Applicability Of White-Balancing Algorithms to Restoring Faded Colour Slides: An Empirical Evaluation

Denis Nikitenko, Michael Wirth
Department of Computing and Information Science, University of Guelph, Guelph, Ontario, Canada
Kataline Trudel
Psychology Department, University of Guelph, Guelph, Ontario, Canada
Email: {dnikiten, mwirth, ktrudel}@uoguelph.ca

***Abstract*— In this paper we investigated the applicability of commonly used white-balancing algorithms to restoring faded photographic colour slides. We have used three sets of synthetic data that simulated colour damage in Kodak Ektachrome slides, as well as three sets of real digitized faded Kodak Ektachrome slides. We have restored all the data sets using nine different algorithms and evaluated restoration results using human participants. In addition, we have conducted an evaluation of the restored synthetic data using a distance metric . We found that while some algorithms provided acceptable restoration of synthetic images, none were found by the human participants to adequately restore real-world data. In addition, we found no correlation between human-based and metric-based evaluation results on 2 out of 3 synthetic data sets. Our results form a strong indication that commonly used white-balancing methods are inadequate for restoring faded colour slides and that simple colour distance metrics do not necessarily correspond to human perception of colour quality.**

***Index Terms*— colour image processing, colour restoration, colour metrics, empirical performance evaluation**

I. Introduction

A. Colour fading in photographs and slides

The digital restoration of historical images, particularly photographic materials, has received considerable attention in recent years. Technological advances in digital imaging have allowed researchers to begin investigating restoration techniques to deal with damage and artifacts unique to different types of photographic prints and films. One of the most common types of damage is the fading of photographic dyes in colour photographs and slides.

Colour fading affects printed photographs, slides, photographic films, and movie films. Many colour dyes used in photography and film-making are quite chemically unstable and can be affected by factors such as light, humidity, temperature, and chemical agents. In addition, other chemicals present in the films (such as colour couplers, stabilizers, etc.) may change with time [1]. As a result, affected dyes are bleached. Bleaching appears to be a linear process in most cases, but depending on the type of film and environmental factors, colour channels may fade differently [1]. The life span of photographic materials ranges from several years to over a century. The best way to preserve them is to store them at low temperature in a no-light area with low humidity. Restoration of faded colour materials by chemicals is impossible because the bleaching of dyes is an irreversible process [1].

There are numerous commercial companies that offer photograph restoration services. They digitize old prints and manually restore them using off-the-shelf commercial image manipulation software. These companies are often capable of producing high quality results. However, when it is necessary to restore large numbers of photographs, for instance when digitizing photographic archives of a library or an art collection, manual methods are unsuitable due to the time and expense involved. There is therefore a need for automated colour restoration methods.

B. Digital colour restoration techniques for photography

One of the earliest works on digital restoration of photographic materials was published by Gschwind in 1989 [2]. This method relies on a model established by artificial bleaching of various types of film in laboratory conditions. Gschwind suggested that the simplest model (in CMY colour space) is [2]:

$$\begin{matrix} Y' \\ M' \\ C' \end{matrix} = \begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix} \cdot \begin{pmatrix} Y \\ M \\ C \end{pmatrix} \quad (1)$$

where Y', M', C' are the optical densities (ODs) of the faded dyes and Y, M, C are the original ODs. Acknowledging the limitations of such a model, he proposed the following "bleaching equation" in order to account for various types of bleaching and staining that occur in photographic materials [2]:

$$\begin{matrix} Y' \\ M' \\ C' \end{matrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix} \cdot \begin{pmatrix} Y \\ M \\ C \\ 1.0 \end{pmatrix} \quad (2)$$

This paper is based on "White-balancing algorithms in colour photograph restoration," by D. Nikitenko, M. Wirth, and K. Trudel, which appeared in the Proceedings IEEE International Conference on Systems, Man and Cybernetics (SMC 2007), Montreal, Canada, 2007.

This work was sustained by a pair of articles published in 1994 by Frey and Gschwind [3], [4]. The former [3] provided several empirically derived mathematical bleaching models for different types of film and discussed film stability. The latter paper [4] gave a more detailed outline of the digital restoration process. The authors point out that when the film type and age are known, the appropriate bleaching model can be used and the restoration is relatively straightforward. When neither information is available, as is often the case, they suggest a user-guided restoration process similar to [2], where the operator suggests the colour adjustments and the system automatically generates the appropriate matrix. The authors also discuss the effects of side absorption by dyes (e.g. absorption of red by the magenta dye) and stress the importance of accurate spectral information on the faded dyes. The methods were later extended in [5] and [1] to work with colour movies.

A similar approach was suggested by Chambah and Besserer [6]. It relies on a linear bleaching model and manual reference colour selection. They subsequently proposed a different method, which consists of colour saturation enhancement and colour balancing using a manually restored reference image [7]. Saturation enhancement is achieved in CIE L*a*b space by stretching the converted 3D point set along the principal axes of the space [7]. The colour balance enhancement method relied on modified versions of retinex white patch (WP) and grey world (GW) algorithms that were originally designed to remove the colour cast caused by an illumination shift [7]. They also proposed a semi-automatic approach that uses principal component analysis (PCA) to represent the data and non-linear colour correlation and colour rotation for hue adjustment [8]. All of these methods are summarized in [9], where three bleaching models, four colour balancing methods, and three histogram manipulation methods are used in order to adjust the "reference colours" for the modified WP+GW scheme [9].

A completely automated approach was proposed by Chambah et al. [10]. It incorporatea the Automatic Colour Equalization method (ACE), which merged the WP and GW algorithms. This method was revised by Rizzi et al. [11] and was further expanded by Rizzi et al. [12]. The latest publication also provided a mechanism for incorporating balancing multiple colour frames with different fading characteristics.

The methods described above represent a significant progress in digital restoration of faded colour dyes. However, all have potential shortcomings. The methods of Gschwind et al. [1]–[5] require knowledge of the film type and age for automated restoration and call for heavy user involvement when such information is not available. The latest approach of Rizzi et al. [10]–[12] is completely automated, but some of the images presented in their results still show a considerable colour cast after processing; furthermore, no failure analysis of the algorithm is reported. Finally, none of the above papers provided details on performance evaluation of the algorithms described therein.

C. Proposed study

The need for automated colour adjustment is not limited to restoration of photographic materials. A somewhat similar situation occurs when taking digital photographs. The human visual system maintains colour invariance under different illumination conditions. Photographic sensors cannot do this and therefore algorithms have had to be developed to compensate for colour changes due to changes in illumination (e.g. bright sunlight vs. tungsten light). They are referred to as "white-balancing" algorithms.

This paper aims to assess the applicability of some of these algorithms to restoring colour in faded photographs and slides. The algorithms were not expected to perform perfectly - changes in image colour due to different illumination conditions is adequately modelled by (1), whereas photographic dye bleaching is more accurately modelled by (2) - but we felt it would be worthwhile to examine the performance of common current "off-the-shelf" colour restoration algorithms.

Performance of colour restoration methods could be evaluated in two ways. One is evaluation by a human observer, or a group of observers (referred to in this work as *subjective* evaluation). The other way would be to compute a numerical metric describing the properties of the image (or *objective* evaluation). Objective evaluation would provide more consistent and replicable results when the goal is to compare relative algorithm performance. Objective evaluation of image processing algorithms is always simplified if "ground truth" data is available. The processed results could then be compared to the ground truth. The algorithms that produced results closest to the ground truth would be considered the best. Several colour distance metrics have been proposed for ground truth-based evaluation of colour image processing algorithms.

However, measuring the absolute performance of colour restoration algorithms is considerably more difficult. Photographs are restored for human use. therefore we would like to know if the images a restoration algorithm produces are aesthetically pleasing (colours are completely natural, no presence of cast), mediocre (still a noticeable cast), etc.. Needless to say, "aesthetically pleasing" and "mediocre" would be rather subjective human judgements of overall colour quality in an image. Numerical metrics that have been shown to correspond well to human perception of colour quality would be invaluable for this purpose. Unfortunately no such metrics are known to the authors, though some ideas for metrics that might indicate the presence of a colour cast have been suggested (for instance, by Gasparini and Schettini [13]).

We have conducted performance evaluation using both objective and subjective evaluation methods. Our aim was to achieve two goals. The first one was to evaluate the performance of white-balancing algorithms by employing human subjects to measure the absolute colour quality

TABLE I.
TESTED ALGORITHMS

Algorithm	Notes
GW	Grey world algorithm
MaxW	Max white algorithm
Ret_unpost	Retinex as implemented by Funt et al. [14] with no post-processing
Ret_post	Retinex as implemented by Funt et al. [14] with post-processing
Stretch	The post-processing stretching algorithm
QM	The combined grey-world/"retinex" approach of Lam [15]. For the reasons given in section II-E, we refer to it as the "quadratic mapping" in this paper
ACASDL	Adjacent Channels Adjustment by Standard Deviation and Luminance algorithm by Lam et al. [16]
SDWGW	Standard deviation-weighted grey world algorithm by Lam et al. [17]
SDWLGW	Standard deviation and luminance-weighted grey world algorithm by Lam et al. [18]

of the restored images. The second goal was to see how well a colour distance metric corresponded to the human judgement.

II. ALGORITHMS

The choice of the algorithms reviewed in this paper was governed by several factors. The number of algorithms was limited by the scope of this paper. The algorithms were chosen because they were well-known and commonly used (or extended common approaches), were all recent, and were generic enough to be potentially applicable to fading-induced colour cast removal. Algorithms such as those relying on specific illuminant properties were excluded for their lack of generality. The algorithms used in this study are listed in Table I together with their abbreviated names.

A. Grey world

One of the most commonly known white-balancing algorithms is the grey world approach, which has its roots in film photography [19]. It assumes that the average intensities of the red, green, and blue channels should be equal. If $Im(i,j)$ is the initial RGB image of size $m \times n$ and $R(i,j)$, $G(i,j)$, and $B(i,j)$ are red, green, and blue channels, respectively, the red and blue channel gains (the green channel is usually left unchanged) are computed as

$$R_{gain} = \mu_G/\mu_R, B_{gain} = \mu_G/\mu_B \tag{3}$$

where μ_R, μ_G,and μ_B are the average intensity values of the R, G, and B channels.

B. Max white

The max white algorithm is another popular approach, which assumes that the maximum values of the RGB channels are white and the brightest white point should correspond to (2^n, 2^n, 2^n), where n is the number of bits per channel (usually 8 or 16). The gain values are computed as:

$$\begin{aligned} R_{gain} &= 2^n/R_{max} \\ G_{gain} &= 2^n/G_{max} \\ B_{gain} &= 2^n/B_{max} \end{aligned} \tag{4}$$

C. Retinex

Retinex is a well-known colour constancy algorithm. Unlike the rest of the algorithms in this study, retinex affects both chromatic properties and the lightness of an image, and is computed locally for small segments of the input image. For brevity, we omit the detailed description of the retinex implementation and refer the reader to Funt, et al. [14], whose retinex implementation we have used in this study.

Funt et al. [14] mention that the output of the retinex algorithm needs to be scaled to the output device. They do not provide implementation details for this post-processing algorithm, to which they refer as a "postlut". We therefore provide our implementation, which is given below.

We felt that because the post-processing step clearly changed the colour properties of the image, it was necessary to evaluate the retinex algorithm alone (Ret_unpost), as well as with the post-processing (Ret_post). Finally, we have also evaluated the post-processing step alone, to isolate any influence that it might have on the final colour of the processed images.

D. Stretch

We have implemented a simple histogram equalization algorithm, which we call Stretch. The individual colour channels on the image are first shifted to the left:

$$\begin{aligned} R_{new} &= R - R_{min} \\ G_{new} &= G - G_{min} \\ B_{new} &= B - B_{min} \end{aligned} \tag{5}$$

and then processed using the MaxW algorithm described above.

E. Combined grey world and "retinex" (quadratic mapping)

An algorithm combining both the grey world and retinex approaches was proposed by Lam [15]. It is worth noting that, for some reason, Lam [15] uses the term "retinex" to describe an algorithm virtually identical to the MaxW approach above and completely different from the localized retinex algorithm implemented by Funt et al. [14]. Our previous paper [20] incorrectly followed this usage. To avoid confusing the reader, we will simply refer to Lam's [15] algorithm as a "quadratic mapping" (QM). Lam bases his method on the observations that [15]:

- GW and "retinex" algorithms tend to produce different results and the corrected image rarely satisfies both assumptions
- Both algorithms adjust the image intensities linearly and, furthermore, there is also a fixed point in the mappings: for pixels with zero intensity, the two mappings would not affect their values.

He proposes a quadratic mapping mapping, which leaves the green channel unchanged and re-defines the red channel (blue channel is adjusted similarly) as [15]:

$$R'(i,j) = \mu R^2(i,j) + \nu R(i,j) \tag{6}$$

To satisfy the grey world assumption, Lam [15] requires that:

$$\mu \sum_{i=1}^{m}\sum_{j=1}^{n} R^2(i,j) + \nu \sum_{i=1}^{m}\sum_{j=1}^{n} R(i,j) = \sum_{i=1}^{m}\sum_{j=1}^{n} R(i,j) \tag{7}$$

and to satisfy the "retinex" assumption, he poses the requirement that:

$$\mu \max_{i,j}(R^2(i,j)) + \nu \max_{i,j}(R(i,j)) = \max_{i,j}(R(i,j)) \tag{8}$$

He then forms a system of two equations, which is solved to find μ and ν for the R channel (values of μ and ν for the B channel are computed analogously):

$$\begin{bmatrix} \sum\sum R^2 & \sum\sum R \\ \max R^2 & \max R \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix} = \begin{bmatrix} \sum\sum R \\ \max R \end{bmatrix} \tag{9}$$

F. Standard deviation-weighted grey world

Standard deviation-weighted grey world (SDWGW) algorithm extends the grey world assumption and was proposed by Lam et al. [17]. It subdivides the image into n blocks and for each one of them calculates standard deviations (σ_R, σ_G, σ_B) and means (μ_R, μ_G, μ_B) of the R, G, and B channels. SDWGD defines standard deviation-weighted averages of each colour channel as [17]:

$$\begin{aligned} SDWA_R &= \sum_{k=1}^{n} \frac{\sigma_R(k)}{\sum_{i=1}^{n}\sigma_R(i)} \times \mu_R(k) \\ SDWA_G &= \sum_{k=1}^{n} \frac{\sigma_G(k)}{\sum_{i=1}^{n}\sigma_G(i)} \times \mu_G(k) \\ SDWA_B &= \sum_{k=1}^{n} \frac{\sigma_B(k)}{\sum_{i=1}^{n}\sigma_B(i)} \times \mu_B(k) \end{aligned} \tag{10}$$

The new amplifier gains are then adjusted as follows:

$$\begin{aligned} R_{gain} &= \frac{SDWA_R + SDWA_G + SDWA_B}{SDWA_R} \\ G_{gain} &= \frac{SDWA_R + SDWA_G + SDWA_B}{SDWA_G} \\ B_{gain} &= \frac{SDWA_R + SDWA_G + SDWA_B}{SDWA_B} \end{aligned} \tag{11}$$

G. Standard deviation and luminance-weighted grey world

Standard deviation and luminance-weighted grey world algorithm (SDLWGW) was also proposed by Lam et al. and appears to be a variation of the SDWGW algorithm [18]. The image is similarly subdivided, but the weights are defined as follows [18]:

$$\begin{aligned} \mu_{L\_R} &= \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{L\_w(i,j)}{\sum_{x=1}^{m}\sum_{y=1}^{n} L\_w(x,y)} \times R_{i,j}(k) \\ \mu_{L\_G} &= \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{L\_w(i,j)}{\sum_{x=1}^{m}\sum_{y=1}^{n} L\_w(x,y)} \times G_{i,j}(k) \\ \mu_{L\_B} &= \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{L\_w(i,j)}{\sum_{x=1}^{m}\sum_{y=1}^{n} L\_w(x,y)} \times B_{i,j}(k) \end{aligned} \tag{12}$$

, where $L_w(i,j)$ is a positive single-peak function value for the luminance value at i-th row, j-th column of the k-th block [18]. Equation (10) is then reformulated and standard deviation and luminance-weighted averages are then defined as follows:

$$\begin{aligned} SDLWA_R &= \sum_{k=1}^{n} \frac{\sigma_R(k)}{\sum_{i=1}^{n}\sigma_R(i)} \times \mu_{L\_R}(k) \\ SDLWA_G &= \sum_{k=1}^{n} \frac{\sigma_G(k)}{\sum_{i=1}^{n}\sigma_G(i)} \times \mu_{L\_G}(k) \\ SDLWA_B &= \sum_{k=1}^{n} \frac{\sigma_B(k)}{\sum_{i=1}^{n}\sigma_B(i)} \times \mu_{L\_B}(k) \end{aligned} \tag{13}$$

The new gains are then computed by substituting $SDLWA_{R,G,B}$ into Equation 11:

$$\begin{aligned} R_{gain} &= \frac{SDLWA_R + SDLWA_G + SDLWA_B}{SDLWA_R} \\ G_{gain} &= \frac{SDLWA_R + SDLWA_G + SDLWA_B}{SDLWA_G} \\ B_{gain} &= \frac{SDLWA_R + SDLWA_G + SDLWA_B}{SDLWA_B} \end{aligned} \tag{14}$$

H. Adjacent Channels Adjustment by Standard Deviation and Luminance

Lam et al. also used SDLWA as defined in (13) in another algorithm called Adjacent Channels Adjustment by Standard Deviation and Luminance algorithm (ACASDL) [16], which computes the new gain values as follows. The new gain for the B channel is adjusted by:

$$B'_{gain} = B_{gain} \times \frac{(SDLWA_R + SDLWA_G)/2}{SDLWA_B} \tag{15}$$

The R gain is then recalculated by first plugging the newly adjusted values of B into (13) and then replacing B, R and G by R, G and B respectively in (15). Similarly, to calculate the G gain, the new values of R and B are plugged into (13) and the gain is calculated by replacing B, R and G by G, B and R respectively in (15).

III. METHODS

A. Metrics

Several different colour distance metrics have been proposed in the literature. We have implemented the CIE L*a*b metric described in [21]. This metric did not require any data-set specific adjustment coefficients that some of the more potentially accurate metrics used (for instance, see [22]). Individual pixel distances are defined as follows:

$$\Delta E^*_{damaged} = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2} \tag{16}$$

Image distances are then defined as follows:

$$\Delta E = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \Delta E^*_{damaged}(i,j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} \Delta E_{reference}(i,j)} \tag{17}$$

where

$$\Delta E_{reference} = \sqrt{(L^*)^2 + (a^*)^2 + (b^*)^2} \tag{18}$$

B. Data

In order to facilitate objective comparison of the white-balancing algorithms, we needed data sets with pairs of faded and ground truth reference images containing exactly the same information. An ideal approach would have been to take a set of photographs, make photographic copies of them, and subject several sets of copied photographs to artificial fading. Due to the lack of the necessary equipment, we have used published data describing fading of slide films to produce "synthetic" faded data sets from an original set of 214 undamaged images. The images featured outdoor scenes and were properly white-balanced.

We multiplied the original images by bleaching matrices published by Frey and Gschwind [3]. The data sets, used for both objective and subjective evaluation, were:

- "Medium" (simulated Kodak Ektachrome 100 faded for 1800 hours in 50% humidity) - least prominent colour degradation
- "Heavy" (KE100, 2600 hrs, 50% hum.) - more colour degradation
- "Humid" (KE100, 1800 hrs, 75% hum.) - the worst degradation

Three additional data sets of actual faded Kodak Ektachrome slides were used, each exhibiting different degrees of fading. The set we will refer to as "Ekt1" displayed the least fading, "Ekt3" - the most, and "Ekt2" could be said to be in the somewhere between the other two in terms of damage severity. There were 27 images in each sample, which were randomly chosen from the corresponding data sets. The majority of images featured outdoor scenes, however several indoor scenes were also present.

C. Objective test methodology

1) Design: We opted for a 3 (degrees of damage severity: Medium, Heavy, and Humid data sets) $\times$ 9 (restoration algorithms) factors design.

2) Procedure and Apparatus: All algorithms were implemented in Matlab by the authors except retinex - we used the Matlab retinex implementation (Frankle-McCann algorithm) by Funt et al. [14] available online[1] and post-processed the result using our Stretch algorithm described above. We have verified the correctness of our post-processing algorithm my running Ret_post on the test images provided by Funt et al. [14] on their website and ensuring that our results matched theirs. For every data set, each of the nine restoration algorithms processed all 214 test images. The distance metrics were then computed for every pair of original/restored images.

D. Subjective test methodology

1) Participants: Our final sample consisted of 39 students from the International Academy of Design and Technology, Toronto.

[1] http://www.cs.sfu.ca/ colour/publications/IST-2000/

2) Design: We opted for a 6 (Medium, Heavy, Humid, Ekt1, Ekt2, and Ekt3 data sets) $\times$ 9 (restoration algorithms) factors design with repeated measure on the second factor, where all the participants who evaluated data from a particular data set scored all of the 9 algorithms. For instance, for the data set Medium, each of the participants evaluated 25 sample images, each processed by all 9 algorithms.

3) Procedure and Apparatus: The test procedure was broken into 6 sessions, one for each data set. The participants were presented with the slides on identical Apple iMac computers, all of which were colour-calibrated. The estimated time to evaluate each slide was one minute and the total number of slides was kept under 30 to keep the running time to approximately 30 minutes. This was done in order to minimize the effects of fatigue and boredom on the participants, as recommended by ITU [23], [24].

For every session, a set of 25 (for Medium and Humid) or 27 (for all others) slides was prepared. The participants' judgement of the first few stimuli was likely to be inconsistent with their judgement of the majority of subsequent stimuli. In accordance with the ITU recommendations [24], we have therefore discarded the initial slides from every session and ensured that the minimum of 15 participants were involved in evaluating each of the data sets. Out of the total 39 participants, 18 volunteered to evaluate data set Medium, 22 - Heavy, 21 - Humid, 26 - Ekt1, 28 - Ekt2, and 26 - Ekt3.

Synthetic data was evaluated in the following order: Heavy, Medium, and Humid. The 3 sessions were run on the same day with breaks between them. The first 5 results from the set Heavy and the first 3 results from the sets Medium and Humid were later discarded. We have also discarded the first 5 results from all of three Ekt* data sets, because we felt that the sets were quite different from each other and the participants might have taken longer to adjust to every set.

For each slide, the original faded image (synthetic or real) was restored using all 9 algorithms. The resulting images were displayed on the slide in 3 rows and labelled "1" to "9". The order in which the images were presented was random and varied for every slide, but a look-up table was kept for each slide set and the images sequences were "unscrambled" after all the evaluation sessions were finished. Thus this was a double-blind study - neither the participants nor the experimenter knew which of the restoration algorithms produced each of the 9 images on each slide. The original faded images were not included in the slides in an attempt to avoid biasing the participants. However, for the three synthetic data sets, each slide also featured the original undamaged image (clearly labelled) to help the participants evaluate restoration quality. For the Ektachrome data sets the ground truth images were not available, since manual restoration of each of the sample images was not possible within the time allocated to this study.

The participants were each given a questionnaire where they were asked to assign a "quality value" between 0 and

TABLE II.
SUMMARY OF SUBJECTIVE EVALUATION RESULTS (HIGHER VALUES INDICATE HIGHER QUALITY)

	Med.	*Heavy*	*Humid*	*Ekt1*	*Ekt2*	*Ekt3*
ACASDL μ	3.38	2.63	1.99	5.07	3.89	2.24
GW μ	3.39	2.57	2.03	4.47	3.73	2.35
GWRet μ	3.41	3.11	4.98	5.47	**7.29**	**4.43**
MaxW μ	3.21	1.50	1.10	4.62	2.26	0.48
SDWGW μ	3.43	2.57	1.89	5.22	3.86	1.76
SDWLGW μ	3.43	2.50	1.89	5.16	3.76	1.81
Ret_unpost μ	2.16	0.96	0.82	4.01	1.41	0.31
Ret_post μ	5.67	6.27	6.02	5.21	1.69	0.34
Stretch μ	**8.61**	**8.72**	**8.17**	5.27	2.55	0.46

TABLE III.
SUMMARY OF OBJECTIVE EVALUATION RESULTS (LOWER VALUES INDICATE HIGHER QUALITY)

	Medium	*Heavy*	*Humid*
ACASDL μ	0.26	0.38	0.51
GW μ	0.26	0.36	0.46
GWRet μ	0.24	0.31	0.23
MaxW μ	0.24	0.39	0.76
SDWGW μ	0.27	0.40	0.58
SDWLGW μ	0.27	0.40	0.58
Ret_unpost μ	0.55	0.64	0.81
Ret_post μ	0.40	0.42	0.45
Stretch μ	**0.02**	**0.04**	**0.09**

9 to every image on every slide. This scale was designed to measure the absolute quality of the image as perceived by humans, where 0 corresponded to a completely "unusable" image and 9 indicated a perfectly restored image (completely natural colours). The participants were instructed to use their personal judgement when assigning scores to each image. Since we were interested in measuring the absolute rather than relative performance of the algorithms, we did not instruct the participants to avoid assigning equal scores to multiple algorithms. Sample questionnaires and the slide sets viewed by the participants are available from the authors upon request.

IV. RESULTS

We explored the mean of the dependent variable (image score for subjectively evaluated data and colour distance for objectively evaluated data) to analyze algorithm performance. The summary of the results is given in Tables II and III.

The dependent variable was not normally distributed, which was expected in the case of image scores, so all the pair-wise algorithm comparisons were conducted using non-parametric Mann-Witney-Wilcoxon tests. Since many of the results were not significantly different, we have omitted the full list of all the significance statistics from this paper. Also, for brevity, we only reported the relevant significance values.

A. Subjective evaluation - synthetic data

For all 3 synthetic data sets, the Stretch algorithm performed notably better than the rest of the algorithms with significance value of less than .001 (Table II). The next highest scoring algorithm with the significance value of less than .001 was Ret_post (Table II). The worst performing algorithm was Ret_unpost with the significance of of less than .001 (Table II).

All of the other algorithms, with the exception of QM, have performed consistently poorly (Table II). Many of the differences between the low-scoring algorithms were not statistically significant with the significance value of more than .05. For data sets Medium and Heavy, QM was judged to be about the same as ACASDL, SDWGW, SDWLGW, MaxW, and GW (Table II). However, it has scored third overall on the Humid data set and it's performance was better than that of ACASDL, SDWGW, SDWLGW, MaxW, and GW with the significance value of less than .001(Table II).

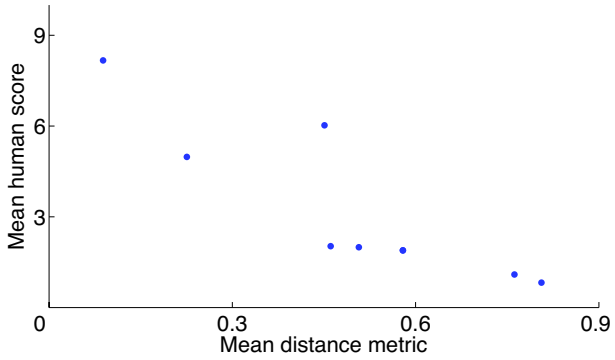


Figure 1. Distance metric vs. human scores: Humid data set

B. Subjective evaluation - Kodak Ektachrome slides

For the Ektachrome data sets the results were quite different. No single algorithm performed significantly better than others on Ekt1. GWRet has the highest mean (Table II). It's performance was significantly different from that of GW, MaxW, ACASDL, SDWGW, SDWLGW, and Ret_unpost with the significance of less that .01. However, its performance was not significantly different from that of Ret_post and Stretch with the significance value of more than .05. For data sets Ekt2 (heavy colour cast) and Ekt3 (very heavy colour cast), QM has, on average, scored higher than all the other algorithms (Table II) with the significance value of less than .001.

C. Objective evaluation and correlation

Objective evaluation results were somewhat consistent with the subjective results. Stretch was still the best-performing algorithm, and Ret_unpost still the worst, with the significance value of less than .001. However, QM was the second best performing algorithm on Heavy and Humid data (Table III) with the significance of less than .001. It can be also noted that Ret_post did not perform as well in the objective evaluation (Table III) as at did in the subjective evaluation (Table II).

We ran a correlational analysis of the means of objective and subjective results on synthetic data for each algorithm. Because the means of objective and subjective results were not normally distributed, non-parametric

measures - Spearman coefficient ρ and Kendall coefficient τ_b - were used. Both indicated no significant correlation between subjective and objective results on Medium and Heavy data, and a significant and strong negative correlation on Humid data ($\rho = -.93, sig. = .001$; $\tau_b = -.979, sig. < .001$). The scatterplot of the mean Humid data can be seen in Figure 1.

V. Discussion

Overall, the results presented in this paper form a strong indication that "off-the-shelf" white-balancing algorithms are not suitable for restoration of faded historical photographic materials. This is quite likely due to the fact that damage due to colour dye fading (modelled by (2) is quite different from a colour cast due to a change in illumination conditions (1), despite the similarity in the outward appearance of both types of "mis-coloured" images.

The low performance of Ret_unpost was to be expected, since the algorithm was meant to be run with post-processing. The reasonably successful performance of Ret_post on synthetic data was most likely due to the strong performance of its post-processing algorithm Stretch, which has significantly outperformed Ret_post. The difference between the performance of Ret_post on synthetic and original data could in part be explained by the fact that it modified the lightness of an image in addition to its colour. Perhaps a metric that only used the "a" and "b" dimensions of the CIE L*a*b space would be more appropriate.

It seems unlikely that the restored images produced by QM from the Humid data set were actually better than those produced from less damaged data in Medium and Heavy sets. We hypothesize that the participants' judgement of colour quality was affected by the context - the more "natural" looking images got somewhat exaggerated high scores, particularly if the majority of the images on each slide showed a heavy colour cast. This might also explain why Stretch and Ret_post have exhibited little performance deterioration as the severity of the synthetic damage increased. As seen from Table II, the average scores of all but the top 3 algorithms worsened as the degree of synthetic damage increased.

QM's unexpectedly high performance on Ekt2 could also be attributed to the "context effect". Its actual absolute performance is more likely to be similar to its average scores on all the other data sets.

The significant differences between the subjective evaluation results on synthetic and real data indicate that the bleaching matrices used to create the synthetic data do not accurately model the damage present in the Kodak Ektachrome slides in the authors' collection. Unfortunately, published bleaching data is extremely scarce. The authors are working on obtaining more bleaching matrices to create more varied synthetic data.

The fairly consistent, albeit rather mediocre, performance of QM might indicate its potential applicability to restoring historical images. However, the algorithm would have to be modified to increase its performance. Consistently good performance of Stretch on synthetic data might indicate its suitability for restoring lightly faded historical data, but further testing would be required to establish this.

We feel that the results presented above highlight the difficulty of evaluating colour restoration algorithms. The lack of correlation between the subjective and objective performance of the algorithms on 2 out of 3 data sets is an indication that the simple distance metric used in this study does not necessarily correspond to human perception of colour. In addition, the we feel that any distance metric has a potential for being misleading. Images with markedly different appearance can be numerically virtually equidistant from the ground truth. For instance, GW and Ret_post have similar mean distances for Humid damage(Table III), but look quite different (Figures 4(d) and 4(j)). Similarly, ACASDL and MaxW have similar mean distances for Heavy damage (Table III), but the ACASDL results have a green/cyan cast, while MaxW results exhibit a pink cast (Figures 3(c) and 3(f)).

This, combined with the current lack of colour metrics that could measure the severity of a colour cast without ground truth data, means that subjective evaluation remains the most accurate method of measuring performance of colour restoration methods. However, subjective evaluation is not without its difficulties. The use of human participants significantly increases the time necessary to complete the evaluation. In addition, the best way to compare performance of colour image processing algorithms is to display 2 results at a time to the participants [24], otherwise the objectivity of the observers is likely to suffer and result in artifacts, such as the "context effect" observed in this study. This would further increase the duration and complexity of the subjective evaluation procedure.

Thus one important development that would allow for fast and accurate objective algorithm evaluation would be the development of colour damage metrics that correspond to the human perception of colour quality and do not require reference images.

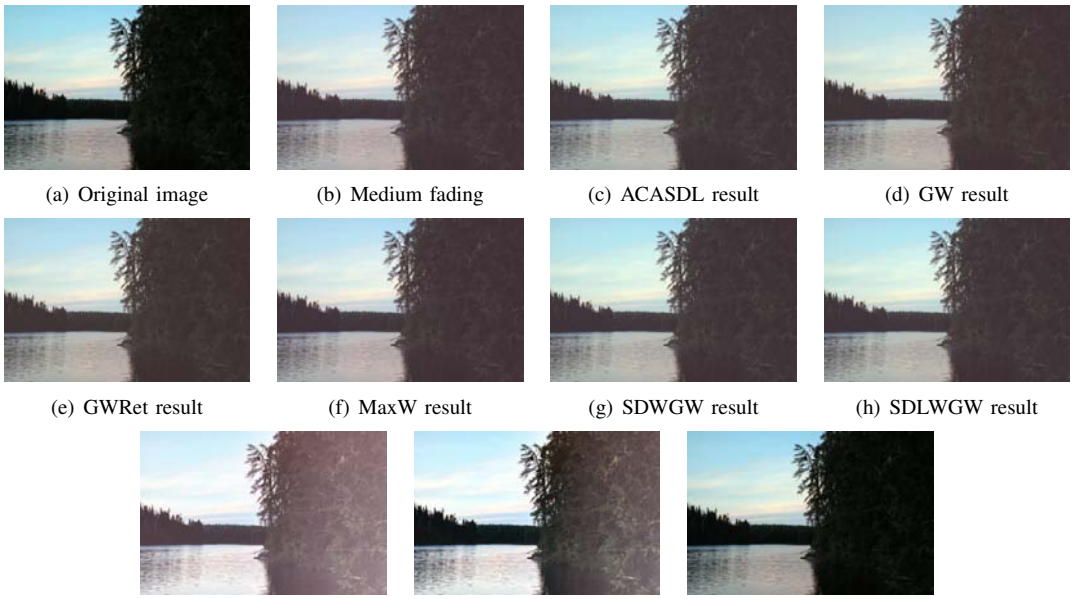
References

[1] L. Rosenthaler and R. Gschwind, "Restoration of movie films by digital image processing," in *IEE Seminar: Digital Restoration of Film and Video Archives*. IEE, 2001, pp. 6/1–6/5.

[2] R. Gschwind, "Restoration of faded color photographs by digital image processing," in *Image Processing III*, 1989, pp. 27–37.

[3] F. Frey and R. Gschwind, "Mathematical bleaching models for photographic three-color materials," *Journal of Imaging Science and Technology*, vol. 38, pp. 513–519, 1994.

[4] R. Gschwind and F. Frey, "Electronic imaging, a tool for the reconstruction of faded color photographs," *Journal of Imaging Science and Technology*, vol. 38, pp. 520–525, 1994.

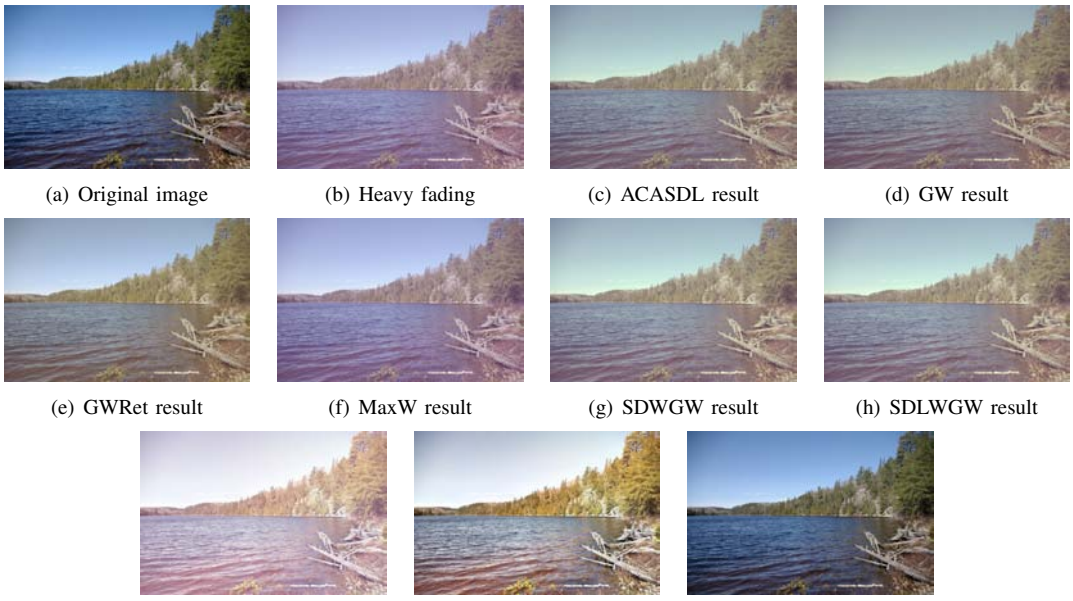


(a) Original image (b) Medium fading (c) ACASDL result (d) GW result

(e) GWRet result (f) MaxW result (g) SDWGW result (h) SDLWGW result

(i) Ret_unpost result (j) Ret_post result (k) Stretch result

Figure 2. Sample restoration results for data set Medium



(a) Original image (b) Heavy fading (c) ACASDL result (d) GW result

(e) GWRet result (f) MaxW result (g) SDWGW result (h) SDLWGW result

(i) Ret_unpost result (j) Ret_post result (k) Stretch result

Figure 3. Sample restoration results for data set Heavy

[5] R. Gschwind, F. S. Frey, and L. Rosenthaler, "Electronic imaging: a tool for the reconstruction of faded color photographs and color movies," in *Proc. SPIE Image and Video Processing III*, 1995, pp. 57–63.

[6] M. Chambah and B. Besserer, "Digital color restoration of faded motion pictures," in *CGIP Conf. Proc.*, 2000, pp. 338–342.

[7] M. Chambah, B. Besserer, and P. Courtellemont, "Recent progress in automatic digital restoration of color motion pictures," in *Proc. SPIE Color Imaging: Device-Independent Color, Color Hardcopy, and Applications VII*, 2001, pp. 98–109.

[8] M. Chambah, B. Besserer, and P. Courtellemont, "Approach to automate digital restoration of faded color film," in *IS&T CGIV 2002*, 2002, pp. 613–618.

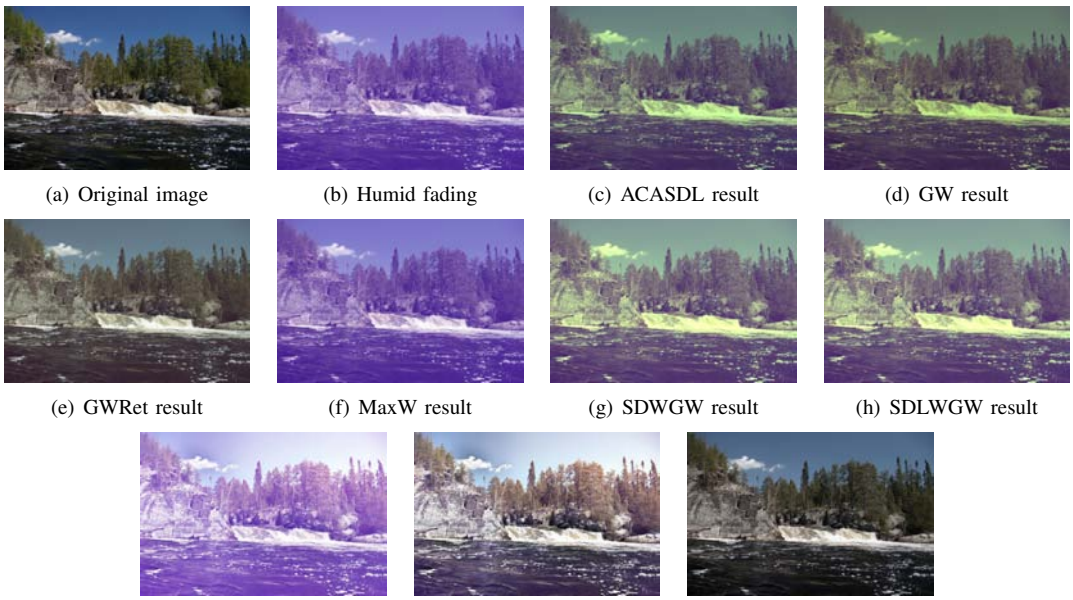
[9] ——, "Latest results in digital color film restoration," *MG&V*, vol. 11, no. 2/3, pp. 363–395, 2002.

[10] M. Chambah, A. Rizzi, C. Gatta, B. Besserer, and D. Marin, "Perceptual approach for unsupervised digital color restoration of cinematographic archives," in *Procs. SPIE / IS&T Electronic Imaging*, vol. 5008, 2003, pp. 138–149.

[11] A. Rizzi, M. Chambah, D. Lenza, B. Besserer, and D. Marini, "Tuning of perceptual technique for digital movie color restoration," in *Proc. SPIE Security, Steganography, and Watermarking of Multimedia Contents VI*, 2004, pp. 1286–1294.

[12] A. Rizzi, C. Gatta, C. Slanzi, G. Ciocca, and R. Schettini, "Unsupervised color film restoration using adaptive color equalization." in *Visual Information and Information Systems, 8th International Conference*, 2005, pp. 1–12.

[13] F. Gasparini and R. Schettini, "Color balancing of digital photos using simple image statistics." *Pattern Recognition*, vol. 37, no. 6, pp. 1201–1217, 2004.

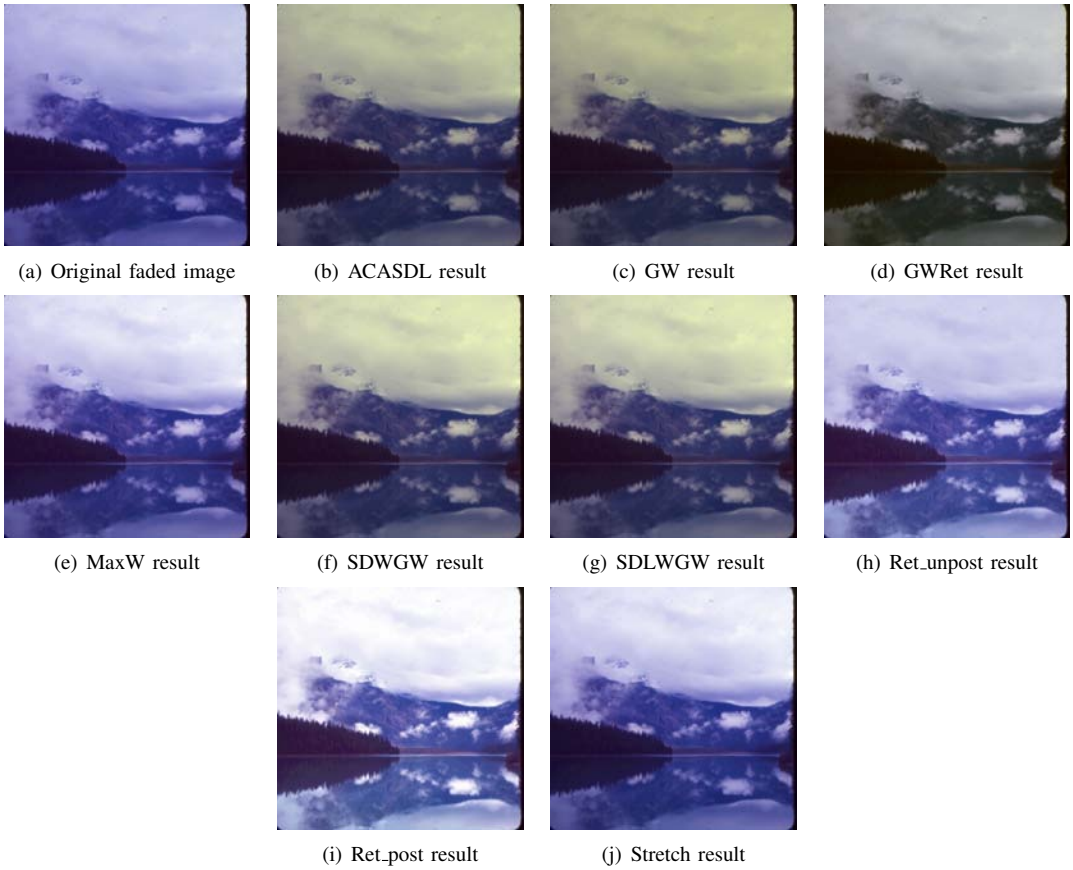


(a) Original image (b) Humid fading (c) ACASDL result (d) GW result

(e) GWRet result (f) MaxW result (g) SDWGW result (h) SDLWGW result

(i) Ret_unpost result (j) Ret_post result (k) Stretch result

Figure 4. Sample restoration results for data set Humid

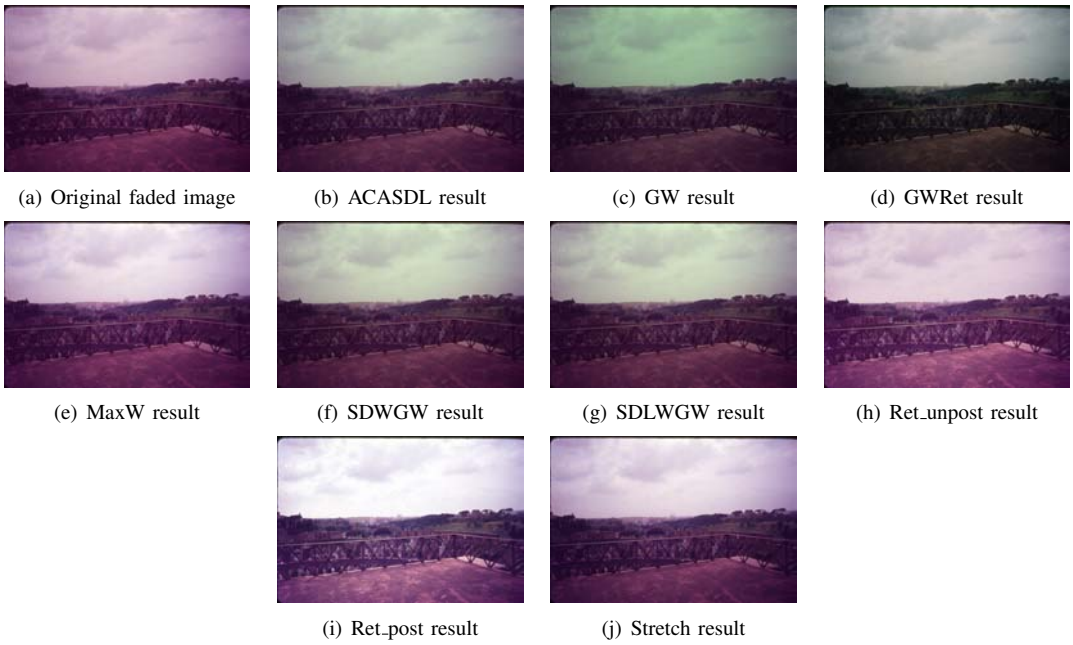


(a) Original faded image (b) ACASDL result (c) GW result (d) GWRet result

(e) MaxW result (f) SDWGW result (g) SDLWGW result (h) Ret_unpost result

(i) Ret_post result (j) Stretch result

Figure 5. Sample restoration results for data set Ekt1

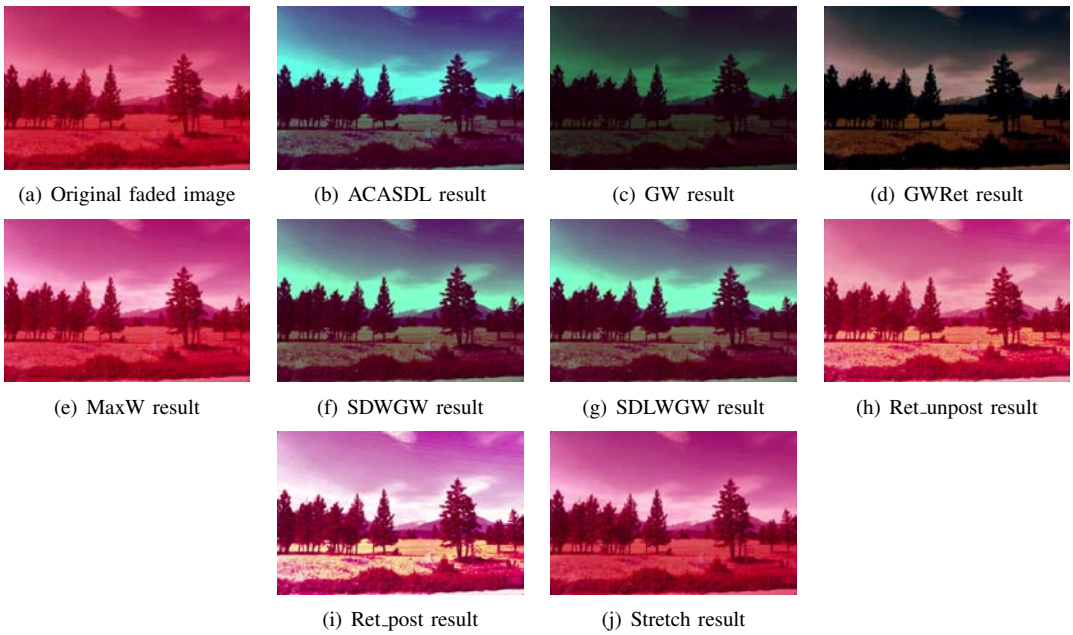


(a) Original faded image (b) ACASDL result (c) GW result (d) GWRet result

(e) MaxW result (f) SDWGW result (g) SDLWGW result (h) Ret_unpost result

(i) Ret_post result (j) Stretch result

Figure 6. Sample restoration results for data set Ekt2



(a) Original faded image (b) ACASDL result (c) GW result (d) GWRet result

(e) MaxW result (f) SDWGW result (g) SDLWGW result (h) Ret_unpost result

(i) Ret_post result (j) Stretch result

Figure 7. Sample restoration results for data set Ekt3

[14] B. V. Funt, F. Ciurea, and J. J. McCann, "Retinex in matlab," in *Color Imaging Conference*, 2000, pp. 112–121.

[15] E. Lam, "Combining gray world and retinex theory for automatic white balance in digital photography," in *ISCE*, 2005, pp. 134–139.

[16] H.-K. Lam, O. C. Au, and C.-W. Wong, "Automatic white balancing using adjacent channels adjustment in rgb domain," in *ICME*, 2004, pp. 979–982.

[17] ——, "Automatic white balancing using standard deviation of rgb components." in *ISCAS (3)*, 2004, pp. 921–924.

[18] ——, "Automatic white balancing using luminance component and standard deviation of rgb components," in *ICASSP '04*, vol. 3, 2004, pp. 493–496.

[19] R. Hunt, *The Reproduction of Colour*. Wiley, 2004.

[20] D. Nikitenko, M. Wirth, and K. Trudel, "White-balancing algorithms in colour photograph restoration," in *IEEE International Conference on Systems, Man and Cybernetics*, 2007, pp. 1037–1042.

[21] *Colorimetry*, 2nd ed. Vienna, Austria: Commission Internationale de l'Eclairage, 1986.

[22] F. Imai, N. Tsumura, and Y. Miyake, "Perceptual color difference metric for complex images based on mahalanobis distance," *Journal of Electronic Imaging*, vol. 10, pp. 385–393, 2001.

[23] ITU, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union - Radiocommunication Sector, Tech. Rep., 2002, recommendation ITU-R BT.500-11.

[24] ——, "Studies toward the unification of picture assessment methodology," International Telecommunication Union - Radiocommunication Sector, Tech. Rep., 1990, iTU-R BT.1082-1.