A TRAINING-BASED NO-REFERENCE IMAGE QUALITY ASSESSMENT ALGORITHM

Huitao Luo

Hewlett-Packard Labs
1501 Page Mill Road, MS 1203, Palo Alto, CA 94304

ABSTRACT

We present a new image quality assessment algorithm that does not rely on reference images. Our general framework is to emulate human quality assessment by first detecting visual components, then assessing quality against an empirical model. We describe an instance of this framework where visual component detection is realized as a face detection method, and quality modeling is realized using radial basis function (RBF) networks. Experiments with this prototype system yielded promising results.

1. INTRODUCTION

Image quality assessment is an area of intense contemporary research interest, but a majority of algorithms described in the literature require the use of an original image as a reference. Although this is useful in applications such as compression, there are other applications where quality assessment is desirable, but a reference image is not available. For example, a *no-reference,* or *blind,* quality assessment algorithm would be of use in digital photography to inform the user that a low- or high-quality photo had been taken; in image management applications, to sort out good from poor photos; and in printing, to encourage (or discourage) the printing of better (or poorer) pictures. It is always easy for a human to judge photo quality without the use of reference photos; but it is far more difficult to design an algorithm that would do the same

Because our understanding of visual perception is still quite limited, not many research works have been reported in the literature. Wang *et al.* and Sheikh *et al.* used image distortion models to estimate degradation in JPEG [1] and JPEG-2000 [2] compressed images, respectively. Nill and Bouzas [3] designed a metric making use of Modulation Transfer Function (MTF) analysis and incorporating features of the human visual system (HVS); but their method is mainly intended for and tested on reconnaissance satellite photos. Chen and Meng [4] used statistics of block region activity scores, with the activity score of a block region being a function of its grayscale standard deviation. Their algorithm is biased, since it tends to give higher scores to images with more high frequency content. Li [5] discussed several objective measures for no-reference quality assessment: edge sharpness, random noise, and structured noise, the last one being relevant to [1, 2]. Several recent papers also discussed quality assessment with known distortion models: Turaga *et al.* [6] on JPEG and Ong *et al.* [7] on JPEG-2000.

We propose a no-reference algorithm for image quality assessment based on object/region detection. This approach is mainly based on the following two observations. First, except for works in noise estimation with known distortion models such as [1, 2, 5], most quality assessment measures [3, 4, 5] are based on heuristics, and only make sense when the domain is limited, or the image content is known. Without assumption about image content, it is difficult to compare the quality measures of two images because different content may contribute differently to quality assessment measures. Second, we believe quality assessment is a process related to image understanding. A human is able to tell if an image is blurred without any reference because he or she understands the content of the image and can judge its visual appearance based on prior experience.In other words, human beings are able to assess image quality because they have acquired empirical appearance model of various visual components.

Based on these observations, we propose a two-step algorithm. First, an object/region detection algorithm is applied to detect certain types of object/region from the image. Second, the spectrum distribution of a detected region is compared with an empirical model to determine a quality score for the detected region. The image quality is then determined based on the scores of the detected region(s). To make this two-step design possible, a basic requirement is that the signal features used to detect an object do not overlap with signal features used in quality assessment. This is intuitively true: we normally are able to identify human faces in an image, independent of their qualities.

We develop an implementation of the proposed algorithm using face detection. Face detection is selected because a human face is normally a region of interest, and the quality of a face region has significant influence on overall image quality. In addition, face detection is a well studied area in object detection, and we have access to a number of

0-7803-8554-3/04/$20.00 ©2004 IEEE.

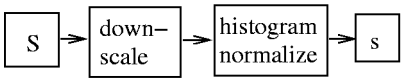

Fig. 1. Preprocessing pipeline for face detection.

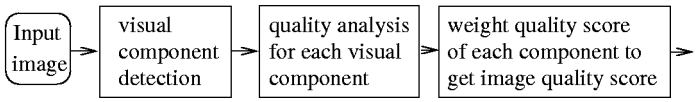

Fig. 2. Quality assessment scheme based on visual component detection.

good face detection algorithms [8, 9].

2. RELATION BETWEEN FACE DETECTION AND QUALITY ASSESSMENT

Intuitively, when a person views an image, his or her ability to find familiar objects in the image is independent of their judgment of the image quality. Of course, this assumes the image is not so severely degraded that its content is no longer recognizable. In this section we show that this observation is valid analytically for corresponding computer algorithms.

In our work, two state-of-the-art face detection algorithms [8, 9] are studied. In the algorithm due to Rowley et al. [8], a neural network is trained to detect face patterns in a region of 20-by-20 pixels. To determine whether an arbitrary image square S is a face region, this square S is downsampled to the size of 20 by 20 and equalized, resulting in a normalized signal s (see Fig 1). The normalized region s is the input to the neural network. In the converting process from S to s, many signal features are lost and thus not used in face detection decision. For example, the downsampling filters out the high frequency components, and the histogram normalization removes considerable brightness and contrast information. A similar proposition holds for the face detector of Viola and Jones [9].In their algorithm, a feature based classifier is trained over a square of 24-by-24 pixels, with each feature defined by a feature template composed of a group of rectangular sub-windows. To detect faces of bigger sizes, V-J algorithm scales up the feature templates to the resolution of a candidate face region, computes features and apply the same classifier. Because each feature is obtained by normalizing each sub-window and computing region means, V-J algorithm is similar to Rowley algorithm in using partial signals information to detect face regions.

On the other hand, many image quality problems are determined by the signal features ignored by the face detectors. For example, blurriness commonly is a problem when there is not enough high frequency components. Noisy is a problem when various high frequency noise signals are added to the original signal. Under/over exposure is a problem when the signal is too dark/bright. This suggests that in our context of quality assessment based on face detection, the features used in object detection are mutually exclusive of features highly relevant to quality assessment. In fact, we applied face detectors [8, 9] on many low quality images containing human faces, and each exhibited a high detection rate. From another perspective, ignoring quality features is just one of the design requirements for the face detection algorithms. A good face detection algorithm should be able to detect faces under various quality conditions. Here although we discuss only face detection algorithms, it will not be difficult to generalize it other object detection problems.

3. QUALITY ASSESSMENT BASED ON VISUAL COMPONENT DETECTION

Based on the analysis in Section 2, we propose an image quality assessment algorithm based on visual component detection. A flowchart of the proposed system appears in Fig. 2. First, the input image is processed by a visual component detection module to identify relevant objects. Next, each detected visual component is processed by a quality assessment module, resulting in a numeric quality score for each detected visual component. Finally, the individual component scores are combined in a weighted sum, resulting in an overall quality score. In this general flowchart, the modules for both visual component detection and quality analysis are both knowledge based, and normally can be solved by machine learning algorithms.

Although the proposed system is suitable for any visual components, in this paper, we use face as a special case to prove feasibility of the proposed algorithm. Face detection is a well studied area, and we will not go into any further details about how it can be accomplished. Our emphasis is on the quality analysis module, which we design using a machine learning method. This is directly motivated by the belief that the quality assessment by humans is a knowledge-based process.

3.1. Learning Based Quality Assessment System

We illustrate our quality assessment design in Fig. 3, and the corresponding model training design in Fig. 4. In both designs, an identical feature computation module is used to extract a group of signal features that a relevant to quality assessment. These features form a high dimension feature vector. In the training stage, a human expert is involved to give each training image (a face image in our case) a quantitative quality score, a machine learning engine is used to find the functional relationship between a training image's feature vector and its labeled score value. When the training

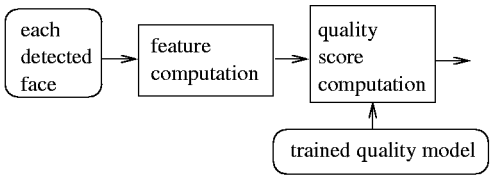

Fig. 3. Quality assessment based on a trained model

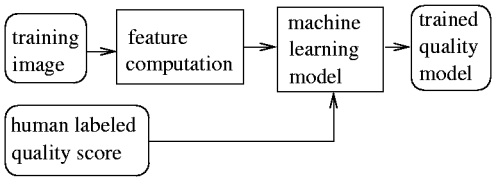

Fig. 4. Knowledge model training

process is finished, this functional relationship is stored as a trained knowledge model. This trained model data is used in the quality assessment system (shown in Fig. 3) to map an input feature vector to a quantitative quality score.

3.2. Signal Feature Design

Given a detected face region, the purpose of a feature computation module is to extract signal features that are relevant to human quality assessment. We solve this problem primarily by using heuristic methods since there is still no good theoretical explanation on human visual system (HVS).

The signal features are designed based on two criteria. The first is the *correlation* criterion. For each candidate feature, we compute its statistical correlation with human quality assessment over a training image database, and only those features exhibiting high correlation values are selected. In addition, this criterion regulates that the correlation between any two selected features should not be high, to avoid redundancy and bias towards certain signal features. The second is *completeness* criterion, which requires that the final selected feature set be *relevant* to most of the known image quality problems. In other words, for any image with a known problem or problems, at least one selected feature is relevant to its quality assessment and is able to serves as the quality indicator. In this work, we are concerned with the following quality problems in consumer photography: blurriness, including bad focus and motion blur; under and over exposure; noisiness, specifically white noise, Gaussian noise, and salt and pepper noise.

Our current feature set contains 10 features, including 1 brightness component (DC component), 7 spectrum components obtained from wavelet decomposition, and 2 noise components. All are extracted from the grayscale channel. The spectrum analysis is designed using a two-level, decimated Wavelet decomposition, which decomposes the detected face region into 7 sub-bands. The power distribution of the signal within each sub-band is computed to generate 7 spectrum components. In addition to spectrum analysis, the two noise components are designed to separate high frequencies noise from high frequency signal. Among them, one (N_1) is computed via direct noise estimation and the other (N_2) is computed by analyzing the spatial distribution of the high frequency signal. The detailed designs are described as follows.

Direct noise estimation: Suppose the detected face region is a $d \times d$ square, a $t \times t$ square template ($t = d/5$) is used to search over the detected face region. At each location, the grayscale standard deviation within the template is computed. We denote the minimum deviation over all locations as the Dev_{min}. The square that generates Dev_{min} is filtered using a Gaussian low-pass filter, and the grayscale standard deviation of the filtered image square is computed as Dev_f. The magnitude of the noise is estimated as $N_1 = \sqrt{(\text{Dev}_{\min} + m)/(\text{Dev}_f + m)}$, where m is a tiny constant to avoid singular cases.

Spatial distribution: This feature is intended to estimate spatial homogeneity of the high frequencies signal. The rationale here is that although both noisy images and high fidelity images have high frequency signal power, the spatial distributions of them tend to be different. In the case of a face region, high frequency noise signal tend to be distributed uniformly, while high frequency image feature signals tend to concentrate around image features (i.e. eyes, mouth, etc). To measure the spatial homogeneity, each high frequency sub-band signal is first thresholded to obtain a binary image. A $t \times t$ square template is used to search over the binary image, and at each location (i, j), the standard deviation within the template is computed and represented as $\text{Dev}(i, j, s)$. The homogeneity of the whole binary image is then defined as $N_2 = \sum_S \sum_J \sum_I \text{Dev}(i, j, s)$, where I, J, S represent the range of variables i, j, s respectively.

3.3. Machine Learning Algorithms

Two algorithms are used in our system: *mixture of Gaussian* and *Redial Basis Function* (RBF) [10]. Mixture of Gaussian is a statistical distribution estimation algorithm, while RBF is a function approximation algorithm. In the RBF implementation, a function F is defined to map the feature vector $V = \{w_0, ..., w_9\}$ of an object region to its quality assessment value g as $g = F(V)$. A RBF network is modeled to approximate this unknown mapping function F. In mixture of Gaussian implementation, a feature vector is defined as the combination of the input and output of the previous mapping function F: $V' = \{g, w_0, ..., w_9\}$. The distribution of this feature vector is modeled using a

mixture of Gaussian model. Suppose this distribution function is G, then the quality assessment is computed as $g' = \arg\max_g(G(g, w_0, ..., w_9))$. Note that g is selected from a set of limited integer numbers.

4. EXPERIMENTS

Our experimental system is built into two parts: a training system and a testing system, both using CMU face detector [8]. For training purpose, 850 images containing human faces are collected, and human experts are invited to assess image quality of each face into five levels (1 to 5, the bigger, the better). One caveat in training design, however, is that the feature set design discussed in Section 3.2 is dependent on the face sizes. Theoretically, a dedicate model should be trained for each different face size. In practice, we quantized the face sizes into 5 discrete levels and trained a model for each of them. Therefore in the training stage, each detected human face region is scaled to 5 different sizes, and a human expert is involved to give each of them a numerical quality score. This process is very laborious. So far, our experimental data is only based on quality labeling data from one human expert.

Like most machine learning systems evaluation, we use both self validation and cross validation tests in the experiment. In a self validation test, all of the 850 human labeled images are used in both training and testing, while in a cross validation, 2/3 of the images are used for training and the rest 1/3 of the images are used for testing. In both tests, the RBF model generates better performance than the mixture of Gaussian model. Therefore we discuss only the RBF performance as follows.

In one experiment, we compared the machine generated image quality score ms with the corresponding human labeled quality score hs. In particular, we computed the difference between the two scores $|\text{ms} - \text{hs}|$ and analyzed its statistics over the testing images. In Table 1, the distribution of this difference is shown in the left side while the right side is its mean value. Note ms generated from RBF is a real number while hs is an integer on the scale of 1 to 5. In calculating the distribution statistics, the difference is rounded to the closest integer. The distribution in Table 1 is shown in cumulative statistics. For example, in the cross validation row, "97%" in column "1" means that on 97% of the testing images, the difference between the RBF machine generated quality score and the human labeled quality score is less than or equal to 1. Obviously, even for cross validation, the trained machine generates quality predications very close to these from the human expert.

In another experiment, we analyzed the relative consistency between machine quality assessment and human expert assessment. Given arbitrarily two testing images, suppose their machine generated quality scores are ms_1 and ms_2, and their human labeled scores are hs_1 and hs_2. Machine assessment scores are considered consistent with human judgment if $(\text{ms}_1 - \text{ms}_2) * (\text{hs}_1 - \text{hs}_2) \geq 0$. In a cross validation setup, 1000 pairs of images are randomly picked and 82.4% of the times machine predictions are consistent with human experts.

	Distribution of difference			mean of difference
	0	1	2	
self validation	67.7%	99%	100%	0.518
cross validation	55%	97%	100%	0.647

Table 1. Quantitative system performance.

5. CONCLUSION

We present a new approach for image quality assessment based on visual component detection and supervised machine learning. Our preliminary experiments based only on face detection and face based quality modeling yielded encouraging results. However, more work needs to be done in areas such as object detection, feature selection and machine learning to better establish this method.

6. REFERENCES

[1] Z. Wang et al., "No-reference perceptual quality assessment of JPEG compressed images," in *ICIP'02*.

[2] H. Sheikh et al., "Blind quality assessment for JPEG2000 compressed images," in *Asilomar Conf'02*.

[3] N. B. Nill and B. H. Bouzas, "Objective image quality measure derived from digital image power spectra," *Optical Engineering*, vol. 31, pp. 813–825, 1992.

[4] Y. Chen and F. Meng, "Image quality measurement based on statistics of activity region," *J. of Chinese Institute. Eng.*, vol. 24, pp. 379–388, 2001.

[5] X. Li, "Blind image quality assessment," in *ICIP'02*.

[6] D. Turaga et al., "No reference PSNR estimation for compressed pictures," *Signal Processing: Image Communication*, vol. 19, 2004.

[7] E. Ong et al., "No-reference JPEG-2000 image quality metric," in *ICME*, 2003.

[8] H. Rowley et al., "Neural network-based face detection," *IEEE PAMI*, vol. 20, pp. 22–38, 1998.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR'01*.

[10] I. Nabney, *Netlab: Algorithms for pattern recognition*, Springer Verlag, 2001.