The final version of this paper is available at https://ieeexplore.ieee.org/document/8516509
DOI: 10.1109/CBMI.2018.8516509

Refining and Evaluating a Video Sharpness Metric

Stefanie Wechtitsch, Werner Bailer, Lucas Paletta
DIGITAL – Institute for Information and Communication Technologies, JOANNEUM RESEARCH
Steyrergasse 17, 8010 Graz, Austria
Email: {firstname.lastname}@joanneum.at

***Abstract*—Automatic quality assessment is an essential step in the professional audiovisual media production process. In this paper we propose a novel sharpness metric taking the specific properties of video into account, and having a higher robustness against variations in image content, interlacing artifacts and noise. Furthermore, a comprehensive user study is presented, where we obtain subjective scores to validate the sharpness metric. We ask 28 viewers for both absolute and relative judgments of sharpness in two separate experiment settings, supported with an eye tracker to obtain the locations used for judgements. Experimental results show that the objective sharpness metric under test is well correlated with human perception. Both results, the absolute and relative subjective ratings confirm the good correlation of the proposed metric with human perception. The analysis of the eye tracking data highlights differences between experts and consumers.**

I. Introduction

Automatic quality metrics and visual quality assessment (QA) is of fundamental importance for numerous video and image processing applications. Typical applications for image quality metrics are situated in the media production and delivery process, as described in [1] for content quality checking by broadcasters being costly and time-consuming. It is obvious that objective techniques are needed to predict visual quality automatically. Substantial effort has been dedicated to the development of new automatic machine vision based quality estimators.

Our work investigates one of the most common distortion types in digital image and video processing: blur. We propose a novel, no-reference sharpness metric that measures to which degree a video appears in focus. We consider sharpness as an effect of blurring, being multiplicative inversely proportional to image blur. The sharpness metric can act as an automatic detector of production insufficiencies (e.g. lens out of focus) or serve as decision maker in verifying the quality if material can be played out for a specific target quality (reuse in new productions, up scaling, etc.) [2].

Objective metrics only provide consistent and reliable results if they correlate well with subjective perception, i.e., if they estimate the quality as perceived by an averaged viewer. 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) [3]. We present an extensive and thorough evaluation study, involving 28 volunteers. In order to get a deeper understanding of subjective judgments and to detect possible variances of the perception of viewers, half of all experiments were done with an eye tracking system.

This paper is organized as follows: In the remainder of this section we first give an overview of recent related work regarding sharpness algorithms and the different approaches that were done. Section II provides a description of the novel sharpness metric based on local gradient feature analysis. Section III presents the experiments, describing the evaluation metrics and the experimental setup with the eye tracking system and the procedure of the user study. Experimental results are provided and discussed in Section IV and finally conclusions are presented in Section V.

Objective quality metrics can be categorized by the requirement of a distortion-free original image (sequence) for comparison. Full-reference metrics need the distorted image (sequence) and the original as input. An evaluation of several full-reference QA algorithms, including sharpness, can be found in [4]. Reduced-reference QA methods work with only partial information of the original visual signal available. Finally, no-reference QA methods measure the image or video quality blindly. In many applications in media production and archiving, the media item under analysis is the master, and no reference is available. As many of these applications have also real-time requirements, no-reference methods with low-computational complexity are required. Thus we particularly consider methods meeting this criteria in the following review.

Ferzli and Karam [5], and more recently, Narvekar and Karam [6] propose spatial domain sharpness metrics, based on the concept of just noticeable blur (JNB) or its extensions, using the analysis of edges and adjacent regions in images proposed by Marziliano et al. [7]. A transform based method, proposed by Sheikh et al. [8], uses statistics of the discrete wavelet transform (DWT) coefficients in natural images to produce quality scores for JPEG2000 compressed images. Wang et al. [9] demonstrate that local phase coherence (LPC) changes and that precisely localized features, e.g. sharp edges, cause a strong LPC in the complex wavelet transform domain. A more recent sharpness metric, calculated in the wavelet transform domain, analyzing the local phase coherence of complex wavelet coefficients is proposed in [10], assuming that blur causes a disruption of local phase near sharp image features. Vu and Chandler [11] proposed a sharpness method with low computational complexity, which measures the log-energy in high frequency discrete wavelet transform sub-bands.

The main shortcomings of existing no-reference sharpness metrics are a low robustness to variations in image content

and to other impairments present. The recent work of Feichtenhofer et al. [12] shows promising progress concerning robustness to content variations. They demonstrate a very good performance that is even competitive to full-reference methods, however, limited to still images.

II. Proposed Sharpness Metric

Since image blur is most noticeable at edges and their slopes, as discussed in related work (e.g. [7], [12]), we extend the idea of Feichtenhofer et al. [12] by considering specific properties of video for our novel sharpness metric. The proposed novel sharpness metric is designed to automatically provide one global measure that determines to which degree an image or image sequence appears in focus. We have chosen this approach because of its simplicity and low computational complexity, to make real-time sharpness estimation possible.

Basically, the sharpness algorithm measures the spread of edges detected by a Sobel filter for both derivatives (x and y direction), as originally proposed by Marziliano et al. [7]. The edge width is defined by the intensity variation of all pixels along the gradient, perpendicular to the edge. The overall image sharpness is calculated by a block-based pooling of local sharpness values of the most significant blocks. The contributions of this work are extensions of the metric for videos and the consideration of other impairments, i.e. noise and interlacing artifacts, in order to minimize their influence to the sharpness result. The measure is a predictor of perceived image sharpness, with focus on only the sharpest areas in the image, utilizing the fact that humans tend to rate sharpness based on the sharpest image regions as well.

A. Robust sharpness metric

The algorithm is applied separately to both fields of the video input images. Due to field sub-sampling, horizontal edges may appear to be sharper. To avoid the influence of very sharp horizontal edges only significant edges with approximately vertical orientation, covering a tolerance angle, are considered. As we operate on fields, interlacing artifacts are avoided and do not impact the sharpness results. In order to minimize the influence of noise (the initial detector result tends to be higher at presence of noise) a median filter is applied to the gray scale input image. The two sharpness values obtained by applying the algorithm on both fields are combined by averaging.

In order to extract edges from the input image, at each pixel the image gradient magnitude is calculated and an adaptive threshold is applied. The adaptive threshold is obtained by using a 0.80 quantile of the gradient histogram H_x while ensuring a minimum threshold τ_{min} for extracted edges. To get a binary edge image, a thinning process is performed by using a second lower threshold, as fraction of the higher one.

A minimum threshold τ_{min} of 55 is set in order to avoid edge maps that are overfilled instead of returning only significant ones. By combining the adaptive and absolute threshold, using

$$I_E(x,y) = \begin{cases} 1, & \text{if } |\nabla(x,y)| \geq \tau, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $|\nabla(x,y)|$ defines the magnitude at each position and $\tau = \max(\tau_{min}, q_{0.8}(H_x))$, we obtain edge images that include a well suited amount of edges independent of the content.

These obtained edge pixels are further used as measuring points and are tested in order to fulfill some requirements. Due to the fact that we operate on fields, horizontal edges could be distorted and do not longer correspond with the original. The gradient orientation of each edge pixel has to be vertical and within a predefined angle tolerance $\alpha_{tol} \in [0°, 45°]$. Noise tends to produce many short and sharp edges within an area, which has an impact on the robustness of the metric. To avoid that the metric is measuring noise edges instead of real edges, each edge pixel has to be associated with an edge with a minimum length (vertical extension), relatively to the actually existing edge lengths l_{min}. Edges with a high contrast are perceived sharper by observers while those with a low contrast might not be even recognized. Therefore, at each edge pixel the contrast of the appropriate edge is tested and has to exceed a fixed predefined minimum value c_{min}.

At every measuring point where the previously mentioned constraints are fulfilled, the edge width is defined by the maximum intensity variation of all pixels along the gradient, perpendicular to the edge. At each remaining edge pixel we compute the edge width by using Equation 2. Since we have applied a median filter at the beginning, the abortion criterion to find the local extrema points has been adapted. If the sign of the gradient is known, the direction, where to search for minimum and maximum perpendicular to the edge is given. For edges with a positive gradient a considerable minimum or maximum of intensity is found by using

$$min(x_{M+i}) = \begin{cases} 1, & \text{if } \; |(|\nabla(x_{M+i-1},y)| - \\ & \quad |\nabla(x_{M+i},y)|)| < \frac{x_M}{4}, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $i \in \mathbb{N}^+$ and X is the intensity vector containing all horizontally neighboring intensities of one edge pixel, so only horizontal rows of the image are taken, $x_M = \text{argmax}_{x_j \in X} |\nabla(x_j, y)|$ is the gradient at the edge pixel, so the maximum gradient of the whole intensity vector X. $max(x_{M-i})$ is defined analogously by plugging $-i$ into Equation 2, and for edges having a negative gradient a local minimum and maximum is found analogously.

This means, a local minimum or maximum is found, when the descent of the edge slope is in/decreased by more than 25% of the maximum gradient of the actual edge. Finally, we compute the edge width for each edge pixel as the distance between minimum and maximum. In order to ensure accuracy of the edge width, the width is corrected by the angle and considering sub-pixel accuracy. Motivated by the observation

that the HVS perceives edges with high contrast as sharper and vice versa, we have refined our metric by adapting the measured edge width dependent on the contrast. This means, edges with a high contrast will be reduced, which consequently results in a higher sharpness and vice versa.

Finally, the field is divided into 32×32 pixel sized blocks to deduce local sharpness values. For each block a representative edge width is calculated statistically from the containing values by requiring a minimum amount of containing edges for robustness. In order to keep the influence of edge locations small we use overlapping blocks. We found that spatial overlaps between 20% and 30% provide the best and most robust results (for the final algorithm 25% has been chosen). A predefined fraction of sharpest blocks is finally contributing to the global image sharpness $s_k \in [0, 1]$, where k is the frame number.

B. Computation of confidence

As stated, the sharpness s_k is computed for each frame k from a video segment of arbitrary length. In order to avoid inappropriate sharpness values obtained by measuring black frames, fade in/out and other frames that might impact the global sharpness we introduce a confidence value $conf_k \in [0, 1]$. It provides the reliability of the sharpness measure for each frame k, where a confidence of 1 indicates highest reliability. In order to ensure robustness, only frames with high confidences will contribute to the final global measure for a video segment.

The value depends on the number of available, measurable edges, the consistency of sharpness over the whole image (represented by blocks) and the vertical extension of the edges. The confidence $conf_j$ is computed by using

$$conf_j = \alpha \frac{b_{meas}}{b_{all}} + \beta(1 - \frac{d_{blocks}}{d_{max}}) + \gamma \frac{e_{v_{max}}}{e_{v_{actMax}}} \qquad (3)$$

where b_{meas} are all blocks containing enough edge pixel to be measurable for our metric, b_{all} are all possible blocks including the overlap, d_{blocks} is the difference of the ordered measured block sharpness values between the first and the last of the finally selected, d_{max} is the chosen maximal difference of two block sharpness values, $e_{v_{max}}$ defines the actually measured maximum edge length in vertical extension and finally, $e_{v_{actMax}}$ is the pre-defined maximum value depending on the image resolution.

III. Evaluation

A. Metrics

For our subjective QA we have chosen to use two of the standard evaluation methods[1] in a slightly adapted way. Assessing individual video sequences in a continuous rating scale is traditionally popular for subjective multimedia quality evaluation. The setup of such an experiment is simple and its evaluation and analysis is straight forward. Thus, we are going to use the Absolute Category Rating (ACR) method in the first part of the experiment, by adapting the 5 to a 6 category scale. We decided to apply a second part to the experiment: Paired comparison is appropriate and efficient as well for the goal of our sharpness evaluation study. Minor differences in sharpness of video sequences may be hard to recognize and can be more effectively dealt with by a paired comparison method. Thus, the subjects were confronted with a Double Stimulus Continuous Quality Scale (DSCQS) method in a second part. Since we wanted to keep the duration of the experiment under 30 minutes and since we are interested in testing a diverse set of videos, we could not show a whole permutation of all possible pairs. 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 watched 32 videos in the first part and 28 videos in the second, comparative part. For half of the subjects we make use of an eye tracking system, as described later in this chapter.

[1]For an extensive classification, review and comparison see [13].

B. Creation of test material

Since, to the best of our knowledge, there are no ground truth databases publicly available containing MOS for varying blurred videos (in contrast to still images), as well as containing other impairments such as noise or interlacing artifacts, we were forced to establish our own reference database. For the subjective sharpness experiment we have considered high definition (HD) video sequences that are of interest in today's production systems, kindly provided by the Flemish public broadcasting organization VRT.

We have created video clips that all had lengths of approximately 8 seconds each. Subjects only need 7-8 seconds of video for forming their quality decision [14]. We have chosen video clips consisting of mainly one single scene, and in general showing constant sharpness over the whole scene. 15 video clips were chosen in order to evenly span a full range of available sharpness. The videos did not contain audio tracks.

The various levels of sharpness were artificially produced by scaling the original videos down to the respective technical resolution, e.g. 50% of 720p resolution (640x360 pixels), followed by an upscaling to the target resolution 720p (an analogously for sharpness levels 33% and 25%). This results in a reference database of 60 videos of 4 different sharpness levels (25 fps frame rate), 32 for the first and 28 videos for the second part. Both datasets spanned the same range of sharpness levels (100%, 50%, 33% and 25% of full resolution).

C. Setup of the experiment

The evaluation tool provides the ability to view the videos in randomized order and accepts discrete or continuous user ratings dependent on the configuration. The tool is used for both parts of the experiment. For the first part, we used a 6 category scale and instead of a moving slider we compromised to provide only one discrete scale for each shown video. Thus only one rating is obtained for a whole video clip. In the second part a pair of videos is shown where each subject

makes a comparative rating on a discrete 5 point scale. The ratings of each subject are collected and for each of the 60 videos a mean opinion score is obtained by computing the mean of all scores. The scales of part 1 and 2 are linearly transformed to be comparable to the sharpness value of the sharpness metric.

All 28 volunteers had near-perfect or corrected to normal vision, and were naive for the purposes of the experiment. 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 QA tasks of images and videos. Due to their answers the subjects were classified by their expert level in two groups, experts and consumers. 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 corresponding to a resolution of 25% of the first. After watching these two video clips the viewers were asked to place them in a range from 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. We get two values on a continuous scale, which can be used for normalizing and rescaling further ratings in order to alleviate the issues of single stimulus testing methods. The subjects were placed approximately 0.75 meters away from the display as proposed in [15]. 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 completing the entire test. The video clips were shown on a 32 inch Samsung LED TV series 6 screen with a native resolution of 1920×1080 pixels.

D. Eye tracking

As already mentioned earlier, half of all subjects judged the sharpness of the video clips while an eye tracking system was applied. For this experiment a SensoMotoric Instruments (SMI) *RED500* static eye tracker with a 500hz sampling rate was used. The eye tracker marks regions that have attracted the viewers' attention, since they have focused them for a certain time (gazing points). Consequently, we capture not only the human perception of sharpness of HD videos, but also their regions of attention during deciding about the sharpness score. Within this experiment, the obtained eye tracking data can be used for validating the novel sharpness metric (in terms of correlation of regions selected for judgment by the automatic metric) and on the other hand we know about the regions of attention, that subject use for judging.

IV. Results

A. Single stimulus results

As mentioned in Section III-B, we have established our own test videos by blurring HD videos with different filter kernel sizes. Thus, the ground truth sharpness is available. Before comparing with the MOS, we have tested our novel sharpness metric for accuracy and consistency by comparing the sharpness results with the ground truth, grouped into 4 different levels of sharpness. The objective sharpness metric is robust and clearly distinguishes between the sharpness levels. 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 highly correlated with the ground truth, since we achieved a Pearson correlation coefficient of 0.856 and a Spearman rank correlation coefficient of even 0.926.

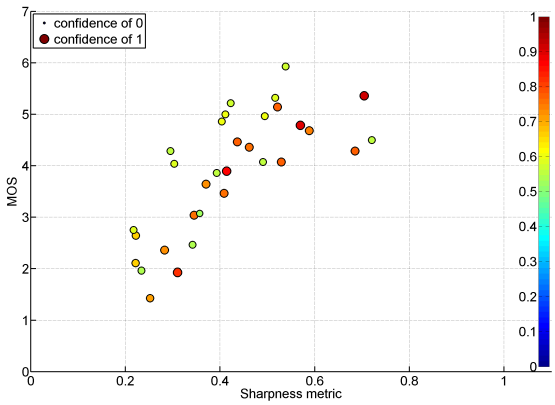

Fig. 1. Comparison of MOS given by 28 subjects and obtained in the single stimulus part of the experiment.

For the ACR, which was used in the first part of the experiment, we can easily compare the subjective perception of all participants with the sharpness metric results. The MOS were obtained by averaging the scores of all 28 users, resulting in a range from 1 to 6. By mapping the MOS to the automatically generated sharpness metric results we obtain the illustration in Figure 1 presenting the evaluation of the first part of the experiment. On the x axis, the results of the automatic sharpness metric is shown and on the y axis the corresponding MOS are listed. The sharpness metric additionally calculates a confidence value, which was introduced in Section II-B. The more edges are present and the more blocks can be used for measuring the sharpness, the higher the confidence. The confidence is visualized by the size of the data points and by the color using the given color bar. The comparison shows, that the sharpness metric is well correlated with subjective perception, resulting in a Pearson correlation coefficient of 0.738 and a Spearman rank correlation coefficient of 0.786.

B. Double stimulus results

In the second part of the experiment, the subjects judged the perceived difference of two subsequently presented video clips. Instead of showing all possible permutations of pairs of the second pool of videos, we just show each video once, in random pairs, until all videos have been shown. Since we had 28 videos for the second part, several of the randomly connected pairs were only shown to one viewer, so that for some of the pairs only one score is available. Since comparative MOS were required, the ratings on the 5

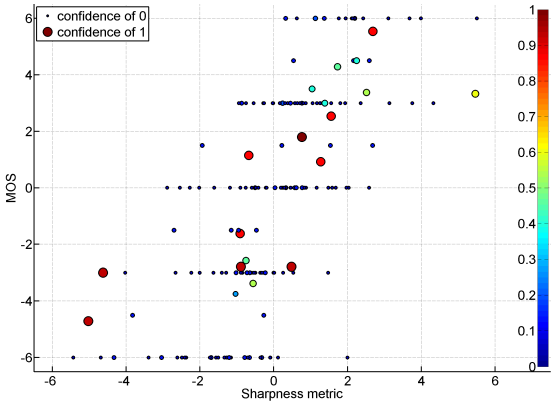

Fig. 2. Comparison of double stimulus MOS to sharpness.

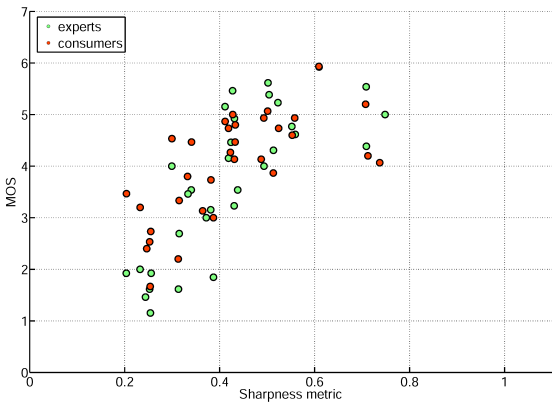

Fig. 3. Comparison of MOS obtained in the single stimulus part of the experiment and grouped in experts and consumers.

point scale were linearly mapped to the possible changes of sharpness and scaled to a range from -6 to 6. For example, if the first clip was less sharp than the second the subject's rating is negative and vice versa. All ratings given for the same video pairs are averaged.

In order to test if the comparative MOS correlates with the performance of the sharpness metric, those values were transformed in the same way. In Figure 2 the results of the DS part are given, where each point corresponds to one specific pair of videos. The confidence is computed by the number of available user ratings and is visualized again by the colors and size of the data points. For the comparative judgments we report a Pearson correlation coefficient of 0.683 and a Spearman rank correlation coefficient of 0.721.

Several single scores have outlier character. Since many points have a very low confidence we investigated only the sharpness changes of the videos independent of the content. All video pairs with equal change of sharpness level are summarized. Since we have investigated the robustness and consistency of the sharpness metric and showing that the results are well correlated with the ground truth, we compare these values with the effective sharpness change, computed by using the ground truth. The resulting Person correlation coefficient of the summarized MOS by sharpness difference with the ground truth is 0.920 and the Spearman rank correlation coefficient is 0.936.

Using a DS method for QA is not well studied and the evaluation process has to be carefully designed when setting up the experiment. The selection of pairs to be shown cannot be random, or the number of allowed video clips is limited. However, with using the DS method, we could show, that the subjects on average were able to distinguish between different levels of sharpness.

C. Analysis of MOS using eye tracking

The eye tracking system has tracked the subjects' eyes and consequently their fixations over the whole experiment. Thus, for each frame of each shown video the fixations can be visualized by a colored circle. In this way, all fixations are easily assignable to the subjects and the information obtained by the questionnaire (as described in III-C). Some observations are quite general and intuitively expected, such as that moving objects and objects entering the scene naturally attract the attention of viewers. The regions of interest are mainly related to the central parts of the frame and, very importantly for us, they do not change significantly for varying sharpness levels.

From the gazing points differences between subjects can be recognized. For example, some subjects focus on faces, while others do not. We can find that viewers with higher expert level tend to focus not on objects relevant for the scene, but rather on highly textured areas, that make it easier to judge sharpness. This lead us to the idea of evaluating the data separately by two groups, experts and consumers. 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.

In Figure 3 the ratings of the first part of the experiment are visualized slightly different from Figure 1, by focusing on the expert level of each subject. When comparing the sharpness metric results with the experts' MOS only, the correlation increases (Pearson: 0.779, Spearman: 0.807), while for the consumer group, the correlation decreases (Pearson: 0.663, Spearman: 0.691). The hypothesis about significant differences between subjects of different expert level could not be confirmed from the ratings. However, some of the outliers are related to the consumer group, which supports our assumption, that subjects with less experience in quality tasks have difficulties in distinguishing minor differences of sharpness. In particular, it seems to be difficult for consumers to distinguish higher sharpness levels. Note that experienced subjects tend to rate video clips with a lower sharpness level more critically, while their ratings for sharper material are more accurate, since obviously, they are able to recognize minor differences. The comparative ratings of the DS part were grouped into experts and consumers as well. Again, the correlation for experts increases to a Pearson correlation coefficient of 0.954 and a Spearman rank correlation coefficient of 0.973, in contrast to a decreasing consumers' correlation (Person: 0.839 and Spearman: 0.900). Compared to the initial (not grouped by

expert level) values (Pearson: 0.920, Spearman: 0.936), this indicates a better ability to distinguish between several levels of sharpness for the experts.

Similarly to the gazing regions used by the expert group, our sharpness metric is designed to measure the sharpness at edges and textures. Furthermore it identifies the sharpest regions, since focusing on the sharpest blocks is necessary for content with e.g. out of focus regions. We have tested the overlap of regions selected by the sharpness metric and fixations of the subjects. We have visualized all blocks that are used by the objective sharpness metric for computing the sharpness of an image. The algorithm selects those blocks that contain the sharpest edges, assuming that humans focus on the sharpest regions. The hypothesis can mostly be confirmed since experts truly focus on sharp edges and textures.

D. Discussion of results

According to these results the accuracy of the MOS can be interpreted as dependent on the expert level, which was also supported by the analysis of the eye tracking data. A small classification error needs to be considered since the expert level is a matter of self-assessment and subjects may have been misplaced. Asking subjects for a calibration judgement on two sequences in the beginning did not allow scaling of the scores in the first part of the experiment, but confirmed the assumption that the single stimulus method presents susceptibility to range effects as reported 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. Diverging user scores may be investigated separately instead of combining them to overall MOS. However, obtaining confidence information in paired comparison-based tests has not been sufficiently studied. Overall, the paired comparison method has potential for subjective tests but (in contrast to other test methodologies) its theoretical and practical frameworks have not been investigated sufficiently in the field of multimedia QA [16]. However, we could show that the comparative rating were better done by the experts rather than by the consumers.

V. Conclusion

We have presented an improved no-reference sharpness metric for video, designed to correlate well with human perception. It shows state-of-the-art performance and resolves some well known issues in existing objective QA metrics. Its major advantages are a more accurate sharpness prediction and a lower susceptibility to diverging image content, as well as more robustness under presence of noise and interlacing. A subjective experiment was performed using two common methodologies, ACR and DSCQS, where subjects had to judge two sets of videos varying in their level of sharpness, down sampled from content originally in HD. The evaluation confirmed a high correlation between the algorithm and the subjects' sharpness assessments. Based on eye tracking data, we assume that the selected edges and further the perceived sharpness mainly correspond with those that humans would select. The gazing locations are clearly dependent on the viewer's experience level. Consumers tend to 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.

Acknowledgment

This work has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013), grant n° 287532, TOSCA-MP, and Horizon 2020 research and innovation programme, grant n° 761802, MARCONI.

References

[1] 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," Proc. SMPTE Tech. Conf., 2009.

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

[3] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on just-noticeable blur and probability summation," IEEE ICIP, 2007.

[4] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *Image Processing, IEEE Transactions on*, vol. 15, no. 11, pp. 3440–3451, 2006.

[5] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb)," *Image Processing, IEEE Transactions on*, vol. 18, no. 4, pp. 717–728, 2009.

[6] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *Image Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2678–2683, 2011.

[7] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Applications to jpeg2000," Signal Proc.: Image Comm., vol. 19, pp. 163–172, 2004.

[8] H. Sheikh, A. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: Jpeg2000," *Image Processing, IEEE Transactions on*, vol. 14, no. 11, pp. 1918–1927, 2005.

[9] Z. Wang, E. P. Simoncelli, and H. Hughes, "Local phase coherence and the perception of blur," *NIPS*, pp. 786–792, 2003.

[10] R. Hassen, Z. Wang, and M. Salama, "No-reference image sharpness assessment based on local phase coherence measurement," *IEEE ICASSP*, pp. 2434–2437, 2010.

[11] P. V. Vu and D. M. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Processing Letters*, vol. 19, no. 7, pp. 423–426, 2012.

[12] C. Feichtenhofer, H. Fassold, and P. Schallauer, "A perceptual image sharpness metric based on local edge gradient analysis," Signal Processing Letters, IEEE, pp. 379–382, 2013.

[13] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, "Objective video quality assessment methods: A classiffcation, review, and performance comparison," in *Broadcasting, IEEE Transactions on*, 2011.

[14] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," *SPIE VCIP*, pp. 8–11, 2003.

[15] "BT.500-11 - Methodology for the subjective assessment of the quality of television pictures," 2002.

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