Digital Data Storage on Microfilm – The MILLENIUM Project: Signal and Information Processing

Christoph Voges, Volker Märgner, Tim Fingscheidt;
Technische Universität Braunschweig, Institute for Communications Technology (IfN); Braunschweig, Germany

Abstract

Digital data storage on microfilm is a highly promising technology for migration-free long-term storage of digital information. The lifetime of the medium microfilm is estimated to be 500 years under appropriate storage conditions. Towards a practical implementation of this concept, the MILLENIUM project was initiated in 2006. The project is funded by the German Federal Ministry of Economics and Technology (BMWi) and is conducted by two research institutions as well as several small and medium-sized enterprises. This paper reflects project achievements with respect to signal and information processing. The complete processing chain for digital data is described. Important parts of this chain are identified and discussed in more detail. Some focus is on frame structures and data organization as well as synchronization, error correction and storage capacity. The developed concepts are ready for implementation.

Introduction

Due to the advances in computer technology, the worldwide existing amount of digital information has been growing continuously during the past decades. Accordingly, a reliable approach for long-term storage of digital information is urgently required. However, to our knowledge, no such solution is available in the market up to now. For many years, microfilm has been used to store analog documents as photographs. The medium microfilm is highly stable and offers – depending on the storage conditions – an estimated lifetime of up to 500 years [1, 2]. With the advances in laser film recording technology for microfilm [3, 4] it has been suggested to use this medium also for digital data (see, e.g., [3, 5, 6]). Recently this field of technology has become more specific including first results and suggestions on channel modelling [7], storage of audio data [8], as well as storage capacity and error correction [9].

Towards a solid technological basis for this emerging technology, the MILLENIUM project was initiated. It is conducted by two research institutions, Technische Universität Braunschweig, Institute for Communications Technology (IfN), and Fraunhofer Institute for Physical Measurement Techniques (IPM), as well as several small and medium-sized enterprises. The project is funded by the Germany Federal Ministry of Economics and Technology (BMWi). The laser recording technology is developed at IPM whereas the research at IfN about signal and information processing is further described within this publication. Such processing also includes the organization of the data on the frame. Besides the storage of the data bits on the medium film, this is a major aspect, since the file information has to be localized within a large number of bits on the film.

This contribution presents the major results of the MILLENIUM project regarding signal and information processing aspects. It is published along with a second paper about the laser exposure hardware as realized within the MILLENIUM project [10]. The next section describes the processing chain for digital data followed by several sections that cover important parts and aspects of this chain in more detail. This includes frame structure and data organization, synchronization, as well as error correction, and data verification. A possible storage capacity for MILLENIUM is figured out followed by a detailed discussion on applications and future developments.

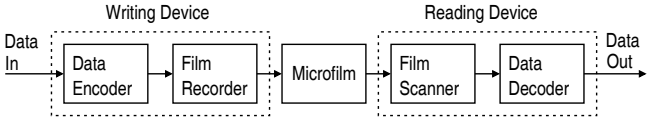

Figure 1. *Processing chain for digital data.*

Processing Chain for Digital Data

The processing chain for digital data includes all components and processes that are required for digital data storage on microfilm. Its main parts are depicted in Figure 1. The original data is first processed by the data encoder and then exposed to the microfilm by the film recorder, in our case a laser film recorder. Both the data encoder and the film recorder form the writing device. The digital data is stored on the film by means of exposure points (cf. [9]). In the laser recording device a modulated laser beam is moved over the film, thereby writing the data pattern consisting of exposure points on the film. For binary modulation the exposure points may directly represent a logical one or zero, respectively. Furthermore, some laser recorders allow amplitude modulation, where each single exposure point represents more information by variation of its amplitude. The horizontal and vertical distance of the exposure points is referred to as the grid space d. Both the grid space d and the number of employed amplitude levels have a significant influence on the storage capacity (see [9] for a detailed analysis). The encoder also introduces redundancy to the original data by applying a forward error correction (FEC) code [11]. This additional information is required for error correction by the FEC decoder as part of the reading device. Furthermore, information for synchronization purposes and for the file system has to be added.

Microfilm material is available as negative black-and-white microfilm and positive color microfilm. After the exposure process, photochemical processing is required [12, 13]. For the investigations within the MILLENIUM project, negative black-and-white microfilm has been used. Although the medium microfilm is quite robust, even a careful and clean handling cannot avoid tiny scratches or dust particles affecting the exposure points. The color microfilm consists of several color layers and can store more information. However, this film material is more expensive and the chemical processing requires more effort.

Table 1: Important system parameters of the Arche [3] and the MILLENIUM [10] laser recorders .

Parameter	Arche	MILLENIUM
Frame dimensions	32×45 mm	32×200 mm
Standard grid space	3 μm	4 μm
Exposure points	10666×15000	8000×50000

The reading device serves to retrieve the original digital data back from the microfilm. It mainly consists of a film scanner generating digital image data and a data decoder that further processes this data. The first step within the decoding process is to identify the position of the exposure points as exactly as possible, e.g., by means of a synchronization pattern. The demodulation process translates the exposure points into digital information that can be processed by the FEC decoder. By exploiting the redundancy introduced by the FEC encoder, the FEC decoder can correct errors, e.g., due to scratches or dust particles as mentioned above. Due to a proper choice and dimensioning of the error correction code, virtually error free recovery of the original data may be possible. For even increased data security, a verification step may be desirable as known, e.g., from CD-R. After all, file system information is required to identify individual files within the bits on the film.

The following sections describe main parts and important aspects of this processing chain in more detail. As practical examples, the Arche laser recorder as well as the new MILLENIUM laser recorder that is optimized for digital data are regarded.

Frame Structure and Data Organization

The laser recorder basically records arrays of exposure points on the film, so-called frames. These frames are defined, e.g., by using an image file as a template where each pixel corresponds to an exposure point on the film. Normally, the frames are separated by a small spacing. A comparison of the main specifications of both the Arche and the MILLENIUM recorders is given in Table 1. Please note that integer multiples of the standard grid space as given in Table 1 are possible without changes to the exposure hardware by omitting intermediate exposure points. Although the standard grid space of the MILLENIUM laser recorder is 1 μm larger, this device was constructed to write smaller exposure points on the film. It is obvious from the frame sizes in Table 1 that the corresponding uncompressed source image files have a considerable size. The film scanner normally involves oversampling, i.e., provides several image pixels for each exposure point. As a result, the file sizes for a completely scanned frame will even be a multiple of the template's image file size. Therefore, it is reasonable to divide the frames into several subframes that can be scanned and processed individually if desired. A further advantage of using subframes is the increased robustness against mechanical deformations of the film material, since the resulting position errors decrease with smaller dimensions of the subframes.

The schematic arrangement of the exposure points as subframes is depicted in Figure 2. The subframes can be arranged as squares, rectangles, or even stripes. Each subframe also contains a synchronization pattern that is necessary to identify the positions of the exposure points, as discussed in the following section. A subframe number is attached to each of the subframes. According to our convention, these subframes are numbered from top to bottom and from left to right. As the data is written onto several frames on a film reel, each frame has to be uniquely identified. For reasons of processing speed it is reasonable that this frame identifier can be read independently from the actual data pattern. This can be achieved by using either human-readable text or a digital barcode, similar to image marks as used for analog microfilm [14]. In order to simplify the processing, it may be reasonable to use larger structures for this barcode as for the actual data patterns. Figure 2 shows two different possibilities to place such a digital frame identifier, either perpendicular or in parallel to the film margins. When using the CCD (charge-coupled device) line array of a standard scanner for film reels, the perpendicular barcode can be captured at once by the CCD array. The parallel option is more difficult to process, since the CCD line array has to be moved over the whole frame to scan the barcode. When searching for a specific frame on a 300 m film reel this is a clear disadvantage. However, the parallel barcode can also be evaluated by a separate photodetector independently of the processing of the CCD image data. For this case, the frequency f_s of data symbols is

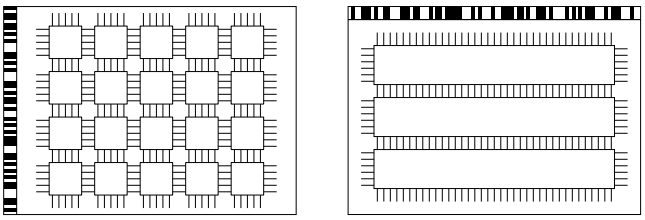

Figure 2. *Example of two possible frame structures: small subframes with perpendicular digital frame identifier (left) and long subframes with parallel digital frame identifier (right).*

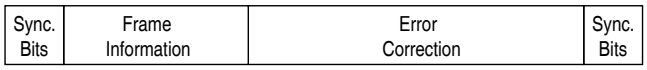

Figure 3. *Data structure of a digital frame identifier.*

$$f_s = \frac{v}{d_s} \tag{1}$$

with the transport speed v and the symbol width d_s. As a realistic example, a symbol width $d_s = 100\,\mu$m (≈33 times the standard grid space of the Arche laser recorder and 25 times the grid space of the MILLENIUM laser recorder) and $v = 1\,$m/s leads to $f_s = 10\,$kHz. Even when considering oversampling, the resulting frequencies can easily be processed by standard analog-to-digital converters. In this way 450 bits can be stored along the side of an Arche frame and even 2000 bits along the side of a MILLENIUM frame. Although a certain part of these bits is required for synchronization and error correction, this is still sufficient to carry frame addressing information. A possible data structure for such a frame identifier is given in Figure 3.

In order to store several files on a film reel consisting of several frames, a data structure has to be defined. An example of such a structure is depicted in Figure 4. Besides the actual data of the files to be stored on the film, file system information is required. This includes the number of files as well as a file header for each of the files to be stored. The file headers are required to store the filenames as well as the start and end addresses of each file. Each address defines a certain position regarding frame number, subframe number and bit number within the indicated subframe. Clearly, the

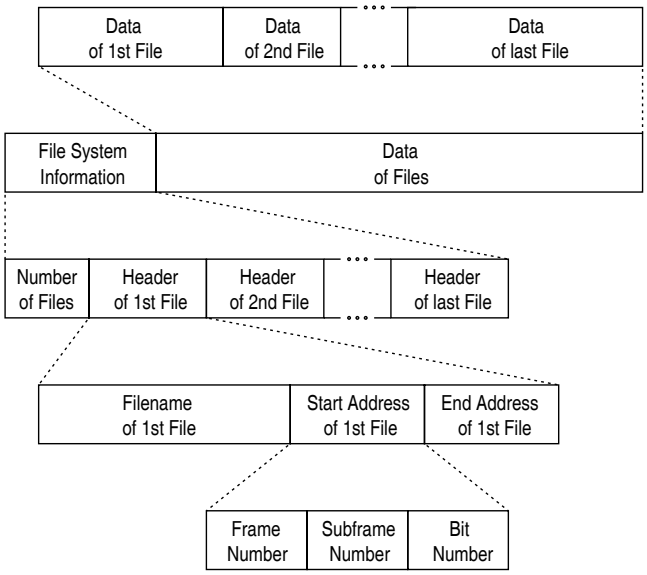

Figure 4. *Suggested overall file system information and data organization of a film reel.*

file system information is essential for data retrieval. To ensure the availability of this information, it may be duplicated, e.g., at the end of the film reel. Note that the suggested file structure allows to store files independently of their specific file format.

Individual FEC coding should be applied to the data of each file as well as the file system information. In the case of convolutional encoding this implies the termination of each encoded entity. Alternatively, when using a block code, the header of each file has to additionally contain information about the file length. The definition of the file headers can also be extended to carry further file information, such as the modification date. However, in order to achieve a more flexible header structure, we suggest to store these additional file attributes in a dedicated file using the XML (extensible markup language) format. This file can also contain metadata about the file contents.

Synchronization

The scanning device provides image information of the data pattern on the microfilm. Before the exposure points can be read from the film material, the position of each data point has to be precisely identified. To complicate matters further, the subframes may be slightly rotated by an unknown angle α. As an example, a subframe with a data pattern is schematically illustrated in Figure 5. The actual data pattern is surrounded by a synchronization pattern with its four parts referred to as the north, south, west, and east synchronization patterns. An alternative concept would be blind synchronization, where the position information is directly extracted from the data points without any synchronization pattern. However, the corresponding detection is more complex and prone to detection errors.

To describe the analysis of a synchronization pattern, the north synchronization pattern is exemplarily regarded. An enlarged north pattern is depicted in Figure 6. The pixels of this grayscale image are stored in the $I \times J$ matrix $\mathbf{A}$ with elements $a_{i,j} \in [0,1]$ and i, j denoting the column and row indices, respectively. Since we are

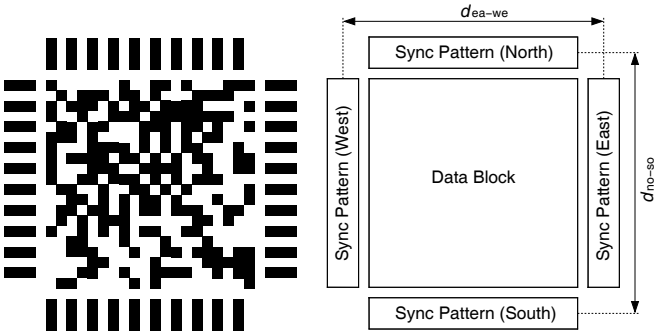

Figure 5. *Schematic illustration of exposure points with synchronization patterns (left) and definition of the nomenclature (right).*

using negative black-and-white microfilm, $a_{i,j} = 1$ corresponds to black and $a_{i,j} = 0$ to white, respectively. Alternatively, when using positive film material, $a_{i,j} = 1$ represents white and $a_{i,j} = 0$ is black. Objective of the synchronization is to identify the horizontal position $s[\ell]$, $1 \le s[\ell] \le I$, in pixel of the ℓ-th data column. The projection histogram $h[k]$ defined as

$$h[k] = \frac{1}{J} \sum_{j=1}^{J} a_{i=k,j} \tag{2}$$

with the discrete argument k is depicted in Figure 6. It is obvious that the local maxima of the projection histogram correspond to the lines of the synchronization pattern. There are N local maxima of interest with $n = 1,\ldots,N$ denoting the n-the local maximum. The position (in pixel) of the n-th local maximum is referred to as $b[n]$. To identify the desired local maxima, the following algorithm is suggested, whereby the values of I and J are assumed to be known system parameters:

1. Find the two outer local maxima $b[1]$ and $b[N]$ above the threshold $\Theta \in [0,1]$.
2. Estimate the approximate position $\tilde{b}[n]$, $n = 2,\ldots,N-1$ of the remaining $N-2$ local maxima based on the result of step 1.
3. For $n = 2,\ldots,N-1$, search the position $b[n]$ of the absolute maxima in $h\left[\tilde{b}[n]-\varepsilon,...,\tilde{b}[n]+\varepsilon\right]$ with the positive integer number $\varepsilon \in \mathbb{N}^{+}$.

The threshold value $\Theta \in [0,1]$ serves to avoid the detection of undesired local maxima that can occur, e.g., due to noise.

The position $b[n]$ of the n-th local maximum is already a good indicator for the position $s[\ell]$ of every second data column for $\ell = 1,3,5,\ldots$. To calculate the positions $s[\ell]$ for all data columns ℓ as good as possible, a linear fit is used as depicted in Figure 7. This analysis has to be carried out for each synchronization pattern separately, leading to the projection histograms $h_{\mathrm{no}}[k]$, $h_{\mathrm{so}}[k]$, $h_{\mathrm{we}}[k]$, and $h_{\mathrm{ea}}[k]$, the local maxima $b_{\mathrm{no}}[n]$, $b_{\mathrm{so}}[n]$, $b_{\mathrm{we}}[n]$, and $b_{\mathrm{ea}}[n]$, as well as the positions $s_{\mathrm{no}}[\ell]$, $s_{\mathrm{so}}[\ell]$, $s_{\mathrm{we}}[\ell]$, and $s_{\mathrm{ea}}[\ell]$. The calculation of each projection histogram has to be performed in analogy to (2). Note that for the west and east patterns, it has to be calculated along the rows instead of the columns.

If the data pattern is rotated by a small angle α, the value of α can be approximated as

$$\alpha = \frac{1}{2}\left[\operatorname{asin}\left(\frac{|s_{\mathrm{ea}}[1]-s_{\mathrm{we}}[1]|}{d_{\mathrm{ea-we}}}\right)+\operatorname{asin}\left(\frac{|s_{\mathrm{no}}[1]-s_{\mathrm{so}}[1]|}{d_{\mathrm{no-so}}}\right)\right] \tag{3}$$

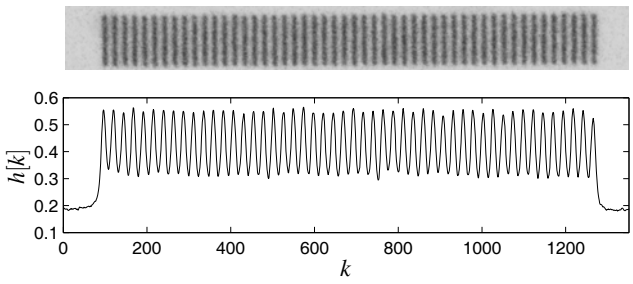

Figure 6. *Image of the north synchronization pattern (top) with the corresponding projection histogram $h[k]$ (bottom).*

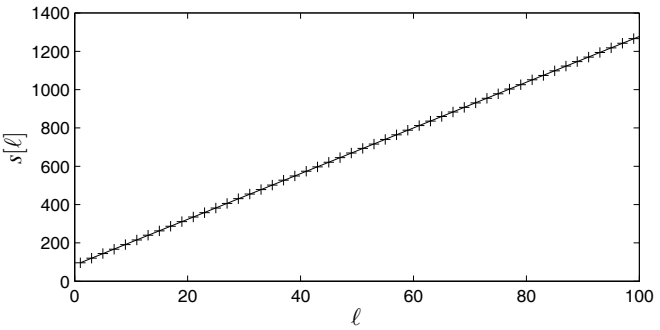

Figure 7. *Calculation of the positions $s[\ell]$ of the ℓ-th data column based on the north synchronization pattern. The crosses show the positions of the detected local maxima $b[n]$ of the projection histogram with $\ell = 2n-1$.*

with $d_{\mathrm{ea-we}}$ and $d_{\mathrm{no-so}}$ defined as depicted in Figure 5. Since $s_{\mathrm{no}}[\ell]$, $s_{\mathrm{so}}[\ell]$, $s_{\mathrm{we}}[\ell]$, and $s_{\mathrm{ea}}[\ell]$ already provide averaged position information, it is sufficient to calculate α for $\ell = 1$ only. After angle correction, the calculation of the positions $s_{\mathrm{no}}[\ell]$, $s_{\mathrm{so}}[\ell]$, $s_{\mathrm{we}}[\ell]$, and $s_{\mathrm{ea}}[\ell]$ has to be repeated, based on the rotated image. Thereby, the accuracy of the results can be improved by using the local maxima of opposite patterns jointly as a basis for the described linear interpolation. The efficiency of the synchronization regarding storage capacity can be calculated as

$$\eta_s = \frac{A_d}{A_t} \tag{4}$$

where A_d denotes the area of the data pattern and A_t the required area for both the data and the synchronization patterns.

Error Correction and Data Verification

For a reliable and virtually error-free reconstruction of the original data, a forward error correction (FEC) code is required. The FEC encoder adds redundancy to the data bits, according to a code rate r defined as the number of net data bits N_n divided by the number of total bits including the redundancy (gross bits) N_g:

$$r = \frac{N_n}{N_g}. \tag{5}$$

The dimensioning of these codes has been discussed in [9].

Even a properly chosen FEC code can normally not avoid residual bit errors. Although this so-called net BER (bit error rate) is generally quite small, e.g., in the range of 10^{-12}, bit errors can still occur. A data verification process can help to further decrease this residual BER. To achieve this, the data on the film is compared to the original data right after the writing process. If a bit error is detected during verification, the defective file can be rewritten to an additional film stripe, since the original data is still available at this stage. In general, it will be reasonable to repeat this process until either zero net BER or a certain specified maximum gross BER is achieved. Since verification has to take place after all chemical processing it is time-consuming and also complicates the processing workflow.

Storage Capacity

The fundamental dependencies regarding storage capacity for digital data on microfilm have been extensively studied in [9] for the Arche laser recorder. As the MILLENIUM recorder allows binary modulation only, the gross storage capacity (i.e., the storage capacity without considering further overhead due to FEC, synchronization, file system, etc.) is only dependent on the grid space d and the frame dimensions. Assuming the specifications given in Table 1, we arrive at a gross storage capacity of about $2 \times 10^9\,\mathrm{bit/m}$ or equivalently $250\,\mathrm{MByte/m}$ for standard 35 mm microfilm assuming no spacing between adjacent frames. Practical tests indicate that a grid space of $d = 4\,\mu\mathrm{m}$ will lead to a sufficiently low gross BER.

Applications and Future Developments

Microfilm-based data storage is the ideal solution for data stocks that have to be stored for a very long time without migration. Such applications are typically found in archives and libraries as well as in government and industry environments. Hybrid archiving of digital and analog data on the same medium is also an outstanding feature of microfilm-based data storage.

With the end of the MILLENIUM project in September 2009, a high-density laser recording device will exist that will be able to record smaller exposure points on microfilm [10] than the Arche laser [3]. This high resolution recording device allows strategies to record large amounts of information on the film. Concepts to store digital data using this recorder have been developed and are ready for implementation. An early prototype based on an automated microscope shows the basic concept of reading the data from the film. Furthermore, this microscope setup helps towards the specification of an optimal reading device. Future developments have to target at a portable reading device with an appropriate user interface. Today, there already exists a variety of commercially available microfilm scanners. However, to our knowledge the optical resolution of these devices is currently not sufficient to keep up with the full capabilities of the laser recording technology for digital data as developed within MILLENIUM.

Today's microfilm service providers as well as many archives already own a variety of scanning or even exposure devices for analog microfilm. Although most of these devices may not be capable of achieving a storage capacity as within MILLENIUM it may be desirable to extend the use of these devices to data storage applications with a lower storage capacity. Since microfilm hardware equipment – also for analog microfilm – is generally quite expensive, this can allow microfilm-based data storage without the financial barrier of high initial investments. To ensure backward compatibility, new scanning devices should be constructed in a way to be capable of processing these lower resolution data microfilms.

Due to advances in technology, the strategies of digital data

storage on microfilm will probably experience revisions from time to time. Also, for different applications other parameter settings might be desirable, e.g., for the FEC, the modulation scheme, or the data structure. Clearly, when dealing with long-term storage applications it should not be acceptable to rely on proprietary solutions. To avoid the emergence of multiple incompatible developments over time, standardization activities would be desirable. An example for an electronic archiving standard is the OAIS (open archival information system) reference model defined in [15].

Conclusions

In this contribution, the current state of the MILLENIUM project regarding signal and information processing for digital data storage on microfilm has been described. Therefore, the whole data processing chain was discussed. Important aspects of this chain have been identified and described in more detail, including frame structure, data organization, and synchronization. Furthermore, future developments and applications have been pointed out.

Clearly, the MILLENIUM project is an important contribution towards the practical realization of digital data storage on microfilm for migration-free long-term storage of digital data. The developed signal and information processing techniques within MILLENIUM allow the reliable storage of digital data. At the end of the project, a high resolution laser recorder will be available that will be capable of writing very small data structures on 35 mm microfilm. Because state of the art microfilm scanners are not able to keep up with the resolution of the MILLENIUM recorder, an early prototype based on an automated microscope will serve as a reading device. The involved algorithms and processes can be applied to a future reading device with certain modifications. Besides an appropriate reading device, future activities would take profit from standardization based on the above described data structures as suggestions and working drafts.

Acknowledgments

The project MILLENIUM is funded by the German Federal Ministry of Economics and Technology (BMWi). We thank our colleagues at Fraunhofer Institute for Physical Measurement Techniques (IPM) in Freiburg, Germany, for the exposure of several test films.

References

[1] Eastman Kodak Company, "KODAK Duplicating (x462), Direct Duplicating (x468), Direct Duplicating Intermediate Microfilm (2470) and Positive Print Duplicating Microfilm (x440) (ESTAR Base), Duplicating Microfilm Datasheet," Rochester, NY, U.S.A., 1999.

[2] ——, "KODAK IMAGELINK HQ, CS, CP and FS Microfilms, Camera Negative Microfilm Data Sheet," Rochester, NY, U.S.A., 1998.

[3] A. Hofmann, W. J. Riedel, K. Sassenscheid, and C. J. Angersbach, "Archivelaser Project: Accurate Long-Term Storage of Analog Originals and Digital Data with Laser Technology on Color Preservation Microfilm," in *Proc. of IS&T Archiving Conference*, Washington, DC, U.S.A., Apr. 2005, pp. 197–200.

[4] C. Normand, R. Gschwind, and W. J. Riedel, "Long-term Preservation of Digital Images on Color Microfilm," in *Proc. of 21st International Conference on Digital Printing Technologies*, Baltimore, MD, U.S.A, Sept. 2005, pp. 353–356.

[5] D. Gubler, L. Rosenthaler, and P. Fornaro, "The Obsolescence of Migration: Long-Term-Storage of Digital Code on Stable Optical Media," in *Proc. of IS&T Archiving Conference*, Ottawa, Canada, May 2006, pp. 135–139.

[6] C. J. Angersbach and K. Sassenscheid, "Long-Term Storage of Digital Data on Microfilm," in *Proc. of IS&T Archiving Conference*, Ottawa, Canada, May 2006, pp. 208–209.

[7] A. Amir, F. Müller, P. Fornaro, R. Gschwind, J. Rosenthal, and L. Rosenthaler, "Towards a Channel Model for Microfilm," in *Proc. of IS&T Archiving Conference*, Bern, Switzerland, June 2008, pp. 207–211.

[8] A. Hofmann and D. Giel, "Long Term Migration Free Storage of Digital Audio Data on Microfilm," in *Proc. of IS&T Archiving Conference*, Bern, Switzerland, June 2008, pp. 184–187.

[9] C. Voges, V. Märgner, and T. Fingscheidt, "Digital Data Storage on Microfilm - Error Correction and Storage Capacity Issues," in *Proc. of IS&T Archiving Conference*, Bern, Switzerland, June 2008, pp. 212–215.

[10] D. M. Giel, A. Hofmann, W. Salzmann, and C. Voges, "Digital Data Storage on Microfilm - The MILLENIUM Project: Hardware Realization," in *Proc. of IS&T Archiving Conference*, Arlington, VA, U.S.A, May 2009.

[11] S. Lin and D. Costello, *Error Control Coding: Fundamentals and Applications*. New Jersey: Pearson Prentice Hall, 2004.

[12] R. E. Jacobson, S. F. Ray, G. G. Attridge, and N. R. Axford, *The manual of photography: photographic and digital imaging*. Oxford: Focal Press, 2000.

[13] J. M. Sturge, *Neblette's handbook of photography and reprography: materials, processes and systems*. New York, U.S.A.: van Nostrand Reinhold Co., 1977.

[14] "Image Mark (Blip) Used in Image Mark Retrieval Systems," ANSI/AIIM MS 8-1988.

[15] "Space data and information transfer systems – Open archival information system – Reference model," ISO 14721:2003.

Author Biography

Christoph Voges received his Dipl.-Ing. degree in Electrical Engineering from Technische Universität Braunschweig, Germany, in 2005 with major subject information technology. During his studies he visited the University of Southampton, UK, as an exchange student. After his degree he joined the Institute for Communications Technology in Braunschweig, Germany. His current research interests include signal processing, coding, and storage technologies. He is particularly working on modulation, channel coding, and channel models for microfilm as a medium for long-term data storage.

Volker Märgner received his Dipl.-Ing. and Dr.-Ing. degrees in Electrical Engineering from Technische Universität Braunschweig, Germany, in 1974 and 1983, respectively. Since 1983 he has been working at Technische Universität Braunschweig. Currently he is a member of the research and teaching staff at the Institute for Communications Technology. His main areas of research are image processing and pattern recognition.

Tim Fingscheidt received his Dipl.-Ing. and Dr.-Ing. degrees in 1993 and 1998, respectively, from RWTH Aachen, Germany. After further research on joint source and channel coding at AT&T Labs, Florham Park, NJ, USA, he joined Siemens COM and later Siemens Corporate Research, where he led teams and standardization activities on speech signal processing. Since 2006 he is Professor for Signal Processing at the Institute for Communications Technology at Technische Universität Braunschweig, Germany. His research interests include source and channel coding, and speech signal processing.