Blind Image Quality Assessment Without Human Training Using Latent Quality Factors

Anish Mittal, Gautam S. Muralidhar, *Student Member, IEEE*, Joydeep Ghosh, *Fellow, IEEE*, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*—We propose a highly unsupervised, training free, no reference image quality assessment (IQA) model that is based on the hypothesis that distorted images have certain latent characteristics that differ from those of "natural" or "pristine" images. These latent characteristics are uncovered by applying a "topic model" to visual words extracted from an assortment of pristine and distorted images. For the latent characteristics to be discriminatory between pristine and distorted images, the choice of the visual words is important. We extract quality-aware visual words that are based on natural scene statistic features [1]. We show that the similarity between the probability of occurrence of the different topics in an unseen image and the distribution of latent topics averaged over a large number of pristine natural images yields a quality measure. This measure correlates well with human difference mean opinion scores on the LIVE IQA database [2].**

***Index Terms*—Distortions, image quality, local artifact, pLSA, topic model.**

I. Introduction

THE past decade has witnessed great advances in multimedia technology with development of a great variety of new handheld devices and smart phones. This has resulted in considerable research to provide best quality-of-experience (QoE) to the end-users. While conventional QoE algorithms primarily focused on optimizing throughput, buffer-lengths, and capacity of delivery networks, perceptual optimization of multimedia services has also fast gained importance, especially in an era of growing video traffic coupled with bandwidth paucity. These perceptual approaches use objective measures of visual quality to deliver the optimum QoE to the end-user.

Full reference (FR) image quality algorithms require both the distorted image and the pristine image, based on which the quality of the distorted image is assessed [1]. No-reference algorithms do not rely on the availability of pristine images. Current state of art no-reference image quality assessment algorithms can predict image quality without knowing the type of distortion the images are afflicted with [1], [4]–[8]. However, these algorithms do require auxiliary information in the form of human opinion scores that are used for learning regression-based models to predict the quality of distorted images. Simulating different kinds of source and channel distortions, and then obtaining human opinion scores is an expensive and time consuming procedure. Further, these methods are limited in application by the distortions they are trained on. Towards this goal, we propose a *highly unsupervised* image quality assessment model that requires no training on human opinion scores. All that is needed is a binary label for each image in our "training set" indicating whether that image is pristine or not. Our approach is based on the hypothesis that distorted images have *"loadings"* or probabilities over latent *"distortion aware"* topics which we refer to as *"latent quality factors"*(LQFs), that differ from the *"loadings"* for *"natural"* or *"pristine"* images. Latent topics are discovered by modeling images as distributions over representative visual words extracted from an assortment of pristine and distorted images. For the LQFs to be discriminatory between pristine and distorted images, the choice of the visual words is important. We form the visual word vocabulary by clustering *"quality-aware"* features that best describe local image distortions [1]. To discover the LQFs, we employ probabilistic latent semantic analysis (pLSA), which was first used to discover meaningful topics that were latent in a large corpora of text documents [9]. Sivic *et al.* [10] subsequently used this model to discover latent object categories from real world images by modeling the images as distributions over visual words in a vocabulary formed by clustering local appearance features such as SIFT features [11]. Using the discovered latent characteristics from pristine and distorted images, we propose a new model of image quality which is based on computing how different the loadings found in an unseen image are when compared to the loadings found in a separate set of pristine images. We show that this quality measure correlates reasonably well with difference mean opinion scores (DMOS) on the LIVE IQA database [2].

II. Proposed Approach

A. Probabilistic Latent Semantic Analysis

We first briefly review the pLSA model of Hofmann [9], which was first employed to discover latent topics embedded within a collection of text documents in a corpus. In our scenario, the corpus is an assortment of pristine and distorted images. Let N be the total number of pristine and distorted images

Manuscript received October 13, 2011; revised November 29, 2011; accepted December 01, 2011. Date of publication December 13, 2011; date of current version December 22, 2011. This work was supported by NSF Grant IIS-1116656 and by Intel and Cisco under the VAWN program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yong Man Ro.

The authors are with the University of Texas at Austin, Austin, TX 78712 USA (e-mail: mittal.anish@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2011.2179293

[1] By "pristine," we mean an image that has not been subjected to any distortions beyond those that normally occur during a quality photo shoot under good conditions. However, no image is truly without distortions, which casts some doubts on the basic assumptions of full reference algorithms [3].

1070-9908/$26.00 © 2011 IEEE

contained in the corpus. Every image in the corpus can be described as an empirical distribution over *"visual words"* from a *"visual word vocabulary"*. Let W be the total number of distinct visual words contained in the vocabulary. Let us suppose that the jth image in the corpus, I_j, is comprised of W_j words with the ith word denoted by w_{ij}. We further assume that there are K latent topics that pervade the collection of images in the corpus, with the kth topic denoted by the indicator variable z_k. Every image can be represented as a distribution over K topics, with a latent topic z_k associated with every word w_{ij} in the image I_j.

The conditional probability of observing a word w_{ij} given an image I_j is obtained by marginalizing over the latent topics z_k i.e., $P(w_{ij}|I_j) = \sum_{k=1}^{K} P(z_k|I_j)P(w_{ij}|z_k)$. Thus, the kth topic is represented by the W-dimensional vector $P(w|z_k)$, and the loadings of image I_j across the topics by K-dimensional vector $P(z_k|I_j)$. The topics that pervade the collection of images, and their loadings given an image, can be inferred by finding the model that best explains the probability distribution of the visual words in the images. This is the maximum likelihood estimate of the model parameters, which can be computed using the expectation-maximization (EM) algorithm described in [9]. Note that the pLSA framework uses the "bag of words" approach as the spatial arrangement of word occurrences is not taken into account.

B. *Quality-Aware Features*

While we do not use perceptually relevant human scores to train our model, we do rely on natural scene statistic (NSS) features to capture perceptually relevant scene properties. Specifically, we use the NSS features introduced in the Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [1] to compute features over every image patch. The principle behind BRISQUE feature design is that natural images obey specific regular statistical properties, which are disrupted by the presence of distortions [12]. Quantifying such deviations from regularity of natural scene statistics is quite useful for assessing the perceptual quality of images [1], [4], [5], [7], [8], [13]. As shown in [1], [4], [5], [7], [8], [13], such characterization is sufficient not only to quantify naturalness, but also to identify the distortions the images are afflicted with. The BRISQUE NSS features naturally blend into the topic modeling framework where the inferred topics emerge out as LQFs that are characteristic of "pristineness" and of the artifacts induced by different distortions.

The BRISQUE features represent statistics of normalized luminance coefficients of images [1]. The BRISQUE features also utilize a model for pair-wise products of neighboring (normalized) luminance values. The BRISQUE feature vector computed over each patch is a 36-dimensional vector.

C. *Construction of Visual Vocabulary*

The approach we take to build the visual word vocabulary is similar to that described by Sivic *et al.* [10], the key and crucial difference being the choice of features used to construct the visual vocabulary—quality based [1] vs local appearance based [11]. The visual words are formed by clustering features computed from multiple patches across all the images in the collection. Each image is divided into overlapping patches of size 64×64, with an overlap of 8×8 between neighboring patches, and local BRISQUE features are computed over each patch. We did not observe a significant difference in performance when the patch size was changed to 32×32, with an overlap of 8×8 between neighboring patches. Feature vectors from all patches across all images are clustered into $W = 400$ visual words using the k-means clustering algorithm with the squared euclidean distance metric. Again, we observed that 400 visual words were sufficient and no improvement in performance was obtained when the visual word count was increased to 1000. This is followed by vector quantization, where every patch is assigned to the nearest cluster center. This yields an empirical distribution over the visual words. Note that the use of visual words has been recently explored for assessing image quality by Ye and Doerman [14]. However, in their approach, visual words were formed using Gabor based local appearance descriptors as opposed to using "quality-aware" visual words. Also, Ye and Doerman used a supervised approach that involved training with DMOS scores, while our approach is based on pLSA, which is a completely unsupervised topic model.

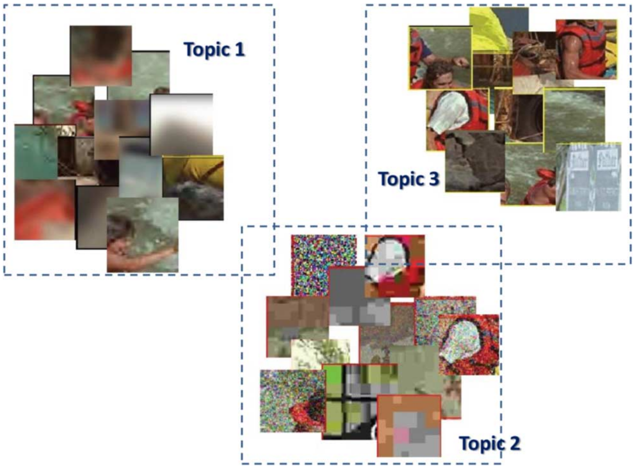

Fig. 1. Examples of image patches assigned to three LQFs discovered by the pLSA model.

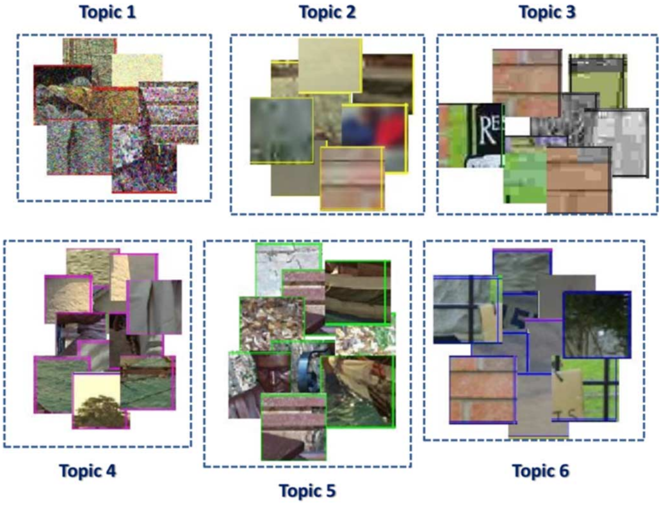

Fig. 2. Examples of image patches assigned to six LQFs discovered by the pLSA model.

TABLE I
MEDIAN SROCC WITH ASSOCIATED STANDARD DEVIATION SCORES ACROSS 1000 TRAIN-TEST EXPERIMENTS ON THE LIVE IQA DATABASE

	JP2k	JPEG	WN	Blur	FF	All
PSNR	0.86 ± 0.05	0.88 ± 0.05	0.94 ± 0.03	0.75 ± 0.10	0.87 ± 0.06	0.86 ± 0.04
BRISQUE	0.90 ± 0.04	0.94 ± 0.03	0.98 ± 0.01	0.95 ± 0.02	0.88 ± 0.05	0.93 ± 0.02
Proposed Approach (3 topics)	*0.84 ± 0.05*	*0.87 ± 0.23*	*0.84 ± 0.30*	*0.87 ± 0.09*	*0.76 ± 0.07*	*0.80 ± 0.12*
Proposed Approach (4 topics)	*0.84 ± 0.07*	*0.87 ± 0.17*	*0.86 ± 0.21*	*0.84 ± 0.15*	*0.76 ± 0.07*	*0.76 ± 0.09*
Proposed Approach (6 topics)	*0.85 ± 0.06*	*0.88 ± 0.08*	*0.80 ± 0.23*	*0.87 ± 0.13*	*0.77 ± 0.08*	*0.80 ± 0.07*

D. Image Quality Inference

The topic specific word distribution $P(w|z)$ learned from an existing collection of images comprising of both pristine and distorted images via the model-fitting procedure (EM) is used to infer the latent quality factors in a new image not contained in the collection. When a new image I_{new} is observed, the $P(z|I_{new})$ can be computed using the "fold-in" heuristic described in [9]. Essentially, for the new image I_{new}, the empirical visual word distribution, i.e., $P(w|I_{new})$ is first computed. Then, the $P(z|I_{new})$ are sought such that the Kullback–Leibler divergence between the empirical visual word distribution $P(w|I_{new})$ and $P(w|I_{new}) = \sum_{k=1}^{K} P(z_k|I_{new})P(w|z_k)$ is minimized. $P(z|I_{new})$ are again estimated by running EM, but this time only the loadings are updated, while $P(w|z)$ estimated during the model fitting procedure is held fixed.

The vector of estimated loadings of the new image I_{new} (i.e., the estimated $P(z|I_{new})$) is now compared to the vector of the estimated loadings of each pristine image in the existing collection. The loadings of the pristine images in the existing collection are obtained during the model fitting procedure that was carried out to learn the topic-specific word distribution $P(w|z)$. The comparison is done by computing the dot product between the two vectors. The average dot product computed across all pristine images in the existing collection is indicative of the image quality. Mathematically, this can be represented as $Q(I_{new}) = 1/N_p \sum_{n=1}^{N_p} P(z|I_{new})'P(z|I_n)$, where $Q(I_{new})$ is the inferred quality of the new image, $'$ is the transpose operator, and I_n is the nth pristine image in the existing collection, which comprises of N_p pristine images. Due to the linearity of the dot product, we can write this as $Q(I_{new}) = P(z|I_{new})'(1/N_p \sum_{n=1}^{N_p} P(z|I_n))$. This expression intuitively suggests that our quality measure can be seen as an estimate of a measure of disruption relative to an "anchor" point learned from pristine images, where the "anchor" refers to the average loadings of the pristine images given by $1/N_p \sum_{n=1}^{N_p} P(z|I_n)$.

III. EXPERIMENTS AND RESULTS

We have conducted our analysis of LQFs and image quality inference on the LIVE IQA database [2], which contains 29 reference images and five distortion types—JPEG, JPEG 2000 (JP2K), Blur, White Noise and Fast Fading (FF). We performed a 1000-fold validation experiment on the LIVE IQA database [2], where in, in each run of the experiment, we randomly select six reference images and their associated distorted versions for performance evaluation, and 23 (different) reference images and their associated distorted versions for learning the LQFs. This ensures that the two sets are completely disjoint and they neither share content, nor do they share specific distortion severities. The EM model-fitting procedure in pLSA is sensitive to the choice of the initial parameters, which are selected at random. To ensure convergence to the best model during the learning process, we ran EM 20 times, with each EM run initialized with different parameters chosen randomly. We then picked the model that yielded the highest log likelihood score. For the analysis of the LQFs and image quality inference, we experimented with two, three, four, and six topics.

A. Analysis of Latent Quality Factors

We first analyzed the loadings that were learned from the pristine and distorted image set. Fig. 1 illustrates examples of image patches assigned to each discovered LQF when the number of latent factors was fixed at 3. Each cluster of image patches in Figs. 1 and 2 contains examples of image patches corresponding to the most probable words in a topic. As can be seen from the figure, the image patches that are most representative of each LQF are different. For example, the set of patches prominent in topic 1 appear to be afflicted with distortions that decrease the energy of the pristine signal (at the same scale or in the same band) due to a low-pass operation such as blur or JP2K, while the set of patches in topic 2 seemingly belong to a set of distortions that increase the energy of the pristine signal, such as white noise or JPEG blocking. Likewise, pristine image patches are most prominent in topic 3. When the number of LQFs is increased to 6, image patches that correspond to white noise and JPEG blocking artifacts are assigned to different LQFs as illustrated by Fig. 2. Also, pristine patches begin to separate into different topics. In contrast, patches with white noise and JPEG blocking artifacts tend to combine with pristine patches when number of topics is reduced to 2.

B. Image Quality Inference

Tables I and II list the median values along with the standard deviation of the Spearman rank ordered correlation coefficient (SROCC) and linear correlation coefficient (LCC), respectively, for our new, completely unsupervised quality assessment measure based on LQFs over 1000 trials for three, four, and six topics. For comparison, we also show the median SROCC and LCC values over 1000 trials for the peak signal to noise ratio (PSNR) metric (a full reference IQA metric), and the supervised BRISQUE metric (a blind IQA metric). The results in Tables I and II clearly show that the proposed quality measure correlates reasonably well with human perception. Although this early model does not yet compete with full reference IQA models and IQA models that are trained on DMOS scores, these results are very promising considering that there is no need to train on DMOS scores thereby avoiding considerable expense.

TABLE II
MEDIAN LCC WITH ASSOCIATED STANDARD DEVIATION SCORES ACROSS 1000 TRAIN-TEST EXPERIMENTS ON THE LIVE IQA DATABASE

	JP2k	JPEG	WN	Blur	FF	All
PSNR	0.88 ± 0.05	0.91 ± 0.05	0.93±0.02	0.79 ± 0.09	0.89 ± 0.05	0.86 ± 0.03
BRISQUE	0.92 ± 0.04	0.96 ± 0.02	0.99±0.00	0.96 ± 0.02	0.93 ± 0.04	0.94 ± 0.02
Proposed Approach (3 topics)	*0.87 ± 0.05*	*0.89 ± 0.05*	*0.88 ± 0.12*	*0.85 ± 0.07*	*0.82 ± 0.06*	*0.78 ± 0.09*
Proposed Approach (4 topics)	*0.87 ± 0.05*	*0.89 ± 0.06*	*0.87 ± 0.08*	*0.85 ± 0.08*	*0.82 ± 0.06*	*0.76 ± 0.08*
Proposed Approach (6 topics)	*0.87 ± 0.05*	*0.90 ± 0.06*	*0.87 ± 0.09*	*0.88 ± 0.07*	*0.84 ± 0.06*	*0.79 ± 0.08*

TABLE III
MEDIAN SROCC AND LCC ACROSS 1000 TRAIN-TEST EXPERIMENTS USING OUR PROPOSED APPROACH FOR THREE TOPICS ON LIVE IQA DATABASE WITH NO JP2K DURING MODEL LEARNING

	JP2k	JPEG	WN	Blur	FF	All
SROCC	*0.84±0.06*	*0.88*±0.16	*0.77±0.24*	*0.86±0.08*	*0.75±0.08*	*0.80±0.09*
LCC	*0.87±0.05*	*0.90±0.04*	*0.87±0.11*	*0.83±0.07*	*0.81±0.07*	*0.79±0.07*

TABLE IV
MEDIAN SROCC AND LCC ACROSS 1000 TRAIN-TEST EXPERIMENTS USING OUR PROPOSED APPROACH FOR THREE TOPICS ON LIVE IQA DATABASE WITH SIFT FEATURES

	JP2k	JPEG	WN	Blur	FF	All
SROCC	*0.00±0.13*	*0.00*±0.11	*0.00±0.11*	*0.02±0.17*	*0.00±0.16*	*0.00±0.06*
LCC	*0.30±0.12*	*0.26±0.14*	*0.30±0.11*	*0.34±0.13*	*0.36±0.13*	*0.15±0.09*

Table III lists the median SROCC and LCC values when the JP2K distortion category is left out while learning the LQFs. These results are encouraging in that the proposed approach performs well on JP2K images contained in the test set even though they had not been seen previously. This can be attributed to the fact that certain latent characteristics of JP2K, such as local blurring, are uncovered due to patches from Gaussian blur images that were used to form the visual words.

C. Importance of Quality-Aware Features

We also experimented by replacing our quality-aware, NSS based BRISQUE features [1] with SIFT features [11] in the pLSA framework. The SIFT features were computed at key points, which were densely sampled at increments of patch size. This resulted in feature computation within overlapping patches of the same size as used in the computation of the BRISQUE features. The SIFT features were clustered to yield a visual word vocabulary comprising of 400 words and the number of topics was set to 3. Table IV lists the median SROCC and LCC values when the SIFT features were used instead of the BRISQUE features. It is evident from Table IV that SIFT features do not capture image quality well owing to a lack of distinction provided by SIFT features across pristine and distorted images. The numbers in Table IV highlight the importance of using well designed, NSS based quality-aware features such as the BRISQUE features [1] for blind image quality assessment.

IV. CONCLUSION AND FUTURE WORK

We presented a completely novel way of determining perceptual image quality based on applying a topic model on image patches represented in a suitable quality-aware space, and then examining the topic distributions for each image. This method obviates the manually intensive process of obtaining DMOS scores. The resulting image quality model can be visualized as a measure of disruption relative to an 'anchor' point learned from pristine images. We have shown that our quality model correlates reasonably well with DMOS scores on the LIVE IQA database [2].

Our future work will be focused on gaining a better understanding of the interplay between the number of topics and inferred image quality. We have already experimented with a more sophisticated topic model such as Latent Dirichlet Allocation (LDA) [15] but the small size of the dataset led to overfitting of hyperparameters yielding poorer performance than pLSA. Future work would involve learning the framework using LDA on a larger size simulated dataset, which will be easy to prepare given that our algorithm does not require human opinion scores.

REFERENCES

[1] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Nov. 2011.

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

[3] A. C. Bovik, "Meditations on video quality," *IEEE Multimedia Commun. E-Lett.*, vol. 4, no. 4, pp. 4–10, 2009.

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

[5] M. Saad and A. C. Bovik, "DCT statistics model-based blind image quality assessment," in *IEEE Int. Conf. Image Processing*, Brussels, , Belgium, Sep. 2011.

[6] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2011.

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

[8] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 583–586, 2010.

[9] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1, pp. 177–196, 2001.

[10] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *IEEE Int. Conf. Computer Vision*, 2005, pp. 370–377.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[12] A. K. Moorthy and A. C. Bovik, "Statistics of natural image distortions," in *IEEE Int. Conf. Acoustics Speech and Signal Processing*, 2010, pp. 962–965.

[13] Z. Wang, "Applications of objective image quality assessment methods," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 137–142, Nov. 2011.

[14] P. Ye and D. Doerman, "No-reference image quality assessment based on visual codebook," in *IEEE Int. Conf. Image Processing*, 2011.

[15] D. M. Blei, A. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.