Human perception of geometric distortions in images

I. Setyawan, R. L. Lagendijk
ICT Group, Department of Mediamatics, Delft University of Technology
Mekelweg 4, 2628CD Delft, The Netherlands
{i.setyawan, r.l.lagendijk}@EWI.tudelft.nl

ABSTRACT

We present in this paper the results of our study on the human perception of geometric distortions in images. The ultimate goal of this study is to devise an objective measurement scheme for geometric distortions in images, which should have a good correspondence to human perception of the distortions. The study is divided into two parts. The first part of the study is the design and implementation of a user-test to measure human perception of geometric distortions in images. The result of this test is then used as a basis to evaluate the performance of the second part of the study, namely the objective quality measurement scheme. Our experiment shows that our objective quality measurement has good correspondence to the result of the user test and performs much better than a PSNR measurement.

Keywords: Geometric distortions, human perception, perceptual test

1. INTRODUCTION

Research on human perception of image quality has been widely performed. Aspects of the image considered in such research are for example color, granularity or sharpness. Another example is to test specific artifacts of a compression algorithm (eg., the blocking artifact of JPEG compression) or watermarking system (eg., the random noise artifact of noise-based watermarking systems). Some examples of the image quality assessment for these distortions can be found for example in[1]. As a result, we already have good understanding of how these aspects influences human perception of quality and we are able to quantify these perceptual aspects in cases where the distortion is near the visibility threshold. We can use the result of this research to improve the performance of various applications dealing with images by designing the systems such that most changes or distortions to the images occur in the areas that have small perceptual impact for human observers. Compression algorithms and watermarking systems are two examples of applications that can take advantage of this knowledge. However, the research on human perception of image quality has not dealt with another type of distortion that an image can undergo, namely geometric distortion (ie., distortions due to geometric operations). As a result, we are currently unable to quantify the perceptual impact of geometric distortions on images.

This paper presents a study of the impact of geometric distortions to human perception of the quality of the distorted images. The goal of this study is to provide a reference point on which to evaluate the performance of our objective geometric distortion measure scheme[2]. The results we obtain from this test are also useful to other researchers performing similar research in this field. Therefore, we also make our test set and test results available for download on our website[3]. The rest of the paper is organized as follows. In Section 2, we present the design of our user test experiment and statistical analysis methods used to process the test results. In Section 3, we present the actual set up of our user test. In Section 4, we present and analyze the result obtained from this user test. In Section 5, we will briefly review our objective geometric distortion measure algorithm, present scores obtained using this method and evaluate its performance based on the subjective test result. Finally, in Section 6 we present our conclusions and provide a roadmap for further research.

2. TEST DESIGN & ANALYSIS METHOD

2.1. Test design

In order to evaluate the perceptual impact of geometric distortion, we have performed a subjective test involving a panel of users, who are asked to evaluate a test set comprising of an original image and various distorted versions of it. The test subjects evaluate one pair of images at a time, comparing 2 images and choosing the one they think is more distorted. This type of experiment is called the paired comparison test. There are two experiment designs for paired

Security, Steganography, and Watermarking of Multimedia Contents VI, edited by Edward J. Delp III, Ping W. Wong, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 5306 © 2004 SPIE and IS&T · 0277-786X/04/$15

comparison test, namely the *balanced* and *incomplete* designs[4,5]. In a balanced design, a test subject has to evaluate all possible comparison pairs taken from the test set. In the incomplete design, a test subject performs comparisons on a subset of the complete test set. The latter design is useful when the number of objects in the test set is very large. In our experiment, we used the balanced paired-comparison design. Our choice for this design is based on three factors. Firstly, the number of objects in our test set is not very large and a test subject can finish the test within a reasonable time frame (as a rule of thumb, we consider a test lasting 60 minutes or less to be reasonable). Secondly, by asking every test subject to evaluate all objects in the test set we will be able to get a more complete picture of the perceptual quality of the images in the test set. Finally, in this design we make sure that each test subject evaluates an identical test set. This makes it easier to evaluate and compare the performance of each test subject.

Let t be the number of objects in the test set. One test subject performing all possible comparison of 2 objects A_i and A_j from the test set, evaluating each pair once, will make $\binom{t}{2}$ paired comparisons in total. The result of the comparisons is usually presented in a $t \times t$ matrix. If ties are not allowed (ie., a test subject must cast his/her vote to one object of the pair), the matrix is also called a two-way preference matrix with entries containing 1's if the object was chosen and 0's otherwise. An example of such matrix for $t = 4$ is shown in Figure 1. Each entry $A_{i,j}$ of the matrix is interpreted as *object A_i is preferred to object A_j*. The indices i and j refer to the rows and columns of the matrix, respectively.

	A_1	A_2	A_3	A_4
A_1	×	1	1	0
A_2	0	×	1	1
A_3	0	0	×	0
A_4	1	0	1	×

Figure 1. An example of a preference matrix

Let a_i be the number of votes object A_i received during the test. In other words, $a_i = \sum_{j=1}^{t} A_{i,j}$, $i \neq j$. We call a_i the *score* of object A_i. It is easy to see that the total score for all objects and the average score among all objects are $\sum_{i=1}^{t} a_i = \frac{1}{2}t(t-1)$ and $\bar{a} = \frac{\sum_{i=1}^{t} a_i}{t} = \frac{1}{2}(t-1)$, respectively. We can extend this result to the case where we have n test subjects performing the paired comparison test. In this case, the test result can also be presented in a preference matrix similar to the one presented in Figure 1. However, each entry $A_{i,j}$ of this matrix now contains the number of test subjects who prefer object A_i to object A_j. If again we disallow ties, the values of $A_{i,j}$ will be integers ranging from 0 to n. We also note that in this case $A_{j,i} = n - A_{i,j}$. Finally, in this case the total and average scores are expressed as $\frac{1}{2}nt(t-1)$ and $\frac{1}{2}n(t-1)$, respectively.

2.2. Statistical analysis of the experiment

After performing paired comparison tests, we obtain a preference matrix for each test set. Now we have to perform an analysis of this test result. We have two main objectives for this analysis. In the first place, we want to obtain the overall ranking of the test objects. The second objective is to see the relative quality differences between the test objects, that is, whether object A_i is perceived to be either similar or very different in quality from object A_j. The analysis we perform on the data to achieve these objectives are the *coefficient of consistency*, the *coefficent of agreement*, and the *significance test on score differences*. Each of these analysis is discussed in the following sections.

2.2.1. Coefficient of consistency

A test subject is consistent when he/she, in evaluating 3 objects A_x, A_y and A_z from the test set, does not make a choice such that $A_x \rightarrow A_y \rightarrow A_z$ but $A_z \rightarrow A_x$. The arrows can be interpreted as "preferred to". Such condition is called a *circular*

triad. While circles involving more than 3 objects are also possible, any such circles can easily be broken up into two or more circular triads. The matrix in Figure 1 has one such triad, namely $A_1 \rightarrow A_2 \rightarrow A_4$ but $A_4 \rightarrow A_1$.

For smaller values of t, one can easily enumerate the circular triads encountered. For larger t, this task becomes very tedious. However, we can compute the number of circular triads, c, from the scores a_i using the following relation[4, 6]

$$c = \frac{t}{24}(t^2 - 1) - \frac{T}{2} \tag{1}$$

where

$$T = \sum_{i=1}^{t} (a_i - \bar{a})^2 \tag{2}$$

The number of circular triads c can be used to define a measure of consistency of the test subjects. There are different approaches to do this[4]. Kendall/Babington-Smith compared the number of circular triads found in the test to the maximum possible number of circular triads. The coefficient of consistence ζ is defined as follows

$$\zeta = 1 - \frac{24c}{t(t^2 - 1)}, \text{ if } t \text{ odd} \tag{3}$$

$$\zeta = 1 - \frac{24c}{t(t^2 - 4)}, \text{ if } t \text{ even} \tag{4}$$

There are no inconsistencies if, and only if, $\zeta = 1$. This number will move to zero as the number of circular triads, thus the inconsistencies, increases.

The coefficient of consistency can be used in the following ways. In the first place, we can use this coefficient to judge the quality of the test subject. Secondly, we can use this coefficient as an indication of the similarity of the test objects. If, on average, the test *subjects* are inconsistent (either for the whole data set or a subset thereof), we can conclude that the test *objects* being evaluated are very similar and thus it is difficult to make consistent judgement. Otherwise, if one particular test *subject* is inconsistent while the other test subjects are – on average – consistent, we may conclude that this particular subject is not performing well. If the consistency of this subject is significantly lower than average, we may consider removing the result obtained by this subject from further analysis.

2.2.2. Coefficient of agreement

Coefficient of agreement shows us the diversity of preferences among n test subjects. Complete agreement is reached when all n test subjects make identical choices during the test. From Section 2.1, we see that if every subject had made the same choice during the test, then half of the entries in the preference matrix will be equal to n, while the other half would be zero. Alternatively, in the worst case situation all entries will be equal to $n/2$ (if n is even) or $(n \pm 1)/2$ if n is odd.

It is obvious that the minimum number of test subjects, n, that we need in order to be able to measure agreement is 2. Each time 2 test subjects make the same decision regarding a pair of test objects A_i and A_j, we say that we have one agreement regarding this pair. In other words, we measure the agreement by counting the number of pairs of test subjects that make the same decision over each pairs of test objects. We do this by computing τ, defined as

$$\tau = \sum_{i=1}^{n} \sum_{j=1}^{n} \binom{A_{ij}}{2}, \qquad i \neq j \tag{5}$$

In Equation (6), $\binom{A_{ij}}{2}$ gives us the number of pairs of test subjects making the same choice regarding objects A_i and A_j.

Thus τ gives us the total number of agreements among n test subjects evaluating t objects. Obviously, when $A_{i,j} = 1$ we do not have any agreement among the subjects and the contribution of this particular $A_{i,j}$ to τ would be zero. If $A_{i,j} = 0$, it means that all test subjects agree *not* to choose A_i over A_j. Although the contribution of this $A_{i,j}$ to τ is also zero, the number of agreements regarding this pair of test objects will be reflected by the value of $A_{j,i}$.

We have $\binom{t}{2}$ pairs of comparisons and $\binom{n}{2}$ possible pairs of subjects, therefore the maximum and minimum number of agreements between the subjects are given by $\max(\tau) = \binom{t}{2}\binom{n}{2}$ and $\min(\tau) = \binom{t}{2}\binom{\lfloor n/2 \rfloor}{2}$, respectively.

We can also express τ in a more computationally convenient way, as follows.

$$\tau = \frac{1}{2}\left[\sum_{i \neq j} \alpha_{i,j}^2 - n\binom{t}{2}\right] \tag{6}$$

Kendall/Babington-Smith[6] defines the coefficient of agreement, u, as follows

$$u = \frac{2\tau}{\max(\tau)} - 1 = \frac{2\tau}{\binom{t}{2}\binom{n}{2}} - 1 \tag{7}$$

The value of $u = 1$ if and only if there is a complete agreement among the test subjects, and decrease when there is less agreement among the test subjects. The minimum value of u is $-1/(n-1)$ if n even or $-1/n$ if n odd. The lowest possible value of u is -1 which can only be achieved when n is 2. In this case we have the strongest form of disagreement between the test subjects, namely that the test subjects completely contradict each other.

We can perform a hypothesis test to test the significance of the value u. The null hypothesis is that all test subjects casted their preference completely at random. The alternative hypothesis is that the value of u is greater than what one would expect if the choices would have been made completely at random. To test the significance of u we use the following statistic, as proposed in[4]

$$X^2 = \frac{4}{n-2}\left[\tau - \frac{1}{2}\binom{t}{2}\binom{n}{2}\frac{(n-3)}{(n-2)}\right] \tag{8}$$

which has χ^2 distribution with $\binom{t}{2}\frac{n(n-1)}{(n-2)^2}$ degrees of freedom.

As n increases the expressions in Equation (10) reduces to a simpler form[7]

$$X^2 = \binom{t}{2}[1 + u(n-1)] \tag{9}$$

with $\binom{t}{2}$ degrees of freedom.

It is important to note that consistency and agreement are 2 different concepts. Therefore, a high u value does not necessarily imply the absence of inconsistencies and vice versa.

The coefficient of agreement also shows whether the test objects, on average, received equal preference from the test subjects. If the overall coefficient of agreement is very low we can expect that the score of each test object will be very close to the average scores of all test objects, i.e., there is no significant difference among the scores. As a consequence, assigning ranks to the objects or drawing conclusion that one object is better (or worse) than the others is pointless since the observed score differences (if any) cannot be used to support the conclusion. On the other hand, strong agreement among the test subjects indicate that there exist significant differences among the scores.

2.2.3. Significance test of the score difference

Significance test of the score difference is performed in order to see whether the perceptual quality of any 2 objects from the test set is perceived as different. In other words, the perceptual quality of object A_i is declared to be different from the quality of object A_j, only if a_i is significantly different from a_j. Otherwise, we have to conclude that the test subjects consider the perceptual quality of the 2 objects is similar.

This problem is equivalent to the problem of dividing the set of scores $S = \{a_1, a_2, \ldots, a_t\}$ into sub-groups such that the variance-normalized *range* (the difference of the largest and lowest values) of the scores within each group,

$$R = \frac{(a_{\max} - a_{\min})}{\sigma_{a_i}} \tag{10}$$

is lower or equal to a certain value $\lceil R_c \rceil$ (in other words, the difference of any 2 scores within the group must be lower or equal to $\lceil R_c \rceil$), which depends on the value of the significance level α. In other words, we want to find R_c such that the probability $P[R \geq R_c]$ is lower or equal to the significance level α. We declare the objects within each group to be not significantly different, while those from different groups are declared to be significantly different. By adjusting the value of α, we can adjust the size of the groups. This in turn controls the probability of false positives (declaring 2 objects to be significantly different when they are not) and false negatives. The larger the groups, the higher the probability of false negatives. On the other hand, the smaller the groups, the higher the probability of false positives.

The distribution of the range R is asymptotically the same as the distribution of variance-normalized range, W_t, of a set of normal random variables with variance = 1 and t samples[4]. Therefore, we can use the following relation to approximate $P[R \geq R_c]$

$$P[W_{t,\alpha} \geq \frac{2R_c - \frac{1}{2}}{\sqrt{nt}}] \tag{11}$$

In Equation (18), $W_{t,\alpha}$ is the value of the upper percentage point of W_t at significance point α. The values of $W_{t,\alpha}$ is tabulated in statistics books for example the one provided in[8]. The value of R is then set as $\lceil R_c \rceil$.

3. TEST PROCEDURE

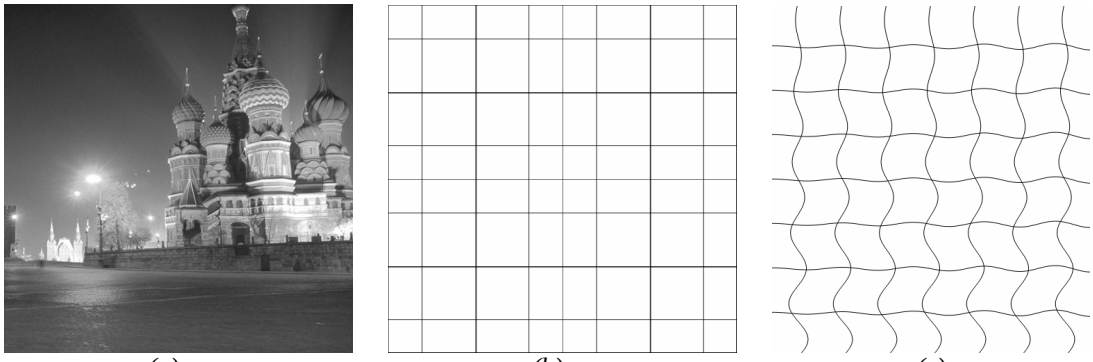
3.1. Test set

We used 2 images, Bird (see Figure 3) and Kremlin (see Figure 2(a)), as basis to build the test set for our experiment. These images are 8-bit grayscale bitmap images with 512 × 512 pixels resolution. The images are chosen primarily due to their content. The Bird image does not have much structures such as straight lines. Furthermore, not every test subject is very familiar with the shape of a bird (in particular the species of bird depicted in the image). So in this case, a subject should have little (if any) "mental picture" of how things should look like. On the other hand, the Kremlin image has a lot of structures and even though a test subject may not be familiar with the Kremlin, he/she should have some prior knowledge of how buildings should look like.

We used 17 different versions of the images. Each version is geometrically distorted in a different way. Thus in our test we have t = 17. The geometric distortions used in the experiment are shown in Table 1. In this table we use the notation A_i, with i = 1, 2, … 17, to identify each image.

The distortions chosen for the test set range from distortions that are perceptually not disturbing to the distortions that are easily visible. The global bending distortions (A_6, A_7, A_8, A_9) are chosen because these kind of distortions are, up to some extent, visually not very disturbing in natural images. However, this distortion severely affects the PSNR value of the distorted images. The sinusoid (stretch-shrink) distortions (A_{10}, A_{11}, A_{12}, A_{13}) distort the image by locally stretching and shrinking the image. Depending on the image content, this kind of distortion may not be perceptually disturbing. The rest of the distortions distorts the image by shifting the pixels to the left/right or upwards/downwards. These distortions are easily visible, even when the severity is low. The distortions (A_2, A_3, A_4, A_5) applies the same distortion severity over the whole image, while the severity of distortions (A_{14}, A_{15}, A_{16}, A_{17}) are varied within the image. Some examples of the geometric distortions used in the experiment are shown in Figures 2(b) and 2(c).

We then proceed to make all possible comparison pairs out of the 17 images, including the comparison of an image with itself. In each pair, we designate the first image as the left image and the other as the right image. This refers to how the images are to be presented to the subjects (see Figure 3). We then repeat each pair once, with the left-right ordering of the images reversed. Thus we have 306 pairs of images for each of the two images for a total of 612 pairs of images in the test set.



(a) *(b)* *(c)*

Figure 2. (a) The Kremlin image, (b) Distortion A_{13} and (c) Distortion A_{16}

Table 1. Geometric distortions used in the experiment

Image	Description
A_1	No distortion (original image)
A_2	Sinusoid, amplitude factor = 0.2, 5 periods
A_3	Sinusoid, amplitude factor = 0.2, 10 periods
A_4	Sinusoid, amplitude factor = 0.5, 5 periods
A_5	Sinusoid, amplitude factor = 0.5, 10 periods
A_6	Global bending, bending factor = 0.8
A_7	Global bending, bending factor = - 0.8
A_8	Global bending, bending factor = 3
A_9	Global bending, bending factor = -3
A_{10}	Sinusoid (stretch-shrink), scaling factor 1, 0.5 period
A_{11}	Sinusoid (stretch-shrink), scaling factor 1, 1 period
A_{12}	Sinusoid (stretch-shrink), scaling factor 3, 0.5 period
A_{13}	Sinusoid (stretch-shrink), scaling factor 3, 1 period
A_{14}	Sinusoid (increasing freq), amplitude factor = 0.2, starting period = 1, freq increase factor = 4
A_{15}	Sinusoid (increasing freq), amplitude factor = 0.2, starting period = 1, freq increase factor = 9
A_{16}	Sinusoid (increasing amplitude), start amplitude factor = 0.1, 5 periods, amplitude increase factor = 4
A_{17}	Sinusoid (increasing amplitude), start amplitude factor = 0.1, 5 periods, amplitude increase factor = 9

3.2. Test subjects

The user test experiment involved 16 subjects, consisting of 12 male (IL, ON, PD, AH, ES, DS, IS, JO, JK, JJ, KK and RH) and 4 female (KC, CL, CE and ID) subjects. The subjects have different backgrounds and levels of familiarity with the field of digital image processing. As discussed in Section 3.1., each user will examine each pair of the test images twice in one test session. Furthermore, subjects IL, DS and IS each performs 3 test sessions. Therefore, in the tables found in Section 4, a number will be added to the subject names to show different test sessions (eg., IL1 shows the result of subject IL from the 1st test, etc.). These repetitions are done to see the difference of test results for one person when the test is repeated. We assume that each repetition of the test (both within a single test session and between test sessions) are independent, thus we have $n = 44$.

3.3. Test procedure

The test is performed on a PC with a 19-inch flatscreen CRT monitor. The resolution is set at 1152 × 864 pixels. The vertical refresh rate of the monitor is set at 75 Hz. To perform the test, we use a graphical user interface as shown in Figure 3. The test subject is then asked to choose which image from the pair is according to them more distorted.

Figure 3. The user interface used in the experiment, showing the Bird image

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	a_i
A_1	×	7	3	0	0	11	21	19	8	8	24	2	10	9	1	0	0	123
A_2	37	×	4	0	0	33	28	29	24	30	36	10	11	12	5	1	1	261
A_3	41	40	×	3	1	42	43	39	39	41	42	25	18	37	9	5	0	425
A_4	44	44	41	×	3	44	44	42	43	43	43	37	39	43	31	15	1	557
A_5	44	44	43	41	×	44	44	44	43	43	44	42	43	44	43	42	24	672
A_6	33	11	2	0	0	×	25	15	17	15	33	4	11	8	1	0	0	175
A_7	23	16	1	0	0	19	×	15	13	21	28	3	11	5	2	0	0	157
A_8	25	15	5	2	0	29	29	×	12	17	27	6	10	9	1	1	0	188
A_9	36	20	5	1	1	27	31	32	×	30	40	8	15	15	2	2	0	265
A_{10}	36	14	3	1	1	29	23	27	14	×	34	6	9	9	0	0	0	206
A_{11}	20	8	2	1	0	11	16	17	4	10	×	4	5	6	1	0	0	105
A_{12}	42	34	19	7	2	40	41	38	36	38	40	×	20	31	9	5	1	403
A_{13}	34	33	26	5	1	33	33	34	29	35	39	24	×	25	17	5	0	373
A_{14}	35	32	7	1	0	36	39	35	29	35	38	13	19	×	6	1	0	326
A_{15}	43	39	35	13	1	43	42	43	42	44	43	35	27	38	×	7	2	497
A_{16}	44	43	39	29	2	44	44	43	42	44	44	39	39	43	37	×	1	577
A_{17}	44	43	44	43	20	44	44	44	44	44	44	43	44	44	42	43	×	674

Figure 4. Preference matrix for the Bird image

4. TEST RESULTS AND ANALYSIS

4.1. User preference matrix

After performing the user test, we obtain the preference matrices for the Bird and Kremlin images. In Figure 4, we show the preference matrix obtained for the Bird image. The preference matrix of the Kremlin image is available for download at our website[3]. The images codes refer to Table 1. The column a_i shows the sum of each row, ie., the score of

each image A_i. Since in our experiment the test subject is asked to choose the image with the *most* distortion, a smaller score a_i means that the image is perceptually better.

4.2. Statistical analysis of the preference matrices

4.2.1. Coefficient of consistency (ζ)

We measured the coefficient of consistency for individual test subjects using Equation (3) since we have $t = 17$. Since each test subject performs the user test twice per session, we use the average value of ζ as an indication of each subject's consistency. The average coefficient of consistency is presented in Table 2.

From Table 2 we can conclude that in general the test subjects are consistent in their decision. We can also see that in general the values of ζ for the Bird image is lower than that of the Kremlin image. This is due to the fact that the Kremlin image contains more structure compared to the Bird image, which helps the test subjects to make consistent decisions. Furthermore, the unfamiliarity of the test subjects to the particular species of bird depicted in the image also makes it more difficult to make consistent decisions.

Table 2. Coefficient of consistency (ζ)

Subject	Bird	Kremlin	Subject	Bird	Kremlin
IL1	0.83	0.93	DS1	0.67	0.87
IL2	0.83	0.91	DS2	0.73	0.92
IL3	0.85	0.95	DS3	0.82	0.93
KC	0.85	0.86	IS1	0.92	0.95
ON	0.94	0.98	IS2	0.94	0.93
PD	0.70	0.87	IS3	0.94	0.97
AH	0.87	0.96	JO	0.93	0.97
CL	0.82	0.90	JK	0.90	0.96
CE	0.83	0.94	JJ	0.85	0.88
ES	0.89	0.94	KK	0.70	0.79
ID	0.66	0.90	RH	0.90	0.95

4.2.2. Coefficient of agreement (*u*)

We measured two types of coefficient of agreements from the preference matrix. The first is the *overall* coefficient of agreement that measures the agreement among all test subjects in the experiment. The second is the *individual* coefficient of agreement, that measures the agreement of a test subject with him-/herself during the 2 repetitions in a test session. A low u value in this case would indicate that the subject is confused and does not have a clear preference of the images being shown.

For the calculation of the overall coefficient of agreement, we have $n = 44$ and $t = 17$. For these values, the maximum and minimum values of u are 1 and -0.0227, respectively. From the preference matrices, we can calculate that the overall coefficient of agreements are $u_{bird} = 0.574$ and $u_{kremlin} = 0.731$. Performing the significance test on both u values using the method described in Section 2.2.2 shows that in both cases, the probability of having a larger u values had the votes been casted at random is smaller than 0.001 (in other words, u is significant at $\alpha = 0.001$). Therefore, we can conclude that in both cases there are strong agreements among the test subjects. However, we can also see that the agreement in the case of the Bird image is much weaker than the Kremlin image, due to the image content.

For the individual coefficient of agreement, we have $n = 2$ and $t = 17$. In this case we have $-1 \leq u \leq 1$. The individual coefficient of agreements are presented in Table 3. As expected we see that all subjects have larger u values for the Kremlin image. The exceptions to this are subject ES, who has the same u values for both images and subjects IS2 and JK who have larger u for the Bird image. After performing the significance test on the values of u, we conclude that all subjects have u values that are significant at $\alpha = 0.05$ for both the Bird and Kremlin images. Therefore we can conclude that the users have clear preferences of the images in the test set.

Table 3. Individual Coefficient of Agreements(u)

Subject	Bird	Kremlin	Subject	Bird	Kremlin
IL1	0.559	0.750	DS1	0.265	0.647
IL2	0.574	0.721	DS2	0.471	0.794
IL3	0.677	0.779	DS3	0.559	0.750
KC	0.662	0.691	IS1	0.721	0.882
ON	0.779	0.868	IS2	0.809	0.721
PD	0.485	0.677	IS3	0.735	0.838
AH	0.559	0.794	JO	0.721	0.853
CL	0.456	0.750	JK	0.824	0.735
CE	0.618	0.691	JJ	0.529	0.691
ES	0.765	0.765	KK	0.368	0.515
ID	0.279	0.691	RH	0.691	0.765

4.2.3. Significance test of score differences

The strong agreements among the test subjects for both images, as shown in the previous section, show that there exist significant differences among the scores of the test objects. We use the procedure described in Section 2.2.4 to find the critical value for the score difference for the images, at significance level $\alpha = 0.05$. From[8] we have $W_{t,\ \alpha} = 4.89$. Substituting this value into Equation (11), we have $R_c = 67.12$ and thus we set $R = 68$. Therefore, only objects having a score difference of more than 68 are to be declared significantly different.

In Figure 5, we present the grouping of the images in the test set based on the significance of the score differences. The images have been sorted from left to right based on their scores, starting from the image with the smallest score (ie., perceived to have the highest quality) to the one with the largest score. The score for each image is shown directly under the image code. Images having score difference smaller than 68 are grouped together. This is represented the shaded boxes under the image code. For example, in Figure 5(a) images A_{14} and A_{13} belong to one group.

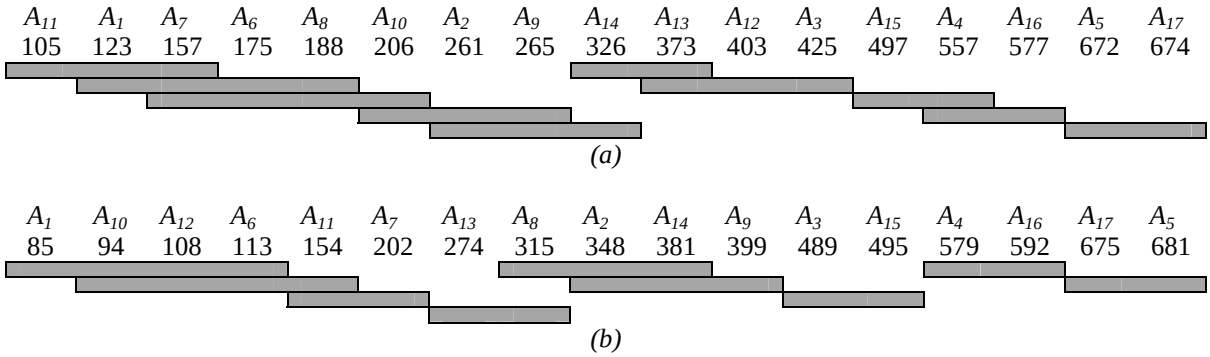


Figure 5. Score grouping for: (a) Bird image and (b) Kremlin image

From Figure 5, we can see that the images occupying the last 6 positions of the ranking for both the Bird and Kremlin images are distorted using the same distortion. Furthermore, they are sorted in the same order (except for images A_5 and A_{17}, but the difference between their scores is not significant). Thus we can conclude that these distortions are perceived similarly by the test subjects, regardless of the image content. These distortions occupy the "lower quality" segment of the ranking so we can also conclude that the distortions are so severe that the image content no longer plays a significant role. For the other images, the influence of image contents on the perceived quality of the distorted images are larger.

Table 4 shows the overall u values for each score group. We expect that when the images in a group do not have significantly different scores, there will not be any clear preference for any of them among the test subjects and therefore the u values should be low. The groups are presented in the 1st and 4th columns using their members as group names. The 3rd and 6th columns of the table show the result of the significance test for u, at significance level $\alpha = 0.05$.

Table 4. Group u values

Bird			Kremlin		
Group	*u*	Significant?	Group	*u*	Significant?
$A_{11}A_1A_7$	0.006	No	$A_1\,A_{10}\,A_{12}\,A_6$	0.008	No
$A_1A_7A_6A_8$	0.061	Yes	$A_{10}\,A_{12}\,A_6\,A_{11}$	0.03	Yes
$A_7A_6A_8A_{10}$	0.041	Yes	$A_{11}\,A_7$	0.011	No
$A_{10}A_2A_9$	0.07	Yes	$A_{13}\,A_8$	0.08	Yes
$A_2A_9A_{14}$	0.085	Yes	$A_8\,A_2\,A_{14}$	0.175	Yes
$A_{14}A_{13}$	-0.004	No	$A_2\,A_{14}\,A_9$	0.054	Yes
$A_{13}A_{12}A_3$	-0.003	No	$A_3\,A_{15}$	-0.021	No
$A_{15}A_4$	0.148	Yes	A_4A_{16}	0.112	Yes
A_4A_{16}	0.08	Yes	$A_{17}\,A_5$	0.011	No
A_5A_{17}	-0.015	No	-	-	-

We can conclude from Table 4 that the *u* values for each group is very low. Some groups even have *u* values that are not significantly larger than the *u* values that would have been achieved had the votes within that group had been casted at random. This results show that indeed the grouping of the images performed based on the significance of score differences has produced groups within which the perceived quality are difficult to distinguish.

4.3. Conclusions

From the analysis of the user test results, we can draw the following conclusions:

1. The test objects are generally perceptually distinguishable by the test subjects. This is supported by the fact that the consistency of the test subjects are relatively high as shown in Table 2. Furthermore, we also see that the individual *u* values (shown Table 3) are also high.
2. There is a general agreement as to the relative perceptual quality of the test images among the test subjects. This is supported by the high overall *u* values for both images. Therefore, we can make a ranking of the images based on their perceived quality.
3. For some images, the relative perceptual quality among them is not clearly distinguishable. We can see this from the grouping of the scores based on the significance test of score differences. This is further supported by the lack of agreement among test subjects regarding the relative quality of images within such groups.

5. OBJECTIVE GEOMETRIC DISTORTION MEASURE

5.1. Overview of the algorithm

The objective geometric distortion measurement is based on the ideas in our previously published work[9] and further developed and described in[2]. The algorithm is based on the hypothesis that the perceptual quality of a geometrically distorted image depends on the homogeneity of the geometric distortion. The less homogenous the geometric distortion, the lower the perceptual quality of the image will be. We proposed a method to measure this homogeneity by approximating the underlying geometric distortion using simple RST/affine approximation. We increase the locality of our approximation until the level of approximation error is lower than a predetermined threshold or until the locality of the approximation reaches a predetermined maximum. The locality is increased using quadtree partitioning of the image. We then determine the score (ie., the quality) of the image based on the resulting quadtree structure.

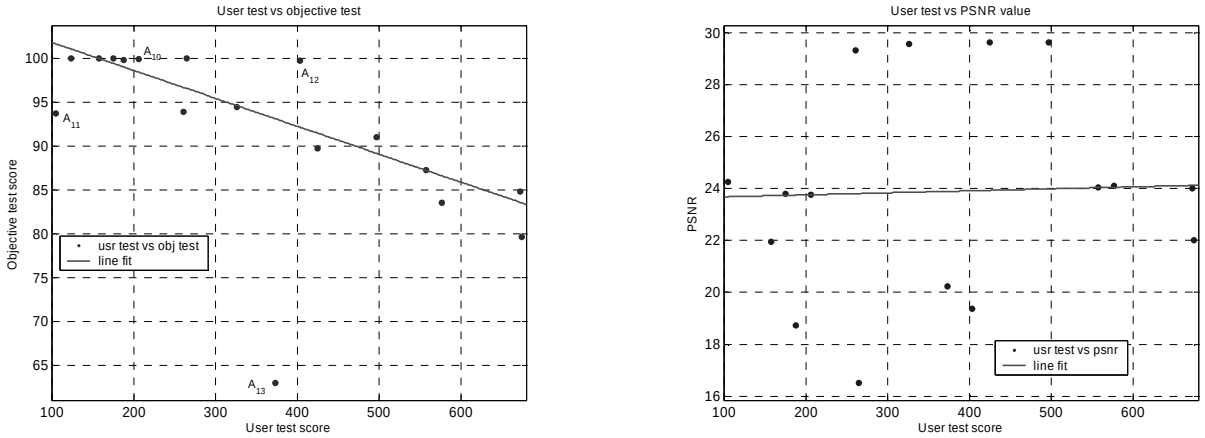
We have implemented some modifications to the algorithm to improve its performance. We briefly discuss the modifications as follows. The first modification is applied to the procedure used in the RST/affine parameter estimation. In the new scheme, we no longer use brute-force search to estimate the local RST/affine transformation parameters. Instead, we now base our scheme on the Optical Flow Estimation algorithm to estimate the RST/affine parameters, as presented in[10, 11]. Using this algorithm significantly improves the speed of the system and also increase the precision of the parameter estimation process, since previously we have to limit the precision with which we sample the range of the RST parameter in order to get a reasonable execution time. The second modification is applied to the scheme used in computing the final score of the distorted image. Previously, we only used the average block size of the quadtree partitioning and to some extent the residual error of the blocks to compute the final score. In the new scheme, we also take into account the estimated RST/affine parameters associated with each block to see how far it differs from the RST/affine parameters when there is no RST/affine distortion. The difference is expressed as the *l*-2 norm distance

between the 2 parameter sets. In the difference calculation, the parameters for Rotation and Scaling are given larger weights compared to the parameters for Translation. The the larger the difference, the lower the score for the block will be. This modification is performed to better fine-tune the performance of the measurement algorithm to better match the result of the subjective test results. This is because even when the RST/affine transformation of a block can be perfectly estimated (ie., zero residual error), such block can still heavily influence the overall perceptual quality of the image if the local RST/affine transformation is severe. In the objective test, the maximum score that can be achieved by an image is 100.

5.2. Performance evaluation of the algorithm

The performance of the objective quality measurement algorithm is evaluated by comparing it to the results of the subjective test and the results of a PSNR measurement. We evaluate whether the ranking produced by our algorithm corresponds well to the ranking produced by the subjective test or whether the ranking is actually more similar to the ranking produced by the PSNR measurement. We also take into account the grouping within the data set when performing this evaluation, thus we consider deviations in the ranking of the objective test (compared to the ranking obtained from the subjective test) to be acceptable as long as this occurs within groups of images whose perceptual quality can not be clearly distinguished. In the evaluation, we look at the *intra-* and *inter-distortion* comparisons. Intra-distortion comparison compares the scores within one type of distortion, but with different parameter sets. The inter-distortion comparison compares the scores of all distortions in the test set.

Our algorithm performs well in performing intra-distortion comparison. That is, images distorted with a more severe parameter set are given lower scores. This is also true for PSNR measurement. To evaluate the inter-distortion comparison performance, we plot the objective test score and PSNR values against the user test scores. The comparison plots for the Bird image are shown in Figure 6. The results for the Kremlin image show similar behaviour.



(a) *(b)*

Figure 6. Result comparisons for the Bird image:
(a) User test vs. Objective test and (b) user test vs. PSNR measurement

From Figure 6 we can conclude that the result of the objective test has a much better correspondence to the user test result compared to the PSNR measurement. In the latter case, the regression line is virtually horizontal. This is also reflected by the correlation coefficient ρ between each pair of data sets. For the Bird image, the value of the correlation coefficient ρ for the {*user test, objective test*} pair is ρ_{uo} = -0.6 while for the {*user test, psnr*} pair the value is ρ_{up} = 0.14. The negative value of ρ_{uo} correctly reflects the fact that larger user test score represents a lower perceptual quality. For the Kremlin image, the values are ρ_{uo} = -0.61 and ρ_{up} = 0.28, respectively.

We also see from Figure 6(a) that image A_{13} does not properly fit to the behaviour of the rest of the data and can be considered an outlier. Removing this image from the set, we get ρ_{uo} = -0.87. In general, we observe that our algorithm cannot handle images distorted by the *sinusoid (stretch-shrink)* distortion (see Table 1), except for image A_{10}. At present, we do not yet have a satisfactory explanation regarding this phenomenon. However, some explanation can be

provided for images A_{10} and A_{11}. The geometric transformation applied to image A_{10} is similar to the one implemented in television broadcasting when it is necessary to convert video frames from one aspect ratio to another. This transformation is perceptually not disturbing unless there is a lot of movement for example camera panning. Therefore, our test subjects give this image a high ranking. In this distortion, the middle part of the image is stretched slightly in the horizontal and vertical direction. The slight increase in image width and height is compensated by shrinking the outer parts of the image. This distortion can be approximated by slightly scaling the whole image. Therefore, our measurement system gives this image a high score. Image A_{11} is given a high ranking by the test subjects due to the unfamiliarity of the subjects to the bird species shown in the picture. Apparently, the test subjects thought that the size of the original bird's head is too large. Therefore, they prefer this image in which the head of the bird is shrinked. However, since this picture actually contains large distortion, our measurement system gives it a low score.

6. CONCLUSION AND FUTURE WORKS

In this paper, we have described the method we use to perform a perceptual user test for geometrically distorted images. We also described the statistical tools we use to analyze the results of the user test. The result of the user test is then used as a basis to validate our objective perceptual quality measurement scheme, which is based on the hypothesis that the perceptual quality of a distorted image depends on the homogeneity of the geometric transformation causing the distortion.

For intra-distortion comparisons, both the PSNR measurement and our measurement scheme work well. For inter-distortion comparisons, our measurement scheme outperforms PSNR measurement. Overall, our measurement scheme has very good correspondence to the subjective test for inter-distortion comparisons. However, detailed comparison between our measurement scheme and the subjective test still show some discrepancies of the ranking and score differences between some images.

In the future, more measurements and user test experiments similar to the one described and analyzed in this chapter should be performed. The data collected from such experiments can than be used to further validate or refine the hypothesis, and to further fine-tune the performance of the objective perceptual quality measurement system.

7. REFERENCES

1. B.W. Keelan, *Handbook of Image Quality: Characterization and Prediction*, Marcel Dekker, Inc., New York, 2002
2. I. Setyawan, D. Delannay, B. Macq and R.L. Lagendijk, *Perceptual Quality Evaluation of Geometrically Distorted Images using Relevant Geometric Transformation Modelling*, in the Proceedings of SPIE, Security and Watermarking of Multimedia Contents V, Vol. 5020, Santa Clara, USA, 2003, pp. 85 – 94
3. www-ict.ewi.tudelft.nl/~iwan/user_test_result.html.
4. H.A. David, *The Method of Paired Comparisons*, 2nd ed., Charles Griffin & Company, Ltd., London, 1988
5. R.E. Bechhofer, T.J. Santner and D.M. Goldsman, *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*, John Wiley & Sons, Ltd., New York, 1995
6. M.G. Kendall, *Rank Correlation Methods*, 4th ed., Charles Griffin & Company, Ltd., London, 1975
7. S. Siegel and N. J. Castellan, Jr., *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed., McGraw-Hill, Boston, 1988
8. E.S. Pearson and H.O. Hartley, *Biometrika Tables for Statisticians*, Vol 1, 3rd ed., Cambridge University Press, 1966.
9. D. Delannay, I. Setyawan, R.L. Lagendijk and B. Macq, *Relevant Modelling and Comparison of Geometric Distortions in Watermarking Systems*, in the Proceedings of SPIE, Application of Digital Image Processing XXV, Vol. 4790, Seattle, USA, 2002, pp. 200-210.
10. A.M. Tekalp, *Digital Video Processing*, Prentice-Hall, Inc., Upper Saddle River, 1995
11. A.M. Tekalp, *Differential Methods*, part of the lecture notes for Digital Video Processing, University of Rochester, New York, USA, 2001