Statistic analysis of millions of digital photos 2017

Dietmar Wueller[1], Reiner Fageth[2]; [1] Image Engineering, Kerpen, Germany, [2] CEWE Stiftung & Co. KGaA, Oldenburg, Germany

Abstract

For cellphone and camera manufacturers, App developers and test labs it is very helpful to know how cameras are used. Under which lighting conditions are images captured? What is captured? Was the flash used? Many more things can be analyzed by looking at the image content and the metadata. The last analysis dates 9 years back to a time when smart phones with integrated cameras started to take off. Since then a lot of the usage conditions of image capture devices have changed. People don't use the flashes anymore, they take a lot more pictures of food than before and a lot of the images are selfies that did not exist. Knowing how a camera is used can help prioritize certain developments, allows to optimize algorithms for image processing and can help archiving images in better ways.

Introduction

Ideally the statistic analysis would be performed directly on the captured images that are stored on the cellphones. Getting access to these images directly is difficult and an issue of data security. Therefore this study was made utilizing images that have been sent to CEWE, a big photo finishing company in Europe. These images are uploaded to be printed in photo books, on calendars or other materials and devices people buy in customized versions with their own images printed on them. For this reason the images cannot be viewed as representative for all images captured but they can be viewed as representative of the images that European people care about. There maybe cultural differences involved in what images people care about. Therefore we need to limit the statement to the source region.

CEWE has a database of app. 1 billion images at a time and a fraction of the stored images (about 1%) were extracted and the metadata was analyzed. In addition approximately 50.000 images were extracted on a random basis and visually inspected and categorized. The images in the database may be the original

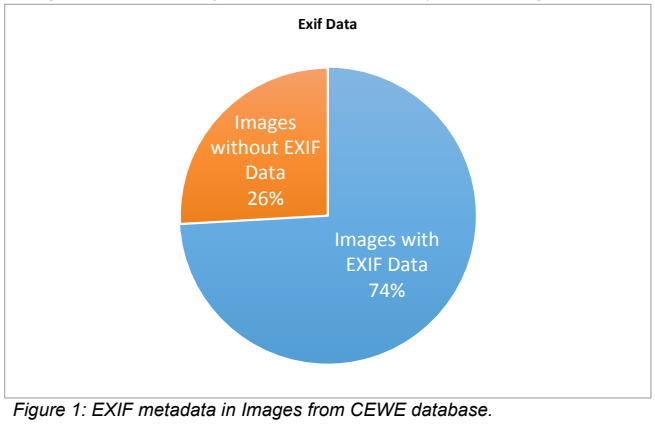

Figure 1: EXIF metadata in Images from CEWE database.

images captured by cameras but some may have also been photoshopped, down sampled, or processed otherwise. In some cases processing destroys the metadata and therefore not all images contain metadata. In our case 74% of the 11 million images contained EXIF [1] metadata.

Data Sources

One of the first questions arising is the source of the images. What type of camera do they originate from?

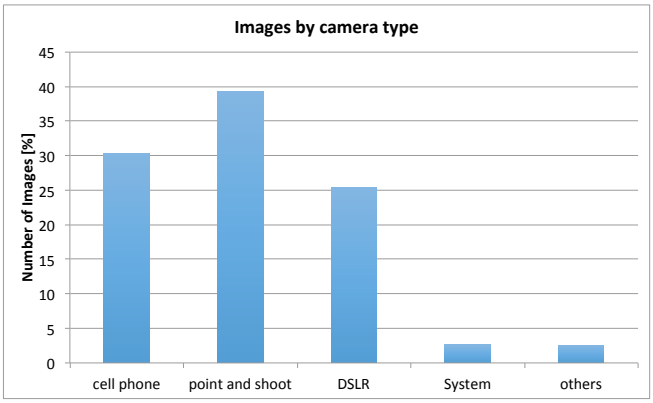

Figure 2: Sources for the images from the database.

It is very interesting to see that still about 40% of the images are captured with point and shoot cameras even though the market for those has gotten pretty small [2]. In Germany the sales of point and shoot cameras has gone down from 6 million in 2012 to 2 million in 2016 and this trend is the same worldwide. Only 30 % of the images in the database originate from cell phones. The authors expected this number to be significantly higher. In contrast to that almost 25% are captured with DSLRs. This indicates that people still shoot a lot of images that are important to them with dedicated cameras. The assumption is that a lot of images people order in printed form are captured during vacations or other official meetings like weddings, birthdays, etc. where people carry dedicated cameras instead of simply using their cell phones. This is a question that may need to be answered in the future.

Interestingly only 2.5% of the images are coming from system cameras (mirrorless with interchangeable lenses). This alternative to DSLR cameras has been on the market for quite some years now and still they don't play a big role in this study.

One of the next questions arising is the pixel count in the images. An important aspect could be the sensor class of the camera used but since people often times can select various pixel counts in the camera settings the sensor class can only be determined in combination with an additional database of all cameras of the past 20 years and their sensors and linking the camera make and model with the CEWE database. This was not available for this study. We only have the pixel counts of the

images in the database. These pixel counts are the ones the users uploaded to the database. They are not altered by the CEWE ordering software and we still wonder why the counts are so low. Have the users set their cameras to the low values? Have they been transferred between devices and down sampled by the software (e.g. WhatsApp does that)? We currently don't have the answer but the strong feeling that the counts are far from what the sensor classes in the cameras are.

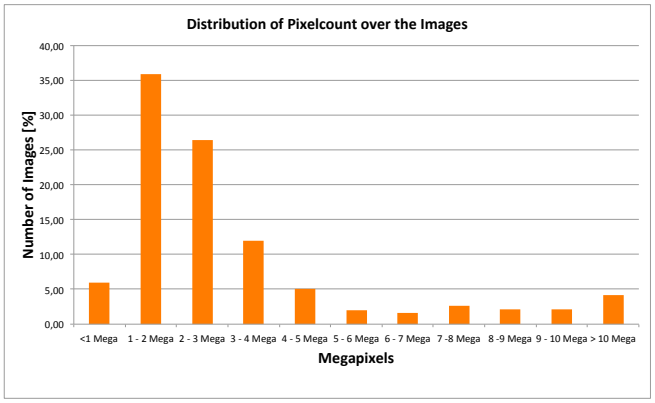

Figure 3: The distribution of images by pixel count.

Light sources

Determining the scene illumination is one of the big challenges for cameras because they do not know if a color cast in the scene originates from a dominant color or the illumination of the scene. Various techniques are used to determine the illumination type in order to get a reliable auto white balance setting for each capture [3]. Our previous study from 2008 [4] already shows that the EXIF tag for the light source is not used often and not used well due to a lag of standardization in this area. Unfortunately this has not changed until now. Still almost 90% of the images do not contain any information about the estimated illuminant.

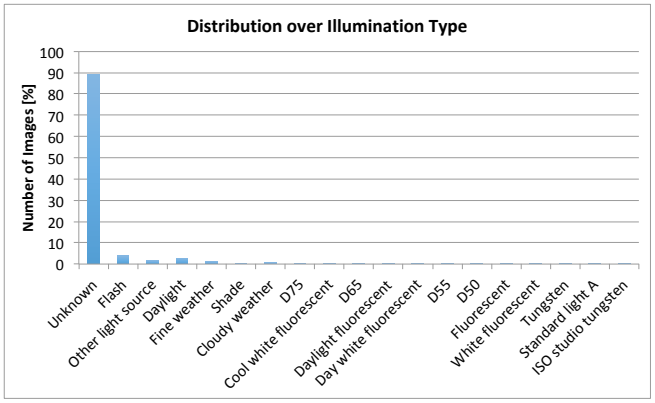

Figure 4: Light source information in the EXIF tags.

Among the remaining images we find a huge variety how manufacturers name the sources. When we remove the images with no information from the count we find that the most used light source is the flash. Given the evaluation for the use of flash later in this document this cannot be true. So we find that in case a flash gets used this tag is filled out more often. In other cases the remaining data does not lead to reliable results about the distribution of illuminants over the captured scenes.

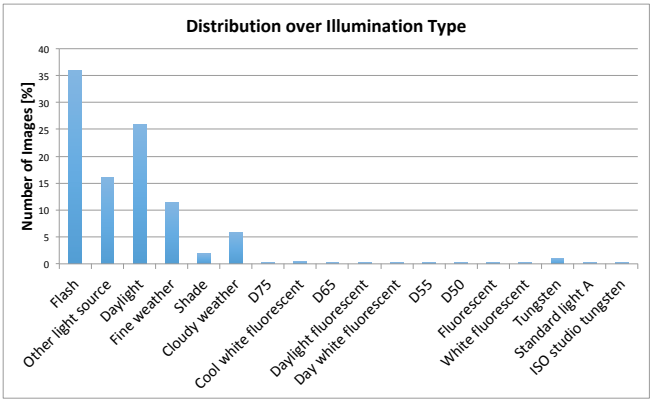

Figure 5: The distribution of illuminants after removing the unknown ones.

The color analysis performed in the previous study [4] has not been repeated because we did not expect a change in this aspect. That experiment verified the grey world assumption and the distribution of dominant colors.

Light levels

Besides the illuminant type or better the spectral distribution of the light source the illumination levels the cameras are used under are very important. This aspect shows the most drastic change compared to the study from 2008. The illumination level under which an image was captured can be derived from the Exposure Value (EV). The EV is calculated from f-stop, Exposure Time and ISO sensitivity setting. It is directly connected with the scene luminance. Assuming an average 18% reflection (grey card) of a standard scene the scene luminance can be translated into scene illumination. Utilizing this a scene illumination of 5 lux corresponds to an EV =1. Every increase of the EV by one doubles the amount of illumination in the scene.

EV	-4	-3	-2	-1	0	1	2	3
Illumination [Lux]	0,16	0,31	0,63	1,25	2,5	5	10	20

EV	4	5	6	7	8	9	10
Illumination [Lux]	40	80	160	320	640	1280	2560

EV	11	12	13	14	15	16
Illumination [Lux]	5120	10240	20480	40960	81920	163840

The scene illumination reflected here is the one that includes all light sources used. This means if a flash or in case of cell phones an LED is used to illuminate the scene in addition to the available light this contributes to the EV.

To find out about how often a flash is used and in which way, a bunch of flash tags are provided. Summarizing the ones with

active flash and the ones with no flash we can see how often the flash is used.

In total 19% of the images are captured using a flash. Looking at the flash use versus camera type a huge variation is visible. With cameras of a newer type like cell phones and system cameras the flash is used a lot less than with point and shoot or DSLR cameras.

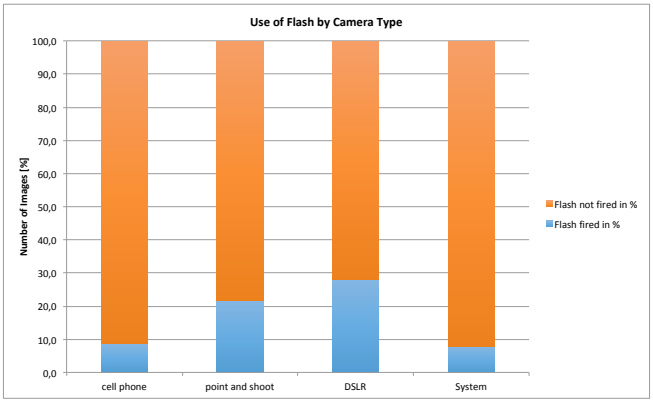

Figure 6: Light source information in the EXIF tags.

The question arises why that is the case? DSLRs and point and shoots have bigger sensors than cell phones, therefore they should need the flash less often. The current assumption is that the reason is a combination of two different things. On the one hand people who are long-term photographers have their style of taking pictures and use the flash more often for the historical reason that film did not have the high sensitivity levels and they still hesitate to use them. On the other hand, and this is clearly noticeable, the automatic functions in the conventional camera types activate the flash from illumination levels that are a lot higher than the ones where cell phones and "modern style" cameras start using flashes.

A similar thing can be seen from the distribution of captured images over the illumination levels. The orange columns are from our study in 2008 where 44% of the images were captured using flash. We can see that whenever the light level fell below 8 the flash was activated. Nowadays cameras are used without flashes even down to 5 lux of illumination. So the whole low light photography started in the recent years. Leading to an upcoming ISO standard to measure low light performance of cameras [5].

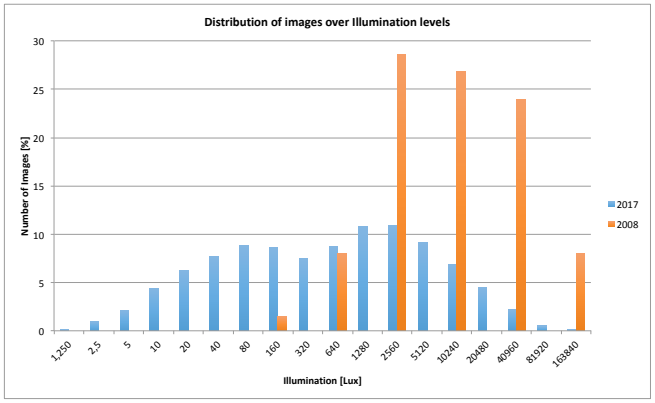

Figure 7: Captured images over the different light levels.

Somewhat expected is the distribution of exposure times and ISO speed values. 1/60 s is the most often used exposure time and ISO 100 is the most often used ISO sensitivity setting in the cameras. For both values the latest technology like electronic shutters have lead to intermediate values besides the typical 1/3 f-stop steps known from traditional photography.

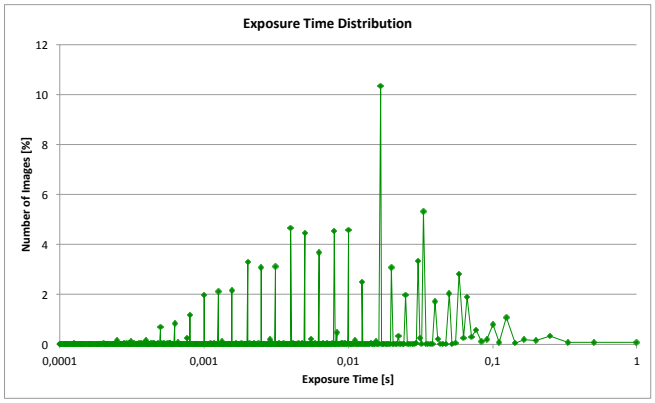

Figure 8: Distribution of exposure times.

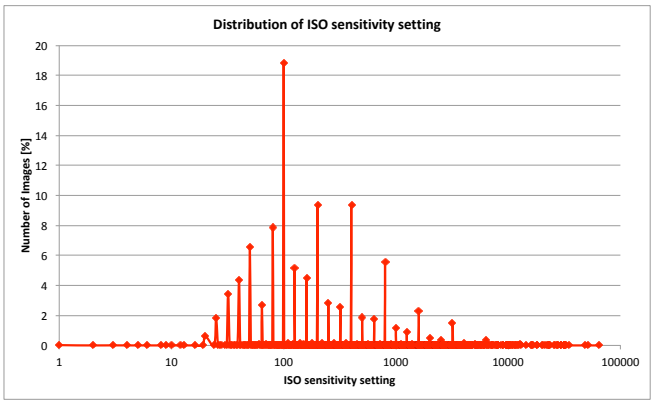

Figure 9: Distribution of ISO sensitivity settings.

Scene Types

Different scenes and objects that are photographed may be processed in different ways. For example a portrait requires an accurate focus on the persons face and nice looking skin tones. A sunset scene should not be perfectly white balanced because this would destroy the atmosphere. The importance of these factors depend on how often a scene of that specific kind is captured. To determine these numbers a set of 355 photo books was taken and manually evaluated.

The person doing that received the instructions that each image should be assigned to one category of the ones mentioned in figure 10. In case an image could be sorted under 2 categories (e.g. a group picture in front of a landscape) the picture should be sorted to the first category on the list where the categories were ordered from left to right as shown in figure 10.

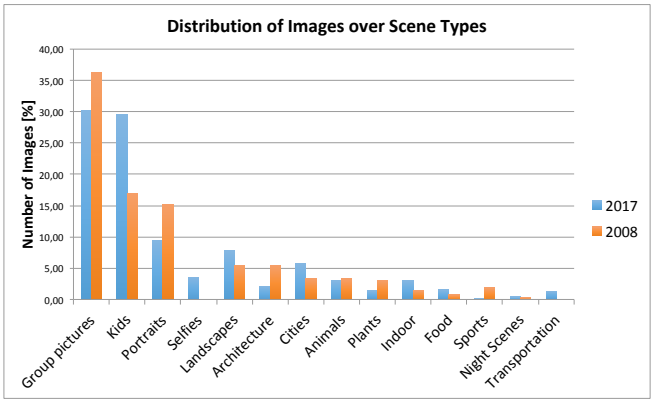

Figure 9: Distribution of scene types over 50,000 images.

Even though people photographs have shifted towards Kids in 2017 the total amount around 70% is about on the same level as 2008. This means that face detection and skin tone adjustment are of utmost importance to photography. What can be seen is that indoor and food photography have increased whereas plants and sports have decreased.

Geographic Information

Last but not least in the current study 9.4% of the images contained GPS data. In 2008 only 0.6% of the images did contain that information.

References

[1] EXIF Metadata Standard Version 2.31, CIPA DC- 008-Translation-2016 Exchangeable image file format for digital still cameras, http://www.cipa.jp/std/documents/e/DC-008-Translation-2016-E.pdf

[2] Foto- und Imagingmarkt 2016 in Deutschland, Photoindustrie-Verband, https://www.piv-imaging.com/imagingmarkt/Foto-und-Imagingmarkt-2016-in-Deutschland-auf-stabilem-Niveau-10010999

[3] Edmund Y. Lam and George S. K. Fung, Automatic White Balancing in Digital Photography, https://pdfs.semanticscholar.org/8fd8/4de7676e347634bdc6c997c72e05132a3755.pdf

[4] Dietmar Wueller, Reiner Fageth, Statistic Analysis of Millions of Digital Photos, Proc. SPIE 6817, Digital Photography IV, March 2008

[5] ISO 19093 Photography — Digital cameras — Measuring low light performance

Author Biography

Dietmar Wueller studied photographic technology at the Cologne University of applied sciences. He is the founder of Image Engineering, an independent test lab that tests cameras for several photographic and computer magazines as well as for manufacturers. Over the past 20 years the company has also developed to one of the world's leading suppliers of test equipment. Dietmar Wueller is the German chair of the DIN standardization committee for photographic equipment and also active in ISO, the IEEE CPIQ (Cellphone Image Quality) group, and other standardization activities.