Evaluation of the visual performance of image processing pipes: information value of subjective image attributes

G Nyman[a], J Häkkinen[a,b,c], E-M Koivisto[a], T Leisti[a], P Lindroos[a], O Orenius,[a] T Virtanen[a], T Vuori[b]

[a]Department of Psychology, University of Helsinki, P.O. Box 9, 00014 Helsinki, Finland;
[b]Nokia Research Center, P.O.Box 407, 00045 Nokia Group, Finland;
[c]Laboratory of Media Technology, Helsinki University of Technology, P.O Box 500 02015 TKK, Finland

ABSTRACT

Subjective image quality data for 9 image processing pipes and 8 image contents (taken with mobile phone camera, 72 natural scene test images altogether) from 14 test subjects were collected. A triplet comparison set-up and a hybrid qualitative/quantitative methodology[1] were applied. MOS data and spontaneous, subjective image quality attributes to each test image were recorded. The use of positive and negative image quality attributes by the experimental subjects suggested a significant difference between the subjective spaces of low and high image quality. The robustness of the attribute data was shown by correlating DMOS data of the test images against their corresponding, average subjective attribute vector length data. The findings demonstrate the information value of spontaneous, subjective image quality attributes in evaluating image quality at variable quality levels. We discuss the implications of these findings for the development of sensitive performance measures and methods in profiling image processing systems and their components, especially at high image quality levels.

Keywords: Image processing pipes, mobile phone camera, subjective image quality, experience, decision making

1. INTRODUCTION

Image quality of mobile devices improves fast and it is necessary to optimize the performance of digital image processing components that critically affect camera performance[2]. Standard methods have been developed [3,4] but at the moment, there is no unique objective and automatic measurement methodology that would match the visual performance of the end-user. In fact, this remains a distant aim, especially in the case of high-quality imaging, that is already an essential aspect of mobile imaging. Because of this, subjective evaluation methods are needed for measuring the image quality contribution of image processing components. Very little help can be expected from the present visual system theory since it does not, as of yet, match the needs of high quality imaging[1]. Hence, we have chosen to study how human observers spontaneously evaluate high image quality, how this could best be measured and in the end, how to support the building of models for the purpose of automatic and "objective" image quality measurements.

1.1 The process of quality perception

Here we consider image quality evaluation by the end-user as a spontaneous decision making process[3] that is based on a complex perceptual, cognitive, and emotional analysis of the image and its content. It is a multidimensional and inherently private process in nature, and it is not possible to define a fixed set of image quality scales that would cover all relevant aspects of subjective quality that an observer might experience when viewing high-quality test images.

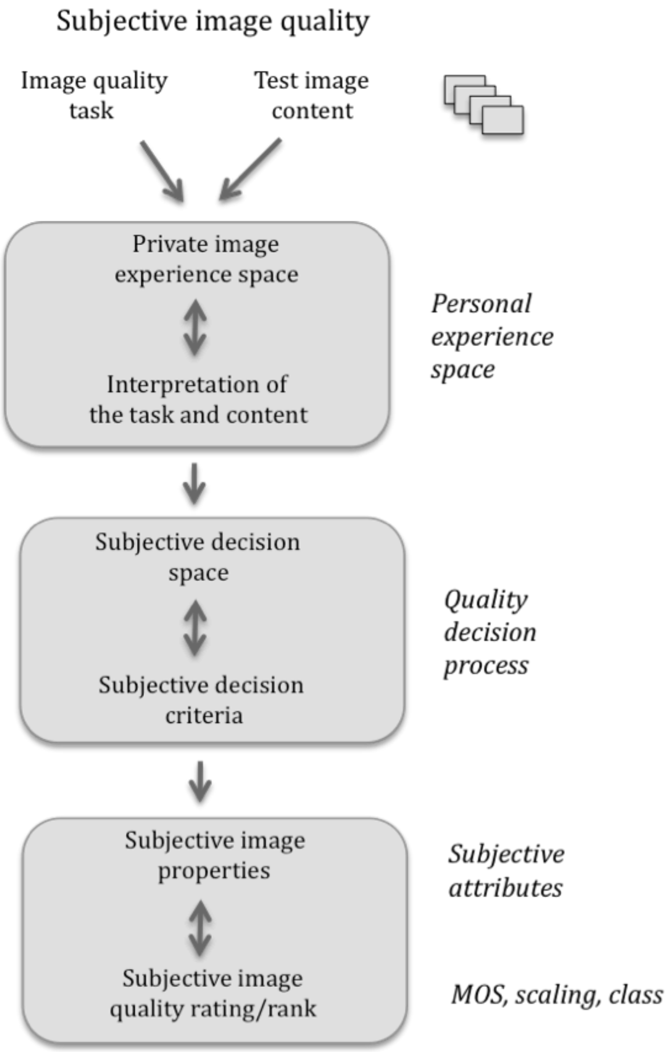
Subjective image quality
Image quality task
Test image content
Private image experience space
Interpretation of the task and content
Personal experience space
Subjective decision space
Subjective decision criteria
Quality decision process
Subjective image properties
Subjective image quality rating/rank
Subjective attributes
MOS, scaling, class

Figure 1. Main components of the subjective image quality decision process.

The image quality experience process can be characterized as follows:

1. Each perceived image (content) becomes perceptually, cognitively, and emotionally interpreted somehow (e.g. visual segregation of image element and objects; memory-based object recognition; multi-level subjective meanings and implications, a compound of bottom-up/top-down processing)[1]. This is a private process.
2. When faced with an image quality evaluation task, the subject undergoes an internal quality decision process (QD) that has multiple decision criteria that are dependent both on the image content and its quality level. This involves both bottom-up and top-down processing.
3. It is not possible to observe QD directly and even the subjects themselves cannot analyze this internal process accurately. However, valuable and systematic, but indirect information can be obtained: knowledge of these multiple decision criteria (and the underlying individual decision space) can be gained when the subjects explain, *why* they have given a specific image quality score to a test image. This information is represented, in the subjects' own language, and in the form of subjective image attributes that they can spontaneously express (e.g. sharp, warm, natural, pale) to the experimenter. In a sense, the subjects participating in an image quality evaluation task must compress their internal perceptual and experiential space into a dimensionally sparse and transformed decision space that is cognitively economical to use when discussing and rating the quality of test images. When a subject

describes his/her quality score, by using a spontaneous attribute like “natural”, he/she reveals something of the decision dimensions and the criteria that he/she has applied in his/her personal decision space. All subjective image quality attributes, even those that appear most direct and physical in nature (e.g. sharp, good contrast, bright pure colors), have this cognitive property.

4. By analyzing the attribute set obtained from the test subjects, in controlled conditions, it is possible to reveal and perhaps model these complex subjective decision phenomena.

Figure 1 summarizes the nature of this complex, interactive and multi-level subjective process. Accordingly, we have here conducted experiments where the subjects were asked to rate the image quality of images that have been processed through different image pipes, and to express their arguments (in the form of spontaneous, subjective quality attributes) for giving a specific quality score to each image. By looking at the relationship between image quality scores and the attribute distributions we hope to gain information of the performance of the image processing pipes and also of how the subjects arrive at the decisions concerning the image quality. We have earlier reported applications of this method (IBQ, Interpretation based quality) to image quality analysis of digital and printed images[1,2,6,7,8] .

2. METHODS

2.1 Materials

Natural image contents were used as stimuli. They were taken with a 5 megapixel mobile phone camera and they represent a wide range of everyday photographic situations. I3A standards were used as guidelines in selecting the test materials. In the experiments, the test images were presented as triplets that were shown in a random order. Altogether 96 triplets were shown on the three Eizo ColorEdge CG241W display set-up. The picture dimensions were 1920x1200, and the viewing angle from a viewing distance of about 1 meter was approximately 40 deg x 25 deg. The room illumination was set to approximately 20 lx on the desk surface and the environment was covered with medium gray curtains.

In order to test the Image Signal Processor pipes (ISPs), the RAW image database was collected. The natural image contents were captured with a commercial 5 megapixel mobile phone camera with LED flash. Only addition to the commercial version was the RAW image output to get the raw image database. The RAW image database was used to run the Image Signal Processor pipes afterwards to obtain each set of images for each processor pipe. Only the automatic exposure algorithm was the same in all test pictures. Altogether 9 different pipes were applied. Most of the pipes were mobile compatible, so they needed to fulfill the mobile requirements of processing power and time etc. There were both software (SW) and hardware (HW) pipes included, and also pipes of both SW and HW. The test image content types could be classified as follows: 3 dark images: Bar (faces), Cathedral, Person (“Pena”); 3 light images: Children, Cherry tree, Cars; 2 Gray images: Beach, Restaurant. Of the test images, 6 were taken outdoors (Person, Cars, Cherry tree, Cathedral, Children, Beach) and two inside (Bar, Restaurant).

2.2. Image quality evaluation procedure

Subjective image quality data for 9 image processing pipes and 8 contents (72 images altogether) were measured using qualitative/quantitative methodology[1,2,6,10] . Triplet comparison was applied (ISO 20462-2) in which the subject gave to each image in the triplet comparison situation, a quality score on the scale 0-100, and then also expressed his/her reasons for giving this score by writing down the relevant attributes for each triplet image. The subjects were instructed to use their own, spontaneous words and expressions to describe the test images. They used plus and minus signs to indicate whether they used the expressions in a positive or negative sense. Each picture was viewed 4 times in this process, and each time, according to the randomizing, in a different triplet. The subjects were instructed to use the scale value “50” as the criteria for acceptable quality. Data from 14 observers are presented.

2.3 Observers

Observers were invited from a panel of students who had participated in an introduction session, where the image quality of everyday photographs (not the actual test stimuli) were discussed. This was done in order to facilitate the use of subjective concepts and language related to such stimuli. No direct suggestions were offered to them, and the panelists were free to discuss openly the quality impressions of the material that was presented. The observers were required not to be professionals on image quality estimation and not doing image quality estimation or image processing as a professional-like hobby. The observers had to be Finnish speaking. Their vision was controlled for the near visual acuity, near contrast vision (near F.A.C.T.) and color vision (Farnsworth D15) before the participation.

3. RESULTS

3.1 MOS data and overall performance of the pipes

Figure 2 shows the distribution of MOS data calculated over all subjects, test images and pipes. The subjects used the scale rather well, peaking close at “50” which they were instructed to use as the threshold value for subjectively acceptable image quality.

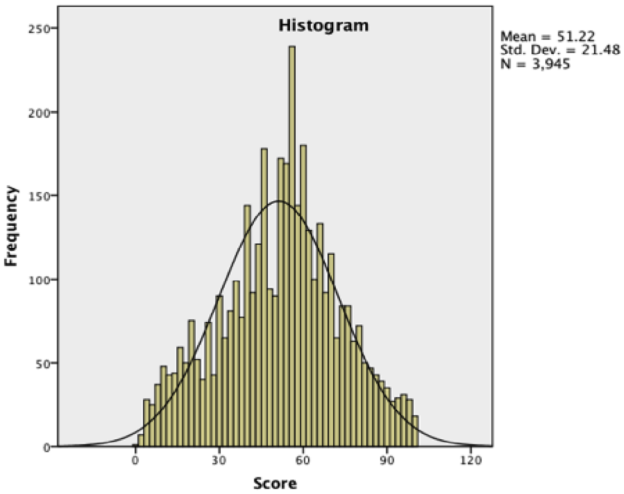

Figure 2. Distribution of image quality scores in the whole material. The distribution is close to normal although it did not pass the Shapiro-Wilk’s test of normality.

Figure 3 shows the average image quality data (MOS) for all nine pipes (numbered from 1 to 9) studied. There were rather large differences between the best and worst performing pipes, having the average MOS scores of 59,3 and 41,8, respectively. When arranged in the order of average performance score from worst to best, it was noticed that of the adjoining pipes, only pipes 1 and 9 were statistically significantly different (p<.05). The two best pipes (8 and 4) did not differ from each other statistically significantly. The S.E. values varied between .89 (pipe 3) to 1.10 (pipe 5). In other words, the subjects were quite consistent in their evaluations. This finding is emphasized by the fact that the data has been calculated over all test image contents.

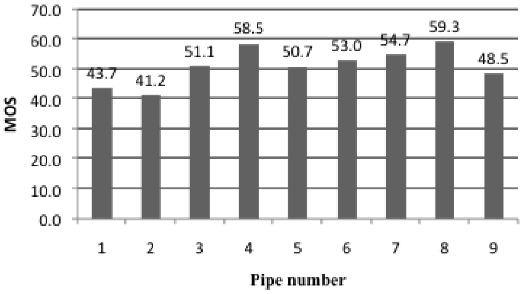

Figure 3. Average quality grade (MOS) values for each pipe, calculated over all contents and observers.

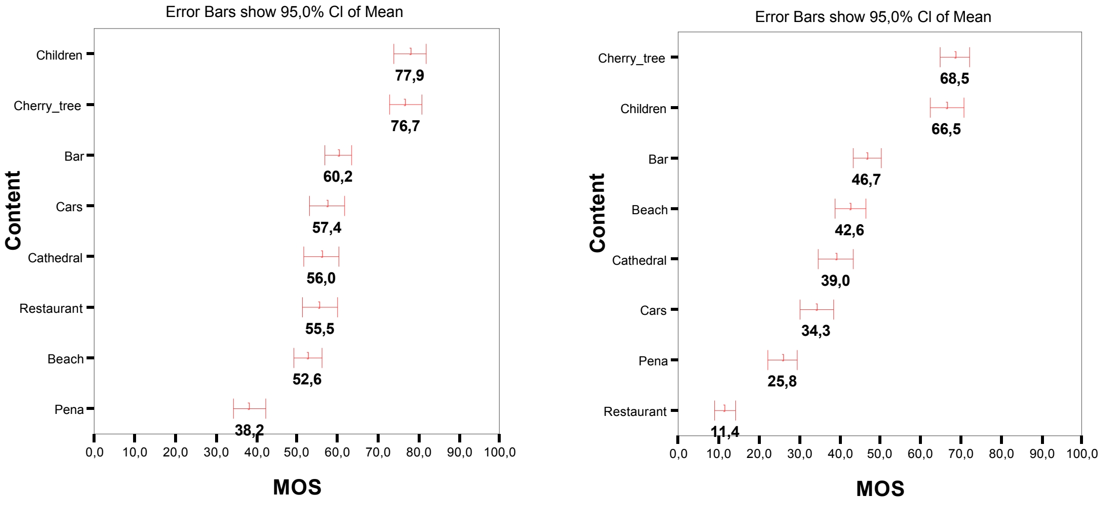

Figure 4. MOS performance for different image contents. The data is for the pipe with best average MOS (left) over all contents and for pipe with the lowest average MOS (right).

Naturally, there was a rather large variation in the MOS values as a function of image contents. As an example, the data of two pipes, the best and worst performing pipes in our study are shown in Figure 4. In both cases, the best overall performance was for the two light images, “Children” and “Cherry tree”. The variability for all contents and for both pipes is reasonable, showing again that the subjects were consistent in their evaluations.

3.2 Quality evaluations: attribute frequency data

When the subjects evaluated the quality of a test image by giving it a score on the scale 0-100, they were instructed to explain their reasons for the score by using spontaneous subjective image quality expressions or attributes (e.g. "unreal", "grainy", "bright", “natural”, “pale” etc). The data obtained from each subject for each test image was coded into attributes in Atlas.ti 5.0 and then transformed into a binary attribute vector form, separately for positive and negative attribute vectors $\mathbf{x}^-$ and $\mathbf{x}^+$. The scaled vector components of these vectors were calculated as proportions where $x^+_i = \sum +attribute_i / \sum total$ and $x^-_i = \sum -attribute_i / \sum total$, where $\sum +attribute_i$ indicates how many times each positive attribute i had occurred in all, and similarly, $\sum -attribute_i$ for each negative attribute i. We first calculated these vectors over all subjects, test images, and pipes; repetitions (situations where the subject repeated the same attribute in an instance where he/she evaluated a certain image) were excluded. $\sum total$ indicates the total number of attributes (positive or negative together) that were used by the subjects so that $\sum (x^+_i + x^-_i) = 1.0$.

Table 1 describes these scaled values (relative attribute frequencies), separately for the positive and negative ones, pooled over all image contents, image pipes, and subjects. It shows, that relative frequency of negative attributes used (that were coded as such) was nearly three times that of the positive ones. Figure 5 shows the data from Table 1, where the relative counts of positive and negative attributes are arranged in the order of their frequency of occurrence. The distribution of negative attributes has a rather smooth form suggesting a rather even use of several subjective attributes to indicate negative image qualities. In other words, the observers have a richer vocabulary for image problems than for high-level qualities. Clearly then, subjectively positive and negative aspects of the test images have been evaluated on quite different grounds.

Table 1. Relative proportions of negative and positive image quality attributes

Negative attributes $x^{-}{}_{i} = \sum\text{-attribute}_i / \sum\text{total}$		**Positive attributes** $x^{+}{}_{i} = \sum\text{+attribute}_i / \sum\text{total}$	
1. Grainy	0,125	1. Sharp	0,129
2. NotSharp	0,088	2. Natural	0,124
3. Red	0,066	3. SatCol	0,037
4. Pale	0,057	4. GoodCol	0,035
5. DarkColors	0,045	5. Warm	0,029
6. Blurry	0,043	6. Bright	0,025
7. Colorless	0,029	$\sum_{pos}$	0,379
8. Yellow	0,029		
9. Gray	0,025		
10. Blue	0,022		
11. UnNatural	0,017		
12. OverExposed	0,017		
13. UnClear	0,016		
14. OverSaturated	0,013		
15. DimDark	0,012		
16. HardContrast	0,010		
17. Cold	0,008		
$\sum_{neg}$	0,621		

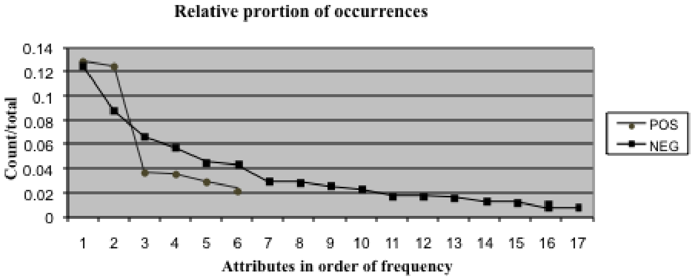

Figure 5. Relative frequencies of attribute occurrences for the negative and positive attributes in the whole test material.

3.3. Pipe performance: DMOS and attribute norm data

It could be suggested, that non-professional subjects are unreliable in a task where they are free to use their own spontaneous descriptions (attributes) in evaluating image quality. To see how the subjective attributes were related to the overall subjective image quality score for each image, we calculated the average vector norm, separately for positive and negative vectors to get an estimate of positive and negative “impact” of each test image. If the subjects were reliable and systematic in using the attributes and the method of collecting these attributes indeed is reliable, we could intuitively expect that the amount of positive attributes would be related to good quality as well as the negative ones would be related to negative quality. This should be reflected in the positive and negative vector norms.

Vectors $\mathbf{x}^-$ and $\mathbf{x}^+$ had the dimensionalities of 17 (negative attributes), and 6 (positive attributes). Average (positive and negative) and scaled vector components for each test image in our image set were computed. For clarity of analysis, we transformed the MOS values (over all subjects and pipes) for each test image into DMOS values by subtracting 50 from the MOS values. Vector norms for each test image were then calculated for $\mathbf{x}^-$ and $\mathbf{x}^+$, over all subjects, and pipes. The norms were also scaled to 1.0 by dividing them by $\sqrt{6}$ and $\sqrt{17}$, for positive and negative vectors, respectively.

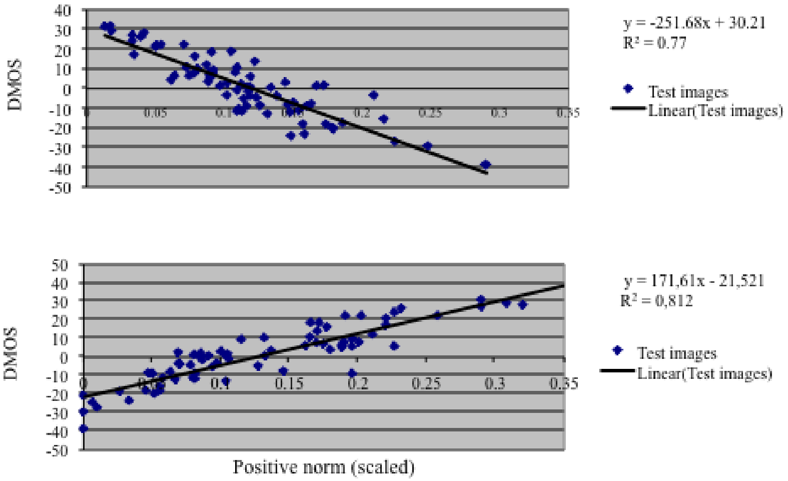

Figure 6. Average DMOS values of test images as a function of negative (above) and positive (below) attribute vector norms (scaled).

The norm data in Figure 6 shows a strong correlation between the vector norms and DMOS data. Apparently the test subjects were consistent in giving negative descriptions (attributes) to low quality image and positive descriptions to higher quality images. Relatively naïve test subjects indeed can be reliable evaluators even in the case of spontaneous and free description of test images. This demonstrates the value of subjective attribute data in profiling image quality. This information can then be used for image tuning and analysis of end-user preferences, for example.

3.4 Quality attributes at different image quality levels

Perception of high and low image quality requires different underlying perceptual mechanisms. To see how the most frequently used subjective attributes contributed to the evaluated overall quality of the image processing pipes, we classified the image quality DMOS values into 5 image quality classes according to the following quality (MOS) score values. Class 1: 0-19, 2: 20-39 , 3: 40-59, 4: 60-79, and 5: 80-100. Top 6 positive and top 6 negative attributes were then selected for further analysis. Figure 7 shows how the proportional contribution of different attributes changes towards the low image quality and high image quality. The contribution of each attribute is expressed as a proportion of times it was mentioned when a test image belonging to this quality class was shown. Clearly, the contribution of negative attributes increased towards the lowest quality class, and similarly the proportion of positive attributes mentioned increases towards the higher quality classes. The leading attributes are “Sharp” for the positive ones, and “Grainy” for the negative ones. The contribution of other attributes is variable. At the intermediate quality class 3 the contributions of positive and negative attributes does not vary much.

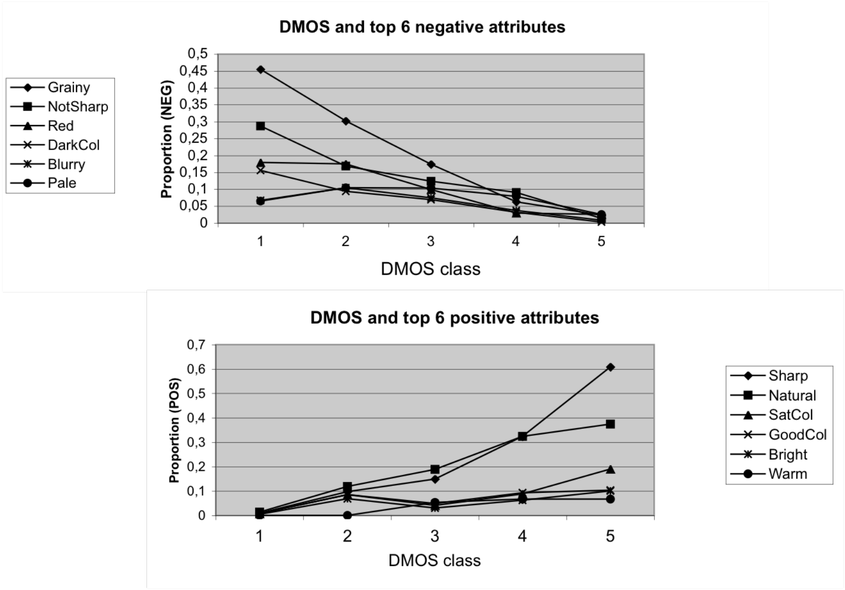

Figure 7. Proportions of the use of negative (top) and positive (bottom) image quality attributes to describe the test images belonging to the image quality classes from 1 (low) to 5 (high). Data has been calculated over all subjects, pipes and image contents.

3.5. Pipe performance

We used again the average attribute data, in the form of the positive and negative vectors, and their computed norms but not scaled as in the data of Figure 6, in order to compare the performance of the pipes with all 8 image contents. Figure 8 shows three example pipes, having the lowest, highest and a medium level overall MOS score. There are clear differences between average pipe performances and each of the pipes shown covers a rather distinct part of this performance space.

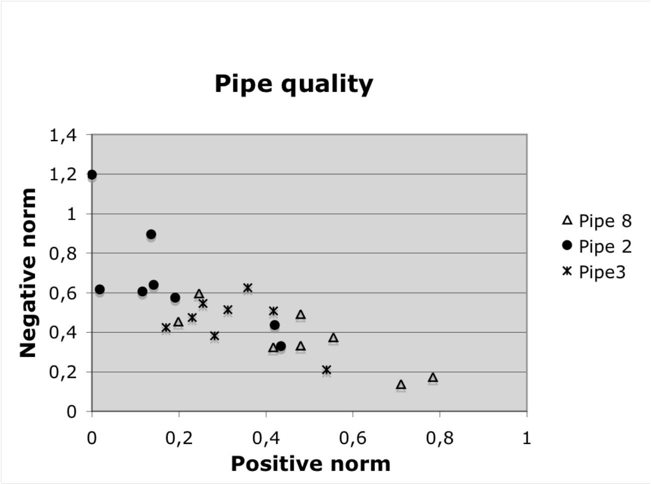

Figure 8. Average performance of three example pipes for all 8 different image contents calculated over all subjects. The data have been plotted as a function of positive and negative norms of the attribute vectors. Each data point represents the average performance of the specific pipe for one image content. The best performing pipe (Pipe 8) is dominated by the positive norm contribution while the weakest one (2) is dominated by the negative norm. For about one third of image contents there is overlap, showing the importance of selecting suitable test image material for differentiating between the pipes.

4. DISCUSSION

The process of subjective image quality evaluation is constrained by the evaluation task, image content, and the private nature of the visual experience that cannot be directly measured. However, reliable information about the underlying subjective dimensions and quality decision criteria can be obtained by allowing the observers to describe spontaneously why they have given the specific scores to the test images. Our present case shows again[1,2,6] how the dimensional structure of the subjective decision space varies as a function of image quality. Hence, as shown by the calculated attribute vector norms, low quality and high quality image processing pipes were characterized by different subjective attributes (decision dimensions) suggesting separate underlying subjective quality spaces. This is a serious challenge to the modeling of visual performance at all quality levels. It also prevents the use of pre-defined image quality measurement scales as tools for mapping the subjective image quality in variable image quality situations. However, useful profiling of the image processing systems can be accomplished by measuring their overall quality performance, their content specific performance and by identifying the relevant positive and negative quality attributes that reflect the subjective decision criteria. This also emphasizes the need for optimal test image contents, in good accordance with the I3A guidelines. Here we have not described the pipe profiles in detail, although we have used this data in our practical pipe performance analysis and description.

High-quality images challenge the performance of the human visual system. With increasing quality demands in the tuning of camera modules and image processing components of mobile phone cameras, also adaptive and sensitive methods and measures are needed to match their performance. The present application of the IBQ method can be seen as an example approach where the end-user is considered as a valuable source of performance information. It can be compared to sensory profiling studies that have a long history in food science and it has also been applied in audio engineering[11]. As to our best knowledge, this type of an approach has not been used in image quality analysis. The multi-dimensional nature of image quality decision requires a testing methodology that is sufficiently open, content sensitive, and subjectively relevant. There is a need for a definition of high image quality and a measurement framework that is rich enough for finding physical and/or computational correlates to it. The other side of the coin is, how to accomplish this so that the measurements can be conducted reliably, fast, and even better, could be done automatically. At present, such a method cannot be directly based on visual system theories, although there are studies that seem to close this underestimated gap between visual system models and natural visual performance[12]. In future, technologically feasible image quality methods must rely on systematically collected and experientially rich subjective data that has high predictive value of end-user preferences and views.

ACKNOWLEDGEMENTS

We thank Nokia for support and inspiring collaboration with image quality methods and materials.

REFERENCES

1. Nyman, G., Radun, J., Leisti, T. and Vuori, T., ”From image fidelity to subjective quality: a hybrid qualitative/quantitative methodology for measuring subjective image quality for different image contents,” Proc.12th International Display Workshops (IDW '05), Takamatsu, Japan, 1825-1828 (2005).

2. Radun, J., Virtanen, T. Nyman G. and Olives, J-L., ”Explaining multivariate image quality – Interpretation-Based Quality Approach,” *Proceedings of ICIS ’06,* Rochester, NY, USA, International Congress of Imaging Science, 119-121 (2006).

3 Engeldrum, P. G., [Psychometric scaling: A toolkit for imaging systems development], Imcotek Press, Winchester, MA, USA (2000).

4. Keelan, B. W., [Handbook of Image Quality], Marcel Dekker inc., New York (2002).

5. I3A, http://www.i3a.org

6. Radun, J., Leisti, T., Nyman, Häkkinen, J., Ojanen, H., Olives, J-L. and Vuori, T., "Content and quality: Interpretation-based estimation of image quality," Transactions on Applied Perception, (4)21, 1-21 (2008).

7. Eerola, T., Kämäräinen, J-K., Lensu, L., Leisti, T., Halonen, R., Kälviäinen, H., Nyman, G. and Oittinen, P.,
"Is there hope for predicting human visual quality experience?" IEEE International Conference on Systems, Man, and Cybernetics, SMC 2008, Singapore (2008).

8. Eerola, T., Kämäräinen, J-K., Lensu, L., Leisti, T., Halonen, R., Kälviäinen, H., Nyman, G. and Oittinen, P., "Finding best measurable quantities for predicting human visual quality experience," IEEE International Conference on Systems, Man, and Cybernetics, SMC 2008, Singapore (2008).

9. Eerola, T., Kämäräinen, J-K., Lensu, L., Leisti, T., Halonen, R., Kälviäinen, H., Nyman, G. and Oittinen, P., "Full reference printed image quality: Measurement framework and statistical evaluation," Journal of Imaging Science and Technology," In press.

10. Radun, J., Leisti, T., Virtanen, T., Häkkinen, J., Nyman, G. and Vuori, T., "Evaluating the multivariate visual quality performance of image processing components," ACM Transaction on applied perception, in press.

11. Olive, E.S. "A multiple regression model for predicting loudspeaker preference using objective measurements: Part I - Listening test results," 116th Audio Engineering Society convention, Preprint 6113, Berlin, May (2004).

12. To, M., Lovell, P. G., Troscianco, T. and Tolhurst, D. J., "Summation of perceptual cues in natural visual scenes," Proc. R. Soc. B 275, 2299-2308 (2008).