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 fully unsupervised no reference image quality assessment (IQA) model that is based on the hypothesis that distorted images have latent characteristics that are indicative of type of artifacts manifested by them, and that the latent characteristics of distorted images differ from the latent characteristics of 'natural' or 'pristine' images. These latent characteristics are uncovered by applying a 'topic model' to suitably chosen quality-aware visual words extracted from the images. The difference between the probability of occurrence of these latent characteristics in unseen images and latent characteristics learned from a large number of pristine natural images yields a quality measure. We show that this measure correlates well with human difference mean opinion scores on the LIVE IQA database [1].**

***Index Terms*—Local artifact, image quality, topic model, pLSA, distortions.**

I. Introduction

The past decade has witnessed great advances in multimedia technology, and a great variety of new devices for capture, storage, compression, transmission, and display of audiovisual stimuli have been developed. This has resulted in considerable research in providing the best quality of experience (QoE) to the end-users. While conventional QoE algorithms primarily focus on optimizing throughput, buffer-lengths, and capacity of delivery networks, perceptual optimization of multimedia services is also fast gaining importance, especially in an era of growing video traffic coupled with bandwidth paucity. These perceptual approaches attempt to deliver the optimum QoE to the end-user by utilizing objective measures of visual quality.

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 [3], [4], [5], [6], [7], [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 *fully unsupervised* image quality assessment model that requires no training on human opinion scores. Our approach is based on the hypothesis that distorted images have certain latent characteristics, which we refer to as latent quality factors (LQFs) that differ from the corresponding LQFs of 'natural' or 'pristine' images. These LQFs are discovered by modeling images as distributions over representative 'quality-aware' visual words, where the visual word vocabulary is formed by clustering 'quality-aware' features that best describe local image distortions [3]. The model we use here, popularly known as probablistic latent semantic analysis (pLSA), 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]. In our proposed approach, the topics or LQFs discovered by the pLSA model correspond to artifacts introduced by different kinds of distortions such as 'blockiness', 'blurriness', and 'graininess'. Using the discovered LQFs from pristine and distorted images, we propose a new model of image quality. Our model is based on computing how different the probability of occurrence of LQFs discovered in an unseen image is when compared to the previously learned LQFs from pristine images. We show that this quality measure correlates reasonably well with difference mean opinion scores (DMOS) on the LIVE IQA database [1].

II. Proposed Approach

A. Probabilistic Latent Semantic Analysis

We first briefly review the pLSA model of Hofmann [9]. Let us suppose that the corpus is a collection of N documents, which in our case are the pristine and distorted images. The visual vocabulary comprises of W visual words with the i^{th} word denoted by w_i. The j^{th} image I_j is assumed to comprise of W_j words with the i^{th} word denoted by w_{ij}. We further assume that there are K LQFs that pervade the images in the corpus, with the k^{th} factor denoted by the indicator variable z_k. In other words, 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 joint probability P(w_{ij},I_j,z_k) is best illustrated in Fig. 1.

The conditional probability of observing a word w_{ij} given an image I_j is is obtained by marginalizing over the latent factors z_k i.e. $P(w_{ij}|I_j) = \sum_{k=1}^{K} P(z_k|I_j)P(w_{ij}|z_k)$. The LQFs

The authors are with University of Texas at Austin, Austin, Texas, USA. E-mail: mittal.anish@gmail.com

[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 [2].

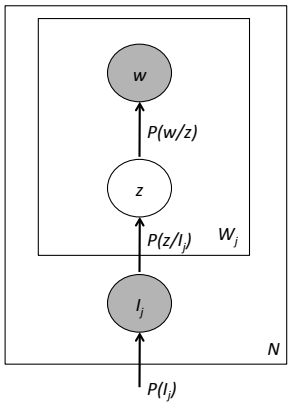

Fig. 1. Graphical representation of the pLSA model.

that pervade the collection of images, and their probabilities 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]. The model-fitting procedure yields the image specific topic distributions $P(z|I)$, and the topic specific word distributions $P(w|z)$. 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 Referenceless Image Spatial QUality Evaluator Engine (RISQUEE) [3] to compute features over every image patch. The principle behind RISQUEE 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 [3], [13], [4], [5], [7], [8]. As shown in [3], [13], [4], [5], [7], [8], such characterization is sufficient not only to quantify naturalness, but also to identify the distortions the images are afflicted with. The RISQUEE 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 RISQUEE features represent statistics of normalized luminance coefficients of images [3]. The RISQUEE features also utilize a model for pair-wise products of neighboring (normalized) luminance values. The RISQUEE 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 [3] 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 RISQUEE 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.

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 mixing coefficients of the latent quality factors $P(z_k|I_{new})$ 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 topic mixing coefficients $P(z|I_{new})$ 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. The mixing coefficients $P(z|I_{new})$ are again estimated by running EM, but this time only the mixing coefficients are updated, while $P(w|z)$ estimated during the model fitting procedure is held fixed.

The vector of estimated topic mixing coefficients of the new image I_{new} (i.e. the estimated $P(z|I_{new})$) is now compared to the vector of the estimated topic mixing coefficients of each pristine image in the existing collection. The topic mixing coefficients 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 n^{th} 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 topic mixing coefficients 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 [1], which contains 29 reference images and 5 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 [1], where in, in each run of the experiment, we randomly select 6 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 2 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, we experimented by fixing the number of factors (topics) to 3 and 6. For image quality inference, we fixed the number of LQFs to 3.

A. Analysis of Latent Quality Factors

We analyzed the LQFs that were learned from the pristine and distorted image set. Fig. 2 illustrates examples of image patches assigned to each discovered LQF when the number of latent factors was fixed to 3. As can be seen from the figure, the image patches that are representative of each LQF are clearly different. For example, one set of patches 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 Gaussian blur or JP2K, while another set of patches 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 all assigned to one quality factor. 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. 3. Also, pristine patches begin to separate out into different topics.

B. Image Quality Inference

Table I and Table II lists the median values 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 comparison, we also show the SROCC and LCC for the peak signal to noise ratio (PSNR) metric, which is a full reference IQA metric . The results in Table I and Table 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 this is a fully unsupervised approach and there is no training using DMOS scores.

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 is competely unsupervised, and 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 [1]. Our future work will be focused on gaining a better understanding of the interplay between the number of topics and inferred image quality and experimenting with a more sophisticated topic model such as Latent Dirichlet Allocation [15].

References

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

[2] A. Bovik, "Meditations on video quality," *IEEE Multimedia Communications E-Letter*, no. 4, pp. 4–10, 2009.

[3] A. Mittal, A. Moorthy, and A. Bovik, "Referenceless image spatial quality evaluation engine," 2011, to be published.

[4] A. Moorthy and A. Bovik, "Blind image quality assessment: From scene statistics to perceptual quality," 2011, to be published.

[5] M. Saad and A. Bovik, "DCT statistics model-based blind image quality assessment," in *International conference of Image Processing*, 2011.

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

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

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

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

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

[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] A. Moorthy and A. Bovik, "Statistics of natural image distortions," in *IEEE International Conference on Acoustics Speech and Signal Processing*. IEEE, 2010, pp. 962–965.

[13] Z. Wang, "Applications of objective image quality assessment methods," *IEEE Signal Processing Magazine*, vol. 28, 2011.

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

	JP2k	JPEG	WN	Blur	FF	All
PSNR	0.894	0.919	0.953	0.835	0.896	0.887
Proposed Approach	*0.816*	*0.879*	*0.899*	*0.861*	*0.768*	*0.823*

TABLE I
MEDIAN SPEARMAN'S RANK ORDERED CORRELATION COEFFICIENT (SROCC) ACROSS 1000 TRAIN-TEST EXPERIMENTS ON THE LIVE IQA DATABASE.

	JP2k	JPEG	WN	Blur	FF	All
PSNR	0.904	0.938	0.898	0.849	0.912	0.887
Proposed Approach	*0.863*	*0.901*	*0.878*	*0.787*	*0.812*	*0.800*

TABLE II
MEDIAN LINEAR CORRELATION COEFFICIENT (LCC) ACROSS 1000 TRAIN-TEST EXPERIMENTS ON THE LIVE IQA DATABASE.

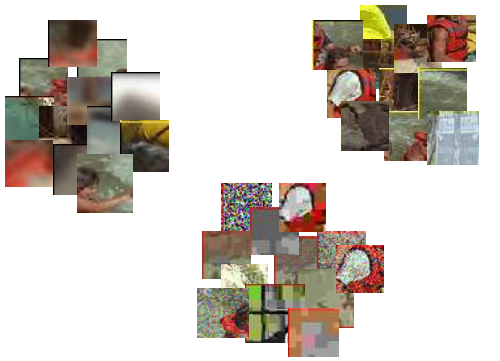

Fig. 2. Examples of image patches assigned to three LQFs discovered by the pLSA model.

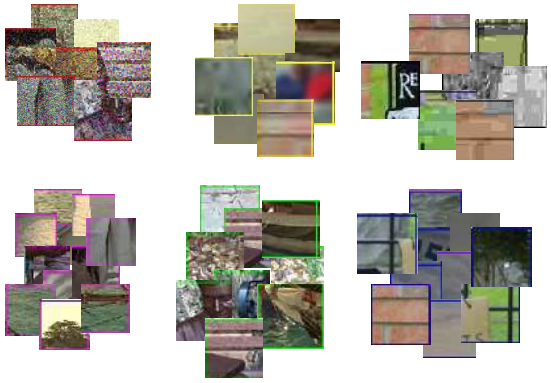

Fig. 3. Examples of image patches assigned to six LQFs discovered by the pLSA model.

[15] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.