# Statistic Analysis of Millions of Digital Photos

Dietmar Wueller[a], Reiner Fageth[b]
[a]Image Engineering Dietmar Wueller, Augustinusstr 9D, 50226 Frechen, Germany
[b]CeWe Color AG, Meerweg 30-32, 26133 Oldenburg, Germany

## ABSTRACT

The analysis of images has always been an important aspect in the quality enhancement of photographs and photographic equipment. Due to the lack of meta data it was mostly limited to images taken by experts under predefined conditions and the analysis was also done by experts or required psychophysical tests.

With digital photography and the EXIF[1] meta data stored in the images, a lot of information can be gained from a semiautomatic or automatic image analysis if one has access to a large number of images. Although home printing is becoming more and more popular, the European market still has a few photofinishing companies who have access to a large number of images. All printed images are stored for a certain period of time adding up to several million images on servers every day.

We have utilized the images to answer numerous questions and think that these answers are useful for increasing image quality by optimizing the image processing algorithms. Test methods can be modified to fit typical user conditions and future developments can be pointed towards ideal directions.

**Keywords:** Image Quality, Exposure Value, EXIF, Meta Data, Image Analysis

## 1. INTRODUCTION

In November 2007, CeWe Color had access to around 4.8 million JPEG images at the stage in which they were sent to the photofinishing company by their customers. The origin of each single image is not known. From experience and from the analysis of the Meta data it is known, however that a large number were taken by digital still cameras. A small number were taken with cell phone cameras, that in many cases do not use the Exif header, and another unknown number were scanned from negatives by the customers. From past experience it is known that most of the images taken with a digital camera are the original images without any post processing. The number of JPEG images originating from digital SLRs in this study is negligible.

The easiest analysis is that of the Exif Meta data itself. A database was filled with the values found in the different tags and the appearance of a specific value was counted for each tag.

Some of the questions asked by the experts, for example: “is the world really gray”, could only be answered using image analysis software which evaluated each pixel in every single image. To ensure an analysis in a reasonable amount of time, it was necessary to keep the evaluation algorithms as simple as possible.

A few questions, such as categorizing the images, required a visual analysis. Since this would have been impossible with the total number of images, a statistic selection of 10.000 images was used to collect the presented data.

After an intensive discussion among the authors and numerous experts, the following questions appeared to be the most interesting ones that were able to be answered by the current study:

1. How many images have embedded EXIF data?
2. Which Exif tags are used how often?
3. How many images consist of how many pixels?
4. How many images were shot at which ISO level?
5. What is the distribution among the different focal lengths?

6. How often was the flash fired?
7. Sorting the images by exposure levels.
8. What is the average compression rate?
9. Are there images that already include GPS information?
10. How many images were shot at which white balance setting?
11. Is the real world really gray or what is the average color?
12. What is the average color in a daylight and in a tungsten image?
13. How many images have a dominant color?
14. How often do images contain colors at the border of the sRGB color space?
15. How many images are under or overexposed?
16. Sorting the images by typical scene types.

Questions 1 – 10 could be answered using the Exif Analysis. Questions 11 – 15 required an image analysis and question number 16 could only be answered using visual inspection.

Another interesting question was the long-term validity of answers provided by these questions. For some of the questions, such as the sorting by scenes or average color, the results provide an answer at least for the next couple of years. The answer to the average pixel count however represents the actual situation, but is expected to change over time due to technological changes in future digital cameras.

## 2. EXIF ANALYSIS

The following answers are based on analysis of the Exif Meta data.

### 2.1 How many images have embedded Exif data?

| **How many images contain Exif information** | Images | Percent |
|---|---|---|
| Evaluated images | 4.788.645 | |
| Images containing Exif information<br>(ISO tag or camera manufacturer tag) | 3.447.579 | 72,0 |
| Images without Exif information<br>(no ISO tag and no camera manufacturer tag) | 1.341.066 | 28,0 |

The presence of Exif information was analyzed by looking into the manufacturer and the ISO tag. If one of them was filled with valid information, the image was rated as one with Exif information.

For the scanned images we did not expect Exif data, although an indication that these images were scans from negatives or pictures would sometimes be helpful. The camera modules of many cell phones also do not provide Exif data. Some manufacturers have not realized how important Exif information can be for archiving and support purposes.

### 2.2 Which Exif tags are used how often?

| **Which Exif tags are used how often?** | Images | Percent |
|---|---:|---:|
| Image with Exif information | 3.447.579 | |
| Exif version | 3.430.131 | 99,5 |
| Camera make | 3.402.731 | 98,7 |
| Camera model | 3.400.491 | 98,6 |
| Camera software | 1.806.142 | 52,4 |
| Date time | 3.413.654 | 99,0 |
| Width | 3.393.993 | 98,4 |
| Height | 3.393.993 | 98,4 |
| Flash used | 3.446.694 | 100,0 |
| Metering mode | 3.334.693 | 96,7 |
| Light source | 2.029.204 | 58,9 |
| Focal length | 3.256.955 | 94,5 |
| Exposure time | 3.338.908 | 96,8 |
| Exposure program | 2.436.825 | 70,7 |
| Exposure Bias Value | 3.357.798 | 97,4 |
| Fnumber | 3.287.745 | 95,4 |
| Max Aperture Value | 3.087.022 | 89,5 |
| ISO | 2.386.274 | 69,2 |
| Orientation | 2.984.946 | 86,6 |

Every image that had a value in the selected tag was counted regardless of the value itself. The pixel count evaluation for example revealed afterwards that some images do not have a meaningful value in that specific tag.

Some tags are used on a regular basis, i.e. the “flash used” tag. Others, i.e. “ISO speed” and “light source”, are not used that often. The reason may be a fixed setting for the specific value in that camera or the manufacturer did not want to include this value in the tag for whatever reason. The light source, for example cannot always be estimated from the white balance setting, furthermore only a few cameras are able to provide GPS information.

### 2.3 How many images consist of how many pixels?

| **How many images consist of how many pixels?** | Images | Percent |
|---|---:|---:|
| Images with width and height information | 3.392.080 | |
| 1mio | 332.142 | 9,8 |
| 1mio - 2mio | 634.649 | 18,7 |
| 2mio - 3mio | 83.344 | 2,5 |
| 3mio - 4mio | 1.032.470 | 30,4 |
| 4mio - 5mio | 310.559 | 9,2 |
| 5mio - 6mio | 577.123 | 17,0 |
| 6mio - 7mio | 118.976 | 3,5 |
| 7mio - 8mio | 228.100 | 6,7 |
| 8mio - 9mio | 12.923 | 0,4 |
| 9mio - 10mio | 20.210 | 0,6 |
| 10mio - 11mio | 34.789 | 1,0 |
| > 11 mio | 6.795 | 0,2 |

The pixel count analysis is based on the images with a valid width and height value and is a snap shot of the situation in November 2007. It is interesting to see that the 8 to 12 Megapixel images still represent a minority among the images taken. Do many consumers switch their cameras to lower pixel counts, or are there only a few cameras with high pixel counts used by customers at this time? A question that remains open.

## 2.4 How many images were shot at which ISO level?

| **How many images were shot at which ISO level?** | Images | Percent |
|---|---|---|
| Total number of images with ISO tag | 2.386.274 | |
| 50 | 269.753 | 11,3 |
| 64 | 152.198 | 6,4 |
| 70 | 11.217 | 0,5 |
| 80 | 252.732 | 10,6 |
| 100 | 827.059 | 34,7 |
| 125 | 69.983 | 2,9 |
| 128 | 12.953 | 0,5 |
| 140 | 51.193 | 2,1 |
| 141 | 9.815 | 0,4 |
| 160 | 95.797 | 4,0 |
| 200 | 225.671 | 9,5 |
| 250 | 25.332 | 1,1 |
| 320 | 42.836 | 1,8 |
| 400 | 152.043 | 6,4 |
| >400 | 41.012 | 1,7 |
| others | 146.680 | 6,1 |

Sixty-five percent of the images were taken at an ISO level of 100 and below. The analysis of this tag has to be viewed in combination with the flash used and the exposure level analysis. With a 46% percent “flash fired” rate and 68% being portrait images, the light level is in many cases high enough to use ISO levels of 100 and below. This is essential because many of today’s consumer cameras need low sensitivity levels to provide low noise images or at least low noise reduction levels.

## 2.5 What is the distribution among the different focal lengths?

| **What is the distribution among the different focal lengths?** | Images | Percent |
|---|---|---|
| Number of evaluated images (with focal length tag) | 1.755.360 | |
| <4mm | 15.039 | 0,9 |
| 4mm - 5mm | 146.352 | 8,3 |
| 5mm - 6mm | 400.919 | 22,8 |
| 6mm - 7mm | 230.645 | 13,1 |
| 7mm - 8mm | 201.180 | 11,5 |
| 8mm - 10mm | 103.790 | 5,9 |
| 10mm - 15mm | 167.251 | 9,5 |
| 15mm - 20mm | 165.813 | 9,4 |
| 20mm - 50mm | 214.185 | 12,2 |
| 50mm - 100mm | 80.833 | 4,6 |
| >100mm | 29.353 | 1,7 |

By looking at the consumer cameras which were available over the last 5 years the average factor to determine the equivalent 35mm focal length can be estimated around 6 which represents a sensor diagonal of app. 7mm. This means around 50% of the images were shot at normal to slightly wide angle focal length.

## 2.6 How often was the flash fired?

| **How often was the flash fired?** | Images | Percent |
|---|---:|---:|
| Images with Exif Meta data | 3.446.694 | |
| With flash | 1.502.499 | 43,6 |
| Without flash | 1.944.195 | 56,4 |
| No value in flash tag | 885 | 0,0 |

How many images could have been taken without flash at higher ISO speed levels to provide more natural images can not be determined from this data.

## 2.7 Sorting the images by exposure levels.

| **Sorting the images by exposure levels.** | Images | Percent |
|---|---:|---:|
| Number of evaluated images<br>(with fnumber, speed and ISO tag) | 908.989 | |
| <6 | 13.434 | 1,5 |
| 6---8 | 72.917 | 8,0 |
| 8--10 | 260.402 | 28,6 |
| 10--12 | 244.491 | 26,9 |
| 12--14 | 217.178 | 23,9 |
| 14--16 | 72.915 | 8,0 |
| 16--18 | 20.384 | 2,2 |
| >18 | 7.268 | 0,8 |

The exposure level[2] (not to be confused with the exposure index described in ISO 12232) could only be determined for images with a valid fnumber, exposure speed and ISO level tag.

Where the exposure level is:

$$2^B = \frac{f^2}{t} + \frac{actual\ sensitivity}{ISO\,100} \qquad (1)$$

with B = exposure value

Typical exposure levels:

| | |
|---|---|
| Indoor with tungsten light | < 7 |
| Outdoor, cloudy sky | 7 – 11 |
| Outdoor, sunny day | >11 |
| Portrait with flash | ≈10 |

### 2.8 What is the average compression rate?

| **What is the average compression rate?** | Images | Percent |
|---|---|---|
| Total images with width and hight information | 3.010.184 | |
| ratio file size compressed / uncompressed | | |
| <0,05 | 20.432 | 0,7 |
| <0,1>0,05 | 51.987 | 1,7 |
| <0,2 >0,1 | 512.007 | 17,0 |
| <0,3>0,2 | 981.372 | 32,6 |
| <0,4>0,3 | 560.443 | 18,6 |
| <0,5>0,4 | 626.390 | 20,8 |
| <0,6>0,5 | 167.747 | 5,6 |
| <0,7>0,6 | 64.726 | 2,2 |
| <0,8>0,7 | 25.107 | 0,8 |
| <1,1>0,8 | 20.405 | 0,7 |

To calculate the compression rate it was necessary to determine the uncompressed file size from the pixel count. The compression rate was calculated by dividing the size of the compressed file by the size of the uncompressed one. Images with a compression rate of more than 1,1 were excluded from the evaluation.

### 2.9 Are there images that already include GPS information?

| **Are there images that already include GPS information?** | Images | Percent |
|---|---|---|
| Total number of images evaluated for GPS information | 2.324.033 | |
| Images with GPS information | 13.355 | 0,6 |
| Images without GPS information | 2.310.678 | 99,4 |

Right now only a few cameras and mobile phones have an integrated GPS receiver. But there are ways to add GPS information to images by using an external sensor and software that compares time and position. A few digital SLRs also have an extra connector for GPS devices. Due to a timing issue this analysis was not done on the same image set as the others.

### 2.10 How many images were shot at which white balance setting?

| **How many images were shot at which white balance setting?** | Images | Percent of total | Percent of images with defined setting |
|---|---|---|---|
| Total number of images with Exif data | 3.393.853 | | |
| Daylight | 39.457 | 1,2 | 70,4 |
| Flash | 7.692 | 0,2 | 13,7 |
| Fluorescent | 3.768 | 0,1 | 6,7 |
| Multisegment (?) | 20 | 0,0 | |
| Multispot (?) | 10 | 0,0 | |
| Std A (?) | 753 | 0,0 | |
| Std B (?) | 20 | 0,0 | |
| Std C (?) | 10 | 0,0 | |
| Tungsten | 5.098 | 0,2 | 9,1 |
| Undefined | 1.972.496 | 58,1 | |
| No value | 1.364.529 | 40,2 | |

Most of the images were taken using auto white balancing. In this case the cameras do not write a value or the "undefined" value into the tag.

## 3. EVALUATION BASED ON A PIXEL BY PIXEL IMAGE ANALYSIS

The analysis was performed using software that was created by the CeWe Color team for studies like this. Each pixel of every image is evaluated for specific characteristics. To ensure a good performance the algorithms used have to be as simple as possible.

### 3.1 Is the real world really gray or what is the average color?

| | Average **digital value** | Average value for **red** | Average value for **green** | Average value for **blue** |
|---|---|---|---|---|
| Average value for all 4.8 million images | 109,27 | 115,84 | 109,86 | 102,11 |
| Average value of images with Exif tag daylight | 113,35 | 118,21 | 114,18 | 107,64 |
| Average value of images with Exif tag tungsten | 94,81 | 102,12 | 93,79 | 88,51 |
| Average value of images with flash on | 100,91 | 112,45 | 99,81 | 90,46 |
| Average value of images with flash off | 113,92 | 117,73 | 115,45 | 108,58 |

The average values were determined by simply averaging all values for all channels or the values of a selected channel.

These values show that the gray world is not really gray. From the fact that 68 % of the images taken are portrait images, a slightly higher red level is something that can be expected. It is also interesting that the tungsten images are significantly darker although they are not underexposed.

### 3.2 How many images have a dominant color?

| **How many images have a dominant color?** | Images | Percent |
|---|---|---|
| Number of evaluated images | 3.258.560 | |
| Red | 404.820 | 12,4 |
| Orange | 238.755 | 7,3 |
| Yellow | 98.563 | 3,0 |
| Green | 14.889 | 0,5 |
| Cyan | 46.221 | 1,4 |
| Blue | 30.852 | 0,9 |
| Magenta | 8.588 | 0,3 |
| No dominant color | 2.415.872 | 74,1 |

For determining color dominance it was necessary to define simple analysis methods based on digital values without color conversions. The definitions stated below were made empirically by the authors.

Images with a dominant color are defined as images where more than 60% of the pixels fulfill the following requirements:

Red: green and blue have the same digital value level (+/- 20%) and the red value is more than 20% higher than green.

Orange: the red value is more than 20% higher than green and the blue value is more than 20% lower than green.

Yellow: green and red have the same digital value level (+/- 20%), green is greater than 119 and the blue value is more than 20% lower than green.

Green: the blue and red values are more than 20% lower than green.

Cyan: green and blue have the same digital value level (+/- 20%) and the red value is more than 20% lower than green.

Blue: the green and red values are more than 20% lower than blue.

Magenta: the blue and red values are more than 20% higher than green.

### 3.3 How often do images contain colors at the border of the sRGB color space?

| **How often do images contain colors at the border of the sRGB color space?** | Images | Percent |
|---|---:|---:|
| Number of evaluated images | 3.336.683 | |
| Images that contain colors at the border of the sRGB color space | 2.934.623 | 88,0 |
| Overexposed images | 42.441 | 1,3 |
| Underexposed images | 32.507 | 1,0 |

Border images: more than 20 pixels with one or two channels at digital value of 255.

Overexposed: > 20% white pixel (three channels at 255).

Underexposed: no more than 100 pixel with luminance value > 180 with luminance L = 0,3red + 0,6green + 0,1blue.

## 4. VISUAL CHARACTERIZATION OF IMAGES

### 4.1 Sorting the images by typical scene types.

| | | |
|---|---:|---:|
| Images used to categorize | 10.000 | |
| Group portraits | 3.628 | 36,3 |
| Children | 1.688 | 16,9 |
| Single Portraits | 1.510 | 15,1 |
| Landscape | 553 | 5,5 |
| Architecture | 541 | 5,4 |
| Urban areas | 340 | 3,4 |
| Animals | 328 | 3,3 |
| Plants | 309 | 3,1 |
| Sports | 186 | 1,9 |
| Indoor | 151 | 1,5 |
| Food | 81 | 0,8 |
| Night images | 28 | 0,3 |
| Others<br>thereof: 156 signs, 142 boats, 107 cars | 657 | 6,6 |
| | | |
| Overexposed (visual impression) | 552 | 5,5 |
| Underexposed (visual impression) | 98 | 1,0 |

Each image was sorted into one single category (except the exposure check). The portraits were counted as such no matter if they were shot indoor or outdoor. This explains the relatively low number of Indoor shots among the total number of images.

## 5. CONCLUSION

The collected data lead to some results that are important to know. The grey world assumption for example may be true for a selection of images that represent numerous different scene types. But since 68% of the images are portrait images the average value is shifted to the red what is also represented in the color dominance evaluation in clause 3.2.

From our experience we see that most cameras can handle typical scenes at exposure values of 8 and higher. But the 10% shot at lower light levels and those images that might have been shot without flash at a higher ISO level if the cameras would perform better make the difference between a standard and a good camera.

For future analysis it might be interesting to look into some of the detail e.g the exposure level for images without flash and those shot at tungsten light etc.

Of course the automatic and semi automatic image statistics can not replace other methods especially the psychophysical analysis for image quality evaluation but it provides an important piece of information that helps to improve the quality of future cameras.

## ACKNOWLEDGMENT

The authors would especially like to thank the CeWe Color team and Carina Matzky for their support. Special thanks also to Henning Hoefel for his help in collecting the data.

## REFERENCES

[1] JEITA, *Exif 2.2*, http://www.exif.org/specifications.html.
[2] DIN 19010 Lichtelektrische Belichtungsmesser, Beuth Verlag, Berlin