

Preserving Original Image Content in High-Resolution Digital Archives

Mark H. McCormick-Goodhart
Smithsonian Institution, Washington D.C.
Joseph Holmes
Consultant for the Museum of Modern Art, New York, NY

Abstract

High-resolution digital image archives have been an unrealistic goal to date because the large file sizes generated for each image have taxed both computer processing power and storage requirements. The new "FlashPix™" file format dramatically alters the relationship between file size and access time, and projects are under way at the Smithsonian Institution and the Museum of Modern Art to evaluate new approaches to creating high quality digital images of objects in our museum collections. Strategies and benefits of high-resolution digital imaging for museums and archives are discussed in this paper.

Introduction

Most museums and archives now have digital imaging projects under way, often to establish a presence on the Internet and to extend the museum community's public outreach effort. Another goal is related to publication and exhibitions. "In house" digital image preparation can eliminate the need to send valuable originals off site for conventional reprographic work. High quality scans can be generated on demand. The entire collection need not be scanned nor does this application necessarily require high storage capacity. The creation of a high-resolution digital archive is a more ambitious goal. High-resolution images can not only be down sampled to "catalog" the collection but also serve the publishing and exhibition needs for reproductions of the originals. Many proponents of digital image projects believe that a digital image database can help preserve collections by reducing damage caused by frequent handling of the original objects. However, the ease of searching a digital image database may also lead to increased use of the original objects due to greater public awareness of the collection. Both points of view are valid, and the actual outcome depends on the nature of the collection, the collections care policies, and the scope and quality of the digital image project.

If digital imaging is to genuinely aid preservation of the original collection, then the scope of the project must be comprehensive. A digital image program that is only partially completed or one that covers only a fraction of the original collection cannot adequately replace the prior methods of cataloging and collection use. Physical use of the originals continues unabated and no lessening of handling damage occurs. Image quality is also important because the lower the quality the more likely that museum and archives patrons will require access to the original items. Highresolution digital images are clearly desirable, but only if cost and associated access time is reasonable. Highly detailed digital images, for example, will frustrate scholars attempting to use them, if the simple act of opening the record prompts the computer to display "egg timers" and "progress bars" that do not rapidly disappear.

The relationship between image quality, storage capacity requirements, image capture costs, access speed, necessary staff expertise, etc., are constantly changing as computer graphics capability improves. In this regard, it is gratifying when major technical advances take place in the field. One such advancement has recently occurred. We are referring to the new image file format "FlashPix™", developed jointly by Live Picture Inc., Eastman Kodak Company, Hewlett-Packard Company, and Microsoft Corporation, and also to the FITS file technology available in the image editing software, "Live Picture", from Live Picture, Inc.;[3]

This technology has major implications for the way museums and archives should now approach digital image projects. The FlashPix format and the IVUE format from which FlashPix™ is derived fundamentally alter the relationship between file size and access time which has severely limited the museum and archive use of high-resolution files to date. Additionally the FITS technology helps to retain the historical integrity of the original object by reducing the need to track multiple versions of digital files as the information is altered or transformed to meet the needs of various exhibition and publishing activities. Simply stated, digital images often get manipulated, degrading their fidelity to the original record. Museums and archives have a responsibility to maintain the integrity of the original record, and hence an image archive that can simultaneously retain the original scan and altered versions of it are highly desirable. The FITS and FlashPix™ technologies greatly facilitate this capability.

File Size, Resolution, and Effective Image Detail

Before planning a digital archive, one of the most important aspects to understand is the relationship between the number of pixels used to represent an image and the level of detail that can be seen in the representation. If an original image

were infinitely sharp and detailed, for each quadrupling of the size of the image file, there would ordinarily be a twofold increase in the resolved image detail visible under sufficiently close inspection. Since photographs have finite resolution, increasing the number of pixels in a scan beyond some point results in a diminishing return on apparent detail. This situation is analogous to reaching "empty magnification" with an analog image. If the goal of the digital image capture is to preserve essentially all of the detail in the original image, then the last quadrupling of the file size will have only a modest effect of increasing image detail, yet it is essential for highest possible quality. As a practical example, to capture nearly all the image detail contained in a well-made four-by-five inch negative or positive film, the scan must capture all of the frequencies where the modulation transfer function of the film image is high. An RGB scan greater than 200MB on a high-quality scanner is required. Reducing the file size to 50MB or even going down to 25MB will create a digital image that meets the needs of many applications, but it cannot fully substitute for the image quality contained in the original film. A 25MB file is not small by today's standards, and even the 200MB scan requires careful post-processing (digital "sharpening") to closely approach the original image sharpness. Image compression can, under some circumstances, greatly reduce file storage requirements with little significant impact on image detail. However, conventional processing of compressed image files would still require working with the full decompressed file size in RAM or virtual memory. The RAM and/or virtual memory obstacles are eliminated when working with files in FlashPix™ format. Also, FlashPix™ currently supports conventional JPEG compression, and is expected to support lossless 3:1 JPEG compression in the near future. For anyone concerned with very fine image quality, digital imaging systems that can readily handle gigantic image files are simply a necessity for the imaging itself to be broadly satisfactory

Color and Tone Reproduction

A digital image archive requires more than just high resolution images. Efforts must be made to ensure that traditional image quality factors including tonal range and color consistency are not sacrificed as data migrates to new storage media and different computer platforms. Otherwise, the original objects must be re-digitized. Unfortunately, color management on the computer is still a relatively immature technology in the sense that serious standardization has not yet taken place, notwithstanding recent advancements (e.g., International Color Consortium [ICC] profiles, Apple ColorSync 2.0, Kodak Precision Color, etc.). Color management systems attempt to translate scanner, monitor, and printer color gamuts by reference to human color perception (e.g., CIELab color space). Color calibration targets are used to help define device characteristics in terms of colorimetric values.

The technology is now to the point where printer and monitor characterization have considerable merit, and monitor profiles and printer profiles should be carefully considered as additional elements that are critical to the realization of high-quality output from a digital archive. Scanner profiles have more limited value at present. The knowledge and skills of the scanner operator still determine the quality of the digital file in the case of scans made from many types of original artwork. For museums and archives, this means that the apparent simplicity of digitizing documents is often responsible for projects being undertaken, but as quality suffers, reality quickly sets in. Digital image capture is not as easy as using the office copier machine. Because the sensors in current scanners do not have the same inherent color sensitivity as the human eye, metamerism is still a problem. For example, a scanner can be profiled to give calorimetric data by characterizing an IT8.7/2 reflection target made on photographic paper, but the calibration is only precise for originals that are also color photographs on this type of paper. This fact became obvious when vintage albumen prints were scanned for a high-resolution digital image pilot project recently conducted at the Smithsonian Institution. A scanner which produced excellent color output from modern chromogenic color photographs, generated images from the albumen prints which were consistently more red than observed visually. The scanner settings required further manual adjustments to correct the unwanted color error. After manual correction, output was excellent.

The High-Resolution Pilot Project at the Smithsonian Institution

Approximately 400,000 images comprise the core photographic and artwork collections at The National Anthropological Archives (NAA) of the Smithsonian Institution. The collections' "broad ethnographic contents include the world's most extensive photographic documentation of American Indian cultures as well as visual records from cultures around the globe amassed by Smithsonian researchers and collectors. The average age of the photographic items in this collection is between 85-90 years old."4 Over one thousand individuals visit the archives each year, and without high quality facsimiles available, direct researcher access to the vintage photographs and manuscripts is the only way to show the information at the present time. The handling of the originals continues to be a principal threat to the longevity of the collection. The high image quality of the vintage prints (many are contact prints from large format negatives) and important text written on the verso of many of the prints requires high resolution image capture. Although many originals are "black and white" photos, almost all convey additional information in terms of color and tone. Color scans are highly desirable.

Recently, a pilot project has been completed to examine image quality issues and to determine if Live Picture technology could solve the problem of poor image access times. Figures l and 2 show 300 dpi and 600 dpi scans of a microcopy resolution test target. The 300 dpi image capture resolved 4.5 line pairs per millimeter in the test target print while 600 dpi captures the target at 9.0 lp/mm. The vintage prints often resolve more than 20 lp/mm so neither scan captures total information content. However, the 600 dpi is

more than sufficient to reveal to the researcher all the image detail seen upon inspection of the original with the unaided eye. Moreover, 300 dpi was satisfactory for recording text on the verso when all handwriting and other textual information was large enough to be easily legible to the unaided eye. Thus, the judicious use of 300 to 600 dpi scans reproduced images and text adequately for this digitizing project. Nevertheless, this scanner setting created uncompressed 24bit color file sizes on the order of 20 to 100 megabytes per image (40 to 200MB per vintage print with verso) when applied to mounted photographic prints typically larger than 8 x 10 inches. Tests showed that a single 80MB image in uncompressed TIFF file format was opened by Adobe Photoshop 4.0 software in approximately 90 seconds. The computer platform was a UMAX 200Mhz Macintosh clone computer with 160MB RAM and 4MB VRAM. The use of large image files for publication and exhibition prints is not deterred by this access time, but for information retrieval for scholarly use, this is unacceptably slow. Tests were then conducted using Live Picture 2.5 software and a FITS file that "composited" 24 IVUE files (12 original prints scanned at 600 dpi, and 12 verso scanned at 300 dpi). The verso of each print was hidden as a layer underneath the corresponding original scene. Over 250MB of JPEG compressed IVUE format data (greater than 1.2 GB when decompressed) were viewed by opening an approximately 1MB FITS file. All FITS and IVUE data resided on a CD-ROM and were accessed using an 8X CDROM reader. All images appeared on screen and were viewable within 12 seconds. Image manipulations (zooming, panning, rotating, flipping) were accomplished in approximately 1 to 4 seconds. Viewing information on the verso was instantaneous by using the computer's mouse to click on the layer viewing icon (activate/deactivate layer control in the FITS file).

The initial scanning project that is envisioned concerns the photographic holdings. These items are more susceptible to deterioration and would also benefit from cold storage, a goal that is unrealistic until a suitable alternative to daily use can be implemented. The photographs comprise approximately 100,000 items. With project planning which includes scanning on multiple workstations and a dedicated staff of four to six workers on two or three shifts, the 100,000 photographs could be digitized in approximately one year. A realistic price using today's technology is approximately $10.00 to $15.00 per image, so the total project cost would be 1.0 to 1.5 million dollars. Cost per image is declining on an almost weekly basis as new image capture devices and storage media reach the market. However, the intrinsic value of the collection and the risk of continued use and storage at room temperature more than justify proceeding with this project as soon as possible.

The Digitization Project at the Museum of Modern Art

The Department of Photography at the Museum of Modern Art in New York is investigating its options for digitizing its entire collection of some 30,000 fine art photographic prints. This collection consists of many kinds of prints, both "black & white" and color, spanning much of the history of photography. The initial approach is to attempt to create an all-purpose archive at very high resolution and with excellent control over the nuances of color and tone, such that the archive's files can be used for essentially any kind of digital output, with any intended use.

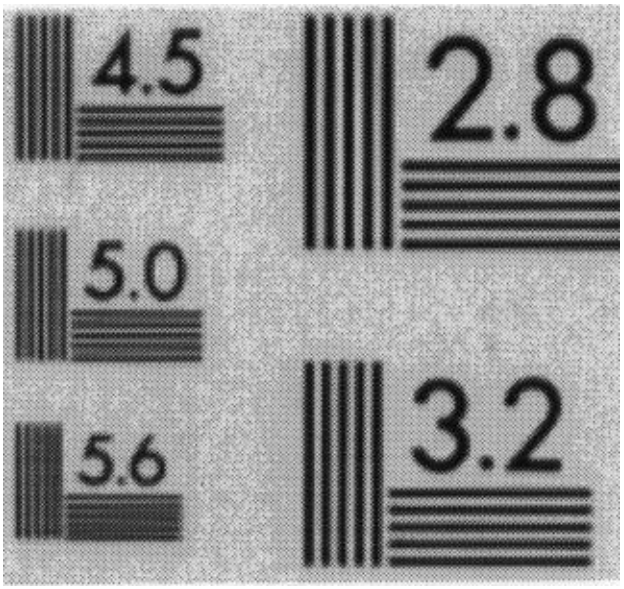

Figure 1. 300 dpi scan of microcopy resolution test target resolved 4.5 line pairs per millimeter.

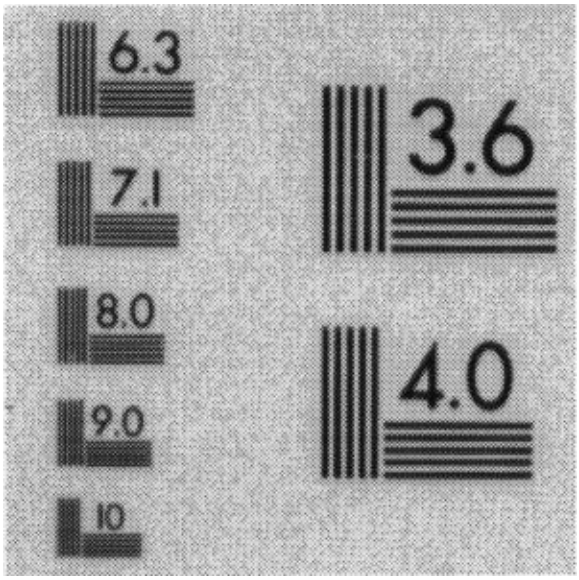

Figure 2. 600 dpi scan of microcopy resolution test target resolved 9.0 line pairs per millimeter.

The most demanding use possible would be the creation of a digital facsimile for display, which, in some cases, would require even more resolution than a fine art poster reproduction. Beyond prints and posters, high-quality book pages (on a par with the finest conventionally produced book pages), small prints for scholarly study, low-resolution files for use on the World Wide Web, thumbnail images for

searching and for curating exhibitions, and so on, would all be produced from the same master, high-resolution files.

This project will endeavor to push the limits of color management to obtain extremely accurate records of original prints, including "black and white" prints of many types with many characteristic appearances, including varying hues, etc. We will attempt to refine the process of scanning prints, usually with a digital camera back, both with and without using ICC camera (i.e. scanner) profiles to the point where a very fine reproduction of an original could be made without any need to look at the original.

Keeping the need to remove original prints from cold storage to a minimum will help to preserve the collection, and in order to make the archive as broadly useful as possible, the archiving strategy will be built around the use of the FlashPix™ file format, the FlashView capability to make global color and tone edits to FlashPix™ files, and the Live Picture FITS capability to make local edits as needed. The FlashPixr™ format solves the critical problems of working with very large image files (for example, over 100MB), and the two editing technologies (both are present in Live Picture software, one is for global edits only and the other allows local brushing in of edits) make a resolution-independent workflow possible.

A resolution-independent workflow allows one image file to be used for output at any resolution up to that of the original scan and including higher interpolated resolutions as well. Further, when combined with the device-independent model of ICC-based color management, a single master digital file can be used for output, not only of any size or level of detail (up to approximately what was in the original print) but for any kind of output as well. The many kinds of output (e.g. monitor display, prints made with inkjet technologies, prints made with offset lithography, prints made with laser exposure of photographic materials, etc.) all vary greatly in the way that they use the numerical color data of an original image file, and therefore a device-independent model for imaging can be hugely beneficial, because it automates the rendering of colors on any given output system with nearly the identical qualities of color seen by the source device or shown on the CRT display.

References:

1. Documentation available on the Internet at: (http://www.livepicture.com/).
2. Documentation available on the Internet at: http://image.cup.external.hp.com/ifori/htdocs/b_sptmenu.html (Hewlett-Packard's Imaging for Internet,a.k.a. "IforI" site)
3. Documentation available on the Internet at: http://www.kodak.com/go/flashpix (Eastman Kodak FlashPix™ site)
4. Homiak, John P., Director, National Anthropological Archives, personal communication.