Low-Cost, On-Demand Film Digitisation and Online Delivery

Matt Garner, matt.garner@findmypast.com

Abstract

Hundreds of millions of pages of microfilmed historical documents are not being digitised at this time due to insufficient individual demand to garner appropriate commercial attention and investment. This paper demonstrates that the cost of digitisation and online delivery can be lowered dramatically using a novel application of recent technological advancements in imaging, data processing and storage. A business model is presented such that an on-demand service can be provided whereby an individual end user can afford to personally sponsor the digitisation and online delivery of an entire reel of film. Such a model could co-exist in the market with the large scale digitisation efforts and provide an invaluable service by removing access barriers to the entire breadth of film archives.

1. Introduction

Hundreds of millions of pages of diverse, genealogically relevant documents from all corners of the earth have been successfully captured during the last century onto microform. Modern advancements in digital imaging, processing, online storage and distribution, however, have created new, completely digital document storage and distribution channels which are technologically superior. Although the volume of digitised records is growing rapidly, only a small percentage of what is available on film is now online.

The boom of online record availability is fuelling a global, rapidly expanding public interest in family history as the convenience of accessing source documents electronically is increasing. While many organisations are involved in the digitisation process, the logical order of scanning employed by each is principally concentrated around those document collections that have the largest general interest (such as census and vital records collections) and thus present the largest immediate return on investment. Due to the high cost of film digitisation and transcription, this effectively creates a fairly fixed break-even point where many previously filmed collections won't qualify for commercial digitisation until there comes a change in present market conditions.

In this paper, I present a selection of results from ongoing research and development concerning a business model using a combination of recent technological advances that have the potential to reduce the cost of digitisation of a roll of film into the range that a single individual could personally and directly finance. In such a model, individual films could be feasibly digitised on a purely *on demand* basis which is a complete reversal of the paradigm of current commercial digitisation efforts. Such a model could be successfully operated within a purpose-built small business setting; however, due to economies of scale, it is not likely that a single end user could adapt the proposed technologies and processes on a small scale and realise cost savings.

2. Identifying the Cost Points of Digitisation and Online Distribution/Delivery

Assuming ready access to the microform material, the first obvious cost is the digital imaging of the film itself. However, additional costs are principally incurred by post-processing and compressing the images, offline storage, labour and administrative overhead. Online delivery additionally requires additional image compression, online storage and considerable amounts of bandwidth, and optionally, transcription.

Numerous advancements have lowered the costs across the board. Advanced optics and high performance CPUs are less expensive than ever before. Bandwidth availability is growing rapidly worldwide as the cost per megabit is dropping exponentially. Cumulatively, these ever-lowering cost points can be leveraged by to substantially lower the cost of digitisation.

3. Commercial Film Digitisation

Film digitisation requires highly specialised, purpose-built equipment. Due to the very small market for automated film digitisation equipment; and also for the fact that this small market is dominated by enterprise and government players, commercial and industrial digitisation machines tend to be extremely expensive. For example, the top-of-the-line NextScan Eclipse™ range of scanners can run from $60,000 (1) to $120,000 or more per machine.

A close examination of a number of different types of microfilm digitisation equipment showed that they all contain fairly simple parts and all essentially operate on the same basic principles. I was greatly intrigued by this and began designing and prototyping a low-cost, automated film scanner as a proof-of-concept exercise.

4. Low-Cost, High-Quality Optics

In accordance with Moore's law, both consumer and commercial digital imaging has advanced tremendously over the last several years (2). Digital cameras with double-digit megapixel size sensors are now readily available. High-end commercial digital cameras also include self-focusing and powerful zoom optics, as well as the ability to be mated with custom lens arrays. They also include software development kits (SDKs) that allow precision, real-time imaging from a PC.

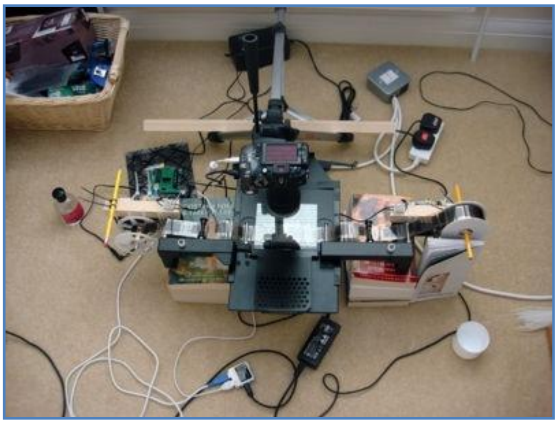


Figure 1. *Mark I* Proof-of-Concept Auto Film Scanner

The early *Mark I* proof-of-concept scanner was built around the Canon S5 IS 8-megapixel professional digital camera. It included a number of USB controlled motors, a single-wavelength LED light source and a high-definition, multi-element macro lens array.

Initial tests proved very promising. Each image frame was being systematically captured into a series of RAW bitmap files. Ultrasonic micro motors driving the lens array allowed for on-the-fly, hardware controlled automatic adjustment on the order of fractions of a millimetre permitting ultra-precise focusing. The shutter speed and aperture can also be continuously and programmatically adjusted to provide the best exposure possible during a film run.

The effective resolution (measured in DPI relative to the surface of the film, being the product of the magnification lenses and native CCD resolution) could be substantially adjusted up or down by modifying the zoom lens position. The maximum effective resolution was near 9600 DPI; however, since the aspect ratio of the CCD device is fixed, zooming to such a high resolution caused undesirable frame cropping. The maximum usable resolution is thus determined by the aspect ratio of the document being imaged. A full frame document, for example a full sheet of newspaper, could be imaged at approximately 2200 DPI, which is essentially the minimum resolution possible. In contrast, terms of a Letter-sized source document, this is equivalent to approximately 200 DPI.

The 8 megapixel S5 IS was later replaced with a 12 megapixel Canon G9. The higher resolution camera increased the maximum Letter-frame resolution to approximately 320 DPI resulting in the output bitmaps of 4000 x 3000 pixels. Similarly, the new Canon G10 14.7-megapixel camera could be installed out-of-the-box further increasing the resolution and output bitmap size to 4415 x 3312 (3) and a projected resolution of 390 DPI.

5. Scanning Automation

A key goal of the research project was the development of complete start-to-end automation which would minimise costly operator supervision. Operational tests revealed that the biggest challenge was thus designing and coding an intelligent motor and camera control algorithm that ensured that every single frame on the film was captured without missing or improperly cropping any frames while still attempting to achieve the highest throughput and efficiency possible. The integrity of the capture process must also be able to be continuously verified. A number of computer vision algorithms were adapted to this purpose.

The camera, while not capturing an image, provides a low resolution, high frame rate live image stream which allows the management application to effectively *watch* the film advance, track film items through the view and to stop the film and capture high resolution images at the appropriate intervals. This eliminates any need to pre-scan a reel to program specific film stops for the camera.

A subset of images from this continuous stream can be recorded and reconstructed into a low resolution, continuous bitmap ribbon. This ribbon, a visual representation of the reel as a whole, can be programmatically checked for consistency as a secondary verification that all frames have been properly captured. If a problem is detected, this data can also be used to advance the film to the proper position for examination by a human operator.

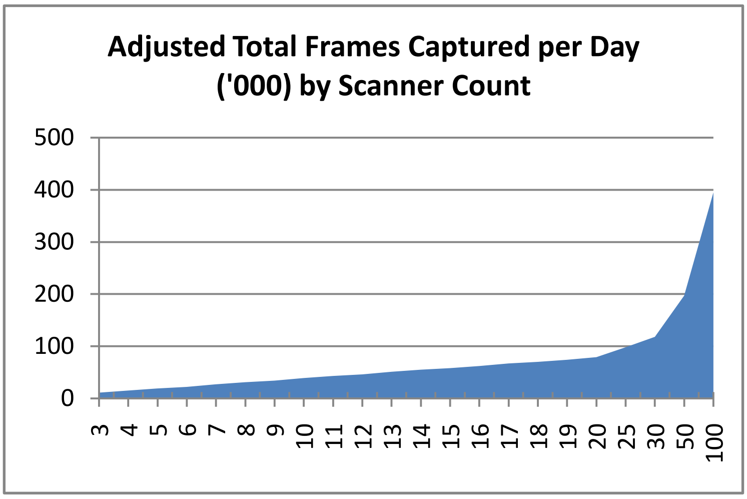
This technique, however, should not be confused with ribbon scanning in earnest, where the reel primarily imaged at high resolution as a continuous stream. Such techniques result in substantially faster frame rates; however, they also require exponentially more expensive processing equipment and high-speed, high-resolution industrial cameras which may cost from $16000 per unit to achieve the same output resolution (4). (I realise that potentially a linear CMOS sensor (i.e. 3000+ pixels by 16 wide) would be better suited at potentially a lower cost; however, further research is outside the scope of this project.)

6. A Massively Parallel Scanner

By developing technology in-house, hefty technology licensing fees can be avoided while still delivering a product of comparable quality. However, there are some caveats that require a further change of paradigm.

The film scanner developed for this project currently only scans 12 pages per minute (ppm) compared to 200ppm on industrial, production scanners. However, the complete system, including the accompanying PC workstation, can be manufactured for well under $2000. Therefore, several dozen, if not hundreds, of identical scanner units can be manufactured for the same price as purchasing a single high-speed unit. While a single reel may take a few hours to read, many reels can be read simultaneously. For instance, 20 machines could have an output of roughly 200ppm at a fraction of the cost.

There are several advantages to having several banks of scanners, rather than investing in single, high capacity units. Since the scanners typically only need supervision during film loading or unloading, a single operator can effectively manage several machines, sequentially moving from one to the next, avoiding idleness. Additionally, plenty of spare parts and even spare units are on hand in case of any device failures. There is added overhead in operating and managing such a system; however, the dramatic cost savings over the traditional approach more than justify the added expense.



Obviously, further industrial and electrical engineering is needed in order to prepare the scanner for industrial use. It must be designed into a single, rugged unit that can survive continuous production usage. It must also be able to be manufactured and serviced at an appropriate scale.

7. Image Post Processing on an Inexpensive Cluster

Once captured, the images require a number of post processing techniques prior to readiness for online delivery. These techniques include channel levelling, conversion to greyscale, image inversion, image rotation, frame border cropping and digital correction of lens effects such as barrel distortion and vignetting. All of these adjustments can be programmatically automated; however, image orientation varies from film to film and may even vary from item to item within a reel. The proper orientation is detected where possible, but image rotation may need to be adjusted by the end user.

A single roll of film may require several gigabytes in its uncompressed form of storage. In preparation for delivery over the Internet, these images must be substantially compressed. In conjunction with the image processing, this is computationally very expensive and may require up to a minute per image depending on the processor speed and compression algorithm used.

In order to tackle this problem, I was able to adapt a highly flexible, cross-platform distributed application platform that I had originally designed for data mining and record linkage. This parallel processing platform operates by using a client program to connect to the local cluster server, download individual job instructions, process a job (image) and save the result to the storage server.

In the ideal environment, as images are captured, they are transmitted over a gigabit LAN from the scanner workstations to a multi-terabyte local network storage server which is, in turn, accessed by the distributed network.

In my lab, I was able to take a dozen older Pentium 4 desktop machines that were due to be disposed of and, using Linux and network booting, quickly convert them into full-time processing cores. The platform management software allows me to manage the entire cluster as a single, powerful machine and run various near real-time image enhancement and compression jobs without requiring a significant investment in high-powered computational equipment or cluster administration.

I was also able to adapt the platform to operate successfully on Amazon Web Services™ EC2 Computing Cloud (although due to bandwidth constraints, this would not be feasible for image processing (5)) as well as configure the client to run as a low-priority background process on across our regular office workgroup of Windows-based workstations.

Figure 2. Part of the Linux Processing Cluster

The net result of the image processing is a folder containing all the sequentially-numbered images from a film roll in JPEG format as well as a secondary folder of thumbnail images also in JPEG format.

8. Online Delivery

Spinning storage is very inexpensive. Single SATA drives are now available at under a penny per gigabyte (6). Ultra high capacity, multi-terabyte RAID systems can be built for under $2000. Such high data density lowers the rack space requirements and, in turn, data centre fees. Fully managed data hosting services ("virtual storage"), such as Amazon's S3 Simple Storage Service™, are fairly inexpensive alternatives as well (7).

Adobe Flex™ technology provides the building blocks of a low-cost, user-friendly graphical user interface (8). At the very least, a user is able to view a virtual reel of thumbnails and then view individual full size frames, simulating the experience of using a film reader directly.

Document transcription opens up a large number of additional user-experience opportunities. However, transcription is outside the scope of this document as the goal is simply to reproduce the film reader experience remotely and electronically at the lowest cost possible.

Moving the processed images from the processing cluster to the online storage presents an additional challenge. Ideally a substantial amount of inexpensive bandwidth is available between the cluster and the web servers; however, this is unlikely to be the case. Therefore, the less expensive alternate is to aggregate multiple image collections on to an external hard drive at the cluster and then physically transport and connect it to the web server cluster on a regularly scheduled rotation. However, the cost of this offline method is directly proportional to geographic proximity of the two locations and cost of the associated manual labour.

9. The On-Demand Model from a User's Perspective

Thus far, I have examined a number of technological processes involved in the digitisation and online delivery of a roll of film. Below is a suggestion of the accompanying delivery process in relation to an end user:

1. An end user searches an online catalogue of the archive, library or repository.
2. The user identifies a film of interest and pays an affordable, up-front fee.
3. The user is told that they will be notified by email in a reasonable period of time (one to two weeks, for example) when the film becomes available online.

4. The film is checked out from the repository and transported to the scanning facility.
5. The film is imaged, enhanced, compressed and copied to online storage.
6. An email is despatched to the user notifying them of film availability.
7. The user logs into the website and accesses a Flex application that displays the images.
8. The original film is returned to the repository.

There are many extensions possible to this model, including the provision of digital rights management (DRM) which may limit, for example, the time the film is available for online access, how many images can be saved to disk or printed, etc., in order to comply with copyright restrictions. The resulting images may be continuously available on online storage to the requesting user, made available to other users or may eventually be moved to an offline digital archive.

10. On-Demand in Practice

Using a combination of the above technologies, as much automation as possible, and with a sufficient volume, I believe it is commercially feasible to lower the price of the high quality digitisation of an entire reel (even a very long one) and delivering it online for even $10 or less per reel. This cost could be directly passed on to the end user and would be comfortably in the range of most family historians. On-demand digitisation would eliminate their potentially lengthy wait for online access to their more obscure materials of interest. This also opens up digitisation options to smaller libraries that otherwise could not justify a large scale digitisation project.

Since most operational expenses are fixed costs based on built-out capacity, according to my fairly conservative financial models, a small operation with ten scanners could run profitably at the $15/reel price point by digitising a total of as few as 25 films per day. With a larger upfront investment/initial build-out and sufficient demand/volume, the price can be reduced further according to economies of scale.

Potentially a larger risk, however, is being overwhelmed by demand. Excessive demand will increase the delay between order and fulfilment due to the fairly fixed nature of the maximum operational capacity. Excessive demand may be mitigated by adjusting the price upwards or by running the scanners on additional shifts.

The on-demand model is complementary to large scale digitisation efforts. Due to the comparatively higher expense per record and lack of transcription/searchability provided by the on-demand model, the large scale digitisation projects will generally yield the most usable product and best value to the end customer. However, since only a limited number of records are commercially viable for large scale treatment, the on-demand method is best suited for providing access to the remainder of the records.

11. On-Demand in the Future

There is, however, a fairly fixed amount of filmed material available as new material is now being captured directly onto digital media. Therefore, the duration of viability of the business model is limited depending on how rapidly the industry as a whole completes digitisation. However, due to the vast amount of pre-digital materials currently available, it is apparent in my financial models that, in most likely scenarios, the entire capital expenditure can be successfully recouped within two years. After the complete return of investment, the business can eventually wind down and wrap up in the black. It is also evident, however, that the longer one waits to enter the marketplace, the lower the ultimate return on investment.

12. Summary

The cost of digitisation and online delivery can be lowered dramatically using a novel application of recent technological advancements in imaging, data processing and storage. Accordingly, the on-demand film digitisation and online delivery model is now commercially viable. Successful application of the model will provide online access to millions of films that are not otherwise likely to become available online for many years. Such would become an invaluable service to family historians worldwide as record accessibility is dramatically improved and both the physical and temporal access barriers to microfilm libraries and archives are overcome. The time is right for this important development to proceed and I am excited to see what lies ahead.

References

1. NextScan Eclipse 300 Roll Film Scanner. *ScannerTraders.* [Online] [Cited: 3 February 2009.] http://www.scannertraders.com/store/index.php?target=products&product_id=29843.

2. **Myhrvold, Nathan.** Moore's Law Corollary: Pixel Power. *New York Times.* [Online] 2006 June 7. [Cited: 19 February 2009.] http://www.nytimes.com/2006/06/07/technology/circuits/07essay.html?_r=1.

3. Canon PowerShot G. *Wikipedia.* [Online] [Cited: 18 February 2009.] http://en.wikipedia.org/wiki/Canon_PowerShot_G.

4. Tohsiba Teli CSC12M25BMP19 CleverDragon. *Aegis Electronic Group.* [Online] [Cited: 18 February 2009.] http://www.aegis-elec.com/products/Teli-CSC12M25BMP19.html.

5. Amazon Compute Cloud (EC2). *Amazon Web Services.* [Online] [Cited: 19 February 2009.] http://aws.amazon.com/ec2/#pricing.

6. Seagate Barracuda 1.5TB Hard Drive ST31500341AS. *TigerDirect.com.* [Online] [Cited: 29 January 2009.] http://www.tigerdirect.com/applications/SearchTools/item-details.asp?EdpNo=4138742&Sku=TSD-1500AS.

7. **Roach, Anne.** Keeping it Spinning: A Background Check of Virtual Storage Providers. *Family History Technology Workshop.* [Online] 2009 March 9. [Cited: 18 February 2009.] http://fht.byu.edu/prev_workshops/workshop08/papers/2/2-3.pdf.

8. Adobe Flex. *Wikipedia.* [Online] [Cited: 19 February 2009.] http://en.wikipedia.org/wiki/Adobe_Flex.