NO-REFERENCE IMAGE SHARPNESS ASSESSMENT BASED ON RANK LEARNING

Yabin Zhang, Haiqiang Wang, Fengfeng Tan, Wenjun Chen, and Zurong Wu

Tencent Media Lab, Shenzhen, China

ABSTRACT

To address the label shortage problem and the fixed-size input constraint of CNN models, we propose a no-reference image sharpness assessment method based on rank learning and effective patch extraction. First, we train a Siamese mobilenet network by learning quality ranks among the synthetically blurred and unsharpen seed images without any human label, which provides effective prior knowledge about the appropriate image sharpness. The extracted single branch finetuned on benchmark datasets is used to predict the subjective rating regarding image sharpness. While the performance of CNN based IQA metrics is compromised due to their fixed-size input constraint, we design a multi-scale gradient guided patch extraction method to increase the quality-sensitive input information and boost the performance. Extensive experimental results on the six public datasets demonstrate that our approach outperforms other state-of-the-art no-reference image sharpness assessment metrics.

Index Terms— Image sharpness, Rank learning, Image quality assessment

1. INTRODUCTION

Perceptual quality assessment is a fundamental problem for multimedia processing and transmission systems. While human visual system is the ultimate receiver of the visual signals, subjective rating process is too laborious, time-consuming, expensive and non-automatic. Over the last two decades, there are plenty of studies on objective quality assessment using computational methods, and the image sharpness is one of the most important quality factors in applications such as streaming and image enhancement.

In the literature, Vu *et al.* [1] proposed a spectral and spatial sharpness (S3) model by measuring the slope of the magnitude spectrum and total spatial variation. Bahrami *et al.* [2] observed that maximum local variation (MLV) is indicative of sharpness and utilize the standard deviation of MLV distribution to measure the sharpness quality. In [3], Li *et al.* proposed an image sharpness model based on sparse representation called SPARISH, where an overcomplete dictionary is trained to capture the high-frequency information changes caused by image blur. In [4], a robust image sharpness evaluation (RISE) method is proposed by learning the multi-scale features in both spatial and spectral domain. Recent works [5, 6] utilized general regression neural network (GRNN) for the estimation of blur kernel and sharpness, respectively. Hosseini *et al.* [7] developed a maximally polynomial (MaxPol) convolution kernel based visual sensitivity model to measure the out-of-focus level of images.

In this paper, we focus on the no-reference quality assessment for image sharpness using deep neural networks. Although deep CNN methods have shown promising results in applications such as object classification and detection, there are still challenges in image quality assessment due to shortage of labeled data. What is more, traditional CNN methods is usually designed to capture the high-level invariant features, where the quality-sensitive information is overlooked or wiped out during large scale resizing required by the fixed-size input constraint of CNN. The main contribution of the proposed method can be summarized as follows. We train a self-supervised Siamese mobilenet model [10, 9] by learning from quality ranks, which learns effective sharpness prior knowledge. We design a multi-scale gradient guided patch extraction method to increase the quality-sensitive input information and boost the performance. With further finetune, our approach achieves state-of-the-art performance on the benchmark datasets.

2. RANK LEARNING BASED METHOD

To relieve the shortage of image labels, our sharpness assessment strategy is the combination of self-supervised rank learning on seed images and finetune on small-scale benchmark datasets as [8, 9]. The ranking of synthetically generated images from the same seed image is easy to obtain and reliable. Given one professional image with carefully tuned sharpness level, either blur or unsharpen operations will degrade its fidelity or the perceptual quality, and the stronger strength will lead to worse perceptual quality.

As shown in Fig. 1, one pair of input images $\{x_a; x_b\}$ are fed to the two branches of Siamese network separately. Firstly, images are cropped into N patches, $\{x_{a,1}, x_{a,2}, ..., x_{a,N}; x_{b,1}, x_{b,2}, ..., x_{b,N}\}$, by the multi-scale patch extraction method. The extracted patches are passed to the backbone CNN and the prediction scores $\{\hat{y}_{a,1}, \hat{y}_{a,2}, ..., \hat{y}_{a,N}; \hat{y}_{b,1}, \hat{y}_{b,2}, ..., \hat{y}_{b,N}\}$ are then averaged as $\{\hat{y}_a; \hat{y}_b\}$. We employ the efficient mobilenet [10] with depth-wise sepa-

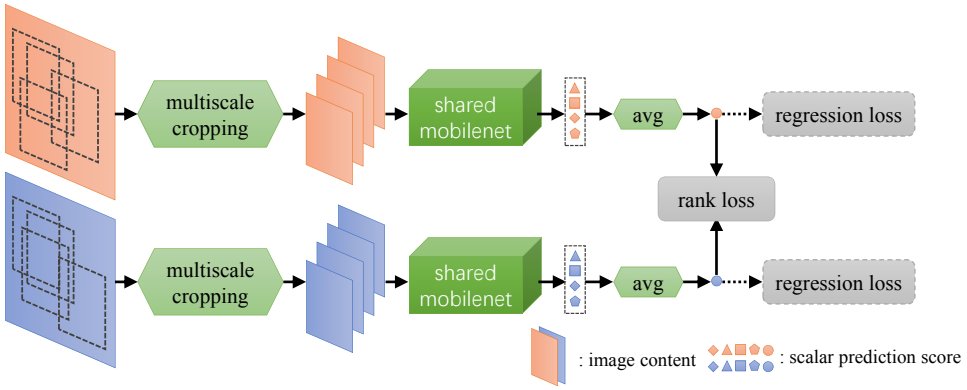


Fig. 1. Proposed rank learning based method with Siamese mobilenet and multi-scale patch extraction.

rable convolutions as the backbone CNN, and the weights of the network is shared in both branches. In the pretraining stage, without additional human labels, only rank loss is used to train the network based on the dynamically generated rank labels. When we finetune the network on small labeled datasets, the ground-truth label $\{y_a; y_b\}$ is available and the regression losses such as $L1$ loss are added to improve the correlation between objective prediction scores and the subjective ratings. After the finetune, the red and blue branches are exactly the same and one single branch will be extracted for the no-reference image sharpness prediction.

2.1. Self-supervised rank learning

To learn the accurate sharpness prior knowledge, we turn to the largest aesthetic visual analysis (AVA) [11] database to obtain the high quality professional images. The source of AVA images is the photograph competition website called DPchallenge, where competitions are held periodically with different topics called challenges. We have selected 207 high-frequency classical challenges from the total 1398 challenges such as *Architecture*, *Best of year*, *Centered Composition*, *Still Life*, *Macro*, *Shapes* and *Texture*, and also removed special challenges that may cause distractions such as *Over-Saturated & Over-Sharpened* and *Motion Blur*. To refine the images from the chosen challenges, we set the mean rating score threshold as 6.0 and the standard deviation threshold is chosen as 0.5, and finally we obtain 4400 high-quality seed images from AVA database.

We assume that the operation $h(x, \phi)$ will degrade the high-quality image monotonically when the control parameter ϕ increases. Assuming there exists the ideal quality prediction model $F(x)$, it indicates that

$$\phi_i < \phi_j \Rightarrow F(h(x, \phi_i)) \geq F(h(x, \phi_j)). \quad (1)$$

Here the chosen $h(x, \phi)$ are the blur operation $h_b(x) = x * k_g(N_b, \sigma_b)$ and the unsharpen operation $h_s = (1 + \alpha) \cdot x - \alpha \cdot x * k_g(N_s, \sigma_s)$, where k_g is the Gaussian kernel with window size N_b / N_s and standard deviation σ_b / σ_s. The control parameter ϕ_b for blur operation is the kernel size N_b while the parameter ϕ_s for unsharpen parameter is α. As we can see, the chosen blur and unsharpen operations are the opposite processes using similar Gaussian filters.

Given one high-quality seed image x and the prediction network denoted as $\hat{F}$, the hinge rank loss for the dynamically generated image pair $x_a = h(x, \phi_i)$ and $x_b = h(x, \phi_j)$ is

$$Loss_{rank} = \sum max(0, \hat{F}(x_a) - \hat{F}(x_b) + \epsilon), \quad (2)$$

where ϵ is the margin parameter and $h(x)$ is randomly selected from the blur and unsharpen operations. In the finetune stage, the rank loss and $L1$ regression loss are combined as

$$Loss = \sum (|y_a - \hat{F}(x_a)| + |y_b - \hat{F}(x_b)|) + \lambda Loss_{rank}, \quad (3)$$

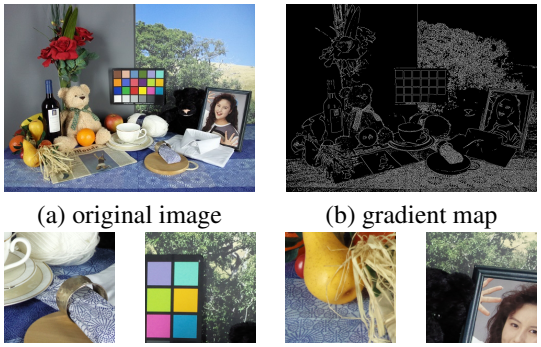
where $\{x_a; x_b\}$ are the input image pair from the training set with ground-truth mean opinion scores (MOS) $\{y_a; y_b\}$.

2.2. Multi-scale image patch extraction

As suggested in [4], gradient information has been shown to be an important structure feature for the image sharpness assessment. We choose the set of multiple patches as the representation of the input image. Instead of simple random crop, the gradient information is used to guide the patch selection process. As shown in Fig. 2 (b), first, the binary gradient map is obtained with Canny edge detection [12]. The patch candidates are generated by sliding the 224×224 window with step size of 32 pixels. Next, we sort the patches based on the amount of corresponding edge pixels, and randomly select N patches from the top 25% patches as the final chosen patch set. As shown in Fig. 2, the extracted patches cannot cover sufficient content patterns for high-resolution images. Therefore, the N patches are extracted from the input image and two additional downsampled images by ratio $\sqrt{2}$ as well.

3. EXPERIMENTAL RESULTS

We conduct the experiments on six widely-used sharpness related datasets including LIVE [13], CSIQ [14], TID2008 [15],



(a) original image (b) gradient map

(c) Extracted patches

Fig. 2. Gradient map guided multiple patch extraction

TID2013 [16], CID2013 [17] and BID [18], where the first four are synthetical datasets and the latter two are real blurred datasets. They contain 145, 150, 100, 125, 473, and 586 blurred images with MOS or difference of MOS (DMOS), respectively. The performance of sharpness metrics is evaluated by Spearman rank order correlation coefficient (SRCC) for the monotonicity of prediction and Pearson linear correlation coefficient (PLCC) for accuracy.

The common evaluation strategy is to randomly choose 80% of each dataset as train set and the rest part as test set, and the median performance of 1000 times train-test operations is reported. However, the time cost for CNN based methods is usually intolerable. In this paper, we choose the alternative k-fold cross validation method to evaluate the performance. Each dataset is randomly split into k groups with approximate same number of images in the non-overlapping way. Each group will work as the test set in turn while the rest $k-1$ groups are chosen as the train set. To avoid the bias and reduce the repeat times, we choose 5-fold cross validation and the average performance of the five-round train-test operations is reported.

3.1. Implementation details

The Siamese mobilenet network is implemented using Pytorch with the Adam optimizer. In self-supervised pretraining stage, the base learning rate is $1e$-3, the weight decay is $1e$-4, the rank loss margin is 0.05 and the batch size is 16. In the finetune stage on each benchmark dataset, the initial learning rate is changed to $1e$-4. The cropping image patch size is 224×224. The patch number N is set to 8 in the training stage. In the evaluation stage, to reduce the discrepancy due to the random crop, N is set to 32.

The strength of blur and unsharpen operations is controlled by the kernel size $N_b \in [3, 5, 7, 11, 15]$ and the $\alpha \in [0.25, 0.5, ..., 4]$, respectively, and the default σ_b and σ_s are fixed to 1. In the loss function, the margin parameter ϵ is set to 0.05 and λ is set to 4 empirically.

3.2. Performance evaluation

We compare the proposed method with eight existing sharpness metrics including S3 [1], MLV [2], Kang's CNN [19], ARISM [20], SPARISH [3], RISE [4], Yu's CNN [6] and MaxPol [7] as summarized in Table 1, where the top two in each row are shown in bold. Since the performance evaluation of some metrics are not available, the overall performance is the averaged performance only on the four synthetic datasets weighted by the amount of their distorted images.

On the traditional synthetically generated blur image datasets, the proposed methods show better performance compared with other methods on CSIQ, TID2008 and TID2013 datasets, and on the LIVE dataset our approach also shows the comparable performance. According to the weighted overall score, our approach yields better overall performance. On the two real blur image datasets, our method has outperformed other existing state-of-the-art methods such as RISE with a relatively large margin.

Our self-supervised rank model learnt from AVA seed images are shown in the penultimate column. Even without finetune on each test dataset, the learnt model still shows very promising generalized performance on the four synthetical blur image datasets and BID dataset.

Datasets	Criterion	B	B+P	B+P+R	B+P+R+M
CID2013	PLCC	0.6326	0.8567	0.8574	0.8629
	SRCC	0.5666	0.8325	0.8238	0.8363
BID	PLCC	0.3157	0.7415	0.7347	0.7509
	SRCC	0.3239	0.6972	0.6825	0.7187
LIVE	PLCC	0.8619	0.9686	0.9727	0.9693
	SRCC	0.8170	0.9557	0.9592	0.9544
CSIQ	PLCC	0.6529	0.9696	0.9708	0.9791
	SRCC	0.5873	0.9400	0.9428	0.9529
TID2008	PLCC	0.8076	0.9553	0.9552	0.9586
	SRCC	0.7470	0.9310	0.9335	0.9489
TID2013	PLCC	0.7882	0.9671	0.9680	0.9646
	SRCC	0.7211	0.9535	0.9551	0.9550
weighted overall	PLCC	0.7764	0.9650	0.9666	0.9680
	SRCC	0.7171	0.9446	0.9472	0.9527

Table 2. Ablation study. (B: baseline with mobilenet; P: pretrained model; R: additional rank loss; M: multi-scale patch extraction) The last row is the weighted overall performance for the four synthetically blur image datasets

3.3. Ablation study

To demonstrate the effectiveness of each component, the ablation study on the public datasets is provided in Table 2 with the same evaluation strategy as those in Table 1 by adopting the 5-fold cross validation. We investigate the impact of the pretrained model, additional rank loss and multi-scale patch extraction. The baseline is the Siamese mobilnet model with $L1$ regression loss and gradient guided patch extraction. "B+P" uses the pretrained rank model for finetune, "B+P+R"

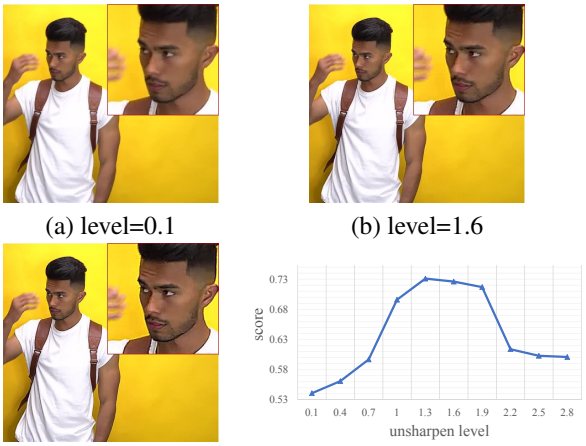
Method	Criterion	S3 [1]	MLV [2]	Kang's CNN [19]	ARISM [20]	SPARISH [3]	RISE [4]	Yu's CNN [6]	MaxPol [7]	Ours w.o. FT	Ours
CID2013	PLCC	0.6863	0.6890	N.A.	0.5523	0.6775	**0.7934**	N.A.	N.A.	0.1843	**0.8629**
	SRCC	0.6460	0.6206	N.A.	0.4719	0.6607	**0.7690**	N.A.	N.A.	0.0854	**0.8363**
BID	PLCC	0.4271	0.3643	N.A.	0.1841	0.3460	**0.6017**	N.A.	N.A.	0.4279	**0.7509**
	SRCC	0.4253	0.3236	N.A.	0.1742	0.3413	**0.5839**	N.A.	N.A.	0.4276	**0.7187**
LIVE	PLCC	0.9434	0.9590	0.9625	0.9560	0.9595	0.9620	**0.9730**	**0.9735**	0.9185	0.9693
	SRCC	0.9436	0.9566	**0.9831**	0.9511	0.9593	0.9493	0.9646	**0.9688**	0.9057	0.9544
CSIQ	PLCC	0.9175	0.9069	0.7743	0.9410	0.9380	0.9463	0.9416	**0.9657**	0.9004	**0.9791**
	SRCC	0.9058	0.9246	0.7806	0.9261	0.9139	0.9279	0.9253	**0.9481**	0.8289	**0.9529**
TID2008	PLCC	0.8555	0.8584	0.8803	0.8430	0.8900	0.9289	**0.9374**	0.9359	0.8238	**0.9586**
	SRCC	0.8480	0.8546	0.8496	0.8505	0.8836	0.9218	0.9189	**0.9394**	0.8213	**0.9489**
TID2013	PLCC	0.8816	0.8830	0.9308	0.8954	0.9020	**0.9419**	0.9221	0.9412	0.7924	**0.9646**
	SRCC	0.8609	0.8785	0.9215	0.8982	0.8940	0.9338	0.9135	**0.9448**	0.8071	**0.9550**
weighted overall	PLCC	0.9042	0.9063	0.8848	0.9090	0.9261	0.9463	0.9449	**0.9563**	0.8612	**0.9680**
	SRCC	0.8944	0.9090	0.8842	0.9064	0.9159	0.9341	0.9322	**0.9514**	0.8417	**0.9527**

Table 1. Performance of our method and other state-of-the-art sharpness metrics on two real and four synthetic datasets

adds the rank loss to the $L1$ regression loss and "B+P+R+M" adds the multi-scale patch extraction method.

As we can see, with the pretrained rank model from AVA images, the performance of "B+P" is improved enormously compared with the baseline model on both synthetically and real blur image datasets. It suggests that the combination rank learning and transfer learning is a good choice for the sharpness assessment applications when the size of available datasets are not large enough. The results also show that the adoption of additional rank loss and multi-scale patch extraction is helpful for the performance improvement.

3.4. Discussion and future work



(a) level=0.1 (b) level=1.6

(c) level=2.8 (d) prediction curve

Fig. 3. Prediction results by self-supervised rank model for video sharpness enhancement

Most of existing works studies the quality of the blurred images in the perspective of fidelity, but the perceptual quality of the properly unsharpened images is likely to be improved for applications such as the user generated micro-videos. Here we show the potential of the proposed method to assess the visual sharpness with the video samples.

In Fig. 3, the chosen 5 seconds length CDN video sample is processed by the standard unsharpen filter from FFmpeg at different levels. The prediction curve in (d) shows the change of average score predicted by the rank model. The video clip is enhanced with more details at medium level, but too many artifacts will be introduced at the very strong unsharpen level, which is consistent with the subjective perception. However, the precision of the video sharpness assessment still needs to be improved with human rating for practical applications. Next step, we will build a crowd-sourced subjective video sharpness dataset which contains thousands of video clips to facilitate the quality prediction for the enhancement purpose.

4. CONCLUSION

We proposed a no-reference image sharpness assessment method based on self-supervised rank learning and multi-scale patch extraction. The Siamese network pretrained with synthetically generated rank labels provides effective image sharpness prior knowledge. The multi-scale gradient guided patch extraction method can increase the quality-sensitive input information for CNN models and further boost the performance. The finetuned single branch from Siamese network achieves the state-of-the-art performance on the widely-used benchmark datasets. In addition, the pretrained rank model is also feasible to predict the perceptual quality of unsharpen videos in the enhancement applications.

5. REFERENCES

[1] Cuong T. Vu, Thien D. Phan, and Damon M. Chandler, "S_3: A spectral and spatial measure of local perceived sharpness in natural images," *IEEE Trans. Image Processing*, vol. 21, no. 3, pp. 934–945, 2012.

[2] Khosro Bahrami and A. C. Kot, "A fast approach for no-reference image sharpness assessment based on maximum local variation," *IEEE Signal Process. Lett.*, vol. 21, no. 6, pp. 751–755, 2014.

[3] Leida Li, Dong Wu, Jinjian Wu, Haoliang Li, Weisi Lin, and Alex C. Kot, "Image sharpness assessment by sparse representation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1085–1097, 2016.

[4] Leida Li, Wenhan Xia, Weisi Lin, Yuming Fang, and Shiqi Wang, "No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1030–1040, 2017.

[5] Ruomei Yan and Ling Shao, "Blind image blur estimation via deep learning," *IEEE Trans. Image Processing*, vol. 25, no. 4, pp. 1910–1921, 2016.

[6] Shaode Yu, ShibinWu, Lei Wang, Fan Jiang, Yaoqin Xie, and Leida Li, "A shallow convolutional neural network for blind image sharpness assessment," *PloS one*, vol. 12, no. 5, 2017.

[7] Mahdi S. Hosseini and Konstantinos N. Plataniotis, "Image sharpness metric based on maxpol convolution kernels," in *ICIP*. 2018, pp. 296–300, IEEE.

[8] Shu Kong, Xiaohui Shen, Zhe L. Lin, Radomír Mech, and Charless C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *ECCV (1)*. 2016, vol. 9905 of *Lecture Notes in Computer Science*, pp. 662–679, Springer.

[9] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *ICCV*. 2017, pp. 1040–1049, IEEE Computer Society.

[10] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[11] Naila Murray, Luca Marchesotti, and Florent Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *CVPR*. 2012, pp. 2408–2415, IEEE Computer Society.

[12] John F. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986.

[13] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[14] Eric C. Larson and Damon M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electronic Imaging*, vol. 19, no. 1, pp. 011006, 2010.

[15] N. Ponomarenko *et al.*, "TID2008-a database for evaluation of full reference visual quality assessment metrics," in *Adv. Mod. Radioelectron*, 2009, vol. 10, pp. 30–45.

[16] Nikolay N. Ponomarenko, Oleg Ieremeiev, Vladimir V. Lukin, Karen O. Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo, "Color image database TID2013: peculiarities and preliminary results," in *EUVIP*. 2013, pp. 106–111, IEEE.

[17] Toni Virtanen, Mikko Nuutinen, Mikko Vaahteranoksa, Pirkko Oittinen, and Jukka Häkkinen, "CID2013: A database for evaluating no-reference image quality assessment algorithms," *IEEE Trans. Image Processing*, vol. 24, no. 1, pp. 390–402, 2015.

[18] Alexandre G. Ciancio, André Luiz N. Targino da Costa, Eduardo A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Trans. Image Processing*, vol. 20, no. 1, pp. 64–75, 2011.

[19] Le Kang, Peng Ye, Yi Li, and David S. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *CVPR*. 2014, pp. 1733–1740, IEEE Computer Society.

[20] Ke Gu, Guangtao Zhai, Weisi Lin, Xiaokang Yang, and Wenjun Zhang, "No-reference image sharpness assessment in autoregressive parameter space," *IEEE Trans. Image Processing*, vol. 24, no. 10, pp. 3218–3231, 2015.