Pilot field trials

Deliverable D6.4.1

TOSCA-MP identifier:	TOSCAMP-D6.4.1-VRT-PilotFieldTrials-v07.docx
Deliverable number:	D6.4.1
Author(s) and company:	Stefanie Wechtitsch (JRS), Mike Matton (VRT)
Internal reviewers:	Alberto Messina (RAI)
Work package / task:	WP06
Document status:	Final
Confidentiality:	Public

Version	Date	Reason of change
1	2013-03-11	document created (structure, input to field trial results)
2	2013-03-26	Revision VRT
3	2013-04-16	Revision JRS, considered comments from VRT
4	2013-04-16	Version for further JRS update
5	2013-04-23	Version for internal review
6	2013-04-23	Response to review comments
7	2013-05-02	Final version

Acknowledgement: The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287532.

Table of Contents

List of Figures

1 Executive Summary

This deliverable describes the results of TOSCA-MP pilot field trials held in February 2013. The field trials consisted of experiments on the subjective evaluation of sharpness. For this purpose, 28 test subjects had to rate videos on sharpness. Two experiments were held: one single stimulus (SS) experiment where the subjects had to state the perceived sharpness on a discrete scale and a second experiment, in which the subjects were presented two videos for which they had to rate the sharpness on a continuous scale.

For half of the test subject, an eye tracking system has been used to track the eye movements of the subject during the sharpness assessment.

The evaluation confirms a high correlation between the objective sharpness measure and the subjective sharpness assessment of the test subjects.

The test subjects were also asked to rate their level of expertise on sharpness evaluation. From the results of the study, we can conclude that expert users tend to be more critical when sharpness is low. From the eye tracking data, it is also confirmed that non-experts tend to focus on high-saliency regions in the video, whereas experts focus on textured regions and edges. The results of this evaluation will be used to support the selection of areas for automatic sharpness detection algorithms and to adapt the sharpness metric to the specific target application.

It will also be necessary to perform more trials with more test subjects in order to confirm the results described in this document.

2 Introduction

2.1 Purpose of this Document

Pilot field trials: Report on pilot user trials of early prototypes for the preparation of the two rounds of field trials.

2.2 Scope of this Document

This documents reports on pilot field trials on sharpness assessment held at VRT in February 2013.

2.3 Reference Documents

Deliverable 2.1: Section 4.4:
Description of the sharpness algorithm until September 2012.
Deliverable 2.2: Section 5.4 (changes reserved since not yet completed)
Description of final sharpness metric used for the pilot field trials.

2.4 Status of this Document

Final version.

3 Evaluation results of the sharpness field trial

3.1 Introduction

In WP2.1 - Automatic visual metadata extraction, we have developed a novel sharpness metric that measures the level of sharpness or how in focus an image or video is [Fassold, 2012], [D2.1,2012]. The application of the sharpness metric is in the professional audiovisual media production and archiving workflow. Automatic quality control ensures that material can be played out for a specific target quality, or that archive material is of sufficient sharpness for reuse in a new production. Furthermore, the sharpness metric is able to detect whether material has been upscaled or supports the decision if the sharpness of the master is high enough for upscaling (e.g., for using archive SD material in an HD production).

Objective metrics only provide consistent and reliable results if they correlate well with subjective perception. Thus the development of quality metrics is typically supported by subjective studies in order to validate the results by experts' or consumers' mean opinion scores (MOS) [Ferzli, 2007].

The main issue is that the human visual system is an extremely complicated system. Humans have a diverging perception of sharpness and quality in general, thus user ratings are difficult to interpret and depend on many factors, such as content properties and interrelations of different quality criteria.

Subjective testing also allows for a better understanding of the mechanisms underlying quality perception, providing useful information for the subsequent modelling phase. Since subjective testing is expensive and time-consuming, it is often performed only for a limited set of test material. The performance of quality metrics can strongly depend on the database and the methodology used for testing. So far, researchers are far from agreeing upon a standard methodology [Redi, 2010]. The ITU [ITU-R, 2002] advised five methodologies mentioned as being reliable for image and video quality assessment, including both double- and single-stimulus approaches.

In order to provide computational models that automatically predict perceptual image sharpness we performed a complete user study, by using two of the methods proposed in literature in a slightly adapted way.

In the following, we describe this extensive and thorough study of evaluating sharpness of video sequences degraded by different levels, which is performed as a two-part experiment, involving 28 volunteers. Two different methods were used. First, subjects were asked to rate the perceived sharpness of several viewed videos without any reference and second, they compared two consecutively shown videos on a discrete comparison scale. The aim of this study was on the one hand to validate our novel sharpness metric against the human perception and on the other hand to get a deeper understanding of subjective judgments and possible differences in the perception of viewers. For half of all tests, an eye tracking system and expert were available in order to find out if humans use different regions of an image for judging the sharpness.

3.2 Sharpness metric

The following results were obtained by applying the developed sharpness metric, which is described in Section 4.4 of [D2.1,2012]. Basically, the sharpness/upscale detector measures how in focus an image or video is by utilizing the spread of edges as basic feature. The edges are detected by a Sobel filter for both, the vertical and horizontal derivatives (as originally proposed by Marziliano et al. [Marziliano, 2002]).

In order to avoid problems due to possible interlacing artefacts, the detector is applied on fields instead of using full frames. To ensure a reliable measure even in case of noisy images/video, a median filter was chosen in a pre-processing phase.
The pixels exhibiting high gradients are used as measuring points for our metric. At every measuring point the edge width is defined by the intensity variation of all pixels along the gradient, perpendicular to the edge, as far as a local min or max is achieved.
The image is then divided into blocks where for each block a representative edge width is calculated. Finally, the overall image sharpness is computed statistically from the most significant block sharpness values.

The extensions and refinements made to the algorithm during the project will be specified in [D2.2, 2013].

3.3 Overview of evaluation methods

The standardized test procedures which are commonly used are described in [ITU-R, 2002] and three of them, which are relevant for our evaluation, are discussed in detail in the following.

In quality assessment tasks, Single Stimulus (SS) methods are often preferred when evaluating objective metrics as they are straightforward to implement and well standardized. Viewers are shown test samples, such as images or videos, then they are asked to assess the perceived quality without having any reference for comparison. Commonly this is done on a continuous rating scale [Redi, 2010]. Furthermore the single stimulus continuous quality evaluation (SSCQE) allows viewers to dynamically rate the quality of an arbitrarily long video sequence. A slider mechanism with an associated quality scale is provided, in order to increase the sampling rate of the subjective quality ratings [Pinson, 2003]. In this way, differences between alternative transmission configurations can be analyzed in a more informative manner [Miras, 2002]. Having subjective scores at a higher sampling rate would be useful for tracking rapid changes in quality and thus would be more useful for evaluating real-time quality monitoring systems [Pinson, 2003]. Several works found that the SSCQE method exhibits some drawbacks. Since there is no reference available the observers have difficulties to give a numerical value for the appropriate quality task. The obtained values will always depend on the quality range spanned by the test material and usually the spread between different assessors is relatively large. Thus, merging results from different user studies, using different testing materials, may be challenging [Redi, 2010]. A further issue is the impact of the 'memory effect'. This effect describes the impact if the rating of a subject is dependent on the previously watched movies, their level of sharpness and the judgments.

In another method of subjective quality assessment, the Double Stimulus Continuous Quality Scale (DSCQS), viewers rate test material by using a reference. They watch multiple sequence pairs, where the original, or ``reference'' sample is always included. Each pair of videos is shown twice, in alternating, randomly chosen order. Subjects are asked to rate both of them on a continuous quality scale ranging from ``bad'' to ``excellent'', which is mapped to a defined scale, not knowing which of the pair is the reference and which the test sequence. By analyzing the difference of both ratings, uncertainties/unreliability of ratings caused by varying content and viewers' experience can be removed. The DSCQS method is preferred if reference and test sequences have only a minor difference in quality [Miras, 2002]. It is widely accepted as an accurate test method providing little dependency on context effects. Such effects may occur when the ratings of the subjects have a high variation due to the severity of impairments or the ordering of videos. Since pairs of videos are randomly shown, these context effects are lowered within this method. Each subject is asked to rate the perceived quality of each sequence after the pair was shown twice. In [Pinson, 2003] the authors found that resulting scores are not significantly impacted by memory-based biases from previously viewed video sequences.

In the Double Stimulus Comparison Scale (DSCS) method, viewers are watching a pair of video sequences of randomized order as in the DSCQS method. The pair is only viewed once in this method, and instead of rating both videos, the difference between the first and second video is rated, indicating whether the video quality of the second clip was better, worse, or the same as the first clip, in a seven point scale [Pinson2, 2003].

Summarized, the subjects in DSCQS and DSCS experiments are watching a pair of videos, where subjects in the former are asked to make two absolute ratings by using a continuous scale. In the latter, subjects are asked to give a relative rating on a discrete scale. Finally, the subjects in SSCQE make absolute ratings on a continuous scale by watching only one video. Usually, a slider is used in order to give continuous ratings over all viewing time [Pinson2, 2003]. The accuracy of the SSCQE method is a controversial issue. Subjects watch a video and rate it without having any reference material, which leads to a wide spread of the opinion scores. Furthermore, contextual effects may be present. Subjects may also loose concentration on the quality task, since they continuously have to move the slider in order to immediately track the changes over the time. As a result the reliability of the obtained scores is impacted [Pinson, 2003]. However, the SSCQE method is very popular and usually used for such context. The setup of an experiment using a SSCQE method and the following evaluation and analysis of the data is simple and the necessary effort is low. Evaluation of video sequences having different

levels of sharpness, where minor differences may be hard to recognize, can be effectively dealt with by paired comparison method. It was also used in [Lee, 2011] because providing preference between two sequences is much easier than assigning scores by using a continuous scale for human observers.

3.4 Experimental setup

As pointed out earlier in Section 3.3, SSCQE is easy to implement but it has important drawbacks. For our subjective quality assessment we have chosen to use two of the standard methods in a slightly adapted way. Assessing individual video sequences in a continuous rating scale is traditionally popular for subjective multimedia quality evaluation, so we are going to use the SSCQE method in the first part of the experiment. Since the accuracy of the SSCQE might be impacted due the continuous moving of the slider we compromised to provide only one discrete scale for each shown video. Thus only one rating obtained by using a 6 point scale is representing the perceived sharpness of a whole video clip. Furthermore we have decided to apply a second part to the experiment: Paired comparison is more appropriate and efficient for the goal of our sharpness evaluation study since minor differences are hard to recognize in a SSCQE method. Thus, the subjects were confronted with a DSCS method in a second part. In this part, a pair of videos is shown where each subject has to make a comparative rate on a discrete 5 point scale. Since we wanted to keep the duration of the experiment under 30 min and since we are interested in testing a diverse set of videos, we could not show a whole permutation of all possible pairs in the second test set. Instead we just show each video once, by randomly selecting a pair of videos until all videos have been shown once. 28 volunteers with varying expert level, ranging from age 20 to 60 participated in the subjective two-part experiment. They had to watch 32 videos in the first part, where the SSCQE method was applied and 28 videos in the second, comparative part. For half of the subjects we make use of an eye tracking system, as described below.

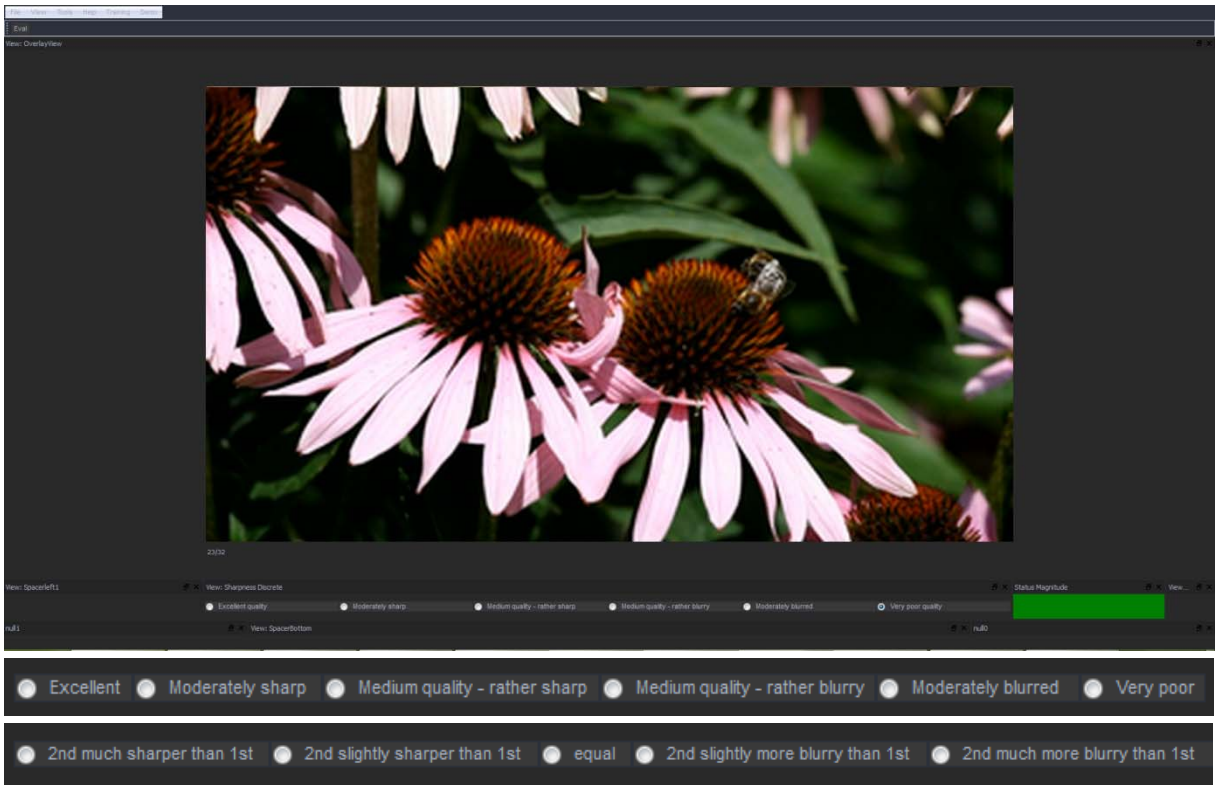


Figure 1: Evaluation tool (top) and rating scale of part 1 (upper bar) and part 2 (lower bar).

3.4.1 Creation of test material

Subjective evaluation experiments are often difficult since the number of publicly available subjective data sets is limited. Since, to our knowledge there are no databases available containing MOS for

varying blurred videos (in contrast to still images), we were forced to develop our own reference database. Material was kindly provided by VRT[1].

For the subjective sharpness experiment we have considered high definition (HD) video sequences that are of interest in today's production systems. Various factors of content affecting the perceived sharpness are considered in the test material for a complete analysis, such as inside/outside scenes, texture variety, faces and other saliency points.

We have created video clips that all had a length of approximately 8 seconds. The authors of [Pinson, 2003] confirmed that on average, subjects only need 7-8 seconds of video for forming their quality decision, and they show that ratings of subjects that have watched videos beyond 8-9 seconds do not differ from ratings that were made after watching a longer scene. Also, the memory effect is not extended much beyond 9 seconds and does not significantly differ from ratings that were made after watching a video after longer time. Therefore, we have chosen video clips with mainly a single scene, and in general showing constant sharpness over the whole scene. Since each video clip has a duration of 8 seconds, the whole experiment has an adequate length so that the test persons can easily stay concentrated.

15 scenes were then chosen and adapted in a further step in order to evenly span a full range of available sharpness. Each video was blurred with 3 different kernels, which results in a test set of 60 videos of 4 different sharpness levels. The 60 generated video clips were arbitrarily divided into two parts, 32 for the first and 28 videos for the second part. So both datasets spanned the same range of sharpness levels (100%, 50%, 33% and 25% of full resolution).

3.4.2 Setup of the experiment

In Figure 1 a screen shot of the used evaluation tool is shown. The tool provides the ability to view the videos in a randomized order and accepts discrete or continuous user ratings dependent on the configuration. The tool is used for both parts of the experiment.

All 28 volunteers had near-perfect or corrected-to-normal vision. Before the experiment they had to complete a questionnaire, where we asked for their age group, their eye defects and their expert level in terms of experience with quality control tasks of images and videos. Due to their answers the subjects were classified by their expert level in 5 groups. Prior to the first part we have presented two video clips of the same content, but with two different levels of sharpness. The first one was raw full HD content and the second was blurred to an extent equivalent to reducing the resolution to 25% of the first. After watching these two video clips the viewers were asked to place them in a range of 0 to 100 without having any instructions or reference for the measure. The reason for this prior assessment was to get an idea of the personal sharpness range of each subject. In many experiments subjects are instructed by showing them example videos with their appropriate sharpness or quality value. We chose to ask for their assessment instead, as we believe that this provides a more reliable way of calibration than presenting a defined calibration, from which viewers might drift away during the experiment.

The subjects were placed in a viewing distance of approximately 3 times the height of the viewing screen. The environment illumination was dimmed and controlled, and we provided a silent environment with as little environmental effects as possible. The subjects had no time limit for giving their rating; however, the majority of the subjects needed 15-25 minutes for the entire test set. The video clips were shown on a 32 inch Samsung LED TV series 6 screen with a native resolution of 1920x1080 pixels.

[1] Flemish public broadcasting company

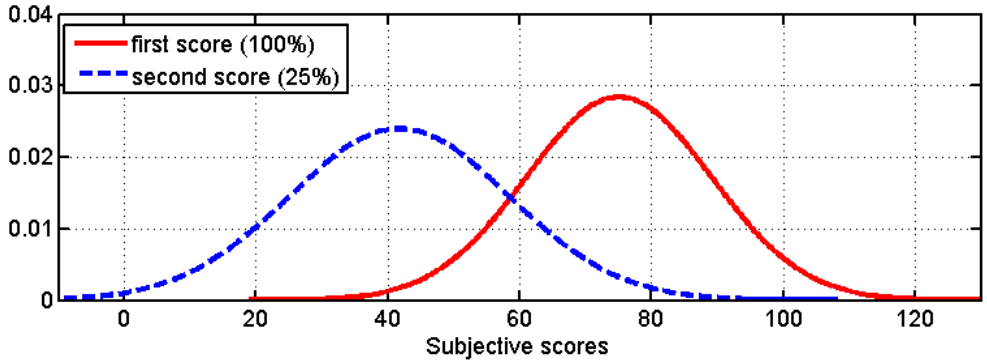


Figure 2: Distribution of initial ratings.

3.4.3 Eye tracking

Half of all subjects judged the sharpness of the video clips while an eye tracking system was used. Thus, we not only captured the perceived sharpness, but also the focus of the viewer while building the judgment of the sharpness. The intent was to track the eyes during the whole experiment, since the obtained gazing maps are easily comparable to the regions of sharpest edges predicted by the sharpness metrics. The eye tracker has marked regions where viewers have focused for a certain time (gazing points) over the whole experiment.

As a result, we have captured not only the human perception of sharpness of HD videos, but also their regions of attention during deciding about the sharpness rating. This has two advantages: On the one hand we can validate the novel sharpness metric within this experiment (in terms of correlation of regions selected for judgment by the automatic metric) and on the other hand we have learned about the regions of attention, that subject with different level of expertise use for judging. In a further step we can easily compare them to the regions used by the sharpness metric.

3.5 Results

Results are presented separately for the two parts of the experiment, the SS and for the comparative DS methods:

3.5.1 Single stimulus results

For the SSCQE method we can easily compare the subjective perception of the subjects with the result of our sharpness metric. In Figure 3 the mapping of all 32 videos considered in the first part of the experiment is visualized, plotting the MOS against the automatic sharpness scores.

The sharpness metric additionally calculates a confidence value, which mainly depends on how many edges are present and consequently how many blocks can be used for measuring the sharpness. This confidence is visualized by the colour and size of the data points. The comparison shows, that the sharpness metric is well correlated with the MOS, resulting in a Pearson correlation coefficient of 0.7382 and a Spearman rank correlation coefficient of even 0.7855. The plot shows a tendency, that the correlation decreases with increasing level of sharpness. This effect can be explained by the high variance of human's perception which is reflected in their ratings and will become clearer later in this section when we further investigate the experience of test persons.

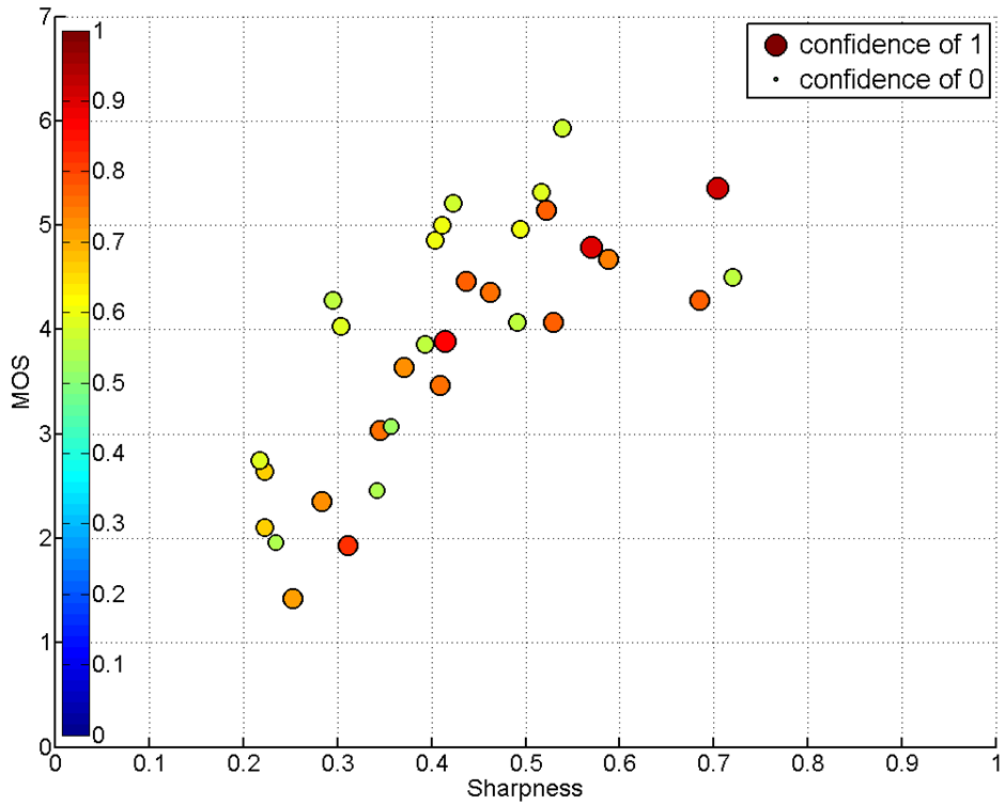


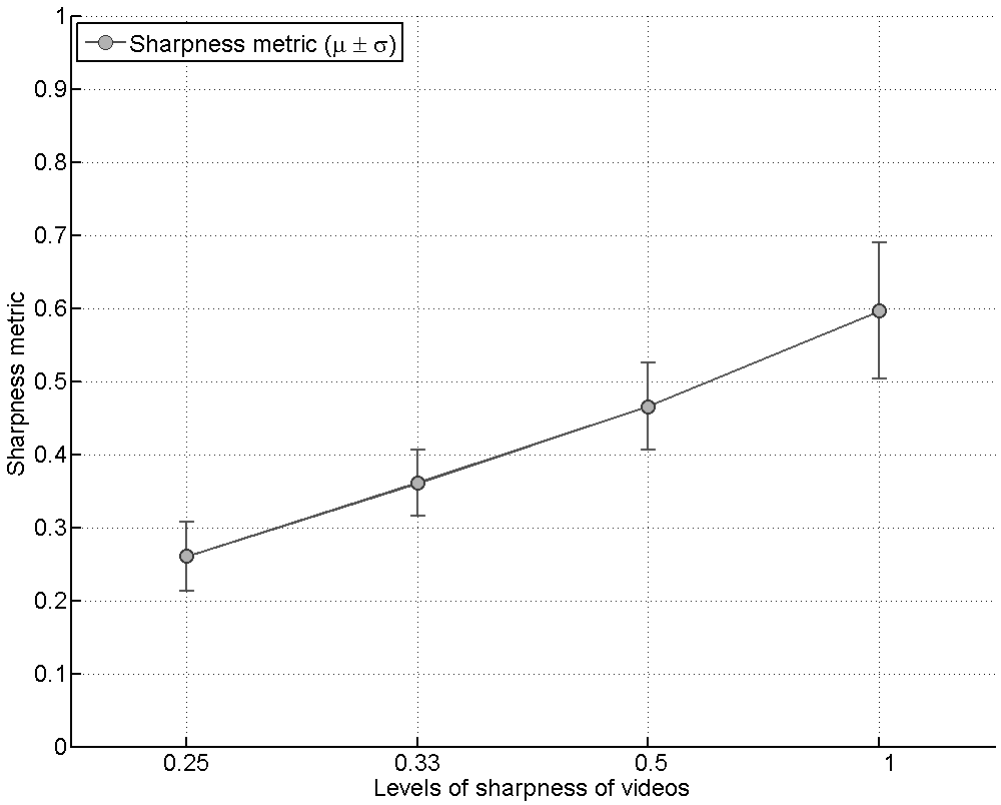
Figure 3: Comparison of single stimulus MOS to sharpness.

The objective sharpness metric is robust and clearly distinguishes between the sharpness levels, which is presented in Figure 4a. The standard deviations of the results separated by the level of sharpness are relatively small and, most importantly, the ranges do not overlap with each other. The results are well correlated with the GT (correlation is 0.8823, Spearman coefficient is 0.8923).

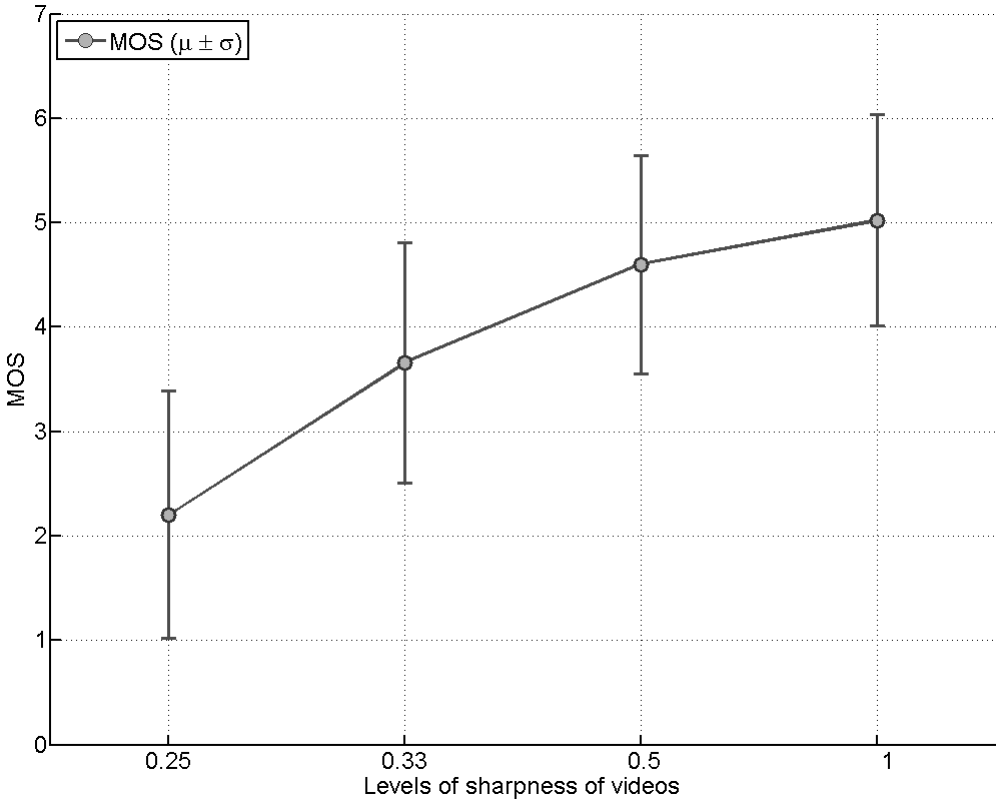
In Figure 4b the correlation between the MOS and the ground truth, i.e., the resolution reduction obtained by blurring the full resolution input data is shown. Due to the high correlations it seems that subjects on average have similar perception of the sharpness. When considering the ratings of the initial examples (100% and 25% sharpness), where each subject was asked to place the videos in a continuous scale, shown in Figure 2, we have to admit that this is not the case, since the personal value ranges of human's sharpness perception seem to vary.

While the differences are in general recognized by the subjects, the standard deviations of the ratings are quite large. The high standard deviation can be explained by the different personal value range of each viewer.

Another reason for the increased standard deviation may be that viewers with less experience in quality tasks may not be able to distinguish between minor differences in sharpness levels, while more experienced viewers do have the ability. Thus, we could not confirm the assumption that the SSCQE method provides unreliable results.

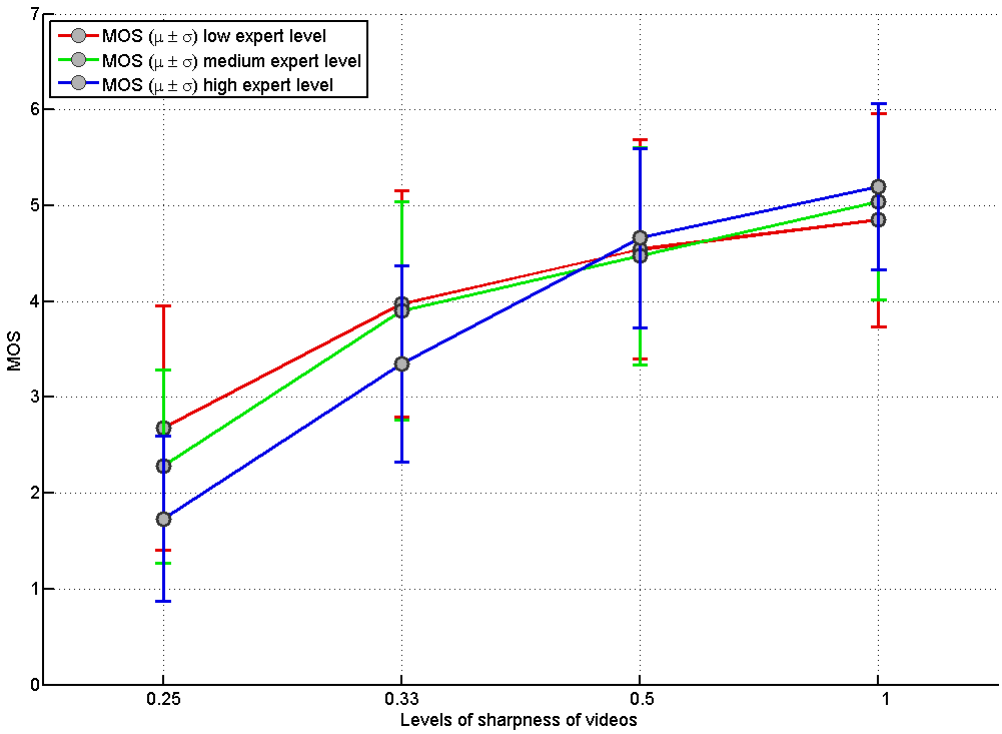


(a)

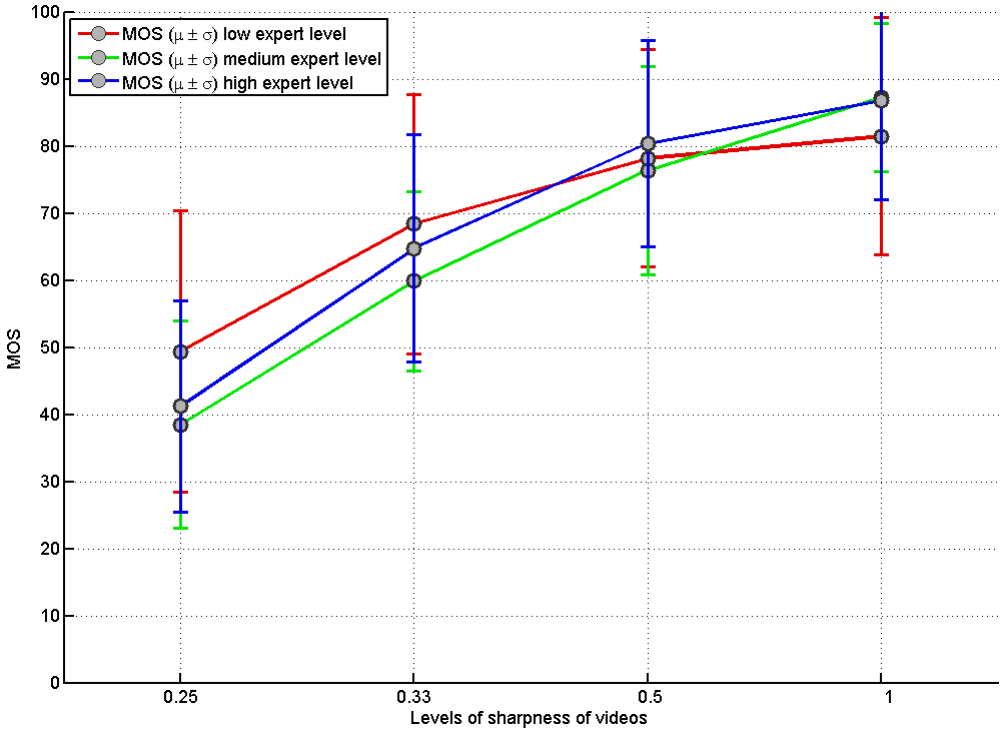


(b)

Figure 4: Sharpness metric results and MOS vs. ground truth.



(a)



(b)

Figure 5: (a) MOS of different expert levels and (b) MOS scaled with the initial ratings.

In Figure 5 the ratings of the single stimulus part of the experiment are visualized in a different way, focusing on the expert level of each subject. The hypothesis about significant differences between subjects of different expert level could not be confirmed from the ratings.

To see a difference in the rating we have scaled the ratings of each subject concerning the first impression of the example videos we have shown.

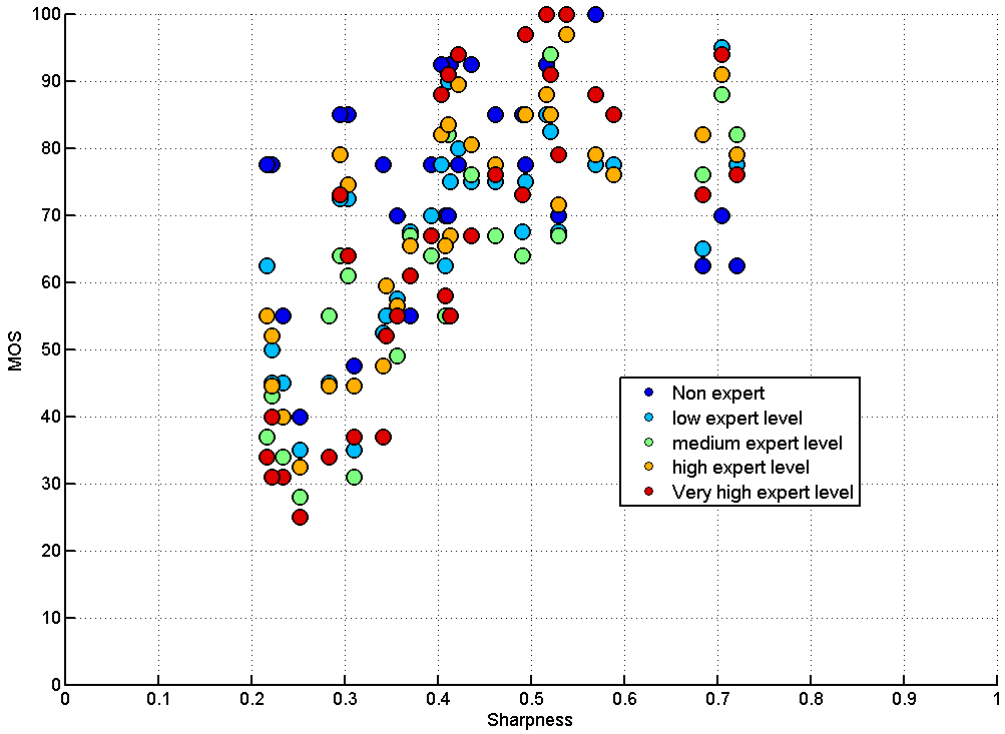


Figure 6: Scaled MOS of part 1 vs. results of sharpness metric.

The resulting MOS and standard deviations for three levels of expertise (low, medium and high) are shown in Figure 5. The correlations (Pearson coefficient) between unscaled MOS and sharpness metric by increasing expert level are 0.3473, 0.6759, 0.7602, 0.7358 and 0.7507 for the highest expert level. For the scaled MOS the correlations with increasing expert level are 0.4113, 0.7155, 0.7491, 0.7402, 0.7435. These correlations are similar when comparing with the ground truth instead of the sharpness metric results.

According to these results the accuracy of the MOS can be interpreted as dependent on the expert level. A small classification error needs to be considered since the expert level is a matter of definition and subjects may have been misplaced due to e.g. subjects' modesty.

In Figure 5 we have noticed that experienced subjects tend to rate video clips with a lower sharpness level more critical, while the ratings for sharper material are similar. The same can be seen in Figure 6. The extension of the points of lower expert levels is more flat or uncorrelated as for the non experts. With increasing expert level the rise of according points becomes steeper, which means that differences were better recognized by the very high expert group, which is also an explanation for the decreasing correlation with increasing level of sharpness.

3.5.2 Double stimulus results

In the second part of the experiment, the subjects judged the perceived difference of two subsequently presented video clips. The video pool for this second part of the experiment consisted of 28 videos. Thus, 14 pairs of videos were shown to the test persons, and 14 times the question has to be answered if the second video was either sharper, equal or less sharp than the first seen video. Instead of showing all possible combinations of pairs of the second pool of videos, we just show each video once, until all videos have been shown. Since the number of possible ways to select 14 couples out of 28 items ends into a combinatorial explosionm several of the randomly connected pairs were only shown to one viewer, such that only one score is available for some of the pairs. In Figure 7 the result of the comparative double stimulus part of the experiment is given. The confidence is computed by the number of available user ratings and is visualized by the colours and size of the data points corresponding to the colorbar and legend of the plot. If at least 12 ratings were available the confidence is set to 1. The comparative scores obtained by using the rating scale shown in Figure 1 were

transformed into a discrete scale of -6 to 6. In order to compare the ratings with the performance of the sharpness metric, also those values were translated and placed in the same scale.

For the comparative judgments we report a correlation of 0.6825 even though several single scores have outlier character.

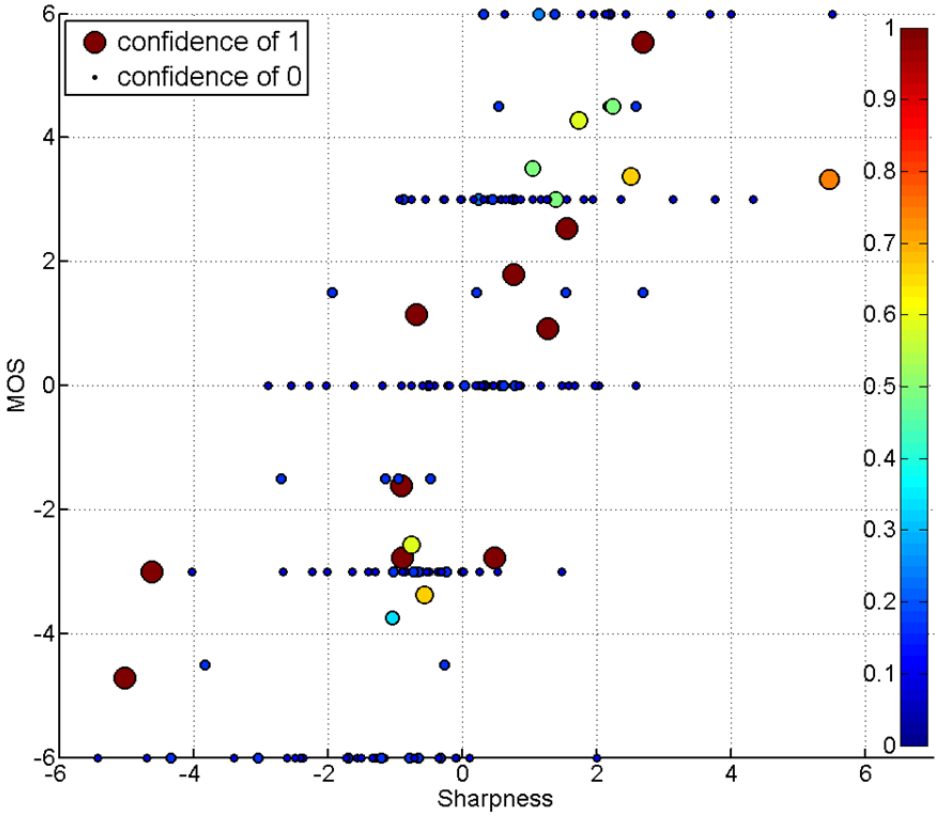


Figure 7: Comparison of double stimulus MOS to sharpness.

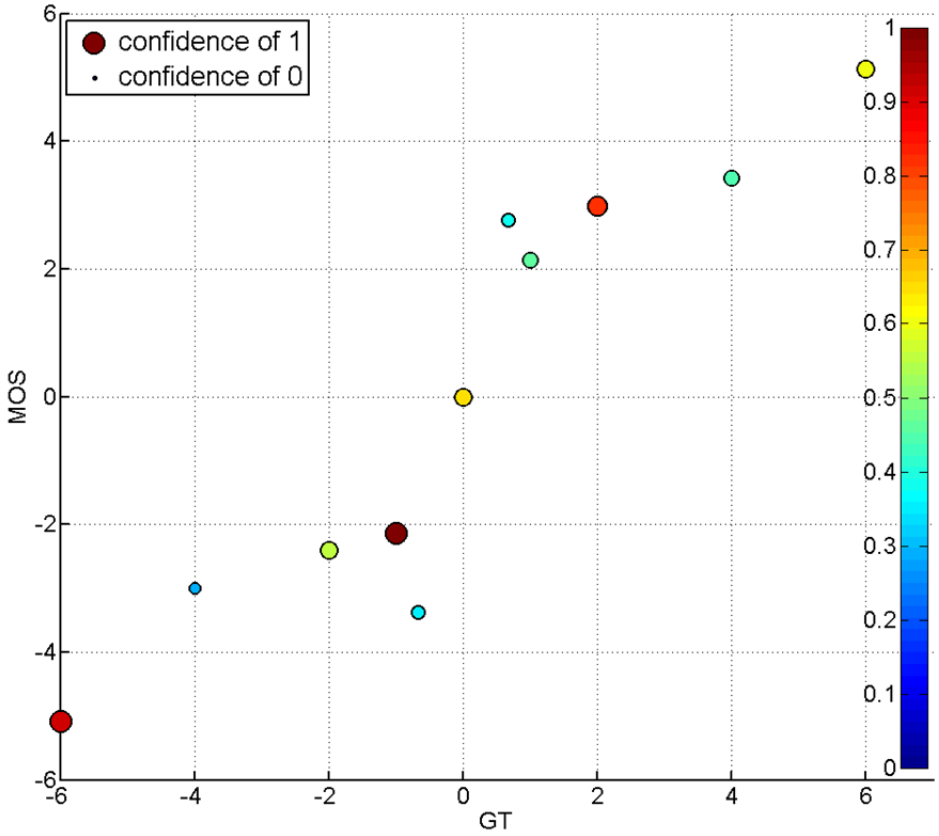


Figure 8: Double stimulus MOS summarized for all video pairs with the same difference of sharpness levels vs. Ground truth (corr. 0.7302).

Since many points have a very low confidence we analyze only the sharpness change. All video pairs with equal difference of sharpness levels are summarized. Summarizing these MOS is simple, but when we want to compare them with the automatic sharpness metric it is difficult to compute corresponding points. Since we have investigated the robustness and consistency of the sharpness metric and showing that the results are well correlated with the ground truth (see Figure 5), we compare these values with the ground truth as well. The result is shown in Figure 8.

The resulting correlation of the summarized MOS by sharpness difference with the ground truth is 0.9202.

Figure 9: Gazing points of a chosen set of subjects.

3.5.3 Analysis of eye tracking data

As mentioned above, half of all tests were complemented by the use of an eye tracking system. While the subjects were watching the video clip, the system was tracking the focused regions within the video clip. In particular subjects with high level of expertise tend to be more critical and we assume that this can be explained by the fact that they used to focus on varying regions of interests for judging.

In Figure 9 the gazing points of six subjects of two different frames of a test video are shown. The green circle corresponds with the gazing location of a viewer with a very high expert level. The left example shows that experts rather focus on those parts of an image where even minor differences in the sharpness and also slightly blurred edges are visible. The purple and the beige circle are located around the face. These two circles belong to viewers with non-expert level. This finding reflects the assumption that faces are naturally attracting in visual tasks, they are salient. The other two circles, the blue and the white one, belong to viewers who placed themselves in a middle level of quality experts. Obviously, they know that in textured areas it is easier to recognize blurred edges.

Figure 10: Comparison of gazing locations of two videos with different sharpness level. The red blocks indicate areas used by the automatic detector.

An interesting scenario arises, when another person comes into focus later in this clip. Immediately, almost all viewers are focusing on the person entering, except for the expert (orange circle). Experts that are asked to judge an image or video concerning the sharpness are undeterred of faces and keep concentrating on edges in the background.

When comparing the videos seen in Figure 10, where two videos with the same content but with different sharpness levels are shown, the regions of interest do not change significantly: The regions people focus on when judging the quality of images/videos do not depend on the level of sharpness and stay similar.

Additionally, we have visualized all blocks that are used by the objective sharpness metric for computing the sharpness of an image. Mainly, the algorithm selects those blocks that contain the sharpest edges, since we have hypothesized, that humans tend to focus on the sharpness regions as well. Partly, the hypothesis can be confirmed by the shown example: The plant hanging down the window in the background is obviously an interesting region for both, subjects and the objective metric.

In order to report reliable results concerning preferred regions of focus we realized that a much higher number of samples are required. We neither could confirm any preferring non-expert users have in viewing tasks nor refute one of the suggestions (faces, motion, texture, etc.). Further analysis of the eye tracking data is planned in future work.

We found that the gazing locations are clearly dependent on the viewer's expert level and are reflected by the scores. Furthermore, the comparison task clearly shows that the ability to detect minor differences of sharpness increases with the expert level, which seems to be independent of the focus regions. In order to find out which parts of a video are chosen to focus on dependent on parameters such as expert level, age and sex, a much higher number of samples is required. In order to make sure that the use of eye tracking does not impact the results, we have also compared the MOS values of the two setups.

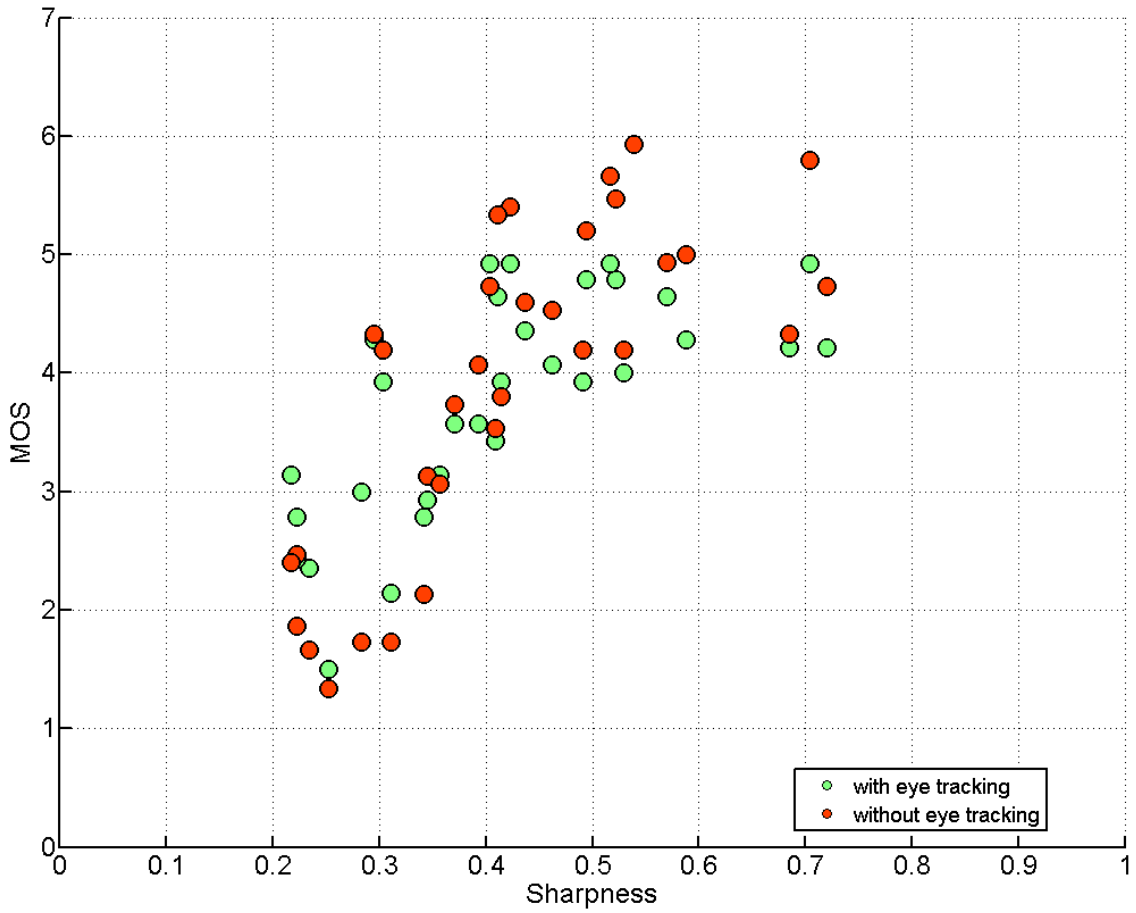


Figure 11: Comparison of MOS obtained in two scenarios (with/without using eye tracking).

To check this, we have compared the MOS values of the two used scenarios, so if an eye tracking system was used or not. No significant difference can be recognized, as shown in Figure 11. The correlation of the MOS obtained from the task where eye tracking was used resulted in correlations of 0.6227, without eye tracking the correlation between MOS and the sharpness metric is 0.7554. In order to analyse the recognition we used the initial ratings, which were tested in the beginning of the experiment and are demonstrated in Figure 2. Thus we scaled the ratings within these values in order to have an artificially, but unique rating scale for each user.

4 Conclusions

In this deliverable, we have presented the results of applying and comparing two common methodologies, SSCQE and DSCS, for the evaluation of a novel algorithm for sharpness of video.

Both methods were compared in subjective experiments with two sets of videos varying in their level of sharpness, down sampled from HD originals.

The evaluation confirmed a high correlation between the algorithm and the subjects' sharpness assessments.

Asking subjects for a calibration judgment on two sequences in the beginning allowed scaling of the scores in the single stimulus part of the experiment. This helped to avoid the range effects described as drawbacks of this method in literature.

The rating procedure of paired comparison is simple so that training of subjects can be performed easily. In addition, the reliability of each subjects' ratings can be judged independently in our methodology, while other subjective data is required for outlier detection in MOS-based methodologies. In terms of rating differences between experts and non-experts, overall no significant differences have been found. However, smaller differences become evident with low-resolution content, where experts tend to be more critical. It will be necessary to perform further validation of the conclusions made in our work by using more diverse contents and a broader range of subjects.

The evaluation has been complemented by the use of an eye tracker, in order to determine where subjects focus when making their sharpness judgments. The results confirm the assumption, that non-experts tend focus on typical high-saliency areas, such as faces or motion areas, while experts select textured regions and edges, which enable them to perceive sharpness degradations more clearly. These results will help to better guide the selection of areas used by automatic algorithms, and the adapt sharpness estimation methods to different target applications.

5 References

[D2.1,2012] Deliverable 2.1, “Automatic Metadata Extraction and Enrichment”, 2012.

[D2.2, 2013] Deliverable 2.2. “Automatic Metadata Extraction and Enrichment”, 2013.

[Schallauer, 2009] P. Schallauer, H. Fassold, M. Winter, W. Bailer, G. Thallinger, and W. Haas, “Automatic content based video quality analysis for media production and delivery processes,” in Proc. SMPTE Tech. Conf., 2009.

[Fassold, 2012] Fassold H.,Wechtitsch S., Hofmann A., BailerW., Schallauer P., Borgotallo R., Messina A., Liu M., Ndjiki-Nya P., and Altendorf P., “Automated visual quality analysis for media production,” in Proc. IEEE ISM, 2012.

[Ferzli, 2007] Rony Ferzli and Lina J. Karam, “A no-reference objective image sharpness metric based on just-noticeable blur and probability summation,” in IEEE ICIP, 2007.

[ITU-R, 2002] ITU-R Rec. BT.500-11, “Methodology for the subjective assessment of the quality of television pictures,” 2002.

[Redi, 2010] J. Redi, H. Liu, H. Alers, R. Zunino, and I. Heynderickx, “Comparing subjective image quality measurement methods for the creation of public databases,” in Proc. Electronic Imaging, 2010.

[Pinson, 2003] M. Pinson and S. Wolf, “Comparing subjective video quality testing methodologies,” SPIE Video Communications and Image Processing Conference, pp. 8–11, 2003.

[Miras, 2002] D. Miras, “A Survey on Network QoS Needs of Advanced Internet Applications,” Tech. Rep., Internet2 QoS Working Group, Nov. 2002.

[Pinson2, 2003] M. Pinson and S. Wolf, “An objective method for combining multiple subjective data sets,” in Proc. of Electronic Imaging, 2003.

[Lee, 2011] J.-S. Lee, F. De Simone, and T. Ebrahimi, “Subjective quality evaluation via paired comparison: Application to scalable video coding,” IEEE Trans. Multimedia, 2011.

6 Glossary

Terms used within the TOSCA-MP project, sorted alphabetically.

DSCQS	Double stimulus continuous quality scale
DSCS	Double stimulus comparison scale
GT	Ground truth
HD	High definition
MOS	Mean opinion score
SS	Single stimulus
SSCQE	Single stimulus continuous quality evaluation

Partner Acronyms

DTO	Technicolor, DE
EBU	European Broadcasting Union, CH
FBK	Fondazione Bruno Kessler, FBK
HHI	Heinrich Hertz Institut, Fraunhofer Gesellschaft zur Förderung der Angewandten Forschung e.V., DE
IRT	Institut für Rundfunktechnik GmbH, DE
K.U.Leuven	Katholieke Universiteit Leuven, BE
JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH, AT
PLY	Playence KG, AT
RAI	Radiotelevisione Italiana S.p.a., IT
VRT	De Vlaamse Radio en Televisieomroeporganisatie NV, BE

Acknowlegdement: The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287532.

