https://doi.org/10.2352/issn.2694-118X.2021.LIM-1

Deep Learning in Image Quality Assessment: Past, Present, and What Lies Ahead

Seyed Ali Amirshahi
The Norwegian Colour and Visual Computing Laboratory, Norwegian University of Science and Technology, Gjøvik, Norway
s.ali.amirshahi@ntnu.no

Abstract

Quality assessment of images plays an important role in different applications in image processing and computer vision. While subjective quality assessment of images is the most accurate approach due to issues objective quality metrics have been the go to approach. Until recently most such metrics have taken advantage of different handcrafted features. Similar (but with a slower speed) to other applications in image processing and computer vision, different machine learning techniques, more specifically Convolutional Neural Networks (CNNs) have been introduced in different tasks related to image quality assessment. In this short paper which is a supplement to a focal talk given with the same title at the London Imaging Meeting (LIM) 2021 we aim to provide a short timeline on how CNNs have been used in the field of image quality assessment so far, how the field could take advantage of CNNs to evaluate the image quality, and what we expect will happen in the near future.

Introduction

For decades Image Quality Assessment (IQA) has been an active field of research [1]. Naturally, the go to approach for assessing the quality of images would be to perform different subjective experiments. While subjective experiments has been the gold standard in the field, such experiments are time consuming and financially expensive. This has resulted in the introduction of different objective Image Quality Metrics (IQMs) which aim to model the subjective judgment of the image quality and are now the go to approach when there is a need for IQA both in the research and industrial community. A common approach for categorising IQMs is how much access we have to the reference image. That is, Full Reference (FR) metrics which have access to the reference image, Reduced Reference (RR) metrics which have access to partial information of the reference image and No Reference (NR) metrics which do not have access or any information of the reference image. Over the years a high number of different IQMs have been proposed resulting in different studies on evaluating the performance of the said metrics [2, 3, 4, 5, 6, 7].

While in recent years the use of Convolutional Neural Networks (CNNs) and other state-of-the-art machine learning techniques have taken over most computer vision and image processing tasks the same could not be claimed in the case of IQA. In fact, until recently most IQMs were based on the use of a few handcrafted features [8, 9]. While such an approach had been closely linked to the lack of a large-scale subjective dataset [10], recently, through online platforms and crowdsourcing [11] few large-scale datasets such as [12, 13, 14] have been introduced. These datasets along with other approaches which we will discuss in the rest of the paper has resulted in the introduction of different CNN based IQMs.

In this paper which is a supplement to a focal take given with the same title we aim to have a short review on how over the years there has been an increase in the number of different CNN based IQMs. Our hope is to provide a story line and link the first studies in the field to its current state and try to have an educated guess on what is waiting for us in the near future.

Initial CNN based IQMs

One of the first if not the first work which used CNNs to evaluate the quality of an images dates back to 2014 [15]. In this NR IQM, Kang et al. use a combination of feature learning and regression and calculate the average score of CNN quality estimates of the patches in the image. Due to the lack of a large-scale dataset with sufficient size for training an entire CNN from scratch, initially most CNN based IQMs were based on using pre-trained CNNs. In this type of approach the features extracted by the CNN were used in evaluating the quality of an image [7, 16, 17, 18]. As an example, DeepBIQ [16] which is also a NR IQM uses features extracted from the Caffe [19] network architecture which is trained on the ImageNet dataset [20]. DeepBIQ then calculates the quality of a given image by averaging the quality scores calculated for multiple regions of the image.

When it comes to the first few FR IQMs which are based on the use of CNN, pre-trained networks play a crucial role. For example, Amirshahi et al. [7, 17] use AlexNet [21] which is pre-trained on the ImageNet dataset to extract deep features from the reference and test images. The comparison of feature maps are then used to evaluate the quality of an image. This ranges from a simple pyramidal approach to compare the strength of feature maps in the test and reference image, similar to what was initially proposed in [22] for calculating the Pyramid Histogram of Orientated Gradients (PHOG) to using traditional IQMs to compare the similarity between corresponding feature maps. While simple, the proposed approaches show a dramatic increase (up to 23%) in the accuracy of IQMs such as Structural Similarity Index (SSIM) [23], Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR), Mean Average Error (MAE), Laplacian Mean Square Error (LMSE), Normalized Absolute Error (NAE), Maximum difference (MD), and Structural Content (SC).

Current CNN based IQMs

As mentioned earlier the size of the labeled data (in the case of IQMs, size of the subjective dataset) we have access to plays an important role in the performance of our CNN [24]. While in recent years a few relatively large-scale subjective datasets have been introduced, unfortunately, compared to other fields of study like image classification and segmentation, the size of the datasets are still too small. While in similar cases generating data using different augmentation techniques is a go to approach, keeping in mind the subjective nature of the image quality scores, in such dataset generating augmented data in order to the increase the size of our dataset is not the first go to option. Nevertheless, to address the lack of a dataset with large enough number of images studies such as [25] artificially augment the datasets. In their study Boss et al. train their network on a set of randomly selected patches from subjectively evaluated images [25]. [26] uses a similar patch-based approach. In their work a CNN model is used to evaluate the quality of an image on a local scale (patches) and then regression is used to evaluate the overall quality of the image.

One of the common methods for categorising different IQMs is to divide the them to single-task [27, 28, 29] and multi-task metrics [30, 31, 32, 33, 34, 35]. As an example, in the case of [28] which is a single-task IQM a fully connected CNN is used while [27] takes advantage of a Generative Adversarial Network (GAN) [36]. When it comes to multi-task metrics, the mentioned IQMs are mostly based on detecting the type of distortion affecting the image quality and then evaluating the image quality based on that. This is done either by using a single network for both tasks or in the case of the Multi-task Rank-Learning Image Quality assessment (MRLIQ) [32] a number of different IQMs for different types of distortions is used.

Different studies have emphasized on the role of attention for evaluating the quality of images and videos [37, 8, 38, 39]. In that order different saliency detection methods have been used to evaluate the quality of images. This ranges from assigning a weight to different regions in the image based on a saliency map generated using saliency detection technique to only calculating the quality of the most salient region in the image [37].

Finally, a common approach in CNN based FR IQMs is the use of Siamese networks [40]. In such an approach the test and reference images are processed in parallel using two different networks with the matching specifications [25]. Ayyoubzadeh and Royat [41] used an attention-based Siamese-Difference neural network to detect the difference between the reference and test images. For the attention mechanism in their approach they used the work by Wang and Shen [42].

What lies ahead

Having access to a large-scale labelled dataset is an important issue when it comes designing and new CNN based IQM. Unfortunately, when it comes to the field of IQA there are only a limited number of subjective datasets available [10]. This is mainly because of the fact that still most subjective datasets are collected under controlled environment in a lab setting which naturally will result in a lower number of rated images and participating observers. In fact, different standards and guidelines have been agreed on in the research community for collecting such data [43, 44]. When it comes to using crowdsourcing for collecting subjective data in an uncontrolled environment such guidelines and standards are not yet available. A new guideline should include subject reliability, difference in viewing condition, display device, visual acuity of the observers, how their cultural background could affect the subjective scores, etc.. Apart from creating large-scale dataset, recent studies have also focused on the possibility of merging different already available datasets [45, 46] which still needs further studies.

When it comes to the IQMs themselves, although current IQMs have shown great performance in evaluating the quality of images, there exist room for improvement. Below some of the future challenges in the field are introduced:

- Current IQMs are mostly focused on images affected by a single distortion and their performance drop when multiple number of distortions are present in the image. In the rare case that an IQM is designed for a multi-distorted images this is done by a predefined set of distortions which the metric is already trained on. This issue is also what has been pointed out in different studies [47, 48] as the future direction they would like to take.
- An advantage of using CNN based IQM is the vast amount of information it provides the users at different convolutional layers and through the feature maps. This is a perfect opportunity to gain a better understanding on how and why such IQMs work and investigate the link between them and the human visual system. As an example, [29] has proposed to focus on the introduction of better sensitivity maps with respect to the human visual system.
- Current IQMs mainly provide a single score to the image that is been evaluated. While such scores could be used as a way to find the distance between the quality of different images, by itself they do not provide the user with interpretable information about the quality of each image. Using CNNs we should be able to not only have a quality score representing the image quality but also provide a descriptive evaluation of the image quality so the user is able to better understand how, where, and to what extent the quality of the image is affected. Such interpretation of the image quality could also be useful in proposing better image enhancement and image processing techniques in general [49].
- With the increase in the size of subjective datasets and advances in machine learning techniques, in the near future we should expect the introduction of personalized IQMs. That is, the IQMs will not only provide the average quality score for all observers but will also be able to predict the quality score given by each individual observer.
- The content of the image plays an important role in the image quality [50]. Over the years, not enough attention has been paid on introducing IQMs which take into account the content of the image [51]. One possible reason for this could be the lack of a dataset which covers a wide range of different content.
- Different studies have pointed out to how a combination of handcrafted features and state-of-the-art machine learning techniques could result in highly accurate IQMs [52].

Conclusion

This short paper which is prepared as a complement to the focal talk given with the same time at the London Imaging Meeting (LIM) 2021 we provided a short storyline on the use of Convolutional Neural Networks for evaluating the quality of images. We provided information about what the field lacks and what lies ahead.

References

[1] Farah Torkamani-Azar and Seyed Ali Amirshahi. A new approach for image quality assessment using svd. In *2007 9th International Symposium on Signal Processing and Its Applications*, pages 1–4. IEEE, 2007.

[2] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006.

[3] Nikolay Ponomarenko, Federica Battisti, Karen Egiazarian, Jaakko Astola, and Vladimir Lukin. Metrics performance comparison for color image database. In *Fourth international workshop on video processing and quality metrics for consumer electronics*, volume 27, page 6, 2009.

[4] Atidel Lahoulou, Ahmed Bouridane, Emmanuel Viennet, and Mourad Haddadi. Full-reference image quality metrics performance evaluation over image quality databases. *Arabian Journal for Science and Engineering*, 38(9):2327–2356, 2013.

[5] Marius Pedersen. Evaluation of 60 full-reference image quality metrics on the CID:IQ. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1588–1592. IEEE, 2015.

[6] Marius Pedersen and Jon Yngve Hardeberg. Full-reference image quality metrics: Classification and evaluation. *Foundations and Trends® in Computer Graphics and Vision*, 7(1):1–80, 2012.

[7] Seyed Ali Amirshahi, Marius Pedersen, and Azeddine Beghdadi. Reviving traditional image quality metrics using CNNs. In *Color and Imaging Conference*, volume 2018, pages 241–246. Society for Imaging Science and Technology, 2018.

[8] Seyed Ali Amirshahi and M-C Larabi. Spatial-temporal video quality metric based on an estimation of qoe. In *2011 Third International Workshop on Quality of Multimedia Experience*, pages 84–89. IEEE, 2011.

[9] Seyed Ali Amirshahi. Towards a perceptual metric for video quality assessment. Master's thesis, Norwegian University of Sciecne and Technology, 2010.

[10] Seyed Ali Amirshahi and Marius Pedersen. Future directions in image quality. In *Color and Imaging Conference*, volume 2019, pages 399–403. Society for Imaging Science and Technology, 2019.

[11] Flávio Ribeiro, Dinei Florencio, and Vítor Nascimento. Crowdsourcing subjective image quality evaluation. In *2011 18th IEEE International Conference on Image Processing*, pages 3097–3100. IEEE, 2011.

[12] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

[13] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KonIQ-10K: Towards an ecologically valid and large-scale IQA database, 2018.

[14] Mykola Ponomarenko, Sheyda Ghanbaralizadeh Bahnemiri, Karen Egiazarian, Oleg Ieremeiev, Vladimir Lukin, Veli-Tapani Peltoketo, and Jussi Hakala. Color image database htid for verification of no-reference metrics: peculiarities and preliminary results. In *9th European Workshop on Visual Information Processing (EUVIP 2021)*, 2021.

[15] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014.

[16] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *arXiv preprint arXiv:1602.05531*, 2016.

[17] Seyed Ali Amirshahi, Marius Pedersen, and Stella X Yu. Image quality assessment by comparing cnn features between images. *Journal of Imaging Science and Technology*, 60(6):60410–1, 2016.

[18] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104–114, 2017.

[19] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[22] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408, 2007.

[23] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[24] S Alireza Golestaneh and Kris Kitani. No-reference image quality assessment via feature fusion and multi-task learning. *arXiv preprint arXiv:2006.03783*, 2020.

[25] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

[26] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1):206–220, 2016.

[27] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 732–741, 2018.

[28] Da Pan, Ping Shi, Ming Hou, Zefeng Ying, Sizhe Fu, and Yuan Zhang. Blind predicting similar quality map for image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6373–6382, 2018.

[29] Jongyoo Kim, Anh-Duc Nguyen, and Sanghoon Lee. Deep cnn-based blind image quality predictor. *IEEE transactions on neural networks and learning systems*, 30(1):11–24, 2018.

[30] Yuge Huang, Xiang Tian, Yaowu Chen, and Rongxin Jiang. Multitask convolutional neural network for no-reference image quality assessment. *Journal of Electronic Imaging*, 27(6):063033, 2018.

[31] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017.

[32] Long Xu, Jia Li, Weisi Lin, Yongbing Zhang, Lin Ma, Yuming Fang, and Yihua Yan. Multi-task rank learning for image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(9):1833–1843, 2016.

[33] Qingbo Wu, Hongliang Li, King N Ngan, Bing Zeng, and Moncef Gabbouj. No reference image quality metric via

distortion identification and multi-channel label transfer. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 530–533. IEEE, 2014.

[34] Hanli Wang, Lingxuan Zuo, and Jie Fu. Distortion recognition for image quality assessment with convolutional neural network. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016.

[35] Sewoong Ahn, Yeji Choi, and Kwangjin Yoon. Deep learning-based distortion sensitivity prediction for full-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 344–353, 2021.

[36] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[37] Meisam Jamshidi Seikavandi and Seyed Ali Amirshahi. Quality is in the salient region of the image. *Electronic Imaging*, 2021.

[38] Seyed Ali Amirshahi, Gregor Uwe Hayn-Leichsenring, Joachim Denzler, and Christoph Redies. Evaluating the rule of thirds in photographs and paintings. *Art & Perception*, 2(1-2):163–182, 2014.

[39] Meisam Jamshidi Seikavandi and Seyed Ali Amirshahi. Evaluating video quality by differentiating between spatial and temporal distortions. In *CVCS*, 2020.

[40] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[41] Seyed Mehdi Ayyoubzadeh and Ali Royat. (asna) an attention-based siamese-difference neural network with surrogate ranking loss function for perceptual image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 388–397, 2021.

[42] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017.

[43] ITURBT Recommendation. 500-11,"methodology for the subjective assessment of the quality of television pictures," recommendation itu-r bt. 500-11. *ITU Telecom. Standardization Sector of ITU*, 7, 2002.

[44] Brian W Keelan and Hitoshi Urabe. Iso 20462: a psychophysical image quality measurement standard. In *Image Quality and System Performance*, volume 5294, pages 181–189. International Society for Optics and Photonics, 2003.

[45] Tomas Mizdos, Marcus Barkowsky, Miroslav Uhrina, and Peter Pocta. How to reuse existing annotated image quality datasets to enlarge available training data with new distortion types. *Multimedia Tools and Applications*, pages 1–23, 2021.

[46] Lukáš Krasula, Yoann Baveye, and Patrick Le Callet. Training objective image and video quality estimators using multiple databases. *IEEE Transactions on Multimedia*, 22(4):961–969, 2019.

[47] Simeng Sun, Tao Yu, Jiahua Xu, Jianxin Lin, Wei Zhou, and Zhibo Chen. Graphiqa: Learning distortion graph representations for blind image quality assessment. *arXiv preprint arXiv:2103.07666*, 2021.

[48] Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma. Continual learning for blind image quality assessment. *arXiv preprint arXiv:2102.09717*, 2021.

[49] Sören Becker, Thomas Wiegand, and Sebastian Bosse. Curiously effective features for image quality prediction. *arXiv preprint arXiv:2106.05946*, 2021.

[50] Michael P Eckert and Andrew P Bradley. Perceptual quality metrics applied to still image compression. *Signal processing*, 70(3):177–200, 1998.

[51] Jenni Radun, Tuomas Leisti, Jukka Häkkinen, Harri Ojanen, Jean-Luc Olives, Tero Vuori, and Göte Nyman. Content and quality: Interpretation-based estimation of image quality. *ACM Transactions on Applied Perception (TAP)*, 4(4):1–15, 2008.

[52] Aladine Chetouani, Maurice Quach, Giuseppe Valenzise, and Frédéric Dufaux. Combination of deep learning-based and handcrafted features for blind image quality assessment. In *9th European Workshop on Visual Information Processing (EUVIP 2021)*, 2021.

Author Biography

Seyed Ali Amirshahi is an Associate Professor at the Norwegian University of Science and Technology (NTNU). His work is focused on image quality assessment and computational aesthetics. He received his PhD from the Friedrich Schiller University of Jena in Germany (2015). Prior to his current position he was a Marie Curie post-doctoral Fellow at NTNU and a visiting researcher at University Sorbonne Paris Nord. Prior to that he was a post-doctoral Fellow at the International Computer Science Institute in Berkeley, CA.