



# Personalizing image enhancement for critical visual tasks: improved legibility of papyri using color processing and visual illusions

Vlad Atanasiu<sup>1,2</sup> · Isabelle Marthot-Santaniello<sup>2</sup>

Received: 28 October 2020 / Revised: 16 May 2021 / Accepted: 26 August 2021 / Published online: 27 December 2021  
© The Author(s) 2021, corrected publication 2022

## Abstract

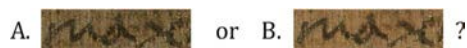
This article develops theoretical, algorithmic, perceptual, and interaction aspects of script legibility enhancement in the visible light spectrum for the purpose of scholarly editing of papyri texts. Novel legibility enhancement algorithms based on color processing and visual illusions are compared to classic methods in a user experience experiment. (1) The proposed methods outperformed the comparison methods. (2) Users exhibited a broad behavioral spectrum, under the influence of factors such as personality and social conditioning, tasks and application domains, expertise level and image quality, and affordances of software, hardware, and interfaces. No single enhancement method satisfied all factor configurations. Therefore, it is suggested to offer users a broad choice of methods to facilitate personalization, contextualization, and complementarity. (3) A distinction is made between casual and critical vision on the basis of signal ambiguity and error consequences. The criteria of a paradigm for enhancing images for critical applications comprise: interpreting images skeptically; approaching enhancement as a system problem; considering all image structures as potential information; and making uncertainty and alternative interpretations explicit, both visually and numerically.

**Keywords** Script legibility · Image enhancement · Color processing · Perceptual image processing · Image quality · Papyrology

*It's obvious—any fool can see it.*  
Homer, *The Iliad*, 7.464 [77, p. 227]

## 1 Introduction

Of the images below, which would you consider more legible,



If you selected B, you behaved like an engineer, confident in the benefits of a high signal-to-noise ratio. It may, therefore, come as a surprise that, in the evaluation reported in this article, three out of eight scholars deciphering these ancient Greek papyri preferred item A, because it is the original document, and originals should not be altered. This is not unreasonable: to question what exactly one is looking at is sensible. We call this attitude *critical vision*, and place it at the core of image enhancement. Our psychophysical and sociological mini-experiment also illustrates how enhancement quality depends on more factors than algorithms alone. These two principles frame the work being presented here.

### 1.1 Objectives

In narrow terms, the scope of this article is the legibility enhancement in the visible light spectrum of papyri documents for scholarly text editing. The broader objectives derive from the challenges, approaches, and lessons of this work. On the technical level, the aim is to demonstrate the utility of color processing and visual illusions in legibility enhancement. To emphasize the power of these approaches,

✉ Vlad Atanasiu  
atanasiu@alum.mit.edu  
Isabelle Marthot-Santaniello  
i.marthot-santaniello@unibas.ch

<sup>1</sup> Department of Informatics, University of Fribourg, Fribourg, Switzerland

<sup>2</sup> Department of Ancient Studies, University of Basel, Basel, Switzerland

the proposed methods will rely on a minimal algorithmic apparatus. On the theoretical level, a paradigm of legibility enhancement for computer-aided vision will be outlined. The central tenet of this paradigm is that critical vision tasks require enhancement technologies capable of making uncertainty explicit and adopting a systemic approach, that encompasses users, tasks, data, methods, tools, and interactions. The methodological objective is to develop methods and paradigms on an empirical, interdisciplinary basis.

## 1.2 Relevance

Papyri make up an extensive but underexploited trove of time capsules. The vast majority of these artifacts are accessible only as images taken in visible light and are often characterized by poor legibility due to degraded physical documents, non-calibrated imaging, and image compression. Digital enhancement promises to reduce the time and uncertainty associated with reading difficult texts from difficult reproductions. This work may be useful in application domains other than papyrology as well, as image enhancement is a ubiquitous stage in image processing. For example, there are similarities between deciphering papyri and examining medical, forensic, and military imagery, with respect to the uncertainty surrounding their interpretation.

## 1.3 Contributions

The present article adds to the limited existing work on papyri enhancement in the visible light spectrum. It promotes the use of color processing and computationally induced visual illusions. It works towards remedying the lack of a holistic approach to legibility enhancement. Making uncertainty explicit is also novel in this domain.

## 1.4 Methodology

The theoretical argument made in this article, the evaluation of the proposed methods, and the design of the ensuing software rely on an in-depth exploratory statistical analysis of data obtained from a user experiment. However, mathematics, algorithms, software, and hardware are never neutral when used by humans; this is especially so when the machine and the human form a closely coupled system, as in the case of legibility enhancement. Therefore, this article presents various points of view regarding how psychological and sociological factors are co-determinant in the design of legibility enhancement methods. The user-centric and context-aware approaches promoted herein are complementary to purely data-driven methods [162].

## 1.5 Applications

The principal application is an aid to human vision for use in transcribing papyri. The developed legibility software is also a tool for publishing better papyri images in print and online. Machine vision, particularly document binarization and text recognition, is the third potential application of the proposed methods, although human-readable texts are not necessarily legible to machines.

## 1.6 Organization

Section 2, “Problematic”, introduces papyrology and scholarly text editing and then defines legibility enhancement through the concept of critical vision. Section 3, “Related work”, reviews the computational methods of image enhancement used in papyrology and for similar tasks in other domains. Section 4, “Methods”, presents the novel legibility enhancement methods. Section 5, “Experiment”, describes a user experiment to empirically understand critical legibility enhancement. Section 7, “Paradigm”, discusses a type of image enhancement that makes uncertainty explicit and summarizes the criteria of a computer-aided critical vision system. Section 8, “Implementation”, shows how the paradigm and empirical findings were translated in software.

The image evaluation dataset is part of the article’s online Supplementary Material. The software resulting from this research is freely available online [16].

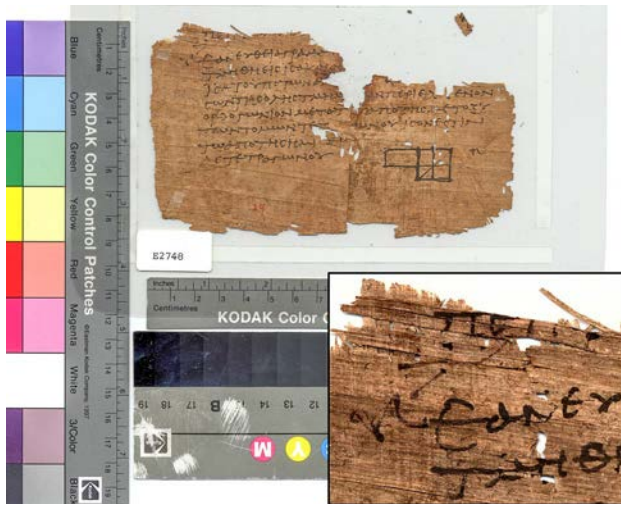
## 2 Problematic

### 2.1 Relevance

*Papyrus* was the quintessential portable writing surface in the Mediterranean area for five thousand years, until the end of the first millennium CE, when paper technology spread. As such, papyrus is a crucial carrier of substantial knowledge about the Egyptian, Greek, Roman, Byzantine, and other civilizations of antiquity. Figures 1 and 2 present surviving samples of two foundational texts of Western mathematics and literature: Euclid’s *Elements* and Homer’s *Iliad*. The number of unpublished papyri is currently estimated to be approximately one to one and a half million, which “will keep papyrologists busy for centuries at least” [182, pp. 644–645].

### 2.2 Data

*Cyperus papyrus L.* is an aquatic plant growing on the banks of the Nile River. By slicing, assemblage, and gluing via its own sap, a fine, flexible, and smooth papyrus



**Fig. 1** Euclid, *Elements*, Book II, Proposition 5; papyrus of the 3rd–4th century CE from Oxyrhynchus, Egypt. (Credit: P.Oxy. I 29, University of Pennsylvania Museum of Archaeology and Anthropology, CC-BY-2.5.)

can be manufactured, that is suitable for writing [34]. The orthogonal fiber pattern and surface roughness were once commercial criteria in the papyrus trade; today, they offer valuable hints to historians on prices, tastes, and document origins. However, the closeness in spatial frequency between papyrus texture and script complicates computational graphonomics. Reading performance is also affected, given the higher perceptual sensitivity to horizontal and vertical gratings [26, pp. 270–271]. The spectral characteristics of papyri satisfy historical and computational aims, such as the recovery of ink traces by means of multispectral imaging. Further characteristic, inconvenient to scholars and scientist alike, include edge irregularity, holes, document fragmentation, and the presence of artifacts surrounding the papyrus, such as color charts, rulers, and shelf marks (Fig. 1).

### 2.3 Task

Papyrology is the study of papyri, although it also encompasses inscribed potsherds, wax tablets, metal foils, and other materials [17, p. xvii]. The aim of papyrology is to elucidate all aspects of the past, from the material to the spiritual. Its main source of information is the text content. Text editing is the process of transforming a poorly legible text in bitmap format into legible electronic text. This is not a purely mechanical task, as considerable paleographical, linguistic, and historical knowledge is required to construct a space of possible readings and then select one or more as the most probable. It involves more than “a pair of sharp eyes and a certain amount of common sense” as any person struggling with an unfamiliar script and text can attest, and it has been aptly compared to the work of a detective or a

puzzle solver [161, pp. 197, 199]. For papyrologists, every word counts, because a single misread character can, literally speaking, change a queen into a fishmonger or teleport a city across space and time. Therefore, variant readings are common and indexed, since 1922, in a dedicated publication [161, pp. 212–213].

The task under consideration can thus be defined as *critical vision* in (a) the objective sense that its outcomes are of *significant importance*, and (b) the subjective sense of an observer exercising *skepticism* in the interpretation of signals. Here, the observers are the text editors, who are circumspect about the reality and meaning of what is seen in images, and their readers, who doubt the editors’ interpretation. The opposite of critical vision is *casual vision*, which is characterized by lower task criticality and levels of observer criticism. In psychological terms the latter is a form of fast, reflexive, and easy cognition (System 1), and the former of slow, analytical, and deliberative cognition (System 2) [83].

The two vision types call for different image processing paradigms. Casual vision is best served by binary images, designed to minimize entropy and produce trustworthy, unambiguous texts. Critical vision requires that ambiguity be maintained. In this case, no image structure is noise; every one is potential information. They provide context to develop concurrent readings and are valuable as archaeological layers of the document’s history. The text on a crumpled paper, to illustrate the argument, wouldn’t be identifiable as a draft if the creases would be removed for the sake of noise reduction.

### 2.4 Strategies

It is not unusual for papyrologists to take part in archaeological excavations, and, like archaeologists, to be sensitive to the context of discoveries and use various techniques to analyze data. A quote from experimental participant FRG illustrates the roles of *contextualization* and *diversification*:

“I found it important to look at the *original*, or a normalized version of the original (*vidiness*), sometimes with the *stretchlim* or *lsv* version of it for the bulk of the transcription work. To understand the state of preservation of the papyrus, the *histeq* or *adaphisteq* files have been very useful. The contrast is very good, and so I always use this type of image at the beginning of a transcription and [for] comparison with the original [ . . . ]. The *negvidiness* and *neglsv* images have been useful to detect differences between ink and holes, so I’ve used them occasionally. *Retinex* I expect to be useful, but not with the “Sammeltafeln” [different contents written on the same papyrus] we have in the Freiburg collection; while it is very good with large papyri, I would



**Fig. 2** Overview and detail of the original papyri images used in the evaluation. For higher resolution images please refer to the article's online Supplementary Material. **Top:** Notice the degradation of the physical documents, the differences in image quality, and the difficulty in comparing details within a limited display space. **Bottom:** Details scaled to equalize character height, demonstrating the

difficulty associated with deciphering such documents. (Credits: (1) Columbia University Library, CC BY-NC 3.0; (2, 4, 5) University of Michigan Library, Public Domain Mark 1.0; (3) Yale University Library, Public Domain Mark 1.0; (6) Courtesy of The Egypt Exploration Society and the University of Oxford Imaging Papyri Project; (7, 8, 9) Istituto Papirologico Vitelli, by permission.)

not use it for the transcription work throughout, only from time to time. *Locallapfilt* I have never found useful.”

The user describes the creation of a coherent, rich, and justifiable interpretation based on multiple perspectives. Implicitly, he indicates that the legibility enhancement system must support the enhancement of a variety of image features, and that methods may be chosen for their good performance on a specific document and task. The strategy fits well with models of visual search [104] and information seeking [102]. In-depth user studies and theoretical models of papyrological reading strategies implemented in expert systems are discussed in [56, 155, 176, 177].

## 2.5 Duration

The text editing of papyri is a delicate and typically lengthy process, comprising document manipulation

during reading, forming hypotheses regarding the content to achieve comprehension, and consulting auxiliary resources such as dictionaries. In addition to legibility, other factors affecting the amount of time required for this work include content difficulty, reader expertise, and work behavior. For example, it took the experienced participant FRG five to six hours split between two sessions to transcribe the papyrus used for training during evaluation, which was equivalent in length and legibility to the test documents presented in Fig. 2. The trainee OC, however, worked on and off for four to five weeks to transcribe a similar papyrus. AW also emphasized how, as time passed, he repeatedly modified his textual interpretation (or “belief”, to use Reverend Bayes’ terminology).

One implication is that dataset size and participant number is limited by necessity in papyrological legibility evaluations. A further consideration is the feasibility and appropriateness of experiments conducted in controlled laboratory settings. Not only is their organization complicated by their length, but the cognitive process of activity followed by rest phases is also impractical to reproduce.

Under these constraints, the “homework” approach is more practical and realistic.

## 3 Related work

### 3.1 Papyrological practice

Three approaches to script enhancement characterize common papyrological practice: interactive image manipulation, spectral decorrelation, and imaging techniques [35, 142].

#### 3.1.1 Interactivity

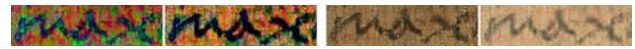
Image editors such as Adobe Photoshop allow easy and quick enhancement of papyri for a wide range of users. The flexibility of interactive image editing promotes solutions adapted to specific document qualities and user interests. Manipulations beyond the most basic tasks, however, require greater expertise, and are time-consuming.

#### 3.1.2 Decorrelation

In the early 2000s, Jon Harman created DStretch, a software for enhancing pre-historical rock art from the American Southwest. This has since become a commonly used tool among papyrologists [71, 142]. DStretch decorrelates spectral bands by using either eigenvectors or principal components [8, 66]. The process can be performed in various color spaces, to highlight different features. An image is obtained with pixels that are ideally well-separated in terms of color values, and represented as pseudocolors. The method yields excellent results for images with a Gaussian distribution, which is characteristic of inscriptions such as on walls, potsherds, or stones. DStretch is also considered user-friendly by papyrologists and can be used in a field-work context, as a smartphone application. One limitation of the decorrelation stretch method is its poorer performance for images with a non-Gaussian distribution, as in the case of papyri reproductions. Moreover, the features emerging from decorrelation are not necessarily meaningful. Figure 3 shows how DStretch induces both achromatic contrast and chromatic noise, in comparison with two of our novel methods.

#### 3.1.3 Imaging

Most papyri images are obtained through color or monochrome photography; these are the images of principal interest in this article. However, the specific needs of paleography have also led to the use or pioneering of many sophisticated techniques. Ultraviolet and infrared imaging, which can reveal ink traces in recycled and degraded documents invisible to unaided human vision, are used to analyze



**Fig. 3** Enhancement with (left to right) DStretch options *rgb* and *lbc*, and the novel methods *vividness* and *lsv* (detail of papyrus 4 of Fig. 2)

individual papyri [5, 60, 181]. With the advent of commercial narrow-band hyperspectral scanners, large-scale digitization has also become possible [41, 93, 94]. Reflectance Transformation Imaging retrieves the three-dimensional texture of the papyrus surface by illuminating it from a multitude of angles, thus providing a level of depth information that is inaccessible through classic photography [142]. X-ray and terahertz-based imaging have been used for the non-destructive reading of subsurface inscriptions in papyri and mummy cartonnage [64, 105]. These methods have been coupled with computed tomography methods for the virtual unrolling of the Herculaneum papyri, carbonized during the famous eruption of Mount Vesuvius in 79 CE [7, 112].

### 3.2 Computer science research

The authors have found very few articles that specifically examine the enhancement of papyri legibility from visible light images. Sparavigna [175] applies an edge detector based on the magnitude of dipole moments to increase the contrast of shape outlines, and thereby further enhancing the effect of a similar edge detection neural mechanism already present in the human visual system, known as the Mach bands [26, pp. 362–363]. The anisotropic filtering of the image in the frequency domain was tested as a preprocessing step to suppress papyri fibers and gaps; unfortunately, this also removed character segments collinear with the fibers.

The best papyri enhancements are obtained using multispectral imaging, as noted above in the papyrological practice overview. Moreover, in conjunction with multispectral images, color processing has also been used for papyri legibility enhancement specifically and for document processing more generally, as will be shown in the next section. We will here widen the survey to encompass other methods, topics, and applications of potential interest to this article, within the field of document processing and beyond, and conclude by discussing the applicability of machine learning.

Because noise—whether biological, mechanical, electrical, or digital—is a fundamental issue in communication, substantial research has aimed at improving document legibility by suppressing a great variety of what are usually considered manifestations of *noise*, including: ink bleed-through [46], see-through [48], foxing [163], shadows [95], termite bites [151], cross-outs [30], photocopied low-contrast carbon copies [36], low resolution raster images [141], and background texture interference [133] (for a history of image

denoising, see [111]). The type of visual *media* can also dictate the typology of enhancement methods (e.g., methods for pictures and for movies differ in whether the time dimension is available as a source of contextual information for optimizing image processing) [24]. In applied science contexts, such as in the photography and video equipment industry, there is interest in developing enhancement methods predicated on an understanding of the *nature of noise*, e.g., optical, mechanical, and electronic noise sources in cameras [139]. Advances in image quality measurement [87, 192] have benefited from research into *visual perception* and *neuroscience* [19, 188], as well as models of *scene statistics* [189]. The role of *tasks* in image enhancement is of particular interest in cultural heritage applications [180]. The *systemic* and *critical* approaches to document analysis advocated in this article have been the subject of exemplary research over more than a half-century in two domains, each with specific aims, constraints, and solutions. First, the legibility of flight documentation and instrumentation plays a critical role in *aviation* performance and safety; here, optimization has been approached mainly through psychophysical experimentation [55]. Second, research on the enhancement of *medical* radiographic images stands out in terms of the extent to which the impact on clinical diagnosis of technologies and perception have been investigated, including the role of visual illusions [23, 158].

In addition to methods, prominent areas of focus in this field are datasets, benchmarks, and ground truthing [117]. Visual confirmation of the attenuation of conspicuous artifacts is a typical means of comparing methods, supplemented with numerical characterization if feasible (i.e. given the availability of reference images or appropriateness of reference-free image quality measurements). User evaluations are rare, and vary from a few participants for historical documents [12, 28] to thousands in online campaigns for industrial applications [63, 144]. Legibility has been systematically studied since the early 20th century in experimental psychology (notably for the design of traffic signage and car plates, flight documentation and instrumentation, and typography for visually impaired [14, 114]), in works on teaching handwriting [159], as well as in relation to optical character recognition [68] and document image quality [4]. Document and image quality differ from writing legibility (e.g., a noise-free, but unfamiliar writing style may be difficult to read), although they conjointly affect the processing of visual information.

While the use of machine learning for image enhancement is well established [178], its application to document image enhancement, in particular historical, is still rare [72, 74, 115, 130, 131]. Notwithstanding results comparable to the state of the art, the challenges it faces are considerable. Historical data is scarce and heterogeneous, and ground truthing is time-consuming, which hinders the generalizability of the

learned models [54]. In respect to the task of critical vision, an implementable machine learning methodology has yet to be devised that can deal with images that are ambiguous, and their interpretations multiple, variable, or unknown. For scholarly applications, the resulting enhancements have to be, furthermore, trustworthy.

## 4 Methods

### 4.1 Justification

The proposed methods are rooted in color science and visual psychology and thus stand in contrast to the mathematics- and data-driven methods typically utilized in the document analysis domain. This section presents the rationale for and benefits of these methods.

*Color image processing* is an interdisciplinary field, owing to the multifarious aspects of color, as diverse as electromagnetic, perceptual, genetic, and linguistic [62, 97]. The perception of shape as the prominent aspect of documents has resulted in color receiving moderate attention in document image analysis. However, if the image resolution is low, or the text and background interfere with each other, as often occurs with papyri, color processing becomes an excellent alternative to spatial and frequency domain processing. Application examples of color-based document processing include historical document image enhancement [133], denoising, and binarization [185], page object segmentation [9], as well text extraction from typographic documents [173], web images [86], and cartographic maps [184]. Figure 4 illustrates the outcomes of the proposed and comparison methods, highlighting interactions between color and shape.

One far-reaching tenet of color research [84, 183] is that the technological, cognitive, and perceptual modalities through which visual reality is apprehended are not necessarily consistent with one another. In the negative polarity method, for example, physically and numerically identical gray-level differences are perceptually distinct. Because such effects, experienced as paradoxical, are the core of some proposed methods, this fact is emphasized through the use of the term *visual illusion* (see definitions, typologies, theories, history, and samples in [147, 164, 165]). Effects appropriate for enhancement were chosen among color illusions, since geometrical illusions are difficult to elicit without degrading legibility, and that illusions of stereopsis, apparent motion, and temporal illusions are tiring to use for extended periods of work.

The use of color and illusions shifts the burden of image enhancement from the computer to the human visual system. This perceptual image processing approach makes possible to create algorithms that are both mathematically



**Fig. 4** Enhancement results on papyri PSI XIII 1298 (15a) r 1 (top) and P.Oxy.XXII 2309 (bottom). From left to right: the original; five standard enhancement methods; and the new methods proposed in this article (italics). Note the different attenuation of the smudge (top row) and papyrus texture (lower row). The image labels refer to the eponymous methods described in Sects. 4, “Methods”, and 5.4 “Algorithms”, and are abbreviations of (from left to right) original image, stretch limits (*stretchlim*), histogram equalization (*histeq*), adaptive histogram equalization (*adpthisteq*), local Laplacian filters (*local-*

*lapfilt*), retina and cortex (*retinex*), lightness, saturation, and value (*lsv*), vividness (*vividness*), negative vividness (*negvividness*), and negative *lsv* (*neglsv*). Retinex and vividness are established names in the literature, while stretchlim, histeq, adapthisteq, and locallapfilt carry over the function names given in the MATLAB programming language, which was the development environment in the work presented herein. All ninety evaluation images are available in the article’s online Supplementary Material.

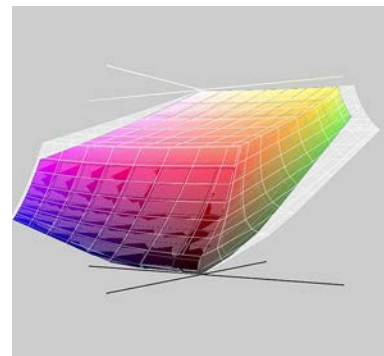


**Fig. 5** Color values at the same location in an image with sRGB (left) and Adobe RGB (1998) (right) embedded color profiles (detail of Fig. 7).

straightforward, and considered by the experimental participants useful. Researchers also have noticed that simplicity is beneficial to legibility enhancement quality [179, pp. 404, 416], and the likelihood of implementation of methods in software and their use by the end-users [100, p. 920].

### 4.2 Chroma contrasting by gamut expansion

In Fig. 5, the picture on the right appears more saturated, and the text is more clearly separated from the background. Yet the color values of both image files are identical. The perceptual difference results from the files being tagged with different color profiles, meaning that the same values



**Fig. 6** The sRGB color gamut volume represented within the Adobe RGB (1998) gamut (gray hull) in the CIELAB color space. The vertical dimension represents lightness,  $L^*$ , while the horizontal axes correspond to the red-green,  $a^*$ , and the blue-yellow,  $b^*$ , opponent colors.

represent different locations in the color space. The embedded profile of the left image is sRGB IEC 61966-2.1, a standard designed for displays with a limited amount of reproducible color, as were typical in the 1990s. The space defined by these colors is called a color gamut (Fig. 6). The

larger Adobe RGB (1998) gamut was designed to meet the needs of photographers and high-fidelity reproduction. Humans perceive far more colors than are present in either of these gamuts.

Numbers alone do not define colors; a coordinate system is also necessary. When a color space is replaced by a wider one, the gamut expansion increases the detectability of differences without affecting the values. The elegance of this legibility enhancement method resides in not requiring any numerical manipulation from software developers.

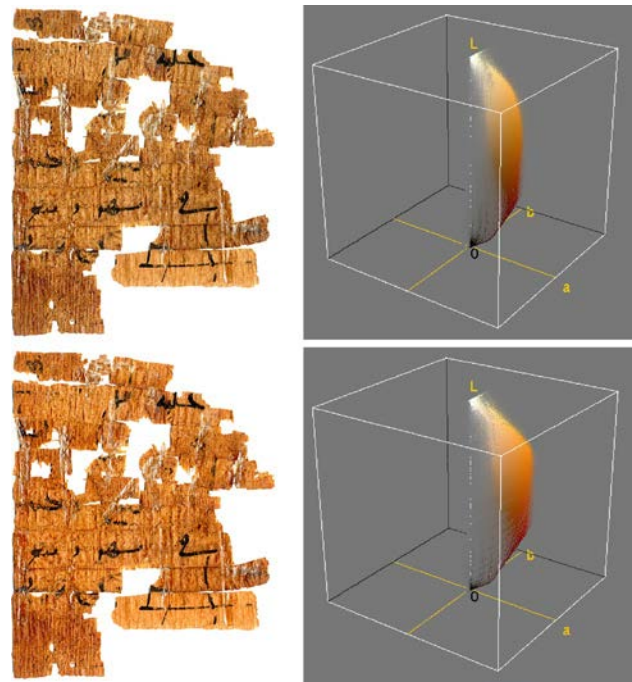
The underlying psychophysics is based on stronger chroma increasing the perceived level of luminance (*brightness*), which is known as the Helmholtz-Kohlrausch effect [45, 52]. This effect is leveraged as an unintended application of gamut mapping to improve the legibility of papyri.

A color space assignment and perceptual effects occur whenever an image file is read and visualized. Therefore being aware of their impact on legibility is important. In the case of papyri, the issue is compounded by so few reproductions having embedded color spaces. The most widespread space, sRGB, is assumed to apply in such situations.

The conversion of numerical values to visible color is done by the color management system of the image viewers, operating systems, and displays. For several reasons, transforming the numerical values to the destination color space may be desirable, instead of solely specifying it. For example, applications might disregard color space profiles, and thus deprive users of the enhancement effect. Nonetheless, the transforms among color spaces (*gamut mapping*) are user- and application-specific. The “*perceptual intent*” shifts out-of-gamut (OOG) color values to in-gamut values, thereby preserving perceptual color appearance; by contrast, the “*colorimetric intent*” strives to preserve numerical fidelity, usually by clipping outlier OOG values to the destination’s gamut boundaries. This is a key topic in color management research [153, 166]. The legibility enhancement method suggested here can be made more robust by using standardized color conversion profiles, such as developed by the International Color Consortium (ICC). The relevant code in the MATLAB programming language is as follows:

```
source_profile = 'sRGB Profile.icc';
destination_profile = 'AdobeRGB1998.icc';
C = makecform('icc', destination_profile, source_profile);
I = imread('input_image.tif');
J = applycform(I, C);
imwrite(J, 'output_image.tif', 'ColorSpace', 'rgb')
```

The results shown in Fig. 7 exhibit the change in chroma and color distribution. On the top is the original image with sRGB profile and its color distribution, and on the bottom the image after conversion to Adobe RGB (1998).



**Fig. 7** Color distribution of a papyrus image with native sRGB color profile (top), and following conversion to Adobe RGB (1998) (bottom). (Credits of papyrus image: Sorbonne Université, Institut de Papyrologie, P. Sorbonne Arabe 201 1500a.)

### 4.3 Lightness contrasting by stretching

*Lightness* refers to perceived reflectance, and is an achromatic component of color, together with brightness [157, pp. 2766–2767]. It is also a principal source of information about shapes for the human visual system, as can be ascertained via image decomposition into achromatic and chromatic colors:

$$\text{Image} = \text{Achromatic} + \text{Chromatic}$$

Hence, increasing the lightness contrast substantially improves legibility. The enhancement formula (*stretchlim*) is

$$L^{*'} = [L^* - \min(L^*)] / [\max(L^*) - \min(L^*)], \quad (1)$$

where  $L^*$  is the lightness in the CIELAB color space, bound to the [0, 100] range and corresponding to the approximate number of perceptually noticeable lightness levels under optimal conditions [51, pp. 24, 202]. CIELAB was developed by the Commission Internationale de l'Éclairage (CIE) as a perceptually uniform color space allowing for the quantitative description and manipulation of color along fundamental phenomenological dimensions, such as lightness, chroma, and hue. Its perceptual properties often make





**Fig. 8** Legibility varies with foreground/background and background/surround polarity.

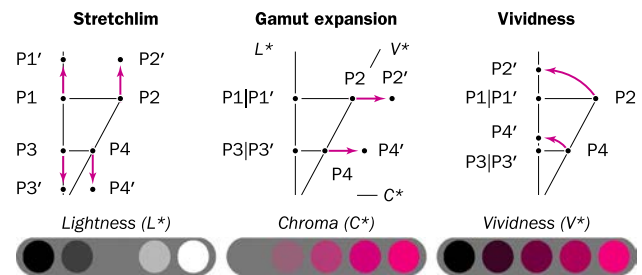
CIELAB a better candidate for color processing than the Red, Green, and Blue (RGB) color space of common digital images. CIELAB is not without its shortcomings and more accurate color appearance models exist and are developed [51, 167]; however, for applications where faithful color reproduction is not the main goal, CIELAB remains adequate. The conversion between RGB and CIELAB is parametrized by a triplet that specifies the illuminant of the scene reproduced in the image [40]. D65 (noon daylight, 6504 K) provided in practice the greatest improvement to papyri legibility.

#### 4.4 Lightness contrasting by negative polarity

The characters in Fig. 8, with CIELAB lightness values of 0 (black) or 100 (white), are set on squares of value 50 (gray). Despite identical numerical difference between figure and ground, the perceived contrast is stronger for light-on-gray configurations (bottom) and weaker for dark-on-gray (top); dark surrounds (left), and diminishing the stimuli (right), increase the effect. This occurs despite the perceptual linearity of CIELAB lightness and the use of calibrated displays.

This contrast illusion, known as *lightness induction*, is modulated by factors including asymmetrical polarity gain, nonlinear focus/surround contrast gain, and spatial frequency [26, 156]. It belongs to the broader category of *color constancy*, the neural and ecological basis of which has been a matter of debate since antiquity [2, 65]. Furthermore, the *irradiation effect*, well known to astronomers, makes light figures in dark surrounds (e.g. stars in the night sky) appear larger [143, 191].

Taken together, these effects let us posit that reversing the polarity of the image (i.e., so that the script appears lighter), will improve legibility. The empirical testing of this hypothesis as applied to papyri images is reported in Sect. 5, “Experiment”. The formula for negative polarity (*neg*),  $L^{*'}$ , is



**Fig. 9** Comparison of three enhancement methods described in this article. The vertical diagram axis represents the CIELAB lightness,  $L^*$ , the horizontal axis the chroma,  $C^*$ , and the diagonal the vividness,  $V^*$ . The variation of color along these axes is illustrated in the bottom part of the figure. For each enhancement type, the diagrams show the location of four exemplary points in a vertical section of the CIELAB color space before and after enhancement. Thus, stretchlim displaces all points towards the extremities of the lightness axis; gamut expansion displaces the points along the chroma axis (except for those situated directly on the lightness axis); finally, vividness applies a rotation transform resulting in all points being collinear with the lightness axis

$$L^{*'} = 100 - L^*, \tag{2}$$

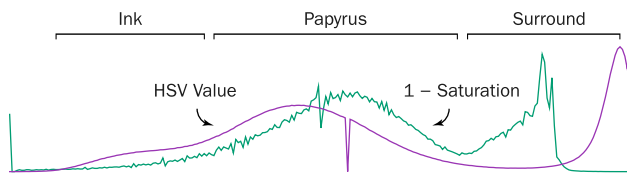
where  $L^*$  is the image lightness in the CIELAB color space. The method is appropriate for papyri, a document type of medium lightness inscribed with dark ink, and typically imaged on a light background (see Fig. 2).

#### 4.5 Selective contrasting by vividness colorization

*Chroma* is a dimension of the CIELAB color space that is orthogonal to lightness and hue (Fig. 9). Chroma modification effects a change between achromatic and chromatic color. Another dimension is *vividness*, which defines a concomitant change in lightness and chroma [22]. Vividness can be used as a model of how a dark ink film applied to the papyrus reduces both the amount of reflected light, thereby changing the lightness, and acts as an achromatic filter. Thus, vividness helps to distinguish ink from papyrus better than lightness. They can be substituted one for the other to improve legibility. Enhancement based on vividness (vividness) preserves the papyrus appearance, because the chroma and hue values remain unchanged (Fig. 3). The process differs from stretchlim and gamut expansion in that the transform is nonlinear and selective: that is, more vivid colors are more strongly emphasized, while achromatic color values do not change. The enhanced lightness,  $L^{*'}$ , is given by

$$L^{*'} = (L^{*2} + a^{*2} + c^{*2})^{-1/2}, \tag{3}$$

where  $L^*$  is the lightness, and  $a^*$  and  $b^*$  are the chromatic components in CIELAB ( $L^{*'}$  values exceeding 100 are



**Fig. 10** The figure depicts an example of the color profile of a papyrus in the HSV color space (the “value” channel and the complement of the “saturation” channel), along with the object classes dominating specific bands of the spectrum (ink pixels, non-inked papyrus surface, and the imaging area surrounding the papyrus).

clipped). The mathematical expression of vividness corresponds to the  $l^2$ -norm of the CIELAB image values.

Because the vividness method modifies the image in a pixel-wise fashion, it increases the distinguishability of ink and papyrus even in the presence of objects such as color charts, which complicate the use of more sophisticated methods such as principal component analysis [27].

#### 4.6 Background attenuation by difference of saturation and value

The hue, saturation, and value (HSV) color space is not a perceptual color space like CIELAB, rather, it is a simple mathematical transformation of the RGB color space [170]. Nonetheless, in HSV the distributions of saturation and value in papyri images roughly overlap (Fig. 10). On the basis of a simple difference between them, it is possible to attenuate the non-inked papyrus surface and, thus, enhance text legibility. Because the HSV value only approximates lightness, we fuse the difference of saturation and value with CIELAB lightness (e.g., by taking the mean) and accordingly obtain an image with sharper edges (Fig. 11).

The algorithm (*lsv*) represents a blind separation of a mixture of distributions. It consists in replacing the CIELAB lightness,  $L^*$ , by pixel-wise operations with  $L^{* \prime}$ , as follows:

$$L^{* \prime} = 100 \cdot {}_0[1 - ({}_0[100 - L^*]^1 + {}_0[V + S - 1]^1)/2]^1, \quad (4)$$

$$V = \max(R, G, B), \quad (5)$$

$$S = [V - \min(R, G, B)]/V, \quad (6)$$

where  $L^{* \prime}$  is the enhanced lightness;  $V$  and  $S$  are the value and saturation image channels in the HSV color space, derived from the  $R$ ,  $G$ , and  $B$  values of the RGB color space; and  ${}_0[\cdot]^1$  denotes normalization to the  $[0, 1]$  range.



**Fig. 11** Papyrus before (left) and after (right) enhancement with the *lsv* method (detail of papyrus 8 in Fig. 2).

### 4.7 Hue contrasting by hue shift

#### 4.7.1 Rationale

The spectral sensitivity of the human visual system is uneven: it peaks in the green band in bright light, and in blue-green in dim light; it is lowest in the red band [26, pp. 28–29, 118–119, 208–209]. As papyri have a brownish tint, a shift towards hues of higher acuity would be expected to improve legibility. The images in Fig. 12 result from applying this rationale. The first is the original, the second has all chromatic information discarded, and the remainder have the means of their CIELAB hue shifted to  $246^\circ$ ,  $162^\circ$ , and  $24^\circ$ , which correspond to the loci of the primary colors [50, p. 343].

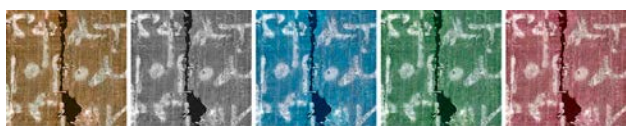
In addition for stronger acuity, a host of perceptual effects combine to make blue a compelling target for a hue shift. The increase in perceived lightness effected by an increase in chroma, as described above, is lowest for yellow and stronger for blue [52]. Moreover, the visual field for blue is broader than that for both green and red, thereby allowing for better comparison between focus and surround [26, pp. 108–109]. A blue text background has also been found to improve performance in analytic tasks [121]. Blue is furthermore robust to the most common colorblindness types [135].

Hue shift can be profitably combined with lightness polarity reversal: (a) lightness induction reinforces hue contrasting; (b) the asymmetry of the color space geometry reduces the dynamic range of chroma in the blue direction with respect to yellow for lightness levels above the medium [127, pp. 133–169].

Achromatic vision can resolve finer details than chromatic vision [128]. Because high spatial frequency is a characteristic of both script and papyri, using only the lightness channel for the decipherment task might make sense. However, the addition of color allows for ink to be better distinguished from stains, shadows, and other entities. Color vision in general improves image segmentation [29, 69].

#### 4.7.2 Implementation

The shift of the papyrus towards blue (*blue* method) is accomplished via the following equation:



**Fig. 12** From left to right: original, lightness channel  $L^*$ , and hue-shifted papyri; all have negative lightness polarity (detail of papyrus 4 in Fig. 2).



**Fig. 13** Hue shift (left) vs sign change (right), plus negative polarity (detail of papyrus 4 in Fig. 2).

$$h'_i = h_i - h_p + h_b, \tag{7}$$

where  $h_i$  and  $h'_i$  are the input and output hue in the cylindrical CIELAB color space, with  $h = \tan^{-1}(b^*/a^*)$  [51, p. 204]; while  $h_p$  is the centroid of the papyrus hue values and  $h_b = 246^\circ$  is the locus of blue in the hue dimension. An initial observation is that this is an ideal method, given the non-homogeneities in CIELAB. Second, the centroid of the papyrus hue is non-trivial to determine in the presence of other scene objects, such as color charts. Third, the location of maximum blue sensitivity varies across individuals, genders, ages, visual ecologies, and cultures, among other factors [49].

A simpler solution is available. This consists in changing the sign of the values in the  $a^*$  and  $b^*$  CIELAB channels:

$$a^{*'} = -a^* \quad \text{and} \quad b^{*'} = -b^*. \tag{8}$$

This is made possible because the brown hue of papyri has blue as its opponent color. It is this method, applied in conjunction with negative lightness, vividness, and gamut expansion, that was employed in the user evaluation. A comparison of the hue shift and sign change methods is shown in Fig. 13.

### 4.8 Dynamic range increase with CIELAB retinex

The retinex theory (from retina and cortex) provides an explanation for the constancy of color perception across a wide range of illumination conditions. Essentially, it expresses lightness as the product of ratios of spatial neighborhoods. As an image enhancement technique, it reveals details in dark and light areas. Many variants have been proposed since Land's and McCann's original work, and the out-comes differ according to image quality, content, and intent [107, 119]. The process is typically performed on each color channel individually, or on a single intensity image derived from the three RGB values, as a substitute



**Fig. 14** Clockwise from top left: **a** Original image, with three faint annotations indicated by an index. **b** Enhancement by retinex in the RGB color space. **c** Retinex enhancement in the CIELAB color space. **d** Blue-shifted and negative of the SE image (detail of papyrus 8 in Fig. 2).

of lightness. Pursuing our thinking within the framework of color processing, we find that performing retinex in a perceptual color space (CIELAB) is beneficial to legibility enhancement. It results in a concomitant increase in contrast in several perceptual dimensions discussed in previous sections: lightness, saturation, vividness, and hue. Consequently, significant features may become visible. In Fig. 14, the presence of annotations in a different ink is striking in the CIELAB retinex process, and further details are revealed by the blue negative. The result, in terms of appearance, insights, and pitfalls, is analogous to watching the *Iliad* unfold in Technicolor rather than in black and white. The *cinematic metaphor of image processing* sums up the rewards and implications of color processing and visual illusions in support of critical vision.

#### 4.8.1 Methodology

Conversion from RGB to CIELAB in the sRGB color space. Multiscale retinex with color restoration (MSRCR) [81, 140, 160] at three scales with Gaussian kernels of standard deviation 15, 80, 250, and a saturation of the 2.5% lowest and highest pixel values applied to (1) the RGB (Fig. 14b) and (2) the  $V^*$  (vividness),  $a^*$  (bluish–yellowish), and  $b^*$  (greenish–reddish) dimensions (Fig. 14c). Output normalization of (2) to the  $[0, 100]$ ,  $[-128, 127]$ , and  $[-128, 127]$  ranges, respectively, and back-conversion to

RGB. (3) Blue negative applied after CIELAB retinex, with  $V^* ' = 100 - V^*$ ,  $a^* ' = -a^*$ ,  $b^* ' = -b^*$  (Fig. 14d).

## 4.9 Cross-spectral colorization

Although the proposed methods are designed for visible light images, their use may be extended to multispectral images.

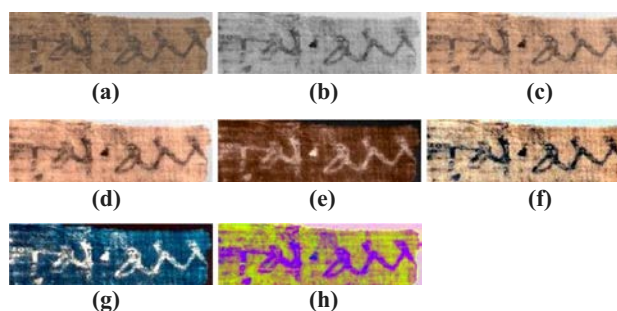
### 4.9.1 Background

Multispectral images consist of a set of single-channel images taken in various bands of the light spectrum; hence, each one is perceptually achromatic. Image fusion, which is a common additional processing step, is based on component analysis, wavelet transform, and other dimensionality reduction techniques [12, 70, 80, 110, 123, 174].

The visualization of multispectral images and images from bands outside the visible spectrum (VIS) may be improved through the use of color processing and by integrating multiple images into a single image. The problem, however, is that non-VIS images have no natural color correspondence. Moreover, the number of dimensions of multispectral images (dozens or even hundreds) usually surpasses that which is perceptible by the trichromatic human visual system.

Many single- and multi-channel “pseudocoloring” solutions have accordingly been developed [39, 187, 195], ranging from analog optical [168] to digital frequency domain coloring [3] to the ubiquitous “colormaps” [1]. However, the perceptual intricacies that must be considered even when color mapping single grayscale images [98, 154] indicate considerable scope for additional research in this area.

In the document processing domain, the pseudocoloring of multispectral and non-VIS images has been used to increase the contrast between text and substrate. For example, George *et al.* [61] used the first three components of the principal and independent component analysis, respectively, of 31 spectral bands of an ostracon as input in the RGB color space. Easton *et al.* [47] inserted the (“invisible”) ultraviolet-dominant grayscale image (long-wave ultraviolet light illumination) into the green and blue dimensions of the (visible) RGB color space, as well as the infrared-dominant grayscale image (tungsten illumination) into the red RGB dimension, thereby facilitating the legibility of the Archimedes parchment palimpsest. Gargano *et al.* [60] applied an overlay technique while working with papyri, which consisted in etching the binarized infrared image (or, alternatively, the red VIS channel) containing the document text into the visible light color image.



**Fig. 15** Example of legibility enhancement by cross-spectral colorization. **a** Papyrus imaged in visible light; **b** infrared image; **c** the infrared, colorized with the chromatic channels of the visible light image and enhanced with the gamut expansion and stretchlim methods; **d** vividness enhancement of the image (c); **e** negative of image (d); **f** MSRCR-RGB retinex of image (c); **g** negative blue-shift of image (f); **h** and the decorrelation stretch of the blue, red, and infrared bands.

### 4.9.2 Method

Our solution does not rely on binarization, which presents problems when applied to papyri [149] and may introduce artifacts and suppress critical contextual information during the interpretation of difficult texts. The rationale is that a single-channel image from the non-visible spectrum (non-VIS), or a fusion of several bands, contains features of interest that are not or poorly identifiable in the visible spectrum. Conversely, the VIS image will contain contextual information, supporting the interpretation of the image. Because the “end-user” of the visualization process is the human visual system, we use the perceptual CIELAB color space rather than the physical RGB space.

The procedure consists in transforming the VIS image from RGB to CIELAB, inserting the non-VIS image into the lightness dimension  $L^*$ , and then converting back to RGB. We thereby obtain an image where the details (rendered as lightness) come from outside the visible spectrum, while the context (rendered as chroma and saturation) comes from within it. The use of color allows for visual discrimination between the two types of information, while the lesser spatial acuity of chromatic perception relative to achromatic perception [154] confers saliency to the script with respect to the background. Since we now have a color image, we may apply the methods presented above for further enhancement. Figure 15 illustrates the improved legibility resulting from the cross-spectral colorization process.

### 4.10 Method integration

The proposed methods lead to better results when used concurrently. A papyri legibility enhancement workflow is presented in Algorithm 1. The vividness method is taken as an

example, with gamut expansion, dynamic range stretching, negative polarity, and blue shift also included.

---

**Algorithm 1.** Papyrus legibility enhancement
 

---

▷ input

**read RGB image values**  
 ▷ color gamut expansion (Section 4.2)

**convert color profile using ICC profiles**  
 · source color space: sRGB IEC 61966-2.1  
 · target color space: Adobe RGB (1998)  
 · source render intent: perceptual  
 · target render intent: perceptual  
 ▷ use a perceptual color space for enhancement (§ 4.3)

**convert color space from RGB to CIELAB**  
 · source color space: sRGB IEC 61966-2.1  
 · whitepoint: D65  
 ▷ vividness enhancement (§ 4.5)

**replace lightness by vividness** (Eq. 3)  
 $L^{*'} = (L^{*2} + a^{*2} + b^{*2})^{-1/2}$   
 ▷ dynamic range increase (§ 4.3)

**stretch dynamic range of lightness to bounds** (Eq. 1)  
 $L^{*'} = [L^* - \min(L^*)] / \{\max[L^* - \min(L^*)]\}$   
 ▷ negative polarity (§ 4.4)

**reverse lightness polarity** (Eq. 2)  
 $L^{*'} = 100 - L^*$   
 ▷ blue shift (§ 4.7)

**change sign of values in chromatic channels** (Eq. 8)  
 $a^{*'} = -a^*$  and  $b^{*'} = -b^*$   
 ▷ back-conversion to primary colors space

**convert color space from CIELAB to RGB**  
 · source color space: sRGB IEC 61966-2.1  
 · whitepoint: D65  
 ▷ manage out-of-gamut values

**clip values to the [0, 1] range**  
 $I' = \min(0, \max(1, I))$   
 ▷ output

**save image** in TIFF or JPEG format  
 ▷ make color profile explicit

**embed ICC color profile in image file**  
 · color space: sRGB IEC 61966-2.1

## 5 Experiment

### 5.1 Objectives

An experiment was carried out with papyrologists as participants in pursuit of four objectives: (1) Test the hypothesis that the proposed methods outperform existing ones. (2) Use the results to support the decision as to which methods to implement. (3) Explore the data to discover further potentially relevant facts and operationalize the findings. (4) Design the experiment to gain insight into the realistic usage conditions of the legibility enhancement software.

### 5.2 Participants

The experiment participants were two female and five male graduate students, and one male faculty member participating in the Eucor Tri-national Papyrology Webinar 2019-2020 held by the Universities of Basel (Switzerland), Freiburg (Germany), and Strasbourg (France). All were trained in Ancient Greek. Participation in the experiment counted towards their grades. The sample is more representative than suggested by its small size, if it is considered relative to the total number of papyrologists (the International Association of Papyrologists had 484 members in May 2020), and its international heterogeneity.

### 5.3 Stimuli

The stimuli consisted of a training and an evaluation set of papyri images. The training set comprised eight images of unpublished papyri from the University of Freiburg Library, whereas the evaluation set comprised the nine publicly available papyri color images from the ICDAR 2019 Competition on Document Image Binarization (DIBCO) [149] (Fig. 2). The quality of the original images in the training set was similar to that of the noisiest evaluation images. The DIBCO images represent typical qualities of papyri documents and reproductions, and the authors participated in their selection process. Each original method was processed with ten of the methods described in the previous section, thus bringing the total of images to 80 and 90 in the training and evaluation set, respectively (the evaluation images are available in the online Supplementary Material).

### 5.4 Algorithms

Four novel enhancement algorithms (*vividness*, *lsv*, *negvividness*, and *neglsv*) and five comparison algorithms (*stretchlim*, *histeq*, *adapthisteq*, *locallapfilt*, and *retinex*) were used. Given the variability of the lightness dynamic range in the test images, and to facilitate a bias-free comparison of methods, all images had their CIELAB lightness normalized in a preprocessing step, while their gamut was also expanded as described above. All methods were implemented in the MATLAB (R2020b) programming environment, except for *retinex*, which was produced with the ImageJ software package. The choice of comparison methods was based on their status as classical enhancement methods (*stretchlim*, *histeq*, *adapthisteq*), having outstanding edge-preserving performance (*locallapfilt*), and being a prime choice for achieving color contrast equalization (*retinex*). — Input images had no embedded color space information, except for P.Oxy. XXII 2309, which was in sRGB. All methods converted the input images to the Adobe RGB (1998) color space with D65 white point and saved them to sRGB to expand the color

gamut. Lightness was obtained from CIELAB. — The novel algorithms were those described in Sect. 4, “Methods”. *Neg-vividness* and *neglsv* apply negative lightness to *vividness* and *lsv*. Due to late-breaking research hue shift and *CIELAB retinex* were not included in the experiment. — *Stretchlim* [118] expands the dynamic range of an image’s CIELAB lightness to its bounds, [0, 100]. — *Histeq* [79, 118] is a global enhancement method that equalizes the spread of the histogram values of an intensity image over the available bit-depth dynamic range via linearization of its cumulative density function, thereby improving its contrast. — *Adapthisteq* (contrast-limited adaptive histogram equalization, CLAHE) [118, 198] is a local enhancement method similar to *histeq*, in that it equalizes the image histogram; however, it does this on small regions (here, 8-by-8 pixels). Accordingly, contrast for local structures is increased. — *Locallapfilt* [118, 134] is a state-of-the-art edge-preserving image enhancing method using pyramids of locally adapted Laplacian filters. Depending on parametrization, it can be used to either smooth or sharpen. Smoothing was used here, with parameters  $\sigma = 0.4$ ,  $\alpha = 2$ , and  $\beta = 1$ , derived empirically for their appropriateness to papyri images.—*Retinex* images were produced with the ImageJ implementation (RGB color space, uniform level, scale 240, scale division 3, and dynamic 1.20) [73].

## 5.5 Procedure

Each participant transcribed the text of a different image from the training set as part of their webinar examination. Afterward, they were asked to rate the processed images according to how useful they were for text transcription, using the following system: X: “*I used only one image for the transcription, this one*”; A: “*Very useful: it was the primary image used for transcription*”; B: “*I sometimes used this image*”; N: “*I didn’t use this image*”. (The results were similar to those obtained from the evaluation set.) This experimental step in real use conditions acquainted participants with the opportunities provided by the various processing methods for scholarly text editing.

During the evaluation step, all participants worked on all images of the evaluation set (Fig. 2). The first task involved rating the images for a hypothetical transcription on the same scale as used for the training set. The second task involved the ranking of the same images by utility. Given the effort and time invested in a scholarly transcription of ancient papyri, requiring participants to provide a transcription of all evaluation images it would have been impractical.

Several further questions were asked of the participants regarding how they organized the images and interacted with the computer to compare the versions (e.g., side by side, overlaid, or printed); whether they zoomed into the images; whether they used a desktop, laptop, tablet, or other device; how many displays were used and what sizes they were;

what the operating system and image display application were used; and whether they could provide any other information about using the processed images for the transcription task.

## 5.6 Setting

The experimental instructions were sent by email to the participants, who downloaded the images to their computers and eventually sent back the completed questionnaires. Given the virtual nature of the papyrology webinar, the participants never physically convened in the same location as the experimenters, thus precluding testing in controlled laboratory conditions. This constraint, however, turned out to be a great opportunity. The homework “field study” enabled *in situ* simulation of the use of then future legibility enhancement software in realistic conditions in a variety of dimensions, and thus led to findings that could not have been obtained in a laboratory setting. In addition to leaving the choice of evaluation environment and time to participants, the setting also maintained their familiarity with the software and hardware employed. Data organization on personal computers, human-computer interaction patterns developed in coexistence with specific operating systems and applications, as well as the display size and number are only some of the factors that would have biased the results if participants had to work in the unfamiliar laboratory settings.

To more accurately simulate real-life conditions, the image filenames were also not anonymized. The identity of these images is indeed known to real users, since they select methods by name in the enhancement software interface. Furthermore, the methods employed in this evaluation produce images that are sufficiently peculiar to make them identifiable at a glance (such as the negative polarity method), which defeats the purpose of anonymization. The differentiated practices that may develop from this knowledge are part of normal user behavior and represent the insights that the evaluation aims to develop. If anonymization was employed, these real behaviors would not have been captured, meaning that the evaluation would have simulated a non-existent condition. An illuminating example of this rationale concerns the special treatment given by the experimental participants to the original image, which was ranked as the most reliable image for the transcription task despite broad acknowledgment of the better script-background contrast in the enhanced images.

Similar considerations to ours have led to remote evaluations in two other legibility enhancement evaluations of historical documents [12, 28]. Large-scale image and video quality assessments with 8100 and 985 participants from multiple countries, rating 1162 and 3000 images respectively, found the results of laboratory and unconstrained online evaluations to be equivalent [63, 144]. Methodologies

for crowdsourcing image quality assessment have also been actively investigated [78, 152], while the “ecological validity” of laboratory and field experimentation has been debated for over half a century in fields such as psychology, sociology, and economics [57, 76, 90, 124].

## 5.7 Frame

The experiment measured the preferences for legibility enhancement methods in the context of specific documents and available hardware and software, developed in real-life conditions by participants as they gained expertise in using these methods for scholarly text transcription. This statement is necessary to clarify how legibility enhancement was framed by the experimental design. It is not (as would be possible with a different design) transcription errors or duration of task completion that are measured to judge the utility of methods for enhancing legibility, but rather their utility as manifested in their rating and ranking by papyrologists.

## 6 Results

Table 1 presents the ratings and rankings of the image processing methods used in the experiment. Providing this data *in extenso* is essential if we are to make visible the statistical features quantified, summarized, and discussed in the next section. For example, while user behavior heterogeneity is apparent from Table 1, it is obfuscated in Table 3 through statistical data aggregation.

### 6.1 How good are the proposed methods?

A visual inspection of Table 1 suggests that, as a group, the proposed methods ranked overall higher than the comparison methods. This impression is substantiated by the distribution of ratings of the two method sets according to their utility: the proposed methods are prevalently of primary use, whereas the comparison methods are mostly of no use (Table 2).

### 6.2 How do the methods differ?

The methods’ effects on legibility were examined by combining the participants’ feedback with a visual inspection of the enhanced images (Fig. 4) and understanding of the algorithms. The original image was typically low-contrast, and some portions of the script were faded or missing. *Stretchlim* (i.e., normalization of lightness) was often insufficient. *Histeq* created high local contrast variation, because it is based on global contrast improvement. *Adapthisteq* strongly enhanced the papyrus structure, which interferes with the script. *Locallapfilt* had very low contrast, and its

smoothness was too strong; parametrization for individual images becomes thus necessary, although this is difficult to do automatically and inconvenient when done manually. *Retinex* performed best out of all methods at revealing the script, creating very sharp edges. However, reading was also made difficult because of interference from the papyrus structure, and there was also a strong vignetting effect that sometimes completely obscured parts of the image. *LSV* had excellent smoothing properties, but often damaged script structures. *Vividness* gives the most balanced results, although the contrast could be improved.

### 6.3 Which methods should be implemented in software?

To determine the relative utility of the individual methods, their ratings were aggregated across all participants and documents in an overall ranking (Table 3). This ranking was obtained with two different ranking methods, after which the outputs were aggregated and optimized. The top-1 enhancement method is vividness, followed by the negative method. The least useful were *locallapfilt* and *histeq*.

#### 6.3.1 Methodology

While ranking has been extensively investigated in statistics [11], we found methods with more useful properties for the task at hand in the domains of operations research, marketing, sports, and voting theory. These methods are non-parametric, and thus facilitate the handling of data with unknown distributions. They take categorical and ordinal data as input, which corresponds to the experimental ratings and rankings. They also differ in terms of their perspectives, with the mean-based methods being competitive and rewarding the individual performance of the enhancement methods, while the median-based method is consensus-based and satisfies the majority of users. These methods are briefly described below; the code was published in [15].

The *Centroids* ranking is based on matrix analysis [108, pp. 127–129]. The  $n$ -by-1 vector  $\mathbf{r}$  of the aggregated rating scores of  $n$  items with individual scores  $s$  (or ranks  $k$ , for which  $s = n + 1 - k$ ) is defined as follows:

$$\mathbf{r} = \text{round}[(\mathbf{S} - \mathbf{S}^T)\mathbf{e}/n], \quad (9)$$

where  $\mathbf{S}$  is an  $n$ -by- $n$  matrix, with  $S_{ij} = \sum(s_i)$  for  $i \neq j$  and  $S_{ij} = 0$  for  $i = j$  (diagonal of zeros),  $\mathbf{S}^T$  is the transpose of  $\mathbf{S}$ , and  $\mathbf{e}$  is an  $n$ -by-1 vector of ones. (Note the row-wise averaging, from which the method derives its name.) Using Table 3A as an example, the items correspond to the enhancement methods; their corresponding ranks are obtained from Table 1, after which application of the above equations results in the values listed in the rating column

**Table 1** Ranking and rating of legibility enhancement methods.

	<b>AW</b> ♂ 2 displays	<b>FG</b> ♀ 3 displays	<b>FRG</b> ♂ 1 display	<b>LG</b> ♂ 2 displays	<b>OC</b> ♂ 2 displays	<b>OR</b> ♀ 1 display	<b>SR</b> ♂ 2 displays	<b>SN</b> ♂ 1 display
columbia.apis. p367.t.0.600	A	<i>original</i> adapthisteq	A	<b>neglsv</b> lsv	A	<i>original</i> retinex	A	A
	B	<b>neglsv</b> retinex <i>original</i> histeq localapfilt	B	<b>neglsv</b> negnorm norm	N	<b>neglsv</b> negnorm norm	B	B
	N	<b>neglsv</b> lsv negnorm norm <i>original</i> histeq localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt retinex adapthisteq histeq	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt lsv negnorm norm stretchlim	N	N
P.Corn. Inv. MSS. A 101. XIII	A	<i>original</i> adapthisteq	A	<b>neglsv</b> lsv	X	<i>original</i> retinex	A	A
	B	<b>neglsv</b> retinex <i>original</i> histeq localapfilt	B	<b>neglsv</b> negnorm norm	N	<b>neglsv</b> negnorm norm	B	B
	N	<b>neglsv</b> lsv negnorm norm <i>original</i> histeq localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt lsv neglsv negnorm norm stretchlim	N	N
P.CYBR inv. 69	A	<i>original</i> adapthisteq	A	<b>neglsv</b> lsv	A	<i>original</i> retinex	A	A
	B	<b>neglsv</b> retinex <i>original</i> histeq localapfilt	B	<b>neglsv</b> negnorm norm	N	<b>neglsv</b> negnorm norm	B	B
	N	<b>neglsv</b> lsv negnorm norm <i>original</i> histeq localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt lsv neglsv negnorm norm stretchlim	N	N
P.Mich. inv. 1318v	A	<i>original</i> adapthisteq	A	<b>neglsv</b> lsv	A	<i>original</i> retinex	A	A
	B	<b>neglsv</b> retinex <i>original</i> histeq localapfilt	B	<b>neglsv</b> negnorm norm	N	<b>neglsv</b> negnorm norm	B	B
	N	<b>neglsv</b> lsv negnorm norm <i>original</i> histeq localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt lsv neglsv negnorm norm stretchlim	N	N
P.Mich. inv. 2755	A	<i>original</i> adapthisteq	A	<b>neglsv</b> lsv	X	<i>original</i> retinex	A	A
	B	<b>neglsv</b> retinex <i>original</i> histeq localapfilt	B	<b>neglsv</b> negnorm norm	N	<b>neglsv</b> negnorm norm	B	B
	N	<b>neglsv</b> lsv negnorm norm <i>original</i> histeq localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt lsv neglsv negnorm norm stretchlim	N	N
P.Oxy. XXII 2309	A	<i>original</i> adapthisteq	A	<b>neglsv</b> lsv	A	<i>original</i> retinex	A	A
	B	<b>neglsv</b> retinex <i>original</i> histeq localapfilt	B	<b>neglsv</b> negnorm norm	N	<b>neglsv</b> negnorm norm	B	B
	N	<b>neglsv</b> lsv negnorm norm <i>original</i> histeq localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt lsv neglsv negnorm norm stretchlim	N	N
PSI XII 1274 r	A	<i>original</i> adapthisteq	A	<b>neglsv</b> lsv	X	<i>original</i> retinex	A	A
	B	<b>neglsv</b> retinex <i>original</i> histeq localapfilt	B	<b>neglsv</b> negnorm norm	N	<b>neglsv</b> negnorm norm	B	B
	N	<b>neglsv</b> lsv negnorm norm <i>original</i> histeq localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt lsv neglsv negnorm norm stretchlim	N	N
PSI XIII 1298 (15a) r 1	A	<i>original</i> adapthisteq	A	<b>neglsv</b> lsv	X	<i>original</i> retinex	A	A
	B	<b>neglsv</b> retinex <i>original</i> histeq localapfilt	B	<b>neglsv</b> negnorm norm	N	<b>neglsv</b> negnorm norm	B	B
	N	<b>neglsv</b> lsv negnorm norm <i>original</i> histeq localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt lsv neglsv negnorm norm stretchlim	N	N
PSI XIV 1376 r	A	<i>original</i> adapthisteq	A	<b>neglsv</b> lsv	A	<i>original</i> retinex	A	A
	B	<b>neglsv</b> retinex <i>original</i> histeq localapfilt	B	<b>neglsv</b> negnorm norm	N	<b>neglsv</b> negnorm norm	B	B
	N	<b>neglsv</b> lsv negnorm norm <i>original</i> histeq localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt	N	<b>neglsv</b> lsv stretchlim histeq <i>original</i> localapfilt lsv neglsv negnorm norm stretchlim	N	N

Legend: “X” denotes methods used to the exclusion of all others, “A” those that were of primary use, “B” those that were of secondary use, and “N”, on gray background, those that were of no use. Brackets indicate tied rankings



**Table 2** Utility of enhancement methods

Methods	Utility category counts			
	Exclusive X	Primary A	Secondary B	No use N
original	4	22	<b>28</b>	18
<i>Proposed methods</i>				
lsv	0	10	30	<b>32</b>
vividness	3	16	<b>41</b>	12
neglsv	0	13	<b>38</b>	21
negvividness	1	12	<b>36</b>	23
<i>Total proposed</i>	4	51	<b>145</b>	88
<i>Comparison methods</i>				
stretchlim	1	1	<b>42</b>	28
histeq	0	0	22	<b>50</b>
adapthisteq	0	9	25	<b>38</b>
locallapfilt	0	1	23	<b>48</b>
retinex	0	8	<b>33</b>	31
<i>Total comparison</i>	1	19	145	<b>195</b>

*Utility category percentages, column-wise*

Category	Proposed methods	Comparisons
X	80	20
A	73	27
B	50	50
N	31	69

of Table 3A. Their ranking is obtained by sorting the rating vector  $r$  in descending order of magnitude.

The *Ratio of Offense and Defense* (ROD) [108, pp. 176–197] takes as input the same score matrix  $S$  used by the Centroids method, from which it derives the ratings vector  $r$  through the following equation:

$$r = o/d, \tag{10}$$

where the methods’ “offensive” strength is  $o = Re$ , while their “defensive” strength is  $d = R^T e$ . (The terminology comes from the world of sports, where this ranking method originates.)

In *Majority Judgment* each elector grades each candidate (instead of choosing a single one), and rates and ranks are obtained from mean scores, resulting in the elections of those candidates that are the most acceptable to the largest number of electors, i.e. consensual [18]. The algorithm begins by iteratively removing the median values  $\{m_1, m_2, \dots, m_p\}$  from the ratings  $\{r_1, r_2, \dots, r_p\}$  of each candidate  $\{c_1, c_2, \dots, c_p\}$ , and concatenating them lexicographically. For example, the ratings vector  $\{7, 9, 9, 11, 11\}$  is reordered by medians into  $\{9, 9, 11, 7, 11\}$ , and concatenated to

**Table 3** Overall ranking of enhancement methods by utility ratings.

A. Ranking by Centroids method				
Rank	Method	Rating	Spread	
1	<i>original</i>	31.1		
2	<b>vividness</b>	29.1		
3	<b>neglsv</b>	11.1		
4	<b>negvividness</b>	10.1		
5	<b>lsv</b>	-2.9		
6	retinex	-3.9		
7	stretchlim	-5.9		
8	adapthisteq	-9.9		
9	locallapfilt	-27.9		
10	histeq	-30.9		

(1) Graphical representation of the ratings distribution.

B. Ranking by Majority Judgement method				
Rank	Method	Rating	Spread	Category
1	<i>original</i>	2.37		Secondary use +
2	<b>vividness</b>	2.26		Secondary use +
3	<b>neglsv</b>	1.71		Secondary use -
4	<b>negvividness</b>	1.68		Secondary use -
5	stretchlim	1.61		Secondary use -
6	retinex	1.57		Secondary use -
7	<b>lsv</b>	1.56		Secondary use -
8	adapthisteq	1.48		No use
9	locallapfilt	1.33		No use
10	histeq	1.31		No use

C. Aggregation of rankings by ROD method				
Rank	Method	Rating	Spread	Optimization
1	<i>original</i>	$\infty$		<i>original</i>
2	<b>vividness</b>	73.908		<b>vividness</b>
3	<b>neglsv</b>	2.091		<b>neglsv</b>
4	<b>negvividness</b>	1.779		<b>negvividness</b>
5	stretchlim	0.483		stretchlim
6	retinex	0.475		retinex
7	<b>lsv</b>	0.460		<b>lsv</b>
8	adapthisteq	0.204		adapthisteq
9	locallapfilt	0.008		locallapfilt
10	histeq	0.000		histeq

generate the numerical value 9.09110711. The resulting scalars  $\{v_1, v_2, \dots, v_n\}$ , called “majority values”, are then sorted in descending order of magnitude to yield the candidate rankings  $\{k_1, k_2, \dots, k_p\}$ , where  $v(k_1) > v(k_2) > \dots > v(k_p)$ . This procedure was applied to the enhancement method utility ratings  $\{X, A, B, N\}$  of Table 1 in order to generate the rankings and ratings presented in Table 3B. While ranking is an ordinal representation and rating is numeric, verbal qualifiers have their own benefits. For example, the utility of the ranked enhancement methods is easily recognizable from

**Table 4** Intraclass ranking of enhancement methods by utility ratings (optimized aggregation by ROD method; ties in brackets)

**A. Within-user ranking** (column-wise aggregation of ratings from Table 1; ties in brackets)

AW	FG	FRG	LG	OC	OR	SR	SN
retinex original vividness adaphisteq neglsv negvividness stretchlim lsv histeq locallapfilt	original lsv vividness neglsv negvividness stretchlim histeq adaphisteq locallapfilt retinex	vividness retinex histeq neglsv original lsv negvividness stretchlim adaphisteq locallapfilt	negvividness neglsv original vividness lsv stretchlim adaphisteq locallapfilt retinex histeq	vividness neglsv stretchlim lsv adaphisteq locallapfilt original negvividness histeq retinex	original retinex neglsv negvividness locallapfilt lsv vividness stretchlim histeq adaphisteq	negvividness lsv neglsv adaphisteq vividness stretchlim retinex original locallapfilt histeq	negvividness lsv original vividness stretchlim neglsv locallapfilt adaphisteq retinex histeq

**B. Within-document ranking** (row-wise aggregation of ratings from Table 1; ties in brackets)

columbia.apis. p367.f.0.600	P.Corn.Inv. MSS.A 101.XIII	P.CtYBR inv.69	P.Mich. inv.1318v	P.Mich. inv.2755	P.Oxy.XXII 2309	PSIXII 1274r	PSIXIII 1298 (15a) r1	PSIXIV 1376r
neglsv original vividness negvividness lsv adaphisteq stretchlim locallapfilt retinex histeq	original negvividness stretchlim retinex adaphisteq vividness histeq lsv neglsv locallapfilt	original vividness negvividness neglsv stretchlim locallapfilt retinex lsv adaphisteq histeq	vividness original negvividness neglsv retinex stretchlim adaphisteq histeq locallapfilt	vividness original stretchlim neglsv retinex negvividness lsv adaphisteq histeq locallapfilt	original vividness negvividness stretchlim retinex neglsv adaphisteq lsv locallapfilt histeq	vividness neglsv original negvividness lsv retinex stretchlim adaphisteq histeq locallapfilt	original lsv neglsv vividness negvividness stretchlim adaphisteq histeq locallapfilt	neglsv lsv original vividness negvividness retinex stretchlim adaphisteq histeq

the category column of Table 3B, which uses the spelled-out names of the four utility classes {Exclusive, Primary, Secondary, No use} corresponding to the ratings. The plus and minus signs following the class names indicate whether a method belongs to the upper or lower range of a utility class.

Empirical evidence suggests that aggregated opinions may outperform individual ones, an approach formalized by *ensemble methods* [103]. When comparing the results of the Centroids and Majority Judgment methods, differences can be seen in ranking and rating, reflecting the methods’ differing outlooks. To obtain a balanced result, the two rating vectors are aggregated, first by deriving the matrices  $R$  of pairwise positive differences from each of the  $n$  (in our case  $n = 2$ ) rating vectors  $r$ :

$$d_{p(ij)} = [r_{p(i)} - r_{p(j)}], \tag{11}$$

$$R_{p(ij)} = \begin{cases} d_{p(ij)}, & \text{for } d_{p(ij)} > 0, \\ 0, & \text{for } d_{p(ij)} \leq 0, \end{cases} \tag{12}$$

where  $p = 1, 2, \dots, n$ . The matrices are then normalized and averaged:

$$\bar{R} = \sum [R_p / \sum (R_p e)] / n, \tag{13}$$



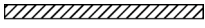



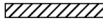












where  $e$  is an  $n$ -by-1 vector of ones. Finally, the aggregated scores vector  $r$  is obtained using the ROD method:  $r = \bar{R}e / \bar{R}e$ . This procedure produced the ratings shown in Table 3C. The approach is akin to graph analysis, as commonly used in, e. g. website ranking methods, since negative differentials between pairwise scores are set to zero, meaning that the matrix  $\bar{R}$  reflects the dominance of each enhancement method.

A further refinement, named *Kemenization*, allows for the aggregated ranking to be optimized [88, pp. 9–23; 108, pp. 175–176]. It consists in switching item  $k_i$  with item  $k_{i+1}$  in the aggregated vector  $r$  if item  $k_{i+1}$  occurs more frequently among the ranks  $i$  of the input vectors  $r_p$ . From the optimization column of Table 3C, we can observe that the procedure results in no modification of the aggregated ranking.

**6.4 Are the evaluated methods sufficient?**

The fact that multispectral images typically provide superior legibility indicates that additional enhancement methods are desirable and sometimes indispensable (e.g., when reading carbonized papyri [125, 171]). Nevertheless, participants attempting further image manipulation, such as in Photoshop, could not usefully improve legibility: “The enhanced pictures were most often so good that this didn’t help as much as I hoped” (SN). Also, one method, i.e. vividness, was substantially better than the others (Table 3). Finally, over-reliance on the original image was apparent: “I like to work on the original version, and I find it’s not always necessary to look at the others” (OR); “I mainly worked with the ‘base’ image (vividness or original). This represented 95% of my work. . . . The other . . . I briefly tried to use them, but quickly abandoned.” (OC). This final comment was reflected in the numerous cases in which methods were deemed not useful (42% of  $n = 720$  in Table 2; swaths of gray in Table 1). Provided that the DIBCO dataset is representative of papyri, the conclusion is that the amount and type of novel and existing methods made available are satisfactory on average.

**Table 5** Agreement of ratings of enhancement methods, measured with Kendall's coefficient of concordance,  $W$ 

A. Inter-class agreement		
Between users	0.376	
Between documents	0.698	
B. Within-user agreement		
1. FG	1	
2. OR	0.727	
3. AW	0.616	
4. SN	0.607	
5. OC	0.500	
6. LG	0.474	
7. SR	0.453	
8. FRG	0.386	
C. Within-document agreement		
1. P.Oxy.XXII 2309	0.376	
2. PSI XIII 1298 (15a) r 1	0.286	
3. P.Corn. Inv. MSS. A 101. XIII	0.283	
4. P.Mich.inv.2755	0.280	
5. columbia.apis.p367.f.0.600	0.280	
6. PSI XIV 1376 r	0.246	
7. P.CtYBR inv. 69	0.243	
8. P.Mich.inv.1318v	0.236	
9. PSI XII 1274 r	0.220	

## 6.5 Enhancement methods are complementary and their utility is context-dependent

Table 1 reveals a substantive variability of method ranking and ratings within participants, between participants, and between documents. Retinex, for example, is the first choice for participant AW, but the last for OC (Table 4). The statistical analysis of agreement and spread quantitatively confirms the striking variability (Table 5 and Fig. 16). Various distributions are present: unimodal, bimodal, and uniform (Fig. 16a). Clearly, *no single method is adequate for every participant-document pairing*. The implication here is that an optimal text transcription can be obtained only by optimizing individual user/document/method configurations, and not by the method that performs best on average; in other words; the choice of enhancement method is best left to the user.

A second insight from Table 1 is provided by participant FG, who stated that all methods were equally useful. Other comments concur: “the images . . . definitely complemented each other” [FG]. Therefore, not only is no single method is the best, but *there is an explicit need for the concurrent use of multiple methods to optimize text transcription*.

The above precept is even more pertinent as image quality varies locally (e.g., AW ranked the method utility differently for the recto and verso of the same papyrus), and local

information is critical to legibility. These findings further support the conclusions drawn from the overall method ranking. Rather than implementing a universal solution for legibility enhancement, *multiple methods should be offered to users*.

These results are unsurprising. The DStretch software, for example, offers more than fifty parameterizable methods. A recent comparison of the performance of eight optical, X-ray, and terahertz-based imaging approaches for recovering text within Egyptian mummy cartonnage concluded that it is only by “carefully selecting, optimizing and combining” them that success may be achieved [64, p. 1]. Similarly, the organizers of the Document Image Binarization competition explicitly state that “no binarization algorithm is good for all kinds of document images” [43, 113]. In industries dealing with color imaging and reproduction, the impossibility of a universal algorithm for color conversion (gamut mapping) is also acknowledged. Various strategies are therefore utilized to select appropriate methods, on the basis of (for example) the user's subjective intention (maintaining saturation or overall color appearance [114, pp. 3–5, 107–109, 194]), the linguistic dimension (preserving the names given to colors [127, pp. 218–219]), or the semantic content (e.g., skin color [127, p. 216]).

### 6.5.1 Comparison

Our findings appear to differ from those obtained by two recent legibility enhancement evaluations of historical documents, which explicitly indicate the homogeneity of agreement between participants. Brenner and Sablatnig [28] produced an annotated dataset of 250 multispectral images from 20 paper and parchment manuscripts, with selected regions of interest graded for their legibility by 20 scholars; the settings were both laboratory and remote, and the interclass correlation coefficient (ICC) [96] was used for statistical analysis. Arsene *et al.*, *Arsene-2018* had seven participants evaluate 15 pseudocolor multispectral fusion methods for their utility in enhancing the legibility of one parchment palimpsest page; a laboratory setting and ICC analysis were used.

There are evident similarities and differences in terms of intent and experimental setup between the above-mentioned works and the present research. We do however consider that the main reason for the behavioral *heterogeneity* observed in our study stems principally from three factors. First, all three evaluations contend with small sample sizes and unconstrained settings, which may amplify variance (while large datasets may hide diversity). Second, we adopted the following methodological “mantras”: do not reject outliers, but question their significance (cf. the case of the original image); examine the raw data (Table 1) before aggregating it (Tables 2, 3, 4, and 5); do not assume normality for the data

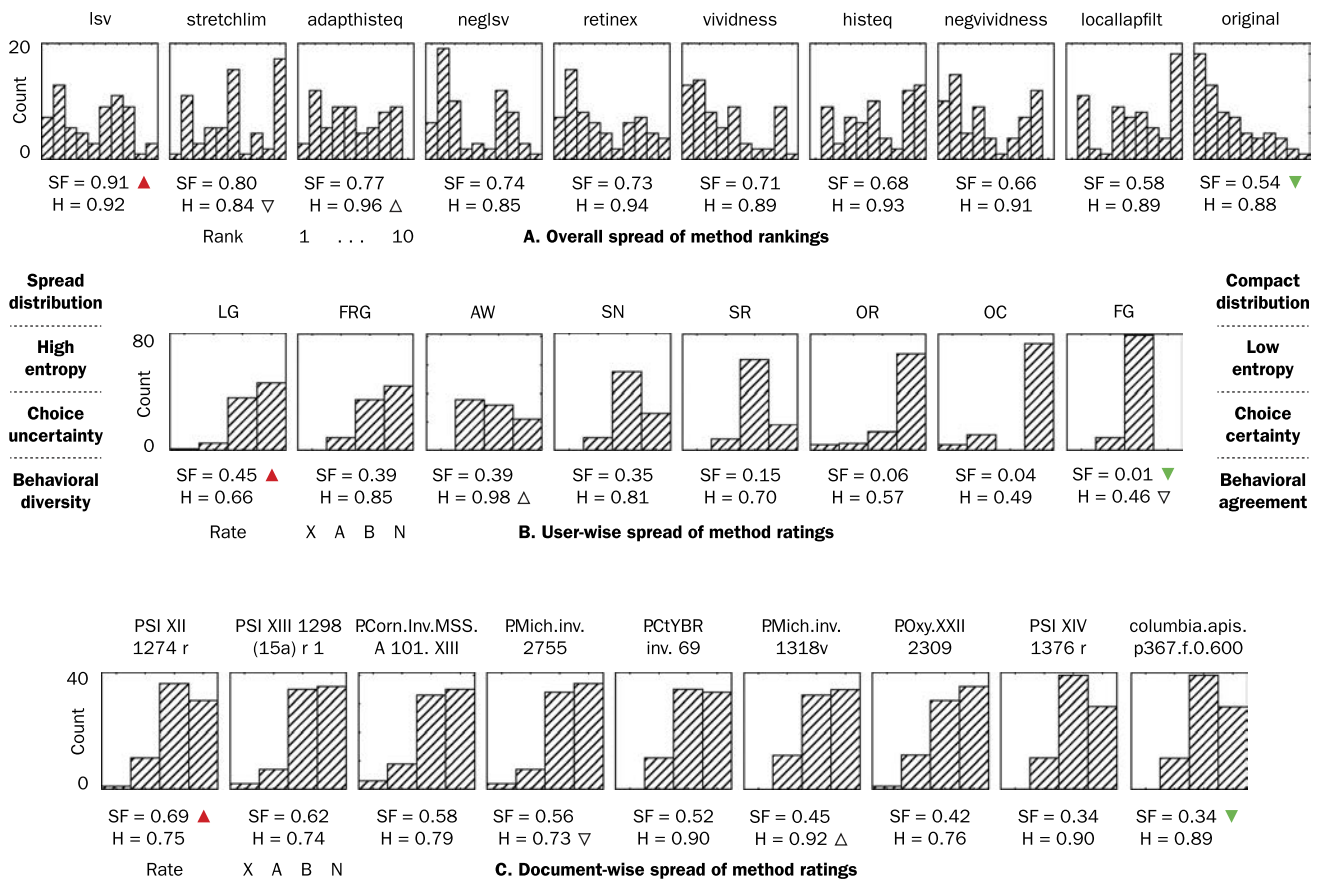


Fig. 16 Spread of enhancement method rankings and ratings, measured with spatial flatness  $SF$  and Shannon entropy  $H$

distribution (while ICC does). Third, the reported behavioral differences may reflect the evaluators' differing outlooks on which experimental aspects to operationalize. The focus is on one hand on personalization and contextualization, and on the other hand on generalization. The choice depends on the situation and affordability: enhancement for critical vision benefits from local optimization, while global optimization is more efficient for casual vision.

### 6.5.2 Methodology

*Kendall's coefficient of concordance*  $W$  is non-parametric and distribution-independent statistic of agreement between multiple rankings, appropriate for the multimodal nature the analyzed data [89, pp. 94–116]. The results in Table 4 were obtained by adjusting for ties and by correcting for small sample sizes, followed by a  $\chi^2$  test of significance for large number of ties. The hypothesis  $H_0$  that the between-user and between-document agreements are due to chance is rejected in favor of the alternative, with  $p = 0.001$  and  $p = 6.124e-9$ , respectively, at an  $\alpha = 0.1$  significance level. The coefficient of concordance,  $W$ , for  $m$  ranked items and  $n$  rankings is defined as

$$W = 12S/[m^2(n^3 - n) + 2 - m \sum (t^3 - t)], \tag{14}$$

where  $S$  is the sum of squares of rank deviations from the mean, and  $t$  is the length of runs of tied ranks. The formula for the  $\chi^2$  value is:

$$\chi^2 = 12S/[m(n^2 + n) + 2 - \sum (t^3 - t)/(n - 1)]. \tag{15}$$

*Spatial flatness*  $SF$  is the complement of the spectral flatness [31, 197], which is itself the ratio of the geometric and arithmetic mean of the magnitude of the Fourier transform  $\mathcal{F}$  of a signal  $x$  of length  $n$ :

$$SF = 1 - n(\prod |\mathcal{F}(x)|)^{1/n} / \sum |\mathcal{F}(x)|. \tag{16}$$

Spatial flatness varies between 0 for an impulse and 1 for a uniform distribution. It is used in signal processing and is implemented as one of the descriptors of the MPEG-7 multimedia content description standard, with applications including classification into noise and tonal sounds [82, pp. 29–32]. We applied this method to characterize histograms. It is a powerful, non-parametric, descriptive statistic that, in a single scalar, conveys information on (1) the coverage

of a range of possible values by actual values in a vector (e.g., to how many different ranks experimental participants assigned a method), (2) the *variance* of the distribution's amplitude (e.g., whether one rank is preferred over the others), and (3) the *homogeneity* of distribution (e.g., whether participants exhibit consistent or eclectic behavior). Such a rich description is made possible by the properties of the frequency domain, where regularities in signals are much easier to investigate than in the spatial domain. For the sake of comparison, Fig. 16 gives the Shannon entropy,  $H$ , of the ranking histograms. The "peakiest" histogram in Fig. 16a has the lowest  $SF$  value, but the lowest  $H$  value is that of a histogram wherein the values are fairly evenly distributed.

## 6.6 From facts to action: an operationalization pattern

The findings on rating variability derived from Table 1 were operationalized via the following design pattern: measurement  $\triangleright$  interpretation  $\triangleright$  implications  $\triangleright$  implementation.

The variability of ratings within columns represents the *within-participant* variability and is indicative of two behavioral attitudes: *eclecticism* in the use of enhancement methods and *adaptability* to document particularities. The prevalence of high eclecticism in the demographic may translate to a decision to accommodate this behavior by simplifying the access of users to the various enhancement methods, e.g., by tiling the interface windows. Conversely, low eclecticism may lead to methods organized in depth, via several menu levels. High adaptability may be considered an incentive to invest in the development of image quality measurement techniques, in order to predict the enhancement methods best suited to a given document, and to rank them dynamically in the software interface. The measurements of agreement in Table 4B show a wide distribution range between eclectic and conservative participants (FG has a uniform stimulus responses), and adaptable and rigid behaviors. The conclusion is that both *depth* and *flat interface designs* are needed, as well as that developing image quality models is desirable.

The principal factors in the level of *within-document rating agreement* in Table 5C are the document and image quality, script legibility, and user experience. This information could aid in evaluating the significance of enhancement method performance. In relating document rankings to their images in Fig. 2, no correlations with either factor are found. For example, moderate agreement characterizes both low and high document or image quality (PSI XIV 1376 r vs columbia.apis.p367.f.0.600), whereas low agreement is obtained for all script cursivity levels (PSI XII 1274 r vs P.Mich.inv.1318v). No conclusions are drawn from these data.

Comparison of *between-user* and *between-document rating agreement* may reveal the dominant factor driving

the observed behavioral heterogeneity in the rating of the enhancement methods, owing either to user variability or to document quality variability. The results would have bearing on whether to invest more in *human-computer interaction* issues or *enhancement algorithms*. The values in Table 5A show a greater agreement in between-document ratings than between-user ratings, and thus support the focus on interaction facilities, more than on enhancement algorithms. This empirical finding is a surprising conclusion to an article devoted to algorithms, and thus merits further exploration.

## 6.7 Approaching legibility enhancement as a system

The preceding findings detailed the *high heterogeneity of user behavior* and the *large number of factors* affecting which legibility enhancement methods users choose to utilize. The following model of the legibility enhancement process aims to help better understand these phenomena.

The SYSTEM takes *bitmap* text as input and outputs *electronic text*. During this process the *entropy* of the text is lowered (i.e., a cryptic text has been read). The system is composed of *interacting elements* and their *properties*. The principal aspects are: (1) the TASK of critical reading, (2) the IMAGES with various degrees of content *affordance*, (3) the TOOL creating an ergonomic environment (such as image viewing software and displays), (4) the METHODS of legibility enhancement that are variously useful for the users, and (5) the USERS themselves, along their level of expertise in interacting with the system components. The BEHAVIOR of users emerges from the processes within the system, and is characterized by an idiosyncratic eclecticism in how they use the methods and the tools, as well as by contextual *adaptability* to the specificities of documents and their digital reproductions. LEGIBILITY is a property not only of images alone, but also of user *expertise* and tool *ergonomy*. Likewise, the *utility* of a legibility enhancement method is a property of the interaction among all elements of the system; therefore, its measurement is relative to the overall context.

## 6.8 Contextual system optimization is preferable

Image interpretation in scholarly transcription, and even more so in medical, forensic, and military contexts, can have significant repercussions on people's lives (e.g., the role of the Rosetta Stone transcription in the Egyptian tourism industry). Optimal performance on a case-by-case basis, rather than considered statistically, is thus critical. This can be achieved only if the "variables" of the legibility system (i.e., methods and tools) are attuned to the "constants" (i.e., user, document, and reproduction). In short, choosing the *globally* optimal method for all combinations of the system states will yield suboptimal performance in most cases; only

when several analysts express themselves on the same document can their opinions be fused into a unique solution that might surpass individual solutions. To use an analogy, biometric security systems are optimized for individual users, whereas physical keys are universally functional for any user; evidently, the average of biometric signals from multiple users will result in very poor authentication performance.

The evidence in respect to legibility enhancement thus favors the *personalization* of methods and tools to specific users or demographics, the *contextualization* to the task, and the adaptation to individual documents and reproductions.

One simple implementation of a legibility enhancement system optimized to local conditions would be to offer a spectrum of methods and then *leave the choice to the users*. More sophisticated techniques are the ranking of methods according to their suitability for the document in question, and adaptation to the behavior of a given user. A fully-fledged context-aware critical visual enhancement system (e.g., autonomous robots) would have at its disposal a multitude of enhancement methods to choose from and combine in the function of the observed distortion types, scene content, environmental conditions, technological resource application domains, and user demographics and individualities.

### 6.8.1 Related work

Digital photo editing is a research area where adaptation to individual and demographic preferences has been proposed as a way to develop image enhancement processes beyond the stage of universal filters [85]. Typical features of interest are the lightness transfer function (S-curve), temperature, and tint, while more complex edits may be modeled using deep learning [91]. Taking user intent and image content into account are some of the challenges of this application type.

Demographics and individual adaptation have long been of interest for medical applications of analog and computational image enhancement. For example, macular degeneration is a common ophthalmologic pathology manifested by a blurred spot in the visual field. The detrimental impact of this condition, including the handicap it presents for reading, may be mitigated by bandpass filtering based on psychophysical models of low vision [137], a technique that could be implemented in an MPEG image decoder for real-time applications [92]. The progressive yellowing of the eye's crystalline lens with aging is another example of the demographic dimension of image enhancement. It leads to a darker field of vision, a condition for which it has been proposed to use lightness enhancement by means of non-linear tone mapping and high-pass filtering (sharpening) [126]. The several types and degrees of colorblindness (affecting about eight percent of males and one percent of females of European ancestry [120, 135, pp. 222–223]) have spurred the development of a further class of demographic-specific

image enhancement techniques intended to simulate color vision deficiency for the normal-sighted [116, 186], as well as enhancing the color perception of the colorblind [53, 150, 196].

One challenge associated with medical-related image enhancement concerns the wide variety and individuality of pathologies that need to be accounted for in software [59]. Furthermore, the suboptimal ergonomics of the hardware in which this software embedded is often a reason why they fail to be used [136]. Consequently, a number of creative design solutions have been proposed, such as the integrated visual and audio representations of images [67]. An additional persistent problem remains the difficulty of evaluating the effectiveness of the experimental enhancement methods outside the laboratory, as part of the daily life of the visually impaired [138]. This aspect has considerable significance, given the far greater disparity for pathological applications than for non-pathological ones between laboratory and real-life environments (in a similar manner as we suggest to be the case for paleographic transcription).

## 6.9 The primacy of the “original” image

Overall, papyrologists ranked original images as the most useful for transcription. Participant SN is unambiguous: “I’m convinced that the original picture is the most important reference point in discussing a text.” Why should this be so, and what might it mean for software development?

In many domains—papyrological, medical, or journalistic—the more removed the information from its source, the less trustworthy it is considered: (cf. witness vs hearsay evidence, primary vs secondary documentary sources). The dogma of the “original” is a socio-professional specificity inculcated early on in papyrological education. As sensible this is in the context of physical documents examined visually, digital reproductions are, however, inherently “manipulated” images, and an image at the exit point of an imaging system is no more faithful than, say, an image calibrated during post-production. Hence, suitable documentation should clarify alterations by the enhancement methods, and thus preserve the confidence of users in that they understand the source of features observed in images.

The “low toner effect” might also have contributed. Psychologists found that, counterintuitively, a job application letter of poor print quality was more trustworthy than a pristine print. The proposed explanation was that low legibility demands more cognitive effort, which is interpreted as a reflection of the high value of the letter content [132, pp. 149–151]. The effect was reproduced with mathematical exercises more often correctly solved when in an uncommon faded typeface [10].

It would however be misleading to rely solely on the overall ranking of enhancement methods (Table 3) and thereby

conclude that the original is the top-performing image under all conditions. For example, between-user variability is substantial: whereas OR consistently rates the original as being of “exclusive use” or “primary use”, OC rejects it as “not useful” (Table 2). The software should thus be able to adapt to user preferences, as well as document specificities.

### 6.10 The unexpected usefulness of negative images

The trivial negative method is noteworthy not only for its excellent performance as a top-2 enhancement method after vividness (Table 2), but also for its *discriminatory qualities*. It helps to ascertain whether a specific pixel cluster might be ink (“I used the negative image to understand whether I was imagining an ink trace or it was real.” [OC]), and to distinguish between ink and holes in the papyrus, a recurrent issue in reading papyri reproductions. One further point of practical relevance regarding this method is that it is applicable to *monochrome images*. Some papyri survive or are remotely accessible only in this format; other (such as the charred papyri) have little chromatic information.

The finding that negative polarity is considered by users to improve legibility is of interest because it is contrary to the prevailing *psychophysical evidence*, which consistently associates better legibility with positive contrast (i.e., dark script on light background) [26, pp. 206–207; 32; 99; 109 145]. The possible cause of the disparity is that these studies typically use binary black and white stimuli, rather than ternary black, gray, and white stimuli, as in the present experiment.

Negative polarity is known to be preferable only for special conditions, such as the degeneration of visual acuity in aging [26, pp. 244–243] or disease [109]. Also, aeronautical displays with negative polarity help to maintain the adaptation of the visual system to a dark environment during night flight [148]. Closer to papyrology is the similar case of readers repeatedly switching between the high luminance of a computer screen and the low luminance of paper documents [20]. These exceptions have a context of *low vision* in common, whether permanent or transitory, that demands increased visual and cognitive efforts. In other words, they are akin to critical reading tasks, such as the decipherment of ancient documents, characterized by high visual and semantic entropy.

In terms of *visual ecology*, positive contrast has dominated the history of writing, albeit mostly for practical reasons such as the cost of colored substrate. Negative polarity texts are often prestige objects, for example, the Rosetta inscription that features a light script with crystalline sparkle on a dark stone [122]; the word “God” calligraphed in gold on a green eight-meter-high panel in the Aya Sofia mosque of Istanbul [25, p. 504]; and the 12th-century “Blue Sutra” of Japan, in gold on blue paper [58, pp. 36, 68–69]. These

examples are also sources of inspiration for how contrast—and thus legibility—might be further improved, using displays with special material properties, such as high-reflectance or fluorescence.

### 6.11 Smooth human–computer interaction is critical

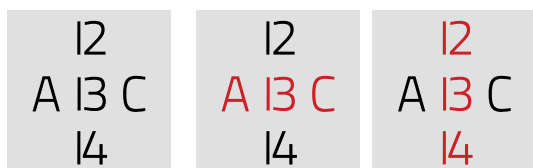
Participant AW reported: “Even with the split-screen option, and using two monitors, I felt like my devices lacked the optimal Graphical User Interface to maximize the usefulness of those pictures. Comparing the same zoomed part of text through several images proved, for example, to be quite costly in terms of time and clicks, and I feel this disturbed the workflow.” Five of eight participants juxtaposed two or more image variants during the transcription work; three used two displays, while one used three displays. Two participants printed the images (as more hard copies can be spread out on a table than on a computer display).

These explicit and implicit statements demonstrate how interaction with the images can impact the advantages derived from image processing. Indeed, without good interaction, there is little enhancement as far as the user is concerned. What, then, is *good interaction*? Observing the participants is informative in this regard: good interaction involves a wide visual field, the synchronization of manipulations across multiple images, the understanding of how different enhancement methods have varying utility for transcription, and a state of mental flow during task completion [42].

### 6.12 Gender may affect legibility enhancement

One female participant in the experiment had perfect agreement in her method ratings, while the second female had the second highest agreement score (Table 5B). If they are not statistical outliers, then the gender disparity in behavior could have implications, given the increased proportion of females in papyrology (202 [57%] out of 356 participants at the International Congress of Papyrology in 2019 were female, compared to 3 [5%] out of 57 at the congress in 1937 [6, 37]). For example, females have been found to have greater chromatic discrimination ability, whereas males are more sensitive to lightness variation [190]. Therefore, the enhancement method through chromatic contrasting might improve legibility more for females, while the negative polarity method could be more appropriate for males. A larger sample size would be needed to further the investigation of the topic.

However, the significance of this section is broader. In addition to gender, there are numerous other perceptual variability factors (age-related, genetic, cultural, and climatic, among others) [49, pp. 149–244] that indicate the utility



**Fig. 17** The shapes in the center of the squares are perfectly legible, but ambiguous: do they represent the letter “B”, the figure “13”, or something else? Coloring helps differentiate between alternatives. Such shapes are said to be “multistable stimuli” [101], and have been the source of well known illusions, e. g. Arcimboldo’s face made of books [164, p. 131]

of demographically tailored and even personalized image enhancements (in a similar fashion to how corrective glasses are fitted individually) rather than those calibrated to the statistical mean of the general population.

## 7 Paradigm

A paradigm of legibility enhancement for critical applications is provided herein, representing a step towards developing computer-assisted critical vision systems. This paradigm helps defining the goals, methods, and limits of legibility enhancement, and is rooted in the problem analysis and experiment carried out in the present research. The preceding sections have circumscribed the notion of critical vision, established that potential information should not be suppressed, and advocated a systems approach in the design of enhancement software. A further element will be discussed below: uncertainty. This element will help operationalize the paradigm and link the field of image enhancement to the vast interdisciplinary research on uncertainty.

### 7.1 Requirements

Because vision is purposeful, we might sometimes stare at things without seeing them when there is no reason to seek them out (“inattentional blindness” [172]). Two converse problems concern our tendency to believe what we see and see what we believe (“confirmation bias” [169, pp. 61–75]). Generally speaking, the ideal enhancement system would inform us about objective uncertainty in images and guard against subjective cognitive biases during their interpretation. Uncertainty is important and its presence should be made explicit. The system should therefore aid in its detection and analysis, represent it visually and numerically, and preserve it for verification. *Uncertainty management should be at the core of the enhancement system and its components, from the algorithms to the interface.*

### 7.2 Examples

One illustrative analogy is the use of white material for pottery reconstruction in archeology to distinguish between missing and original parts. An example of script polysemy and the role of context in reading is shown in Fig. 17, wherein the central shapes can be interpreted as either letters or numbers depending on the reading direction. Mundane cases where an individual seeks to mentally solve visual ambiguities include faded maps or letters and frightening shadows at night, while imagery-based clinical diagnosis and military decisions represent domains in which uncertainty is both pervasive and critical. For the facial identification of suspects, police organizations have developed mechanical and computational aids for using witness recollections to produce faces composited from a catalog of elements, such as eyes, nose, and mouth, assembled by likelihood into a coherent whole [44].

### 7.3 Rationale

A paradigm called *conservative preprocessing* was proposed by Chen, Lopresti, and Nagy in a rare work on the theory of image enhancement [38]. It consists in preserving the original by means of reversible transformations and the use of derivative representations as a proxy. Other researchers have noted that applying the concept of noise suppression outside the casual reading task might be misplaced: for example, bleed-through removal may unwittingly obliterate the only trace of text that is inaccessible because the page is glued to the support, was not scanned, or belongs to a missing second or third page [180]. Such invasive enhancement can thus destroy historical, forensic, and conservation-related information in documents and reproductions [60, 129]. Problems of this kind are common in digital library projects. Technical factors can also make enhancement problematic; for example, because band-pass filtering is known to frequently suppress script parts [133, p. 445], while neural networks can “hallucinate” character-like shapes [21, p. 1221], should the user not be made aware of the possibility of erroneous readings?

### 7.4 Conceptualization

Let us here define uncertainty in the context of image analysis and consider possible solutions. As discussed above, image enhancement is the outcome of interacting stimuli, observers, tasks, contexts, and technologies. Multiple concurrent enhancements therefore appear to be better adapted to the high combinatorial variability potential of such a SYSTEM.



To clarify the uncertainty reduction PROCESS, we draw on cognitive psychology (in particular, decision-making in critical contexts, such as those encountered by medical emergency services, firefighters, police, and military personnel [169]). The process consists of an *interpretation* stage, in which possible stimuli of interest are detected and separated from the surrounds (building “situational awareness”), various options as to their identification are formulated (integration into coherent “gestalts”), and their likelihood is estimated (multidimensional weighting and/or ranking); this is followed by a *decision* stage, in which a choice of interpretations is made. For our proposed method, it is noteworthy that (1) *alternatives* to the visual stimulus are produced, (2) the visualization is *mental*, rather than physical, and (3) it involves mental *manipulations* of the alternative images.

Establishing a taxonomy of common FACTORS leading to uncertainty helps us to find appropriate remedies. The inability to detect visual targets (*lack* of information) and vaguely defined targets (*fuzzy* information) are types of uncertainty that arise early in the interpretation process and that may be countered by more and sharper imagery; that is, alternative views. Decisions regarding *implausible* interpretations may be solved by considering the target’s compatibility with its visual context, which shows the utility of retaining context and even using alternative backgrounds. The uncertainty developing from the need to choose between *contradictory* interpretations demands an increased capability to compare alternatives. Entities need not necessarily be noisy to make interpretation difficult (viz. readily distinguishable characters in unknown scripts); to deal with *novelties*, information exogenous to the image and observer must be supplied. Finally, an *overabundance* of potential interpretations may become overwhelming and result in “decision inertia” [169, pp. 37–41]. Information visualization theory suggests for such cases filtering the amount of information, through selection or ranking [168].

In terms of LIMITATIONS, the number of items processable at a given time in human memory is however finite and can degrade under time pressure and stress; similarly, the visual field size is also limited and increasingly blurred as it moves away from the fovea. Thus, methods designed to *increase* the number of alternatives must be accompanied by mechanisms to *limit* visible alternatives and *manage* all available stock.

## 7.5 Operationalization

We propose to make uncertainty materially explicit through the physical rendering and cueing of mental interpretations of images (as a literal application of the notion that “seeing is believing”). When the alternatives are known, highlighting or inpainting of objects may be used to find fits between

potential targets and a given context, as in a puzzle game where the pieces are materializations of mental images. For unknown alternatives, solutions may be elicited by purposive or random alterations at the global or local level, statically or as a movie, similar to the shimmering reflection on an undulating water surface, or searching for optimal shapes by sketching, or tilting a hologram to observe it from different perspectives (our deliberate use of metaphors is a method for problem solving, as well as for communication [75, 106]). Machine learning may help shape prediction and synthesis, while human-machine interaction techniques are used to guide the image exploration. The result is much more than an image enhancement algorithm: it is a *computer-assisted critical vision system*. Current research on such systems includes applications for, e.g., medical radiographic diagnosis [158, pp. 359–414].

This method operates as illustrated using Fig. 17, in which two images were derived from the original; each of these make explicit a different interpretation through the perceptual differentiation of distinct semantic classes. If semantic grouping became unfeasible, the random alternation of individual shapes between two colors would be the next best option (also revealing a less obvious solution when only the three vertical bars are red, in that they may represent the upper-case letter “I”, the lower-case letter “i”, or the figure “1”). The inpainting technique presented in this example consists in removing the gaps between adjacent central shapes to strengthen their identity as the letter “B”. In the real case of papyri, the most common character allographs of interest may be statically inpainted or flickered. This is a complex and possibly contentious intervention in the image structure; thus, in our implementation, we avoided segmentation by opting to present variants of the whole scene.

## 7.6 Frameworks

The presented method for making uncertainty explicit in view of helping decision-making may be understood in information theoretical terms as finding a balance between maximal (= quandary) and minimal choice (= no or unknown alternative) by comparing image alternatives with different entropies for different entities within the images. The heterogeneity of user behavior observed in our experiment suggests that the optimum is contextually determined by a combination of factors.

Bayesian theory offers an additional useful theoretical framework, in that establishing the identity of the visual target is dependent on both the prior assumptions of the observed shape having the hypothesized identity and the likelihood of observing the hypothetical entity in the given context. In this sense, the explicit presentation of multiple

alternatives for comparison is a means of updating and optimizing one's beliefs about the identity of visual stimuli.

## 7.7 Criteria

Our work yielded a set of core criteria defining our paradigm of legibility enhancement for critical applications:

1. *Critical vision*: Legibility enhancement for critical applications is a matter of critical vision. It consists in exerting skepticism about the interpretation of images.

2. *Potential information*: Any image structure should be viewed as potential information that should not be suppressed in critical vision applications; interpretations arise from connections made between image parts.

3. *Uncertainty explicitation*: Support for critical vision is realized by making alternative image interpretations materially explicit. Visual target identity depends on both the shape and its context.

4. *Systems approach*: Legibility enhancement for critical applications is a system problem. Optimization depends on tasks, data, users, tools, and their interactions.

## 8 Implementation

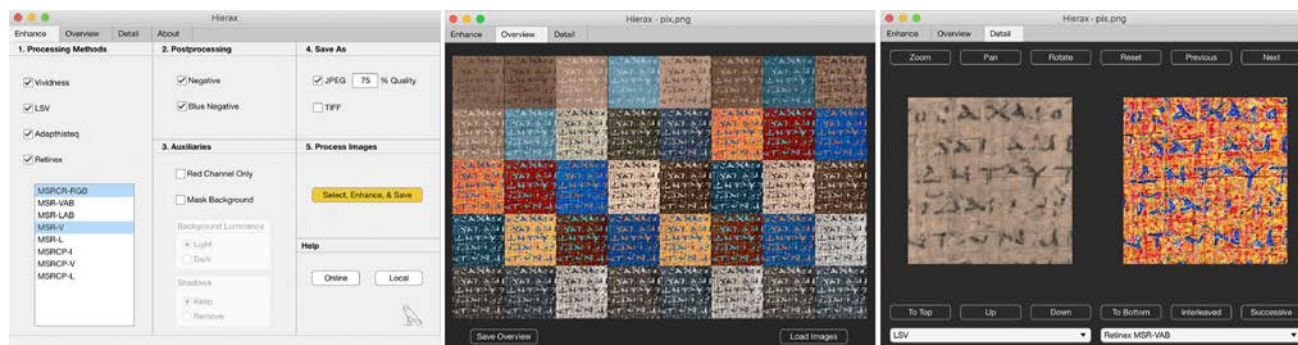
The research findings were implemented in the open-source software Hierax [16]. To account for the heterogeneity of user behavior and the complementarity of enhancements, a range of methods were offered: our novel methods, and *adapthisteq* and *retinex*. Since interaction with the images was found to affect the usefulness of enhancements, the software includes an image viewer. Our design methodology conforms to elements of the so-called information visualization “mantra” for graphical user interfaces: overview, filter, detail, and context [168]. One interface tab presents an overview of one image enhanced using the selected methods

(Fig. 18); through the spatial juxtaposition of patterns, the resulting mosaic facilitates understanding of the characteristics of different enhancement methods. A second tab allows for detailed side-by-side comparison, with zooming, panning, and rotation being synchronized; the diptych view is intended for the interpretation of details in close coupling of vision and motor control. Rapid switching between images via the keyboard is a temporal image analysis modality that leverages attentional capture through flicker, introduced to help detect potential features of interest [193]. The mosaic, diptych, and flicker image analysis techniques implement the paradigm of uncertainty explicitation at the whole scene level. These facilities were appreciated by the paleographers who beta-tested the software. Hierax has already been employed in a real-world setting, to enhance the 70 papyri of the University of Basel Library for a forthcoming online catalog.

## 9 Conclusions

This work aimed to improve the legibility of ancient papyri for text editions. Novel enhancement methods were developed on the basis of color processing and visual illusions. A user experiment demonstrated that these methods outperform classical and more complex enhancement methods. Future work could examine whether the new methods could also improve reading by machines.

The experiment also yielded unexpected findings. Critical for the software, user behavior was found to be *heterogeneous* in type, and not only variable in intensity. Thus, better performance might be achieved by *personalization*, *contextualization*, and *complementarity*, instead of searching for an overall optimum that may be spurious or counterproductive. Experiments with more users and images should be conducted to strengthen, nuance, and perhaps diversify the findings.



**Fig. 18** The graphical user interface of the Hierax software for papyri legibility enhancement, which implements the methods described in this article. Method selection and parametrization (left); overview of all processed images (center); image comparison and interaction panel (right)

The findings also lead to a paradigm of legibility enhancement for critical applications, in view of a future computer-aided critical vision system.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10032-021-00386-0>.

**Acknowledgements** This article is an output of the project n° PZ00P1\_174149, *Reuniting fragments, identifying scribes and characterizing scripts: the digital paleography of Greek and Coptic papyri*, funded by the Swiss National Science Foundation. The experimental participants are heartily thanked for their valuable efforts. Gorge Nagy, Andreas Fischer, Irvin Schick, and Bernice Rogowitz, as well as the three anonymous reviewers, offered much appreciated critique. The authors wish to acknowledge the proofreaders' help in improving the manuscript readability: Jacqueline M. Huwyler, Thomas Hardy of Science Journal Editors, and K. F. of Proofed. — Dedicated by V. A. to Aries Arditì and Ruth Rosenholtz, for the vision from the lighthouse.

**Author Contributions** V. A. conceived the algorithms, programmed the software, designed the experiment, analyzed the results, and wrote the article. I. M.-S., project leader, defined, planned, and supervised the research (data collection, guidance, tests, experiment coordination, critical revision of the manuscript).

**Funding Information** Open Access funding provided by Université de Fribourg.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abidi, B.R., Zheng, Y., Gribok, A.V., Abidi, M.A.: Improving weapon detection in single energy X-ray images through pseudo-coloring. *IEEE Trans. Syst. Man Cybern. C (Appl. Rev.)* **36**(6), 784–796 (2006)
- Adelson, E.H.: Lightness perception and lightness illusions. In: Gazzaniga, M.S. (ed.) *The New Cognitive Neurosciences*, 2nd edn., pp. 339–351. MIT Press, Cambridge, MA (2000)
- Afruz, J., Wilson, V., Umbaugh, S.E.: Frequency domain pseudo-color to enhance ultrasound images. *Comput. Inform. Sci.* **3**(4), 24–34 (2010)
- Alaei, A., Raveaux, R., Conte, D., Stantic, B.: “Quality” vs. “Readability” in document images: statistical analysis of human perception. In: *Proceedings of the 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 24–27 April 2018, Vienna, Austria, pp. 363–368 (2018)
- Alexopoulou, A.A., Kaminari, A.A., Panagopoulos, A., Pöhlmann, E.: Multispectral documentation and image processing analysis of the papyrus of tomb II at Daphne, Greece. *J. Archaeol. Sci.* **40**(2), 1242–1249 (2013)
- Allberry, C.R.C., et al.: *Actes du Ve Congrès International de papyrologie*, Oxford, 30 Août – 3 Septembre 1937. Fondation égyptologique Reine Élisabeth, Bruxelles (1938)
- Allegra, D., Ciliberto, E., Ciliberto, P., Milotta, F.L.M., Petrillo, G., Stanco, F., Trombatore, C.: Virtual unrolling using X-ray computed tomography. In: *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, 31 August–4 September 2015, Nice, France, pp. 2864–2868 (2015)
- Alley, R.E.: *Algorithm Theoretical Basis Document for Decorrelation Stretch*, ASTER-AST06. Jet Propulsion Laboratory, Pasadena, CA (1996). <http://www.dstretch.com/DecorrelationStretch.pdf>
- Altamura, O., Berardi, M., Ceci, M., Malerba, D., Varlaro, A.: Using colour information to understand censorship cards of film archives. *Intl. J. Doc. Anal. Recogn.* **9**(1), 281–297 (2007)
- Alter, A.L., Oppenheimer, D.M., Epley, N., Eyre, R.N.: Overcoming intuition: metacognitive difficulty activates analytic reasoning. *J. Exp. Psychol.* **136**(4), 569–576 (2007)
- Alvo, M., Yu, P.H.L.H.: *Statistical Methods for Ranking Data*. Springer, New York, NY (2014)
- Arsene, C.T.C., Church, S., Dickinson, M.: High performance software in multidimensional reduction methods for image processing with application to ancient manuscripts. *Manuscript Cultures* **11**, 73–96 (2018)
- Asano, Y., Fairchild, M.D., Blondé, L.: Individual colorimetric observer model. *PLoS ONE* **11**(2), e0145671 (2016)
- Atanasiu, V.: *Ugraphia: The Utopia of a Perfectly Legible Script*. /grafematik/ Collection of extended abstracts of the Conference Graphemics in the 21st century: From graphemes to knowledge, 14–15 June 2018, Brest, France, pp. 113–140 (2018)
- Atanasiu, V.: Non-parametric rating and ranking functions. *MATLAB Central File Exchange* (2020). <https://www.mathworks.com/matlabcentral/fileexchange/78384>
- Atanasiu, V., Marthot-Santaniello, I.: Hierax legibility enhancement software (2020). <https://hierax.ch>
- Bagnall, R.S.: *The Oxford Handbook of Papyrology*. Oxford Univ. Press, Oxford (2009)
- Balinski, M., Rida, L.: *Majority Judgment Measuring, Ranking, and Electing*. MIT Press, Cambridge, MA (2010)
- Barten, P.G.J.: *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. SPIE Optical Engineering Press, Bellingham, WA (1999)
- Bauer, D., Cavonius, C.: Improving the legibility of visual display units through contrast reversal. In: Grandjean, E., Vigliani, E. (eds.) *Ergonomic aspects of visual display terminals*, pp. 137–142. Taylor & Francis, London (1980)
- Belthangady, C., Royer, L.A.: Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat. Methods* **16**, 1215–1225 (2019)
- Berns, R.S.: Extending CIELAB: Vividness,  $V^*_{ab}$ , Depth,  $D^*_{ab}$ , and Clarity,  $T^*_{ab}$ . *Color Res. Appl.* **39**(4), 322–330 (2014)
- Beutel, J., Van Metter, R.L., Kundel, H.L. (eds.): *Handbook of Medical Imaging*, vol. 1. SPIE, Bellingham, WA (2000)
- Bhowmik, S., Sarkar, R., Nasipuri, M., Doermann, D.: Text and non-text separation in offline document images: a survey. *Intl. J. Doc. Anal. Recogn.* **21**(1–2), 1–20 (2018)
- Blair, S.: *Islamic Calligraphy*. Edinburgh Univ. Press, Edinburgh (2006)
- Boff, K.R., Lincoln, J.E. (eds.): *Engineering Data Compendium: Human Perception and Performance*, vol. 1. Harry Armstrong Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio, OH (1988)

27. Bouwmans, T., Javed, S., Zhang, H., Lin, Z., Otazo, R.: On the applications of robust PCA in image and video processing. *Proc. IEEE* **106**(8), 1427–1457 (2018)
28. Brenner, S., Sablatnig, R.: Subjective assessments of legibility in ancient manuscript images: the SALAMI dataset. In: *Proceedings of the ICPR International Workshops and Challenges*, 10–15 January 2021, Virtual Event, part VII, pp. 68–82 (2021)
29. Breuil, C., Jennings, B.J., Barthelmé, S., Guyader, N., Kingdom, F.A.A.: Color improves edge classification in human vision. *PLoS Comput. Biol.* **15**(10), e1007398 (2019)
30. Brink, A., van der Klauw, H., Schomaker, L.: Automatic removal of crossed-out handwritten text and the effect on writer verification and identification. In: *Proceedings of the SPIE-IS&T Electronic Imaging Conference, Document Recognition and Retrieval XV*, 29–31 January 2008, San Jose, CA, USA, vol. 6815: 68150A (2008)
31. Broersen, P.M.T.: *Automatic Autocorrelation and Spectral Analysis*. Springer, London (2006)
32. Buchner, A., Mayr, S., Brandt, M.: The advantage of positive text-background polarity is due to high display luminance. *Ergonomics* **52**(7), 882–886 (2009)
33. Buckle, C.E., Udawatta, V., Straus, C.M.: Now you see it, now you don't: visual illusions in radiology. *Radiographics* **33**(7), 2087–2102 (2013)
34. Bülow-Jacobsen, A.: Writing materials in the ancient world. In: Bagnall, R.S. (ed.) *The Oxford Handbook of Papyrology*, pp. 3–29. Oxford Univ. Press, Oxford (2009)
35. Bülow-Jacobsen, A.: Photography of Papyri and Ostraca. In: Caputo, C., Lougovaya, J. (eds.) *Using Ostraca in the Ancient World*, pp. 59–83. De Gruyter, New Discoveries and Methodologies, Berlin (2021)
36. Cannon, M., Hochberg, J., Kelly, P.: Quality assessment and restoration of typewritten document images. *Int. J. Doc. Anal. Recogn.* **2**(2–3), 80–88 (1999)
37. Capasso, M., Davoli, P.: Abstracts of the 29th International Congress of Papyrology, 28 July–8 August 2019, Lecce, Italy. Università del Salento, Lecce (2019)
38. Chen, J., Lopresti, D., Nagy, G.: Conservative preprocessing of document images. *Int. J. Doc. Anal. Recogn.* **19**(4), 321–333 (2016)
39. Christen, M., Vitacco, D.A., Huber, L., Harboe, J., Fabrikant, Brugger, P.: Colorful brains: 14 years of display practice in functional neuroimaging. *Neuroimage* **73**(6), 30–39 (2013)
40. CIE: Colorimetry, TR 15:2004. CIE, Vienna (2004)
41. Ciortan, I., Deborah, H., George, S., Hardeberg, J.Y.: Color and hyperspectral image segmentation for historical documents. In: *Proceedings of the 2015 Digital Heritage Conference*, 28 September–2 October 2015, Granada, Spain, pp. 199–206 (2015)
42. Csikszentmihalyi, M.: *Flow: The Psychology of Optimal Experience*. HarperPerennial, New York, NY (2008)
43. DIB: Document Image Binarization (2021). <https://dib.cin.ufpe.br>
44. Davies, G., Valentine, T.: Facial Composites: Forensic Utility and Psychological Research. In: Lindsay, R.C.L., Ross, D.F., Read, J.D., Togli, M.P. (eds.) *Handbook of Eyewitness Psychology: Volume 2, Memory for People*, pp. 59–82. Psychology Press, New York, NY (2014)
45. Donofrio, R.L.: Review paper: the Helmholtz–Kohler effect. *J. Soc. Inform. Disp.* **19**(10), 658–664 (2011)
46. Drira, F., LeBourgeois, F.: Mean-Shift segmentation and PDE-based nonlinear diffusion: toward a common variational framework for foreground/background document image segmentation. *Int. J. Doc. Anal. Recogn.* **20**(3), 201–222 (2017)
47. Easton, R.L., Knox, K.T., Christens-Barry, W.A.: Multispectral imaging of the Archimedes palimpsest. In: *Proceedings of the 32nd Applied Imagery Pattern Recognition Workshop (IEEE-AIPR'03)*, 15–17 October 2003, Washington, DC, USA, pp. 111–116 (2003)
48. Elhedda, W., Mehri, M., Mahjoub, M.A.: Hyperkernel-based intuitionistic fuzzy c-means for denoising color archival document images. *Int. J. Doc. Anal. Recogn.* **23**(3), 161–181 (2020)
49. Elliot, A.J., Fairchild, M.D., Franklin, A. (eds.): *Handbook of Color Psychology*. Cambridge Univ. Press, Cambridge (2015)
50. Fairchild, M.D.: Refinement of the RLAB color space. *Color Color Res. Appl.* **21**(5), 338–346 (1996)
51. Fairchild, M.D.: *Color Appearance Models*, 3rd edn. Wiley, Chichester (2013)
52. Fairchild, M.D., Pirrotta, E.: Predicting the lightness of chromatic object colors using CIELAB. *Color. Res. Appl.* **16**(6), 385–393 (1991)
53. Farup, I.: Individualised Halo-Free Gradient-Domain Colour Image Daltonisation. *J. Imag.* **6**(116), 1–10 (2020)
54. Fiorucci, M., Khoroshiltseva, M., Pontil, M., Travaglia, A., Del Bue, A., James, S.: Machine learning for cultural heritage: a survey. *Pattern Recogn. Lett.* **133**(5), 102–108 (2020)
55. Fitts, P.M., (ed.), (1947). *Psychological Research on Equipment Design*, Report no. 19. Army Air Forces, Washington D.C
56. de la Flor, G., Luff, P., Jirotko, M., Pybus, J., Kirkham, R., Carusi, A.: The case of the disappearing ox: seeing through digital images to an analysis of ancient texts. In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2010)*, 10–15 April 2010, Atlanta, GA, USA, pp. 473–482 (2010)
57. Fréchet, G.R., Schotter, A.: *Handbook of Experimental Economic Methodology*. Oxford University Press, Oxford (2015)
58. Fu, S., Lowry, G.D., Yonemura, A.: *From Concept to Context: Approaches to Asian and Islamic Calligraphy*. Smithsonian Institution, Washington, DC (1986)
59. Gao, X.W., Loomes, M.: A new approach to image enhancement for the visually impaired. In: *Proceedings of the IS&T International Symposium on Electronic Imaging 2016, Color Imaging XXI: Displaying, Processing, Hardcopy, and Applications*, 14–18 February 2016, San Francisco, CA, USA, COLOR-325: 1–7 (2016)
60. Gargano, M., Bertani, D., Greco, M., Cupitt, J., Gadia, D., Rizzi, A.: A perceptual approach to the fusion of visible and NIR images in the examination of ancient documents. *J. Cult. Heritage* **16**(4), 518–525 (2015)
61. George, S., Grecicosei, A.M., Waaler, E., Hardeberg, J.Y.: Spectral image analysis and visualisation of the khirbet qeiyafa ostracon. In: *Proceedings of the 6th International Conference on Image and Signal Processing (ICISP 2014)*, 30 June–2 July 2014, Cherbourg, France, 272–279 (2014)
62. Gevers, T., Gijssen, A., van de Weijer, J., Geusebroek, J.-M.: *Color in Computer Vision: Fundamentals and Applications*. Wiley, Chichester (2012)
63. Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Process.* **25**(1), 372–387 (2016)
64. Gibson, A., Piquette, K.E., Bergmann, U., Christens-Barry, W., Davis, G., Endrizzi, M., Fan, S., et al.: An assessment of multimodal imaging of subsurface text in mummy cartonnage using surrogate papyrus phantoms. *Herit. Sci.* **6**, 7 (2018)
65. Gilchrist, A.: *Seeing Black and White*. Oxford Univ. Press, Oxford (2006)
66. Gillespie, A.R., Kahle, A.B., Walker, R.E.: Color enhancement of highly correlated images: I. Decorrelation and HSI contrast stretches. *Remote Sens. Environ.* **20**, 209–235 (1986)
67. Gonzalez, A., Benavente, R., Penacchio, O., Vazquez-Corral, J., Vanrell, M., Parraga, C.A.: *Coloresia: An Interactive Colour*

- Perception Device for the Visually Impaired. In: Sappa, A.D., Vitrià, J. (eds.) *Multimodal Interaction in Image and Video Applications*, pp. 47–66. Springer, Berlin (2013)
68. Govindaraju, V., Srihari, S.N.: Image quality and readability. In: *Proceedings of the International Conference on Image Processing (ICIP)*, 23–26 October 1995, Washington, DC, USA, vol. 3, pp. 324–327 (1995)
  69. Hansen, T., Gegenfurtner, K.R.: Independence of color and luminance edges in natural scenes. *Vis. Neurosci.* **26**, 35–49 (2009)
  70. Hariharan, H., Koschan, A., Abidi, B., Gribok, A., Abidi, M.: Fusion of visible and infrared images using empirical mode decomposition to improve face recognition. In: *Proceedings of the International Conference on Image Processing*, 8–11 October 2006, Atlanta, GA, USA, pp. 2049–2052 (2006)
  71. Harman, J.: Using decorrelation stretch to enhance rock art images. *DStretch: Rock Art Digital Enhancement* (2008). <http://www.dstretch.com/AlgorithmDescription.html>
  72. He, S., Schomaker, L.: DeepOtsu: Document enhancement and binarization using iterative deep learning. *Pattern Recogn.* **91**(7), 379–390 (2019)
  73. Hernández, F.-J.: Retinex for ImageJ (2010). <https://web.archive.org/web/20130323053256/http://www.dentistry.bham.ac.uk/landinig/software/retinex/retinex.html>
  74. Hidalgo, H.L., España, S., Castro, M.H., Pérez, H.A.: Enhancement and cleaning of handwritten data by using neural networks. In: *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, 7–9 June 2005, Estoril, Portugal, part I, pp. 376–383 (2005)
  75. Hofstadter, D.: *Fluid Analogies Research Group: Fluid Concepts and Creative Analogies*. BasicBooks, New York, NY (1995)
  76. Holleman, G.A., Hooge, I.T.C., Kemner, C., Hessels, R.S.: The real-world approach and its problems: a critique of the term ecological validity. *Front. Psychol.* **11**, 721 (2020)
  77. Homer: *The Iliad*, trans. Robert Fagles. New York, NY: Penguin (1990)
  78. Hoßfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia, P.: Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. *IEEE Trans. Multimedia* **16**(2), 541–558 (2014)
  79. Hummel, R.: Image enhancement by histogram transformation. *Comput. Graphics Image Process.* **6**, 184–195 (1977)
  80. James, A.P., Dasarathy, B.V.: Medical image fusion: a survey of the state of the art. *Inform. Fusion* **19**, 4–19 (2014)
  81. Jobson, D.J., Rahman, Z., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.* **6**(7), 965–976 (1997)
  82. Hyoung-Gook, K., Moreau, N., Sikora, T.: *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. Wiley, Chichester (2005)
  83. Kahneman, D.: *Thinking, Fast and Slow*. Penguin, London (2011)
  84. Kambara, A.: Effects of experiencing visual illusions and susceptibility to biases in one’s social judgments. *Percept. Mot. Skills* **7**(4), 1–6 (2017)
  85. Kapoor, A., Caicedo, J.C., Lischinski, D., Kang, S.B.: Collaborative personalization of image enhancement. *Int. J. Comput. Vis.* **108**(1–2), 148–164 (2014)
  86. Karatzas, D., Antonacopoulos, A.: Colour text segmentation in web images based on human perception. *Image Vis. Comput.* **25**, 564–577 (2007)
  87. Keelan, B.W.: *Handbook of Image Quality: Characterization and Prediction*. Marcel Dekker, New York, NY (2002)
  88. Kemeny, J.G., Snell, J.L.: *Mathematical Models in the Social Sciences*. MIT Press, Cambridge, MA (1972)
  89. Kendall, M.G.: *Rank Correlation Methods*, 4th edn. McGraw-Hill, New York, NY (1970)
  90. Kihlstrom, J.F.: Ecological validity and “ecological validity.” *Perspect. Psychol. Sci.* **16**(2), 466–471 (2021)
  91. Kim, H.U., Koh, Y.J., Kim, C.S.: PieNet: Personalized Image Enhancement Network. In: *Proceedings of the Computer Vision (ECCV 2020)*, 23–28 August 2020, Glasgow, UK, pp. 374–390 (2020)
  92. Kim, J., Vora, A., Peli, E.: MPEG-based image enhancement for the visually impaired. *Opt. Eng.* **43**(6), 1318–1328 (2004)
  93. Kim, S.J., Deng, F., Brown, M.S.: Visual enhancement of old documents with hyperspectral imaging. *Pattern Recogn.* **44**, 1461–1469 (2011)
  94. Kim, S.J., Zhuo, S., Deng, F., Fu, C.-W., Brown, M.S.: Interactive visualization of hyperspectral images of historical documents. *IEEE Trans. Visual. Comput. Graph.* **16**(6), 1441–1448 (2010)
  95. Kligler, N., Katz, S., Tal, A.: Document enhancement using visibility detection. In: *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18–23 June 2018, Salt Lake City, UT, USA, pp. 2374–2382 (2018)
  96. Koo, T.K., Li, M.Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropractic Med.* **15**(2), 155–163 (2016)
  97. Koschan, A., Abidi, M.: *Digital Image Processing*. Wiley, Chichester (2008)
  98. Kovési, P.: Good colour maps. How to design them (2015). [arXiv:1509.03700](https://arxiv.org/abs/1509.03700)
  99. Kowalik, G., Nielek, R.: Senior programmers: characteristics of elderly users from stack overflow. In: *Proceedings of the 8th International Conference on Social Informatics (SocInfo 2016)*, 11–14 November 2016, Bellevue, WA, USA, part II, pp. 87–96 (2016)
  100. Krupinski, E.A.: Practical applications of perceptual research. In: Beutel, J., Van Metter, R.L., Kundel, H.L. (eds.) *Handbook of Medical Imaging*, Bellingham, vol. 1, pp. 895–929. SPIE, WA (2009)
  101. Kruse, P., Stadler, M. (eds.): *Ambiguity in Mind and Nature: Multistable Cognitive Phenomena*. Springer, Berlin (1995)
  102. Kuhlthau, C.C.: *Seeking Meaning: A Process Approach to Library and Information Services*. Libraries Unlimited, Westport, CT (2004)
  103. Kuncheva, L.I.: *Combining pattern classifiers: methods and algorithms*, 2nd edn. Wiley, New Jersey, NJ (2014)
  104. Kundel, H.L.: Visual search in medical images. In: Beutel, J., Van Metter, R.L., Kundel, H.L. (eds.) *Handbook of Medical Imaging*, vol. 1, pp. 837–858. SPIE, Bellingham, WA (2000)
  105. Labaune, J., Jackson, J.B., Pagès-Camagna, S., Menu, M., Mourou, G.A.: Terahertz investigation of Egyptian artifacts. In: *Proceedings of the 35th International Conference on Infrared, Millimeter, and Terahertz Waves*, 5–10 September 2010, Rome, Italy, pp. 1–3 (2010)
  106. Lakoff, G., Johnson, M.: *Metaphors We Live By*. Univ. of Chicago Press, Chicago, IL (1980)
  107. Land, E.H.: The retinex theory of color vision. *Sci. Am.* **237**(6), 108–128 (1977)
  108. Langville, A.N., Meyer, C.D.: *Who’s #1? The Science of Rating and Ranking*. Princeton Univ. Press, Princeton, NJ (2012)
  109. Legge, G.E., Rubin, G.S., Schleske, M.M.: Contrast-polarity effects in low-vision reading. In: Woo, G.C. (ed.) *Low Vision: Principles and Applications*, pp. 288–307. Springer, New York, NY (1987)
  110. Lettner, M., Sablatnig, R.: Spatial and spectral based segmentation of text in multispectral images of ancient documents. In: *Proceedings of the 10th International Conference on Document Analysis and Recognition*, 26–19 July 2009, Barcelona, Spain, pp. 813–817 (2009)

111. Li, X.: Image denoising: past, present, and future. In: Gunturk, B.K., Li, X. (eds.) *Image restoration: fundamentals and advances*, pp. 1–19. CRC Press, Boca Raton, FL (2013)
112. Lin, Y., Seales, W.B.: Opaque document imaging: building images of inaccessible texts. In: *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV' 05)*, 17–21 October 2005, Beijing, China, vol. 1, pp. 662–669
113. Lins, R.D., Almeida, M.M., Bernardino, R.B., Jesus, D., Oliveira, J.M.: Assessing binarization techniques for document images. In: *Proceedings of the ACM Document Engineering Conference (DocEng' 17)*, 4–7 September 2017, Valletta, Malta, pp. 183–192 (2017)
114. Lund, O.: Knowledge construction in typography: the case of legibility research and the legibility of sans serif typefaces. PhD thesis, Univ. of Reading, Reading (1999)
115. Lombardi, F., Marinai, S.: Deep learning for historical document analysis and recognition: a survey. *J. Imag.* **6**(10), 110 (2020)
116. Machado, G.M., Oliveira, M.M., Fernandes, L.A.F.: A physiologically-based model for simulation of color vision deficiency. *IEEE Trans. Visual. Comput. Graph.* **15**(6), 1291–1298 (2009)
117. Marinai, S., Karatzas, D.: Report from the AND 2009 working group on noisy text datasets. *Intl. J. Doc. Anal. Recogn.* **14**(2), 113–116 (2011)
118. MathWorks: Contrast adjustment. Matlab: Image Processing Toolbox (2020). <https://www.mathworks.com/help/images/contrastadjustment.html>
119. McCann, J.J.: Retinex at 50: color theory and spatial algorithms: a review. *J. Elect. Imag.* **26**(3), 031204 (2017)
120. McIntyre, D.: *Colour blindness: Causes and Effects*. Dalton Publishing, Chester (2002)
121. Mehta, R., Zhu, R.: Blue or red? Exploring the effect of color on cognitive task performances. *Science* **323**(5918), 1226–1229 (2009)
122. Miller, E., Lee, N.J., Uprichard, K., Daniels, V.: The examination and conservation of the Rosetta stone at the British Museum. In: Roy, A., Smith, P. (eds.) *Tradition and Innovation: Advances in Conservation*, pp. 128–132. The Intl. Institute for Conservation of Historic and Artistic Works, London (2000)
123. Mindermann, S.: Hyperspectral imaging for readability enhancement of historic manuscripts. M.A. thesis, Technical Univ. München, München (2018)
124. Mitchell, G.: Revisiting truth or triviality: the external validity of research in the psychological laboratory. *Perspect. Psychol. Sci.* **7**(2), 109–117 (2012)
125. Mocella, V., Brun, E., Ferrero, C., Delattre, D.: Revealing letters in rolled Herculaneum papyri by X-ray phase-contrast imaging. *Nat. Commun.* **6**, 5895 (2015)
126. Moriyama, D., Azetsu, T., Ueda, Ch., Suetake, N., Uchino, E.: Image enhancement with lightness correction and image sharpening based on characteristics of vision for elderly persons. *Opt. Rev.* **27**(4), 352–360 (2020)
127. Morović, J.: *Color Gamut Mapping*. Wiley, Chichester (2008)
128. Mullen, K.T.: The contrast sensitivity of human colour vision red-green and blue-yellow chromatic gratings. *J. Physiol.* **359**, 381–400 (1985)
129. Mussell, J.: *The Nineteenth-Century Press in the Digital Age*. Palgrave Macmillan, London (2012)
130. Neji, H., Noguera-Iso, J., Lacasta, J., Ben Halima, M., Alimi, A.: Adversarial autoencoders for denoising digitized historical documents: the use case of incunabula. In: *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR 2019)*, 20–25 September 2019, Sydney, Australia, pp. 31–34 (2019)
131. Nguyen, C.K., Nguyen, C.T., Hotta, S., Nakagawa, M.: A character attention generative adversarial network for degraded historical document restoration. In: *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR 2019)*, 20–25 September 2019, Sydney, Australia, pp. 420–425 (2019)
132. Oppenheimer, D.M.: Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Appl. Cogn. Psychol.* **20**, 139–156 (2006)
133. Pan, X.-B., Brady, M., Bowman, A.K., Crowther, C., Tomlin, R.S.O.: Enhancement and feature extraction for images of incised and ink texts. *Image Vis. Comput.* **22**, 443–451 (2004)
134. Paris, S., Hasinoff, S.W., Kautz, J.: Local laplacian filters: edge-aware image processing with a laplacian pyramid. *Commun. ACM* **58**(3), 81–91 (2015)
135. Parry, N.R.A.: Color vision deficiencies. In: Elliot, A.J., Fairchild, M.D., Franklin, A. (eds.) *Handbook of Color Psychology*, pp. 216–242. Cambridge Univ. Press, Cambridge (2015)
136. Peli, E.: Limitations of image enhancement for the visually impaired. *Optom. Vis. Sci.* **69**(1), 15–24 (1992)
137. Peli, E., Lee, E., Trempe, C.L., Buzney, S.: Image enhancement for the visually impaired: the effects of enhancement on face recognition. *J. Opt. Soc. Am. A* **11**(7), 1929–1939 (1994)
138. Peli, E., Woods, R.L.: Image enhancement for impaired vision: the challenge of evaluation. *Int. J. Artif. Intell. Tools* **18**(3), 415–438 (2009)
139. Perry, S.: Image and video noise: an industry perspective. In: Bertalmío, M. (ed.) *Denoising of photographic images and video: fundamentals, open challenges and new trends*, pp. 217–230. Springer, Cham (2018)
140. Petro, A.B., Sbert, C., Morel, J.-M.: Multiscale Retinex. *Image Process. On Line* **4**, 71–88 (2014)
141. Pham, T.-A., Delalandre, M.: Post-processing coding artefacts for JPEG documents. *Int. J. Doc. Anal. Recogn.* **20**(3), 189–200 (2017)
142. Piquette, K.E.: Revealing the material world of ancient writing: digital techniques and theoretical considerations. In: Hoogendijk, F.A.J., van Gompel, S.M.T. (eds.) *The Materiality of Texts from Ancient Egypt: New Approaches to the Study of Textual Material from the Early Pharaonic to the Late Antique Period*, pp. 94–118. Brill, Leiden (2018)
143. Plateau, J.: *Mémoire sur l'irradiation*. Hayez, Bruxelles (1839)
144. Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., Kuo, C.-C.J.: Image database TID2013: Peculiarities, results and perspectives. *Image Commun.* **30**(1), 57–77 (2015)
145. Pons, C., Mazade, R., Jin, J., Dul, M.W., Zaidi, Q., Alonso, J.M.: Neuronal mechanisms underlying differences in spatial resolution between darks and lights in human vision. *J. Vis.* **17**(14), 5–24 (2017)
146. Popowicz, A., Smolka, B.: Overview of Grayscale Image Colorization Techniques. In: Celebi, M.E., Lecca, M., Smolka, B. (eds.) *Color Image and Video Enhancement*, pp. 345–370. Springer, Cham (2015)
147. Da Pos, O., Zambianchi, E.: *Visual Illusions and Effects: A Collection*. Guerini, Milano (1996)
148. Poston, A.M.: A literature review of cockpit lighting. Technical memorandum 10-64, US Army Human Engineering Laboratory, Aberdeen Proving Ground, MD (1974)
149. Pratikakis, I., Zagoris, K., Karagiannis, X., Tsochatzidis, L., Mondal, T., Marthot-Santaniello, I.: ICDAR 2019 Competition on Document Image Binarization (DIBCO 2019). In: *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*, 20–25 September 2019, Sydney, Australia, pp. 1547–1556 (2019)
150. Rasche, K., Geist, R., Westall, J.: Detail preserving reproduction of color images for monochromats and dichromats. *IEEE Comput. Graph. Appl.* **26**(3), 22–30 (2005)

151. Raj, V., Arunkumar, C.: Content restoration of degraded termite bitten document images. *Int. J. Eng. Res. Appl.* **4**(5), 151–155 (2014)
152. Ribeiro, F., Florencio, D., Nascimento, V.: Crowdsourcing subjective image quality evaluation. In: *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP 2011)*, 11–14 September 2011, Brussels, Belgium, pp. 3097–3100 (2011)
153. Rodney, A.: *Color Management for Photographers*. Focal Point, Oxford (2005)
154. Rogowitz, B.E., Treinish, L.A.: Data visualization: the end of the rainbow. *IEEE Spectrum* **12**, 52–59 (1998)
155. Roued-Cunliffe, H.: A decision support system for the reading of ancient documents. Ph.D. thesis, Faculty of Classics, Univ. of Oxford, Oxford (2011)
156. Rudd, M.E.: Lightness computation by the human visual system. *J. Elect. Imag.* **26**(3), 031209 (2017)
157. Rudd, M.E., Popa, D.: Stevens's brightness law, contrast gain control, and edge integration in achromatic color perception: a unified model. *J. Opt. Soc. Am.* **24**(9), 2766–2782 (2007)
158. Samei, E., Krupinski, E.A. (eds.): *The Handbook of Medical Image Perception and Techniques*, 2nd edn. Cambridge Univ. Press, Cambridge (2019)
159. Sassoon, R.: *Handwriting of the Twentieth Century*. Intellect, Bristol (2007)
160. Schivre, G.: *Multiscale Retinex*. MATLAB Central File Exchange (2020). <https://www.mathworks.com/matlabcentral/fileexchange/71386-multiscale-retinex>
161. Schubert, P.: Editing a Papyrus. In: Bagnall, R.S. (ed.) *The Oxford Handbook of Papyrology*, pp. 197–215. Oxford Univ. Press, Oxford (2009)
162. Schwartz-Shea, P., Yanow, D.: *Interpretive research design: concepts and processes*. Routledge, New York, NY (2012)
163. Seuret, M., Chen, K., Eichenberger, N., Liwicki, M., Ingold, R.: Gradient-domain degradations for improving historical documents images layout analysis. In: *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR)*, 23–26 August 2015, Nancy, France, pp. 1006–1010 (2015)
164. Shapiro, A., Todorović, D. (eds.): *The Oxford Compendium of Visual Illusions*. Oxford Univ. Press, Oxford (2017)
165. Shapiro, A.G., Hedjar, L.: Color illusion as a spatial binding problem. *Curr. Opin. Behav. Sci.* **30**, 149–155 (2019)
166. Sharma, A.: *Understanding color management*, 2nd edn. Wiley, Hoboken, NJ (2018)
167. Sharma, G., Rodríguez-Pardo, C.E.: The dark side of CIELAB. In: *Proceedings of the SPIE-IS&T Electronic Imaging, Color Imaging XVII: Displaying, Processing, Hardcopy, and Applications*, vol. 8292: 82920D (2012)
168. Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: *Proceedings of the 1996 IEEE Symposium on Visual Languages*, 3–6 September 1996, Boulder, CO, USA, pp. 336–343 (1996)
169. Shortland, N.D., Alison, L.J., Moran, J.M. (2019). *Conflict: How Soldiers Make Impossible Decisions*. Oxford Univ. Press, Oxford (2019)
170. Smith, A.R.: Color gamut transform pairs. In: *SIGGRAPH '78: Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques*, 23–25 August 1978, Atlanta, GA, USA, pp. 12–19 (1978)
171. Sider, D.: *The Library of the Villa Dei Papii at Herculaneum*. Getty Publications, Los Angeles, CA (2005)
172. Simons, D.J., Chabris, C.F.: Gorillas in our midst: Sustained inattention blindness for dynamic events. *Perception* **28**(9), 1059–1074 (1999)
173. Sobottka, K., Kronenberg, H., Perroud, T., Bunke, H.: Text extraction from colored book and journal covers. *Int. J. Doc. Anal. Recogn.* **2**(4), 163–176 (2000)
174. Sonnad, S.: A survey on fusion of multispectral and panchromatic images for high spatial and spectral information. In: *Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 23–25 March 2016, Chennai, India, pp. 177–180 (2016)
175. Sparavigna, A.C.: Digital restoration of ancient papyri (2009). [arXiv:0903.5045](https://arxiv.org/abs/0903.5045)
176. Tarte, S.: Papyrological investigations: transferring perception and interpretation into the digital world. *Liter. Ling. Comput.* **26**(2), 233–247 (2011)
177. Terras, M.M.: *Image to interpretation: an intelligent system to aid historians in reading the Vindolanda texts*. Oxford Univ. Press, Oxford (2006)
178. Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., Lin, C.-W.: Deep learning on image denoising: an overview. *Neural Netw.* **131**(11), 251–275 (2020)
179. Tonazzini, A.: Color space transformations for analysis and enhancement of ancient degraded manuscripts. *Pattern Recognit. Image Anal.* **20**(3), 404–417 (2010)
180. Tonazzini, A., Gerace, I., Martinelli, F.: Document image restoration and analysis as separation of mixtures of patterns: from linear to nonlinear models. In: Gunturk, B.K., Li, X. (eds.) *Image Restoration: Fundamentals and Advances*, pp. 285–310. CRC Press, Boca Raton, FL (2012)
181. Tournié, A., Fleischer, K., Bukreeva, I., Palermo, F., Perino, M., Cedola, A., Andraud, C., Ranocchia, G.: Ancient Greek text concealed on the back of unrolled papyrus revealed through short-wave-infrared hyperspectral imaging. *Sci. Adv.* **5**(10), eaav8936 (2019)
182. Van Minnen, P.: The future of papyrology. In: Bagnall, R.S. (ed.) *The Oxford handbook of papyrology*, pp. 644–660. Oxford Univ. Press, Oxford (2009)
183. VanderPlas, S., Hofmann, H.: Signs of the sine illusion: why we need to care. *J. Comput. Graph. Stat.* **24**(4), 1170–1190 (2015)
184. Velázquez, A., Levachkine, S.: Text/graphics separation and recognition in raster-scanned color cartographic maps. In: *Proceedings of the International Workshop on Graphics Recognition, GREC 2003: Graphics Recognition. Recent Advances and Perspectives*, Lecture Notes in Computer Science, vol 3088, pp. 63–74 (2004)
185. Ventzas, D., Ntogas, N., Ventza, M.M.: Digital restoration by denoising and binarization of historical manuscripts images. In: Ventzas, D. (ed.) *Advanced Image Acquisition, Processing Techniques and Applications*, pp. 73–108. Rijeka, InTech (2012)
186. Viénot, F., Brettel, H., Mollon, J.D.: Digital video colourmaps for checking the legibility of displays by dichromats. *Color. Res. Appl.* **24**(4), 243–252 (1999)
187. Vilaseca, M., Pujol, J., Arjona, M., Martínez-Verdú, F.M.: Color visualization system for near-infrared multispectral images. *J. Imag. Sci. Technol.* **49**(3), 246–255 (2005)
188. Wang, Z., Bovik, A.C.: *Modern Image Quality Assessment*. Morgan & Claypool, San Rafael, CA (2006)
189. Wang, Z., Simoncelli, E.P.: Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In: *Proceedings of the SPIE-IS&T Electronic Imaging Conference, Human Vision and Electronic Imaging X*, 18 March 2005, San Jose, CA, USA, pp. 149–159 (2005)
190. Webster, M.A.: Individual differences in color vision. In: Elliot, A.J., Fairchild, M.D., Franklin, A. (eds.) *Handbook of Color Psychology*, pp. 197–215. Cambridge Univ. Press, Cambridge (2015)
191. Westheimer, G.: Irradiation, border location and the shifted-chessboard pattern. *Perception* **36**(4), 483–494 (2007)

192. Wu, H.R., Rao, K.R. (eds.): *Digital Video Image Quality and Perceptual Coding*. CRC Press, Boca Raton, FL (2006)
193. Yemelyanov, K.M., Lin, Sh.Sch., Luis, W.Q., Pugh, E.N. Jr, Eng-heta, N.: Bio-inspired display of polarization information using selected visual cues. In: *Proceedings of the SPIE Optical Science and Technology*, 5158, Polarization Science and Remote Sensing, 3–8 August 2003, San Diego, California, USA, pp. 71–84 (2003)
194. Yu, F.T.S., Tai, A., Chen, H.: Spatial filtered pseudocolor holographic imaging. *J. Opt.* **9**(5), 269–273 (1978)
195. Zheng, Y.: An overview of night vision colorization techniques using multispectral images: From color fusion to color mapping. In: *International Conference Audio Language Image Processing*, 16–18 July 2012, Shanghai, China, pp. 134–143 (2012)
196. Zhu, Zh., Toyoura, M., Go, K., Kashiwagi, K., Fujishiro, I., Wong, T.T., Mao, X.: Personalized Image Recoloring for Color Vision Deficiency Compensation. *IEEE Transactions on Multimedia*, in press (2021)
197. Ziyadee, T.: Unsupervised denoising via Wiener entropy masking in the STFT domain. In: *Proceedings of the 2014 IEEE Military Communications Conference*, 6–8 October 2014, Baltimore, MD, USA, pp. 467–472 (2014)
198. Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: *Graphic Gems IV*, pp. 474–485. Academic Press Professional, San Diego, CA (1994)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



## Supplementary Material

Presented herein are the ninety images used in the evaluation of papyri legibility enhancement experiment. The methods are described in the article, Sections 4 Methods and 5.1 Algorithms.

TSVP →



original



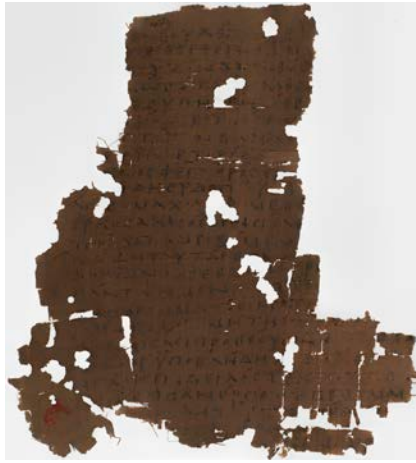
stretchlim



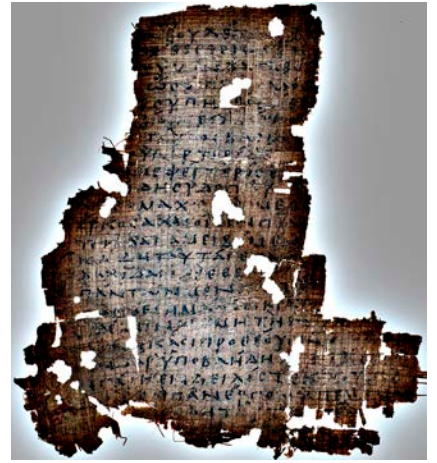
histeq



adapthisteq



locallapfilt



retinex



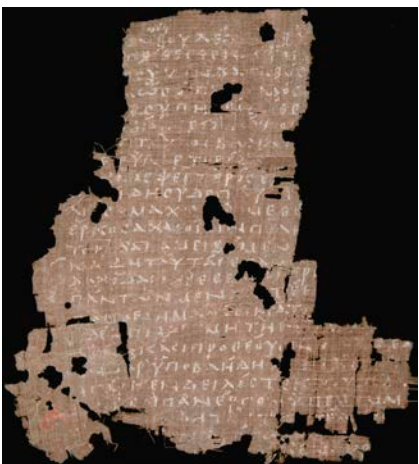
lsv



vividness



neglsv



negvividness

SM2

**columbia.apis.p367.f.0.600**

Credit of original papyrus reproduction:  
Columbia University Library, CC BY-  
NC 3.0



original



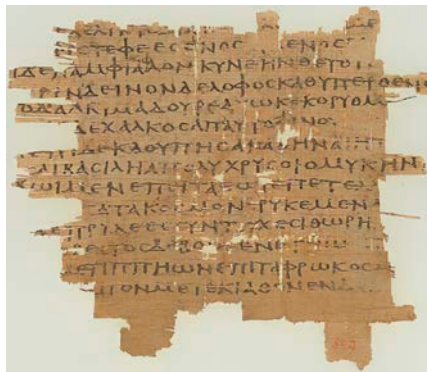
stretchlim



histeq



adapthisteq



locallapfilt



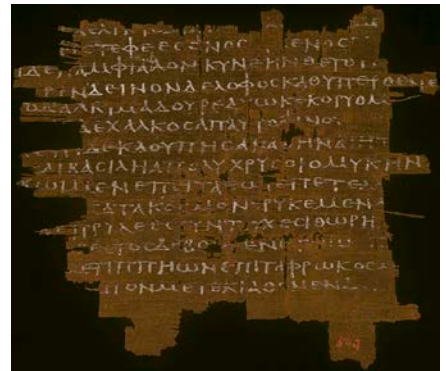
retinex



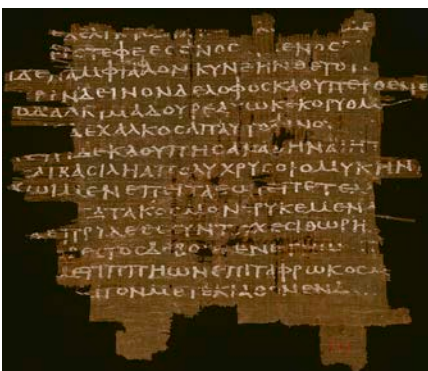
lsv



vividness



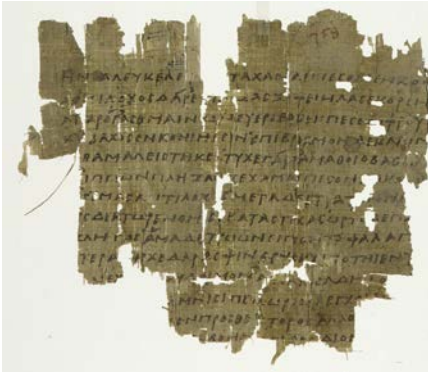
neglsv



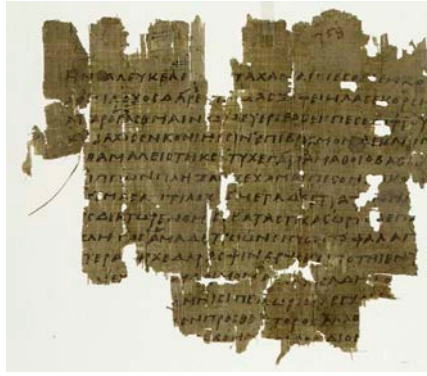
negvividness

P.Corn. Inv. MSS.A 101. XIII

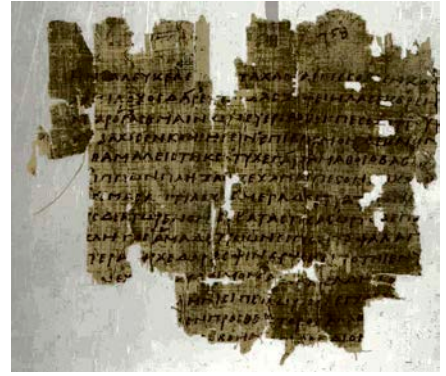
Credit of original papyrus reproduction:  
University of Michigan Library, Public  
Domain Mark 1.0



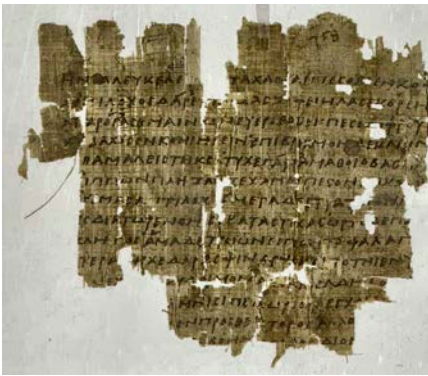
original



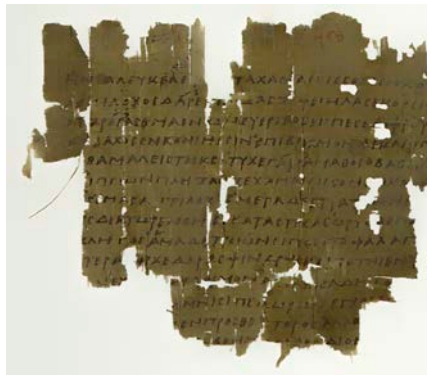
stretchlim



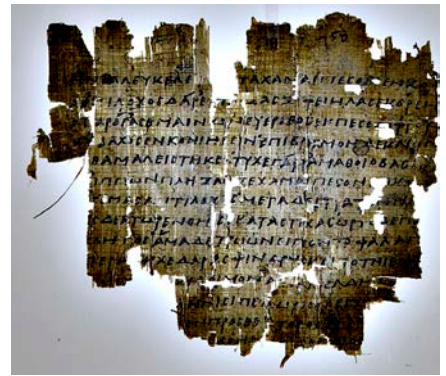
histeq



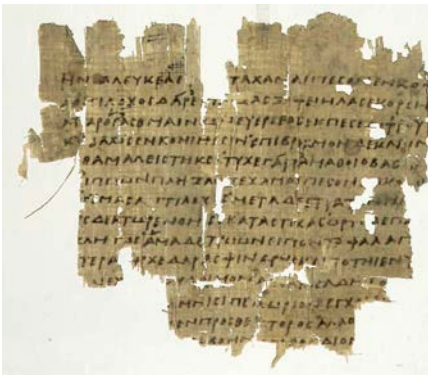
adapthisteq



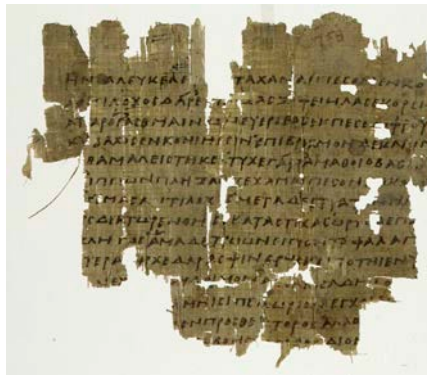
locallapfilt



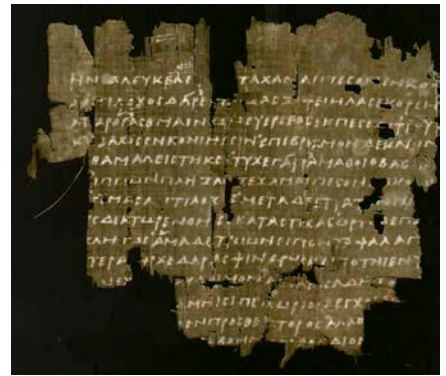
retinex



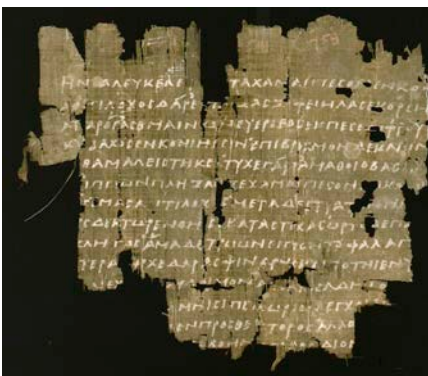
lsv



vividness



neglsv



negvividness



original



stretchlim



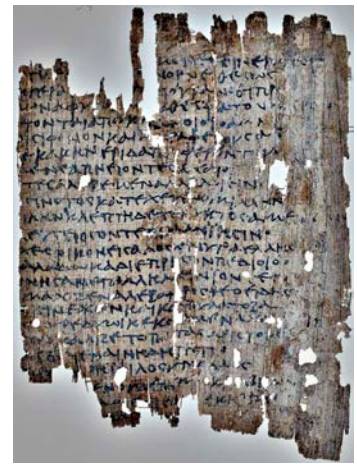
histeq



adapthisteq



locallapfilt



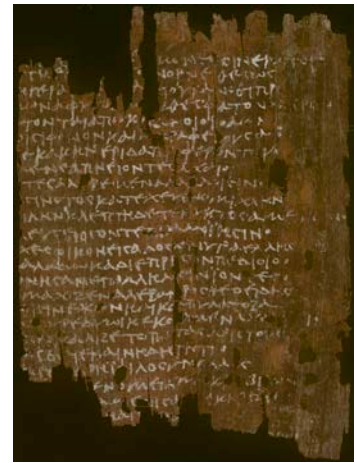
retinex



lsv



vividness



neglsv



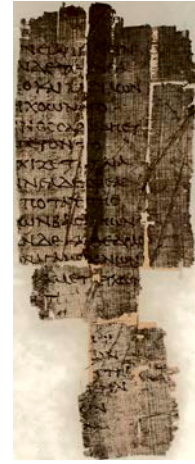
negvividness



original



stretchlim



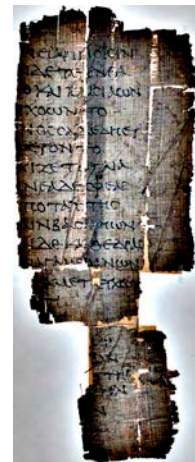
histeq



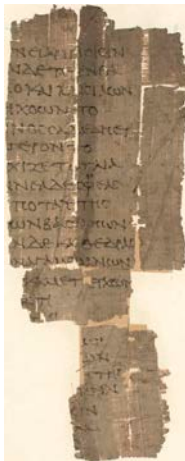
adapthisteq



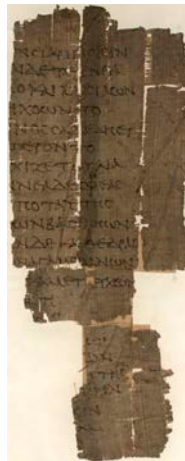
locallapfilt



retinex



lsv



vividness



neglsv



negvividness

**P.Mich.inv.2755**

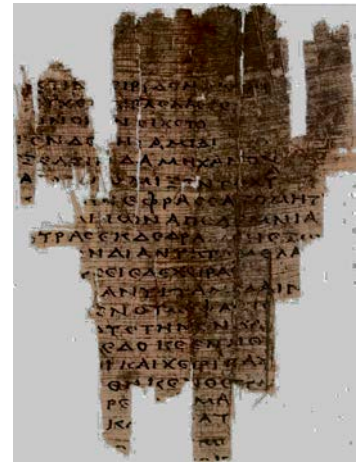
Credit of original papyrus reproduction:  
University of Michigan Library, Public  
Domain Mark 1.0



original



stretchlim



histeq



adapthisteq



locallapflt



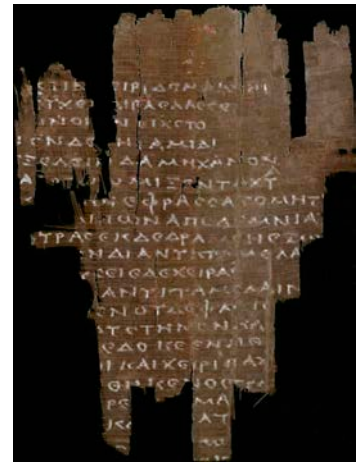
retinex



lsv



vividness



neglsv



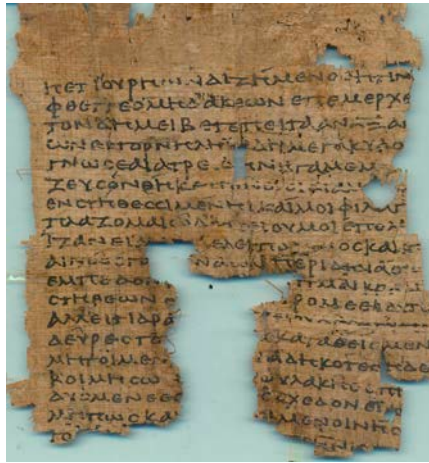
negvividness

### P.Oxy.XXII 2309

Credit of original papyrus reproduction:  
Courtesy of The Egypt Exploration Society and the University of Oxford Imaging Papyri Project



original



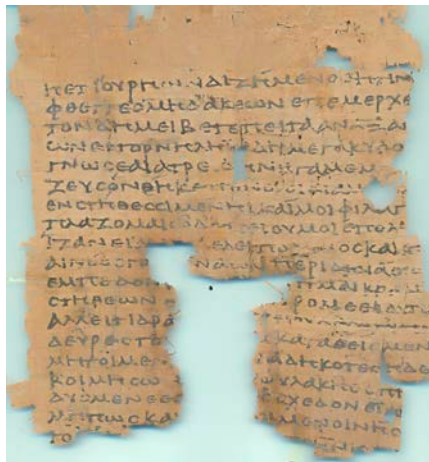
stretchlim



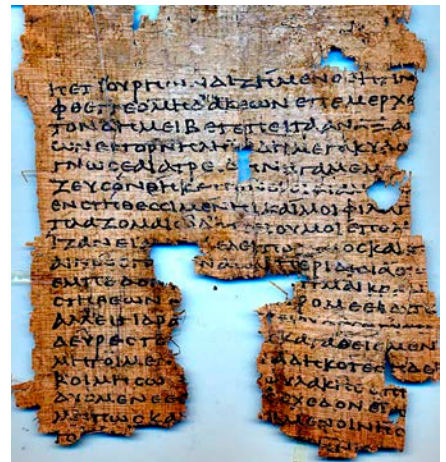
histeq



adapthisteq



locallapfilt



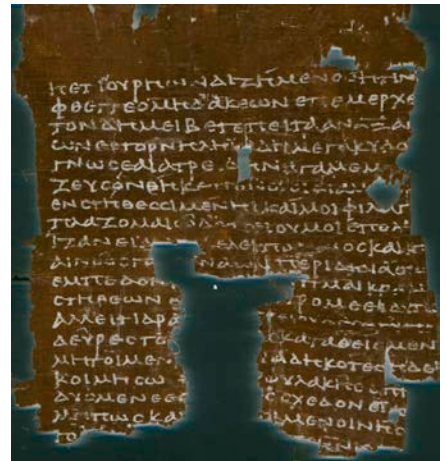
retinex



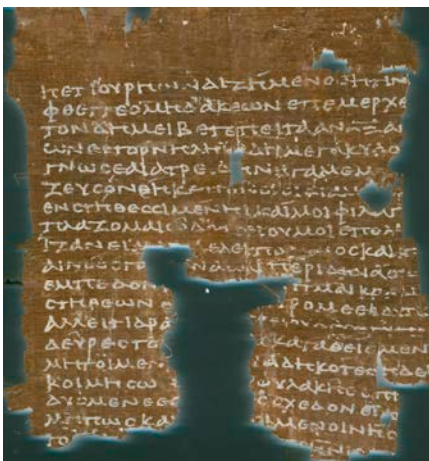
lsv



vividness



neglsv

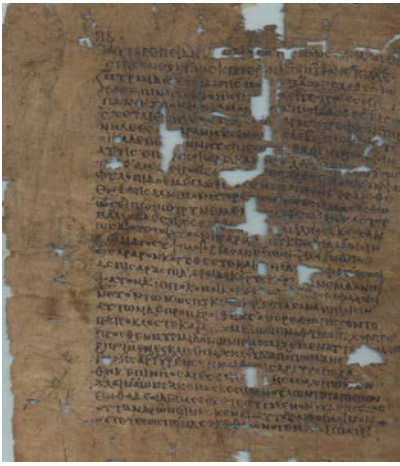


negvividness

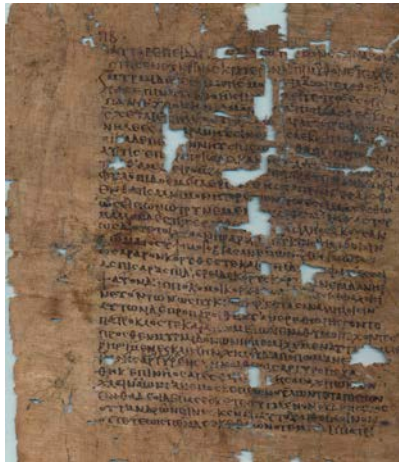
### PSI XII 1274 r

Credit of original papyrus reproduction: Istituto Papirologico Vitelli, by permission

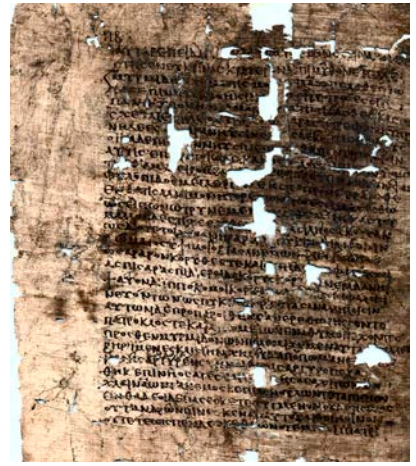




original



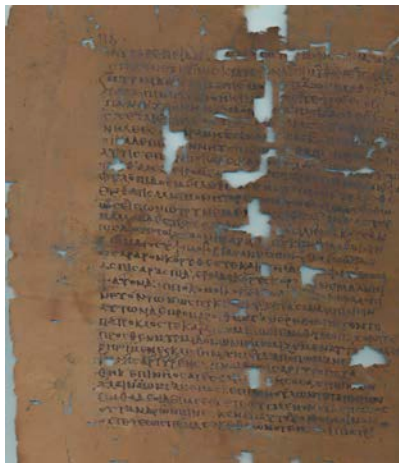
stretchlim



histeq



adapthisteq



locallapfilt



retinex



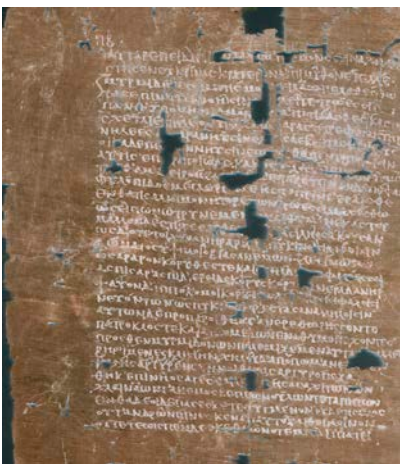
lsv



vividness



neglsv



negvividness

### PSI XIII 1298 (15a) r 1

Credit of original papyrus reproduction:  
Istituto Papirologico Vitelli, by  
permission



original



stretchlim



histeq



adapthisteq



locallapfilt



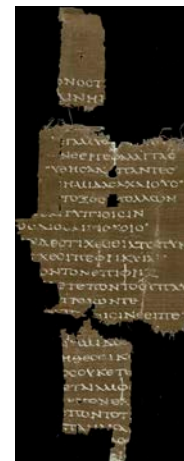
retinex



lsv



vividness



neglsv



negvividness

SM10

### PSIX 1376 r

Credit of original papyrus reproduction: Istituto Papirologico Vitelli, by permission