TECHNICAL REPORT

ISO/TR 12033

First edition
2009-12-01

Document management — Electronic imaging — Guidance for the selection of document image compression methods

Gestion de documents — Imagerie électronique — Directives pour le choix des méthodes de compression d'image

Reference number
ISO/TR 12033:2009(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In exceptional circumstances, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide by a simple majority vote of its participating members to publish a Technical Report. A Technical Report is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TR 12033 was prepared by Technical Committee ISO/TC 171, *Document management applications*, Subcommittee SC 2, *Application issues*.

This first edition of ISO/TR 12033 cancels and replaces ISO/TS 12033:2001, which has been technically revised.

Introduction

With respect to the rapid increase of applications using digitization techniques, the role of compression methods has become a factor of growing importance for the management of the volumes of stored data.

The effects of the available compression methods vary greatly, depending on the source documents. For example, an electronic image management (EIM) system configured for scanning and storing continuous tone images will have different image compression requirements as compared to an application involving only text.

Practical methods for analysing user requirements for image compression in order to select accurate and optimal image compression schemes are complex. This Technical Report was issued in order to guide users and system developers in their selection of these methods.

Document management — Electronic imaging — Guidance for the selection of document image compression methods

1 Scope

This Technical Report gives information to enable a user or electronic image management (EIM) integrator to make an informed decision on selecting compression methods for digital images of business documents. It provides technical guidance to analyse the type of documents and which compression methods are most suitable for particular documents in order to optimize their storage and use.

For the user, this Technical Report provides information on image compression methods incorporated in hardware or software in order to help the user during the selection of equipment in which the methods are embedded.

For the equipment or software designer, this Technical Report provides planning information.

This Technical Report is applicable only to still images in bit map mode. It only takes into account compression algorithms based on well-tested mathematical work.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 12651:1999, *Electronic imaging — Vocabulary*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 12651 and the following apply.

3.1
compression
process of removing redundancies in digital data to reduce the amount that should be stored or transmitted

NOTE Lossless compression removes only enough redundancy so that the original data can be recreated exactly as it was. Lossy compression sacrifices additional data to achieve greater compression. This is typically useful for greyscale or colour image compression, where details that are not perceptible, or are minimally perceptible, to the human eye can be eliminated, normally with a dramatic increase in compression. It is advisable that lossy compression not be used for documents containing textual information and not be used for long term archival of any type of documents.

3.2
resolution
number of pixels per unit of length

3.3
dots per inch
dpi
number of dots that a scanner (printer) can scan (print) per inch both horizontally or vertically

3.4
brightness
visual sensation that enables an observer to detect luminance

3.5
contrast
ratio of on pixel brightness to off pixel brightness

3.6
bit level
number of bits used to define a pixel

3.7
luminance
Y
luminous flux emitted from a surface

NOTE The former term was photometric brightness.

3.8
chrominance
Cr
Cb
colour portion of the video signal including hue and saturation but not brightness

NOTE Low chroma means the colour picture looks pale or washed out; high chroma means intense colour; black, grey and white have a chrominance equal to zero.

3.9
ITU-T Group 3 and Group 4
compression algorithms standards defined by the ITU-T in Recommendations T.4 and T.6

3.10
Joint Photographic Experts Group
JPEG
name of the committee that developed the ISO/IEC 10918 series which shares the same popular name

NOTE The “J” refers to the joint development with the ITU-T.

3.11
Comité Consultatif International Télégraphique et Téléphonique
former name of the International Telecommunication Union (ITU) standardization body

3.12
compression ratio
relationship of the total bits used to represent the original to the total number of encoded bits

3.13
Joint Bi-level Image Experts Group
JBIG
name of the sub committee that developed ISO/IEC 11544

NOTE The joint committee is with ITU-T. JBIG and JPEG are managed by ISO/IEC JTC1/SC 29/Working Group 1.

4 General

In a document imaging system, users are concerned about the quality of archived images, for two reasons:

a) it can affect the imaging system's future in the medium or even long-term;

b) it is necessary to choose the imaging tools based on an evolving technology.

The digitization process, which by nature transforms an image conveying comprehensible information into a dematerialized one, changes the observer's perception of that image. The observer may consider the image as being improved, though more frequently he considers it degraded. In fact, images undergo a number of successive transformations at different points during the digitization process. At each of these stages, attempts are made to keep the image within acceptable legibility limits, but also to restrict its size to within acceptable economic limits.

The specific role of one of the digitization stages — compression — is to reduce the size of the image. Some compression methods are reversible in that the decompression algorithm restores the initial digital information. These methods are lossless and have no impact on the quality of the image as it is perceived by the human eye. Other methods are lossy, and may cause degradation perceptible to the eye. By adjusting certain parameters, the user can bring a lossy method within acceptable limits; because the acceptance of a lossy method is a subjective judgement. Any image or document, on which a computerized treatment may be applied, should not be compressed with such a method. This is one of the major reasons not to use lossy compression for long-term archiving, as future usage of the image or document is unknown.

While numerous compression methods are described in technical literature, few are stable according to industrial standards. These are based on a limited number of principles:

— dominance of certain patterns,

— pattern repetition, and

— noticeable mathematical properties.

In any individual method, the number of parameters the user can modify is small.

The choice of a method and compression parameters are in large part determined by two considerations:

a) the characteristics of the document;

b) the period of time the document is to be retained (retention time).

Obviously, the graphical contents of a document play a key role in determining the method and its parameters. However, other factors characterizing the application context are also very important (see Table 1).

The graphical content of the document is important to the compression process. A business document that can be copied or faxed as “pure black and pure white” (even if the original was blue ink on yellow paper) are probably best compressed with the technologies developed by the ITU-T for a facsimile. Colour or grey scale photos are probably best compressed using one of the JPEG technologies. But if the photo has been converted to variable size black dots (like many “half-tone” newspaper photos), then JBIG is a superior compression technology.

5 Type of document and digitization parameters

5.1 General

A document is a set of organized information intended for presentation to a human user. Documents can be a single page or a set of pages, and can contain arbitrary content types, such as character content, graphical content, and various types of image content.

The following document content may be found in various types of documents. The classification list hereafter is somewhat arbitrary, but for a given application, these distinctions may be used to understand how to handle a given document.

5.2 Type of documents

This clause focuses on only those documents that are most likely to be archived electronically. These documents include:

— black text on white background, or more technically, dark text on light background (even if the ink happens to be blue or red or other single colour, on whatever colour paper);

— photographs, i.e. black and white or colour;

— mixed documents containing both text and photographs reproduced by a printing process, i.e. black and white or colour.

5.3 Document classification and digitization

5.3.1 General

For the purpose of determining a compression scheme, documents may be described in the following five ways. For each type of document, digitization methods are briefly described.

5.3.2 Black and white documents

Digitizing pages printed in black and white or more generally in bi-tonal mode (primarily text with a unique foreground on a unique background) generates bi-level images where each pixel is represented by a bit.

The most important digitization parameter is resolution.

Resolution should be determined according to visual perception needs and on the limits of the complete imaging process. Human eyes will not see noticeable differences on documents digitized at more than 300 dpi. This is the most commonly used resolution to keep quality unaltered. Any resolution under 300 dpi will have visible effects on the digitized document. A resolution over 300 dpi may be needed when computerized treatment is done on the document. Also, 300 dpi is the resolution limit of the human eye and should be considered as the needed resolution at the visual size, i.e. if the zooming factor to visualize is 4, a resolution of 1 200 dpi on the original size will provide 300 dpi on the visual size.

There are also other parameters, related to image processing, which vary according to the kind of image. If, for example, the images to be digitized are text, then it is advisable to produce black characters that are sharply defined against a white background. The brightness (adjusting the colour of a pixel against a threshold) and contrast parameters (adjusting the colour of a pixel against that of the surrounding pixels) should be adjusted for this purpose.

5.3.3 Grey scale documents

This form of representation is applied to photographic documents, printed on paper from a black and white film.

Digitization changes an initially continuous document into a matrix of pixels whose intensity is encoded in a range of levels. Thus, 8-bit encoding produces 256 grey scales.

The number of grey scales or the bit level should be determined according to visual perception needs and the limits of the complete imaging process. Quality tests have demonstrated that human eyes will not see noticeable differences on grey scale images coded with more than 8 bits. Therefore 8 bits encoding is the most common value.

5.3.4 Pseudo-grey or halftone documents

This category includes images that simulate grey using a variable arrangement of black and white pixels. There are two possibilities:

— the source document is a photographic reproduction in a text; it was produced using a printing technique and is, itself, a pseudo-grey document (a raster image can simulate grey by a pattern of black and white pixels);

— the source document is a photographic original, but was digitized in pseudo-grey for performance reasons, for example to reduce the storage volume or transmission times on a network.

5.3.5 Colour documents

This form of representation is applied to photographic documents, printed on paper from a colour photographic original. Another application is digital colour capture of business documents where yellow highlights, colour boxes, pencil, red pen, etc., is a part of the information capture integrity.

Colour documents are intended to be restored in colour, but may also be reproduced in grey scale.

Colour representation is based on the neuro-physiological properties of the human eye, notably the “visual trivariance” principle, which states that all colours can be produced by combining the three primary colours. Thus, a colour can be represented by three coordinates in a vector space based on primary colours, or by linear combinations of these coordinates.

The colour space most frequently adopted for electronic displays uses an additive of red, green and blue colours. These colours are differentiated by the retinal cones in the eye. Another colour space decouples the variables into one “luminance” variable, and two “chrominance” variables. This colour space is used to transmit television signal.

The most frequently used colour space in printing is of cyan, magenta, and yellow colours. A printed digital image emits light indirectly by reflecting light that falls upon it. For example, a page printed in yellow absorbs (hence this is called a subtractive colour space) the blue component of white light and reflects the remaining red and green components, thereby creating a similar effect as a monitor emitting red and green light. Hence the printing industry mixes cyan, magenta, and yellow inks to create all other colours. Combining these subtractive primary colours will generate black, but in practice black ink is used, hence the term “CMYK” colour space, the last character “K” standing for black.

In a digitized colour image, each pixel is represented by assembling three components corresponding to the primary colours. The bit level adopted for a component determines the quality of hues; the standard of 8 bits per component can represent $256^3 = 16$ million different colours. Representations on a total of 8 bits sent by data communication networks are also fairly frequent.