Erschienen in: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). - Piscataway, NJ : IEEE, 2017. - ISBN 978-1-5386-4024-1
https://dx.doi.org/10.1109/QoMEX.2017.7965673

The Konstanz Natural Video Database (KoNViD-1k)

Vlad Hosu[1], Franz Hahn[1], Mohsen Jenadeleh[1,2], Hanhe Lin[1], Hui Men[1], Tamás Szirányi[3], Shujun Li[4], Dietmar Saupe[1]
[1]Department of Computer and Information Science, University of Konstanz, Germany
[2]Faculty of Computer Science and Engineering, Shahid Beheshti University, G. C, Tehran, Iran
[3]Institute for Computer Science and Control, Hungarian Academy of Sciences, Hungary
[4]Department of Computer Science, University of Surrey, United Kingdom

***Abstract*—Subjective video quality assessment (VQA) strongly depends on semantics, context, and the types of visual distortions. Currently, all existing VQA databases include only a small number of video sequences with artificial distortions. The development and evaluation of objective quality assessment methods would benefit from having larger datasets of real-world video sequences with corresponding subjective mean opinion scores (MOS), in particular for deep learning purposes. In addition, the training and validation of any VQA method intended to be 'general purpose' requires a large dataset of video sequences that are representative of the whole spectrum of available video content and all types of distortions. We report our work on KoNViD-1k, a subjectively annotated VQA database consisting of 1,200 public-domain video sequences, fairly sampled from a large public video dataset, YFCC100m. We present the challenges and choices we have made in creating such a database aimed at 'in the wild' authentic distortions, depicting a wide variety of content.**

***Keywords*—Video database; authentic video; video quality assessment; fair sampling; crowdsourcing.**

I. Introduction

Most of the Internet traffic today stems from user-generated videos on sharing web-sites and social networks. Video sequences pass through several stages of processing before they reach consumers, which often deteriorate visual quality. Moreover, the vast amount of user-generated video content and the increased diversity of end user devices (ranging from smaller and power-constrained mobile devices to large displays such as 4K Ultra HDTVs and TV walls) calls for a broad range of video quality to be supported. Adapting video quality to different use cases has become an important topic for researchers, content providers and distributors [1].

Automatic and accurate prediction of video quality is a basic operation for many video processing applications such as video quality monitoring in transmission protocols, video quality filtering in sharing services, automatic and recommended camera parameter settings during video capturing, and video enhancement. Specifically, no-reference methods attempt to judge the quality of a video sequence without any additional information about the original recorded scene. Such blind methods may apply machine learning techniques to learn from large amounts of annotated data. However, current video quality assessment (VQA) databases contain only a small number of video sequences with little content diversity, thus offering limited support for designing and evaluating no-reference VQA methods effectively and fairly.

Additionally, these databases were mostly designed to include only artificially distorted video sequences to simulate quality loss in compression, transmission, and other parts of the video processing and distribution pipeline. Some databases capture imagery with a variety of cameras to encompass authentic video acquisition distortions, however, with content restricted to a small number of physical scenes.

Winkler [2] proposed several criteria for quantitative comparisons of source content, test conditions, and subjective ratings, applying them to 27 image and video databases. Most collections have not been found satisfactory in terms of content range and uniformity. Only few databases showed good uniformity for test conditions (image/video quality), but not over the whole quality range. Also the distortion variety was found lacking in most databases covering mainly compression and transmission, but not the many other types of natural distortions found "in the wild" [3].

To overcome these limitations we introduce KoNViD-1k, a large publicly available database of video sequences based on YFCC100m (Yahoo Flickr Creative Commons 100 Million) dataset [4] with a diverse set of video content. In this paper we report the filtering mechanisms and sampling procedures necessary to construct high-quality VQA databases of this kind, focusing on their usefulness in a variety of applications.

In the next section, we describe the database creation procedure and the set of attributes we have considered to maximise its diversity. Additional information regarding the removal of non-natural video sequences and sampling techniques are provided as well. Next, in Sec. III, we review our crowdsourcing-based process of collecting subjective mean opinion scores (MOS) and detail our results as well as crowd worker statistics. In Sec. IV we relate our database characteristics and creation methodology with other existing works and outline the differences, before discussing conclusions of our work and considering possible future work.

II. Database creation

Starting with a large initial collection of video sequences, the goal of our database was to ensure diversity in several dimensions that could impact video quality. Thus, after removing outliers, we sampled across six attributes that were calculated on the entire collection, in a more uniform manner to ensure diverse coverage of selected samples. We call this methodology "fair sampling".

A. Video collection

We began with a well-known public database of video sequences and images, YFCC100m [4], containing 793,436

Creative Commons (CC) licensed videos. We selected videos based on practical requirements, such that they:

- Were still available for download
- Played at more than 15 frames per second (FPS)
- Lasted longer than 8 seconds
- Did not have a "No Derivative Works" CC attribute
- Had a resolution higher than 960×540 (W $\times$ H)
- Were in landscape layout

This filtering yielded a subset of 144,889 videos for further processing. The entire collection was downloaded from the Flickr servers for later processing. For longer videos, only the first 30 seconds were stored.

B. Attribute computation

Winkler [2] suggested three attributes related to temporal, color, and spatial aspects to measure the content diversity of video databases. Guided by his research, we chose six video attributes and used them not only for the analysis of the resulting database, but for its creation as well. For each attribute we relied on the best-performing technique available in the literature (to the best of our knowledge). All attributes, except the one related to colorfulness, were computed on grayscale frames.

Since the computational complexity of some of the selected metrics is high relative to the number of frames to be analyzed, we created cropped and scaled versions of the videos to run them on. In some cases, only a subset of frames were used for processing.

1) Blur: The blur of a frame was assessed by the cumulative probability of blur detection (CPBD) metric [5]. Intuitively, the technique measures the probability of blur based on the distribution of edge widths. The CPBD metric works on individual frames. We applied it to videos by averaging the CPBD values on one frame every second over the entire duration of the video. We made this choice because of the high computational cost of running the blur measure.

2) Colorfulness: For this attribute we used Hasler and Suesstrunk's metric reported in [6]. With the RGB channels of a frame as matrices R, G, and B, one computes two matrices $\mathrm{rg} = \mathrm{R} - \mathrm{G}$ and $\mathrm{yb} = \frac{1}{2}(\mathrm{R} + \mathrm{G}) - \mathrm{B}$. Then, the metric is calculated as $\sqrt{\sigma_{\mathrm{rg}}^2 + \sigma_{\mathrm{yb}}^2} + \frac{3}{10}\sqrt{\mu_{\mathrm{rg}}^2 + \mu_{\mathrm{yb}}^2}$, where $\sigma_\cdot^2$ and $\mu_\cdot$ denote the variance and mean of the values in their respective matrices. Finally, the average value over all frames yields the colorfulness metric of a video.

3) Contrast: Frame contrast was measured simply by the standard deviation of pixel grayscale intensities [7]. The average frame-level standard deviation then gave the contrast of a video.

4) Spatial information: The spatial information (SI) was obtained by applying a Sobel filter to each frame to extract the gradient magnitude for each pixel and then computing its standard deviation [8]. The average standard deviation over all frames yielded the SI of the video.

5) Temporal information: Similar to SI, the temporal information (TI) is the mean of frame-wise standard deviations of pixel-wise frame difference [8].

TABLE I: Attribute thresholds for filtering outlier videos.

	Attribute	Lowest value	Highest value
1	Blur amount	0.05	0.88
2	Colorfulness	4.37	123.00
3	Contrast	7.51	97.48
4	Spatial information	7.70	187.76
5	Temporal information	3.07	56.81
6	VNIQE	3.58	23.08

6) Video quality: We used the Natural Image Quality Evaluator (NIQE) [9] as a proxy to assess video quality by computing the mean NIQE value of all frames. This method (VNIQE) does not require knowledge of distortion types nor quality ratings for training. The NIQE of a frame is simply the distance between certain ideal features and a particular frame's features. It can be interpreted as the degree of frame-level deviation from naturalness, according to the NIQE model.

C. Filtering

Based on the six attributes, videos were selected such that they are suitable for VQA. We removed videos depicting non-natural scenes like screen recordings or stop-motion sequences, as well as overly dark or bright videos.

1) Filtering extremes: Most of these situations were encountered at extremes of the attribute values. For instance, very low TI videos were often found to be screen-text recordings due to little change between frames. Dark videos have low contrast and SI. Uniformly and brightly colored videos have a high level of colorfulness.

Therefore, we decided to remove videos that have extreme attribute values. The filtering thresholds were empirically chosen based on a qualitative inspection such that most of the filtered videos show obvious artificial content. The selected thresholds are available in Table I.

2) Filtering stop-motion videos: With regard to removing stop-motion videos, we relied on two observations; namely they show periodical changes in TI, while a high percentage of consecutive frames show no change at all. We used the difference of TI of consecutive frames to quantify both factors.

With respect to the periodicity, we found local maxima of TI with an inter-peak distance greater than 1 second. By computing the distances d_i between consecutive local maxima together with their mean μ and variance σ^2, we extracted a measure of regularity as $R = \mu/\sigma^2$. If the variance σ^2 is low, then the regularity R is high. The mean of the distances is the average spacing between peaks. If the peaks are further apart the formula tolerates smaller changes in the peak timing, and R is higher. $R > 1/300$ was found to be a good indicator for detecting stop-motion videos. Secondly, if more than 30% of consecutive frames showed no difference in TI, we assumed to be dealing with a non-natural video sequence.

These criteria further eliminated about 500 videos from our video collection, leading to a database of 124,865 videos, which we call KoNViD-125k. The distributions of the normalized attribute values are displayed in Fig. 1 on the left.

D. Sampling

Our eventual goal is to sample a set of 10,000 diverse videos from the KoNViD-125k filtered collection and to have

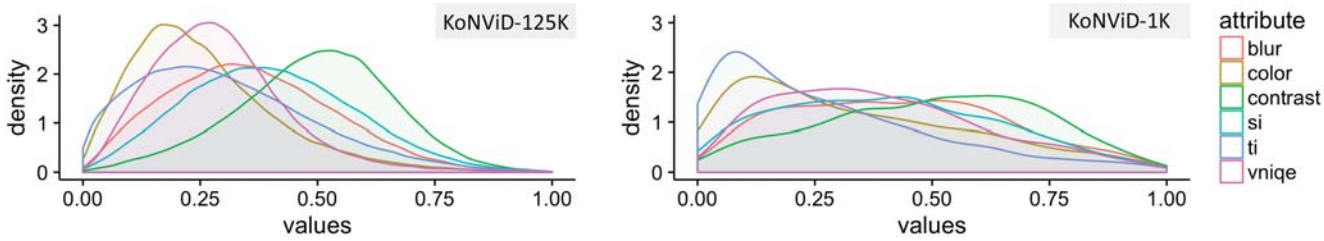

Fig. 1: Distribution of attribute values over the larger filtered dataset and the sampled 1,200 videos.

the sample subjectively annotated by human observers. In this paper, we focus on what sampling procedures we should use and understand how well they would work for constructing VQA databases of this kind.

We devised a "fair-sampling" strategy with which we generated a subset of 10,000 videos. From the resulting set we took a random collection of 1,200 videos, which forms the KoNViD-1k database [10]. We performed subjective studies to assess the visual quality of the videos in this database (see Sec. III for details). Consequently, we arrived at a better understanding of the diversity of the 10,000 fair-sampled videos, and the efficiency of our approach.

A "fair sampling" mechanism should produce a broader diversity of video properties than a random sampling mechanism. Our videos are represented as points in a 6-dimensional attribute space. We can think of the KoNViD-125k collection as a sample of $M = 124,865$ points of a multivariate and approximately normal distribution, for which random subsampling of 10,000 items would yield a subset with a similar normal distribution. The sampling procedure is engineered to give a more uniform 6-dimensional sample distribution.

Each attribute relates to a particular subjective property. Most videos having extreme values for one or more attributes show severe quality degradations. Random sampling is unlikely to select these "unusual" videos, which are as important as the "normal" ones. A preliminary qualitative inspection of several hundred videos (both randomly and fairly sampled) suggested that our strategy creates a balanced mix of videos.

With respect to the sampling procedure, the method of Vonikakis et al. [11] can ensure a uniform distribution for each attribute independently. However, we are also interested in sampling videos with joint distortions, having extreme values in several attributes simultaneously. Thus, we applied a different approach.

Note that our attributes are correlated as shown in Fig. 2. For instance, contrast and spatial information (SI) have a 0.62 correlation, whereas VNIQE and SI have a negative correlation of -0.43. Thus, as a preprocessing step we applied a principal component analysis (PCA), that shows that 37.1%, 57.7%, 73.4%, 85.8%, and 95.2% of the variance of the data is explained by the 1st to 5th principal components, respectively. We decorrelated the attributes by taking five components, maintaining all but 5% of the signal energy.

One way to solve the subsampling problem is to design a sampling method that favors videos that are part of a low-density region in the attribute space as follows.

Let $\rho : \mathbb{R}^5 \to \mathbb{R}$ denote the probability density function of

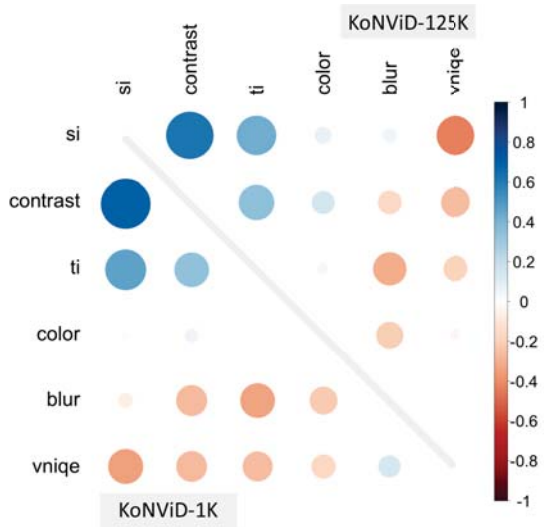

Fig. 2: Correlations between attribute values. Correlations considering KoNViD-125k dataset are above the main diagonal, and for KoNViD-1k below. Larger and darker circles represent a stronger absolute correlation coefficient. Red hues encode negative correlations, whereas blues are positive.

the 5-dimensional PCA attribute vectors v_i for natural videos, estimated from a video collection such as our KoNViD-125k. We used a k-NN method to estimate the density in the neighbourhood of a video in the PCA attribute space. For the i-th video attribute v_i we first found its k-th nearest neighbour $v_{k(i)}$ (we chose k = 500) and computed the distance $||v_i - v_{k(i)}||$. This distance is the smallest radius of a hypersphere about v_i that encompasses k videos. The density $\rho(v_i)$ was then taken inversely proportional to the volume $||v_i - v_{k(i)}||^n$, with $n = 5$, the number of dimensions.

The task then is to assign suitable sampling probabilities $p_i, i = 1, \ldots, M$ for the set of videos in KoNViD-125k having attribute vectors $v_i, i = 1, \ldots, M$. Then, 10,000 subsamples will be drawn with corresponding probabilities p_i and without replacement. A natural choice is to set the probabilities p_i proportional to the inverse of the attribute densities, $1/\rho(v_i)$, at the corresponding attribute vectors, v_i, for all $i = 1, \ldots, M$.

From the 10,000 fairly sampled videos, we randomly subsampled 1,200 to form the KoNViD-1k dataset. A Gaussian kernel density estimation along each dimension shows the distribution of the samples (see Fig. 1, right sub-figure). We can see that attribute values are more evenly spread in most dimensions, except for the temporal information. This might be caused by local correlations in the data that have not been removed by PCA.

TABLE II: Database information and comparison.

	KoNViD-1k	KoNViD-125k
Unique authors	480	8418
Average videos per user	7.3	7.5
Max videos per user	11	1500
Videos with user-tags	620 (52%)	62646 (50%)
Total unique tags	2669	47336

E. Database analysis

We have shown that our fair sampling strategy has led to a better diversity with respect to six attributes. The diversity of the KoNViD-1k collection extends to other content-related characteristics such as Flickr meta-data tags and authorship. Table II summarizes related statistics.

The KoNViD-1k videos are encoded at three predominant frame rates: 24, 25 and 30 FPS corresponding to 27%, 5% and 68% of the items respectively. There are a total of 12 resolutions, with the largest percentage of videos having a frame size of 1280×720 pixels (85% of the videos), followed by 1920×1080 (9%). Most of the videos (97%) have an audio channel. The proportions of all these characteristics are similar between KoNViD-1k and KoNViD-125k.

III. Subjective quality assessment

In the KoNViD-1k database we make available a variety of meta-data together with subjective quality scores. Due to the large number of videos we crowdsourced the subjective scores using the widely used CrowdFlower platform (https://www.crowdflower.com/).

A. Crowdsourcing VQA

Quality control is a key component when designing an experiment for the crowd, so we considered how to set up our procedure carefully. Initially, each worker was instructed according to VQEG recommendations [12], which were modified to fit our single stimulus presentation technique. In the instructions workers were informed about types of degradation (e.g., related to motion, color, brightness, and details) and about how they would be asked to evaluate the overall quality of each video. Next, examples of videos with "Good", "Fair" and "Bad" quality were displayed for anchoring. Further, workers were instructed on the steps required to rate a video. Initially, a button was displayed below the video, which started playing the video muted. Once the video was finished playing a rating scale was displayed and workers had to select one of five categories to proceed. Only if a worker had watched and rated all 10 videos on a page, could he proceed to the next. By this design we hoped to ensure a better engagement of the workers and a better quality control.

In order to control for the quality of workers' performance, it is common to use gold standard questions, which is also a feature provided by CrowdFlower. Since there is no ground truth for our database, we have devised a plan to filter unreliable workers. We randomly sampled a subset of 100 videos from our pilot for an uncontrolled (no gold standard questions) crowdsourcing experiment of 50 ACR scores per video. From these, we computed the 95% confidence intervals of the MOS values to select those videos with a size of the confidence interval smaller than 0.5 on the ACR scale. The resulting 65 highly agreed upon videos and their MOS values were used as the ground truth for test questions for the evaluation of the entire data set.

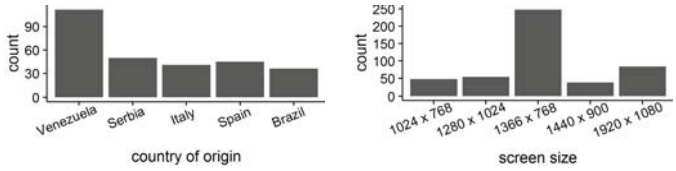

Fig. 3: Crowd-worker statistics on country of origin and screen resolution. First five groups are shown.

We employed the same setup for KoNViD-1k as above, with the addition of the test questions for quality control. We considered the rounded MOS ± 1 as eligible answers to these test questions. Workers that fell below 70% accuracy on test questions were removed from the experiment along with the data they had generated. Moreover, we only allowed CrowdFlower workers of Level 1 and above (more than 70% accuracy in all previous tasks) to participate in our experiment. Due to the fixed number of 65 test questions, workers could rate at most 550 videos in batches of 10 per page (10 questions for the quiz page, 55 batches with one test question each).

B. Crowdsourcing results and analysis

In our study we required a 95% confidence interval for the MOS values, averaged over all stimuli, with a length not exceeding 0.5 on the 5-point ACR scale. This was achieved by a minimum of 50 judgments per video. This setup resulted in a total of 642 workers from 64 countries participating in our experiment. On average each video received 114 votes (including test questions) with a mean accuracy of 94% on the test questions. See Fig. 3 for a histogram of common user statistics. With regards to screen size, 94% of workers had a resolution above 1024×600 pixels (width $\times$ height), which allowed full size display of the videos. It is to be noted that we did not enforce 1:1 pixel display. However, nearly two-thirds of workers did in fact view the videos unzoomed at 1:1, while 20% used a ratio of 0.9 to 0.75 and 14% displayed the videos at a zoom of up to 2 (including font scaling).

When comparing video quality ratings across different studies, commonly MOS, standard deviations and confidence intervals are considered [8]. However, it has been shown that when comparing different rating scales, the design and discretisation of rating scales have a strong influence on the standard deviations of the opinion scores (SOS). In [13] a quadratic model is proposed for the dependence of the variance σ^2 on the MOS values. For the 5-point ACR scale the function $\sigma^2(\text{MOS}) = a(\text{MOS} - 1)(5 - \text{MOS})$ is fitted to empirical data, which yields the SOS parameter a. The parameter a quantifies the variance of the user ratings more appropriately than the average over all stimuli. Moreover, it characterises application categories and correlates with task difficulty [14]. For VQA, the SOS parameter a was reported to fall in the range $[0.11, 0.21]$ [13].

We compared the standard deviations of the crowdsourced ratings of our KoNViD-1k to the CVD2014 (normalized from a 0-99 scale to a 5-point ACR scale) and IRCCyN-IVC-1080i databases [15], [16], where the subjective scores were

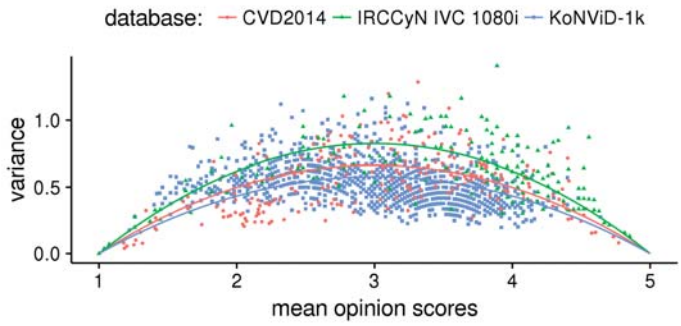

Fig. 4: Standard deviation of quality ratings as a function of MOS values for VQA in three databases with algebraic regression curves. The SOS hypothesis a values for our experiment on KoNViD-1k, CVD and IRCCyN are 0.14, 0.17 and 0.21 respectively. Smaller values imply better agreement. Each dot represents one stimulus (video).

gathered in a lab setting, see Fig. 4. With $a = 0.14$ our SOS hypothesis parameter is lower than those of the two compared databases (0.17 for CVD2014 and 0.21 for IRCCyN-IVC-1080i, respectively). This suggests that our task was simpler than the compared lab studies and resulted, as is desirable, in lower variances of worker ratings with smaller confidence intervals for the MOS values.

IV. Related work

In recent years, a broad range of VQA databases have been released [2]. The first comprehensive VQA database, dubbed EPFL-PoliMI, was published in [17]. The diversity of the items in this database was guaranteed by selecting the scenes that are representative of different levels of spatial and temporal complexity, which are the same as the SI and TI used in our KoNViD-1k database. LIVE [18], another well-known database, generated four types of distortions based on reference videos. It was further extended to LIVE Mobile [19], which also modelled distortions in heavily trafficked wireless networks, containing dynamical changing distortions including frame-freezes and temporally varying compression rates. Other databases similar to LIVE have been released such as IVP [20], CSIQ [21], MCL-V [22]. The aforementioned VQA databases, have the following drawbacks. First, all these VQA databases include videos that were artificially created from a small number of distortion-free reference videos. The distortion types range from compression artifacts e.g. MPEG-2, H.264 to transmission-based distortions such as those induced by packet-loss over IP networks or error-prone wireless systems, additive Gaussian noise, and others. The authenticity and representativeness of said distortions are far from that of distorted videos "in the wild". Second, the limited number of reference videos cannot guarantee the content diversity of these databases, and large-scale machine learning techniques cannot reliably learn from such a limited number of distorted videos. To address the drawback of artificial distortions in VQA databases, the CVD2014 database features different types of cameras to produce distortions related to the video acquisition process. A total of 78 cameras captured the same five selected physical scenes. However, the small number of cameras and scenes it contains limits the diversity of video content and quality distortions.

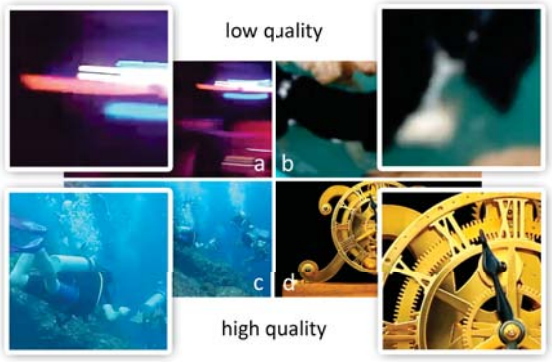

Fig. 5: Example extreme quality videos: a. MOS 1.26 (lowest quality score in KoNViD-1k), b. MOS 1.52, c. MOS 4.12, d. 4.64 (highest score). Low scores are usually caused by strong motion-blur due to shake, out-of-focus, or compression artifacts. High-scoring videos are sharp and show few distortions.

These deficiencies of current video databases would be difficult to assess quantitatively. However, an indirect confirmation is given by the fact that the performance of two established objective VQA algorithms on our KoNViD-1k database was significantly worse than on the traditional databases that were used for their development, even when the techniques were trained on our natural video dataset [23].

All of the above VQA databases provide subjective scores, namely MOS or difference mean opinion scores (DMOS), via lab-based studies which are time-consuming and expensive. Recently, it has been shown that reliable measures of quality of experience can be generated by crowdsourcing for images [24], [3] and videos [25]. It has been suggested that crowdsourcing workers can produce reliable VQA annotations by using paired comparisons and converting the results to DMOS [26].

V. Conclusions, Limitations, and Future Work

We created a fairly sampled, subjectively annotated video database, showing authentic distortions. For this purpose we first collected videos from the YFCC100m dataset following some practical minimum requirements (total of 144,889 videos). In order to guarantee the diversity in terms of content and multi-dimensional quality of the sampled dataset, six attributes (blurriness, colorfulness, contrast, temporal information, spatial information and VNIQE) were computed. Based on the values of the six attributes, we further filtered the extreme videos using empirical thresholds for each attribute. From the filtered collection of videos, we then devised a methodology based on k-NN density estimation to uniformly sample a subset of 10,000 videos. A random subset of 1,200 videos was further randomly sampled to produce a new VQA database, namely KoNViD-1k. Subjective scores of all videos in the KoNViD-1k database were obtained by a well-designed crowdsourcing experiment of ACR judgments. Compare with other two databases [15], [16], the standard deviations of the crowdsourced ratings for KoNViD-1k is the smallest, showing the highest agreement between workers.

Our KoNViD-1k database consists of diverse videos from YFCC100m published by different users using various cameras with different shooting skills; hence the number of high quality videos is quite small. This will limit the diversity of quality

in our database, thus further influence the quality assessment performance. When we extend our KoNViD-1k to 10,000 videos we will consider including more high quality ones.

Since the videos were filtered and sampled based on six attributes, the metric we chose for each attribute has influenced the diversity of the database. As a pilot, the methods for computing the attributes were chosen mostly for their feasibility and low computational complexity. We will consider alternative methods and empirically validate their effectiveness in representing the intended subjective attribute and to evaluate the linearity of the corresponding subjective scale, e.g. contrast or blur, by crowdsourcing experiments. We will also consider including additional attributes, such as overall brightness, camera shake, and color appropriateness.

Our choice of using quality ratings as test questions is a topic debated in the multimedia crowdsourcing community, as it may filter sincere users. Consensus seems to be that objective questions such as content related ones are a better way to perform reliability checks [27].

For the purpose of reducing the computational complexity of some of the metrics when computing the six attributes, we created cropped and scaled versions of the original videos to run them on, and chose a subset of frames to be processed in some cases. It remains to be checked that the resulting attribute values are entirely representative of the originals.

This is the largest quality assessment database of authentic Internet videos to date. Our strategy to carefully sample and subjectively annotate videos produces a highly representative set of videos (content types and quality distortions). We hope our efforts will lead the VQA community to greater levels of ecological validity for existing benchmarks and provide opportunities for developing new VQA methods. For instance, tapping into deep learning is possible when substantial training data is available. Our efforts for KoNViD-1k and the upcoming KoNViD-10k will pave the way to "deeper" VQA techniques.

Acknowledgments

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

References

[1] S.-F. Chang and A. Vetro, "Video adaptation: Concepts, technologies, and open issues," *Proc. of the IEEE*, vol. 93, no. 1, pp. 148–158, 2005.

[2] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.

[3] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

[4] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[5] 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.

[6] 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.

[7] E. Peli, "Contrast in complex images," *Journal of the Optical Society of America, A*, vol. 7, no. 10, pp. 2032–2040, 1990.

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

[9] 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.

[10] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The Konstanz Natural Video Database KoNViD-1k," http://database.mmsp-kn.de/, 2017.

[11] V. Vonikakis, R. Subramanian, and S. Winkler, "Shaping datasets: Optimal data selection for specific target distributions across dimensions," in *Proceedings of 2016 International Conference on Image Processing*. IEEE, 2016, pp. 3753–3757.

[12] ITU-T, "Objective perceptual assessment of video quality: Full reference television," Tutorial, ITU-T Telecommunication Standardization Bureau, 2004.

[13] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *Third International Workshop on Quality of Multimedia Experience*, 2011, pp. 131–136.

[14] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, 2015.

[15] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014 a database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073–3086, 2016.

[16] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: A resolution dependent paradigm," in *Proceedings of 2008 International Workshop on Image Media Quality and its Applications (IMQA 2008)*, 2008.

[17] F. D. Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 2430–2433.

[18] 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.

[19] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[20] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan, "IVP Subjective Quality Video Database," The Chinese University of Hong Kong, http://ivp.ee.cuhk.edu.hk/research/database/subjective/, 2011.

[21] P. V. Vu and D. M. Chandler, "ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *Journal of Electronic Imaging*, vol. 23, no. 1, pp. 013 016–013 016, 2014.

[22] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C. J. Kuo, "MCL-V: A streaming video quality assessment database," *Journal of Visual Communication and Image Representation*, vol. 30, pp. 1–9, 2015.

[23] H. Men, H. Lin, and D. Saupe, "Empirical evaluation of no-reference VQA methods on a natural video quality database," in *QoMEX 2017: International Conference on Quality of Multimedia Experience*, 2017.

[24] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *Proceedings of 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 3097–3100.

[25] C. C. Wu, K. T. Chen, Y. C. Chang, and C. L. Lei, "Crowdsourcing multimedia QoE evaluation: A trusted framework," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1121–1137, Aug 2013.

[26] 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: International Conference on Quality of Multimedia Experience*, 2016.

[27] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best practices and recommendations for crowdsourced QoE – Lessons learned from the Qualinet Task Force 'Crowdsourcing'," COST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services (QUALINET), 2014.