MTHE 493 Thesis

Text Restoration Using Probabilistic Methods

Authors:
Amy Cotton - 10092691

Emma Hansen - 10090532

Luca Castelli - 10099808

Richard Hum - 10092437

Supervisor:
Dr. Abdol-Reza Mansouri

April 7, 2017

Acknowledgements

We would like to thank:

- Professor Mansouri for guiding through the theoretical background and teaching us all of the rich mathematical concepts and patiently answering our questions,
- Our fellow students in Apple Math for putting up with our frustration when the code produced cows instead of zebras,
- G. Winkler and D. Mumford et. al for providing the basis for the mathematical theory, and
- Amy Cotton, for frequently bringing snacks and being a complete PowerPoint Wizard.

Abstract

This report combines the results of Geman and Geman's Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images (1984) and Mumford, Wu, and Zhu's Filters, Random Fields and Maximum Entropy (1998) to restore images. Images are modelled as Markov random fields (MRF) and the equivalence between Markov random fields and Gibbs random fields is exploited. An adequate energy function is defined. Given an image with white additive noise, the posterior probability is an MRF with respect to a neighbourhood structure. By introducing simulated annealing the problem of image restoration becomes a maximum a posteriori (MAP) estimation problem of an ideal image given a noisy image. The performance of the restoration procedure is improved by utilizing the FRAME algorithm to learn suitable image priors. The FRAME methodology makes use of the maximum entropy principle, filters and Markov random fields. Filters are selected to capture characteristics of images, histograms are generated from the filtered images, and a gradient descent procedure is followed while making use of the Gibbs Sampler.

Index Terms - Gibbs Field, Gibbs Sampler, Markov Kernel, Maximum Entropy, Simulated Annealing, FRAME Algorithm

Contents

1 Introduction

1.1 Motivation

Led by advancements in technology, the digitization of documents has become very popular, and the use of paper documents has dwindled [1]. Often, the quality of these digitized documents is not great. A text restoration solution combining both the results of Geman and Geman (1984) and Mumford, Wu, and Zhu (1998) is proposed in this report.

There are two main reasons digitized documents experience lower quality. First, as paper documents age, both the paper and ink fade [2]. Second, it is common for the actual process of scanning documents to result in loss of quality, see Figure 1. As such, the need for a method to restore text quality arises.

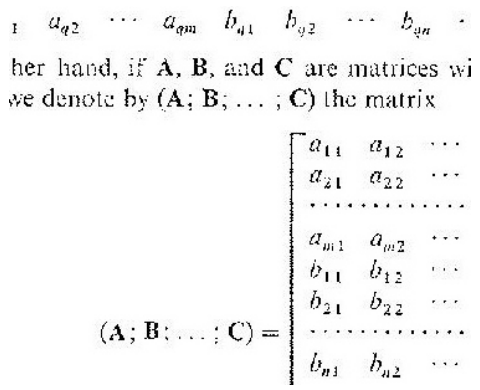


Figure 1: Example of a scanned document with poor quality [3].

Text image denoising is highly relevant across many different areas. Possible applications include enhancing the text of:

- medical records,
- equipment manuals,
- old or damaged books, and
- pictures taken of text.

1.2 Design Problem

Our main goal was to design a system that accurately and efficiently improved the quality of poorly scanned images of English text. To do so, not only must the mathematical theory be considered, but also how the solution will be implemented, and how it will affect the users.

As many of the documents being archived digitally are the intellectual property, or personal information, of people and because of copyright laws, precautions must be taken to ensure the security of information. Canada's copyright laws state that "if the original is rare or unpublished and is (i) deteriorating, damaged or lost, or (ii) at risk of deteriorating

or becoming damaged or lost" then libraries, archives, and museums are allowed to make a copy of the information [4]. Since archiving is moving towards digital storage, this means that personal information will be stored digitally, where it is more accessible to hackers than it was to thieves.

Implementing the system poses another challenge. There is significant cost associated with converting hard-copy documents to digital ones, both in money spent on training and technology, and in time spent scanning the documents [5]. Part of this cost comes from ensuring the safety of the employees doing the archiving, since studies have shown that there is a correlation between prolonged use of back-lit screens and eye strain. Thus, policy must be implemented to ensure the safety of employees [6].

1.3 Approach

The problem of restoring or denoising an image is non-trivial and is a popular research topic. Common methods involve applying linear filtering and making use of wavelet transforms [7]. The approach taken in the proposed solution is through probabilistic methods. The technique involves modelling images as Markov random fields. The idea is to study a sample of text images to learn the features that characterize the space of text images. Mumford, Wu, Zhu's *Filters, Random Fields, and Maximum Entropy* algorithm (FRAME) outlines a process to do so using the Maximum Entropy Principle. Geman and Geman's Gibbs Sampler plays a critical role in the FRAME algorithm by allowing one to sample from a Gibbs distribution.

1.4 Objectives

The implementation of our proposed solution was tackled in two phases:

1. Implementing Geman and Geman's Gibbs Sampler to perform restoration of images with added Gaussian noise.

2. Implementing Mumford, Wu, and Zhu's FRAME algorithm using text images as inputs to learn a prior probability distribution.

Resulting in the successful restoration of a text image.

2 Design Solution

This project focuses on the design of a system to restore text images and not on the implementation of said system in a commercial application. Thus, the process outlined below is what was used in the design of the system.

Image restoration was treated as a probabilistic estimation problem where images were modelled as a family of random variables using the following observation model.

$$O = I + N \tag{1}$$

where O is the observed image, I is the ideal underlying image, and N is assumed to be white Gaussian noise. The assumption of white Gaussian noise was made because additive

noise results from a variety of sources, and the cumulative effect of these sources can be approximated by a Gaussian distribution. This is due to the Central Limit Theorem which states that for any identically independently distributed process $\{U_i\}_{i=1}^{\infty}$ with mean μ and variance σ^2, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_i - \mu)$ converges in distribution to the Gaussian distribution with zero mean and variance σ^2, as $n \to \infty$ [8].

Assume that the observed image is o and $\mathcal{I}$ is the set of all images. Our goal now becomes maximizing the posterior probability, i.e. the probability that the ideal image is recovered given the observed image. The posterior probability is given by $P(X = x|O = o)$, where $x \epsilon \mathcal{I}$, which can be re-written using Bayes Rule to give Equation 2.

$$P(X = x|O = o) = \frac{P(O = o|X = x)P(X = x)}{P(O = o)} \tag{2}$$

Since we are trying to maximize the posterior probability over x, the denominator of Equation 2 irrelevant. Thus the goal of this design project becomes maximizing the numerator of Bayes Rule, shown in Equation 3.

$$\underset{x}{\text{argmax}} P(X = x|O = o) = \underset{x}{\text{argmax}} \underbrace{P(O = o|X = x)}_{\text{likelihood}} \underbrace{P(X = x)}_{\text{prior}} \tag{3}$$

This newly defined problem raises two questions which have been divided into two phases for the project. Phase 1 is to find the global maximizer x^* that solves Equation 3 for a basic prior probability. This is completed using simulated annealing and constructing an inhomogeneous Markov chain with invariant probability measure that converges to the Dirac on the global maximizer of the posterior probability. This process is described in more detail in Section 2.1.

Phase 2 focuses on estimating a prior probability for the set of images that are text documents. Filters are used to capture important characteristics of clear text images. An example of a text image characteristic is high contrast between the text and background. We wish to find the maximum entropy distribution that matches the observed characteristics of the images in the training set, described in detail in Section 2.2.

2.1 Phase 1

2.1.1 Defining the Image Model

A random field Π on a finite set X is a probability measure that satisfies $\Pi(x) > 0$ for all $x \epsilon X$. A Markov random field is a random field with respect to a neighbourhood structure δ if for all x ϵ X where the condition in Equation 4 is satisfied.

$$\Pi(X_s = x_s|X_t = x_t, t \neq s) = \Pi(X_s = x_s|X_t = x_t, t \epsilon \delta\{s\}) \tag{4}$$

The foundation of the mathematics behind this probabilistic approach to image restoration is modelling the images as arrays of random variables with dependencies on a defined neighbourhood structure. In an example of an image model, seen in Figure 2, the neighbourhood structure is $t_0, ..., t_3$.

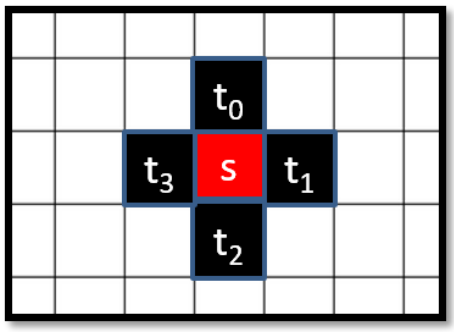


Figure 2: Illustration of a neighbourhood.

In the context of images, assuming Figure 2 is a Markov random field, the conditional probability of site s taking the value x_s, given the value of the other pixels, is equivalent to the probability of the value at this site knowing only the defined neighbourhood, x_{t_0}, x_{t_1}, x_{t_2}, and x_{t_3}.

This design considers greyscale images, therefore the random variables at each site s take values between 0 and 255. We consider an observed image $O = (O_{i,j})_{i,j}$ where $(O_{i,j})$ are random variables. With the observation model discussed above, $O = I + N$, where $I = (I_{i,j})$ is the ideal image, and $N = (N_{i,j})$ is the Gaussian noise process.

Probability measures of the structure of Equation 5 are called Gibbs fields said to be associated with the energy function H, defined in Equation 6. Note that a Gibbs field is a random field.

$$x \longmapsto \Pi(x) = \frac{exp(-H(x))}{\sum_x exp(-H(x))} = Z^{-1} exp(-H(x)) \tag{5}$$

$$H(x) = \frac{1}{2(Var(x))^2} \sum_{s=1}^{\sigma} (o_s - x_s)^2 + \lambda \sum_{s=1}^{\sigma} \sum_{s' \epsilon \delta\{s'\}} (o_s - x'_s)^2 \tag{6}$$

Gibbs fields are essential to the design of this image restoration system since they imply a way to write the joint probability of the random variables as functions over cliques in the graph. By looking at Gibbs random fields rather than Markov random fields you can use Gibbs sampling through conditional distributions rather than the full joint distribution, which is an unfeasible task due to the distribution's size. The Hammersley-Clifford Theorem stated below, Theorem 2.1, allows us to apply the desirable properties of a Gibbs random field to the image model, a Markov random field.

Theorem 2.1 (The Hammersley-Clifford Theorem) *Let a neighbourhood system δ on a set s be given. The a random field is a Markov random field for δ if and only if it is a neighbour Gibbs field for δ (i.e. Π is a Gibbs field for a neighbour potential U).*

2.1.2 Simulated Annealing

Much like in the annealing of metals, simulated annealing "heats up" the energy function and then "cools" it according to a cooling schedule. This process allows for the convergence

to global minimums, instead of local minimums. To ensure the even "cooling" of the energy function, a visiting scheme is used which steps through each element of the image array.

Define a primitive Markov kernel to be a Markov kernel P such that for $r \in \mathbb{N}$, $(P(x,y))^r > 0$ for all $x, y \in X$. Then, the following can be said:

- Every primitive Markov kernel P has a unique invariant distribution ($\mu P = \mu$).
- Given a primitive Markov kernel with invariant distribution μ, $\nu P^n \to \mu$ uniformly in all distributions ν, as $n \to \infty$.

The following theorems are used in the development of Equations 7 and 8.

Theorem 2.2 *Let $(P_n)_{n\geq 1}$ be primitive Markov kernels, each with an invariant measure μ_n. Assume further that the following conditions are satisfied:*

1. $\Sigma_n ||\mu_n - \mu_{n+1}|| \leq \infty$
2. $c(P_i ... P_n) \to 0$ *as* $n \to \infty$ *for all* $i \geq 1$

Then $\mu_\infty = \lim_{n\to\infty} \mu_n$ *exists and uniformly in all initial distributions* ν, $\mu_\infty = \lim_{n\to\infty} \nu P_1 ... P_n$.

where $c(P_i ... P_n)$ are called the Dobrushin contraction coefficients.

Theorem 2.3 *For all $x \in \mathcal{I}$, $\lim_{n\to\infty} \nu P^{(n)}(x) = \Pi(x)$ uniformly in all distribution ν. Where, $\mathcal{I}$ is the set of images and $\Pi(x)$ is the Gibbs distribution.*

Let $\Pi(x) = Z^{-1} \exp(-H(x))$ be a Gibbs random field on $\mathcal{I}$ and introduce a parameter $\beta > 0$, where β is the inverse temperature. We define the Gibbs field at β to be $\Pi_\beta(x) = Z^{-1} \exp(-\beta H(x))$. A cooling schedule is an increasing sequence $(\beta(n))_{n\in\mathbb{N}}$ with $\beta(n) \to \infty$ as $n \to \infty$. For every $n \in \mathbb{N}$, a Markov kernel is defined by:

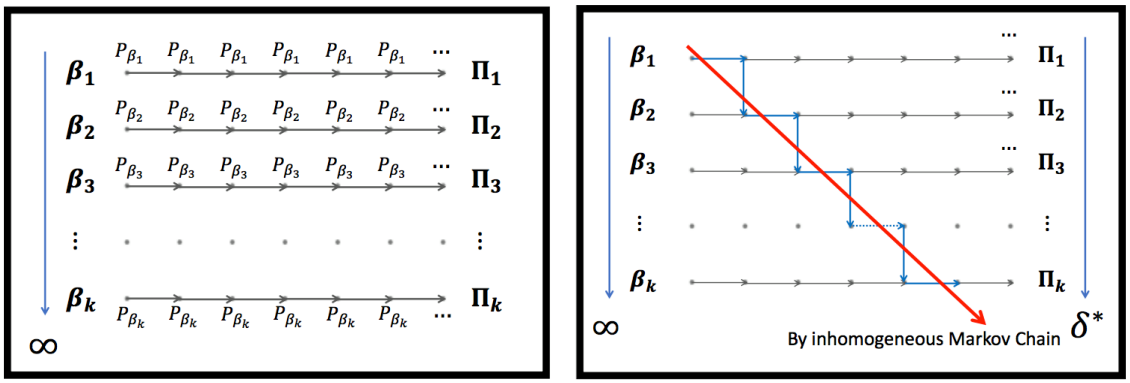
$$P_n(x,y) = \left(\Pi_{\{1\}}^{\beta(n)} \cdots \Pi_{\{\sigma\}}^{\beta(n)}\right)(x,y) \tag{7}$$

where $\Pi_{\{s\}}^{\beta(n)}$ is the single site local characteristic of $\Pi^{\beta(n)}$ for site s.

Then, if the cooling schedule is such that $\beta(n) \propto \ln(n)$, uniformly on all distributions ν we obtain:

$$\lim_{n\to\infty} (\nu P_1 \cdots P_n)(x) = \begin{cases} \frac{1}{|M|}, x \in M \\ 0, else \end{cases} \tag{8}$$

where M is the set of global minimizers of H. Thus, under the assumption that there is only one global minimizer, the ideal image, this converges Dirac at the global minimizer. Figure 3 gives an illustration of the simulated annealing process, where 3a shows Equation 7 in each row, and 3b shows the process used to step to the next row to get the convergence of the Gibbs field.



(a) (b)

Figure 3: Visual representation of simulated annealing.

For another explanation of the construction of the inhomogeneous Markov chain, shown in Figure 3: suppose you think of a die that has "infinite" sides, each with a primitive Markov kernel. When you roll the die and see what value you get, you shave or bias the die based on that value. Then you roll again and continue the process for an infinite number of times. Then by Theorems 2.2 and 2.3 you are guaranteed to have these primitive Markov kernels converge uniformly under all distributions to the Gibbs distribution. When the die is rolled infinite times, these Gibbs distribution converge uniformly to the Dirac, and you eventually end up with the ideal image.

Phase 1 concludes with the construction of this inhomogeneous Markov chain that converges to the Dirac, which is the global maximizer of the posterior probability given a known prior probability distribution.

2.2 Phase 2

2.2.1 Maximum Entropy

Maximum entropy is an important concept in statistics used for constructing probability distributions on sets of random variables since maximizing entropy minimizes bias in the distribution. As we have modelled our images as random variables, the maximum entropy principle provides a good method of constructing a prior probability distribution on our class of images. Given unknown functions $\phi(x)$ with known probability distributions, and letting Ω be the set of all probability distributions $p(x)$ satisfying the constraints, the maximum entropy principle states that a good choice of probability distribution is the one that has maximum entropy subject to the constraints. This is represented by Equation 9.

$$p(x;\Lambda) = \frac{1}{Z(\Lambda)} \exp(-\Sigma_{n=1}^{N} \lambda_n \phi_n(x)) \tag{9}$$

Where $\Lambda = (\lambda_1, \lambda_2, ..., \lambda_n)$ is the Lagrange parameter, and $Z(\Lambda)$ is the partition function. In the case of image restoration, the unknown functions are the histograms of the filtered images.

The λ_i are determined numerically using Equation 10.

$$\frac{d\lambda^{\alpha}}{dt} = E_{P(I;\Lambda_k)}(H^{syn}(\alpha)) - H^{obs}(\alpha) \tag{10}$$

where $E_{P(I;\Lambda_k)}(H^syn(\alpha))$ is the expected histogram of the filtered image and $H^{obs}(\alpha)$ is the histogram of the observed image. Thus, by minimizing the distance between the histograms of the observed and filtered images, the constraints λ_i are minimized and the entropy maximized.

3 Implementation

3.1 Image Input/Output Interface

Both the Gibbs Sampler and the FRAME algorithm require an input image $I(v)$ where v is a particular pixel of the image. A greyscale image was read into Python and converted to a matrix of greyscale 0 to 255 pixel values. The matrix was then read into C and represented as an array $img(x, y)$, where x represents the row and y the column of the read matrix.

Similarly, once the algorithm finished running, Python was once again used to reconstruct the image given an image matrix.

3.2 Gibbs Sampler

Gibbs Sampling is a Markov Chain Monte Carlo algorithm. It allows us to sample values from a probability distribution. Generally speaking, sampling from a joint probability distribution is a difficult task. For instance, a 100×100 image would have a joint probability distribution with 100×100 random variables; it is unfeasible to sample from such a distribution. Gibbs Sampling provides a technique to efficiently approximate the joint distribution by sampling from conditional distributions instead.

Gibbs Sampling is an essential component of both the restoration algorithm and the FRAME algorithm for learning image priors. It was a challenge to implement mainly due to its large run-time complexity.

3.2.1 Gibbs Sampler Algorithm

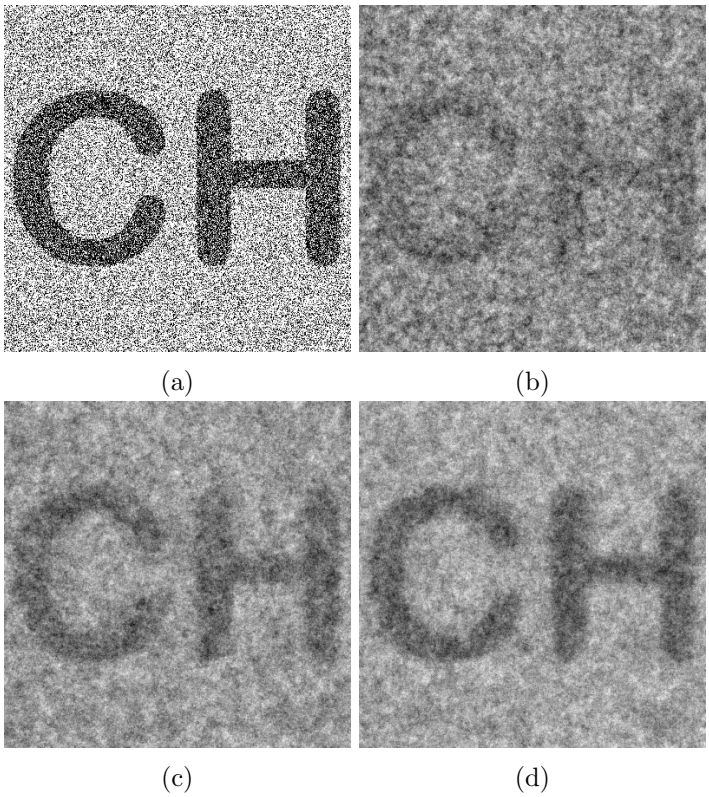
Given image $I(v)$, flip counter $\rightarrow 0$
Repeat
    Increment β according to cooling schedule
    Randomly pick a location v under the uniform distribution
    For $val = 0, ..., 255$
        Calculate $P(I(v) = val|I(-v))$ by $P(val) = Z^{-1}e^{-\beta H(val)}$
    Randomly flip $I(v) = val$ under $P(val|I(-v))$
    flip counter = flip counter + 1
Until flip counter = number of pixels in the image $\times$ W sweeps

3.2.2 Algorithm Discussion

The algorithm itself is not very complicated and can be understood relatively easy; the challenge was achieving a feasible run-time. The algorithm requires sweeping through the image

matrix, $I(v)$, W times. Moreover, at every sweep and at every matrix entry v, the algorithm requires 255×2 further sweeps to calculate the conditional probability $P(val|I(-v))$. This amounts to a high degree of run-time complexity. For instance, if a 100×100 image is used and the algorithm is run for $W = 500$ sweeps, this would amount to $500 \times 100 \times 100 \times 255 \times 2$ total loop iterations.

Theoretically, the more sweeps, the closer the sample is to the true Gibbs distribution. However, the algorithm's run-time complexity increases significantly with W. It was important to determine a minimal value for W that was large enough such that the sampling generated a good approximation of the true Gibbs distribution. The number of sweeps W was determined through iterative testing at varying sweeps and visual result comparison. Figure 4 shows the result of running the Gibbs Sampler for varying sweep values.



(a) (b)

(c) (d)

Figure 4: 4a) input image 4b) 100 sweeps 4c) 500 sweeps 4d) 1000 sweeps

Notice the image has not converged to a good enough approximation of the input image after running 100 sweeps. At 500 sweeps you can see the result is much better than at 100 sweeps, and very similar to the result at 1000 sweeps.

A final note on the Gibbs Sampler is that it can be parallelized since the algorithm works with conditional distributions in local neighbourhoods.

3.3 FRAME

The FRAME algorithm for learning image priors was also a challenge to implement. Naturally, since it makes use of the Gibbs Sampler, its run-time complexity was also very large.

3.3.1 The FRAME Algorithm

Input a texture image I^{obs}
Select a group of k filters $S_k = \{F^1, F^2, ..., F^k\}$
Compute $\{H^{obs(\alpha)}, \alpha = 1, ..., k\}$
Initialize $\lambda_i^\alpha = 0$, $i = 0, ..., 255$, $\alpha = 1, ..., k$.
Initialize I^{syn} as white uniform noise texture
Repeat
    Calculate $H^{syn(\alpha)}$, $\alpha = 1, ..., k$ from I^{syn}
    Update λ^α, $\alpha = 1, ..., k$ by Eq.(5), $P(I; \Lambda_k, S_k)$ is updated
    Apply Gibbs Sampler to flip I^{syn} for W sweeps under $P(I; \Lambda_k, S_k)$.
Until $\frac{1}{2}\sum_1^{255} |H_i^{obs(\alpha)} - H_i^{syn(\alpha)}| \leq \epsilon$ for $\alpha = 1, ..., k$

3.3.2 Algorithm Discussion

The algorithm incorporates multiple image, filter and histograms arrays. The run-time complexity of the algorithm linearly increases with the number of filters used. It was important to find a balance between the number of filters used and the performance of the algorithm.

3.3.3 Filter Selection

Three different filter types were considered:

- Dirac: captures pixel intensity counts
- Gaussian: smooths pixels; represented by $Gaussian(\mu, \sigma)$ where μ is the mean and σ is the standard deviation
- Gabor: captures edge features at varying angles; represented by $Gabor(\sigma, \theta)$ where σ is the standard deviation and θ is the angle

Six specific filters were chosen: Dirac, Gaussian(0,1), Gabor(2,0°), Gabor(2,45°), Gabor(2,90°), and Gabor(2,135°). These filters were chosen intuitively based on what would characterize text well. Then, the filters were applied to text to experimentally validate the chosen filters. Figure 5 shows the resulted filtered images after individually applying each filter.

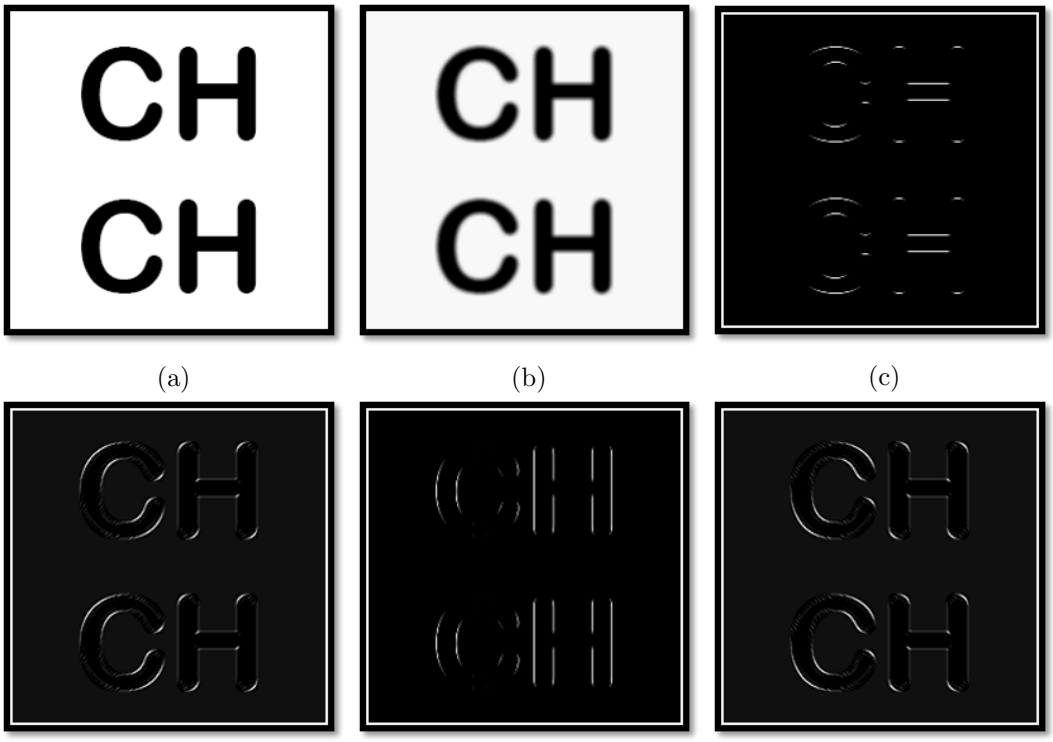


Figure 5: 5a input image 5b) Gaussian(0,1) 5c) Gabor(2,0°) 5d) Gabor(2,45°) 5e) Gabor(2,90°) 5f) Gabor(2,135°).

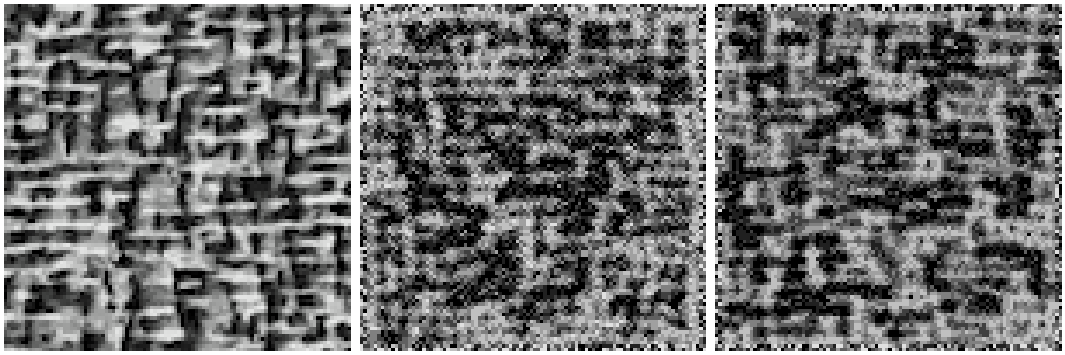
Upon applying the six filters simultaneously to the input image 5a one can see that the six filters do effectively capture the image's features, seen in Figure 6.

Figure 6: Complete filtered characterization

4 Results and Analysis

4.1 Results

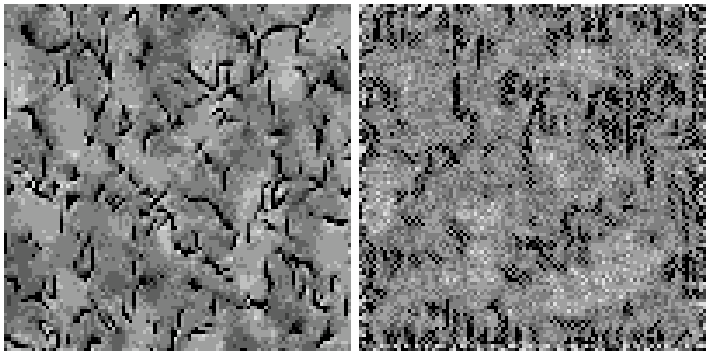
We present the results of our FRAME algorithm below, where various textures have been synthesized. It was important to test our algorithm on different textures to ensure it was working properly. The first texture is shown below in Figure 7a. The resultant synthesized textures with 40 and 60 FRAME iterations are shown in Figures 7b and 7c, respectively. You can see that even after only 40 iterations the pattern is emerging.



(a) (b) (c)

Figure 7: Synthesis of the knit texture.

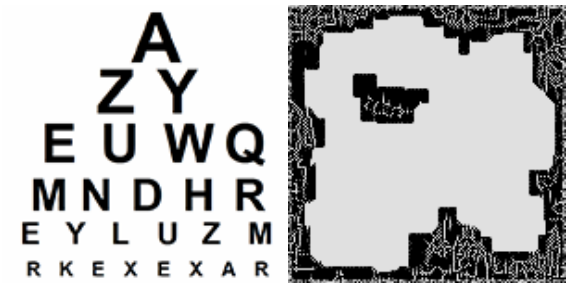
A second texture is shown below in Figure 8a. The resultant synthesized texture with 50 FRAME run iterations is shown in Figure 8b.



(a) (b)

Figure 8: Synthesis of a rock face.

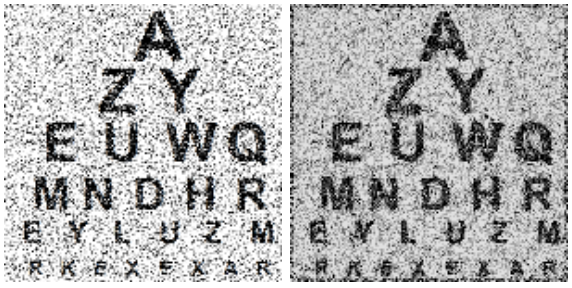
Although the synthesized images are not exactly the same as the input textures, they display characteristics and features that match. Having tested the FRAME algorithm on easily recognizable textures, next a text image was run through the algorithm with 50 iterations. The resultant synthesized text image is shown below in Figure 9b



(a) (b)

Figure 9: Synthesis of the Snellen chart.

Clearly, the synthesized Snellen chart in Figure 9b looks nothing like the original in Figure 9a. However, the important features of the text have been captured. The synthesized image consists of mostly clumps of black regions, with large regions of white. This is consistent with our input text image. Also, there are many horizontal and vertical edges, along with a few diagonal ones. This accurately captures the features of our Snellen chart, as it is composed of mostly horizontal and vertical edges.



(a) (b)

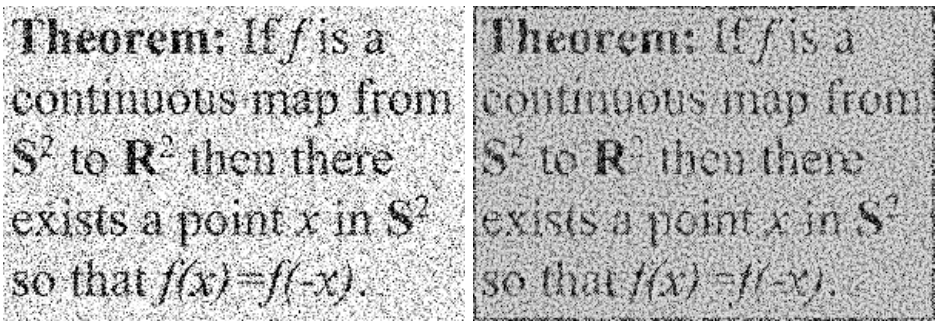
Figure 10: Restoration of a noisy version of the Snellen chart.

As the final part of the project, we used the synthesized texture as our image prior, and attempted to restore a noisy image. The noisy image can be seen in Figure 10a, and the restored image can be seen in Figure 10b. The restoration was completed with 200 Gibbs sweeps, as it was determined that anything beyond that was giving diminishing returns. It is clear to see that the restored image is cleaner and easier to read. For example, the letter "H" on the fourth line had its middle section reconstructed, as well as the left side of the "U" on the fifth line. Not only do the results qualitatively improve, they also quantitatively improve. Below in Table 1, it can be seen that the signal-to-noise ratio of the restored image is greater than that of the noisy image, indicating that the restoration was successful.

Table 1: SNR Analysis

	SNR Value
Original Image	3.5787
Restored Image	3.9380

In Figure 11 below, an example of an incorrect prior is shown. The left image shows the noisy image, and the right image is the attempted restoration. It is clear to see that the text is much less clear, and harder to read.



(a) (b)

Figure 11: Restoration of an image using an incorrect prior.

4.2 Tool Analysis

Two coding languages were considered to develop the program: C and MATLAB. Although MATLAB has a robust directory of built in functions and allows for easier visualization of data, it has a much slower run time than C does due to dynamically allocated variables and calls to external libraries. In C, on the other hand, it is more difficult to develop code, but the run time is significantly lower. Because of the improvements in run time, C was chosen.

Tying into an archival software application, if MATLAB was used as the language, anyone wishing to use this software would also have to have a MATLAB license. This would increase the cost of the software, an individual MATLAB license costs $2,150 [9].

As mentioned in Section 3.3.3, the Gabor filter angles were chosen to be 0°, 45°, 90°, and 135°. This was because the majority of characters in text documents can be approximated with horizontal, vertical and ±45° angles, such as in stencil fonts. This was deemed acceptable since the goal of text restoration is to preserve the information conveyed by the text, and not to preserve the exact font used.

In the development of the image prior, two algorithms were considered: Gibbs and Metropolis. The Gibbs algorithm, discussed above is simple to implement, but slow to run, whereas the Metropolis algorithm is a generalization of the Gibbs Sampler and often has a faster run time [10]. The Gibbs Sampler was chosen, despite the longer run time, because of ease of implementation and reliability.

4.3 Design Improvements

Improvements for the Gibbs Sampler could involve parallelization. This was not possible given the resources the team had available but this could be leveraged to greatly improve run-time.

Further improvements for our design would be to continue following Mumford, Zhu, and Wu and optimize filter selection for Phase 2. This means selecting filters that we know capture the most characteristics. Looking at the current images, it appears that the filters

we use do not adequately capture small edges and features in the image. This could be improved in the future by using a combination of smaller and larger filters to capture both large and small features. It would also be beneficial to optimize the code, since it takes approximately 10 hours to get a single output for Phase 2. This greatly hinders progress in the testing and debug stages.

An additional improvement could be to use multiple images in our texture synthesis stages. Currently only a single image is used in the synthesis, making our prior very specific. Using multiple images would help the algorithm learn a more general texture of English text.

Limitations for our proposed solution are that this only works for text documents, it would be beneficial if it could work for graphs and other symbols such as sheet music. Also, the resulting PDFs from our system are not "machine-readable" meaning the text cannot be searched.

5 Engineering Considerations

5.1 Standards

Digital archive systems, if implemented in a national record centre, hospital, or other institutions that deals with personal information must comply with the Personal Information Protection and Electronic Documents Act (PIPEDA). To do so, a compliance officer should be appointed and a review of the organizations current practices concerning the collection and storage of personal information should be conducted [11]. An electronic privacy policy should be developed based on the results of the review. PIPEDA also recommends updating forms used to collect information and setting up a protocol to deal with opt-outs [11]. All employees, volunteers, and directors should be trained on the new privacy policy and sign confidentiality agreements if they deal with any personal information [11].

5.2 Social Factors

As society moves away from hard copies and towards a digital world the necessity of recording old documents digitally is becoming more apparent. The digitization of documents makes them more accessible to the general public [12]. This results in many societal benefits, such as remote access to university library documents for students and faculty. If weather conditions are dangerous, staying at home to complete their work is now an option. Another benefit to digitization is online copies of documents, like medical records, making them more portable and less likely to be lost [13]. Digitization also poses a threat to sensitive information if cyber security is breached by hackers. Keeping digital records poses a risk of lost information from corrupted files and system failures [5]. Thus, there is a trade-off of between having accessible, portable information and its safety.

An image restoration system would be useful for devices like smart-phones and tablets. Studies have shown a correlation between eyestrain and the use of back-lit tablets, such as e-readers [6]. This can be avoided through the use of e-readers whose screens mimic the surface of paper [6]. The increased use of tablets, cell phones, and e-readers also have a negative impact on the posture of those using the devices [14]. With new technology being

put out on the market rapidly there has been an increase in electronic waste, referred to as E-waste. E-waste is connected health risks from contact with harmful materials and inhalation of toxic fumes [15].

Many studies on reading habits conducted in the United States over the past few years have shown significant decline in daily reading of individuals under the age of 24. Specifically those aged 18 to 24 read 55% less than the total adult population [16]. That being said, the number of people choosing to read on alternate devices such as e-readers, computers, and cell phones were not included in these studies [16]. In fact, studies have been conducted that show the use of e-readers improves reading among boys and those typically reluctant to read [17]. As well, e-readers have been shown to improve reading speed and comprehension in children with dyslexia [18]. Although the improved reading results are linked to there being fewer words per line when reading on e-readers, the digitization of documents can allow customization of the number of words per line, hence increasing the resources available to those with dyslexia [18]. Despite the health risks associated with reading devices, this design project could lead to an increase in voluntary reading [19].

5.3 Environmental Factors

On average, every person in the United States uses 9 trees, equivalent to 680lbs, worth of paper per year [20]. Pulp and paper mills produce effluents that flow into lakes and rivers. This waste has been proven to have negative effects on fish reproduction performance and metabolism [21]. Negative effects on ecosystems caused by paper production can be mitigated by society moving towards digital documentation.

Electronic books do not require paper or ink, and can be distributed an unlimited number of times with no added costs or inputs [22]. Due to the initial large carbon footprint of tablets, such as iPads, a user must read 32 - 42 books on the tablet for the average amount of carbon dioxide produced per book to match that of a paper book [23]. That being said, a life cycle analysis that was conducted on e-readers showed that they produced 75% less kilograms of carbon dioxide in their lifetime than paper books [24].

A possible argument is that manufacturing digital storage devices cause more harm than the paper alternative. However, the storage infrastructure is already there, and is continuously being manufactured independent of the publishing world. Another consideration is that the resources required to store the books, such as maintaining adequate temperature and humidity, requires a significant amount of energy. In Ontario, 65% of electricity is generated via nuclear [25]. Although nuclear is one of the most eco-friendly power generation methods, there are still negative impacts to the environment in terms of running the plants and transporting the electricity.

For the image restoration system, two types of waste were considered: paper waste and electronic waste. Paper is susceptible to degradation but paper books remain relevant longer than electronics. This is because, if the properly taken care of, books can survive for decades whereas the average life span of a tablet or smart-phone is 4.6 - 5.1 years [26]. This rapid increase of new technology results in an increase in E-waste and poses a threat to the environment from the accumulation of chemicals in soil, water, and food [15]. Thus, the implementation of this application can contribute to the negative effects of E-waste. On the other hand, if the image restoration system is used to scan and destroy libraries and

archives there will be an in increase in paper waste. If recycled properly, this paper waste can be reused and save trees. If every American recycled one-tenth of their newspapers 25,000,000 trees would be a year [27].

5.4 Economic Analysis

The average cost of a paper book is approximately $14.00 - $25.00 whereas an E-book can cost anywhere between $0.99 - $12.99 [5]. There is also the initial cost of $400.00 - $1500.00 for the actual reading device [5]. If this is considered a sunk cost then digitizing novels is economically beneficial to society. When companies or individuals turn towards digitization, another consideration is the feasibility of implementing a new software system. Existing systems in place for hard-copy file storage may be functioning for the stakeholders and buying a new digital system can be costly both in implementation and training of employees.

The economics of this solution is split into two parts: the development of the software and the implementation of the software. Since this software will have to be relatively complex (read images, modify images, write images), the development of it will cost between $100,000 and $250,000 [28]. Implementation depends on the size of the project and the existing infrastructure. A larger project will require more personnel, larger storage capacity, and a longer transfer time, and thus will cost more than a smaller scale project. For example, to implement the *Eloquent Archives* software program, the initial purchase is between $4,000 and $35,000 depending on the size of the projects, with an annual maintenance cost of 15% of the purchase cost [29].

More difficult to quantify is the economic impact of the restoration program used to pirate print documents. This is detrimental to both publishing companies and authors as there is very little that can be done to stop people from borrowing a library book, scanning it at home, and returning it to the library. An indirect economic impact that is difficult to quantify is how companies and individuals utilize space efficiently [30]. If they are converting and destroying the physical copies of manuals, employee documents, or personal files to digital this can result in vasts amounts of space, that was being used to archive these documents, becoming available. Utilizing rented space in an effective business way can result in economic benefit to a company [30].

5.5 Trade-offs

When considering specific applications of the restoration program trade-offs can be made in robustness, speed, and clarity. An individual may not care as much about the the quality of the document, and may place more emphasis on the speed of the program. Corporate use for documents of a specific style, say papers that follow the same page layout, would not need a high level of robustness in recognizing different layouts, and thus could be optimized for a quicker run-time. Increased robustness would mean a larger base of image priors, thus resulting in a longer development time. These trade-offs could be accommodated by marketing different versions of the software.

5.6 Ethical Factors

Digitization and restoration software is making it easier for people to access information, but also making it easier to pirate intellectual property such books, papers, and textbooks. If a restoration system was to be implemented, further work would have to be done on security of the restored information. This could be accomplished releasing different licences: a corporate license and a general license. The corporate license would allow an unlimited number of pages to be scanned into one document so that archives can preserve the full document, whereas the general license would put a cap on the number of pages allowed to be scanned into one document to hinder the scanning of full books.

Since this technology could be used to restore poorly scanned versions of pirated books, it poses an ethical dilemma for both the designers and the users. One must consider the trade-off between the potential for abuse and the benefits it gives legitimate users when implementing the system for public use [31].

6 Conclusion

In conclusion, the issue of poorly scanned documents was identified. It was determined that probabilistic methods would be successful in restoring these images. Engineering tools were used to implement and verify the results obtained. After an analysis of the engineering considerations, it was determined that before a wide scale restoration system could be implemented policies would have to be put into place to ensure the safety of the information being stored.

Index

References

[1] S. Markovitch and P. Willmott, "Accelerating the digitization of business processes," May 2014. http://www.mckinsey.com/business-functions/business-technology/our-insights/accelerating-the-digitization-of-business-processes.

[2] M. Seery, "Saving paper," *Royal Society of Chemistry*, March 2013. http://www.rsc.org/education/eic/issues/2013March/paper-conservation-cellulose-acid-hydrolysis.asp.

[3] R. W. Brockett, "Finite dimensional linear systems," *John Wiley and Sons, Inc*, 1970.

[4] J. M. Besek, P. S. Loengard, and J. C. Ginsburg, "Maintaining the integrity of digital archives," *Kernochan Center for Law, Media and the Arts*. http://web.law.columbia.edu/sites/default/files/microsites/kernochan/files/MELLON-Final-report.pdf.

[5] S. McKay, "Digitization in an archival environment," *ELECTRONIC JOURNAL OF ACADEMIC AND SPECIAL LIBRARIANSHIP*, vol. 4, no. 1, 2003. http://web.law.columbia.edu/sites/default/files/microsites/kernochan/files/MELLON-Final-report.pdf.

[6] S. Benedetto, V. Drai-Zerbib, M. Pedrotti, G. Tissier, and T. Baccino, "E-readers and visual fatigue," *PLOS*, December 2013. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0083676#s3.

[7] "Survery of image denoising techniques," *University of Nevada, Reno*, 2004. https://www.cse.unr.edu/ fredh/papers/conf/034-asoidt/paper.pdf.

[8] F. Alajaji and P.-N. Chen, "Information theory for single-user systems," 2016. http://www.mast.queensu.ca/ math474/it-lecture-notes.pdf.

[9] MATLAB, "Pricing and licensing," 2017. https://www.mathworks.com/pricing-licensing.html?prodcode=ML&intendeduse=comm.

[10] G. Winkler, *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer-Verlag Berlin Heidelberg, 2003.

[11] D. Edwards, "Personal information protection and electronic documents act," 2005. http://www.omafra.gov.on.ca/english/rural/facts/05-049.htm#5.

[12] "Digitization," *Smithsonian Institution Archives*. http://siarchives.si.edu/services/digitization.

[13] "Digitizing healthcare: Why having an electronic medical record matters," *Possible Health*, March 2015. http://possiblehealth.org/blog/electronic-medical-record/.

[14] Yassierli and A. Juraida, "Effects of netbook and tablet usage postures on the development of fatigue, discomfort and pain," *Journal of Engineering and Technological Sciences*, vol. 48, pp. 243 – 253, 2016. http://journals.itb.ac.id/index.php/jets/article/view/1605/1279.

[15] "Children's environmental helth: Electronic waste," *World Health Organization*, 2015. http://www.who.int/ceh/risks/ewaste/en/.

[16] J. E. Moyer, ""teens today don't read books anymore": A study of differences in interest and comprehension based on reading modalities: Part 1, introduction and methodology," November 2010. http://www.yalsa.ala.org/jrlya/2010/11/teens-today-dont-read-books-anymore-a-study-of-differences-in-interest-and-comprehension-based-on-reading-modalities-part-1-introduction-and-methodology/.

[17] Å. K. Tveita and A. Mangena, "A joker in the class: Teenage readers' attitudes and preferences to reading on different devices," *Library & Information Science Research*, vol. 36, pp. 179–184, November 2014. http://www.sciencedirect.com/science/article/pii/S0740818814000516.

[18] M. H. Schneps, J. M. Thomson, C. Chen, G. Sonnert, and M. Pomplun, "E-readers are more effective than paper for some with dyslexia," *PLOS*, September 2013. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0075634.

[19] A. G. Larson, "Reading habits among students and its effect on academic performance: A study of students of koforidua polytechnic," *Library Philosophy and Practice*, May 2014. http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=2908&context=libphilprac.

[20] S. Kaufman, "Paper facts," *Cleveland State University*, 2008. http://cua6.urban.csuohio.edu/ sanda/syl/envpolmaterials/GREEN[Accessed 29 November 2016].

[21] M. Hewitt, "A decade of research on the environmental impacts of pulp and paper mill effluents in canada: Sources and characteristics of bioactive substances," *Toxicology and Environmental Health*, 2007.

[22] R. D. Selby, K. P. Carter, and S. H. Gage *International Journal of Sustainability in Higher Education*, vol. 15, no. 2, pp. 142–156, 2014. http://journals2.scholarsportal.info/pdf/14676370/v15i0002/142_scet.xml.

[23] G. P. Initiative, "Environmental impacts of e-books," 2011.

[24] G. L. Kozak, "Printed scholarly books and e-book reading devices: A comparative life cycle assessment of two book options," 2003.

[25] "Electricity generated in ontario," *Canadian Nuclear Society*, 2016. https://www.cns-snc.ca/media/ontarioelectricity/ontarioelectricity.html.

[26] "The average life span of 7 popular tech products," *Specout by Graphiq*, 2015. http://smartphones.specout.com/stories/9635/average-lifespan-tech-products.

[27] U. of Southern Indiana, "Paper recycling facts," https://www.usi.edu/recycle/paper-recycling-facts/.

[28] J. Flackett, "How much does it cost to build software application?," 2015. https://www.linkedin.com/pulse/how-much-does-cost-build-software-application-dr-john-flackett.

[29] L. Spiro, "Archival management software," *Council on Library and Information Resources*, 2009. https://www.clir.org/pubs/reports/spiro/spiro/spiro_Jan13.pdf.

[30] L. Cameron, "The impact of digitization on business models in copyright-driven industries: A review of the economic issues," *National Research Council Committee on the Impact of Copyright Policy on Innovation in the Digitl Era*, 2013.

[31] Ernesto, "Anti-piracy patent stops students from sharing textbooks," *Possible Health*, June 2012. https://torrentfreak.com/anti-piracy-patent-prevents-students-from-sharing-books-120610/.