC), 2010 IEEE Congress on , vol., no., p.1-8, 18-23 July 2010

HILL, T. P. Random-number guessing and the first digit phenomenon. Psychological Reports, v. 62, p. 967-971, 1988.

HILL, T. P. Base-Invariance Implies Benford's Law. Mathematical Society, v. 123, n. 3, p. 887-895, 1995a.

HILL, T. P. The Significant-Digit Phenomenon. The American mathematical monthly, v. 102, n. 4, p. 322-327, 1995b.

HILL, T. P. A Statistical Derivation of the Significant-Digit Law. Statistical Science, v. 10, n. 4, p. 354–363, 1995c.

HSÜ, E. H. An experimental study on "mental numbers" and a new application. Journal of General Psychology, v. 38, p. 57-67, 1948.

JOLION, J. M. Images and benford’s law. Journal of Mathematical Imaging and Vision, v. 14, n. 1, p. 73-81, 2001.

KRAKAR, Z.; ZGELA, M. Application of Benford's Law in Payment Systems Auditing. Journal of Information and Organizational Sciences, v. 33, n. 1, p. 39-51, 2009.

KNUTH, D. The Art of Computing Programming. [S.l.]: Addison-Wesley, v. 2, 1969. 219-229 p.

LÉVY, P. L’addition des variables aléatoires définies sur une circonférence. Bulletin de la Société Mathématique de France, v. 67, n. 1, p. 1-41, 1939.

LUQUE, B.; LACASA, L. The first-digit frequencies of prime numbers and Riemann zeta zeros. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science, v. 465, n. 2107, p. 2197-2216, 2009.

MORGAN, J. A. et al. Letters to the Editor. The American Statistician, v. 26, n. 3, p.62-66, 1972.

NIGRINI, M. J. The detection of income tax evasion through an analysis of digital distributions. University of Cincinnati. [S.l.]: [s.n.]. 1992.

NIGRINI, M. J. A taxpayer compliance application of Benford's law. Journal of the American Taxation Association, v. 18, n. 1, p. 72-91, 1996.

NIGRINI, M. J.; MITTERMAIER, L. The Use of Benford's Law as an Aid in Analytical Procedures. Auditing: A Journal of Practice & Theory, v. 16, n. 2, p. 52-67, 1997.

NEWCOMB, S. Note on the Frequency of Use of the Different Digits in Natural Numbers. American Journal of Mathematics, v. 4, n. 1, p. 39-40, 1881.

PÉREZ-GONZÁLEZ, F.; HEILEMAN, G.; ABDALLAH, C. T. Benford's Law in Image Processing. Proc. IEEE International Conference on Image Processing. [S.l.]: [s.n.]. 2007. p. 405–408.

PETITT, A. N.; STEVENS, M. A., The Kolmogorov-Smirnov Goodness-of-Fit Statistics for Discrete and Grouped Data, Technometrics, 19 (1977).

PINKHAM, R. S. On the distribution of first significant digits. The Annals of Mathematical Statistics, v. 32, n. 4, p. 1223–1230, 1961.

POINCARÉ, J. H. Calcul des Probabilités. [S.l.]: [s.n.], 1912. p. 313-320

QADIR, G.; ZHAO, X.; HO, A.T.S. Estimating JPEG2000 compression for image forensics using Benford's Law. Optics, Photonics, and Digital Technologies for Multimedia Applications, SPIE, 2010

QADIR, G.; ZHAO, X.; HO, A.T.S.; CASEY, M. Image forensic of glare feature for improving image retrieval using Benford's Law, in Proc. ISCAS, 2011, p.2661-2664., 2011

RAIMI, R. A. The First Digit Problem. The American Methematical Monthly, v. 83, n. 7, p. 521-538, 1976.

ROBBINS, H. On the equidistribution of sums of independent random variables. Proceedings of the American Mathematical Society, v. 4, p. 786-799, 1953.

SANCHES, J.; MARQUES, J. S. Image reconstruction using the benford law. International Conference on Image Processing. [S.l.]: IEEE. 2006. p. 2029–2032.

SCHAEFER, G.; STICH, M. UCID - An Uncompressed Colour Image Database. Nottingham Trent University. [S.l.]. 2003.

SHAO, L.; MA, B.-Q. Empirical Mantissa Distributions of Pulsars. arxiv:1005.1702v1 (Preprint), p. 1-15, 2011.

SHI, J. & TOMASI, C., Good Features to Track, IEEE Conference on Computer Vision and Pattern Recognition , p. 593-600. 1994.

SNYDER, M. A.; CURRY, J. H.; DOUGHERTY, A. M. Stochastic aspects of one-dimensional discrete dynamical systems: Benford's law. Physical Review E, v. 64, n. 2, p. 1-5, 2001.

STEELE, M.; CHASELING, J. Powers of Discrete Goodness-of-Fit Test Statistics for a Uniform Null Against a Selection of Alternative Distributions. Communications in Statistics: Simulation and Computation, v. 35, n. 4, p. 1067-1075, 2006.

TAO, T. Benford's law, Zipf's law, and the Pareto distribution. What's New, 2009. Disponivel em: <http://terrytao.wordpress.com/2009/07/03/benfords-law-zipfs-law-and-the-pareto-distribution/>. Acesso em: Agosto 2011.

THOMAS, J. K. Unusual patterns in reported earnings. Accounting Review, v. 64, n. 4, p. 773-787, 1989.

TOLLE, C. R.; BUDZIEN, J. L.; LAVIOLETTE, R. A. Do dynamical systems follow Benford's law? Chaos: An Interdisciplinary Journal of Nonlinear Science, v. 10, n. 2, p. 587–607, 2000.

WEISSTEIN, E. W. Newton-Cotes Formula: Mathworld-a wolfram web resource. 2012. Disponível em: <http://mathworld.wolfram.com/Newton-CotesFormulas.html>. Acesso em: 19 Janeiro 2012.

WEYL, H. Über die Gleichverteilung von Zahlen mod. Eins. Mathematische Annalen, v. 77, n. 3, p. 313-352, 1916.

WONG, S. C. Y. Testing Benford's Law with the First Two Significant Digits, 2010.