

Quantifying Perceptual Image Quality

D. Amnon Silverstein and Joyce E. Farrell
Imaging Technology Department, Hewlett Packard Laboratories
Palo Alto, California

Abstract

This paper describes a more efficient paired comparison method that reduces the number of trials necessary for converting a table of paired comparisons into scaler data. Instead of comparing every pair of samples (the complete method), a partial method is used that makes more comparisons between closer samples than between more distant samples. A sorting algorithm is used to efficiently order the samples with paired comparisons, and each comparison is recorded. When the sorting is completed, more trials will have been conducted between closer samples than between distant samples. A regression is used to scale the resulting comparison matrix into a one dimensional perceptual quality estimate.

Introduction

To quantify subjective image quality, experimenters usually rely on one of several methods. In a direct method, the subjects are required to quantify their subjective impression of quality (with a number or graphic scale, for example) and (with some assumptions) this data can be averaged between observers. These methods suffer from several drawbacks, an overview of which can be found in Riskey[1]. One of the major problems is that this metric is unit-less. The subjective scaling depends on many factors, and when a sample is scaled in one experiment, it will almost always have a different value then when it is scaled in another experiment.

Another method relies on threshold judgments. These methods use the assumption that image fidelity is the same as image quality. Quality is then reduced to the detectability of differences between an original image and a distorted image. One way to quantify the detectability is by measuring the percent of subjects who can correctly identify which of two images has been distorted. When image quality varies monotonically with some adjustable parameter, the parameter can be adjusted so the distortion is at the threshold of visibility. This method has several problems when used to evaluate image quality. First, the threshold for detecting a distortion does not generally predict the perceived image quality and threshold measurements can not be extrapolated to predict super-threshold quality[2]. A second problem is that it is not always possible to adjust the distortions to threshold level. For example, in hardcopy images, if we wanted to compare the quality of images from two printers, the quality of the samples can not typically be adjusted to be at a threshold level.

The method of pair-wise comparisons generates reliable and informative data about the relative quality of two images. Two image samples are compared to each other by several subjects. The percentage of the time one sample is preferred over the other is used as an index of the relative quality of the two samples. The disadvantage of this method is that it requires many comparisons, typically 10 or so for every pair of samples.

One of the major advantages of this method is that the data can be converted to scaler data (with some additional assumptions, which will be discussed). Under the scaling assumptions however, not every pair of comparisons yields equally useful data. The efficiency of the method of paired comparisons can be improved by carefully selecting a subset of the pairs for comparison.

Thurstone's Law of Comparative Judgements.

Consider a set of test samples that are judged against one another pair-wise, across a set of subjects, so that an N X N matrix C of times-preferred is compiled, where N is the number of samples. Each element $C_{i,j}$ represents the number of times sample i was judged to have higher quality than sample j. When subjects do not agree on which sample was better, there is said to be "confusion" between the two samples.

Thurstone's case V method of comparative judgements[3] can be applied to determine the relative qualities of all of the samples if:

1. Each sample has a single value that can describe its quality, q_i.
2. Each observer estimates the quality of this sample with a value from a normal distribution around this actual quality.
3. Each sample has the same perceptual variance.
4. Each comparison is independent.

If these assumptions are valid, then the quality of each sample i can be described by a scaler value q_i with units of standard-deviations of preference. The distance between two samples $d'_{i,j}$ in units of the standard deviation can be estimated with the inverse cumulative-normal function (Z-score).

1. $$d'_{i,j} = q_i - q_j \approx \sqrt{2}\, Z\left(\frac{C_{i,j}}{C_{i,j} + C_{j,i}}\right)$$

One Dimensional Scaling

If each sample is compared against every other sample a sufficient number of times, then we can use a method to determine an estimate of each samples quality value. Based on our first assumption, all of the samples can be positioned on a one dimensional quality line. We can estimate the distance of a sample to the mean of all samples by taking the mean distance between the sample and all other samples[4].

2. $$d'_{i,mean} \approx \frac{\sum d'_{i,j}}{N}$$

Unfortunately, this approach suffers from several problems. The certainty of the distance estimates varies inversely with their magnitudes. When the distance is large and there are not enough subjects, there may be unanimous agreement between all of the subjects. If this happens, then the distance is estimated as infinite, so the mean can not be computed.

One solution is to use a weighted average, based on the certainty of the distance estimates[5]. The infinite distance estimates can be forced to have finite distances by assuming that the actual distance was the minimum distance possible such that there was a 50% chance that the subjects would have judged the samples unanimously. This will typically be an underestimate of the actual distance, and the underestimate is more severe when there are fewer trials.

In general, this method only provides approximately correct results when the number of trials in each comparison is large. The number of trials required to do each comparison goes up with the number of samples at the rate:

3. $$Trials = \frac{(N^2 - N)}{2}$$

When the number of samples is large, the distance between the extreme samples tends to get larger as well, so a very large number of trials is required to get accurate results.

The new method we describe here does not conduct trials between every pair of samples, and it conducts *fewer* trials between distant samples. The method described above needs more trials for these distant samples, in order to obtain accurate results.

An alternative approach is to use regression[6]. A set of estimated q values is adjusted to maximize the probability that a comparison matrix C could have randomly resulted from the experiment if the estimated values were the actual values. A regression is required when the comparison matrix is sparse.

The probability that a specific comparison matrix C would have been the result of an experiment where each pair of samples had an actual distance of $d'_{i,j}$ is:

4. $$P_{experiment} = \prod_{i,j} \begin{bmatrix} C_{i,j} + C_{j,i} \\ C_{i,j} \end{bmatrix} P_{i,j}{}^{C_{i,j}} (1 - P_{i,j})^{Cj,i}$$

where $P_{i,j}$ is the expected percentage of subjects to prefer sample i over sample j. $C_{i,j}$ is the number of subjects who preferred sample i over sample j, and $C_{j,i}$ is the number of subjects who preferred sample j over sample i. The first part of the equation inside of the product is the combination function that represents the number of different ways $C_{i,j}$ subjects can be chosen from the total population ($C_{i,j} + C_{j,i}$). The second part of the equation is the probability that any one sequence of preference decisions would be made. The chance of any specific comparison matrix occurring is the product of the probabilities of each element of that matrix occurring.

The expected percentage preference is a function of the d' distance between the two samples. It is the square root of two times the cumulative normal function.

5. $$d'_{i,j} = q_i - q_j$$

6. $$P_{i,j} = \sqrt{2} \int_{-\infty}^{d'_{i,j}} \frac{1}{\sqrt{2\pi}} \exp(\frac{-x^2}{2})$$

We conduct a search for the set of q values that maximizes the probability of obtaining the comparison matrix C. In practice, since the probability of an exact set of data originating from any set of q values is always so small as to be difficult to compute, it is useful to compute the log of $P_{experiment}$. We then minimize the negative log probability instead of maximizing equation 6. In other words, we use the negative log of equation 6 as the stress function.

A second difficulty in the minimization procedure is the problem with local minima. If approximate values for q are found, it may not be possible to find a better value for any individual q_i, but a better solution can exist if several values are adjusted at the same time. This makes it difficult for the regression tool we used to find a very good solution (we used the leastsq program from Matlab[7]).

Instead of regressing the absolute values of q, it is better to find the optimal values for the distances between the nearest q values. To do this, we used an iterative procedure. We first used the previously described Z-score averaging method to estimate the correct ordering. We then used leastsq to search for a set of distances between nearest neighbors that reduced the stress. We then used these improved distance estimates to re-order the samples, and iterated, until no further reduction in stress could be found. This method converges very quickly, and produces a better estimate of the q values than the averaging method. It also has the advantage that it does not have trouble with unanimous decisions.

Since a unanimous matrix entry has a finite chance of occurring from a given set of q values, we do not usually need a special procedure to deal with unanimous matrix entries. For example, consider 3 samples; A, B, and C. If the comparison between A and C is unanimous, the best fitting $d'_{a,c}$ would be infinite. But if there is some comparison between A and B, and between B and C, then equation 4 will only maximized when there is some finite distance between A and C.

A difficulty arises when there are two separate classes of samples that are never confused. For example, there may be confusions between A and B, but no confusions with either A and C or B and C. In this case, A and B form one class of samples, and C is a second class.

It is also possible to have overlapping classes. For example, all subjects may agree that A has lower quality than B and that B has lower quality than C, but there could be confusion between A and C. In this case, A and C form one class, and B forms a second class.

Overlapping classes could happen from random chance if very few trials were conducted. When more trials are conducted, the existence of overlapping classes would indicate a probable violation in one or more of the case V assumptions. In the case of partial methods (which will be described), few trials may have been conducted between A and B and between B and C, and many trials may have been conducted between A and C, and this could cause the case of overlapping classes.

In the case of non-overlapping classes, equation 4 will be maximized when the classes are infinitely far apart. In this case, not enough trials were conducted to estimate the distance between the two classes. If more trials can not be conducted, a 50% probability lower-bound on the distance may be computed by calculating the closest distance at which the unanimous decisions would have occurred half of the time. This can be approximated by switching ½ of a trial between the lowest sample in the higher class, and the highest sample in the lower class before applying the regression.

Using Sorting Methods for Paired Comparisons

The regression method allows us to find quality values for partial comparison matrices. A partial comparison matrix is one in which not every pair of samples is compared. The number of comparisons needed goes up very fast as the number of samples is increased, as can be seen in equation 3. Further, not all comparisons provide the same amount of information. Comparisons between very distant samples do not provide as accurate an estimate of distance as nearby samples. By strategically choosing the comparisons, a partial comparison matrix can be more efficient than a complete matrix.

For example, consider a set of samples A, B, C, and D spaced evenly 1 SD of quality apart. In the case of A and B, the samples are separated by a d' of 1. In this case, by equation 6 we would expect about 1 subject in 7 to misjudge the quality of the two, so a small number of trials could produce a reasonably good estimate of the distance. But for samples A and D which are separated by a d' of 3, we would only expect about 1 subject out of 20,000 to misjudge the samples. In this case, any practical number of trials would always give us the same unanimous result. After several hundred trials, we would only know that the distance apart was at least a d' of 2.

The heuristic we use here is to try to concentrate the number of trials on comparisons between closer samples. In the example, it is easy to see that we could obtain a good estimate of the distance between A B, B C, and C D with far fewer trials than it would take to estimate the distance A D directly. The distance between A and D can then be obtained based on the small distance measurements using the regression procedure previously described.

Therefore, it is seems useful to conduct more trials between closer samples. However, we do not know the quality distance between the samples until we have conducted the experiment so how can we conduct more trials between samples that are close? Previous methods have used an initial experiment to find the approximate distances, and then more trials were conducted between closer samples[5]. In our method, we estimate the distances as we conduct the experiment. This has the advantage that a single method is used to conduct all of the trials. Further, all of the previously obtained information is used for each new trial.

A paired comparison procedure can be used to implement a sorting of the samples. There are several efficient sorting algorithms based on comparing two elements at a time, that require $N \log_2 N$ rather than N^2 comparisons between samples. Each comparison can be recorded to form a partial comparison matrix. The advantage of doing this is that a sorting algorithm must include comparisons between nearest samples. This assures that there will be one test between each nearest set of samples, and fewer checks between more distant samples.

It is important to note that a method that used an $N \log_2 N$ sorting procedure would have fewer trials and thus less data than the N^2 complete method. Therefore, it could not be analyzed to provide as accurate an estimate of the original quality values. However, it would provide more information per trial than the complete method.

One way to compare samples while sorting is to use a binary tree sorting method. A binary tree is formed, with samples as the nodes. Each node of the tree is a partitioning element for a left sub-tree and a right sub-tree. The left sub-tree consists of nodes that were all judged to be lower in quality, and the right sub-tree consists of nodes that were judged to be higher in quality. To add a new sample to a tree, the new sample is compared to the root node. If there are no nodes in the tree, the sample is added as the root node. Otherwise, if the new sample is judged to be higher in quality, it is then added to the right sub-tree and if it is judged to be lower in quality, it is then added to the left sub-tree recursively. The samples can be added to the tree in a random order. As the samples are added, a comparison matrix is constructed.

To improve the efficiency of the sorting, it is useful to balance the tree after each comparison. This is done by rebuilding the tree so that it is as short as possible and has as few nodes at the bottom as is possible. Note that there may be many ways to build a tree that is as short as possible. Nodes that are higher in the tree have a greater chance of being tested. To reduce the number of separate classes (as described earlier), it is possible to construct a tree that maximizes the trials that could possibly link two classes. For example, in Figure 1 part 3, either F or C could be placed at the top to produce a balanced tree. If previous trials had left F in a separate class from the other samples, it would be better to place F at the top of the tree so it would be used in more trials.

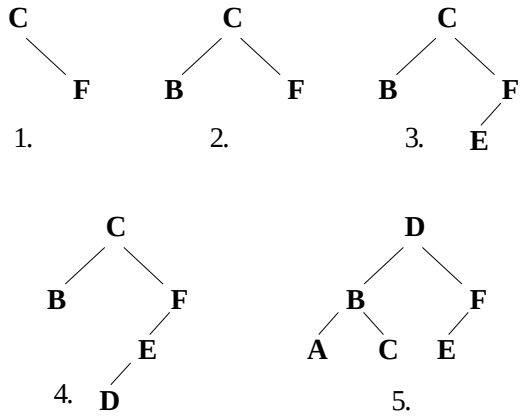

Figure 1 An example of a binary tree sorting

An example of a binary tree being used to sort samples A,B,C,D, E, F, with quality in that order.
1. The first sample, C, is chosen at random. The sample F is then compared to it, and judged as having higher quality.
2. Sample B is then compared to C, and since it is judged to have lower quality, it is not compared to F.
3. Sample E is compared to C and judged to have higher quality, and then to F and judged to have lower quality.
4. After D is added with the same procedure, the tree has become unbalanced.
5. The tree is balanced, and then A is added.

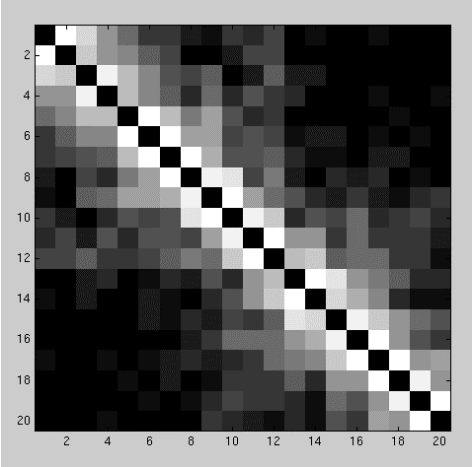

Figure 2 Density plot of the number of trials with the tree method

The figure shows the number of trials in the comparison matrix from a simulated experiment with 20 test samples and 10 subjects. Each square represents the number of times that pair of samples was compared, with white representing 10 comparisons and black representing no comparisons. The rows and columns are ordered in ascending quality units, which are known since the data is from a simulated experiment. As can be seen, many more trials are conducted between samples with similar quality (near the diagonal), than are between samples with very different quality (in the upper right and lower left corners). If the complete matrix method had been used, the figure would be solid white, since 10 trials would have been conducted between each pair. This simulation followed the same methods described in the following section, and this section contains more details about the simulation.

This process can be repeated for each subject. The probability that a set of samples will be compared is inversely related to their distance in units of d'. The end result is a comparison matrix with many more trials between close samples and few trials between distant samples. In this way, the method adapts to the distribution of the samples on the quality dimension, to conduct trials in the places that provide more information.

The tree method may be difficult to conduct with hardcopy samples, due to the complex order of presentation. In this case, alternative sorting methods may simplify the record keeping and presentation order. Some version of the Quick Sort or Heap Sort algorithms[8] may provide an efficient and easy to implement method.

Efficiency of the Tree Method

We can demonstrate the improvement in efficiency by means of a Monte Carlo simulation. We simulated an experiment where there were 20 samples. Each sample had a quality value q, that was chosen from a random even interval that spanned 40 standard deviations. All of the assumptions described in the first section were simulated. After the set of q values were chosen, two experiments were simulated.

To simulate an observer making a comparison between two samples q_i and q_j under the assumptions described in section 1, a normal random value with unit standard deviation was added to the two values, and the larger of the two was judged as having higher quality. After a set of simulated comparisons was made, the regression procedure previously described was used to estimate the original q values from the comparison matrix.

We used two methods to choose which comparisons should be made. In the first method, we used the complete matrix procedure. Every pair was compared an equal number of times. In the second method, we used the binary tree method described. The mean-squared error (MSE) between the estimate and the actual q value was then computed.

Figure 3 shows the MSE as a function of the number of trials for the two methods. The first point of the upper curve shows the MSE of the estimated quality after the complete matrix method was used with 5 trials between each pair (950 trials total). The first point in the lower curve shows the MSE with the binary tree method. The samples were sorted using the tree method 15 times (926 trials total). The Binary tree method reduced the MSE of this first point by a factor of 2 ± .3. The error bars in the figure show the estimated error in the shape of the curves (since the simulation was only run 10 times).

The MSE for the binary tree method is lower when the same number of trials are used. Likewise, for the same MSE, the binary tree method requires fewer trials. It can be seen that the binary tree method produced about the same MSE with 15 runs (926 trials) MSE = 2.2 ± .3, as the complete matrix did with 40 runs (7600 trials) MSE = 2.1 ± .3.

Violations of the Assumptions

The assumptions can fail in many instances. Quality can be multidimensional when there is no single quality line that can fit the data. The distribution of the perceived quality of samples might not form a normal distribution, or

might have different standard deviations for different samples. There may be a memory effect, where the judgements are not really independent.

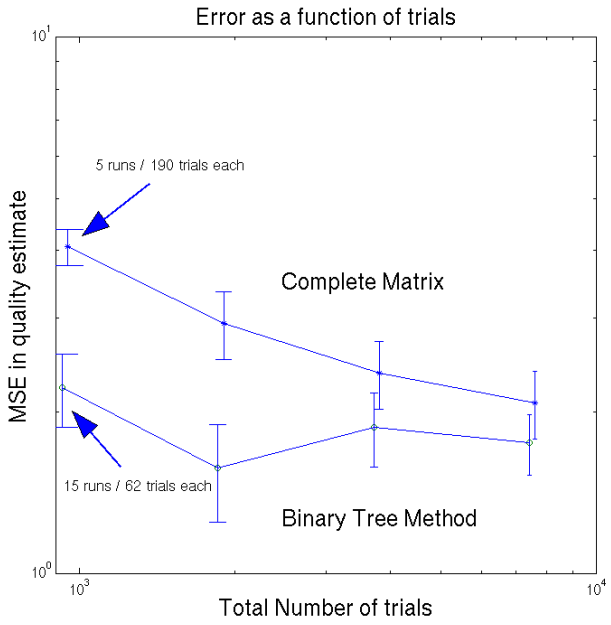

Figure 3 Error as a function of trials

This plot shows the result of running a simulated experiment to determine the quality values of 20 samples. The error bars in the figure show the estimated error in the shape of the curves (since the simulation was only run 10 times

Violations of the assumptions will result in a fit that has a larger than typical residual stress, which should be tested. One method that can be used to estimate the typical stress is to use a Monte Carlo simulation as above. The residual stress from the fit to real data is compared to the residual stress from a fit to modeled data[2]. If the actual data has significantly larger residual stress, one or more of the class V assumptions was probably violated.

References

1. D. R. Riskey, "Use and abuses of category scales in sensory measurement," *Journal of Sensory Studies* **1** 217-236 (1986)
2. D. A. Silverstein and J. E. Farrell "The relationship between image fidelity and image quality," *Proceedings of the IEEE International Conference on Image Processing*, Lausanne Switzerland, 881-884 1996
3. L.L. Thurstone, *The Measurement of Values*, University of Chicago Press, 1959
4. F. Mosteller "Remarks on the method of paired comparisons: III A test of significance when equal standard deviations and equal correlations are assumed," *Psychometrika* **16**, 207-218, 1951
5. C. J. Bartleson, *Optical Radiation Measurements volume 5: Visual Measurements*, Academic Press, Orlando, 1984
6. N. Draper and H. Smith *Applied Regression Analysis*, John Wiley and Sons, New York, 1981
7. Matlab: *The Language of Technical Computing*, The Math Works, 1997
8. W. H. Press, S. A. Teukolsky, W. T Vellerling, B. P. Flannery Numerical Recipes in C, Cambridge Press, 1992