Beyond Human Opinion Scores: Blind Image Quality Assessment based on Synthetic Scores

Peng Ye[1], Jayant Kumar[2], and David Doermann[1]

[1]Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA
[2]Xerox Research Center Webster, NY, USA

[1]{pengye,doermann}@umiacs.umd.edu [2]jayant.kumar@xerox.com

Abstract

State-of-the-art general purpose Blind Image Quality Assessment (BIQA) models rely on examples of distorted images and corresponding human opinion scores to learn a regression function that maps image features to a quality score. These types of models are considered "opinion-aware" (OA) BIQA models. A large set of human scored training examples is usually required to train a reliable OA-BIQA model. However, obtaining human opinion scores through subjective testing is often expensive and time-consuming. It is therefore desirable to develop "opinion-free" (OF) BIQA models that do not require human opinion scores for training.

This paper proposes BLISS (Blind Learning of Image Quality using Synthetic Scores). BLISS is a simple, yet effective method for extending OA-BIQA models to OF-BIQA models. Instead of training on human opinion scores, we propose to train BIQA models on synthetic scores derived from Full-Reference (FR) IQA measures. State-of-the-art FR measures yield high correlation with human opinion scores and can serve as approximations to human opinion scores. Unsupervised rank aggregation is applied to combine different FR measures to generate a synthetic score, which serves as a better "gold standard". Extensive experiments on standard IQA datasets show that BLISS significantly outperforms previous OF-BIQA methods and is comparable to state-of-the-art OA-BIQA methods.

1. Introduction

With the development and popularity of digital imaging devices, digital images have become an important vehicle for representing and communicating information. Unfortunately, digital images may be degraded at various stages of their life cycle and these degradations may lead to the loss of visual information, the poor experience of human viewers and difficulties for image processing and analysis at subsequent stages. The problem of visual information quality assessment arises in numerous image/video processing and computer vision applications, including image compression, image transmission and image retrieval and it plays an important role in these applications.

This paper addresses the problem of general-purpose blind image quality assessment (BIQA). Unlike full-reference IQA (FR-IQA) methods [12, 17, 19, 22], where an undistorted reference image is used to quantify the difference between a distorted image and its corresponding ideal version, BIQA does not require information from the reference image and can be used in applications when reference images are not available. BIQA methods can be broadly classified into two categories: distortion-specific (DS) approaches and general-purpose approaches. DS approaches usually target one or two types of distortions and specific properties of these distortions are examined and embedded in IQA system designs. Meanwhile, general-purpose approaches do not investigate any particular type of distortion but rather they build a general computational model to work universally for different types of distortions.

Current state-of-the-art general purpose BIQA methods [6, 9, 11, 20, 21] rely on examples of distorted images and corresponding human opinion scores to learn a regression function that maps image features to quality scores. This type of model is considered "opinion-aware" (OA) because human opinion scores are provided for the distorted images. A large set of training images with scores is required to train a reliable OA-BIQA model, but obtaining human opinion scores can be time-consuming and expensive. To overcome this limitation, there has been an increasing interest in learning "opinion-free" (OF) BIQA models [7, 8, 18], which do not require human opinion scores for training.

This paper proposes a simple yet effective method for

1063-6919/14 $31.00 © 2014 IEEE
DOI 10.1109/CVPR.2014.540

extending OA-BIQA models to OF-BIQA models. Instead of training on human opinion scores, we propose to train BIQA models on synthetic scores derived from FR-IQA measures. FR measures quantify the differences between distorted images and their undistorted reference images and are easy to obtain. State-of-the-art FR measures yield high correlation with human opinion scores, so they can be used to approximate human opinion scores for training BIQA models. Different FR measures may quantify visual quality in different ways and no single method typically gives the best performance in all situations. We apply unsupervised rank aggregation to combine FR measures for generating a better baseline with which to train. Extensive experiments on three standard IQA datasets show that the proposed method significantly outperforms previous OF-BIQA methods. Furthermore, models trained on the synthetic scores (including FR measures and combined synthetic scores) are comparable to models trained on human opinion scores. This observation implies that we may replace human opinion scores with synthetic scores in training BIQA models without performance loss. The strategy of training on synthetic scores helps to overcome the bottleneck arising from limited training data due to the lack of expensive human opinion scores and allows to use a larger set of data for training.

Our contributions are two-fold. First, we use FR measures to replace human opinion scores for training BIQA models. This is an extremely flexible strategy and can be used with any well established BIQA model. Second, we develop an effective method to combine FR measures in an unsupervised way. The combined synthetic scores yield high correlation with human opinion scores and outperform each individual FR measure anticipated in the combination.

The remainder of this paper describes our BLISS (Blind Learning of Image Quality using Synthetic Scores) system in detail. In Section 2, we briefly review previous work on OF-BIQA and FR measure combination methods. Section 3 describes our unsupervised score combination method and Section 4 provides experimental results on three different IQA datasets. Finally, Section 5 concludes our work.

2. Related Work

2.1. OF-BIQA Models

The first OF-BIQA model seen in the literature was the TMIQ model introduced by Mittal et al. [7]. TMIQ applies probabilistic latent semantic analysis (pLSA) to quality-aware visual words extracted from a large set of pristine and distorted images to uncover latent characteristics or "topics" that are essential for visual quality. The topic mixing coefficients are estimated for the pristine images. Then, given a test image, its estimated topic mixing coefficients are compared to those for the pristine images and their differences are used to infer the quality for the test image. This method performs poorly compared to state-of-the-art OA-BIQA models.

Later Mittal et al. introduced another OF-BIQA model – NIQE [8]. NIQE builds a multivariate Gaussian (MVG) model for natural scene statistic (NSS) features of sharp image regions extracted from pristine images. For a test image, the distance between the MVG constructed from the NSS features of the test image and the MVG model constructed from pristine images is computed as the quality measure. NIQE significantly outperforms TMIQ, yet it does not require distorted images for training and thus is "distortion unaware" and "completely blind". NIQE was shown to perform well on the five types of distortions in the LIVE dataset. This method however may not work universally well on all types of distortions, and when it fails, it is hard to adjust the model to improve the performance since the model does not incorporate examples of distorted images during training.

Xue et al. proposed a quality-aware clustering (QAC) method [18] for OF-BIQA. QAC assigns each image patch a quality score based on a FR measure, then applies clustering to patches at different quality levels. Each cluster centroid is associated with a quality score. For a test image, overlapped patches are extracted, then each patch is compared to the quality aware cluster centroids and the quality score of its nearest neighbor is assigned to the patch. The final quality score for the test image is the weighted average of the patch level quality score.

All these previous OF-BIQA models are inferior to state-of-the-art OA-BIQA models. Our method can be applied to extend existing OA-BIQA models to OF-BIQA models and achieve comparable performance.

2.2. Combining Multiple Full-Reference Measures

FR-measure combination methods aim to combine multiple types of FR-measures to yield a better quality measure [5]. Different FR measures quantify visual quality from different aspects and typically no single method will give the best performance in all situations. Therefore combining multiple FR measures may produce a better IQA measure which outperforms each individual FR measure used in the combination.

Liu et al. [5] introduced a supervised FR measure combination method. A nonlinear regression function is learned to map a feature vector formed by multiple FR measures to a human opinion score. This method requires human opinion scores to train the regression model and thus is not suitable to use in the "opinion-free" scenario.

We have developed an unsupervised FR measure combination method. Given a set of images, we first apply unsupervised rank aggregation to obtain a single consensus ranking based on multiple FR measures. We then adjust a given FR measure based on the consensus ranking to gen-

erate combined synthetic scores. To the best of our knowledge, this is the first work that approaches FR measure combination problem in an unsupervised way.

3. Unsupervised FR measure combination

In this section, we describe our unsupervised FR measure combination method, which involves two steps: generating consensus ranking and score adjustment, and is summarized in Fig. 1.

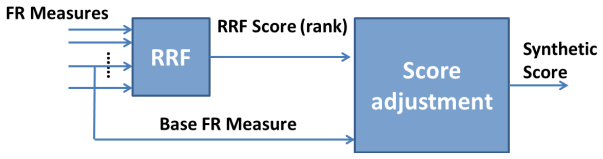

Figure 1: Overview of the proposed unsupervised FR measure combination method.

3.1. Full-Reference Measures

FR IQA measures are computed based on the differences between distorted images and their undistorted reference images. Five different FR measures are used in our experiments, including GMSD [19], VIF [12], FSIM, FSIMC [22] and WSSIM [17]. **GMSD** (Gradient Magnitude Similarity Deviation) computes the pixel-wise gradient magnitude similarity (GMS) and uses this to generate the final measure as the standard deviation of the GMS map. **VIF** (Visual Information Fidelity Index) models image sources using a wavelet domain Gaussian Scale Mixture (GSM) model. It measures the information shared between the source image and the distorted image based on an image distortion channel and a visual distortion model. **FSIM** (Feature Similarity Index) measures similarities of distorted and reference images based on two low-level features: gradient magnitude and phase congruency. Phase congruency is also used to guide the pooling step. **FSIMC** is the FSIM with color information incorporated. **IW-SSIM** (Information content Weighted SSIM) is an enhanced version of the Structural Similarity Index (SSIM) [16] measure, that performs pooling over the SSIM map using weights that are proportional to the local information content.

These five FR measures achieve state-of-the-art performance on standard IQA datasets, so we select them to use in our system.

3.2. Combining Full-Reference Measures

Different FR measures usually lie in different ranges, therefore their values are not comparable and we cannot combine them by simply averaging. Suppose we have K different FR measures and N training images $I_i, i = 1, ..., N$. The first step in combining these FR measures is to construct a consensus ranking via unsupervised rank aggregation. The rank aggregation problem is crucial for many applications such as meta-search, crowdsourcing and social choice. There are many well established methods for this problem. We have selected and used the *Reciprocal Rank Fusion* (RRF) to generate the consensus ranking.

SROCC	GMSD	VIF	FSIM	FSIMC	WSSIM	RRF
GMSD	–	0.9637	0.9896	0.9908	0.9769	0.9911
VIF	0.9637	–	0.9765	0.9745	0.9781	0.9844
FSIM	0.9896	0.9765	–	0.9986	0.9902	0.9978
FSIMC	0.9908	0.9745	0.9986	–	0.9897	0.9976
WSSIM	0.9769	0.9781	0.9902	0.9897	–	0.9939
RRF	0.9911	0.9844	0.9978	0.9976	0.9939	–

Table 1: Pair-wise SROCC between FR measures and RRF-score.

RRF [2] was initially proposed for combining the document rankings from multiple Information Retrieval (IR) systems. Despite its simplicity, RRF is one of the top performing unsupervised rank aggregation methods. According to a recent study in [14], RRF outperforms other competing unsupervised rank aggregation methods on the MQ-agg datasets. The RRF score for image I_i is given by

$$RRFscore(I_i) = \sum_{k=1}^{K} \frac{1}{\gamma + r_k(i)} \tag{1}$$

where $r_k(i)$ is the rank of I_i given by the k-th FR measure and $\gamma = 60$ is a constant. According to [2], the constant γ mitigates the impact of high rankings by outlier systems. The rank of I_i given by the $RRFscore(I_i)$ is denoted as t_i. It is worth noting that VIF, FSIM, FSIMC and WSSIM are quality indexes for which a higher value indicates good quality, while for GMSD, a smaller value indicates good quality. We can negate the GMSD value to make it a quality index.

The pair-wise spearman rank order correlation coefficient (SROCC) between FR measures and RRF scores are shown in Table 1. It can be seen that the RRF score has a high SROCC with each individual FR measure. This means that the rank given by the RRF is consistent with all five FR measures.

We note that the $RRFscore$ cannot be directly used as an image quality score, because the $RRFscore(I_i)$ is a quality indicator of I_i relative to other images in the dataset. It does not directly reflect the quality of I_i. In order to generate a valid quality measure, we propose to adjust the score of a base FR measure according to the RRF rank. Suppose the score of a base FR measure for I_i is y_i and a higher score indicates better quality and a smaller rank t_i. The final combined quality score s is obtained by minimizing the

following objective function.

$$L(s) = \sum_{i=1}^{N}(s_i - y_i)^2 + \lambda \sum_{i<j}(s_i - s_j)\mathbf{1}(t_i > t_j) + (s_j - s_i)\mathbf{1}(t_i < t_j) \quad (2)$$

where $\lambda = \frac{(max(y)-min(y))\lambda_0}{N}$ is a constant balancing factor and $\mathbf{1}(x) = 1$ if x is true, $\mathbf{1}(x) = 0$ otherwise. The first term in the above equation tends to minimize the mean squared error between the s and y. The second term penalizes the inconsistency of pair-wise preferences between s and t. An optimal s can be found by setting the derivative of $L(s)$ with respect to s equal to 0, which yields a simple closed form solution as follows:

$$\frac{\partial L(s)}{\partial s} = 0 \Rightarrow s_i = y_i - \frac{(max(y) - min(y))\lambda_0}{2N} n_i \quad (3)$$

where $n_i = |\{j : t_j < t_i\}| - |\{j : t_j > t_i\}|$. The proposed method requires the sorting of K FR measures for N images, therefore the computational complexity is $O(KNlog(N))$.

The success of the combination method relies on the uniformity of the score distribution. This condition is a typical property of image quality datasets, since if the quality distribution is imbalanced, it would not be a good benchmark for evaluating IQA systems and it would be hard to use the dataset to train any BIQA model to achieve good performance. The proposed method works the best when all FR measures involved in the combination have similar performance. We can compute the pair-wise SROCC between different FR measures, and FR measures that have low correlation with other measures may be removed. We show experimentally in Section 4 that the synthetic scores defined in Eq. 3 have higher correlation with human opinion scores compared to their base FR measures on multiple IQA datasets.

3.3. Training

Once the combined synthetic scores are computed, we can use them to replace human opinion scores for training a BIQA model, and the original OA-BIQA models will become "opinion-free". For training BIQA models, we use Support Vector Regression (SVR). The computational complexity during testing is determined by the base BIQA method. No additional overhead will be introduced by BLISS. We can also use a single FR measure for training, but as will be shown later, moderate performance improvements can be achieved by training on combined synthetic scores.

4. Experiments

4.1. Experimental Protocol

Datasets: Three IQA databases were used in the experiments to demonstrate the effectiveness of the proposed method.

(1) LIVE [13]: The LIVE dataset contains a total of 779 distorted images derived from 29 reference images. Each reference image is distorted by five different distortions – JP2k compression (JP2K), JPEG compression (JPEG), White Gaussian (WN), Gaussian blur (BLUR) and Fast Fading (FF) at 7-8 different levels.

(2) TID2008 [10]: The TID2008 dataset contains 1700 distorted images derived from 25 reference images. A total of 17 different distortions at four degradation levels are included in this dataset. In our experiments, we only examine the four common distortions that are shared by the LIVE dataset, i.e. JP2k, JPEG, WN and BLUR.

(3) CSIQ [4]: The CSIQ dataset consists of 30 reference images and their distorted versions with 6 different types of distortions at 4 to 5 different levels. For the CSIQ dataset, we consider the same four types of distortions – JP2k, JPEG, WN and BLUR.

Evaluation: The performances of IQA measures are evaluated using Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC). It is a common practice to evaluate the FR-IQA measures with a curve-fitting procedure [15], since different IQA measures may not lie in the same range. A similar procedure may also be applied to OF-BIQA models. A logistic regression function, $Q_p = \beta_1(\frac{1}{2} - \frac{1}{exp(\beta_2(Q-\beta_3))}) + \beta_4 Q + \beta_5$, is used to map the original IQA measures to the range of human opinion scores. In our experiments, we randomly select 80% of the reference images and their associated distorted versions for training to obtain $\beta_i, i = 1, ..., 5$ and use the remaining 20% of the reference images and their associated distorted versions for testing. This procedure is repeated 1000 times and the median values of LCC and SROCC are reported.

4.2. Implementation Details

Training Set Construction: We downloaded 100 high resolution images under the Attribution License from flickr.com. The topics of these images include animal, building, indoor scene, forest, human, plant, man-made object, food, sports, etc. The 100 images form our reference image set. Then from each reference image, we generate distorted images with four types of distortions including JPEG and JPEG2k compression, white Gaussian noise and Gaussian blurring. For each distortion, 8 distortion levels are considered. A total of 3300 images are generated to form our training set including 3200 distorted images and 100 reference images. FR measures and combined synthetic scores are computed as groundtruth for the training set. In particular, these scores are mapped to the range of $[0, 100]$ (the lower the better) by a linear function to make it consistent with the range of DMOS in the LIVE dataset and the quality scores of reference images are set to -1.

Base BIQA model: We use CORNIA [20] as the base

BS	*cbsize*	*C*	ϵ	λ_0
5	10000	100	1	4

Table 2: Parameters used in our experiments.

BIQA method because it gives state-of-the-art performance with the use of a linear regression function. Our training set contains 3300 images and training a nonlinear SVR would be time consuming. To speed up the training process, we use the fast liblinear library [3].

Base FR measure: Among the five types of FR measures anticipated in the score combination, we select GMSD as the base FR measure because GMSD [19] yields high linear correlation with human opinion scores without applying any nonlinear fitting and it is very efficient to compute.

Parameters: Several parameters have to be specified for our experiments. (1) In CORNIA feature extraction: *BS* - patch size; *cbsize* - codebook size. (2) For learning the regression function using liblinear: *C* - cost in the loss function; ϵ - parameter in ϵ-insensitive loss function used in ϵ-SVR. The solver we used in liblinear is the L2-regularized L2-loss support vector regression (primal). (3) In Eq. 2: λ_0 the balancing factor. Table 2 shows the values of these parameters.

4.3. Evaluation

4.3.1 Comparison with FR and OF-BIQA Algorithms

We compare our method with previous FR measures: PSNR (Peak Signal to Noise Ratio) and SSIM [16] and state-of-the-art OF-BIQA methods: QAC [18] and NIQE [8]. We test on the four types of distortions which are shared by the LIVE, CISQ and TID2008 datasets (JPEG2K, JPEG, WN and Gaussian BLUR). BLISS is trained on our 3300 flickr image dataset. Test results on each subset and all four subsets combined are reported. BLISS-S is trained using GMSD, which yields the best performance among all the five FR measures. BLISS-C is trained on the combination of five FR measures using Eq. 3. It is worth noting that the same parameters specified in Table 2 are used for experiments on all three datasets.

Results on the LIVE, the CSIQ and the TID2008 datasets are presented in Tables 3, 4 and 5 respectively. We can see that BLISS significantly outperforms the other two competing OF-BIQA models. BLISS-C slightly outperforms BLISS-S. Table 6 shows results from a two sample T-test with 5% significance level, showing the combined score for training outperforms the use of a single FR measure. We also compute the standard deviation (STD) of the SROCC and LCC obtained from 1000-fold cross-validation experiments on the LIVE dataset. As is shown in Table 7, BLISS-C and BLISS-S tend to have smaller STD compared to other methods. This demonstrates the consistency of the proposed method.

SROCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.870	0.885	0.942	0.761	0.867
SSIM	0.939	0.946	0.964	0.907	0.910
NIQE	0.924	0.944	**0.972**	0.939	0.922
QAC	0.868	0.938	0.952	0.918	0.877
BLISS-S	0.911	0.935	0.965	0.954	0.935
BLISS-C	**0.928**	**0.946**	0.970	**0.959**	**0.943**
LCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.873	0.876	0.926	0.766	0.853
SSIM	0.921	0.955	0.982	0.891	0.900
NIQE	0.931	0.957	0.955	0.950	0.919
QAC	0.851	0.943	0.924	0.919	0.863
BLISS-S	0.911	0.958	0.974	0.958	0.933
BLISS-C	**0.933**	**0.965**	**0.976**	**0.967**	**0.939**

Table 3: Results on LIVE.

SROCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.910	0.891	0.933	0.809	0.885
SSIM	0.962	0.954	0.912	0.960	0.934
NIQE	0.925	0.883	0.835	0.907	0.887
QAC	0.888	**0.912**	**0.865**	0.852	0.858
BLISS-S	0.935	0.889	0.815	0.913	0.899
BLISS-C	**0.949**	0.910	0.848	**0.917**	**0.918**
LCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.861	0.887	0.946	0.771	0.856
SSIM	0.906	0.982	0.910	0.945	0.930
NIQE	0.934	0.945	0.834	0.929	0.904
QAC	0.896	0.947	**0.911**	0.861	0.890
BLISS-S	0.951	0.952	0.833	0.944	0.927
BLISS-C	**0.965**	**0.959**	0.863	**0.945**	**0.938**

Table 4: Results on CSIQ.

SROCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.838	0.887	0.917	0.929	0.869
SSIM	0.962	0.932	0.847	0.959	0.905
NIQE	0.887	0.875	**0.817**	0.845	0.795
QAC	0.890	0.887	0.717	0.856	0.861
BLISS-S	0.919	0.922	0.779	0.869	0.898
BLISS-C	**0.923**	**0.926**	0.807	**0.880**	**0.899**
LCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.888	0.880	0.945	0.914	0.845
SSIM	0.971	0.964	0.816	0.954	0.902
NIQE	0.911	0.921	**0.796**	0.849	0.804
QAC	0.878	0.917	0.736	0.842	0.842
BLISS-S	**0.945**	**0.955**	0.748	0.875	0.910
BLISS-C	0.941	0.952	0.770	**0.880**	**0.917**

Table 5: Results on TID2008.

SROCC	PSNR	SSIM	NIQE	QAC	BLISS-S	BLISS-C
PSNR	0	-1	-1	-1	-1	-1
SSIM	1	0	-1	1	-1	-1
NIQE	1	1	0	1	-1	-1
QAC	1	-1	-1	0	-1	-1
BLISS-S	1	1	1	1	0	-1
BLISS-C	1	1	1	1	1	0

Table 6: Results of the two sample T-test performed between SROCC values obtained by different measures. 1 (-1) implies the algorithm in the row is statistically superior (inferior) to the algorithm in the column. 0 indicates the algorithm in the row is statistically equivalent to the algorithm in the column.

	PSNR	SSIM	NIQE	QAC	BLISS-S	BLISS-C
SROCC	0.0328	0.0175	0.0180	0.0237	0.0164	0.0143
LCC	0.0302	0.0181	0.0160	0.0243	0.0152	0.0137

Table 7: Standard deviation of SROCC and LCC for 1000 iterations of experiments on LIVE.

SROCC	CORNIA	BRISQUE	LCC	CORNIA	BRISQUE
DMOS	0.881	0.882	DMOS	0.883	0.892
SS	**0.905**	0.897	SS	**0.925**	0.893

Table 8: Train on LIVE and test on TID2008

SROCC	CORNIA	BRISQUE	LCC	CORNIA	BRISQUE
DMOS	0.899	0.899	DMOS	0.914	0.927
SS	**0.908**	0.895	SS	**0.928**	0.912

Table 9: Train on LIVE and test on CSIQ

4.3.2 Comparison with OA-BIQA Algorithms

In the second set of experiments, we use human opinion scores (DMOS) and synthetic scores (SS) to train two state-of-the-art BIQA models BRISQE [6] and CORNIA [20] respectively. We train these models on images with JP2k, JPEG, WN and GBLUR distortions in the LIVE dataset and test on the images with the same four types of distortions in the TID2008 dataset and the CSIQ dataset. The median SROCC and LCC evaluated on all four types of distortions from 1000-fold cross-validation experiments are reported in Tables 8 and 9. In this experiment, CORNIA is trained using a linear SVR with the parameters specified in Table 2 and BRISQUE is trained using SVR with a RBF kernel[1]. As is shown in Tables 8 and 9, models trained on the synthetic scores[2] are comparable to models trained on the human opinion score. BLISS works well with both CORNIA and BRISQUE. The best performance is achieved by training on the synthetic scores. This result implies that we can replace human opinion scores with synthetic scores without loss of performance.

4.3.3 Comparison of the combined synthetic score and FR measures

Evaluation on LIVE To demonstrate the effectiveness of our score combination method, we tested the five FR measures and the combined measures on the LIVE dataset [13]. Table 10 shows the LCC and SROCC obtained using each FR measure independently and the synthetic score based on the corresponding FR measures[3]. As is shown in this table, by exploiting the overall rank information, combined measures consistently improve over each individual measure. All five FR measures have high SROCC on LIVE, but the combined measures slightly outperform their base measures in SROCC. The LCC values are significantly improved. The conventional method for improving the LCC of a FR measure relies on fitting a nonlinear logistic function, but human opinion scores are required to find the optimal parameters in the logistics function. The proposed method improves LCC in a fully unsupervised way. One key factor to the success of BLISS is that BLISS use synthetic scores that have high linear correlation with human opinion scores to train the BIQA model.

SROCC	GMSD	VIF	FSIM	FSIMC	WSSIM
original	0.960	0.964	0.963	0.965	0.957
SS, $\lambda_0 = 1$	0.967	0.970	0.968	0.968	0.966
SS, $\lambda_0 = 4$	0.969	0.970	0.969	0.968	0.968
LCC	GMSD	VIF	FSIM	FSIMC	WSSIM
original	0.942	0.941	0.859	0.860	0.803
SS, $\lambda_0 = 1$	0.965	0.958	0.956	0.956	0.945
SS, $\lambda_0 = 4$	0.967	0.963	0.967	0.967	0.966

Table 10: Test FR measures on LIVE (779 distorted images): 'original'–correlation between original FR measures and DMOS; 'SS'–correlation between synthetic scores and DMOS.

Next we examined the effect of the balancing constant λ_0. Figs. 2 and 3 show how the SROCC and LCC of the combined scores change with different values of λ_0. *FR*-SS represents the synthetic score with *FR* as the base measure. $\lambda_0 = 0$ corresponds to using the original FR measures. When λ_0 is very small, the synthetic score is dominated by the base FR measure and the performance is primarily determined by the base FR measure. As we increase the value of λ_0, the importance of the rank information increases. When $\lambda_0 \geq 1$, the value of LCC and SROCC is not very sensitive to the value of λ_0.

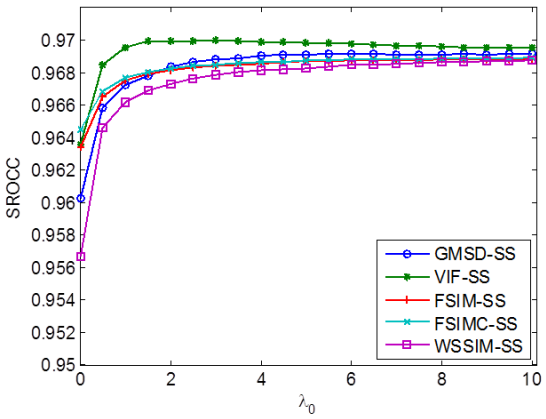

Figure 2: Effect of λ_0 on SROCC (Tested on LIVE).

To demonstrate that the proposed method is robust to

[1]Parameters for training BRISQUE model using libsvm are suggested by the author as "-b 1 -s 3 -g 0.05 -c 1024 -p 1".

[2]Note that CORNIA+SS is equivalent to BLISS-C in Tables 3, 4 and 5.

[3]No nonlinear fitting procedure is applied in this experiment.

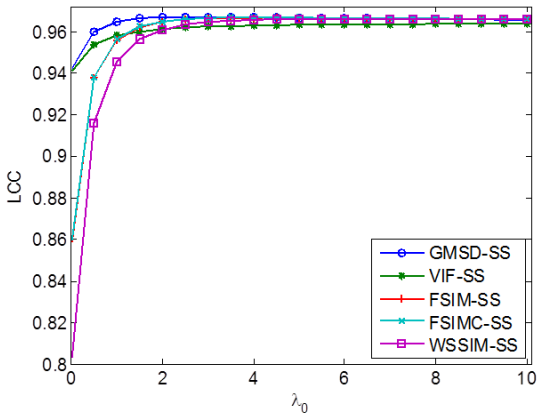

Figure 3: Effect of λ_0 on LCC (Tested on LIVE).

changes in the dataset size, we randomly sample a subset of the LIVE dataset and apply our method on the subset. This process is repeated 1000 times and median values of SROCC and LCC are presented in Figs. 4 and 5. We can see that the performance decreases only slightly as we reduce the dataset size to 10% of original size.

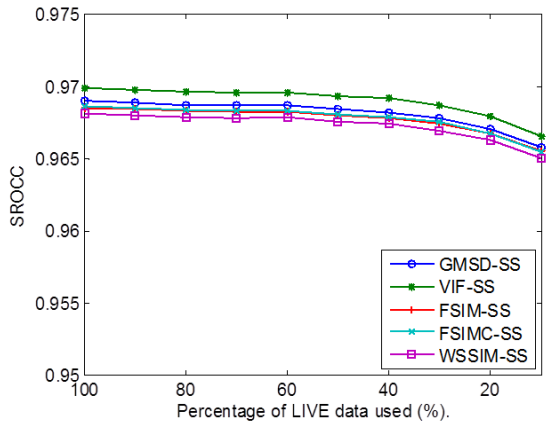

Figure 4: Effect of dataset size on LCC (Tested on LIVE, $\lambda_0 = 4$).

Evaluation on TID2008 All the five FR measures have similar performance in terms of SROCC on the LIVE dataset. However, on the TID2008 dataset, the performance of the five FR measures varies a lot. We compute the pairwise SROCC between FR measures and look at the average value of the SROCC. For GMSD, VIF, FISM, FSIMC and WSSIM, the average SROCCs are 0.891, 0.784, 0.930, 0.930, 0.900 respectively. It is obvious that VIF is not consistent with the other four types of FR measures. Therefore, VIF is discarded for computing the RRF score. Table 11 presents the evaluation results on 1700 distorted images in the TID2008 dataset. We see that GMSD performs the best

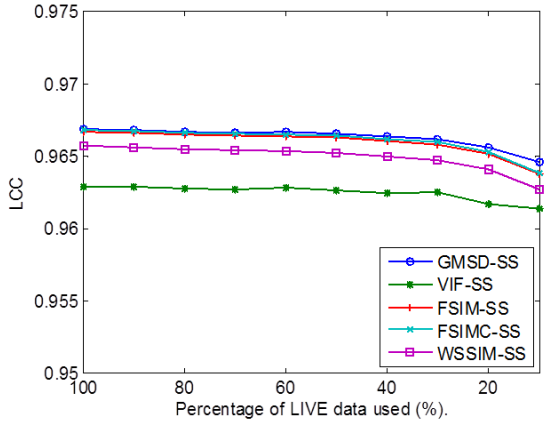

Figure 5: Effect of dataset size on SROCC (Tested on LIVE, $\lambda_0 = 4$).

SROCC	GMSD	FSIM	FSIMC	WSSIM
original	0.891	0.881	0.884	0.856
SS, $\lambda_0 = 1$	0.898	0.892	0.892	0.887
SS, $\lambda_0 = 4$	0.896	0.893	0.893	0.892
LCC	GMSD	FSIM	FSIMC	WSSIM
original	0.872	0.830	0.834	0.809
SS, $\lambda_0 = 1$	0.885	0.884	0.884	0.884
SS, $\lambda_0 = 4$	0.878	0.878	0.878	0.878

Table 11: Test FR measures on TID2008 (1700 distorted images): 'original'–correlation between original FR measures and DMOS; 'SS'–correlation between synthetic scores and DMOS.

among all FR measures and the synthetic scores slightly outperform GMSD.

4.4. The ambiguity of human opinion score

As shown in our experimental results in Section 4.3.2, often times models trained on the synthetic scores can outperform models trained on human opinion scores. This result may be explained by the inherent ambiguity of human opinion scores. The mean opinion score (MOS) test is the most widely used subjective test for obtaining groundtruth data for image quality dataset. However, there are many known problems with the MOS test and pair-wise comparison based tests have been proposed as an alternative to the MOS test [1]. How to properly design the subjective IQA test is still an open problem. We may consider the current ground-truth labels in the IQA dataset as a noisy approximation to the unknown "gold-standard".

Furthermore, human opinion scores in different datasets were obtained under different experimental conditions. Therefore, the MOS labels in different datasets may not be consistent. On the other hand, FR measures are objective measures that capture the inherent properties of image distortion which do not vary from dataset to dataset. It is

therefore possible to train a better prediction model using synthetic scores.

5. Conclusions

An unsupervised method is presented which combines multiple FR measures into a single synthetic score. The combined synthetic score outperforms each individual FR measure. We use the combined synthetic score or a single FR measure to replace the human opinion score in training conventional OA-BIQA model. In both cases, the results are obtained at significantly reduced cost. The BIQA models trained on synthetic scores are comparable to models trained on human opinion scores and significantly outperform previous OF-BIQA models.

Acknowledgment

The partial support of this research by DARPA through BBN/DARPA Award HR0011-08-C-0004 under subcontract 9500009235, the US Government through NSF Awards IIS-0812111 and IIS-1262122 is gratefully acknowledged.

References

[1] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. In *Proceedings of ACM Multimedia*, pages 491–500, 2009.

[2] G. V. Cormack, C. L. A. Clarke, and S. Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009.

[3] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, pages 1871–1874, 2008.

[4] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010.

[5] T.-J. Liu, W. Lin, and C.-C. J. Kuo. Image quality assessment using multi-method fusion. *IEEE Transactions on Image Processing*, 22(5):1793–1807, 2013.

[6] A. Mittal, A. Moorthy, and A. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[7] A. Mittal, G. S. Muralidhar, and A. C. Bovik. Blind image quality assessment without human training using latent quality factors. *IEEE Signal Processing Letters*, 19 (2):75–78, 2012.

[8] A. Mittal, G. S. Muralidhar, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20 (3):209–212, 2013.

[9] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, Dec. 2011.

[10] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radio Electronics*, 10:30–45, 2009.

[11] M. Saad, A. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, Aug. 2012.

[12] H. R. Sheikh, A. C. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, Dec. 2005.

[13] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. LIVE image quality assessment database release 2. Online, http://live.ece.utexas.edu/research/quality.

[14] M. N. Volkovs and R. S. Zemel. Supervised CRF framework for preference aggregation. In *International Conference on Information and Knowledge Management (CIKM)*, pages 89–98, 2013.

[15] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment – Phase II, Aug. 2003.

[16] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[17] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, 2011.

[18] W. Xue, L. Zhang, and X. Mou. Learning without human scores for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 995–1002, 2013.

[19] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Trans. on Image Processing*, 23(2):684–695, 2014.

[20] 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)*, pages 1098–1105, 2012.

[21] P. Ye, J. Kumar, L. Kang, and D. Doermann. Real-time no-reference image quality assessment based on filter learning. In *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 987–994, 2013.

[22] L. Zhang, D. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.