Assessing Image Quality on Import From Camera

K. Armin Samii
Computer Science Undergraduate
Univ. of Calif. at Santa Cruz
ksamii@ucsc.edu

Allison Carlisle
Bioinformatics/Computer Science Undergraduate
Univ. of Calif. at Santa Cruz
acarlisl@ucsc.edu

Uliana Popov
Computer Science Graduate
Univ. of Calif. at Santa Cruz
uliana@soe.ucsc.edu

James Davis
Computer Science Professor
Univ. of Calif. at Santa Cruz
davis@soe.ucsc.edu

February 20, 2011

Abstract

In this paper, we propose a novel method for ranking images as they are imported from a users camera. By gathering sets of similar images, we rank images in relation to others in the set and provide the user with the best images from every scene photographed. Our goal is to filter out images which are technically flawed and provide the photographer with a high-quality representative subset of the imported images. The filtered set chosen by our algorithm matches the user's choices with high accuracy (XX%).

1 Introduction

To focus our research and determine which image qualities are relevant, we assume three general steps a photographer takes between shooting and using a picture, which we call the Photographer's Process:

1. Remove: Sort through the images being imported and select the ones most suited for retouching, flag them, and remove the rest.

2. Retouch: Modify the raw files just selected to stylize and enhance them.

3. Retrieve: Select the retouched images most suitable to a given task.

Our research focuses on the first step of the Photographer's Process. By quantifying qualities which cause a photograph to be rejected unanimously, we have created an automated system which rejects over XX% of the images a human would reject, based on XX data sets of XX images tested on XX users each. We have created a scale for rating each image's quality on a scale from one to ten. Over XX% of our ratings were within one of the average user ranking, which was tested on the same dataset of XX images and XX different users.

2 Related Work

Previous work in quality assessment explores single features such as blur[10] and color harmony[4][5]. Others applied the assessment techniques to different types of images- compressed, iconic [9] [23], [22], [1].

[20], [8], [12], and [15] classified images by whether they were taken by professionals or by home users.

[3] derived various classification of photographs through computations using EXIF data rather than pixel data.

[14] focused on exploring visual attention and sensitivity, and so [19] , and [16].

[18] assess quality of images and video using statistical models in an information-theoretic setting, and derive a novel QA algorithm.

Various publications have focused on the second step of the Photographer's Process, optimizing quality through retouching[2][13]. More research has focused on assessing quality for the purposes of photograph retrieval[24][21][11][7][6]. The methods for image retrieval are not as useful for image removal because they assume a finalized photograph, rather than looking at the photograph's potential to be enhanced.

Various approaches have been proposed to rank any set of images based on aesthetic[...refs] or technical quality[...refs] but these do not make use of the reference-based algorithms which are possible when working with images directly from a camera import. While some works assess both technical and aesthetic qualities[...refs], we believe the two are separate problems. Aesthetics are largely preference-based and an automated algorithm may not be trusted by a photographer, whereas technical quality, although dependent on properties of the Human Visual System, is more readily automated without consideration of personal preference (and thus no human input is required).

Therefore, our work assesses an image based on blur, noise, exposure, and the relationship between colors. Past research has quantified these artifacts globally[...refs] and independently (no-reference algorithms) [...refs]. Our approach combines local-feature algorithms [...refs] and reference-based algorithms [...refs] to extract similar foreground content in near-duplicate photographs and compare the foreground data to find a relative ranking. By focusing on the foreground, we reduce the computational errors that may arise when looking globally. For example, background blur and underexposure are not signs of a poor photograph[...ref], but may be mistaken as such[...refs, somebody who had this mistake].

Similarly, by separating foreground and background content, we can better measure exposure balance and weight noise levels which closer match the Human Visual System (HVS)[...refs, the foreground-background noise level paper] and assess exposure levels based on foreground-background histogram comparison, similar to[...refs?].

3 Quantifying Image Quality

Each selected factor will provide a ranking between zero and nine, with larger numbers indicating higher quality. Because an image which ranks low on any part would be considered poor, we have propose an algorithm which penalizes low scores more than it rewards high scores, because our primarily goal is to weed out poor images. We use an inverse-logarithmic scale for this overall ranking:

$$\sum_{i=1}^{n} \frac{\log^{-1}(W_i Q_i)}{rn}$$

Where r is the range of rankings (9 in our work), Q_i represents the rating of module i, and W_i represents the weight of that module. Weights are assigned as follows:

- $W_{exposure} = 40\%$
- $W_{blur} = 30\%$
- $W_{noise} = 15\%$
- $W_{color} = 10\%$
- $W_{gray} = 5\%$

3.1 Content Recognition

We obtain a bounding box around the most salient foreground object, in a method similar to [#]. Rather than using the precision that [#] uses, we aim for accuracy, and observed faster and equally accurate results with this method. We achieve high accuracy because of the simplicity, though secondary objects are sometimes ignored or mistaken as foreground. We have found that this does not heavily affect the outcome of our algorithm.

3.2 Similar-Image Clustering

To find similar images, we perform three steps of increasing complexity for high confidence. First, we use the timestamp to obtain a similarity index between all pairs of images. Based on the algorithm explained in [17], which finds gaps in timestamps, we have derived a formula which finds temporal closeness between two images:

$$T_{i,j} = \frac{G}{A_g}$$

Where G is the log of the time gap between the two images:

$$G_{i,j} = \log(|T_i - T_j|)$$

And A_g is the average gap within a window:

$$A_g = \sum_{a=i-w}^{j+w} G_{a,i}$$

We have empirically chosen a window size of $w = 8$. We use the calculated similarity index $T_{i,j}$ as a weight for the next two steps.

The first content-based step is fast 9-segment test similar to [...refs]. We divide the images into a 3x3 grid and compare the average color of each square. Because this is sensitive to exposure differences and movement, we gather interest points using SOME ALGORITHM FROM NASA. We then compute a ranking based on the number of matched interest points.

The two content-based algorithms, weighted by the timestamp, clusters similar images with high accuracy.

3.3 Blur Detection

Using the bounding box, the blur-detection algorithm quantifies the contrast between the edges and background. It ignores the rest of the image. In a method similar to [...refs], it obtains a value for the sharpness. This algorithm does not perform well when ranking diverse images, due to differences in scale and levels of acceptable background blur, so constraining it to the bounding box increases accuracy. Furthermore, we compare these rankings to the near-duplicates to remove any photograph-specific artifacts which we may not have accounted for. This increases the accuracy of the algorithm with just one photographs, and does even better when there is a sequence of similar images.

3.4 Noise Detection

We use a binary-weight scheme to weigh the noise, in a weak combination of local and global quality assessment. Noise within the bounding box has more negative weight than noise in the background. Our algorithm is similar to [...refs], which has results good enough to support itself.

3.5 Exposure

Because local-vs-global exposure methods are more accurate [...refs], we compare the foreground and background exposure to determine quality. Based on the idea that the the mean value of image luminosity is an indication of how well-exposed an image is[...refs], images are divided into different categories of further analysis through mean luminosity ranges. (ALLISON: I DON'T GET THIS, WHAT ARE THE DIFFERENT CATEGORIES OF FURTHER ANALYSIS?) Luminosity is a measure of how bright the human visual system perceives a color. As the human eye is most sensitive to green and red ranges, those color portions are weighted more heavily to a color's overall perceived intensity. The calculation used by the exposure analyst is

$$p = .59r + .3g + .11b$$

where p is the perceived intensity, and r, g, and b are the red, green, and blue components, respectively[...refs].

The categories of mean luminosity value (henceforth referred to as the mean) correspond roughly to a parabolic mapping of exposure values, with the best images typically falling between 130 and 140, and the further towards the extreme high and low means, the worse quality the image is. These divisions are used as the starting point of analysis. Within each category, different measures will indicate either a positive or negative overall impact on image quality. But the measurements can signify different changes in the various categories. For example, in an overall dark mean, having more extreme bright areas makes the image more balanced, while in a bright image they usually indicate that it has been overexposed.

The other measures of image quality are also based on luminosity, and were determined by taking various measurements on a deliberately chosen pool of images that had a wide variety of exposure problems. The measures are as follows: clipping, highlights, lowlights, the upper 60th and 98th percentiles, the lower 60th and 98th percentiles, and variance. Each of these measures are calculated on both the foreground and the background of the image.

Clipping is an indication of how much information loss there is in the image due to extreme shadows and bright spots. Each of these measures is taken in relation to the total number of pixels being measured.

Overexposure of an image can lead to large areas in the photo in which the human eye cannot discern any form (note that this definition can also indicate areas of a solid white value that contains absolutely no data about the form present). However, as the program is primarily concerned about how humans perceive images, the number of pixels in the highest five perceived values are used to calculate the highlights present in the image. Some highlighting can be desirable in an image, but the amount that is too much varies depending on the image mean.

Lowlights can be caused by underexposure, but they can also be a product of poor lighting (eg, photographing a shadowed object in extremely bright conditions). Again, some lowlights are desirable for contrast in an image, but too much leads to an undecipherable image.

The percentiles of the image are determined for both the bright and the dark sides of luminosity. For example, to calculate the lower 60th percentile, the value is found at which the sum of the number of pixels that are darker than the given value is equal to sixty percent of the pixels in the image. The upper percentiles are found by the summation of the pixels that are brighter than the given value. These four measures give a good indication of the spread of the luminosity, eg, how sharp the transitions between the extreme and middle values are.

The most extreme mean luminosity values are images of very poor quality, and are rated as very poor exposure quality, based solely on the mean value, although having a bright 98th high percentile can provide enough contrast to make a discernible image, but certainly nothing of quality. Means between 80 and 170 encompass nearly all images of medium through excellent exposure, and thus require the most analysis. Means below 100 are also part of the low quality batch of images, they have a fairly clear relationship with the 98th high percentile and are thus analyzed very simply.

Means between 100 and 120 are further subdivided by the amount of highlighting present. WRITE SOME MORE STUFF

Means between 120 and 138 were found to have the highest concentration of high ratings. WRITE SOME MORE STUFF

The images on the most extreme end of the bright scale were not very populated in the training image set, but were hypothesized to follow a similar curve as the dark side, with the redeeming quality being the presence of dark values.

3.6 Color Harmony

The "ColorCritic" module is inspired by the types of harmonies based on the Color Harmonization paper by Cohen-Or, Sorkine, Gal, Leyvand, and Xu /citeCohen-Or:2006:CH:1179352.1141933 Their program shifts the colors in a photo to be part of one of the seven types of harmony, but needs a user to direct the placement and type of the harmony template. The ColorCritic uses the aforementioned color templates as a way of measuring what types of harmony are present in an image.

ColorCritic uses the I,v,L,I,Y, and X harmonies (T and N type harmonies were omitted), for each harmony type, the hue is found with the most occurrences of supporting harmony (other pixels that fall in the harmonic range) as well as the percent of the image that has harmonious pixels. Similar to the Exposure Analyst, ColorCritic divides it's images into several categories before further analysis. However, because people tend to like colorful, intense images, ColorCritic divides the images by the color saturation in the image using the results from Mechanical Turk rankings to determine where the exact divisions are. The Harmony types are used to differentiate between various qualities of images withing the saturation ranges.

The first Harmony checked is the average and great image's I type. If there is a large amount of I harmony, the difference between I and Y harmonies is checked. Because I and Y overlap, you need to check how much of an increase if between the two. If it is very significant, Y harmony is most likely. When the difference is slight, you have a good I harmony (and thus high rating). If there is a medium difference, the I harmony is weak. Rating of the images can be determined on medium to great harmony. A similar check is done in X versus Y harmony. A last pass is done to check the L type on the great images, and a pass checking the v and i harmonies is done on the average images.

The poor images are less well-defined but, the X harmony was the best place to start, and then a pass was done with i.

4 Results

We have run our algorithm on several publicly available datasets as well as our own. To determine the success of the ImageSorter, we compared our results with a ground truth from studies that were run through Amazon Mechanical Turk.

Separate studies were run for several sets of previously ranked images (Ke, Tang, and Jing's data set of 48 images as well as Luo and Xiaoou's data set), as well as our own set of test images (459 images). A random sampling of 100 images from Luo et. al. was ranked on Turk, rather than the entire set. For each raking, a minimum of 10 valid Turk responses were used (presence of an invalid input was used to determine the validity of Turk responses) on each image to generate the ground truth ranking. Turkers were asked to rank each image on a scale of 1 to 10 to rate image quality. Ke's rankings were scaled linearly to fit on a 1 to 10 scale.

Enumerated below are the results and comparisons to similar works.

4.1 Binary Classification

With a dataset of 4,000 users obtained through Amazon Mechanical Turk, we have found that our algorithm can distinguish between professional and non-professional images well. Barsky [24] and Luo *et. al*[15] obtained 93% accuracy when classifying images into the two categories. When classifying "non-professional" as ratings below five and "professional" as ratings above five (ratings of exactly five are ignored), our rating system matches a user's with XX% accuracy.

4.2 Quality Filtering

When asking a user to choose which images to keep and which to discard, our algorithm correctly discarded 82% of the images, and incorrectly discarded 18%. When limiting both the algorithm and the user to discarding a fixed number of images (ten), we improved these numbers to 92% accuracy with 6% false-negatives. These results rank favorably with [...ref].

4.3 Image Ranking

Because this was not the goal of our work, we cannot compare to the state of the art[...ref] which received XX% accuracy, but we did manage a no-reference accuracy of within 30% of the user's rating. Further, when we introduced a single set of twenty near-duplicates to a user and asked them to rank from best to worst, their top three and bottom three matched our results with 97% accuracy.

4.4 Comparison with Ke[12]

When comparing Ke's data set to our ground truth, we see that their ground truth is based on different factors. Whereas our work matches Amazon Mechanical Turk users' ratings with an accuracy of 80%, Ke's only matches it with an accuracy of 45%. (Here, accuracy is defined as being within one standard deviation of the users' rankings.) CHARTS AND GRAPHS TO FINALLY MAKE IT CLEAR.

5 Conclusion and Future Work

We propose a method of integrating current research to apply it to the field of photography. By ranking an image's quality in relation to other images in the set, rather than an absolute scale, we obtain an accurate sorting of images. We improve upon past research which provides an absolute ranking by applying it to a relative scale. Images are sorted both relatively within the set of similar images and globally within the set of all input images.

Our work focuses on a small portion of the Photographer's Process. We would like to see our idea of relative rankings applied to the second and third steps (retouching and retrieving). Retouching can be made relative by looking at each subset and

References

[1] Tamara Lee Berg and David Forsyth. Automatic ranking of iconic images. Technical Report UCB/EECS-2007-13, EECS Department, University of California, Berkeley, Jan 2007.

[2] Subhabrata Bhattacahrya, Rahul Sukthankar, and Mubarak Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the international conference on Multimedia*, MM '10, pages 271–280, New York, NY, USA, 2010. ACM.

[3] M. Boutell and J. Luo. Bayesian fusion of camera metadata cues in semantic scene classification. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–623 – II–630 Vol.2, 2004.

[4] Kenneth E. Burchett. Color harmony attributes. *Color Research & Application*, 16(4):275–278, 1991.

[5] Kenneth E. Burchett. Color harmony. *Color Research & Application*, 27(1):28–31, 2002.

[6] Wei-Ta Chu and Chia-Hung Lin. Consumer photo management and browsing facilitated by near-duplicate detection with feature filtering. *Journal of Visual Communication and Image Representation*, 21(3):256 – 268, 2010.

[7] Jingyu Cui, Fang Wen, and Xiaoou Tang. Real time google and live image search re-ranking. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 729–732, New York, NY, USA, 2008. ACM.

[8] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Wang. Studying aesthetics in photographic images using a computational approach. In Ale Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision ECCV 2006*, volume 3953 of *Lecture Notes in Computer Science*, pages 288–301. Springer Berlin / Heidelberg, 2006. 10.1007/11744078_23.

[9] A.M. Eskicioglu and P.S. Fisher. Image quality measures and their performance. *Communications, IEEE Transactions on*, 43(12):2959 –2965, December 1995.

[10] Ping Hsu and Bing-Yu Chen. Blurred image detection and classification. In Shinichi Satoh, Frank Nack, and Minoru Etoh, editors, *Advances in Multimedia Modeling*, volume 4903 of *Lecture Notes in Computer Science*, pages 277–286. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-77409-9_26.

[11] Shao-Hang Kao, Wei-Yen Day, and Pu-Jen Cheng. An aesthetic-based approach to re-ranking web images. In Pu-Jen Cheng, Min-Yen Kan, Wai Lam, and Preslav Nakov, editors, *Information Retrieval Technology*, volume 6458 of *Lecture Notes in Computer Science*, pages 610–623. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-17187-1_57.

[12] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality

assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 419 – 426, June 2006.

[13] Johannes Kopf, Boris Neubert, Billy Chen, Michael Cohen, Daniel Cohen-Or, Oliver Deussen, Matt Uyttendaele, and Dani Lischinski. Deep photo: model-based photograph enhancement and viewing. *ACM Trans. Graph.*, 27:116:1–116:10, December 2008.

[14] Zhongkang Lu, W. Lin, X. Yang, EePing Ong, and Susu Yao. Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. *Image Processing, IEEE Transactions on*, 14(11):1928 –1942, November 2005.

[15] Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision ECCV 2008*, volume 5304 of *Lecture Notes in Computer Science*, pages 386–399. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-88690-7_29.

[16] Vitaly Pimenov. Fast image matching with visual attention and surf descriptors. In *Proceedings of the 19th International Conference on Computer Graphics and Vision*, pages 49–56, 2009.

[17] J.C. Platt, M. Czerwinski, and B.A. Field. Phototoc: automatic clustering for browsing personal photographs. In *Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 1, pages 6 – 10 Vol.1, December 2003.

[18] H.R. Sheikh, A.C. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *Image Processing, IEEE Transactions on*, 14(12):2117 –2128, December 2005.

[19] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, and Shaohui Liu. Photo assessment based on computational visual attention model. In *Proceedings of the seventeen ACM international conference on Multimedia*, MM '09, pages 541–544, New York, NY, USA, 2009. ACM.

[20] Hanghang Tong, Mingjing Li, Hong-Jiang Zhang, Jingrui He, and Changshui Zhang. Classification of digital photos taken by photographers or home users. In Kiyoharu Aizawa, Yuichi Nakamura, and Shinichi Satoh, editors, *Advances in Multimedia Information Processing - PCM 2004*, volume 3331 of *Lecture Notes in Computer Science*, pages 198–205. Springer Berlin / Heidelberg, 2005. 10.1007/978-3-540-30541-5_25.

[21] Roelof van Zwol, Börkur Sigurbjornsson, Ramu Adapala, Lluis Garcia Pueyo, Abhinav Katiyar, Kaushal Kurapati, Mridul Muralidharan, Sudar Muthu, Vanessa Murdock, Polly Ng, Anand Ramani, Anuj Sahai, Sriram Thiru Sathish, Hari Vasudev, and Upendra Vuyyuru. Faceted exploration of image search results. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 961–970, New York, NY, USA, 2010. ACM.

[22] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600 –612, April 2004.

[23] Zhou Wang, H.R. Sheikh, and A.C. Bovik. No-reference perceptual quality assessment of jpeg compressed images. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–477 – I–480 vol.1, 2002.

[24] Che-Hua Yeh, Yuan-Chen Ho, Brian A. Barsky, and Ming Ouhyoung. Personalized photograph ranking and selection system. In *Proceedings of the international conference on Multimedia*, MM '10, pages 211–220, New York, NY, USA, 2010. ACM.