2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)

KADID-10k: A Large-scale Artificially Distorted IQA Database

Hanhe Lin, Vlad Hosu, and Dietmar Saupe
Department of Computer and Information Science, University of Konstanz, Germany
Email: {hanhe.lin, vlad.hosu, dietmar.saupe}@uni-konstanz.de

***Abstract*—Current artificially distorted image quality assessment (IQA) databases are small in size and limited in content. Larger IQA databases that are diverse in content could benefit the development of deep learning for IQA. We create two datasets, the Konstanz Artificially Distorted Image quality Database (KADID-10k) and the Konstanz Artificially Distorted Image quality Set (KADIS-700k). The former contains 81 pristine images, each degraded by 25 distortions in 5 levels. The latter has 140,000 pristine images, with 5 degraded versions each, where the distortions are chosen randomly. We conduct a subjective IQA crowdsourcing study on KADID-10k to yield 30 degradation category ratings (DCRs) per image. We believe that the annotated set KADID-10k, together with the unlabelled set KADIS-700k, can enable the full potential of deep learning based IQA methods by means of weakly-supervised learning.**

***Index Terms*—image quality assessment, image quality dataset, crowdsourcing**

I. Introduction

Objective image quality assessment (IQA), i.e., to automatically estimate the perceptual quality of a distorted image, has been a long-standing research topic. Objective IQA methods are divided into three categories based on the availability of pristine reference images: full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference IQA (NR-IQA). To develop and evaluate these methods, a number of benchmark databases have been proposed, some of which are compared in Table I.

Recently, deep Convolutional Neural Networks (CNN) have dramatically improved the state-of-the-art in image classification [1] and for many other computer vision tasks. However, developing a CNN-based IQA method is still challenging due to the lack of sufficient data for training. For example, the state-of-the-art CNNs like InceptionResNet [1], having hundreds of millions of parameters, require massive amounts of data to train from scratch, whereas the current largest artificially distorted IQA database, TID2013 [2], contains only 3,000 rated images.

To help solve this problem, we have created a large-scale artificially distorted IQA database, KADID-10k. It consists of 81 pristine images, where each pristine image was degraded by 25 distortions in 5 levels. For each distorted image, 30 reliable degradation category ratings were obtained by crowdsourcing, performed by 2,209 crowd workers. Compared to TID2013, KADID-10k is three times as large.

In addition, we have created KADIS-700k. It contains 140,000 pristine images, along with methods to degrade each image by a randomly selected distortion in 5 levels. The development and evaluation of IQA methods, especially weakly supervised [3] deep learning based methods, could significantly benefit from these datasets. Specifically, KADIS-700k would enable "inaccurate" weak supervision [3] of models that are refined on the subjective scores of KADID-10k. Both datasets are available in [4].

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projektnummer 251654672 TRR 161 (Project A05).

978-1-5386-8212-8/19/$31.00 ©2019 IEEE

II. Dataset creation

We collected the pristine images for both datasets from Pixabay.com, an international website for sharing photos and videos. These images were released under the Pixabay License, thus are free to be edited and redistributed. Moreover, each uploaded image had been shown to Pixabay users, who could cast their votes for accepting or declining it according to its perceptual quality. For each image, up to twenty independent votes were collected to make a decision. Therefore, the quality rating process provided by Pixabay provides a reasonable indication that the released images are pristine.

We downloaded 654,706 images with resolution greater than 1500×1200. All the images were rescaled to the same resolution as that in TID2013 (512×384), maintaining the pixel aspect ratios, followed by cropping if required. From these images we manually selected 81 high quality images as pristine images for KADID-10k. From the remaining images, we randomly sampled 140,000 images as pristine images for KADIS-700k.

For both of our datasets, we have applied existing implementations of image distortion methods and proposed several new types as well. In total there are 25 types of distortions, which can be grouped into: blurs (Gaussian, lens, motion), color related (diffusion, shifting, quantization, over-saturation and de-saturation), compression (JPEG2000, JPEG), noise related (white, white with color, impulse, multiplicative, white noise + denoise), brightness changes (brighten, darken, shifting the mean), spatial distortions (jitter, non-eccentricity patch, pixelate, quantization, color blocking), sharpness and contrast. For more information, please refer to [4]. We manually set the parameter values that control the distortion amount such that the visual quality of the distorted images varies perceptually linearly with the distortions parameter, from an expected rating

TABLE I: Comparison of existing benchmark IQA databases.

Database	Year	Content	No. of distorted images	Distortion type	No. of distortion types	No. of rated images	Ratings per image	Subjective study environment
IVC [5]	2005	10	185	artificial	4	185	15	lab
LIVE [6]	2006	29	779	artificial	5	779	23	lab
CSIQ [7]	2009	30	866	artificial	6	866	5∼7	lab
TID2013 [2]	2013	25	3,000	artificial	24	3,000	9	lab
CID2013 [8]	2013	8	474	authentic	12∼14	480	31	lab
LIVE In the Wild [9]	2016	1,169	1,169	authentic	N/A	1,169	175	crowdsourcing
Waterloo Exploration [10]	2016	4,744	94,880	artificial	4	0	0	N/A
KonIQ-10k [11]	2018	10,073	10,073	authentic	N/A	10,073	120	crowdsourcing
KADID-10k	2019	81	10,125	artificial	25	10,125	30	crowdsourcing
KADIS-700k	2019	140,000	700,000	artificial	25	0	0	N/A

of 1 (bad) to 5 (excellent). The distortion parameter values were chosen based on a small set of images, and applied the same for the remaining images in our database.

III. Subjective image quality assessment

We performed a subjective IQA study on figure-eight.com, a crowdsourcing platform. The experiment first presented workers with a set of instructions, including the procedure to rate an image and examples of different ratings. We used a standard degradation category ratings (DCR) method [12]. Specifically, given the pristine image on the left side, the crowd workers were asked to rate the distorted image on the right side in relation to the reference pristine image on a 5-point scale, i.e., imperceptible (5), perceptible but not annoying, slightly annoying, annoying, and very annoying (1). To control the quality of crowd workers, we annotated a number of "annoying" images and images with "imperceptible" degradation as test questions according to distortion settings. The test questions served two purposes. Firstly, they filter out unqualified crowd workers. Before starting the actual experiment, workers took a quiz with test questions only. Only those with an accuracy surpassing 70% were eligible to continue. Secondly, hidden test questions were presented throughout the experiments to motivate workers to continuously pay full attention. We collected 30 DCRs for each image.

IV. Results and analysis

Using the sampling, image processing, and crowdsourcing as described, we produced two datasets. KADID-10k contains 81 pristine images and $81 \cdot 25 \cdot 5 = 10125$ distorted images with 30 quality scores each; KADIS-700k contains 140,000 pristine images and 700,000 distorted images (500,000 for training, 100,000 for validation, and 100,000 for testing).

A. IQA evaluation

We have evaluated eleven FR-IQA methods and seven NR-IQA methods on KADID-10k, see Table II. For FR-IQA, we report the performance on entire database. For NR-IQA, the database was randomly split into training set (60%), validation set (20%) and test set (20%) according to reference images so there was no content overlapping between the three sets. A support vector regression (SVR) model (RBF kernel) was trained and evaluated for each NR-IQA method, with 10 repetitions. Furthermore, we fine-tuned InceptionResNetV2 [1] with the pre-trained weights on ImageNet. By changing the output layer to linear layer with one neuron, we trained 20 epochs with mean squared error loss. The model that gave the best performance on validation set was used for evaluation.

TABLE II: Performance comparison on KADID-10k.

	Method	PLCC	SROCC	KROCC
FR-IQA	SSIM [13]	0.723	0.724	0.537
	MSSSIM [14]	0.801	0.802	0.609
	IWSSIM [15]	0.846	0.850	0.666
	MDSI [16]	0.873	0.872	0.682
	VSI [17]	**0.878**	**0.879**	**0.691**
	FSIM [18]	0.851	0.854	0.665
	GMSD [19]	0.847	0.847	0.664
	SFF [20]	0.862	0.862	0.675
	SCQI [21]	0.853	0.854	0.662
	ADD-GSIM [22]	0.817	0.818	0.621
	SR-SIM [23]	0.834	0.839	0.652
NR-IQA	BIQI [24]	0.460	0.431	0.299
	BLIINDS-II [25]	0.559	0.527	0.375
	BRISQUE [26]	0.554	0.519	0.368
	CORNIA* [27]	0.580	0.541	0.384
	DIIVINE [28]	0.532	0.489	0.341
	HOSA* [29]	0.653	0.609	0.438
	SSEQ [30]	0.463	0.424	0.295
	InceptionResNetV2 (fine-tune)	**0.734**	**0.731**	**0.546**

* To reduce time complexity, CORNIA and HOSA features were restricted to 100 dimensions using PCA.

We report Pearson linear correlation coefficients (PLCC), Spearman rank order correlation coefficients (SROCC), and Kendall rank order correlation coefficients (KROCC) in Table II. Clearly, even the best FR-IQA method, VSI, gives a performance far from satisfactory, not to mention NR-IQA methods. Our fine-tuned InceptionResNetV2 outperformed the other NR-IQA methods. Using KADIS-700k, the performance can be further improved by weakly supervised learning [3].

B. Reliability analysis

We had included in our crowdsourcing experiments a few images from the TID2013 IQA database [2]. Thus, we can compare our subjective IQA results with those obtained in the original experiments on TID2013. In detail, 8 pristine images along with their distorted images (24 distortions, 5 levels per image) contained in TID2013 [2] were chosen to conduct the same subjective study as for KADID-10k. The SROCC between our crowdsourcing study and original study is 0.923. It is similar to what was reported in [2], where the SROCC between the lab study and an internet study was 0.934.

We also evaluated the quality of our experimental results by calculating common reliability measures, the Intra-class

Correlation Coefficient (ICC), inter-group correlations, and corresponding error measures.

Hosu et al. [31] reported reliability measures for IQA experiments using absolute category ratings (single stimulus). They showed that the ICC ranges from 0.4 in crowdsourcing experiments, to 0.58 for domain experts (online experiments), and 0.68 in the lab on the CID2013 database [8]. The ICC computed on KADID-10k is 0.66, at the higher end of the reported reliability range. However, our experiments are paired comparisons, which are considered generally easier to perform, and might results in a higher ICC.

With respect to inter-group correlations, computed by bootstrapping with resampling (100 times) from 30 ratings per pair collected in KADID-10k, similar to [31], we obtained a very high agreement of 0.982 SROCC, and low mean differences of 0.148 MAE and 0.193 RMSE between groups.

V. Conclusion

IQA is missing truly massive datasets for deep learning. We introduced such a dataset for semi-supervised learning, consisting of 700,000 (KADIS-700k) artificially degraded images without subjective scores, and 10,000 images of a similar kind that have been subjectively scored (KADID-10k).

Our collection of pristine images (KADIS-140k), is 30 times as large as the largest existing, the Waterloo dataset pristine images. We distorted images both by reproducing distortions from the literature (TID2013) and introducing new ones that relate to predominant defects in the wild. Our artificially distorted dataset (KADIS-700k), is 7 times as large as the entire Waterloo dataset. Moreover, our subjectively scored dataset (KADID-10k) is reliably annotated, and 3 times larger than the best existing TID2013.

We compared several NR-IQA and FR-IQA methods on KADID-10k. The best performance was 0.879 SROCC, achieved by VSI, showing there is room for improvement, even for FR-IQA methods.

References

[1] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 4, 2017, p. 12.

[2] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[3] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2017.

[4] H. Lin, V. Hosu, and D. Saupe, "The KADID-10K Image Database," 2019, http://database.mmsp-kn.de/kadid-10k-database.html.

[5] P. Le Callet and F. Autrusseau, "Subjective quality assessment IRCCyN/IVC database," 2005, http://www.irccyn.ec-nantes.fr/ivcdb/.

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

[7] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010.

[8] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, "CID2013: A database for evaluating no-reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 390–402, 2015.

[9] 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.

[10] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, and L. Zhang, "Group MAD competition – a new methodology to compare objective image quality models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1664–1673.

[11] H. Lin, V. Hosu, and D. Saupe, "KonIQ-10k: Towards an ecologically valid and large-scale IQA database," *arXiv:1803.08489*, 2018.

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

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[14] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. IEEE, 2003, pp. 1398–1402.

[15] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[16] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet, "Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator," *IEEE Access*, vol. 4, pp. 5579–5590, 2016.

[17] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[18] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[19] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.

[20] H.-W. Chang, H. Yang, Y. Gan, and M.-H. Wang, "Sparse feature fidelity for perceptual image quality assessment." *IEEE Transaction on Image Processing*, vol. 22, no. 10, pp. 4007–4018, 2013.

[21] S.-H. Bae and M. Kim, "A novel image quality assessment with globally and locally consilient visual quality perception," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2392–2406, 2016.

[22] K. Gu, S. Wang, G. Zhai, W. Lin, X. Yang, and W. Zhang, "Analysis of distortion distribution for pooling in image quality prediction," *IEEE Transactions on Broadcasting*, vol. 62, no. 2, pp. 446–456, 2016.

[23] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 1473–1476.

[24] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, 2010.

[25] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[26] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[27] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1098–1105.

[28] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[29] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[30] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856–863, 2014.

[31] V. Hosu, H. Lin, and D. Saupe, "Expertise screening in crowdsourcing image quality," in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–6.