Empirical Evaluation of No-Reference VQA Methods on a Natural Video Quality Database

Hui Men, Hanhe Lin, Dietmar Saupe
Department of Computer and Information Science, University of Konstanz, Germany
Email: {hui.3.men, hanhe.lin, dietmar.saupe}@uni-konstanz.de

***Abstract*—No-Reference (NR) Video Quality Assessment (VQA) is a challenging task since it predicts the visual quality of a video sequence without comparison to some original reference video. Several NR-VQA methods have been proposed. However, all of them were designed and tested on databases with artificially distorted videos. Therefore, it remained an open question how well these NR-VQA methods perform for natural videos. We evaluated two popular VQA methods on our newly built natural VQA database KoNViD-1k. In addition, we found that merely combining five simple VQA-related features, i.e., contrast, colorfulness, blurriness, spatial information, and temporal information, already gave a performance about as well as those of the established NR-VQA methods. However, for all methods we found that they are unsatisfying when assessing natural videos (correlation coefficients below 0.6). These findings show that NR-VQA is not yet matured and in need of further substantial improvement.**

***Keywords*—empirical evaluation, no-reference, video quality assessment, feature combination**

I. Introduction

Objective Video Quality Assessment (VQA), automatically quantifying the quality of videos without human judgments, can be classified into three types: Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR). FR-VQA and RR-VQA methods now have achieved promising performance, although hard to realize in some applications because of the full or partial availability of the corresponding reference. Therefore, NR-VQA methods have more applicability and difficulty simultaneously, since they must predict video quality without a reference. Several NR-VQA methods were proposed and achieved good performance. However, these were evaluated on video databases with content that was artificially distorted from pristine sources. Therefore, there arises a question: will they still perform well on natural videos?

In this paper, we evaluate two of the most popular NR VQA methods (i.e., Video BLIINDS and VIIDEO) on a natural VQA database. Additionally, we propose a NR-VQA model combining five features of a video, including contrast, colorfulness, blurriness, Spatial Information (SI) and Temporal Information (TI), which outperforms VIIDEO, and has performance close to Video BLIINDS.

II. The Natural Video Quality Database

The experiments in this paper were implemented on our newly built natural VQA database KoNViD-1k consisting of 1,200 public-domain video sequences [1]. The videos in this database were fairly sampled from a large public video dataset, Flickr's YFCC100M, which contains 793,436 Creative Commons licensed videos. Their associated subjective Mean Opinion Scores (MOS) of 5-level absolute category rating scale ranging from "Bad" to "Excellent" were obtained by crowdsourcing using the Crowdflower.com platform. Examples of videos with "Excellent", "Good", "Fair", "Poor", and "Bad" quality came first as an anchor. To rate a video, workers needed to click a "Play" button to start the video. After the video was finished playing, the rating scale was displayed and workers selected one of the five categories to proceed.

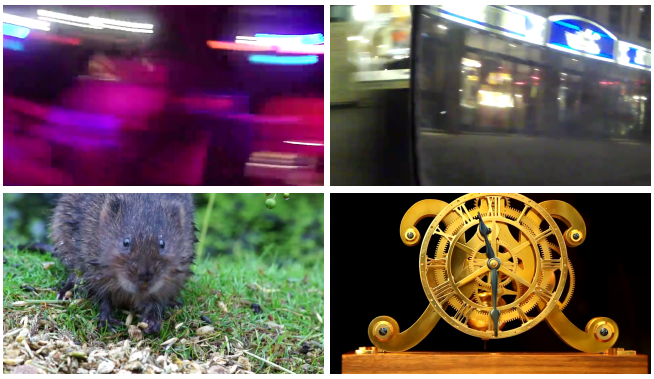

Fig. 1: Example frames of several video sequences contained in the natural VQA database. Top left: MOS 1.26 (lowest MOS in this database). Top right: MOS 1.66. Bottom left: MOS 4.4. Bottom right: MOS 4.64 (highest MOS in this database).

Crowdsourcing was chosen for subjective quality assessment because due to the size of the database, lab studies are not feasible and it had been shown in a previous study that MOS values gained by crowdsourcing can be reliable, more economical, efficient, and effective than scores collected in lab studies [2].

Unlike traditionally used databases for VQA, our database ensures the naturalness of the videos by not including any artificially distorted video versions as well as by filtering unnatural ones (e.g., screen recording, slide shows, stop-motion videos, etc.). A few example frames are shown in Fig. 1. For more details on the natural VQA database KoNViD-1k see [1].

III. No-Reference VQA Methods

Video BLIINDS [3] is a NR-VQA method making use of two types of features. One is related to motion, namely motion coherency and egomotion. The other type of features is based

QoMEX2017 – Erfurt, Germany; 978-1-5386-4024-1/17/$31.00 ©2017 IEEE

on a model for natural video statistics and includes spatial features, temporal information, and frame-based NIQE features [4]. In order to predict video qualities, Video BLIINDS uses these features to train a linear kernel support vector regressor on the LIVE VQA database [5].

VIIDEO [6], a completely blind VQA method, does not require any pristine reference video or human judgments for training. The method assumes that for a video of good quality, its local statistics of frame differences processed by local mean removal and divisive contrast normalization, resulting in the Transformed Frame Differences (TFD), should follow a generalized Gaussian distribution. Furthermore, products of pairs of adjacent coefficients in TFD computed along horizontal, vertical and diagonal spatial orientations are observed to follow an asymmetric generalized Gaussian distribution. Besides features based on these two assumptions, VIIDEO further models inter sub-band correlations over local and global time spans to extract another feature. The resulting quality prediction then is completely blind.

IV. Feature Combination VQA Model

Intuitively, contrast, blurriness, colorfulness, and motion may be expected to have strong impact on video quality from a visual perceptual aspect. Therefore, we define a multi-feature vector to blindly predict video quality by combining five features, i.e., blurriness, colorfulness, contrast, SI, and TI.

The blurriness of a frame is assessed by measuring the probability of blur based on the distribution of edge widths [7]. To reduce the computational cost, we apply it on one frame out of every second over the entire video. Regarding colorfulness, with the RGB channels of a frame as matrices R, G, and B, one computes two matrices $\mathrm{rg} = \mathrm{R} - \mathrm{G}$ and $\mathrm{yb} = \frac{1}{2}(\mathrm{R} + \mathrm{G}) - \mathrm{B}$, and defines the final metric as $\sqrt{\sigma_{\mathrm{rg}}^2 + \sigma_{\mathrm{yb}}^2} + \frac{3}{10}\sqrt{\mu_{\mathrm{rg}}^2 + \mu_{\mathrm{yb}}^2}$, with σ^2 and μ being corresponding variances and means [8]. The average over all frames yields the colorfulness of a video. The contrast of a video is computed as the average frame contrast, which can be simply measured by the standard deviation of pixel gray-scale intensities [9]. The SI of a video is computed by averaging standard deviation over all frames' SI, each of which is obtained by applying a Sobel filter to extract the gradient magnitude and then computing its standard deviation [10]. Similar to SI, the TI is the mean of frame-wise standard deviations of pixel-wise frame difference [10].

In our Feature Combination (FC) model, we extract the aforementioned five features from each video and use a regression model for prediction.

V. Experimental Results

Based on the predicted scores and MOS gained by human judgments, four criteria are adopted to evaluate the VQA methods, namely the Pearson Linear Correlation Coefficient (PLCC), the Spearman Rank Order Correlation Coefficient (SROCC), the Kendall Rank Order Correlation Coefficient (KROCC), and the Root Mean Square Error (RMSE). Note that for RMSE, larger values imply worse precision while for the other three criteria (which range from $[-1, 1]$), the larger their values are, the better the prediction performance is.

For the evaluation of the video BLIINDS and our FC model, we extracted two different feature descriptors (46 features for video BLIINDS, and 5 features for FC model) for each video. A Support Vector Regression was used to predict objective scores, where the final performances were obtained by k-fold (k = 5) cross-validation with 10 repetitions. Note that both methods, video BLIINDS and the FC model, were trained on the natural video database.

For VIIDEO, the totally blind VQA method, which directly predicts video quality by extracting statistical features without a training phase based on human judgments, we just applied it to directly obtain the predicted scores.

Empirical results in Table I show that video BLIINDS gave the best performance with the highest PLCC, SROCC, KROCC and the lowest RMSE. Our simple FC Model, however, was comparable. On the other hand, VIIDEO clearly showed a significantly lower performance. Note that Video BLIINDS and VIIDEO showed much better performances on the LIVE Database than on natural videos.

TABLE I: NR VQA performance comparison.

	Natural VQA Database			LIVE Database	
	V.BLIINDS	VIIDEO	FC Model	V.BLIINDS [3]	VIIDEO [6]
SROCC	**0.572**	0.031	0.472	0.759	0.624
PLCC	**0.565**	-0.015	0.492	0.881	0.651
KROCC	**0.401**	0.020	0.330	-	-
RMSE	**0.526**	0.639	0.556	-	-

VI. Conclusion and Future Work

The evaluations of two popular VQA methods showed that assessing quality for natural videos is far more difficult and complicated than for manually distorted ones. These two methods partially rely on measuring the difference between the statistical features of the tested videos and the ones of ideal quality. The prediction on a database with artificially distorted test items achieved rather satisfying performance as reviewed in TABLE I. However, real videos in the wild usually contain more than one type of distortion, and automatic quality assessment becomes more challenging, yielding unsatisfactorily low correlation coefficients. Our FC model which extracts features from videos themselves without referring to an idealized model of naturalness thereby avoids the problems that VIIDEO is facing with natural imagery. Despite its simplicity it already achieved performance near that of Video BLIINDS and may serve as a promising starting point for further development.

In future work we will extend our current model by adding features such as noisiness and sharpness, which will further increase the correlation with subjective scores. For VQA of natural videos like those in KoNViD-1k visual aesthetics and how much people "like" a given content may have an influence. In fact, in currently ongoing work for image quality assessment (IQA) it could be shown that including specific aesthetics features in NR-IQA algorithms can improve performance. Video popularity ("likes") are included in the KoNViD-1k database and will also be considered in future work.

Acknowledgment

We thank the German Research Foundation (DFG) for financial support within project A05 of SFB/Transregio 161.

REFERENCES

[1] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The Konstanz natural video database (KoNViD-1k)," in *QoMEX 2017: 9th International Conference on Quality of Multimedia Experience*, 2017.

[2] D. Saupe, F. Hahn, V. Hosu, I. Zingman, M. Rana, and S. Li, "Crowd workers proven useful: A comparative study of subjective video quality assessment," in *QoMEX 2016: 8th International Conference on Quality of Multimedia Experience*, 2016.

[3] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[4] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[5] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[6] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2016.

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

[8] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Electronic Imaging 2003*. International Society for Optics and Photonics, 2003, pp. 87–95.

[9] E. Peli, "Contrast in complex images," *JOSA A*, vol. 7, no. 10, pp. 2032–2040, 1990.

[10] ITU-T, "Subjective video quality assessment methods for multimedia applications," ITU-T Recommendation P.910, 2008.